Oct 18 2025

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Sargantana: An Open RISC-V Processor for HW/SW Co-Design of Domain-Specific Accelerators

Narcís Rodas, Max Doblas, Víctor Soria-Pardos, Santiago Marco-Sola, Miquel Moretó

BZL

Financiado por la Unión Europea
NextGenerationEU

MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

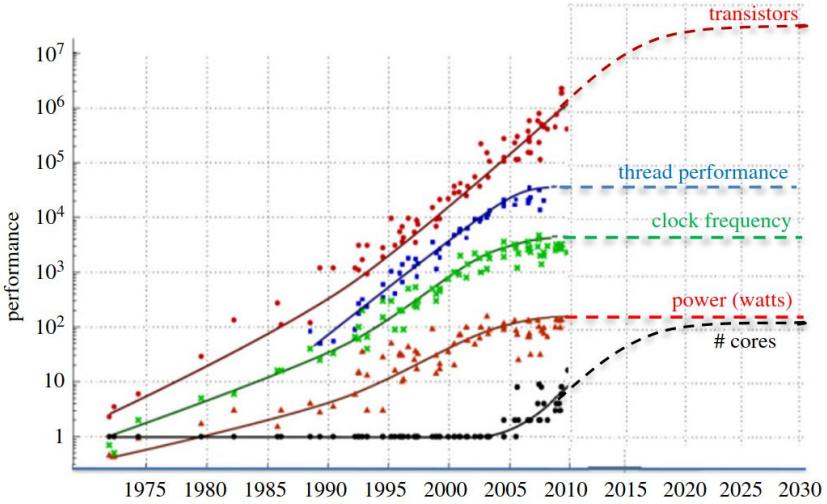
Plan de Recuperación, Transformación y Resiliencia

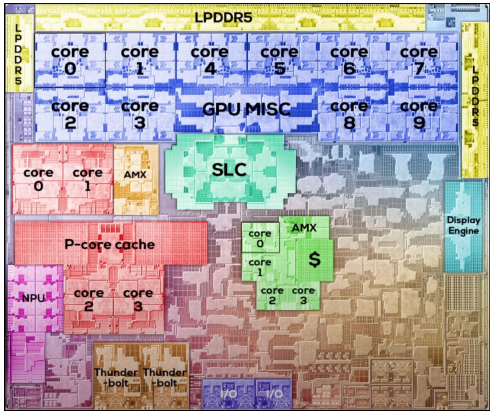
Index

Index

1.1. Context

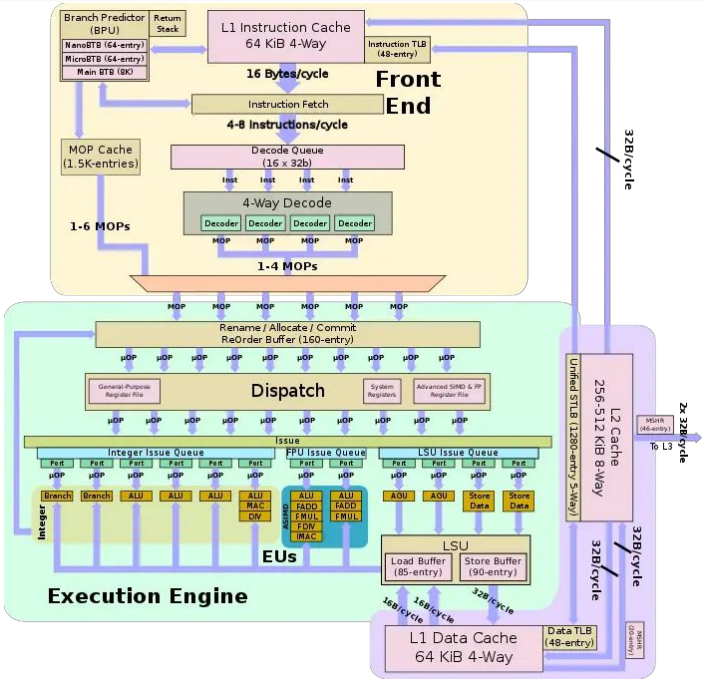
- **End of Moore's Law**. Transistor scaling no longer guarantees performance or efficiency gains.
- **Breakdown of Dennard scaling**. Power density and thermal limits constrain frequency growth.
- **Performance growth** now depends on architecture design more than ever.
- **Shift toward heterogeneity**. CPUs, GPUs, and domain-specific accelerators (DSAs) integrated in the same SoC.
- **DSAs are key to modern HPC** and AI systems, providing order-of-magnitude speedups and energy efficiency.





