



HAL
open science

An Energy-Aware Scheduler for Dynamically Reconfigurable Multi-Core Systems

Robin Bonamy, Sébastien Bilavarn, Fabrice Muller

► **To cite this version:**

Robin Bonamy, Sébastien Bilavarn, Fabrice Muller. An Energy-Aware Scheduler for Dynamically Reconfigurable Multi-Core Systems. 6th International Workshop on Reconfigurable Communication-centric Systems-on-Chip, ReCoSoC 2015, Jun 2015, Bremen, Germany. pp.1-6, 10.1109/ReCoSoC.2015.7238084 . hal-01192796

HAL Id: hal-01192796

<https://hal.archives-ouvertes.fr/hal-01192796>

Submitted on 14 Mar 2016

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.

An Energy-Aware Scheduler for Dynamically Reconfigurable Multi-Core Systems

Robin Bonamy, Sébastien Bilavarn, Fabrice Muller
LEAT, University of Nice Sophia Antipolis, CNRS, France
robin.bonamy@unice.fr, sebastien.bilavarn@unice.fr, fabrice.muller@unice.fr

Abstract—This paper describes an energy-aware scheduling approach intended for use in heterogeneous multiprocessors supporting hardware acceleration with Dynamic and Partial Reconfiguration. Scheduler decisions rely on pragmatic power and energy models to map the load across cores and reconfigurable regions with regards to the actual power costs. Results on a multithreaded H.264/AVC profile decoder with three possible hardware functions on a Xilinx Zynq based platform report energy gains up to 44.1% over full software execution and 49.6% over static hardware / software execution, while ensuring real-time decoding requirement.

I. INTRODUCTION

The use of heterogeneous multiprocessor System-on-Chips has increased because of their potential to address energy efficiency, power and heat density problems. Despite the promises, the increasing level of heterogeneity greatly affects the design and mapping complexity in upcoming systems and applications. Indeed it is no longer simply a question of mapping efficiently concurrent processes to the best cores, but also to consider dynamic aspects such as power management or hardware acceleration and their impact on energy efficiency.

Hardware acceleration is a relatively well-known player in the energy performance equation which is regaining attention today with the advent of Dynamic and Partial Reconfiguration (DPR). Partial reconfiguration is a technique related to FPGAs that can be used to extend their inherent flexibility: it allows to reprogram specific regions with new functionality while other regions continue to run. Drastic reduction of hardware resource utilization results in less static power thus significant better energy efficiency, adding to the inherent benefits of dedicated hardware. However, a variety of parameters such as FPGA partitioning, accelerator parallelism or software execution strongly affect the actual processing efficiency, and other techniques such as blanking or DPR based clock gating can also be used to further decrease power. As a result, the quantity and scope of decisions in a dynamically reconfigurable multi-core system highly complexifies the scheduler's job. It is however critical to provide good support at this level since bad decisions can affect energy efficiency to the point of total ineffectiveness. We address in the following the definition and evaluation of an energy-aware scheduling and mapping approach for multiprocessor systems supporting hardware acceleration with DPR, and report representative application results and achievements that are denoting promising prospects in this field.

The outline of the paper is the following. We first review existing works in the field of heterogeneous multi-core scheduling, pointing out relative novelty in the use of DPR for that matter. In section 3, we introduce pragmatic power and energy models underlying the proposed scheduling procedure. We then describe the decision schemes used in the energy-aware scheduling and mapping process. Detailed results on a multithreaded H.264/AVC decoder on a Xilinx Zynq based platform are analyzed and discussed. Finally, main conclusions from these results are presented and future directions for extension and exploitation are proposed.

II. OVERVIEW OF HETEROGENEOUS MULTIPROCESSOR STRATEGIES

With the rise of multi-core heterogeneous architectures, there has been many works addressing efficient scheduling application workloads with multiple cores, possibly of different types. Early investigations focused primarily on improving load balancing to achieve better performance (throughput, instructions-per-cycle, etc.). Recently the challenge of heterogeneous thread scheduling and global power management switched on delivering higher power-performance levels (performance per Watt). Many works explored the energy efficiency benefits of heterogeneity, that are more extensively discussed for instance in [1]. There is a broad consensus on the fact that exploiting heterogeneity is essential for a better use of energy, the necessary counterpart being that it greatly complexifies the scheduler.

