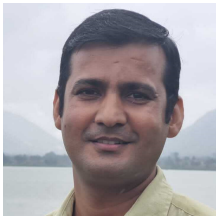
Ashish Panwar

panwarit014@gmail.com
https://apanwariisc.github.io



About Me

I enjoy working on fundamental research challenges in computer systems, often motivated by the continuously evolving hardware and software ecosystem. My primary areas of interest include memory management, operating systems and systems for machine learning—particularly the large language models.

Employment History

Since March'25	**Principal Researcher.** Microsoft Research, India.
July'22 – Feb'25	**Senior Researcher.** Microsoft Research, India.
Oct'16 – July'18	**Member of Technical Staff.** Advanced Technology Group (ATG), NetApp, India.
Aug'15 – Oct'16	**Software Engineer.** Intel India Pvt. Ltd.
Jan'12 – July'12	**Assistant System Engineer.** Tata Consultancy Services Ltd., India.

Education

Aug'18 – July'22	**Ph.D. in Computer Science**, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India. **CGPA: 9/10** Thesis: *Operating System Support for Efficient Virtual Memory*
Aug'12 – July'15	**M.Sc. (Engg.) in Computer Science**, Department of Computer Science and Automation, Indian Institute of Science, Bangalore, India. **CGPA: 6.3/8**
Aug'07 – May'11	**B. Tech. in Information Technology**, Meerut Institute of Engineering and Technology (MIET), Meerut, UP, India. **Percentage: 75%.**

Research Publications (full conference papers)

- **[ISMM'25] EMD: Fair and Efficient Dynamic Memory De-bloating of Transparent Huge Pages**
  Parth Gangar, Ashish Panwar, K. Gopinath
  In proceedings of the ACM SIGPLAN International Symposium on Memory Management (ISMM) 2025.
- **[ASPLOS'25] POD-Attention: Unlocking Full Prefill-Decode Overlap for Faster LLM Inference**
  Aditya K Kamath, Ramya Prabhu, Jayashree Mohan, Simon Peter, Ramachandran Ramjee, Ashish Panwar
  In proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2025. [Artifact Evaluated, Best Artifact Award]
  Submissions: 510, Accepted: 105, Acceptance rate: 20.59%
- **[ASPLOS'25] vAttention: Dynamic Memory Management for Serving LLMs without PagedAttention**
  Ramya Prabhu, Ajay Nayak, Jayashree Mohan, Ramachandran Ramjee, Ashish Panwar
  In proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2025. [Artifact Evaluated]
  Submissions: 510, Accepted: 105, Acceptance rate: 20.59%
- **[OSDI'24] Taming Throughput-Latency Trade-off in LLM Inference with Sarathi-Serve**
  Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, Ramachandran Ramjee
  In proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 2024. [Artifact Evaluated]
  Submissions: 282, Accepted: 49, Acceptance rate: 17.38%

- **[MLSys'24] VIDUR: A Large-Scale Simulation Framework for LLM Inference**
  Amey Agrawal, Nitin Kedia, Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, Alexey Tumanov [Source Available]
  In proceedings of the 7th Annual Conference on Machine Learning and Systems (MLSys), 2024.

- **[PETS'24] SIGMA: Secure GPT Inference with Function Secret Sharing**
  Kanav Gupta, Neha Jawalkar, Ananta Mukherjee, Nishanth Chandran, Divya Gupta, Ashish Panwar, Rahul Sharma
  In Proceedings of the 24th Privacy Enhancing Technologies Symposium (PETS), 2024.
  Submissions: 250, Accepted: 49, Acceptance rate: 19.60%

- **[MICRO'21] Trident: Harnessing Architectural Resources for All Page Sizes in x86 Processors**
  Ashish Panwar, Venkat Sri Sai Ram*, Arkaprava Basu*
  In proceedings of the 54th IEEE/ACM International Symposium on Microarchitecture (MICRO), 2021.
  * Joint first authors. [Artifact Evaluated]
  Submissions: 423, Accepted: 94, Acceptance rate: 22.22%

- **[PACT'21] nuKSM: NUMA-aware Memory De-duplication on Multi-socket Servers**
  Akash Panda, Ashish Panwar, Arkaprava Basu
  In proceedings of the 30th International Conference on Parallel Architectures and Compilation Techniques (PACT), 2021. [Artifact Evaluated]
  Submissions: 96, Accepted: 25, Acceptance rate: 26.04%

- **[ASPLOS'21] Fast Local Page-Tables for Virtualized NUMA Servers with vMitosis**
  Ashish Panwar, Reto Achermann, K. Gopinath, Abhishek Bhattacharjee, Arkaprava Basu, Jayneel Gandhi
  In proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2021. [Artifact Evaluated]
  Submissions: 398, Accepted: 75, Acceptance rate: 18.84%

- **[ASPLOS'20] Mitosis: Transparently Self-Replicating Page-Tables for Large-Memory Machines**
  Reto Achermann, Ashish Panwar, Abhishek Bhattacharjee, Timothy Roscoe, Jayneel Gandhi
  In proceedings of the 25th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2020. [Artifact Evaluated]
  Submissions: 476, Accepted: 86, Acceptance rate: 18.07%

- **[ASPLOS'19] HawkEye: Efficient Fine-grained OS Support for Huge Pages**
  Ashish Panwar, Sorav Bansal, K. Gopinath
  In proceedings of the 24th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2019. [Source Available]
  Submissions: 350, Accepted: 74, Acceptance rate: 21.14%

- **[ASPLOS'18] Making Huge Pages Actually Useful**
  Ashish Panwar, Aravinda Prasad, K. Gopinath
  In proceedings of the 23rd ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2018. [Source Available]
  Submissions: 307, Accepted: 56, Acceptance rate: 18.24%

- **[HiPC'15] Towards Practical Page Placement for a Green Memory Manager**
  Ashish Panwar, K. Gopinath
  In proceedings of the 22nd IEEE International Conference on High Performance Computing (HiPC), 2015.

