Conglong Li

conglong.li@gmail.com; conglong.li@microsoft.com
https://conglongli.github.io/

Research Interests

In general, I work on improving performance and resource efficiency of various computer systems via data analysis and algorithms/policies optimization. Currently I am focusing on optimizing AI systems platform for deep learning training and inference.

Education

2014–2020 **Ph.D. in Computer Science**, *Carnegie Mellon University*.
Advisor: David G. Andersen.
Thesis: Learned Adaptive Accuracy-Cost Optimization for Machine Learning Systems.

2013–2014 **M.S. in Computer Science**, *Rice University*.
Advisor: Alan L. Cox.
Thesis: GD-Wheel: A Cost-Aware Replacement Policy for Key-Value Stores.

2009–2013 **B.S. in Computer Science**, *Rice University*.
magna cum laude, distinction in research and creative work.
GPA 4.04, 1st in Dept. of Computer Science.

Work Experience

2020–Present **Researcher**, *Microsoft*, Bellevue, WA.
Working on AI systems platform optimization under Web Experiences Platform Org (Search, Ads, News, Edge, Maps). Currently focusing on the DeepSpeed project (https://github.com/microsoft/DeepSpeed).

Summer 2019 **Research Intern**, *Microsoft*, Bellevue, WA.
Worked on improving approximate nearest neighbor search performance (this project started in 2018). Designed ML models (GBDT and neural networks) to predict the search termination condition for each query. Evaluations demonstrate up to 7.1 times speedup under the same accuracy targets. Project ended up with a paper published at SIGMOD 2020.

Summer 2017 **Software Engineer Intern**, *Microsoft*, Bellevue, WA.
Worked on designing caching strategies for Bing Ads. Designed ML models (GBDT) to provide intelligent cache refresh decisions. Simulations on production traces demonstrate a potential 35.2 to 106.1 million dollars net profit gain in a quarter. Transferred the project to developing team to ship it in product. Project ended up with a paper published at WWW 2018.

Summer 2016 **Research Intern**, *Microsoft*, Redmond, WA.
Worked on designing caching strategies for Bing Ads. Designed domain-specific caching heuristics to save ads selection cost and improve net profit. Simulations on production traces demonstrate a potential 20.7 to 70.5 million dollars net profit gain in a quarter. Project ended up with a paper published at SoCC 2017.

Publications *(Google Scholar Profile)*

SIGMOD 2020 Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination.
Conglong Li, Minjia Zhang, David G. Andersen, Yuxiong He.
In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.*

MLSys 2019 Scaling Video Analytics on Constrained Edge Nodes.
Christopher Canel, Thomas Kim, Giulio Zhou, **Conglong Li**, Hyeontaek Lim, David G. Andersen, Michael Kaminsky, Subramanya R. Dulloor.
In *Proceedings of Machine Learning and Systems 2019.*

WWW 2018 Better Caching in Search Advertising Systems with Rapid Refresh Predictions.
Conglong Li, David G. Andersen, Qiang Fu, Sameh Elnikety, Yuxiong He.
In *Proceedings of the 2018 World Wide Web Conference.*

SoCC 2017 Workload Analysis and Caching Strategies for Search Advertising Systems.
Conglong Li, David G. Andersen, Qiang Fu, Sameh Elnikety, Yuxiong He.
In *Proceedings of the 2017 Symposium on Cloud Computing.*

ANCS 2017 Using Indirect Routing to Recover from Network Traffic Scheduling Estimation Error.
Conglong Li, Matthew K. Mukerjee, David G. Andersen, Srinivasan Seshan, Michael Kaminsky, George Porter, Alex C. Snoeren.
In *2017 ACM/IEEE Symposium on Architectures for Networking and Communications Systems.*

CoNEXT 2015 Scheduling Techniques for Hybrid Circuit/Packet Networks.
He Liu, Matthew K. Mukerjee, **Conglong Li**, Nicolas Feltman, George Papen, Stefan Savage, Srinivasan Seshan, Geoffrey M. Voelker, David G. Andersen, Michael Kaminsky, George Porter, Alex C. Snoeren.
Nominated for Best Paper.
In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies.*

EuroSys 2015 GD-Wheel: A Cost-Aware Replacement Policy for Key-Value Stores.
Conglong Li, Alan L. Cox.
In *Proceedings of the Tenth European Conference on Computer Systems.*

ACM TACO 2013 Vol. 10(4) Reducing DRAM Row Activations with Eager Read/Write Clustering.
Myeongjae Jeon, **Conglong Li**, Alan L. Cox, Scott Rixner.
In *ACM Transactions on Architecture and Code Optimization.*

Professional Service

MLSys 2020	External Reviewer (1 paper) for *Proceedings of Machine Learning and Systems 2020.*
PACT 2019	External Reviewer (2 papers) for *28th International Conference on Parallel Architectures and Compilation Techniques.*
Middleware 2018	External Reviewer (1 paper) for *Proceedings of the 19th International Middleware Conference.*
ICAC 2018	External Reviewer (1 paper) for *2018 IEEE International Conference on Autonomic Computing.*
IEEE CLOUD 2018	External Reviewer (1 paper) for *2018 IEEE 11th International Conference on Cloud Computing.*

Skills

Programming	Python, C, C++, C#, Java.
Speaking	English, Chinese (native), Japanese (JLPT N1).