AI Accelerators Dependability: Assessment and Performance Isolation Strategies

Marcello Cinque

macinque@unina.it

F. Boccola, R. Della Corte, L. De Simone, N. Di Giacomo, F. Pizzo, S. Toscano

DEpendable and Secure Software Engineering and Real-Time Systems (DESSERT) · Federico II

dessertlab.github.io

90th IFIP Meeting – UNC, Charlotte, USA – June 27th, 2026

Motivation

- AI systems are integrated in mission- and safety-critical applications

Autonomous Vehicles

… and biometric recognition

UAVs and Satellites

- Need to **assess AI systems dependability** when used in **critical enviroments** (e.g. ISO 26262, IEC/TR 5469)

ISO/IEC TR 5469 – Functiona Safety and AI Systems

The technical report still does not solve the **gap** between **functional safety standards** and **AI**.

But it tells:

- When AI can sit inside a safety function
- What properties you need to demonstrate
- How you can demonstrate them in the lifecycle of other standards (e.g., ISO 26262)
  - E.g., statistical evaluation, fault injection, fault isolation, …

Our on-going contributions

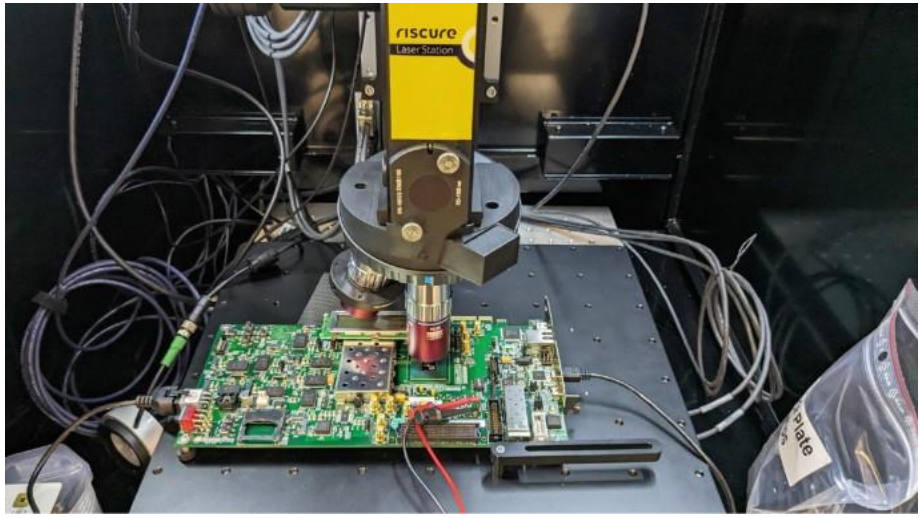
- Fault injection in AI accelerators
- Workload isolation strategies
- AI benchmarking framework

Our on-going contributions

- **Fault injection in AI accelerators**
- Workload isolation strategies
- AI benchmarking framework

