Ram Srivatsa Kannan

Software Engineer | Tech Lead

San Jose ramsrivats@gmail.com 7347413242 ramsrivatsa.github.io ramsrivatsa
ramsrivatsa

Background and Experience

Designing and leading **mission-critical software platforms** that operate at scale across teams and organizations. Specialized in applying distributed systems and database fundamentals to drive architectural clarity, operational excellence, and long-term system evolution.

Act as a **technical owner and decision-maker** for complex initiatives, leading zero-to-one efforts and guiding cross-functional teams of 10+ engineers. Balance hands-on system design and implementation with mentorship, stakeholder alignment, and delivery of durable, production-grade solutions.

Education

Ph.D **University of Michigan, Ann Arbor**, Computer Science and Engineering — Sept 2013 – Dec 2018

- **Thesis:** Enabling Fairness in Cloud Computing Infrastructure

BS **MIT, Anna University**, Information Technology — Sept 2009 – May 2013

- **Thesis:** SCOC IP design for Application Specific Architectures

Experience

Netflix Inc. Online data stores team – SQL, NoSQL, Caching, and Abstractions — Los Gatos, CA, May 2023 – Now

- **Scaling Stateful Services for Big Bet Live Events** Conceived, architected, and led a **zero-to-one** platform for automated capacity planning and autoscaling of stateful data systems (Cassandra, EVCache, Kafka) powering Netflix's largest live events (e.g., Jake Paul vs. Mike Tyson, NFL Christmas games). Established an event-driven workflow that eliminated weeks of manual preparation across teams, reduced operational risk, and created a repeatable standard for live-event readiness.

  Replaced ad-hoc, human-driven load testing, cluster sizing, hardware provisioning, and cloud scaling with a unified automation framework adopted across the data platform organization, setting a repeatable standard for future live-event readiness.
- **Building Relational Database Abstractions** Conceived and led a zero-to-one relational datastore abstraction that provides a SQL-aware, policy-driven access layer between applications and relational databases, centralizing authentication, routing, connection management, and query handling. Authored the foundational architecture and execution roadmap, defining design principles, phased delivery, and adoption strategy to guide org-wide implementation and company-wide adoption.
- **Right-Sizing the NoSQL Fleet** Led an effort to right-size Netflix's NoSQL database fleet by designing a telemetry-driven capacity model that identified and vertically scaled down over-provisioned clusters. Extended an open-source capacity planner to translate real production workload signals into precise CPU, memory, and storage requirements per cluster, resulting in a **25% reduction in infrastructure costs** and multi-million-dollar annual savings.
- **Aurora Migration:** Served as the technical owner for **Paving the Path for Aurora PostgreSQL** initiative, focused on migrating other relational databases at Netflix to AWS Aurora PostgreSQL. Defined the target architecture and led the design of end-to-end, automated migration workflows, including deterministic cutover and rollback mechanisms. Drove alignment across data, security, application, and program teams to meet performance, reliability, and compliance requirements, and established a repeatable migration standard adopted across Netflix. This work was featured in my presentation at AWS re:Invent.

- **Relational Database Fleet Management:** Owned the reliability and evolution of Netflix's relational database fleet (1,000 production and test clusters spanning Aurora/RDS PostgreSQL, CockroachDB, and MySQL). Defined operational standards for upgrades, red-black refreshes, and self-service provisioning; led fleet-wide performance and security improvements, including in-house CA certificate rotation. This work was presented at Roachfest.

Yugabyte Inc. Core Database R&D Team — Sunnyvale, CA — Aug 2020 – Feb 2023

- **Automatic Shard Splitting in YugabyteDB** Implemented shard-pruning at the query layer to eliminate unnecessary cross-shard RPCs to improve read performance in sharded databases.
  - Github Commit – cb8d9c
  - Github Commit – 57bee5
- **Selective Update Optimization** Developed an update-path optimization for sharded tables that uses UPDATE query keys to avoid unnecessary index lookups, reducing RPC traffic and significantly improving write performance.
  - Github Commit – cf9418
- **Safe Handling of OOM Scenarios in YugabyteDB** Built a post-crash analysis and recovery mechanism that hooks into PostgreSQL's interrupt handling to isolate OOM-causing queries, preventing full-query aborts and improving database stability and throughput.
  - Github Commit – 5738e0f
- **Optimizing Nested Loop (NL) joins by batching RPCs** Addressed inefficient execution of NL joins in YugabyteDB by implementing batched RPC execution. This improved the performance of NL joins in YugabyteDB by up to $7x$.
  - Github Issue – 7836

Uber Technologies Inc, Developer Platforms Team — Palo Alto, CA — March 2019 – Aug 2020

- **Shadow: Safe Rollout System** Designed and implemented a shadow-traffic system to validate microservices against live production traffic in isolated environments, improving rollout safety and test coverage without impacting system state.

Selected Talks and Publications

How Netflix autopilots migration from Amazon RDS to Aurora at scale — AWS re:Invent 2024
AWS re:Invent 2025 Slides

How Netflix enables CockroachDB-as-a-service — Roachfest 2024
Roachfest 2024
Promo Full talk

Caliper: Interference Estimator for Multi-tenant Environments Sharing Architectural Resource — ACM TACO 2019
ACM Transactions on Architecture and Code Optimization

GrandSLAm - Guaranteeing SLAs for Microservices executing in Serverless Computing Infrastructures — Eurosys 2019
European Conference on Computer Systems, March 2019

Proctor: Identifying Interference in Shared datacenters — ISPASS 2018
International Symposium on Performance Analysis of Systems and Software, April 2019

Prophet: Precise QoS Prediction on Non-Preemptive Accelerators to Improve Utilization in Warehouse Scale Computers — ASPLOS 2017
International Conference on Architectural Support for Programming Languages and Operating Systems, April 2017