Vima Gupta

vimagupta.github.io

Email: vgupta345@gatech.edu
+14703349450 linkedin.com/in/vima-gupta

Education

Georgia Institute of Technology — Atlanta, GA
PhD Computer Science, specializing in Systems for AI; advised by Dr. Anand Iyer — *Aug'23-Dec'27*
M.S. Computer Science, specializing in Computing Systems (thesis track) — *Jan'21-May'23*
Relevant Coursework: Systems for Machine Learning, Advanced Operating Systems, Statistical Machine Learning

Birla Institute of Technology and Science (BITS), Pilani — Pilani, India
Bachelors of Engineering in Electrical & Electronics Engineering — *Aug'14-May'18*

Publications

- **Vima Gupta**, Oytun Kuday Duran, Nandan Suresh Meda, Ikhyun An, Ganesh Ananthanarayanan, Anand Padmanabha Iyer, "SANDHI: Fine-grained Merging for Memory Efficient Multi-Model Serving", **SOSP 2026** [Paper]
- Muyan Hu, Ahan Gupta, Jiachen Yuan, **Vima Gupta**, Taeksang Kim, Xin Xu, Janardhan Kulkarni, Ofer Dekel, Vikram Adve, Charith Mendis, "VTC: DNN Compilation with Virtual Tensors for Data Movement Elimination," **OSDI 2026**. [Paper]
- **Vima Gupta**, Jae Hyung Ju, Kartik Sinha, Ada Gavrilovska, Anand Padmanabha Iyer, "LYNX - Efficient MoE inference for workload-agnostic LLM serving" (Under Review) [Paper]
- Joel Mathew Cherian, Ashutosh Muralidhara Bharadwaj, **Vima Gupta**, Anand Padmanabha Iyer, "CHAI: CacHe Attention Inference for Text2Video." (Arxiv'26) [Paper]
- Ziyi Yang, Kalit Inani, Keshav Kabra, **Vima Gupta**, Anand Padmanabha Iyer, "SAFuzz: Semantic-Guided Adaptive Fuzzing for LLM-Generated Code." (Arxiv'26) [Paper]
- **Vima Gupta** and Sashank Varma, "Understanding Infinity: Neural Network Models of Becoming a Cardinal Principle Knower" (**CogSci '24**) [Paper]
- **Vima Gupta** and Sashank Varma, "Learning to count: a neural network model of the successor function" Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44, 2022. [Poster]
- **Vima Gupta** and Rahul Singhal, "Performance analysis of a visible light vehicle-to-vehicle wireless communication system" 2019, International Conference on Microwave Integrated Circuits, Photonics and Wireless Networks (IMICPW). IEEE, 2019 **(Best Paper Award)** [Paper]

Research Internships

Autonomous Systems Group, Nokia Bell Labs — *Jun'26 – Aug'26*
Research Advisors: Dr. Manzoor A. Khan, Buvana Ramanan;

- Built a system that harnesses spare capacity in RAN accelerators through workload-aware sharding of MoEs.

Systems Research Group, Microsoft Research — *May'25 – Aug'25*
Research Advisor: Dr. Roshan Dathathri, Dr. Shan Lu;

- Developed adaptive pool-level KV-aware scheduler for multi-instance, multi-region LLM inference, optimizing SLO attainment and GPU utilization across agentic workloads, multi-turn conversation graphs.

MAIA Frameworks Team & Microsoft Research — *May'24 – July'24*
Research Advisors: Dr. Ondřej Čertík, Dr. Janardhana Kulkarni, Dr. Abhinav Jangda;

- Designed adaptive CUDA kernel system that dynamically tunes launch parameters, reducing latency by up to **30%** versus static heuristics on SOTA flash-attention kernels validated on large-scale GPU clusters like H100s and A100s.

ML Compilers Group, Cerebras Systems — *May'22 – Jul'22*
Research Advisor: Ke Deng

- Compiler transformation for Wafer-scale inference chip through compile-time MLIR graph rewrite optimizations.

Research Experience

Designing Efficient AI Inference Systems — *Aug'23 – Now*
Advised by Prof. Anand Iyer, NEXS Lab — Atlanta, GA

- Optimizing resource and energy usage in agentic systems through hardware-harness codesign for local coding agents.
- Designing systems for per-model and multi-model inference pipelines by alleviating GPU memory bottlenecks.

