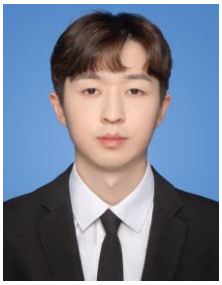
Curriculum Vitae

PERSONAL DETAILS



Name: Yicheng Feng (冯一诚) Date of Birth: March 1999 Github: https://fwyc0573.github.io
E-mail: yichengfengg@gmail.com Qualifications: Ph.D. Candidate English Proficiency: IELTS 6.5

EDUCATION BACKGROUND

Ph.D. Candidate, 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, Supervisor: Xiaofei Wang, 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

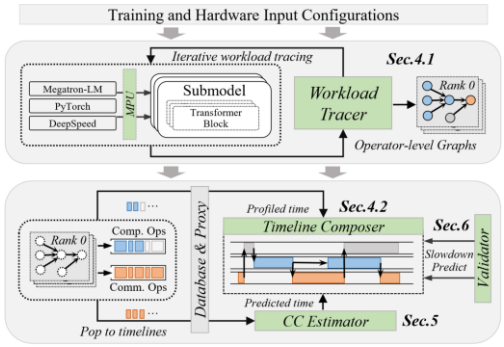
ML System & Cloud Computing: Currently focused on machine learning systems, especially simulation of training and inference in large-scale scenarios. Also interested in cloud-native technologies.

RESEARCH / PROGRAM EXPERIENCES

NetX Lab, Computer Science and Engineering, The Chinese University of Hong Kong

ML System Simulation Jan. 2023 – Present

Echo: Simulating Distributed Training at Scale

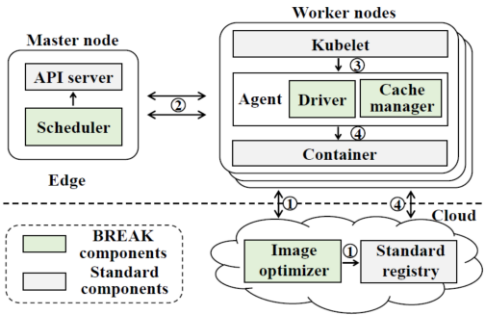


- **Background:** Simulation offers unique values for both enumeration and extrapolation purposes, and is becoming increasingly important for managing the massive machine learning clusters and large-scale distributed training jobs.
- **Overview:** We build Echo to tackle three key challenges in large-scale 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) accurately estimating the collective communication without high overheads of discrete-event based network simulation, and (3) accounting for the interference-induced computation slowdown from overlapping communication and computation kernels on the same device. Echo delivers on average 8% error in training step—~3x lower than state-of-the-art simulators—for GPT-175B on a 96-GPU H800 cluster with 3D parallelism on Megatron-LM under 2 minutes.
- **Paper:** *https://arxiv.org/abs/2412.12487*, under recycle, preparing for *ACM ASPLOS'26 Spring*.
- **Code:** *https://github.com/NetX-lab/Echo*

PPIO Edge Clouds Co., Ltd., China. (Industry-academia Cooperation)

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



- **Background:** Containers and containers orchestration platforms (COPs) have become increasingly popular in edge computing. However, the edge features, such as high latency links and resource-constrained, also bring unique challenges to fast container deployment.
- **Overview:** We rethink container's key design of layer-based structure and analyze the unique challenges and great potential that resides in edge clouds. We propose *BREAK, an accelerating framework for efficient container deployment among edge clouds:*

  1) We design a container image refactoring solution which is compatible with all current container engines and registry. It makes optimization which also preserves the convenient stack-of-layers structure of images;

2) We implement a customized K8s scheduler which extends the awareness of network performance, disk space, and container layer cache to make a suitable container placement for fast deployment;

3) We design a distributed shared layer-stack cache and make cooperative container deployment through push-based P2P layer transfer among geo-nearby edge clouds to accelerate deployment;

4) We dissected the problem of layer extraction and introduce storage-driver tailored for container runtime which allow chaotic, asynchronous, and parallel extraction while avoiding redundant extraction.

- Paper: accepted by *IEEE INFOCOM* (1st author), CCF-A, Dec. 2023; *APnet* (1st author), CCF-C, March 2023.
- **Code:** *https://github.com/fwyc0573/PreliminaryBooster*

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

- **Background:** Deploying Latency-Critical (LC) services and Best-Effort (BE) services in a mixed manner is expected to improve resource utilization in edge clouds. Yet, co-locating LC and BE services on edges faces unique challenges. Unlike cloud datacenters, edge clouds are heterogeneous, resource-constrained and distributed, which leads to more fierce competition for edge resources and makes it difficult to balance fluctuating co-located workloads.
- **Overview:** We introduce Tango, a harmonious scheduling framework for K8s-based edge-cloud systems with mixed services. Tango further enhances K8s with automatic scaling and traffic scheduling capabilities. Specifically, we make the following contributions:

  1) Tango incorporates novel components and mechanisms for elastic resource allocation, including (i) resource usage regulations, (ii) a dynamic vertical pod autoscaler (D-VPA) which add a low-level control flow to the resources of K8s Cgroup, and (iii) a QoS re-assurance mechanism;

  2) Two traffic scheduling algorithms are tailored: (i) a distributed agile network flows algorithm for LC requests and (ii) a centralized scheduling algorithm combining GNN and DRL for BE requests;

  3) Compared to the state-of-the-art approach, experiments on large-scale hybrid edge clouds show that Tango can improve the system resource utilization and throughput while guaranteeing the QoS of LC services.
- **Paper:** accepted by *ACM ICPP* (1st author), CCF-B, April. 2023.
- **Code***: https://github.com/fwyc0573/Tango*

A Large-scale Holistic Measurement of Crowdsourced Edge Cloud Platform

- **Overview:** In this paper, we perform the first-of-its-kind measurement on a large-scale crowdsourced edge platform, which covers over 10,000 edge servers, 100,000 users and 10,000,000 requests. The measurement takes a holistic view: 1) edge cloud servers, 2) containerized services, and 3) user requests features.
- **Paper:** accepted by *Springer WWWJ* (1st author), CCF-B; Accepted by *IEEE IWQoS* (2nd author), CCF-B.
- **Code:** *https://github.com/fwyc0573/MeasurementBasedModelingGenerators*

PUBLICATIONS

- **Echo: Simulating Distributed Training at Scale**
  Under recycle, 1st author, preparing for ACM ASPLOS'26 Spring
- **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 (preliminary work)
- **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
- **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
- **5 China patents** (2 in process, first author except supervisor), 2 CSCD journals, 1 software copyright

HONORS AND AWARDS

- **Outstanding Graduate of Tianjin University**, Meritorious Student of Tianjin University
- **Outstanding Graduate of Jiangsu Province**, Jiangnan University First/Second Class Academic Scholarship