Projet

Analyse énergétique d'un processeur vectoriel pour des calculs de DNN
Rapport intermédiaire

ELE6307 - Machines neuronales : architectures et applications

Hiver 2022
Département de génie électrique
École Polytechnique de Montréal

Dernière mise à jour: 3 avril 2022

Yoan **Fournier** 1958736

Victor **Gaudreau-Blouin** 1958297

Introduction

Afin de modéliser l'efficacité énergétique du coprocesseur vectoriel ARA, le simulateur Timeloop a été utilisé. Ce simulateur utilise une structure hiérarchique pour modéliser une architecture. Une structure détaillant le problème à résoudre par l'architecture est aussi spécifiée. La commande *timeloop-mapper* permet ensuite de trouver automatiquement la correspondance idéale du problème sur l'architecture matérielle.

L'architecture simplifiée de ARA, comme mentionnée dans la proposition du projet, peut se modéliser en termes de *Lanes*. Chaque *Lanes* à sa plus simple expression consiste en une banque de 8 registres ainsi qu'une unité arithmétique capable d'effectuer des multiplications et des additions. Bien que sur ARA, l'unité d'addition est séparée de l'unité de multiplication, les deux unités peuvent fonctionner durant le même cycle sur des données indépendantes. Dans le cadre de la modélisation, on simplifie en utilisant la structure *intmac* fournie par l'outil Timeloop-Accelergy. Nous considérons que la différence causée par cette simplification est minimale. En effet, les opérations dans une Lane sont pipelinées et les opérations de multiplication et addition peuvent être exécutées en même temps. La seule différence principale est la latence causée par le pipelinage des opérations. Cependant, ce n'est pas une métrique de performance étudiée dans le cadre de ce projet.

L'architecture utilisée est présentée à la figure 1. On y observe une mémoire DRAM principale reliée à la mémoire cache L1 de 64 KB. La cache L1 est reliée à un nombre paramétrable de *Lanes*. Chaque Lane comporte sa banque de 8 registres et son unité de multiply-accumulate.

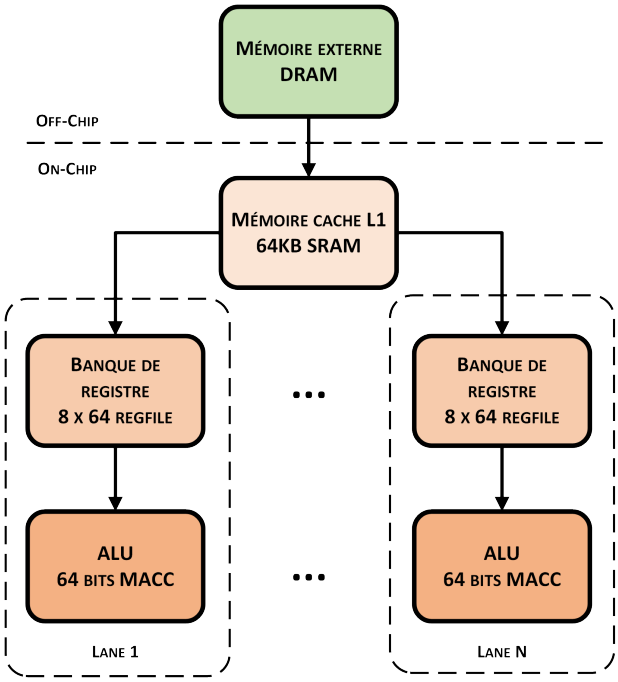

Figure 1 – Architecture de ARA sur Timeloop

Afin d'étudier la performance de l'architecture en fonction du nombre de *Lanes*, le problème utilisé est la première couche convolutionnelle du modèle VGG16. Cette couche comporte 64 filtres avec un *kernel* de 3×3. L'entrée de la couche est une image de dimension $224 \times 224 \times 3$.

