MINJIA ZHANG

Bellevue, Washington, USA 98052
(+1)614-940-0520 ⋄ minjiaz@microsoft.com ⋄ zhangminjia.me

EDUCATION

Doctor of Philosophy (Ph.D.) *September 2010 - September 2016*
Ohio State University, Columbus, OH, USA
Computer Science and Engineering
Advisor: Michael D. Bond; Committee: P. Sadayappan, Atanas Rountev, Radu Teodorescu

Master of Science (M.S.) *September 2008 - August 2010*
Huazhong University of Science and Technology, Wuhan, China
Computer Science and Engineering
Advisor: Hai Jin

Bachelor of Science (B.S.) *July 2004 - June 2008*
Huazhong University of Science and Technology, Wuhan, China
Computer Science and Engineering

RESEARCH EXPERIENCE

- Microsoft, Redmond, WA 10/2016–Present
  Principal Researcher 08/2020–Present
  - Efficient, effective, and easy-to-use DL training and inference library
    - Working on DeepSpeed (https://www.deepspeed.ai/), an open source deep learning optimization library that makes massive scale training and inference easy, efficient, and effective for everyone.
    - Working on democratizing large-scale AI training [USENIX ATC 2021].
    - Working on DL training through heterogeneous memory [HPCA 2021, NVMW 2021].
    - Working on sparse training [NeurIPS 2020].
    - Working on expedited DL inference optimization [NeurIPS 2020, IPDPS 2021, ICLR 2021]

  Senior RSDE 10/2016–08/2020
  - Ultra Fast & High Throughput ML/DL Acceleration Library on CPUs
    - Collaboratively designed and built the fastest library for RNN-based DL models on CPUs that achieved 10X lower latency, 10X-100X more throughput and 10X-100X cost reduction than existing DL frameworks such as TensorFlow and CNTK [USENIX ATC 2018].
    - Helped shipping 10+ DL models into production with great latency and cost reduction. Enabling these models in production are critical to advance and empower Microsofts intelligent search service [OpML'19].
  - Efficient DNN Training Methods
    - Collaboratively built the DeepSpeed library.
    - Speeding up the training process Transformer networks by up to 2.3 times through progressive layer drop [NeurIPS'20].
    - Support large model DNN training through heterogeneous memory [HPCA' 21].

  - Deep Learning Model Optimization
    * Speeding up the decoding process of NLP models through graph-based decoder by an order of magnitude [NeurIPS 2018].
    * DL model compression through structured sparsity [ICLR 2018].
  - Efficient DNN Compilation
    * Speeding up the learning to compile method by 1.3–3.9 times through adaptive tensor program compilation [NeurIPS 2020].
  - Large-Scale Vector Search
    * Capacity-optimized multi-store ANN algorithm that allows vector search to benefit from both DRAM and SSDs simultaneously with high accuracy and low latency [CIKM 2019].
    * Learning-based approach to early terminate ANN search for billion-scale datasets [SIGMOD 2020].
    * Support efficient billion-point nearest neighbor search through heterogeneous memory [NeurIPS 2020].

- Microsoft Research, Redmond, WA 05/2016–08/2016
  Research Intern Mentors: Kathryn McKinley, Yuxiong He, Sameh Elnikety
  - Worked on supporting global snapshot transactions in SQL Data Warehouse.
  - Worked on the design of supporting global snapshot transactions using logical counter and centralized time authority.
  - Implemented and integrated global snapshot transactions in SQL DW.

- Programming Language and Software System Lab, Ohio State University 05/2011–05/2016
  Research Assistant Advisor: Michael D. Bond
  Hardware is becoming more parallel. Writing concurrent programs to utilize parallel hardware is hard and error-prone. Exploring and building efficient runtime systems and runtime analysis to make complex, concurrent programs more reliable and scalable. To achieve the best performance and scalability, exploited optimizations opportunities in managed runtime: JIT compiler, garbage collector, adaptive system, and utilized both static analysis and dynamic analysis.
  - Transactional memory helps improve the programmability of writing concurrent programs. Designed and built software transactional memory systems with strong isolation, low overhead, and strong progress guarantees. Demonstrated that the system was significantly faster than existing three state-of-the-art STMs [PPoPP 2015].
  - Introduced a novel relaxed dependence tracking mechanism for capturing and enforcing cross-thread dependences. Built RT, and two client analysis based on RT: a dependence recorder and an STM system. RT reduced 49% execution time on average and achieved 6X maximally speedup [CC 2016].
  - Worked on data ownership locking, non-blocking synchronization, static analysis to eliminate redundant instrumentation in dependence tracking with biased reader-writer lock (Octet) [OOPSAL 2013].
  - Worked on the speculation approach to support the SBRS memory model. [ASPLOS 2015].
  - Worked on the online & offline profiling in runtime system in hybrid tracking [PPoPP 2016].
  - Strong memory models that throw consistency exceptions trade one problem (undefined semantics) for another (poor availability). Showed how to improve availability while still pro-