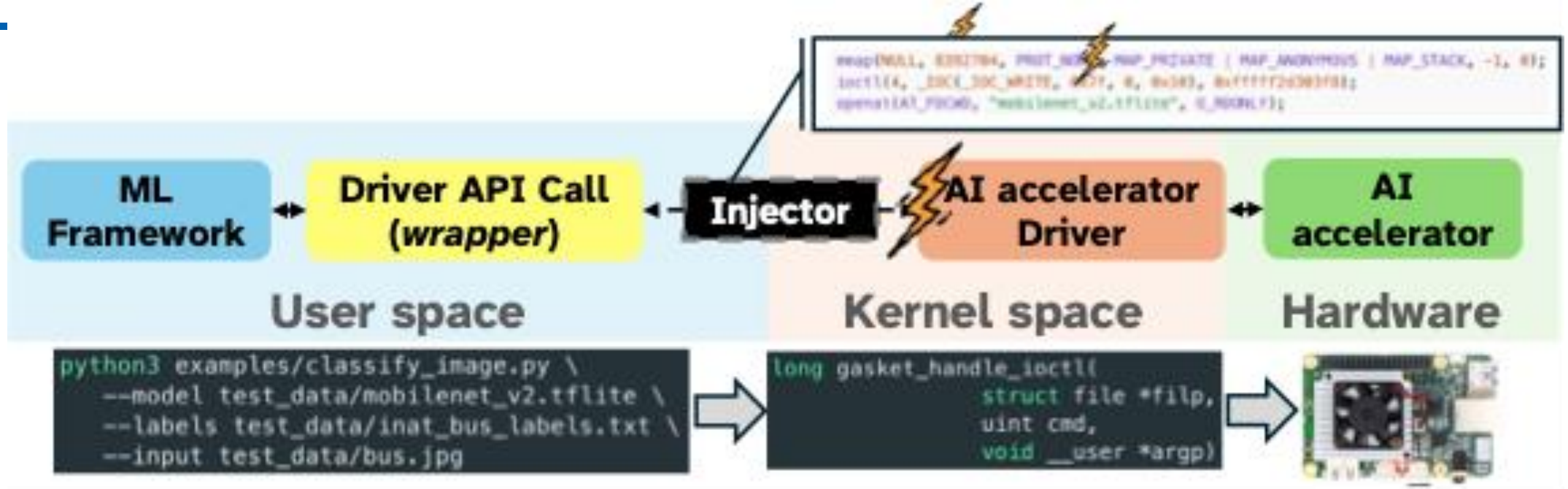
The State of the Art

- Low-level **hardware faults** (RTL/Physical): Accurate but costly/impractical
  - *Examples:* beam/radiation testing [VLSI '24], laser fault injection (LFI), RTL simulators [CLUSTER '24], FPGA emulation [VLSI '24]
- ML framework-level **software perturbations**: Scalable but miss hardware-software interactions
  - *Examples:* TensorFI [DSN '20], PyTorchFI [DSN-W '20] , LLTFI (LLVM-based injectors) [ISSRE '22]



[DSN '26]