Of the work addressing this problem, [2] proposed a scheduler for a system of processors based on execution prediction to map future processing needs to the most suited processor. Their method applies to single-ISA heterogeneity supporting differing voltages and frequencies. [3] extended a symbiotic scheduling heuristic [4], originally developed to enhance throughput and lower response time, for chip multiprocessors with simultaneous multithreading cores. They report up to 7.4% savings in energy, 10.3% savings in energy-delay product, and 35% savings in power. [5] is another contribution considering both scheduling and power management to address process variations in CMPs. They conducted a design exploration, proposed a number of schedulers to satisfy different objectives, and developed a linear programming solution for power management. The authors in [6] examined the scalability problem for manycores, comparing basic scheduling heuristics and proposed scheduling and power management algorithms for heterogeneous systems scaling up to 256 cores. A recent study [7] addressed a study including the ARM big.LITTLE architecture, associating an energy

efficient processor cluster (Cortex-A7) with a higher performance processor (Cortex-A15). They report that different class of resource allocation heuristics (race-to-idle vs. never-idle) have very different results on different platforms, indicating that the efficiency of a strategy greatly depends on platform characteristics and can go as far as getting inefficient in some cases.

Aside from the question of core specialization, heterogeneity also extends to the aspects arising at run-time due for example to power management [8]. The dynamic use of different power states (P-states, sleep states) available for each core matches the problem of low power scheduling which had a long history of research over the last 20 years. Surveys have been described for example in [9] and [10] from the abundant literature. Our own previous experimental studies in this field [11] came to similar conclusions as [7] and [12] concerning the importance of platform characteristics and application knowledge on the efficiency of a strategy. Likewise, results indicate that custom strategies, i.e. more specialized schedulers dedicated to an application or application domain, can reach significant levels of energy gains (in the 5% to 50% range for video processing applications on representative platforms) compared to existing OS and platform-based strategies.

From this large literature, existing works often consider a specific perspective on the heterogeneous scheduling problem: a type of architecture, one precise objective (manycore scaling, process variability), limited heterogeneity (similar cores, DVFS), etc. In addition, a type of heterogeneity remains uncovered regarding the recent advancement of graphic and reconfigurable processing units: the possibility of hardware (or accelerated) execution of tasks. [1] is thus extending heterogeneity to hardware acceleration enabled with the progresses of Dynamic and Partial Reconfiguration (DPR) and High-Level Synthesis (HLS). Hardware acceleration delivers orders of magnitude performance and energy efficiency compared to software execution, and the flexibility introduced with DPR can improve these benefits. The underlying heterogeneous mapping and scheduling problem have started lately to be re-investigated and this paper extends the conclusions of [13] on the definition of low power scheduling policies able to support efficient execution of dynamic hardware and software tasks.

III. POWER MODELING

The remainder of this paper addresses a scheduling method capable of taking relevant run-time decisions to reduce the energy use while satisfying performance constraints. As such it requires that the power impact of resource allocation decisions can be realistically predicted. This section describes the underlying formal characterization, which results in great part from previous work reported in [13], and is divided in two main components: a platform and an application model.

A. Platform model

The platform is described by enumerating the execution units (EU) available and providing information on the corresponding power characteristics. To cope with the target reconfigurable multi-core platform, two main execution units are considered : a) processor cores and b) reconfigurable regions (RR).

The power model used for CPU cores is based on the type of core and is defined by the related static power $P_{core_j}^{static}$, the idle power $P_{core_j}^{idle}$ and the run power $P_{core_j}^{run}$, where j is the id of the execution unit.

As a RR needs to be configured before a task can run, which takes time and power that has also to be assessed, the reconfiguration controller can be modeled by its power P^{reconf} . In turn, the reconfiguration time depends on the area of the RR to configure ($T_{RR_j}^{reconf}$).

The power model for the RRs is defined by the static power consumption of the resources (e.g. slices, BRAM, clock tree, etc) in the given RR ($P_{RR_j}^{static}$). The related idle and run power depend on the actual task to be configured on this region. So the corresponding idle and run power are not defined in the platform model, instead they are characterized in the following application model.

B. Application model

An application is characterized by a set of tasks (\mathcal{G}) with their data and execution dependencies modeled by a task flow graph. Each task can have one or several implementations available, at least every task has a software implementation. An implementation is a description of how the task is being executed on a software or hardware EU. The model of an implementation reflects the task id (i), the EU id (j), execution time ($T_{i,j}$), and idle / run power consumption for hardware execution (P_{i,RR_j}^{idle} , P_{i,RR_j}^{run}). In case of software execution, the task idle / run power numbers are derived from the core idle and run power described in the platform model.

