Minjia Zhang

Urbana, Illinois, USA 61801

📱 (+1) 614 940 0520 • ✉ minjiaz@illinois.edu
🌐 https://minjiazhang.github.io/

Research Interests

High-performance systems for deep learning and machine learning: parallel, distributed, heterogeneous system design, and algorithms, with applications to NLP and multi-modality models

Education

Ohio State University **Columbus, OH**
Ph.D. in Computer Science *2010–2016*
Advisor: Michael D. Bond
PhD's Dissertation: Efficient and Scalable Runtime Support for Parallelism
Committee Members: P. Sadayappan, Atanas Rountev, Radu Teodorescu

Huazhong University of Science and Technology **Wuhan, China**
M.S. in Computer Science *2008–2010*

Huazhong University of Science and Technology **Wuhan, China**
B.S. in Computer Science *2004–2008*

Academic Appointments

University of Illinois Urbana-Champaign **Urbana, IL**
Assistant Professor *2024–Present*

Microsoft AI and Research **Redmond, WA**
Principal Researcher *2020–2024*

Microsoft Research Redmond **Redmond, WA**
Senior Researcher *2016–2020*

Microsoft Research Redmond **Redmond, WA**
Research Intern *2016 Summer*

Microsoft Research Redmond **Redmond, WA**
Research Intern *2015 Fall*

Microsoft Research Redmond **Redmond, WA**
Research Intern *2015 Summer*

Honors and Awards

2025: Google ML and Systems Junior Faculty Award
2025: NSF CAREER Award
2024: AMD AI & HPC Award
2024: ICLR Outstanding Paper Honorable Mention
2020, 2018, 2017: Microsoft Excellence Awards
2017: Selected for Microsoft CTO Kevin Scott's one of the three "Cool Tech" Showcase (DeepCPU)
2015: OOPSLA Distinguished Paper Award
2015: OOPSLA Distinguished Artifact Award

2015: Bronze Medal, SPLASH Student Research Competition
2015: NSF Travel Award for PPoPP and SPLASH
2013: Silver Medal, PLDI Student Research Competition
2010,2011,2013: Ohio State University Fellowship
2008: First-Class Chinese National Scholarship (top 0.2%)
2007: Highest Honor Student (Teyiu Award) at HUST (top 2%)
2005–2008: Merit Scholarship for all 8 semesters at HUST
2004: Exempted from National College Entrance Examination (top 0.3%)
2004: 1st Prize, Chinese National Math Olympiad

Refereed Conference

ASPLOS 2026: Xinyu Lian, Masahiro Tanaka, Olatunji Ruwase, **Minjia Zhang**, *SuperOffload: Unleashing the Power of Large-Scale LLM Training on Superchips* (acceptance rate: 20/208 = 9.8%)

SIGMOD 2026: Jingyi Xi*, Chenghao Mo*, Ben Karsin, Artem Chirkin, Mingqin Li, **Minjia Zhang**, *VecFlow: A High-Performance Vector Data Management System for Filtered-Search on GPUs* (acceptance rate: XX/XX = XX%)

SC 2025: Yueming Yuan, Ahan Gupta, Jianping Li, Sajal Dash, Feiyi Wang, **Minjia Zhang**, *X-MoE: Enabling Scalable Training for Emerging Mixture-of-Experts Architectures on HPC Platforms* (acceptance rate: 136/643 = 21.2%), **Nominated for Best Student Paper Award**

ICCV 2025: Nick Gong, Zhen Zhu, **Minjia Zhang**, *InstantEdit: Text-Guided Few-Step Image Editing with Piecewise Rectified Flow* (acceptance rate: 2698/1129 = 24%)

ACL Finding 2025: Xiao Wang, Mengjue Tan, Qiao Jin, Guangzhi Xiong, Yu Hu, Aidong Zhang, Zhiyong Lu, **Minjia Zhang**, *"MedCite: Can Language Models Generate Verifiable Text for Medicine?"* (acceptance rate: XX/XX = XX%)

ACL Finding 2025: Akshat Sharma, Hangliang Ding, Jianping Li, Neel Dani, **Minjia Zhang**, *"MiniKV: Pushing the Limits of 2-Bit KV Cache via Compression and System Co-Design for Efficient Long Context Inference"* (acceptance rate: XX/XX = XX%)

