XULONG TANG

210 S. Bouquet Street, Sennott Square 6115, Pittsburgh, PA, 15232
Tel: (412) 624-8419
Email: tax6@pitt.edu
Homepage: http://xzt102.github.io/

EDUCATION EXPERIENCE

2014.8-2019.7 **Pennsylvania State University**
Ph.D. in Computer Science and Engineering
Advisor: Dr. Mahmut Taylan Kandemir

2014.1-2014.7 **College of William and Mary**
Ph.D. in Computer Science
Transfer to Pennsylvania State University in 2014 fall

2010.8-2013.12 **University of Science and Technology of China**
M.E. in Computer Science and Technology
Advisor: Dr. Hong An

2006.8-2010.7 **Harbin Institute of Technology**
B.E. in Computer Science and Technology
Advisor: Dr. Chunqi Sun

RESEARCH EXPERIENCE

2019 - present **University of Pittsburgh**
Assistant Professor

- Optimizing emerging applications on single- and multi- GPU systems.
  - Exploring application (e.g., deep learning applications) characteristics and algorithm innovations.
  - Developing efficient compiler optimizations and runtime management.
  - Designing architectures and system features toward next-generation GPU systems.
- Building efficient quantum computing simulation Eco-system.
  - Leveraging innovative system optimizations to simulate large and complex quantum circuits.
  - Developing both front-end and back-end quantum compiler supports for complex quantum circuits.
  - Building software-hardware co-optimizations for domain-specific (e.g., quantum chemical applications) quantum applications.
  - Exploring heterogeneous system designs that consists of quantum devices and classical devices.
- A Software and Hardware Co-Design for Addressing the Performance Bottlenecks in Secure Non-Volatile Memory.
  - Developing zero-copy page remapping and parallel remapping and computation techniques for SGX.
  - Developing compiler-assisted user controlled paging.
  - Designing high performance persistence support in secure NVM.

2014 - 2019 **Pennsylvania State University**
Research Assistant/Teaching Assistant
Advisor: Dr. Mahmut Taylan Kandemir, Dr. Chita R. Das

- Optimize GPU dynamic parallelism for irregular applications
- Investigate compiler-assisted optimizations for computation assignment and data access on manycore platforms

2017 Fall **Advanced Micro Devices (AMD Research)**
Co-op Engineer
Mentor: Bradford M. Beckmann, Sooraj Puthoor
- Participate in the project of prototyping the next generation GPUs. Explore efficient runtime task management on GPUs
- Reduce oversubscribing of command queues in GPUs

2015 Summer **SAMSUNG Research America (SRA)**
Research Intern
Mentor: Liangjun Zhang
- Model the memory hierarchy of high-performance, low-power mobile GPUs

2014 Spring **College of William and Mary**. *Compilers and Adaptive Programming Systems Lab*
Research Assistant
Advisor: Dr. Xipeng Shen
- Investigate the reasons of performance degradation on integrated CPU-GPU processors

2010 - 2013 **ICT of Chinese Academy of Science, Beijing**
Research Assistant
Advisor: Dr. Dongrui Fan
- Build a two-layer video codec benchmark suite
- Redesign x264 codec into a fine-grain pipelined version to achieve task-level parallelism

2010 - 2011 **University of Science and Technology of China (USTC)**
Research Assistant
Advisor: Dr. Hong An
- Propose adaptive scheduling based on characterization of dynamic GPU behaviors

PUBLICATIONS

Pitt Students from My Group.

* The authors contribute equally.

[C1]. Bingyao Li, Jieming Yin, Anup Holey, Youtao Zhang, Jun Yang, **Xulong Tang**, "Trans-FW: Short Circuiting Page Table Walk in Multi-GPU Systems via Remote Forwarding.", *In Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 91/364 = 25%* **(HPCA 2023)**

[C2]. Yue Dai, Youtao Zhang, **Xulong Tang**, "CEGMA: Coordinated Elastic Graph Matching Acceleration for Graph Matching Networks.", *In Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 91/364 = 25%* **(HPCA 2023)**

[C3]. Mehrnoosh Raoufi, Jun Yang, **Xulong Tang**, Youtao Zhang, "AB-ORAM: Constructing Adjustable Buckets for Space Reduction in Ring ORAM.", *In Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 91/364 = 25%* **(HPCA 2023)**

