Harvard John A. Paulson
School of Engineering
and Applied Sciences

NYU TANDON SCHOOL OF ENGINEERING

Jeff (Jun) ZHANG, *Ph.D.* — *Curriculum vitae* (April 2022)

Contact Information

Harvard Architecture, Circuits, and Compilers Group
Maxwell Dworkin 311, 33 Oxford St
Cambridge, MA 02138

Mobile: +1-516-697-9088
E-mail: jeffzhang@g.harvard.edu
Google Scholar

Research Interests

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

Education

New York University, New York

Ph.D., Electrical and Computer Engineering, **2015 - 2020**

- Dissertation: *Towards Energy-Efficient and Reliable Deep Learning Inference.*
- Thesis Contribution: ***1 Book Chapter, 2 US Patents, 11 Peer-reviewed Publications***
- Advisor: Prof. Siddharth Garg

Hunan University, Changsha, China

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

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

Research Experience

Harvard University, Cambridge, MA
Postdoctoral Fellow **2020 -**

- Algorithm-hardware Co-design and Implementation of AI Accelerators and SoCs.
- Advisors: Prof. David Brooks, Prof. Gu-Yeon Wei

Microsoft Research, Redmond, WA
Research Intern, HoloLens AI Hardware Team **Fall 2018**

- Novel 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 **Summer 2018**

- Reinforcement Learning based SSD and I/O Management for Datacenter Ecosystem.
- Mentors: Vijay Balakrishnan, Dr. Zvika Guz

Awards and Honors

ACM/IEEE Design, Automation and Test in Europe

- Best Paper Award Candidate, 2022

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, 2016, 2015

Ministry of Education of China

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

Hunan University

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

Book Chapters

B.1 Hanif, M., Khalid, F., Rehman, S., Zhang, J., Liu, K., Theocharides, T., Artussi, A., Garg, S., and Shafique, M. *“Robust Computing for Machine Learning-Based Systems"*, in Dependable Embedded Systems. Springer. 2021.

US Patents

U.2 Ambardekar A., Zarar, S., Zhang, J. *Selectively controlling memory power for schedule computations*, US Patent App. 16/355,086, 2020.

U.1 Zarar, S., Ambardekar A., Zhang, J. *Compression-encoding scheduled inputs for matrix computations*, US Patent App. 16/261,199, 2020.

Conference Proceedings

Note: **[C.4–11] are published at NYU and included in the Ph.D. thesis.**

C.19 Jia, T., Mantovani., P., Cassel dos Santos, M., Giri., D., Zuckerman., J., Loscalzo., E.J., Cochet., M., Swaminathan., K., Tombesi., G., Zhang, J., Chandramoorthy., N., Wellman., J.D., Tien, K., Carloni., L.P., Shepard., K., Brooks., D., Wei., G., Bose., P. *A 12nm Agile-Designed SoC for Swarm-Based Perception with Heterogeneous IP Blocks, a Reconfigurable Memory Hierarchy, and an 800MHz Multi-Plane NoC.*
[Under Review]

C.18 Tan, C., Tambe, T., Zhang, J., Fang, B., Geng, T., Wei, G., Brooks, D., Tumeo, A., Gopalakrishnan, G., Li, A. *ASAP-Automatic Synthesis of Area-Efficient and Precision-Aware CGRA.* ACM 36th International Conference on Supercomputing (ICS 2022). (Acceptance rate: 23.6%)

C.17 Zhang, W., Zhang, J., Xie, M., Liu, T., Wang, W., and Pan, C. *M2M-Routing: Environmental Adaptive Multi-agent Reinforcement Learning based Multi-hop Routing Policy for Self-Powered IoT Systems.* ACM/IEEE 25th Design, Automation and Test in Europe (DATE 2022). Antwerp, Belgium. Mar., 2022. (Acceptance rate: 25%)

C.16 Zhang, S., Wang, R., Ma, D., Zhang, J., Yin, X., and Jiao, X. *Energy-Efficient Brain-Inspired Hyperdimensional Computing Using Voltage Scaling.* ACM/IEEE 25th Design, Automation and Test in Europe (DATE 2022, Interactive Presentation). Antwerp, Belgium. Mar., 2022. (Acceptance rate: 34%) **Best Paper Award Candidate.**

C.15 Gupta, U., Hsia, S., Zhang, J., Wilkening, M., Pombra, J., Lee, H., Wei, G., Wu, C., and Brooks, D. *RecPipe: Co-designing Models and Hardware to Jointly Optimize Recommendation Quality and Performance.* IEEE/ACM 54th International Symposium on Microarchitecture (MICRO 2021). Oct., 2021. (Acceptance rate: 94/430=21.9%)
Artifact Evaluation: *Available, Functional, Reproduced*.

