Yihua Cheng

+1 773-690-8622 yihua98@uchicago.edu https://apostac.github.io/about.html ApostaC

Research interest

Computer networks; Stream processing systems; Video codecs and video streaming systems; Deep learning systems.

Education

Ph.D. **University of Chicago**, Computer Science — 2020.9 — present
- Advised by Junchen Jiang

B.S. **Peking University**, Yuanpei College — 2016.9 — 2020.6

Industry Experience

Conviva Inc., Research Intern

CA, USA
2023.6—present
2022.6—2022.12
2020.10—2021.3

Time-state analytics: new generation data analytics with complex stateful queries
- Created a benchmark suite for state-of-the-art stream processing systems on time-state analytics workloads.
- Initial design and implementation for Conviva's new-generation data processing engine using Rust.
- Optimizing the data processing engine to get more than 6× higher throughput compared to the state-of-the-art stream processing systems (having the same cost).

Modeling QoE for video-on-demand services
- Conduct large-scale measurement of user's QoE sensitivity to quality incidents in VoD services.
- Analyze the correlation between video content and the user's QoE sensitivity.
- Propose a new algorithm using the user's sensitivity measurement to improve the overall QoE for all users.

Microsoft Research Asia, Star Leap Program

Beijing, China
2021.5—2021.9

RL-based congestion control algorithm for multi-party video conference
- Developed the multi-party video conference simulation tool for RL training and testing. (The tool is also used internally for later projects)
- Contributed to the ringmaster (`https://github.com/microsoft/ringmaster`) codebase to support multi-party conference emulation.
- Help train and evaluate the RL congestion control model.

Alibaba, research intern

Hangzhou, China
2020.5—2020.9

Real-time video platform for DingTalk and TaobaoLive
- Participate in building and optimizing the real-time video application testbed in the company.
- Implement Screen Content Coding Extension (SCC) feature in the video application SDK.

Data-center congestion control algorithm with programmable switches
- Implement the prototype of the congestion control algorithm within the Linux kernel.
- Implement the tool for automatic hyper-parameter tuning for the congestion control algorithm.
- Deploy Memcached, OpenStack, and Ceph to the testing cluster and generate the performance report.

Academic Research Projects

Ph.D. program, University of Chicago — 2020.9—present

- Data-driven video streaming QoE optimization
- Loss resilient real-time video through neural video codecs
- Faster LLM serving with KVCache and token streaming

Summer Intern, Johns Hopkins University — 2019.6—2019.9

- High-performance GPU-based packet classification

SOAR Group, Peking University — 2018.2—2020.6

- Bandwidth prediction for the cellular network in extreme-high mobility scenarios

Publications

GRACE: Loss-Resilient Real-Time Video through Neural Codecs — **NSDI 2024**

Yihua Cheng, Ziyi Zhang, Hanchen Li, Anton Arapin, Yue Zhang, Yuhan Liu, Kuntai Du, Xu Zhang, Francis Y. Yan, Amrita Mazumdar, Nick Feamster, Junchen Jiang

CacheGen: KV Cache Compression and Streaming for Fast Language Model Serving — ***Preprint***

Yuhan Liu, Hanchen Li, **Yihua Cheng**, Siddhant Ray, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, Junchen Jiang

Earth+: on-board satellite imagery compression leveraging historical earth observations — ***Preprint***

Kuntai Du, **Yihua Cheng**, Peder Olsen, Shadi Noghabi, Ranveer Chandra, Junchen Jiang

Online Profiling and Adaptation of Quality Sensitivity for Internet Video — **SoCC 2023**

Yihua Cheng, Hui Zhang, Junchen Jiang

Raising the Level of Abstraction for Time-State Analytics With the Timeline Framework — **CIDR 2023**

Henry Milner, **Yihua Cheng**, Jibin Zhan, Hui Zhang, Vyas Sekar, Junchen Jiang, Ion Stoica

Optimizing Real-Time Video Experience with Data Scalable Codec — **Sigcomm EMS 2023**

Hanchen Li, **Yihua Cheng**, Ziyi Zhang, Qizheng Zhang, Anton Arapin, Nick Feamster, Amrita Mazumdar

POLYCORN: Data-driven Cross-layer Multipath Networking for High-speed Railway through Composable Schedulerlets — **NSDI 2023**

Yunzhe Ni, Feng Qian, Taide Liu, **Yihua Cheng**, Zhiyao Ma, Jing Wang, Zhongfeng Wang, Gang Huang, Xuanzhe Liu, Chenren Xu

An Active-Passive Measurement Study of TCP Performance over LTE on High-speed Rails — **Mobicom 2019**

Jing Wang, Yufan Zheng, Yunzhe Ni, Chenren Xu, Feng Qian, Wangyang Li, Wantong Jiang, **Yihua Cheng**, Zhuo Cheng, Yuanjie Li, Xiufeng Xie, Yi Sun, Zhongfeng Wang

Additional Experience And Awards

Project mentor (2019 Autumn): Mentoring the project "High-speed packet classifier" in the course Computer Networks (Honor track) at Peking University

Peking University Student Cluster Team (2017-2020): 1^{st} prize in Asia Student Cluster Competition 2019; 6^{th} place in SC19 Student Cluster Competition; 6^{th} place in SC17 Student Cluster Competition

Technologies

Languages: Python, C++, Rust, C, CUDA, Java, Scala, SQL

Skills: Linux programming (vi, tmux, bash); Parallel programming (MPI, CUDA); Neural network development (PyTorch)