ROHAN JUNEJA

COM3, 11 Research Link ◇ Singapore 119391
+65 8132 4894 ◇ rohan@comp.nus.edu.sg ◇ https://rohanjuneja.github.io/

Research Interests

I work at the intersection of VLSI design, computer architecture, and AI, developing hardware-software co-design solutions to address performance and efficiency bottlenecks in the post Dennard era. My research focuses on reconfigurable architectures and GPU-based accelerators for irregular and mixed-precision workloads driven by sparsity and quantization, key enablers of AI model compression. I design LLVM-based compiler toolchains to map compressed models efficiently onto custom accelerators, and validate these designs through real-world chip tapeouts. By advancing performance, energy efficiency, and reliability, my work aims to lower the carbon footprint of modern computing systems.

Education

- **National University of Singapore**
  Ph.D. in Computer Science (CGPA: 4.58/5) *January 2021 - Present*

  Advisors: Prof. Li-Shiuan Peh, Prof. Tulika Mitra

  Thesis: Scalable Architectures for Sparse and Quantized AI Models: Bridging Efficiency and Computational Complexity

- **IIIT Delhi**
  B.Tech in Electronics and Communications Engineering (CGPA: 8.41/10) *May 2014 - May 2018*

  Advisors: Prof. Lam Siew Kei, Prof. Sujay Deb

  Thesis: Securing untrusted memories in embedded systems

Professional Experience

- **IBM Research** Yorktown Heights, New York
  Incoming PhD Research Intern *Oct 2025 - Dec 2025*
- **National University of Singapore** Singapore
  PhD Researcher *Jan 2021 - Present*
  - Designed deep learning edge accelerators for transfer learning and attention-based models, targeting low-power, high-throughput inference at the edge.
  - Contributed to PACE, a Coarse-Grained Reconfigurable Array (CGRA) integrated with a RISC-V Rocket core, achieving 360 GOPS/W. Involved in both hardware design and development of the LLVM-based PACE compiler.
  - Utilized FPGA-based emulation for pre-silicon validation of PACE, enabling functional verification and performance optimization prior to fabrication.
  - Developed reconfigurable architectures optimized for sparsity and multi-precision support, improving scalability and energy efficiency across deep learning workloads.
  - Designed hardware-aware quantization techniques for both TPUs and GPUs, targeting efficient inference of LLM and AI models. Developed CUDA kernels using the CUTLASS library to exploit hardware characteristics and maximize throughput.
- **Advanced Micro Devices** Singapore
  PhD Research Intern *May 2022 - July 2022*
  - Designed an accelerator for Ethereum's Beacon Chain using parallel processing of Tamper-Evident Plan (TEP) trees to optimize transaction validation and execution.
  - Optimized performance modeling and profiling techniques for blockchain workloads.

- **Renesas Electronics Corporation** Singapore
  PhD Engineering Intern *Jan 2022 - April 2022*
  - Worked on Renesas' Dynamically Reconfigurable Processor (DRP), focusing on deep learning acceleration and dynamic architecture reconfiguration.
  - Engineered a smart gym application, leveraging DRP-based acceleration to optimize computational efficiency and responsiveness in embedded fitness solutions.
- **Qualcomm** Bengaluru, India
  CPU Design Engineer *July 2018 - January 2021*
  - Worked as a CPU design engineer for Qualcomm Snapdragon Processors.
  - Delivered multi-clock domain and Low Power (UPF) RTL for ARM Kryo cores in Snapdragon 765G, as well as medium-, high-tier, and compute chips.
  - Responsible for restructuring memory model RTL to support partial power gating.
  - Gained experience with Power Manager IP, DCVS, Low Power Modes using ARM's P-channel, and boot RTL in Snapdragon CPUs.
  - Experienced in writing SystemVerilog assertions, code coverage and functional coverage closure.
  - Experience with C, C++, Python, SystemVerilog, Synthesis flows, performance profiling, and optimizations.

