

## CONTEXT

**DIANE Project** (France 2030) – led by Lacroix, with IETR, STMicroelectronics, and CraftAI.



### Project Objectives:

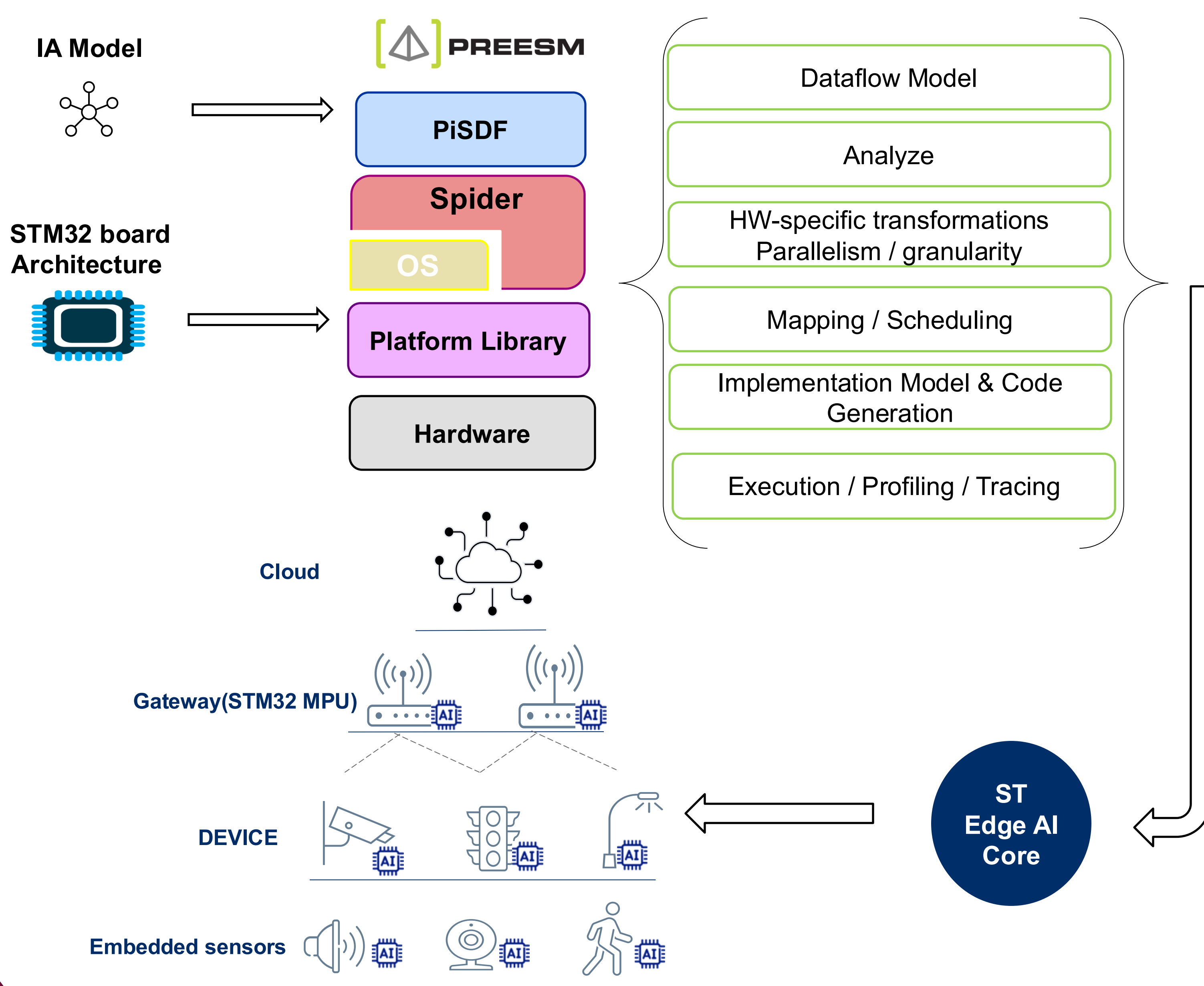
- Migration of AI algorithms from **cloud** → **edge** for smart city services.
- Reduce latency and bandwidth consumption.
- Improve data security.