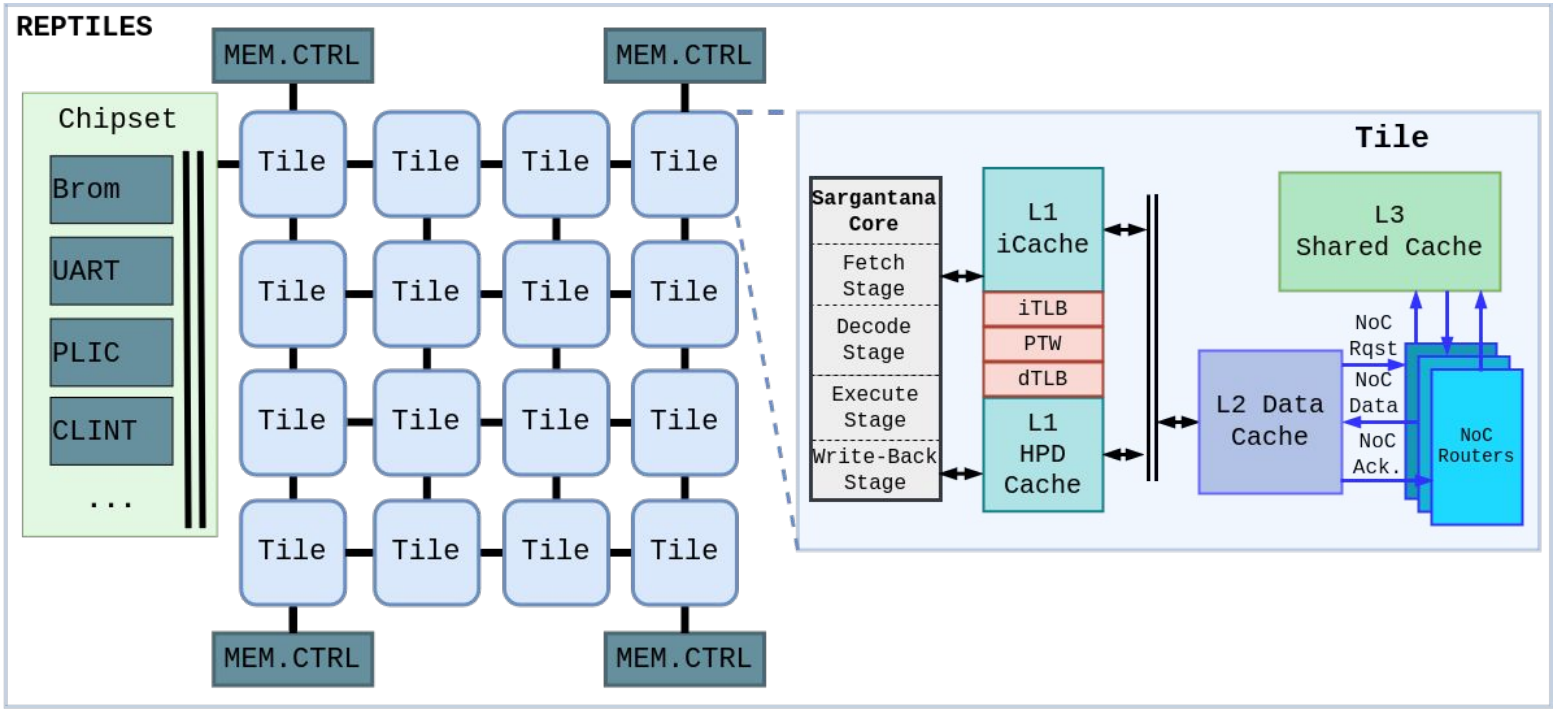
1.2. Motivation & Gap

- **Modern RTL Processors.**
  - **Difficult to integrate** new extensions or accelerators.
  - **RTL simulations are accurate** but continue to grow in complexity.
  - Mostly **closed designs** with limited access and flexibility.
- **Cycle-Level Simulators** (e.g., gem5).
  - **Simpler and easier** to modify than real hardware.
  - Produce **less realistic results** compared to actual systems.
  - **Lack physical design analysis** and hardware-level validation.



1.3 Our Goal

Our **goal** is to design and implement Sargantana, an **open-source RISC-V processor** that provides a **modular and Linux-capable** environment for the development, prototyping, and **evaluation of domain-specific accelerators**.