[C4]. Sheng Li*, Geng Yuan*, Yue Dai*, Youtao Zhang, Yanzhi Wang, **Xulong Tang**, "SmartFRZ: An Efficient Training Framework using Attention-Based Layer Freezing.", *Eleventh International Conference on Learning Representations.* **(ICLR 2023)**

[C5]. Yilun Zhao, Yanan Guo, Yuan Yao, Amanda Dumi, Devin M Mulvey, Shiv Upadhyay, Youtao Zhang, Kenneth D Jordan, Jun Yang, **Xulong Tang**, "Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation Using GPUs.", *In Proceedings of the 28th IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 80/273 = 29%* **(HPCA 2022)**

[C6]. Bingyao Li*, Qi Xue*, Geng Yuan*, Sheng Li, Xiaolong Ma, Yanzhi Wang, **Xulong Tang**, "Optimizing Data Layout for Training Deep Neural Networks.", *In Proceedings of the WWW '22: Companion Proceedings of the Web Conference 2022.* **(WWW 2022 workshop)**

[C7]. Yifan Gong, Geng Yuan, Zheng Zhan, Wei Niu, Zhengang Li, Pu Zhao, Yuxuan Cai, Sijia Liu, Bin Ren, Xue Lin, **Xulong Tang**, Yanzhi Wang, "Automatic Mapping of the Best-Suited DNN Pruning Schemes for Real-Time Mobile Acceleration.", *In Proceedings of ACM Transactions on Embedded Computing Systems.* **(TECS 2022)**

[C8]. Geng Yuan, Sung-En Chang, Qing Jin, Alec Lu, Yanyu Li, Yushu Wu, Zhenglun Kong, Yanyue Xie, Peiyan Dong, Minghai Qin, Xiaolong Ma, **Xulong Tang**, Zhenman Fang, Yanzhi Wang, "You Already Have It: A Generator-Free Low-Precision DNN Training Framework using Stochastic Rounding.", *In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference. Acceptance Ratio: 1645/5804 = 28%* **(ECCV 2022)**

[C9]. Geng Yuan, Yanyu Li, Sheng Li, Zhenglun Kong, Sergey Tulyakov, **Xulong Tang**, Yanzhi Wang, Jian Ren, "Layer Freezing & Data Sieving: Missing Pieces of a Generic Framework for Sparse Training.", *In Proceedings of the 36 the Conference on Neural Information Processing Systems. Acceptance Ratio: 25.6%* **(NeurIPS 2022)**

[C10]. Mahmut Taylan Kandemir, **Xulong Tang**, Jagadish Kotra, Mustafa Karakoy, "Fine-Granular Computation and Data Layout Reorganization for Improving Locality.", *In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design. Acceptance Ratio: 22%* **(ICCAD 2022)**

[C11]. Yajuan Du, Mingyang Liu, Yuqi Yang, Mingzhe Zhang, **Xulong Tang**, "Enhancing GPU Performance via Neighboring Directory Table Based Inter-TLB Sharing.", *In Proceedings of the IEEE International Conference on Computer Design. Acceptance Ratio: 69/228 = 30.2%* **(ICCD 2022)**

[C12]. Sebastien Ollivier, Sheng Li, Yue Tang, Chayanika Chaudhuri, Peipei Zhou, **Xulong Tang**, Jingtong Hu, Alex K. Jones, "Sustainable AI Processing at the Edge.", *In Proceedings of the IEEE Micro.* **(IEEE Micro)**

[C13]. F. Yu, Z. Xu, T. Shen, D. Stamoulis, L. Shangguan, D. Wang, M. Zhang, **X. Tang**, R. Madhok, C. Zhao, X. Li, N. Karianakis, D. Lymberopoulos, C. Liu, A. Li, Y. Chen, and X. Chen, "Rethinking Latency-aware DNN Design with GPU Tail Effect Analysis.", *Poster accepted in the 17th European Conference on Computer Systems (EuroSys).* **(EuroSys 2022 poster)**

[C14]. Yue Dai, **Xulong Tang**, Youtao Zhang, "An Efficient Segmented Quantization for Graph Neural Networks.", *In Proceedings of CCF Transactions on High Performance Computing.*

[C15]. Zhendong Wang, Xiaoming Zeng, **Xulong Tang**, Danfeng Zhang, Xing Hu, Yang Hu, "Demystifying Archhints for Model Extraction: An Attack in Unified Memory System.", *Arxiv.*