viding well-defined semantics for programs with data races, including how to relax semantics in a principled way for better availability. [ISMM 2017].

- Microsoft Research, Redmond, WA 09/2015–11/2015
  Research Intern Mentors: Kathryn McKinley, Sameh Elnikety, Yuxiong He, Srikumar Rangarajan
  - Designed and implemented global-snapshot in SQL Server and Azure SQL.
  - System demo was selected to appear at the SQL Server MVP Tell & Show event.
- Microsoft Research, Redmond, WA 05/2015–08/2015 Research Intern Mentors: Kathryn McKinley, Sameh Elnikety, Yuxiong He
  - Designed and implemented session-consistency for read-scaling.
  - Supported session-consistency by modifying the HaDr and log-shipping subsystems in SQL Server.
  - Integrated the session-consistency support, a custom load-balancer, and the EQ frontend to provide session-consistent reads in the production code.
- Network Based Computing Lab, Ohio State University 09/2010–03/201
  Research Assistant Mentor: Dhabaleswar K. Panda
  Memcached is an important component for large-scale data processing. However, memcached is written with sockets and do not deliver best performance on datacenters with high performance networks. Performed case studies on the new design of memcached with high performance interconnects and demonstrated its associated benefits.
  - Examined the challenges in redesigning the network and I/O part of memcached with InfiniBand and RDMA. Changed Memcached to utilize one-sided RDMA reads and writes over InfiniBand. Improved the throughput significantly and reduced memccheds latency to the scale of μs. [ICPP 2011]
- Service Computing Technology and System Lab & Cluster and Grid Computing Lab, HUST 09/2008–08/2010
  Research Assistant Advisor: Hai Jin Mentors: Song Wu, Xuanhua Shi
  VNIX is a three-layer platform for management of virtualization resource in a distributed and cloud computing environment. Added new functionalities and optimizations to help administrators to manage a large number of distributed VMs. [ICPADS 2010]
  - Implemented the front end of virtual network management in VNIX.
  - Reduced the average data transferring size in live migration by 44.1% by compressing dirtied memory with LZO compression algorithm.
- Center for Biomedical Imaging and Bioinformatics, HUST 11/2006–09/2007
  Undergraduate Researcher Mentors: Enmin Song, Renchao Jin
  The human genome project constantly requires searching common segmentations of tens of thousands of genes. It is therefore important to find a fast algorithm to do common genome segmentation search.
  - Proposed and implemented several substring algorithms (including Ukkonens suffix tree algorithm) to solve this problem, and proved that the genome classification problem could be solved in linear time complexity.

REFEREED CONFERENCE

1. **NSDI 2022** John Thorpe, Pengzhan Zhao, Jonathan Eyolfson, Yifan Qiao, Zhihao Jia, **Minjia Zhang**, Ravi Netravali, Guoqing Harry Xu, "Bamboo: Making Preemptible Instances Resilient for Affordable Training of Large DNNs", (acceptance rates: 50/272=18.4%)

2. **SC 2022** Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, **Minjia Zhang**, Olatunji Ruwase, Reza Yazdani Aminabadi, Shaden Smith, Yuxiong He "Enabling Efficient Inference of Transformer Models at Unprecedented Scale", (acceptance rates: 81/320=25.3%)

3. **ICML 2022** Samyam Rajbhandari, Conglong Li, Zhewei Yao, **Minjia Zhang**, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, Yuxiong He "Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale", (acceptance rates: 1117/5630=21.9%)

4. **DAC 2022** Soobee Lee, Minindu Weerakoon, Jonghyun Choi, **Minjia Zhang**, Di Wang, Myeongjae Jeon, "Hierarchical Memory for Continual Learning", (acceptance rates: 20-25%)