Our proposal: mind the driver

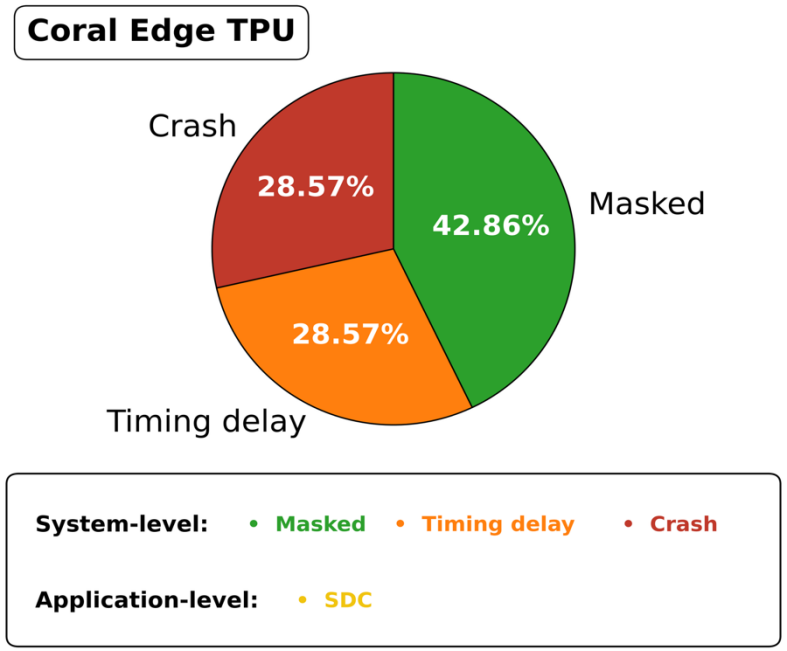


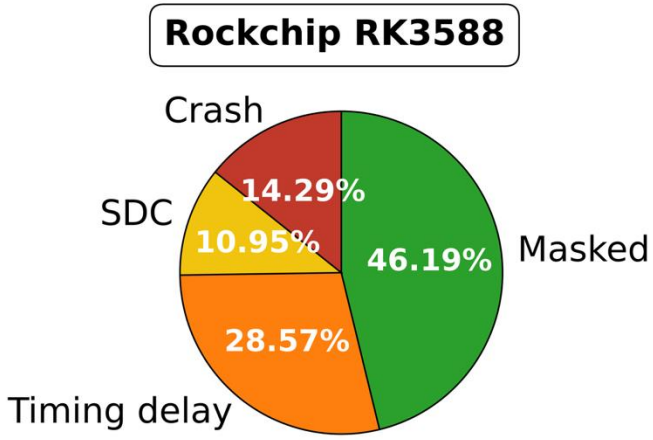
- Focuses on the critical host-device software interface where hardware faults are either contained or amplified
- Extends resilience evaluation to system-visible failures (e.g., hangs, crash escalations, sync breakdowns)
- Unveil **unseen failures modes** agnostic to ML frameworks and AI accelerators

M. Cinque, L. De Simone, N. Di Giacomo. Mind the Driver: A fault injection methodology for AI systems. DSN 2026 Posters track

Experimental Results on two different AI accelerators

RQ: Which failures are due to driver timing errors?





Note:

Watchdog prevents crashes but causes SDCs. Accuracy drops as the application consumes invalid data.

Our on-going contributions

- Fault injection in AI accelerators
- **Workload isolation strategies**
- AI benchmarking framework

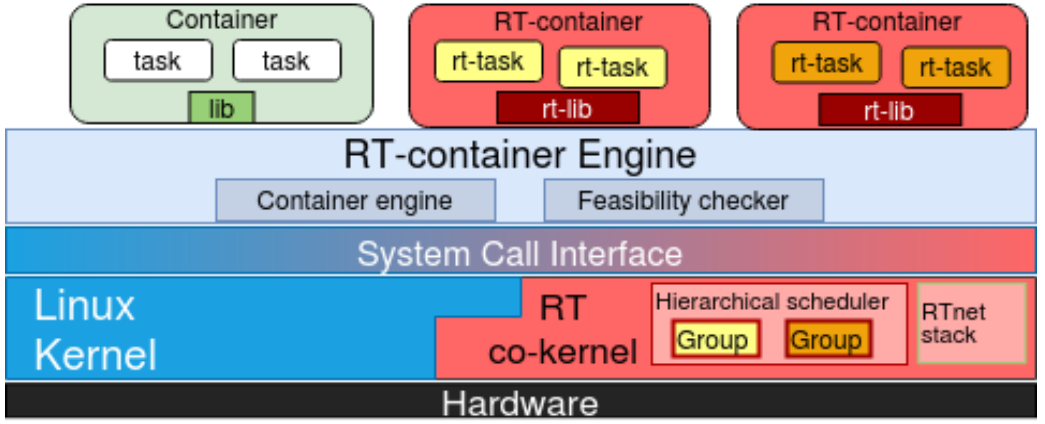
How to isolate AI workloads?

A plethora of approaches for for **real-time containers**:

- **RT-Linux:** running containers on a **PREEMPT_RT** patched kernel with the **SCHED_DEADLINE** policy for group scheduling (rt-cgroups), affinity, isolcpu, …
- **Dual-kernels**: mapping containers on Xenomai or RTAI[1]
- **Sandboxes:** run as lightweight VMs (RunX, firecracker, gvisor, partitioned containers[2], …)
- **Baremetal**: run on accelerators (e.g., Zephyr app on an RPU)

1. M. Barletta, M. Cinque, R. Della Corte, L. De Simone, "Achieving isolation in mixed-criticality industrial edge systems with real-time containers", in ECRTS 2022

2. M. Barletta, M. Cinque, R. Della Corte, G. Farina, L. De Simone, D. Ottaviano. "Partitioned Containers: Towards Safe Clouds for Industrial Applications", DSN 2023 – Disrupt Track



How to isolate AI workloads?

