Github

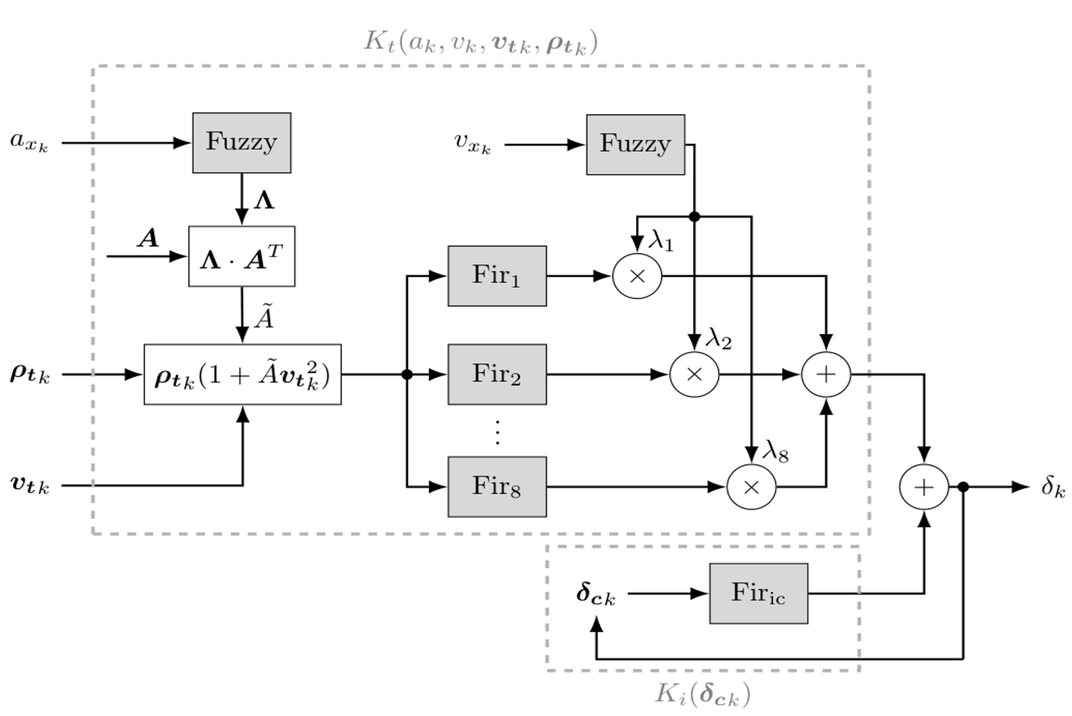
Model-Structured Neural Networks on Edge Hardware: A Preliminary Study

Team

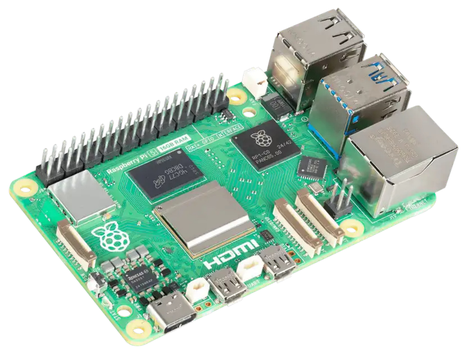
Giovanni Maria Francesco La Scala, Sebastiano Taddei, Francesco Baroni,
Gioele Defrancesco, Filippo Faccini, Gastone Pietro Rosati Papini

Abstract: Deploying neural networks on **edge devices** is hard because of limited resources, but it is important for **real-time** robotics. This preliminary work compares ways to run a compact model-structured neural network on a **Raspberry Pi 5** and finds that a hand-optimized C++ implementation performs better than standard C++/ONNX/PyTorch deployment methods.