MALT: Qubit Mapping and Routing for Shuttling-based Quantum Computers — *Jan'22 – May'23*
Master's Thesis (GRA); Advised by Dr. Thomas Conte, Tinker Lab — Atlanta, GA

- Extended MaxSAT techniques to improve the robustness and latency for quibit mapping and routing algorithms for shuttling-based trapped-ion quantum computer with dynamically evolving connectivity. (Paper under review)

Work Experience

PACE: Physical Activity and Care for Everyone *May'21 – Dec'21*
Co-founder, CREATE-X Atlanta, GA

- Developed a real-time exercise-feedback system for Android using Google's Mediapipe, including pose-detection models, error-tolerance heuristics, and extensible exercise-library interface for remote physical-training applications.
- Performed structured user-needs and market analysis to guide system design; built the product website and co-developed the cross-platform mobile prototype (iOS/Android) showcased during the CREATE-X incubator demo.

Arm Embedded Technologies *May'18 – Dec'20*
Design Engineer Bengaluru, India

- Led a sub-team of three interns to design an IoT subsystem for the open-source ecosystem. Synthesis, floorplanning and PnR for high performance cores, ultra low power machine learning accelerators and octa-core clusters.
- Youngest engineer selected consecutively for technical poster presentation at Arm's Global Engineering Conference.

Achievements and Extra-curricular activities

- Awarded the **CRNCH Fellowship** (Center for Research into Novel Compute Hierarchies) for 2026-27 cycle.
- Received the MLSys'26, OSDI'26, SOSP'26 student grant to attend conference in Seattle USA, Prague, Czechia.
- Received the OSDI'24 travel grant to attend USENIX OSDI 2024, ATC 2024 conference in Santa Clara, California.
- Won 3rd position at Klaus Poster Symposium, held across College of Computing, Georgia Institute of Technology.
- Awarded the **EDIC fellowship** (declined) at EPFL, Lausanne, one among fifty candidates selected across the world.
- Awarded the **Adobe Research Women in Technology scholarship** 2022 from candidates across North America
- Student Organizations: Secretary at Quantum computing Association (2021), India Club Finance Leader (2021), English Drama Club Co-ordinator (2016-2017), Logistics head at Department of Controls (2015-2017).
- Awarded bronze medal for basketball in BITS Open Sports Meet, 2015
- Awarded 'Most Outgoing Student of the Year' in high school, 2012
- Secured All India 3rd rank a national quizzing competition, 'Kaho What's My Idea' hosted by Derek'O'Brien, 2011

Talks

- **SANDHI: Fine-grained Merging for Memory Efficient Multi-Model Serving**, Paper Talk, SOSP'26
- **Repurposing RAN Edge Accelerators for MoE Inference** End-of-Internship Talk, Nokia Bell Labs
- **HEDGE: Resource-Efficient and Lossless Expert Reduction for MoE Inference** Poster, OSDI'26
- **Efficient MoE Inference via expert-remapping**, Invited Talk, System Innovation Group, Microsoft
- **LYNX: Workload-Agnostic Expert Remapping for Efficient MoE Inference** MLSys'26 (YPS)
- **Alleviating Memory Capacity Bottlenecks in Multi-Model Serving**, Systems Seminar, Georgia Tech
- **Adaptive Scheduling for Multi-model LLM Serving** End-of-Internship Talk, Microsoft Research
- **MoE Inference for Low-Latency Serving** Systems Research Group, Microsoft Research
- **Kairos: Adaptive Kernel Dispatch for Latency-Sensitive LLMs**, End-of-Internship Talk, Microsoft Research
- **Panini: Context-aware Waste-minimizing Speculative Decoding for LLMs**, CoC Poster Symposium
- **Effective qubit mapping, routing and scheduling for trapped-ion architectures**, CRNCH Summit
- **Multi-core jigsaw puzzle, floorplanning and PnR**, Global Engineering Conference 2020, Arm
- **A novel approach to clocking Arm's bigger cores**, Global Engineering Conference 2019, Arm Birmingham

Mentorship

Oytun Kuday Duran, Ikhyun An, Nandan Meda Suresh *May'25 – present*
PhD Student, Master's Students

- Developed noise-sensitive heuristics to enable merging of co-located LLMs and alleviate capacity bottlenecks.

Jae Hyung Ju, Kartik Sinha *Jan'24 – Dec'25*
PhD Student, Master's Student

- Co-designed scalable expert reduction techniques to accelerate low-latency MoE serving.

Skills and Teaching Experience

Programming skills – C++, Python, C, OpenMP, OpenMPI, Assembly, MATLAB

Python Libraries and software suites – PyTorch, Numpy, Matplotlib, Tensorflow, Streamlit, Qemu, Libvirt, Vtune

Graduate Teaching Assistant — Head GTA (CS8803) for advanced graduate research-based class on Systems for AI managing supervising eight research projects. Held office hours and delivered tutorials on LLM fundamentals and data visualization; CS 6476 Computer Vision (OMSCS): Designed and graded assignments for 500+ students.