# RAMYA PRABHU

Phone: (+91) 702-254-9671 ◇ Email: ramrag0107@gmail.com
Homepage: https://the-mind-palace.github.io/
Google Scholar ◇ Github ◇ LinkedIn ◇ Microsoft Research Profile

## EDUCATION

**PES University** *June 2023 - June 2019*
B.E. in Computer Science and Engineering GPA: 8.97/10.0
*Awarded MRD Scholarship - merit scholarships for students ranked in the* ***top 20%*** *of the department*

## RESEARCH INTERESTS

I am interested in building performant computing systems. My recent work has been with optimizing performance for LLM inference.

## PUBLICATIONS

[1] **Ramya Prabhu**, A. Nayak, J. Mohan, R. Ramjee, and A. Panwar, "*vAttention*: Dynamic memory management for serving llms without pagedattention," *30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2024. [Online]. Available: https://arxiv.org/abs/2405.04437.

[2] A. K. Kamath, **Ramya Prabhu**, J. Mohan, S. Peter, R. Ramjee, and A. Panwar, "Pod-attention: Unlocking full prefill-decode overlap for faster llm inference," *conference name*, 2024. arXiv: 2410.18038 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2410.18038.

## RESEARCH EXPERIENCE

**Research Fellow - Microsoft Research India, Bengaluru** Jul 2023 - Present
*Supervisors: Dr. Ramachandran Ramjee, Dr. Ashish Panwar, and Dr. Jayashree Mohan*

- Developed **vAttention** [1], a scheme for attention KV cache management which improved prefill throughput by **1.29x** and decode throughput by **1.99x** and is being considered for adoption in popular inference stacks like vLLM and DeepSpeed [**ASPLOS 25**]
- Assisted in profiling and experimenting for **POD-Attention** [2], a GPU kernel for attention that efficiently utilizes compute and memory resources by overlapping computation. It sped up attention computation by up to **75%**
- Optimised MoE inference in both offline and online scenarios, through a scheme that was able to improve the throughput of GShard based model architectures by upto **66%**

**Research Intern - Intel Labs** Dec 2022 - Jun 2023
*Supervisors: Sreenivas Subramoney and Anant Nori*

- Worked on optimizing address translation subroutine on Intel CPUs [power and performance]
- Profiled and analysed performance bottlenecks in Intel CPUs. Hacked into industry-grade hardware simulator to test and validate solutions
- Internship resulted in **2 patents [one filed, one in the process]**

**Research Intern - IIT Bombay** May 2022 - Sept 2022
*Supervisors: Dr. Biswabandan Panda*

- Addressed the mitigation the bottleneck that DRAM bandwidth constraints impose on a server system

- Created a prototype that dynamically predicts the criticality of an Instruction Pointer that out-performed the SOTA criticality prediction schemes
- Implemented and tested state-of-the-art criticality schemes and prefetchers to analyze their efficacy for the given system

**Research Intern - PES University** May 2021 - Sept 2021

*Supervisors: Dr. Subramaniam Kalambur*

- Ran experiments to profile and analyse NUMA systems
- Built a dataset to predict memory policy for client workloads

## ACHIEVEMENTS AND ACCOLADES

| | |
|---|---|
| Manupatra Out-of-the-Box Prize Awardee, awarded by OPENNYAI | *2022* |
| ExploreCS Research Scholarship, awarded by Google | *2022* |
| MRD Scholarship, awarded by PES University - Awarded to the **top 20%** of the students in the department | *2020,2022,* |
| KVPY Scholarship, awarded by DST, Gov of India | *2019* |

## SKILLS/HOBBIES

| | |
|---|---|
| **Programming Languages** | Python, C/C++ |
| **Frameworks and Tools** | PyTorch, vLLM, Vim |
| **Hobbies** | Painting and singing |