Frank Sifei Luan

312-804-2781 lsf@berkeley.edu franklsf

Education

University of California, Berkeley Berkeley, CA
Ph.D. in Computer Science 2020 – 2024
Research topics: AI systems, distributed data processing, cloud computing

The Univerity of Chicago Chicago, IL
B.S. in Computer Science and B.A. in Statistics 2013 – 2017

Experience

xAI Palo Alto, CA
Member of Technical Staff 2025 – Present
- Led a Data Scaling team of 12 engineers working across pre-training data, data infrastructure, batch inference, synthetic data generation, multi-agent systems, and agentic search.
- Led a project to build web search from scratch: crawling, indexing, ranking, evaluation, and agentic RL.
- Co-created Grokipedia, scaling the real-time editing agents to grow the encyclopedia to 6 million articles.
- Built a batch inference framework on Ray to support O(T)-token-scale LLM-based distributed data processing and synthetic data generation.
- Built an embedding search system to support analytics and data mining from O($100T$) pre-training tokens.

Anthropic San Francisco, CA
Member of Technical Staff 2024 – 2025
- Built the scaling infrastructure for LLM pre-training and post-training on GPU, TPU and Trainium clusters.

Wayo San Francisco, CA
Co-founder & CTO 2023 – 2024
- Created an AI-powered global product sourcing company and raised $2M in venture capital funding.
- Built and led a team of 10 engineers, designers and product managers to launch a marketplace with $1M+ GMV.

Anyscale San Francisco, CA
Software Engineering Intern 2021
- Integrated Exoshuffle, a research system, into Ray Data, an open-source data processing framework for ML.

Facebook Menlo Park, CA
Software Engineer 2017 – 2019
- Created ML-based software engineering tools for natural language code search, code recommendation, code completion, code synthesis, code review, and automatic bug fixing.
- Deployed these tools on codebases of 10B+ LOC, improving the productivity of 50K+ engineers.
- Published 3 academic papers and presented at top-tier academic conferences for software engineering.

Facebook Menlo Park, CA
Software Engineering Intern 2016
- Implemented several new features in the Facebook Pages SMB Platform, totaling 20K+ LOC.

SketchMe Menlo Park, CA
Co-founder & CTO 2014 – 2015
- Created a social app company and raised $1.5M in venture capital funding.
- Built a team of 5 engineers and designers to launch a mobile app with 10K+ users.

Honors & Awards

CloudSort World Record 2022
The Sort Benchmark Committee
A foundational big data benchmark. The most cost-efficient way to sort 100 TB of data a public cloud [3, 4].

IEEE Software Magazine Best Paper Award 2021
IEEE Computer Society
Awarded to the article "AI in Software Engineering at Facebook" [7].

ACM SIGPLAN Distinguished Paper Award 2019
The ACM Special Interest Group on Programming Languages
Awarded to the paper "Aroma: Code Recommendation via Structural Code Search" [10].

Champion of the Midwest Trading Competition 2016
The University of Chicago
Developed the best-performing automated trading algorithms for three securities markets.

ACM International Collegiate Programming Contest Regional Finalist 2014
Association for Computing Machinery
Mid-central USA Regional Contest Finalist.

Invited Talks

Dissertation Talk: An Extensible Architecture for Distributed Heterogeneous Processing 2024
University of California, Berkeley Berkeley, CA

Ray Data: Efficient Heterogeneous Execution with the Streaming Batch Model 2024
Paul G. Allen School of Computer Science & Engineering, University of Washington Seattle, WA

Exoshuffle: An Extensible Shuffle Architecture 2024
Meta Menlo Park, CA

The Ray Dataplane: History and the CloudSort World Record 2023
Ray Summit [video] San Francisco, CA

Exoshuffle: Large-Scale Shuffle at the Application Level 2022
Google Mountain View, CA

Using ML for Code Discovery at Facebook 2019
Curry On London (co-located with ECOOP) [video] London, UK

Using Machine Learning for Developer Productivity 2019
F8 Developer Conference [video] San Jose, CA

Publications

[1] Frank Sifei Luan. "An Extensible Architecture for Distributed Heterogeneous Processing". PhD thesis. EECS Department, University of California, Berkeley, Dec. 2024.

