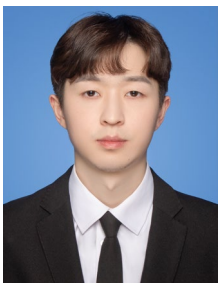
Curriculum Vitae

PERSONAL DETAILS



Name: Yicheng Feng (冯一诚) Date of Birth: March 1999 Homepage: fwyc0573.github.io

E-mail: yichengfengg@gmail.com Qualifications: Ph.D. Candidate English Proficiency: IELTS 6.5

EDUCATION

Ph.D. Student, Computer Science and Engineering, Computer Science & Engineering, The Chinese University of Hong Kong, Supervisor: Hong Xu. Aug. 2024 - Present

Master, College of Intelligence and Computing, Software Engineering, Tianjin University, Lab: Tanklab, Ranking: 1/228. Sept. 2021 - Jan. 2024

Bachelor, School of Artificial Intelligence and Computer, Digital Media Technology, Jiangnan University, GPA: 3.73/4.0, Average Mark: 90.4/100, Ranking: 5/133. Sept. 2017 - June 2021

RESEARCH INTERESTS

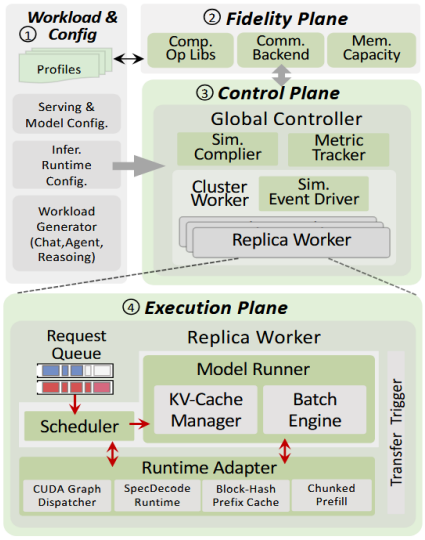
ML System & Cloud Computing: Currently focused on ML systems, especially simulation of training and inference in large-scale scenarios. Also interested in cloud-native technologies.

RESEARCH EXPERIENCE

NetX Lab, The Chinese University of Hong Kong & StepFun Co., Ltd., China

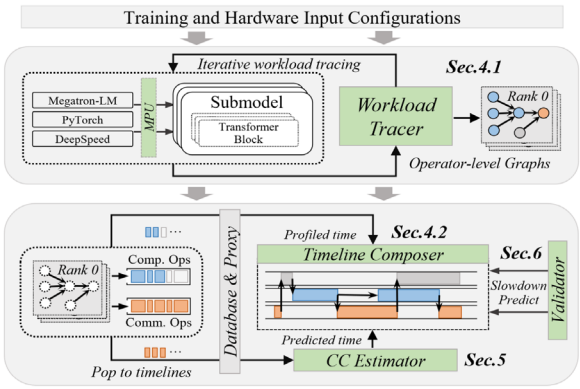
AI Infra Interns, Group of Training/Inference Framework Jan. 2024 – Present

Frontier: Towards Comprehensive and Accurate LLM Inference Simulation



- **Overview:** We build Frontier, a comprehensive and accurate LLM inference simulator covering modern serving architectures (co-location, PDD, AFD) and runtime optimizations (e.g., speculative decoding, CUDA Graph, prefix caching, parallelism). Its Fidelity Plane delivers high-precision predictions for computation, communication, and memory costs. A disaggregated orchestration framework with role-specific workers and rich primitives and events natively models pipeline stages, MoE straggler effects, disaggregated KV-cache/activation transfer. Together, these enable Frontier to faithfully simulate multi-phase reasoning serving and agentic RL rollouts, supporting large-scale Pareto frontier exploration under realistic dynamics.
- **Paper:** *https://arxiv.org/pdf/2605.21312,* code**:** *will release soon.*

