SENSQ 6115, 210 S. Bouquet Street
Pittsburgh, PA, 15232
Contact: (757) 532-5183
tax6@pitt.edu
http://xzt102.github.io/
Google Scholar

Xulong Tang

APPOINTMENTS

2025 – Now **Associate Professor**
University of Pittsburgh, Department of Computer Science

2025 – Now **Visiting Scholar**
AMD Research and Advanced Development (RAD)

2019 – 2025 **Assistant Professor**
University of Pittsburgh, Department of Computer Science

EDUCATION

2014 – 2019 **Ph.D. in Computer Science and Engineering**, Pennsylvania State University
Advisors: Dr. Mahmut Taylan Kandemir and Dr. Chita R. Das

2010 – 2013 **M.S. in Computer Science and Technology**, University of Science and Technology of China
Advisor: Dr. Hong An

2006 – 2010 **B.S. in Computer Science and Technology**, Harbin Institute of Technology
Advisor: Dr. Chunqi Sun

AWARDS AND HONORS

2023 ICLR 2023 Spotlight Award

2020 Intel Research Award

2016 MICRO 2016 Best Paper Nomination

2016 PSU Outstanding Research Assistant

2015 PSU Outstanding Teaching Assistant

RESEARCH

PUBLICATIONS

CONFERENCES

CVPR'26 Sheng Li, Connelly Barnes, Mamshad Nayeem Rizve, Hongwu Peng, Zhengang Li, Ohi Dibua, Alireza Ganjdanesh, **Xulong Tang**, Yan Kang, Yifan Gong, "Content-Aware Dynamic Patchification for Efficient Video Diffusion.", *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

arXiv Yueqi Wang, Bingyao Li, Mohamed Tarek Ibn Ziad, Lieven Eeckhout, Jun Yang, Aamer Jaleel, **Xulong Tang**, "Chunk-Level KV Cache Reuse for Efficient RAG Serving."

arXiv Sheng Li, Yang Sui, Yue Wu, Zhuoran Song, Bo Yuan, **Xulong Tang**, Yue Dai, "Accelerating 3D Gaussian Splatting with Tensor Cores."

arXiv Aditya Pawar, Yingheng Li, **Xulong Tang**, Youtao Zhang, Jun Yang, "Swap-Free Quantum LDPC Code Mapping on Near-Term Local Architecture."

arXiv Yingheng Li, **Xulong Tang**, Paul Hovland, Ji Liu, "Non-Clifford Fusion: T-Gate Optimization for Quantum Simulation."

arXiv Tianyu Wang, Sheng Li, Bingyao Li, Yue Dai, Ao Li, Geng Yuan, Yufei Ding, Youtao Zhang, **Xulong Tang**, "Improving GPU Multi-Tenancy Through Dynamic Multi-Instance GPU Reconfiguration."

arXiv Sheng Li, Geng Yuan, Yue Dai, Tianyu Wang, Yawen Wu, Alex K. Jones, Jingtong Hu, Geng Yuan, Yanzhi Wang, Bo Yuan, Yufei Ding, **Xulong Tang**, "EdgeOL: Efficient in-situ Online Learning on Edge Devices."

ICCAD'25 Rongchao Dong, Zewei Mo, Yingheng Li, Yue Dai, Aditya Pawar, Jun Yang, Youtao Zhang, **Xulong Tang**, "STMC: Small-Tile Multiple-Copy Compilation for Reliable Measurement-Based Quantum Computing.", *In Proceedings of the 44th ACM/IEEE International Conference on Computer-Aided Design*.

ICML'25 Yue Dai, Liang Liu, **Xulong Tang**, Youtao Zhang, Jun Yang, "MemFreezing: A Novel Adversarial Attack on Temporal Graph Neural Networks under Limited Future Knowledge.", *In Proceedings of the 42nd International Conference on Machine Learning*.