IV. SCHEDULER SPECIFICATION

The scheduling procedure involves two steps: a waiting task list is first determined and sorted according to a specific strategy (e.g. Earliest Deadline First), then a task implementation is mapped on a given execution unit according to an energy efficiency criteria. In the following, this process is further referred to as Energy Aware Heterogeneous Scheduler (EAHS).

A. Task Scheduling

Most common task scheduling strategies are a) FIFO (First In First Out) the first task arriving in the waiting state will be the first scheduled, b) EDF (Earliest Deadline First) which consists of sorting tasks to first schedule the one that must finish first and finally c) priority sort, a priority is assigned to each task and the highest priority task is executed first.

However these strict definitions may be alleviated in case of heterogeneous scheduling. Considering a scenario where the top task in the waiting list is blocked, waiting for matching execution unit to be free, then all other tasks in the waiting list are blocked despite the fact that they may be run on other free EUs. This will increase the application delay and execution units under-utilization. Thus we propose a *soft* scheduling approach which consists of using one of the previously mentioned strategies, but letting the possibility to bypass a task blocked because a corresponding mapping is not possible for the moment (e.g. either the reconfiguration controller is busy or compatible execution unit(s) not free).

B. Task Mapping

When managing task allocation on the architecture resources, the choice of an implementation has a major impact on reducing energy consumption. A cost function is thus defined to help identifying the most energy efficient mapping from the different possibilities. This cost is computed each time a task has to be mapped, for all possible implementations, which is based on energy and execution time estimations of the implementation being processed.

Energy estimation is given in equation (1):

$$\forall j ; E_j^{cost} = \begin{cases} E_{i,j}^{run} & \text{when } \rho_{i,j} = 0 \\ E_{i,j}^{run} + E_j^{conf} & \text{when } \rho_{i,j} = 1. \end{cases} \quad (1)$$

where

$$\begin{aligned} E_{i,j}^{run} &= P_{i,j}^{run} \times T_{i,j} \\ E_j^{conf} &= T_{RR_j}^{reconf} \times P^{reconf} \end{aligned}$$

and $\rho_{i,j}$ represents the need to perform a reconfiguration. For instance if the same implementation is already configured, $\rho_{i,j} = 0$ means that RR_j can be used directly to run task i , otherwise $\rho_{i,j} = 1$ and a reconfiguration is required before execution. For all non reconfigurable EUs: $\rho_{i,j} = 0$.

To execute a new task, the scheduler has to check first that the execution unit EU_j currently evaluated is free. If not, the task can be implemented later and this delay must be considered for decision. Task execution time is thus estimated in equation (2):

$$\forall j ; T_j^{cost} = T_{i,j} + T_{RR_j}^{reconf} \times \rho_{i,j} + T_j^{busy} \quad (2)$$

where T_j^{busy} represents the time during which the EU_j is busy, running another task i' . It is derived from the execution time $T_{i',j}$, the begin time of the task currently running, and the current scheduling time.

The final cost for EU_j is computed in equation (3), where α is a user parameter (real value between 0 and 1) to promote performance (α close to 0) or energy efficiency scheduling (α close to 1).

$$\forall j ; Cost_j = \alpha \frac{E_j^{cost}}{\max(E^{cost})} + (1 - \alpha) \frac{T_j^{cost}}{\max(T^{cost})} \quad (3)$$

The cost function is evaluated for all available implementations and execution units for the current top task in the waiting list. The lowest $Cost_j$ value is selected and the task is implemented on EU_j . However, if EU_j is busy the task is kept waiting and will be implemented later.

V. APPLICATION STUDY

The application considered in this case study is a H.264/AVC profile video decoder. The input specification code used is a version derived from the ITU-T reference code [16] to better cope with hardware design constraints. From the code profiling, five major functions are highlighted and define the task set \mathcal{G} as

$$\mathcal{G} = \{Exp_Golomb, MB_Header, Inv_CAVLC, Inv_QTr, Inv_Pred, DB_Filter\}. \quad (4)$$

An ESL design methodology [15] is used to provide real implementations for the possible hardware functions. The deblocking filter (DB_Filter), inverse CAVLC (Inv_CAVLC), and inverse quantization and transform block (Inv_QTr) contribute together to 76% of the global execution time on a single CPU core. They represent the three functionalities of the decoder that are generated from high level synthesis for hardware execution and can be either software or hardware executed.

