

SRI HARSHA

System Architecture Architect | AI/Agentic Strategist | Telecom Systems

+91 9880122928 | reach.sriharsha@gmail.com | Bengaluru, India | US Patent Holder (US011350319B2)
[linkedin.com/in/sri-harsha-52445811](https://www.linkedin.com/in/sri-harsha-52445811)

PROFESSIONAL SUMMARY

22+ years building telecom and AI systems — from kernel-level packet processing to agentic AI platforms. Architect who has taken 3 products from R&D to commercial deployment: world's first commercial 3GPP Femto Gateway (SFR France), world's first 5G-NSA deployment (T-Mobile), and supporting Reliance Jio's greenfield all-IP LTE network.

Unique career arc as a product architect across every layer of the stack — bare-metal C/C++ with DPDK, cloud-native Kubernetes microservices in Golang, and now full-stack AI/LLM/Agentic infrastructure — with each layer shipped into commercial products, not internal tooling. Consistently the one who takes a product from concept and spec to field deployment: authoring PRDs/FSDs/DSDs, owning design reviews, and driving MVPs through to production.

Currently driving strategic AI adoption at Ribbon Communications: built on-premises LLM clusters (12 GPUs, 230GB VRAM), deployed RAG chatbots saving \$100K+/year, and engineered agentic RFx systems that cut sales response time from months to days.

KEY HIGHLIGHTS

22+ Years

Total Experience

3 Products

R&D to Commercial

3B Events/Hr

Real-time Processing

US Patent

QoS Optimization

12 GPU / 230GB

On-Prem AI Cluster

WORK EXPERIENCE

Ribbon Communications — Research Labs | System Architecture Architect | 2021 – Present

Core architect across Ribbon's **Acumen** platform — the company's umbrella product for network intelligence and AI, spanning its **Analytics**, **AI Platform**, and **Builder** pillars — alongside a standalone cross-organizational PoC and org-wide strategic advisory scope.

- **Acumen Analytics** (formerly Ribbon Analytics Platform): Core architect on Ribbon's cloud-native, real-time big-data analytics platform — architected systems processing **3 billion events/hour** (designed to scale to **400 billion**), serving **2.5M customers** at Airtel India. Generates KPIs across both **IMS and wireless network layers** — **VoLTE, VoWiFi, and VoNR** — using **proprietary algorithms and ML-based clustering and most-probable-root-cause identification**, giving telecom operators visibility through every layer of their network.
 - **5G VoNR Analytics**: Extended to the full 5G stack (AMF, AUSF, SMF, gNodeB, UPF, PCF, UDM, CHF, NRF, NSSF, SCP); became a quasi-NWDAF; first VoNR deployment at Airtel India.
 - **MINT (Mobile Inband Telemetry)**: Intel Tofino switch + P4 programming, processing terabits/second; RESTful feedback/command loop with OAM systems; rich metadata for ML-based network analysis.
 - **XDR Generation**: Passive monitoring across the 5G network producing eXtended Detection & Response records.
- **Acumen AI Platform**: Second pillar of Acumen — architected Ribbon's self-service AI platform (FastAPI + Celery + Flask + MCP Servers on distributed Kubernetes), a scaled-down Amazon Bedrock equivalent with user self-registration, custom knowledge bases, and self-issued API keys for external agentic workflows.
 - **Acumen AI Cluster**: Built a 5-node Kubernetes cluster — 12 physical GPUs (8× L40S + 4× Tesla M60), 230GB VRAM, 192 virtual GPU slices.
 - **RAG Chatbot**: Pioneered on-premises LLM deployment — 100+GB knowledge base, 10M+ tokens served, 92% retrieval hit rate, **\$100K+ annual savings**, full data sovereignty for thousands of users.
 - **RFx Responder**: Agentic AI system addressing ~15,000 technical questions in minutes; reduced sales first response **from months to days**; accuracy improved **75% → 87%**.
 - **LLM Fine-Tuning**: Built pipelines using Unsloth/LoRA/QLoRA/RLHF/RLAIF across Llama, GPT-OSS, and DeepSeek; created custom LLMs and SLMs for chat and automated test-case generation.
 - **LLM Observability**: Arize-Phoenix and Langfuse for tracing, evaluation, and dataset/ground-truth creation.

