

Swapnil Surdi

San Jose, California

• +1 (213) 294 8993 • swapnil.surdi@icloud.com • [in surdi](#) • [swapnilsurdi](#)
•  Google Scholar - 66E21KUAAAAJ

Summary

Senior Software Engineer with 8+ years architecting scalable backend systems and cloud-native microservices. Delivered 99.9% uptime for platforms handling 100K+ daily requests and cut API latency from 12 seconds to below one second through optimized AWS architecture. Skilled in AI-enabled RAG pipelines with MCPs, guardrails, evaluations, and LLMOps workflows. Expert in LLM-driven development and automation. Proven track record in 0-to-1 products and rapid prototyping in regulated environments (ISO 13485, HIPAA)

Skills

Languages: Python, Go, TypeScript/JavaScript, Java

Backend & APIs: Node.js, Django, FastAPI, SpringBoot, GraphQL, REST, Kafka, RabbitMQ

Cloud & Infrastructure: AWS (ECS, Fargate, Lambda, DynamoDB, S3, CloudFront), Docker, Kubernetes

AI/ML Systems: MCP, LLM Integration, LangChain, RAG Pipelines, Agentic Workflows, Guardrails, Evaluation

Data: PostgreSQL, MongoDB, Redis, DynamoDB, Oracle

Compliance: ISO 13485, HIPAA, IEC 62304

Experience

Senior Software Engineer

Sept 2021 – July 2025

Los Angeles, CA

- Treatment Technologies & Insights Inc
- Architected 0-to-1 HIPAA-compliant microservices platform on AWS Fargate supporting 4 enterprise healthcare clients, designing core services (Profile, Health Data, Organizations, Trials) with standardized error handling, distributed tracing, and auto-scaling achieving 99.9% uptime.
 - Built an agentic RAG system with MCP tools for documentation and compliance workflows, enabling multi-hop reasoning across the codebase, Confluence, and Jira using 768-dimensional embeddings in Qdrant, which accelerated documentation retrieval and streamlined compliance verification
 - Reduced patient data API latency from 12s to sub-second (P95) through DynamoDB partition key redesign, batch processing, and Redis caching at API Gateway layer for 1000+ patient datasets.
 - Designed cross-platform dynamic survey engine with version-controlled JSON schemas, enabling new client onboarding in days instead of months while supporting 12+ languages including RTL, with unified rendering across React, iOS, and Android.
 - Built AWS S3/CloudFront CDN architecture for medical imaging (OHIF viewer), reducing 500MB study load times from 90+ seconds to under 25 seconds through worker-based processing and edge caching.
 - Implemented distributed error tracking with partial UUID correlation across frontend and backend, reducing MTTR by 70% through centralized watchdog service.
 - Led ISO 13485 eQMS implementation on Confluence/Jira, automating CAPA workflows and SDLC documentation, shortening release cycles for Class I medical device certification.
 - Developed Go-based testing framework enabling parallel end-to-end testing and rate limiting validation, achieving 85% test coverage.
 - Designed JWT-based multi-tenant configuration system enabling single-deployment architecture, reducing infrastructure costs by 60%.
 - Mentored three junior engineers on system design and AWS best practices, conducted architecture reviews, and set team coding standards, which improved code quality and accelerated feature delivery

Lead Network Engineer

Aug 2020 – Sept 2021

Los Angeles, CA

- Beamlink Inc.
- Architected mesh networking backbone for disaster communication using OLSR protocol with optimized routing intervals for battery efficiency.
 - Built Python Django coordination server managing 4G tower discovery, internet routing, and packet prioritization (voice, control, data).
 - Designed an alpha node election algorithm for dynamic internet gateway selection with automatic failover across the mesh network, ensuring continuous connectivity during node failures
 - Contributed to patent (US20240129000A1) for antenna system with MIMO configuration.

Application Developer	Oct 2016 – June 2018
Apprely Technologies	Pune, India
● Built financial platform with Kafka/RabbitMQ for asynchronous processing, integrating QuickBooks, Zoho, and Plaid/Stripe payment gateways.	
● Developed Android applications with Django REST backends for healthcare appointment management and GPS-based outdoor navigation.	
Associate Software Engineer	Oct 2015 – Oct 2016
Accenture	Pune, India
● Automated Salesforce regression testing for Virgin Media UK using Selenium, processing 60K+ tests and reducing 2 months manual QA to under 1 hour. Awarded Ace Prize.	

Selected Projects

SmartContext (AI Personal Assistant)	In Development
● Building context-aware AI assistant leveraging Qdrant embeddings for persistent user context and multi-tool orchestration.	
MCP-Cache	https://github.com/Chapus/mcp-cache
● Published NPM utility for MCP agents optimizing LLM token usage through intelligent caching of large tool outputs.	
Launchlab	https://github.com/swapnilsurdi/launchlab
● Built Docker-based homelab orchestration with one-liner deployment for self-hosted services (Jellyfin, Immich, Paperless-ngx).	

Education

Master of Science, Astronautical Engineering	2020
University of Southern California	GPA: 3.5
Bachelor of Engineering, Electronics & Telecommunications	2015
Savitribai Phule Pune University	GPA: 3.5

Certifications: Six Sigma Green Belt (USC Marshall), HIPAA Compliance, GDPR Data Security, QA Training (Accenture)

Patents & Publications

- Patent (Pending): *Fixed Base Station Antenna System using MIMO Configuration* (US20240129000A1).
- Publication: *Space situational awareness through blockchain technology*, Journal of Space Safety Engineering, 2020.