Yang You

Office: COM2-0201, NUS, Singapore
Email: youy@comp.nus.edu.sg
Phone: +65 86-488-197
https://ai.comp.nus.edu.sg/

Current Appointment (starting from 08/2020)

Assistant Professor (Tenure-Track) — National University of Singapore
6 fully-funded PhD & multiple postdoc positions available — *Department of Computer Science*

Education (08/2015 — 07/2020)

PhD in Computer Science — University of California, Berkeley
Advised by Prof. James Demmel — *Thesis: Parallel & Distributed Machine Learning Algorithms*

Research Interest

- **High Performance Computing**: Scalable Algorithms, Parallel Computing, Distributed Systems
- **Machine Learning**: Deep Learning, Optimization Algorithm, Matrix Computations

Selected Awards

[09/2020] Top Reviewers of International Conference on Machine Learning (ICML 2020)
[04/2020] Lotfi A. Zadeh Prize (This award recognizes a Berkeley graduating PhD student who has made outstanding contributions to soft computing and its applications)
[11/2019] Best Student Paper Finalist of SC (ACM/IEEE Supercomputing Conference)
[11/2019] Best Paper Candidate at ICDM (International Conference on Data Mining)
[08/2018] Best Paper Award of ICPP (1 out of 313 submissions: 0.3%, plenary presentation) [Link]
[11/2017] ACM/IEEE George Michael Memorial HPC Fellowship: the only PhD fellowship on ACM website
Media Coverage: [ACM] [Berkeley] [China] [EurekAlert] [IEEE] [insideHPC]
[07/2015] Outstanding Graduate of Tsinghua University (ranked 1st of 134 students, top 3 got the awards) [Link]
[07/2015] Outstanding Graduate of Beijing (ranked 1st of 134 students, top 4 got the awards) [Link]
[07/2015] Outstanding Graduate of Tsinghua CS Department (ranked 1st of 134 students, top 20 got the awards) [Link]
[07/2015] Best Thesis Award of Tsinghua University (10 out of 134 students: 7%) [Link]
[05/2015] Best Paper Award of IPDPS (4 out of 496 submissions: 0.8%, plenary presentation) [Link]
[10/2014] Siebel Scholar (35,000 USD for 1 year), 85 top students from world's leading universities [link]
[10/2011] National Scholarships of China (ranked 1st of 52 students, top 2 got the award) [Link]
[10/2010] National Scholarships of China (ranked 1st of 52 students, top 2 got the award) [Link]

Peer-Reviewed Publications

- **[KAIS'20] Y. You**, Y. He, S. Rajbhandari, W. Wang, C.-J. Hsieh, K. Keutzer, J. Demmel. Fast LSTM by dynamic decomposition on cloud and distributed systems, Journal of Knowledge and Information Systems. [pdf]
- **[ICLR'20] Y. You**, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, C.-J. Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. **It helped Google establish new state-of-the-art results on GLUE, RACE, and SQuAD benchmarks. It is being used by state-of-the-art models like Albert, ELECTRA, SHA-RNN.** [pdf].
- **[HPC Asia'20]** A. Wongpanich, **Y. You**, J. Demmel. Rethinking the Value of Asynchronous Solvers for Distributed Deep Learning. International Conference on High Performance Computing in Asia-Pacific Region. [pdf]
- **[SC'19] Y. You**, J. Hseu, C. Ying, J. Demmel, K. Keutzer, C.-J. Hsieh. Large-Batch Training for LSTM and Beyond, International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing), November 17-22, Denver, USA. 20.9% (72/344) direct acceptance rate for regular papers. **Best Student Paper Finalist**. [pdf]
- **[ICDM'19] Y. You**, Y. He, S. Rajbhandari, W. Wang, C.-J. Hsieh, K. Keutzer, J. Demmel. Fast LSTM Inference by Dynamic Decomposition on Cloud Systems, IEEE International Conference on Data Mining,