USENIX ATC 2025: Xinyu Lian, Sam Ade Jacobs, Lev Kurilenko, Masahiro Tanaka, Stas Bekman, Olatunji Ruwase, **Minjia Zhang**, *"Universal Checkpointing: A Flexible and Efficient Distributed Checkpointing System for Large-Scale DNN Training with Reconfigurable Parallelism"* (acceptance rate: 100/634 = 15.7%)

MLSys 2025: Beichen Huang*, Yueming Yuan*, Zelei Shao*, **Minjia Zhang**, *"MiLo: Efficient Quantized MoE Inference with Mixture of Low-Rank Compensators"* (acceptance rate: 61/271 = 22%)

SenSys 2025: Zhiwei Ren, Junbo Li, **Minjia Zhang**, Di Wang, Xiaoran Fan, Longfei Shangguan, *"Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications"* (acceptance rate: 46/245 = 18%)

HPCA 2025: Shuangyan Yang, **Minjia Zhang**, Dong Li, *"Buffalo: Enabling Large-Scale GNN Training via Memory-Efficient Bucketization"* (acceptance rate: 112/534 = 21%)

HPCA 2025: Zihan Liu, Xinhao Luo, Junxian Guo, Wentao Ni, Yangjie Zhou, Yue Guan, Cong Guo, Weihao Cui, Yu Feng, Minyi Guo, Yuhao Zhu, **Minjia Zhang**, Jingwen Leng, Chen Jin, *"VQ-LLM: High-performance Code Generation for Vector Quantization Augmented LLM Inference"* (acceptance rate: 112/534 = 21%)

ICSE 2025: Xinyu Lian, Yinfang Chen, Runxiang Cheng, Jie Huang, Parth Thakkar, **Minjia Zhang**, Tianyin Xu, *"Large Language Models as Configuration Validators"* (acceptance rate: 22%)

NeurIPS 2024 D&B: Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, **Minjia Zhang**, Qing Li, Baobao Chang, *"UltraEdit: Instruction-based Fine-Grained Image Editing at Scale"* (acceptance rate: 460/1820 = 25.3%)

PODC 2024: Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, **Minjia Zhang**, Reza Yazdani Aminabadi, Shuaiwen Leon Song, Samyam Rajbhandari, Yuxiong He, *"DeepSpeed-Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models"* (acceptance rate: 21.3%)

SIGMOD 2024: Yongye Su, Yinqi Sun, **Minjia Zhang**, Jianguo Wang, *"Vexless: A Serverless Vector Data Management System Using Cloud Function"*

Nature Methods 2024: Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Peter K Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, Leon Song, **Minjia Zhang**, Conglong Li, Shiyang Chen, Yuxiong He, Mohammed AlQuraishi, *"OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization"*

ICLR 2024 (Oral): Suyu Ge, Yunan Zhang, Liyuan Liu, **Minjia Zhang**, Jiawei Han, Jianfeng Gao, *"Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs"* (acceptance rate: 85/7304 = 1.16%)

AAAI 2024: Conglong Li, Zhewei Yao, Xiaoxia Wu, **Minjia Zhang**, Connor Holmes, Cheng Li, Yuxiong He, *"DeepSpeed Data Efficiency: Improving Deep Learning Model Quality and Training Efficiency via Efficient Data Sampling and Routing"* (acceptance rate: 2342/12100 = 23.75%)

NSDI 2024: Jiangfei Duan, Ziang Song, Xupeng Miao, Xiaoli Xi, Dahua Lin, Harry Xu, **Minjia Zhang**, Zhihao Jia, *"Parcae: Proactive, Liveput-Optimized DNN Training on Preemptible Instances"* (acceptance rate: 40/227 = 17.6%)

ECAI 2023: **Minjia Zhang**, Niranjan Uma Naresh, Yuxiong He, *"Revisiting the Efficiency-Accuracy Tradeoff in Adapting Transformer Models via Adversarial Fine-Tuning"* (acceptance rate: 392/1632 = 24%)

ICLR 2023: Yucheng Lu, Conglong Li, **Minjia Zhang**, Christopher De Sa, Yuxiong He, *"0/1 Adam: Maximizing Communication Efficiency for Large-scale Training"* (acceptance rate: 1574/4956 = 32%)

