LEI ZHANG

601 108th Ave NE, Bellevue, WA 98004
https://www.geraldleizhang.com & geraldleizhang@gmail.com

RESEARCH INTEREST

My research interests are broadly in distributed systems.

- **Reliability and observability** distributed tracing, PL for systems towards root cause analysis
- **Distributed caching** heterogeneous memory management, CDN caching, performance quantification
- **Systems for ML** reliability, observability, performance analysis, RDMA-based AI infra for LLM, Collective Communication Library

EMPLOYMENT

ByteDance Inc. Sep. 2023 – current
Research Scientist

Princeton University Jan. 2022 – Aug. 2023
Postdoc Research Associate

Facebook Inc. May 2018 – Aug. 2018
Ph.D. Software Engineer Intern

EDUCATION

Emory University Aug. 2018 – Dec 2021
Ph.D. Computer Science
- Advisor: Prof. Ymir Vigfusson
- Transferred from Georgia Tech as a post-qualified Ph.D. candidate

Georgia Institute of Technology Aug. 2015 – Aug. 2018
M.S. Computer Science
- Advisor: Prof. Karsten Schwan (deceased)

Tsinghua University Aug. 2011 – July 2015
B.E. Computer Science

PUBLICATIONS

Under review

Tracing Dependencies in Collective Communication Towards Reliable LLM Training Under review

Towards Bandwidth-adaptive Live Volumetric Video Conferencing Under review

A Lightweight Telemetry System with Service Tracing for Locating Network Slowdowns Under review

Automatic Instrumentation for Fine-grained Observability in Distributed Systems Under review

Papers

Minder: Faulty Machine Detection for Large-scale Distributed Model Training NSDI'25
*Yangtao Deng, Xiang Shi, Zhuo Jiang, Xingjian Zhang, **Lei Zhang**, Zhang Zhang, Bo Li, Zuquan Song, Hang Zhu, Gaohong Liu, Fuliang Li, Shuguang Wang, Haibin Lin, Jianxi Ye, Minlan Yu*

Latenseer: Causal Modeling of End-to-end Latency Distributions by Harnessing Distributed Tracing SOCC'23
*Yazhuo Zhang, Rebecca Isaacs, Yao Yue, Juncheng Yang, **Lei Zhang**, Ymir Vigfusson*

The Benefit of Hindsight: Tracing Edge-Cases in Distributed Systems NSDI'23
Lei Zhang, *Zhiqiang Xie, Vaastav Anand, Ymir Vigfusson, Jonathan Mace*

When is the Cache Warm? Manufacturing a Rule of Thumb HotCloud'20
Lei Zhang, *Juncheng Yang, Anna Blasiak, Mike McCall, Ymir Vigfusson*

Optimal Data Placement for Heterogeneous Cache, Memory, and Storage Systems SIGMETRICS'20
Lei Zhang, *Reza Karimi, Irfan Ahmad, Ymir Vigfusson* ***Best Student Paper***

Deceptive Secret Sharing DSN'18
Lei Zhang, *Douglas M. Blough*

Systematic Data Placement Optimizationin in Multi-Cloud Storage for Complex Requirements IEEE ToC'16
Maomeng Su, ***Lei Zhang***, *Yongwei Wu, Kang Chen, Keqin Li*

Thesis

Measurement and Analysis Methods of Performance Problems in Distributed Systems Doctoral Thesis

Awards

Kenneth C. Sevcik Outstanding Student Paper Award SIGMETRICS'20

Bronze Medal 24th, 25th China Mathematical Olympiad

Professional Services

Program Committee USENIX ATC'25

Program Committee ACM EuroSys'25

External Program Committee ACM SIGMETRICS'23

Program Committee ACM SOCC'22, 23, 24

External Reviewer USENIX ATC'21, HotCloud'20, HotStorage'20

External Reviewer ACM Eurosys'19, SOCC'18, SRDS'18

Research Projects

Reliability and Observability

Tracing Edge-cases in Distributed Systems

Today's distributed tracing frameworks are limited to low sampling rate ($< 0.1\%$) and ill-equipped to troubleshoot rare edge-case requests. We propose retroactive sampling to trace 100% data and retrieve trace data only after problems are detected. We design and build Hindsight, a lightweight always-on tracing system to practically track edge-case system symptoms with low overhead. Hindsight is shown to provide nanosecond-level tracing overhead and is efficient to capture real-world system issues.