November 8-11, Beijing, China. 9.08% (95/1046) accptance rate for regular papers. **Best Paper Candidate**. [pdf]

- [**TPDS'19**] **Y. You**, Z. Zhang, C.-J. Hsieh, J. Demmel, K. Keutzer. Fast Deep Neural Network Training on Distributed Systems and Cloud TPUs, IEEE Transactions on Parallel and Distributed Systems, h5-index=76, ISSN: 1045-9219, DOI: 10.1109/TPDS.2019.2913833
- [**ICPP'18**] **Y. You**, Z. Zhang, C.-J. Hsieh, J. Demmel, K. Keutzer. ImageNet Training in Minutes, 47th International Conference on Parallel Processing. August 13th - 16th, Eugene, USA. **Best Paper Award (1 out of 313 submissions: 0.3%); The most cited HPC conference paper (HPDC, ICS, ICPP, IPDPS, PPOPP, SC, etc.) published between 2018 and 2020** (according to Google Scholar). [pdf] [code]
- [**ICS'18**] **Y. You**, J. Demmel, C.-J. Hsieh, R. Vuduc. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems, ACM International Conference on Supercomputing (ICS), June 12-15, Beijing, China. 18.7% (36/193) acceptance rate [pdf]
- [**SysML'18**] **Y. You**, Z. Zhang, C.-J. Hsieh, J. Demmel, K. Keutzer. Speeding up ImageNet Training on Supercomputers, System Machine Learning Conference, Feb 15, Stanford, USA. The first year of this conference only accepts 2-page paper [pdf]
- [**BMC Genomics'18**] Y. Zhao, C. Sun, D. Zhao, **Y. You**, et al. PGAP-X: extension on pan-genome analysis pipeline, BMC Genomics, DOI: 10.1186/s12864-017-4337-7 [pdf]
- [**SC'17**] **Y. You**, A. Buluc, J. Demmel. Scaling Deep Learning on GPU and Knights Landing Clusters, International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing), November 12-17, Denver, USA. 18.7% (61/327) acceptance rate [pdf]
- [**ICPP'17**] **Y. You**, J. Demmel. Runtime Data Layout Scheduling for Machine Learning Dataset, 46th International Conference on Parallel Processing. 28.4% (60/211) acceptance rate. [pdf]
- [**J-STARS'17**] W. Li, H. Fu, **Y. You**, L. Yu, J. Fang. Parallel Multiclass Support Vector Machine for Remote Sensing Data Classification on Multicore and Many-Core Architectures. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017 h5-index=45. DOI: 10.1109/JSTARS.2017.2713126 [pdf]
- [**TPDS'16**] **Y. You**, J. Demmel, K. Czechowski, L. Song, R. Vuduc. Design and Implementation of a Communication-Optimal Classifier for Distributed Kernel Support Vector Machines, IEEE Transactions on Parallel and Distributed Systems, h5-index=76, DOI: 10.1109/TPDS.2016.2608823 [pdf]
- [**NeurIPS'16**] **Y. You**, X. Lian, J. Liu, H. Yu, I. Dhillon, J. Demmel, C.-J. Hsieh. Asynchronous Parallel Greedy Coordinate Descent, Conference on Neural Information Processing Systems, Dec 05-10, Barcelona, Spain. 22.7% (568/2500) acceptance rate [pdf] [link]
- [**JPDC'16**] **Y. You**, H. Fu, D. Bader, G. Yang. Designing and Implementing a Heuristic Cross-Architecture Combination for Graph Traversal, Journal of Parallel and Distributed Computing, h5-index=36, DOI: 10.1016/j.jpdc.2016.05.007 [pdf]
- [**IPDPS'15**] **Y. You**, J. Demmel, K. Czechowski, L. Song, R. Vuduc. CA-SVM: Communication-Avoiding Support Vector Machines on Distributed Systems. **Best Paper Award (4 out of 496 submissions: 0.8%)** of IEEE International Parallel and Distributed Processing Symposium, May 25-29, Hyderabad, INDIA. DOI: 10.1109/IPDPS.2015.117 [pdf] [code]
- [**JPDC'15**] **Y. You**, H. Fu, S. Song, A. Randles, D. Kerbyson, A. Marquez, G. Yang, A. Hoisie. Scaling Support Vector Machines on the Modern HPC Platforms, Journal of Parallel and Distributed Computing, h5-index=36, DOI: 10.1016/j.jpdc.2014.09.005 [pdf]
- [**IPDPS'14**] **Y. You**, S. Song, H. Fu, A. Marquez, M. Dehnavi, K. Barker, K. Cameron, A. Randles, G. Yang. MIC-SVM: Designing A Highly Efficient Support Vector Machine For Advanced Modern Multi-Core and Many-Core Architectures. IEEE Parallel and Distributed Processing Symposium, May 19-23, Phoenix, USA. 21% (114/541) overall acceptance rate; 17.5% acceptance rate for software track. DOI: 10.1109/IPDPS.2014.88 [pdf] [code]
- [**ICPP'14**] **Y. You**, D. Bader, M. Dehnavi. Designing a Heuristic Cross-Architecture Combination for Breadth-First Search, 43rd International Conference on Parallel Processing, Sep 9-12, Minneapolis, USA. 36% (54/150) acceptance rate. DOI: 10.1109/ICPP.2014.16 [pdf]
- [**IJHPCA'14**] **Y. You**, H. Fu, S. Song, M. Dehnavi, L. Gan, X. Huang, G. Yang. Evaluating the Many-core and Multi-core architectures through accelerating LWC stencil on Multi-core and Many-core architectures. International Journal of High Performance Computing Application (2013 SCI IF=1.625), **21% (5/24) acceptance rate**. DOI: 10.1177/1094342014524807 [pdf]

