Tremplin Recherche 2025-2026

English version below

- **Titre du projet:** Exécution de réseaux de neurones sous contraintes temps réel sur une carte NVIDIA GPU
- **Laboratoire:** Laboratoire Gaspard Monge (LIGM,UMR 8049 CNRS)
- **Équipe**: Logiciel, Réseau, Temps réel (LRT)
- **Nom et adresse e-mail des tuteurs et tutrices:**
  - ► Yasmina ABDEDDAÏM, yasmina.abdeddaim@esiee.fr
  - ► Mourad DRIDI, mourad.dridi@esiee.fr
- **Filières visées :** Informatique, Datascience et intelligence artificielle, Systèmes embarqués, Artificial Intelligence and Cybersecurity.
- **Année d'étude :** E4 ou E5
- **Selon les résultats du tremplin recherche, le tremplin peut être poursuivi en stage E4 ou E5.**

Présentation générale du sujet : Un système critique est un système pour lequel une erreur dans le fonctionnement peut avoir des conséquences dramatiques. Ces systèmes nécessitent de pouvoir prouver formellement leur bon fonctionnent avant leur utilisation. Dans le cas où le système critique utilise des algorithmes d'Intelligence Artificielles (IA), par exemple un véhicule autonome qui utilise des réseaux de neurones afin de détecter des objets, prouver le fonctionnement correcte du système est une tâche complexe à cause du non déterministe de certains algorithmes d'IA et de la complexité des architectures GPU sur lesquelles sont souvent exécutés les algorithmes d'IA.

Durant ce tremplin recherche, nous nous intersaisons au problème d'implémentation d'algorithmes d'IA sur une architecture GPU NVIDIA. Plus précisément nous nous intéressons au problème d'ordonnancement temps réel de réseaux de neurones implémentés en langage CUDA sur une architecture GPU. Cette thématique de recherche émerge dans le domaine des systèmes critiques (exemples [1, 2, 3]) suite au besoin croissant d'assurer le bon fonctionnement de systèmes qui utilisent des algorithmes d'IA.

Objectif du projet : La théorie de l'ordonnancement temps réel propose des modèles et des algorithmes qui permettent de montrer

qu’un système vérifie dans le pire scenario d’exécution des contraintes temporelles strictes. Suite à un précédent tremplin recherche et un stage recherche, nous disposons d’un outils qui permet d’exécuter plusieurs réseaux de neurones (en parallèle à d’autres programmes) provenant du framework Darknet [4] qui doivent respecter des contrainte de temps. Cet outil permet de définir :

1. La partition du GPU sur laquelle est exécuté chaque réseau
2. La périodicité de chaque réseau
3. La priorité de chaque réseau

Nous souhaitons continuer ce travail en proposons différents algorithmes d’ordonnancement temps réel dans le but de trouver la stratégie qui nous permet de respecter les contraintes temps réel.

Les étapes du tremplin recherche sont :

1. Familiarisation avec l’outil proposé précédemment
2. Amélioration de certaines fonctionnalités
3. Proposition d’un modèle temps réel en se basant sur un ensemble d’expérimentation
4. Proposition de politiques d’ordonnancement temps réel et leur implémentation dans l’outil
5. Comparaison entre les résultats théoriques de l’ordonnancement temps réel et les résultats obtenus en pratique.

Bibliographie

[1] *Y. Abdeddaïm, M. Dridi, J.Dumont, Research directions for real-time implementation of AI algorithms, Real Time Syst. 61(2): 253-258 (2025)*

[2] A. Zou, J. Li, C. D. Gill, and X. Zhang, “RTGPU: real-time GPU scheduling of hard deadline parallel tasks with fine-grain utilization,” IEEE Trans. Parallel Distributed Syst., vol. 34, no. 5, pp. 1450–1465, 2023.

[3] J. Bakita and J. H. Anderson, “Hardware compute partitioning on nvidia

gpus,” in 2023 IEEE 29th Real-Time and Embedded Technology and Applications Symposium (RTAS), 2023, pp. 54–66

[4] https://pjreddie.com/darknet/

****

- **Project title:** Execution of neural networks under real-time constraints on an NVIDIA GPU card
- **Laboratory:** Gaspard Monge Laboratory (LIGM, UMR 8049 CNRS)
- **Team:** Software, Networks, Real Time (LRT)
- **Names and email addresses of supervisors:**
  - Yasmina ABDEDDAÏM, yasmina.abdeddaim@esiee.fr
  - Mourad DRIDI, mourad.dridi@esiee.fr
- **Targeted fields of study:** Computer Science, Data Science and Artificial Intelligence, Embedded Systems, Artificial Intelligence and Cybersecurity.
- **Year of study:** E4 or E5

Depending on the results of the research programme, the project may be continued as an E4 or E5 internship.

General overview of the topic: A critical system is a system in which an error can lead to dramatic consequences. These systems require formal proof of their proper functioning before use. In cases where the critical system uses artificial intelligence (AI) algorithms, for example an autonomous vehicle that uses neural networks to detect objects, proving that the system is functioning correctly is a complex task due to the non-deterministic nature of certain AI algorithms and the complexity of the GPU architectures on which AI algorithms are often executed. During this research programme, we focus on the problem of implemention of AI algorithms on an NVIDIA GPU architecture. More specifically, we are interested in the problem of real-time scheduling of neural networks implemented in CUDA on a GPU architecture. This research topic is emerging in the field of critical systems (examples [1, 2, 3]) due to the growing need to ensure the proper functioning of systems that use AI algorithms.

Project objective: Real-time scheduling theory proposes models and algorithms that demonstrate that a system verifies strict timing constraints in the worst-case execution scenario. Following a previous research programme and research internship, we have a tool that allows us to run several neural networks (in parallel with other programs) from the Darknet framework [4] that must satisfy tming constraints. This tool allows us to define:

1. The GPU partition on which each network is executed
2. The periodicity of each network
3. The priority of each network

We aim to continue this work by proposing different real-time scheduling algorithms with the aim of finding the strategy that allows us to meet real-time constraints. The stages of the research programme are:

1. Familiarization with the previously proposed tool
2. Improvement of certain functionalities
3. Proposal of a real-time model based on a set of experiments
4. Proposal of real-time scheduling policies and their implementation in the tool
5. Comparison between the theoretical results of real-time scheduling and the results obtained in practice.