This video decoder supports the possibility of slice decomposition of frames as defined in the H.264/AVC standard. A slice represents an independent zone of a frame, therefore decoding one slice (of a frame) is independent from another (slice of the same frame). This way, the decoder can process different slices of a frame in parallel and this application is an interesting case study for multi-core scheduling. We have thus considered a decomposition of the image where four streams process four slices of a same frame.

A. Modeling

The platform used in this case study is a Xilinx ZC702 [17] which is composed of a dual core ARM Cortex-A9 and a Xilinx Artix-7 equivalent FPGA. This development board allows power measurement for both FPGA and CPU core power supply. Thus measurement series are carried out under the following default setup conditions: 1V core voltage, 22°C room temperature, 667MHz ARM core frequency and 100MHz FPGA. The power values exposed in table I are setup from direct measurements on the ZC702 evaluation board.

TABLE I. MODEL PARAMETERS FOR A ZYNQ ZC702 PLATFORM.

Model	Value	Model	Value
P_{core}^{static}	89.82 mW	P^{reconf}	72 mW
P_{core}^{idle}	55.83 mW	$T_{RR_1}^{reconf}$	1.36 ms
P_{core}^{run}	119.4 mW	$T_{RR_2}^{reconf}$	1.36 ms

The task graph corresponding to the H.264 use case is shown in Figure 1. In this application configuration, reconfig-

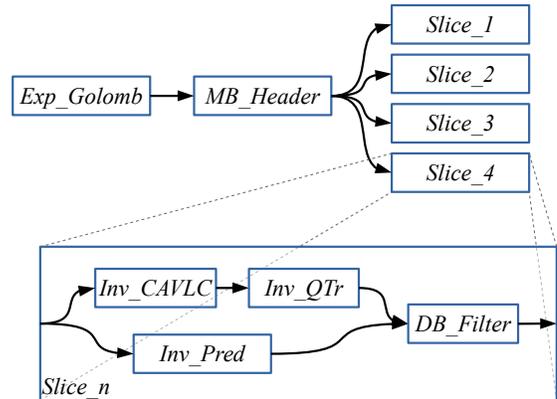


Fig. 1. H.264/AVC profile decoder task graph.

urable regions defined for DPR execution are sized to host all hardware tasks and maximize the use of parallelism, with the methodology described in [18]. The two resulting RRs (RR_1 and RR_2) are then implemented and characterized in terms of

power and execution time on the ZC702 platform, leading to the task model parameters reported in table II.

TABLE II. MODEL PARAMETERS FOR H.264/AVC PROFILE DECODER ON A XILINX ZYNQ PLATFORM.

Task i	EU j	$P_{i,j}^{run}$ (mW)	$P_{i,j}^{idle}$ (mW)	$T_{i,j}$ (ms)
<i>Exp_Golomb</i>	<i>core1</i>	119.4	-	6
	<i>core2</i>			
<i>MB_Header</i>	<i>core1</i>	119.4	-	5.9
	<i>core2</i>			
<i>Inv_CAVLC</i>	<i>core1</i>	119.4	-	6.6
	<i>RR1</i>			
	<i>RR2</i>			
<i>Inv_QTr</i>	<i>core1</i>	119.4	-	3.1
	<i>core2</i>			
	<i>RR1</i>			
<i>Inv_Pred</i>	<i>RR2</i>	5.8	34.2	1.5
	<i>core1</i>	119.4	-	3.2
<i>core2</i>				
<i>DB_Filter</i>	<i>core1</i>	119.4	-	10.5
	<i>core2</i>			
	<i>RR1</i>			
	<i>RR2</i>	6.4	44.3	0.9

B. Scheduling Results

The proposed scheduler is compared against four common scheduling strategies. User parameter α is set to 0.8 in a way to promote energy efficiency and to introduce a few performance concerns. Analysis on the H.264 decoder is carried out using a SystemC/TLM simulator [18] and results are reported in Table III and Figure 2. FPGA area is expressed in terms of Xilinx slice resources along with execution time and energy consumed to decode one frame. An energy-delay metric is also computed to provide additional measure and comparison of the energy performance trade-off [14]. Two solutions are highlighted for an EDF full software scheduling execution. The first one is tailored for the proposed platform (two cores) but can not achieve real-time video decoding (58.7ms per frame). As the application hyperperiod should be less than 40ms, the platform model is extended for a total of four ARM Cortex-A9 cores.

Another reference result is the scheduling profile obtained using two cores and static hardware accelerators (EDF Static). Indeed for this application configuration and task parallelism, a very large reconfigurable area would be required (39536 slices), which can be hard to implement on a real FPGA. Therefore another solution is introduced (EDF SmartStatic) which implements only one static accelerator for each possible hardware function (*DB_Filter*, *Inv_CAVLC*, *Inv_QTr*) to sequentially map different task invocation to the unique corresponding accelerator.