[C16]. Bingyao Li, Jieming Yin, Youtao Zhang, **Xulong Tang**, "Improving Address Translation in Multi-GPUs via Sharing and Spilling Aware TLB Design.", *In Proceedings of the 54th IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 94/430 = 21.8%* **(MICRO 2021)**

[C17]. Weizheng Xu, Ashutosh Pattnaik, Geng Yuan, Yanzhi Wang, Youtao Zhang, **Xulong Tang**, "ScaleDNN: Data Movement Aware DNN Training on Multi-GPU", *In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design. Acceptance Ratio: 121/514 = 23.5%* **(ICCAD 2021)**

[C18]. Fuxun Yu, Shawn Bray, Di Wang, Longfei Shangguan, **Xulong Tang**, Chenchen Liu, Xiang Chen, "Automated Runtime-Aware Scheduling for Multi-Tenant DNN Inference on GPU", *In Proceedings of the IEEE/ACM International Conference on Computer-Aided Design. Acceptance Ratio: 121/514 = 23.5%* **(ICCAD 2021)**

[C19]. **Xulong Tang**, Mahmut Taylan Kandemir, Mustafa Karakoy, "Mix and Match: Reorganizing Tasks for Enhancing Data Locality", *In Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS Journal). Acceptance Ratio: 15/124 = 12.1%* **(SIGMETRICS 2021)**

[C20]. Mahmut Taylan Kandemir, **Xulong Tang**, Hui Zhao, Jihyun Ryoo, Mustafa Karakoy, "Distance-in-Time versus Distance-in-Space", *In proceedings of 42nd annual ACM SIGPLAN conference on Programming Language Design and Implementation. Acceptance Ratio: 87/320 = 27%* **(PLDI 2021)**

[C21]. Huaipan Jiang, Haibo Zhang, **Xulong Tang**, Vineetha Govindaraj, Jack Sampson, Mahmut Taylan Kandemir, Danfeng Zhang, "Fluid: A Framework for Approximate Concurrency via Controlled Dependency Relaxation", *In proceedings of 42nd annual ACM SIGPLAN conference on Programming Language Design and Implementation. Acceptance Ratio: 87/320 = 27%* **(PLDI 2021)**

[C22]. Shixiong Jing, Qinkun Bao, Pei Wang, **Xulong Tang**, Dinghao Wu, "Characterizing AI Model Inference

Applications Running in SGX Environment", *In Proceedings of the 15th International Conference on Networking, Architecture, and Storage.* **(NAS 2021)**

[C23]. Xinyi Zhang, Yawen Wu, Peipei Zhou, **Xulong Tang**, Jingtong Hu, "Algorithm-Hardware Co-design of Attention Mechanism on FPGA Devices", *In Proceedings of the International Conference on Hardware/Software Codesign and System Synthesis.* **(CODES+ISSS 2021)**

[C24]. Geng Yuan, Peiyan Dong, Mengshu Sun, Wei Niu, Zhengang Li, Yuxuan Cai, Jun Liu, Weiwen Jiang, Xue Lin, Bin Ren, **Xulong Tang**, Yanzhi Wang, "Work in Progress: Mobile or FPGA? A Comprehensive Evaluation on Energy Efficiency and a Unified Optimization Framework", *In proceedings of IEEE 27th Real-Time and Embedded Technology and Applications Symposium.* **(RTAS 2021)**

[C25]. Weizheng Xu, Youtao Zhang, **Xulong Tang**, "Parallelizing DNN Training on GPUs: Challenges and Opportunities", *In Proceedings of the WWW '21: Companion Proceedings of the Web Conference 2021.* **(WWW 2021 workshop)**

[C26]. Yuxuan Cai, Geng Yuan, Hongjia Li, Wei Niu, Yanyu Li, **Xulong Tang**, Bin Ren, and Yanzhi Wang, "A Compression-Compilation Co-Design Framework Towards Real-Time Object Detection on Mobile Devices", *The Thirty-Fifth AAAI Conference on Artificial Intelligence.* **(AAAI 2021)**

[C27]. Yuxuan Cai, Hongjia Li, Geng Yuan, Wei Niu, Yanyu Li, **Xulong Tang**, Bin Ren, and Yanzhi Wang, "YOLObile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design", *The Thirty-Fifth AAAI Conference on Artificial Intelligence. Acceptance Ratio: 20.9%* **(AAAI 2021)**

[C28]. Mahmut Taylan Kandemir, **Xulong Tang**, Jihyun Ryoo, Mustafa Karakoy, "Compiler Support for Near Data Computing", *Proceedings of the ACM SIGPLAN Annual Symposium Principles and Practice of Parallel Programming. Acceptance Ratio: 48/150 = 32%* **(PPOPP 2021)**

