Sri Harsha

SOFTWARE ENGINEERING ARCHITECT

PHONE: +91 9880122928 | EMAILL: REACH.SRIHARSHA@GMAIL.COM | BENGALURU - 560077

Professional Summary

- Highly **accomplished and self-motivated** professional with over 20 years of experience in product development, specializing in telecommunications and AI/ML domain.
- Proficiency with Agentic coding tools like Github Co-pilot, windsurf and aware of limitation, merits of them and how to use them in various stages of software development to deployment.
- Demonstrated aptitude for **rapid learning** and effective **application of new technologies** within enterprise environments.
- Currently **driving the strategic adoption** of **AI** applications and **Agentic** technologies within the organization.
- Experience in building full stack **AI/LLM/ML** infrastructure and application solution. With **AIOps** approach.
- Expertise in **adopting open-source** applications into enterprise solutions, which accelerates product development and also achieves cost optimization.
- Extensive expertise in the design and development of cloud-native, microservices-based systems.
- Engineered the S2a (SaMOG) Wi-Fi Offload gateway solution with various features.
- Architected and implemented 5G-NSA solutions for Cisco MME/PGW for world's first 5G -NSA deployment T-Mobile.
- Led the design and development of policy and user plane management functionalities for Cisco's 5G SMF solution. Implemented 5G QoS modeling within the Cisco 5G SMF.
- Provided critical leadership to the Technical Assistance Center (TAC) team for Jio, supporting the world's first greenfield all-IP 4G/LTE network deployment.
- Pioneered the development to deployment of the first 3GPP-compliant HNbGW for SFR, France.

Work Profile

Ribbon Communications | System Architecture Architect, Ribbon Research Labs **2021-present**

Technical Capabilities: (For detailed project check Project section)

- Architected high-performance systems capable of processing 3 billion events hourly, establishing robust analytics infrastructure for enterprise-scale data. Customer appreciation Link
- Network Analytics system design to scale for 5G VoNR analytics, which provides visibility and observability into operator's mobile networks.
- Pioneered integration of emerging technologies including **P4 programming** and **eBPF** for kernel-level monitoring to enhance product capabilities.
- Developed LLM-based applications using **RAG**, document classification, sentiment analysis, model fine-tuning, and prompt engineering.
- Directed on-premises LLM deployments by selecting right hardware vendor with **NVIDIA M60, A100, H100 and L40S GPUs**, enabling advanced AI capabilities which help to **achieve data sovereignty**.
- Engineered and deployed an innovative **RFC/RFx response system** using AI agentic architecture, reducing sales first response time from months to days.
- Implemented **AI-powered chatbot solutions** across product lines with **100+ GB** data, saving over $100,000 annually and improving team productivity. This served **1000s of users** with more than **10 Million Tokens**
- Fine-tuned open-source models using **Unsloth** and **LoRA** techniques to create custom LLMs for task-specific applications. Quantized LLMs to be hosted with smaller memory foot print.
- Implemented **low-code/no-code agentic frameworks** like Langflow to democratize AI development and accelerate prototyping. Solution proposal and video link

- Currently developing a **self-service AI Kubernetes cluster** across heterogeneous GPU distributed hardware. This provides platform to create AI agents smoothly. Automating the work flows using agentic implementations.
- Scaled development and test optimization using **agentic coding tools** across multiple teams.
- As architect for various solutions responsible for generating Product Requirement Specifications (PRDs), Feature Specification Documents (FSD) and involve rigorous Design Specification review.

Management & Strategic Skills:

- Spearheaded development of next-gen telecommunications products, **driving innovation** in core technology.
- Led **strategic planning** of analytics solutions for 5G technologies, including simulator and hardware finalization.
- Conducted **market analysis, competitive intelligence, product viability assessments**, and MVP definition.
- Served as a key member of the **AI/ML steering committee**, guiding enterprise-wide AI initiatives and roadmaps.
- Introducing new technologies like **Codeium**, **Windsurf**, and **Cursor**, driving adoption across org.
- Collaborated **cross-functionally** with clients, Product Line Management, sales teams, and partners to develop tailored solutions.
- Proven capability to **deliver PoCs and MVPs** for customer AI/ML requirements.

