ZHEN XIE

(first name is pronounced like "J-EH-N")

Engineering Building N14,
Department of Computer Science,
Binghamton, NY 13902-6000

http://zhen-xie.com
Phone: (607) 777-4595
Email: zxie3@binghamton.edu

PARTICULARS

EMPLOYMENT

The State University of New York (SUNY) at Binghamton	Binghamton, NY, USA
Assistant Professor in Department of Computer Science	*Aug. 2023 - Present*
Argonne National Laboratory	Lemont, IL, USA
Postdoctor in Leadership Computing Facility	*Aug. 2021 - Jul. 2023*
University of California	Merced, CA, USA
Postdoctor in Electrical Engineering and Computer Science	*Aug. 2019 - Jul. 2021*

EDUCATION

University of Chinese Academy of Sciences	Beijing, China
Doctor Degree in Computer Science	*Sep. 2013- Jun. 2019*
Wuhan University Of Technology	Wuhan, China
Bachelor Degree in Computer Science and Technology	*Sep. 2009 - Jun. 2013*

RESEARCH INTERESTS

My research area is High-Performance Computing (HPC) with a focus on the interaction between machine learning algorithms and system-level performance optimization. I am working on (i) **System for Machine Learning**: building modern ML/DL algorithms and systems on heterogeneous and parallel HPC architectures (e.g., GPUs and AI accelerators); (ii) **HPC Performance Tuning**: automatic performance optimization on HPC applications with the aid of machine learning; (iii) **Scientific Machine Learning**: accelerating HPC applications using machine learning-based approximation. My work has been published in multiple top-tier conferences and journals, including PPoPP, SC, ICS, EuroSys, Euro-Par, TPDS, and TACO, and has received some awards, such as ACM Gordon Bell Special Prize.

PUBLICATIONS

CONFERENCE PUBLICATION

1. **TrainBF: High-Performance DNN Training Engine using BFloat16 on AI Accelerators.**
   Zhen Xie, Siddhisanket Raskar, Murali Emani, and Venkatram Vishwanath;
   29th International European Conference on Parallel and Distributed Computing [**Euro-Par'23**].
2. **Merchandiser: Data Placement on Heterogeneous Memory for Task-Parallel HPC Applications with Load-Balance Awareness.**
   Zhen Xie, Jie Liu, Jiajia Li, and Dong Li;
   27th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming [**PPoPP'23**] (acceptance rate: 23.7%).
3. **LB-HM: Load Balance-Aware Data Placement on Heterogeneous Memory for Task-Parallel HPC Application.**
   Zhen Xie, Jie Liu, Sam Ma, Jiajia Li, and Dong Li;
   Poster in 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming [**PPoPP'22**].
4. **MD-HM: Memoization-based Molecular Dynamics Simulations on Big Memory System.**
   Zhen Xie, Wenqian Dong, Jie Liu, Ivy Peng, Yanbao Ma, and Dong Li;
   35th ACM International Conference on Supercomputing [**ICS'21**] (acceptance rate: 24.2%).
5. **Tahoe: Tree Structure-Aware High Performance Inference Engine for Decision Tree Ensemble on GPU.**
   Zhen Xie, Wenqian Dong, Jiawen Liu, Hang Liu and Dong Li;
   16th ACM European Conference on Computer Systems, 2021 [**EuroSys'21**] (acceptance rate: 20.9%).