C.14 Zhang, J., Agostini, N., Song, S., Tan, C., Limaye, A., Amatya, V., Manzano, J., Minutoli, M., Castellana, V., Tumeo, A., Wei, G., and Brooks, D. *Towards Automatic and Agile AI/ML Accelerator Design with End-to-End Synthesis.* IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021, Special Session). Jul., 2021.

C.13 Tan, C., Agostini, N., Zhang, J., Minutoli, M., Castellana, V., Xie, C., Geng, T., Li, A., Barker, K., and Tumeo, A. *OpenCGRA: Democratizing Coarse-Grained Reconfigurable Arrays.* IEEE 32nd International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021, Invited Paper). Jul., 2021.

C.12 Zhang, W., Liu, T., Zhang, J., and Pan, C. *SAC: A Novel Multi-hop Routing Policy in Hybrid Distributed IoT System based on Multi-agent Reinforcement Learning.* IEEE 22nd International Symposium on Quality Electronic Design (ISQED 2021). Apr., 2021.

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.* USENIX 12th Workshop on Hot Topics in Cloud Computing (HotCloud 2020), co-located with USENIX Annual Technical Conference (ATC 2020). 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%)
Interviewed by *the Next Platform.*

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.**

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 (EUC 2013). 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

Note: **[J.6–9] are published at NYU; and [J.7–9] are included in the Ph.D. thesis.**

J.13 Curzel, S., Agostini, N., Castellana, V., Manzano, J., Zhang, J., Brooks, D., Wei, G., Ferrandi, F., and Tumeo, A., *End-to-end Synthesis of Dynamically Scheduled Machine Learning Accelerators.* IEEE Transactions on Computers, Special Issue on Hardware Acceleration of Machine Learning. [Under Review]

J.12 Zhang, W., Pan, C., Liu, T., Zhang, J., Sookhak, M., and Xie, M., *Self-sustaining IoT Systems with Optimized Topology and Intelligent Routing.* ACM Transactions on Cyber-Physical Systems (TCPS). [Under Review]

J.11 Agostini, N., Curzel, S., Zhang, J., Limaye, A., Tan, C., Amatya, V., Minutoli, M., Castellana, V., Manzano, J., Brooks, D., Wei, G., and Tumeo, A., *Bridging Python to Silicon Smartly: The SODA Toolchain.* IEEE Micro, Special Issue on Compiling for Accelerators. [Major Revision]

J.10 Yin, M., Veldanda, A., Trivedi, A., Zhang, J., Pfeiffer, K., Hu, Y., Garg, S., Erkip, E., Righetti, L., and Rangan, S., *Millimeter Wave Wireless Assisted Robot Navigation with Link State Classification.* IEEE Open Journal of the Communications Society (OJCOMS). 2022.

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. **Two US Patents Granted with Microsoft.**

J.8 Zhang, J., Ghodsi, Z., Rangineni, K., 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.
Top 5 Accessed Articles in March 2021.

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.

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.

J.3 Peng, S., Ouyang, A., and 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.

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 Invited Talks

Note: **[P.2, P.4–9] are talks related to P.h.D thesis contributions.**

P.10 *SODALITE: Software Defined Accelerators from Learning Tools Environment.*
Presenter, NSF Workshop on Machine Learning Hardware Breakthroughs Towards Green AI and Ubiquitous On-Device Intelligence, Nov., 2020

P.9 *Energy Efficient and Reliable Deep Learning Accelerator Design.*
Guest Lecture, ECE 8405, Villanova University, Apr. 2021
Presenter, NSF Workshop on Machine Learning Hardware Breakthroughs Towards Green AI and Ubiquitous On-Device Intelligence, Nov., 2020
Guest Lecture, ECE 538, The University of New Mexico, Oct., 2020
Harvard Architecture, Circuits, and Compilers Group, Aug., 2020
ACM SIGDA/IEEE CEDA DAC PhD Forum. Online. July., 2020.
ACM SIGDA/IEEE CEDA DATE PhD Forum. Grenoble, France. Mar., 2020.
TTTC's E.J. McCluskey Doctoral Thesis Competition. San Diego, CA. Apr., 2020.
Microsoft HoloLens Team. Redmond, WA, USA. Aug. 28, 2018.

