الجزائرية الديمقراطية الشعبية الجمهورية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة العليا للإعلام الآلي - 08 ماي 1945 – بسيدي بلعباس
Ecole Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbes

Mémoire de Fin d'étude

Pour l'obtention du diplôme d'**ingénieur d'état**

Filière : **Informatique**

Spécialité : **Ingénierie des Systèmes Informatiques (ISI)**

Thème

Etude de l'interférence et son impact sur les performances GPUs NVIDIA

Présenté par **:**

- Nesrine Roumaissa ROUHA

Soutenu le : **14/09/2025**

Devant le jury composé de :

- Dr. Bensenane Hamdane	Président
- Dr. Houssam-Eddine Zahaf	Encadrant
- Pr. BENSLIMANE Sidi Mohamed	Encadrant
- Dr. Amrane Abdelkader	Examinateur

Année Universitaire : 2024 / 2025

Abstract

This work investigates the temporal determinism of GPU execution in real-time contexts, focusing on interference and resource occupancy in convolutional neural network (CNN) inference. We propose an experimental pipeline that automatically generates parameterized CUDA kernels, executes them on the NVIDIA Jetson AGX Orin platform, and profiles their behavior using Nsight Compute. The approach combines analytic exploration of configurations with automated profiling to systematically capture execution time variability, occupancy limits, and memory hierarchy effects. A case study on the MNIST dataset demonstrates the feasibility of reproducing CNN inference directly in CUDA, validating the methodology and providing insights into the impact of tiling, shared memory usage, and resource contention on temporal predictability.

Keywords : CUDA, embedded GPU, temporal determinism, interferences, resource occupancy, automated profiling, Nsight Compute, kernel generation, convolution, reverse engineering, temporal variability, Jetson Orin, Deep Learning, MNIST.

Résumé

L'intégration des GPU dans des systèmes temps réel est freinée par la variabilité des temps d'exécution et les interférences entre kernels. Ce travail propose un pipeline expérimental complet, allant de la génération automatique de kernels CUDA à leur profilage systématique sur GPU embarqué (Jetson Orin).

Un générateur analytique, couplé à un noyau exécutable qui permet d'explorer des configurations compatibles avec les contraintes architecturales et de mesurer leur stabilité temporelle grâce à deux modes. L'orchestration automatisée, reposant sur *NVIDIA Nsight Compute*, assure la collecte de métriques architecturales et temporelles dans un cadre reproductible.

Un cas d'étude sur le dataset MNIST valide la méthodologie et démontre sa pertinence pour analyser les compromis entre performance, occupation des ressources et déterminisme temporel des kernels CUDA.

Mots-clés : CUDA, GPU embarqué, déterminisme temporel, interférences, occupation des ressources, profilage automatisé, Nsight Compute, génération de kernels, convolution, reverse engineering, variabilité temporelle, Jetson Orin, Deep Learning, MNIST.