

Yuke Wang

Phone: (+1) 805-259-9421

Email: yw236@rice.edu

yukew@amazon.com

yuke.wang96@gmail.com

Homepage: <https://wang-yuke.com>

[\[Google Scholar\]](#)[\[Github\]](#)[\[Linkedin\]](#)

EDUCATION

- | | |
|-------------|--|
| 2018 – 2024 | Ph.D. in Computer Science University of California, Santa Barbara, USA Advisor: Dr. Yufei Ding |
| 2014 – 2018 | B.E. in Software Engineering University of Electronic Science and Technology of China, China Advisor: Dr. Yu Tang |

EMPLOYMENT

- | | |
|-------------|---|
| Fall 2025 | Assistant Professor in Computer Science Rice University, USA. |
| 2024 - 2025 | Postdoc Scientist Amazon AWS AI, USA. Supervisor: Xiang Song |
| Summer 2023 | Research Intern Microsoft Research, USA. Supervisor: Saeed Maleki |
| Summer 2022 | Research Intern NVIDIA Research, USA. Supervisor: Michael Garland |
| Summer 2021 | High-Performance Engineering Intern NVIDIA, USA. Supervisor: Mehrzaad Samadi |
| Summer 2020 | Research Intern Alibaba DAMO Academy, USA. Supervisor: Yuan Xie |

AREAS OF RESEARCH

Yuke's research interests include **Deep-Learning (DL) Systems**, and **GPU-based Parallel and Distributed Computing**. His Ph.D. research spans deep neural networks (DNNs), graph neural networks (GNNs), and deep reinforcement learning (DRL) and their system-level optimization and acceleration on GPUs. The ultimate goal of Yuke's research is to facilitate *efficient, scalable, and secure* deep learning in the future.

- *Efficient* DL: **GNNAdvisor** [OSDI'21], **QGTC** [PPoPP'22], **TC-GNN** [ATC'23].
- *Scalable* DL: **MGG** [OSDI'23], **El-Rec** [SC'22], **RAP** [ASPLOS'24].
- *Secure* DL: **ZENO** [ASPLOS'24], **Faith** [ATC'22], **UAG** [AAAI'21].

PUBLICATIONS

Selected

| | |
|------------------|---|
| OSDI'23 | Yuke Wang, Boyuan Feng, Zheng Wang, Tong Geng, Ang Li, Kevin Barker, Yufei Ding, " MGG: Accelerating Graph Neural Networks with Fine-grained intra-kernel Communication-Computation Pipelining on Multi-GPU Platforms ", the USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2023. |
| USENIX ATC'23 | Yuke Wang, Boyuan Feng, Zheng Wang, Guyue Huang, Yufei Ding, " TC-GNN: Bridging Sparse GNN Computation and Dense Tensor Cores on GPUs ", the USENIX Annual Technical Conference (ATC), 2023. |
| PPoPP'22 | Yuke Wang, Boyuan Feng, Yufei Ding, " QGTC: Accelerating Quantized Graph Neural Networks via GPU Tensor Core ", the ACM SIGPLAN Symposium on Principles & Practice of Parallel Programming (PPoPP), 2022. |
| OSDI'21 | Yuke Wang, Boyuan Feng, Gushu Li, Shuangchen Li, Lei Deng, Yuan Xie, Yufei Ding, " GNNAdvisor: An Adaptive and Efficient Runtime System for GNN Acceleration on GPUs ", the USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2021 . |
| SC'21 | Boyuan Feng*, Yuke Wang*, Tong Geng, Ang Li, Yufei Ding, " APNN-TC: Accelerating Arbitrary-Precision Neural Networks on Tensor Cores ", the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2021. (* : equal contribution) |
| CCGrid'21 | Yuke Wang, Boyuan Feng, Gushu Li, Georgios Tzimpragos, Lei Deng, Yuan Xie, Yufei Ding, " TiAcc: Triangle-inequality based Hardware Accelerator for K-means on FPGAs ", the IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 2021. |
| IPDPS'21 | Yuke Wang, Boyuan Feng, Yufei Ding, " DSXplore: Optimizing Convolutional Neural Networks via Sliding-Channel Convolutions ", the IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2021. |
| TCAD'21 | Yuke Wang, Boyuan Feng, Gushu Li, Lei Deng, Yuan Xie, Yufei Ding, " STPAcc: Structural TI-based Pruning for Accelerating Distance-related Algorithms on CPU-FPGA Platforms ", IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. |

- CIKM'21 Yuke Wang, Boyuan Feng, Xueqiao Peng, Yufei Ding, "An Efficient Quantitative Approach for Optimizing Convolutional Neural Network", The ACM Conference on Information and Knowledge Management (CIKM), 2021 .
- ICTAI'20 Boyuan Feng*, Yuke Wang*, Xu Li, Shu Yang, Xueqiao Peng, Yufei Ding, "SGQuant: Squeezing the Last Bit on Graph Neural Networks with Specialized Quantization", the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2020. (*: **equal contribution**).

