Marc Chengliang Zhang

PRESENT POSITION

Ph.D Candidate

Dept. of Computer Science & 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 recipient: 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

PUBLICATIONS

Jun Yi, <u>Chengliang Zhang</u>, Wei Wang, Cheng Li, Feng Yan, "Not All Explorations Are Equal: Harnessing Heterogeneous Profiling Cost for Efficient MLaaS Training," in the *Proceedings of IEEE International Parallel and Distributed Processing Symposium (IPDPS'20)*, New Orleans, LA, May 2020.

<u>Chengliang Zhang</u>, 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 2019 (20% acceptance rate).

<u>Chengliang Zhang</u>, 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).

Yinghao Yu, <u>Chengliang Zhang</u>, Wei Wang, Jun Zhang, Khaled Letaief, "Towards Dependency-Aware Cache Management for Data Analytics Applications," in the *IEEE Transactions on Cloud Computing*.

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*.

RESEARCH EXPERIENCE

Research Scientist Intern @ Bell Labs Germany

- Serverless Computing
- Federated Learning

Fast Secure Federated Learning System

- Inter-enterprise federated learning with Homomorphic Encryption
- Accelerate training and inference of Secure Federated Learning
- Mitigate encryption and communication overhead

MArk: ML Serving on Public Cloud

- Serve machine learning inference on public cloud
- Cost-effective and SLO-aware
- Characterization of ML serving and its performance cloud services
- Combine FaaS and laaS to reduce over-provisioning
- Characterization of hardware accelerators like GPU and TPU

Speculative Synchronization

- Distributed data parallel training
- Relaxed consistency can increase throughput but hurt update quality
- Re-synchronize if the parameter copy is too stale to produce beneficial updates

SKILLS

 \diamond Python \diamond Java, Scala \diamond C++ \diamond Keras \diamond TensorFlow \diamond MXNet

♦ Hadoop / Spark
♦ Mandarin: native
♦ English: IELTS 8

REFERENCES

Wei Wang, Assistant Professor

Hong Kong University of Science and Technology

Room 3524, CSE Department, HKUST Clear Water Bay, Kowloon, Hong Kong

Phone: +852–2358–6972 Email: weiwa@cse.ust.hk

Web: https://www.cse.ust.hk/~weiwa/

Feng Yan, Assistant Professor

University of Nevada, Reno

SEM 233, CSE Department Reno, NV 89557

Phone: +1-775-784-6448 Email: fyan@unr.edu

Web: https://wolfweb.unr.edu/~fyan/