[C29]. Yuxuan Cai, Hongjia Li, Geng Yuan, Wei Niu, Yanyu Li, **Xulong Tang**, Bin Ren, and Yanzhi Wang, "YOLObile: Real-Time Object Detection on Mobile Devices via Compression-Compilation Co-Design", *In NeurIPS 2020 Workshop on Machine Learning for Autonomous Driving.* **(NeurIPS 2020 workshop)**

[C30]. **Xulong Tang**, Ziyu Zhang, Weizheng Xu, Mahmut Taylan Kandemir, Rami Melhem, Jun Yang, "Enhancing Address Translations in Throughput Processors via Compression", *In proceedings of the 29th International Conference on Parallel Architectures and Compilation Techniques. Acceptance Ratio: 35/135 = 25.9%* **(PACT 2020)**

[C31]. Zhendong Wang, Zihang Jiang, Zhen Wang, **Xulong Tang**, Cong Liu, Yang Hu, "Enabling Latency-aware Data Initialization for Integrated CPU/GPU Heterogeneous Platform", *published in the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems.* **(TCAD 2020)**

[C32]. **Xulong Tang**, Mahmut Taylan Kandemir, Mustafa Karakoy, Meena Arunachalam, "Co-Optimizing Memory-Level Parallelism and Cache-Level Parallelism", *In proceedings of 40th annual ACM SIGPLAN conference on Programming Language Design and Implementation. Acceptance Ratio: 76/274 = 27.7%* **(PLDI 2019)**

[C33]. **Xulong Tang**, Ashutosh Pattnaik, Onur Kayiran, Adwait Jog, Mahmut Taylan Kandemir, Chita Das, "Quantifying Data Locality in Dynamic Parallelism in GPUs", *In proceedings of ACM Measurement and Analysis of Computing Systems. Acceptance Ratio: 6/67 = 8.9%* **(SIGMETRICS 2019)**

[C34]. **Xulong Tang**, Mahmut Taylan Kandemir, Hui Zhao , Myoungsoo Jung, Mustafa Karakoy, "Computing with Near Data", *In proceedings of ACM Measurement and Analysis of Computing Systems. Acceptance Ratio: 6/67 = 8.9%* **(SIGMETRICS 2019)**

[C35]. Ashutosh Pattnaik, **Xulong Tang**,Onur Kayiran, Adwait Jog, Asit Mishra, Mahmut T. Kandemir, Anand Sivasubramaniam, Chita R. Das, "Opportunistic Computing in GPU Architectures", *In proceedings of 46th International Symposium on Computer Architecture. Acceptance Ratio: 62/365 = 16.9%* **(ISCA 2019)**

[C36]. Mustafa Karakoy, Orhan Kislal, **Xulong Tang**, Mahmut Taylan Kandemir, Meena Arunachalam, "Architecture-Aware Approximate Computing", *In proceedings of ACM Measurement and Analysis of Computing Systems. Acceptance Ratio: 6/67 = 8.9%* **(SIGMETRICS 2019)**

[C37]. Jihyun Ryoo, Mengran Fan, **Xulong Tang**, Huaipan Jiang, Meena Arunachalam, Sharada Naveen, Mahmut Taylan Kandemir, "Architecture-Centric Bottleneck Analysis for Deep Neural Network Applications", *In proceedings of the 26TH IEEE International Conference on High Performance Computing, Data, and Analytics.* **(HiPC 2019)**

[C38]. Jihyun Ryoo, Orhan Kislal, **Xulong Tang**, Mahmut T. Kandemir, "Quantifying and Optimizing Data Access Parallelism on Manycores", *In proceedings of 26th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems.* **(MASCOTS 2018)**

[C39]. Orhan Kislal, Jagadish B. Kotra, **Xulong Tang**, Mahmut T. Kandemir, Myoungsoo Jung, "Enhancing Computation-to-Core Assignment with Physical Location Information", *In proceedings of 39th annual ACM SIGPLAN conference on Programming Language Design and Implementation. Acceptance Ratio: 55/254 = 22.4%* **(PLDI 2018)**

[C40]. Sooraj Puthoor, **Xulong Tang**, Joseph Gross, Bradford M Beckmann, "Oversubscribed Command Queues in GPUs.", *In proceedings of the 11th Workshop on General Purpose GPUs in conjunction with PPoPP 2018.* **(PPoPP 2018)**

