Marc Chengliang Zhang

PRESENT POSITION

Ph.D Candidate
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong

Office: Room CYT 3007
Phone: +852–6216–2287
Email: czhangbn@cse.ust.hk
Web: https://marcoszh.github.io/

RESEARCH INTERESTS

My interests cover **big data analytics systems** and **cloud computing**, with a special focus on **machine learning systems**. I enjoy identifying fundamental system design and performance issues in large-scale ML systems for both training and inference, and searching for general and efficient solutions.

EDUCATION

Hong Kong University of Science and Technology, Hong Kong SAR
Department of Computer Science and Engineering

◊ **Ph.D.** Computer Science and Engineering — September 2016 - present
◊ *Supervisor:* Wei Wang
◊ Hong Kong PhD Fellowship awardee, a prestigious and highly selective fellowship.

Harbin Institute of Technology, Harbin, China
School of Computer Science and Technology

◊ **B.Eng.** Software Engineering — September 2012 - June 2016
◊ *Honors:* National Scholarship (Top 2%), People's Scholarship, Fuji Xerox Scholarship

PUBLICATIONS

Chengliang Zhang, Minchen Yu, Wei Wang, Feng Yan, "MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving," in the *Proceedings of USENIX Annual Technical Conference (ATC'19)*, Renton, WA, July 2018 (20% acceptance rate).

Chengliang Zhang, Huangshi Tian, Wei Wang, Feng Yan, "Stay Fresh: Speculative Synchronization for Fast Distributed Machine Learning," in the *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS'18)*, Vienna, Austria, July 2018 (20% acceptance rate).

Preprints

Chengliang Zhang, Minchen Yu, Wei Wang, Feng Yan, "Towards Cost-Effective and SLO-Aware Machine Learning Inference Serving on Public Cloud," to be submitted to *IEEE Transactions on Parallel and Distributed Systems*.

Yinghao Yu, Chengliang Zhang, Wei Wang, Jun Zhang, Khaled Letaief, "Towards Dependency-Aware Cache Management for Data Analytics Applications," submitted to *IEEE Transactions on Cloud Computing*, currently under review.

Research Experience

Fast Secure Federated Learning System

I am currently working on accelerating the training and inference process of secure federated learning. We partner with a large commercial bank, and focus on enabling inter-enterprise federated learning with the help of Homomorphic Encryption. Specifically, I am working on how to mitigate the encryption and communication overhead.

MArk: ML Serving on Public Cloud

We aspired to serve machine learning models on public cloud with both SLO compliance and cost-effectiveness. We first characterized ML serving and its performance on diverse cloud services, and then designed MArk based on our insights. MArk uses predictive autoscaling to maintain high utilization, and eliminates overprovisioning by utilizing flexible, yet expensive FaaS. MArk further brings down the cost by adopting ML accelerators judiciously.

Speculative Synchronization

Asynchronous parallel can improve distributed ML training's throughput. However, the introduced inconsistency leads to low quality updates. We proposed Speculative Synchronization to exploit the trade-off between training throughput and update quality, which allows a worker to abort the current computation and synchronize, if it is confident that the benefits of fresher parameter outweight the loss of completed computation.