5. **AAAI 2022 Minjia Zhang**,Niranjan Uma Naresh, Yuxiong He, "Adversarial Data Augmentation for Task-Specific Knowledge Distillation of Pre-Trained Transformers", in the Thirty-Sixth AAAI Conference on Artificial Intelligence (acceptance rates: 1349/9251=15%)

6. **WSDM 2022 Minjia Zhang**, Wenhan Wang, Yuxiong He, "GraSP: Optimizing Graph-based Nearest Neighbor Search with Subgraph Sampling and Pruning", in the Fifteenth International Conference on Web Search and Data Mining (acceptance rate: 159/786=20.2%)

7. **NeurIPS 2021** Connor Holmes, **Minjia Zhang**, Yuxiong He, Bo Wu, "NxMTransformer: Semi-Structured Sparsification for Natural Language Understanding via ADMM", in the thirty-fifth Annual Conference on Neural Information Processing Systems (acceptance rate: 2372/9122=26%)

8. **USENIX ATC 2021** Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, **Minjia Zhang**, Dong Li, Yuxiong He, "Optimizer-Offload: Democratizing Billion-Scale Model Training", in the 2021 USENIX Annual Technical Conference (acceptance rate: 64/341=18.7%)

9. **ICLR 2021 Minjia Zhang***, Menghao Li*, Chi Wang, Minqin Li, "DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation", in the 9th International Conference on Learning Representations

10. **IPDPS 2021 Minjia Zhang***, Zehua Hu*, Minqin Li, "DUET: Compiler-Aware Subgraph Scheduling for Tensor Programs on a Coupled CPU-GPU Architecture", in the 35th IEEE International Parallel & Distributed Processing Symposium, Portland, Oregon, USA (acceptance rate: 105/462=22.7%)

11. **HPCA 2021** Jie Ren, Jiaolin Luo, Kai Wu, **Minjia Zhang**, Hyeran Jeon, Dong Li, "Efficient Tensor Migration and Allocation on Heterogeneous Memory Systems for Deep Learning", in the 27th IEEE International Symposium on High-Performance Computer Architecture, Seoul, South Korea (acceptance rate: 63/258=24.4%)

12. **NeurIPS 2020 Minjia Zhang**, Yuxiong He, "Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping", in the thirty-fourth Annual Conference on Neural Information Processing Systems, Vancouver Canada, December 2020. (acceptance rate: 1900/9454=20%)

13. **NeurIPS 2020** Jie Ren, **Minjia Zhang**, Dong Li, "HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory", in the thirty-fourth Annual Conference on Neural Information Processing Systems, Vancouver Canada, December 2020. (acceptance rate: 1900/9454=20%)

14. **NeurIPS 2020** Menghao Li*, **Minjia Zhang***, Chi Wang, Minqin Li, "AdaTune: Adaptive Tensor Program Compilation Made Efficient", in the thirty-fourth Annual Conference on Neu-

ral Information Processing Systems, Vancouver Canada, December 2020. *Equal contribution. (acceptance rate: 1900/9454=20%)

15. **SIGMOD 2020** Conglong Li, **Minjia Zhang**, Yuxiong He, David Anderson, "Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination", in the 2020 ACM International Conference of Management of Data, Portland, OR, USA (acceptance rate: 123/458=26.9%)

16. **CIKM 2019 Minjia Zhang**, Yuxiong He, "GRIP: Multi-Store Capacity-Optimized High-Performance Nearest Neighbor Search for Vector Search Engine", in the 28th ACM International Conference on Information and Knowledge Management, Beijing, China (acceptance rate: 200/1030=19.4%)

17. **USENIX OpML 2019 Minjia Zhang**, Samyam Rajbandari, Wenhan Wang, Elton Zheng, Olatunji Ruwase, Jeff Rasley, Jason Li, Junhua Wang, Yuxiong He, "Accelerating Large Scale Deep Learning Inference through DeepCPU at Microsoft", in the 2019 USENIX Conference on Operational Machine Learning.

