Received: 14 February 2020 | Revised: 23 February 2020 | Accepted: 23 February 2020
DOI: 10.1002/cpe.5736

WILEY

EDITORIAL

The 27th International Heterogeneity in Computing Workshop and the 16th International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms

Alexey L. Lastovetsky | **Ravi Reddy Manumachu**

School of Computer Science, University College Dublin, Dublin, Ireland

Correspondence
Ravi Reddy Manumachu, School of Computer Science, University College Dublin, Dublin, Ireland.
Email: ravi.manumachu@ucd.ie

1 | INTRODUCTION

Heterogeneity is now a pervasive characteristic of computing. From the macrolevel, where networks of distributed computers composed of diverse node architectures are interconnected with potentially heterogeneous networks, to the microlevel, where deeper memory hierarchies and various accelerator architectures are increasingly common, the impact of heterogeneity on all computing tasks is profound.

High-performance computing (HPC) clusters, clouds, and data centers today exhibit heterogeneity at various levels of their software stacks and hardware topologies. These platforms commonly feature tight integration of multicore CPU processors and accelerators such as graphical processing units (GPUs), field programmable gate arrays (FPGAs), Intel Xeon Phis, and so on, empowering them to provide not just unprecedented computational power but also to address the critical concern of energy efficiency. The TOP500 list[1] showcases the snapshot of dominant hardware trends in HPC. It currently features around 150 systems that contain multicore CPUs integrated with accelerators/coprocessors. While multicore CPU space is dominated by three processors, Intel Xeon E5 (Broadwell), Intel Xeon Gold, and Intel Platinum, the accelerators are quite diverse: GPUs that include Tesla V100, Tesla P100, Tesla K80, from NVIDIA, Vega 20 from AMD and Many-cores such as Intel Xeon Phi 5120D, Intel Xeon Phi 7120P, Intel Xeon Phi 5110P, Matrix-2000, PEZY-SC2, and so on. Many heterogeneous platforms (of similar flavor as those in this list) are fueling rapid advances not just in scientific application areas but also in the data science fields of big data analytics, deep learning, and so on.

The perennial programming challenges on such platforms have been to maximize their efficiency and resource utilization since the platforms are ever-changing with novel and multifarious parallel architectures. New ideas, innovative algorithms, and specialized programming environments and tools are constantly needed to address the challenges.

The Heterogeneity in Computing Workshop (HCW) and International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar) have been the flagship forums bringing together researchers to discuss these challenges and the solutions. The wide range of topics deliberated includes, to name a few, heterogeneous parallel programming paradigms, algorithms, models and tools for performance and energy optimization on heterogeneous platforms, and fault tolerance of parallel computations on heterogeneous platforms.

The accepted articles in the workshops this year covered topics, techniques, and applications, exhibiting lucidly the depth, breadth, and growth of the heterogeneous computing field. Two promising developments are, however, apparent. The first is the slow but steady adoption of FPGAs as yet another acceleration technology competing with GPUs and Intel Xeon Phis in HPC and data science. This is evidenced from the increasing number of publications submitted to the two workshops featuring FPGAs for accelerating software algorithms in the fields of cryptography, anomaly detection, and so on. The second is the growing awareness of energy of computing as a serious environmental concern and a grand technological challenge. Energy is now considered a fundamental design constraint along with performance in all computing settings. The number of submissions focusing on single-objective optimization for performance with energy constraints and biobjective optimization for both performance and energy on heterogeneous platforms is steadily increasing.

This special issue contains five selected articles from the HCW'2018 and HeteroPar'2018 workshops. We hope you find the articles, whose summaries are below, informative, interesting, and thought-provoking.

2 | HCW AND HETEROPAR WORKSHOPS: SUMMARY

2.1 | HCW'2018

The 27th International Heterogeneity in Computing Workshop (HCW'2018) was held on 21 May in Vancouver, BC, Canada, in conjunction with the International Parallel and Distributed Processing Symposium (IPDPS) annual series of international conferences. Seven articles were presented at the workshop covering topics ranging from scheduling scientific workflows on cloud platforms to optimization of parallel applications on FPGA platforms. Out of these seven articles, two were submitted and accepted for publication in the special issue.

2.2 | HeteroPar'2018

The 16th edition of the International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms Workshop (HeteroPar'2018) was held on 27 August in Turin, Italy. For the 10th time, this workshop was organized in conjunction with the Euro-Par annual series of international conferences. Twenty-six articles were submitted for review from 16 countries. Each paper secured three reviews from members of the program committee. After a thorough peer-reviewing process, 10 articles (an acceptance ratio of 38%) were selected for presentation at the workshop. The topics included realistic simulations of file replication strategies, anomaly detection using FPGA, optical coherence tomography accelerated using GPUs, application-centric parallel memories, perturbations in heterogeneous systems, FPGA-accelerated change-point detection, merging publish-subscribe pattern and shared memory, a modular precision format, fast heuristic-based GPU compilation and benchmarking latest GPU, and tensor cores. Out of the 10 articles presented at the workshop, three were submitted and accepted for publication in the special issue.

