Raja Gond

Research Fellow | Microsoft Research India

rajagond.github.io raja.gond@outlook.com github.com/rajagond Google Scholar

Education

May 2023 | **Indian Institute of Technology Bombay (IIT Bombay)** **Mumbai, MH, India**
Jul 2019 | Bachelor of Technology (Honors) in Computer Science and Engineering CGPA: 8.6/10.0

Preprint and Working Paper

[II] **Samanvaya: Compute-Communication Overlap For Efficient MoE Inference** [Working Paper]
Raja Gond, Prajwal Singhania, Nipun Kwatra, Ramachandran Ramjee

[I] **emucxl: an emulation framework for CXL-based disaggregated memory applications** [PDF, code]
Raja Gond, Purushottam Kulkarni
arXiv preprint arXiv:2404.08311 [2024]

Research Experience

Present | **Microsoft Research | AI Infrastructure** [🌐] **Bangalore, India**
Jul 2023 | *Research Fellow | Advisors: Dr. Nipun Kwatra, Dr. Ramachandran Ramjee*

Samanvaya: Compute-Communication Overlap for Efficient Inference in Mixtures of Experts (MoE) Models

- Proposed a fine-grained overlap method that effectively hides communication costs in **MoE models**
- Implemented **Expert Parallelism in vLLM** and highlighted its benefits over Tensor Parallelism for MoE
- Developed a lightweight signaling mechanism to initiate Direct Memory Access (DMA)-based partial GPU-GPU communication, which frees all SMs to be used by compute kernel and allows effective overlap
- Demonstrated up to a 20% reduction in MoE MLP time for Mixtral 22B in microbenchmarks on 8 H100s
- Working to resolve expert load-balancing issues that hinder our gains in end-to-end performance

Compute-Communication Overlap for Efficient Inference in Dense Large Language Models (LLMs)

- Developed a method that decomposes computation and hides communication in Tensor Parallelism for LLMs, reducing communication overhead by **15%** on GPT-3 microbenchmarks on A100 GPUs with NVLink
- Explored prior overlap solutions to identify issues caused when applying them to new models and GPUs

Sep 2024 | **Virginia Tech | Department of Computer Science** **Remote**
Jul 2024 | *Research Collaborator | Advisor: Prof. Huaicheng Li*

Damon-CXL: Two-tier memory management for Compute Express Link (CXL) memory

- Integrated DAMON-based memory management patches into the linux and reviewed the source code
- Analyzed Redis performance on emulated CXL memory using **YCSB** benchmarks and compared results with vanilla linux memory management configurations to identify improvements and bottlenecks

Jun 2023 | **IIT Bombay | SynerG Lab, Computer Science and Engineering** [🌐] **Mumbai, India**
Jan 2023 | *Undergraduate Researcher | Advisor: Prof. Purushottam Kulkarni*

emucxl: Emulation Framework and Access Library for CXL-Based Disaggregated Memory Systems

- Developed a user-space library coupled with a **NUMA-based CXL emulation backend** for standardized CXL memory access that enables rapid prototyping of disaggregated memory solutions
- Conducted a literature survey on **CXL stds** and showed emucxl capabilities through practical use cases

Aug 2022 | *Undergraduate Researcher | Advisors: Prof. Purushottam Kulkarni, Prof. Umesh Bellur*

R&D Project: Persistent Memory (PMem) Applications [PDF, code]

- Designed and implemented a robust reader-writer program on **Non-Volatile Memory** using advanced array and pointer techniques, which provides fault tolerance and efficient data access
- Explored **Persistent Memory Development Kit** libraries to understand PMem capabilities and analyzed performance differences between traditional and PMem-based **Redis** using real-world benchmarks

Awards

Microsoft Global Hackathon 2023: Executive Challenge First Prize 🎖

- Hack for the Microsoft Cloud in the Era of AI (Idea: Microsoft Confidential) September 2023
  Collaborated closely with the Hackathon teammates spread across global Microsoft offices to develop an innovative solution that enhances cloud infrastructure capabilities and presented it to the **Microsoft Cloud + AI leadership**

Industry Experience

Technology Analyst Intern May'22 - July'22
Investment Management Division, Morgan Stanley

- Designed and implemented a **Java** utility library for translating MT Swift payment messages generated by a trading platform into enriched MX messages, facilitating and streamlining the migration process to new Swift messaging standards
- Integrated MX format verification and conducted in-depth analysis of MT formats, MX equivalents, and translation
- Received an offer for a **full-time position** with the team upon graduation, based on exemplary internship performance

Teaching and Mentorship

Undergraduate Teaching Assistant *Dept. of Computer Science and Engineering, IIT Bombay* May'22 - April'23

- Computer Networks + Lab (CS224/CS252) | Instructor: *Prof. Bhaskaran Raman* Spring'23
  Responsible for evaluating lab assignments, explaining concepts, and resolving doubts for **over 200 CSE sophomores**
- Operating Systems + Lab (CS347/CS333) | Instructors: *Prof. Purushottam Kulkarni, Prof. Umesh Bellur* Fall'22
  Designed and managed lab assignments, addressed student doubts during lab sessions and online, proctored theory and lab exams, and evaluated answer scripts and lab coding assignments, for a batch of **over 180 CSE juniors**
- Computer Systems (Bootcamp) | Instructors: *Prof. Mythili Vutukuru, Prof. Purushottam Kulkarni* Summer'22
  Involved in the design of weekly assignments and asynchronous doubt-solving to aid self-paced learning for students