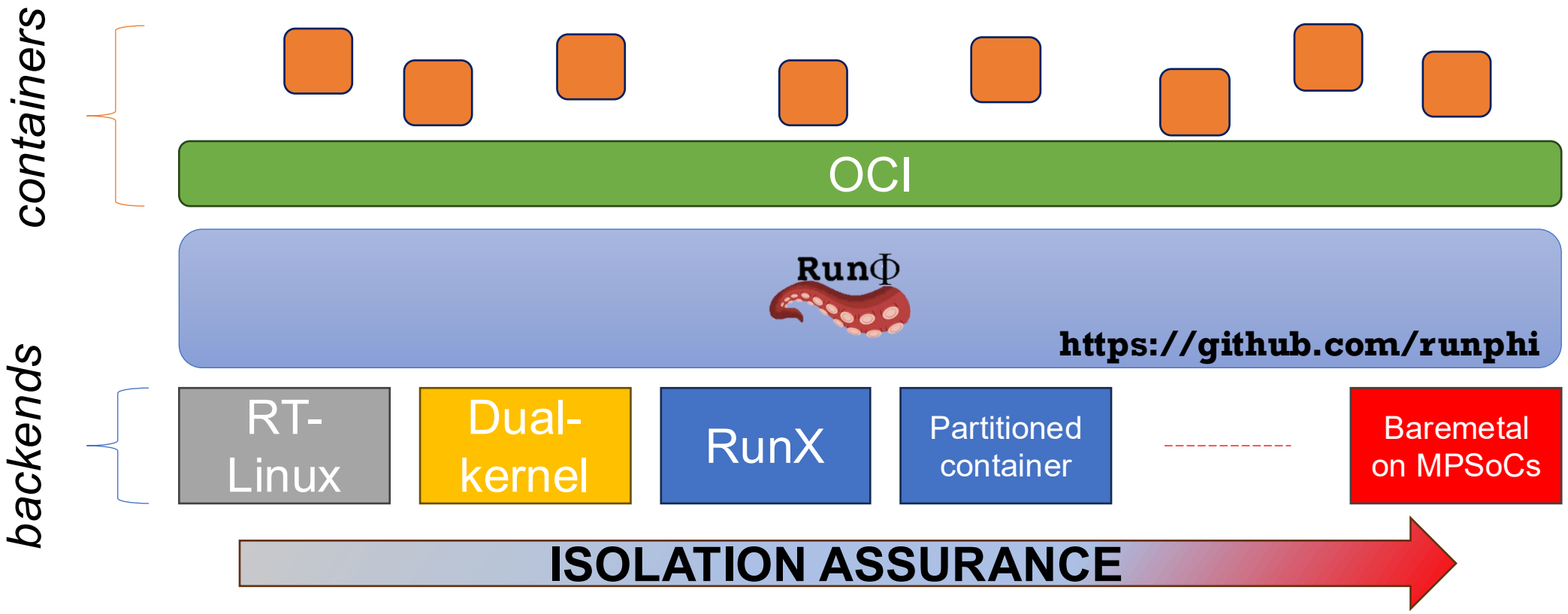
A plethora of approaches for for **real-time containers**:

- **RT-Linux:** running containers on a **PREEMPT_RT** patched kernel with the **SCHED_DEADLINE** policy for group scheduling (rt-cgroups), affinity, isolcpu, …
- **Dual-kernels**: mapping containers on Xenomai or RTAI[1]
- **Sandboxes:** run as lightweight VMs (RunX, firecracker, gvisor, partitioned containers[2], …)
- **Baremetal:** run on accelerators (e.g., Zephyr app on an RPU)

Challenge

Can we **transparently map** a container on all this different "backends", based on criticality requirements?

