NEW YORK UNIVERSITY

Jeff (Jun) ZHANG *Curriculum vitae*

CONTACT INFORMATION

NYU Center for Cyber Security
370 Jay St, 10th Floor
Brooklyn, 11201, NY

Mobile: +1-516-697-9088
E-mail: jeffjunzhang@nyu.edu
Google Scholar

RESEARCH INTERESTS

Deep Learning, Computer Architecture, EDA, Embedded and Real-Time System

EDUCATION

New York University, New York

Ph.D. Candidate, Electrical and Computer Engineering, July 2020

- Adviser: **Prof. Siddharth Garg**
- Thesis Topic: Energy-Efficient and Reliable Deep Learning Acceleration.

Hunan University, Changsha, China

M.Eng., 2013, B.Eng., 2011,

- *Magna Cum Laude*, with Honors in Engineering.

INDUSTRY EXPERIENCE

Microsoft Research, Redmond, WA
Research Intern, Hololens AI Hardware Team **August 2018 to November 2018**

- Proposed/Optimized Architectures for Energy-Efficient Deep Learning Accelerators.
- Mentors: Dr. Shuayb Zarar, Dr. Amol Ambardekar.

Samsung Semiconductor Inc., San Jose, CA
Research Intern, Memory Platform Lab **May 2018 to August 2018**

- Applied Reinforcement Learning to Storage (SSD) and I/O Management for Datacenter Performance and Ecosystem.
- Mentors: Vijay Balakrishnan, Dr. Zvika Guz.

AWARDS AND HONORS

ACM SIGDA/IEEE CEDA DATE PhD Forum
- Shortlist for Best Presentation Award, 2020

TTTC's E.J. McCluskey Doctoral Thesis Competition
- Semifinalists at IEEE VLSI Test Symposium, 2020

IEEE VLSI Test Symposium
- Best Paper Award Nomination, 2018

New York University
- Ernst Weber Fellowship, 2015, 2016

Ministry of Education of China
- National Scholarship for graduate students (top 1%), 2012
- National Scholarship (top 2%), 2008, 2010

Hunan University
- Excellent Graduate Student, 2012
- Hunan University Fellowships for Master's Studies (top 10%), 2011
- Outstanding Graduates of Hunan Province (top 1%), 2011
- Pacemaker to Merit Student (top 0.1%), **HIGHEST honor**, 2009
- The First Class Scholarship (top 5%), 2009
- Merit Student, Excellent Student Cadre (top 4%), 2008, 2010, 2011

ONGOING PROJECTS

Leveraging model diversity for high QoS deep learning serving in the clouds.
Collaboration with Microsoft Research September 2019 to Now

Action-perception loops over 5G millimeter wave wireless for cooperative manipulation.
Collaboration with OPPO Research, NYU Wireless November 2019 to Now

CONFERENCE PROCEEDINGS

C.11 Zhang, J., Elnikety, S., Zarar, S., Gupta, A., and Garg, S. *Model-Switching: Dealing with Fluctuating Workloads in Machine-Learning-as-a-Service Systems.* 12th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud'20). Boston. July., 2020. (Acceptance rate: 22/95=23%)

C.10 Zhang, J., Raj, P., Zarar, S., Ambardekar, A., and Garg, S. *CompAct: On-chip Compression of Activations for Low Power Systolic Array Based CNN Acceleration.* ACM International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES) in conjunction with (ESWEEK 2019). New York. Oct., 2019. (Acceptance rate: 16/75=21%)
Also appears at ACM Transactions on Embedded Computing Systems (TECS).

C.9 Zhang, J., Liu, K., Khalid, F., Hanif, M., Rehman, S., Theocharides, T., Artussi, A., Shafique, M., and Garg, S. *Building Robust Machine Learning Systems: Current Progress, Research Challenges, and Opportunities.* ACM/IEEE 56th Design Automation Conference (DAC 2019, Special Session). Las Vegas. Jun., 2019.

