Avinash Kumar

avinkumar@utexas.edu • https://avinkumarut.github.io
linkedin.com/in/avinkumar2020/

Education

The University of Texas at Austin **Aug 2025 – May 2028**
PhD in Electrical and Computer Engineering, CGPA: 4.0/4.0
Focus: Computer Architecture, Computer Systems and Embedded Systems
Amazon AI PhD Fellowship
Professor Edith Clarke Fellowship

The University of Texas at Austin **Aug 2023 – May 2025**
Masters of Science in Electrical and Computer Engineering, CGPA: 4.0/4.0
Focus: Computer Architecture, Computer Systems and Embedded Systems
George J. Heuer, Jr. Ph.D. Endowed Graduate Fellowship

National Institute of Technology, Karnataka **Jul 2016 – Jun 2020**
Bachelor of Technology in Electrical and Electronics Engineering, CGPA: 9.73/10
Salutatorian in Electrical Engineering for academic excellence.

Publications

- **Avinash Kumar**, Sujay Sanghavi, and Poulami Das, "Test-Time Speculation", *Under Review* at Conference on Neural Information Processing Systems (***NeurIPS***) 2026.
- **Avinash Kumar**, Sujay Sanghavi, and Poulami Das, "HiSpec: Hierarchical Speculative Decoding for LLMs", *Under Review* at Conference on Neural Information Processing Systems (***NeurIPS***) 2026.
- **Avinash Kumar** and Poulami Das, "ExSEL: Adaptive Expert Selection For Efficient Inference With Mixture-of-Experts", *Under Review* at International Conference on Architectural Support for Programming Languages and Operating Systems (***ASPLOS***) 2027.
- **Avinash Kumar**, Shashank Nag, Jason Clemons, Lizy John, and Poulami Das, "HELIOS: Adaptive Model And Early-Exit Selection for Efficient LLM Inference Serving", *Accepted* at Conference on the Machine Learning and Systems (***MLSys***) 2026.
- Shvetank Prakash, Andrew Cheng, ... **Avinash Kumar**, ... Amir Yazdanbakhsh, Vijay Janapa Reddi, "QuArch: A Benchmark for Evaluating LLM Reasoning in Computer Architecture", *Accepted* at International Conference on Machine Learning (***ICML***) 2026.
- Ravi Ghadia, **Avinash Kumar**, Gaurav Jain, Prashant Nair, Poulami Das, "Dialogue Without Limits: Constant-Sized KV Caches for Extended Responses in LLMs", *Accepted* at the International Conference on Machine Learning (***ICML***) 2025.
- **Avinash Kumar**, Eneet Kaur, Poulami Das, "Understanding Real Time Decoding for Photonic Quantum Computers", *Accepted* at the NSF Workshop on Quantum Operating Systems and Real-Time Control at the International Symposium on Microarchitecture (***MICRO***) 2024.
- CA Valliappan, **Avinash Kumar**, Renuka Mannem, GR Karthik, Prasanta Kumar Ghosh, "An Improved Air Tissue Boundary Segmentation Technique for Real Time Magnetic Resonance Imaging Video Using Segnet" *Accepted* at the International Conference on Acoustics, Speech and Signal Processing (***ICASSP***) 2019.

Awards and Achievements

- 🏆 **ArchReasoning Challenge:** Winner of the *Persistent Challenge* at ***ISCA*** 2025, awarded for eliciting the highest failure rate across state-of-the-art AI reasoning models on computer architecture tasks.
- **Student Travel Grant:** $1,100 awarded for project proposal submitted to the gem5 (an architecture simulation tool) bootcamp from the University of California, Davis.
- **Student Travel Grant:** $400 awarded by the Yale Quantum Institute for accepted paper and poster in the NSF Workshop on Quantum Operating Systems and Real-Time Control at the International Symposium on Microarchitecture (***MICRO***) 2024.
- **Summer Research Fellowship** from the Indian Academcy of Science (IAS), Indian National Science Academy (IANS) and the National Academy of Sciences, India (NASI), Summer 2018.
- **Institute of Engineers' Certificate of Merit**, awarded annually to the top-ranked student in Electrical Engineering at NIT Karnataka, Surathkal (2017–2018).
- **Best Performer Award:** Recognized at NVIDIA, for ASIC-I (2021), II (2022) and III (2023) levels.