Other

- USENIX
ATC'24 Zheng Wang, Yuke Wang, Boyuan Feng, Guyue Huang, Dheevatsa Mudigere, Bharath Muthiah, Ang Li, Yufei Ding. "OPER: Optimality-Guided Embedding Table Parallelization for Large-scale Recommendation Model.", the USENIX Annual Technical Conference (ATC), 2024.
- ASPLOS'24 Zheng Wang, Yuke Wang, Jiaqi Deng, Da Zheng, Ang Li, Yufei Ding. "RAP: Resource-aware Automated GPU Sharing for Multi-GPU Recommendation Model Training and Input Preprocessing.", ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024.
- ASPLOS'24 Boyuan Feng, Zheng Wang, Yuke Wang, Shu Yang, Yufei Ding. "ZENO: A Type-based Optimization Framework for Zero Knowledge Neural Network Inference", ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024.
- ISCA'23 Hezi Zhang, Anbang Wu, Yuke Wang, Gushu Li, Hassan Shapourian, Alireza Shabani, Yufei Ding. "A Compilation Framework for Photonic One-Way Quantum Computation. ", International Symposium on Computer Architecture, 2023.
- MLSys'23 Guyue Huang, Yang Bai, Liu Liu, Yuke Wang, Bei Yu, Yufei Ding, Yuan Xie. "ALCOP: Automatic Load-Compute Pipelining in Deep Learning Compiler for AI-GPUs.", Sixth Conference on Machine Learning and Systems, 2023.
- USENIX
ATC'22 Boyuan Feng, Tianqi Tang, Yuke Wang, Zhaodong Chen, Zheng Wang, Shu Yang, Yuan Xie, Yufei Ding. "Faith: An Efficient Framework for Transformer Verification on GPUs", the USENIX Annual Technical Conference (ATC), 2021.
- SC'22 Zheng Wang, Yuke Wang, Boyuan Feng, Dheevatsa Mudigere, Bharath Muthiah, Yufei Ding. "EL-Rec: Efficient Large-scale Recommendation Model Training via Tensor-Train Embedding Table", the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC), 2022.
- USENIX
ATC'21 Boyuan Feng, Yuke Wang, Gushu Li, Yuan Xie, Yufei Ding, "Palleon: A Runtime System for Efficient Video Processing toward Dynamic Class Skew", the USENIX Annual Technical Conference (ATC), 2021.
- PPoPP'21 Boyuan Feng, Yuke Wang, Guoyang Chen, Weifeng Zhang, Yuan Xie, Yufei Ding, "EGEMM-TC: Accelerating Scientific Computing on Tensor Cores with Extended Precision", the ACM SIGPLAN Symposium on Principles & Practice of Parallel Programming (PPoPP), 2020.

| | |
|------------------|--|
| USENIX ATC'21 | Boyuan Feng, <u>Yuke Wang</u> , Gushu Li, Yuan Xie, Yufei Ding, " <i>Palleon: A Runtime System for Efficient Video Processing toward Dynamic Class Skew</i> ", the USENIX Annual Technical Conference (ATC), 2021. |
| AAAI'21 | Boyuan Feng, <u>Yuke Wang</u> , Yufei Ding, " <i>UAG: Uncertainty-aware Attention Graph Neural Network for Defending Adversarial Attacks</i> ", the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), 2021. |
| ICASSP'21 | Boyuan Feng, <u>Yuke Wang</u> , Yufei Ding, " <i>SAGA: Sparse Adversarial Attack on EEG-based Brain Computer Interface</i> ", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. |
| ICML'20 | Liu Liu, Lei Deng, Zhaodong Chen, <u>Yuke Wang</u> , Shuangchen Li, Jingwei Zhang, Yihua Yang, Zhenyu Gu, <u>Xing Hu</u> , Yufei Ding, Yuan Xie, " <i>Boosting Deep Neural Network Efficiency with Dual-Module Inference</i> ", the International Conference on Machine Learning (ICML), 2020. |

PROFESSIONAL SERVICE

| | |
|-----------|--|
| [05/2024] | Transactions on Parallel and Distributed Systems Reviewer |
| [05/2024] | APSys'24 Program Committee |
| [03/2023] | Journal of Supercomputing Paper Reviewer |
| [02/2023] | IEEE Transactions on Neural Networks and Learning Systems Reviewer |
| [02/2023] | PLDI'23 Artifact Evaluation Committee |
| [11/2022] | ECOOP'23 Artifact Evaluation Committee |
| [11/2022] | PPoPP'23 Artifact Evaluation Committee |
| [10/2022] | CGO'23 Artifact Evaluation Committee |
| [10/2022] | MLSys'23 External Review Committee |
| [10/2022] | IEEE Transactions on Computers Reviewer |
| [09/2022] | USENIX Security'23 Artifact Evaluation Committee |
| [09/2022] | ASPLOS'23 Artifact Evaluation Committee |
| [08/2022] | POPL'23 Artifact Evaluation Committee |
| [08/2022] | ACM Computing Survey Reviewer |
| [07/2022] | MICRO'22 Artifact Evaluation Committee |
| [06/2022] | SIGCOMM'22 Artifact Evaluation Committee |
| [04/2022] | ISSTA'22 Artifact Evaluation Committee |
| [04/2022] | OSDI'22 Artifact Evaluation Committee |
| [04/2022] | USENIX ATC'22 Artifact Evaluation Committee |
| [01/2022] | PLDI'22 Artifact Evaluation Committee |
| [01/2022] | EuroSys'22 Artifact Evaluation Committee |