Echo: Simulating Distributed Training at Scale



- **Overview:** We build Echo to tackle three key challenges in large-scale training simulation: (1) tracing the runtime training workloads at each device in an ex-situ fashion so we can use a single device to obtain the actual execution graphs of 1K-GPU training, (2) estimating the collective communication without high overheads of discrete-event based network simulation, and (3) accounting for the interference-induced computation slowdown from overlapping comm and comp kernels on the same device. Echo delivers on average 8% error. For GPT-175B on a 256-GPU H800 cluster with 3D parallelism on Megatron-LM under 3 minutes.
- **Paper:** *https://arxiv.org/abs/2412.12487,* code**:** https://github.com/NetX-lab/Echo *(keep updating)*

Dynamic Compute and Network Orchestration for Disaggregated RL

- **Overview:** We build a dynamic orchestration framework for disaggregated RL that jointly optimizes compute and network resources to overcome severe generation bottlenecks. The system features (1) an adaptive compute scheduler that employs proactive parallelism switching and reactive request migration to resolve intra-step workload imbalances caused by unpredictable response lengths. (2) To address phase-varying traffic demands, we co-designed a reconfigurable hybrid optical-electrical network that materializes stage-aware topologies on demand for training, generation, and weight synchronization. Evaluations demonstrate up to a 1.42x throughput increase on a 64-H800 testbed and up to 2.06x higher cost-efficiency at scale.
- **Paper:** *https://arxiv.org/abs/2601.01209*accepted by *ACM SIGCOMM* (2nd author), CCF-A, May. 2026,

PPIO Edge Clouds Co., Ltd., China.

Edge-cloud Development Interns, Group of IaaS, Department of Technology Center Jan. 2022 – Jan. 2024

BREAK: A Holistic Approach for Efficient Container Deployment among Edge Clouds

- **Overview:** We build BREAK, a K8s-compatible framework that accelerates container deployment on resource-constrained edge clouds. The core idea is to rethink the layer-based image structure for edge scenarios: (1) we built a container image refactoring solution that optimizes image composition while preserving the standard stack-of-layers format, and (2) introduces a storage-driver tailored for container runtime that enables chaotic, asynchronous, parallel layer extraction.

Tango: Harmonious Management and Scheduling for Mixed Services Co-located among Distributed Edge Clouds

- **Overview:** Designed and implemented Tango, a K8s-based scheduling framework that enables efficient co-location of Latency-Critical (LC) and Best-Effort (BE) services on distributed edge clouds. The core contribution is a Dynamic Vertical Pod Autoscaler that bypasses K8s's native resource management by injecting Cgroup-level control flows, enabling fine-grained, real-time elastic resource allocation with QoS re-assurance. The framework further integrates lightweight traffic scheduling for both LC and BE requests.

PUBLICATIONS

- **Frontier: Towards Comprehensive and Accurate LLM Inference Simulation**
  Under review, 1st author, short paper was accepted by SOSP Workshop PACMI '25
- **Echo: Simulating Distributed Training at Scale**
  Under recycle, 1st author
- **Dynamic Compute and Network Orchestration for Disaggregated RL**
  Accepted, 2nd author, ACM Special Interest Group on Data Communication (SIGCOMM), CCF-A, 2026
- **BREAK: A Holistic Approach for Efficient Container Deployment among Edge Clouds**
  Accepted, 1st author, IEEE International Conference on Computer Communication (INFOCOM), CCF-A, 2024
  Accepted, 1st author, ACM Asia-Pacific Workshop on Networking (APnet), CCF-C, 2023
- **Tango: Harmonious Management and Scheduling for Mixed Services Co-located among Distributed Edge Clouds**
  Accepted, 1st author, ACM International Conference on Parallel Processing (ICPP), CCF-B, 2023
  Accepted, 2nd author, IEEE Transactions on Services Computing (TSC), CCF-A, 2023
- **A Large-scale Holistic Measurement of Crowdsourced Edge Cloud Platform**
  Accepted, 1st author, Springer World Wide Web Journal (WWWJ), CCF-B, 2023 (conference-based extending)
  Accepted, 2nd author, IEEE International Workshop on Quality of Service (IWQoS), CCF-B, 2023

HONORS AND AWARDS

- **SOSP 2025 Student Scholarship & Travel Grant Award**
- **Outstanding Graduate of Jiangsu Province**, Jiangnan University First/Second Class Academic Scholarship
- **Outstanding Graduate of Tianjin University**, Meritorious Student of Tianjin University