

Proposition de sujet de stage de fin d'étude

Prototypage de stratégies de génération de code pour des outils de synthèse matérielle de haut niveau (HLS)

Encadrants: Christophe Alias et Matthieu Moy

Lieu: Laboratoire de l'Informatique du Parallélisme (LIP)
École Normale Supérieure de Lyon

Depuis le début des années 2000, la limite de miniaturisation des transistors force à multiplier les unités de calcul (processeurs, processeurs spécialisés) des superordinateurs pour améliorer les performances [3], ce qui augmente la consommation énergétique et donc le coût du calcul dans des proportions qui ne sont plus tenables. De nouveaux paradigmes de calcul basse consommation doivent être inventés. Une solution est de cabler directement les algorithmes principaux du calcul sur des circuits reconfigurables FPGA (Field Programmable Gate Array) [2]. Pour exécuter une application sur des FPGA, de nombreux verrous doivent être levés, dont la traduction automatique d'un algorithme de calcul en un circuit efficace (HLS, High-level synthesis).

La traduction d'un programme en circuit se fait en deux étapes. Tout d'abord, le *front-end* produit une représentation intermédiaire adaptée à la synthèse de circuit, typiquement un réseau de processus [1] qui capture le parallélisme de l'application et les communications entre unités parallèles. Ensuite, le *back-end* traduit chaque composant du réseau de processus en matériel en assurant une bonne réutilisation des ressources. Au final, le circuit produit peut être vu comme un gigantesque réseau de processus pipeliné, acceptant des entrées et produisant des sorties à intervalles réguliers.

L'équipe CASH nouvellement créée travaille sur des approches innovantes d'extraction du parallélisme vers une représentation intermédiaire. La génération de code finale vers le FPGA est déléguée à un outil de synthèse externe : nous utilisons donc des compilateurs dits « source-to-source », qui lisent du code C séquentiel, en extraient le parallélisme, puis génèrent du code qui explicite le parallélisme.

Beaucoup de variantes sont possibles au niveau de la génération de code C, et la manière de générer le code peut avoir un impact considérable sur les performances finales du circuit.

L'objectif de ce stage est d'expérimenter les stratégies de génération de code en écrivant du code C manuellement, ou en générant le code systématiquement à partir d'une représentation intermédiaire arbitraire pour déterminer la meilleure stratégie. En d'autres termes, il s'agit de prototyper plusieurs manières de générer le code et d'étudier les performances du circuit généré.

Plan indicatif

Voici une description plus précise des tâches attendues.

- 1. Prise en main d'un outil de synthèse de haut niveau (HLS) industriel.

- **2.** Écriture d'un programme parallèle flot de donnée en utilisant la variante de C utilisée en entrée de l'outil de HLS
- **3.** Étude des performances du circuit généré
- **4.** Modification du programme écrit en **2.** et étude de l'impact sur les performance **3.**.

Encadrement

Ce stage sera co-encadrée par Christophe Alias (CR1 Inria, ENS-Lyon) et Matthieu Moy (MCF HDR UCBL).

Christophe Alias (<http://perso.ens-lyon.fr/christophe.alias/>) s'intéresse à la synthèse de circuit haut niveau dans le modèle polyédrique depuis plus de 8 ans. Il a co-encadré deux thèses (Alexandru Plesco avec Alain Darté et Tanguy Risset, et Guillaume Iooss avec Sanjay Rajopadhye). Dans le même temps, il a écrit un compilateur de réseaux de processus, transféré dans la startup Xtremlogic qu'il a co-fondé en 2014 avec Alexandru Plesco. Actuellement Christophe Alias est en concours scientifique à 20% dans XtremLogic et n'a plus de charge d'encadrement depuis la soutenance de Guillaume Iooss en juillet 2016.

Matthieu Moy (<https://matthieu-moy.fr>) travaille sur la simulation en SystemC depuis une quinzaine d'années (en partenariat avec STMicroelectronics et en particulier dans le cadre de la HLS), et a déjà encadré plusieurs thèses et post-doctorants sur le sujet. Plus récemment, il s'est intéressé aux calculs de pire temps d'exécution de logiciel et de pire temps de traversée de réseaux sur puces dans le cadre de systèmes temps-réel critiques. Il est titulaire de l'habilitation à diriger des recherches depuis 2014. Anciennement responsable de l'équipe Synchrone du laboratoire Verimag, il a intégré le LIP en septembre 2017. Il co-encadre aujourd'hui 4 thèses.

Compétences souhaitées

Notions solides en parallélisme, notions en architecture des ordinateurs. Maîtrise du langage C bases solides en C++.

Références

- [1] Christophe Alias and Alexandru Plesco. Data-aware Process Networks. Research Report RR-8735, Inria - Research Centre Grenoble – Rhône-Alpes, June 2015.
- [2] Altera Corporation. Altera FPGAs achieve compelling performance-per-watt in cloud data center acceleration using CNN algorithms. <http://www.prnewswire.com/news-releases/altera-fpgas-achieve-compelling-performance-per-watt-in-cloud-data-center-acceleration-using.html>, 2015.
- [3] Nor Zaidi Haron and Said Hamdioui. Why is cmos scaling coming to an end? In *Design and Test Workshop, 2008. IDT 2008. 3rd International*, pages 98–103. IEEE, 2008.