Hong Zhang

465 Soda Hall, UC Berkeley,
Berkeley, CA, 94720-1776 USA
☎ *+1 (415) 660-0797*
✉ *hongzhangblaze@gmail.com*
🔗 *https://hongzhangblaze.github.io*

Research Interests

I am broadly interested in computer systems and networking, with special focuses on distributed data analytics and ML systems, data center networking, and serverless computing. I develop high-performance, scalable systems and scheduling algorithms for big data and ML applications.

Research & Work Experience

2019.3 - Ongoing **Postdoctoral Scholar**, *RISELab, Electrical Engineering and Computer Science Department*, UC Berkeley, Berkeley, CA, USA.
Advisor: Prof. Ion Stoica

Education

2013.8 - 2019.3 **Ph.D. in Computer Science and Engineering**, *The Hong Kong University of Science and Technology (HKUST)*, Hong Kong, China.
Advisor: Prof. Kai Chen
Thesis: "Towards Efficient and Practical Network Optimization for Big Data Analytics"

2010.9 - 2013.7 **M.S. in Communication and Information System**, *Huazhong University of Science and Technology (HUST)*, Wuhan, China.
Advisor: Prof. Hongbo Jiang

2006.9 - 2010.7 **B.S. in Electronics and Information Engineering**, *Huazhong University of Science and Technology (HUST)*, Wuhan, China.
Advanced Class (60 selected from over 2000 engineering students)

Awards and Honors

2013-2019	HKUST Postgraduate Scholarship
2017	HKUST Research Travel Grant
2017	ACM SIGCOMM Student Grant
2016	**Google PHD Fellowship in Systems and Networking**
2015	ACM EuroSys Student Grant
2011	Student Grant for Infocom Student Activities
2010-2013	Postgraduate Exempted from Admission Exam with Full Scholarship, HUST

Selected Projects

2020.11-Present **SerFlex: Online adaptive ML serving system against bursty and unpredictable workload.** Developing an adaptive ML serving system that can timley react to bursty and unpredictable workloads to meet per-request latency requirements [**Ongoing**].

2020.10-Present **NetHint: Cooperative network optimization for big data and ML applications in public cloud.** Developing a network abstraction and an interactive mechanism between cloud provider and tenants to cooperatively enhance the performance of big data and ML applications [**Under submission**].

2019.6-2020.9 **Caerus: Timely task scheduling for serverless analytics.** Developed a task execution framework for serverless analytics. It optimizes both execution cost and job completion time by fully exploiting task execution dependencies [**NSDI'21**].

2017.9-Present **DeepScheduler: Optimizing parameter synchronization for ML training systems.** Developing a scheduling framework for model training systems. It fully exploits the allreduce communication pattern to speed up distributed ML training [**APNet'20, Under submission**].

2016.2-2017.2 **Hermes: Network load balancing system for big data applications.** Developed a resilient load balancing system that can gracefully handle uncertainties (e.g., congestions and failures) for big data applications in a practical, readily-deployable fashion [**SIGCOMM'17**].

2015.6-2016.2 **CODA: Automatic network optimization for big data applications.** Developed a network scheduler that can automatically identify and exploit application semantics (e.g., communication and execution dependencies) without manually updating applications [**SIGCOMM'16**].

2013.10-2015.2 **Amoeba: Deadline-aware networked system for inter-data center data transfers.** Developed a deadline-based network abstraction and a deadline-aware networked system to guarantee deadlines for inter-data center data transfers [**EuroSys'15, ToN'17**].

Publications

Peer-reviewed publications

[1] **Hong Zhang**, Yupeng Tang, Anurag Khandelwal, Jingrong Chen and Ion Stoica, "Caerus: NIMBLE Task Scheduling for Serverless Analytics" in *Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation* (**NSDI**), 2021.

[2] Xinchen Wan, **Hong Zhang**, Hao Wang, Shuihai Hu, Junxue Zhang and Kai Chen, "RAT - Resilient Allreduce Tree for Distributed Machine Learning" in *Proceedings of the 4th Asia-Pacific Workshop on Networking* (**APNet**), 2020, doi: 10.1145/3411029.3411037

[3] **Hong Zhang**, Kai Chen and Mosharaf Chowdhury, "Pas de Deux: Shape the Circuits, and Shape the Apps Too!" in *Proceedings of the 2nd Asia-Pacific Workshop on Networking* (**APNet**), 2018, doi: 10.1145/3232565.3232568

[4] **Hong Zhang**, Junxue Zhang, Wei Bai, Kai Chen, Mosharaf Chowdhury, "Resilient Datacenter Load Balancing in the Wild" in *Proceedings of the ACM SIGCOMM 2017 Conference* (**SIGCOMM**), 2017, doi: 10.1145/3098822.3098841

[5] **Hong Zhang**, Kai Chen, Wei Bai, Dongsu Han, Chen Tian, Hao Wang, Haibing Guan, Ming Zhang, "Guaranteeing Deadlines for Inter-Datacenter Transfers" in *IEEE/ACM Transactions on Networking* (**ToN**), 2017, doi: 0.1109/TNET.2016.2594235, ISSN: 1063-6692, Impact factor: 3.56

