Early Performance and Energy Prediction of Neural Networks Deployed on Multi-Core Platforms

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

▶ **To cite this version:**

Quentin Dariol, Sébastien Le Nours, Sébastien Pillement, Ralf Stemmer, Domenik Helms, et al.. Early Performance and Energy Prediction of Neural Networks Deployed on Multi-Core Platforms. GRETSI 2023 XXIXème Colloque Francophone de Traitement du Signal et des Images, Aug 2023, GRENOBLE, France. 2023-08, pp.ID PAPER 1144. hal-04186902

HAL Id: hal-04186902
https://hal.science/hal-04186902

Submitted on 24 Aug 2023

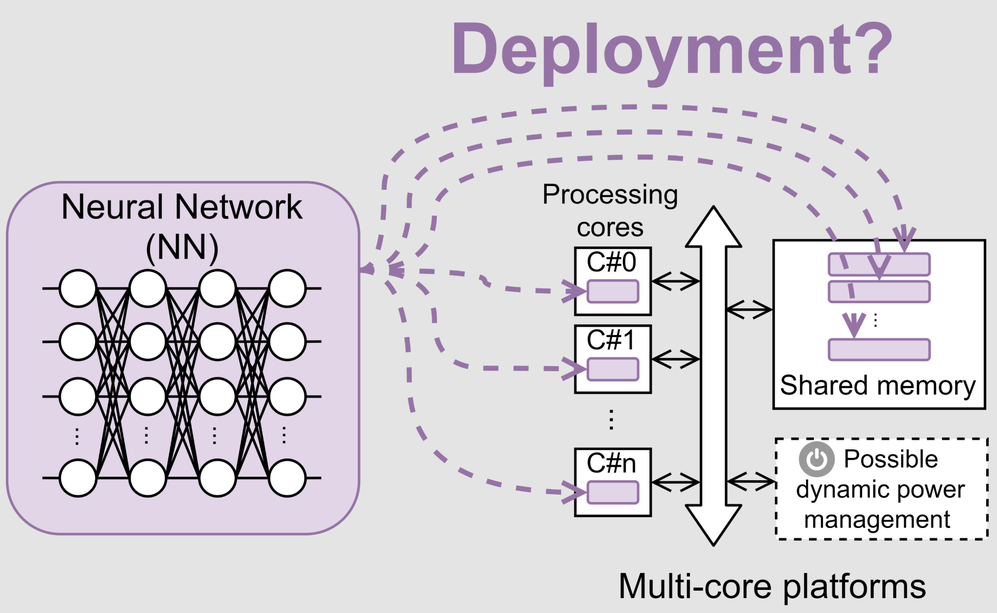
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

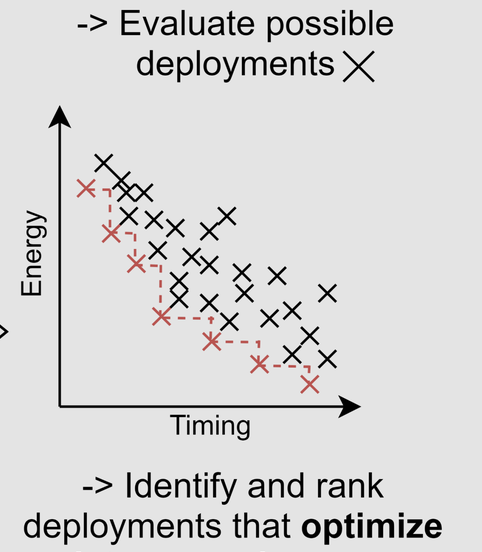
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Early Performance and Energy Prediction of Neural Networks Deployed on Multi-Core Platforms

Quentin Dariol[1,2], Sébastien Le Nours[1], Sébastien Pillement[1], Ralf Stemmer[2], Domenik Helms[2], Kim Grüttner[2]
1: Nantes Université – Institut d'Electronique et des Technologies du NumériQue (IETR) UMR CNRS 6164, Nantes, France | Contact: quentin.dariol@dlr.de
2: German Aerospace Center - Institute for Systems Engineering for Future Mobility (DLR SE) , Oldenburg, Germany

