Integrated Programmable-Array accelerator to design heterogeneous ultra-low power manycore architectures

Laboratoires : Université de Bretagne Sud/Lab-STICC (UMR 6285, CNRS), Lorient, France
University of Bologna / Micrel Lab., Bologna, Italy

Context

Today's technological advances allow us to produce very complex multi-core architectures containing hundreds of processors. However, programming infrastructure does not evolve at the pace demanded by technological advances and market pressure. It is still necessary to find new techniques, new architectures and new tools to help designers to efficiently implement complex applications on sophisticated platforms and make use of the underlying hardware. Moreover, in order to combine the ever increasing performance requirements with an extremely tight energy budget, systems are moving towards heterogeneous architectures as the main design paradigm. In this context, designers use hardware accelerators. The objectives of this work are: 1) to explore heterogeneous many-cores architectures integrating reconfigurable hardware accelerators [5]; 2) develop associated programming models to address the growing complexity of application development. The proposed approach will allow programmers to easily deploy applications on dynamically reconfigurable heterogeneous many-cores architectures. In this context, an OpenMP-based programming model and a many-core architecture model integrating a CGRA (Coarse Grained Reconfigurable Array) [1] [2] [3] [4] [5] will be defined. An automated design flow and HW / SW module for dynamic reconfiguration of the CGRA will be provided. The results will be validated on a virtual platform and a hardware prototype, using signal processing and image processing applications.

The main block of the targeted many-core architectures is a multi-core cluster containing strongly coupled shared memories. Notable examples are the STHORM architecture of ST, Kalray MPPA, Plurality HAL, Adapteva Epiphany, or GPUs like Fermi from Nvidia. This type of cluster allows to combine short latency communication and high bandwidth between a certain number of cores (typically 16). Replicating clusters and interconnecting them hierarchically across a network-on-a-chip can scale to a large number of cores. As an example, the STHORM architecture has 4 clusters and 69 processors, the Kalray MPPA architecture has 16 clusters and 256 processors.

The benefits of hardware acceleration of critical kernels for a given application domain are known. It is therefore necessary to study the evolution of these clusters in terms of heterogeneity. The two key elements of the work we propose are an integration of a programmable accelerator coupled to the multi-core and a shared-memory communication scheme. Several years of multi-core programming have provided many parallel applications, based on abstract, standard and portable programming models (e.g. OpenMP, OpenCL). It is therefore important to investigate new approaches to hardware acceleration that are consistent with multi-core programming models.

From a programming point of view, in the same way that "threads" are a good abstraction of the processor in most parallel programming models, the hardware accelerators here will be abstracted in the form of hardware tasks, an important step to simplify the development of applications for multi-core architectures with hardware accelerators.

A third key element in the novelty of the proposed approach is the integration of reconfigurable accelerators into our highly coupled shared memory cluster. The use of a CGRA will allow to reconfigure an accelerator to satisfy the needs of an application, combining several hardware tasks with a low cost reconfiguration.

PhD. Program.

The work program will begin with an in-depth study of the theoretical and technical aspects covered by the work of the thesis. First, a rigorous review of the state of the art will make possible to apprehend the existing CGRAs both from the hardware point of view and from the point of view of the associated software tools. Université de Bretagne Sud/Lab-STICC has the skills and expertise in this field. Reuse of existing design methodologies and associated software tools will be preferred. In a second step, many-core architectures and their associated programming models will be studied. The

expertise of University of Bologna in this field will be especially useful for carrying out this mission. In a third phase, the infrastructure resulting from a previous collaboration between the Université de Bretagne Sud/Lab-STICC and the University of Bologna to combine many-cores architectures and integrated hardware accelerators will serve as a starting point for the work plan [6].

The second major step of the thesis work will be to specify this new paradigm that combines many-cores and CGRA. A dedicated architecture will be defined to allow efficient coupling between the many-cores part and the CGRA. A design flow will be proposed and developed to facilitate the programming approach of this type of complex platform. This flow will include the programming model, the compilation tool chain and the HW/SW management system to concretely execute an application on the target platform. Mechanisms for the on-the-fly management of hardware tasks on the accelerator will also be explored. Special attention will be paid to the dynamic reconfiguration of the CGRA through the design of a specific module. The results in terms of performance, area and energy consumption will be studied.

Références

[1] Peyret, T. « Architecture matérielle et flot de programmation associé pour la conception de systèmes numériques tolérants aux fautes », Thèse de Doctorat, Université de Bretagne Sud, Dec. 2014

[2] Peyret, T., Corre, G., Thevenin, M., Martin, K., Coussy, P., "Efficient application mapping on CGRAs based on backward simultaneous scheduling/ binding and dynamic graph transformations", In 2014 IEEE 25th Int. Conf. on Application-Specific Systems, Architectures and Processors ASAP'14

[3] Peyret, T., Corre, G., Thevenin, M., Martin, K., Coussy, P., "An Automated Design Approach to Map Applications on CGRAs", In Great Lakes Symposium on VLSI GLSVLSI'14

[4] S. Das, K. Martin, P. Coussy, D. Rossi, and L. Benini. "Efficient mapping of CDFG onto coarse-grained reconfigurable array architectures", In 22nd Asia and South Pacific Design Automation Conference, 2017.

[5] S. Das, T. Peyret, K. Martin, G. Corre, M. Thevenin, and P. Coussy. "A scalable design approach to efficiently map applications on cgras", In 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pages 655–660, July 2016.

[6] Satyajit Das, Kevin J. M. Martin, Philippe Coussy, Davide Rossi, Luca Benini. "A 142MOPS/mW Integrated Programmable Array Accelerator for Smart Visual Processing", Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), 2017

Contact UBS / Lab-STICC, France	**Contact University of Bologna, Italy**
Prof. Philippe COUSSY	Prof. Luca BENINI
philippe.coussy@univ-ubs.fr	luca.benini@unibo.it