

Leiming Yu

☎ (617) 515-1913 | ✉ leimingyu830@gmail.com | 🌐 <https://leimingyu.github.io>

Summary

I am an expert in General Purpose Computing on GPUs. I'm passionate about parallel computing. I have a solid background in computer architecture, machine learning, deep learning, and medical imaging. I am interested in expanding my expertise in software and hardware stacks on GPU design and development.

Experience

Qualcomm

Boxborough, MA

Staff Engineer (Graphics Research Team)

Sep 2022 – Present

- Explore architecture designs for the future Adreno GPUs

Analogic

Peabody, MA

Imaging Engineer

July 2019 – Sep 2022

- Accelerated Computed Tomography (CT) related algorithms on GPUs
 - Reduced the runtime of Metal Artifact Reduction by 12x
 - Optimized the runtime of extended Field of View by 70x
- Developed and deployed deep learning models to denoise Low Dose CT images
- Developed a CUDA stream implementation to run Correction and Reconstruction on a single GPU
- GPU Evaluation, Image Quality Evaluation, Operating System Hardening

Mitsubishi Electric Research Lab

Cambridge, MA

Internship

July 2016 - Dec 2016

- Optimized Model Predictive Control solvers on NVIDIA Jetson TX1
- Developed efficient SGEMV kernels that outperform cuBLAS by 2x
- Developed mpcCUDA, GPU-accelerated Model Predictive Control solvers in Matlab

Mathworks

Natick, MA

Internship

May 2012 - Aug 2012

- Accelerated PSK Demodulator/Modulator on GPU
- Accelerated LDPC Decoder for Large Parity Check Matrix on GPU
- Improved the parfor section in commViterbiSystemGPU demo
- Accelerated Turbodecoder on Matlab Distributed Computing Server

Skill Highlights

- Expertise in CUDA and OpenCL for General Purpose Computing on GPUs.
- Experience in C/C++/OpenMP/MPI
- Experience with scripting languages: Python, Shell, R
- Experience with DevOps: Docker

Peer Review

- Journal of Parallel and Distributed Computing (JPDC), 2019
- IEEE Transactions on Computers (TCSI), 2019
- Simulation Modelling Practice and Theory (Elsevier Journal), 2018
- Parallel, Distributed and Network-based Processing (PDP), 2016

Awards and Honors

- Best Poster: HPC Day 2017
- Best Poster: HPC Day 2016
- Student Travel Grant: IISWC 2017
- Student Travel Grant: PPoPP 2015
- GE Hackathon Challenge: *Most Robust* Winner 2017

Education

Northeastern University

Ph.D. in Computer Engineering

Boston, MA

Jan 2011 - April 2019

University of Bridgeport

M.S in Electrical Engineering

Bridgeport, CT

Jan 2008 - Dec 2010

Shanghai Maritime University

B.S. in Electrical Engineering

Shanghai, China

Sep 2002 - Sep 2006

Certificates

- Heterogeneous Parallel Programming
- Intro to Parallel Programming
- Introduction to TensorFlow for Artificial Intelligence, Machine Learning and Deep Learning

Talks and Presentations

- Poster on Neural Network Denoiser for MCX, SPIE Photonics West, 2019
- Poster on MCX denoising using neural networks, HPC Day, Northeastern University, 2018
- Poster on fast MCX for heterogeneous computing systems, COE PhD Research Expo, 2018
- Tutorial on Monte Carlo eXtreme (MCX) in OpenCL, MCX Workshop, 2017
- Poster Winner on Concurrent Kernel Execution, HPC Day, UMass Dartmouth, 2017
- GTC Talk, "Portable Performance for MCX in 3D Turbid Media for Single and Multiple GPUs", 2016
- Poster Winner on Monte Carlo eXtreme (MCX), HPC Day, UMass Dartmouth, 2016

Teaching Experience

- Invited lectures on GPU Programming for Philips (Andover, MA), 2017
- Lecturer for GPU Class, Northeastern University, 2015-2017
- Teaching Assistant for GPU Class, Northeastern University, 2013
- Teaching Assistant for Audio Processing Lab and Digital Processing Lab, UB, 2008-2010

Academia

September 2011-April 2019	Research Assistant	Northeastern University (College of Engineering)
September 2008-Spring 2010	Graduate Assistant	University of Bridgeport (School of Engineering)

January 2018-April 2019

- Machine-learning Based Interference-aware Scheduler for GPU Clusters
 - Automated interference feature selection using Principal Feature Analysis for GPU workloads
 - Proposed interference sensitivity analysis for concurrently scheduled workloads
 - Improved the First-Come-First-Serve policy by 16%
 - Achieved 10% better throughput than a state-of-art similarity-based scheduler
- Monte Carlo Photon Transportation Simulation Denoising using Neural Network Models
 - Developed a RNN to learn the noise and customized the U-Net model to learn the degradation contour
 - Improved the SNR by 20 dB over the noise-adaptive NLM filter for homogenous media simulation