Mapping containers on backends: RunPHI



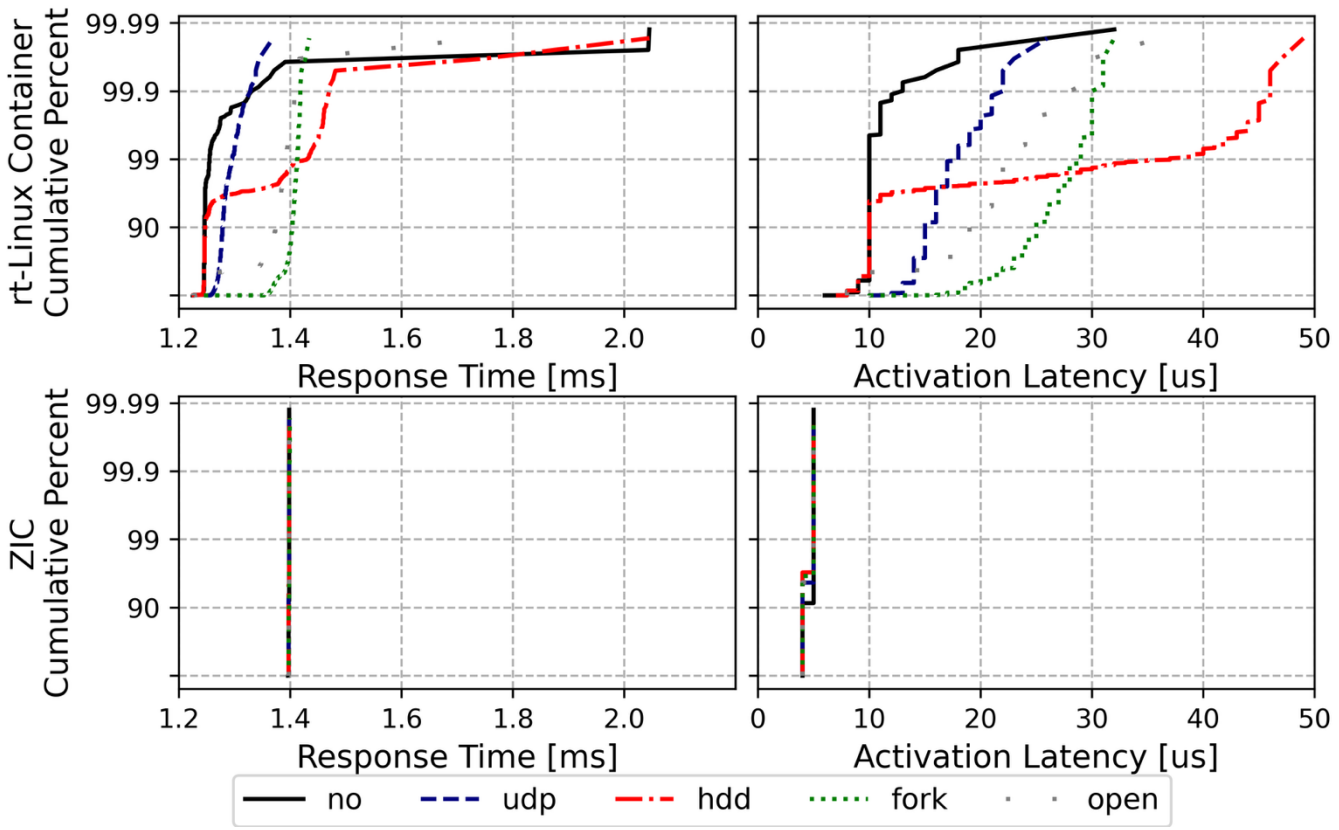
M. Barletta, M. Cinque, L. D. Simone, R. D. Corte, G. Farina and D. Ottaviano, RunPHI: Enabling Mixed-criticality Containers via Partitioning Hypervisors in Industry 4.0, ISSREW 2022

D. Ottaviano, M. Barletta, F. Boccola, Zero-Interference Containers: A Framework to Orchestrate Mixed-Criticality Applications, DSN 2025

Benefits of RunPHI

- Transparency
  - `docker run` works to run the same RT-POSIX container on Linux or on a bare-metal partition
- Mixed-Criticality-native
- System-level Diversity for free

