ONERA
THE FRENCH AEROSPACE LAB

www.onera.fr

PROPOSITION DE POST-DOCTORAT

Intitulé : Génération automatique de code pour l'algèbre linéaire rapide sur GPU

Référence : **PDOC-DAAA-2020-01**
(à rappeler dans toute correspondance)

Début du contrat : 01/01/2021	**Date limite de candidature :** 30/09/2020

Durée : 12 mois, éventuellement renouvelable une fois - Salaire net : environ 25 k€ annuel

Mots clés

Génération automatique de code, algèbre linéaire rapide, calcul sur GPU, CFD

Profil et compétences recherchées

Docteur en sciences dans le domaine de la simulation numérique ou des langages de programmation/compilation,

Expérience de la programmation parallèle en mémoire partagée sur CPU multi-cœurs ou GPU,

Compétences en algèbre linéaire numérique fortement appréciées.

Présentation du projet post-doctoral, contexte et objectif

L'ONERA travaille à la génération automatique de code pour un futur logiciel de simulation en mécanique des fluides [1]. Le but de ces travaux est en particulier de fournir des paradigmes de programmation de haut niveau évitant au développeur de réaliser lui-même les optimisations nécessaires pour obtenir des performances HPC élevées. L'ONERA cherche aussi à assurer automatiquement la portabilité vers différents langages et différentes architectures matérielles actuelles (CPU multi-coeurs, GPU) et futures.

Par ailleurs, Google AI travaille à une approche générique pour la génération de code qui consiste à utiliser une approche hiérarchique à plusieurs niveaux de représentation ("intermediate representation") [2]. L'une des ambitions de cette approche est d'obtenir un code optimal pour l'exécution sur différents types d'architectures matérielles.

Ce travail en collaboration entre l'ONERA et Google AI, visera à la résolution rapide de problèmes d'algèbre linéaire issus des équations de la mécanique des fluides en écoulements compressibles discrétisées sur des maillages structurés. Nous chercherons à générer automatiquement un code optimisé pour l'exécution sur GPU. Pour cela nous mettrons en place et étendrons potentiellement le formalisme du dialecte LinALG du framework MLIR [3, 4] sur des problèmes d'algèbre linéaire de difficulté croissante. Pour guider la génération automatique ultérieure, nous devrons réaliser une analyse des goulots d'étranglement à chaque niveau de parallélisme (warps, blocs, threads, vectorisation). Nous étudierons les décisions d'optimisation qui peuvent être prises pour générer du code, les choix des différents niveaux de parallélisme, les choix d'ordonnancement des boucles, les choix de tuilage (cache-blocking) ou de vectorisation de boucles. Après l'étude de cas simples, nous étendrons l'approche à la résolution par méthode de Gauss-Seidel (itération de l'application à un vecteur d'une décomposition LU de la matrice initiale), dans une approche de type wavefront [5] qui sera adaptée aux spécificités des GPU.

[1] B. Maugars, S. Bourasseau, C. Content————————————, B. Michel, B. Berthoul, P. Raud, L. Hascoët, Algorithmic Differentiation for an efficient CFD solver, soumis à Journal of Computational Physics
[2] https://pliss2019.github.io/albert_cohen_slides.pdf
[3] https://www.tensorflow.org/mlir
[4] Ulysse Beaugnon, Basile Clément, Nicolas Tollenaere, Albert Cohen, On the Representation of

Partially Specified Implementations and its Application to the Optimization of Linear Algebra Kernels on GPU, ArXiv https://arxiv.org/pdf/1904.03383

[5] R. Barrett, M.W. Berry, T.F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. Van der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, SIAM, Jan 1, 1994

Collaborations extérieures

INRIA, Google AI

Laboratoire d'accueil à l'ONERA

Département : DAAA

Lieu (centre ONERA) : Châtillon

Contact : Denis Gueyffier

Tél. : 01 46 73 37 43 Email : denis.gueyffier@onera.fr