Journal of Parallel and Distributed Computing 168 (2022) 137–138

Contents lists available at ScienceDirect

Journal of Parallel and Distributed Computing

www.elsevier.com/locate/jpdc

Editorial

Special Issue on Computer Architecture and High-Performance Computing

Computer Architecture and High-Performance Computing is an evolving research topic with new architectures for multicore processors and accelerators that require novel techniques to consider all the features available in modern devices today.

To couple with all these advances, high-performance computing requires support from languages and compilers, modeling and simulation, energy-aware techniques, performance measurements and analysis, among many other issues.

The purpose of this special issue is to collect the main trends and innovations related to computer architectures, networking and distributed systems, parallel applications and algorithms, performance evaluation and modeling, and system software, concerning high-performance computing. It contains revised and extended versions of nine papers presented at SBAC-PAD 2020, the 32nd IEEE International Symposium on Computer Architecture and High-Performance Computing, held on September 8–11, 2020, in Porto, Portugal. From the 87 full papers submitted, 30 were selected to be presented at the conference, and from those, 9 were selected for this special issue.

The paper "A Robotic Middleware Combining High Performance and High Reliability", in the context of autonomous systems, presents the design of robotic middleware to combine both high performance and high reliability in order to achieve system reliability and product safety. Furthermore, authors designed a lightweight service discovery, to achieve high performance and a weak centralized mechanism for high reliability. Experiments show that communication latency significantly outperforms state-of-the-art robotic middleware by up to 41%.

Shared memory machines make concurrent data structure design a critical factor for the design of high-performance applications or parallel systems. The paper "Using Skip Graphs for NUMA Locality", presents a NUMA-aware concurrent data structure design based on a data-partitioned, concurrent skip graph indexed by thread-local sequential maps, which obtains significant quantitative and qualitative improvements on NUMA locality, as well as reduced contention for synchronized memory accesses.

The paper "Reducing Response Latency of Composite Functions-as-a-Service through Scheduling" models resource management in FaaS as a scheduling problem that combines: sequencing of invocations; deployment; and allocation of invocations to deployed environments. Results show that if the setup times are long compared to invocation times, algorithms that use information about the composition of functions consistently outperform greedy, myopic algorithms, leading to a significant decrease in response latency.

The paper "Mapping series-parallel streaming applications on hierarchical platforms with reliability and energy constraints" addresses streaming applications that have a series-parallel dependence graph, where data is continuously generated and must be processed on the fly. A dynamic-programming algorithm for the special case of linear chains is presented, which is optimal for a special class of schedules. Simulations on realistic settings demonstrate the good performance of the proposed algorithm, in particular, significant energy savings were obtained.

The paper "Decoupling GPGPU voltage-frequency scaling for deep-learning applications" addresses the problem of reducing energy consumption for the tasks of training and inference in the context of deep learning. The paper introduces a new methodology to fully characterize the impact of non-conventional DVFS on GPUs. The approach was evaluated on two AMD devices and the results show that it is possible to safely decrease the GPU voltage, allowing for a significant reduction of the energy consumption (up to 38%) on the training procedure of CNN models, with no degradation of the accuracy of the network.

The paper "Targeting a Light-Weight and Multi-Channel Approach for Distributed Stream Processing" addresses the problem of high-throughput data streams that arise in areas such as real-time event monitoring, complex dataflow processing, and big data analytics. It presents a new distributed stream processing engine, called Asynchronous Iterative Routing (AIR), which implements a lightweight, dynamic sharding protocol. AIR expedites direct and asynchronous communication among all the worker nodes via a channel-like communication protocol on top of the Message Passing Interface (MPI), thus avoiding the need for a dedicated driver node. It is shown that AIR scales out particularly well to multicore HPC architectures, and significantly outperforms existing distributed stream processing engines.

The paper "A Fast and Concise Parallel Implementation of the 8x8 2D Forward and Inverse DCTs using Halide" presents the implementation of the Discrete Cosine Transformation (DCT) using Halide. It is shown that the Halide code is significantly easier to maintain and port to new architectures than the existing code, where a Halide code with 200 lines replaces the original code with 20,000 lines to do the same JPEG encoding and decoding. The Halide implementation is compared for ARMv8 NEON and x86-64 SIMD extensions and shows a 5–25 percent performance improvement over the SIMD code in libjpeg-turbo for decoding and a 10–40 percent improvement for encoding.

The paper "Optically Connected Memory for Disaggregated Data Centers" proposes and evaluates an Optically Connected Memory (OCM) architecture that disaggregates the main memory from the

https://doi.org/10.1016/j.jpdc.2022.06.013
0743-7315/© 2022 Published by Elsevier Inc.

computation nodes in data centers, for next-generation data centers. Recent advances in integrated photonics enable the implementation of reconfigurable, high-bandwidth, and low energy-per-bit interconnects. Results show that OCM is capable of interconnecting four DDR4 memory channels to a computing node using two fibers with 1.02 pJ energy-per-bit consumption, and OCM performs up to 5.5 $\times$ faster than a disaggregated memory with 40G PCIe NIC connectors to computing nodes.

The paper "Efficient and Portable GEMM-based Convolution Operators for Deep Neural Network Training on Multicore Processors" aims to develop high-performance realizations of the convolution operators, which concentrate a significant portion of the computational cost of Convolutional Neural Networks. In this paper, the authors extend their previous work to the full training process of CNNs on multicore processors, proposing new high-performance strategies to tackle the convolution operators that are present in the more complex backward pass of the training process, while maintaining the portability of the realizations. Results show that, in comparison with baseline implementations, the use of the new convolution operators using pre-allocated memory can accelerate the training by a factor of about 6–25%, provided there is sufficient memory available. And, the operator variants that do not rely on persistent memory, can save up to 70% of memory.

The guest editors of this special issue wish to thank the authors for their contributions, the track chairs for their efforts in reviewing and in the selection process, and all reviewers for their useful suggestions. Finally, we would like to express our gratitude to Professor Viktor Prasanna (Editor-in-Chief) and Professor Albert Y. Zomaya (Associate Editor-in-Chief, Special Issues) for the opportunity to edit this special issue.

Jorge G. Barbosa *
Faculty of Engineering, University of Porto, Portugal

Lúcia M.A. Drummond
Universidade Federal Fluminense, Brazil

Laurent Lefèvre
Inria, ENS Lyon, University of Lyon, France

* Corresponding editor.