18. **NeurIPS 2018 Minjia Zhang**, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, Yuxiong He, Navigating with Graph Representations for Fast and Scalable Decoding of Neural Language Models, In the thirty-second Annual Conference on Neural Information Processing Systems, Montral CANADA, December 2018. (acceptance rate: 1010/4854=20.8%)

19. **USENIX ATC 2018 Minjia Zhang***, Samyam Rajbhandari*, Wenhan Wang, Yuxiong He, DeepCPU: Serving RNN-based Deep Learning Models 10x Faster, in 2018 USENIX Annual Technical Conference. *Equal contribution. (acceptance rate: 76/378=20.1%)

20. **ICLR 2018** Wei Wen, Yuxiong He, Samyam Rajbhandari, **Minjia Zhang**, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen, Hai Li, Learning Intrinsic Sparse Structures within Long Short-Term Memory, in the 6th International Conference on Learning Representations. (acceptance rate: 337/937=36%)

21. **ISMM 2017 Minjia Zhang**, Swarnendu Biswas, Michael Bond, Avoiding Consistency Exceptions Under Strong Memory Consistency Models, In 2017 ACM SIGPLAN International Symposium on Memory Management, June 2017, Barcelona, Spain

22. **CC 2017** Swarnendu Biswas, Man Cao, **Minjia Zhang**, Michael Bond and Ben Wood, Lightweight Data Race Detection for Production Runs, In 26th International Conference onCompiler Construction, February 2017.

23. **CC 2017 Minjia Zhang**, Swarnendu Biswas, Michael D. Bond, Relaxed Dependence Tracking for Parallel Runtime Support, In 25th International Conference onCompiler Construction, March 2016.

24. **PPoPP 2017 Minjia Zhang**, Swarnendu Biswas, Michael D. Bond, POSTER: On the Problem of Consistency Exceptions in the Context of Strong Memory Models, In ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, February, 2017, Austin, Texas, USA

25. **PPoPP 2016** Man Cao, **Minjia Zhang**, Aritra Sengupta, and Michael D. Bond, Drinking from Both Glasses: Combining Pessimistic and Optimistic Tracking of Cross-Thread Dependences, In ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, March 2016.

26. **OOPSLA 2015** Swarnendu Biswas, **Minjia Zhang**, Michael D. Bond, and Brandon Lucia, Valor: Efficient, Software-Only Region Conflict Exceptions (Distinguished Artifact Award, Distinguished Paper Award), In ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications, October 2015.

27. **SPLASH 2015 Companion Minjia Zhang**, SIRe: An Efficient Snapshot Isolation-based Memory Model for Detecting and Tolerating Region Conflicts, SPLASH '15 Companion, October, 2015,

Pittsburgh, PA, USA

28. **PPoPP 2015** Minjia Zhang, Jipeng Huang, Man Cao, and Michael D. Bond, Low-Overhead Software Transactional Memory with Progress Guarantees and Strong Semantics, In ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, February 2015.

29. **ASPLOS 2015** Aritra Sengupta, Swarnendu Biswas, Minjia Zhang, Michael D. Bond, and Milind Kulkarni, Hybrid Static-Dynamic Analysis for Statically Bounded Region Serializability, In ACM Conference on Architectural Support for Programming Languages and Operating Systems, March 2015.

30. **OOPSLA 2013** Michael D. Bond, Milind Kulkarni, Man Cao, **Minjia Zhang**, Meisam Fathi Salmi, Swarnendu Biswas, Aritra Sengupta, and Jipeng Huang Octet: Capturing and Controlling Cross-Thread Dependences Efficiently , In ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA), October 2013.

31. **ICPP 2011** Jithin Jose, Hari Subramoni, Miao Luo, **Minjia Zhang**, Jian Huang, Md. Wasi-ur-Rahman, Nusrat S. Islam, Xiangyong Ouyang, Hao Wang, Sayantan Sur, and D. K. Panda, Memcached Design on High Performance RDMA Capable Interconnects, In International Conference on Parallel Processing (ICPP), September 2011.

32. **ICPADS 2010 Minjia Zhang**, Hai Jin, Song Wu, Xuanhua Shi "VirtCFT: A Transparent VM-level Fault-Tolerant System for Virtual Clusters", In International Conference on Parallel and Distributed System (ICPADS), Dec 2010.