PPoPP 2023: Zhen Peng, **Minjia Zhang**, Kai Li, Ruoming Jin, Bin Ren, *"iQAN: Fast and Accurate Vector Search with Efficient Intra-Query Parallelism on Multi-Core Architectures"* (acceptance rate: 31/131 = 23.6%)

ASPLOS 2023: Shuangyan Yang, **Minjia Zhang**, Wenqian Dong, Dong Li, *"Betty: Enabling Large-Scale GNN Training with Batch-Level Graph Partitioning"* (acceptance rate: 128/598 = 21%)

NSDI 2023: 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 rate: 50/272 = 18.4%)

MobiCom 2023: Xinyue Ma, Suyeon Jeong, **Minjia Zhang**, Di Wang, Jonghyun Choi, Myeonjae Jeon, *"Cost-effective On-device Continual Learning over Memory Hierarchy with Miro"* (acceptance rate: 440/2972 = 15%)

NeurIPS 2022 (Oral): Xiaoxia Wu*, Zhewei Yao*, **Minjia Zhang***, Conglong Li, Yuxiong He, *"Extreme Compression for Pre-trained Transformers Made Simple and Efficient"* (acceptance rate: 183/10411=1.76%)

NeurIPS 2022 (Spotlight): Conglong Li, **Minjia Zhang**, Yuxiong He, *"The Stability-Efficiency Dilemma: Investigating Sequence Length Warmup for Training GPT Models"* (acceptance rate: 2665/10411=25.6%)

NeurIPS 2022 (Spotlight): Zhewei Yao, Reza Yazdani Aminabadi, **Minjia Zhang**, Xiaoxia Wu, Conglong Li, Yuxiong He, *"ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers"* (acceptance rate: 2665/10411=25.6%)

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 rate: 81/320=25.3%)

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 rate: 1117/5630=21.9%)

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

AAAI 2022: **Minjia Zhang**, Niranjan Uma Naresh, Yuxiong He, *"Adversarial Data Augmentation for Task-Specific Knowledge Distillation of Pre-Trained Transformers"* (acceptance rate: 1349/9251 = 15%)

WSDM 2022: **Minjia Zhang**, Wenhan Wang, Yuxiong He, *"GraSP: Optimizing Graph-based Nearest Neighbor*

Search with Subgraph Sampling and Pruning" (acceptance rate: 159/786 = 20.2%)

NeurIPS 2021: Connor Holmes, **Minjia Zhang**, Yuxiong He, Bo Wu, *"NxMTransformer: Semi-Structured Sparsification for Natural Language Understanding via ADMM"* (acceptance rate: 2372/9122 = 26%)

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"* (acceptance rate: 64/341 = 18.7%)

ICLR 2021: **Minjia Zhang**, Menghao Li, Chi Wang, Minqin Li, *"DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation"* (acceptance rate: 860/2997 = 28.7%)

IPDPS 2021: **Minjia Zhang**, Zehua Hu, Minqin Li, *"DUET: Compiler-Aware Subgraph Scheduling for Tensor Programs on a Coupled CPU-GPU Architecture"* (acceptance rate: 105/462 = 22.7%)

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"* (acceptance rate: 63/258 = 24.4%)

NeurIPS 2020: **Minjia Zhang**, Yuxiong He, *"Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping"* (acceptance rate: 1900/9454 = 20%)

NeurIPS 2020: Jie Ren, **Minjia Zhang**, Dong Li, *"HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory"* (acceptance rate: 1900/9454 = 20%)

NeurIPS 2020: Menghao Li, **Minjia Zhang**, Chi Wang, Minqin Li, *"AdaTune: Adaptive Tensor Program Compilation Made Efficient"* *Equal contribution (acceptance rate: 1900/9454 = 20%)

SIGMOD 2020: Conglong Li, **Minjia Zhang**, Yuxiong He, David Anderson, *"Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination"* (acceptance rate: 123/458 = 26.9%)

CIKM 2019: **Minjia Zhang**, Yuxiong He, *"GRIP: Multi-Store Capacity-Optimized High-Performance Nearest Neighbor Search for Vector Search Engine"* (acceptance rate: 200/1030 = 19.4%)

