HENGGANG CUI

1856 Fairhill Rd, Allison Park, PA, 15101
(651)403-3366 ⋄ cuihenggang@gmail.com

INTERESTS

Deep learning and large-scale machine learning systems.

EDUCATION

Carnegie Mellon University, Pittsburgh, PA.

Ph.D., Electrical and Computer Engineering (GPA: 3.93) — May, 2017

- Advisor: Greg Ganger
- Research Topic: Large-Scale Machine Learning Systems

Master of Science, Electrical and Computer Engineering — May, 2016

Tsinghua University, Beijing, China

Bachelor of Science, Electronic Information Science and Technology — July 2012

SKILLS

Programming languages	C++, Python, C, Java, CUDA, Scala
Software Systems	Git, Linux, Shell, LaTeX, Matlab, SVN
Big data Systems	TensorFlow, Caffe, Spark, MPI, GraphLab, HBase, OpenTSDB, Hadoop

INDUSTRY EXPERIENCE

Uber Advanced Technologies Group, Pittsburgh, PA.

Machine Learning Specialist — June, 2017 to *now*

- Develop deep learning models for self-driving vehicles.

PHD THESIS RESEARCH

Thesis: Exploiting Application Characteristics for Efficient System Support for Large-Scale Machine Learning

- **Committee**: Greg Ganger, Phil Gibbons, Garth Gibson, Eric Xing, and Derek Murray
- **Thesis Subprojects**: As follows:

GeePS: Specialized Parameter Server for Deep Learning on GPUs

Published at EuroSys'16, and open-sourced at https://github.com/cuihenggang/geeps

- Designed GeePS, a parameter server system for distributed deep learning on GPU machines.
- Extended Caffe, a popular deep learning system, to run on distributed GPU machines, by using GeePS.
- Achieved good scalability from single-machine Caffe (13× more throughput with 16 machines), by overlapping communication with computation.
- Supported DNNs that cannot fit in GPU memory, by swapping data to/from CPU memory in the background.

IterStore: Efficient Parameter Server for Iterative ML

Published at SoCC'14 and open-sourced at https://github.com/cuihenggang/iterstore

- Observed the iterativeness characteristic of many ML applications: same access sequence every iterations.
- Designed the *virtual iteration* method for collecting the repeating access sequence.
- Proposed five parameter server optimizations exploiting the collected info, including prefetching, marshalling-free contiguous data layout, locality and numa-aware data placement, and specialized caching policy.
- Experimented with real ML applications (matrix factorization, LDA, PageRank), showing up to 50× speedups.

MLtuner: Automatic ML tuning

Work under submission

- Designed the MLtuner system that automatically tunes ML task tunables, including learning rate, training batch size, and data staleness bound.
- Used efficient snapshotting and optimization-guided online trial-and-error, allowing MLtuner to find good tunable settings with little overhead.

- Experimented with real ML tasks, including deep learning and matrix factorization, showing that MLtuner automatically enables performance within 40–178% of optimal tunable settings.

Stale Synchronous Parallel (SSP): Trading Data Freshness for Speed
Published at NIPS'13 and ATC'14

- Designed a parameter server system that supports SSP, a new consistency model for synchronizing parallel ML workers, with tunable data staleness bound.
- Experimented with many real ML applications, exploring the tradeoffs of data freshness and speed.

INTERNSHIP

HP Labs, advised by Kimberly Keeton May 2014 to Aug 2014

Aperture: Ingest-Time Transformation for Big Time Series Data
Published at SoCC'15, with patent US20170322987A1

- Designed Aperture, a database system for big time series data analytics.
- Proposed the *ingest-time transformation* approach that allows Aperture to answer analytical queries more efficiently with transformed data.
- Experimented with many real use cases (correlation search, IP occurrence query, and anomaly detection), showing query latency reductions by over an order of magnitude, with minimal impact on ingest throughput.

PUBLICATIONS

1 **Henggang Cui**, Hao Zhang, Gregory R. Ganger, Phillip B. Gibbons, and Eric P. Xing. GeePS: Scalable Deep Learning on Distributed GPUs with a GPU-Specialized Parameter Server. In *ACM European Conference on Computer Systems (EuroSys'16)*, 2016.

2 Aaron Harlap, **Henggang Cui**, Wei Dai, Jinliang Wei, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, and Eric P. Xing. Addressing the Straggler Problem for Iterative Convergent Parallel ML. In *ACM Symposium on Cloud Computing (SoCC'16)*, 2016.

3 **Henggang Cui**, Kimberly Keeton, Indrajit Roy, Krishnamurthy Viswanathan, and Gregory R. Ganger. Using Data Transformations for Low-latency Time Series Analysis. In *ACM Symposium on Cloud Computing (SoCC'15)*, 2015.

4 Jinliang Wei, Wei Dai, Aurick Qiao, Qirong Ho, **Henggang Cui**, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing. Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics. In *ACM Symposium on Cloud Computing (SoCC'15)*, 2015.

5 **Henggang Cui**, Alexey Tumanov, Jinliang Wei, Lianghong Xu, Wei Dai, Jesse Haber-Kucharsky, Qirong Ho, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, and Eric P. Xing. Exploiting Iterative-ness for Parallel ML Computations. In *ACM Symposium on Cloud Computing (SoCC'14)*, 2014.

6 **Henggang Cui**, James Cipar, Qirong Ho, Jin Kyu Kim, Seunghak Lee, Abhimanu Kumar, Jinliang Wei, Wei Dai, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, and Eric P. Xing. Exploiting Bounded Staleness to Speed Up Big Data Analytics. In *USENIX Annual Technical Conference (ATC'14)*, 2014.

7 Qirong Ho, James Cipar, **Henggang Cui**, Jin Kyu Kim, Seunghak Lee, Phillip B. Gibbons, Garth A. Gibson, Gregory R. Ganger, and Eric P. Xing. More Effective Distributed ML via a Stale Synchronous Parallel Parameter Server. In *Neural Information Processing Systems (NIPS'13)*, 2013.

8 **Henggang Cui**, Danielle Rasooly, Moises R. N. Ribeiro, and Leonid Kazovsky. Optically Cross-Braced Hypercube: A Reconfigurable Physical Layer for Interconnects and Server-Centric Datacenters. In *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, 2012.

9 Dan Li, **Henggang Cui**, Yan Hu, Yong Xia, and Xin Wang. Scalable Data Center Multicast using Multi-class Bloom Filter. In *19th IEEE International Conference on Network Protocols (ICNP'11)*, 2011.

PROFESSION SERVICES

Program Committee: ACM Symposium on Cloud Computing 2018
Program Committee: ACM Symposium on Cloud Computing 2017
Program Committee: Computing Conference 2017