Wenchen Han

wenchen.han.22@ucl.ac.uk +44 7548535016 charleshan24.github.io/CharlesHan.github.io/

EDUCATION

University College London *Sept 2022 – (Exp.) Sept 2026*

- Ph.D. candidate, Dept. of Computer Science. Supervisors: Ran Ben Basat, Brad Karp

Peking University *Sept 2018 – June 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.

STATEMENTS and KEYWORDS

My passion lies at the intersection of building efficient AI infrastructure and algorithmic innovations. My work spans both distributed LLM training and inference systems, where I am driven by inefficiencies of existing systems and seek algorithmic or system innovations to address them. Examples of relevant projects include co-designing gradient compression with all-reduce topologies to achieve scalable DDP collectives in LLM training, devising *progressive split-stream KV Cache transfer* with quantization for accelerating disaggregated LLM inference, etc.

- System designs for LLM inference & training; CUDA Quantization; LLM deployment vllm; Networked systems.

SELECTED RESEARCH AND INTERN EXPERIENCES

Research Intern, Lynx: Accelerating KV Transfer for LLM Inference *Edinburgh, UK*
SIR team, Huawei R&D UK. Manuscript B@SIGCOMM. *June 2025 - Jan 2026*

- Re-thought KV transfer as *divisible*: leveraged unequal bit significance to enable partial-state decoding.
- Designed progressive split-stream KV transfer in Lynx for disaggregated LLM inference, that overlaps KV transfer communication with decoding computation over partial KV state.
- Built speculative split-stream execution to achieve INT8/BF16-equivalent accuracy and INT4-equivalent low TTFT.
- Implemented in NPU kernel and integrated into vLLM serving pipeline, achieving 1.43× speed-up and BF16 accuracy.
- Submitted our work to SIGCOMM '26 and planned to open-source our artifact soon.

Research Assistant, Gradient Quantization for DDP LLM Training *University College London*
DynamiQ. Publication A@HotNets, *Manuscript C*@SIGCOMM. *Mar 2023 – June 2025*

- Devised a scalable gradient compression system for accelerating all-reduce collectives in LLM fine-tuning.
- Co-designed DynamiQ quantization with all-reduce topologies (e.g., butterfly), achieving better scalability.
- Implemented DynamiQ efficiently in **CUDA + NCCL + PyTorch DDP** hooks.
- Evaluated DynamiQ (5bit) against 4 LLM fine-tuning workloads, achieving 34% TTA acceleration over MXFP8.
- (HotNets '24) Presented issues and general principles for the designs and evaluation of gradient compression systems to maximize their utility, demonstrating them with three use cases.

Software Engineer Intern, Data Center Networking *Beijing, China*
ByteDance, Publication G@NSDI. *Sept 2020 – Mar 2021*

- Worked on developing Tiara, a P4+FPGA+CPU based stateful L4 load balancing system, achieving 1.6Tbps throughput while supporting 80M concurrent flows.
- Designed and implemented a high-performance (lock-free) control plane using DPDK+C making LB decisions and offloading connection tables to FPGA-NICs, achieving 4Mops per CPU core and microsecond-level latency.
- Collaborated with cross-functional hardware and software teams to deploy Tiara in production data centers.

OTHER RESEARCH EXPERIENCES

Research Assistant, Distributed Programmable Networks *PKU and UCL*
Fast Reaction Algorithms for Network Coordination In Switches. Manuscript D. *June 2021 – Feb 2023*

- Built **FRANCIS**, a P4-based framework that facilitates running message-passing algorithms (MPA) in distributed programmable switches to achieve fast reaction to network events.
- Developed **FRANCIS** prototypes for real-world use cases, namely clock synchronization, multicast, and routing.
- Delivered 100 × μs reaction time (18× faster than state-of-the-arts), improving application performance.

Research assistant, Network Measurement Algorithms *Peking University*
Double Anonymous Sketch. Publication F@SIGMOD. *Mar 2022 – Aug 2022*

- Developed a generic "strong-unbiased" *Double Anonymous Sketch* achieving fairness in global Top-K flow detection.
- Conducted extensive simulations and achieved significantly higher F1 Score than Waving Sketch and USS.

Research Assistant. Programmable in-network Caching. *University College London*
The SQUID project. Publication E@CoNEXT. *Sept 2022 – Mar 2023*

- Proposed a data plane (P4) algorithm atop **SQUID**, a sampled quantiles q-MAX algorithm, for in-network caching systems—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.

SKILLS

- **Programming Languages:** C/C++, Python, Golang, CUDA/GPU programming.
- **Software Engineering:** Distributed systems, Kubernetes, Docker, Bash, CMake/Makefile, Git, DPDK, RDMA.
- **GPU Computing:** CUDA profiling and optimization, CUDA-accelerated quantization, NCCL, MXFP8/4, etc.
- **ML Tools:** PyTorch and C++ Torch; familiarity with TensorFlow, vLLM.
- **LLM systems:** LLM deployment with vLLM; Gradient and KV quantization; Collective communication (NCCL); `lm-eval`; Disaggregated PD; Speculative decoding; Prefix Caching; etc.

PUBLICATIONS

A. Beyond Throughput and Compression Ratios: Towards High E2E Utility of Gradient Compression
Wenchen Han, Shay Vargaftik, Michael Mitzenmacher, Brad Karp, Ran Ben Basat. In HotNets '24.

B. Lynx: Progressive Speculative Quantization for accelerating KV Transfer in Long-Context Inference
Wenchen Han, Gingfung Matthew Yeung, Marco Barletta, William Toner, Amory Hoste, Adam Barker. Under review in SIGCOMM'26.

C. DynamiQ: Accelerating Gradient Synchronization using Compressed Multi-hop All-reduce
Wenchen Han, Shay Vargaftik, Michael Mitzenmacher, Ran Ben Basat. Under review in SIGCOMM '26.

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

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

F. SQUID: Faster Analytics via Sampled Quantiles Data-structure
(Alphabetical order) Ran Ben Basat, Gil Einziger, ***Wenchen Han***, Bilal Tayh. In CoNEXT 2024.

G. Tiara: A Scalable and Efficient Hardware Acceleration Architecture for Stateful Layer-4 Load Balancing
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. In NSDI '22

H. Achieving Top-K-fairness for Finding Global Top-K Frequent Items
Yikai Zhao, Wei Zhou, ***Wenchen Han***, Zheng Zhong, Yinda Zhang, Xiuqi Zheng, Tong Yang, Bin Cui. In TKDE 2025

I. Bounded Memory in Distributed Networks

(Alphabetical order) Ran Ben Basat, Keren Censor-Hillel, Yi-Jun Chang, ***Wenchen Han***, Dean Leitersodorf, Gregory, Schwartzman. In SPAA 2025.

TALKS

- **Towards High End-to-end Utility of Gradient Compression for AI Training.** In *HotNets 2024*.
- **Reduced Communication Is not All You Need: Towards High End-to-end Utility of Gradient Compression.** In *Coseners 2024*.
- **FRANCIS: Fast Reaction Algorithms for Network Coordination In Switches.** In *Coseners 2023*.

HONORS AND AWARDS

- **John Hopcroft Turing Class Award**, PKU (8th / 58) *Oct. 2021*
- **John Hopcroft Turing Class Award**, PKU (11th / 58) *Nov. 2020*
- **Silver Medals** in APIO 2017 and NOI 2017.

PROFESSIONAL SERVICES

- **CoNEXT '25** shadow TPC.