USENIX OpML 2019: **Minjia Zhang**, Samyam Rajbhandari, Wenhan Wang, Elton Zheng, Olatunji Ruwase, Jeff Rasley, Jason Li, Junhua Wang, Yuxiong He, *"Accelerating Large Scale Deep Learning Inference through DeepCPU at Microsoft"*

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"* (acceptance rate: 1010/4854 = 20.8%)

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

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"* (acceptance rate: 337/937 = 36%)

ISMM 2017: **Minjia Zhang**, Swarnendu Biswas, Michael Bond, *"Avoiding Consistency Exceptions Under Strong Memory Consistency Models"*

CC 2017: Swarnendu Biswas, Man Cao, **Minjia Zhang**, Michael Bond, Ben Wood, *"Lightweight Data Race Detection for Production Runs"*

CC 2017: **Minjia Zhang**, Swarnendu Biswas, Michael D. Bond, *"Relaxed Dependence Tracking for Parallel Runtime Support"*

PPoPP 2017: **Minjia Zhang**, Swarnendu Biswas, Michael D. Bond, *"POSTER: On the Problem of Consistency Exceptions in the Context of Strong Memory Models"*

PPoPP 2016: Man Cao, **Minjia Zhang**, Aritra Sengupta, Michael D. Bond, *"Drinking from Both Glasses: Combining Pessimistic and Optimistic Tracking of Cross-Thread Dependences"*

OOPSLA 2015: Swarnendu Biswas, **Minjia Zhang**, Michael D. Bond, Brandon Lucia, *"Valor: Efficient, Software-Only Region Conflict Exceptions"* (**Distinguished Artifact Award, Distinguished Paper Award**)

SPLASH 2015 Companion: **Minjia Zhang**, *"SIRe: An Efficient Snapshot Isolation-based Memory Model for Detecting and Tolerating Region Conflicts"*

PPoPP 2015: Minjia Zhang, Jipeng Huang, Man Cao, Michael D. Bond, *"Low-Overhead Software Transactional*

Memory with Progress Guarantees and Strong Semantics"

ASPLOS 2015: Aritra Sengupta, Swarnendu Biswas, **Minjia Zhang**, Michael D. Bond, Milind Kulkarni, *"Hybrid Static-Dynamic Analysis for Statically Bounded Region Serializability"*

OOPSLA 2013: Michael D. Bond, Milind Kulkarni, Man Cao, **Minjia Zhang**, Meisam Fathi Salmi, Swarnendu Biswas, Aritra Sengupta, Jipeng Huang, *"Octet: Capturing and Controlling Cross-Thread Dependences Efficiently"*

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, D. K. Panda, *"Memcached Design on High Performance RDMA Capable Interconnects"*

ICPADS 2010: **Minjia Zhang**, Hai Jin, Song Wu, Xuanhua Shi, *"VirtCFT: A Transparent VM-level Fault-Tolerant System for Virtual Clusters"*

Journal Articles

TPAMI 2025: Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, **Minjia Zhang**, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, Shuaiwen Leon Song, *"RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model"*, in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025

Performance Evaluation 2025: Syed Zawad, Xiaolong Ma, Jun Yi, Cheng Li, **Minjia Zhang**, Lei Yang, Feng Yan, Yuxiong He, *"FedCust: Offloading Hyperparameter Customization for Federated Learning"*, in *Performance Evaluation: An International Journal*, 2025

IEEE DEB 2023: **Minjia Zhang**, Jie Ren, Zhen Peng, Ruoming Jin, Dong Li, Bin Ren, *"Exploiting Modern Hardware Architectures for High-Dimensional Vector Search at Speed and Scale"*, in *IEEE Data Engineering Bulletin*, 2023

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

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

Workshop Papers

NeurIPS 2023 (AI4Science): Shuaiwen Song, Bonnie Kruft, **Minjia Zhang**, Conglong Li, Shiyang Chen, Chengming Zhang, Masahiro Tanaka, Xiaoxia Wu, Mohammed AlQuraishi, Gustaf Ahdritz, Christina Floristean, Rick Stevens, Venkatram Vishwanath, Arvind Ramanathan, Sam Foreman, Kyle Hippe, Prasanna Balaprakash, Yuxiong He, *"DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies"*, in the *NeurIPS 2023 Workshop on AI for Science (AI4Science)*

