Wencong Xiao

xiaowencong@gmail.com +86 18611201750
T2 #12328, No.5 Danling Street, Haidian, Beijing, China, 100080
5th year Ph.D. candicate in joint Ph.D. program of MSRA and BUAA

Education

2014 – 2019 **Ph.D., Beihang University, in Distributed System**
Joint Ph.D. program with Microsoft Research Asia
Supervisors: Lidong Zhou (MSRA), Wei Li (Beihang University)

2010 – 2014 **B.S., Computer Science, Beihang University**
Thesis title: *Job Performance Study on Big Data Platform*.

Internship Experience

2013.07 - 2019.03 Microsoft Research Asia.
System Research Group, Mentor: Ming Wu, Lidong Zhou

2016.07 - 2016.10 Microsoft Research Redmond.
System Research Group, Mentor: Lidong Zhou

Research Interests

Distributed system, machine learning system, resource management, graph computing

Publications (selected)

Conference paper

Gandiva: Introspective Cluster Scheduling for Deep Learning **OSDI'18**
Wencong Xiao, Romil Bhardwaj, Ramachandran Ramjee, Muthian Sivathanu, Nipun Kwatra, Zhenhua Han, Pratyush Patel, Xuan Peng, Hanyu Zhao, Quanlu Zhang, Fan Yang and Lidong Zhou

Tux^2: Distributed Graph Computation for Machine Learning **NSDI'17**
Wencong Xiao, Jilong Xue, Youshan Miao, Zhen Li, Cheng Chen, Ming Wu, Wei Li and Lidong Zhou

Scheduling CPU for GPU-based Deep Learning Jobs **SoCC'18 Poster**
Wencong Xiao, Zhenhua Han, Hanyu Zhao, Xuan Peng, Quanlu Zhang, Fan Yang and Lidong Zhou

Optimization Mapping for Deep Learning **SOSP'17 AISys**
Wencong Xiao, Cheng Chen, Youshan Miao, Jilong Xue and Ming Wu

All You Need to Know about Scheduling Deep Learning Jobs **SOSP'17 SRC**
Wencong Xiao, Fan Yang and Lidong Zhou

Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN **ATC'19**
Myeongjae Jeon, Shivaram Venkataraman, Amar Phanishayee, Junjie Qian, **Wencong Xiao** and Fan Yang

SeerNet: Predicting Convolutional Neural Network Feature-Map Sparsity through Low-Bit Quantization **CVPR'19**
Shijie Cao, Lingxiao Ma, **Wencong Xiao**, Chen Zhang, Yunxin Liu, Lintao Zhang, Lanshun Nie and Zhi Yang

Efficient and Effective Sparse LSTM on FPGA with Bank-Balanced Sparsity **FPGA'19**
Shijie Cao, Chen Zhang, Zhuliang Yao, **Wencong Xiao**, Lanshun Nie, Dechen Zhan, Lintao Zhang, Hsiao-Wuen Hon, Ming Wu and Yunxing Liu

Balanced Sparsity for Efficient DNN Inference on GPU **AAAI'19**
Zhuliang Yao, Shijie Cao, **Wencong Xiao**, Chen Zhang and Lanshun Nie

KV-Direct: High-Performance In-Memory Key-Value Store with Programmable NIC **SOSP'17**
Bojie Li, Zhenyuan Ruan, **Wencong Xiao**, Yuanwei Lu, Yongqiang Xiong, Andrew Putnam, Enhong Chen and Lintao Zhang

GraM: Scaling Graph Computation to the Trillions **SoCC'15**
Ming Wu, Fan Yang, Jilong Xue, **Wencong Xiao**, Youshan Miao, Lan Wei, Haoxiang Lin, Yafei Dai and Lidong Zhou

Journal paper

BeamRaster: A Practical Fast Massive MU-MIMO System with Pre-computed Precoders **TMC**
Meng Meng, **Wencong Xiao**, Tong He, Yuechen Tao, Kun Tan, Jiansong Zhang and Wenjie Wang

Research Experiences (selected)

GPU Cluster Resource Management for Deep Learning. 2017.1 – 2018.5
Supervised by Fan Yang and Lidong Zhou (*published in* **OSDI'18**)

Deep learning (DL) training is feedback-driven exploration with wide heterogeneity in terms of GPU usage. Gandiva is built to address such challenges in cluster management by leveraging the intra-job predictability feature with introspective scheduling. Achievement:

- Identifies three unique features in DL jobs: feedback-driven exploration in progress, performance heterogeneity in resource affinity, intra-job predictability in periodicity
- Co-designs scheduler and frameworks (e.g., Tensorflow) for introspective scheduling
- Proposes low-level primitives for DL scheduling: time-slicing, packing, migration, etc.
- Accelerates AutoML hyper-parameter exploration up to 13.6x and improves GPU cluster utilization by 26%

Distributed Graph Computation for Machine Learning. 2015.11 – 2016.7
Supervised by Ming Wu and Lidong Zhou (*published in* **NSDI'17**)

Machine learning (ML) algorithms (e.g., Logistic Regression) exhibit graph traversal patterns that naturally fit in graph engine. Tux^2 is built to leverage the elegance of graph engines in easy programming, structure-aware optimization, and great scalability, while maintain the ML features. Tux^2 achieves up to 10x performance speedup comparing with PowerLyra/PowerGraph and Petuum/ParameterServer. Tux^2 extends graph engine with innovations in three dimensions:

- Scheduling: stale synchronous parallel model for trade-off between convergence and efficiency
- Data representation: heterogeneous data model for flexible and efficient optimization
- Programming: a novel MEGA graph model to easily implement ML algorithms

High Performance Graph Computing over RDMA. 2014.9 – 2015.10
Supervised by Ming Wu (*published in* **SoCC'15**)

Developed GraM, an efficient and scalable graph engine for graph algorithms (e.g., PageRank). It scales up to multi-core while scales out in a cluster, significantly beating state-of-art graph engines often over an order of magnitude on typical graph algorithms. Besides, GraM is capable to process PageRank on a trillion-edge graph with 64 servers in 140 seconds, setting a new milestone for graph computing. GraM exploits the multi-core CPU architecture and RDMA-based NIC with key designs:

- Uses a unified message-passing model for both scale up and out
- Benefits from a special designed multi-core aware RDMA-based communication stack with computation and communication overlapping
- Adopts auto-adaptive configuration trade-off in scale cost and parallelism benefit

Awards

2018 Ph.D. National scholarship award.
OSDI'18 scholarship award.
SoCC'18 scholarship award.
2017 Microsoft research fellowship nomination award.
NSDI'17 scholarship award.
2016 Microsoft research rising star award.
2014 Outstanding undergraduate student award of Beijing China.

Skills

Programming	C++, Python, Java, C#, Bash, LaTeX.
System analysis	Performance tuning, outlier diagnostics, bottleneck investigation.
Open-source System	YARN, Kubernetes, Tensorflow, PyTorch, Spark, PowerGraph.