Jinliang Wei

GHC 7505, Carnegie Mellon University – 5000 Forbes Ave, Pittsburgh, PA 15213
www.cs.cmu.edu/~jinlianw jinlianw@cs.cmu.edu (765) 427-5257

SUMMARY & OBJECTIVE

Jinliang is a final-year Ph.D. student in Computer Science at Carnegie Mellon University. His doctoral research is focused on programmable and efficient distributed machine learning systems. He is looking for full-time software engineering and research positions in computer systems starting in Oct 2019, preferably related to his research interest.

EDUCATION

Carnegie Mellon University **Pittsburgh, PA**
Ph.D. Candidate in Computer Science *Aug 2012 – Sep 2019 (Expected)*
M.S. in Computer Science *Aug 2012 – Dec 2018*
Advisors: Garth A. Gibson, Eric P. Xing
Thesis: Efficient and Programmable Distributed Shared Memory Systems for Machine Learning Training

Purdue University **West Lafayette, IN**
B.S. in Computer Engineering, Minor in Mathematics with Distinction (GPA: 3.80/4.0) *Aug 2008 – Dec 2011*

ACADEMIC EXPERIENCE

Computer Science Department, Carnegie Mellon University **Pittsburgh, PA**
Graduate Research Assistant *Aug 2012 – Present*

Major thesis research projects are highlighted.

- **Dynamic Scheduling for Dynamic Control Flow in Deep Learning Systems**
  Proposed to dynamically schedule operators on distributed devices to reduce GPU memory consumption during neural network training by leveraging runtime information (e.g., which branch is actually executed). Implemented a tensor swapping mechanism that reduces the GPU memory consumption by up to 50% and enables training with 2× larger batch size, for models such as transformer.

  Publications: [LearningSys @ NeurIPS'18]

- **Automatic Parallelization of Imperative Machine Learning Programs for Distributed Training**
  Developed Orion that automatically parallelizes serial, imperative machine learning programs (in Julia) for distributed training. By leveraging compiler static dependence analysis, Orion parallelizes ML programs while preserving data dependence. Orion-parallelized programs achieve considerably faster convergence rate than manual data parallelism and comparable convergence to manual model parallelism while substantially reducing programmer effort.

  Publications: [EuroSys'19],[SysML'18]
  Code: github.com/jinliangwei/orion

- **Bösen Parameter Server for Data-Parallel Machine Learning Training**
  Led the development of Bösen and invented a network communication management mechanism that schedules network communication based on bandwidth availability and update value magnitude. Network management improves the convergence time of data-parallel training programs by $2-3\times$ compared to state-of-the-art bulk-synchronous-parallel and bounded-staleness systems.
  LightLDA which was developed based on Bösen trained a topic model with 1 trillion model parameters and 200 billion tokens. Caffe was integrated with Bösen for data-parallel neural network training. Bösen is a major pillar of Petuum, which led to a ML startup in 2016.

  Publications: [SoCC'15],[AAAI'15],[WWW'15],[KDD'15],[ATC'17]
  Code: github.com/jinliangwei/parameter_server

- **Benchmarking and Performance Optimization of Apache Spark for Machine Learning Training**
  Evaluated Apache Spark for machine learning model-parallel training with model parameters distributed as RDDs (as opposed to driver program local variables) and identified three major performance bottlenecks: 1) repeated shuffling of model parameters due to model parallelism; 2) marshalling overhead due to Python-JVM interfacing; and 3) scheduling overhead due to growing lineage caused by repeated model parameter updates. Mentored an M.S. student to address the growing lineage overhead.

- **LazyTable & IterStore Parameter Servers**
  Contributed to LazyTable and IterStore. IterStore achieved up to 10× faster computation throughput on machine learning applications and comparable performance on graph analytics applications, compared to state-of-the-art graph processing systems, such as PowerGraph, leveraging the iterative computation pattern.

  Publications: [SoCC'14][ATC'14]

School of Electrical and Computer Engineering, Purdue University **West Lafayette, IN**
Undergrad Research Assistant (w/ Prof. Sanjay Rao, Dr. Xin Sun) *May 2011 – March 2012*

- **Enterprise Routing Design Visualization**

Undergrad Research Assistant (w/ Prof. Vijay Raghunathan, Dr. Mohammad S. Hossain) *May 2010 – Aug 2010*

- **Wireless Sensor Network Interference Avoidance**

School of Mechanical Computer Engineering, Purdue University **West Lafayette, IN**
Undergraduate Student Programmer (w/ Prof. Carl Wassgren, Dr. Avik Sarkar) *Aug 2009 – May 2010*

- **Distributed Discrete Element Method (DEM) using MPI**

INDUSTRIAL EXPERIENCE

Research in Software Engineering (RiSE), Microsoft Research **Redmond, WA**
Research Intern w/ Saeed Maleki, Madan Musuvathi, Todd Mytkowicz *May 2017 – Aug 2017*
Distributed a parallel SGD algorithm (SymSGD) that retains SGD's sequential semantics; implemented optimizations that substantially reduce the computation and communication bottleneck due to combining parallel model copies by leveraging feature sparsity; our distributed logistic regression scales to 3TB of data on 20 machines and converges significantly faster than data parallelism.

