CCF Transactions on High Performance Computing (2021) 3:1–3
https://doi.org/10.1007/s42514-020-00062-5

EDITORIAL

Editorial for the special issue on reliability and power efficiency for HPC

Jifeng He[1] · Chenggang Wu[2] · Huawei Li[2] · Yang Guo[3] · Tao Li[4]

Published online: 20 January 2021

High Performance Computing (HPC) performs more complex tasks with the application of parallel and distributed algorithms than computing on a single node. And HPC continuously advances in traditional domains of science and engineering. However, the emergence of novel applications calls for the lower latency of the network, which pushed the horizon of edge computing. Today, the diversity of HPC systems is more extensive, and rapid changes in hardware platforms and program environments increasingly challenge the high concurrency exploitation, hybrid resource management, energy efficiency, performance tuning, scalability and fault-tolerance.

We have nine invited papers selected for this special issue based on a peer-review procedure, which cover a few different aspects that relate to energy-efficient designs on FPGA and framework for resource or task scheduling.

The first part of the special issue focuses on the energy-efficient circuit designs. We have two papers to implement the neural network accelerators on FPGA, and three papers that discuss run-time reconfigurable physical unclonable function unit, security verification resources allocation framework, and reliability estimation of gate-level approximate arithmetic circuits, which aim to achieve high energy and resource efficiency.

- The paper written by Dong Wen et al. proposes an energy-efficient convolutional neural network (CNN) accelerator for speech classification based on FPGA and quantization. The accelerator not only owns high power-efficiency but also provides reliable speech classification performance, higher throughput and better time-delay overhead. Related methods on quantization scheme and layer-by-layer hardware pipeline design are also presented, and comparison between the proposed architecture and high-performance CPU and other specific accelerators shows its performance superiority to existing hardware platforms.
- The paper written by Zhe Han et al. proposed a scalable fully pipelined binarized neural networks (BNN) architecture, which targeted on maximizing throughput and keeping energy and resource efficiency in large FPGA. It utilized the resource with sharing on-chip memory and balancing the computation resources and achieved excellent performance by exploiting multi-levels parallelism and balancing pipeline stages. The evaluation on Xilinx Ultra-Scale XCKU115 shows that the proposed architecture achieves 2.24 × –11.24 × performance and 2.43 × –11.79 × resource efficiency improvement compared with other BNN accelerators.
- The paper written by Shen Hou et al. proposes a dynamically configurable LFSR-based Physical Unclonable Function (PUF). It does not only resist modeling attacks, but is also sufficiently lightweight to fit the low-end internet of things and embedded devices. High entropy source and large CRP space are achieved by dynamically configuring the LFSR. Both experiments on simulation and FPGA prove the effectiveness of the proposed design.
- The paper written by Haoyi Wang et al. presented a security game framework to guide the security verification

✉ Tao Li
litao@nankai.edu.cn

Jifeng He
jifeng@sei.ecnu.edu.cn

Chenggang Wu
wucg@ict.ac.cn

Huawei Li
lihuawei@ict.ac.cn

Yang Guo
guoyang@nudt.edu.cn

[1] East China Normal University, Shanghai, China

[2] State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[3] National University of Defense Technology, Changsha, China

[4] Nankai University, Tianjin, China

resources allocation. The framework utilizes the Trojan vulnerability measurement as player utilities, and it could work at all circuit verification levels. A new Stackelberg security game specific to hardware security is also proposed. The new game model minimizes the defender utility loss with limited verification resources restriction. The paper also proposed RTL security vulnerability measurement to measure each data propagation path vulnerability quantitively and efficiently.

- The paper written by Jianhui Jiang et al. presents three gate-level approximate arithmetic circuit reliability estimation methods based on the probability gate model. The proposed fusion algorithm considers the effect of each fanout node on the reliability of the circuit separately and then uses a linear model to obtain the circuit reliability. The results on benchmark circuits show that the methods achieve higher accuracy and efficiency than the existing methods.

The second part of the special issue, consisting of two frameworks and an application implementation on the high-throughput cluster, focuses on the resource management and performance tuning for computing system. The heterogeneity of many-core processor also brings about new techniques to the frameworks. In addition, there is a paper to review the key challenges, mechanisms, and evaluations of FT-Matrix DSP series, which are important co-processors in modern computing systems.