- **[ICPADS'14]** L. Gan, H. Fu, W. Xue, Y. Xu, C. Yang, X. Wang, Z. Lv, **Y. You**, G. Yang, and K. Ou. Scaling and Analyzing the Stencil Performance on Multi-Core and Many-Core Architectures. IEEE International Conference on Parallel and Distributed Systems (ICPADS). DOI: 10.1109/PADSW.2014.7097797 [pdf]
- **[IPDPS-W'13]** **Y. You**, H. Fu, X. Huang, G. Song, L. Gan, W. Yu, G. Yang. Accelerating the 3D Elastic Wave Forward Modeling on GPU and MIC. IEEE Parallel and Distributed Processing Symposium **Workshops**, May 20-24, Boston, USA. One of the **best papers** of AsHES workshop. DOI: 10.1109/IPDPSW.2013.216 [pdf]

Technical Reports

- **Y. You**, I. Gitman, B. Ginsburg. Scaling SGD Batch Size to 32K for ImageNet Training. NeurIPS workshop. Widely used in industry. **Available in Intel Caffe, NVIDIA Caffe, Facebook Caffe2 (PyTorch), and Google's distributed TensorFlow**. [pdf]

Impact of Our Research

- **[11/02/2017]** We broke the ImageNet training speed record (48 minutes). [link]
- **[11/07/2017]** We broke the ImageNet training speed record (31 minutes). [link]
- **[12/12/2017]** We broke the ImageNet training speed record (14 minutes). [link]
- **[07/30/2018]** Tencent (Jia et al.) used our LARS optimizer to break ImageNet training speed record (6.6 minutes). [link]
- **[11/13/2018]** Sony (Mikami et al.) used our LARS optimizer to break ImageNet training speed record (3.7 minutes). [link]
- **[11/16/2018]** Google (Ying et al.) used our LARS optimizer to break ImageNet training speed record (2.2 minutes). [link]
- **[03/29/2018]** Fujitsu (Yamazaki et al.) used our LARS optimizer to break ImageNet training speed record (74.7 seconds). [link]
- **[04/01/2019]** We broke the BERT training speed record (76 minutes). [link]
- **[07/10/2019]** Google (Kumar et al.) used our LARS optimizer to break ImageNet training speed record (67.1 seconds). [link]
- **[08/13/2019]** NVIDIA used our training recipe to break BERT training speed record (53 minutes). [link]
- **[09/26/2019]** Google (Lan et al.) used our LAMB optimizer to establish new state-of-the-art results on GLUE, RACE, and SQuAD benchmarks. [link]
- **[10/02/2019]** Our LARS optimizer was added to MLPerf Training Benchmark, which is an industry standard to evaluate fastest deep learning systems and implementations. [link]
- **[02/13/2020]** Our LARS optimizer helped Geoffrey Hinton's team (Chen et al.) to achieve state-of-the-art ImageNet classification results by semi-supervised learning. [link]

Experience

UC Berkeley Computer Science Division — Berkeley, CA, USA
Graduate Student Researcher (GSR) — *08/2015 – 07/2020*

