CCIT Report #814 October 2012

THE EFFECT OF COMMUNICATION AND SYNCHRONIZATION ON AMDAHL'S LAW IN MULTICORE SYSTEMS

L. Yavits, A. Morad, R. Ginosar

Abstract—This work analyses the effects of sequential-to-parallel synchronization and inter-core communication on multicore performance, speedup and scaling. A modification of Amdahl's law is formulated, to reflect the finding that parallel speedup is lower than originally predicted, due to these effects. In applications with high inter-core communication requirements, the workload should be executed on a small number of cores, and applications of high sequential-to-parallel synchronization requirements may better be executed by the sequential core, even when f, the Amdahl's fraction of parallelization, is very close to 1. 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.

Index Terms— Multicore, Analytical Performance Models, Amdahl's law.

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 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 the data exchange among the parallel cores during parallel execution (*inter-core communication*). This paper investigates the effect of these synchronization and communication elements on multicore performance, speedup and scaling.

Many multicore performance models tend to ignore the impact of sequential-to-parallel synchronization and inter-core communication on performance and power. We propose an analytic 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 the level of parallelism. Second, a smaller number of larger cores tends to be more efficient performance-wise than a larger number of smaller cores as implied by Amdahl's law, and this effect becomes more predominant with growing parallelism. Third, for low arithmetic intensity [22] tasks, parallel execution may result in lower speedup, or even no speedup relative to Amdahl's law [1]. We also arrive at a number of conclusions as to an 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 very close to 1. For low arithmetic intensity tasks, the workload had better be assigned to the sequential core, even if the parallelization fraction f is very close to 1. Both conclusions stand in contradiction to Amdahl's law, which implies that the higher f is, the more cores should be used.

Extending Amdahl's law in the multicore era and modeling multicore performance has 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] [6][7] extended the model by considering also cache area and power, and suggested an optimization framework. Eyerman and Eeckhout introduced a model that accounted for critical sections of the parallelizable fraction of a program [7]. Gunther *et al.* considered the effects of resource sharing and limitations on such models [10].

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. investigated the tradeoff between the number of CMP cores and cache architecture, including cache size, L2 and L3 cache effects, shared vs. private and hybrid caches, and uniform vs. non-uniform caches [4]. Chung et al. extended Hill and Marty's model by considering the constraints of off-chip bandwidth, and by introducing accelerators achieving better performance/power ratio [9]. Sun and Chen [18] took a different approach to multicore performance modeling, based

- *Leonid Yavits (*), E-mail: yavits@tx.technion.ac.il.*
- *Amir Morad (*), E-mail: amirm@tx.technion.ac.il.*
- *Ran Ginosar (*), E-mail: ran@ee.technion.ac.il.*

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