TONG (TONY) GENG

Homepage: tonytgeng.com
902 Battelle Blvd, Richland, WA 99354
(+1) 857 770 8848 ⋄ tonygeng521@gmail.com

RESEARCH INTERESTS

Neural Network, Graph Intelligence, Computer Architecture, Reconfigurable Architecture

EDUCATION

Boston University *Sep 2017 - Jan 2021*
Computer Engineering, PhD GPA: 4.0/4.0
Dissertation: *FPGA-based High-performance Neural Network Acceleration*

Eindhoven University of Technology *Jan 2016 - 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: 8.3/10.0
Thesis: *Scratchpad memory system design with access-pattern aware auto-load mechanism*

Zhejiang University *Sep 2009 - Aug 2013*
Electronic Engineering and Information Technology, Bachelor of Engineering GPA: 87.8/100.0

RESEARCH AND WORK EXPERIENCE

Pacific Northwest National Laboratory Richland, WA
Postdoctoral Research Associate (Outstanding Postdoc Award) *Jan 2021 - Present*

- **Architecture Design to Address Data Locality Problems in Graph Neural Networks**
  - Design and implement *I-GCN [MICRO21]* architecture to accelerate GNN inference with the support of on-the-fly graph-component-aware data locality enhancement.
- **Algorithm-Accelerator Co-design for Graph Neural Network Acceleration**
  - Design and implement *GCoD [HPCA22, ICCAD21]*, the first GNN algorithm and accelerator co-design framework which relieves data locality problems in GNNs by regularizing graph structures during training and uses a dedicated two-pronged accelerator to process regularized graphs.
- **System Design of Next-generation HPC Platforms with Hierarchical Heterogeneity**
  - Design and implement *ARENA [TPDS21]*, a reconfigurable compute-flow system architecture and prototype it on PNNL's Junction cluster with Xilinx SN1000 SmartNICs and Versal ACAP.
  - Study Neural Network acceleration with hierarchically heterogeneous architecture.
- **Coarse-Grained Reconfigurable Arrays Architecture Design**
  - Design and implement *DRIPS [HPCA22]*, a coarse-grained, dynamically and partially reconfigurable architecture for data-dependent streaming applications with the support of on-the-fly workload rebalancing. We also present a unified compiler framework to facilitate the mapping of a given streaming application onto the DRIPS CGRA architecture.

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

- **Architecture Design to Address Load Imbalance Problems in Graph Neural Networks**
  - Design and implement *AWB-GCN [MICRO20]* architecture to accelerate GNN inference and SpMMs which matrices follow power-law non-zero distributions with the support of on-the-fly workload rebalancing/autotuning.

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

- **Architecture Design to Prune Superfluous Operations in Binary Neural Network**
  - Design and implement *O3BNN & LP-BNN [TPDS21, SC19, ICS19, ASAP19]* architectures to accelerate BNN inference with lightweight runtime superfluous operation detection and pruning.

Boston University Boston, MA

Graduate Research Assistant *Sep 2017 - Jan 2021*

- **FPGA-cluster System Design to Achieve Scalable and High-performance CNN Training**
  - Design and implement *FPDeep [TC20, FCCM18, FPL18]*, the first FPGA-cluster-based CNN training framework which is able to map CNN training to distributed FPGA clusters efficiently using fine-grained model-parallelism and with near-ideal workload balancing.
  - FPDeep supports highly scalable CNN training and addresses the poor generalization problems resulting from the growth of mini-batch size.
  - Demonstrate competitiveness of FPGA clusters with GPU clusters for DNN training.
- **CGRA-based Accelerator Design for Mixed-Precision QNN Acceleration**
  - Design and implement *CQNN [HPEC20]*, a binary-CGRA-based QNN accelerator that supports on-the-fly precision conversion through runtime NOC reconfiguration and binary computational component grouping.
- **Algorithm-Architecture Co-Design for Recurrent Neural Network**
  - Design and implement *CSB-RNN [ICS20]*, an optimized full-stack RNN framework with a novel compressed-structured-block pruning technique and a novel hardware architecture with a dedicated compiler with the support of dynamic workload rebalancing.
- **FPGA-based Full-scale Molecular Dynamics Simulation** *[SC19, FCCM21, ASAP19]*

Graduate Teaching Assistant *Fall 2018 & Spring 2019*

EC311 - Introduction to Logic Design (two semester) Evaluation: 4.6/5.0

Guest Lecturer *Spring 2020*

EC700 - Advanced Topics in Electrical Computer Engineering (two lectures)