6. **IA-SpGEMM: an Input-aware Auto-tuning Framework for Parallel Sparse Matrix-Matrix Multiplication.**
   Zhen Xie, Guangming Tan, Weifeng Liu, and Ninghui Sun;
   33rd ACM on International Conference on Supercomputing [**ICS'19**] (acceptance rate: 23.3%);
7. **Enabling Energy-Efficient DNN Training on Hybrid GPU-FPGA Accelerators.**
   Xin He, Jiawen Liu, Zhen Xie, Hao Chen, Guoyang Chen, Weifeng Zhang, and Dong Li;
   35th ACM International Conference on Supercomputing [**ICS'21**] (acceptance rate: 24.2%).
8. **Smart-PGSim: Using Neural Network to Accelerate AC-OPF Power Grid Simulation.**
   Wenqian Dong, Zhen Xie, Gokcen Kestor, and Dong Li;
   32nd ACM/IEEE International Conference for High Performance Computing [**SC'20**] (acceptance rate: 22.3%);
9. **Adaptive Neural Network-Based Approximation to Accelerate Eulerian Fluid Simulation.**
   Wenqian Dong, Jie Liu, Zhen Xie, and Dong Li;
   31st ACM/IEEE International Conference for High Performance Computing [**SC'19**] (acceptance rate: 22.6%).
10. **Flame: A Self-Adaptive Auto-Labeling System for Heterogeneous Mobile Processors.**
    Jiawen Liu, Jie Liu, Zhen Xie, and Dong Li;
    ACM/IEEE Symposium on Edge Computing [**SEC'21**].
11. **RIANN: Real-time Incremental Learning with Approximate Nearest Neighbor on Mobile Devices.**
    Jiawen Liu, Zhen Xie, Dimitrios Nikolopoulos, and Dong Li;
    USENIX Conference on Operational Machine Learning [**USENIX OpML'20**].
12. **Modeling Traffic of Big Data Platform for Large Scale Datacenter Networks.**
    Zhen Xie, Zheng Cao, Zhan Wang, Dawei Zang, En Shao, and Ninghui Sun;
    22nd IEEE International Conference on Parallel and Distributed Systems [**ICPADS'16**] (acceptance rate: 29.9%);

GORDEN BELL PRIZE

13. **GenSLMs: Genome-Scale Language Models Reveal SARS-CoV-2 Evolutionary Dynamics.**
    Maxim Zvyagin, Alexander Brace, ..., Zhen Xie, ..., Venkatram Vishwanath, and Arvind Ramanathan
    ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research, [**Gordon Bell'22**].

JOURNAL PUBLICATION

14. **A Pattern Based SpGEMM Library for Multi-core and Many-core Architectures.**
    Zhen Xie, Guangming Tan, Weifeng Liu, and Ninghui Sun;
    IEEE Transactions on Parallel and Distributed Systems (**TPDS**), 2021;
15. **TLB-pilot: Mitigating TLB Contention Attack on GPUs with Microarchitecture-Aware Scheduling.**
    Bang Di, Daokun Hu, Zhen Xie, Jianhua Sun, Hao Chen, Jinkui Ren, and Dong Li;
    ACM Transactions on Architecture and Code Optimization (**TACO**), 2021;
16. **Revealing bottlenecks and predicting optimal performance of Sparse Matrix-Vector and Convolution using the Probability-Process-Ram model.**
    Zhen Xie, Guangming Tan, and Ninghui Sun;
    Computer Research and Development, 2020;
17. **PRF : A Process-RAM-Feedback Performance Model to Reveal Bottlenecks and Propose Optimizations.**
    Zhen Xie, Guangming Tan, and Ninghui Sun;
    High Technology Letters, 2019;

WORKSHOP PUBLICATION

18. **Transfer Learning Across Heterogeneous Features For Efficient Tensor Program Generation.**
    Gaurav Verma, Siddhisanket Raskar, Zhen Xie, Abid M Malik, Murali Emani, Barbara Chapman;
    2nd International Workshop on Extreme Heterogeneity Solutions, [**ExHET'23**].

19. **Throughput-oriented and Accuracy-aware DNN Training with BFloat16 on GPU.**
Zhen Xie, Sid Raskar, Murali Emani;
Workshop on Scalable Deep Learning over Parallel and Distributed Infrastructures (ScaDL), [**IPDPS-W'22**].
20. **A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads.**
Murali Emani, Zhen Xie, Sid Raskar, etc.
13th IEEE International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems, [**PMBS'22**].
21. **Flame: A Self-Adaptive Auto-Labeling System for Heterogeneous Mobile Processors.**
Jiawen Liu, Jie Liu, Zhen Xie, and Dong Li;
On-Device Intelligence Workshop at Machine Learning and Systems Conference, [**MLSys-W'20**].