Short Papers and Pre-prints

- **[Pre-print. Arxiv'26] LLM-42: Enabling Determinism in LLM Inference with Verified Speculation**
  Raja Gond, Aditya K Kamath, Ramachandran Ramjee, Ashish Panwar

- **[OpSysRev'25] Efficient LLM Inference via Chunked Prefills**
  Arney Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S Gulavani, Alexey Tumanov, Ramachandran Ramjee
  In ACM SIGOPS Operating Systems Review, 2025.

❖ **[Pre-print. Arxiv'25] PyGraph: Robust Compiler Support for CUDA Graphs in PyTorch**
Abhishek Ghosh, Ajay Nayak, Ashish Panwar, Arkaprava Basu

❖ **[IEEE CAL'24] Address Scaling: Architectural Support for Fine-Grained Thread-Safe Metadata Management**
Deepanjali Mishra, Konstantinos Kanellopoulos, Ashish Panwar, Akshitha Sriraman, Vivek Seshadri, Onur Mutlu, Todd C Mowry
In IEEE Computer Architecture Letters, 2024.

❖ **[ApSys'16] A Case for Protecting Huge Pages from the Kernel**
Ashish Panwar, Naman Patel, K. Gopinath
In proceedings of the 7th ACM SIGOPS Asia-Pacific Workshop on Systems (APSys), 2016.

Patents

◆ **Efficient LLM Inference by Piggybacking Decodes with Chunked Prefills**
Inventors: *Jayashree Mohan, Ashish Panwar, Nipun Kwatra, Bhargav S. Gulavani, Ramachandran Ramjee, Amey Agrawal*
US Patent App. 18/416,564

Technical Talks

✹ "Enabling Determinism in LLM Inference" at CDS, IISc, 2024.
✹ "From Basics to Breakthroughs: Tales from LLM Systems Research" at the workshop on Present and Future Computing Systems, CSA, IISc, 2024.
✹ "Efficiently Serving Large Language Models" at the Center for Networked Intelligence, IISc, 2024.
✹ "Trident: Harnessing Architectural Resources for All Page Sizes in x86 Processors" at *MICRO, 2021 (virtual).*
✹ "Fast Local Page-Tables for Virtualized NUMA Servers with vMitosis" at *ASPLOS, 2021 (virtual).*
✹ "System Software Enhancements for Efficient Memory Management" at *India Design Review, Semiconductor Research Corporation (SRC), Bangalore, India, January, 2020.*
✹ "OS and Hypervisor Support for Self-Replicating Page-Tables" at *VMware Research, Palo Alto, US, 2019.*
✹ "Making Huge Pages Actually Useful"
  – *ACM Inter-Research-Institute Student Symposium, Kochi, India, February, 2019.*
  – *NetApp, Sunnyvale, US, April, 2018.*
  – *Qualcomm, Bangalore, India, March, 2018.*
  – *ASPLOS, Williamsburg, Virginia, US, 2018.*
✹ "Towards Practical Page Placement for a Green Memory Manager" at *HiPC, Bangalore, India, 2015.*

Honors and Awards

✦ Recipient of the *Alumni Medal for Best PhD Thesis* 2022-2023 at CSA, IISc.
✦ Recipient of the *Prime Minister's Fellowship Scheme for Doctoral Research* (2019), co-sponsored by CII, Government of India, and Microsoft Research India (1 of the 3 recipients in computer science across India).
✦ Recipient of the *Quantum Leaper* award at NetApp, 2017.
✦ Recipient of the *Star Performer* award during the *Initial Learning Program* at TCS, 2012.

Teaching Experience

◗ *Operating Systems*, co-taught with Prof. Vinod Ganapathy at CSA, Indian Institute of Science (Jan-April, 2023). Topics covered: Scheduling and memory management. I have also designed three programming assignments for this course in 2022 and 2023 (based on fast checkpoint/restore, fair scheduling and user-level memory swapping).

In addition, I have served as a teaching assistant for the following courses:

- *Operating Systems*, graduate course at CSA, Indian Institute of Science (Feb-June, 2021)
- *Compiler Design*, graduate course at CSA, Indian Institute of Science (Jan-April, 2020)
- *Operating Systems*, graduate course at CSA, Indian Institute of Science (Jan-April, 2020)
- *Programming and Data Structures*, undergraduate course at Indian Institute of Science (Aug-Dec, 2019)

Professional Service

I have served as a reviewer on the following program committees:

- IEEE/ACM International Symposium on Computer Architecture (ISCA), 2026.
- ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2025.
- ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2024.
- [Heavy member] Usenix Annual Technical Conference (ATC), 2025.
- [Light member] Usenix Annual Technical Conference (ATC), 2024.
- ACM SIGPLAN International Symposium on Memory management (ISMM), 2024.
- [External] IEEE International Symposium on Performance Analysis of System and Software (ISPASS), 2025.
- [External] ACM Transactions on Computer Systems (TOCS), 2023.
- [External] ACM Transactions on Architecture and Code Optimization (TACO), 2022.

References

Arkaprava Basu
Associate Professor
Indian Institute of Science
arkapravab@iisc.ac.in

K. Gopinath
Professor (superannuated)
Indian Institute of Science
gopi@iisc.ac.in

Abhishek Bhattacharjee
Professor
Yale University, USA
abhishek@cs.yale.edu

Sorav Bansal
Professor
Indian Institute of Technology Delhi
sbansal@iitd.ac.in

Mark D Hill
Professor Emeritus of Computer Science
University of Wisconsin-Madison
markhill@cs.wisc.edu