## PROBLEM STATEMENT

- **Resource-constrained sensors:** limited memory, low compute, low energy.
- **Embedded AI platforms:** STM32MP25 / STM32N6 with AI accelerators.
- **Neural network adaptation:** memory, compute, and energy optimized.
- **Heterogeneous IoT scheduling:** task placement on CPU / GPU / NPU.

## PHD OBJECTIVE

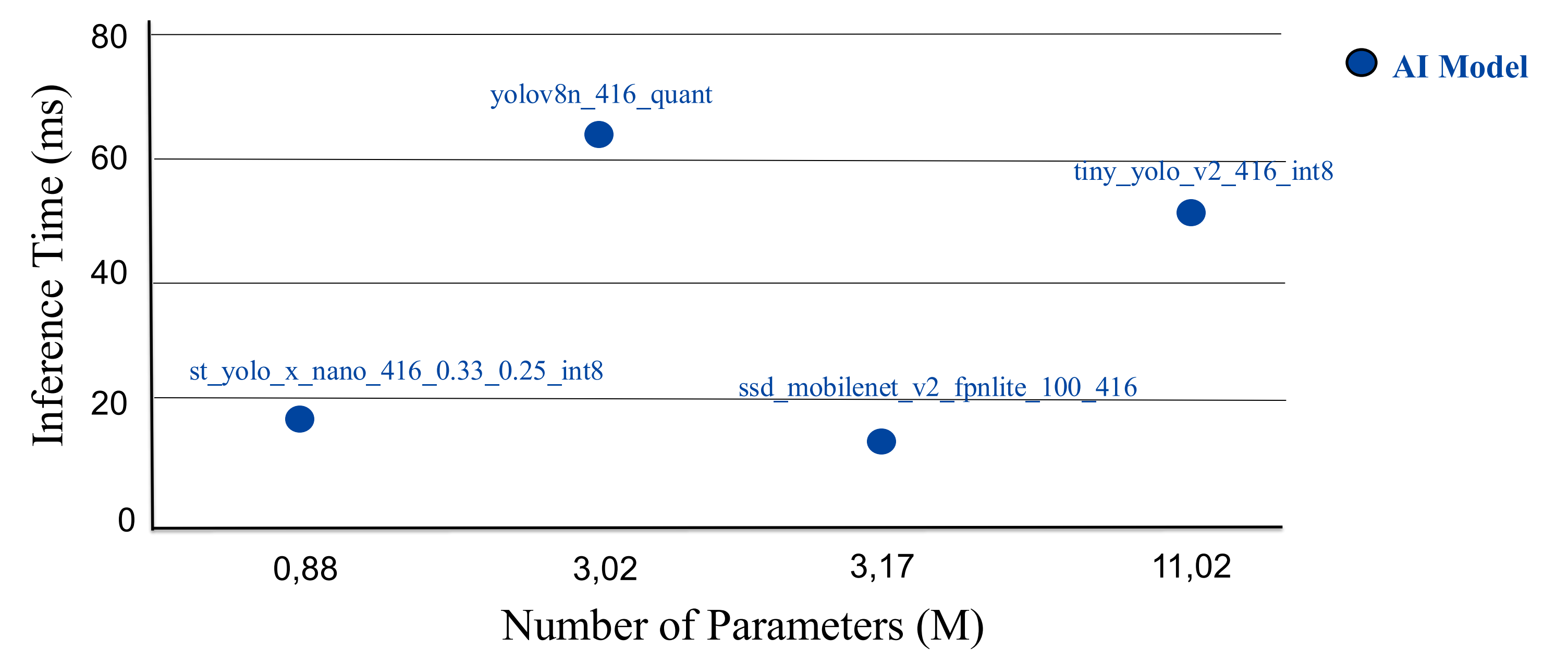
**Dataflow [1]-based AI implementation:** methods and heuristics for running algorithms close to the data using PREESM [2].



