Raveesh Garg

Curriculum Vitae

266 Ferst Drive NW
Klaus Adv Computing Bldg Rm 3305
Atlanta, GA, 30332
🌐 *sites.gatech.edu/raveesh*
in *Raveesh Garg*

Research Interests

Accelerators for Artificial Intelligence and Scientific Computing
Programmable Spatial Accelerators
Computer Architecture

Education

2021-Present **Georgia Institute of Technology**, *PhD in Electrical and Computer Engineering*, Atlanta, GA, USA

- Advisor: *Dr. Tushar Krishna*
- Research Area: *Architecture and Mapping Support for Exploiting Inter-Operation Data Reuse in AI, HPC and Graph Applications on Spatial Accelerators.*
- GPA - *4/4*

2019-2021 **Georgia Institute of Technology**, *Master of Science in Electrical and Computer Engineering*, Atlanta, GA, USA

- Advisor: *Dr. Tushar Krishna*
- Master's Thesis: *Understanding the Design Space of Dataflows for Graph Neural Network Accelerators.*
- GPA - *4/4*

2015-2019 **Birla Institute of Technology and Science, Pilani**, *Bachelor of Engineering in Electronics & Instrumentation Engineering*, Pilani, Rajasthan, India

- GPA - *9.29/10*

Skills

Programming Verilog, C/C++, Assembly Language, Python, MATLAB/Octave

Simulators and EDA Tools gem5 garnet on-chip network simulator, Structural Simulation Toolkit (SST), SESC SuperScalar simulator, Xilinx ISE and Vivado, ModelSim, Icarus iverilog, Cadence Encounter RTL Compiler, SPICE, Cadence Virtuoso, Synopsys Design Vision, Cadence Innovus.

Publications and Pre-prints

IPDPS 2025 *Raveesh Garg*, Michael Pellauer, Sivasankaran Rajamanickam, and Tushar Krishna. "CELLO: Co-designing Schedule and Hybrid Implicit/Explicit Buffer for Complex Tensor Reuse", 39th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2025).

ASPLOS 2023 Francisco Muñoz-Martínez, *Raveesh Garg*, José L. Abellán, Michael Pellauer, Manuel E. Acacio, and Tushar Krishna. "Flexagon: A Multi-Dataflow Sparse-Sparse Matrix Multiplication Accelerator for Efficient DNN Processing", in Proceedings of the 28th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '23,

IPDPS 2022 **(Best Paper Nominee - Top 5/474)** *Raveesh Garg*, Eric Qin, Francisco Muñoz-Martínez, Robert Guirado, Akshay Jain, Sergi Abadal, José L Abellán, Manuel E Acacio, Eduard Alarcón, Sivasankaran Rajamanickam, and Tushar Krishna. "Understanding the Design-Space of Sparse/Dense Multiphase GNN dataflows on Spatial Accelerators", 36th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2022)

arXiv 2025 *Raveesh Garg*, Michael Pellauer, and Tushar Krishna. "HARP: A Taxonomy for Heterogeneous and Hierarchical Processors for Mixed-reuse Workloads."arXiv preprint arXiv:2502.13113(2025)

arxiv 2024 *Raveesh Garg*, Hyoukjun Kwon, Eric Qin, Yu-Hsin Chen, Tushar Krishna, and Linaghzhen Lai, "Pipeorgan: Efficient Inter-operation Pipelining with Flexible Spatial Organization and Interconnects," arXiv preprint arXiv:2405.01736 (2024)

arXiv 2022 Eric Qin, *Raveesh Garg*, Abhimanyu Bambhaniya, Michael Pellauer, Angshuman Parashar, Sivasankaran Rajamanickam, Cong Hao, and Tushar Krishna. "Enabling Flexibility for Sparse Tensor Acceleration via Heterogeneity." arXiv preprint arXiv:2201.08916 (2022).

IEEE INDICON 2018 *Raveesh Garg* and Karri Babu Ravi Teja, "A High-Speed Pipelined Architecture for Block Motion Estimation Using Hexagon- Based Search Algorithm," 2018 15th IEEE India Council International Conference (INDICON), Coimbatore, India, 2018

Internship Experience

May 2024 - Aug 2024 **IBM Research**, *Research Scientist Intern*, Yorktown Heights, New York, USA

- ○ Research Project: Worked on mapping strategies for RaPiD-core compiler.

Aug 2022 - Nov 2022 **Meta Reality Labs**, *Part-time Student Researcher*, (Remote) Atlanta, GA, USA

- ○ Research Project: Mapping exploration for AR/VR DNN workloads on multi-accelerator system with focus on inter-layer pipelining.

