

# Cheng Li

1308 W. Main St., Urbana, Illinois 61801

cli99.netlify.com

cli99@illinois.edu

github.com/rai-project

4194967797

## RESEARCH INTEREST

My research lies in the field of GPU-accelerated applications, with an emphasis on Deep Learning (DL). My work has focused on understanding and optimizing Deep Learning workloads. In the process, I have developed several open-source tools to benchmark, profile, and summarize Deep Learning training and inference across hardware and software stacks.

## EDUCATION

### University of Illinois Urbana-Champaign

*Ph.D. in Computer Science*

GPA: 3.95/4.0

Champaign, IL

*Expected August 2020*

Thesis: Performance Benchmarking, Analysis and Optimization of Deep Learning Inference

### University of Michigan

*M.S. in Computer Science and Engineering*

GPA: 3.96/4.0

Ann Arbor, MI

*May 2015*

### Shanghai Jiao Tong University

*B.S. in Electrical Engineering*

GPA: 3.85/4.0

Shanghai, China

*August 2013*

### University of Michigan

*B.S. in Computer Engineering*

GPA: 3.63/4.0

Ann Arbor, MI

*May 2013*

## WORK EXPERIENCE

### Alibaba Group

*Research Intern*

Sunnyvale, CA

*May - August 2019*

- Extended MLModelScope with automatic cross-stack analysis capability.
- Used MLModelScope to benchmark and characterize public, MLPerf and AI Matrix models across systems of interest.
- Performed model/framework/system advising using the data collected, and explore its applicability in the Alibaba Cloud.

### IBM Thomas J. Watson Research Center

*Research Intern*

Yorktown Heights, NY

*May - August 2018*

- Evaluated existing techniques for Deep Learning performance estimation on different models and systems, and understood the sources of inaccuracy.
- Developed an analysis tool that generates layer benchmarks, finds patterns within models, and performs performance prediction for Deep Learning models across hardware.

### 9th Programming and Tuning Massively Parallel Systems and AI School

*Teaching Assistant*

Barcelona, Spain

*July 2018*

- Designed GPU labs and projects for the summer school students.
- Advised the students during the summer school's hackathon.

### IBM Thomas J. Watson Research Center

*Research Intern*

Yorktown Heights, NY

*May - August 2017*

- Developed **MLModelScope** a hardware/software agnostic platform for consistent benchmarking and analysis of Deep Learning inference at scale.
- Profiled and optimized the GPU-accelerated alternating least square(ALS) algorithm for Matrix Factorization.

### University of Illinois Urbana-Champaign

*Lead Teaching Assistant for CS483 - Applied Parallel Programming*

Champaign, IL

*August - December 2016*

- Designed GPU labs, exams, and projects for a class of 200 students. Maintained the assignment and the project submission systems - **WebGPU** and **RAI**.

## RECENT PROJECTS

---

### XSP: Understanding DL Performance

- XSP is an across-stack profiling design that innovatively leverages distributed tracing to aggregate profile data from different profiling sources and construct a holistic and hierarchical view of DL model execution.
- XSP introduces a leveled and iterative measurement approach that accurately captures the latencies at all levels of the HW/SW stack despite the profiling overhead.
- We implement the design for GPUs and couple it with an automated analysis pipeline that enables systematic characterization and comparison.

### Benanz: DL Optimization Advising

- We propose a “lower-bound” latency metric for DL models on GPUs based on the observation that the latency of a DL model is bounded by the latencies of the cuDNN and cuBLAS API calls corresponding to the model layers.
- Benanza is a sustainable and extensible benchmarking and analysis design that automatically generates micro-benchmarks given a set of models, computes their “lower-bound” latencies using the benchmark data, and informs optimizations of their execution on GPUs.

### MLModelScope: DL Benchmarking

- MLModelScope is a framework- and hardware-agnostic distributed platform for benchmarking and profiling DL models across datasets/frameworks/systems.
- MLModelScope proposes a specification to define DL model evaluations and techniques to provision the evaluation workflow using the user-specified HW/SW stack.
- MLModelScope is implemented as an open-source project with support for all major frameworks and hardware architectures.

### DLBricks: Reducing DL Benchmarking Effort

- DLBricks is a composable benchmark generation design that reduces the effort of developing, maintaining, and running DL benchmarks on CPUs.
- DLBricks decomposes DL models into a set of unique runnable networks and constructs the original model's performance using the performance of the generated benchmarks.

### TrIMS: Transparent and Isolated Model Sharing for DL Inference

- TrIMS is a generic memory sharing technique that enables constant data to be shared across processes or containers while still maintaining isolation between users.
- TrIMS mitigates the DL model loading overhead and increases the hardware resource utilization in inference by sharing models across all levels of the memory hierarchy in the cloud environment — GPU, CPU, local storage, and remote storage.

### TOPS: Accelerating Reduction and Scan Using Tensor Core Units

- TOPS is a library of collectives expressed as matrix multiplication operations on Tensor Cores Units (TCU, specialized hardware for matrix multiplication).
- It is the first to broaden the class of algorithms expressible as TCU operations and show benefits of the mapping in terms of program simplicity, efficiency, and performance.
- We implemented reduction and scan using NVIDIA V100 Tensor Cores and achieved up to 100× and 3× speedup compared to state-of-the-art methods while decreasing the power consumption by up to 22% and 16% correspondingly.