January 2017-December 2017

- Model-based Concurrent Kernel Execution (CKE) Analysis for GPUs
 - Proposed a block size tuning scheme based on the similarity of GPU kernels
 - Developed GPU-based data transfer model, kernel execution model and resource contention model
 - Explored the design space of CKE within a single GPU context
 - Attained less than 12% CKE prediction error and a close-to-optimal dispatching solution

January 2015-December 2015

- Monte Carlo Photon Simulation (MCX) in OpenCL
 - Developed an algorithm to calculate the kernel thread number that ensures resources are fully occupied
 - Leveraged just-in-time compilation to reduce divergent branches for higher execution efficiency
 - Developed thread-level and device-level load-balancing strategies to fully utilize CPUs/GPUs

June 2013-December 2014

- GPU-accelerated Hidden Markov Model (HMM) for Speech Recognition
 - Explored the task-level and data-level parallelism in HMM
 - Applied advanced GPU programming features (HyperQ and cuBLAS) for acceleration
 - Achieved 9x speedup compared to an optimized CPU version using an NVIDIA GTX 680 GPU
- Parallel IIR on GPU
 - Explored data-level parallelism by decomposing an IIR filter into multiple second-order IIR filters
 - Utilized the SHFL instruction to perform the inter-thread operations and attained 2x speedup
- Fiber Scattering Simulation on a GPU Cluster
 - Developed an efficient single GPU implementation (memory system, math intrinsics, CKE)
 - Developed an efficient GPU cluster solution using MPI and GPU implementation
 - Achieved 28x speedup compared to the MPI + OpenMP solution

September 2012- June 2016

- Database Administrator for Puerto Rico Testsite for Exploring Contamination Threats (NIH)
 - Used EarthSoft EQuIS for data processing and Microsoft SQL server as the database engine
 - Built a web server using EQuIS Enterprise to automate data reporting for distributed users
 - Applied PCA, K-means, Hierarchical Clustering to identify biomarkers in urine data

September 2011-2012

- K-means Clustering for Spectrum Sensing on GPU/CPU
 - Implemented GPU-accelerated K-means to identify empty wireless band for Spectrum Sensing
 - Attained 70x speedup over the Matlab implementation

Fall 2008- 2010

- Teaching Assistant for Audio Processing Lab and Digital Signal Processing Lab

Publications

Journals and Book Chapters

[J8] Ardakani, Matin Raayai, Leiming Yu, David R. Kaeli, and Qianqian Fang. "Framework for denoising Monte Carlo photon transport simulations using deep learning." *Journal of Biomedical Optics* 27, no. 8 (2022).

[J7] Gong, Xun, Xiang Gong, Leiming Yu, and David Kaeli. "HAWS: Accelerating GPU Wavefront Execution through Selective Out-of-order Execution." *ACM Transactions on Architecture and Code Optimization (TACO)* 16, no. 2 (2019): 15.

[J6] Yaoshen Yuan, Leiming Yu, Zafer Doğan, and Qianqian Fang. "*Graphics Processing Units-accelerated Adaptive Nonlocal Means Filter for Denoising Three-dimensional Monte Carlo Photon Transport Simulations.*" *Journal of biomedical optics* 23, no. 12 (2018): 121618.

[J5] Leiming Yu, Fanny Nina-Paravecino, David Kaeli, and Qianqian Fang. "*Scalable and Massively Parallel Monte Carlo Photon Transport Simulations for Heterogeneous Computing Platforms.*" *Journal of biomedical optics* 23, no. 1 (2018): 010504.

[J4] Fanny Nina-Paravecino, Leiming Yu, Qianqian Fang, and David Kaeli. "*High-performance Monte Carlo Simulations for Photon Migration and Applications in Optical Brain Functional Imaging.*" In *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, pp. 67-85. Springer, Cham, 2017.

[J3] Yan Zhang, Hideyo Inouye, Michael Crowley, Leiming Yu, David Kaeli, and Lee Makowski. "*Diffraction Pattern Simulation of Cellulose Fibrils Using Distributed and Quantized Pair Distances.*" *Journal of Applied Crystallography* 49, no. 6 (2016): 2244-2248.

[J2] Xiangyu Li, Leiming Yu, David Kaeli, Yuanyuan Yao, Pogu Wang, Roger Giese, Vicent Yusa and Akram Alshawabkeh, "*A Framework for Big Metabolomic Data Management and Analysis*", *IARIA Journal*. Vol 9, 2016.

[J1] Xiangyu Li, Leiming Yu, David Kaeli, Yuanyuan Yao, Pogu Wang, Roger Giese, and Akram Alshawabkeh. "Big Data Analysis on Puerto Rico Testsite for Exploring Contamination Threats." ALLDATA 2015 (2015): 36.

