Special Issue on Commercial Products 2021

Hsien-Hsin S. Lee, *Facebook, Inc., Cambridge, MA, 02142, USA*

It gives me much pleasure to introduce this Special Issue on Commercial Products 2021. This issue is intended to serve as a venue for communicating the latest technological innovation in microprocessors, accelerators, system architectures, and large-scale system optimization technologies from industries. Aside from our external call-for-papers, we also proactively recruited papers including those thoroughly reviewed and recommended by the 48th ACM/IEEE International Symposium on Computer Architecture (ISCA-48) Industry Track but could not be included in the final program due to the limited number of papers the conference could accommodate. At the end, we accepted five papers to appear in this Special Issue.

In "Kunpeng 920: The First 7-nm Chiplet-Based 64-Core ARM SoC for Cloud Services," Xia *et al.* from Huawei Hisilicon describe an ARM-based Kunpeng 920 SoC design as a high-performance solution in supporting Huawei's services with cloud-edge collaborative computing models. In this article, they describe a LEGO-style, chiplet-based SoC architecture with CC-NUMA fabric. Their chiplet design based on TSMC CoWoS integration enables low cost and design reuse for various applications across computing, communication, and AI.

In "I-DVFS: Instantaneous Frequency Switch During Dynamic Voltage and Frequency Scaling," Gendler *et al.* discuss a low-latency mechanism called I-DVFS to perform instantaneous frequency switching during DVFS, which has been employed by Intel's Tiger Lake and Ice Lake CPUs. The frequency transition during DVFS is decoupled from IP execution, allowing the IP blocks to remain functional without stall, frequenting the use of DVFS and achieving high power efficiency.

In "ACCL: Architecting Highly Scalable Distributed Training Systems With Highly Efficient Collective Communication Library," Dong *et al.* from Alibaba Group describe the Alibaba Collective Communication Library (ACCL), a high-efficiency collective communication library for scalable distributed model training in datacenters. The article proposes a hybrid algorithm to facilitate the simultaneous use of multiple heterogeneous interconnects in one collective operation to improve the allreduce operations and the end-to-end performance of model training.

In "Low-Precision Hardware Architectures Meet Recommendation Model Inference at Scale," Deng *et al.* from Facebook study low precision arithmetic techniques for improving the efficiency of recommendation inference serving while meeting the stringent accuracy requirements for Facebook's production recommender systems. This article discusses a suite of tools that assist the automation of quantization model search and its iterative workflow. Also presented are several lessons learned including quantization support, balanced FP and INT performance, and their implications to nonlinear activation functions.

Finally, "Datacenter-Scale Analysis and Optimization of GPU Machine Learning Workloads" is another Facebook article that aims at performance optimization for GPU servers running deep learning training. In this work, Wesolowski *et al.* show how to use introspective tools to profile and collect system-wide metrics for various machine learning workflows for training jobs and then use them to identify the performance optimization opportunities across the entire execution stack for GPU fleet in Facebook's datacenters.

ACKNOWLEDGMENTS

I would like to thank all the authors who submitted their work to this special issue to share their learning with the community. I would also like to thank all the expert reviewers who agreed to review these manuscripts and provided their valuable feedback. Finally, I would like to thank *IEEE Micro* Editor-in-Chief Prof. Lizy K. John for her guidance in the entire process. I hope you enjoy the papers we selected for this issue.

HSIEN-HSIN S. LEE is an Area Research Lead at Facebook AI Research. His work investigates inter-domain optimization across computer architecture, machine learning, and semiconductor technologies. He is a Fellow of IEEE. Contact him at lee.sean@gmail.com.

0272-1732 © 2021 IEEE
Digital Object Identifier 10.1109/MM.2021.3103240
Date of publication 13 August 2021; date of current version 14 September 2021.