Eindhoven University of Technology Eindhoven, the Netherlands

Research Engineer/PhD Student *Sep 2015 - Dec 2016*

- **Fault-tolerant computer architecture**
- **Reliability (Architectural Vulnerability Factor) Modeling**
- **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**

PUBLICATIONS

1. [**HPCA 2022**] H.You*, ***T.Geng****, Y.Zhang, A.Li, Y.Lin: *GCoD: Graph Convolutional Network Acceleration via Dedicated Algorithm and Accelerator Co-Design*, The 28th IEEE International Symposium on High-Performance Computer Architecture.
2. [**HPCA 2022**] C.Tan, N.B.Agostini, ***T.Geng***, C.Xie, J.Li, A.Li, K.Barker, A.Tumeo: *DRIPS: Dynamic Rebalancing of Pipelined Streaming Applications on CGRAs*, The 28th IEEE International Symposium on High-Performance Computer Architecture.
3. [**MICRO 2021**] ***T.Geng***, C.Wu, ..., M.Herbordt, Y.Lin, A.Li: *I-GCN: A Graph Convolutional Network Accelerator with Runtime Locality Enhancement through Islandization*, the 54th IEEE/ACM International Symposium on Microarchitecture.

4. [**TPDS 2021**] ***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.

5. [**HPEC 2021**] ***T.Geng***, C.Wu, C.Tan, ..., M.Herbordt, A.Li: *A Survey: Handling Irregularities in Neural Network Acceleration with FPGAs*, IEEE High Performance Extreme Computing Conference.

6. [**SC 2021**] B.Feng, Y.Wang, ***T.Geng***, A.Li, Y.Ding: *APNN-TC: Accelerating Arbitrary Precision Neural Networks on Ampere GPU Tensor Cores*, Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis.

7. [**TPDS 2021**] C.Tan, C.Xie, ***T.Geng***, ..., K.Barker, A.Li: *ARENA: Asynchronous Reconfigurable Accelerator Ring to Enable Data-Centric Parallel Computing*, IEEE Transactions on Parallel and Distributed Systems.

8. [**ICCD 2021**] C.Tan, ***T.Geng***, C.Xie, N.Agostini, J.Li, A.Li, K.Barker, A.Tumeo: *DynPaC: Coarse-Grained, Dynamic, and Partially Reconfigurable Array for Streaming Applications*, the 39th IEEE International Conference on Computer Design. ***(Best Paper Award)***

9. [**ICCAD 2021**] Y.Zhang, H.You, Y.Fu, ***T.Geng***, A.Li, Y.Lin: *G-CoS: GNN-Accelerator Co-Search Towards Both Better Accuracy and Efficiency*, 2021 International Conference On Computer Aided Design.

10. [**ICCAD 2021**] D.Manu, ..., ***T.Geng***, A.Li, C.Ding, W.Jiang, L.Yang: *BFL-DISCO: Federated Generative Adversarial Network for Graph-based Molecule Drug Discovery*, 2021 International Conference On Computer Aided Design.

11. [**ICCAD 2021**] H.Peng, ..., ***T.Geng***, A.Li, J.Bi, M.Song, W.Jiang, H.Liu, C.Ding: *Optimizing FPGA-based Accelerator Design for Large-Scale Molecular Similarity Search*, 2021 International Conference On Computer Aided Design.

12. [**FCCM 2021**] C.Wu, ***T.Geng***, S.Bandara, C.Yang, V.Sachdeva, W.Sherman, M.Herbordt: *Upgrade of FPGA Range-Limited Molecular Dynamics to Handle Hundreds of Processors*, the 30th IEEE International Symposium On Field-Programmable Custom Computing Machines.

13. [**MICPRO 2021**] Y. Li, ***T.Geng***, A. Li, H. Yu: *BCNN: Binary Complex Neural Network*, Microprocssors and Microsystems.

14. [**BDMA 2021**] Y. Li, ***T.Geng***, A. Li, H. Yu: *GAAF: Searching Activation Functions for Binary Neural Networks through Genetic Algorithm*, Journal of Big Data Mining and Analytics.

15. [**CPE 2021**] P. Haghi, ..., ***T.Geng***, ..., A. Skjellum, M. C. Herbordt: *Reconfigurable Switches for High Performance and Flexible MPI Collectives*, Concurrency and Computation: Practice and Experience.

16. [**ISQED 2021**] H.Peng, S.Huang, ***T.Geng***, A.Li, W.Jiang, H.Liu, S.Wang, C.Ding: *Accelerating Transformer-based Deep Learning Models on FPGAs using Column Balanced Block Pruning*, the 22nd International Symposium on Quality Electronic Design.