JOURNAL ARTICLES

1. **TECS** Reza Yazdani, Olatunji Ruwase, **Minjia Zhang**, Yuxiong He, Jose-Maria Arnau, Antonio Gonzalez, "SHARP: An Adaptable, Energy-Efficient Accelerator for Recurrent Neural Network", In the ACM Transactions on Embedded Computing Systems 2022.

2. **TOPC 2017** Man Cao, **Minjia Zhang**, Aritra Sengupta, Swarnendu Biswas, and Michael D. Bond, Hybridizing and Relaxing Dependence Tracking for Efficient Parallel Runtime Support, In ACM Transactions on Parallel Computing, April 2017.

WORKSHOP PAPERS

1. **EMDC 2022** Yongbo Yu, Fuxun Yu, Zirui Xu, Di Wang, **Minjia Zhang**, Ang Li, Shawn Bray, Chenchen Liu and Xiang Chen, "Powering Multi-Task Federated Learning with Competitive GPU Resource Sharing", in the Second International Workshop on the Efficiency of Modern Data Centers.

2. **MLSys 2022** Fuxun Yu, Yongbo Yu, Di Wang, **Minjia Zhang**, Longfei Shangguan, Tolga Soyata, Chenchen Liu and Xiang Chen, "A Survey on Multi-Tenant DL Inference on GPU", in the MLSYS'22 workshop on Cloud Intelligence/AIOps.

3. **NVMW 2021** Jie Ren, **Minjia Zhang**, Dong Li, "HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory", in the 12th Non-Volatile Memories Workshop, San Diego, USA

4. **NVMW 2021** Jie Ren, Jiaolin Luo, Kai Wu, **Minjia Zhang**, Hyeran Jeon, Dong Li, "HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory", in the 12th Non-Volatile Memories Workshop, San Diego, USA

5. **WODET 2014** Man Cao, **Minjia Zhang**, and Michael D. Bond, Drinking from Both Glasses: Adaptively Combining Pessimistic and Optimistic Synchronization for Efficient Parallel Runtime Support, In the 5th Workshop on Determinism and Correctness in Parallel Programming, March 2014.

TECHNICAL REPORTS AND PREPRINT

1. Samyam Rajbhandari, Conglong Li, Zhewei Yao, **Minjia Zhang**, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, Yuxiong He, "DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale", https://arxiv.org/abs/2201.05596
2. Dantong Zhu, Minjia Zhang, "Understanding and Generalizing Monotonic Proximity Graphs for Approximate Nearest Neighbor Search", https://arxiv.org/pdf/2107.13052.pdf
3. Reza Yazdani, Olatunji Ruwase, **Minjia Zhang**, Yuxiong He, Jose-Maria Arnau, Antonio Gonzalez, "LSTM-Sharp: An Adaptable, Energy-Efficient Hardware Accelerator for Long Short-Term Memory", https://arxiv.org/abs/1911.01258
4. **Minjia Zhang**, Swarnendu Biswass, Michael D. Bond, All That Glitters is Not Gold: Improving Availability and Practicality of Exception-Based Memory Models, Technical Report #OSU-CISRC-4/16-TR01, Computer Science & Engineering, Ohio State University, Apr 2016.
5. **Minjia Zhang**, Swarnendu Biswass, Michael D. Bond, Optimizing Parallel Runtime Support with Asynchronous Coordination, Technical Report #OSU-CISRC-11/15-TR23, Computer Science & Engineering, Ohio State University, Nov 2015.
6. Swarnendu Biswas, **Minjia Zhang**, Michael D. Bond, and Brandon Lucia, Efficient, Software-Only Data Race Exceptions, Technical Report #OSU-CISRC-3/15-TR04, Computer Science & Engineering, Ohio State University, Mar 2015.
7. Swarnendu Biswas, **Minjia Zhang**, and Michael D. Bond, Lightweight Data Race Detection for Production Runs, Technical Report #OSU-CISRC-1/15-TR01. Computer Science & Engineering, Ohio State University, Jan 2015.
8. **Minjia Zhang**, Jipeng. Huang, Man Cao, and Mike D. Bond. LarkTM: Efficient, strongly atomic software transactional memory. Technical Report #OSU-CISRC-11/12-TR17, Computer Science & Engineering, Ohio State University, Nov 2012.