- **Sentry Integration:** Deployed for production issue tracking across products.
- **Acumen Builder:** Third pillar of Acumen — core member conceptualizing and productizing Ribbon's visual **no-code/low-code agentic workflow engine** (Langflow-class capability, built in-house). Composable "lego-brick" building blocks — knowledge bases, agents, and MCP integrations — spanning multiple LLM providers (**Azure, AWS Bedrock, and on-prem OpenAI-spec APIs**), letting enterprise teams assemble their own agents. Took the product from concept through to production deployment at **Altice**, generating **double-digit-million-dollar revenue**.
- **TM Forum Drone Network Assurance PoC** (*DTW Ignite 2026, Copenhagen — Catalyst project finalist*): Architected and built a full-stack PoC demonstrating **closed-loop drone telemetry over live 5G networks** — real-time flight-path ingestion and simulation, cell-aware drone positioning against live gNodeB topology, and a rules-driven proximity engine that feeds drone-proximity events back into the network as live 5G network-assurance actions (coverage-quality forecasts, cell-boundary/congestion alerts, slice requests) — closing the loop between airborne telemetry and network response. A joint **Ribbon/Wavelo + CGI Catalyst** that cut across network, product, and platform teams; demonstrated live to telecom operators at DTW Ignite 2026, Copenhagen, and selected as a project finalist.
- **Market Strategy, MWC & CXO AI Advisory:** Competitive intelligence, product viability assessments, and MVP definition — including next-gen analytics prototyping on open-source free5GC. Owns delivery of customer-facing POC demos and MVPs for Mobile World Congress and other global events, engaging VPs/C-suite/Sales leadership through full dev-to-deploy cycles. Trusted AI strategy advisor to VPs, SVPs, PMs, and CXOs on build-vs-buy decisions and investment prioritization; key member of the AI/ML Steering Committee driving enterprise GenAI adoption (introduced Codeium, Windsurf, Cursor, Copilot); prepares executive AI strategy decks and technology roadmap presentations.

Cisco Systems India Pvt Ltd | Software Engineer IV — Mobility Unit | 2015 – 2021

- **5G SMF (Cloud-Native, Industry-First):** Core team member who built Cisco's industry-first cloud-native 5G Core with a full Service-Based Interface (SBI) architecture — entire product built ground-up in **Golang**, with every layer of the technology new to telecom at the time. Delivered **T-Mobile USA's first countrywide 5G (SA) network**. Owned the full **QoS Model** — interfacing with PCF, UPF, AMF, and UDM — the heart of the 5G Standalone control plane. Architected session management, session handover, and sticky-session routing for stateless processing, scaling to **16,000 calls/second**. Led delivery for a team of **8 engineers**.
- **5G-NSA (T-Mobile):** Architected 5G-NSA solution on MME/PGW for world's first 5G-NSA deployment; implemented path update procedures and extended QoS attributes.
- **MPTCP Prototype:** Prototyped Multipath TCP on DPDK for high-throughput networking; ported TCP stack to DPDK; explored scheduling algorithms (Reno, Vegas, Lowest RTT).
- **SaMOG WiFi Gateway:** Designed S2a WiFi Offload features (Trusted WiFi in LTE, roaming, web auth); deployed across Indian colleges and public WiFi.
- **Call Flow Diagram Tool:** Built sequence diagram generation from distributed K8s pods — single flow view across pods.
- **IEC-5 Startup — 5G-SPACE:** Semifinalist among 400+ submissions (top 20) in Cisco CTO organization.

