A Measurement-based Performance Evaluation Framework for Neural Networks on MPSoCs

Quentin Dariol, Sébastien Le Nours, Sébastien Pillement, Ralf Stemmer, Kim Grüttner, Domenik Helms

A Measurement-based Performance Evaluation Framework for Neural Networks on MPSoCs

Quentin DARIOL, Sébastien LE NOURS, Sébastien PILLEMENT (IETR),
Ralf STEMMER, Kim GRÜTTNER, Domenik HELMS (OFFIS e. V.)
Email: quentin.dariol@etu.univ-nantes.fr

CONTEXT - OBJECTIVE

❖ **Difficulty in timing prediction for neural networks executed on MPSoC**

➢ Hard to predict variability of execution time due to dependency to input data and contention on shared resources,

➢ Required exploration of partitioning and mapping/scheduling of neural network for efficient implementation.

❖ **Current limitations of existing real-time analysis methods for neural networks**

➢ Most approaches focus on systematic implementations to measure execution time while exploring mapping/schedulings, and few high level models have been proposed to predict execution time while taking variability inducing phenomena in consideration.

➢ Variability on parallel software execution can be captured by probabilistic model.

❖ **Objective**

➢ Evaluate the efficiency of probabilistic SystemC models for the analysis of non-functional properties of neural networks on MPSoC.

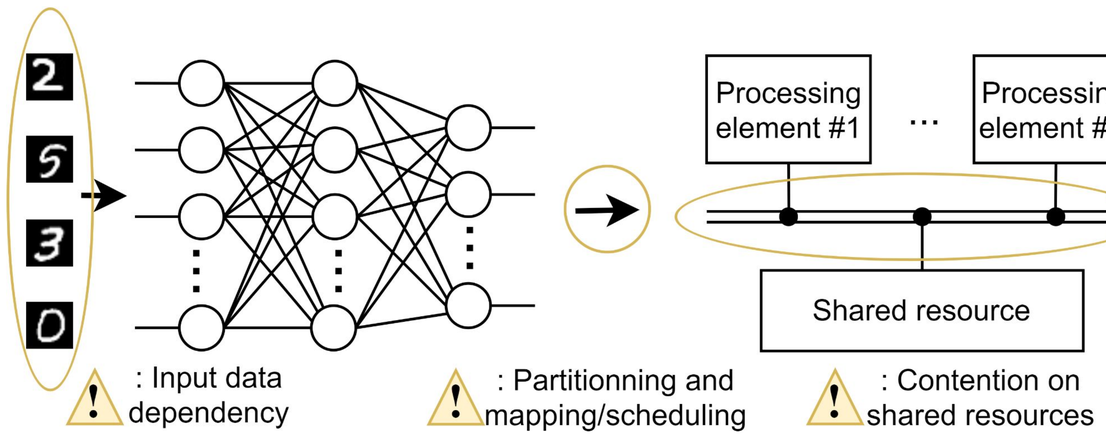

Fig 1. Causes of difficulty in performance prediction of neural networks on MPSoC