PATENTS

1. **Minjia Zhang**, Yuxiong He, "Multi-layer Semantic Search", U.S. Patent, MS# 406007-US-NP, 2019
2. **Minjia Zhang**, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, Yuxiong He, Graph Representations for Identifying a Next Word, US 2019 / 0377792 A1
3. **Minjia Zhang**, Samyam Rajbhandari, Wenhan Wang, Yuxiong He, Deep Learning Model Scheduling, US 2019 / 0311245 A1

INVITED TALKS AND PRESENTATIONS

- Invited Talks
    1. Invited talk by Saurabh Tangri on "Extreme Compression for Pre-trained Transformers Made Simple and Efficient" at Intel AI Group, July 28th 2022
    2. Invited lecture by Zhihao Jia on "DeepSpeed: The library to accelerate training and inference of DNN at scale" at CMU, April 18th 2022
    3. Invited lecture on "DeepSpeed: The library to accelerate training and inference of DNN at scale" at the Efficient Large-Scale AI Workshop as a part of MSR Project Green., April 15th 2022
    4. Invited lecture by Myeongjae Jeon on "DeepSpeed: The library to accelerate training and inference of DNN at scale" at UNIST, April 13th 2022

5. Invited lecture on "New algorithms for Approximate Nearest Neighbor Search Systems at Scale" at Kent State University, October 20, 2022
6. Invited keynote speech on "DL Inference and Training Optimization Towards Speed and Scale" at Tsinghua AIR 2021
7. Invited keynote speech on "DL Inference and Training Optimization Towards Speed and Scale" at EMDC 2021
8. Invited talk on "DL Inference Optimization Towards Speed & Scale" at the ICT Young Scholars' Forum, 2020, Beijing, China
9. Invited talk on "TVM@Microsoft" at the TVM and Deep Learning Compilation Conference 2019, Seattle, Washington, US
10. Invited talk on DeepCPU: Deep Learning Serving Optimizations on CPUs at the Deep Learning workshop at Microsoft TechFest 2018, March 2018, Redmond, WA, USA
11. Invited talk on DeepCPU: Deep Learning Serving Optimizations on CPUs at Microsoft Research Talk Series, February 2018, Redmond, WA, USA
12. Invited talk on DeepCPU: Deep Learning Serving Optimizations on CPUs at Microsoft Machine Learning, AI & Data Science Conference (MLADS) December 2017, Redmond, WA, USA

- Presentations
  1. Presented work on adversarial data augmentation for knowledge distillation at AAAI 2022
  2. Presented work on graph sampling and pruning for nearest neighbor search at WSDM 2022
  3. Presented work on "DUET: Compiler-Aware Subgraph Scheduling for Tensor Programs on a Coupled CPU-GPU Architecture" at IPDPS 2021.
  4. Presented work on "DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation" at ICLR 2021
  5. Presented work on "Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping" at NeurIPS 2020
  6. Presented work on "AdaTune: Adaptive Tensor Program Compilation Made Efficient" at NeurIPS 2020
  7. Presented work on "GRIP: Multi-Store Capacity-Optimized High-Performance Nearest Neighbor Search for Vector Search Engine" at CIKM 2019, Beijing, China
  8. Presented work on Accelerating Large Scale Deep Learning Inference through DeepCPU at Microsoft at 2019 USENIX OpML, May 2019, Santa Clara, CA, USA
  9. Presented work on DeepCPU: Serving RNN-based Deep Learning Models 10x Faster at 2018 USENIX Annual Technical Conference, July 2018, Boston, MA, USA
  10. Presented work on detecting and tolerating region conflicts to support region snapshot isolation at ACM Student Research Competition, OOPSLA 2015, Pittsburg, PA, USA
  11. Presented work on low-overhead and scalable software transactional memory with strong progress guarantees at the 20st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), 2015, San Francisco, CA, USA
  12. Presented work on building efficient, strongly atomic, and scalable software transactional memory at ACM Student Research Competition, PLDI 2013, Seattle, WA, USA

