Sudarshan Mehta

+1-669-281-9888 | sudarshan.mehtacs@gmail.com | linkedin.com/in/sudarshan-mehtacs | github.com/sudarshan-mehta | Santa Clara, CA

PROFESSIONAL SUMMARY

Software engineer with 6+ years building high-throughput distributed systems and low-latency trading infrastructure. Cut system latency by 95% (2 min to 100ms), scaled microservice architectures to 25,000+ RPS, and shipped ML-driven personalization systems on AWS. Researching LLM inference optimization at the hardware-software boundary. M.S. Computer Science, Santa Clara University (GPA 3.85).

CORE COMPETENCIES

Distributed Systems | Low-Latency Architecture | ML Infrastructure | LLM & GenAI | Apache Kafka & Kubernetes | Data Pipeline Orchestration | System Design at Scale | Java / Python / Go / C++

WORK EXPERIENCE

Amazon **Software Engineer, Console Analytics** Jul 2025 - Present | Palo Alto, CA

- Built a dataset pipeline using Helios generating fractional engagement metrics that drive personalized experiences across AWS Console.
- Designed and shipped a dynamic UI generation system where a chat agent renders contextual UIs in real time using Amazon Q (Aurora), delivered end-to-end on AWS.
- Simplified a legacy architecture to a single-cell design, improving reliability and deployment velocity. Orchestrated multi-stage workflows via Apache Airflow, Lambda, DynamoDB, and Step Functions.

Morgan Stanley **Software Development Engineer I / II** Aug 2021 - Sep 2023 | Bangalore, India

- Engineered a dynamic pricing grid using maturity-based linear interpolation, cutting fixed income trade pricing latency from 2 minutes to under 100ms (95% improvement).
- Managed 42 microservices on Kubernetes with load balancing, sustaining 25,000+ RPS at peak load. Built Rice, a low-latency Scala pricing engine for real-time fixed income trades.
- Owned Taser (Trade Approval System) end-to-end on Azure -- reduced approval latency by 40%, raised unit test coverage from 35% to 80%, and built a full CI/CD pipeline.

CLSA Technology & Services **Software Development Engineer I** Jul 2018 - Aug 2021 | Pune, India

- Built Stingray, a high-touch low-latency equity trading system from scratch, connected to 11 exchanges worldwide on AWS. Cut system load time 94% (3 min to under 10 sec) via static data caching.
- Built a custom log analysis tool (similar to Elasticsearch) that maps interconnected trade IDs, triggers alerts, and surfaces system insights across the trading stack.

PROJECTS

TuneX Open Source -- Fine-tuning platform to train any open-source LLM on custom datasets. No ML expertise required. Built with Python and React.

Smart NIC Tokenization Research -- Offloading LLM tokenization to NVIDIA smart NICs via Triton Inference Server to reduce time-to-first-token in inference pipelines. (Santa Clara University)

Custom Relational Database (C++) -- Built a database engine from scratch: storage, query parser, CRUD, JOINs, and B-tree indexing.

EDUCATION

Santa Clara University -- M.S. Computer Science & Engineering Sep 2023 - Mar 2025

GPA: 3.85 / 4.0 | Research: LLM tokenization offloading to Smart NICs | ACM, Hack for Humanity, Competitive Programming Club

MIT Academy of Engineering -- B.E. Information Technology Jul 2014 - Jul 2018

SKILLS

Languages	Java, Python, Go, C++, Scala, C#, TypeScript
Frameworks	Spring Boot, Flask, TensorFlow, Keras, NVIDIA Triton, React, Angular, Flutter
Cloud & Infra	AWS (Lambda, DynamoDB, Step Functions), Azure, Docker, Kubernetes, Apache Kafka, Apache Airflow
Databases	PostgreSQL, MySQL, MongoDB, DynamoDB, IBM Db2, Firebase