Anish **Saxena**

PhD Student · School of Computer Science · Georgia Institute of Technology
1210, Klaus Advanced Computing Building, 266 Ferst Dr NW, Atlanta, GA 30332
(+1) 470-232-5725 | anish.saxena@outlook.com | anish-saxena.github.io | Anish-Saxena | Anish-Saxena

Education

Georgia Institute of Technology — *Atlanta, USA*
Doctor of Philosophy, Computer Science · **GPA: 3.93/4.00** — *2021 - 2026 (expected)*
- Advisor: Prof. Moinuddin K. Qureshi
- Research: Efficient AI, Datacenter Systems, and Hardware Security

Indian Institute of Technology Kanpur — *Kanpur, India*
Bachelor of Technology, Mechanical Engineering · **CPI: 9.1/10.0** — *2017 - 2021*
- Advisor: Prof. Biswabandan Panda
- Minor in Computer Systems

Publications

Conference Papers

CXL-Centric Memory System for Scalable Servers
Albert Cho, **Anish Saxena***, Moinuddin Qureshi, Alexandros Daglis* — **Equal contribution*
To appear in the 36^{th} International Conference for High Performance Computing, Networking, Storage, and Analysis **(SC)**, Atlanta, USA, November 2024.

ImPress: Securing DRAM Against Data-Disturbance Errors via Implicit Row-Press Mitigation
***Anish Saxena**, Aamer Jaleel, Moinuddin Qureshi*
To appear in the 57^{th} IEEE/ACM International Symposium on Microarchitecture **(MICRO)**, Austin, USA, November 2024.

Rubix: Low Overhead Secure Rowhammer Mitigations via Randomized Memory Mapping
***Anish Saxena**, Saurav Mathur, Moinuddin Qureshi*
29^{th} ACM International Conference on Architectural Support for Programming Languages and Operating Systems **(ASPLOS)**, San Diego, USA, April 2024.

START: Scalable Tracking for Any Rowhammer Threshold
***Anish Saxena** and Moinuddin Qureshi*
30^{th} IEEE International Symposium on High-Performance Computer Architecture **(HPCA)**, Edinburgh, Scotland, March 2024.

PT-Guard: Integrity-Protected Page Tables against Breakthrough Rowhammer Attacks
***Anish Saxena**, Gururaj Saileshwar, Jonas Juffinger, Andreas Kogler, Daniel Gruss, Moinuddin Qureshi*
53^{rd} IEEE/IFIP Conference on Dependable Systems and Networks **(DSN)**, Porto, Portugal, June 2023.

AQUA: Scalable Rowhammer Mitigation by Quarantining Aggressor Rows at Runtime
***Anish Saxena**, Gururaj Saileshwar, Prashant Nair, Moinuddin Qureshi*
55^{th} Annual IEEE/ACM International Symposium on Microarchitecture **(MICRO)**, Chicago, USA, October 2022.

Journal Paper

Scalable Multi-node Fast Fourier Transform on GPUs
*Manthan Verma, Soumyadeep Chatterjee, Gaurav Garg, Bharatkumar Sharma, Nishant Arya, Shashi Kumar, **Anish Saxena**, Mahendra K. Verma*
4^{th} Edition of **SN Computer Science**, 625, Springer Nature Singapore, 2023.

Workshop Paper

DABANGG: A Case for Noise Resilient Flush-Based Cache Attacks
Anish Saxena *and Biswabandan Panda*
16^{th} IEEE Workshop on Offensive Technologies **(WOOT)**, San Francisco, USA, May 2022.

Under Review

Citadel: Rethinking Memory Allocation to Prevent Inter-Domain Rowhammer Exploits
Anish Saxena*, Walter Wang, Alexandros Daglis*
Submitted to the 52^{nd} Intenational Symposium on Computer Architecture **(ISCA)**, Tokyo, Japan, June, 2025.

Honors

2024 **NVIDIA Graduate Fellow**, Awarded to 10 students worldwide, 600 applications *USA*
2019 **Semiconductor Research Corporation (SRC) Member**, Sole undergrad from India *India*
2017 **Aditya Birla Group Scholar**, Awarded to 15 students from IIT and BITS *Mumbai*
2017 **All India Rank 1828**, Joint Entrance Examination Advanced, 175,000 students *India*
2017 **KVPY Fellowship**, Awarded by IISc Bangalore and Government of India *Bangalore*

Experience

Future Architectures & Systems Lab, Georgia Tech *Prof. Moinuddin K. Qureshi*
Graduate Research Assistant *Aug. 2021 - present*
- Leveraging computer architecture and systems to tackle challenges in LLM serving, datacenters, and security.
- Published top-tier first-author papers (MICRO, HPCA, ASPLOS, SC) and delivered seminars and invited talks.