ADVISING AND MENTORING

1. Yucheng Lu, Ph.D. student at University of Cornell, maximizing the communication efficiency of DNN training with 0/1 Adam, June 2021-Feb 2022
2. Connor Holmes, Ph.D. student at University of Colorado at Boulder, DNN sparsification (published at NeurIPS 2021), May 2021-Present
3. Zhen Peng, Ph.D. student at University of William and Marry, ultra-fast graph-based ANN search, co-advised with Bin Ren and Ruoming Jin, Sep 2019-June 2021
4. Hongyi Wang, Ph.D. student at University of Wisconsin Madison, efficient DNN training, June 2020-September 2020
5. Jie Ren, Ph.D. student at University of California at Merced, DL training and inference via heterogeneous memory (published at HPCA 2020/NeurIPS 2020/USENIX ATC 2021), June 2020-September 2020
6. Connor Holmes, Ph.D. student at University of Colorado at Boulder, Exploiting sparsity in DNN inference, June 2020-September 2020
7. Dantong Zhu, Ph.D. student at Georgia Institute of Technology, Monotonic relative nearest neighbor graph for ANN search, January 2020-June 2020
8. Zehua Hu, M.S. student at Beijing University, Graph partitioning of TVM relay IR for heterogeneous DL serving (published at IPDPS 2021), July 2019-March 2021
9. Menghao Li, M.S. student at Beijing University, Bayesian optimization for Optimizing the autotuning process of DL compiler (published at NeurIPS 2020/ICLR 2021), February 2020-March 2021
10. Conglong Li, Ph.D. student at Carnegie Melon University, learning-based early termination for ANN search (published at SIGMOD 2020), May 2019-August 2019
11. Stephen Zhou, Ph.D., student at Massachusetts Institute of Technology, Automatic model optimization, June 2018-August 2018

HONORS AND REWARDS

1. Microsoft Excellence Award 2020
2. Microsoft Excellence Award 2018
3. Microsoft Excellence Award 2017
4. Microsoft CTO Kevin Scott selected DeepCPU as one of three topics for "Cool Tech" showcase
5. Bronze medal of the Student Research Competition in SPLASH 2015
6. Distinguished Paper Award in OOPSLA 2015
7. Distinguished Artifact Award in OOPSLA 2015
8. ACM Artifact Evaluation Stamp for Low-Overhead STM in PPoPP 2015
9. NSF Travel Award for presenting at PPoPP 2015 and SPLASH 2015
10. Silver medal of the Student Research Competition in PLDI 2013
11. Awarded Ohio State University Fellowship in 2010, 2011, 2013
12. Awarded Chinese National Scholarship (top 2%) in 2008
13. Awarded Teyiu (Honor) Student at HUST (top 2%) in 2007

14. Awarded merit certificate and scholarship for 8 out of 8 semesters in 2005, 2006, 2007, 2008 for academic performance
15. Exempted from 2004 National College Entrance Examination (NCEE) in China for excellent academic performance($< 0.3\%$)
16. Awarded 1st Prize in Chinese National Math Olympiad Competition in 2004

PROFESSIONAL SERVICES

- **Chair**
  1. Session Chair: The 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2019)
  2. Publicitiy Co-Chair: The ACM SIGPLAN Conference on Programming Language Design and Implementation 2019 (PLDI 2019)
- **Program Committee**
  1. Sub-Committee: Microsoft E+D Research Council.
  2. Program Committee: The 34th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2020)
  3. The 12th International Conference on Mobile, Hybrid, and On-line Learning (eLmL 2020)
  4. Program Committee: The 33th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2019)
  5. Program Committee: The 32th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2018)
  6. Shadow Program Committee: The 23th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2018)
  7. Artifact Evaluation Committee: The ACM SIGPLAN Conference on Programming Language Design and Implementation 2017 (PLDI 2017)
  8. Artifact Evaluation Committee: The ACM SIGPLAN conference on Systems, Programming, Languages and Applications: Software for Humanity (SPLASH 2015)
  9. Artifact Evaluation Committee: The ACM SIGPLAN Conference on Programming Language Design and Implementation 2015 (PLDI 2015)
- **Reviewer**
  1. Reviewer: European Conference on Computer Vision 2022 (ECCV 2022)
  2. Reviewer: The International Conference on Machine Learning (ICML 2022)
  3. External Reviewer: The 2022 USENIX Annual Technical Conference (USENIX ATC 2022)
  4. Reviewer: The 2022 ML Reproducibility Challenge (RC 2022)
  5. Reviewer: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)
  6. Reviewer: The 10th International Conference on Learning Representations (ICLR 2022)
  7. Reviewer: The 36th AAAI Conference on Artificial Intelligence (AAAI 2022)
  8. Reviewer: The International Conference on Machine Learning (ICML 2021)

