Sankeerth Durvasula

sankeerth@cs.toronto.edu @Kwaehp linkedin
https://www.cs.toronto.edu/~sankeerth/

Research Interests

Computer Architecture, Deep Learning, Visual Computing/Graphics, Systems for Deep Learning Acceleration

Education

2020 – · · · · ◇ **Ph.D. Computer Science** GPA: *3.8/4.0*
University of Toronto
Supervisor: Prof. N. Vijaykumar

2017 – 2018 ◇ **Masters in Tech., Electrical Engineering** cGPA: *9.16/10*
Indian Institute of Technology Madras
Supervisor: Prof. V. Kamakoti

2013 – 2017 ◇ **Bachelors in Tech., Electrical Engineering** cGPA: *9.16/10*
Indian Institute of Technology Madras

Work Experience

2021 – · · · · ◇ **Vector Institute**, Toronto Canada
Research Affiliate

Jul-Nov 2024 ◇ **NVIDIA Research, Architecture Research Group (ARG)**, Austin USA
Research Intern
- Worked on acceleration of dictionary-compressed large language models
- Built a strong baseline for speeding up weight-matrix dequantization step that is inefficient in modern GPUs
- Achieved speedup of up to 2.9X on compressed-LLM inference

2020 - ◇ **Teaching Assistantship**
- CSC258: Computer Organization - University of Toronto
- CSC2231: Topics in Visual and Mobile Computing Systems - University of Toronto

2018 – 2020 ◇ **Goldman Sachs Pvt. Ltd.**,
Senior Analyst, Global Compliance

Skills

Programming Languages/Frameworks: Writing performant code in C++, CUDA, C, JAX, PyTorch

Talks

- ◇ Presented *Towards Achieving Speed-of-light inference Speeds on Dictionary Compressed LLMs* Architecture Research Group, Nvidia Research, 2024.
- ◇ Presented *ACE*: *Automatic Concurrent Execution of GPU kernels*, at the Programmable Architectures and Compilation Techniques (PACT) conference, 2024.

Talks (continued)

- Presented *Distributed Training of Neural Radiance Fields*: *A Performance Characterization* at the International Symposium on Performance Analysis of Systems and Software (ISPASS) conference, 2024.
- Presented *EvConv*: *Fast cnn inference on event camera inputs for high-speed robot perception*, at the International Conference on Robotics and Systems (IROS), 2023.
- Presented *Voxelcache*: *Accelerating online mapping in robotics and 3d reconstruction tasks* at the Programmable Architectures and Compilation Techniques (PACT) conference, 2022.
- Presented *Efficient automatic differentiation for GPU based differentiable simulators* at the Student Research Competition (SRC) Finalists Round, MICRO 2023.
- Presented *HIWE: Hierarchical Importance Weighted Encoding* at Intel, Vision Research Group, 2023.

Service

- Reviewer for International Conference for Robotics and Applications (ICRA) 2022.
- Secondary reviewer for ACM Microarchitecture (ACM MICRO) Conference 2024.
- Secondary reviewer for Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2024.
- Secondary reviewer for ACM Microarchitecture (ACM MICRO) Conference 2023.
- Secondary reviewer for Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2023.
- Secondary reviewer for Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2022.
- Secondary reviewer for International Symposium for Computer Architecture (ISCA) 2023.
- Student organizer for ACM MICRO 2023 conference held in Toronto, Canada.

Awards

October 2023 ◇ 2^{nd} place at Student Research Competition at ACM MICRO 2023.

2023 ◇ Recipient of the 2023 Wolfond Scholarship in Wireless Information Technology.

2022 ◇ Recipient of the 2022 Wolfond Scholarship in Wireless Information Technology.

2013 ◇ One of the 30 students shortlisted for the Indian National Math Olympiad (INMO-2013).

2012 ◇ Selected to be a KVPY (Kishore Vaignyanik Protsahan Yojana) scholar. Ranked 15 out of over 100,000 applicants.

Research Articles

Conference Proceedings / Journal Articles

1. S. Durvasula, A. Zhao, P. Sanjaya, G. Guan, R. Liang, and N. Vijaykumar, "Scar: Sub-core and atomic reduction for raster-based rendering pipelines," in *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2025.
2. C. Giannoula, P. Yang, I. Fernandez, J. Yang, S. Durvasula, Y. Li, M. Sadrosadati, J. Luna, O. Mutlu, and G. Pekhimenko, "Pygim: An efficient graph neural network library for real processing-in-memory architectures," in *special interest group on performance evaluation (SIGMETRICS)*, 2025.

3. S. Durvasula, J. Zhao, R. Kiguru, Y. Guan, and N. Vijaykumar, "Ace: Efficient gpu kernel concurrency for input-dependent irregular computational graphs," in *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2024.

4. C. Li, R. Liang, H. Fan, Z. Zhengen, S. Durvasula, and N. Vijaykumar, "Disorf: A distributed online nerf training and rendering framework for mobile robots international conference on robotics," in *Robotics and Automation Letters (RA-L)*, 2024.

5. J. Zhao, L. Zhang, S. Durvasula, F. Chen, N. Jain, S. Panneer, and N. Vijaykumar, "Distributed training of neural radiance fields: A performance characterization," *International Symposium on Performance Analysis of Systems and Software (2-page abstract, presented at proceedings) (ISPASS)*, 2024.

6. S. Durvasula, Y. Guan, and N. Vijaykumar, "Ev-conv: Fast cnn inference on event camera inputs for high-speed robot perception," *IEEE Robotics and Automation Letters (RA-L). Presented at IROS*, 2023.

7. S. Durvasula, R. Kiguru, S. Mathur, J. Xu, J. Lin, and N. Vijaykumar, "Voxelcache: Accelerating online mapping in robotics and 3d reconstruction tasks," in *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2022.

Posters

1. S. Durvasula, Y. Guan, and N. Vijaykumar, "Ev-conv: Fast cnn inference on event camera inputs for high-speed robot perception," *IEEE International Conference on Robotics and Systems (IROS)*, 2023.

2. S. Durvasula and N. Vijaykumar, "Efficient automatic differentiation for gpu-based differentiable simulators," *Student Research Competition at ACM MICRO*, 2023.

3. S. Durvasula and N. Vijaykumar, "Accelerating simulation engines for deep reinforcement learning with concurrent kernel execution," *Student Research Competition at IEEE/ACM Programmable Architectures and Compilation Techniques (PACT)*, 2022.