|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Aditya K Kamath | | | Website: [akkamath.github.io](https://akkamath.github.io/)  Email: [akkamath@cs.washington.edu](mailto:akkamath@cs.washington.edu) | |
| Research direction: My projects revolve around reducing the impact of data movement. I’m currently working on (1) improving LLM inference latencies by designing better attention kernels and (2) reducing CPU-GPU data transfer in GNNs and DLRMs. | | | | |
| ACADEMIC QUALIFICATIONS | | | | |
| Year | **Degree** | | **Institute, City** | |
| 2021 – Present | **Ph.D.** in Computer Science | | University of Washington, Seattle | |
| 2021 - 2024 | **M.S.** in Computer Science | | University of Washington, Seattle | |
| 2015 – 2019 | **B.Tech.** in Computer Science | | National Institute of Technology Karnataka, Surathkal | |
| PROFESSIONAL EXPERIENCE | | | | |
| University of Washington | Graduate Research Assistant | | | (Sep 2021 – Present) | |
| UW logos | UW Brand | * Working on **reducing data movement** in contemporary applications [**ISCA ‘24**]. * Writing **CUDA** kernels to reduce data movement, and worked with the **gem5** simulator. * Part of the Computer Systems Lab advised by [Prof. Simon Peter](https://homes.cs.washington.edu/~simpeter/). | | | |
| Microsoft Research | Research Intern | | | (Jun 2024 – Sep 2024) | |
| Microsoft Research Limited Fund - Cambridgeshire Community Foundation | * Worked on POD-Attention; **improving LLM inference** by designing a **better attention kernel** for hybrid batches containing chunked prefills and decodes. [**ASPLOS ‘25**] * Mentored by [Dr. Ashish Panwar](https://apanwariisc.github.io/). | | | |
| AMD Research | GPU-Centric Collectives Distributed Systems Research Intern | | | (Jun 2022 – Sep 2022) | |
|  | * Worked on improving **GPU-initiated collective communication**. * Improved[**ROC\_SHMEM**](https://github.com/ROCm-Developer-Tools/ROC_SHMEM) **All-to-All** communication collective using **CUDA/HIP**. * Worked with the parallel and distributed programming team. | | | |
| Indian Institute of Science | Research Assistant | | | (Jun 2019 – Aug 2021) | |
| Image result for iisc logo | * Worked on enhancing **race detection** in **GPUs**. **[ISCA ‘20, SOSP ‘21]** * Applied **NVM** to parallel architectures, i.e., GPU-enhanced persistent KVS and DB. **[ASPLOS ‘22, ‘23]** * Worked under the guidance of [Prof. Arkaprava Basu](https://www.csa.iisc.ac.in/~arkapravab/). | | | |
| NOTABLE PUBLICATIONS | | | | |
| 1. [ASPLOS ‘25] POD-Attention: Unlocking Full Prefill-Decode Overlap for Faster LLM Inference [[Paper](https://akkamath.github.io/files/ASPLOS25_POD.pdf)]   Ribbon with solid fillDistinguished Artifact Award  Aditya K Kamath, Ramya Prabhu, Jayashree Mohan, Simon Peter, Ramachandran Ramjee, Ashish Panwar  *30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*   1. [ISCA ‘24] (MC)^2 : Lazy MemCopy at the Memory Controller [[Paper](https://akkamath.github.io/files/ISCA24_MCSquare.pdf)] [[Video](https://www.youtube.com/watch?v=gMgQGgUI1xA)]   Aditya K Kamath, Simon Peter  *51st IEEE/ACM International Symposium on Computer Architecture*   1. [ISCA ‘24] Scalable, Programmable and Dense: The HammerBlade Open-Source RISC-V Manycore   Dai Cheol Jung, Max Ruttenberg, Paul Gao, Scott Davidson, Daniel Petrisko, Kangli Li, Aditya K Kamath, et. al.  *51st IEEE/ACM International Symposium on Computer Architecture*   1. [ASPLOS ‘23] Scoped Buffered Persistency Model for GPUs [[Paper](https://akkamath.github.io/files/ASPLOS23_SBRP.pdf)] [[Video](https://www.youtube.com/watch?v=cdCSMuIT3D0)]   Shweta Pandey\*, Aditya K Kamath\*, Arkaprava Basu  *28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*   1. [ASPLOS ‘22] GPM: Leveraging Persistent Memory from a GPU [[Paper](https://akkamath.github.io/files/ASPLOS22_GPM.pdf)] [[Video](https://www.youtube.com/watch?v=WER5mZPYFSc)]   Shweta Pandey\*, Aditya K Kamath\*, Arkaprava Basu  *27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*   1. [SOSP ‘21] i****GUARD: In-GPU Advanced Race Detection****  [[Paper]](https://akkamath.github.io/files/SOSP21_iGUARD.pdf)  [[Video]](https://www.youtube.com/watch?v=UxLNZHoxRjY)   Aditya K Kamath and Arkaprava Basu  ACM SIGOPS 28th Symposium on Operating Systems Principles   1. ****[ISCA ‘20]** **ScoRD: A Scoped Race Detector for GPUs**** [[Paper]](https://akkamath.github.io/files/ISCA20_ScoRD.pdf)  [[Video]](https://www.csa.iisc.ac.in/~arkapravab/papers/ScoRD_talk.mp4)   Aditya K. Kamath\*, Alvin A George\*, Arkaprava Basu  47th IEEE/ACM International Symposium on Computer Architecture  \*Authors contributed equally | | | | |
| WORKSHOPS AND TALKS | | | | |
| * POD-Attention: Unlocking Full Prefill-Decode Overlap for Faster LLM Inference   + Microsoft Research India (AI Infrastructure Group) * [(MC)^2: Lazy MemCopy at the Memory Controller](https://www.youtube.com/watch?v=gMgQGgUI1xA)   + Cornell University (Networked and Operating Systems Group)   + Indian Institute of Science (Computer Systems Lab)   + International Symposium on Computer Architecture (ISCA)   + University of Washington (Systems Lab) * [GPM: Leveraging Persistent Memory from a GPU](https://www.youtube.com/watch?v=WER5mZPYFSc)   + University of California San Diego (Non-Volatile Memories Workshop) | | | | (Sept ‘24)  (Nov ‘24)  (Aug ‘24)  (July ‘24)  (May ‘24)  (May ‘22) |
| TEACHING EXPERIENCE | | | | |
| Undergraduate Teaching Assistant at NITK Surathkal | | | (2018 - 2019) | |
| * Taught a lesson on the functioning of a cache and modern cache replacement policies. * Taught a lesson on Persistent Memory and possible future uses. * Taught a lesson on importance of simulation in systems research, and how to use Intel PIN tool for tracing. * Designed a project for students to create a working cache simulator. | | | | |
| VOLUNTEER SERVICE | | | | |
| * Grad Admission Reader (2022) at University of Washington: Reviewed applications of graduate school applicants. * Pre-Application Mentorship Program (2022, 2023) at University of Washington: Guided students from historically marginalized groups through the graduate application process, revising their SOP and resume. * Head Placement Coordinator at NITK: Responsible for directing the entire NITK campus hiring process for 2019. Managed dozens of Placement Coordinators and coordinated with HRs of hundreds of companies. * Co-Head of Algorithms Group of Web Enthusiasts’ Club at NITK: Organised competitive coding events in college. Gave talks on the basics of algorithms and optimisations. | | | | |
| TECHNICAL SKILLS | | | | |
| * Programming Languages: | | C, C++, CUDA, Python | | |
| * Simulator Experience: | | gem5, GPGPU-Sim, SST, ns-3, ChampSim | | |
| * Relevant Courses: | | Computer Organization and Architecture, High Performance Computing, Heterogeneous Parallel Computing, Data Structures and Algorithms, Operating Systems | | |