Guest Editorial

Pavan Balaji, *Senior Member, IEEE*, Jidong Zhai, and Min Si

This special section presents the state-of-the-art technologies and the challenges of parallel and distributed computing techniques for artificial intelligence (AI), machine learning (ML), and deep learning (DL). AI, ML, and DL have established themselves in a multitude of domains because of their ability to process and model unstructured input data.

We thank all the authors for their submissions. The selection process from the 102 submissions involved multiple stages. In the first review round, we had seven minor revisions, 34 major revisions, and one paper that was transferred to the regular track of TPDS. In the second revision round, 15 papers were accepted together with 18 papers under a minor revision decision. In the third revision round, 17 papers were accepted and one paper was transferred to the regular track of TPDS. In total, 32 papers were accepted for the special section, leading to a final acceptance percentage of 31 percent.

The papers in the special section cover diverse topics related to the convergence of parallel and distributed computing and AI/ML/DL. Of the accepted papers, five present federated learning techniques on distributed systems: "Biscotti: A Blockchain System for Private and Secure Federated Learning," "Mutual Information Driven Federated Learning," "Accelerating Federated Learning over Reliability-Agnostic Clients in Mobile Edge Computing Systems," "An Efficiency-boosting Client Selection Scheme for Federated Learning with Fairness Guarantee," and "FedSCR: Structure-based Communication Reduction for Federated Learning."

Five papers focus on edge computing for mobile and embedded systems: "Learning Spatiotemporal Failure Dependencies for Resilient Edge Computing Services," "Accelerating Gossip-based Deep Learning in Heterogeneous Edge Computing Platforms," "Distributed Task Migration Optimization in MEC by Extending Multi-agent Deep Reinforcement Learning Approach," "Systematically Landing Machine Learning onto Market-Scale Mobile Malware Detection," and "A Game-based Approach for Cost-aware Task Assignment with QoS Constraint in Collaborative Edge and Cloud Environments."

Five papers discuss convolutional neural networks: "The Case for Strong Scaling in Deep Learning: Training Large 3D CNNs with Hybrid Parallelism," "A Hybrid Fuzzy Convolutional Neural Network Based Mechanism for Photovoltaic Cell Defect Detection with Electroluminescence Images," "Model Parallelism Optimization for Distributed Inference via Decoupled CNN Structure," "FT-CNN: Algorithm-Based Fault Tolerance for Convolutional Neural Networks," and "SmartTuning: Selecting HyperParameters of a ConvNet System for Fast Training and Small Working Memory."

Four papers present research on parallel patterns, synchronous and scalability analysis: "The Scalability for Parallel Machine Learning Training Algorithm: Dataset Matters," "A Runtime and Non-Intrusive Approach to Optimize EDP by tuning Threads and CPU Frequency for OpenMP Applications," "Breaking (Global) Barriers in Parallel Stochastic Optimization with Wait-Avoiding Group Averaging," and "Petrel: Heterogeneity-aware Distributed Deep Learning via Hybrid Synchronization" [1].

Four papers discuss benchmarking techniques, AutoML, network compression, and privacy-preserving: "iMLBench: A Machine Learning Benchmark Suite for CPU-GPU Integrated Architectures," "A Distributed Framework For EA-Based NAS," "Parallel Blockwise Knowledge Distillation for Deep Neural Network Compression," and "Privacy-Preserving Computation Offloading for Parallel Deep Neural Networks Training".

Three papers are about scheduling and resource allocation: "Fine-grained Powercap Allocation for Power-constrained Systems based on Multi-objective Machine Learning," "DLBooster: Boosting End-to-End Deep Learning Workflow with Data Preprocessing and Scheduling Codesign," and "A Probabilistic Machine Learning Approach to Scheduling Parallel Loops with Bayesian Optimization."

Two papers describe work on sparse computation: "SGD_Tucker: A Novel Stochastic Optimization Strategy for Parallel Sparse Tucker Decomposition" and "Adaptive SpMV/SpMSpV on GPUs for Input Vectors of Varied Sparsity."

Two papers present research on FPGAs: "Improving HW/SW Adaptability for Accelerating CNNs on FPGAs through A Dynamic/Static Co-Reconfiguration Approach" and "Efficient Methods for Mapping Neural Machine Translator on FPGAs."

Two papers explore AI acceleration using GPUs: "Accelerating Binarized Neural Networks via Bit-Tensor-Cores in Turing GPUs" and "EDGES: An Efficient Distributed GraphEmbedding System on GPU clusters."

We thank all our committee members for their hard work and many contributions that have made the special section a reality in this special year 2020. We also thank the community for your interest in this special section. We hope you can use this information to advance your research in parallel and distributed computing for AI/ML/DL or in many other fields that rely on parallel computing for AI for insight, advances, and breakthroughs.

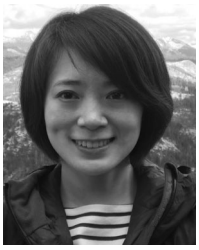
Digital Object Identifier no. 10.1109/TPDS.2020.3047357

Reference

[1] Q. Zhou *et al.*, "Petrel: Heterogeneity-aware distributed deep learning via hybrid synchronization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 5, pp. 1030–1043, May 2021.

Pavan Balaji (Senior Member, IEEE) holds appointments as a senior computer scientist and group lead at the Argonne National Laboratory, where he leads two groups: Programming Models and Runtime Systems and Future Architectures for AI. His research interests include parallel programming models and runtime systems for communication and I/O on extreme-scale supercomputing systems, modern system architecture, cloud computing systems, data-intensive computing, deep learning, and big-data sciences. He has more than 200 publications in these areas and has delivered more than 200 talks and tutorials at various conferences and research institutes. He is the recipient of several awards including the IEEE TCSC Award for Excellence in Scalable Computing (Middle Career), in 2015; TEDxMidwest Emerging Leader Award, in 2013; U.S. Department of Energy Early Career Award, in 2012; Crain's Chicago 40 under 40 Award, in 2012; Los Alamos National Laboratory Director's Technical Achievement Award, in 2005; Ohio State University Outstanding Researcher Award, in 2005; best paper awards at PACT 2019, ACM HPDC 2018, IEEE ScalCom 2013, Euro PVM/MPI 2009, ISC 2009, IEEE Cluster 2008, Euro PVM/MPI 2008, ISC 2008; Best Paper Finalist at IEEE/ACM SC 2014; Best Poster Award at IEEE ICPADS 2018; Best Poster Finalist at IEEE/ACM SC 2014; and Best Student Poster Award at ICPP 2018. He has served as a chair or editor for more than 100 journals, conferences and workshops, and as a technical program committee member in numerous conferences and workshops. He is a distinguished member of ACM.

Jidong Zhai is a tenured associate professor with the Department of Computer Science and Technology, Tsinghua University. He is the recipient of a IEEE TPDS Award for Editorial Excellence, NSFC Young Career Award, and CCF-IEEE CS Young Computer Scientist Award. His research interests include high-performance computing, performance evaluation, and compiler optimization. He is on the editorial board of *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Cloud Computing*, and *Journal of Parallel and Distributed Computing*.



Min Si is an assistant computer scientist at Argonne National Laboratory. She is the recipient of the 2018 IEEE-CS Technical Consortium on High Performance Computing (TCHPC) Early Career Researchers Award for excellence in high performance computing and won the Karsten Schwan Best Paper Award at HPDC 2018. Her research interests include high-performance computing, runtime systems, and parallel programming models.