

Anirudh Sunil

+1-510-676-8390 | asunil3@illinois.edu | linkedin.com/in/anirudh-sunil | github.com/anirudhs1010

EDUCATION

University of Illinois Urbana-Champaign

Urbana, IL

Bachelor of Science in Mathematics and Computer Science

Dec. 2026

Coursework: Computer Architecture, Data Structures, Algorithms, Graph Theory, Probability and Statistics, Systems Programming, Programming Languages and Compilers

EXPERIENCE

Amazon

Sunnyvale, CA

Software Development Engineer Intern

May 2026 – August 2026

- Built a distributed Java Remote Strategy Service handling $\sim 9K$ TPS with a 120 ms production latency SLA for recommendation widgets on Amazon Sports & Outdoors product pages.
- Architected a modular strategy framework supporting N strategy IDs per implementation, reducing new strategy integration to 1 annotated class with automatic runtime discovery.
- Migrated the service to AWS Fargate, provisioning 8+ production infrastructure components including CloudAuth, PrivateLink, autoscaling, load balancing, and monitoring.
- Resolved 20+ dependency, authentication, and build issues from previous iteration for successful production deployment.

University of Chicago

Remote

Research Assistant Intern

Jan. 2026 – May 2026

- Improve scientific artifact discoverability and reproducibility for Trovi by integrating research artifacts for 10k users.
- Normalize metadata and write concise summaries to enhance data accessibility and utility for 100+ JSON requests.
- Refine search and analytics for coverage and trends to support the Chameleon project with 5 Github open source contributions.

National Center for Supercomputing Applications @ UIUC

Urbana, IL

Data Analyst Intern

Aug. 2025 – Dec. 2025

- Preprocessed geospatial and socioeconomic feature data for the Amazon in Brazil, enhancing data quality for analysis.
- Conducted data modeling, analysis, and predictive assessments to provide insights for decision-making processes.
- Created Graphical Neural Network (GNN) with Python, facilitating better understanding of complex data sets.
- Assisted in the development, training, validation, and testing of machine learning algorithms, improving model accuracy.

Warmly, (YC S20) (acquired by Hubspot)

San Francisco, CA

Software Engineer Intern

Jun. 2025 – Aug. 2025

- Implemented Salesflow email integration with TypeScript to optimize automation of customer outreach time by 87%.
- Integrated CRM system with React, Postgres, and SQL to simulate sales team interactions with Redis server for authentication, improving data management.

PROJECTS

Paged KV-Cache gRPC Service for LLM Inference | *Rust, PyTorch, CUDA, gRPC* | [Link](#)

- Built a Rust gRPC paged KV-cache service for LLM inference with 16-token FP16 pages, configurable $32\text{--}head \times 64\text{--}dimension$ KV layout, and a 1,024-page cache capacity.
- Integrated the cache with Hugging Face SmolLM2-1.7B, running 24 transformer layers on an NVIDIA Tesla T4 GPU through a PyTorch/CUDA inference harness.
- Transferred complete per-layer KV state between Python and Rust over gRPC, measuring 56.5 ms to upload and 49.2 ms to restore the full cache.
- Designed the single-node architecture as a foundation for GPU-resident PagedAttention, prefix-cache reuse, eviction policies, and multi-node inference serving.

Rankd | *NextJS, Tailwind CSS, Supabase* | [Link](#)

- Designed a web app that allows users to play a ranking game for companies and brainrot characters.
- Implemented ELO Ranking algorithm that updates leaderboard continuously; scaled to handle 150+ users.
- Updated security measures for POST requests in database and decreased request latency by 25%.

DISTINCTIONS

- Stanford - CodeInPlace: CS106A Section Leader/Instructor
- Andreessen Horowitz - a16z: alpha speedrun finalist

SKILLS

Programming Languages: C, C++, Python, Java, SQL, JavaScript, TypeScript, HTML, CSS

Tools and Frameworks: React, NextJS, NodeJS, LangChain, Keywords AI API, PostgreSQL, Redis, AWS, Kubernetes, Burp Suite, OWASP ZAP, DirBuster, Fusion360