Résultats

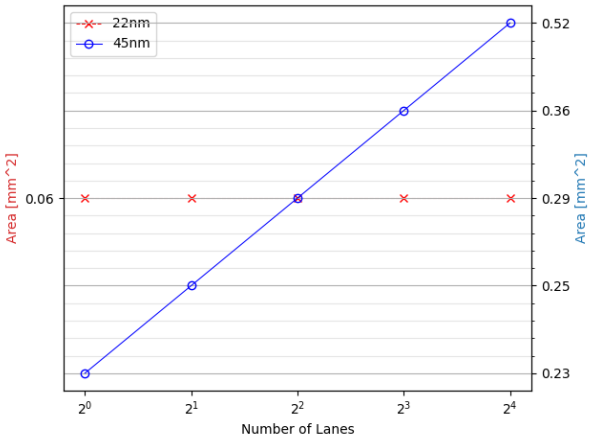

FIGURE 2 – Aire de ARA en fonction du nombre de *Lanes* pour la couche 1 de VGG16

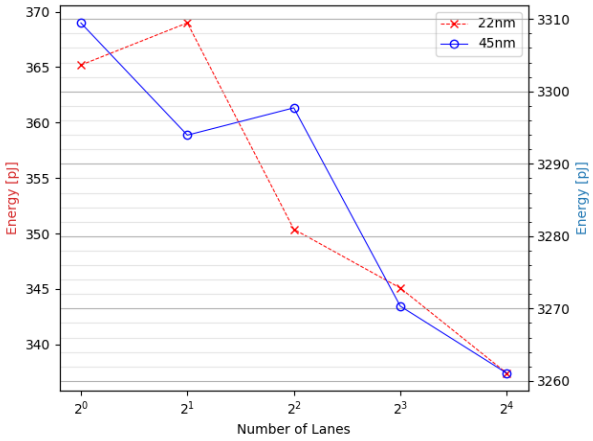

FIGURE 4 – Énergie de ARA en fonction du nombre de *Lanes* pour la couche 1 de VGG16

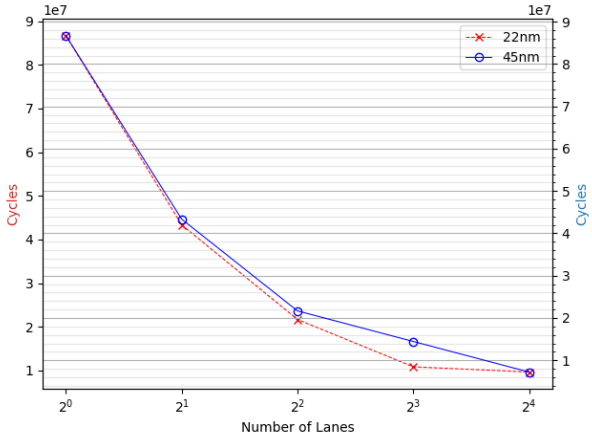

FIGURE 3 – Cycles de ARA en fonction du nombre de *Lanes* pour la couche 1 de VGG16

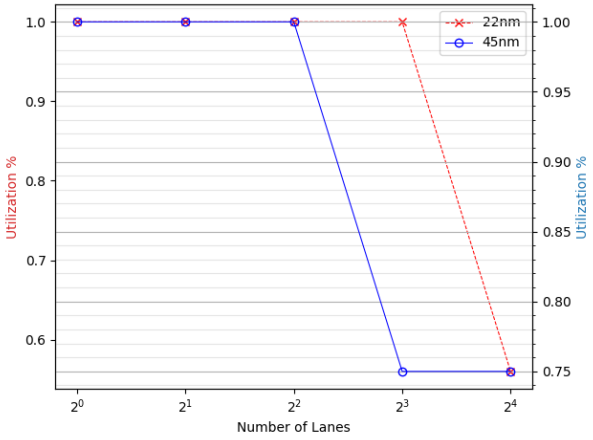

FIGURE 5 – Utilisation de ARA en fonction du nombre de *Lanes* pour la couche 1 de VGG16

Suite du projet