DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

**Sujet de thèse:** **Réseau d'interconnexion auto-adaptable et fiable pour architectures d'accélérateurs pour l'IA**

**Mots clefs :** architectures, réseaux de neurones, inférence dynamique, NoC auto-adaptable, NoC fiable

**Contexte technique et scientifique**

Les algorithmes d'apprentissage profond tels que les réseaux neuronaux convolutifs (CNN, Convolutional Neural Networks) sont largement utilisés dans de nombreuses applications usuelles telles que la reconnaissance et la classification d'images. Ces CNN sont composés de plusieurs couches de neurones qui nécessitent plusieurs mégaoctets (voire Giga pour les modèles très avancés) de paramètres pour des milliards d'opérations en une seule passe d'inférence [1], nécessitant ainsi d'importants mouvements de données. De nombreux travaux de recherche ont proposé des architectures dédiées améliorant l'inférence des CNN, mais peu ont proposé un travail d'optimisation sur le réseau d'interconnexion. En effet, l'intégration de réseaux efficaces dans les architectures de calcul pour l'IA pose des problèmes majeurs de consommation d'énergie, de garantie de service et de coûts de communication. Le traitement des CNNs doit non seulement assurer un parallélisme élevé pour un haut débit mais aussi optimiser le mouvement des données de l'ensemble du système pour atteindre une haute efficacité énergétique. De plus, les accélérateurs matériels de réseaux neuronaux actuels ne sont pas optimisés pour les applications de nouvelle génération, où les charges de travail changent en temps réel et les transferts de données sont irréguliers. Ces comportements dynamiques augmentent également en raison de la mise en œuvre de nouvelles techniques d'optimisation, comme le pruning [2] pour réduire le nombre des paramètres des réseaux, pour en améliorer l'efficacité énergétique, et notamment sur les réseaux neuronaux profonds (DNN, Deep Neural Networks). Dans ce contexte d'applications dynamiques, l'équilibrage de la charge et le dimensionnement des ressources deviennent des tâches difficiles à anticiper et à réaliser. De plus, la congestion et les blocages dans les transferts de données peuvent facilement compromettre les performances de la qualité de service (QoS, Quality of Service). Enfin, les évolutions technologiques entraînent une plus grande sensibilité des architectures aux perturbations de l'environnement et au vieillissement des composants, et notamment pour le support de communications qui doit donc être fiabilisé.

Dans ce contexte, le travail de la thèse visera à développer un réseau sur puce (NoC, Network-on-Chip) dynamique et tolérant aux fautes pour supporter l'implémentation d'algorithmes d'IA évolutifs et flexibles. Il s'appuiera sur une gestion optimisée du flux de données pour un traitement à faible consommation et efficace des accès aux paramètres des CNN. Le défi majeur de cette thèse est donc d'assurer le caractère auto-adaptable dans le réseau d'interconnexion pour des architectures IA moderne, tout en garantissant la QoS et en étant fiable.

Pour adresser l'ensemble de ces challenges simultanément, la thèse devra proposer des techniques de gestion du réseau innovantes elle-même basées sur l'IA afin de pouvoir s'adapter aux changements applicatifs et à l'apparition de fautes et/ou de congestions au sein du réseau. L'auto-adaptabilité du réseau devra être supportée par des techniques matérielles (telles que des mécanismes de surveillances légères [3] et de reconfiguration dynamique locale de l'architecture pour une utilisation intelligente des ressources) et logicielles (telles que des mécanismes de tolérance aux fautes [4] et des heuristiques d'optimisations [5] afin d'équilibrer la charge du trafic et de contourner les zones du réseau encombrées ou avec des fautes). Ces techniques peuvent utiliser des méthodes hybrides [6] qui fournissent un apprentissage hors ligne pour surveiller la charge de travail et la consommation d'énergie et une décision en ligne pour lancer la réaffectation des applications au sein de l'architecture.

list
Laboratoire d'Intégration des Systèmes et des Technologies

leti
Laboratoire d'Electronique et de Technologie de l'Information

Commissariat à l'Energie Atomique et aux Energies Alternatives
Institut Carnot CEA LIST
Centre de Saclay | Nano-Innov Bât 862 | PC 172
91191 Gif sur Yvette Cedex