Software execution requires a large amount of energy (35mJ) while static implementations benefit from dedicated hardware efficiency with 21.1mJ (EDF Static) and 19.5mJ (EDF SmartStatic). However EDF SmartStatic is slower than EDF Static (35.4ms vs. 19.3ms) because less accelerator instances limit the exploitation of task parallelism. The outcome is visible in the energy-delay product: EDF Static obtains the best score (407) followed by EDF SmartStatic (690) and EDF Software (1350).

As for the DPR solution (two cores, two RRs), EAHS has a score which is close to EDF Static (478). In this case, energy consumption dropped from 21.1mJ to 16mJ compared to EDF

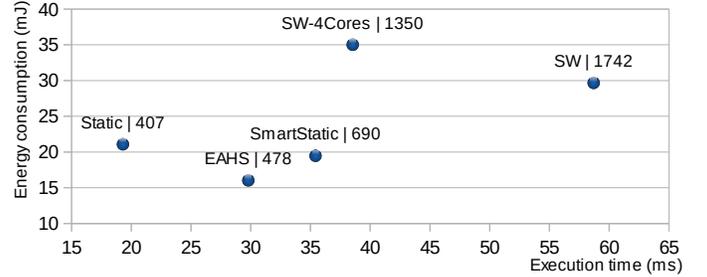


Fig. 2. Energy consumption versus time for different scheduling solutions along with energy-delay product.

Static, but execution time increased due to reconfigurations and fewer hardware resources. In addition, it is worth noting that EAHS provides a solution which consumes 3.4mJ less and is 5.6ms faster than EDF SmartStatic thanks to a better use of reconfigurable resources.

Although these results are indicative of energy efficiency, they are not completely rigorous regarding the effective execution constraint which is to process frames periodically, every 40ms typically. In this case, the objective is not so much to minimize the performance energy product but to meet execution deadlines at the lowest energy cost. Therefore, previous energy results are adjusted in Table IV for an execution constraint of 40ms.

The result is that EDF Static consumes more energy than the four other solutions, close to 40mJ. It is due to the large static and idle power consumption arising from the amount of hardware accelerators when they are not in use. This solution which had the best energy-delay previously is in fact the worst if we consider the effect of idle power consumption. EAHS takes advantage of less reconfigurable resources in this case and provides the best energy efficiency with 20.06mJ, which is about 50% better than EDF Static and 6.8% more efficient than EDF SmartStatic.

The corresponding EAHS scheduling profile is presented in Figure 3, showing the usage of each execution unit during time for two 40ms periods. As defined in the platform model, two Cortex-A9 cores are used together with two reconfigurable regions. Reconfigurations are clearly visible (in pink) between different executions of tasks on the two RRs. All four invocations of each hardware accelerator are successively configured. The first four invocations of *Inv_CAVLC* are implemented using both RRs, so that they can run in parallel by pair. Then one RR is used for four sequential invocations of *Inv_QTr* and the other is configured to run four sequential invocations of *DB_Filter*. We see clearly in this application example the ability of the scheduler to maximize the use of hardware execution (more energy efficient), to take advantage of parallel execution, and to minimize the number of reconfigurations (having non-negligible energy overheads), which are essential conditions ensuring DPR energy efficiency in practical applications.

VI. CONCLUSION

This paper described a hardware / software scheduling strategy for energy efficient application execution on hetero-

Low Power Electronics. Digest of Technical Papers., IEEE Symposium (pp. 8-11), 1994.

- [15] T. Damak, I. Werda, S. Bilavarn, N. Masmoudi. 2013, Fast Prototyping H.264 Deblocking Filter Using ESL tools, Transactions on Systems, Signals & Devices, Issues on Communications and Signal Processing, pp. 345362, 2013.
- [16] 2005. ISO/IEC 14496-10, Advanced Video Coding for Generic Audio-visual Services, ITU-T Recommendation H.264, Version 4. (2005).
- [17] Xilinx Zynq-7000 All Programmable SoC ZC702 Evaluation Kit, <http://www.xilinx.com/products/boards-and-kits/ek-z7-zc702-g.html> (2015).
- [18] F. Duhem, F. Muller, R. Bonamy, S. Bilavarn, FoRTReSS: a flow for design space exploration of partially reconfigurable systems, Design Automation for Embedded Systems, Springer Verlag, feb 2015.