NVIDIA Research, USA *Dr. Po-An Tsai*
Research Intern, Architecture Research Group *May. 2024 - Aug. 2024*
- Developed roofline model to analyze serving of Mixture-of-Expert Large Language Models (MoE-LLMs).
- Improved throughput of MoE-LLMs by $3\times$ in long context serving scenarios by restricting set of usable experts.
- Designing scheduling techniques to preserve model quality when using adaptive Branch-Train-Mix models.

AMD Research, USA *Dr. Shaizeen Aga*
Research Intern, Applications-driven Architecture Team *May. 2023 - Aug. 2023*
- Developed roofline model to analyze training of Large Language Models (LLM) on thousands of GPUs.
- Optimized training efficiency for systems with varying compute, capacity, and bandwidth design points.
- Designed a DDR-centric GPU design with better LLM training scalability compared to HBM-centric designs.

Micron Technology, USA *Patrick Estep*
Research Intern, Advanced Memory Group *May. 2022 - Aug. 2022*
- Architected CXL-enabled memory systems to accelerate performance of Google's datacenter workloads.
- Developed a multi-core emulation mechanism to replay datacenter workload traces on native hardware.

NVIDIA Corporation, India *Bharatkumar Sharma*
HPC GPU Advocate Intern, OpenACC and GPU Hackathons Team *May 2021 - Aug. 2021*
- Created multi-node GPU programming guides for HPC with topology-aware communication and profiling.
- The open-source tutorials and code are accessible at github.com/openhackathons-org/nways_multi_gpu/.

Intel Labs, India *Anant Nori*
Research Intern, Processor Architecture Research Lab *May 2020 - Sep. 2020*
- Improved performance of non-inclusive caches by extending reuse-distance based cache policies.
- Accelerated in-house cycle-level simulator and collected memory traces to perform functional simulations.

CAR3S Group, IIT Kanpur — *Prof. Biswabandan Panda*
Group Member and SRC Student Member — *Apr. 2019 - Jun. 2020*
- Devised DABANGG (WOOT'22), refinements that enable accurate and noise-resilient cache attacks.
- Developed tracing techniques for ARM architectures and analyzed cache compression for mobile SoCs.

New York Office, IIT Kanpur — *Prof. Manindra Agrawal*
Computer Systems Intern — *May 2018 - Jul. 2018*
- Led a team of 4 to develop the infrastructure stack of a multi-node microservices-based Kubernetes cluster.
- Configured Spinnaker-based CI/ CD pipeline and integrated canary analysis and stress testing capabilities.

Patents

[Submitted] Efficient Long-Context Mixture-of-Expert LLM Serving — *Dr. Po-An Tsai*
NVIDIA Research — *Aug. 2024*
- Proposed a serving framework to maximize weight reuse in expert GEMMs by restricting set of usable experts.

[Submitted] Tiered Memory for Dynamic KV-Cache Scheduling — *Dr. Aamer Jaleel*
NVIDIA Research — *Aug. 2024*
- Proposed a dyanamic KV recompute-vs-storage scheme to maximize TTFT for tiered memory hierarchies.

[Submitted] Efficient Materialized Views in CXL-enabled Databases — *Dr. Pratik Mishra*
AMD Research — *Aug. 2023*
- Proposed a CXL-centric infrastructure which reduces compute in databases by reusing common sub-queries.

