TONG GENG

875 Beacon St, Apt 2, Boston MA 02215
(+1) 857 770 8848 ⋄ tonygeng521@gmail.com

EDUCATION

Boston University *Sep 2017 - Present*
Computer Engineering, PhD GPA: 4.0/4.0

Eindhoven University of Technology *Sep 2015 - Dec 2016*
Electronic Systems, PhD (Transfer) GPA: 10.0/10.0

Eindhoven University of Technology *Sep 2013 - Aug 2015*
Electronic Systems, Master of Science GPA: 8.0/10.0

Zhejiang University *Sep 2009 - Aug 2013*
Electronic Engineering and Information Technology, Bachelor of Engineering GPA: 85.8/100.0

RESEARCH AND WORK EXPERIENCE

Pacific Northwest National Laboratory Richland, WA
Post Doctorate Research Associate *To Start Jan 2021*

- **Architecture Design for Next-generation Reconfigurable HPC Platforms**
  - Investigate the design approaches of mapping current Computation-Flow-Architecture to a multi-FPGA cluster platform
  - Perform design exploration on optimizing the task circulation and data-fetching networks
- **Novel Memory Architecture Design for Machine Learning**
- **FPGA-based Hardware Accelerator Design for Graph Convolutional Network**

PhD Intern, Machine Learning *May 2019 - Aug 2019*

- **Architecture Design of Graph Convolution Network (GCN): AWB-GCN**
  - Proposed a novel architecture, AWB-SPMM, to accelerate Sparse Matrix Multiplication (SPMM) kernels with power-law Non-Zero distributions
  - Proposed an efficient accelerator design, AWB-GCN, which provides over 21x faster GCN inference on Intel D5005 FPGA than PyTorch-Geometric-based RTX8000 GPU implementations
  - Featured in PNNL news: https://www.pnnl.gov/news-media/sharing-load-speeds-machine-learning

PhD Intern, Machine Learning *May 2018 - Aug 2018*

- **Architecture Design of Binary Neural Network (BNN) Inference: O3BNN & LP-BNN**
  - Proposed architectures, O3BNN & LP-BNN, to accelerate BNN inference with runtime pruning
  - The proposed designs realize ultra-low latency BNN inference: latency of AlexNet and VGGNet-19 are $22\mu s$ and $355\mu s$ respectively.

Boston University Boston, MA
Graduate Research Assistant *Sep 2017 - Present*

- **FPGA cluster-based acceleration of CNN training: FPDeep**
  - Maps CNN training to distributed FPGA clusters efficiently using hybrid model- & layer- parallelism and with perfect workload balancing
  - Supports highly scalable CNN training and address the poor generalization problems resulted from the growth of mini-batch size
- **ADMM-based RNN Acceleration: ACSB-RNN**
- **CGRA-based QNN Acceleration: CQNN**
- **FPGA cluster-based Molecular Dynamics simulation**
- **Embedded FPGA-based In-Switch processing of MPI Collectives**

Eindhoven University of Technology Eindhoven, the Netherlands
Research Engineer/PhD Student *Sep 2015 - Dec 2016*

- **Fault-tolerant computer architecture**
- **Reliability (Architectural Vulnerability Factor) Modeling of CPU**
- **SIMD processor architecture design and optimization for real-time CNN inference**

Master Thesis Project *Sep 2014 - Aug 2015*

- **Scratchpad memory system design with access-pattern aware auto-load mechanism**

PC EXPERIENCE AND PAPER REVIEW

- Program Committees: PPOPP conference AEC (Artifact Evaluation Committee)
- Paper reviews: Transaction on Computer, Transaction on Reconfigurable Technology and Systems, Parallel Computing, MICPRO, PACT, FCCM, FPL, FPT, PPOPP, DSD, CASES, HPEC, HEART

AWARDS AND ACHIEVEMENTS

- 2019 - Travel grant to attend International Conference on Supercomputing (ICS) 2019
- 2017-2018 - Distinguished Computer Engineering Fellowship at Boston University
- 2013-2015 - Amandus H. Lundqvist Scholarship at Eindhoven University of Technology

TECHNICAL SKILLS

- **Program Languages:** Python, C++, C, Verilog, VHDL, SystemVerilog, HLS, System C, LaTeX
- **Software/Tools/OS:** Windows, Linux, Xilinx Vivado, Altera Quartus, Xilinx Vitis, Xilinx SDAccel, Cadence, Matlab, VS, Labview, Modelsim