9. Reviewer: The IEEE/CVF International Conference on Computer Vision (ICCV 2021)
10. Reviewer: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)
11. Reviewer: The 9th International Conference on Learning Representations (ICLR 2021)
12. Reviewer: The 35th AAAI Conference on Artificial Intelligence (AAAI 2021)
13. Reviewer: The 26th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021)
14. Reviewer: The 34th Conference on Neural Information Processing Systems (NeurIPS 2020)
15. Reviewer: The 8th International Conference on Learning Representations (ICLR 2020)
16. Reviewer: The 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)
17. Reviewer: NeurIPS 2019 Reproducibiklity Challenge
18. Reviewer: The ACM SIGPLAN Conference on Programming Language Design and Implementation 2019 (PLDI 2019)
19. Reviewer: The 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2019)
20. Subreviewer: The 19th International Middleware Conference 2018
21. Subreviewer: The 15th IEEE International Conference on Autonomic Computing 2018
22. Subreviewer: The IEEE International Conference on Cloud Computing (CLOUD) 2018
23. Subreviewer: The 24th IEEE International Conference on High Performance Computing, Data, and Analytics 2017
24. Subreviewer: The 14th IEEE International Conference on Autonomic Computing 2017
25. Subreviewer: The 7th Workshop on the Theory of Transactional Memory WTTM 2015
26. Journal Reviewer: IEEE Access 2020
27. Journal Reviewer: Journal of Systems and Software 2020
28. Journal Reviewer: IEEE Transaction on Cloud Computing 2019
29. Journal Reviewer: ACM Transaction on Privacy and Security 2019
30. Journal Reviewer: Concurrency and Computation: Practice and Experience 2018
31. Journal Reviewer: Journal of Computer Science 2017
32. Journal Reviewer: Concurrency and Computation: Practice and Experience 2017

SELECTED PRESS COVERAGE

1. **Microsoft Research Blog**, DeepSpeed: Advancing MoE inference and training to power next-generation AI scale, Jan 19, 2022
2. **Microsoft Research Blog**, DeepSpeed powers 8x larger MoE model training with high performance, August 18, 2021
3. **Microsoft Research Blog**, DeepSpeed: Accelerating large-scale model inference and training via system optimizations and compression, May 24, 2021

4. **Towards Data Science**, Microsoft ZeRO-Offload: Democratizing Billion-Scale Model Training, Jan 28, 2021
5. **Medium**, ZeRO-Offload: Training Multi-Billion Parameter Models on a Single GPU, Jan 27, 2021
6. **The Batch**, Toward 1 Trillion Parameters, Sep 16, 2020
7. **Analytics India Magazine**, Microsoft Releases Latest Version Of DeepSpeed, Its Python Library For Deep Learning Optimisation, Sep 15, 2020
8. **siliconANGLE**, Microsoft AI tool enables extremely large models with a trillion parameters, Sep 11, 2020
9. **Microsoft Research Blog**, DeepSpeed: Extreme-scale model training for everyone, Sep 10, 2020
10. **VentureBeaet**, Microsofts updated DeepSpeed can train trillion-parameter AI models with fewer GPUs, Sep 10, 2020
11. **DeepSpeed.ai**, Microsoft DeepSpeed achieves the fastest BERT training time, May 27, 2020
12. **Microsoft Research Blog**, Research Collection: Tools and Data to Advance the State of the Art, May 19, 2020
13. **Microsoft Research Blog**, ZeRO-2 & DeepSpeed: Shattering barriers of deep learning speed & scale, May 19, 2020
14. **Microsoft Research Blog**, ZeRO & DeepSpeed: New system optimizations enable training models with over 100 billion parameters, Feb 13, 2020
15. **WinBuzzer**, Microsoft DeepSpeed with Zero Can Train 100 Billion Parameter AI Models, Feb 11, 2020
16. **MSPoweruser**, Meet Microsoft DeepSpeed, a new deep learning library that can train massive 100-billion-parameter models, Feb 10, 2020
17. **VentureBeat**, Microsoft trains worlds largest Transformer language model, Feb 10, 2020
18. **InfoWorld**, Microsoft speeds up PyTorch with DeepSpeed, Feb 10, 2020