KAI WU

5200 N Lake Road ◇ SE2 Room 213C ◇ Merced, CA 95348
(517)-763-1599 ◇ kwu42@ucmerced.edu ◇ kaikylewu.com

RESEARCH INTERESTS

My research area is computer systems with a focus on memory and storage systems. I design high-performance computer systems for high performance computing (HPC), cloud, database and AI systems. My recent work focuses on non-volatile memory (persistent memory).

Computer system | High performance computing | Large-scale distributed system | Cache/Memory/Storage system characterization and performance optimization | Runtime scheduling | System resilience and reliability | Data processing framework | OLTP/OLAP/HTAP| Key-value store

EDUCATION

University of California, Merced, CA, USA *June 2016 - Present*
Ph.D., in Electrical Engineering and Computer Sciences

Michigan State University, East Lansing, MI, USA *Aug 2014 - May 2016*
M.S., in Computer Science and Engineering

Harbin Normal University, Harbin, CHINA *Aug 2010 - Jun 2014*
B.S., Digital Media Technology

RESEARCH EXPERIENCE

University of California, Merced June 2016 - Present
Research Assistant *Work with with Prof. Dong Li*

Data management (placement & migration) on heterogeneous memory-based supercomputing systems

- Proposed a runtime system to guide data movement for *MPI*-based HPC applications on heterogeneous memory based large-scale supercomputer (up to 128 nodes); this design applied a sampling-based online memory access profiling on hardware performance counters for detecting hot data; achieved 28% performance improvement over state-of-the-art solutions. (Published in SC'17 & JCST)
- Proposed a runtime data management system for *OpenMP-task* applications on heterogeneous memory-based large-scale supercomputer; this design is featured with a hybrid performance model combines machine learning modeling and analytical modeling to capture complex memory access patterns of tasks; outperformed by two state-of-the-art systems by 24%. (Published in SC'18)

High-performance data consistence on persistent memory

- Designed and developed a runtime system that improves the performance of the cache-line flushing mechanism on Intel Optane DC persistent memory to achieve high performance data consistence; optimization techniques include concurrency control of cache line flushing, proactive cache line flush, and coalescing cache line flushing; achieved up to 49.8% speedups. (Published in PACT'20)
- Characterized the recomputation ability of HPC applications on persistent memory without crash consistency protection, and developed a framework that uses a systematic approach to decide how to selectively persist application data objects to significantly increase the possibility of successful recomputation. (Published in CLUSTER'20 & MCHPC'18)
- Designed an algorithm-based method to establish lightweight crash consistency on the persistent memory of HPC applications; the proposed method only needs to slightly extend the application data structure or sparsely flush cache blocks. (Published in CLUSTER'17)

- Designed a high-performance transaction system for OLTP workloads on persistent memory; optimizations include avoiding log, random writes, small writes, and fragmentation.

System evaluation of emerging Non-volatile memory devices

- Evaluated the performance of memory-intensive HPC applications on Intel Optane DC persistent memory; identified two bottlenecks arising from the asymmetric bandwidth and threads scaling limitation on Optane, and proposed two optimization techniques. (Published in IPDPS'20)
- Evaluated the performance of I/O intensive HPC applications on NVMe SSD and verified efficiency of current I/O mechanisms (e.g., POSIX I/O, MPI I/O, page cache, etc.). (Publised in NAS'17)

Bytedance Inc.(System infrastructure lab) May 2020 - Present
Research Intern

- Researched on hybrid transactional/analytical processing (HTAP) system; explored adaptive data compaction algorithms to effectively process petabytes of data.

Lawrence Livermore National Laboratory May 2018 - Aug 2018
Research Intern *Work with Dr.Maya B Gokhale*

- Explored pre-fetch and eviction optimizations using the user faulted approach for efficient access to persistent memory for data-intensive applications such as out-of-core graph applications and near-earth asteroid detection applications; achieved up to 2.5x speedups. (Publised in MCHPC'19)

Los Alamos National Laboratory May 2017 - Aug 2017
Research Intern *Work with Dr.Nathan DeBardeleben and Dr.Qiang Guan*

- Built an analytical model to predict the fault injection result of the application running in large-scale based on fault injection results of the application running in small-scale and serial. (Publised in ICPP'18 and SC'17)

PUBLICATIONS

Referred Conference & Workshop Papers

[PACT'20] Kai Wu, Ivy B. Peng, Jie Ren and Dong Li. "Ribbon: High Performance Cache Line Flushing for Persistent Memory". In 29th International Conference on Parallel Architectures and Compilation Techniques, 2020.

[IPDPS'20] Ivy B. Peng, Kai Wu, Jie Ren, Dong Li and Maya Gokhale. "Demystifying the Performance of HPC Scientific Applications on NVM-based Memory Systems". In 34rd IEEE International Parallel and Distributed Processing Symposium, 2020.

[CLUSTER'20] Jie Ren, Kai Wu and Dong Li. "Exploring Non-Volatility of Non-Volatile Memory for High Performance Computing Under Failures". In 22th IEEE Cluster Conference, 2020.

[MCHPC'19] Ivy B. Peng, Marty McFadden, Eric Green, Keita Iwabuchi, Kai Wu, Dong Li, Roger Pearce and Maya Gokhale. "UMap: Enabling Application-driven Optimizations for Page Management". In Workshop on Memory Centric High Performance Computing held in conjunction with SC19.

