J Supercomput (2015) 71:2783–2786
DOI 10.1007/s11227-015-1487-8

Advances in multicore systems architectures

H. Sarbazi-Azad[1,2] · **N. Bagherzadeh**[3] · **G. Jaberipour**[4]

Published online: 22 July 2015

1 Introduction

The ever increasing computation power demand and rapid advances in very large integration and communication technology have led to the development of high-performance multi- and many-core computing systems. These systems enjoy parallelism at thread, instruction, task, and program levels. Moreover, in almost all cases, they have to meet tight power consumption requirements.

Multicores are dominating all aspects of computing ranging from mobile devices to desktops and supercomputers. Some of the major areas of research and development are: on-chip networking, mapping and scheduling tasks, low-power design considerations, parallel programing tools, and multicore algorithms.

This special issue presents recent research results on multicore systems; contributions cover different topics such as architecture, programming tools and applications, as well as other related aspects of multicore systems design.

Contributed papers include some of the high-quality papers accepted at the CADS 2013 conference that took place in Tehran, which were selected based on the evaluation scores and presentation quality at the conference. We invited authors of these papers to extend and submit their work for this special issue. Additionally, we solicited high-quality papers from known researchers in the field as a result of our wide dis-

✉ H. Sarbazi-Azad
azad@ipm.ir

1 Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

2 School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

3 Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, USA

4 Department of Electrical and Computer Engineering, Shahid-Beheshti University, Tehran, Iran

tribution of call for papers for this special issue. The review process required three rounds of diligent reviews and revisions resulting in the final accepted manuscripts for publication in the journal. In total, six papers were accepted; a brief overview of each paper is presented next.

2 The papers

2.1 Integrated circuit-packet switching NoC with efficient circuit setup mechanism

In this paper, the authors address one of the major problems of circuit-switching based on circuit setup time overhead, by an efficient and fast algorithm using time-division multiplexing (TDM) scheme. They show performance improvement by reserving circuits for anticipated messages, and hence completely hide circuit setup time. To address the low resource utilization problem, they integrate the proposed circuit-switching into a packet-switched NoC and use unused circuit resources to transfer packet-switched data. Their results show considerable reduction in NoC power consumption and packet latency.

2.2 Power-efficient prefetching on GPGPUs

The graphics processing unit (GPU) is the most promising candidate platform for achieving faster improvements in peak processing speed, low latency and high performance. The highly programmable and multithreaded nature of GPUs makes them a remarkable candidate for general purpose computing. The authors of this paper focus on improving performance by better hiding long waiting time for transferring data from the slow global memory. Furthermore, they show that the proposed method can reduce power and energy. Reduction in access time to off-chip data has a noticeable role in reducing waiting time and the percentage of unutilized elements. This paper discusses that using processing elements in a suitable manner to prefetch data during stall times bridging the memory gap in an energy efficient manner leads to lower power and energy consumption. Simulation results show that one potentially can improve instruction per cycle (IPC), power, energy and energy efficiency by up to 24.76, 22.47, 24.72 and 36.01 %, respectively.

2.3 Design and analysis of a mesh-based wireless network-on-chip

Network-on-chip (NoC) architecture is regarded as a solution for future on-chip interconnects, but performance advantages of conventional NoC architectures are limited by the long latency and high power consumption of long-distance communication among cores. To solve these limitations, the authors of this paper employed an on-chip wireless communication to provide express links as an alternative for transferring long-distance data. This paper presents a hybrid NoC architecture utilizing both wired and wireless communication approaches. They also devised a deadlock free routing

algorithm that is capable of making efficient use of the incorporated wireless links. Moreover, simulated annealing optimization techniques were applied to find optimal locations for wireless routers. Cycle-accurate simulation results confirmed a significant improvement in transfer latency.

2.4 PS directory: a scalable multilevel directory cache for CMPs

As the number of cores increases in current and future chip-multiprocessor (CMP) generations, coherence protocols must rely on novel hardware structures to scale in terms of performance, power, and area. Systems that use directory information for coherence purposes are currently the most scalable alternative. This paper studies the important differences between the directory behavior of private and shared blocks. The proposed PS directory, is a two-level directory cache that keeps the reduced number of frequently accessed shared entries in a small and fast first-level cache, namely shared cache, and uses a larger and slower second-level private cache to track the large amount of private blocks. Experimental results demonstrate that for a 16-core CMP as compared to a conventional directory, the PS directory improves performance by 14 % while reducing silicon area and energy consumption by 34 and 27 %, respectively. Also, compared to the state-of-the-art multi-grain directory, the PS directory apart from increasing performance reduces power by 18.7 %, and provides more scalability in terms of area.

2.5 In-order delivery approach for 2D and 3D NoCs

In many applications, it is critical to guarantee in-order delivery of packets. Since NoCs packets may use different paths, in-order delivery constraint cannot be met without support. To guarantee in-order delivery, traditional approaches either use dimension-order routing or employ reordering buffers at network interfaces. Dimension-order routing degrades the performance considerably while the usage of reordering buffers imposes large area overhead. In this paper, authors present a mechanism allowing packets to be routed through multiple paths in the network, helping to balance the traffic load, while guaranteeing in-order delivery. The proposed method combines the advantages of both deterministic and adaptive routing algorithms. Their basic idea is to use different deterministic algorithms for independent flows. This approach neither requires reordering buffers at destinations nor limits packets to use a single path. The concept is investigated in both 2D and 3D mesh networks.

2.6 A statistical performance analyzer framework for OpenCL kernels on Nvidia GPUs

Understanding performance bottlenecks of applications in high-performance computing can lead to dramatic improvements of applications performance. In this paper, authors provide a statistical performance analyzer framework that not only helps finding bottlenecks but also shows additional information that is commonly not available

using a profiler. Recently, OpenCL has been proposed to be used in a variety of platforms (e.g., CPUs and GPUs); therefore, a program written in one platform can be imported to other platforms with minimal effort. Authors selected OpenCL to design their performance model for Nvidia GPUs. To construct the model, the values of GPU performance counters for selected benchmarks were measured. The proposed method in this paper can be leveraged to characterize unknown applications based on their performance similarities with an existing database of benchmarks to predict their likely performance bottlenecks.

3 Conclusion

In the last decade, research in various areas of multicore systems architecture design have reached the level where many new ideas have been proposed and implemented by researchers in the field. Because of the continued great interest in this area of research, this special issue was formed to address practical and innovative contributions by researchers. We hope that by gathering these selected pieces of work in a special issue of a well-known journal will serve as a contribution to the expansion of such exchanges and will promote additional research contributions.

Acknowledgments We would like to express our deep gratitude to the Editor-in-Chief, Prof. Hamid R. Arabnia, for hosting this special issue in the *Journal of Supercomputing* and for his support and helpful advices during various stages of preparing this special issue. Our sincere appreciation also goes to Sudha Subramanian and her colleagues in the editorial office, for their excellent job during the course of this project. We thank the authors for their contributions, with tribute to those whose works were not selected for inclusion in this special issue. Last but not least, we would like to deeply acknowledge and appreciate the work of many reviewers who have provided invaluable evaluations and recommendations.