

SAI PRAKASH

Senior Backend Engineer | Distributed Systems Architect | Cloud & ML Infrastructure Specialist

Kakkanad, Kochi | saiprakash50@gmail.com | linkedin.com/in/saisayoojyam

PROFESSIONAL SUMMARY

Senior Backend Engineer with 7 years of expertise architecting scalable distributed systems, cloud-native applications, and ML/AI infrastructure. Specialized in backend architectures that bridge hardware, software, and cloud platforms, with recent deep focus on production computer vision, RAG pipelines, and agentic AI systems. Track record of taking systems from proof-of-concept to optimized, horizontally scalable production deployments — including cutting end-to-end API latency by over 55% through architectural re-design.

CORE COMPETENCIES

Backend Architecture & Development

Systems: Distributed Systems Design | Microservices Architecture | RESTful API Design

Pipelines: Real-time Data Pipelines | Event-Driven Architecture | Message Queue Systems (Kafka, Celery)

Performance: High-Performance Computing | Low-Latency Systems | Scalability & Performance Optimization

Cloud Infrastructure & DevOps

AWS: Lambda, EKS, EC2, ALB, ASG, AMI, RDS, IoT Core, CDK, API Gateway, CloudFormation, S3

Azure: Functions, DevOps, Pipelines, Resource Management

Containers: Kubernetes, Docker, Docker Compose, Portainer, Serverless Framework

CI/CD: Automated Deployment Pipelines, Infrastructure as Code, GitOps

ML & AI Infrastructure

Computer Vision: 3D Reconstruction (COLMAP, OpenMVS), Segmentation, OCR, VLM-based Image Captioning

LLM/RAG: RAG Systems, Vector Databases, LangGraph Agentic Orchestration, Local LLM Deployment

ML Ops: ML Model Serving & Orchestration, GPU Infrastructure Management, AI Agent Development

Hardware-Software Integration

Edge & IoT: Edge Computing, IoT Backend Systems, Embedded Linux (Yocto), Home Automation (Home Assistant)

Real-Time: WebRTC & STUN/TURN, Real-Time Communication Protocols, PLC Data Integration

Languages & Frameworks: Python (Django, Flask, FastAPI), C++, SQL, React.js

Databases & Storage: PostgreSQL, MySQL, SQLite, Redis, Amazon RDS, Vector Databases, Time-Series Databases

Development Tools: Git, CMake, VS Code, Postman, Sentry, Azure DevOps, Tailscale

PROFESSIONAL EXPERIENCE

Senior Software Engineer | **Cubet** | Sep 2026 – Present

FindMe — Face-Match Event Video Search (PoC)

- Building a proof-of-concept that lets event attendees find the video clips they appear in, using face recognition
- **Curator side:** detects every face in an event video, generates embeddings for each person, and stores them in **pgvector**
- **Attendee side:** registration by QR code or access code with phone + OTP verification; attendees enroll 3 face photos, which become the embedding matched against the event video to surface their clips
- Event-driven architecture. Stack: Python backend with a dedicated face-recognition module, Celery queues, PostgreSQL + pgvector, React frontend — all on Docker Compose

Senior Product Engineer | **Myelin Foundry** | Aug 2024 – Sep 2026

2D-to-3D Point Cloud & Object Fitting System — Scalable CV Infrastructure (Latest Project)

- Architected an end-to-end 3D reconstruction pipeline converting 2D images into dense point clouds using **COLMAP** and **OpenMVS**, with downstream segmentation and volumetric measurement of segmented objects (e.g., luggage-to-car-boot fitting)
- Built the initial system as a distributed microservices architecture across **6 containers** plus a dedicated segmentation machine, using **PostgreSQL, Redis, and Celery** for task orchestration and state management
- Re-architected the entire workflow from a containerized, message-queue-based design to a **bare-metal deployment**, eliminating inter-container API overhead, Celery queuing latency, and Redis dependency to meet strict latency requirements
- Reduced end-to-end processing time per API request from **~60 seconds to under 25 seconds** (58%+ latency reduction) through this architectural simplification