Motivation





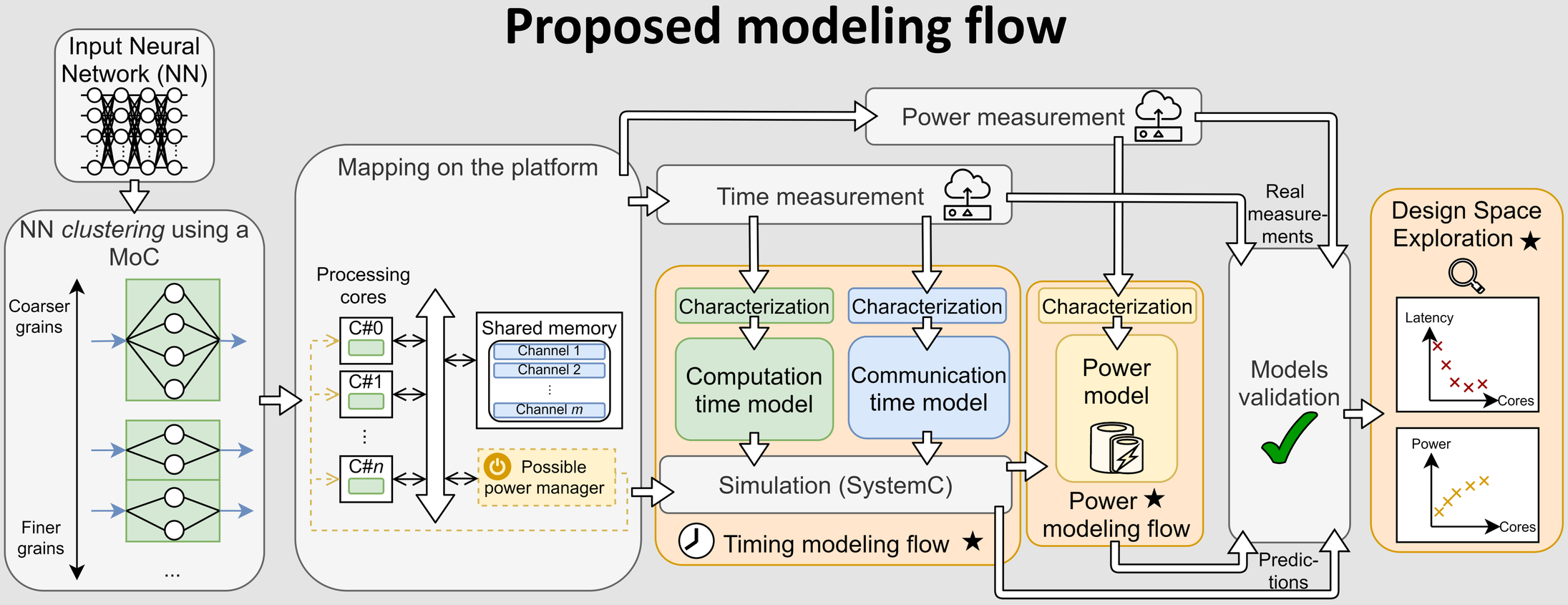
Context

- Raise of interest for **NNs deployed onto edge embedded platforms**, and more specifically **multi-core** platforms.
- **Early prediction** of non-func. properties (e.g. **latency, energy**) of NN deployments is necessary to **optimize execution** and **meet user defined constraints.**

Challenges - How to propose **fast yet accurate models** with scalability in consideration of:

- **Contention for shared resources,**
- **Computation/communication workload variability** between deployments,
- Influence of **power management** strategies (e.g. clock gating).

Proposed modeling flow



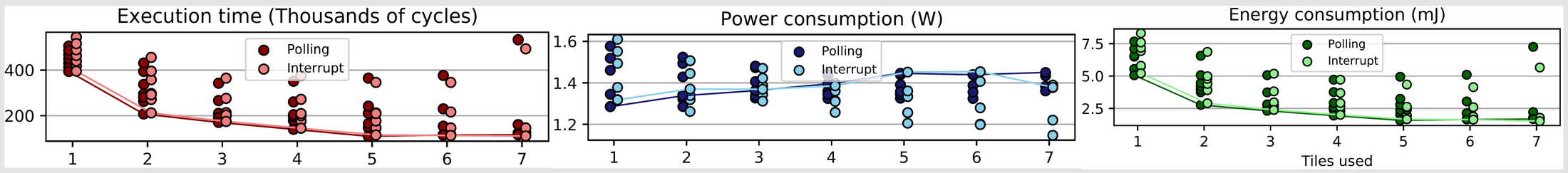
→ The proposed modeling flow relies on **calibration through measurements**, **analytical models** and **high level simulation** to **offer fast-yet-accurate evaluation**.

Results

- **27 mappings** of **4 NNs including 1 CNN** tested with and without power management.
- **Timing** properties predicted with **>97% accuracy**, and **energy** properties with **>95% accuracy**.
- More information on the project can be found on our **open source Git repository:**

https://gitlab.univ-nantes.fr/lenours-s/pssim4ai

→ Example use of the modeling flow to perform **Design Space Exploration (DSE) under performance and energy constraints:**



Poster presented during the GRETSI'23 colloquium, 28.08.23 – 01.09.23, Grenoble, France.
This work was funded by the WISE consortium, France in the project pSSim4AI and by the Federal Ministry of Education and Research (BMBF, Germany) in the project Scale4Edge (16ME0465).