- Workload Isolation



Our on-going contributions

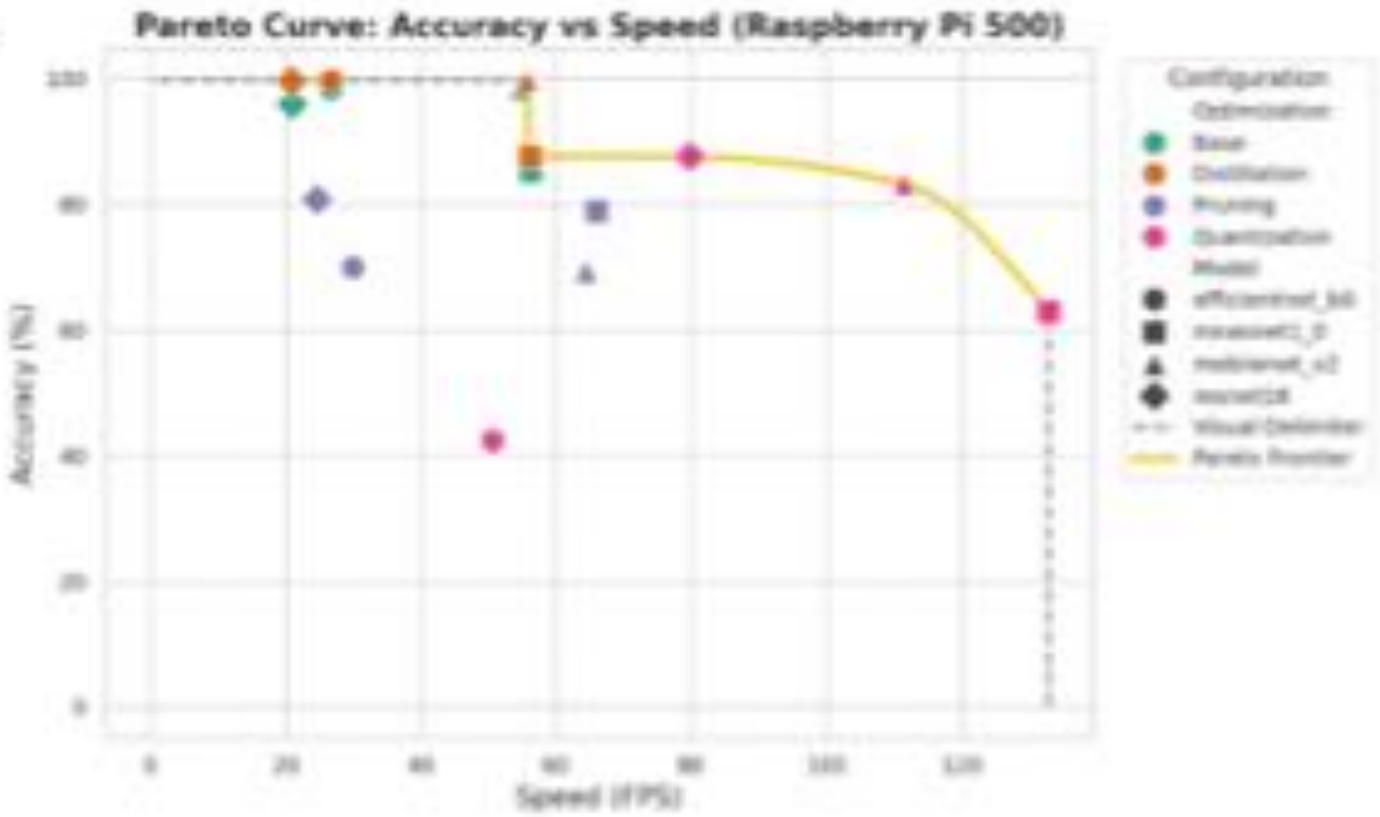
- Fault injection in AI accelerators
- Workload isolation strategies
- **AI benchmarking framework**

MargisBench

- A specialized Command-Line Interface (CLI) to automate the **entire AI end-to-end benchmarking pipeline**.
- Core Objectives:
  - Compare Deep Learning models and their configuration
  - Seamless deployment across diverse edge platforms and AI accelerators
  - Tracking of Inference Time, Static Memory Allocation, and Accuracy

F. Brunello, S. Cangiano, M. Cinque. MargisBench: A Framework to Assess AI models Optimizations and Isolation at the Edge, to appear at DAIES@Safecomp 2026

