GEONHWA JEONG

Klaus Advanced Computing Building 3305, 266 Ferst Dr NW, Atlanta, GA 30332, USA.
☎ +1 470-309-8607 ✉ geonhwa.jeong@gatech.edu 🌍 ghjeong12.github.io

RESEARCH INTERESTS

Computer architecture, HW/SW co-design, Domain-specific accelerators, Compiler optimization, Datacenter-scale computing, Deep learning, Model compression

EDUCATION

Ph.D. in Computer Science Aug. 2019 - May 2024 (expected)
Advisor: Prof. Tushar Krishna *GPA: 4.0/4.0*

Georgia Institute of Technology (Georgia Tech)

Master of Science in Computer Science Aug. 2019 - Dec. 2021
Specialization: Machine Learning *GPA: 4.0/4.0*

Georgia Institute of Technology (Georgia Tech)

Bachelor of Science in Creative IT Engineering Mar. 2013 - Feb. 2019
Double Major in Computer Science and Engineering *Summa Cum Laude*

Pohang University of Science and Technology (POSTECH)

RESEARCH EXPERIENCES

Synergy Lab at Georgia Tech Nov. 2019 - Current
Research Assistant (Advisor: Prof. Tushar Krishna) *Atlanta, USA*

First of all, I worked on optimizing the dataflow considering both software and hardware sides to fully exploit available parallelism in spatial accelerator [*MICRO'20, PACT'21, TPDS'22*]. Also, I led a project to efficiently integrate a systolic array based matrix engine in CPU to accelerate deep learning workloads [*DAC'21*]. Extending the previous work to accelerate sparse DNNs with structured sparsity, I introduced new flexible sparse/dense matrix engine in CPUs [*HPCA'23*]. Currently, I am conducting research on 1) leveraging a set of structured sparsity patterns using approximation for tensor workloads 2) exploiting both quantization and sparsification with structured decomposition for large language models.

High Performance Architecture Lab at Georgia Tech Sep. 2019 - Nov. 2019
Research Student (Advisor: Prof. Hyesoon Kim) *Atlanta, USA*

The Unified Virtual Memory with GPU and CPU not only gets rid of programmers burden a lot but also enables running a program with the working set size larger than the GPU capacity. I worked on efficient page prefetching mechanisms with various workloads.

Compiler Optimization Research Lab at POSTECH Sep. 2017 - June 2018
Research Student (Advisor: Prof. Hanjun Kim) *Pohang, Republic of Korea*

I was engaged in three projects including implementation of IoT platform for various types of users (manufacturer, service developer, user), development of programmable magnetic blocks to teach computational thinking to kids and development of hot function/basic block detector using LLVM compiler.

Database and Data Mining Lab at POSTECH Mar. 2015 - Nov. 2015
Research Student (Advisor: Prof. Wook-Shin Han) *Pohang, Republic of Korea*

I was involved in a team to develop a new graph database system to process large streaming data. I participated in the initial design of the query processing engine and the development of memory management system to support the engine [*SIGMOD'18*].

WORK EXPERIENCES

NVIDIA Research May 2023 - Aug. 2023
Research Intern *Westford, MA, USA*

I worked on a project to exploit both sparsity and reduced precision for Large Language Models. I filed a patent for the work.

NVIDIA Research May 2022 - Aug. 2022
Research Intern *Santa Clara, CA, USA*

I worked on a project to accelerate DNN model inferences using tensor approximation while minimizing accuracy drop. I filed a patent for the work and submitted a paper.

Facebook May 2021 - Aug. 2021
Research Engineering Intern *Seattle, WA, USA*

I worked on workload characterization for data compression in datacenter-scale services at Facebook and explored HW offloading opportunities [*ISPASS'22*].

Intel Labs May 2020 - Aug. 2020
Graduate Technical Intern *Hillsboro, OR, USA*

I implemented a performance model for a new architectural feature of CPU and built a framework to validate the model by comparing against RTL results.

VoyagerX Feb. 2019 - July 2019
Software Engineering Intern *Seoul, Republic of Korea*

I developed a mobile application for automatic meeting notes with speaker diarization using a deep learning model to extract feature vectors from voice data.

Samsung Research July 2018 - Aug. 2018
Field Training *Seoul, Republic of Korea*

I was involved in a team developing a cloud management system, which is used by the members of Samsung Research to develop their own programs, in GitOps manner to keep the more robust system.

Korea Augmentation to the US Army Nov. 2015 - Aug. 2017
Service Member *Seoul, Republic of Korea*

I served mandatory military service as a movement specialist with the United States Army and participated in three combined exercises (KR 17, UFG 17, and KR 18).

Kakao Corp. Jan. 2015 - Feb. 2015
Intern *Seongnam, Republic of Korea*

I developed an abuse detection system to automatically identify abusers to prevent general users from being exposed to inappropriate posts and comments.