- Designed a **custom AWS AMI** running the optimized workflow as a native system service, with auto-scaling via **ALB and ASG** for elastic, low-latency scale-out
- Implemented custom scaling logic accounting for compute-intensive workloads (each request consuming ~1/3 of a g7.4xlarge instance), enabling accurate right-sizing of concurrent capacity
- Built a centralized S3-based artifact store for pipeline outputs, enabling downstream retrieval and reuse without recomputation
- Achieved stable sub-25-second latency per request while running 3 concurrent instances in production, validating scalability under real load

Automated RAG Pipeline with Agentic Routing

- Built an automated RAG pipeline ingesting multi-source documents into a vector database, preserving source location and page-number metadata for traceable retrieval
- Implemented an image-handling pipeline for embedded PDF content, routing images to OCR or a VLM to generate descriptive captions, enriching metadata for more accurate downstream responses
- Designed an agentic orchestration layer using LangGraph, with agents that classify incoming queries as conversational, document retrieval, or vector search, routing each to the appropriate handling path
- Deployed and ran the entire pipeline locally on an RTX 4070, demonstrating efficient on-device LLM/VLM inference without cloud dependency

Custom Learning Management System (LMS) — Quiz & Learning Platform

- Designed and deployed a custom LMS featuring quiz-taking and structured learning modules, built end-to-end with a React frontend and FastAPI backend
- Used SQLite for lightweight data persistence suited to the application's scale and deployment simplicity
- Secured the platform with SSL and deployed it under the company domain, owning the full lifecycle from development to production hosting

VR 360° Real-Time Streaming System — Backend Infrastructure

- Architected and developed the complete backend communication infrastructure in C++ for a VR streaming platform, handling vehicle registration, session management, and real-time call routing
- Designed and implemented WebRTC signaling server with STUN/TURN integration, achieving sub-200ms latency for 360° video streaming over cellular networks
- Built custom CMake-based compilation pipeline for Qualcomm's Yocto Linux environment, optimizing embedded system performance for resource-constrained hardware
- Engineered video processing backend for 360° equirectangular streaming at 1080p (2:1 aspect ratio), implementing efficient encoding strategies for bandwidth-constrained scenarios
- Developed backend integrations for live monitoring dashboards and YouTube Live streaming APIs, enabling real-time telemetry and broadcast capabilities

Industrial Anomaly Detection Platform — Distributed Backend Architecture

- Designed and implemented a fault-tolerant, event-driven microservices architecture using Apache Kafka for real-time industrial IoT data processing
- Built scalable backend infrastructure processing thousands of PLC telemetry events per second from distributed manufacturing equipment
- Architected ML model serving layer with dedicated microservices for each anomaly detection model, implementing async communication patterns and state management
- Developed PostgreSQL-backed REST APIs and WebSocket servers for real-time dashboard updates and alert notifications
- Designed the system for horizontal scalability, enabling seamless addition of new detection models and data sources

RAG-Based Operator Assistance System — AI Infrastructure

- Architected end-to-end RAG system for AI-powered machinery troubleshooting, integrating voice recognition, document retrieval, and LLM inference
- Built document ingestion pipeline supporting multilingual technical manuals (Chinese and English), implementing OCR, translation APIs, and text chunking strategies
- Deployed local LLM infrastructure on GPU servers, optimizing inference performance and implementing request queuing for concurrent user queries
- Designed vector database integration for semantic document search, achieving high-precision retrieval for technical documentation

Software Engineer | **Corestrat Labs** | Sep 2023 – July 2024

Algorithmic Trading Platform — Cloud Backend Architecture

- Architected serverless trading bot backend using Azure Functions, designing event-driven architecture for real-time market data processing
- Developed core trading algorithm engine with backtesting capabilities, implementing complex order execution logic and risk management systems
- Built scalable MySQL database schema for high-frequency trading data, optimizing for write-heavy workloads and complex analytical queries
- Developed backend integration with React.js frontend, exposing WebSocket APIs for real-time trading updates
- Managed Azure infrastructure deployment using Infrastructure as Code, implementing CI/CD pipelines for automated testing and deployment

