Sandeep **Kumar**

Research Scientist & Team Lead — Memory Management Across the Stack (CXL, Tiered Memory, Linux Kernel) @ Intel — Bengaluru, India

+91 82773 61995 | sandeep007734@gmail.com | sandeep007734.github.io | sandeep007734 | sandeep007734

Summary

Systems researcher and team lead in full-stack **memory management** — from hardware (**CXL**, tiered and heterogeneous memory, Intel IAA/DSA accelerators, CPU–GPU memory) and page-table telemetry, through **Linux kernel** memory management (LKML patch), to runtimes and workloads: DNN training, LLM inference, vector databases, virtual machines, and cloud. I lead Intel's **System Memory Architecture** team, translating workload characterization into kernel prototypes and platform decisions that deliver **up to 2× better server memory TCO** and **59–83% less fast memory** for DNN training. Published at **EuroSys**, **USENIX ATC**, **ISMM**, and **DATE**; multiple filed patents.

Technical Skills

Memory Hardware	CXL, tiered & heterogeneous memory (DRAM, PMEM, CXL), NUMA, CPU–GPU memory, Intel IAA/DSA
Linux Kernel	Memory management (page tables, page migration, zswap); system calls, kernel–userspace interfaces
Workloads	DNN training, LLM inference & KV-cache (vLLM), vector databases, virtual machines, FaaS
Performance	Linux `perf`, Intel VTune, `pmu-tools`; workload characterization; benchmark design
Programming	C, C++, Python, shell scripting
Leadership	Team lead; mentoring; cross-org collaboration with CPU architects; university collaborations

Ongoing Projects — Memory Systems for AI

CPU-Assisted Agentic Inference *Intel*

- Chartering a pathfinding effort to use **Intel Xeon** compute and accelerators for agentic inference, evolving CPU memory-tier deployments toward balanced CPU–GPU execution.

KV-Cache Offload Optimization *Intel*

- Defining the optimization roadmap for **KV-cache offload** across GPU HBM, host DRAM, and CXL, with Xeon-accelerated compression and decompression.

Rethinking Memory Subsystem for CXL Memory Systems *Intel*

- Driving systems pathfinding for **CXL memory**, replacing DRAM-centric assumptions with designs suited to its latency, bandwidth, capacity, and access characteristics.

Experience

Intel *Bengaluru, India*

Research Scientist & Team Lead, System Memory Architecture *March 2022 – Present*

- Lead the System Memory Architecture team, owning memory-management research from hardware to workloads: next-generation memory (**CXL**), accelerators (Intel IAA/DSA), and **CPU–GPU memory**. Partner with CPU architects on future memory and profiling features; mentor researchers and drive university collaborations.
- *Telemetry:* Co-built *Telescope*, page-table-based telemetry for terabyte-scale workloads: **over 90% precision and recall at 0.9% single-CPU utilization** on 5 TB footprints, up to 34% higher tiering throughput; patch posted on LKML.
- *Kernel:* Designed and built multi-tier compressed memory management in the **Linux kernel**, delivering **up to 2× better memory TCO** for server workloads.
- *Virtualization & serverless:* Contributed to memory tiering inside **virtual machines** and to memory-restoration templates that speed up **FaaS warm starts**.
- *ML framework:* Co-built proactive memory tiering for CPU-based **PyTorch** DNN training, **cutting the fast-memory footprint by 59–83%** at 1–16% overhead.

Intel Labs *Bengaluru, India*

Graduate Research Intern *July 2020 – January 2021*

- Investigated page-table management for heterogeneous (DRAM + persistent memory) systems, resulting in a publication (ISMM'21) and a granted patent on page-table profiling.

Indian Institute of Science *Bengaluru, India*
Research Associate *September 2014 – July 2017*

- Built workload-driven autotuning for Hadoop MapReduce with noisy-gradient methods, demonstrating cloud-scale gains.

Dell R&D *Bengaluru, India*
Software Development Engineer *July 2013 – June 2014*

Selected Publications

1. *Speeding up FaaS Warm Starts with Memory Restoration Templates*
   Aravinda Prasad, Sreenivas Subramoney, **Sandeep Kumar**, and Antti Kervinen. In the **Workshop on Disruptive Memory Systems (DIMES)**, 2026.
2. *TierScape: Harnessing Multiple Compressed Tiers to Tame Server Memory TCO*
   Sandeep Kumar, Aravinda Prasad, and Sreenivas Subramoney. In **EuroSys**, 2026.
3. *TierTrain: Proactive Memory Tiering for CPU-Based DNN Training*
   Sathvik Swaminathan, **Sandeep Kumar**, Aravinda Prasad, and Sreenivas Subramoney. In **ISMM**, 2025.
4. *Efficient Memory Tiering in a Virtual Machine*
   Chandra Prakash, **Sandeep Kumar**, Aravinda Prasad, and Sreenivas Subramoney. **arXiv preprint**, 2025.
5. *Telescope: Telemetry for Gargantuan Memory Footprint Applications*
   Alan Nair, **Sandeep Kumar**, Aravinda Prasad, Ying Huang, Andy Rudoff, Sreenivas Subramoney. In **USENIX ATC**, 2024.
6. *Page Table Management for Heterogeneous Memory Systems*
   Sandeep Kumar, Aravinda Prasad, Smruti R. Sarangi, and Sreenivas Subramoney. In **ISMM**, 2021.

Full list: sandeep007734.github.io/publications

Patents

1. Methods and Apparatus to Profile Page Tables for Memory Management (*granted*).
   Aravinda Prasad, Sandeep Kumar, Sreenivas Subramoney, Andy Rudoff.
2. Context-Aware Memory Tiering for Machine Learning Training (*filed*).
   Sathvik Swaminathan, Sandeep Kumar, Aravinda Prasad, Sreenivas Subramoney.

*Additional patent applications in the filing pipeline.

Education

Ph.D., Computer Science, Indian Institute of Technology Delhi, *New Delhi, India* 2017 – 2024
M.E., Computer Science (2-year master's), Indian Institute of Science, *Bengaluru, India* 2011 – 2013
B.Tech., Computer Science, Guru Gobind Singh Indraprastha University, *New Delhi, India* 2007 – 2011

Languages

Languages **English**: Full professional proficiency | **Hindi**: Native

References

Sreenivas Subramoney, Senior Distinguished Engineer, *NVIDIA, Bengaluru, India*
Prof. Smruti R. Sarangi, Professor, Department of Computer Science, *IIT Delhi, India*
Prof. K. Gopinath, Professor, Computer Science and Automation, *IISc Bengaluru, India*