SELECTED PUBLICATIONS

1. ***T.Geng***, A.Li, T.Wang, C.Wu, Y.Li, ..., M.Herbordt: AWB-GCN: *A Hardware Accelerator of Graph-Convolution-Network through Runtime Workload Rebalancing*, the 53rd IEEE/ACM International Symposium on Microarchitecture (**MICRO** 2020)
2. ***T.Geng***, T.Wang, C.Wu, Y.Li, ..., A.Li, M.Herbordt: *O3BNN-R: An Out-Of-Order Architecture for High-Performance and Regularized BNN inference*, IEEE Transactions on Parallel and Distributed Systems (**TPDS**)
3. ***T.Geng****, T.Wang*, A.Li, X.Jin, M.Herbordt: *FPDeep: Scalable Acceleration of CNN Training on Deeply-Pipelined FPGA Clusters*, IEEE Transactions on Computers (**TC**)
4. ***T.Geng***, C.Wu, C.Tan, B.Fang, A.Li, M.Herbordt: *CQNN: a CGRA-based QNN Framework*, IEEE High Performance Extreme Computing Conference (**HPEC** 2020)
5. ***T.Geng****, R.Shi*, P.Dong*, ..., M.Herbordt, A.Li, Y.Wang: *CSB-RNN: A Faster-than-Realtime RNN Acceleration Framework with Compressed Structured Blocks*, the 34th ACM International Conference on Supercomputing (**ICS** 2020)
6. P.Haghi, ***T.Geng***, T.Wang, A. Guo, M.Herbordt: *FP-AMG: FPGA-Based Acceleration Framework for Algebraic Multigrid Solvers*, the 29th IEEE International Symposium On Field-Programmable Custom Computing Machines (**FCCM** 2020)
7. A.Li, ***T.Geng***, T.Wang, M.Herbordt, S.Song, K.Barker: *BSTC: A Novel BinarizedSoft-Tensor-Core Design for Accelerating Bit-Based Approximated Neural Nets*, Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (**SC** 2019)
8. C.Yang, ***T.Geng***, T.Wang, ..., M.Herbordt: *Fully integrated FPGA molecular dynamics simulations*, Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (**SC** 2019)
9. ***T.Geng***, T.Wang, C.Wu, C.Yang, W.Wu, A.Li, M.Herbordt: *O3BNN: An Out-Of-Order Architecture for High-Performance Binarized Neural Network Inference with Fine-Grained Pruning*, the 33th ACM International Conference on Supercomputing (**ICS** 2019)

10. ***T.Geng***, T.Wang, ..., M.Herbordt: *LP-BNN: Ultra-low-Latency BNN Inference with Layer Parallelism*, the 30th IEEE International Conference on Application specic Systems, Architectures and Processors (**ASAP** 2019)

11. T.Wang, ***T.Geng***, X.Jin, M.Herbordt: *FP-AMR: A Reconfigurable Fabric Framework for Block-Structured Adaptive Mesh Refinement Applications*, the 28th IEEE International Symposium On Field-Programmable Custom Computing Machines (**FCCM** 2019)

12. Q.Xiong, C.Yang, R.Xu, R.Patel, ***T.Geng***, A.Skjellum, M.Herbordt: *GhostSZ: A Transparent SZ Lossy Compression Framework with FPGAs*, the 28th IEEE International Symposium On Field-Programmable Custom Computing Machines (**FCCM** 2019)

13. C.Yang, ***T.Geng***, T.Wang, J.Sheng, ... M.Herbordt: *Molecular Dynamics Range-Limited Force Evaluation Optimized for FPGAs*, the 30th IEEE International Conference on Application specic Systems, Architectures and Processors (**ASAP** 2019)

14. T.Wang, ***T.Geng***, X.Jin, M.Herbordt: *Accelerating AP3M-Based Computational Astrophysics Simulations with Reconfigurable Clusters*, the 30th IEEE International Conference on Application specic Systems, Architectures and Processors (**ASAP** 2019)

15. ***T.Geng***, E.Diken, T.Wang, L.Jozwiak, M.Herbordt: *An Access-Pattern-Aware On-Chip Vector Memory System with Automatic Loading for SIMD Architecture*, IEEE High Performance Extreme Computing Conference (**HPEC** 2018)

16. ***T.Geng***, T.Wang, A.Sanaullah, C.Yang, R.Patel, M.Herbordt: *A Framework for Acceleration of CNN Training on Deeply-Pipelined FPGA Clusters with Work and Weight Load Balancing*, the 28th International Conference on Field-Programmable Logic and Applications (**FPL** 2018)

17. ***T.Geng***, T.Wang, A.Sanaullah, C.Yang, R.Xu, R.Patel, M.Herbordt: *FPDeep: Acceleration and Load Balancing of CNN Training on FPGA Clusters*, the 27th IEEE International Symposium On Field-Programmable Custom Computing Machines (**FCCM** 2018)

18. Z.Xiang, T.Wang, ***T.Geng***, ..., M.Herbordt: *Soft-Core, Multiple-Lane, FPGAbased ADCs for a Liquid Helium Environment*, IEEE High Performance Extreme Computing Conference (**HPEC** 2018)

19. ***T.Geng***, L.Waeijen, M.Peemen, H.Corporaal, Y.He: *MacSim: A MAC-Enabled HighPerformance SIMD Architecture for Deep Learning*, the 19th Euromicro Conference on Digital System Design (**DSD** 2016)

20. Y.He, M.Peemen, L.Waeijen, ..., H.Corporaal, ***T.Geng***: *A Configurable SIMD Architecture with Explicit Datapath for CNN*, International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (**SAMOS** 2016)