Zhuoran Zhao

Email : zhuoran@utexas.edu
Mobile : +1-512-751-1819
Personal Website: https://zoranzhao.github.io
GitHub: https://github.com/zoranzhao
LinkedIn: https://www.linkedin.com/in/zoranzhao

Summary

My current research interests mainly include Machine Learning (ML) compiler, ML inference runtime and software/hardware co-design for high-concurrency ML serving systems. During my PhD, I spent most of my time in the area of electronic system-level (ESL) design and modeling, mainly focusing on distributed runtime/middleware and system-level performance modeling for edge computing systems.

Skills

- **Programming languages**: C/C++, Python
- **Tools and frameworks**: PyTorch, TorchInductor, Triton, Apache Thrift etc.
- 5-year industry project experiences on large-scale distributed recommender systems, ML inference runtime/compiler and GPU performance optimization

Education

- **Ph.D. in Electrical and Computer Engineering;** Dec. 2014 – May 2019
  University of Texas at Austin; *Austin, Texas*
  Advisor: Prof. Andreas Gerstlauer
- **M.S. in Electrical and Computer Engineering;** Aug. 2012 – Dec. 2014
  University of Texas at Austin; GPA: 3.93/4.00 *Austin, Texas*
- **B.S. in Electrical Engineering;** Sep. 2008 – Jun. 2012
  Zhejiang University; GPA: 3.95/4.00 *Zhejiang, China*
  Honored Minor: Advanced Honor Class of Engineering Education (ACEE)

Full-time Experience

- **Facebook** Menlo Park, CA
  Staff Research Scientist *Oct. 2019 - Present*
  - **Ads ML Ranking Infrastructure (Oct. 2019 - Jun. 2023)**: Tech Lead on PyTorch/GPU enablement and massive adoption for Ads ML ranking models
  - **PyTorch Accelerator Enablement (Jun. 2023 - Present)**: Working in the domain of PyTorch GPU inference runtime and compiler, enabling TorchInductor on large-scale production ranking models, enabling ahead-of-time TorchInductor (AOTInductor) on AMD GPU for both Meta internal workloads and open-source communities

- **University of Texas at Austin** Austin, Texas
  Graduate Research Assistant/Teaching Assistant *Aug. 2012 - May 2019*
  - Research project focusing on a portable and lightweight runtime framework for locally distributed CNN/DNN inference in resource-constrained IoT edge clusters, developed in C [2].
  - Research project focusing on a source-level network/system co-simulation framework for distributed embedded/mobile computing cluster prototyping, developed in C++ with LLVM, OMNeT++ and SystemC framework [1].
  - Teaching Assistant: EE382N Embedded System Design and Modeling (Fall 2015), EE319K Introduction to Embedded Systems (Fall 2012)

Selected Publications

[1] Zhuoran Zhao, K. Mirzazad and A. Gerstlauer, **"Network-level Design Space Exploration of Resource-constrained Networks-of-Systems,"** *ACM Transactions on Embedded Computing Systems (TECS), 2020.*

[2] Zhuoran Zhao, K. Mirzazad and A. Gerstlauer, **"DeepThings: Distributed Adaptive Deep Learning Inference on Resource-Constrained IoT Edge Clusters,"** *CODES+ISSS, special issue of IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2018.*

[3] Zhuoran Zhao, V. Tsoutsouras, D. Soudris, A. Gerstlauer, **"Network/System Co-Simulation for Design Space Exploration of IoT Applications,"** *Proceedings of the International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS), 2017.*

[4] Zhuoran Zhao, A. Gerstlauer and Lizy K. John, **"Source-Level Performance, Energy, Reliability, Power and Thermal (PERPT) Simulation,"** *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2017.*

[5] Zhuoran Zhao, D. Lee and A. Gerstlauer, **"Host-Compiled Reliability Modeling for Fast Estimation of Architectural Vulnerabilities,"** *In Silicon Errors in Logic, System Effects Workshop (SELSE), 2015*

[6] S. Chakravarty, Zhuoran Zhao, A. Gerstlauer, **"Automated, Retargetable Back-Annotation for Host-Compiled Performance and Power Modeling,"** *Proceedings of the IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2013.*

[7] L. Guckert, M. O'Connor, S. K. Ravindranath, Zhuoran Zhao and V. J. Reddi, **"A Case for Persistent Caching of Compiled JavaScript Code in Mobile Web Browsers,"** *In Workshop On Architectural And Microarchitectural Support For Binary Translation (AMAS-BT), 2013*

Professional Service

- **Technical Program Committee (TPC) Member:**
  - Design Automation Conference (DAC) 2021 (Session manuscript reviewer)
  - Design Automation Conference (DAC) 2022, 2023 (Session manuscript reviewer and presentation co-chair)
- **Reviewer:**
  - IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
  - IEEE Transactions on Parallel and Distributed Systems (TPDS)
  - IEEE Internet of Things Journal (IoT-J)
  - Design, Automation and Test in Europe (DATE) Conference, 2018
  - IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2018
- **Teaching:**
  - Teaching Assistant: EE382N Embedded System Design and Modeling, 2016
  - Teaching Assistant: EE319K Introduction to Embedded System, 2012

Honors and Awards

- Best in Session Award for the presentation "Automated, Retargetable Back-Annotation for Host-Compiled Power and Performance Modeling," in Semiconductor Research Corporation (SRC) TECHCON, Sep 11, 2013
- National Scholarship in China (2%), 2009, 2010