CCIT Report #791 June 2011

Multi-Amdahl: Optimal Resource Sharing with Multiple Program Execution Segments

Tsahee Zidenberg, Isaac Keslassy, and Uri Weiser

***Abstract*—This paper presents Multi-Amdahl, a resource allocation analytical tool for heterogeneous systems. Our model includes multiple program execution segments, where each one is accelerated by a specific hardware unit. The acceleration speedup of the specific hardware unit is a function of a limited resource, such as the unit area, power, or energy. Using the Lagrange theorem we discover the optimal resource distribution between all specific units. We then illustrate this general Multi-Amdahl technique using several examples of area and power allocation among several cores and accelerators.**

I. Introduction

In the past few years, chip designers have increasingly taken into account resource constraints, most notably power, as a design goal. Their focus has shifted from improving performance to improving performance within a limited power envelope. Heterogeneous cores have been suggested for performance/power ratio improvement. These units are designed for specific workloads, trading efficiency for flexibility. The shift towards special-purpose hardware can be seen in today's CPU products, which add a graphic accelerator to the general-purpose cores [1], [2], [3]. Another example of this trend can be seen inside the general-purpose core; special-purpose logic is added, supporting specific computation, such as CRC or cryptography [4].

In multicore environment it is necessary to correctly balance the performance of parallel and serial code segments to overcome the Amdahl law ceiling [5]. Parallel code runs most efficiently when splitting the available area into many processors, while the serial code can only run on a single processor. The difference between the requirements of the two sections created the *asymmetric-cores* approach [6]. An optimal point for the asymmetry can be found using Amdahl's law for balancing the importance of parallel and serial execution, along with the fact that all processors share a common resource [7].

As hardware becomes more specialized and diverse, Amdahl's law becomes *multidimensional*. In this environment not only some code segments are accelerated while others are not, but also different accelerated segments might rely on different types of accelerators. Some of these segments could represent high-level computation sections, such as matrix multiplication and FFT, while others could represent low-level computation sections, such as sections with many floating-point instructions.

A limited resource, such as power or area, is shared among the various specific units on the chip. Our target is to find the optimal way to distribute that resource between the specific HW units, balancing the efficiency of different hardware units with their importance and performance.

This paper presents two main contributions:

- We propose Multi-Amdahl, an analytical tool to optimize resource allocation among n different specific HW units running n segments of execution code. The architect may impose constraints on the design, such as total area, power, or energy, and expect optimal outcome, such as maximum speedup. In our model, we take into account the differences in efficiency and scalability of hardware units, and the workload distribution among the different segments.
- Initial results and intuitions obtained from Multi-Amdahl are presented. These results suggest that the opportunities that exist in heterogeneity might surpass the cost of inflexibility. In other words, the occasional use of an accelerator might exceed the costs resulting from its frequent inactivity.

The paper structure is as follows: Section II covers the related work, Section III presents the Multi-Amdahl technique. Our technique is compared with Amdahl's law in Section IV. In Sections V and VI, we present a few examples for extending the basic model. Section VII presents results from our model and their implications, and Section VIII concludes and points to potential future extensions of our work.

II. Related work

The move towards accelerators

Venkatesh et al. [8] explored the problem of "dark silicon". Today, threshold voltage no longer scales when technology advances. The result is that a shrinking percent of the chip may be activated simultaneously. To achieve peak performance, we must make sure that minimal power is spent for each function, so that more functions could be executed in parallel. The article suggested an architecture with many heterogeneous cores, each designed for optimal power efficiency of a different software function. Those automatically-generated units provided up to 30x efficiency (work/J) over general-purpose MIPS. The article also suggested "patchable" versions of those units with lower (16x) improvements.

Shee et al. [9] tested a few architectures for a heterogeneous chip built especially to encode JPG images. Each core was optimized for one specific pipeline stage. Mostly, optimization was done by removing unnecessary components from a

All writers are with the department of Electrical Engineering, Technion, Haifa 32000, Israel. As usual, this technical report has not been peer-reviewed and does not count as a publication. It simply presents a few preliminary results.