- Performance Benchmark and Optimization for Deep Neural Networks
- Communication Avoiding Machine Learning Algorithms on Distributed systems
- Communication-Efficient Solver for Kernel Ridge Regression (600× speedup without losing accuracy)
- Fast DNN Training for ImageNet on CPUs: AlexNet in 11 minutes and ResNet-50 in 15 minutes

Google Brain — Mountain View, CA, USA
Student Researcher — *09/2019 – 05/2020*

- Optimize TensorFlow for Large-Scale Deep Learning on TPU Pod

Google Brain — Mountain View, CA, USA
Student Researcher — *01/2019 – 05/2019*

- Optimize TensorFlow for Large-Scale Deep Learning on TPU Pod

Intel Labs — Santa Clara, CA, USA
Research Intern — *08/2018 – 12/2018*

- Fast and Efficient LSTM Training

Google Brain Mountain View, CA, USA
Software Engineering Intern *05/2018 – 08/2018*
- Optimize TensorFlow for Large-Scale Deep Learning on TPU Pod

Microsoft Research Redmond, WA, USA
Research Intern *01/2018 – 05/2018*
- Fast LSTM Inference on Cloud System
- Design and Implement Approaches based on SVD and Tensor Decomposition
- Achieved up to 30× Speedup and 20× Parameter Reduction

NVIDIA Santa Clara, CA, USA
Deep Learning Intern *05/2017 – 08/2017*
- Scaling SGD Batch Size to 32K for ImageNet training by ResNet50 model
- Achieve 3× speedup over standard AlexNet-ImageNet Training on DGX station
- Enables multiple solvers on each GPU, which achieves 1.4× speedup over 1-solver-per-GPU

IBM T. J. Watson Research Center Yorktown, NY, USA
Research Intern *05/2016 – 08/2016*
- Design communication-optimized GPU-enabled learning algorithms
- Improve the communication efficiency of Elastic Averaging SGD
- Evaluate collective operations on GPUs (e.g., NCCL)

High Performance Computing Lab, Georgia Institute of Technology Atlanta, GA, USA
Research Assistant (Exchange Student) *05/2014 – 08/2014*
- Convert a communication-intensive algorithm (SMO) to a communication avoiding algorithm (CA-SVM)
- CA-SVM achieves 7× average speedup over the original algorithm with only 1.3% average losses in accuracy
- CA-SVM keeps 95.3% weak scaling efficiency when we increase the number of processors from 96 to 1536

High Performance Computing Lab, Georgia Institute of Technology Atlanta, GA, USA
Research Assistant (Exchange Student) *10/2013 – 11/2013*
- Adaptive method based on regression, which supports the runtime combination technique
- Cross-architecture combination, which achieves 8.5×, 2.6×, and 2.2× average speedup over MIC, CPU and GPU
- Pairwise comparison between CPU, GPU and MIC, which helps users select the best architectures

