Accélération matérielle FPGA sur plateforme embarquée appliquée aux réseaux de neurones

Lieu et contacts

Laboratoire TIMA, 46 avenue Félix Viallet, Grenoble, France.
Contact : Adrien Prost-Boucle, adrien.prost-boucle@univ-grenoble-alpes.fr
Contact : Frédéric Pétrot, frederic.petrot@univ-grenoble-alpes.fr

(Version française)

Mots-clés : *Système embarqué, Accélération Matérielle, FPGA, Réseaux de Neurones*

Contexte : Nous avons des accélérateurs matériels de réseaux de neurones pour circuits FPGA. Aujourd'hui, ils sont exécutés sur carte FPGA branchée en PCI-Express dans un ordinateur. Nous souhaitons étendre le support aux plateforme embarquées de type Zynq (technologie Xilinx, système CPU+FPGA sur une même puce), et à d'autres cartes qui ne supportent pas la connexion en PCI-Express.

Objectifs : Le support des technologies Zynq-7000 a été testé de façon préliminaire par des stagiaires précédents. Ce support doit être ajouté de façon plus solide dans nos outils générateurs de réseaux, étendu à des technologies de FPGA plus récentes et plus conséquentes (Zynq Ultrascale), et évalué par des campagnes de test en conditions réelles avec un ou plusieurs benchmarks connus.

Le stage suivra principalement les étapes suivantes :

— ajouter le support des accélérateurs matériels en FPGA embarqués sur la même puce, pour les puces Zynq,
— lancer des campagnes d'exécution des réseaux de neurones sur différentes cartes avec puces Zynq (Zybo, ZC706, ZCU102),
— mettre en place une application de démonstration sur carte avec reconnaissance d'images en temps réel avec webcam, sous système Petalinux.

Dans un second temps, selon le candidat et le temps disponible, les objectifs seront étendus aux items suivants :

— d'une part avec le support de la suite de synthèse logique libre (logiciel Yosys), pour pouvoir synthétser des accélérateurs matériels sur la carte elle-même,
— et d'autre part avec la mise en place un environnement d'exécution de réseaux pour des cartes FPGA plus grosses, notamment avec communications via Ethernet ou via une version plus récente du PCI-Express.

Prérequis :

— Maîtrise de la programmation C/C++
— Solides connaissances en architectures de circuits numériques
— Des bases en langage VHDL
— Une première expérience avec les outils Xilinx Vivado/Vitis serait un plus
— Une première expérience en logiciel *bare metal* serait un plus

Références :

— High-Efficiency Convolutional Ternary Neural Networks with Custom Adder Trees and Weight Compression (2019)
`https://cnrs.hal.science/hal-01686718v2`

(English version)

Title : FPGA hardware acceleration on embedded platform applied to neural networks

Keywords : *Embedded System, Hardware Acceleration, FPGA, Neural Networks*

Context : We have hardware accelerators of neural networks for FPGA hardware targets. Nowadays, these accelerators are executed on PCI-Express boards connected within a workstation. We want to extend the support to embedded platforms such as Zynq chips (Xilinx technology, CPU+FPGA system on a chip), and to other boards that do not support PCI-Express connexion.

Objectives : The support of Zynq-7000 targets has been experimented by previous interns. This support has to be integrated within our tools to generate and operate neural networks, extended to more recent and more powerful FPGA technologies (Zynq Ultrascale), and benchmarked with test campaigns in real conditions with one or several known benchmarks.

The internship will follow the following steps :

— add support for hardware accelerators in embedded FPGA on same chip, for Zynq chips,
— launch execution campaigns of neural networks on different FPGA boards with Zynq chips (Zybo, ZC706, ZCU102),
— bring up a demonstration application on board for image recognition in real time from avec webcam, under Petalinux system.

In a second step, depending on the candidate and available time, objectives will be extended to following items :

— support the open-source synthesis toolsuite (tools Yosys and nextpnr), in order to synthesize hardware accelerators on the board itself,
— prepare hardware context to target larger FPGA boards, in particular with Ethernet interfaces or more recent PCI-Express version.

Prerequisites :

— Mastering of programming with C/C++
— Solid knowledge of digital circuit architectures
— Basic knowledge of VHDL language
— Experience with Xilinx tools Vivado/Vitis would be a plus
— Experience with embedded platforms and drivers would be a plus

References :

— High-Efficiency Convolutional Ternary Neural Networks with Custom Adder Trees and Weight Compression (2019)
`https://cnrs.hal.science/hal-01686718v2`