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

MLSys 2022 Workshop: Fuxun Yu, Yongbo Yu, Di Wang, **Minjia Zhang**, Longfei Shangguan, Tolga Soyata, Chenchen Liu, Xiang Chen, *"A Survey on Multi-Tenant DL Inference on GPU"*, in the *MLSys 2022 Workshop on Cloud Intelligence / AIOps*

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 (NVMW)*, San Diego, USA, 2021

NVMW 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 *12th Non-Volatile Memories Workshop (NVMW)*, San Diego, USA, 2021

WODET 2014: Man Cao, **Minjia Zhang**, 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 (WODET), March 2014

Patents

U.S. Patent 2019: **Minjia Zhang**, Yuxiong He, *"Multi-layer Semantic Search"*, U.S. Patent 20200311077

U.S. Patent 2018: **Minjia Zhang**, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, Yuxiong He, *"Graph Representations for Identifying a Next Word"*, U.S. Patent 20190377792

U.S. Patent 2018: **Minjia Zhang**, Samyam Rajbhandari, Wenhan Wang, Yuxiong He, *"Deep Learning Model Scheduling"*, U.S. Patent, 20190311245

Invited Talks and Presentations

Invited Talks

Feb 2025: Invited talk by Deming Chen on *"Towards Efficient and Scalable Systems for Training Large-Scale AI-based Scientific Models"* at the AMD-Xilinx Heterogeneous Compute Cluster (HACC) Seminar

Sep 2024: Invited talk on *"Efficient and Scalable Machine Learning Systems for Training Large-Scale DL Models on Parallel and Distributed Hardware"* at the Meta Monetization AI Speaker Series

Jun 2024: Invited talk on *"Towards Efficient System and Algorithm for Large-Scale Scientific Discovery"* at the European Trillion Parameter Consortium (TPC) Kickoff Workshop, Barcelona

Dec 2023: Invited panelist at the Efficient Natural Language and Speech Processing (ENLSP-III) Workshop

Jul 2022: Invited talk by Saurabh Tangri on *"Extreme Compression for Pre-trained Transformers Made Simple and Efficient"* at Intel AI Group

Apr 2022: Invited talk on *"DeepSpeed: The Library to Accelerate Training and Inference of DNN at Scale"* at the Efficient Large-Scale AI Workshop, MSR Project Green

2021: Invited talk on *"DL Inference and Training Optimization Towards Speed and Scale"* at Tsinghua AIR

Apr 2021: Invited keynote on *"DL Inference and Training Optimization Towards Speed and Scale"* at EMDC 2021

2020: Invited talk on *"DL Inference Optimization Towards Speed & Scale"* at the ICT Young Scholars' Forum, Beijing, China

Dec 2019: Invited talk on *"TVM@Microsoft"* at the TVM and Deep Learning Compilation Conference, Seattle, WA, USA

Mar 2018: Invited talk on *"DeepCPU: Deep Learning Serving Optimizations on CPUs"* at the Deep Learning Workshop, Microsoft TechFest, Redmond, WA, USA

Feb 2018: Invited talk on *"DeepCPU: Deep Learning Serving Optimizations on CPUs"* at the Microsoft Research Talk Series, Redmond, WA, USA

Dec 2017: Invited talk on *"DeepCPU: Deep Learning Serving Optimizations on CPUs"* at Microsoft MLADS, Redmond, WA, USA

Presentations

IIDAI 2025: Presented a lighting talk at IIDAI Annual Meeting on long context extension of hybrid models

NeurIPS 2022: Presented work on extreme model compression

AAAI 2022: Presented work on adversarial data augmentation for knowledge distillation

WSDM 2022: Presented work on graph sampling and pruning for nearest neighbor search

IPDPS 2021: Presented *"DUET: Compiler-Aware Subgraph Scheduling for Tensor Programs on a Coupled CPU-GPU Architecture"*

ICLR 2021: Presented *"DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation"*

NeurIPS 2020: Presented *"Accelerating Training of Transformer-Based Language Models with Progressive Layer Dropping"*

NeurIPS 2020: Presented *"AdaTune: Adaptive Tensor Program Compilation Made Efficient"*