[C41]. **Xulong Tang**, Orhan Kislal, Mahmut Kandemir, Mustafa Karakoy, "Data Movement Aware Computation Partitioning", *In proceedings of The 50th Annual IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 61/327 = 18.6%* **(MICRO 2017)**

[C42]. Akbar Sharifi, Wei Ding, Diana Guttman, Hui Zhao, **Xulong Tang**, Mahmut Kandemir, Chita Das, "DEMM: a Dynamic Energy-saving mechanism for Multicore", *In proceedings of The 25th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. Acceptance Ratio: 26/84 = 30.9%* **(MASCOTS 2017)**

[C43]. Orhan Kislal, Jagadish Kotra, **Xulong Tang**, Mahmut Taylan Kandemir, Myoungsoo Jung, "POSTER: Location-Aware Computation Mapping for Manycore Processors", *In proceedings of The 26th International Conference on Parallel Architectures and Compilation Techniques.* **(PACT 2017)**

[C44]. **Xulong Tang** , Ashutosh Pattnaik, Huaipan Jiang, Onur Kayiran, Adwait Jog, Sreepathi Pai, Mohamed Ibrahim, Mahmut Kandemir, Chita Das, "Controlled Kernel Launch for Dynamic Parallelism in GPUs", *In Proceedings of 23th International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 50/224 = 22.3%* **(HPCA 2017)**

[C45]. **Xulong Tang** , Mahmut Kandemir, Praveen Yedlapalli, Jagadish Kotra, "Improving Bank-Level Parallelism for Irregular Applications", *In Proceedings of 49th Annual IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 61/283 = 21.6%* **(MICRO 2016) Best Paper Nomination.**

[C46]. Ashutosh Pattnaik, **Xulong Tang**, Adwait Jog, Onur Kayiran, Asit K. Mishra, Mahmut T. Kandemir, Onur Mutlu, Chita R. Das, "Scheduling Techniques for GPU Architectures with Processing-In-Memory Capabilities", *In Proceedings of 25th International Conference on Parallel Architectures and Compilation Techniques. Acceptance Ratio: 31/139 = 22.3%* **(PACT 2016)**

[C47]. Onur Kayiran, Adwait Jog, Ashutosh Pattnaik, Rachata Ausavarungnirun, **Xulong Tang**, Mahmut T. Kandemir, Gabriel H. Loh, Onur Mutlu, Chita R. Das, "μC-States: Fine-grained GPU Datapath Power Management", *In Proceedings of 25th International Conference on Parallel Architectures and Compilation Techniques. Acceptance Ratio: 31/139 = 22.3%* **(PACT 2016)**

[C48]. Wei Ding, **Xulong Tang**, Mahmut Taylan Kandemir, Yuanrui Zhang, Emre Kultursay "Optimizing Off-Chip Accesses in Manycores", *In Proceedings of 36th annual ACM SIGPLAN conference on Programming Language Design and Implementation. Acceptance Ratio: 58/303 = 19.1%* **(PLDI 2015)**

[C49]. Mahmut Taylan Kandemir, Hui Zhao, **Xulong Tang**, Mustafa Karaköy, "Memory Row Reuse Distance and its Role in Optimizing Application Performance", *In Proceedings of ACM International Conference on Measurement and Modeling of Computer Systems. Acceptance Ratio: 32/239 = 13.3%* **(SIGMETRICS 2015)**

[C50]. **Xulong Tang**, Hong An, Gongjin Sun, Dongrui Fan, "A Video Coding Benchmark Suite for Evaluation of Processor Capability", *In Proceedings of 14th IEEE/ACIS International Conference on Software Engineering,*

Artificial Intelligence, Networking and Parallel/Distributed Computing. **(SNPD 2013)**

[C51]. Gu Liu, Hong An, Xiaoqiang Li, Wei Zhou, Xuechao Wei, **Xulong Tang**, "FlexBFS: A Parallelism-aware Implementation of Breadth-First Search on GPU", *Accepted as a poster by 17th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming.* **(PPoPP 2012)**

PATENTS

2022	"Systems and methods for optimizing quantum circuit simulation using graphics processing units", UPB-025PR.
2019	"Multi-kernel wavefront scheduler", US20190370059A1.