## PUBLICATIONS

---

1. **The Design and Implementation of a Scalable DL Benchmarking Platform** (IEEE CLOUD'20)  
*Cheng Li\*, Abdul Dakkak\*, Jinjun Xiong, Wen-Mei Hwu*
2. **DLSpec: A Deep Learning Task Exchange Specification** (USENIX OpML'20)  
*Cheng Li\*, Abdul Dakkak\*, Jinjun Xiong, Wen-Mei Hwu*
3. **XSP: Across-Stack Profiling and Analysis of Machine Learning Models on GPUs** (IPDPS'20, Best Paper Nomination)  
*Cheng Li\*, Abdul Dakkak\*, Jinjun Xiong, Wei Wei, Lingjie Xu, Wen-Mei Hwu*
4. **Benanza: Automatic uBenchmark Generation to Compute "Lower-bound" Latency and Inform Optimizations of Deep Learning Models on GPUs** (IPDPS'20)  
*Cheng Li\*, Abdul Dakkak\*, Jinjun Xiong, Wen-Mei Hwu*
5. **DLBricks: Composable Benchmark Generation to Reduce Deep Learning Benchmarking Effort on CPUs** (ICPE'20)  
*Cheng Li, Abdul Dakkak, Jinjun Xiong, Wen-Mei Hwu*
6. **AI Matrix: A Deep Learning Benchmark for Alibaba Data Centers** (arXiv'19)  
*Wei Zhang, Wei Wei, Lingjie Xu, Lingling Jin, Cheng Li*
7. **MLModelScope: Evaluate and Introspect Cognitive Pipelines** (IEEE Services'19)  
*Cheng Li, Abdul Dakkak, Jinjun Xiong, Wen-Mei Hwu*
8. **TrIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference in Function as a Service Environments** (IEEE CLOUD'19)  
*Abdul Dakkak, Cheng Li, Simon Garcia de Gonzalo, Jinjun Xiong, Wen-Mei Hwu*
9. **Accelerating Reduction and Scan Using Tensor Core Units** (ICS'19)

Abdul Dakkak, **Cheng Li**, Jinjun Xiong, Isaac Gelado, Wen-Mei Hwu

10. **Evaluating Characteristics of CUDA Communication Primitives on High-Bandwidth Interconnects (ICPE'19, Best Paper)**  
Carl Pearson, Abdul Dakkak, Sarah Hashash, **Cheng Li**, I-Hsin Chung, Jinjun Xiong, Wen-Mei Hwu
11. **Accelerating Reduction Using Tensor Core Units** (HPCaML'19)  
Abdul Dakkak, **Cheng Li**, Jinjun Xiong, Wen-Mei Hwu
12. **SCOPE: C3SR Systems Characterization and Benchmarking Framework** (arXiv'18)  
Carl Pearson, Abdul Dakkak, **Cheng Li**, Sarah Hashash, Jinjun Xiong, Wen-mei Hwu
13. **Matrix Factorization on GPUs with Memory Optimization and Approximate Computing** (ICPP'18)  
Wei Tan, Shiyu Chang, Liana Fong, **Cheng Li**, Zijun Wang, LiangLiang Cao
14. **RAI: A Scalable Project Submission System for Parallel Programming Courses** (IPDPSW'17)  
Abdul Dakkak, Carl Pearson, **Cheng Li**, Wen-mei Hwu
15. **KLAP: Kernel Launch Aggregation and Promotion for Optimizing Dynamic Parallelism** (MICRO'16)  
Izzat El Hajj, Juan Gomez-Luna, **Cheng Li**, Li-Wen Chang, Dejan Milojicic, Wen-mei Hwu
16. **DjiNN and Tonic: DNN as a Service and Its Implications for Future Warehouse Scale Computers** (ISCA'15)  
Johann Hauswald, Yiping Kang, Michael A. Laurenzano, Quan Chen, **Cheng Li**, Trevor Mudge, Ronald G. Dreslinski, Jason Mars, Lingjia Tang
17. **Sirius: An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers** (ASPLOS'15)  
Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang, **Cheng Li**, Austin Rovinski, Arjun Khurana, Ronald G. Dreslinski, Trevor Mudge, Vinicius Petrucci1, Lingjia Tang, Jason Mars
18. **Stochastic circuits for real-time image-processing applications** (DAC'13)  
Armin Alaghi, **Cheng Li**, John P. Hayes

## TALKS & POSTERS

---

### Super Computing 2019

Across-stack Profiling and Analysis of ML Models on GPUs

Denver, CO

November 18, 2019

### Tutorial at IISWC 2019

Challenges and Solutions for End-to-End and Across Stack ML Benchmarking

Orlando, FL

November 3, 2019

### HotChips 2019

MLModelScope: Evaluate and Profile ML Models at Scale and Across Stack

Palo Alto, CA

August 18, 2019

### Tutorial at ISCA 2019

Benchmarking Deep Learning Systems

Phoenix, AZ

June 22, 2019

### Tutorial at ASPLOS 2019

Benchmarking Deep Learning Systems

Providence, RI

April 14, 2019

### NVIDIA GPU Technology Conference 2019

TOPS, TRIMS and MLModelScope

San Jose, CA

March 22, 2019

### Super Computing 2018

MLModelScope

Dallas, TX

November 11, 2018

### IBM AI Research Week 2018

MLModelScope

Boston, MA

October 11, 2018

### NVIDIA GPU Technology Conference 2017

RAI: A Scalable Submission System for GPU Applications

San Jose, CA

March 22, 2017

## PROGRAMMING LANGUAGES

---

C/C++, Go, CUDA, Python, JavaScript, Bash, Mathematica

## MEMBERSHIP

---

IEEE, ACM, CRA-W (Computing Research Association-Women), WCS (Women in Computer Science)