Automatic Instrumentation for Distributed Tracing

Today's distributed tracing systems are designed to track coarse-grained information in large-scale systems, which leaves the debugging and troubleshooting hardness to developers. We propose an automatic mechanism to leverage static analysis for automatic fine-grained tracing instrumentation, which can provide straightforward debugging information and ease today's troubleshooting process.

Distributed caching

Optimal Data Placement for Heterogeneous Memory Systems

In recent years, modern memory hardware (e.g. NVMe) and architectures (e.g. NUMA) drive distributed memory/cache systems less hierarchical, which changes the picture of large-scale in-memory system design. We proposed design principles to fit the modern memory scenarios to fully leverage the efficiency of those new techniques. We build CHOPT, a data-driven offline optimal placement algorithm. CHOPT exposes that state-of-the-art distributed cache can be further improved by up to 44.8%, and provides practical heuristics towards online cache algorithm design.

Understanding Distributed Cache Warmup

Cache warmup can significantly improve application performance, but understanding cache warmup process is hard. We provide a practical quantification of cache warmup process. We derive a rule-of-thumb formula to estimate cache warmup time. Our method can provide high accuracy on a wide range of real-world cache workloads.

Fine-grained Storage Utilization Management for Facebook's Video Cache *(intern project)*

I develop a utilization monitor of downstream storage usage for different customers and different video files. The monitor tolerates lags of relied tracing services through approximation. I develop a service to maintain priorities of fine-grained utilization and automatically make eviction decisions, to ensure efficient resource utilization and fairness.

Formalized Data Placement Optimization in Multi-cloud Storage

Optimizing cloud storage objectives at the application level is a challenging task, especially when it comes to fulfilling ad-hoc user configurations and considering erasure coding. We developed Triones, a systematic model that is based on erasure coding and formalized data placement optimization for multi-cloud storage configuration. Triones can efficiently optimize cloud application's performance and service quality, e.g. fault-tolerance, latency, and costs.

Systems for ML

Faulty Machine Detection for Large-scale Distributed Model Training

Large-scale distributed model training can involve thousands of machines and is highly susceptible to faults, which can halt training for hours and occur multiple times per day. To automate and accelerate fault detection, we built Minder, a system that identifies faulty machines by recognizing abnormal patterns in monitoring metrics before failures disrupt training. Deployed in production for over a year, Minder achieves fault detection within 3.6 seconds on average, with a precision of 0.904 and an F1-score of 0.893.

Tracing Dependencies in Collective Communication Towards Reliable LLM Training

Reliability is critical in LLM training, yet many issues—particularly those involving complex collective communications—remain hard to diagnose, leading to inefficiency and wasted resources. To address this, we present a lightweight distributed tracing and root cause analysis system that exposes internal communication states and dependencies previously hidden in collective communication libraries. Deployed in production for over six months, the system achieved a 100% detection rate for communication-related anomalies, identifying root causes within 20 seconds in 60% of cases.

Others

Live Volumetric Video Streaming

Live volumetric video streaming is not available yet, because it's challenging to deal with high data volume and computational complexity, to meet the requirement of low latency and bandwidth. We propose a new design to directly transmit adaptive selected point clouds to client side and prove such

mechanism can ease the overall latency and computation effort. We build the first bandwidth-adaptive live volumetric video streaming system which is practical on real-world hardware and network scenarios. We show that our system can provide high streaming quality compared with state-of-the-art (non-live) volumetric video streaming systems.

Data-driven Malware Detection

I develop a systematic data-driven approach to attribute malware behaviors to specific actors, tools, and intents. I build a system to parse and clean raw system and network logs from a real malware execution dataset, identify unique system patterns to define malware's distinct features. I design LSM-Tree-like data structure to effectively manage large-scale malware execution traces effectively, and a clustering algorithm for identifying related malware families.

Deceptive Secret Sharing

Secret sharing is challenging for sensitive data where even subset of data pieces is confidential. We propose deceptive secret sharing to combine deception into secret sharing schemes and prevent attacks such as insider attacks. We build a distributed storage prototype for notable enhanced security with limited storage and computational overhead.

MENTORSHIP

Rajrup Ghosh (Ph.D., USC), Jingyuan Chen (Ph.D., Princeton), Yazhuo Zhang (Ph.D., Emory), Tao Zhou (Undergrad, Emory)