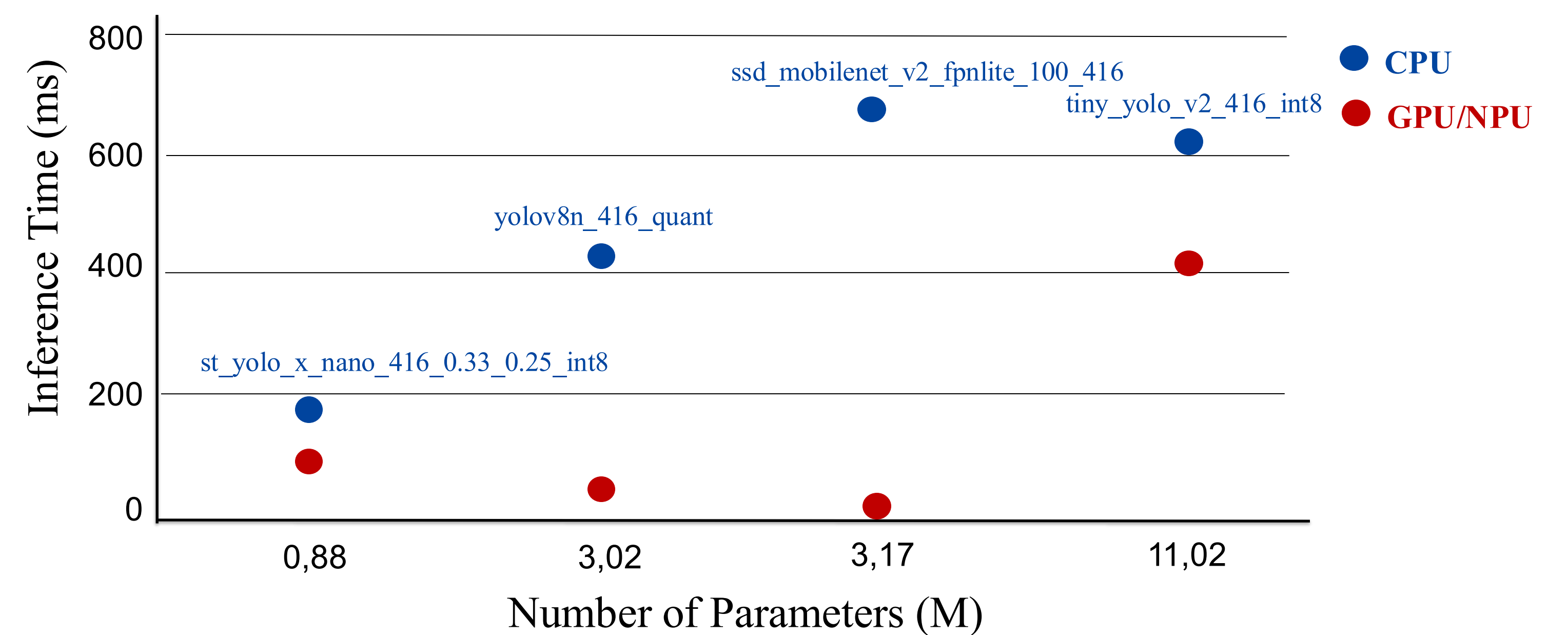
## RESULTS

### Preliminary Experimental Results:

- Inference time does **not scale linearly** with model parameters.
- Leveraging **GPU/NPU acceleration** substantially reduces inference latency.



**Figure 1** : Impact of model parameters on inference time (STM32N6570-DK)



**Figure 2** : Impact of model parameters and processing unit on inference time ( STM32MP25 )

## PERSPECTIVES

- Comparative modeling: **time, memory, energy**
- **Multi-board & multi-core** execution exploration
- Neural network modeling with **PREESM [2]**
- Microcontroller architecture modeling with **PREESM [2]**
- **Heuristics for scheduling** on heterogeneous platforms

## REFERENCE

- [1] A. Honorat, T. Bourgoïn, H. Miomandre, K. Desnos, D. Ménard, J.-F. Nezan, "Influence of Dataflow Graph Moldable Parameters on Optimization Criteria," *DASIP 20*
- [2] M. Pelcat, K. Desnos, J. Heulot, C. Guy, J.-F. Nezan, S. Aridhi, "PREESM: A Dataflow-Based Rapid Prototyping Framework for Simplifying Multicore DSP Programming," *EDERC 2014*.