Integer Linear Programming for Design Space Exploration in Heterogenous MPSoC

Dammak Bouthaina*, Mouna Baklouti†, Smail Niar* and Mohamed Abid†
*LAMIH, University of Valenciennes, France
Email: {bouthaina.damak, smail.niar}@univ-valenciennes.fr
†ENIS, Sfax, Tunisia
Email: {mouna.baklouti, mohamed.abid}@enis.rnu.tn

I. Introduction

In today's embedded systems market there is a trend towards integrating more and more services. A modern MultiProcessor System-On-Chip (MPSOC) may run on the same time a large number of applications from various domains. To full fill the consumer expectation, Heterogeneous MPSOC (Ht-MPSOC) has become a promising solution for modern embedded systems, exploiting the integration capacity provided by semiconductor technology and, at the same time, meeting the applications performance-constraint. The current trend towards Ht-MPSOC design is the use of application-specific instructions which provides a parallel execution to further enhance the total computational throughput of the system. In typical modern Ht-MPSOC, a large number of applications are running simultaneously, and different compute intensive kernels (called tasks) can be extracted to be implemented as specific hardware accelerators and invoked by application specific instructions. These dedicated hardware accelerators incur area and power overheads. Thus, as the total available area for implementation of specific hardware accelerator is limited, designer may not be able to exploit the full potential of all the hardware accelerators for the running applications. To tackle this problem, traditional solution propose the utilization of a larger FPGA without considering the effect of static energy dissipation.

In this context, we propose the accelerator's sharing technique that allows a better hardware utilization. Mobile and embedded applications contain a large portion of very similar kernels of code, like matrix multiplication and convolution operations. This offers a range of possible configurations that share the hardware accelerators among the processors. This means that different computational tasks can be implemented on the same hardware accelerator. This technique reduces area and power consumption and preserves performance. The fabric sharing comes at the cost of possible time overheads due to possible conflicts between cores. The designer has to strike the right balance between the number of private and shared accelerators in one side and the accelerators sharing-degree on the other side. This leads to a large architectural space exploration. In other words, our space exploration needs to consider a) the way to implement each computational task (software or hardware) and b) the sharing-degree that denotes the number of processors sharing a hardware accelerator.

For this purpose, we develop analytical model to estimate FPGA area consumption and execution time. Based on this analytical model, we propose an ILP model to identify the configurations of the different hardware accelerators leading to the optimal architecture in terms of FPGA area utilization and execution time.

II. Sharing Technique Approach

Because of logic-area cost, an MPSOC on which N applications are running on the different processors has a less opportunity to implement all the application-specific instructions. Recent application tasks are based on same frequently used kernel operations such as matrix multiplication and convolution operations. For an Ht-MPSOC, without considering the hardware-sharing of similar kernels which are executed on different processors, custom accelerators will be implemented for different custom instructions that provide the same computations. The proposed sharing approach offers a range of possible specific instructions shared among different tasks. This means that different computational tasks can share several accelerators. Figure 1 summarizes the proposed sharing approach. App0, App1 and App2 have same heavy computational tasks that can be executed simultaneously on three private accelerators associated to three application-specific instructions. However based on our proposed technique of sharing accelerators, only one accelerator is implemented and shared among P0, P1 and P2. The sharing degree defines the number of processors issuing the same accelerator.

III. Linear Programming Approach

Te proposed model explores the possible accelerator configurations of Ht-MPSOC architecture associated to the different application-specific instructions and select the optimal configurations leading to a minimum area consumption and respecting a required time constraint.
We define $T = (T_1...T_i...T_p)$ the sequence of most computational tasks executed on one or different processors and n_i denotes the number of applications that contain the task T_i. The different T_i tasks are candidates for hardware implementations and are expected to provide the desired performance. We denote $(C_i^0...C_i^{n_i})$ the set of the possible implementations