Mentorship, Community Building & Organizational growth:

- Shortlisted, **interviewed, and selected interns** both in Indian and US universities.
- **Mentored interns** on industry projects, tools, and processes to help them complete keystone projects successfully.
- As **Moderator** for internal **AI/ML forum**, helped to solve daily questions, right usage of the tools and helping the organization to adapt Gen AI faster.

Project Profile:

Demo, MVP and PoC for the Mobile World Congress, Global Tech Events and Customer Events:

I am responsible for delivering Demo, Minimal Viable Product (MVP) for major Global Tech events and Customer Demos as part of Ribbon Research Labs.

1. **Stakeholder Engagement:** Served as the primary technical contact for managing communication and gathering requirements from diverse stakeholders, include VPs, C-suite members, Sales Leadership and Product Line Management(PLMs).
2. **MVP/Demo Delivery:** Led the complete development and deployment cycles for high profile MVPs and technical Demos, ensuring alignment with strategic goals for presentation at major Global Tech events like Mobile World Congress(MWC) and Customer Demos.
3. **Technical Translation:** Expertise in translating complex business requirements and high-level concepts into concrete, function and presentation-ready demoable products and supporting materials.
4. **Project Coordination:** Coordinated cross-functional teams and managed meeting schedules to ensure timely and effective conversion of requirements into polished final deliverables.

Retrieval-Augmented Generation (RAG) Chatbot

Production-grade Retrieval-Augmented Generation (RAG) Chatbot built on a Kubernetes/Docker based microservices foundation.

1. **Data Ingestion:** Ingestion pipeline supports various formats (PDF, DOCX, XLSX, HTML, PPT) via LangChain Document Loaders. This pipeline is enhanced by Vision-Based LLMs to extract and describe information from embedded images (e.g., architecture diagrams, user manuals), capturing crucial finer details.
2. **Vector Database & Indexing:** Utilizes Milvus for high-performance vector storage, supporting L2 and Cosine similarity. Indexing employs IVF_FLAT for efficient searching across dynamic fields. The dynamic fields helped in filtering the records based on versatile application requirements.
3. Embedding Pipeline: Leverages Hugging Face BGE (Baidu Gan Embedding) models to generate normalized vectors. The pipeline is managed using a memory-efficient singleton pattern.

4. **Retrieval & Performance:** Employs advanced, context-aware RAG techniques with optimized chunking strategies, achieving a hit rate of 92%. Retrieved results undergo sophisticated re-ranking before integration into the LLM response generation process.

5. **LLM Orchestration:** Features a dynamic LLM layer using various models (Llama, Gemma, Phi, GPT-OSS, Mistral) selected based on query categorization. Orchestration is managed by Ollama for development/experimentation and vLLM for production deployment. Pipelines are defined using LangChain Expression Language (LCEL).

6. **Observability:** Arize-Phoenix and Langfuse were used for the observability of the LLM transaction. The observability include tracing with session correlation for LangChains LCEL calls with token usage tracking and latency metrics. Langfuse also used to build ground truth dataset as well as creating curated dataset which later used in fine-tuning, essentially creating active feed-back loop.

Acumen AI Cluster: NVIDIA GPU based Kubernetes Cluster

Acumen AI Cluster brought up from grounds up using NVIDIA GPUs across different machines. This is being utilized for internal projects and precursor to Production grade AI cluster. AI cluster actively used for demo and experimental AI and ML work loads across various teams. The salient features of the AI cluster.

1. A **5-Node Kubernetes cluster** with 1 control place and 4 GPU based Nodes running on Ubuntu 22.04 and Debian with container runtime. Calico CNI provides the network for pod-to-pod communication.

2. **12 Physical NVIDIA GPUs (8x L40S + 4x Tesla M60)** delivering **230 GB of GPU VRAM** for large scale model inference, training and fine tuning activities

3. **192 Virtual GPU slices** via NVIDIA time-slicing technology, this enables multi-tenant GPU sharing with 16x oversubscription per physical GPU. **NVIDIA GPU Operator** for seamless **GPU Orchestration.** This is leveraged using **CUDA run time 12.7/12.9.**

4. **A High-density compute** with **440 CPU Cores, 1.4 TB system RAM** across the workers. With 12+TB of storage with Supermicro NVMe SSDs.

