# Timing Prediction of Deep Neural Networks on Multi-Core Platforms with Memory Hierarchy Effects

Antoine Cantin[1,2], Sébastien Le Nours[1], Sébastien Pillement[1],
Domenik Helms[2], Kim Grüttner[2], Ralf Stemmer[2]

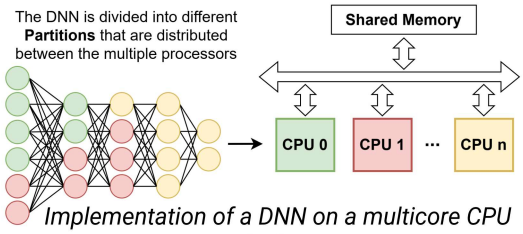
[1]Nantes Univ, CNRS, IETR UMR 6164, F-44000 Nantes, France
[2]German Aerospace Center (DLR), Oldenburg, Germany
Contact : antoine.cantin@etu.univ-nantes.fr

## Context

### Problem statement

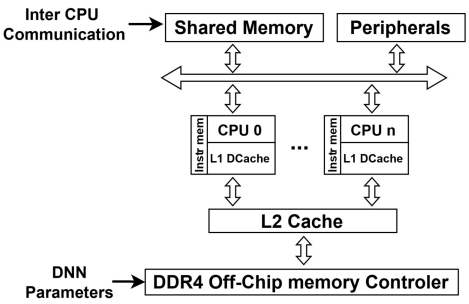
The inference of **Deep Neural Networks (DNNs)** on a resource-limited **multicore CPU** platform must respect strict energy, memory size and timing constraints. Thus, optimizations must be conducted to enhance performance.

These systems have high complexity and massive design space so modeling and **evaluation methods are needed to find the best solutions.**




*Implementation of a DNN on a multicore CPU platform*

## Proposed Modeling and evaluation flow




*Model of Architecture*

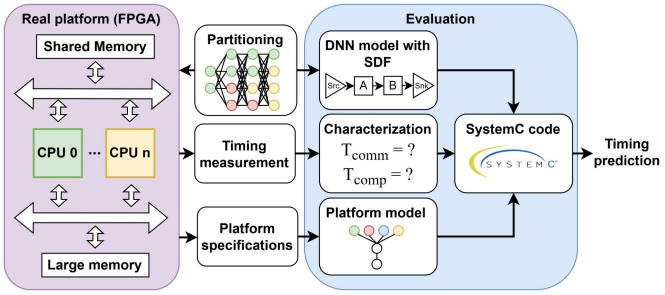
### Model of Architecture (MoA)

The chosen architecture is constituted of **multiple CPU cores** communicating through a shared bus and a shared memory. To store the parameters of a DNN, an external large-sized DDR4 memory is used. A **L1 + L2 cache hierarchy** is also added to improve DDR access performance.

State of the Art multicore platforms such as RK3399Pro HMPSoC, TI AM5K2E0x multicore or NxP MSC8156 DSP have similar architectures.

### New timing prediction framework

- This flow can be used to **predict the execution time** of a DNN on a given platform.
- The MoA is implemented on a ZCU102 board with 1 to 7 32 bits Microblazes and multiple Xilinx IPs.

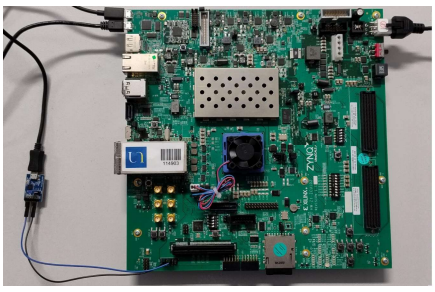



*Overview of the proposed flow*

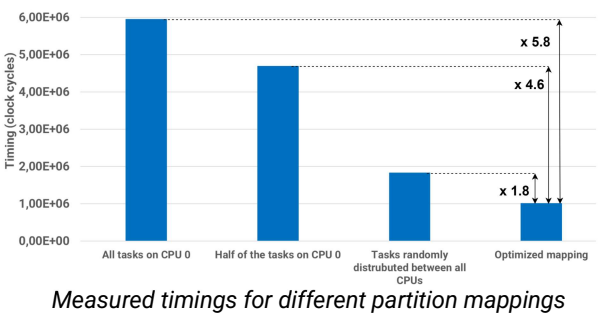
### Evaluation with a hybrid methodology

- A model of the network is built following the rules of the Synchronous Data Flow model of computation
- Communication and computation **analytical models** are calibrated through **measurements**
- An abstract representation of the platform is created from specifications
- Everything feeds a **SystemC simulation** that produces the timing prediction
- This prediction can then be integrated in a **Design Space Exploration** (DSE) flow that search for optimized solutions
- The main goal is to optimize the deployement of a DNN on the platform : the partitioning, the mapping of the partitions on the CPUs or the mapping of data in memory

## Experimental platform



*ZCU102 board with the measurement infrastructure*




*Measured timings for different partition mappings*

## Future work

- Update analytical models for a complex memory hierarchy. The mathematical models that describes the computation time have to take in account the latency induced by caches and by the concurrent access of multiple CPUs to the DDR.
- Test the flow with ResNet
- Implement power consumption prediction

12/06/2025

This work is supported by France 2030 Priority research program and equipment for artificial intelligence PEPR AI, under the ref ANR-23-PEIA-0009.