CIKM 2019: Presented *"GRIP: Multi-Store Capacity-Optimized High-Performance Nearest Neighbor Search for Vector Search Engine"*, Beijing, China

USENIX OpML 2019: Presented *"Accelerating Large Scale Deep Learning Inference through DeepCPU at Microsoft"*, Santa Clara, CA, USA

USENIX ATC 2018: Presented *"DeepCPU: Serving RNN-based Deep Learning Models 10x Faster"*, Boston, MA, USA

OOPSLA 2015: Presented work on detecting and tolerating region conflicts at ACM Student Research Competition, Pittsburgh, PA, USA

PPoPP 2015: Presented work on low-overhead and scalable software transactional memory at the 20th ACM SIGPLAN PPoPP, San Francisco, CA, USA

PLDI 2013: Presented work on efficient and strongly atomic STM at ACM Student Research Competition, Seattle, WA, USA

Teaching

Courses

2025 Spring: CS 498 Machine Learning Systems, UIUC. Designed and offered the first dedicated MLSys course at UIUC; topics included LLM distributed training, inference, and compresion algorithms.

2024 Fall: CS 598 AIE – AI Efficiency: System & Algorithms, UIUC

2024 Spring: CS 598 AIE – AI Efficiency: System & Algorithms, UIUC

Guest Lectures

Nov 2024: Invited guest lecture by Arvind Krishnamurthy on *"Mixture-of-Experts in the Era of LLMs"* at the University of Washington

Oct 2022: Invested guest lecture by Ruoming Jin on *"New Algorithms for Approximate Nearest Neighbor Search Systems at Scale"* at Kent State University

Apr 2022: Invited guest lecture by Zhihao Jia on *"DeepSpeed: The Library to Accelerate Training and Inference of DNN at Scale"* at Carnegie Mellon University (CMU)

Apr 2022: Invited guest lecture by Myeongjae Jeon on *"DeepSpeed: The Library to Accelerate Training and Inference of DNN at Scale"* at Ulsan National Institute of Science and Technology (UNIST)

Graduate Committee Service

Doctoral Preliminary Exam Committee

Mar 2025: Zhenrui Yue (Ph.D., UIUC)

Apr 2025: Yinfang Chen (Ph.D., UIUC)

May 2025: Boyuan Tian (Ph.D., UIUC)

Field/Qualifying Exam Committees

Mar 2025: Haoyang Zhang (Ph.D., UIUC)

Mar 2025: Shreesha Gopalakrishna Bhat (Ph.D., UIUC)

Mar 2025: Shashwat Jaiswal (Ph.D., UIUC)

Feb 2025: Jiyu Hu (Ph.D., UIUC)

Feb 2025: Mingyue Tang (Ph.D., UIUC)

Oct 2024: Eashan Gupta (Ph.D., UIUC)

Sep 2024: Saif Ur Rahman (Ph.D., UIUC)

Thesis Committees

2025: Akshat Sharma (M.S., UIUC), *"Pushing the Limits of Long Context LLM Inference via KV Cache Compression"*

2025: Xiao Wang (M.S., UIUC), *"Enhancing the Verifiability of Large Language Model based Medical Question Answering Systems"*

2022: Soobee Lee (M.S., UNIST), *"Hierarchical Episodic Memory for Continual Learning"*

Advising and Mentoring

Chengming Zhang (Ph.D., Indiana University Bloomington). Long sequence support for scientific applications (published at PODC 2024 and NeurIPS AI4Science Workshop 2023). Co-advised with Leon Song. *Apr–Sep 2023*

Xinyue Ma (Ph.D., UNIST). *"Cost-effective On-device Continual Learning over Memory Hierarchy with Miro"* (MobiCom 2023). Co-advised with Myeongjae Jeon. *Apr 2022–Jun 2023*

Ziang Song (M.S., Johns Hopkins University). Spot-instance training. Co-advised with Zhihao Jia (CMU). *Jun 2022–June 2023*

Jiangfei Duan (Ph.D., Chinese University of Hong Kong). Spot-instance training. Co-advised with Zhihao Jia (CMU). *Jun 2022–June 2023*

Suyeon Jeong (M.S., UNIST). *"A Design Space Evaluation of Replay-based Continual Learning with Memory Hierarchy"*. Co-advised with Myeongjae Jeon. *Apr–Dec 2022*