Samsung Research Institute, Bangalore | Chief Engineer | 2014 – 2015

- **Reliance Jio:** Led TAC team supporting world's first and largest all-IP 4G/LTE network — 125K gNodeBs with Evolved Packet Core.
- **EPC Operations:** OAM of MME, SGW, PGW; managed KPIs for 100,000+ eNodeBs; managed team of 7 engineers.

Nokia Solutions & Networks | Product Architect / OAM Expert | 2008 – 2014

- **World's First Femto Gateway:** Architected and delivered world's first commercial 3GPP Femto Gateway (HNBGW) for SFR, France — lab to field deployment.
- **FCAPS @ Scale:** 60,000 concurrent Femto APs with 99.999% availability using active/backup redundancy.
- **Agile Transformation:** Led transformation of 120-member cross-functional team; Scrum Master for multiple teams.
- **CI/CD Pioneer:** Handcrafted CI/CD system using Perl and Expect before Jenkins existed; automated testing framework with 5,000+ test cases.
- **Health Monitor Console:** Observability platform predating Prometheus/Grafana — BU-level innovation finalist.
- **IMS CSCF Redesign:** Converted bare-metal IMS CSCF to cloud-native using DPDK; achieved 400% SIP packet processing efficiency.

- **Features:** Multiple-CN, IuFlex, AGPS, Location Locking, Area-based Charging, RADIUS/EAP-SIM (RFC2865/RFC4186), SOAP Web Service, MySQL scaling to 50M entries, flow-based logging, NETCONF/YANG, SNMP, TR-069/TR-196.

Motorola India Pvt Ltd | Software Engineer | 2004 – 2008

- **3G Flex & Feature Database:** Core developer for P2K phone platform — NV memory and file system management for 3G handsets delivered across 5 continents.
- **Scaling:** Scaled from few to 172 simultaneous customers at Razr peak — 6X customer scaling with 4X delivery performance.
- **OTA Updates:** Architected Over-The-Air update system for 3G phones; built E2E Validation Tool (VC++) and Advanced Flex Tool.

TECHNICAL SKILLS

AI / ML / Agentic: LLMs, SLMs, RAG, LoRA, QLoRA, RLHF, Unsloth, vLLM, Ollama, LangChain, LangGraph, LlamaIndex, Langflow, Ragflow, Dify, HuggingFace, PyTorch, OpenAI API, Claude Code, GitHub Copilot, Windsurf, Cursor, Codeium, Fine-tuning Pipelines, Document Classification, Sentiment Analysis, Prompt Engineering, Custom LLM/SLM Creation, TTS/STT Models, MLOps, Synthetic Data Pipelines, MCP Servers, Agentic Workflows, ReAct Patterns, Agent Memory, Context Engineering, LLM Selection Strategy, Embedding Model Selection

Telecom & Networking: 5G Core (SMF, AMF, UPF, PCF, AUSF, UDM, CHF, NRF, NSSF, SCP), 4G/LTE (MME, SGW, PGW), 3G/UMTS, VoLTE, VoNR, IMS (HSS, CSCF, DRA), Femto Gateway (HNBGW), SaMOG, EPC, NWDAF, FCAPS, OAM, 3GPP Specifications, IETF Standards, free5GC

Programming Languages: C/C++ (20+ yr), Python (10+ yr), Golang (8+ yr), Perl (15+ yr), Shell Scripting, JavaScript, PHP, VC++

Protocols (40+): IP/IPv4/IPv6, TCP/UDP, GTPv2, RADIUS, Diameter, EAP-SIM, EAP-AKA, IPSec/IKEv2, gRPC/Protobuf, SNMP, NETCONF/YANG, TR-069, SIP, MPTCP, HNBAP, RANAP, GRE/EoGRE/PMIP, SOAP, REST

Cloud & Infrastructure: Kubernetes, Docker, Cloud-Native Microservices, AWS Bedrock, GCP, Azure, Distributed Systems, HA (99.999%), Containerization

Systems & Performance: DPDK, P4 Programming (Intel Tofino), eBPF, Multi-core Programming, Linux Internals, Real-time Packet Processing, Terabits/sec Processing