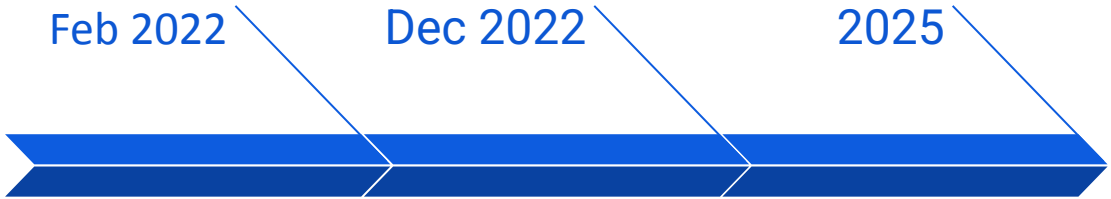
1.4 Contributions

1. We present **Sargantana full-fledged edge-class RISC-V core** designed as a research platform for DSA.
   - Linux capable
   - Simple and modular
2. We demonstrate **easy integration of custom ISA extensions**.
   - ***Mix-GEMM*** for Deep Neural Network (DNN) inference
   - ***GMX*** for efficient sequence alignment
3. We demonstrate **easy integration of loosely-coupled accelerators.**
   - ***WFAsic*** for genome analysis
4. **Open Source Sargantana code** in the GitHub repository.
   - https://github.com/bsc-loca/sargantana

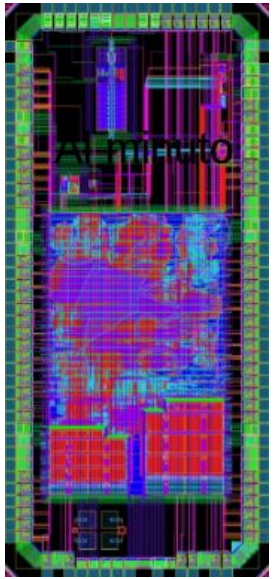
Index

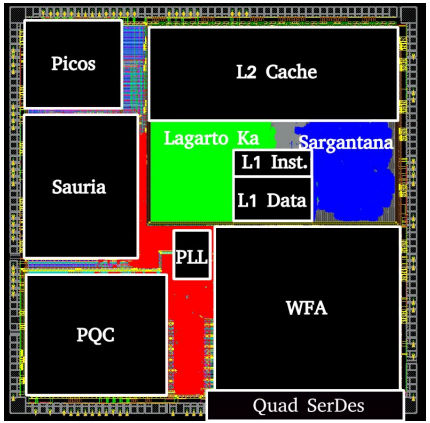
2. Sargantana Architecture Overview

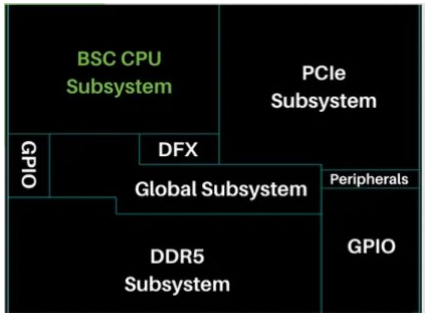
- **RV64GBV (IMAFDV)** User-Level ISA v2.3 and Privileged ISA v1.11.
- **Register renaming** with 64 physical registers.
- **Non-blocking** memory pipeline with up to 16 misses on flight.
  - Connected to the OpenHW CV-HPDcache.
- Can run up to **1.5 GHz** in a 7nm technology node.



Sargantana	Kameleon	Cinco Ranch
Custom ISA Extensions	LKa + Sargantana + Accelerators	LOx + LKa VPU + Sargantana
1.2 GHz, 2.9mm2 in GF 22nm	1.2 GHz, 9mm2 in GF 22nm	PLL 0.6-2.8GHz 16mm2 (3.2mm2) in Intel3