Yongye Su (Ph.D., Purdue University). Serverless vector search (published at SIGMOD 2024). Co-advised with Jianguo Wang. *Jul 2022–Sep 2023*

Shuangyan Yang (Ph.D., UC Merced). Large-scale GNN training (published at ASPLOS 2023). Co-advised with Dong Li

Yucheng Lu (Ph.D., Cornell University). Communication-efficient DNN training with 0/1 Adam (ICML 2023). *Jun 2021–Feb 2022*

Soobee Lee (M.S., UNIST). Hierarchical episodic memory for continual learning (DAC 2022). Co-advised with Myeongjae Jeon. *Sep 2020–Nov 2021*

Connor Holmes (Ph.D., University of Colorado Boulder). DNN sparsification (NeurIPS 2021). *May 2021–May 2022*

Zhen Peng (Ph.D., William & Mary). Ultra-fast graph-based ANN search (PPoPP 2023). Co-advised with Bin Ren and Ruoming Jin. *Sep 2019–Jun 2021*

Hongyi Wang (Ph.D., University of Wisconsin Madison). Efficient DNN training. *Jun–Sep 2020*

Jie Ren (Ph.D., UC Merced). DL training/inference via heterogeneous memory (HPCA 2020, NeurIPS 2020, USENIX ATC 2021). *Jun–Sep 2020*

Connor Holmes (Ph.D., University of Colorado Boulder). Exploiting sparsity in DNN inference. *Jun–Sep 2020*

Dantong Zhu (Ph.D., Georgia Tech). Monotonic relative nearest neighbor graph for ANN search (arXiv). *Jan–Jun 2020*

Zehua Hu (M.S., Peking University). Graph partitioning of TVM Relay IR for heterogeneous DL serving (IPDPS 2021). *Jul 2019–Mar 2021*

Menghao Li (M.S., Peking University). Bayesian optimization for DL compiler auto-tuning (NeurIPS 2020, ICLR 2021). *Feb 2020–Mar 2021*

Conglong Li (Ph.D., Carnegie Mellon University). Learning-based early termination for ANN search (SIGMOD 2020). *May–Aug 2019*

Stephen Zhou (Ph.D., MIT). Automatic model optimization. *Jun–Aug 2018*

Professional Services

Chair

2026: Sponsorship Chair, The 32nd IEEE International Symposium on High Performance Computer Architecture (HPCA 2026)

2025: Area Chair, The 39th Annual Conference on Neural Information Processing Systems (NeurIPS 2025)

2025: Session Chair (Large Language Model), The 30th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2025)

2025: Artifact Evaluation Chair, The Eighth Conference on Machine Learning and Systems (MLSys 2025)

2024: Area Chair, The 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024)

2019: Session Chair (Machine Learning), The 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2019)

2019: Publicity Co-Chair, ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2019)

Program Committee

2025: Program Committee, the Program Committee of the 4th Workshop on Practical Adoption Challenges of ML for Systems (PACMI '25)

2025: Program Committee, SIGCOMM

2025: Program Committee, USENIX Annual Technical Conference (USENIX ATC 2025)

2025: Program Committee, ACM SIGPLAN PPoPP 2025

2025: Program Committee, AAAI 2025

2023: Program Committee, IEEE IPDPS 2025

2024: Program Committee, USENIX ATC 2024

2024: Program Committee, MLSys 2024

2023: Program Committee, MLSys 2023

2023: Program Committee, ASPLOS 2023

2023: Program Committee (Industry/Applications Track), IEEE ICDE 2023

2023: Program Committee, IEEE IPDPS 2023

2021: Program Committee, IEEE IPDPS 2021

2020: Program Committee, IEEE IPDPS 2020

2019: Program Committee, IEEE IPDPS 2019

2018: Program Committee, IEEE IPDPS 2018

2018: Shadow Program Committee, ASPLOS 2018

2017: Artifact Evaluation Committee, PLDI 2017

2015: Artifact Evaluation Committee, SPLASH 2015

2015: Artifact Evaluation Committee, PLDI 2015

Conference Reviewer

2025: Reviewer for ICLR, CVPR, ICCV, AAAI

2024: Reviewer for VLDB, ECCV, ICML, MLSys, CVPR, AAAI, ICLR.