PUBLICATIONS

[1] **Geonhwa Jeong**, Bikash Sharma, Nick Terrell, Abhishek Dhanotia, Zhiwei Zhao, Niket Agarwal, Arun Kejariwal, Tushar Krishna "Characterization of Data Compression in Datacenter," in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software* ***(ISPASS)****, Apr. 2023.*

[2] **Geonhwa Jeong**, Sana Damani, Abhimanyu Bambhaniya, Eric Qin, Christopher J. Hughes, Sreenivas Subramoney, Hyesoon Kim, and Tushar Krishna, "VEGETA: Vertically-Integrated Extensions for Sparse/Dense GEMM Tile Acceleration on CPUs," in *Proceedings of the 29th IEEE International Symposium on High-Performance Computer Architecture* ***(HPCA)****, Feb. 2023.*

[3] **Geonhwa Jeong**, Bikash Sharma, Nick Terrell, Abhishek Dhanotia, Zhiwei Zhao, Niket Agarwal, Arun Kejariwal, and Tushar Krishna, "Understanding Data Compression in Warehouse-Scale Datacenter Services," in *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software* ***(ISPASS)****, May 2022.*

[4] Gordon E. Moon, Hyoukjun Kwon, **Geonhwa Jeong**, Prasanth Chatarasi, Sivasankaran Rajamanickam, Tushar Krishna, "Evaluating Spatial Accelerator Architectures with Tiled Matrix-Matrix Multiplication," *IEEE Transactions on Parallel and Distributed Systems* ***(TPDS)****, Apr. 2022.*

[5] **Geonhwa Jeong**, Eric Qin, Ananda Samajdar, Christopher J. Hughes, Sreenivas Subramoney, Hyesoon Kim, Tushar Krishna, "RASA: Efficient Register-Aware Systolic Array Matrix Engine for CPU," in *Proceedings of the 58th Annual Design Automation Conference* ***(DAC)****, Dec. 2021.*

[6] **Geonhwa Jeong**, Gokcen Kestor, Prasanth Chatarasi, Angshuman Parashar, Po-An Tsai, Siva Rajamanickam, Roberto Gioiosa, and Tushar Krishna, "Union: A Unified HW-SW Co-Design Ecosystem in MLIR for Evaluating Tensor Operations on Spatial Accelerators," in *Proceedings of 30th International Conference on Parallel Architectures and Compilation Techniques* ***(PACT)****, Sep. 2021.*

[7] Eric Qin, **Geonhwa Jeong**, William Won, Sheng-Chun Kao, Hyoukjun Kwon, Sudarshan Srinivasan, Dipankar Das, Gordon E. Moon, Sivasankaran Rajamanickam, Tushar Krishna, "Extending Sparse Tensor Accelerators to Support Multiple Compression For-

mats," in *Proceedings of the 35th IEEE International Parallel & Distributed Processing Symposium* ***(IPDPS)****, May 2021.*

[8] Jan Moritz Joseph, Lennart Bamberg, **Geonhwa Jeong**, Ruei-Ting Chien, Rainer Leupers, Alberto Garcia-Ortiz, Tushar Krishna, Thilo Pionteck, "Bridging the Frequency Gap in Heterogeneous 3D SoCs through Technology-Specific NoC Router Architectures," in *Proceedings of the 26th Asia and South Pacific Design Automation Conference* ***(ASP-DAC)****, Jan. 2021.*

[9] Sheng-Chun Kao, **Geonhwa Jeong**, Tushar Krishna, "ConfuciuX: Autonomous Hardware Resource Assignment for DNN Accelerators using Reinforcement Learning," in *Proceedings of 53rd Annual IEEE/ACM International Symposium on Microarchitecture* ***(MICRO)****, Oct. 2020.*

[10] Kyoungmin Kim, In Seo, Wook-shin Han, Jeong-Hoon Lee, Sungpack Hong, Hassan Chafi, Hyungyu Shin, **Geonhwa Jeong**, "TurboFlux: A Fast Continuous Subgraph Matching System for Streaming Graph Data," in *Proceedings of the 44th International Conference on Management of Data* ***(SIGMOD)****, June 2018.*

HONORS AND AWARDS

Scholarship from Kwanjeong Educational Foundation	Sep. 2019 - Current
ISCA'23 Student Travel Grant	June 2023
ISPASS'23 Student Travel Grant	May 2023
ISPASS'23 Best Paper Candidate	May 2023
ASP-DAC'21 Best Paper Candidate	Jan 2021
HPCA'20 Student Travel Grant	Feb. 2020
National Scholarship from ICT Creative Consilience Program	Mar. 2013 - Feb. 2019
Army Achievement Medal from the US Army	Aug. 2017

SKILLS

Programming	C/C++, SystemC, Java, Verilog, Ocaml
Scripting	Python, Javascript, PHP, Perl
Others	PyTorch, Tensorflow, Hadoop, Spark

TEACHINGS

Advanced Computer Architecture (Head TA)	Fall 2023
Advanced Computer Architecture (TA)	Fall 2022

PROFESSIONAL SERVICES

IEEE TCAS'23, TVLSI'23 Reviewer
ISCA'23 Submission Chair
IEEE Micro'22, CAL'22 Reviewer
PPoPP'22 Artifact Evaluation Committee Member
PACT'21 Artifact Evaluation Committee Member