Sridhar Mallareddy

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

Summary

Platform and Backend Engineer specializing in high-scale distributed systems, event-driven data pipelines, and cloud-native infrastructure across AWS and Azure. Proven track record in microservices orchestration, performance tuning, and distributed database optimization to reduce compute costs and improve system reliability. Core stack: JavaScript, TypeScript, Python, SQL, Kubernetes, distributed messaging, and NoSQL/OLAP data stores.

Experience

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

- Scaled core messaging infrastructure from 6k to 13k+ concurrent sessions by removing async processing bottlenecks across **RabbitMQ**, **Redis**, and **MongoDB**; reduced overall platform error rates from 6% to 0.05%.
- Redesigned MongoDB sharding across 7 high-traffic collections, eliminating full-cluster scatter queries (previously 86% of traffic) using access-pattern compound keys; raised shard utilization from 33% to 75% and cut query latency by 5–10×.
- Fixed MongoS connection fan-out at 13k req/s without adding database capacity by routing 250+ application pods through node-local proxy configurations, dropping connection load by 8× and meeting target SLAs.
- Built a real-time analytics pipeline (**MongoDB → Kafka → AWS Glue → Apache Hudi → S3/Athena**); introduced data-freshness SLIs and automated recovery to fix pipeline lag, bringing dashboard data freshness down from 6 hours stale to sub-minute.
- Replaced CPU-based autoscaling with event-driven **KEDA** scaling based on queue depth, event-loop lag, and request rates; isolated workloads into dedicated node pools to cut overall compute costs by 27%.
- Stabilized a production **Trino** cluster by tuning JVM heap allocations, G1GC settings, per-query memory caps, and worker concurrency; eliminated OOM crashes and dropped dashboard query latency from 20s+ to under 3s.
- Resolved RabbitMQ CPU load spikes above 150 by tuning Erlang scheduler settings and splitting a monolithic 32-node cluster into four isolated 8-node clusters with `ha-two` mirroring policies.
- Eliminated deployment restart cascades in Kubernetes by configuring service-aware liveness and readiness probes tailored to actual boot and recovery timelines across workloads.

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

- Optimized ML-model performance on Samsung Neural Accelerator Platform (SNAP) through ablation studies and added **OpenCL** support to the inference framework.
- Built a performance-analysis tool to identify and resolve model-execution bottlenecks.

Technical Skills

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

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 optimize LLC latency versus power budget; achieved a 7.3% throughput boost at the thermal threshold with migration overhead under 52μs.

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; teaching assistant for Computer Architecture, Real-Time OS, Introduction to ML, and HCI.