- [11/2021] ASPLOS'22 Artifact Evaluation Committee
- [10/2021] SOSPP'21 Graduate Student Mentor
- [10/2021] Artificial Intelligence Review Paper Reviewer
- [10/2021] Journal of Supercomputing Paper Reviewer
- [08/2021] SOSPP'21 Artifact Evaluation Committee
- [07/2021] MICRO'21 Artifact Evaluation Committee
- [07/2021] SC'21 Artifact Evaluation Committee
- [10/2020] AAAI'21 Paper Reviewer Committee

AWARDS

- [05/2023] Graduate Division Dissertation Fellowship of UCSB
- [07/2022] 2022 USENIX Student Travel Grant for OSDI'22/USENIX ATC'22
- [06/2022] 2021-2022 Graduate Student External Award in CS Department of UCSB
- [11/2021] 2022-2023 NVIDIA Graduate Fellowship (Top 10 out of global applicants)
- [10/2021] 2021 ACM PACT Student Research Competition (First Prize Winner)
- [09/2021] 2021 SIGIR Student Travel Grant
- [06/2021] 2020-2021 Outstanding Publication Award in CS Department of UCSB
- [06/2020] 2020 Summer GSR recipient in CS Department of UCSB
- [06/2019] 2019 Summer GSR recipient in CS Department of UCSB
- [10/2017] Outstanding Graduates Award of UESTC
- [10/2017] First-class People's Scholarship (2/20 in the Elite Program)
- [04/2017] Interdisciplinary Contest In Modeling (ICM) [Honorable Mention]
- [04/2017] Suzhou Industrial Zone Scholarship (2/20 in the Elite Program)
- [10/2016] International Software Testing Qualifications Board (Certified Tester)
- [04/2016] First-class People's Scholarship (4/116)

TEACHING EXPERIENCE

- [09/2019] Teaching Assistant of CS160 (Translation of Programming Languages)
- [07/2019] Teaching Assistant of CS8 (Python Programming Language)
- [01/2019] Teaching Assistant of CS16 (C++ Programming Language)

OPENSOURCE PROJECT

| | |
|------------|---|
| MGG | Accelerating Graph Neural Networks with Fine-grained intra-kernel Communication-Computation Pipelining on Multi-GPU Platforms. https://github.com/YukeWang96/MGG_OSIDI23.git |
| TC-GNN | Bridging Sparse GNN Computation and Dense Tensor Cores on GPUs. https://github.com/YukeWang96/TC-GNN_ATC23.git |
| QGTC | Accelerating Quantized GNN via GPU Tensor Core. https://github.com/YukeWang96/QGTC_PPoPP22.git |
| GNNAdvisor | An Adaptive and Efficient Runtime System for GNN Acceleration on GPUs. https://github.com/YukeWang96/GNNAdvisor_OSIDI21.git |
| APNN-TC | Arbitrary Precision Neural Networks on Ampere GPU Tensor Cores. https://github.com/YukeWang96/APNN-TC_SC21.git |
| DSXplore | Convolutional Neural Networks via Sliding-Channel Convolutions. https://github.com/YukeWang96/DSXplore_IPDPS21.git |

STUDENT MENTORING

| | |
|---------------|--|
| Xiaoya Zhou | Accelerating the Large Language Model through Systemic Optimizations. (Undergrad at UCSB) [04/2023-Now] |
| Anshuman Dash | Automating the Optimization Flow of Graph Neural Networks via Dynamic Compilation. (Undergrad at UCSB) [09/2022-12/2022] |
| Qijun Zhang | Optimizing the Computation Efficiency of the Large-Scale Deep Learning via Holistic System Design. (Now as Ph.D. at HKUST) [06/2022-09/2022] |
| Xueqiao Peng | Optimizing Convolutional Neural Network with Quantitative Approach. (published at CIKM'21) (Now as Ph.D. at Ohio State University) [06/2020-09/2020] |

TALKS

| | |
|-----------|--|
| [10/2023] | Guest Lecture at the University of Rochester ECE403, hosted by Tong Geng. |
| [07/2023] | Invited Talk on Graph Learning Acceleration at CUHK and CityUHK, hosted by Hong Xu and Qiang Su. |
| [11/2022] | Gesture Lecture at NCSU CS591, hosted by Xipeng Shen. |
| [11/2022] | Technical Talk at AWS AI at Santa Clara, hosted by Yida Wang. |
| [10/2022] | SAMPLE Talk at the University of Washington, hosted by Zihao Ye. |
| [04/2022] | NVIDIA GTC'22. |