Parallel Computing 40 (2014) 1–16

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

The effect of communication and synchronization on Amdahl's law in multicore systems

L. Yavits *, A. Morad, R. Ginosar

Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 14 January 2013
Received in revised form 10 November 2013
Accepted 11 November 2013
Available online 21 November 2013

Keywords:
Multicore
Analytical Performance Models
Amdahl's law

ABSTRACT

This work analyses the effects of sequential-to-parallel synchronization and inter-core communication on multicore performance, speedup and scaling from Amdahl's law perspective. Analytical modeling supported by simulation leads to a modification of Amdahl's law, reflecting lower than originally predicted speedup, due to these effects. In applications with high degree of data sharing, leading to intense inter-core connectivity requirements, the workload should be executed on a smaller number of larger cores. Applications requiring intense sequential-to-parallel synchronization, even highly parallelizable ones, may better be executed by the sequential core. To improve the scalability and performance speedup of a multicore, it is as important to address the synchronization and connectivity intensities of parallel algorithms as their parallelization factor.

1. Introduction and related work

Consider a multicore comprising a general purpose core responsible for the sequential fraction of the code (the *sequential core*) and a number of parallel cores that execute the parallelizable fraction of the program. Parallelization incurs (1) data exchange between the sequential core and the parallel cores at the beginning and the end of each parallel section of a program (*sequential-to-parallel synchronization*), and (2) data exchange among the parallel cores during parallel execution (*inter-core communication*). This paper investigates the effect of sequential-to-parallel synchronization and inter-core communication on multicore performance and scaling from Amdahl's law perspective.

Some multicore performance models tend to ignore the impact of sequential-to-parallel synchronization and inter-core communication on performance and power. We propose an analytical model that accounts for these effects. Based on this model, we reach a number of conclusions that affect multicore design. First, we show that impact of sequential-to-parallel synchronization and inter-core communication on performance becomes more significant with parallelism. Second, we find that a smaller number of larger cores tends to be more efficient performance-wise than a larger number of smaller cores as implied by Hill and Marty [2], and this effect becomes more predominant with growing parallelism. Third, for low arithmetic intensity (the ratio of arithmetic operations to memory traffic [20]) tasks, parallel execution may result in lower speedup, or even no speedup relative to Amdahl's law. There are also implications as to the optimal task scheduling: for tasks with high inter-core communication requirements, the workload should be assigned to a small number of cores, even if the parallelization fraction f is close to 1. For low arithmetic intensity tasks, the workload may better be assigned to the sequential core, even if the parallelization fraction f is close to 1. Both conclusions stand in contradiction to Amdahl's law [1], which implies that the higher f is, the more cores should be used.

* Corresponding author. Tel.: +972 544 496 863.
E-mail addresses: yavits@tx.technion.ac.il (L. Yavits), amirm@tx.technion.ac.il (A. Morad), ran@ee.technion.ac.il (R. Ginosar).

0167-8191/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.parco.2013.11.001

Implications of multicore design on Amdahl's law and modeling multicore performance have been thoroughly studied. Hill and Marty augmented Amdahl's law with a corollary to multicore architecture by constructing a model for multicore performance and speedup [2]. Cassidy et al. [5–7]. extended the model by also considering cache, and suggested an area allocation optimization framework. Eyerman and Eeckhout introduced a model that accounts for critical sections of the parallelizable fraction of a program [8]. Gunther et al. [10] considered the effects of resource sharing and resource limitations on such models.

In this paper, we analyze the work of Hill and Marty, Cassidy et al., Eyerman and Eeckhout and Gunther et al., introduce a novel model that considers the effects of sequential-to-parallel synchronization and inter-core communication, and compare all five models in a unified framework.

Other studies also discussed multicore scalability and performance. Oh et al. [4] investigated the tradeoff between the number of CMP cores and the cache size for variety of cache organizations. Chung et al. [9] extended Hill and Marty's model by considering the constraints of off-chip bandwidth, and by introducing accelerators achieving better performance/power ratio. Sun and Chen [16] took a different approach to multicore performance modeling, based on Sun and Ni [21] law rather than Amdahl's law. Rogers et al. [11] studied the limitations of multicore scaling imposed by memory bandwidth constraints and created an analytical model for memory traffic. Loh [12] augmented Hill and Marty's model by accounting for the cost (area) of the "uncore", namely the parts of the chip outside the core. Synchronization and communication overhead was originally studied by Flatt and Kennedy [29]. Morad et al. [15] incorporated this overhead into the multicore performance model.

The rest of this paper is organized as follows. Section 2 starts by discussing factors that limit speedup, continues to describe Hill and Marty's, Cassidy et al., Eyerman and Eeckhout's and Gunther et al. models, and introduces our models for symmetric and asymmetric multicore architectures. Section 3 describes simulations that validate the analytical models. Section 4 offers a discussion of the results, and Section 5 presents the summary.

2. Multicore performance model

The section starts by presenting four existing models in a unified framework and proceeds to introduce a novel model for symmetric and asymmetric multicore architectures.

2.1. Beyond Amdahl's law – speedup limiting factors

Our multicore performance model accounts for the effects of inter-core communication and sequential-to-parallel data synchronization, as follows. Consider the asymmetric multicore model shown in Fig. 1, the symmetric multicore model in Fig. 2 and the sequential/parallel execution model of Fig. 3. Examples of asymmetric (even heterogeneous) systems include a CPU with its own memory, accompanied by a GPU with graphics memory, and the CELL processor [24], and are analyzed in [15]. Tiled architectures [25–27] are examples of symmetric systems.

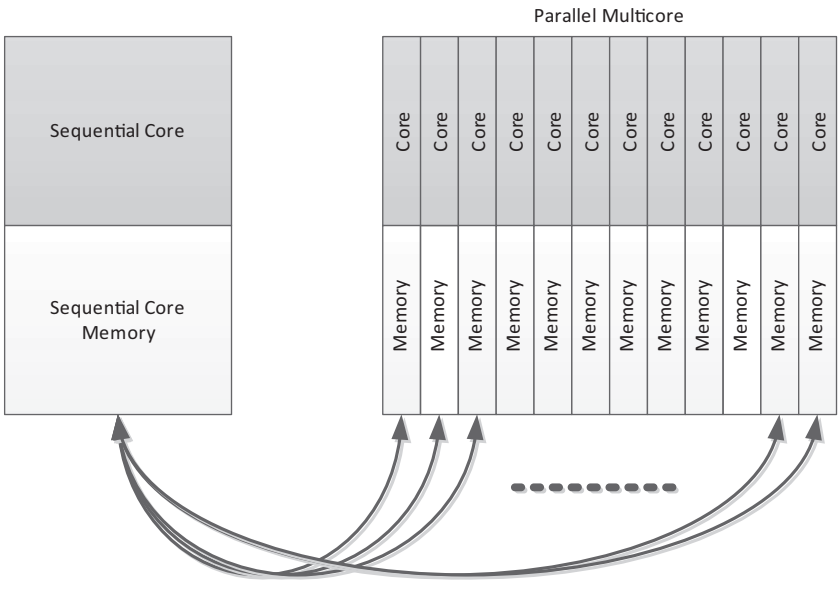

Fig. 1. Asymmetric multicore model.