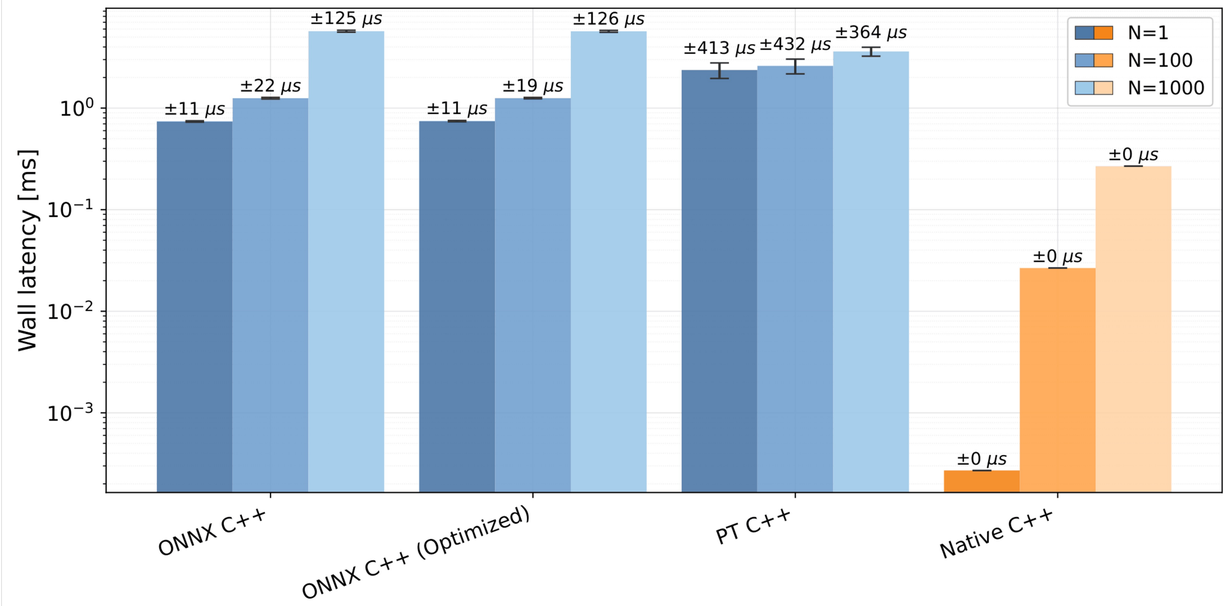
Problem Statement & Methodology

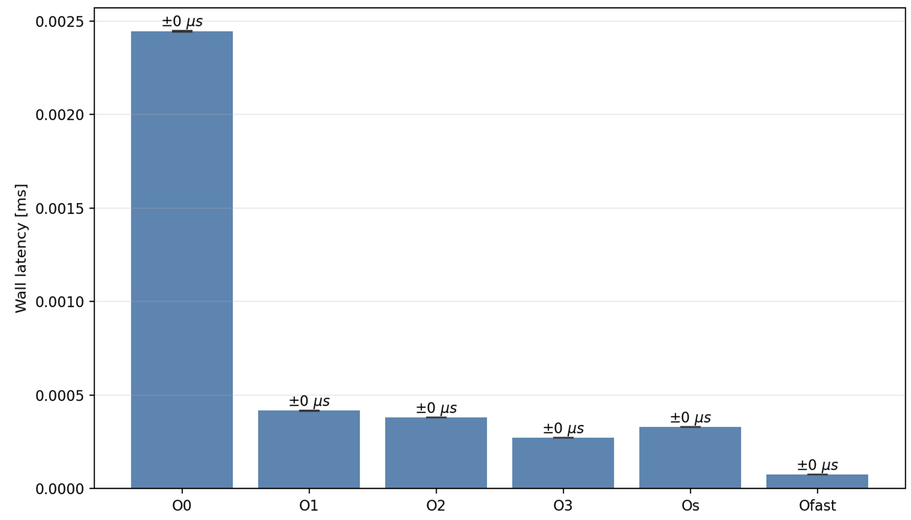


- Comparison of four **MSNN** deployments: **ONNX**, optimized ONNX, **LibTorch**, and native **C++.**
- Evaluation metrics: inference **latency**, **RAM** usage, and **Flash** footprint.
- Benchmarks executed on **Raspberry Pi 5** with different batch sizes.



Results





Implementation	RAM [MB]	Binary [MB]	Dynamic Libraries [MB]
ONNX C++	49.496	0.500	2.711
ONNX C++ Optimized	49.758	0.500	2.711
PT C++	168.903	0.575	310.190
Native C++	23.216	0.498	2.508

Native C++ achieved the lowest latency, outperforming framework-based solutions. Compiler optimizations further improved native backend performance.

Future Works

- Analysis on **different hardware** platforms using different MSNN architectures;
- Develop **code generation** into nnodely framework suitable for embedded platforms at compile time;
- Compare with lightweight methods (i.e. TFLite);
- Investigate the trade-offs between **onboard** computation and **cloud-based** inference by comparing energy consumption and latency for different MSNNs.

References

[1] M. Bechtel, Q. Weng, and H. Yun, “Deeppicarmicro: Applying tinymlto autonomous cyber physical systems,” 2022. [Online]. Available:https://arxiv.org/abs/2208.11212
[2] M. Da Lio, M. Piccinini, and F. Biral, “Robust and sample-efficient es-timation of vehicle lateral velocity using neural networks with explain-able structure informed by kinematic principles,” IEEE Transactions onIntelligent Transportation Systems, vol. 24, no. 12, pp. 13 670–13 684,2023.
[3] M. Da Lio, D. Bortoluzzi, and G. P. Rosati Papini, “Modelling longitudinal vehicle dynamics with neural networks,” Vehicle SystemDynamics, vol. 58, no. 11, pp. 1675–1693, 2020. [Online]. Available:https://doi.org/10.1080/00423114.2019.1638947

This work was supported under the FIS 2 Call, Grant Assignment Decree No. 1236 adopted on 01/08/2023 by the Italian Ministry of University and Research (MUR), for the project FIS-2023-03684 ’Structured neural network framework for modeling and control of autonomous systems – Neu4mes’, CUP E53C24003800001.

Poster Presented at the 8th International Workshop on Robotics Acceleration with Computing Hardware and Systems (RoboARCH 2026)
Co-located with IEEE International Conference on Robotics & Automation (ICRA) 2026, Vienna, Austria