ICS'25 Xiaoyu Hao, Sen Zhang, Liang Qiao, Qingcai Jiang, Jun Shi, Junshi Chen, Hong An, **Xulong Tang**, Hao Shu, Honghui Yuan "CIExplorer: Microarchitecture-Aware Exploration for Tightly Integrated Custom Instruction.", *In Proceedings of the ACM International Conference on Supercomputing 2025*.

ISCA'25 Yingheng Li, Yue Dai, Aditya Pawar, Rongchao Dong, Jun Yang, Youtao Zhang, **Xulong Tang**, "Reinforcement Learning-Guided Graph State Generation in Photonic Quantum Computers.", *In Proceedings of the 52nd ACM International Symposium on Computer Architecture*.

MLSys'25 Zaifeng Pan, Yitong Ding , Yue Guan , Zheng Wang , Zhongkai Yu, **Xulong Tang**, Yida Wang, Yufei Ding, "FastTree: Optimizing Attention Kernel and Runtime for Tree-Structured LLM Inference.", *In Proceedings of the Eighth Annual Conference on Machine Learning and Systems*.

ASPLOS'25 Yue Dai, **Xulong Tang**, Youtao Zhang, "Cascade: A Dependency-aware Efficient Training Framework for Temporal Graph Neural Network.", *In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.

ASPLOS'25 Liang Qiao, Jun Shi, Xiaoyu Hao, Xi Fang, Sen Zhang, Minfan Zhao, Ziqi Zhu, Junshi Chen, Hong An, **Xulong Tang**, Bing Li, Honghui Yuan, Xinyang Wang, "Pruner: A Draft-then-Verify Exploration Mechanism to Accelerate Tensor Program Tuning.", *In Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*.

ICLR'25 Sheng Li, Qitao Tan, Yue Dai, Zhenglun Kong, Tianyu Wang, Jun Liu, Ao Li, Ninghao Liu, Yufei Ding, **Xulong Tang**, Geng Yuan, "Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning.", *In Proceedings of the 13th International Conference on Learning Representations*.

HPCA'25 Yueqi Wang, Bingyao Li, Mohamed Tarek Ibn Ziad, Lieven Eeckhout, Jun Yang, Aamer Jaleel, **Xulong Tang**, "OASIS: Object-Aware Page Management for Multi-GPU Systems.", *In Proceedings of the 31st IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 21.2%. 13 pages*.

MICRO'24 Bingyao Li, Yueqi Wang, Lieven Eeckhout, Jun Yang, Aamer Jaleel, **Xulong Tang**, "STAR: Sub-Entry Sharing-Aware TLB for Multi-Instance GPU.", *In Proceedings of the 57th IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 22.7%. 13 pages.*

ASPLOS'24 Yingheng Li, Aditya Pawar, Zewei Mo, Youtao Zhang, Jun Yang, **Xulong Tang**, "FMCC: Flexible Measurement-based Quantum Computation over Cluster State.", *In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. Acceptance Ratio: 63/340 = 18.5%. 13 pages.*

ASPLOS'24 Aditya Pawar, Yingheng Li, Zewei Mo, Yanan Guo, **Xulong Tang**, Youtao Zhang, Jun Yang, "QRCC: Evaluating Large Quantum Circuits on Small Quantum Computers through Integrated Qubit Reuse and Circuit Cutting.", *In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. Acceptance Ratio: 63/340 = 18.5%. 13 pages.*

DAC'24 Zewei Mo, Yingheng Li, **Xulong Tang**, Jun Yang and Youtao Zhang, "FCM: Wire Cutting For Fusion Reduction in Measurement-based Quantum Computing.", *In Proceedings of the 61th Design Automation Conference. Acceptance Ratio: 75/410 = 18.3%. 6 pages.* Link.

DAC'24 Yifan Gong, Yushu Wu, PU ZHAO, zheng zhan, Liangkai Liu, Chao Wu, **Xulong Tang**, and Yanzhi Wang, "LOTUS: learning-based online thermal and latency variation management for two-stage detectors on edge devices.", *In Proceedings of the 61th Design Automation Conference. Acceptance Ratio: 75/410 = 18.3%. 6 pages.*

ICLR'24 Sheng Li, Chao Wu, Ao Li, Yanzhi Wang, **Xulong Tang**, Geng Yuan, "Waxing-and-Waning: a Generic Similarity-based Framework for Efficient Self-Supervised Learning.", *In Proceedings of the 12th International Conference on Learning Representations. Acceptance Ratio: 2260/7404 = 30.52%. 9 pages.* Link.

HPCA'24 Yueqi Wang*, Bingyao Li*, Aamer Jaleel, Jun Yang, **Xulong Tang**, "GRIT: Enhancing Multi-GPU Performance with Fine-Grained Dynamic Page Placement.", *In Proceedings of the 30th IEEE International Symposium on High-Performance Computer Architecture. Acceptance Ratio: 20.2%. 13 pages.* DOI.

ICCD'23 Yue Dai, **Xulong Tang**, Youtao Zhang, "FlexGM: An Adaptive Runtime System to Accelerate Graph Matching Networks on GPUs.", *In Proceedings of the 41st IEEE International Conference on Computer Design. Acceptance Ratio: 30.2%. 8 pages.* DOI.

MICRO'23 Bingyao Li, Yanan Guo, Yueqi Wang, Aamer Jaleel, Jun Yang, **Xulong Tang**, "IDYLL: Enhancing Page Translation in Multi-GPUs via Light Weight PTE Invalidations.", *In Proceedings of the 56th IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 101/424 = 23.8%. 12 pages.* DOI.

MICRO'23 Zhengang Li, Geng Yuan, Tomoharu Yamauchi, Zabihi Masoud, Yanyue Xie, Peiyan Dong, **Xulong Tang**, Nobuyuki Yoshikawa, Devesh Tiwari, Yanzhi Wang, Olivia Chen, "SupeRBNN: Randomized Binary Neural Network Using Adiabatic Superconductor Josephson Devices.", *In Proceedings of the 56th IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 101/424 = 23.8%. 12 pages.* DOI.

DAC'23 Mehrnoosh Raoufi, Jun Yang, **Xulong Tang**, Youtao Zhang, "EP-ORAM: Efficient NVM-Friendly Path Eviction for Ring ORAM in Hybrid Memory.", *In Proceedings of the 60th Design Automation Conference. Acceptance Ratio: 263/1156 = 22.7%. 6 pages.* DOI.

DAC'23 Yingheng Li, Aditya Pawar, Mohadeseh Azari, Yanan Guo, Youtao Zhang, Jun Yang, Kaushik Parasuram Seshadreesan, **Xulong Tang**, "Orchestrating Measurement-Based Quantum Computation over Photonic Quantum Processors.", *In Proceedings of the 60th Design Automation Conference. Acceptance Ratio: 263/1156 = 22.7%. 6 pages.* DOI.

DAC'23 Bingyao Li, Yueqi Wang, **Xulong Tang**, "Orchestrated Scheduling and Partitioning for Improved Address Translation in GPUs.", *In Proceedings of the 60th Design Automation Conference. Acceptance Ratio: 263/1156 = 22.7%. 6 pages.* DOI.

HPCA'23 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%. 12 pages.* DOI.

HPCA'23 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%. 12 pages.* DOI.

HPCA'23 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%. 12 pages.* DOI.

ICLR'23 Spotlight 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. Acceptance Ratio: 1204/4955 = 24.30%. Spotlight Ratio: 280/4955 = 5.65%. 9 pages.* Link.

HPCA'22 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%. 13 pages.* DOI.

WWW'22 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'22. 7 pages.* DOI.

ECCV'22 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'22: 17th European Conference. Acceptance Ratio: 1645/5804 = 28%. 14 pages.* DOI.

NeurIPS'22 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%. 10 pages.* Link.

ICCAD'22 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%. 8 pages.* DOI.

ICCD'22 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%. 7 pages.* DOI.

EuroSys'22 F. Yu, Z. Xu, T. Shen, D. Stamoulis, L. Shangguan, D. Wang, M. Zhang, **Xulong 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). Acceptance Ratio: 241/1308 = 18%. 12 pages.* Link.

MICRO'21 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%. 12 pages.* DOI.

ICCAD'21 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%. 8 pages.* DOI.

ICCAD'21 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%. 8 pages.* DOI.

SIGMETRICS '21 **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%. 21 pages.* DOI.

PLDI'21 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%. 14 pages.* DOI.

PLDI'21 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%. 14 pages.* DOI.

NAS'21 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. 4 pages.* DOI.

ATS'21 Zhendong Wang, Rujia Wang, Zihang Jiang, **Xulong Tang**, Shouyi Yin, Yang Hu, "Towards a Secure Integrated Heterogeneous Platform via Cooperative CPU/GPU Encryption.", *In Proceedings of the Asian Test Symposium'21. 6 pages.* DOI.

ODES+ISSS '21 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. 23 pages.* DOI.

RTAS'21 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. 4 pages.* DOI.

WWW'21 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'21. 4 pages.* DOI.

AAAI'21 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. Acceptance Ratio:'.9%. 4 pages.* DOI.

AAAI'21 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:'.9%. 7 pages.* DOI.

PPOPP'21 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%. 13 pages.* DOI.

NeurIPS'20 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'20 Workshop on Machine Learning for Autonomous Driving.*

PACT'20 **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%. 12 pages.* DOI.

PLDI'19 **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%. 13 pages.* DOI.

SIGMETRICS '19 **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%. 21 pages.* DOI.

SIGMETRICS '19 **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%. 27 pages.* DOI.

SIGMETRICS '19 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%. 22 pages.* DOI.

ISCA'19 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%. 13 pages.* DOI.

HiPC'19 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. Acceptance Ratio: 39/171 = 23%. 10 pages.* DOI.

MASCOTS'18 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. 12 pages.* DOI.

PLDI'18 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%. 14 pages.* DOI.

PPoPP'18 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'18. Acceptance Ratio: 20%. 10 pages.* DOI.

MICRO'17 **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%. 13 pages.* DOI.

MASCOTS'17 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%. 10 pages.* DOI.

PACT'17 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. 2 pages.* DOI.

HPCA'17 **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%. 12 pages.* DOI.

MICRO'16 Best Paper Nomination **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%. 11 pages.* DOI.

PACT'16 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%. 12 pages.* DOI.

PACT'16 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%. 12 pages.* DOI.

PLDI'15 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%. 12 pages.* DOI.

SIGMETRICS '15 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%. 12 pages.* DOI.

SNPD'13 **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. 15 pages. DOI.*

PPoPP'12 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. DOI.*

JOURNALS

TECS'22 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. Acceptance Ratio: 23%. 21 pages. DOI.*

IEEE Micro 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. 8 pages. DOI.*

CCF Yue Dai, **Xulong Tang**, Youtao Zhang, "An Efficient Segmented Quantization for Graph Neural Networks.", *In Proceedings of CCF Transactions on High Performance Computing. 13 pages. DOI.*

TCAD'20 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. 12 pages. DOI.*

PATENT

2022 "Systems and methods for optimizing quantum circuit simulation using graphics processing units", UPB-025PR.

2019 "Multi-kernel wavefront scheduler", US20190370059A1.

GRANT

2025 "ExpandQISE: Track 2: URI-PQI Collaboration - Application Of Quantum Fundamentals To Advance Research and Workforce Development" (NSF)

2024 "CCF: Small: Hardware-Software Co-design for Privacy Protection on Deep Learning-based Recommendation Systems" (NSF)

2024 "Evaluating Large Quantum Circuit on Small Quantum Devices — An Integrated Circuit Cutting and Qubit Reuse Approach" (PQI)

2023 "CSR: Small: Expediting Continual Online Learning on Edge Platforms through Software-Hardware Co-designs" (NSF)

2022 "SHF: Small: Expediting the Execution of Machine Learning Applications on Multi-GPU Infrastructure with Architecture Awareness and Runtime Support" (NSF)

2022 "Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation Using GPUs" (Pitt Momentum)

2021 "Embracing Heterogeneity in Modern GPUs" (Pitt Momentum)

2020 "FoMR: A Software and Hardware Codesign for Addressing the Performance Bottlenecks in Secure NVM" (NSF)

TALKS

2026 "Unleashing Next-Generation GPU Computing for the AI Era.", *University of Delaware*

2025 "Unleashing Multi-GPU Computing to the Next-Level.", *University of California, Los Angeles*

2025 "Unleashing Multi-GPU Computing to the Next-Level.", *University of Southern California*

2025 "Unleashing Multi-GPU Computing to the Next-Level.", *University of California, Irvine*

2025 "Expediting Continual Online Learning on Edge Platforms through Software-Hardware Co-designs.", *NSF CSR PI Meeting 2025*

2025 "Toward High-fidelity, Scalable, and Accessible Quantum Computing System.", *ModSim workshop by University of Washington and Brookhaven National Laboratory, 2025*

2025 "Toward Fault-tolerant and Scalable Quantum Computing.", *QCE 2025*

2025 "Unleashing Multi-GPU Computing to the Next-Level.", *ISCA 2025 Forum*

2025 "Reinforcement Learning-Guided Graph State Generation in Photonic Quantum Computers.", *ISCA 2025*

2024 "Toward High-Fidelity, Scalable, and Accessible Quantum Computing Systems.", *Pitt SCI Dean's Spotlight*

2024 "Compilation in Measurement-Based Quantum Computing.", *Pitt Quantum Institute 2024*

2023 "Towards Efficient and Scalable Computing for Multi-GPUs.", *Shanghai Jiao Tong University*

2023 "Towards Efficient and Scalable Computing for Multi-GPUs.", *University of Science and Technology of China*

2023 "Towards Efficient and Scalable Computing for Multi-GPUs.", *Sun Yat-sen University*

2023 "Towards Efficient and Scalable Computing for Multi-GPUs.", *The Hong Kong University of Science and Technology*

2023 "CEGMA: Coordinated Elastic Graph Matching Acceleration for Graph Matching Networks.", *HPCA 2023*

2021 "Optimizing Quantum Circuit Simulation/Emulation on HPC Platforms.", *Pitt Quantum Institute 2021*

2021 "Embracing Heterogeneity in Modern GPUs.", *Pitt Momentum 2021*

2021 "Mix and Match: Reorganizing Tasks for Enhancing Data Locality.", *SIGMETRICS 2021*

2021 "Enhancing address translation in GPUs through compression.", *Intel*

2020 "Enhancing Address Translations in Throughput Processors via Compression.", *PACT 2020*

2019 "Co-Optimizing Memory-Level Parallelism and Cache-Level Parallelism.", *PLDI 2019*

2019 "Computing with Near Data.", *SIGMETRICS 2019*

2019 "Irregularity-aware Computation and Data Management in Manycore Systems.", *University of California, Davis*

2019 "Irregularity-aware Computation and Data Management in Manycore Systems.", *University of Pittsburgh*

2018 "Quantifying and Optimizing Data Access Parallelism on Manycores.", *MASCOTS 2018*

2018 "Scheduling in the cloud.", *MASCOTS 2018*

2018 "Enhancing Computation-to-Core Assignment with Physical Location Information.", *PLDI 2018*

2017 "Data Movement Aware Computation Partitioning.", *MICRO 2017*

2017 "DEMM: a Dynamic Energy-saving mechanism for Multicore.", *MASCOTS 2017*

2017 "Controlled Kernel Launch for Dynamic Parallelism in GPUs.", *HPCA 2017*

2016 "Improving Bank-Level Parallelism for Irregular Applications.", *MICRO 2016*

2015 "Memory Row Reuse Distance and its Role in Optimizing Application Performance.", *SIGMETRICS 2015*

TEACHING AND MENTORING

UNDERGRADUATE COURSES TAUGHT

CS1541 Introduction to Computer Architecture
Semesters taught: Spring 2021, Fall 2023, Spring 2025

CS1645 High Performance Computing Systems
Semesters taught: Fall 2024

GRADUATE COURSES TAUGHT

CS2410 Computer Architecture
Semesters taught: Spring 2021, Spring 2022, Spring 2023, Spring 2024, Spring 2025

CS2210 Compiler Design
Semesters taught: Spring 2020

CS2045 High Performance Computing Systems
Semesters taught: Fall 2024

CS3410 Advanced Topics in Computer Architecture
Semesters taught: Fall 2022, Fall 2025

PHD STUDENTS

Sheng Li Computer Science, 08/2021 - present

Yueqi Wang Computer Science, 08/2022 - present

Yingheng Li Computer Science, 08/2022 - present

Haoqing Yang Computer Science, 08/2024 - present

Buxin Tu Computer Science, 08/2026 - present (incoming)

Zhou Ye Computer Science, 08/2026 - present (incoming)

ALUMNI

Bingyao Li Ph.D. 2025 → Assistant Professor, CSE, University of California, Riverside

Yue Dai Ph.D. 2025 → Assistant Professor, CS, Illinois Institute of Technology

Mehrnoosh Raoufi Ph.D. 2024 → Oracle

Tianao Ge Visiting Scholar 2024 → HKUST GZ

Yuan Yao M.S. 2023 → Bloomberg LP

Ziyu Zhang M.S. 2022 → Apple

Yilun Zhao Visiting Scholar 2022 → Ph.D. @ ICT-CAS

Qi Xue B.S. 2022 → M.S. @ UPenn

PHD STUDENT COMMITTEE MEMBERSHIP

- Bingyao Li (Chair: Xulong Tang)
- Mehrnoosh Raoufi (Chair: Youtao Zhang)
- Chi Zhang (Chair: Youtao Zhang)
- Marika Schubert (Chair: Alan George)
- Luke Kljucaric (Chair: Alan George)
- Boyuan Yang (Chair: Wei Gao)
- Lei Zhao (Chair: Youtao Zhang)
- Yue Dai (Chair: Youtao Zhang)
- Yue Tang (Chair: Jingtong Hu)
- Yawen Wu (Chair: Jingtong Hu)
- Zhenjiang Fan (Chair: Stephen Lee)
- Longhao Li (Chair: Taieb Znati)
- Jinming Zhuang (Chair: Peipei Zhou)

MASTERS STUDENTS SUPERVISED

- Zhongxuan Song
- Weizheng Xu
- Zachary Michael Smith
- Thomas Matthew Dicarlo
- Antony Paul
- Yuan Yao
- Ziyu Zhang
- Yilun Zhao
- Ziwei Quan

UNDERGRADUATE STUDENTS SUPERVISED

- Christopher Hinson
- Derrick Hicks
- Qi Xue

SERVICE

PROFESSIONAL SERVICES

Steering Committee
ASPLOS Steering Committee Member

Chairship
General Co-Chair, ASPLOS 2026
Program Vice Chair, IEEE Micro Top Picks 2025
Publication Chair, ASPLOS 2025
Submission Chair, AIM 2017

Associate Editor
ACM Transactions on Architecture and Code Optimization (TACO)
IEEE Transactions on Computers (TC)

Program Committee
HPCA (2020, 2021, 2023, 2024, 2025, 2026)
ISCA (2020, 2022, 2023, 2024, 2025, 2026)
MICRO (2020, 2021, 2022, 2023, 2024, 2025)
ASPLOS (2020, 2024)
PLDI (2021, 2023)
SIGMETRICS (2024)
NeurIPS (2024)
ASP-DAC, PACT
NAS (2019, 2021, 2022)

Journal Reviewer
Transactions on Parallel and Distributed Systems (TPDS)
International Journal of Computational Science and Engineering (IJCSE)
Electronics and Telecommunications Research Institute Journal (ETRIJ)

Advances in Science Technology and Engineering Systems Journal (ASTESJ)
IEEE Computer Architecture Letters
IEEE Access Journal
Future Generation Computer Systems (FGCS)