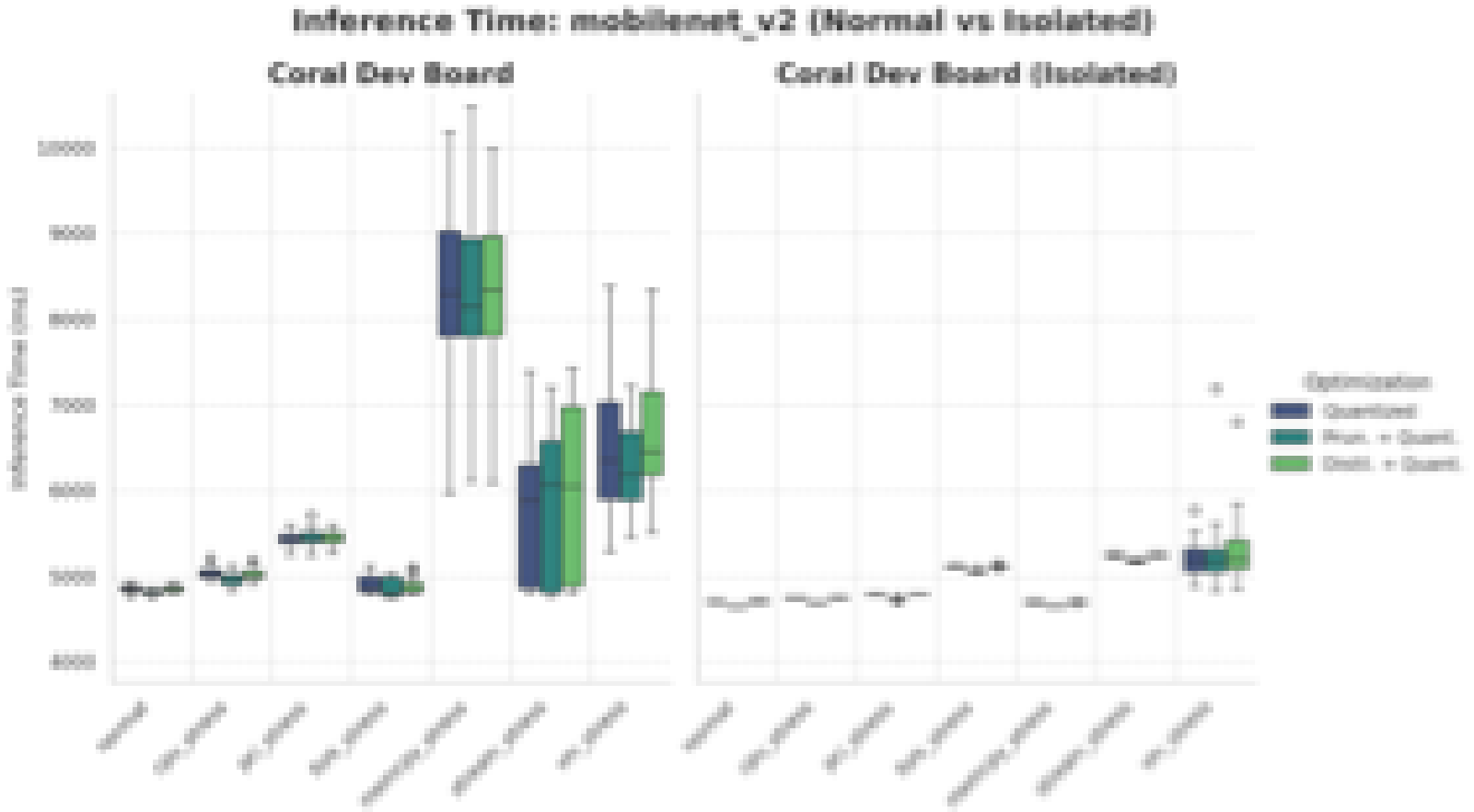
Example result: selecting the best model configuration



Configurations on the pareto curve are the best trade-off between performance and accuracy

Example result 2: Isolation strategies assessment

- CPU pinning and memory access regulation are effective to contain co-located interference



CaPriC

PhD School on
Cyber-**P**hysical **C**loud

Anacapri, Capri Island, Italy
October 19-23, 2026
https://capric-school.github.io/

UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

DIETI
DIPARTIMENTO
DI ECCELLENZA
2023 - 2027

MICS
Made in Italy
Circolare e Sostenibile

Thank you!

macinque@unina.it

DEpendable and Secure Software Engineering and Real-Time Systems (DESSERT) · Federico II

dessertlab.github.io