May 2022 - Aug 2022 **Meta Reality Labs**, *Research Scientist Intern*, Sunnyvale, California, USA

- ○ Research Project: Cost modeling mappings for AR/VR DNN workloads on multi-accelerator system with focus on inter-layer pipelining.

Research Projects

August 2024 - December 2024 **Characterizing Hierarchical and Heterogeneous Architectures for mixed-reuse applications**

- ○ Proposed a new taxonomy for hierarchical and heterogeneous accelerators and used it to characterize various kinds of architectures for mixed-reuse applications like LLMs. Used Timeloop as the cost model.

May 2023 - November 2024 **Accelerator Microarchitecture for HPC Applications**

- ○ Working on microarchitecture of an accelerator for HPC applications, targeting applications where different operations have low intra-operation reuse.
- ○ Specifically worked on architecture of the hybrid implicit/explicit on-chip buffer mechanism to enable maximum on-chip data reuse across tensor operations.

May 2022 - April 2024 **Flexible inter-operation Pipelining for Energy Efficient DNN Accelerators**

- Developed a cost model to study the memory footprint and latency for various intra-layer and inter-layer CNN and GEMM mappings.
- Worked on mapping strategies for inter-operation pipelining (aka fusion) in DNN accelerators, for edge application domains, for example, AR/VR. Focused on reduction in on-communication between producer and consumer.

Nov 2021 - Oct 2022 **Exploiting Inter-Operation Data Reuse in HPC Applications**

- Proposed a systematic methodology for identifying inter-operation reuse patterns in a complex graph of einsums and to determine the mapping of these einsums.
- Targeted applications like Conjugate Gradient which have complex einsum dependency graphs with SpMM and highly skewed dense GEMMs.

Aug 2021 - July 2022 **Multi-Dataflow Accelerators for Sparse Workloads**

- Contributed to the designing the architecture of a reconfigurable sparse accelerator with a unified engine supporting Inner-product, Outer-product and Gustavson's dataflow.
- Contributed to the designing the architecture for a heterogeneous sparse accelerator with multiple sub-accelerator engines capable of processing multiple dataflows.

Sept 2020 - Oct 2021 (Best Paper Nominee at IPDPS 2022) **Dataflow Design-Space Exploration for GNN Accelerators**

- Proposed a taxonomy for description of dataflows capturing the dataflows of individual phases SpMM and DenseGEMM and pipelined parallelism between the two phases and encoded it into a simulation framework OMEGA.
- OMEGA uses STONNE simulator to model GEMM and SpMM individually and an analytical model to compute pipelined statistics from individual kernel statistics.

Workshop, Tutorials and Talks

ASPLOS 2023 **Tutorial: Enabling Detailed Cycle-Level Simulation of AI and HPC Applications with Detailed Memory Hierarchy using STONNE, OMEGA and SST-STONNE**, *(Co-organizer and Presenter)*

Discussed dataflow design-space exploration of Graph Neural Network mappings and demonstrated the OMEGA framework that models the metrics for GNN dataflows.

ModSim 2022 **Workshop on Modeling & Simulation of Systems and Applications 2022**

SST-STONNE: Enabling cycle-level simulation of flexible spatial accelerators for DNNs and GNNs with a detailed memory hierarchy.

ASPLOS 2022 **Young Architect Workshop 2022**

A Communication-Centric Dataflow Accelerator for High-Performance Conjugate Gradient.

ASPLOS 2022 **Tutorial: STONNE+OMEGA: Cycle-level Simulation of Dense/Sparse DNN and GNN Accelerators**, *(Co-organizer and Presenter)*

Discussed dataflow design-space exploration of Graph Neural Network mappings and a demonstrated the OMEGA framework that models the metrics for GNN dataflows.

SIAM PP22 **Minisymposium: Co-Design of Data Flow Accelerators for Scientific Simulations and Machine Learning**, *(Presenter)*

Discussed the design-space of dataflows for multiphase kernels with sparse and dense computations like GNNs in a minisymposium at SIAM PP22.

Honors and Awards

IPDPS 2022 Best Paper Award Nomination (First authored). Top 5/474 submissions.

Service

IEEE TVLSI 2025 Reviewer for the journal IEEE Transactions on Very Large Scale Integration (VLSI) Systems

ACM TACO 2024 Reviewer for the journal ACM Transactions on Architecture and Code Optimization

ISCA 2023 PC Meeting Student Volunteer for 2023 International Symposium on Computer Architecture (ISCA'23)

HPCA 2022 Artifact Evaluation PC Reviewer for the 2022 IEEE International Symposium on High Performance Computer Architecture (HPCA'22)