Direction de la Recherche Technologique
Département Architecture Conception et Logiciels Embarqués

**Environnement de recherche**

La thèse est proposée par le Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) en collaboration avec l'Institut d'Electronique et des Technologies du numéRiques site de l'Université de Nantes. Le CEA est un acteur majeur en matière de recherche, de développement et d'innovation. Cet organisme de recherche technologique intervient dans trois grands domaines : l'énergie, les technologies pour l'information et la santé, et la défense. Reconnu comme un expert dans ses domaines de compétences, le CEA est pleinement inséré dans l'espace européen de la recherche et exerce une présence croissante au niveau international. Le Laboratoire d'Intégration des Systèmes et des Technologies (LIST) a notamment pour mission de contribuer au transfert de technologies et de favoriser l'innovation dans le domaine des systèmes embarqués.

Le laboratoire d'accueil est le Laboratoire Environnement de Conception et Architecture (LECA) qui est chargé de concevoir des architectures de calcul sur puce innovantes et flexibles répondant aux enjeux de performance, de coût, consommation énergétique, sûreté et sécurité, avec une emphase sur les architectures pour les systèmes embarqués critiques, et les accélérateurs dédiés au calcul neuronal (IA-DNN/CNN). Ce laboratoire est basé à Palaiseau (au sud de Paris (91)) et fait partie de l'institut CEA-LIST (http://www-list.cea.fr/index.php/). Le doctorant sera rattaché à l'équipe ASIC de l'IETR sur le site de Polytech'Nantes (https://www.ietr.fr). L'équipe développe des activités de recherche sur la conception et l'optimisation de systèmes embarqués fiables et dynamiques, et possède une expertise reconnue dans la conception d'architectures.

| | | | |
|---|---|---|---|
| Encadrantes : | Hana KRICHENE | CEA | Email : hana.krichene@cea.fr |
| | Chiara SANDIONIGI | CEA | Email : chiara.sandionigi@cea.fr |
| Directeur de thèse : | Sébastien PILLEMENT | | Université de Nantes, IETR, CNRS (UMR 6164)<br>Email : Sebastien.Pillement@univ-nantes.fr |
| Qualifications : | Master en informatique/électronique. Bonne connaissance en réseaux de neurones et programmation pour l'embarqué. Des bonnes qualités d'analyse et d'expérimentation seront très appréciés. | | |
| Candidature : | Envoyez à Hana KRICHENE votre dossier comprenant:<br>- Un CV détaillé,<br>- Une lettre de motivation,<br>- Notes et rang sur les 3 dernières années,<br>- Les noms de 2 références qui pourraient vous recommander | | |
| Contact : | Hana KRICHENE<br>CEA DRT/DSCIN/LECA<br>Institut Carnot CEA List - CEA Saclay - Nano-INNOV<br>Bat 862 - PC 172 F91191 Gif-Sur-Yvette Cedex<br>Téléphone: +33 1 69 08 36 37<br>Email: hana.krichene@cea.fr | | |
| Date de début : | Septembre 2022 | | |

**Références :**


[1] LeCun, Y. et al., "Deep learning", Nature, vol. 521, 2015.

[2] Bondarenko, A et al., "Neurons vs Weights Pruning in Artificial Neural Networks". Environment Technology Resources Proceedings of the International Scientific and Practical Conference. vol 3.166, 2015

[3] Das, R. et al., "Catnap: Energy proportional multiple network-on-chip", International Symposium on Computer Architecture, 2013.

[4] Nain, Z. et al., "A Network Adaptive Fault-Tolerant Routing Algorithm for Demanding Latency and Throughput Applications of Network-On-A-Chip Designs", Electronics: Deep Learning-Based Routing for Network-on-a-Chip (NoC): Opportunities, Challenges, and Solutions, 2020.

[5] Kennedy, J., Eberhart, R., "Particle Swarm Optimization", IEEE International Conference on Neural Networks, 1995.

[6] Paul, S. et al., "Adaptive Task Allocation and Scheduling on NoC based Multicore Platforms with Multitasking Processors", ACM Trans. on Embedded Computing Systems (TECS), 2020.