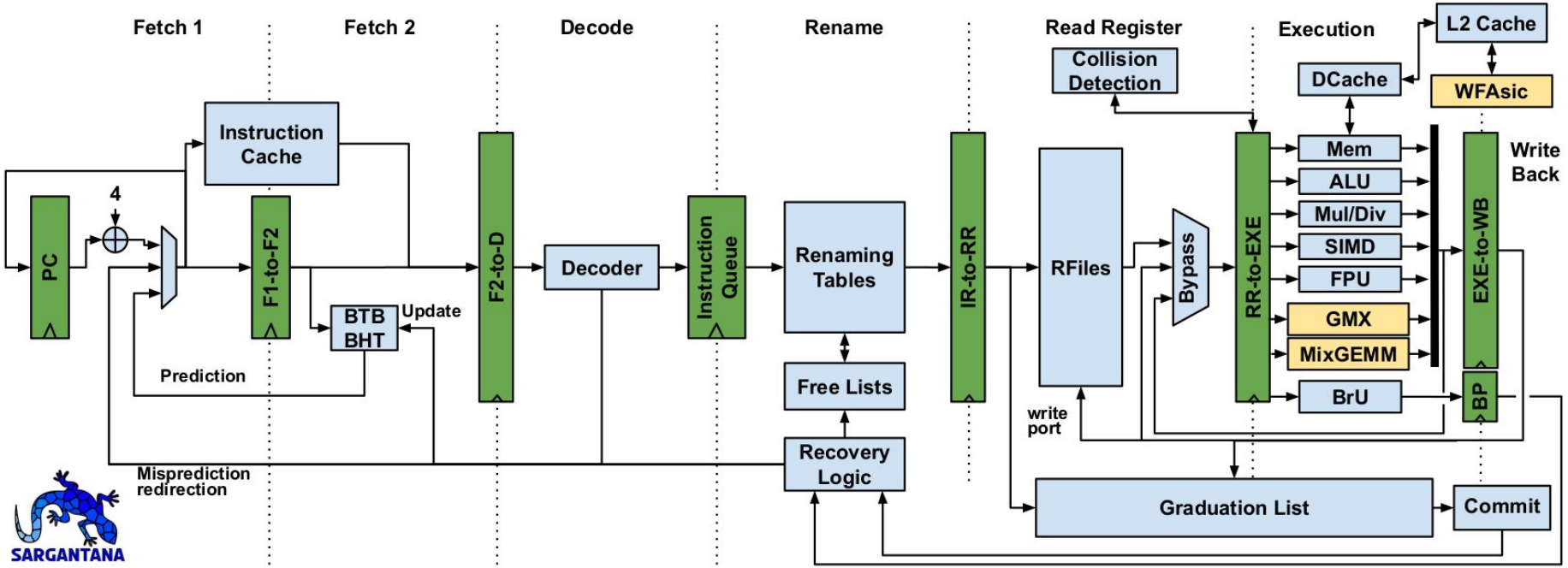






2.1 Sargantana Pipeline

- 7-stage single-issue in-order core with out-of-order writeback



2.2. Integration Features for Domain-Specific Accelerators

- **ISA extensions:**
  - **Register renaming** and **out-of-order writeback** simplify the addition of new functional units and ISA extensions.
  - **Scoreboard logic** manages instruction dependencies efficiently, enabling easy accelerator integration.
  - These mechanisms **reduce data dependencies** and allow multi-cycle instructions without stalling the pipeline.
- **Loosely coupled accelerators:**
  - Connection to higher cache levels through AXI interfaces and use **memory-mapped regions** for communication and control.
  - The **modular pipeline** eases experimentation with new execution paths and co-processors.

Index

3.1. Custom ISA Extensions

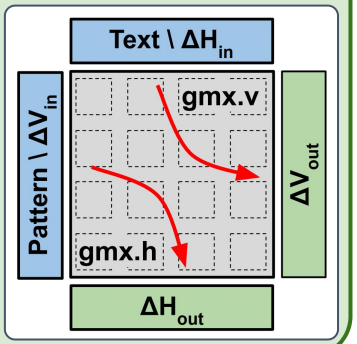
- Custom ISA Extensions are an area-efficient way to add functionality.
- Usually require less software modifications to run compared to loosely-coupled DSAs.

GMX

(MICRO 2023)

BioInformatics. A set of ISA extensions that enable efficient sequence alignment computations based on dynamic programming

- **>42x** long read alignment acceleration on CPU
- Competitive performance per area w.r.t. DSA
- **0.022 mm²** in 22nm FD-SOI
- Around two integer multipliers

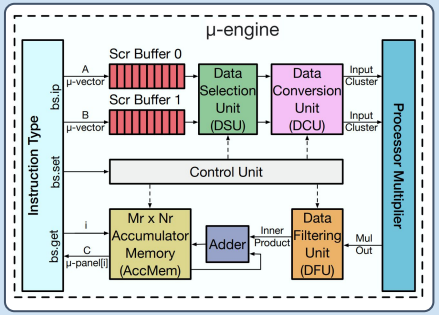


Mix-GEMM

(HPCA 2023)