Department Academic Mentor *Student Mentorship Program, IIT Bombay* July'22 - April'23

- Selected out of **70+** applicants through a rigorous procedure based on Statement of Purpose, interviews, and peer reviews
- Mentored students with academic or general concerns to help ease their transition into the CSE department

Selected Academic Projects

SCLP: Compiler for C-like Language Spring'22
Guide: Prof. Uday Khedker | Implementation of Programming Languages

- Built a compiler to generate Abstract Syntax Tree (AST), Three Address Code, and corresponding assembly Code (ASM)
- Implemented the scanner using **Lex**, the parser using **Yacc** and constructed the object-oriented AST representation in **C++**, enabling the efficient processing of arithmetic and relational expressions, loops, and control flow statements

Custom Shell and Feature Extension of xv6 Operating System Fall'21
Guide: Prof. Mythili Vutukuru | Operating Systems

- Implemented custom shell supporting serial, parallel, and background command execution with signal handling
- Designed and implemented a **priority-based** scheduling algorithm in xv6 that improves the efficiency of task execution
- Enhanced xv6 memory management by integrating **lazy page allocation** to significantly improve memory utilization

Understanding Linux Kernel Internals Through Custom Module Implementation Spring'23
Guide: Prof. Purushottam Kulkarni | Topics in Virtualization and Cloud Computing

- Designed kernel modules to explore **kernel internals** having process listing and heap analysis functionalities
- Enhanced modules to determine kernel stack pointers, map address spaces, and measure memory allocations

3D Visualization and Analysis of Seismic Volumes Spring'23
Guide: Prof. Prabhu Ramachandran | Parallel Scientific Computing and Visualization

- Developed a visualization tool using the **Mayavi** and **TraitsUI** Python libraries for interactive geological analysis
- Enhanced subsurface geological investigation through advanced geophysical analysis and multi-dimensional plotting

Justice System and Prison Overflow Spring'23
Guide: Prof. Om P. Damani | System Dynamics: Modeling & Simulation for Development

- Conducted a literature survey to identify factors contributing to prison overflow and developed a **system dynamics model** to simulate impact on prison population dynamics that provides insights for reforms to mitigate overcrowding

Robust Mastermind Player Spring'21
Guide: Prof. Ashutosh Gupta | Logic for Computer Science

- Formulated and implemented a player for the logic-based game Mastermind using **SAT** solving techniques and the Z3 Theorem Prover, which gives accurate performance even against adversary's inconsistent or unreliable feedback

Talks

Compute and Communication trade-offs for scalable Large Language Models (LLMs)

- Host: Prof. Purushottam Kulkarni, SynerG Lab, IIT Bombay — January 2024

AI-Infrastructure Reading Group, Microsoft Research India Lab

- Flux: Fast Software-based Communication Overlap On GPUs Through Kernel Fusion — August 2024
- Splitwise: Efficient generative LLM inference using phase splitting — April 2024

Other Projects

Network Simulation — Course Project - Spring'21

- Implemented a File Transfer Protocol in **C** and analyzed throughput variations of TCP variants using Wireshark and **NS3**

Online Computing and Development Environment (IDE) — Course Project - Fall'20

- Developed a **Django**-based multi-language online IDE with real-time testing, file storage, and library/package support

Data Prefetchers and Cache Replacement Interaction — Course Project - Fall'21

- Compared cache replacement policies (LRU, Hawkeye) combined with prefetchers (PACMan, IPCP) across various traces

Multi-cycle RISC Processor — Course Project - Spring'21

- Implemented an **8-register, 16-bit** multi-cycle processor with sync write and async read operations in **VHDL**

Real-Time Application Monitor — Course Project - Spring'22

- Developed an app to monitor system resources, with **Telegraf** for data collection and a time-series database for storage

Key Coursework

Systems	Topics in Virtualization and Cloud Computing, Operating Systems, Computer Networks, Parallel Scientific Computing and Visualization, Database and Information Systems, Implementation of Programming Languages, Computer Architecture, Principal of Systems and Data Security, Digital Logic Design, Introduction to GPU Programming (Online)
AI/ML	AI/ML, Foundations of Reinforcement Learning, Automatic Speech Recognition

Technical Skills

Programming	C/C++, CUDA, Python, Java, MATLAB, Bash, SQL, Assembly,
Software & Tools	PyTorch, LaTeX, Git, Lex, Yacc, Mayavi, TraitsUI, ChampSim, NS-3
Tools/Frameworks	HTML, CSS, JavaScript, Angular, Django

Extracurricular Activities

- Completed **80 hours** of community service at Social Development under the National Service Scheme (NSS), IITB — 2020
- Associated with **Parivartan**, an initiative of the NSS, involving writing blogs on sustainable development — 2019
- Awarded the **National Cadet Corps (NCC) 'A'** certificate for completing training in the Junior Division Air Wing — 2017
- Attended the **Annual Training Camp-311, NCC**, which included rigorous physical training, drills, and sports — 2016