Data & Databases: ClickHouse, Hadoop, Parquet, Milvus (Vector DB), Redis, PostgreSQL, MySQL, Chroma

Observability: Prometheus, Grafana, Arize-Phoenix (LLM Evals/Tracing/Dataset Creation), Langfuse (LLM Observability/Eval Sets/Prompt Tracking), Langsmith, Sentry, OpenTracing

DevOps & Tools: Git, SVN, ClearCase, Jenkins, CI/CD, JIRA, Confluence, GCC/G++, GDB, Valgrind, Wireshark, strongSwan

AI LAB & EXPERIMENTATION COOKBOOK

Active personal AI experimentation lab — a living cookbook of hands-on projects exploring cutting-edge AI/ML. GitHub: github.com/reachsriharsha/mycookbook

- **Voice AI:** Built real-time voice agent using LiveKit + Sarvam Saarika STT v2.5 + Ollama Gemma3:4B + Sarvam Bulbul TTS v2. Tested multiple speakers for Indian English voice quality. Gemma video/audio fine-tuning.
- **LLM Fine-tuning Lab:** Comprehensive Unsloth cookbook — LoRA/QLoRA on Llama 3.2 3B, PDF-to-QA pipelines (PyMuPDF), GGUF export for Ollama, ROUGE/BERTScore evaluation. Classification fine-tuning, Lamini experiments.
- **RAG & Compliance AI:** RFQ compliance assessment with DeepSeek-R1:8B + LangChain + Chroma (multi-stage retrieval: cosine similarity, cross-encoder re-ranking, fuzzy matching). ExcelRAGProcessor for intelligent table chunking. Milvus backup/restore tooling.
- **Agentic Frameworks:** Custom Langflow components (KB List Fetcher, Ollama integration, regex processors). Go-to-Celery bridge. 460+ section Self-Service AI Platform architectural wiki.
- **Full-Stack:** Firebase + FastAPI RBAC authentication. OTP auth with token refresh. Data Explorer (Flask). LangChain dynamic prompting techniques.
- **NLP & Deep Learning:** LSTM networks, text summarization, POS tagging, NLTK, HuggingFace text-to-image pipelines.

PRODUCT & LEADERSHIP

Product Lifecycle (Concept to Market), PRD/FSD/DSD Authoring, Revision Design Specifications, MVP Definition, Market Analysis & Competitive Intelligence, Agile Transformation (120 members), Scrum Master, Cross-functional Team Leadership, TAC Management, Mentoring Engineers, Customer Interface & Field Deployment.

CERTIFICATIONS

Introduction to LangGraph — LangChain Academy | December 2025

Foundation: Introduction to LangChain — Python — LangChain Academy | December 2025

Agentic AI! — DeepLearning.AI | October 2025

Building AI Voice Agents for Production — DeepLearning.AI | September 2025

Advanced NLP Techniques: LoRA for Fine-Tuning Llama3 LLMs — Udemy | February 2025

Machine Learning for Absolute Beginners — Level 1 — Udemy | February 2023

PATENT & AWARDS

US Patent US011350319B2: QoS optimization for inter-RAT handover

- IEC-5 Semifinalist — 5G-SPACE venture, top 20 of 400+ at Cisco
- TM Forum DTW Ignite 2026 Catalyst Finalist — Drone Network Assurance PoC (Ribbon/Wavelo + CGI)
- BU Innovation Finalist (x2) — Real Time Analyzer for Logs & Health Monitor Console at Nokia
- Most Inspiring Person — Femto Gateway team at Nokia
- Spot Awards — CI framework development & FGW feature integration
- Star Scrum Team Award — Nokia

EDUCATION

M.S. Software Systems — BITS Pilani (Distance)

B.E. Computer Science & Engineering — AIT, Chickmagaluru (VTU)

PERSONAL

Date of Birth: March 28, 1983 | Languages: Kannada, English, Hindi | Interests: Technology exploration, photography, traveling, reading, music