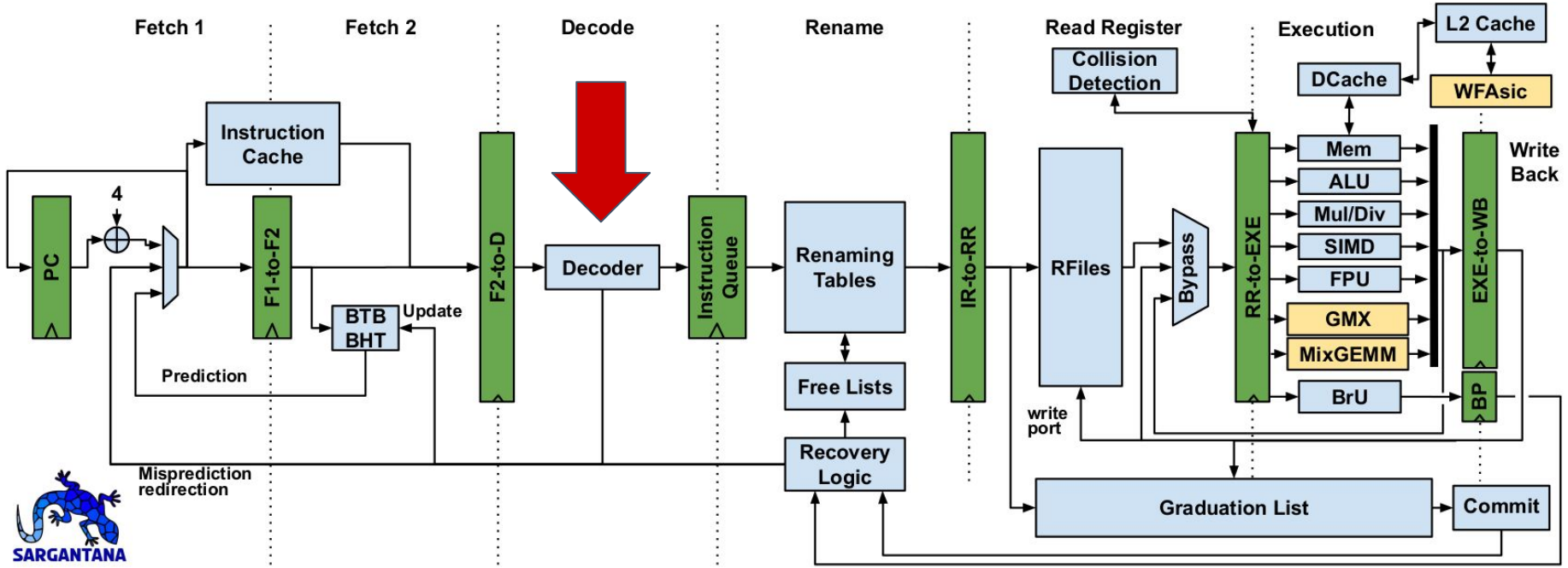
A hardware-software codesigned architecture capable of efficiently computing quantized DNN convolutional kernels based on byte and sub-byte data sizes.

