YASH LALA

yashlala.com ⬦ github.com/yashlala ⬦ linkedin.com/in/yashlala

(510)-400-5572 ⬦ yash.lala@yale.edu ⬦ Palo Alto, CA

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 - Present
PhD Student — *Advisor: Anurag Khandelwal*

- Researching techniques for accelerating OS memory migration. Project aims to facilitate 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. Particular focus on optimizing high throughput, low-latency runtimes, such as memory migration systems and RDMA SmartNIC network stacks. Publication to appear at **SOSP '25**.

NVIDIA — May 2025 - Aug 2025
Systems Software Intern — *Advisor: Pasha Shamis*

- Developed network performance models for Mixture-of-Experts (MoE) communication phases in ML models. Conducted performance analysis of existing MoE library implementations (such as NCCL, DeepEP, etc), then optimized the libraries for NVIDIA's cluster hardware. Gained experience with LLM workload network traffic patterns, GPU performance debugging (particularly with regards to GPU-initiated RDMA network communication).

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

- Volunteered during the school year, employed full-time to work on kernel patches over the summer. Developed kernel mechanisms for for transparent memory disaggregation. Worked heavily with kernel programming and debugging tools, such as QEMU + GDB, serial port debugging, and perf.
- Independently developed patches for the Linux kernel's swap subsystems, with the goal of merging these changes upstream. Patchset extends the cpuset controller to allow per-cgroup control of active swap devices. 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 latency. The patch reduces 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 for an intensive summer program. Led discussion sections, prepared discussion material and assignments, graded papers, and advised students.

Veritas Technologies LLC — June 2021 - Sept 2021
SDE Intern

- Worked on large-scale data consolidation and backup devices (NetBackup Flex platform). Implemented functionality allowing Flex nodes to automatically discover new backup nodes over the datacenter network, then to securely assimilate them into a backup cluster. Primarily worked with Ansible, Docker, and various glue languages.

PROJECTS

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

- Paper to appear at **SOSP 2025**.

NDN Multicast — May 2022 - Present

- Worked on extending routing protocols for NDN (Named Data Networks). 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.

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 benchmark MovieLens 1M and MovieLens 20M datasets. Paper **accepted to AAAI Student track**.

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 punch-card based IBM 1401.