Hongkuan Zhou

📞 +1-(323)-352-4237 | ✉ tedzhouhk@gmail.com | 🜚 hongkuan
🏠 tedzhouhk.com | in hongkuan

EDUCATION

University of Michigan-Shanghai Jiao Tong University Joint Institute Shanghai, China
B.S. in Electrical Computer Engineering *Aug. 2013 – May 2017*

University of Southern California Los Angeles, US
M.S. in Computer Engineering   GPA: 4.00/4.00 *Aug. 2017 – May 2019*

University of Southern California Los Angeles, US
Ph.D. in Computer Engineering *Aug. 2019 – May 2024*
Advisor: Professor Viktor K. Prasanna   Current GPA: 4.00/4.00 *(estimated)*

RESEARCH AND WORKING EXPERIENCE

Research Assistant Dec. 2017 – Present
FPGA/Parallel Computing Lab, USC *Los Angeles, US*

- Research interests: software-hardware co-design for acceleration and applications of large-scale static and temporal Graph Neural Networks (GNNs).
- Developed scalable static and dynamic GNN training solutions on single-GPU, single-machine multiple-GPU, and distributed GPU clusters.
- Developed static and dynamic GNN inference solutions for real-time and large-scale GNN applications on CPU, GPU, and FGPAs.
- Developed GNN-based solutions for knowledge graph reasoning and wireless network optimization.

Applied Scientist Intern May 2021 – Aug. 2021
Amazon Web Service *Palo Alto, US*

- Designed and implemented TGL – a general single-machine multiple-GPU training framework for Temporal GNNs training on large-scale dynamic graphs.
- TGL achieved more than 10× speedup compared with state-of-the-art systems and was deployed to fraud detection system in production.

Applied Scientist Intern May 2022 – Aug. 2022
Amazon Web Service *Palo Alto, US*

- Designed and implemented DistTGL – a scalable training solution to memory-based Temporal GNNs on distributed GPU clusters.
- DistTGL achieved 2-8× speedup in training time and siginificantly better accuracy compared with single-machine multiple-GPU approach.

PUBLICATIONS

GNN Acceleration

- Third Author, "An Efficient Distributed Graph Engine for Deep Learning on Graphs", *(under double-blind review)*, 2023
- **First** Author, "DistTGL: Distributed Memory-based Temporal Graph Neural Network Training", *International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, 2023
- **First** Author, "TGL: A General Framework for Temporal GNN Training on Billion-Scale Graphs", *International Conference on Very Large Data Bases (VLDB)*, 2022 [Paper] [Code]
- **Co-First** Author, "Model-Architecture Co-Design for High Performance Temporal GNN Inference on FPGA", *International Parallel and Distributed Processing Symposium (IPDPS)*, 2022 [Paper] [Code]
- **First** Author, "Accelerating Large Scale Real-Time GNN Inference using Channel Pruning", *International Conference on Very Large Data Bases (VLDB)*, 2021 [Paper] [Code]
- **Co-First** Author, "Accurate, Efficient and Scalable Training of Graph Neural Networks", *Journal of Parallel and Distributed Computing (JPDC)*, 2020 [Paper] [Code]
- **Co-First** Author, "GraphSAINT: Graph Sampling Based Inductive Learning Method", *International Conference on Learning Representations (ICLR)*, 2020 [Paper] [Code]

- **Co-First** Author "Accurate, Efficient and Scalable Graph Embedding", *International Parallel and Distributed Processing Symposium (IPDPS)*, 2019 [Paper] [Code]

GNN Applications
- **First** Author, "HTNet: Dynamic WLAN Performance Prediction using Heterogenous Temporal GNN", *International Conference on Computer Communications (INFOCOM)*, 2023 [Paper] [Code]
- Second Author, "Throughput Optimization in Heterogeneous MIMO Networks: a GNN-based Approach", *International Workshop on Graph Neural Networking (GNNet)*, 2022 [Paper]
- **First** Author, "SeDyT: A General Framework for Multi-Step Event Forecasting via Sequence Modeling on Dynamic Entity Embeddings", *International Conference on Information and Knowledge Management (CIKM)*, 2021 [Paper] [Code]

Miscellaneous
- **First** Author "Design and Implementation of Knowledge Base for Runtime Management of Software Defined Hardware", *High Performance Extreme Computing Conference (HPEC)*, 2019 **(Best Student Paper Nominee)** [Paper]

SERVICE

Website Chair: HiPC 2023
Reviewer: FCCM, PAD, ASAP, IPDPS, BigData, HPEC, ReConfig, TPDS, SBAC-PAD, IJCAI, CCGrid, NeurIPS, Supercomputing, TKDE, IPM, NeuroComputing

TECHNICAL SKILLS

Computer Languages: C/C++, Python, Matlab, Verilog, SQL
Developer Tools: PyTorch, Tensorflow, DGL, Bash, Git, LaTeX
Language: Mandarin Chinese (native), English (proficient), Japanese (elementary)