RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

ONERA
THE FRENCH AEROSPACE LAB

www.onera.fr

PROPOSITION DE STAGE EN COURS D'ETUDES

Référence : **DTIS-2026-54** *(à rappeler dans toute correspondance)*	Lieu : Toulouse
Département/Dir./Serv. : DTIS/COVNI	Tél. : 05 62 25 22 11
Responsable(s) du stage : Benjamin Lesage, Cédric Cazanove	Email. : benjamin.lesage@onera.fr, cedric.cazanove@onera.fr

DESCRIPTION DU STAGE

Thématique(s) : Ingénierie des systèmes et des logiciels

Type de stage : ☒ Fin d'études bac+5 ☒ Master 2 ☐ Bac+2 à bac+4 ☐ Autres

Intitulé : Assessing the Benefits of Resource Management Techniques on High-Performance Embedded platforms

Sujet : Timeliness, ensuring computation results are available at the right time, is a crucial property for Critical Real-Time systems, to ensure the system will behave as expected. Timing and Schedulability analyses traditionally ensure the system components satisfy their temporal constraints. Those rely on some characterisation of the underlying platform, i.e. its behaviour under different workloads, and of the software under test (SUT), especially its worst-case execution times. Complex platform challenge characterisation efforts as they exhibit poor predictability; more complex hardware features, tailored towards improving the average execution time of SUT, tend to result in higher, hard to analyse, execution time variability.

High-Performance Computing (HPC) platforms offer hybrid architectures with general-purpose and specialised accelerators. GPU-based accelerators in particular, as found on the NVIDIA AGX Xavier, are well suited to supporting Machine Learning and Inference tasks. However, the embedding of such HPC platforms in Real-Time Systems raises many issues as they are tailored for speed and highly parallel workloads. Concurrent interfering workloads, competing for a small set of shared computing and memory resources, can induce great execution time variability which is hard to understand and bound.

Different approaches have been explored to understand and mitigate the impact of interferences for a SUT running on HPC platforms, i.e. to mitigate the impact of contending tasks on shared resources. Beyond classical Resource Management (RM) techniques, which often involve trade-offs in transparency, performance overhead, and applicability to embedded contexts, an emerging direction lies in platform modeling, data-driven and machine learning analysis. The internship will focus on exploring these methods by deploying microbenchmarks and workloads on a real GPU-based platform, collecting relevant performance metrics, and applying statistical and machine learning techniques to build models that reflect the platform's behavior under different conditions. The objective is to derive guidelines for supporting the integration of such complex architectures in critical real-time systems.

Reference:

Cédric Cazanove, Benjamin Lesage, Frédéric Boniol, and Jérôme Ermont. Assessing Unchecked Factors for Certification: An Experimental Approach for GPU Cache Parameters. WCET 2024.

Est-il possible d'envisager un travail en binôme ? Non

Méthodes à mettre en oeuvre :

☐ Recherche théorique	☐ Travail de synthèse
☒ Recherche appliquée	☐ Travail de documentation
☒ Recherche expérimentale	☐ Participation à une réalisation

Possibilité de prolongation en thèse : A renseigner

Durée du stage :	Minimum : 2 mois	Maximum : 5 mois
Période souhaitée : 5 mois		

PROFIL DU STAGIAIRE	
Connaissances et niveau requis : Hardware Architecture; Low level programming languages (C, C++); CUDA Framework (optional); Linux-based platforms; Task/Thread Scheduling; Real-Time systems' constraints; Embedded computing; Machine Learning; Modelisation	Ecoles ou établissements souhaités :