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

MEMOIRE

En Vue de l'obtention du diplôme de **Master**

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 thesis is part of the Héritage project, which aims to ensure temporal determinism in heterogeneous real-time systems leveraging GPUs. The rise of Convolutional Neural Networks (CNNs) in critical applications, such as embedded vision and autonomous vehicles, makes it necessary to control execution time variability induced by hardware and software interferences.

After presenting the fundamental concepts of GPU architecture and its constraints in real-time environments, a comprehensive scientific review is conducted. This review analyzes several approaches from the literature, ranging from memory transfer optimization and spatial partitioning of computing units, to the management of CPU–GPU memory interference and asynchronous multi-stream execution. The studies are compared according to their efficiency, adaptability, and limitations in a mixed-criticality context.

The analysis reveals the absence of integrated solutions that combine isolation, dynamic scheduling, and precise control of interferences. These findings motivate the proposal, in the forthcoming engineering thesis, of an experimental methodology based on reverse engineering of CNN layers in CUDA, systematic profiling with Nsight Compute, and fine-grained evaluation of hardware impacts on temporal stability. The ultimate goal is to design optimization strategies that guarantee predictable execution on embedded GPUs

Keywords : real-time systems, Deep Learning, GPU, CUDA, temporal determinism, interference, occupancy, shared memory, Nsight Compute, CNN, MNIST.

Résumé

Ce mémoire s'inscrit dans le cadre du projet **Héritage**, qui vise à garantir le **déterminisme temporel** dans les systèmes temps réel hétérogènes exploitant des GPU. L'essor des réseaux de neurones convolutifs (CNN) dans les applications critiques, telles que la vision embarquée et les véhicules autonomes, impose de maîtriser les variabilités de temps d'exécution induites par les interférences matérielles et logicielles. Après avoir présenté les notions de base sur l'architecture GPU et ses contraintes en environnement temps réel, une **revue scientifique exhaustive** est menée. Celle-ci analyse plusieurs approches issues de la littérature, allant de l'optimisation des transferts mémoire et du partitionnement spatial des unités de calcul, à la gestion de l'interférence mémoire CPU–GPU et à l'exécution asynchrone multi-flux. Les travaux sont comparés selon leur efficacité, leur adaptabilité et leurs limites dans un contexte de criticité mixte. L'analyse révèle l'absence de solutions intégrées combinant isolation, ordonnancement dynamique et contrôle précis des interférences. Ces constats motivent la proposition, dans le cadre du mémoire d'ingénieur à venir, d'une méthodologie expérimentale basée sur le **reverse engineering** de couches CNN en CUDA, le profilage systématique par Nsight Compute et l'évaluation fine des impacts matériels sur la stabilité temporelle, en vue de concevoir des stratégies d'optimisation garantissant une exécution prédictible sur GPU embarqués.

Mots-clés : systèmes temps réel, Deep Learning, GPU, CUDA, déterminisme temporel, interférence, occupation, mémoire partagée, Nsight Compute, CNN, MNIST.