GRANT

2022	"SHF: Small: Expediting the Execution of Machine Learning Applications on Multi-GPU Infrastructure with Architecture Awareness and Runtime Support". Funded by National Science Foundation. PI. $599,922
2022	"Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation Using GPUs". Funded by Pitt Momentum. PI. $25,000
2020	"FoMR: A Software and Hardware Codesign for Addressing the Performance Bottlenecks in Secure NVM". Funded by National Science Foundation. co-PI. $330,000
2020	"Embracing Heterogeneity in Modern GPUs". Funded by Pitt Momentum. PI. $16,000
2019	Pitt startup funding package for tenure-stream assistant professor.

TEACHING

2023 Spring	Instructor, CS2410 - Computer Architecture - at Pitt
2022 Fall	Instructor, CS3410 - Advanced Topics in Computer Architecture - at Pitt
2022 Spring	Instructor, CS2410 - Computer Architecture - at Pitt
2021 Spring	Instructor, CS1541 - Introduction to Computer Architecture - at Pitt
2021 Spring	Instructor, CS2410 - Computer Architecture - at Pitt
2020 Spring	Instructor, CS2210 - Compiler Design - at Pitt
2018 Fall	Co-instructor of CMPEN 431 - Introduction to Computer Architecture - at Penn State
2016 Spring	Guest Lecture, CSE 521 - Design and Implementation of Compilers - at Penn State
2015 Spring	Teaching Assistant of CMPEN 431 - Introduction to Computer Architecture - at Penn State
2014 Fall	Teaching Assistant of CMPEN 431 - Introduction to Computer Architecture - at Penn State
2014 Spring	Teaching Assistant of CS 210 - Introduction to Python - at College of William and Mary
2011 Summer	Teaching Assistant of Introduction to Computer System - at USTC

TALKS

- Embracing Heterogeneity in Modern GPUs. *Pitt Momentum 2021*
- Mix and Match: Reorganizing Tasks for Enhancing Data Locality. *SIGMETRICS 2021*
- Optimizing Quantum Circuit Simulation/Emulation on HPC Platforms. *Pitt Quantum Institute 2020*
- Enhancing address translation in GPUs through compression. *Intel*
- Enhancing Address Translations in Throughput Processors via Compression. *PACT 2020*
- Co-Optimizing Memory-Level Parallelism and Cache-Level Parallelism. *PLDI 2019*
- Computing with Near Data. *SIGMETRICS 2019*

- Irregularity-aware Computation and Data Management in Manycore Systems. *Job talk at multiple universities, Spring 2019*
- Quantifying and Optimizing Data Access Parallelism on Manycores. *MASCOTS 2018*
- Scheduling in the Cloud. *MASCOTS 2018*
- Enhancing Computation-to-Core Assignment with Physical Location Information. *PLDI 2018*
- Data Movement Aware Computation Partitioning. *MICRO 2017*
- DEMM: a Dynamic Energy-saving mechanism for Multicore. *MASCOTS 2017*
- Controlled Kernel Launch for Dynamic Parallelism in GPUs. *HPCA 2017*
- Improving Bank-Level Parallelism for Irregular Applications. *MICRO 2016*
- Memory Row Reuse Distance and its Role in Optimizing Application Performance. *SIGMETRICS 2015*

AWARDS AND HONORS

2019	NSF Travel Grants / SIGMETRICS'2019
	ACM Travel Grants / PLDI'40
2018	NSF Travel Grants / PLDI'39
2017	NSF Travel Grants / MICRO'50
	NSF Travel Grants / HPCA'23
2016	**Best Paper Nomination of MICRO'49**
	NSF Travel Grants / MICRO'49
2015	NSF Travel Grants / PLDI'36

PROFESSIONAL SERVICES

Program Committee	Artifact Evaluation Committee of PPoPP'19, PPoPP'18 Committee member of HPCA (2020, 2021, 2023) ISCA (2020, 2022, 2023) MICRO (2020, 2021, 2022) ASPLOS (2020) PLDI (2021, 2023) NAS (2019, 2021, 2022)
Journal Reviewer	Transactions on Parallel and Distributed Systems (TPDS) International Journal of Computational Science and Engineering (IJCSE) Transactions on Architecture and Code Optimization (TACO) Electronics and Telecommunications Research Institute Journal (ETRIJ) Advances in Science Technology and Engineering Systems Journal (ASTESJ) IEEE Transactions on Computers IEEE Computer Architecture Letters IEEE Access Journal Transactions on Computers Future Generation Computer Systems (FGCS)
Other Activities	Submission chair of AIM 2017 workshop