5. **Production AI/ML Stack:** Ollama and vLLM for serving LLM request, Jupyter PyTorch/CUDA notebooks for the ML workloads along with Milvus for vector database and Langfuse for the LLM observability. NodePort exposed services allow the external API for the LLM services and notebooks.

Acumen AI Platform: A Self-Serving AI platform

Acumen AI Platform is built over the Acumen AI cluster, the platform enables user to register themselves, create their own knowledge bases and chat with the knowledge bases. The platform is built with the following stack which runs on a distributed Kubernetes cluster

1. Python FastAPI Backend: The back end is conceptualized as API providing various services such as User Management, Knowledge Base, Applications, Vector DB interfacing, File and URL ingestion task orchestration etc..

2. Celery Based Distributed workloads: The back end orchestrates the work of various applications using the celery based distributed workload over Kubernetes cluster. Depending on type of work separate queues are formed to handle workloads. The workloads are automatically scaled up and down based on various factors.

3. Flask based Front end: Flask based front end with html, css and bootstrap based simple application creation. This allowed users to log in and access various applications. Users can create their own API keys form this platform and use these keys get access to the services exposed by Acumen AI platform to the external agentic workflows.

4. MCP Servers: The platform caters various Model Context Protocol (MCP) servers and the tools which will help to cater the automated agentic flows.

RFx Responder:

RFx responder provides answer for the customer questions either given product is “compliant”, “not-compliant” or partially compliant.

RFx Responder was developed as an experimental application of RAG AI paradigm, specifically extending utility beyond traditional chatbots. Its primary function is to provide first-pass answers for complex questionnaires such as RFP, RFI, RFQ, DDQ. The tool achieved its precision by simultaneously searching both legacy RFx databases and Custom Knowledge bases. To ensure efficiency, it incorporates parallel processing for question evaluation using celery framework. The results were presented with simple HTML user interface with all supporting references accessible in a separate dedicated view. The tool uses the knowledge base based on Acumen AI Platform.

The impact of the tool is immediate, the Sales Engineering team able to provide a response to customers in days instead of weeks or months. The initial accuracy achieved was 75%, which gradually improved to 87%. This not only helped the seasoned Sales team members but gave knowledge to the new members as well. They can learn about the capabilities of the product from day one, instead of chasing around right point of contact and reading through 100s of pages of documentation.

RFx Responder provides stats using java script chart libraries with answers, which will give the one-shot view of how many features of customer request readily compliant and what new things we need to be delivered and how many are partially compliant. This gives very clear picture within minutes for Sales, Product and Program managers what capabilities we lack for given RFx in our products and solutions.

Fine-Tuning LLMs:

Implemented the fine-tuning pipe-line using unsloth with various models such as llama, gpt-oss, deepseek etc. Initially Fine-Tuning was done as exploratory activity. Later realizing this can add value to our products, this was adapted to create customized LLMs and SLMs for applications like Chat and automated test case generation using the Feature Specification documents. The quality of answers and the kind of output was steered using right set of prompts to achieve many applications and automation in the CI-CD pipelines. Techniques adopted for fine tuning include LoRA, QLoRA, Reinforcement Learning with Human Feedback(RLHF) and Reinforcement Learning with AI feedback. These techniques are used based on merit and the right set of models.

Fine-Tuning itself has been automated and exposed as service to interna teams

1. data-set generation by providing the documentation as input.
2. The Fine-Tuning pipeline does use synthetic-data-kit, which generates the meaningful question and answer dataset using local vLLM.
3. Data cleaning and then final data is improvement is achieved with human in the middle flows.
4. The standardized Hugging Face dataset format generated.
5. Pipeline can accept configuration for the Quantization, EPOCs and various other parameters of the methodology.
6. Pipeline executes the fine tuning and stores the newly fine tuned model to be served on Ollama and vLLM.
7. The tool reports the details of the fine tuning output with loss details and other interesting parameters as graphs.
8. The fine-tuning log is also analyzed by the LLM and nice report is prepared along with the fine tuning model, which helps to decide if the fine-tuning is successful or not.

Acumen AI Builder: Langflow based low-code AI builder for Agentic workflows.

