

Poster abstract: Fast and Energy-driven Design Space Exploration for Heterogeneous Architectures

Baptiste Roux, Matthieu Gautier, Olivier Sentieys, Jean-Philippe Delahaye

▶ **To cite this version:**

Baptiste Roux, Matthieu Gautier, Olivier Sentieys, Jean-Philippe Delahaye. Poster abstract: Fast and Energy-driven Design Space Exploration for Heterogeneous Architectures. FCCM 2018 - 26th IEEE International Symposium on Field-Programmable Custom Computing Machines, Apr 2017, Napa, United States. hal-01809560

HAL Id: hal-01809560
https://hal.science/hal-01809560

Submitted on 6 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines

Fast and Energy-driven Design Space Exploration for Heterogeneous Architectures

Baptiste Roux*, Matthieu Gautier†, Olivier Sentieys*, Jean-Philippe Delahaye‡
*Inria, Irisa, University of Rennes 1, France
†University of Rennes 1, Irisa, Inria, France
‡DGA MI, French MoD, France
Email: baptiste.roux@inria.fr

I. Introduction

In the last years, the integration of specialized hardware accelerators in Multiprocessor System-on-Chip (MpSoC) led to a new kind of architectures combining both software (SW) and hardware (HW) computational resources. For these new Heterogeneous MpSoC (HMpSoC) architectures, performance and energy consumption depend on a large set of parameters such as the HW/SW partitioning, the type of HW implementation or the communication cost. Design Space Exploration (DSE) consists in adjusting these parameters while monitoring a set of metrics (execution time, power, energy efficiency) to find the best mapping of the application on the targeted architecture. With the shift from performance-aware to energy-aware designs, DSE frameworks started to integrate state-of-the-art power models. These power modeling tools require simulations of the application, which drastically increases the exploration time. This work introduces a DSE method based on an analytical power model to circumvent the computation time bottleneck of state-of-the-art power models.

II. Approach

The proposed DSE, depicted in Fig. 1, proposes to optimize the HW/SW partitioning and mapping under user-defined objectives, especially an energy constraint. The flow targets tiling-based parallel applications and relies on an analytical power model that provides the DSE framework with the execution time and energy of a HW/SW configuration. The power model parameters are obtained with the measurements of a tiny subset of the design space, which are then injected into two extraction functions to obtain analytical formulations of the execution time and the energy consumption of the computation kernel. The partitioning problem constraints are defined as a set of inequalities with Boolean, integer (discrete) and non-integer (continuous) variables within a Mixed Integer Linear Programming (MILP) framework. Then, solutions can be efficiently determined using commercial or open source solvers. The optimization is defined with an linear objective function. The intersection of the inequality constraints represents a polyhedron of the feasible solution. The objective function defines a direction into the solution space and the optimal solution is found at the intersection between the objective function and the feasible solutions. The goal of the proposed DSE method is to find the best configuration that minimizes the user objective (e.g. execution time or total energy consumption of the application).

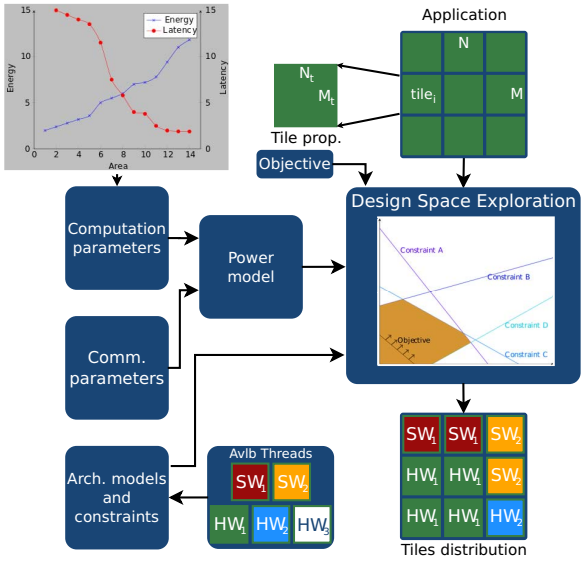


Fig. 1: Overview of the proposed DSE.

III. Results

This methodology was tested on two application kernels: a matrix multiplication and a Stencil computation. The Zynq-based heterogeneous architecture was used for the experimentation. The communication parameters were extracted with μBenchmark methodology[1]. The computation parameters were extracted with a Least Squared Root (LSR) algorithm from measurements of a subset of the configurations.

These experiments show that the acceleration and energy saving obtained are at minimum of 12 % compared to the full HW approach. Furthermore, the MILP resolution time takes less than 0.6 sec on an *intel i7 Haswell-ult* processor running at 2.10 GHz and the estimation accuracy is within 5 % to 10 %, which is quite acceptable for real hardware measurements. These results open new opportunity for future computer-aided design tools. Such method could be included in a complete framework with a multi-step exploration to accelerate every computation kernel within an application and to obtain an energy-efficient mapping of a full application on heterogeneous multiprocessor architecture.

[1]B. Roux et al. "Communication-Based Power Modelling for Heterogeneous Multiprocessor Architectures". In: *IEEE 10th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (mcSoC)*. Sept. 2016, pp. 209–216. DOI: 10.1109/MCSoC.2016.27.

978-1-5386-4037-1/17 $31.00 © 2017 IEEE
DOI 10.1109/FCCM.2017.31

IEEE computer society