Publication Record

Conferences

1. Nexus Machine: An Active Message Inspired Reconfigurable Architecture for Irregular Workloads *MICRO 2025*
   Rohan Juneja, Pranav Dangi, Thilini Kaushalya, Tulika Mitra, Li-Shiuan Peh
2. Building an Open CGRA Ecosystem for Agile Innovation *ICCAD 2025*
   Rohan Juneja, Pranav Dangi, Thilini Kaushalya, Zhaoying Li, Dhananjaya Wijerathne, Li-Shiuan Peh, Tulika Mitra
3. HALO: Hardware-aware quantization with low critical-path-delay weights for LLM acceleration *AAAI 2026 [Under Review]*
   Rohan Juneja, Shivam Aggarwal, Safeen Huda, Tulika Mitra, Li-Shiuan Peh
4. Reliable and Sustainable Acceleration through Reconfigurable Hardware *ASPLOS 2026 [Under Review]*
   Rohan Juneja*, Pranav Dangi*, Lieven Eeckhout, Tulika Mitra
   *Equal contribution
5. EdgeWizard: A Motion-Resilient EEG Wearable with On-Device Processing Using Reconfigurable Fabric *Sensys 2026 [Under Review]*
   Rohan Juneja*, Teck Lun Goh*, Vishruti Ranjan, Shivam Aggarwal, Zhaoying Li, Tulika Mitra, Li-Shiuan Peh
   *Equal contribution
6. A Data-Driven Dynamic Execution Orchestration Architecture *ASPLOS 2026*
   Pranav Dangi, Zhenyu Bai, **Rohan Juneja**, Zhaoying Li, Zhanglu Yan, Huiying Lan, Tulika Mitra
7. Enhancing CGRA Efficiency through Aligned Compute and Communication Provisioning *ASPLOS 2025*
   Zhaoying Li, Pranav Dangi, Chenyang Yin, Thilini Kaushalya, **Rohan Juneja**, Cheng Tan, Zhenyu Bai, Tulika Mitra
8. ZeD: A Generalized Accelerator for Variably Sparse Matrix Computations in ML *PACT 2025*
   Pranav Dangi, Zhenyu Bai, **Rohan Juneja**, Dhananjaya Wijerathne, Tulika Mitra
9. NOVA: NoC-based Vector Unit for Mapping Attention Layers on a CNN Accelerator *DATE 2024*
   Mohit Upadhyay, **Rohan Juneja**, Weng-Fai Wong, Li-Shiuan Peh
10. FLEX: Introducing FLEXible Execution on CGRA with Spatio-Temporal Vector Dataflow *ICCAD 2023*
    Thilini Kaushalya, Dan Wu, **Rohan Juneja**, Dhananjaya Wijerathne, Tulika Mitra, Li-Shiuan Peh

11. REACT: A Heterogeneous Reconfigurable Neural Network Accelerator with Software Configurable NoCs for Training and Inference on Wearables *DAC 2022*
    Mohit Upadhyay, **Rohan Juneja**, Bo Wang, Jun Zhou, Weng-Fai Wong, Li-Shiuan Peh
12. Cache-Aware Dynamic Skewed Tree for Fast Memory Authentication *ASP-DAC 2021*
    Saru Vig, **Rohan Juneja**, Siew Kei Lam
13. DISSECT: Dynamic Skew-and-Split Tree for Memory Authentication *DATE 2020*
    Saru Vig, **Rohan Juneja**, Siew Kei Lam, Guiyuan Jian
14. Dynamic NoC Platform for Varied Application Needs *ISQED 2018*
    Sidhartha Shankar, Hemanta K. Mondal, **Rohan Juneja**, Sri Harsha Gade, Sujay Deb

Journals

1. CTScan: A CGRA-based Platform for Emulation of Power Side-Channel Attacks on Edge CPUs *TRETS 2025*
   Yaswanth Tavva, **Rohan Juneja**, Trevor E. Carlson, Li-Shiuan Peh
2. Framework for Fast Memory Authentication using Dynamically Skewed Integrity Tree *TVLSI 2019*
   Saru Vig, **Rohan Juneja**, Guiyuan Jiang, Siew Kei Lam, Changhai Ou

Chip Tapeouts

1. A 360 GOPS/W CGRA in a RISC-V SoC with Multi-Hop Routers and Idle-State Instructions for Edge Computing Applications *ISOCC 2024*
   Vishnu Nambiar, Yi Sheng Chong, Thilini Kaushalya, Dhananjaya Wijerathne, Zhaoying Li, **Rohan Juneja**, Li-Shiuan Peh, Tulika Mitra, Anh Tuan Do
2. PACE: A Scalable and Energy Efficient CGRA in a RISC-V SoC for Edge Computing Applications *HotChips 2024*
   Vishnu Nambiar, Yi Sheng Chong, Thilini Kaushalya, Dhananjaya Wijerathne, Zhaoying Li, **Rohan Juneja**, Li-Shiuan Peh, Tulika Mitra, Anh Tuan Do

Patents

1. A Reconfigurable Execution Unit for Collective Routing and Computation of Multiple Operations for Hardware Acceleration *Patent 10202402819X, 2025*
   Rohan Juneja, Zhaoying Li, Pranav Dangi, Tulika Mitra

Teaching Experience

- **National University of Singapore (NUS)**
  Teaching Assistant, Introduction to Operating Systems *January 2023 – April 2023*
  - Independently led weekly tutorial sessions for a cohort of 48 undergraduate students, facilitating an in-depth understanding of core operating systems concepts.
  - Provided detailed feedback on student assignments and exams, fostering continuous improvement and strengthening analytical skills.
- **National University of Singapore (NUS)**
  Teaching Assistant, Introduction to Operating Systems *August 2022 – November 2022*
  - Delivered interactive tutorial sessions tailored to a class of 48 students, emphasizing problem-solving and practical application of OS principles.
  - Supported Prof. Weng-Fai Wong in streamlining assessments, including the design, evaluation, and grading of assignments and examinations.
- **IIIT Delhi**
  Teaching Assistant, GPU Computing *January 2018 – April 2018*
  - Conducted weekly lab sessions on CUDA programming, equipping students with practical skills in GPU computing and parallel programming.
  - Designed and graded hands-on programming assignments to evaluate comprehension and encourage problem-based learning.
  - Collaborated with Prof. Ojaswa Sharma to develop supplemental course materials that bridged theory with real-world GPU applications.