Department of Computer Science, Tsinghua University Beijing, China
Research Assistant *09/2012 – 07/2015*
- Design and implement MIC-SVM, a highly parallel support vector machines for x86 many-core architectures
- Adaptive support for input patterns and data parallelism to fully utilize the multi-level parallelism
- MIC-SVM achieves 4.4-84× and 18-47× speedups against LIBSVM on MIC and Ivy Bridge CPUs respectively

Institute of High Performance Computing, Tsinghua University Beijing, China
Research Assistant *06/2011 – 09/2011*
- Developed a distributed system for automated software deployment and user data storage

Teaching

Designing, Visualizing and Understanding Deep Neural Networks Berkeley, CA, USA
UC Berkeley CS194-129 (funding from Google) *08/2016 – 12/2016*
- Algorithms, Applications, and Implementations of Deep Learning Techniques
- Head TA/GSI of Prof. John Canny

Operating Systems and Systems Programming Berkeley, CA, USA
UC Berkeley CS162 *08/2018 – 12/2018*
- Theory, Algorithms, and Implementations of Operating Systems
- TA/GSI of Prof. Ion Stoica

Contributions to Open-Source Software

[Asyn SVM]: the fastest implementation for Kernel Support Vector Machines on shared systems as of 2016

[CA-SVM]: a Communication-Avoiding approach for Kernel Support Vector Machines on distributed systems

[MIC-SVM]: an efficient design of Sequential Minimal Optimization approach for SVM on shared-memory systems

[NVIDIA-Caffe]: I enabled multiple solvers on each GPU, which achieves 1.4× speedup over 1-solver-per-GPU

[NVIDIA-Caffe]: I developed the LARS algorithm with B. Ginsburg and I. Gitman for large-batch training

[Intel-Caffe]: I helped Intel team implement large-batch DNN training algorithms

[Tensorflow]: I helped Sameer Kumar and Chris Ying implement large-batch DNN training algorithms

Academic Services

- [**AAAI'21**] Member of the Program Committee for the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) [link].
- [**SC'21**] Technical Papers Committee of International Conference for High Performance Computing, Networking, Storage and Analysis (formerly Supercomputing) [link].
- [**ISC'21**] Program Committee of ISC High Performance, formerly known as the International Supercomputing Conference. [link].
- [**IPDPS'21**] Program Committee of IEEE International Parallel and Distributed Processing Symposium. [link].
- [**ICLR'21**] Reviewer of International Conference on Learning Representations [link].
- [**TPDS**] regular reviewer of IEEE Transactions on Parallel and Distributed Systems (reviewed tens of papers) [link].
- [**NeurIPS'20**] Reviewer of Conference on Neural Information Processing Systems [link].
- [**HiPC'20**] PC member of International Conference on High Performance Computing, Data, and Analytics [link].
- [**TPDS-SS**] Program Committee Member of IEEE Transactions on Parallel and Distributed Systems (Special Section on Parallel and Distributed Computing Techniques for AI, ML, and DL), 2020 h5-index=76 [link].
- [**IEEE Access'20**] Reviewer of IEEE Access [link].
- [**TKDE'20**] Twice reviewer of IEEE Transactions on Knowledge and Data Engineering [link].
- [**ICML'20**] Reviewer of International Conference on Machine Learning [link].
- [**TBD'20**] Reviewer of IEEE Transactions on Big Data [link].
- [**IPDPS'20**] External Reviewer of International Parallel and Distributed Processing Symposium [link].
- [**T-SP'19**] Reviewer of IEEE Transactions on Signal Processing [link].
- [**TOPC'19**] Reviewer of ACM Transactions on Parallel Computing [link].
- [**NEUNET'19**] Reviewer of Journal of Neural Networks [link].
- [**CLUS'19**] Reviewer of Cluster Computing: the Journal of Networks, Software Tools and Applications [link].
- [**IBM'19**] Reviewer of IBM Journal of Research & Development [link].
- [**ICPP'18**] Reviewer of International Conference on Parallel Processing [link].
- [**NCAA**] 2 times reviewer of Neural Computing and Applications [link].
- [**FCGS**] Reviewer of Future Generation Computer Systems, h5-index=63 [link].
- [**CCGRID'18**] Reviewer of IEEE International Symposium on Cluster Computing and the Grid [link].
- [**JMLR'17**] Reviewer of Journal of Machine Learning Research, h5-index=70 [link].
- [**JPDC**] Two times reviewer of Journal of Parallel and Distributed Computing, h5-index=36 [link].
- [**IJCAI'17**] **Senior Program Committee member** of International Joint Conference on Artificial Intelligence. Melbourne, Victoria, Australia, August 19 - 25, 2017 [link].
- [**IPDPS'17**] Sub-Reviewer in Algorithms Track of IEEE International Parallel and Distributed Processing Symposium. Orlando, Florida, USA, May 29 - June 2, 2017 [link].
- [**APDCM'16**] Reviewer of 18th Workshop on Advances in Parallel and Distributed Computational Models. Chicago, Illinois, USA, May 23 - 27, 2016 [link].