- up to 13.6 GOPS
- up to 1.3 TOPS/W
- 1% of SoC area
- 2.3% of SoC power



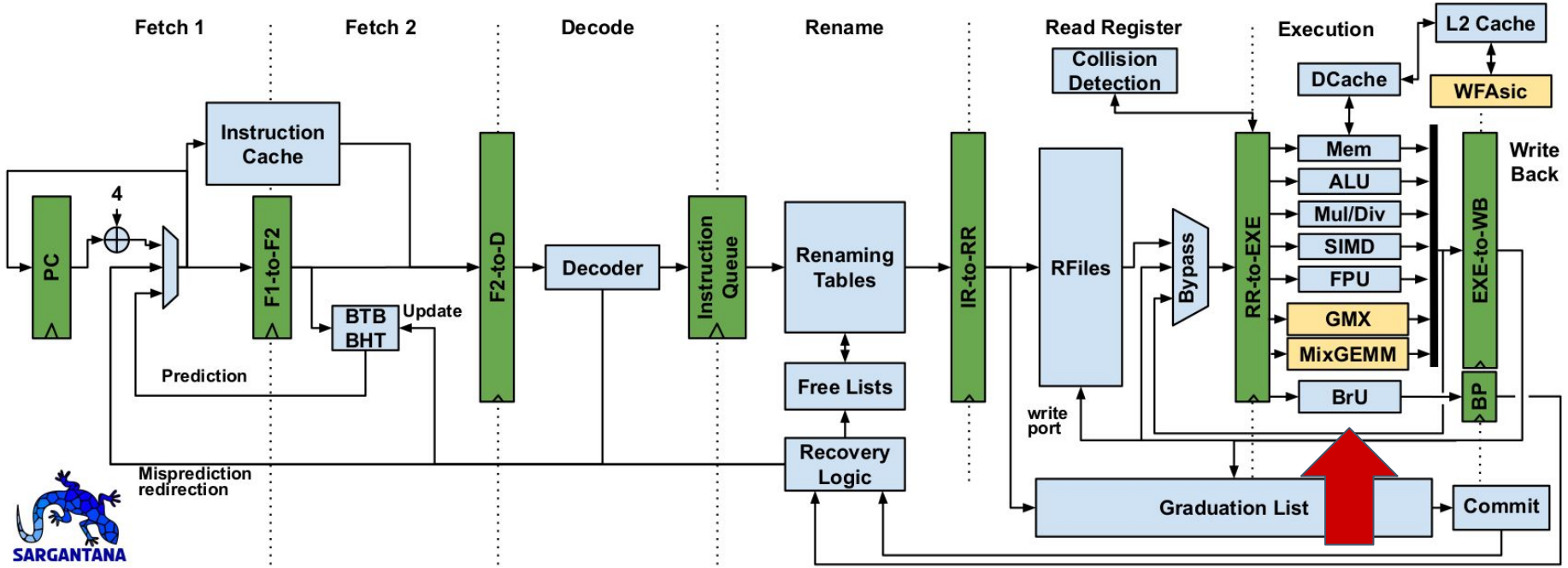
3.1. Custom ISA Extensions (HOW TO INTEGRATE)

① Add new instruction encoding in the decoder and set the instruction fields.
② Add new functional unit or logic of the instruction in the execution stage.
③ If latency > 1 cycle, add the latency in the scoreboard to avoid collisions.



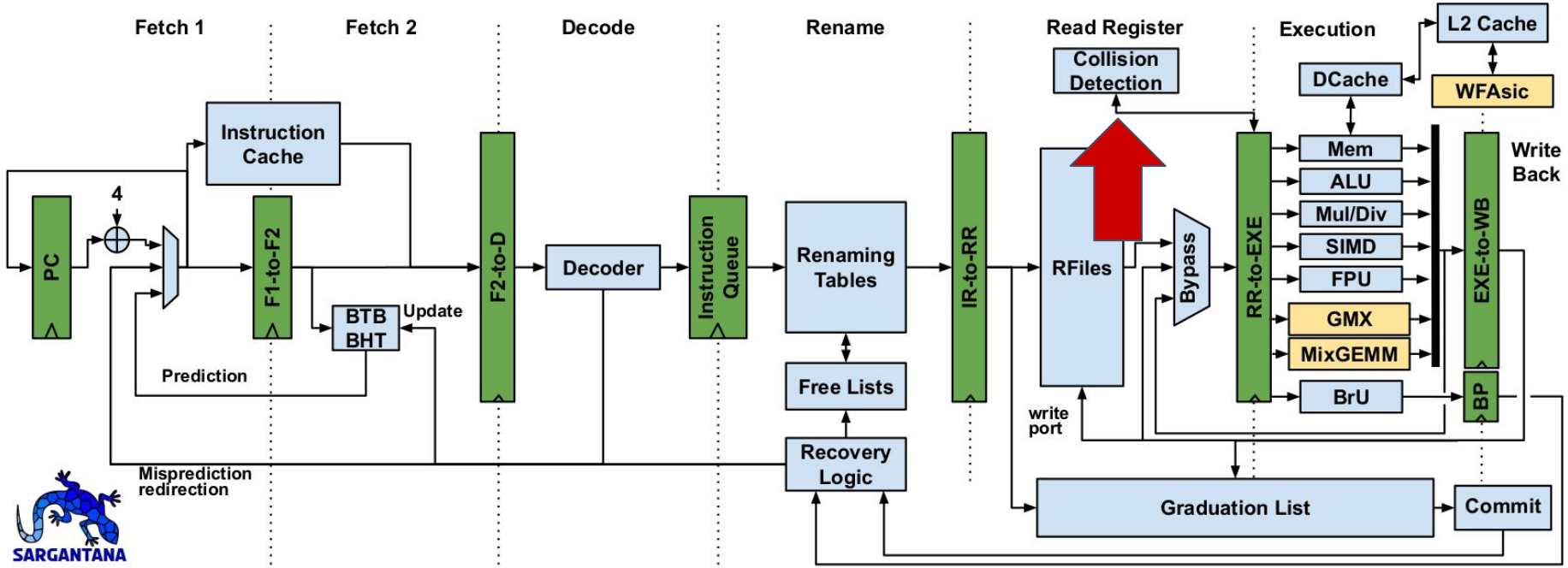
3.1. Custom ISA Extensions (HOW TO INTEGRATE)