[SC'18] Kai Wu, Jie Ren and Dong Li. "Runtime Data Management on Non-Volatile Memory-based Heterogeneous Memory for Task-Parallel Programs". In 30th ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, 2018.

[ICPP'18] Kai Wu, Wenqian Dong, Qiang Guan, Nathan Debardeleben and Dong Li. "Modeling Application Resilience in Large Scale Parallel Execution". In 47th International Conference on Parallel Processing, 2018.

[MCHPC'18] Jie Ren, Kai Wu and Dong Li. "Understanding Application Recomputability without Crash Consistency in Non-Volatile Memory". In Workshop on Memory Centric High Performance Computing held in conjunction with SC18.

[SC'17] Kai Wu, Yingchao Huang and Dong Li. "Unimem: Runtime Data Management in Non-Volatile Memory-based Heterogeneous Main Memory". In 29th ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, 2017.

[CLUSTER'17] Shuo Yang, Kai Wu, Yifan Qiao, Dong Li and Jidong Zhai. "Algorithm-Directed Crash Consistence in Non-Volatile Memory for HPC". In 19th IEEE Cluster Conference, 2017.

[NAS'17] Wei Liu, Kai Wu, Jialin Liu, Feng Chen and Dong Li. "Performance Evaluation and Modeling of HPC I/O on Non-Volatile Memory". In 12th International Conference on Networking, Architecture, and Storage, 2017.

Referred Poster Papers

[NVMW'19] Kai Wu, Jie Ren and Dong Li. "Architecture-Aware, High Performance Transaction for Persistent Memory". In 10th Workshop on Non-Volatile Memory, 2018.

[NVMW'19] Kai Wu, Jie Ren and Dong Li. "Runtime Data Management on Non-Volatile Memory-based Heterogeneous Memory for Task-Parallel Programs". In 10th Workshop on Non-Volatile Memory, 2018.

[NVMW'19] Jie Ren, Kai Wu and Dong Li. "EasyCrash: Exploring Non-Volatility of Non-Volatile Memory for High Performance Computing Under Failures". In 10th Workshop on Non-Volatile Memory, 2018.

[NVMW'18] Kai Wu and Dong Li. "Unimem: Runtime Data Management in Non-Volatile Memory-based Heterogeneous Main Memory". In 9th Workshop on Non-Volatile Memory, 2018.

[NVMW'18] Jie Ren, Kai Wu and Dong Li. "Algorithm-Directed Crash Consistence in Non-Volatile Memory for HPC". In 9th Workshop on Non-Volatile Memory, 2018.

[SC'17] Kai Wu, Qiang Guan, Nathan DeBardeleben and Dong Li. "Characterization and Comparison of Application Resilience for Serial and Parallel Executions". In 29th ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC), 2017.

Work Under Submission

Kai Wu and Dong Li. "Unimem: Runtime Data Management on Non-Volatile Memory-based Heterogeneous Main Memory for High Performance Computing".

Kai Wu, Jie Ren, Ivy Peng and Dong Li. "ArchTM: Architecture-Aware, High Performance Transaction for Persistent Memory".

Jie Ren, Jiaolin Luo, Kai Wu, Minjia Zhang, Hyeran Jeon and Dong Li. "Efficient Tensor Migration and Allocation on Heterogeneous Memory Systems for Deep Learning".

Jie Ren, Jiaolin Luo, Ivy Peng, Kai Wu and Dong Li. "Performance Analysis and Optimization of Electromagnetic Particle-In-Cell Method on Emerging Persistent Memory-based Platform".

Preprints

Kai Wu, Frank Ober, Shari Hamlin, Qiang Guan and Dong Li, "Early Evaluation of Intel Optane Non-Volatile Memory with HPC I/O Workloads". Technical Report, PASA Lab, UC Merced.

Kai Wu, Yingchao Huang and Dong Li. "High Performance Data Persistence in Non-Volatile Memory for Resilient High Performance Computing". Technical Report, PASA Lab, UC Merced.

PROFESSIONAL SERVICES

Reviewers: NPC'20, CLUSTER'20, NPC'19, ICPP'19, SC'18, IPDPS'17, CLUSTER'17, HPCC'17, NAS'17.
Student Volunteer: SC'20 SC'19, SC'18, SC'16
Graduate Student Representative (UC Merced EECS), Jan 2020 - Present

AWARDS

Student Travel Award: NVMW'20, SC'19, NVMW'19, SC'18, OSDI'18, ASPLOS'18, NVMW'18, CLUSTER'17, NVMW'17, SC'16
Graduate Travel Fellowship, University of California, Merced, 2018 & 2020
Bobcat Graduate Research Fellowship, University of California, Merced 2017
First-Prize, LanQiao Cup National Collegiate Programming Contest, C/C++ group, 2013
Honorable Mention, ACM/ICPC International Collegiate Programming Contest China Tonghua Invitational Contest, 2013
Honorable Mention, ACM/ICPC International Collegiate Programming Contest China Hei Longjiang Province Contest, 2013
China National Scholarship, 2013

PROGRAMMING SKILLS

C/C++, Python, Fortran
Linux kernel programming
GPU(CUDA), MPI, OpenMP, PMDK(Persistent memory programming)

TEACHING

University of California, Merced 2016 Summer & 2018 Fall & 2019 Fall & 2020 Spring
Teaching Assistant

- CSE15 - Discrete Mathematics
- CSE20 - Introduction to Computing I (Java Basic)
- CSE179 - Parallel Computing