3 | ACCEPTED PAPERS FOR THE SPECIAL ISSUE: SUMMARY

The wide gulf between the growth of arithmetic performance and memory bandwidth performance inherent in heterogeneous platforms today motivates the work of Thomas et al.[2] They present a customized precision format that splits the mantissa (significand) of the standard IEEE floating point precision format into segments, which enables accessing only part of the significand information. This strategy while preserving the exponent range of the floating point precision format enables much faster memory access if reduced accuracy in the memory operations is acceptable. The chief objectives of this strategy are (i) radically decoupling the data storage format from the processing format, (ii) designing a "modular precision ecosystem" that accommodates more flexibility in terms of customized data formats and memory access, and (iii) developing innovative algorithms and applications that dynamically adapt data access accuracy to the numerical requirements.

Modern heterogeneous systems currently feature accelerators that offer massive parallelism for compute-intensive applications, but often suffer from memory bandwidth limitations. Giulio et al[3] propose an approach to build application-centric configurable parallel memories for applications that suffer from memory bandwidth limitations on modern heterogeneous platforms. Their methodology is an application-to-accelerator workflow containing the following main steps: (i) analysis of the application memory access traces to extract parallel accesses, (ii) configuring and building custom application-specific parallel memory, (iii) generation of the parallel-memory accelerator in hardware, and (iv) embedding the accelerator in the original host code.

Accurate prediction of utilizations of resources is a key to optimization of mapreduce (and bigdata) applications executed in large-scale Hadoop clusters. Lei et al[4] introduce a method to address this problem. Their method consists of two steps: (i) a simulator that performs performance simulation of Hadoop applications, where the execution times of the maps and reduces in the applications are used to train a resource utilization model and (ii) predictor of utilizations of resources based on the results from the simulations in the first step.

Scientific applications are characterized by compute-intensive parallel loops. The performance of these loops is impacted by load imbalances arising from the irregularity of loop accesses due to the inherent nature of the application and the system due to fluctuating network latencies and bandwidths. Ali et al[5] propose an approach, simulation-assisted scheduling algorithm selection (SimAS), that selects the best dynamic loop self-scheduling (DLS) technique on heterogeneous HPC systems in the event of such load imbalances.

Scheduling the tasks of a sequential application on a heterogeneous CPU-GPU platform while minimizing for both execution time and energy consumption is a promising new field of research. Massinissa et al[6] present an optimal algorithm and a fast 2-approximation algorithm to solve a task scheduling problem with energy constraint and communication delays, where the execution times and energy consumptions of the tasks are represented by constants. The sequential application is represented by a linear chain of tasks. The objective is to minimize the total execution time (makespan) while respecting an energy bound. While the optimal algorithm solves the problem without energy constraint, the 2-approximation algorithm solves the problem with energy constraint.

ORCID

Ravi Reddy Manumachu https://orcid.org/0000-0001-9181-3290

REFERENCES

1. Top500. *Top 500. The List - November 2019*; 2020.
2. Grützmacher T, Cojean T, Flegar G, Göbel F, Anzt H. A customized precision format based on mantissa segmentation for accelerating sparse linear algebra. *Concurr Comput Pract Exp*. 2019;e5418-e5418. https://doi.org/10.1002/cpe.5418.
3. Stramondo G, Ciobanu CB, Laat C, Varbanescu AL. Designing and building application-centric parallel memories. *Concurr Comput Pract Exp*. 2019;e5485-e5485. https://doi.org/10.1002/cpe.5485.
4. Yu L, Teng F, Ning S, Li Y, Cui Z, Du S. A two steps method of resources utilization predication for large Hadoop data center. *Concurr Comput Pract Exp*. 2020;e5634-e5634. https://doi.org/10.1002/cpe.5634.
5. Mohammed A, Ciorba FM, Sim AS. A simulation-assisted approach for the scheduling algorithm selection under perturbations. *Concurr Comput Pract Exp*. 2020;e5648-e5648. https://doi.org/10.1002/cpe.5648.
6. Ait AM, Zaourar L, Munier A. Efficient algorithm for scheduling parallel applications on hybrid multicore machines with communications delays and energy constraint. *Concurr Comput Pract Exp*. 2020;e5573-e5573. https://doi.org/10.1002/cpe.5573.

How to cite this article: Lastovetsky AL, Reddy Manumachu R. The 27th International Heterogeneity in Computing Workshop and the 16th International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms. *Concurrency Computat Pract Exper*. 2020;e5736. https://doi.org/10.1002/cpe.5736