Media Coverage on Research

- [**i-programmer**] ImageNet Training Record - 24 Minutes, Sep 21, 2017 [link], [copy].
- [**EurekAlert**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**ScienceDaily**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**NSF**] Supercomputing speeds up deep learning training, Nov 13, 2017 [link], [copy].
- [**Intel**] Solving Science and Engineering Problems with Supercomputers and AI, Nov 15, 2017 [link], [copy].
- [**Berkeley**] EECS-affiliated team break record for fastest deep learning training, Nov 15, 2017 [link].
- [**R&D Magazine**] Supercomputing Speeds Up Deep Learning Training, Nov 15, 2017 [link], [copy].
- [**fourthventricle**] Supercomputing Speeds Up Deep Learning Training, Nov 17, 2017 [link], [copy].
- [**techxplore**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- [**TACC**] Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].

- **[Science NewsLine]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[Topix]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[Technology News]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[Get Knows]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017, [link].
- **[Technology Networks]** Deep Learning Training Accelerated by Super Computing, Nov 14, 2017 [link], [copy].
- **[Primeur Magazine]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[World IT]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[Parallel State]** Supercomputing Speeds Up Deep Learning Training, Nov 13, 2017 [link], [copy].
- **[CACM]** Supercomputing Speeds Up Deep Learning Training, Nov 17, 2017 [link], [copy].
- **[Intel Software]** Intel CPUs for Deep Learning Training, Nov 17, 2017 [link], [copy].
- **[The Next Web]** Facebook's nerds bested by Japans in the race to train AI, Nov 20, 2017 [link], [copy].
- **[Synced]** New Google Brain Optimizer Reduces BERT Pre-Training Time From Days to Minutes, Apr 4, 2019 [link].
- **[Google]** Google's scalable supercomputers for machine learning, Cloud TPU Pods, are now publicly available in beta, May 7, 2019 [link].