LangFlow is an open source, low-code platform designed to simplify the process of building, testing and deploying applications powered by Large Language Models (LLMs). This provides a drag-and-drop visual interface that allows the users from developers to non-programmers to orchestrate complex AI workflows without writing extensive code. LangFlow is like n8n, RagFlow or Defy, which allows modern agentic systems to be built on SaaS based services as well as accessing local services. It provides handy components like vector databases, Open AI component, Anthropic, AWS, etc… which can be linked as nodes in the flow to create customized, automated agentic behavior.

The LangFlow open-source tool has been modified architecturally to scale to the Ribbon's Acumen environment. Lots of custom components were built which are specific to Ribbon and its customers which allow the data exposure to the LLM Agentic Flows. This tool utilizes the Acumen AI Platform to give host of applications and data access to build modern LLM based agentic work flows.

Ribbon Analytics:

Ribbon Analytics is a comprehensive big-data platform to provide deep, actional insights into Real-Time Communications network for service providers and enterprises. The tool works in tandem with network passive taps and proprietary sensors, hence a non-invasive network analytics solution.

This product has FraudProtect, RoboProtect and various other applications. For Wireless Operators it allows analyze VoLTE call quality along with network monitoring KPI sets. Features such as tracing calls and ladder diagram are built in. This helps operators to address the Quality Assurance and also measure Quality of Experience with identification of one-way call, re-established calls analyzing the SIP messages. With set of Machine Learning (ML) algorithms inbuilt, allows to correlate the errors in the network with right kind of network element in the topology.

Example: Most Probable Cause identification, An application which works with LogAnalyzer (another application) which does temporal correlation of the logs. This time stamped longs then matched with the eXtended Data Record (XDR) to identify problems in the network. This helps to find the needle in haystack.

The Ribbon Analytics was enhanced to support full 5G stack. This involves creating XDRs which are based on passive monitoring if the whole 5G network includes elements such as AMF,AUSF, SMF, gNodeB, UPF, PCF,UDM, CHF, NRF, NSSF and SCP. This allowed Ribbon to foray into the VoNR world and first of its kind deployment in Airtel India. This served peak of 2.5 Million customers in a cluster with supporting 3 billion events per hour. This gave the clear picture into the 5G network and VoNR deployments to monitor errors, performance and debugging the issues. The Ribbon analytics enhanced to become quasi-NWDAF in the 5G network.

Mobile INband Telemetry: Process and Analyze Terra Bits per second:

Mobile INband Telemetry in short MINT uses the Intel Tofino switch to program the packet processing using p4 programming language. This has been orchestrated with RESTful APIs which gets feedback and commands from the OAM systems based on alarms and the system signals. The capacity of the Tofino switch with the p4 allows processing of Terra Bits of data per second and hence creating huge meta data for processing and analyzing the network traffic. This rich set of data helped to analyze the traffic using ML algorithms to identify lots of patterns helping the operators to analyze the network at very lower costs.

Cisco Systems India Pvt Limited | Software Engineer 2015-2021

- ***WiFi Gateways***: Design and implementation of various features in S2a (SaMOG) product. A WiFi offload entity to support WiFi deployment for telecom operators.
- ***MPTCP***: Prototype of cutting edge MPTCP protocol to support high data throughput. Porting TCP stack to DPDK application.
- ***LTE to 5G Transition***: Implement 5G-NSA solution on MME and PGW to support initial 5G deployment with legacy 4G/LTE systems.
- ***5G SMF***: Part of the team successfully delivered Cisco's industry first Cloud Native Stand Alone 5G Core network. Played various roles from conceptualization to field deployment. Owned heart of the SMF solution QoS Modelling, which is responsible for coordinating whole 5G sessions. Along with QoS modeling handled the communication between SMF and UPF and SMF and CHF.

Samsung Research Institute | Chief Engineer 2014-2015

- Leading TAC team for OAM of Samsung Core network, which is deployed for Jio, world's first greenfield all IP EPC.
- Handled day-today OAM issues and helped the both signalling and data-plane team in analyzing issues and brining the operations smoother.
- Managed the KPIs sets for the whole network of more than 100,000 (1 lakh+) eNodeBs and helped to get the network running.

