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 : *Electronique*

Par
Mohamed Amine ZHIRI

Reconfigurable Network-on-Chip for AI Dataflow Accelerators

Thèse présentée et soutenue à Paris-Saclay, le 20 novembre 2026
Unité de recherche : IETR UMR 6164

Rapporteurs avant soutenance :

M. Abdoulaye GAMATIÉ Directeur de Recherche, CNRS LIRMM, Montpellier
M. Cédric KILLIAN Professeur des Universités, Université de Saint-Etienne

Composition du Jury :

Président :		
Examinateurs :	Daniel CHILLET	Professeur des Universités, Université de Rennes
	M. Abdoulaye GAMATIÉ	Directeur de Recherche, CNRS LIRMM, Montpellier
	M. Cédric KILLIAN	Professeur des Universités, Université de Saint-Etienne
Dir. de thèse :	M. Sébastien PILLEMENT	Professeur des Universités, Nantes Université
Enc. de thèse :	Mme. Hana KRICHENE	Ingénieur de Recherche, CEA-List, Paris-Saclay
	Mme. Chiara SANDIONIGI	Ingénieur de Recherche, CEA-List, Grenoble

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Réseau sur puce reconfigurable pour les accélérateurs dataflow d'IA

Mot clés : Réseau sur puce, Accélérateur d'IA, Dataflow de DNN, Architecture Reconfigurable, Exécution Multi-DNN

Résumé : Les applications modernes basées sur les réseaux de neurones profonds (DNN) présentent des exigences de communication hétérogènes, caractérisées par des besoins en bande passante variés, faisant de l'efficacité des transferts de données un facteur essentiel pour les performances des accélérateurs d'intelligence artificielle (IA). En outre, l'exécution multi-DNN introduit un défi supplémentaire, car plusieurs réseaux partagent simultanément les ressources d'un même accélérateur. Les infrastructures de communication offrant une adaptabilité limitée ne permettent donc pas de répondre efficacement à ces exigences hétérogènes et dynamiques.

Cette thèse répond à ces défis à travers deux contributions complémentaires. HT-NoC est un réseau sur puce multi-dataflow à débit adaptatif, qui s'ajuste aux exigences de communication des différentes couches de DNN. SPADA est un mécanisme matériel de partitionnement spatial qui alloue dynamiquement les ressources de l'accélérateur en fonction des caractéristiques des couches, permettant ainsi l'exécution simultanée de plusieurs couches de DNN. Ensemble, HT-NoC et SPADA constituent des mécanismes matériels permettant de répondre aux exigences de communication hétérogènes et d'assurer une exécution multi-DNN efficace dans les accélérateurs d'IA.

Title: Reconfigurable Network-on-Chip for AI Dataflow Accelerators

Keywords: Network-on-Chip, AI Accelerators, DNN Dataflows, Reconfigurable Architecture, Multi-DNN Execution

Abstract:

Modern Deep Neural Network (DNN) workloads exhibit heterogeneous communication requirements, with diverse bandwidth demands that make efficient data movement critical to the performance of AI accelerators. Furthermore, multi-DNN execution introduce an additional challenge, as multiple workloads concurrently share accelerator resources. Communication infrastructures with limited adaptability are therefore insufficient to efficiently accommodate such heterogeneous and dynamic requirements.

This thesis addresses these challenges through two complementary contributions. HT-NoC is a throughput-adaptive, multi-dataflow network-on-chip that adapts its reconfigurable communication modes to the characteristics and communication requirements of different DNN layers. SPADA is a hardware-assisted spatial partitioning mechanism that dynamically allocates accelerator resources according to layer characteristics, enabling the concurrent execution of multiple DNN layers. Together, HT-NoC and SPADA provide hardware mechanisms for addressing heterogeneous communication requirements and multi-DNN execution in AI accelerators.