Conference Proceedings

[C10] Zamyatin, Alexander A., Leiming Yu, and David Rozas. "3D residual convolutional neural network for low dose CT denoising." In Medical Imaging 2022: Physics of Medical Imaging, vol. 12031, SPIE, 2022.

[C9] Dong, Shi, Zlatan Feric, Leiming Yu, David Kaeli, John Meeker, Ingrid Y. Padilla, Jose Cordero, Carmen Velez Vega, Zaira Rosario, and Akram Alshawabkeh. "An Efficient Data Management Framework for Puerto Rico Testsite for Exploring Contamination Threats (PROTECT)." In 2018 IEEE International Conference on Big Data (Big Data), pp. 5316-5318. IEEE, 2018.

[C8] Leiming Yu, Fanny Nina-Paravecino, David Kaeli, and Qianqian Fang. "Fast Monte Carlo Photon Transport Simulations for Heterogeneous Computing Systems." In Clinical and Translational Biophotonics, pp. JTh3A-38. Optical Society of America, 2018.

[C7] Yaoshen Yuan, Leiming Yu, and Qianqian Fang. "Denoising in Monte Carlo Photon Transport Simulation Using GPU-accelerated Adaptive Non-Local Mean Filter." In Optical Tomography and Spectroscopy, pp. JTh3A-41. Optical Society of America, 2018.

[C6] Leiming Yu, Xun Gong, Yifan Sun, Qianqian Fang, Norm Rubin, and David Kaeli. "Moka: Model-based Concurrent Kernel Analysis." In 2017 IEEE International Symposium on Workload Characterization (IISWC), pp. 197-206. IEEE, 2017.

[C5] Yifan Sun, Xiang Gong, Amir Kavyan Ziabari, Leiming Yu, Xiangyu Li, Saoni Mukherjee, Carter McCardwell, Alejandro Villegas, and David Kaeli. "Hetero-mark, A Benchmark Suite for CPU-GPU Collaborative Computing." In 2016 IEEE International Symposium on Workload Characterization (IISWC), pp. 1-10. IEEE, 2016.

[C4] Yash Ukidave, Fanny Nina-Paravecino, Leiming Yu, Charu Kalra, Amir Momeni, Zhongliang Chen, Nick Materise, Brett Daley, Perhaad Mistry, and David Kaeli. "Nupar: A Benchmark Suite for Modern GPU Architectures." In Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering, pp. 253-264. ACM, 2015.

[C3] Yan Zhang, Leiming Yu, David Kaeli, and Lee Makowski. "Fast Simulation of X-ray Diffraction Patterns from Cellulose Fibrils Using GPUs." In Bioengineering Conference (NEBEC), 2014 40th Annual Northeast, pp. 1-2. IEEE, 2014.

[C2] Leiming Yu and B.D. Barkana. "Speech Disorders: An Analysis of Hypernasal Speech Using Signal Processing Techniques", Proceedings of the 2009 ASEE NE American Society for Engineering Education Conference, April 3-4, 2009.

[C1] Leiming Yu and B.D. Barkana. "Classifying Hypernasality Using the Pitch and Formants", Proceedings of the 6th International Conference on Information Technology – New Generations, ITNG 2009.

Workshop Proceedings

[W6] Leiming Yu, Abraham Goldsmith, and Stefano Di Cairano. "Efficient Convex Optimization on GPUs for Embedded Model Predictive Control." In Proceedings of the General Purpose GPUs, pp. 12-21. ACM, 2017.

[W5] Patrick Reilly, Leiming Yu, and David Kaeli. "Accelerating Machine Learning Algorithms in Python", Boston Area Architecture Workshop, 2017.

[W4] Saoni Mukherjee, Xiang Gong, Leiming Yu, Carter McCardwell, Yash Ukidave, Tuan Dao, Fanny Nina Paravecino, and David Kaeli. "Exploring The Features of OpenCL 2.0." In Proceedings of the 3rd International Workshop on OpenCL, p. 5. ACM, 2015.

[W3] Leiming Yu, Yan Zhang, Xiang Gong, Nilay Roy, Lee Makowski, and David Kaeli. "*High Performance Computing of Fiber Scattering Simulation*." In Proceedings of the 8th Workshop on General Purpose Processing using GPUs, pp. 90-98. ACM, 2015.

[W2] Leiming Yu, John Magrath, Ajey Pandey, Matthew Sears, and David Kaeli. "*Speech Recognition on Modern Graphic Processing Units*", Proceedings of the 6th Annual Boston Area Architecture Workshop.2015.

[W1] Leiming Yu, Ukidave Yash, and David Kaeli. "*GPU-accelerated HMM for Speech Recognition*." In Parallel Processing Workshops (ICCPW), 2014 43rd International Conference on, pp. 395-402. IEEE, 2014.