CONTRIBUTION

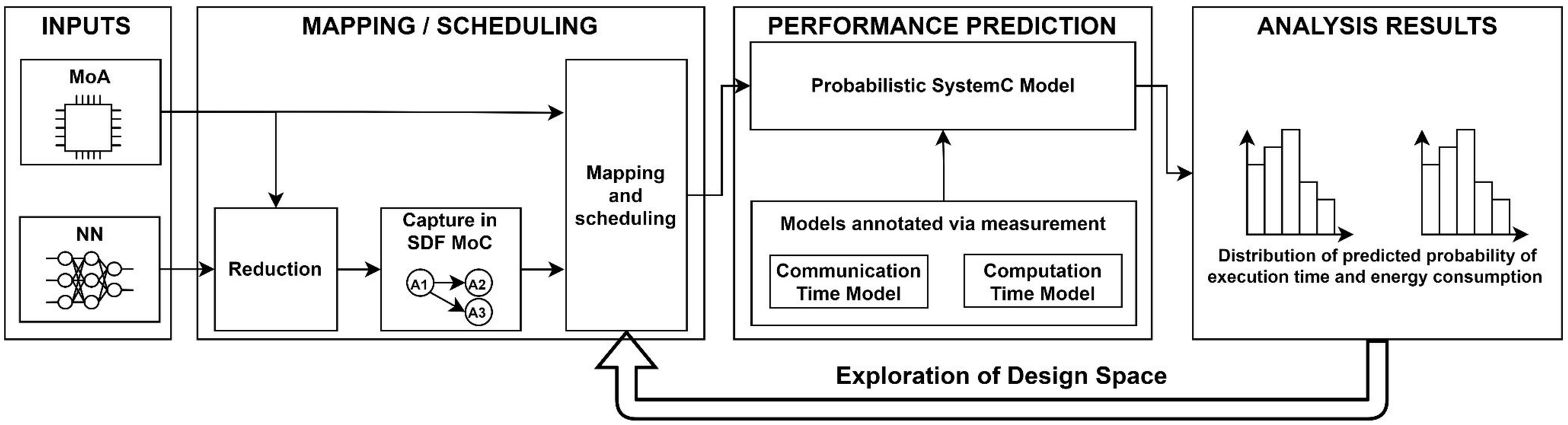

Fig 2. Workflow proposed for performance prediction of neural networks on MPSoC

❖ **Evaluation workflow**

➢ Update for neural network applications of our previous workflow for timing prediction for MPSoC tested on video processing applications [1] [2].

➢ Performance prediction performed by scalable probabilistic models annotated via measured timings, obtained through implementation on FPGA,

➢ Exploration of several mapping/schedulings optimizing timing and energy for considered application.

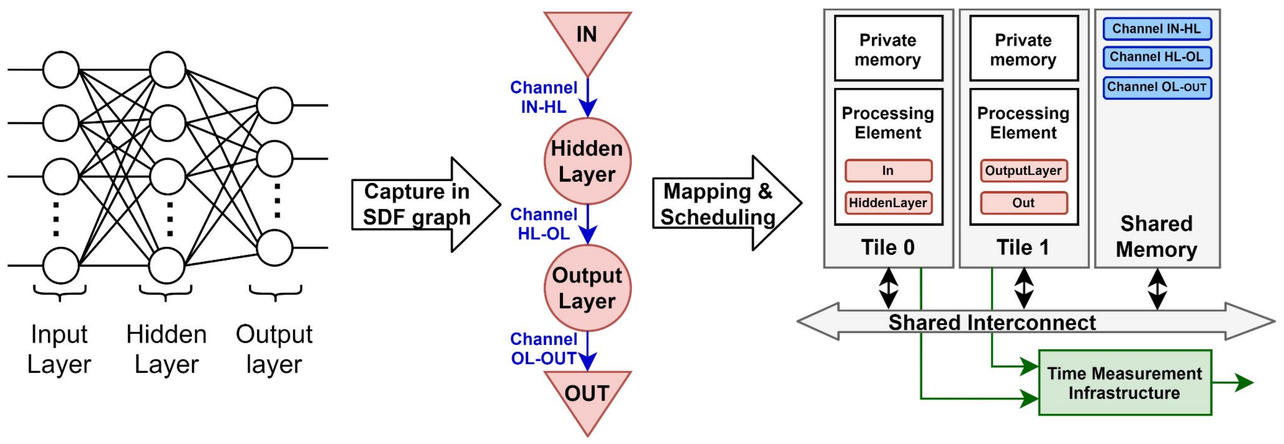

Fig 3. Capture of a neural network application as SDF graph and mapping/scheduling on targeted MPSoC platform

❖ **Case study**

➢ Elementary multi-layer perceptrons (such as MNIST use-case) captured using LibFANN [3] and modeled in Synchronous Dataflow Graph (SDFG),

➢ Model of Architecture (MoA) consisting of a shared memory and several tiles composed of one processing element with local memory.

PROSPECTS

❖ **Prospects**

➢ Update of Model of Architecture to include a DDR memory and mechanisms such as caches or DMAs on tiles to handle the important memory needs of neural network applications.

➢ Adapt the workflow for Convolutional Neural Network (CNN) applications,

➢ Allow exploration of several grains to describe neural network applications using the SDF Model of Computation (MoC), to explore more possible optimized implementations.

[1] Ralf Stemmer, Hai-Dang Vu, Kim Gr¨uttner, Sebastien Le Nours, Wolfgang Nebel, et al.. Towards Probabilistic Timing Analysis for SDFGs on Tile Based Heterogeneous MPSoCs. 10th European Congress on Embedded Real Time Software and Systems (ERTS 2020), Jan 2020, Toulouse, France.
[2] Hai-Dang Vu. Fast and Accurate Performance Models for Probabilistic Timing Analysis of SDFGs on MPSoCs. Electronics. Universite de Nantes, 2021. English.
[3] Nissen, Steffen. (2003). Implementation of a Fast Artificial Neural Network library (FANN).