CCIT Report #812 September 2012

Generalized MultiAmdahl: Optimization of Heterogeneous Multi-Accelerator SoC

Amir Morad, Tomer Y. Morad, Leonid Yavits, Ran Ginosar and Uri Weiser

Abstract— Consider a workload comprising a consecutive sequence of program execution segments, where each segment can either be executed on general purpose processor or offloaded to a hardware accelerator. An analytical optimization framework based on MultiAmdhal framework and Lagrange multipliers, for selecting the optimal set of accelerators and for allocating resources among them under constrained area is proposed. Due to the practical implementation of accelerators, the optimal architecture under area constraints may exclude some of the accelerators. As the fraction of the workload that can be accelerated decreases, resources (e.g. area) may shift from accelerators into the general purpose processor. The framework can be extended in a number of ways, spanning from SoC partitioning, bandwidth to power distribution, energy and other constrained resources.

Index Terms— MultiAmdahl, Chip Multiprocessors, Modeling of computer architecture.

1 Introduction

Large scale multiprocessor SoCs may be composed of several heterogeneous processing elements. These elements, spanning from DCT and motion estimation to GPUs, are customized hardware accelerators that are specifically designed to speed up certain tasks. Today's SoC architects have at their disposal a wide range of such heterogeneous accelerators to utilize. It is up to the system architect to efficiently allocate the resources among the accelerators, while considering physical limitations (area, power) of the design. To reach an optimal solution, the architect should take into account the efficiency of these units under resource constraints as well as the specifics of the workload.

Several previous studies use analytic models to derive optimal resource allocation of multiprocessors. Zidenberg et al. [9] presented the MultiAmdahl model that considers the implications of accelerating portions of the workload on the overall execution. Cassidy et al. [2] optimized processor area vs. L2 cache area vs. number of cores for a parallel-execution area-constrained symmetric multicore using Lagrange multipliers. MultiAmdahl differs from previously-published models [2], [3], [5], [8] by describing a system with several heterogeneous hardware units operating in sequence, directly modeling various design constraints and accounting for their impact.

The MultiAmdahl model provides the framework to resolve the following question: given an architecture composed of a predefined set of accelerators, what is the optimal resource allocation among the accelerators. The architect therefore must first select, prior to the optimization stage, a set of accelerators to use, and then allocate the resources among the selected accelerators.

At the optimal resource allocation point, all accelerators reach an equilibrium state in which if some area is taken from one accelerator and given to another, the overall runtime increases. While searching for such optimal point, one must take into account all accelerators' characteristics, total area constraints and the workload specifics. Thus, any a-priori, ad-hoc selection criteria based on performance threshold or otherwise (let alone universal criteria) electing a subset of the accelerators may yield only sub-optimal results. Hence, the key contribution of this paper is to enable the SoC architect to select an optimal subset of accelerators and allocate the area among them, without having to iterate in an ad-hoc fashion. In this paper we thus generalize the original MultiAmdahl framework to allow the optimal selection of a subset of available accelerators given the workload and the resource budget.

2 Optimization of Multi-Accelerator SoC

The MultiAmdahl [9] model optimally distributes a limited resource, such as chip area or power, among the units of a heterogeneous chip. The model assumes a workload consisting of M+1 execution segments. Each segment i of the workload requires t_i seconds to run on a reference processor. Assume that each segment i can be accelerated by using a special purpose accelerator, where each accelerator's performance speedup $Perf_j(a_j)$, relative to performance on the reference processor, is a function of the area assigned to it. The acceleration function $f_i(a_i)$ represents the inverted performance speedup of the i^{th} accelerator having area a_i:

$$f_i(a_i) = \frac{1}{Perf_i(a_i)} \quad (1)$$

- *Amir Morad (*), E-mail: amirm@tx.technion.ac.il.*
- *Tomer Y. Morad (*), E-mail: tomerm@tx.technion.ac.il.*
- *Leonid Yavits (*), E-mail: yavits@tx.technion.ac.il.*
- *Ran Ginosar (*), E-mail: ran@ee.technion.ac.il.*
- *Uri C. Weiser (*), E-mail: uri.weiser@ee.technion.ac.il.*

(*) Authors are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.

Manuscript received (July 5th, 2012).