Wenchen Han

wenchen.han.22@ucl.ac.uk | (+44) 07548535016 | Personal website

EDUCATION

Department of Computer Science, University College London. Sep. 2022 – (Exp.) Sep. 2027

Ph.D. student in Systems and Networking group at UCL. Supervisors: Ran Ben Basat, Brad Karp.

School of EECS, Peking University (PKU) Sep. 2018 - Jun. 2022

B.S. in Intelligence Science and Technology, Turing Class

- GPA: 3.791/4.0, Rank: 4/82, Major GPA: 3.814/4.0
- Thesis: Hierarchical Aggregation for Efficient and Scalable Federated Learning in WAN.

RESEARCH INTERESTS

- *Algorithmic design for distributed ML systems and networked systems, programmable data plane.*

PUBLICATIONS

1. Yikai Zhao*, **Wenchen Han***, Zheng Zhong*, Yinda Zhang, Tong Yang, Bin Cui. Double-Anonymous Sketch: Achieving Fairness for Finding Global Top-K Frequent Items. In SIGMOD 2023.

2. Chaoliang Zeng, Layong Luo, Teng Zhang, Zilong Wang, Luyang Li, **Wenchen Han**, Nan Chen, Lebing Wan, Lichao Liu, Zhipeng Ding, Xiongfei Geng, Tao Feng, Feng Ning, Kai Chen, Chuanxiong Guo. Tiara: A Scalable and Efficient Hardware Acceleration Architecture for Stateful Layer-4 Load Balancing. In NSDI 2022.

* indicates equal contributions.

MANUSCRIPTS

1. **Wenchen Han**, Vic Feng, Gregory Schwartzman, Yuliang Li, Michael Mitzenmacher, Minlan Yu, Ran Ben Basat. FRANCIS: Fast Reactions Algorithms for Network Coordination In Switches. Under submission.

2. Ran Ben Basat, Gil Einziger, **Wenchen Han**, Bilal Tayh. SQUID: Faster Analytics via Sampled Quantiles Data-structure. *Under review.*

3. Ran Ben Basat, Keren Censor-Hillel, Yi-Jun Chang, **Wenchen Han**, Dean Leitersodorf, Gregory, Schwartzman. Bounded Memory in Distributed Networks. Under submission.

INDUSTRY EXPERIENCES

Data Center Networking Research Intern, ByteDance Inc., China Sep. 2020-Feb. 2021

Advisor: Teng Zhang.

- **Tiara: P4-FPGA-DPDK L4 Load Balancing System.**
  - Worked on designing a P4 + FPGA (fast path) + CPU (slow path) system for L4 load balancing to achieve both high throughput (1.6Tbps) and better scalability (supporting 80M concurrent flows).
  - Designed and implemented an efficient and scalable control plane + slow path components based on DPDK to make LB decisions and to offload connection table into FPGA-NICs for fast path forwarding, achieving a > 4Mpps/CPU-core throughput.

RESEARCH EXPERIENCES

Scalable and All-reduce-compatible Quantization-based Gradient Compression (GC) for DistML.

UCL. *Working with Ran Ben Basat, Shay Vargaftik, Michael mitzenmacher and Brad Karp.*

- Proposed scalable GC algorithms for different all-reduce topologies to minimize communication overhead.
- For ring topology, designed efficient algorithms to handle and reduce overflows during gradient aggregation

(integer summation) and to achieve better scalability w.r.t. the number of workers.

- Extended NCCL to support our algorithms, achieving ~60% less communication overhead compared with FP16.
- (Ongoing) Proposed a new BIT-structured tree topology to achieve (asymptotically) optimal scalability compared with ring for quantization-based compression.

Fast Reaction Algorithms for Network Coordination In Switches.

*UCL. Working with Ran Ben Basat, Minlan Yu, and Michael Mitzenmacher *.*

- Proposed to run distributed algorithms (DA) on P4 switches to achieve fast reaction to **network events**.
- Developed a general framework to facilitate the implementation of DA-based network events reaction in P4.
- Motivated our design choice with 3 real-world use cases, and showed that existing systems exhibit long delays in network events reaction, causing performance degradation.
- Designed DA protocols for each use cases respectively, and implemented them in Tofino to achieve 100× μs reaction time with minimum message overheads.

Double Anonymous Sketch for Fair Global Top-K Heavy Flow Detection.

PKU. Advisor: Tong Yang.

- Proposed a concept called fairness, and observed that under distributed load-imbalanced scenarios, the accuracy of prior solutions would drop significantly due to unfairness in aggregation.
- Developed a generic "strongly-unbiased" sketch called Double Anonymous Sketch that can be adapted to any existing sketch to achieve fairness in global top-K detection.
- Conducted extensive simulations and achieved significantly higher F1 Score than Waving Sketch and USS.

Sampled Quantiles q-MAX Algorithm for Score-based Caching in the Switches' Data Plane.

UCL. Working with Ran Ben Basat and Gil Einziger.

- Proposed a data plane (P4) algorithm atop SQUID, a sampled quantiles q-MAX algorithm, for in-network caching system, being the first to support a wide spectrum of caching policies and achieve real-time cache update.
- Implemented a prototype of SQUID-P4 and demonstrated that it achieves a near-optimal cache-hit ratio.

*: Some contributors are not listed here.

OTHER EXPERIENCES

Improving Triplet Loss for Metric Learning in Face Recognition. Jul. 2019 – Aug. 2019

Face Recognition Research Intern; Advisor: Shaoran Xiao; Megvii.

- Proposed and Implemented "Multi-batch Triplet Loss" for better convergence speed for face recognition.
- Deployed for the real-world practice by Megvii.

TALKS

- FRANCIS: Fast Reaction Algorithms for Network Coordination In Switches. In Coseners 2023.

ACADEMIC HONORS

- John Hopcroft Turing Class Award, PKU (8th / 58) Oct. 2021
- John Hopcroft Turing Class Award, PKU (11th / 58) Nov. 2020

SKILLS

- Programming Languages: C/C++, Python, CUDA, P4 (Tofino) data plane programming.
- Tools: NCCL, Pytorch, DPDK, NS-3 simulator, Mininet, etc.