17. [**ASAP 2021**] H.Peng, ..., ***T.Geng***, ..., C.Ding: *Binary Complex Neural Network Acceleration on FPGA*, IEEE International Conference on Application specific Systems, Architectures and Processors.

18. [**ASAP 2021**] C.Tan, ***T.Geng***, ..., A.Tumeo: *OpenCGRA: Democratizing Coarse-Grained Reconfigurable Arrays*, IEEE International Conference on Application specific Systems, Architectures and Processors.

19. [**HPEC 2021**] P.Haghi, A.Guo, ***T.Geng***, ..., M.Herbordt: *Workload Imbalance in HPC Applications: Effect on Performance of In-Network Processing*, IEEE High Performance Extreme Computing Conference. ***(Best Student Paper Award)***

20. [**HPEC 2021**] C.Wu, S.Bandara, ***T.Geng***, ..., M.Herbordt: *System-Level Modeling of GPU/FPGA Clusters for Molecular Dynamics Simulations*, IEEE High Performance Extreme Computing Conference.

21. [**MICRO 2020**] ***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.

22. [**HPEC 2020**] ***T.Geng***, C.Wu, C.Tan, B.Fang, A.Li, M.Herbordt: *CQNN: a CGRA-based QNN Framework*, IEEE High Performance Extreme Computing Conference.

23. [**TC 2020**] T.Wang*, ***T.Geng****, A.Li, X.Jin, M.Herbordt: *FPDeep: Scalable Acceleration of CNN Training on Deeply-Pipelined FPGA Clusters*, IEEE Transactions on Computers.

24. [**ICS 2020**] R.Shi*, P.Dong*, ***T.Geng****, ..., 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.

25. [**FCCM 2020**] 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.

26. [**FPT 2020**] P.Haghi, A. Guo, ***T.Geng***, ..., M.Herbordt: *A Reconfigurable Compute-in-the-Network FPGA Assistant for High-Level Collective Support with Distributed Matrix Multiply Case Study*, 2020 International Conference on Field-Programmable Technology.

27. [**HPEC 2020**] C.Wu, ***T.Geng***, V.Sachdeva, ..., M.Herbordt: *A Communication-Efficient Multi-Chip Design for Range-Limited Molecular Dynamics*, IEEE High Performance Extreme Computing Conference.

28. [**HPEC 2020**] P.Haghi, A.Guo, ..., ***T.Geng***, J.Broaddus, R.Marshall, A.Skjellum, M.Herbordt: *FPGAs in the Network and Novel Communicator Support Accelerate MPI Collectives*, IEEE High Performance Extreme Computing Conference.

29. [**ICS 2019**] ***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.

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

31. [**SC 2019**] 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.

32. [**SC 2019**] 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.

33. [**ASAP 2019**] 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 specific Systems, Architectures and Processors.

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

35. [**FCCM 2019**] 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.

36. [**FCCM 2019**] 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.

37. [**FPL 2018**] ***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.

38. [**FCCM 2018**] ***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.

39. [**HPEC 2018**] ***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.

40. [**HPEC 2018**] 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.

41. [**DSD 2016**] ***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.

42. [**SAMOS 2016**] 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.

*shared equally in 1st author credit

PC EXPERIENCE AND PAPER REVIEW

- TPC/ERC: HPCA 2022, IPDPS 2021, FPL 2021, ASAP 2021, PPOPP 2019, HPC-FPGA-Cluster 2021
- Journal Paper Reviews: TC, TPDS, TRETS, TCAD, CAL, CSUR, ParCo, MICPRO

AWARDS AND ACHIEVEMENTS

- 2021 Outstanding Postdoc Award at PNNL
- 2021 Best Paper Award, The 39th IEEE International Conference on Computer Design (ICCD)
- 2021 Best Student Paper Award, High Performance Extreme Computing (HPEC)
- 2019 Travel grant to attend International Conference on Supercomputing (ICS)
- 2017-2018 Distinguished Computer Engineering Fellowship at Boston University
- 2013-2015 Amandus H. Lundqvist Scholarship at Eindhoven University of Technology

MEDIA REFERENCES

- AWB-GCN work featured in PNNL news: https://www.pnnl.gov/news-media/sharing-load-speeds-machine-learning
- O3BNN work featured in DOE news: https://www.eurekalert.org/features/doe/2020-11/dnnl-tio112320.php

LEADERSHIP

- 2011-2013 Captain of the College Basketball Team, Zhejaing University
- 2011-2012 Vice President of Student Union, Zhejiang University
- 2010-2011 Chair of Athletics Committee, Student Union, Zhejiang University