RESEARCH EXPERIENCE

A Comprehensive Evaluation of Novel AI Accelerators for Deep Learning Workloads
Oct 2021- Jun 2022

- **Work:** We introduce an overview of dataflow-based novel AI accelerators from SambaNova, Cerebras, Graphcore, and Groq. We present a first-of-a-kind evaluation of these accelerators with diverse workloads, such as Deep Learning (DL) primitives, benchmark models, and scientific machine learning applications.
- **Outcome:** This work has led to many key insights, challenges, and opportunities in integrating these novel AI accelerators in supercomputing systems.

TrainBF: High-Performance DNN Training Engine using BFloat16 on AI Accelerators
Aug 2021- May 2022

- **Work:** We present a high-performance DNN training engine to ensure competitive model training accuracy as single-precision training while maximizing the performance benefits of half-precision (BFloat16 format).
- **Outcome:** TrainBF outperforms the state-of-the-art mixed-precision training approach by 1.24×, 1.37×, and 1.16× on NVIDIA A100 GPU, AMD MI100 GPU, and a novel dataflow processor SambaNova, respectively.

MD-HM: Memoization-based Molecular Dynamics Simulations on Big Memory System
May 2020- April 2021

- **Work:** We trade memory capacity for computation capability to improve MD performance by the lookup table-based memoization technique. We introduce MD-HM, a memoization-based MD simulation framework customized for the big memory system.
- **Outcome:** MD-HM uses a new two-phase LSM-tree to optimize read/write performance. Evaluating with nine MD simulations, we show that MD-HM outperforms the state-of-the-art LAMMPS simulation with an average speedup of 7.6× based on Intel Optane-based big memory system. This work is published in ICS'21.

Tahoe: Tree Structure-Aware High Performance Inference Engine for Decision Tree Ensemble on GPU
Sep 2019- April 2020

- **Work:** Introduce an inference engine, Tahoe, for decision tree ensemble on GPU; Tahoe is adaptive to various tree structures by re-arranging node and tree layout in memory to improve memory access efficiency and avoid load imbalance, and by using the optimal data placement strategy to make best use of shared memory and reduce parallel reduction overhead.
- **Outcome** Tahoe consistently outperforms the state-of-the-art industry-quality inference engine FIL. Compared with FIL, Tahoe leads to an average of 5.31x, 3.67x and 4.05x speedup on Kepler, Pascal and Volta GPUs, respectively; This work is published in Eurosys'2021.

Smart-PGsim: Using Neural Network to Accelerate AC-OPF Power Grid Simulation
May 2019- June 2020

- **Work:** Accelerate a power-grid application using neural networks; Identify code regions to be approximated and extract features for neural network construction; Reformulate the solving method in the power grid simulation; Design an interactive learning model for multitask prediction; Enable physics-informed learning using domain knowledge; Implement neural network models with frameworks PyTorch.
- **Outcome:** Reduces simulation time by an average of 2.60× (up to 3.28×) without losing the optimality of the solution; This work is published in SC'20.

Smart-fluidnet: Adaptive Neural Network-Based Approximation to Accelerate Eulerian Fluid Simulation
Sep 2018- April 2019

- **Work:** Adaptively accelerate the Eulerian fluid simulation to meet the requirements on execution time and simulation quality; Ensemble multiple neural network models to build an adaptive runtime system to reach the user's requirement on simulation quality; Implement neural network models with Keras.
- **Outcome:** Smart-fluidnet is 1.46× and 590× faster than a state-of-the-art neural network model and the original fluid simulation respectively; This work is published in SC'19.

ACADEMIC HONORS

- Postdoctoral Fellowship at UC Merced, 2019 and 2020.
- China Merit Student Award for graduate students (Ph.D. students), 2014 and 2016.
- Prize of Excellence in Programmable Acceleration Card (PAC) competition at Intel, 2013
- The Excellent Graduation Thesis, 2013.
- The Outstanding Graduates of Wuhan University Of Technology, 2013.
- Second prize, "Wuhan University ACM National Software Challenge Competition", 2011.
- China National Scholarship for undergraduate students at Wuhan University Of Technology, 2011.