Nokia Solution & Networks | Product Architect 2008-2014

- Played key role in successfully developing and deploying the industry's first commercial 3GPP Femto Gateway, transitioning the solution from laboratory prototype to field implementation.
- Led comprehensive agile transformation initiative for a 120-member cross-functional team, implementing Scrum methodologies and enhancing development efficiency and product delivery timelines.
- Pioneered automated continuous integration and deployment infrastructure before industry standardization, (CI/CD) orchestrating an advanced testing framework capable of executing 5,000+ test cases within a three-hour window.
- Drove product evolution from initial proof-of-concept through full commercialization, employing startup methodologies to rapidly iterate and deliver market-ready solutions.
- Architected and implemented comprehensive FCAPS (Fault, Configuration, Accounting, Performance, Security) management system supporting 60,000 concurrent Femto Access Point connections with active/backup redundancy architecture to ensure five-nines (99.999%) service availability.

Motorola India Pvt Limited | Software Engineer **2004-2008**

- Responsible for Feature Database for Motorola P2K Platform.
- Design and implemented a highly automated system to scale and cater 6X number of customers.
- Desing and deliver Over the Air update for flex files.

Education

BITS, Pilani, *MS Software Systems* **(CGPA: 7.26)** **2011**

Bachelor of Engineering, *Information Science and Engineering,* **AIT Chickmagluru (76%)** **2004**

Skills & abilities

- **Programming:** Previously C/C++ and Golang, now python, javascript
- **Full Stack:** python-based framework for back end, full stack developer for complete project PoC and MVP
- **Networking:** Hands on with Linux/Networking. Docker and Kubernetes networking.
- **Data Handling:** Handled large scale data with Hadoop, Parquet, clickhouse, nosql like redis and sql mostly postgres ,
- **Could Technologies:** Design and implementation of Kubernetes-based micro-service based applications. gRPC, protobuf, Grafana, git, ReSTful APIs, devOPS, open tracing etc…
- **Hyper Scalers:** Worked with AWS bedrock and hands on with GCP and Azure
- **3GPP Telecom**: Extensive experience from 3G, 4G/LTE to 5G. Implementation of 5G interfaces which support various version of ReSTful calls.
- **AI/ML/Data Engineering:** Python, AI/ML, LLMs, SLM. various transformers, vector db, streamlit, hugging face, Ollama, langflow, langgraph, langchain, llama index, langfuse and AI Ops, ML Ops.Open AI, TTS(test to speech) and STT (speech to text) models.
- **Pre & Post Training of LLMs**: Creation of fine tuning data set, Fine Tuning with un-sloth. Building customized LLMs for customized tasks.
- **Monitoring tools**: blackbox, Prometheus , Grafana, Arize Phoenix, Langsmith, Langfuse,
- **AI/ML frameworks**: PyTorch, langflow, n8n, ragflow, defy, ollama, vLLM.

Patents

US011350319B2: OPTIMIZED QUALITY OF SERVICE ENFORCEMENT FOR HANDOVER BETWEEN DIFFERENT RADIO ACCESS TYPES

Initiatives

2025: Ideating and realizing AI/ML based applications chat bot, sentiment analysis, classification, dynamic knowledge article generation, this helped to get access to various knowledge base**(Ribbon)**.

2024: Ideating and realizing RFQ Responder, which accelerated the sales teams response to days from months. **(Ribbon)**.

2021: Next generation analytics product to cater 5G market using open source free5gc **(Ribbon)**.

2020: Call flow sequence diagram generation from distributed k8 pods to provide single flow view. This helped Dev, QA and customer teams to pinpoint issue within minutes during feature testing **(Cisco).**

2019: Trained in Cisco Design level thinking, a startup experience within Cisco CTO org. Our proposal selected semifinalist among 400+ submissions for driving innovation and startup culture within Cisco in IEC-5 **(Cisco)**.

2013: Part of the team handcrafted CI-CD system before Jenkins was normal in industry using perl and expect framework **(NSN).**

2012: Desing and prototype of Health Monitor Console for Femto Gateway, an observability platform before Prometheus and Grafana existed. This helped to monitor distributed performance setups at one place. **(NSN).**