COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Électronique*

Par
Quentin DARIOL

Early Timing and Energy Prediction and Optimization of Artificial Neural Networks on Multi-Core Platforms

Thèse présentée et soutenue à Nantes, France, le 27/11/2023
Unité de recherche : IETR UMR CNRS 6164

Rapporteurs avant soutenance :

Prof. Dr.-Ing. Matthias JUNG	Full professor, Würzburg Universität, Germany
Dr. Angeliki KRITIKAKOU	Associate professor - HDR, IRISA/INRIA, Université de Rennes, France

Composition du Jury :

Examinateurs :	Dr. Kim GRÜTTNER	Head of department, German Aerospace Center (DLR), Germany
	Prof. Dr.-Ing. Matthias JUNG	Full professor, Würzburg Universität, Germany
	Dr. Angeliki KRITIKAKOU	Associate professor - HDR, IRISA/INRIA, Univ. Rennes, France
	Prof. Dr. Frédéric PÉTROT	Full professor, TIMA, Université Grenoble Alpes, France
	Prof. Dr. Gregor SCHIELE	Full professor, Duisburg-Essen Universität, Germany
Dir. de thèse :	Prof. Dr. Sébastien PILLEMENT	Full professor, IETR, Nantes Université, France
Encadrant :	Dr. Sébastien LE NOURS	Associate professor - HDR, IETR, Nantes Université, France

Invité :

Dr. Domenik HELMS	Principal scientist, German Aerospace Center (DLR), Germany

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Prédiction et optimisation des propriétés temporelles et de l'énergie des réseaux de neurones artificiels implémentés sur les plateformes multicœurs

Mot clés : Intelligence artificielle embarquée, conception au niveau système, prédiction des propriétés temporelles et de l'énergie

Résumé : Le besoin de mettre en oeuvre les Réseaux de Neurones artificiels (NNs) sur des plates-formes multicœurs embarquées est devenu fondamental. La prédiction des propriétés temporelles (temps d'inférence, latence, débit) et énergétiques au plus tôt dans le processus de conception est nécessaire pour trouver des solutions qui optimisent l'utilisation des ressources et respectent les contraintes imposées au système. Une difficulté majeure de cette modélisation vient de la nécessité de décrire correctement l'influence du partage de ressources (processeur, mémoire, bus de communication) au sein des plateformes multicoeurs. Dans cette thèse, nous présentons un flot complet de prédiction et d'optimisation des propriétés temporelles et de l'énergie qui combine plusieurs approches de modélisation. Ce flot conduit à optimiser l'occupation des ressources sans dégrader les performances des NNs mis en oeuvre. Les prédictions sont confrontées à des expérimentations sur cibles réelles. Les modèles proposés ont une précision de plus de 97% sur le temps et 93% sur l'énergie sur 54 mappings de 4 NNs, avec un temps de prédiction de 20s par mapping. Nous montrons comment utiliser les modèles pour explorer efficacement l'espace de conception et trouver des solutions optimisées qui satisfont les contraintes imposées au système.

Title: Early Timing and Energy Prediction and Optimization of Artificial Neural Networks on Multi-Core Platforms

Keywords: Embedded artificial intelligence, system level design, timing and energy prediction

Abstract: The need to implement artificial Neural Networks (NNs) on embedded multi-core platforms has become fundamental. Predicting timing properties (inference time, latency, throughput) and energy as early as possible in the design process is necessary to find solutions that optimize resource use and respect the constraints imposed on the system. A major modeling difficulty comes from the need to correctly describe the influence of resource sharing (processor, memory, communication bus) within multi-core platforms. In this thesis, we present a complete flow for predicting and optimizing timing properties and energy, combining several modeling approaches. This flow leads to optimized resource occupancy without degrading the performance of implemented NNs. Predictions are compared with measurements on real targets. The proposed models have an accuracy of over 97% on timing and 93% on energy for 54 mappings of 4 NNs, with a prediction time of 20s per mapping. We show how to use the models to efficiently explore the design space and find optimized solutions that satisfy the constraints imposed on the system.