Research Experience

Research Associate, Adavanced Micro Devices "AMD" **May 2025 – April 2026**

Advisor: Dr. Bobbie Manne

- Characterized the fine-grained power-profile of LLMs across different phases and model architectures.
- Designed a power-excursion management framework that selectively boosts phases of LLM execution where additional power headroom yields tangible performance gains.

Research Associate, Adavanced Micro Devices "AMD" **Aug 2024 – May 2025**

Advisor: Dr. Jagadish Kotra

- Investigated model aware hardware optimizations to accelerate LLMs, while supporting confidential computing features (such as integrity checks) on future AMD Instinct GPUs.
- Characterized address-translation overheads for LLMs, and Recommendation Models, and proposed hardware solutions to improve overall system efficiency.

Graduate Research Assistant, University of Texas at Austin **Jan 2024 – Present**

Advisor: Prof. Poulami Das

- Developed **Test-Time Speculation**, an online distillation framework that addresses acceptance-length degradation in speculative decoding for long-response tasks by continuously adapting the draft model; in collaboration with Prof. Sujay Sanghavi.
- Developed **HiSpec**, a hierarchical speculative decoding framework that leverages intermediate layers of early-exit models for low-overhead verification, addressing the verification bottleneck in speculative decoding; in collaboration with Prof. Sujay Sanghavi.
- Collaborated with the Edge Computing group led by Prof. Vijay Reddi to create **QuArch**, a dataset of 2,600+ problems establishing the first comprehensive benchmark for LLM reasoning in computer architecture.

Industry Experience

Senior GPU Architect, NVIDIA **Jul 2020 - Aug 2023**

Manager: Raghavendra Bhat

- Collaborated with cross-functional teams to analyze and optimize GPU performance and power efficiency.
- Evaluated system trade-offs and configurations for AI features such as deep learning upscaling and frame generation using cycle-level GPU modeling.

Skills

- **Programming Languages:** Python, C, C++, CUDA, Bash.
- **Tools & Frameworks:** Gem5, Intel PIN, Nvidia - Nsight, AMD - ROCProfiler, PyTorch, vLLM.
- **Others:** P4 (perforce), Github, Microsoft Excel, GPUView, Omniperf, Nvidia-Smi, Amd-Smi.

Teaching Experience

Graduate Teaching Assistant, University of Texas at Austin **Jan 2024 - May 2024**

Instructor: Prof. Poulami Das

- Head teaching assistant for **Computer Architecture - ECE 460N**.
- Facilitated teaching sessions for a senior undergraduate course with **70+ students**, covering fundamental topics in computer architecture including instruction set architectures (ISA), microcode, pipelining, branch prediction, and memory systems.

Mentoring Experience

University of Texas at Austin **Aug 2023 - Present**

- Invited to the **Graduate student panel** for ECE 107H student Leadership seminar at UT Austin.
- Mentoring two graduate students (Ravi Ghadia. Sourish Wawdhane) on systems research.

Recent Academic Projects

Prefetch filtering for Server workloads **Jan 2024 - Apr 2024**

Advisor : Prof. Calvin Lin

- Developed a prefetch filter for the Triage prefetcher using Google's memory trace workloads. Our filter was able to improve memory bandwidth by up to 22% while improving LLC miss rate by ∼0.4%.
- Simulations were performed using DynamoRIO, Google's dynamic instrumentation tool.

Scalable GPU benchmarking using N-Body simulation **Jan 2024 - Apr 2024**

Advisor : Prof. Lizy JOhn

- Utilized N-body simulation to develop a dynamic (timeless) GPU benchmark, which had both varying compute and varying time unlike traditional benchmarks which have fixed compute and variable time.
- Unlike CPU HINT, our benchmarking (evaluated on NVIDIA A100, H100 and RTX 5000) also stress-tested the GPU for control instructions, whereas CPU HINT focused mainly on memory and compute capabilities.