P.8 *Model-Switching: Dealing with Fluctuating Workloads in MLaaS Systems.*
Invited Talk, Microsoft Advertising Group. Sunnyvale, CA, USA. Apr. 14, 2020.

P.7 *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.6 *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.5 *Energy-Efficient and Fault-Tolerant Hardware Accelerators for Deep Learning.*
NYU WIRELESS Open House. Brooklyn, NY, USA. Jan. 25, 2019.

P.4 *SparseTPU: Exploiting Sparsity for Energy-Efficiency in Systolic Arrays.*
Microsoft Research. Redmond, WA, USA. Nov. 16, 2018.

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

P.2 *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.

Research Mentoring

- Vikas Natesh, Thierry Tambe (PhD - Harvard) **Spring 2021**
  Project: Exploring the use of Non-volatile Memories in 3DIC Architectures.
- Akshaj Veldanda, Mingsheng Yin (PhD - NYU) **Fall 2020**
  Project: Cooperative Action-perception Manipulation over 5G.
  Publication: [J.11] IEEE OJ-COMS

- Parul Raj (MSc - NYU) **Spring 2019**
  Project: Systolic Array Based CNN Accelerator on FPGA.
  Publication: [C.10] ESWeek'19, [J.9] ACM TECS'19, NYU Master thesis
  First Employer: Design Verification Engineer at Apple

- Monish Narendra Kapadia, Karan Mamaniya (MSc - NYU) **Spring 2019**
  Project: Thermal Analysis for Deep Learning Accelerators.

- Rafael Jin (MSc - NYU) **Spring 2018**
  Project: Optimization on Bandwidth-constraint Systolic Array for CNN.
  Publication: NYU Master thesis
  First Employer: Software Engineer at Akuna Capital

- Fengyang Jiang (MSc - NYU) **Fall 2017**
  Project: Physical Design Optimization for CNN Acceleration.
  First Employer: Ph.D. student at Pennsylvania State University

- Kartheek Rangineni (Undergraduate - IIT/NYU) **Summer 2017**
  Project: Efficient Deep Learning Accelerator Design.
  Publication: [C.7] DAC'18, [J.8] IEEE Design & Test'19
  First Employer: Firmware Engineer at Intel

Professional Experience

Harvard University, Cambridge, Machachusetts
Teaching Assistant **Spring 2022**
- Design of VLSI Circuits and Systems.

Harvard University, Cambridge, Machachusetts
Course Assistant **Spring 2021**
- Topics in Mixed-Signal Integrated Circuits.

New York University, Brooklyn, New York
Teaching Assistant **2016 - 2017**
- Fall 2016, 2017, Computer Architecture.
- Spring 2017, Introduction to VLSI.

Oklahoma State Universiy, Stillwater, Oklahoma
Visiting Student **Spring 2015**
- Compiler Optimization for Embedded Non-Volatile Memories.
- Advisor: Professor Jingtong Hu.

Columbia University, New York, New York
Visiting Student **Spring 2014**

Chongqing University, Chongqing, China
Research Assistant **2013**
- Performance-Aware Fault Tolerant Design for Multiprocessors.
- Advisor: Professor Edwin Sha.

SZZT Electronics Shenzhen Co., Ltd., Shenzhen, China
Undergraduate Intern **2010 - 2011**
- Encryption driver development for PIN PAD.
- System design for financial self-service terminals.

Academic Service

Organizing Committee

- The NOPE Workshop at ASPLOS, 2022
- ACM CADathlon@ICCAD, 2022

TPC Member, Invited Reviewer

- The Conference on Machine Learning and Systems (MLSys), 2022
- ACM/IEEE Design Automation Conference (DAC), 2022, 2021, 2020
- IEEE International Conference on Computer Design (ICCD), 2022, 2021
- ACM Transactions on Design Automation of Electronic Systems, 2022
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 2021, 2020, 2019
- IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 2021
- IEEE International Symposium on Workload Characterization (IISWC), 2021
- International Symposium on Computer Architecture (ISCA), 2021
- International Parallel & Distributed Processing Symposium (IPDPS), 2021
- IEEE Transactions on Circuits and Systems, 2021
- IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2021, 2019, 2018
- Journal of Low Power Electronics and Applications, 2021
- Concurrency and Computation: Practice and Experience, 2021
- USENIX OSDI Artifact Evaluation Committee, 2020
- ACE Transactions on Architecture and Code Optimization, 2020
- IEEE Embedded Systems Letters, 2020, 2019
- IEEE Open Journal of Circuits and Systems, 2020
- IEEE Transactions on Very Large Scale Integration Systems, 2019, 2018
- 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