Siddharth Singh
University of Maryland, College Park

Email-id : **ssingh37@umd.edu**

EDUCATION

Degree	**University/Institute**	**Year**	**GPA**
Ph.D. in Computer Science	University of Maryland	2020-	3.96/4.0
B.Tech (Hons.) + M.Tech, Computer Science	IIT Kharagpur	2015-2020	9.60/10.0

RESEARCH INTERESTS

- High Performance Computing, Parallel Deep Learning, Heterogeneous Computing, Structure from Motion

SKILLS AND EXPERTISE

- Languages : Python, C++, C, Java
- Software : MPI, NCCL, CUDA, OpenCL, Pytorch, HPCToolkit, Hatchet, OpenMP, Git, Github, SQL, Spark

PUBLICATIONS

- Prajwal Singhania, **Siddharth Singh**, Shwai He, Soheil Feizi, and Abhinav Bhatele. Loki: Low-rank keys for efficient sparse attention. In *Advances in Neural Information Processing Systems (NeurIPS, to appear)*, 2024
- Abhimanyu Hans, John Kirchenbauer, Yuxin Wen, Neel Jain, Hamid Kazemi, Prajwal Singhania, **Siddharth Singh**, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a goldfish, don't memorize! mitigating memorization in generative llms. In *Advances in Neural Information Processing Systems (NeurIPS, to appear)*, 2024
- **Siddharth Singh**, Prajwal Singhania, Aditya Ranjan, John Kirchenbauer, Jonas Geiping, Yuxin Wen, Neel Jain, Abhimanyu Hans, Manli Shu, Aditya Tomar, Tom Goldstein, and Abhinav Bhatele. Democratizing ai: Open-source scalable llm training on gpu-based supercomputers. In *Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis. (to appear)*, SC '24. ACM/IEEE, 2024
- **Siddharth Singh**, Olatunji Ruwase, Ammar Ahmad Awan, Samyam Rajbhandari, Yuxiong He, and Abhinav Bhatele. A hybrid tensor-expert-data parallelism approach to optimize mixture-of-experts training. In *Proceedings of the ACM International Conference on Supercomputing*, ICS '23. ACM, 2023
- **Siddharth Singh** and Abhinav Bhatele. Exploiting sparsity in pruned neural networks to optimize large model training. In *Proceedings of the IEEE International Parallel Distributed Processing Symposium*, IPDPS '23. IEEE Computer Society, May 2023
- **Siddharth Singh** and Abhinav Bhatele. Axonn: An asynchronous, message-driven parallel framework for extreme-scale deep learning. In *Proc. IEEE International Parallel Distributed Processing Symposium*, IPDPS '22. IEEE Computer Society, May 2022
- Anirban Ghose, **Siddharth Singh**, Vivek Kulaharia, Lokesh Dokara, Srijeeta Maity, and Soumyajit Dey. Pyschedcl: Leveraging concurrency in heterogeneous data-parallel systems. *IEEE TC*, 2021
- Pinkesh Badjatiya, Mausoom Sarkar, Nikaash Puri, Jayakumar Subramanian, Abhishek Sinha, **Siddharth Singh**, and Balaji Krishnamurthy. Status-quo policy gradient in multi-agent reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021
- **Siddharth Singh***, Shalmoli Ghosh*, Prajwal Singhania*, Koustav Rudra, and Saptarshi Ghosh. Stance detection in web and social media: A comparative study. In *Proc. 10th Conference and Labs of the Evaluation Forum (CLEF)*, 2019 (*equal contribution)

RESEARCH EXPERIENCE

- **Nvidia, Applied Deep Learning (Megatron-LM)** *(June 2023 - August 2023)*
  Summer Research Intern
  - Developed a 4-bit quantization algorithm for the key-value cache to reduce memory consumption and latency of LLM inference.
  - Developed a fused implementation of the algorithm with the flash attention kernel in CUDA.
- **Microsoft Research, DeepSpeed.ai** *(June 2022 - August 2022)*
  Summer Research Intern
  - Worked on scaling DeepSpeed's support for Mixture of Expert models.

  - Integrated expert parallelism, tensor parallelism and ZeRO to enable training of larger base models.
  - Introduced novel communication optimizations to minimize communication in expert parallelism and tensor parallelism.

- **University of Maryland, Parallel Software and Systems Group** *(August 2020 -)*
  Graduate Research Assistant, Advisor: Abhinav Bhatele
  - Working on developing algorithms and techniques to optimize the parallel training of large multi-billion parameter models at scale.
  - My work has been nominated as a finalist for the ACM Gordon Bell competition 2024.

- **IIT Kharagpur, High Performance Real-time Computing Lab** *(August 2019 - April 2020)*
  Master's Thesis Project, Advisor: Soumyajit Dey
  - Contributed to the design and development of an OpenCL based framework aimed at rapid prototyping and intelligent scheduling of computation graphs (e.g. neural networks) on heterogeneous GPU-CPU environments.
  - Developed and implemented a fine grained scheduling algorithm that sped up the inferencing of a transformer architecture by 20 percent by jointly scheduling on a CPU and GPU.

- **Wadhwani AI, Mumbai, India** *(May 2019 - July 2019)*
  Summer Research Intern
  - Developed a Blender3D based rendering simulator to generate realistic synthetic data of babies in indoor environments, which is used by the organisation to benchmark their algorithms.
  - Engineered a solution based on structure-from-motion algorithms for obtaining SMIL based deformable human meshes from a video recording of a subject baby that produced a relative error of 9% on weight estimation on simulated data.

ACHIEVEMENTS

- Led a team at UMD which was nominated as a finalist for the ACM Gordon Bell Prize, 2024.
- Received the Outstanding Graduate Research Assistant Award at the University of Maryland in AY 23-24.
- Awarded the Dean's Fellowship at the University of Maryland in 2020.
- Was ranked 4th in the Department of Computer Science, IIT Kharagpur wrt GPA.
- Second Runner Up, PAN IIT Hackathon 2019.
- Received the Best Thesis Award for my Master's Thesis Project at IIT Kharagpur.
- Secured a national rank of 405 among 1.3M students appearing for JEE Advanced 2015.

COURSE PROJECTS

- **Algorithmic Evolutionary Biology** - Parallelized the popular FastRFS algorithm using OpenMP to obtain a speedup of 8x.
- **Social Computing** - Implemented the Mondrian Algorithm to k-anonymize Barry Becker's 1994 Census Database on UCI.
- **Information Retrieval** - Trained and compared standard text classification models - Kim's CNN, LSTMs, BERT on a dataset crawled from movies on IMDB from 1946-2018 to predict ratings from plots.
- **Natural Language Processing** - Developed a tag co-occurence based clustering metric to recognise temporal changes in the usage of Quora tags (eg. :- 'Beef' tag changing from culinary to political in India around 2015).
- **Image Processing** - Devised an image processing based pipeline to automatically read the transistor numbers from images of transistors.
- **Computer Networks** - Implemented a reliable rudimentary TCP-like transport layer protocol over UDP with several TCP functionalities including reliable transmission, flow control and congestion control.
- **Database Management Systems** - Developed a Django and SQL based web app for centralized scheduling of deadlines and class tests in a university.
- **Operating Systems** - Implemented a RAM Resident File System with Unix like functionalities and API calls.
- **Computer Organisation and Architecture** - Developed single cycle and multi cycle CPUs on Verilog with a Reduced Instruction Set (RISC).
- **Compilers** - Programmed a full fledged compiler to generate assembly level instructions for a custom Matlab like matrix manipulation library.