Nikhil Baskar

8667877361 | vishalnikhil0307@gmail.com |blog|github.com/Nikhil0307 | linkedin | portfolio

Summary

Software Engineer specializing in distributed systems and high-performance infrastructure.Software Engineer with ~3 years of experience in designing and scaling distributed systems and caching architectures. Proven ability to build low-latency, high-throughput systems and resolve performance bottlenecks. Hands-on with gRPC, Go, Python, and modern serialization protocols. Strong ownership, system design skills, and a passion for backend engineering where scale and reliability matter.

Skills

Languages: Python, Go, Java, SQL, Bash
Frameworks & Tools: gRPC, REST API, Kubernetes, Docker, Redis, RabbitMQ, CherryPy, Spring
Databases & Storage: MySQL, MongoDB
Infrastructure & Cloud: AWS(S3, EC2, DynamoDB), CI/CD, Prometheus, Grafana, Git
System Design: Distributed Systems, Event-Driven Architecture, Fault Tolerance, Distributed Messaging, Horizontal Scaling
Performance Optimization: Scalability, Load Balancing, Caching Strategies, Observability, Profiling & Benchmarking

Experience

Member Technical Staff, Zoho Corporation – Chennai **06/2023 – Present.**

- **Orchestration & Scalability:** Designed and implemented a Kubernetes-based orchestration layer that simplified AI/ML workload management, reducing deployment time by 40%, cutting manual configurations by 60%, and improving developer productivity by 50%.
- **Infrastructure & Deployment:** Migrated monolithic service to Kubernetes-based microservices, enabling auto-scaling, rolling updates, and zero-downtime deployments, improving scalability by 3x and reducing downtime by 90%.
- **High-Performance Distributed Storage:** Built a gRPC-based distributed database, supporting 1M+ requests per day, functioning as a distributed cache with pluggable modules like similarity caching, TTL-based eviction, pre-fetching, and write-back caching, reducing cache misses by 35% and improving read speeds by 70%.
- **Optimized Docker Workflows:** Migrated to multi-stage Docker builds, reducing image sizes by 40%, cutting build time from 10 minutes to 4 minutes, and improving deployment efficiency by 3x.
- **Message Queue Orchestration:** Engineered a dynamic RabbitMQ consumer manager during Kubernetes migration, enabling auto-scaling based on queue metrics, eliminating OOM failures for long-running processing tasks, improving system throughput by 2.5x while maintaining 99.9% message delivery reliability.

Project Trainee, Zoho Corporation – Chennai **08/2022 – 05/2023.**

- **ML Model Optimization:** Pioneered ONNX adoption, eliminating dependency conflicts across 150+ ML models, resulting in 30% faster model deployment and reducing API response time by 25%.
- **Low-Latency Model Serving:** Converted ML models to ONNX format, optimizing 200K+ inference requests per day, reducing average latency from 150ms to 100ms while maintaining 99.8% accuracy.
- **In-Browser ML Inference:** Enabled client-side model inference by loading lightweight ML models into browsers, reducing server load by 50% and improving response times by 2x.
- **Efficient In-Memory Caching:** Developed an LRU-based in-memory caching system, reducing datastore queries by 40%, cutting latency by over 50%, and handling 2M+ cache lookups per day with 99.9% cache hit rate.

University Project

CookBook - Recipe Web Application

- Built a Recipe web application using Struts, Servlets, and Spring MVC, combining legacy systems with modern Spring technology to deliver a scalable and maintainable backend
- Designed and implemented a caching mechanism to reduce backend calls, integrated Apache Lucene to enhance search functionality, and utilized the ELK stack for faster and more efficient search result processing

Education

Veltech Multitech Dr Rangarajan Dr Sakunathula Engineering College – B.e in Computer Science **07/2019 - 05/2023**