SERVICE

- Conference Reviewer: ICCD'23, ICCD'22, LCTES'21, ICS'21, IPDPS'21, IPDPS'20, NPC'20, IPDPS'19, ICPP'19, PPOPP'19, Cluster'19, NPC'19, SC'18, CCGrid'17, etc.
- PC member: EuroSys'23, AI4S'23
- Appointed Journal Reviewers: TPDS, TECS, JPDC, and JHPC.
- Student Volunteer at ICS'19, ICS'18, ICPADS'16.

TEACHING EXPERIENCE

- **Lecturer.** CS-457-01/CS-557-01: Intro To Distributed Systems, Fall 2023, Binghamton University.
- **Teaching Assistant.** CSE 375: Principle and Practice of Parallel Computing, Prof. Haixiang Lin, Fall 2014 and Spring 2015, University of Chinese Academy of Sciences.
- **Teaching Assistant.** CSE 163: Introduction to Computer Concepts and Programming, Prof. Zhimin Tang, Spring 2014, University of Chinese Academy of Sciences.
- **Teaching Assistant.** CSE 347: Data Mining, Prof. Ying Liu, Fall 2013, University of Chinese Academy of Sciences.

SELECTED PRESENTATIONS

1. [**SDL Workshop**] Distributed Deep Learning.
   In Simulation, Data, and Learning Workshop for AI, Lemont, IL, October, 2021.
2. [**Argonne Seminar**] Performance Optimization of ML and HPC applications on Heterogeneous Systems.
   In Argonne National Laboratory, Virtual conference, August, 2021.
3. [**ICS'21**] MD-HM: Memoization-based Molecular Dynamics Simulations on Big Memory System.
4. Performance Prediction and Optimization of Floating Point Operating Patterns, Invited talk at China University Of Petroleum, Beijing, Jun, 2019.
5. [**ICS'19**] IA-SpGEMM: an Input-aware Auto-tuning Framework for Parallel Sparse Matrix-Matrix Multiplication.
   In 33rd ACM International Conference on Supercomputing, Phoenix, AZ, June, 2019.
6. [**HPCaML'19**] Auto-tuning Parallel Sparse Matrix-Matrix Multiplication by Deep Learning
   In the First International Workshop on the Intersection of High Performance Computing and Machine Learning, HPCaML, Washington DC, February, 2019.

OPEN SOURCE COMMUNITY

- IA-SpGEMM: An Input Auto-tuning Sparse General Matrix-Matrix Multiplication on Multicore and Manycore Architecture. Link
- Tahoe: a high-performance GPU inference for decision tree ensemble. Link
- Smart-Fluidnet: a framework that automates model generation for fluid dynamic simulation. Link
- ALCF: Simulation, Data, and Learning Workshop for AI. Link

REFERENCES

FROM ACADEMIA

Prof. Dong Li
Associate Professor
Dept. of Electrical Engg. & Comp. Sc.
University of California, Merced
Merced, CA, 95340
dli35@ucmerced.edu

Prof. Jiajia Li
Assistant Professor
Dept. of Computer Science
North Carolina State University
Raleigh, NC, 27695
jiajia.li@ncsu.edu

Prof. Hang Liu
Presidential Fellowship AP
Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
hliu77@stevens.edu

FROM NATIONAL LABORATORY

Dr. Prasanna Balaprakash
Computer Science Leader
Mathematics and Computer Science
Argonne National Laboratory
Lemont, IL 60439
pbalapra@anl.gov

Dr. Murali Krishna Emani
Computer Scientist
Argonne Leadership Computing Facility
Argonne National Laboratory
Lemont, IL 60439
memani@anl.gov

Dr. Rajeev Thakur
Deputy Director & IEEE Fellow
Data Science and Learning
Argonne National Laboratory
Lemont, IL 60439
thakur@anl.gov