① Add new instruction encoding in the decoder and set the instruction fields.
② Add new functional unit or logic of the instruction in the execution stage.
③ If latency > 1 cycle, add the latency in the scoreboard to avoid collisions.



3.1. Custom ISA Extensions (HOW TO INTEGRATE)

① Add new instruction encoding in the decoder and set the instruction fields.
② Add new functional unit or logic of the instruction in the execution stage.
③ If latency > 1 cycle, add the latency in the scoreboard to avoid collisions.



3.2. Loosely-coupled accelerators

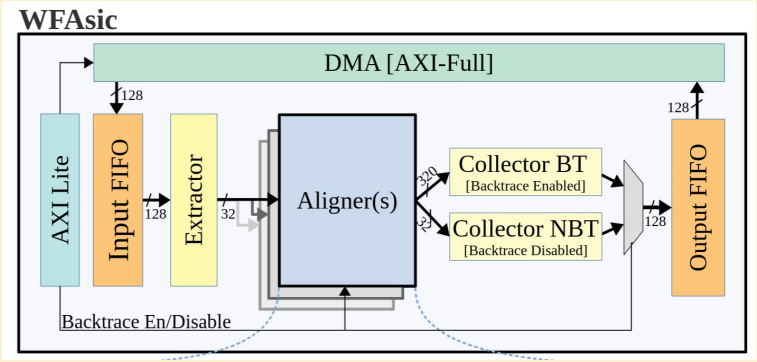
- Loosely coupled accelerators provide **higher flexibility** and enable tailored control logic for **improved performance**.

WFAsic

(ICPP 2023)

BioInformatics. Accelerator for exact pairwise alignment based on the WFA algorithm. Implements the extend and compute operators of WFA in hardware.

- **Up to 1076x** long-read alignment acceleration on CPU
- Higher energy efficiency than GPU or FPGA solutions
- **1.6 mm²** in GF 22nm

