YASH LALA

yashlala.com ◇ github.com/yashlala ◇ linkedin.com/in/yashlala
(510)-400-5572 ◇ yash@yashlala.com ◇ Palo Alto, CA

OVERVIEW

I'm a third-year PhD student at Yale, broadly interested in systems for efficient data-center scale computing. My recent work focuses on far-memory systems.

I have 4+ years of kernel programming expertise, including large-scale projects such as reimplementing the memory-management subsystem in my most recent publication (SOSP '25). I have demonstrated expertise in developing high-throughput, low-latency kernel datapaths — such as a Linux-based system capable of migrating 200 Gib/s of memory at <270 ns per page per core. I'm deeply familiar with CPU microarchitecture and its interactions with the OS, and I excel in principled, data-driven, bottleneck analyses.

I'm currently expanding my research area towards cluster-scale optimizations, especially for AI/ML systems. NVIDIA gave me experience with ML workloads; I spent a summer developing network performance models for these workloads and optimizing MoE communication libraries for GPU clusters. The resulting insights have spurred my new set of projects focusing on system optimizations at large scales (intra-cluster failure recovery) and small scales (faster embedding similarity searches within GPUs).

EDUCATION

Yale Ph.D. in Computer Science — 2023 - Present
UCLA B.S. in Computer Science — GPA: 3.782, 2018 - 2022

PROFESSIONAL EXPERIENCE

NOVA Lab, Yale CS Department — Aug 2023 - Expected 2028
PhD Student — *Advisor: Anurag Khandelwal*

- Broadly researching systems for efficient data-center scale computing. Recent work focuses on far-memory systems.
- Developed techniques for accelerating OS memory migration, aimed towards facilitating high-throughput RDMA-based far memory runtimes. Developed extensive knowledge of the the Linux kernel's memory management, RDMA, page placement, IPI, and swap subsystems. My particular focus was on optimizing high throughput, low-latency runtimes, such as memory migration systems and RDMA SmartNIC network stacks. Publication to appear at **SOSP '25**.
- Developing compact data structures for efficient range queries in high-dimensional spaces. Ongoing project.

NVIDIA — May 2025 - Aug 2025
Systems Software Intern — *Supervisor: Misha Smelyanskiy*

- Developed performance models for Mixture-of-Experts (MoE) communication phases for AI/ML models on NVIDIA's GB200 cluster hardware. Estimated ROI for future communication optimizations, and localized network performance bottlenecks in distributed systems of GPUs.
- Examined MoE token to expert mapping distributions to quantify the workload changes required for optimal hardware utilization. Explored algorithm-network codesign opportunities.
- Profiled existing GPU MoE library implementations (such as NCCL, DeepEP, etc). Optimized GPU MoE libraries for NVIDIA hardware by better utilizing tiered Multi-Node NVLink and Infiniband datacenter networks.
- Gained experience with LLM workload network traffic patterns and GPU performance debugging (particularly in GPU-initiated RDMA network communication).

SOLAR Lab, UCLA CS Department — Sept 2021 - Sept 2022
Student Researcher — *Supervisor: Harry Xu*

- Developed Linux kernel mechanisms for transparent far memory, with the goal of merging my changes upstream. The patchset extends the cpuset controller to allow per-cgroup control of active swap devices. The associated refactoring has positive implications for swap throughput, and makes it easy to manage frontswap-based remote memory systems. Patchset available at github.com/yashlala/canvas-linux.
- Developed a patchset to improve the Linux kernel's physical page allocation tail latencies by refilling the percpu low-order free page lists asynchronously using RCU.

CSSI Program, UCLA CS Department — July 2022
Tutor Undergrad (TA)

- Taught introductory data science to high school students. Led discussion sections, prepared assignments, graded papers, and advised students.

Veritas Technologies LLC June 2021 - Sept 2021
SDE Intern

- Worked on NetBackup Flex, a platform for large-scale data consolidation and backup. Implemented automatic node discovery and cluster assimilation over datacenter networks. Primarily worked with Ansible, Docker, and various glue languages.

PUBLICATIONS

Scalable Far Memory: Balancing Faults and Evictions August 2023 - August 2025

- Adapted the Linux Kernel to create a far-memory system capable of offloading excess application memory to arbitrary storage backends, for cluster-wide memory utilization improvements. Unlike prior work, this system scales to highly threaded, memory-intensive applications with minimal (near-ideal) application slowdown. Results are applicable to tiered memory systems and all far-memory backends (eg. NVMe, zswap, and RDMA).
- Rearchitected the Linux memory management subsystems to handle high-throughput page faulting and eviction. System achieves >200 Gib/s page migration throughput with only two CPU cores.
- Extensive low-level performance work around the OS-CPU interface, extending well into microarchitectural details. Gained expertise in: x86 IPI overheads and queuing characteristics; IOTLB-related virtualization overheads; TLB and page-walker coherence on x86; RDMA NIC control path throughput bottlenecks due to driver bugs; hypervisor-mediated IPI delivery costs; and lockless programming within the kernel.
- Developed principles for scalable memory migration. Results generalize to a Library-OS based re-implementation.
- Co-first author. Paper to appear at **SOSP 2025**.

GRU4RecBE: Session Based Recommendations with Features March 2021 - June 2021

- Developed session-based recommendation system in PyTorch which extends the GRU4REC architecture with rich item features extracted from the pre-trained BERT architecture. Non-attentive model outperforms state-of-the-art session-based models over the MovieLens benchmark datasets. Paper **accepted to AAAI Student track**.

PROJECTS

NDN Multicast Feb 2022 - June 2022

- Extended routing protocols for NDN (Named Data Network) under Lixia Zhang. Extended NLSR (a link-state routing algorithm for NDN) to allow for efficient multicast delivery of NDN Interest packets. Student paper available at yashlala.com/nlsr-poster.pdf.

MISCELLANEOUS

Enjoys teaching: led student seminar series, volunteered with ACM TeachLA and as a docent for the Computer History Museum. Certified as a operator for the punched-card based IBM 1401.