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 ML; advised by Dr. Ada Gavrilovska and Dr. Anand Iyer
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

- **V. Gupta**, K. Sinha, A. Gavrilovska, A. Iyer "Lynx: Enabling Efficient MoE Inference through Dynamic Batch-Aware Expert Selection" (Under submission) [Paper]
- **V. Gupta**, A. Austin, E. Pinto, J. Young and T. Conte, "Effective qubit mapping routing and scheduling for Trapped-Ion shuttling architectures" (Under submission)
- **V. Gupta** and S. Varma, "Understanding Infinity: Neural Network Models of Becoming a Cardinal Principle Knower" (CogSci '24) [Paper]
- **V. Gupta** and S. 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]
- **V. Gupta** and R. 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 Experience

Lynx: Enabling Efficient MoE Inference through Dynamic Batch-Aware Expert Selection *Jan'24 – Nov'24*
Research Advisors: Dr. Anand Iyer, Dr. Ada Gavrilovska

- Designing mechanisms for pushing the pareto-frontier for accuracy (quality of output generated across tasks) versus system footprint (latency, throughput and KV cache memory) trade-offs for MoE based large language models for online serving.

Kairos: Adaptive Kernel Dispatch for Latency-Sensitive LLM Workloads *May'24 – July'24*
Research Advisors: Dr. Ondrej Certik, Dr. Janardhana Kulkarni, Dr. Abhinav Jangda; Microsoft Research

- Existing frameworks like Flash Attention use kernel dispatch parameters to avoid intractable grid search over a large input space, resulting in up to 10-30% latency overhead.
- Kairos aims to overcome this challenge by adaptively selecting kernel dispatch parameters, enabling more efficient and scalable performance optimization for complex CUDA kernels.

Galadriel: High throughput speculative decoding with latency bounds *Aug'23 – Dec'23*
Research Advisors: Dr. Kexin Rong, Dr. Alexey Tumanov

- Designing scheduling policies for speeding up LLM inference for lower latency request and while gaining higher throughput.
- Understanding how small models can serve as auto-complete for LLM deployment by increasing its semantic awareness.

Work Experience

Microsoft Research *May'24 – July'24*
Research Intern, AI Frameworks Redmond, WA

- Kernel optimizations for attention mechanism through adaptively determining optimal dispatch parameters across LLM serving techniques on latest Llama architecture models.
- Showcased latency gains upto 35% for SOTA Flash Attention 2 kernel on A100 GPUs across context lengths.

Cerebras Systems *May'22 – July'22*
ML Frameworks Intern, Backend Atlanta, GA

- Converted the block sparse attention graph in BigBird, an NLP transformer, to match with existing highly optimized full attention kernel, from Tensorflow to MLIR lowering, at compile time for improved performance.

- The transformation was implemented through an MLIR graph match and rewrite pattern, automated in C++.

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

- Developed an exercise library for Android application to enable remote physical training using Google's Mediapipe to give real-time feedback through pose detection.
- Conducted market research and designed the product website, and contributed towards iOS and Android application development towards our demo for CREATE-X, start-up incubator.

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 in a customer facing role.
- Youngest engineer selected consecutively to present innovative work on system design at Arm's Global Engineering Conference.

Achievements and Extra-curricular activities

- Won 3rd position at Klaus Poster Symposium, held across College of Computing, Georgia Institute of Technology.
- Awarded the **EDIC fellowship** 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

Skills and Teaching Experience

Programming skills – C++ (DSA and OOP), Python, C, OpenMP, OpenMPI, Assembly, MATLAB, Agile practices
Python Libraries and software suites – PyTorch, Numpy, Matplotlib, Tensorflow, Streamlit, Qemu, Libvirt, Vtune
Graduate Teaching Assistant – Computer Vision (OMSCS 6476): Designed and graded assignments for a class of 500+ students.