Kahfi S. Zulkifli

Charlottesville, Virginia   kahfiz@virginia.edu   kahfizulkifli.github.io   in kahfizulkifli

Education

University of Virginia, USA 2024 — Present
Ph.D. in Computer Science

Bandung Institute of Technology (ITB), Indonesia 2019 — 2024
B.Eng. in Computer Science

Publications

Verifying Computational Graphs in Production-Grade Distributed Machine Learning Frameworks
Kahfi S. Zulkifli, Wenbo Qian, Shaowei Zhu, Yuan Zhou, Zhen Zhang, Chang Lou
under submission, manuscript ready upon request

Verifying Semantic Equivalence of Large Models with Equality Saturation
Kahfi S. Zulkifli*, Wenbo Qian*, Shaowei Zhu, Yuan Zhou, Zhen Zhang, Chang Lou *(*equal contribution)*
EuroMLSys Workshop (co-located with EuroSys'25)

Heimdall: Optimizing Storage I/O Admission with Extensive Machine Learning Pipeline
Daniar H. Kurniawan, Rani Ayu Putri, Peiran Qin, ***Kahfi S. Zulkifli***, Ray A. O. Sinurat, Janki Bhimani, Sandeep Madireddy, Achmad Imam Kistijantoro, Haryadi S. Gunawi
Twentieth European Conference on Computer Systems (EuroSys '25)

EVStore: Storage and Caching Capabilities for Scaling Embedding Tables in Deep Recommendation Systems
Daniar H. Kurniawan, Ruipu Wang, ***Kahfi S. Zulkifli***, Fandi Wiranata, John Bent, Ymir Vigfusson, Haryadi S. Gunawi
28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23)

Research Experience

Detecting silent errors in distributed machine learning models May 2024 — Present

- Collaborated with researchers and developers from **AWS**
- Designed semantic equivalence framework on verifying computational graphs based on egglog, an e-graph engine
- Detected 17 old bugs in popular machine learning frameworks and **reported 5 new real-world bugs in Amazon Neuronx** to developers
- Verified equivalence of Llama-3.1 and Mixtral-8x7B models under several minutes on a commodity machine

Reducing tail latency in solid state drives (SSDs) with machine learning Sep 2022 — May 2024

- Designed 16 machine learning models that has accuracy between 70%-90% with AutoML for predicting storage performance in FEMU, an SSD emulator
- Reduced the inference latency of machine learning models down to 10ns, 50x faster than existing models
- Modified Ceph, an existing object-storage library with machine learning models that show up to 40% improvement at p99 latency and deployed 10 clusters, each with 2 OSDs across 20 nodes in FEMU

Optimizing model inference latency of deep recommendation systems (DRS) Aug 2021 — Aug 2022

- Designed new caching algorithms based on groupability, a novel property that measures the probability of an embedding value present in an inference request, which increased perfect hit rate up to 30%
- Helped integrate EVStore system, which reduced average latency by 23% and p90 latency by 27%, increased the throughput by 4x at only 0.2% loss in accuracy

Industry Experience

National Land Agency Business Intelligence Dashboard (Webgis Indonesia) June 2023 — Dec 2023

- This project aims to analyze huge land property dataset by utilizing a business intelligence dashboard
- Maintained the Extract-Transform-Load (ETL) pipeline reaching hundreds of gigabytes from Apache Hive to the ELK Stack (Elastic Search, Logstash, and Kibana) and analyzing the data in real time with Superset
- Helped integrate the whole architecture, from data warehouse to the business intelligence dashboard, which helped 2 national government agencies to analyze huge amounts of data

National Geospatial Data Warehouse (Webgis Indonesia) June 2022 — Aug 2022

- Purpose of this project is to analyze real-time climate and earthquake data in Indonesia
- Maintained an Extract-Transform-Load (ETL) pipeline using Open Talend Studio and PostgreSQL
- Built a geospatial dashboard using MapLibre JavaScript, used by multiple national government agencies

Mosaik.id (Radya Laps Harapan Bangsa) January 2022 — June 2022

- Designed a Safe Internet Mobile App to promote Safe Internet usage for students in remote areas
- Developed the Backend Architecture, database, CI-CD and deployed the API on Azure single-handedly

Honors and Awards

Travel Grants

- SOSP ’24

Talks

Verifying Semantic Equivalence of Large Models with Equality Saturation

- EuroMLSys ’25, Rotterdam, The Netherlands

Technical Skills

Languages Python, Java, C, C++, C#, Javascript

Software Meta’s Deep Learning Recommendation Model, Ceph-RADOS, AWS Neuron SDK