Jiajia Li

High Performance Computing Group
Pacific Northwest National Laboratory (PNNL)
+1 (404) 940 4603 • Jiajia.Li@pnnl.gov • jiajiali.org

WORK EXPERIENCE

Pacific Northwest National Laboratory (PNNL)
Research Scientist *Aug 2018-Now*

IBM Thomas J. Watson Research Center
Research Intern, Mentor: Dr. Jee Choi and Dr. Dong Chen *May-Aug 2016*

Intel Parallel Computing Research Lab
Research Intern, Mentor: Dr. Mikhail Smelyanskiy *May-Jul 2015*

EDUCATION

Georgia Institute of Technology **High Performance Computing**
Ph.D., Advisor: Prof. Richard Vuduc *2013 – 2018*

University of Chinese Academy of Sciences (UCAS) **Computer Architecture**
Doctor of Engineering, Advisors: Prof. Mingyu Chen and Guangming Tan *2008 – 2013*

Dalian University of Technology **Information and Computing Science**
Bachelor of Sciences, As an Accelerated Student *2005 – 2008*

PUBLICATIONS

∘ **Jiajia Li**, Yuchen Ma, Xiaolong Wu, Ang Li, Kevin Barker. PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite. Technical Report. 2019. (Under review).

∘ **Jiajia Li**, Bora Ucar, Umit Catalyurek, Kevin Barker, Richard Vuduc. Efficient and Effective Sparse Tensor Reordering. Technical Report. 2019. (Under review).

∘ Ke Meng, **Jiajia Li**, Guangming Tan. A Pattern Based Algorithmic Autotuner for Graph Processing on GPUs. Principles and Practice of Parallel Programming. PPoPP'19. (Accepted, **Best Paper Award Finalist**)

∘ Jeffrey S. Young, Eric Hein, Srinivas Eswar, Patrick Lavin, **Jiajia Li**, Jason Riedy, Richard Vuduc, Thomas M. Conte. A Microbenchmark Characterization of the Emu Chick. Technical Report. 2019.

∘ Eric Hein, Srinivas Eswar, Abdurrahman Yasar, **Jiajia Li**, Jeffrey S. Young, Tom Conte, Umit V. Catalyurek, Rich Vuduc, Jason Riedy, Bora Ucar. Programming Strategies for Irregular Algorithms on the Emu Chick. Technical Report. 2019.

∘ **Jiajia Li**. Scalable Tensor Decompositions in High Performance Computing Environments. PhD Dissertation. Georgia Institute of Technology, Atlanta, GA, USA. July 2018.

∘ **Jiajia Li**, Jimeng Sun, Richard Vuduc. HiCOO: Hierarchical Storage of Sparse Tensors. ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis. SC'18. (**Best Student Paper Award**)

∘ Yuchen Ma, **Jiajia Li**, Xiaolong Wu, Chenggang Yan, Jimeng Sun, Richard Vuduc. Optimizing Sparse Tensor Times Matrix on GPUs. Journal of Parallel and Distributed Computing (Special Issue on Systems for Learning, Inferencing, and Discovering). 2018.

∘ Eric Hein, Tom Conte, Jeffrey Young, Srinivas Eswar, **Jiajia Li**, Patrick Lavin, Richard Vuduc, Jason

Riedy. An Initial Characterization of the Emu Chick. 2018 IEEE International Parallel and Distributed Processing Symposium Workshops. IPDPSW. 2018.

∘ Yue Zhao, **Jiajia Li**, Chunhua Liao, Xipeng Shen. Bridging the Gap between Deep Learning and Sparse Matrix Format Selection. 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. PPoPP'18.

∘ Guangming Tan, Junhong Liu, **Jiajia Li**. Design and Implementation of Adaptive SpMV Library for Multicore and Manycore Architecture. ACM Transactions on Mathematical Software. 2018.

∘ **Jiajia Li**, Jee Choi, Ioakeim Perros, Jimeng Sun, Richard Vuduc. Model-Driven Sparse CP Decomposition for Higher-Order Tensors. 31st IEEE International Parallel & Distributed Processing Symposium. IPDPS. 2017.

∘ Xiuxia Zhang, Guangming Tan, Shuangbai Xue, **Jiajia Li**, Keren Zhou, Mingyu Chen. Understanding the GPU Microarchitecture to Achieve Bare-Metal Performance Tuning. PPoPP'17.

∘ **Jiajia Li**, Yuchen Ma, Chenggang Yan, Richard Vuduc. Optimizing Sparse Tensor Times Matrix on multi-core and many-core architectures. The sixth Workshop on Irregular Applications: Architectures and Algorithms (IA^3), co-located with SC. 2016.