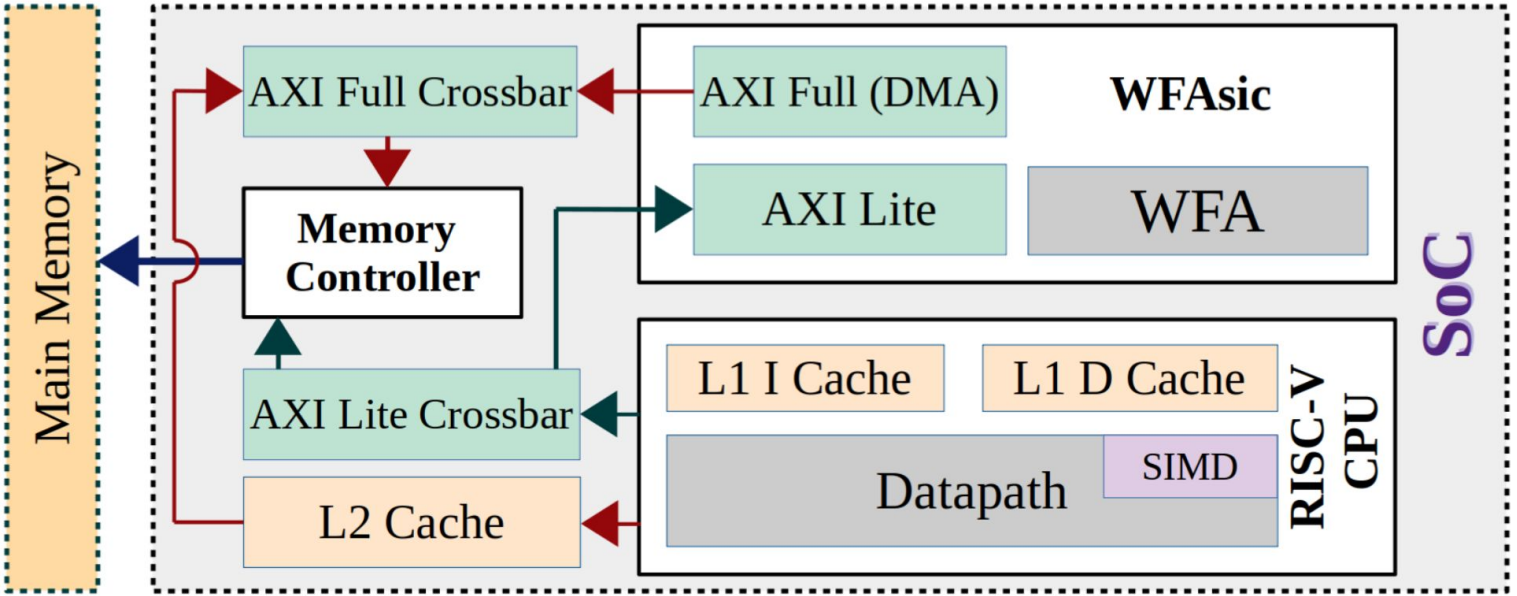


3.2. Loosely-coupled accelerators (HOW TO INTEGRATE)

Steps to connect a loosely-coupled accelerator

① Modify the memory map of the core to add a region for your accelerator

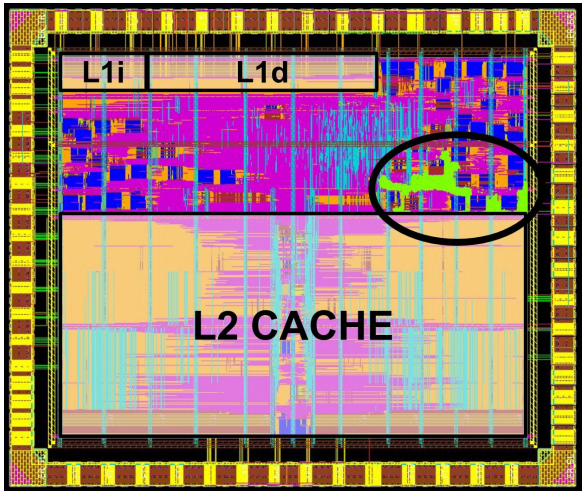
② Connect the accelerator via AXI interface



3.3. Physical design

Implementation of the Sargantana Core in 22nm GlobalFoundries targeting 1 GHz.

- Sargantana infrastructure allows an easy and viable way to perform physical synthesis.
- We show area results of ISA extensions and loosely-coupled accelerators.
  - ISA extensions are area efficient only represent 0.8% - 2.4% of Sargantana core.
  - Loosely-coupled accelerators occupy a similar area than the Sargantana core.
    - Achieves a higher performance results due to its specialization.



Design	Area (mm^2)	Power (mW)
Sargantana	1.24	397.44
Mix-GEMM 1-issue	0.01	3.60
Mix-GEMM 2-issue	0.03	8.60
GMX	0.02	8.47
WFAsic	1.60	312.00

Cornerstone for state-of-the-art DSA evaluation

- The three accelerators integrated in Sargantana (Mix-GEMM, GMX, and WFAsic) represent state-of-the-art designs in DNN inference, sequence alignment, and genome analysis.
  - All published on top venues with strong results
  - Relevant application areas (e.g., IA, genomics)
  - Proves Sargantana's maturity: from RTL → FPGA → synthesis
  - Simple, modular, Linux-capable research platform
  - Enables rapid DSA prototyping and evaluation

- We believe Sargantana can serve as a cornerstone for future DSA research, enabling agile prototyping, benchmarking, and co-design of next-generation accelerators.

4. Conclusions

- DSAs are the current solution for performance growth in HPC
  - Integrating DSAs in modern processors is hard !!
- The 3 accelerators integrated in Sargantana (Mix-GEMM, GMX, and WFAsic) represent state-of-the-art DSA designs.
  - All published on top venues with strong results (MICRO and HPCA).
  - Relevant application areas (IA, bioinformatics, and genomics)
  - Proves Sargantana's maturity: from RTL → FPGA → synthesis
  - Simple, modular, Linux-capable research platform
  - Enables rapid DSA prototyping and evaluation
- We believe that Sargantana will enable researchers to more easily test, prototype and evaluate their designs in an environment that gives them accurate results in terms of performance and efficiency.
  - https://github.com/bsc-loca/sargantana

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank You!

Cátedra Chip UPC
de arquitecturas avanzadas
y sistemas fotónicos

España | digital 2026
PERTE Chip

BZL
Financiado por la Unión Europea
NextGenerationEU
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia