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

2019 - present **University of Pittsburgh**
Assistant professor in Department of Computer Science
School of Computing and Information

2014 - 2019 **Pennsylvania State University**
Ph.D. in Computer Science and Engineering
Advisor: Dr. Mahmut Taylan Kandemir

2014 Spring **College of William and Mary**
Ph.D. in Computer Science
Transfer to Pennsylvania State University in 2014 fall
Advisor: Dr. Xipeng Shen

2010 - 2013 **University of Science and Technology of China**
M.E. in Computer Science and Technology
Advisor: Dr. Hong An

2006 - 2010 **Harbin Institute of Technology**
B.E. in Computer Science and Technology
Advisor: Dr. Chunqi Sun

RESEARCH EXPERIENCE

2019 - present **University of Pittsburgh**
Assistant Professor
- DNN acceleration on modern computing platforms
- GPUs: applications, compiling, runtime and architecture
- Compiler-assisted massive parallel computing
- Heterogeneous computing architectures/systems

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

[C1]. Bingyao Li, Jieming Yin, Youtao Zhang, **Xulong Tang** "To Appear", *In Proceedings of the 54th IEEE/ACM International Symposium on Microarchitecture. Acceptance Ratio: 94/430 = 21.8%* **(MICRO 2021)**

[C2]. 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)**

[C3]. 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)**

[C4]. **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)**

[C5]. 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)**

[C6]. 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)**

[C7]. 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)**

[C8]. 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)**

[C9]. 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)**

[C10]. 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)**

[C11]. 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)**

[C12]. 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)**

[C13]. 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)**

[C14]. **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)**

[C15]. 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)**

[C16]. **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)**

[C17]. **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)**

[C18]. **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)**

[C19]. 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)**

[C20]. 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)**

[C21]. 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)

[C22]. 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)**

[C23]. 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)**

[C24]. 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)**

[C25]. **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)**

[C26]. 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)**

[C27]. 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)**

[C28]. **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)**

[C29]. **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.**

[C30]. 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)**

[C31]. 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)**

[C32]. 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)**

[C33]. 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)**

[C34]. **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)**

[C35]. 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

2019 "Multi-kernel wavefront scheduler", US20190370059A1.

GRANT

2020 "FoMR: A Software and Hardware Codesign for Addressing the Performance Bottlenecks in Secure NVM". Funded by National Science Foundation. co-PI.

2020 "Embracing Heterogeneity in Modern GPUs". Funded by Pitt Momentum. PI.

2019 Pitt startup funding package for tenure-stream assistant professor.

TEACHING

2021 Spring	Instructor, CS1541 - Introduction to Computer Architecture - at Pitt
2021 Spring	Instructor, CS2419 - 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

- 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 NAS 2019, ASP-DAC 2020, HPCA 2020, ASPLOS 2020, ISCA 2020, MICRO 2020, PACT 2020, NAS 2020, HPCA 2021, PLDI 2021, MICRO 2021, NAS 2021
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