∘ **Jiajia Li**, Casey Battaglino, Ioakeim Perros, Jimeng Sun, Richard Vuduc. An Input-Adaptive and In-Place Approach to Dense Tensor-Times-Matrix Multiply. The International Conference for High Performance Computing, Networking, Storage and Analysis (SC) 2015

∘ Casey Battaglino, **Jiajia Li**, Ioakeim Perros, Jimeng Sun, Richard Vuduc. Tensors in Data Analysis: Methods, Applications, and Software. 2015. (To be submitted)

∘ **Jiajia Li**, Zhonghai Zhang, Guangming Tan, David Bader. SMAT: A Cross-Platform Input Adaptive Auto-Tuner for Sparse Matrix-Vector Multiplication. Technical Report. 2014

∘ **Jiajia Li**. Research on Sparse Matrix Vector Multiplication Auto-tuning Method. PhD Thesis. The University of Chinese Academy of Sciences, Beijing, China. 2013

∘ **Jiajia Li**, Guangming Tan, Mingyu Chen, Ninghui Sun. SMAT: An Input Adaptive Auto-Tuner for Sparse Matrix-Vector Multiplication. Programming Language Design and Implementation (PLDI) 2013

∘ **Jiajia Li**, Xingjian Li, Guangming Tan, Mingyu Chen, Ninghui Sun. An Optimized Large-Scale Hybrid DGEMM Design for CPUs and ATI GPUs. International Conference on Supercomputing (ICS) 2012

∘ **Jiajia Li**, Xiuxia Zhang, Guangming Tan, Mingyu Chen. The Study of Choosing the Best Storage Format of Sparse Matrix Vector Multiplication, Journal of Computer Research and Development. (IN CHINESE)

∘ **Jiajia Li**, Xiuxia Zhang, Guangming Tan, Mingyu Chen. Algebraic Multi-grid Optimization Study on GPU. HPC China (IN CHINESE) 2011

∘ **Jiajia Li**, Guangming Tan, Mingyu Chen. Automatically Tuned Dynamic Programming with an Algorithm-by-Blocks. 16th International Conference on Parallel and Distributed Systems (ICPADS) 2010

HONORS AND AWARDS

2018: ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC'18) Best Student Paper Award.

2018: SIAM ALA'18 Student Travel Grant.

2018: GaTech CoC Graduate Student Council Travel Grant.

2017: IBM PhD Fellowship for 2017-2018. [Link]

2017: Travel grant from ATIP Workshop, co-located with SC'17 [Link]

2017: Travel grant from IPAM for Big Data Meets Computation Workshop 2017 [Link]

2016: Selected students to attend IEEE-WIE Women's Leadership Summit 2016

2013: ZhuLiYueHua Award for the Excellent PhD Students of Chinese Academy of Sciences (Top 0.2%)

2011: Xia Peisu Scholarship of Institute of Computing Technology (Top 1%)

2011: Outstanding Research Assistant of the Computer Architecture Laboratory at UCAS

2010: Outstanding Student of the Computer Architecture Laboratory at UCAS

ACTIVITIES

Nov 2018: As a local chair of the 25th International European Conference on Parallel and Distributed Computing (Euro-Par'19)

Oct 2018: As the web chair of International Conference on Parallel Architectures and Compilation Techniques (PACT'19)

Sep 2018: As a co-chair of The First International Workshop on the Intersection of High Performance Computing and Machine Learning (HPCaML'19), held in conjunction with International Symposium on Code Generation and Optimization (CGO'19)

Aug 2018: As a co-organizer of SIAM Conference on Computational Science and Engineering (SIAM CSE'19) Minisymposium "High Performance Sparse Matrix, Tensor, and Graph Kernels"

Jun 2018: As an external PC member of The 32nd ACM International Conference on Supercomputing (ICS'18)

Oct 2017: As a PC member of Student Research Competition (SRC) of ASPLOS'18.

Aug 2017: As a Program Committee Member of IPDPS'18

2014-now: As a Reviewer of Parallel Computing Journal (PARCO), CCF Transactions on High Performance Computing (THPC), The Transactions on Parallel and Distributed Systems(TPDS), the Frontiers of Computer Science, IEEE Transactions on Neural Networks and Learning Systems(TNNLS), Algorithmica Journal, The 32nd ACM International Conference on Supercomputing (ICS'18), Journal of Low Power Electronics and Applications, Journal of Parallel and Distributed Computing (JPDC), The 47th International Conference on Parallel Processing (ICPP'18), the 21st IEEE International Conference on Parallel and Distributed Systems (ICPADS'15)

Aug 2014-now: As an organizer of Hot CSE seminar, a PhD academic seminar in GT CSE.

Nov 2013-now: As a Volunteer Librarian of Repetitive Stress Injury (RSI) Lending Library of GT College of Computing.

Spring 2017: As a Teach Assistant of "Intro to High-Performance Computing (OMSCS) (CSE 6220)"

Apr 2016: As a volunteer judge for Undergraduate Research Opportunities Program 11th Annual Undergraduate Research Spring Symposium.