[6] **Hong Zhang**, Li Chen, Bairen Yi, Kai Chen, Mosharaf Chowdhury and Yanhui Geng, "Toward Automatically Identifying and Scheduling Coflows in the Dark" in *Proceedings of the ACM SIGCOMM 2016 Conference* (**SIGCOMM**), 2016, doi: 10.1145/2934872.2934880

[7] **Hong Zhang**, Hongbo Jiang, Bo Li, Fangming Liu, A. Vasilakos and Jiangchuan Liu, "A Framework for Truthful Online Auctions in Cloud Computing with Heterogeneous User Demands" in *IEEE/ACM Transactions on Computers* (**TC**), 2016, doi: 10.1109/TC.2015.2435784, ISSN:0018-9340, Impact factor: 3.75

[8] **Hong Zhang**, Kai Chen, Wei Bai, Dongsu Han, Chen Tian, Hao Wang, Haibing Guan, Ming Zhang, "Guaranteeing Deadlines for Inter-Datacenter Transfers" in *Proceedings of the 10th European Conference on Computer Systems* (**EuroSys**), 2015.

[9] **Hong Zhang**, Bo Li, Hongbo Jiang, Fangming Liu, A. Vasilakos and Jiangchuan Liu, "A Framework for Truthful Online Auctions in Cloud Computing with Heterogeneous User Demands" in *Proceedings of the 32nd Annual IEEE International Conference on Computer Communications* (**INFOCOM**), 2013, doi: 10.1109/INFCOM.2013.6566946, ISSN: 0743166X

Preprints

[10] **Hong Zhang**, Jingrong Chen, Junxue Zhang, Jiacheng Xia, Kai Chen, Junchen Jiang, Ion Stoica and Junhuan Sun, "De-colocated Scheduling for Distributed Deep Learning" *Under submission*

[11] Jingrong Chen, **Hong Zhang**, Wei Zhang, Liang Luo, Jeffrey Chase, Ion Stoica and Danyang Zhuo, "NetHint: White-Box Networking for Multi-Tenant Data Centers" *Under submission*

[12] Junxue Zhang, Chaoliang Zeng, **Hong Zhang**, Shuihai Hu, Mo Chen and Kai Chen, "LiteFlow – Neural Networks in Datapath" *Under submission*

Selected Talks

- Caerus: NIMBLE Task Scheduling for Serverless Analytics
  - NSDI, April 2021, Virtual Event
  - Duke Systems and Networking Seminar, Feb 2021, Virtual Event
- Just-on-time Job Scheduling on Serverless Architecture
  - Monthly Seminar of Google's Data Processing Group, February 2020, Sunnyvale
  - RISELab Winter Retreat, January 2020, Monterey
- Pas de Deux: Shape the Circuits, and Shape the Apps Too!
  - ACM APNet, July 2018, Beijing
- Network Scheduling for Big Data Applications
  - Berkeley RISELab Seminar, August 2018, Berkeley
  - HKUST CSE Department Seminar , March 2018, Hong Kong
- Resilient Datacenter Load Balancing in the Wild
  - ACM SIGCOMM, August 2017, Los Angeles
- Toward Automatically Identifying and Scheduling Coflows in the Dark
  - ACM SIGCOMM, August 2016, Florianopolis
- Guaranteeing Deadlines for Inter-Datacenter Transfers
  - ACM EuroSys, April 2015, Bordeaux
- A Framework for Truthful Online Auctions in Cloud Computing with Heterogeneous User Demands
  - IEEE INFOCOM, April 2013, Turin

Professional Activities

Technical Program Committee

SIGCOMM Poster and Demo 2019

Reviewer

IEEE/ACM Transactions on Networking
Transactions on Mobile Computing
Journal on Selected Areas in Communications
IEEE Transactions on Cloud Computing
IEEE INFOCOM
ACM TOMPECS
The Conference on Web and Internet Economics
MOBIQUITOUS
Journal of Parallel and Distributed Computing
Optical Switching and Networking
IEEE Communications Letters

Teaching

2017 Spring	Teaching Assistant, HKUST COMP 2021: Unix & Script Programming
2016 Fall	Teaching Assistant, HKUST COMP 3511: Operating Systems
2016 Spring	Teaching Assistant, HKUST COMP 3511: Operating Systems
2015 Spring	Teaching Assistant, HKUST COMP 2021: Unix & Script Programming
2014 Fall	Teaching Assistant, HKUST COMP 3511: Operating Systems
2014 Spring	Teaching Assistant, HKUST COMP 2611: Computer Organization

Student Advising

2019.9-Ongoing	Yupeng Tang, (Currently Ph.D. student, Yale University)
2017.12-2021.9	Jingrong Chen, (Currently Ph.D. student, Duke University)
2016.7-2017.6	Junxue Zhang, (Currently Ph.D. student, HKUST)
2015.9-2016.2	Bairen Yi, (Currently Software Engineer, ByteDance)
2016.3-2016.5	Final Year Project: Praveg Maheshwari, Sonali Vrat, and Justinas Lingys (Undergraduates in HKUST)