Sridhar Mallareddy

sridharmallareddy@gmail.com | /sridharmallareddy | /imsridhar | sridharmallareddy.pages.dev

Summary

Platform / Infrastructure engineer with 3+ years specializing in performance engineering and reliability of large-scale distributed systems on Kubernetes (EKS/AKS). Track record of diagnosing the non-obvious bottleneck — across message brokers, databases, and query engines — and fixing it systematically rather than adding capacity. Currently at Kore.AI, scaling a multi-tenant conversational-AI platform to 15k+ concurrent sessions across AWS and Azure.

Experience

Kore.AI Jun 2023 – Present
Software Engineer, Platform Infrastructure Performance *Hyderabad, IN*

- Scaled platform from 6k to 15k+ concurrent users by diagnosing four sequential bottlenecks across **RabbitMQ**, **Redis**, and **MongoDB** that had resisted all prior scale-out attempts; platform error rate dropped from 6% to 0.05%. MongoDB work included shard key redesign across 7 collections (86% of critical queries were doing full-cluster scatter) — compound keys aligned to access patterns improved shard utilisation from 33% to 75%, 5–10× query improvement on highest-traffic collections.
- Stabilised production **Trino** cluster after recurring crashes every ∼1 week — profiled workload patterns to right-size JVM heap, per-query memory caps, worker concurrency limits, and GC tuning (G1GC region sizing, pause targets) aligned to actual query mix; zero unplanned restarts since, dashboard query time from 20+ seconds to under 3.
- Replaced CPU **HPA** with **KEDA**-driven autoscaling on per-pool saturation signals (RMQ queue depth, event-loop lag, RPS per pod) and split one node group into workload-specific pools — reduced infrastructure cost ∼**27%** per scaling event.
- Built MongoDB → **Kafka** → **AWS Glue** → **Apache Hudi** → S3/Athena analytics pipeline; added data freshness SLI after a silent failure left customer dashboards 6 hours stale with zero error signals.
- Led zero-downtime VM-to-**Kubernetes** migration (graduated traffic shift, flag-based rollback); reclassified and redesigned health probe contracts per deployment type — stateless services, workers, and stateful components each got probe thresholds and endpoints matched to their actual startup and readiness semantics, eliminating cascading restart incidents that had carried over from the VM model.

Samsung R&D May 2022 – Jul 2022
Software Engineer Intern, OnDevice AI *Bangalore, IN*

- Optimised ML model performance on Samsung Neural Accelerator Platform (SNAP) via ablation studies; added **OpenCL** support to the inference framework.

Research

Computer Systems Group, IIIT Hyderabad | *Prof. Deepak Gangadharan*

- Proposed a framework to quantify data age under transient faults and thermal constraints in single- and multicore systems with minimal computational overhead.

Parallel Computing Systems Group, University of Amsterdam | *Prof. Anuj Pathania*

- Built a performance-prediction model for dynamic thread migration to optimise LLC latency vs. power budget — 7.3% throughput boost at thermal threshold, migration overhead under 52μs.

Technical Skills

Languages: C/C++, Node.js, JavaScript, TypeScript, Python, Bash/Shell, SQL, Rust
Orchestration: Kubernetes (EKS, AKS), Helm, KEDA, Karpenter, Istio, Docker, Terraform
Datastores: MongoDB (sharded), Redis, RabbitMQ, Kafka, PostgreSQL
Data / Analytics: Trino, Apache Hudi, AWS Glue, S3, Athena
Observability: Prometheus, Grafana, Datadog, OpenTelemetry, OpenMetrics, JVM/G1GC tuning

Education

International Institute of Information Technology Hyderabad, IN
MS, Real-Time Systems & Computer Systems — Funded Research & Thesis *Jul 2021 – May 2023*

International Institute of Information Technology Hyderabad, IN
B.Tech (Honors), Computer Science & Engineering *Jul 2018 – May 2022*

- Dean's List. Research Award. TA for Computer Architecture, Real-Time OS, Intro to ML, HCI.