[2] Frank Sifei Luan, Ziming Mao, Ron Yifeng Wang, Charlotte Lin, Amog Kamsetty, Hao Chen, Cheng Su, Balaji Veeramani, Scott Lee, SangBin Cho, Eric Liang, Ion Stoica, and Stephanie Wang. *Ray Data: Efficient Heterogeneous Execution with the Streaming Batch Model*. Dec. 2024.

[3] Frank Sifei Luan, Stephanie Wang, Samyukta Yagati, Sean Kim, Kenneth Lien, Isaac Ong, Tony Hong, SangBin Cho, Eric Liang, and Ion Stoica. *Exoshuffle-CloudSort*. 2023. arXiv: `2301.03734 [cs.DC]`.

[4] Frank Sifei Luan, Stephanie Wang, Samyukta Yagati, Sean Kim, Kenneth Lien, Isaac Ong, Tony Hong, Sangbin Cho, Eric Liang, and Ion Stoica. "Exoshuffle: An Extensible Shuffle Architecture". In: *Proceedings of the ACM SIGCOMM 2023 Conference*. ACM SIGCOMM '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 564–577.

[5] Zongheng Yang, Zhanghao Wu, Michael Luo, Wei-Lin Chiang, Romil Bhardwaj, Woosuk Kwon, Siyuan Zhuang, Frank Sifei Luan, Gautam Mittal, Scott Shenker, and Ion Stoica. "SkyPilot: An Intercloud Broker for Sky Computing". In: *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA: USENIX Association, Apr. 2023, pp. 437–455.

[6] Zongheng Yang, Wei-Lin Chiang, Sifei Luan, Gautam Mittal, Michael Luo, and Ion Stoica. "Balsa: Learning a Query Optimizer Without Expert Demonstrations". In: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD '22. Philadelphia, PA, USA: Association for Computing Machinery, 2022, pp. 931–944.

[7] Johannes Bader, Sonia Seohyun Kim, Frank Sifei Luan, Satish Chandra, and Erik Meijer. "AI in Software Engineering at Facebook". In: *IEEE Software* 38.4 (2021), pp. 52–61.

[8] Stephanie Wang, Eric Liang, Edward Oakes, Ben Hindman, Frank Sifei Luan, Audrey Cheng, and Ion Stoica. "Ownership: A Distributed Futures System for Fine-Grained Tasks". In: *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, 2021, pp. 671–686.

[9] Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Xi Chen, and Ion Stoica. "NeuroCard: one cardinality estimator for all tables". In: *Proc. VLDB Endow.* 14.1 (2020), pp. 61–73.

[10] Sifei Luan, Di Yang, Celeste Barnaby, Koushik Sen, and Satish Chandra. "Aroma: code recommendation via structural code search". In: *Proc. ACM Program. Lang.* 3.OOPSLA (Oct. 2019).

[11] Saksham Sachdev, Hongyu Li, Sifei Luan, Seohyun Kim, Koushik Sen, and Satish Chandra. "Retrieval on source code: a neural code search". In: *Proceedings of the 2nd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*. MAPL 2018. Philadelphia, PA, USA: Association for Computing Machinery, 2018, pp. 31–41.

Academic Services

Reviewer 2024
Conference on Neural Information Processing Systems (NeurIPS)

Reviewer 2023
Conference on Neural Information Processing Systems (NeurIPS)

Artifact Evaluation Committee 2022
The European Conference on Computer Systems (EuroSys)

Artifact Evaluation Committee 2021
Symposium on Operating Systems Principles (SOSP)

Reviewer 2021
Conference on Machine Learning and Systems (MLSys)

Skills

Programming languages (industry experience):
C, C++, Java, JavaScript, Objective-C, OCaml, Python, Ruby, Rust, SQL, TypeScript

Frameworks (research/industry experience):

- *Machine learning:* CUDA, JAX, PyTorch, Ray, TensorFlow, XLA
- *Data processing:* Arrow, ClickHouse, Dask, Flink, Pandas, PostgreSQL, Spark
- *Infrastructure:* AWS, Azure, Docker, Google Cloud, Kubernetes, OCI, Terraform

Private Pilot Certificate, Instrument Rating
Federal Aviation Administration