Peizhen Guo

Department of Computer Science,
Yale University, New Haven, CT 06511
Tel: 203-645-1204
E-mail: peizhen.guo@yale.edu
Website: http://cs.yale.edu/homes/guo-peizhen/

EDUCATION

Yale University New Haven, CT 08/2015 – 05/2021(expected)
PhD in Computer Science
Supervisor: Wenjun Hu, Assistant Professor

Tsinghua University Beijing, China 09/2011 – 07/2015
BS, Major in Electronic Engineering
GPA: 90.42/100

Hong Kong University of Science and Technology Hong Kong, China 02/2014 – 05/2014
Non-degree Exchange Student
GPA: 3.914/4.0

FIELDS OF INTEREST

- **Mobile and Edge Computing, Distributed Systems, and System Support for Deep Learning**.

PUBLICATIONS

- **[*NSDI'21*]** **Peizhen Guo**, Bo Hu, Wenjun Hu. **Mistify: Automating DNN Model Porting for On-Device Inference at the Edge.**
- **[*SoCC'19*]** Rui Li, **Peizhen Guo**, Bo Hu, Wenjun Hu. **Libra and the Art of Task Sizing in Big-Data Analytic Systems.**
- **[*MobiCom'18*]** **Peizhen Guo**, Bo Hu, Rui Li, Wenjun Hu. **FoggyCache: Cross-Device Approximate Computation Reuse.**
- **[*ASPLOS'18*]** **Peizhen Guo**, Wenjun Hu. **Potluck: Cross-Application Approximate Deduplication for Computation-Intensive Mobile Applications.**
- **[*Multimedia Systems'15*]** Liang Zheng, Shengjin Wang, **Peizhen Guo**, Hanyue Liang, Qi Tian. **Tensor Index for Large Scale Image Retrieval.**

WORK EXPERIENCE

Facebook, Inc. 06/2020-09/2020
PhD SDE Intern, Presto team in backend data infrastructure Boston, MA
- Designed and implemented runtime adaptive query optimization framework for the Presto query engine.
- Added tracking logic for broadcast memory usage and enforced broadcast memory limit on all Presto queries.

Facebook, Inc. 06/2018-09/2018
PhD SDE Intern, Spark team in backend infrastructure Menlo Park, CA
- Proposed and deployed a data skew detection module in core Spark.
- Designed and implemented the skew join operator in Spark SQL.
- Enabled cost-based optimization for the broadcast join operator in Spark SQL.

VMware, Inc. 06/2017-09/2017
R&D Intern, Data center converged infrastructure Palo Alto, CA
- Explored opportunities and open problems towards power efficient virtualized infrastructure.
- Developed a mechanism for collaborative tuning of VM and application internals for optimal power efficiency.

RESEARCH EXPERIENCE

Systems & Networking

Department of CS, Yale University. Advisor: Prof. Wenjun Hu 08/2015-present

Automating DNN Model Porting for Ubiquitous On-Device Deep Learning Inference

- Addressed scalability challenge of porting pre-trained DNN models to ubiquitous deployment endpoints.
- Designed abstractions and algorithms to enable automatic model tailoring at scale, towards a diverse range of resource and performance specifications with minimal manual efforts.
- Implemented a model porting system to perform as an intermediary that decoupled DNN design and deployment.

DNN Model Semantic Understanding, Indexing, and Query System (under submission)

- Addressed the challenge that existing DNN model repositories lack the ability to understand DNN internals and provide fine-grained query support.
- Designed algorithms to measure semantic correlation of DNN models with provable guarantee.
- Built a query system that indexed DNN models according to their semantic correlation and resource profile, which further supported model query using high-level user preferences.

Approximate Computation Reuse as a New Deduplication Paradigm

- Addressed approximate computation reuse as an overlooked opportunity to exploit the error-tolerance nature of the emerging computation-intensive workloads, such as deep learning and graphic rendering.
- Designed algorithms for approximate caching of high-dimensional data, functional equivalence analysis of program DAGs, automatic DAG rewriting, and secure computation reuse with bounded error rate.
- Implemented approximate computation reuse as a decentralized service for edge computing scenario, which is deployed on Android and Ubuntu Linux desktop OS, showing over 10x performance enhancement.

Auto-Tuning Fine-Grained Parallelism in Data-Analytics Systems

- Demonstrated that job partitioning was non-trivial and hugely influenced the performance of the system.
- Proposed a stochastic control based job partitioning algorithm that adaptively matched task size with machine capability in real-time.
- Deployed the algorithm in Spark and HDFS framework and achieved up to 3x performance enhancement.

Department of EE, Tsinghua University. Advisor: Prof. Jun Bi 09/2014-02/2015

Fertile, Flexible, and Future-proof Enterprise Network Architecture Design

- Redesigned the data plane of OpenVSwitch to enable fertile flow entries with limited hardware resources
- Extended function of traditional OpenVSwitch for flexible stateful forwarding
- Realized the system as an extensible framework open to future protocols

Department of CS, USC. Advisor: Prof. Minlan Yu 07/2014-09/2014

Software-based flexible traffic measurement for cloud-scale attack detection

- Implemented a dynamic packet sampling mechanism without hardware change
- Developed a mathematic model to describe global relative measurement error vs. measurement cost
- Deployed per-flow sampling module for accurate Heavy-Hitter detection

Department of CSE, UCLA. Advisor: Prof. Lixia Zhang 05/2014-08/2014

Named Data Network Signature Logging System

- Developed a mechanism for long-lived data verification when the public key was outdated
- Modified Merkle Hash Tree tamper-evident data structure with regard to NDN property
- Designed and implemented a P2P protocol for synchronization among loggers and among loggers' auditors

Department of ECE, HKUST. Advisor: Prof. Bo Li 02/2014-06/2014

Software-defined task scheduling in datacenter network

- Proposed a model to characterize the traffic pattern in datacenter network by real-time sampled flow statistics
- Presented a heuristic flow scheduling scheme which adjusts scheduling behavior based on current traffic patterns

Large-scale Image Retrieval

Department of EE, Tsinghua University Advisor: Prof. S. Wang 09/2013-12/2013

Tensor-indexed large-scale image retrieval

- Proposed a tensor-index retrieving framework to enhance performance
- Added local compatibility condition to joint image features
- Implemented an image retrieval system that outperformed state-of-art performance with less time

TEACHING/MENTORING EXPERIENCE

As Teaching Assistant

• *Computational Tools for Data Science*,	Fall 2016
• *Building Distributed Systems*,	Fall 2017
• *Great Ideas in Computer Science*,	Spring 2018

Mentoring

• *Sanat Khurana*,	EECS'19
• *Julia McClellan*,	EE'21

Guest lectures

• *ECE590 / COMPSCI590 – Edge Computing,*	Fall 2018 at Duke University
• *ECE590 / COMPSCI590 – Edge Computing,*	Spring 2020 at Duke University

TECHNICAL SKILLS

- Languages: C/C++, Python, Java, Scala, Go, Matlab, Assembly, Verilog, LaTeX, Bash
- Tools: TensorFlow, Spark, Presto, Hadoop, Akka

COMPETITIONS AND HONORS

National College Student Physics Competition

- Won second prize in non-physics-major group

The 13th Teamwork AI Programming Competition of Tsinghua University

- Rank top 16

Tsinghua University Scholarship

- For distinguished contribution in art and organization