C.8 Zhang, J., Garg, S. *FATE: Fast and Accurate Timing Error Prediction Framework for Low Power DNN Accelerator Design.* ACM/IEEE 37th Internation Conference on Computer Aided Design (ICCAD 2018). San Diego. Nov., 2018. (Acceptance rate: 98/396 = 24.7%)

C.7 Zhang, J., Rangineni, K., Ghodsi, Z., and Garg, S. *ThUnderVolt: Enabling Aggressive Voltage Underscaling and Timing Error Resilience for Energy Efficient Deep Neural Network Accelerators.* ACM/IEEE 55th Design Automation Conference (DAC 2018). San Francisco. Jun., 2018. (Acceptance rate: 168/691=24.3%) **(Citations: 32)**

C.6 Zhang, J., Gu, T., Basu., K., and Garg, S. *Analyzing and Mitigating the Impact of Permanent Faults on a Systolic Array Based Neural Network Accelerator.* IEEE VLSI Test Symposium (VTS 2018) . San Francisco. Apr., 2018. **Best Paper Award Nomination (Citations: 20)**

C.5 Zhang, J., Garg, S. *BandiTS: Dynamic Timing Speculation Using Multi-Armed Bandit Based Optimization.* ACM/IEEE 20th Design, Automation and Test in Europe (DATE 2017). Lausanne, Switzerland. Mar., 2017. (Acceptance rate: 24%)

C.4 Yasin, A., Zhang, J., Chen, H., Garg, S., Roy, S., and Chakrabory, K. *Synergistic Timing Speculation for Multi-threaded Programs.* ACM/IEEE 53th Design Automation Conference (DAC 2016). Austin. Jun., 2016. (Acceptance rate: 152/876=17%)

C.3 Cui, X., Zhang, J., Wu, K., and Sha, E. *Efficient Feasibility Analysis of DAG Scheduling with Real-Time Constraints in the Presence of Faults.* IEEE 19th Asia and South Pacific Design Automation Conference (ASP-DAC 2014). Singapore. Jan., 2014.

C.2 Zhang, J., Sha, E., Zhuge, Q., Yi, J., and Wu, K. *Efficient Fault-Tolerant Scheduling on Multiprocessor Systems via Replication and Deallocation.* IEEE/IFIP 11th International Conference on Embedded and Ubiquitous Computing (EUC2013). Zhangjiajie, China. Nov., 2013. **Distinguished Paper**
Also appears at International Journal of Embedded Systems (IJES).

C.1 Zhang, J., Deng, T., Gao, Q., Zhuge, Q., and Sha, E. *Optimizing Data Allocation for Loops on Embedded Systems with Scratch-Pad Memory.* IEEE 18th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2012). Seoul. Aug., 2012.

JOURNAL PUBLICATIONS

J.9 Zhang, J., Raj, P., Zarar, S., Ambardekar, A., and Garg, S. *CompAct: On-chip Compression of Activations for Low Power Systolic Array Based CNN Acceleration.* ACM Transactions on Embedded Computing Systems (TECS), Special Issue on Papers from ESWeek 2019.

J.8 Zhang, J., Ghodsi, Z. and Garg, S. *Enabling Timing Error Resilience for Low-Power Systolic-Array Based Deep Learning Accelerators.* IEEE Design & Test, Special Issue on Robust and Resource-Constrained ML. 2019.

J.7 Zhang, J., Basu, K. and Garg, S. *Fault-Tolerant Systolic Array Based Accelerators for Deep Neural Network Execution.* IEEE Design & Test. 2019.

J.6 Cui, X., Zhang, J., Wu, K., Garg, S. and Karri, R. *Split Manufacturing Based Register Transfer Level Obfuscation.* ACM Journal on Emerging Technologies in Computing. 2019.

J.5 Wang, Y., Li, K., Zhang, J., and Li, K. *Energy Optimization for Data Allocation with Hybrid SRAM+ NVM SPM.* IEEE Transactions on Circuits and Systems I: Regular Papers. 2017. **(Citations: 15)**

J.4 Sha, E., Wang, L., Zhuge, Q., Zhang, J., and Liu, J. *Power Efficiency for Hardware/software Partitioning with Time and Area Constraints on MPSoCs.* International Journal of Parallel Programming (IJPP). Special Issue on Top Papers from IFIP 10th Network and Parallel Computing. 2015. Springer. **(Citations: 26)**

J.3 Peng, S., Ouyang, A., Zhang, J.. *An Adaptive Invasive Weed Optimization Algorithm.* International Journal of Pattern Recognition and Artificial Intelligence. 2015.

J.2 Zhang, J., Sha, E., Zhuge, Q., Yi, J., and Wu, K. *Efficient Fault-Tolerant Scheduling on Multiprocessor Systems via Replication and Deallocation.* International Journal of Embedded Systems (IJES). Distinguish paper from IEEE 10th Embedded and Ubiquitous Computing. 2014. **(Citations: 16)**

J.1 Zhang, J., Deng, T., Gao, Q., Zhuge, Q., and Sha, E. *Optimizing Data Placement of Loops for Energy Minimization with Multiple Types of Memories.* Journal of Signal Processing Systems (JSPS). 2013. Springer.

