ZHEN XIE

Building 240 Room 4E17,
9700 S. Cass Ave,
Lemont, IL-60439

http://zhen-xie.com
Phone: (209) 446-7069
Email: zhen.xie@anl.gov

PARTICULARS

EMPLOYMENT

Argonne National Laboratory — Lemont, IL, USA
Postdoctor in Leadership Computing Facility — *Aug. 2021 - Present*

University of California — Merced, CA, USA
Postdoctor in Electrical Engineering and Computer Science — *Aug. 2019 - Aug. 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, numerical computation solvers, and system-level performance optimization. I am working on (i) **Automatic Performance Tuning**: Automatically speed up HPC and AI/DL applications on various parallel architectures, (ii) **Heterogeneous Computing and Memory Management**: optimizing computing and memory utilization on heterogeneous architectures, and (iii) **Scientific Machine Learning**: accelerating HPC applications using machine learning-based approximation.

PUBLICATIONS

CONFERENCE PUBLICATION

1. **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 ACM 26th SIGPLAN Symposium on Principles and Practice of Parallel Programming [**PPoPP'22**].
2. **MD-HM: Memoization-based Molecular Dynamics Simulations on Big Memory System.**
   Zhen Xie, Wenqian Dong, Jie Liu, Ivy Peng, Yanbao Ma, and Dong Li;
   ACM 35th International Conference on Supercomputing [**ICS'21**] (acceptance rate: 24.2%).
3. **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;
   ACM 16th European Conference on Computer Systems, 2021 [**EuroSys'21**] (acceptance rate: 20.9%).
4. **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;
   ACM 35th International Conference on Supercomputing [**ICS'21**] (acceptance rate: 24.2%).
5. **Smart-PGSim: Using Neural Network to Accelerate AC-OPF Power Grid Simulation.**
   Wenqian Dong, Zhen Xie, Gokcen Kestor, and Dong Li;
   ACM/IEEE 32nd International Conference for High Performance Computing [**SC'20**] (acceptance rate: 22.3%);
6. **Adaptive Neural Network-Based Approximation to Accelerate Eulerian Fluid Simulation.**
   Wenqian Dong, Jie Liu, Zhen Xie, and Dong Li;
   ACM/IEEE 31st International Conference for High Performance Computing [**SC'19**] (acceptance rate: 22.6%).

7. **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**].

8. **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**].

9. **IA-SpGEMM: an Input-aware Auto-tuning Framework for Parallel Sparse Matrix-Matrix Multiplication.**
Zhen Xie, Guangming Tan, Weifeng Liu, and Ninghui Sun;
ACM 33rd on International Conference on Supercomputing [**ICS'19**] (acceptance rate: 23.3%);

10. **Modeling Traffic of Big Data Platform for Large Scale Datacenter Networks.**
Zhen Xie, Zheng Cao, Zhan Wang, Dawei Zang, En Shao, and Ninghui Sun;
IEEE 22nd International Conference on Parallel and Distributed Systems [**ICPADS'16**] (acceptance rate: 29.9%);

JOURNAL PUBLICATION

11. **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;

12. **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;

13. **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;

14. **PRF : A Process-RAM-Feedback Performance Model to Reveal Bottlenecks and Propose Optimizations.**
Zhen Xie, Guangming Tan, and Ninghui Sun;
High Technology Letters, 2019;

ACADEMIC HONORS

- Postdoctoral Fellowship at UC Merced, 2019 and 2020.
- China Merit Student Award for graduate students (Ph.D. student), 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.
  2011 Wuhan University ACM Challenge National software station

SERVICE

- Conference Reviwer for 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.
- Appointed Journal Reviewers for TPDS, TECS, and JHPC.
- Student Volunteer at ICS'19, ICS'18, ICPADS'16.

RESEARCH EXPERIENCE

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; Evaluate Tahoe with 15 common datasets on three generations of GPU based on Kepler, Pascal and Volta microarchitectures; 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 5.31x, 3.67x and 4.05x speedup (up to 9.58x, 8.77x, and 10.14x) for high parallelism tasks and 2.34x, 1.52x and 1.45x speedup (up to 5.08x, 3.82x, and 3.17x) for low parallelism tasks 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 (particularly, the AC-OPF problem); 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.

IA-SpGEMM: an Input-aware Auto-tuning Framework for Parallel Sparse MatrixMatrix Multiplication
April 2018- Jan 2019

- **Work:** Sparse matrix-matrix multiplication (SpGEMM) is a sparse kernel that is used in a number of scientific applications. We propose IA-SpGEMM, an input-aware auto-tuning Framework for SpGEMM, that provides a unified programming interface in the CSR format and automatically determines the best format and algorithm for arbitrary sparse matrices.
- **Outcome** We evaluate our framework on CPUs and a GPU, and the results show that IA-SpGEMM is on average 3.27x and 13.17x faster than MKL on an Intel and an AMD platform, respectively, and is 2.23x faster than cuSPARSE on an NVIDIA GPU; This work is published in ICS'19.

TEACHING EXPERIENCE

- **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.

TALKS

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.
   In 35th ACM International Conference on Supercomputing, Virtual conference, June, 2021.
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.

WORKSHOPS

- **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**].
- **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**].

OPEN SOURCE COMMUNITY

- IA-SpGEMM: a An Input Auto-tuning Sparse General Matrix-Matrix Multiplication on Multicore and Manycore Architure. 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. Ninghui Sun;
Professor and Academician
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China 100190
snh@ict.ac.cn

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

FROM NATIONAL LABORATORY

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

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