Oct 2015: As a Reviewer of "The Transactions on Parallel and Distributed Systems"

Sep 2015: As a Reviewer of "The 21st IEEE International Conference on Parallel and Distributed Systems (ICPADS'15)"

Fall 2014: As a Teach Assistant of "High-Performance Computing: Tools and Applications (CSE 6230)"

Oct 2013, 2014, 2015: As a Volunteer Reviewer of "President's Undergraduate Research Awards (PUMA)" and "National Center for Women & IT (NCWIT) Award"

May 2012: As a Teach Assistant of "Parallel Computer Architecture" class of Dragonstar Project

2012: As a Teacher for One-Day Training of "Parallel Computing on GPU using CUDA" in Sun Yat-sen University

RESEARCH EXPERIENCE

Sparse Tensor Algorithms Optimization and Applications **08/2018 – now**

HPC Group, PNNL *Research Scientist*

- Optimize sparse tensor algorithms on new computer architectures, including Emu, Nvidia GPUs, and AMD GPUs, supported by "CENATE: The Center for Advanced Technology Evaluation" project.
- Build a sparse tensor operation benchmark suite, supported by CENATE project.
- Speedup sparse tensor kernels for quantum Chemistry, supported by NWChemEx project.

Optimizing Tensor Algorithms **01/2015 – 07/2018**

HPC Garage *Graduate Research Assistant*

- Propose a **sparse** tensor format (HiCOO) for tensor decomposition on multicore CPUs. (Published in SC'18.)
- Building a **sparse** tensor operation library (ParTI!) for tensor decompositions on multicore CPUs and GPUs with MATLAB interface. (Part of this work has been published in JPDC, IAˆ3 @ SC'16 and Tensor-Learn @ NIPS'16.)
- Implemented distributed sparse CP-APR algorithm for non-negative **sparse** tensors using MPI, achieved 90× speedup on 320 IBM Power8 cores. (Jointed work with IBM Data Centric Systems Research Group.)
- Optimizing matricized tensor times Khatri-Rao product (MTTKRP) and CP decomposition for high-order **sparse** tensors on multicore CPUs by proposing a novel memoization algorithm. (Published in IPDPS'17)
- Collecting sparse tensors from real-world applications to build a **sparse** tensor dataset [FROSTT], collaborating with UMN, IBM, and Intel.
- Built an input-adaptive and in-place approach to **dense** tensor-times-matrix multiply (InTensLi) by eliminating data transformation, and achieved 4-13× speedups over state-of-the-art libraries. (Published in SC'15)

SMAT: An Application- & Architecture-Aware Auto-tuner of SpMV **10/2011 – 12/2013**

High Performance Computer Research Center, HPC LAB *Graduate Research Assistant*

Compared with the functions in MKL library, SMAT ran faster by more than **3 times**.

- Extracted a set of parameters to represent SpMV's performance characteristics according to the observations of 2373 matrices in UF sparse matrix collection.
- Applied a machine learning method to formulate a decision tree prediction model to search for the optimal SpMV kernel.
- Provided an unified interface based on CSR format, when choosing from different formats and SpMV implementations.
- Achieved the performance up to 75 and 33 GFLOP/s in single- and double-precision respectively on Intel platform with 12 threads. Based on SMAT, AMG algorithm showed above 20% performance improvement.
- Extended the SMAT framework to GPU and Intel Xeon Phi platforms.

Algebraic Multigrid (AMG) Optimization on GPU **12/2010 – 05/2011**

CARCH and Institute of Applied Physics and Computational Mathematics, *Graduate Research Assistant*

AMG-GPU achieved ***2*× speedup** using Jacobi iterative method, compared with CPU version.

- Applied CSR-SpMV optimization on GPU to optimize SpMV kernels in AMG.
- Concluded that due to the diverse features of AMG on different grid levels, different optimization methods should be dynamicly applied on each level.

Hybrid DGEMM Design for a Heterogeneous CPU-GPU Architecture **04/2010 – 12/2010**

CARCH *Graduate Research Assistant*

Achieved ***844 GFLOPS*** with ***80%*** floating-point efficiency on a system comprising Intel CPU and ATI GPU.

- Developed a new software pipelining for the DGEMM kernel running on CPU-GPU heterogeneous architecture.
- Compared with AMD ACML-GPU library, the optimized DGEMM improved performance by more than 2×.
- Concluded that the major scaling bottleneck is resource contention, especially PCIe and host memory contention.

Dynamic Programming Auto-tuning **09/2009 – 04/2010**

CARCH *Graduate Research Assistant*

Achieved speedup as ***10*×** for scalar program and further ***4*× *(double) or 2*× *(single)*** for SIMDization program.

- Proposed an algorithm-by-blocks for dynamic programming.
- Built an automatically tuned system to optimize dynamic programming on general-purpose processors.

SKILLS

Platforms: Linux, Windows, Mac OS

Programming Language: C/C++, Python, Matlab, Shell, Scala, HTML

Parallel Programming: MPI, OpenMP, CUDA, Apache Spark

HOBBIES

Traveling, Singing, Billiards, Swimming, Running