POSTERS, PREPRINTS AND TALKS

P.9 Zhang, J., and Garg, S. *Energy Efficient and Reliable Deep Learning Accelerator Design.* ACM SIGDA/IEEE CEDA DATE PhD Forum. Grenoble, France. Mar. 9, 2020.
TTTC's E.J. McCluskey Doctoral Thesis Competition. San Diego, CA. Apr. 7, 2020.

P.8 Zhang, J., and Garg, S. *Leveraging Model Diversity for High QoS Deep Learning Inference in the Clouds.* Workshop on Hardware and Algorithms for Learning On-a-Chip (HALO 2019) in conjuction with ICCAD 2019. Westminster, CO. Nov. 7, 2019.

P.7 Zhang, J., Zarar., S., Ambardekar, A., and Garg, S. *CompAct TPU: Enabling Compressed Activation Memories for Low-Power DNN Acceleration.* Work-in-Progress Session of the ACM/IEEE 56th Design Automation Conference (DAC 2019). Las Vegas. Jun. 2–6, 2019.

P.6 Zhang, J., Garg, S. *Energy-Efficient and Fault-Tolerant Hardware Accelerators for Deep Learning. NYU WIRELESS Open House*. Brooklyn, NY, USA. Jan. 25, 2019.

P.5 *SparseTPU: Exploiting Sparsity for Energy-Efficiency in Systolic Arrays.* Microsoft Research. Redmond, WA, USA. Nov. 16, 2018.

P.4 *Energy Efficient and Error Resilience Deep Learning Accelerators Design.* Microsoft HoloLens Team. Redmond, WA, USA. Aug. 28, 2018.

P.3 *RL-IOD: Minimize SSD Read Tail Latency.* Samsung Semiconductor Memory Platform Lab. San Jose, CA, USA. Aug. 8, 2018.

P.2 Zhang, J., Ghodsi, Z., Rangineni., K., and Garg, S. *Enabling Extreme Energy Efficiency Via Timing Speculation for Deep Neural Network Accelerators.* Workshop of the 1st Computational Intelligence & Soft Computing (CISC 2017) in conjunction with PACT 2017. Portland, Oregon, USA. Sep. 10, 2017.

P.1 Massad, ME., Zhang, J., Garg, S. and Tripunitara, MV. *Logic Locking for Secure Outsourced Chip Fabrication: A New Attack and Provably Secure Defense Mechanism.* arXiv:1703.10187. 2017.

PROFESSIONAL EXPERIENCE

New York University, Brooklyn, New York
Teaching Assistant **September 2016 to December 2017**

- Fall 2016, 2017, Computer Architecture.
- Spring 2017, Introduction to VLSI.

Oklahoma State Universiy, Stillwater, Oklahoma
Visiting Student **January 2015 to May 2015**

- Compiler Optimization for Embedded Non-Volatile Memories.
- Advisor: Professor Jingtong Hu.

Chongqing University, Chongqing, China
Research Assistant **July 2011 to December 2014**

- Performance-Aware Fault Tolerant Design for Multiprocessors.
- Advisor: Professor Edwin Sha.

SZZT Electronics Shenzhen Co., Ltd., Shenzhen, China
Undergraduate Intern **September 2010 to September 2011**

- Driver development for PIN PAD encryption.
- System design for financial self-service terminals.

SERVICE

Reviewer

- ACE Transactions on Architecture and Code Optimization, 2020
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 2019
- ACM/IEEE Design Automation Conference, 2020, 2019
- IEEE Transactions on Very Large Scale Integration Systems, 2019, 2018
- IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, 2018
- IEEE Embedded Systems Letters, 2019
- Journal of Systems Architecture, 2019
- IEEE Access, 2018
- Journal of Pattern Recognition and Artificial Intelligence, 2017
- IEEE Design & Test, 2016
- International Journal of Parallel Programming, 2016

SKILLS

Programming: C/C++, PYTHON (NUMPY, NUMBA), LUA, BASH, OPENMP, VERILOG, TCL
Framework & Tools: MATLAB, TORCH/PYTORCH, TENSORFLOW, KERAS, MODELSIM, CADENCE GENUS/VIRTUOSO, XILINX ISE, VIVADO HLS, ALTERA QUARTUS, DOCKER