Talks

2024 **CoaXiaL: CXL-centric Server Systems**, paper talk at 36^{th} SC conference — *Atlanta*
2024 **ImPress: Implicit RowPress Mitigation**, paper talk at 57^{th} MICRO conference — *Austin*
2024 **High-Throughput Mixture-of-Expert LLM Serving**, end-of-intern talk at Nvidia — *Westford*
2024 **Rubix: Randomized Memory Mappings**, paper talk at 29^{th} ASPLOS conference — *San Diego*
2024 **START: Scalable Rowhammer Tracking**, paper talk at 30^{th} HPCA conference — *Edinburgh*
2024 **Scalable Rowhammer Protection at Ultra-low Cost**, invited guest talk — *IIT Bombay*
2023 **Memory System Design for Scalable LLM Training**, end-of-intern talk at AMD — *Santa Clara*
2023 **PT-Guard: Integrity-Protected Page Tables**, paper talk at 53^{rd} DSN conference — *Portugal*
2023 **Secure and Scalable Rowhammer Defenses**, invited guest talk at SAFARI Group — *ETH Zurich*
2022 **AQUA Rowhammer Mitigation**, paper talk at the 55^{th} MICRO conference — *Chicago*
2022 **Rowhammer Attacks and Defenses**, invited guest talk — *IIT Bombay*
2022 **Analyzing Google Datacenter Workload Traces**, invited talk at Intel Labs — *Bangalore*
2022 **Replaying Google Datacenter Workload Traces**, end-of-intern talk at Micron — *Dallas*
2022 **DABANGG Attack**, paper talk at the 16^{th} WOOT workshop — *San Francisco*
2021 **CUDA Programming**, guest lecture, course on High Performance Computing — *IIT Kanpur*
2020 **Noise-resilient Flush Attacks**, CAOS reading group to students and faculty — *IIT Kanpur*
2020 **Microarchitectural Security**, talk and demo at SRC Annual Design Review — *Bangalore*
2019 **Flush-based Attacks**, guest lecture, course on Secure Memory Systems — *IIT Kanpur*

Mentorship

Walter Wang — *Georgia Tech*
PhD Student — *Aug. 2023 - present*
- Developing a new memory allocator in Linux that is resilient to inter-domain Rowhammer exploits.

Caroline Huang — *Georgia Tech*
PhD Student — *Aug. 2023 - Nov. 2024*
- Designed adaptive Refresh Management (RFM) scheduling policies that scale to low Rowhammer threshold.

Hritvik Taneja — *Georgia Tech*
PhD Student — *Jan. 2024 - Apr. 2024*
- Analyzed impact of quantization on error propagation in privacy-preserving Fully Homomorphic Encryption.

Saurav Mathur — *Georgia Tech*
PhD Student — *Jan. 2023 - Apr. 2023*
- Developed randomized memory mapping to minimize hot-rows, enabling scalable Rowhammer mitigations.

Sibi Renganatth Sudhakar — *Georgia Tech*
MS Student → ASIC Low Power Designer at NVIDIA — *Jan. 2023 - Apr. 2023*
- Analyzed memory reservation primitives to enable isolation-based protection against Rowhammer attacks.

Service

2024	**IEEE Computer Architecture Letters**, Reviewer	*GaTech*
2024	**SCS Graduate Student Association**, Co-Chair of Events, Co-Chair of Faculty Affairs	*GaTech*
2023	**Secure and Reliable Computer Architecture**, Head TA	*GaTech*
2023	**SGA Graduate Student Wellness Committee**, Lead, Bill of Rights Team	*GaTech*
2023	**HPCA 2024 Conference**, Artifact Evaluation Committee	*Edinburgh*
2023	**Secure and Reliable Computer Architecture**, Head TA	*GaTech*
2022	**Introduction to Quantum Computing**, Course Development TA	*GaTech*
2020	**Systems Reading Group**, Leader	*IIT Kanpur*
2019	**Programming Club**, Coordinator	*IIT Kanpur*

Coursework

- Advanced OS: Distributed Systems^{A}
- High Performance CompArch^{A}
- High Performance Computing^{A*}
- Topics in OS: Programming NVM
- Computer Organization^{A}
- Introduction to Economics^{A*}
- Machine Learning^{A}
- Parallel CompArch^{A}
- High Performance Programming^{A}
- Operating Systems^{A}
- Data Structures & Algorithms
- Financial Markets^{o}
- Secure & Reliable CompArch^{A}
- Computer Architecture^{A*}
- Modern Cryptology^{A}
- Quantum Computing^{A}
- Non Classical Logic
- Linear Algebra

A^{*}: grade for exceptional performance *A: grade* *o: online course*

Miscellaneous

- Mentored MS and PhD students on technical projects and advised them on navigating grad school. *2021-present*
- Gave an invited talk, as a "pioneering alumni" of IIT Kanpur, on research career opportunities in CS. *2023*
- Mentor to 15 students at IIT Kanpur; helped them navigate academic and career challenges in college. *2020*
- Represented CAR3S group in departmental seminars and maintained the group's digital presence. *2019, 2020*