2023: Reviewer for NeurIPS, ECAI, ICCV, CVPR, ICLR, AAAI-23.

2022: Reviewer for NeurIPS, ECCV, ICML, USENIX ATC (external), ML Reproducibility Challenge (RC), CVPR, ICLR, AAAI.

2021: Reviewer for NeurIPS, ICML, ICCV, CVPR, ICLR, AAAI, ASPLOS.

2020: Reviewer for NeurIPS, ICLR.

2019: Reviewer for NeurIPS, NeurIPS Reproducibility Challenge, PLDI, ASPLOS.

2018: Subreviewer for Middleware, IEEE ICAC, IEEE CLOUD.

2017: Subreviewer for IEEE HiPC, IEEE ICAC.

2015: Subreviewer for WTTM.

Journal Reviewer

2024: ACM Transactions on Database Systems (TODS)

2023: Transactions on Machine Learning Research (TMLR)
2022: Transactions on Machine Learning Research (TMLR)
2020: IEEE Access
2020: Journal of Systems and Software
2019: IEEE Transactions on Cloud Computing
2019: ACM Transactions on Privacy and Security
2018: Concurrency and Computation: Practice and Experience
2017: Journal of Computer Science
2017: Concurrency and Computation: Practice and Experience

University/Industry Services

2025: Graduate Admission Committee, Department of Computer Science, UIUC
2024: Graduate Admission Committee, Department of Computer Science, UIUC
2022–2024: Sub-Committee Member, Microsoft E+D Research Council

Selected Press Coverage

Feb 2025: **UIUC News**, *CS professor Minjia Zhang receives NSF Career Award*.

Sep 2023: **Microsoft Research Blog**, *Announcing the DeepSpeed4Science Initiative: Enabling large-scale scientific discovery through sophisticated AI system technologies*.

Jul 2022: **TheSequence**, *A Model Compression Library You Need to Know About*.

Jul 2022: **Microsoft Research Blog**, *DeepSpeed Compression: A composable library for extreme compression and zero-cost quantization*.

Jan 2022: **Microsoft Research Blog**, *DeepSpeed: Advancing MoE inference and training to power next-generation AI scale*.

Aug 2021: **Microsoft Research Blog**, *DeepSpeed powers 8x larger MoE model training with high performance*.

May 2021: **Microsoft Research Blog**, *DeepSpeed: Accelerating large-scale model inference and training via system optimizations and compression*.

Jan 2021: **Towards Data Science**, *Microsoft ZeRO-Offload: Democratizing Billion-Scale Model Training*.

Jan 2021: **Medium**, *ZeRO-Offload: Training Multi-Billion Parameter Models on a Single GPU*.

Sep 2020: **The Batch**, *Toward 1 Trillion Parameters*.

Sep 2020: **Analytics India Magazine**, *Microsoft Releases Latest Version Of DeepSpeed, Its Python Library For Deep Learning Optimisation*.

Sep 2020: **siliconANGLE**, *Microsoft AI tool enables "extremely large" models with a trillion parameters*.

Sep 2020: **Microsoft Research Blog**, *DeepSpeed: Extreme-scale model training for everyone*.

Sep 2020: **VentureBeat**, *Microsoft's updated DeepSpeed can train trillion-parameter AI models with fewer GPUs*.

May 2020: **DeepSpeed.ai**, *Microsoft DeepSpeed achieves the fastest BERT training time*.

May 2020: **Microsoft Research Blog**, *Research Collection: Tools and Data to Advance the State of the Art*.

May 2020: **Microsoft Research Blog**, *ZeRO-2 & DeepSpeed: Shattering barriers of deep learning speed & scale*.

Feb 2020: **Microsoft Research Blog**, *ZeRO & DeepSpeed: New system optimizations enable training models with over 100 billion parameters*.

Feb 2020: **WinBuzzer**, *Microsoft DeepSpeed with Zero Can Train 100 Billion Parameter AI Models*.

Feb 2020: **MSPoweruser**, *Meet Microsoft DeepSpeed, a new deep learning library that can train massive 100-billion-parameter models*.

Feb 2020: **VentureBeat**, *Microsoft trains world's largest Transformer language model*.

Feb 2020: **InfoWorld**, *Microsoft speeds up PyTorch with DeepSpeed*.