- The paper written by Zichen Xu et al. proposes a control framework, CROP, to save power in database relational operations. In contrast to today's heuristic-based power tuning techniques, CROP uses a controller design based on control theory to minimize overshoot and ensure the shortest settling time. CROP adapts a fuzzy classifier to tune the sensitivity of the whole system control. The prototype of CROP wraps these functions in a container hierarchy. CROP is evaluated with various database benchmarks. Results show that Crop achieves up to 51.3% additional energy savings, compared to existing state of the practice methods.
- The paper written by Yibin Tang et al. investigates a realistic scenario when an on-line scheduler is needed to meet the requirement of latency even when the edge computing resources and communication speed are dynamically fluctuating, while protecting the privacy of users as well. It presents a real-time task scheduling method for privacy protection of neural networks in mobile-cloud systems, which can flexibly and dynamically allocate computation resources for the neural network applications, while satisfy different constraints of QoR and QoS as well. The approximate computing feature of neural networks and the trade-off for neural network propagation paths are also explored. The experiments on two sets of neural networks show that it significantly improves the energy efficiency of real-time neural networks on edge devices.
- The paper written by Dongrui Fan et al. presents a scalable and efficient implementation of graph traverse on High-throughput cluster (HTCs). HTCs adopt High-Throughput many-core architecture, which has the characteristics of high concurrency, strong real-time, and low-power consumption. Asynchronous virtual ring method, thread caching scheme and vertex ID reordering are proposed to improve graph traverse performance on HTCs. Evaluation shows its good scalability and performance superiority to existing work under the same cluster scale.
- The paper written by Yaohua Wang and Yang Guo et al. reviews two milestone Digital Signal Processors (DSPs): FT-Matrix and FT-Matrix2, which are designed by National University of Defense Technology with the purpose of advancing DSPs into the era of higher performance computing, AI, and even beyond. The key challenges, mechanisms, and evaluations of FT-Matrix DSP series are demonstrated. Possible future directions for enabling DSPs for a wider scope of applications are also described.

We would like to take this chance to thank all the authors and the reviewers for their brilliant contribution to this special issue of CCF THPC. Only with their great efforts, we can put together the nine research papers that discuss different topics, and present different ideas that help to optimize the resource management and performance tuning with different underlying architectures.

Jifeng He is a Chinese computer scientist and a permanent professor of East China Normal University. He was elected to the Chinese Academy of Sciences in 2005 and won the title of National Excellent Member of Chinese Communist Party in 2016, respectively. His research interests include sound methods for the specification of computer systems, communications, application and standards, and techniques for designing and implementing those specifications in software and/or hardware with high reliability.

Chenggang Wu is currently a professor at the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences. He has been engaged in theresearch of computer system security, program analysis and virtualization technology for a long time, aiming at improving the security and reliability of computers with systematic solutions. He served as the general co-chair of the International academic Conference CGO 2013, the Steering Committee member of CGO, the Program co-chair of the APPT 2013, and the member of the program Committee of PPoPP2017, PPoPP2018, CGO2015, CGO2016, CGO2017, and PLDI2012.

Huawei Li is currently a professor at the State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences. Her research interests include testing of very large-scale integration/ SoC circuits, approximate computing architecture and machine learning accelerators. She currently serves as the Secretary General of the China Computer Federation (CCF) Technical Committee on Integrated Circuit Design, serves as an Associate Editor of the IEEE TRANSACTION ON VERY LARGE-SCALE INTEGRATION (VLSI) SYSTEMS, and serves as the Steering Committee Chair of the IEEE Asian Test Symposium (ATS).

Yang Guo received his Ph.D. degree from National University of Defense Technology, Changsha, China in 1999. Currently he is a professor at the university, where he leads the digital signal processor group and is the director of the Integrated Circuits. He has authored or co-authored more than 50 publications on journals and conference proceedings. His primary research interests include low power VLSI circuits, microprocessor design and verification, electronic design automation (EDA) techniques for VLSI circuits.

Tao Li is currently a professor of Nankai University, China. He was a visiting professor at the University of Minnesota at Twin Cities, USA, from 2013 to 2014. His present research interests include heterogeneous computing, deep learning, intelligent internet of things. He is a distinguished member of CCF.