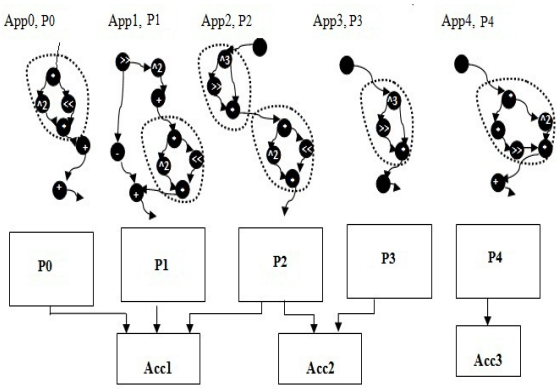

Fig. 1. Sharing Technique for HT-MPSOC Architecture

of Ti. C_i^0 corresponds to the software implementation and $C_i^{n_i}$ corresponds to the implementation with n_i private hardware accelerators. For $C_i^j, j \in \{1..n_i - 1\}$, j hardware accelerator occurrences are implemented on the FPGA and shared for the n_i processors that have T_i in their programs. The goal of the ILP model is to minimize the Ht-MPSOC area objective function considering a performance constraint. The area objective function is as follow

$$Area = \sum_{i=1}^{p} \sum_{j=0}^{n_i} (x_i^j * a_i^j) \tag{1}$$

Where x_i^j is a binary variable set to 1 if the configuration j of the pattern T_i is chosen to implement the corresponding accelerator. For a pattern i, only one configuration is selected. This constraint is expressed as follow: Each processor has a required performance $TLimit_i$ imposed to the objective function and is expressed as follow:

$$\forall i \in \{1..p\}, \sum_{j=1}^{n_i} x_i^j = 1 \tag{2}$$

$$t_i \leqslant TLimit_i, \forall i \in \{1..n\} \tag{3}$$

t_i defines the execution time of processor i and is expressed as follow:

$$t_i = t_i^0 - \sum_{k=1}^{p} \sum_{l=0}^{n_k} (x_k^l * s_i^k * R_k^l) \tag{4}$$

TheR_k^l is the execution-time-reduction when implementing the task T_k with configuration j. The presented equations formulate the configuration of the different patterns into an ILP problem. The implementation of each accelerator depends on the area/performance cost. The area cost is modelled in the objective function and the required performance is imposed as a constraint. Our model explores all the possible accelerator configuration to obtain the optimal architecture.

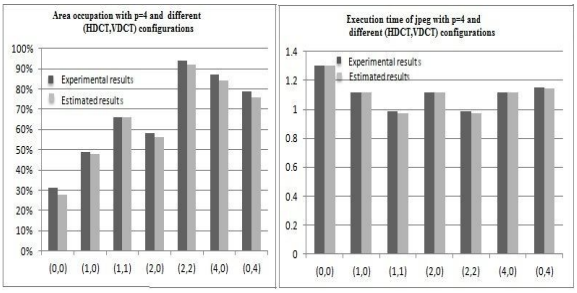

Fig. 2. Implementation and estimated results for the jpeg application, for different (HDCT,VDCT) sharing degree

IV. Results

We perform our experimental evaluation on the Xilinx platform and the jpeg encoder application. First, based on the profiling results, we extract the computational tasks (HDCT and VDCT) to be implemented as accelerators. The extracted patterns are synthesized and the area and execution time of the different configurations are estimated based on proposed analytical equations. The estimated area and execution time informations are inserted into our ILP model to generate the optimal architecture. Secondly, we verified the accuracy of the estimated area and execution time by implementing the different architecture configurations on the XILINX platform and evaluating their performances. Our experiments are based on the microblaze processors running at 125 MHz frequency that execute the jpeg application to encode bmp images. we compare in figure 2 the obtained experimental results to the estimated ones. The area consumption and execution time estimation accuracies are respectively about 3% and 1%. Hence, we can rely on our model to explore more complex architectures.

V. Conclusion

In our work, we propose propose a new ILP model for optimizing resource allocation of FPGA-based Ht-MPSOC accelerators satisfying performance constraints. The proposed model explores the private and shared accelerator design space configurations. The aim here is to identify the configurations that minimizes FPGA resources by the utilization of shared accelerators. Future research will focus on several areas. First, we plan to apply the same methodology to explore more applications and more complex architectures. Second, we plan to extend our problem formulation, to model non-symmetric Ht-MPSoC.