Senior Application Engineer | **Myelin Foundry** | May 2020 – Sep 2023

Senior Application Engineer (Apr 2023 – Sep 2023)

- Led backend architecture and AWS infrastructure design for multiple IoT and real-time streaming projects
- Architected distributed systems using AWS IoT Core, implementing MQTT-based communication protocols for thousands of connected devices
- Built real-time data streaming pipelines using Kinesis, processing and routing telemetry data from edge devices to cloud storage and analytics platforms
- Managed technical team of 4 engineers, conducting architecture reviews and mentoring on distributed systems design patterns

Application Engineer (May 2022 – Apr 2023)

- Managed comprehensive AWS infrastructure including RDS database clusters, Route 53 DNS management, and EC2 auto-scaling groups
- Led performance optimization initiatives, reducing API response times by 40% through database query optimization and caching strategies
- Built monitoring and alerting infrastructure using CloudWatch, implementing custom metrics and automated incident response

Product Engineer (May 2020 – Apr 2022)

- Developed edge computing backend systems integrating AWS IoT, Lambda, and local compute on ESP32 microcontrollers
- Built Docker-based containerized applications for edge deployment, managing orchestration and updates remotely
- Implemented Kubernetes clusters for container orchestration, managing multi-node deployments and service discovery

System Administrator | **Opt IT Technologies** | Dec 2019 – May 2020

- Managed cloud infrastructure across AWS and Azure platforms, implementing security best practices and access control policies
- Administered hybrid Windows/Linux environments, automating system maintenance and monitoring tasks

IT Service Engineer | **Zinage Systems** | Jun 2016 – Jul 2018

- Analyzed network infrastructure and performed troubleshooting for enterprise-scale network systems
- Managed Active Directory services and group policy configurations for large organizations

PERSONAL PROJECTS

Algorithmic Trading Bot (Nifty 50 Index)

- Built an algorithmic trading system analyzing stock and options market data to generate buy/sell suggestions for Nifty 50 index trades
- Implemented data capture from TradingView, with analysis and visualization pipelines using SimpleCharts
- Self-hosted the system on a desktop PC, using Portainer for container/compose management
- Enabled secure remote access for clients via Tailscale mesh VPN, allowing distributed access without exposing infrastructure publicly

FREELANCE & INDEPENDENT WORK

- **Freelance Web Designer & Developer** — designed and delivered custom websites for independent clients, handling both front-end design and back-end development
- **Freelance Photographer & Drone Photographer** — provided photography and aerial drone photography services alongside web development work

TECHNICAL HIGHLIGHTS

Backend Architecture Expertise

- Designed and deployed 10+ production-grade distributed systems handling millions of requests daily
- Expert in event-driven architectures, microservices patterns, and API gateway implementations

Cloud-Native Development

- Deep expertise in AWS and Azure serverless architectures, container orchestration, and managed services
- Built CI/CD pipelines reducing deployment time from hours to minutes

AI/ML Infrastructure

- Production experience across the AI stack: computer vision (3D reconstruction, segmentation), RAG systems with agentic routing, and local LLM/VLM deployment
- Implementing ML model serving infrastructure with GPU optimization

Leadership & Collaboration

- Led cross-functional teams in architecting and delivering complex backend systems
- Mentored junior engineers in distributed systems design, cloud architecture, and best practices

INTERESTS

- **Home Automation / IoT:** Designed and wired a full home automation setup integrating smart bulbs, sensors, and cameras, unified through Home Assistant
- Built a pet-monitoring system using OpenClaw with camera integration — used to track a cat's recovery post-surgery, sending real-time status updates via Telegram
- **Drone Building:** Hobbyist drone builder, with hands-on experience in hardware assembly and flight systems
- Distributed systems patterns, cloud-native architectures, real-time data processing, and hardware-software integration challenges