HP Labs **Palo Alto, CA**
Research Associate on Software Research *May 2015 – Aug 2015*
Designed and implemented a distributed key-value store on non-volatile memory based on Masstree for The Machine.

Conviva Inc. **San Mateo, CA**
Software Engineering Intern *April 2012 – Aug 2012*

PUBLICATIONS

Priority-based Parameter Propagation for Distributed DNN Training [SysML'19]
Anand Jayarajan, **Jinliang Wei**, Garth A. Gibson, Alexandra Fedorova, Gennady Pekhimenko

Automating Dependence-Aware Parallelization of Machine Learning Training on Distributed Shared Memory [EuroSys'19]
Jinliang Wei, Garth A. Gibson, Phillip B. Gibbons, Eric P. Xing

Dynamic Scheduling for Dynamic Control Flow in Deep Learning Systems [LearningSys @ NeurIPS'18]
Jinliang Wei, Garth A. Gibson, Vijay Vasudevan, Eric P. Xing

Efficient and Programmable Machine Learning on Distributed Shared Memory via Static Analysis [SysML'18]
Jinliang Wei, Garth A. Gibson, Eric P. Xing

Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters [ATC'17]
Hao Zhang, Zeyu Zheng, Shizhen Xu, Wei Dai, Qirong Ho, Xiaodan Liang, Zhiting Hu, **Jinliang Wei**, Pengtao Xie, Eric P. Xing

Addressing the Straggler Problem for Iterative Convergent Parallel ML [SoCC'16]
Aaron Harlap, Henggang Cui, Wei Dai, **Jinliang Wei**, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing

Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics [SoCC'15, Best Paper Award]
Jinliang Wei, Wei Dai, Aurick Qiao, Henggang Cui, Qirong Ho, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing

Petuum: A New Platform for Distributed Machine Learning on Big Data [KDD'15]

Eric P. Xing, Qirong Ho, Wei Dai, Jin Kyu Kim, **Jinliang Wei**, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanu Kumar, Yaoliang Yu

LightLDA: Big Topic Models on Modest Compute Clusters [WWW'15]
Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, **Jinliang Wei**, Xun Zheng, Eric P. Xing, Tie-Yan Liu, Wei-Ying Ma

High-Performance Distributed ML at Scale through Parameter Server Consistency Models [AAAI'15]
Wei Dai, Abhimanu Kumar, **Jinliang Wei**, Qirong Ho, Garth A. Gibson, Eric P. Xing

Exploiting Iterative-ness for Parallel ML Computations [SoCC'14]
Henggang Cui, Alexey Tumanov, **Jinliang Wei**, Liangho Xu, Wei Dai, Jesse Haber-Kucharsky, Qirong Ho, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, Eric P. Xing

Exploiting Bounded Staleness to Speed Up Big Data Analytics [ATC'14]
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, Eric P. Xing

A Software Toolkit for Visualizing Enterprise Routing Design [SafeConfig'11]
Xin Sun, **Jinliang Wei**, Sanjay Rao, Geoffrey Xie

TEACHING EXPERIENCE (TEACHING ASSISTANT)

15-719 – Advanced Cloud Computing (CMU) **Spring 2017**
Instructors: Prof. Garth Gibson, Prof. Greg Ganger, Prof. Majd Sakr
Designed and led a course project that uses Apache Spark for ETL processing up to 1TB of Common Crawl data and for model-parallel training of collaborative filtering; led recitation.

15-415/615 – Database Applications (CMU) **Fall 2015**
Instructors: Christos Faloutsos, Andy Pavlo

ECE 207 – Electronic Measurement Techniques (Purdue) **Fall 2011**

ENGR 131/132 – Transforming Ideas to Innovation (Purdue) **Fall 2009 - Spring 2011**

CONFERENCE TALKS

Automating Dependence-Aware Parallelization of Machine Learning Training on Distributed Shared Memory **EuroSys 2019**

Managed Communication and Consistency for Fast Data-Parallel Iterative Analytics **SoCC 2015**

AWARDS & HONORS

Best Paper Award **SoCC 2015**

Student Travel Grant **SoCC 2015**

Student Travel Grant **SafeConfig 2011**

Charles W. Brown Scholarship **Purdue University, 08/2011 – 12/2011**

Eli Shay EE Scholarship **Purdue University, 08/2010 – 05/2011**

SKILLS

Programming C, C++, Julia, Python, Shell, Java
Big Data Systems TensorFlow, Apache Spark