Selected Public Talks

- **[11/19/2019]** Large-Batch Training for LSTM and Beyond. ACM/IEEE Supercomputing Conference (Denver, Colorado, USA).
- **[11/09/2019]** Fast LSTM Inference by Dynamic Decomposition on Cloud Systems. IEEE International Conference on Data Mining (Beijing, China).
- **[05/28/2019]** Fast and Accurate Distributed Deep Neural Networks Learning. Jiangmen Venture Capital (livestream).
- **[05/09/2019]** Real-Time Distributed DNN Learning: Training ImageNet in Minutes. Nanjing University Youth Forum (Nanjing, Jiangsu, China).
- **[05/06/2019]** Fast and Accurate Deep Neural Networks Training. Shanghai Jiaotong University (Shanghai, China).
- **[11/16/2018]** AutoTuning for Large-Batch Deep Learning. SC Deep Learning workshop (Dallas, TX, USA).
- **[11/01/2018]** AutoTuning for Large-Batch Training. BeBop meeting (Berkeley, CA, USA).
- **[08/14/2018]** Can we Train ImageNet in Minutes? ICPP 2018 (Eugene, OR, USA).
- **[06/22/2018]** Speed up Deep Learning on Distributed Systems. Alibaba (Beijing, China).
- **[06/15/2018]** Communication-Avoiding Kernel Ridge Regression. ICS'18 (Beijing, China).
- **[06/07/2018]** Speed up Deep Learning on Latest Supercomputers. China Academy of Science (Shanghai, China).
- **[06/05/2018]** Fast Deep Learning on Latest Supercomputers. Sun Yat-sen University (Guangzhou, Guangdong, China).
- **[04/04/2018]** ImageNet Training in Minutes: Scaling SGD Batch Size to 32K. Matrix Computations Seminar (Berkeley, CA, USA).
- **[03/26/2018]** Fast Deep Learning by Large-Batch Training. Microsoft Research (Redmond, WA, USA).
- **[11/15/2017]** Fast and Scalable Deep Learning on GPU and KNL Clusters. Supercomputing'17 (Denver, CO, USA).
- **[11/08/2017]** Fast and Scalable Deep Learning on KNL Clusters. Intel PCL (Santa Clara, CA, USA).
- **[08/17/2017]** Scaling the Batch Size to 32K without Losing Accuracy. Imperial College (London, UK).
- **[08/16/2017]** Runtime Data Layout Scheduling for Deep Learning Applications. ICPP'17 (Bristol, UK).
- **[08/04/2017]** Large Batch Training for GPUs. NVIDIA (Santa Clara, CA, USA).
- **[12/12/2016]** Asynchronous Coordinate Gradient Descent on Shared Memory System. Pierre and Marie Curie University (Paris, France).
- **[11/09/2016]** Asynchronous Parallel Greedy Coordinate Descent. Matrix Computations Seminar (Berkeley, CA, USA).

- **[08/11/2016]** Asynchronous Optimization Method for Multi-GPU System. IBM Watson Research Center (Yorktown, New York, USA).
- **[04/13/2016]** Communication-Efficient Support Vector Machines. SIAM PP'16 (Paris, France).
- **[02/10/2015]** CA-SVM: Communication-Avoiding Support Vector Machines on Distributed Systems. BeBop meeting (Berkeley, CA, USA).
- **[05/21/2014]** MIC-SVM: Designing A Highly Efficient Support Vector Machine For Advanced Modern Multi-Core and Many-Core Architectures. IPDPS'14 (Phoenix, AZ, USA).
- **[05/20/2013]** Accelerating the 3D Elastic Wave Forward Model on GPU and MIC. IPDPS AsHES workshop (Boston, MA, USA).

Mentoring and Service

- **Since 2015 Fall:** Mentor for 1 or 2 UC Berkeley EECS undergraduate students in research/study per semester.
- **2018 Spring:** UC Berkeley Computer Science Division Student Core Committee for Faculty Hiring.
- **Since 2016 Spring:** Student volunteer and host for incoming UC Berkeley EECS PhD students each year.
- **2012 Fall - 2015 Spring:** Student leader in Tsinghua University youth league for organizing technology talks and exhibitions.
- **2010 Fall - 2011 Fall:** Chief student leader in CAU's EECS honors program.

Skills

- General: **C/C++, Matlab**, **Python**, **Java**, **Scala**, **Lua** and **Shell script**
- Multi-Core GPUs, CPUs and MIC: **CUDA**, **OpenMP**, **Pthreads** and **Intel Cilk**
- Distributed Systems: **MPI**, **Hadoop**, and **Apache Spark**
- Tools: **Caffe**, **TensorFlow**

Certificates

- Crash Course in Deep Learning with Google TensorFlow and Python (**Udemy**)