

AYUSH GUPTA

[email](#) [linkedin](#)

OVERVIEW

Staff Software Engineer and Founding Engineer specializing in the architecture of high-scale, AI-integrated real-time platforms serving over **500k+ monthly active users**. Proven track record of taking complex systems from 0-to-1, spearheading multi-tenant **Next.js 16** applications, low-latency **WebRTC** state machines, and high-throughput **Triton Inference** clusters (**1280+ RPS**) with edge-local **ONNX/WASM** intelligence. Expert in full-stack orchestration (**TypeScript**, **Python/FastAPI**, **Node.js**, **React Native**), event-driven backends, and multi-cloud GitOps infrastructure (**Azure AKS**, **ArgoCD**, **OpenTelemetry**). Committed to a **"Systems First"** philosophy: building self-healing, observable architectures with integrated AI safety, RAG pipelines, and Web3 (Solana) payment infrastructure.

SKILLS

Frontend	Next.js 16 (Parallel/Intercepting Routes), TypeScript, Redux Toolkit, React Native CLI / Expo (Android/iOS), WebRTC, Framer Motion
Backend	Python, FastAPI, Node.js, .NET (ASP.NET Core), Express, Golang, gRPC, CQRS, Event-Driven
AI Engineering	LLM APIs (Anthropic, OpenAI), RAG, Triton Inference Server, vLLM, ONNX Runtime, LangChain, LangGraph, AutoGen, Prompt Engineering, Model Quantization
Web3	Solana (web3.js), SPL Tokens, Helius Webhooks, Phantom/Wallet Integration
DevOps	Azure AKS, ArgoCD (GitOps), Azure Front Door (AFD), AWS, OpenTelemetry (OTel), Prometheus, Grafana
Data	PostgreSQL (query tuning), Redis (Lua scripting), Memgraph (Graph DB), Cache API (Service Worker), Persistence Layer
Real-Time	SignalR Hub Sync, Socket.io, P2P Data Channels, MediaStream Manipulation, TURN/STUN
Leadership	Founding Team Leadership, Technical Mentorship, Architectural Governance, Product Strategy

EXPERIENCE

Staff Software Engineer (Founding Engineer) Vooz Inc	Sep 2024 - Present <i>Remote</i>
---	-------------------------------------

- **Architectural Foundation:** Engineered a multi-tenant Next.js environment utilizing **Parallel and Intercepting Routes** to deliver a seamless "modal-less" navigation experience, scaling to **500k+ MAU** with optimized SSR/caching delivery.
- **Real-Time Video Coordination & Resilience:** Designed and maintained the **Vooz Video Coordinator**, a high-availability state machine managing WebRTC handshakes and **Media Recovery Logic** for hot-swapping hardware mid-call; established a **SignalR Hub Synchronization** protocol recovering user state from abnormal WebSocket closes (Code 1006) for 99.9% session continuity.
- **Group Audio Platform (0→1):** Shipped **Hangouts**, host-owned group audio rooms on a **P2P WebRTC mesh**, holding all live room state in **Redis** behind nine atomic **Lua** scripts; time-scored membership makes crashed clients and half-open sockets self-expire with no reconciliation job.
- **Service Decomposition:** Split a monolithic **.NET** API into independently scaled **REST** and **Realtime** services behind one ingress, decoupling WebSocket connection density from REST capacity so hub headroom no longer forces redundant REST pods.
- **Database Performance:** Resolved production **PostgreSQL** CPU saturation (**~97% → nominal**) by rewriting the six hottest queries by **pg_stat_statements** share: **OR→UNION** decomposition, covering indexes built **CONCURRENTLY**, and a username cache eliminating **443M** redundant lookups.
- **AI Inference at Scale:** Spearheaded the migration from standalone containers to a high-throughput **Triton Inference Server** cluster, implementing **Dynamic Batching** to achieve **1280+ RPS** with full **OpenTelemetry (OTel)** tracing.
- **Multi-Rail Payment Infrastructure:** Architected payments across crypto and fiat: **Solana** (Node.js microservice, **Helius Webhooks**, atomic confirmation) alongside **CCBill** and **Square** (Web Payments SDK, **3-D Secure** buyer verification, Apple Pay) behind one provider-switchable gateway.
- **LLM Integration & Orchestration:** Integrated LLM APIs (Anthropic, OpenAI) to power advanced conversational features, automated content moderation, and semantic search queries across the platform database.
- **Infrastructure & Edge logic:** Orchestrated multi-cloud deployments (AKS/DigitalOcean) secured by **ArgoCD (GitOps)** and **Azure Front Door (AFD)**, using cookie-based edge logic for optimized routing between marketing and core app layers.
- **Client-Side Intelligence & AI Safety:** Engineered a browser-local inference engine (**ONNX Runtime**, WebGPU/WASM in a Web Worker) scanning remote video feeds in real time, delivered via the **Cache API**; paired with backend NSFW services in a dual-track pipeline for a unified blur/report response.

Software Developer (Lead Role) Aimica Ltd	Jan 2023 - Sep 2024 <i>Remote</i>
--	--------------------------------------

- Led a team of 3 developers in building AI-powered mobile applications using React Native (Expo), driving architectural decisions and code quality standards.
- Designed and implemented scalable RAG & Multimodal AI systems integrating Claude, OpenAI GPT, LLaMA, image generation, and open-source AI models.
- Developed backend services (Node.js/Python) to support AI orchestration, caching, and semantic retrieval workflows.
- Built custom native modules (Swift/Java) to bridge AI features with platform-specific capabilities.

Mobile Developer

Dec 2021 - Dec 2022

Infoedge Software Solutions

Noida, India

- Optimized critical React Native flows for high-traffic applications, resolving memory leaks and improving UI frame rates on low-end devices.
- Collaborated with backend teams to define strictly typed REST API contracts, ensuring data consistency across mobile and server layers.

EDUCATION

B.Tech in Computer Science Engineering, Dr. A.P.J. Abdul Kalam Technical University

2018 - 2022

TECHNICAL PROJECTS

Triton & vLLM Distributed AI Inference Cluster. High-throughput GPU inference backend powering 1280+ RPS across production LLM and vision workloads. Engineered with Dynamic Batching, Prometheus metrics, and OpenTelemetry distributed tracing. (Triton, vLLM, OTel, Python).

Zero-Latency Edge ONNX Safety Engine. Browser-native inference engine evaluating real-time video feeds on-device using WebGPU execution providers and Cache API, eliminating cloud inference overhead. (ONNX, WebGPU, WASM, JavaScript).

Medical Paper NLP Categoriser. Developed and trained a deep learning model using **TensorFlow** on PubMed RCT datasets. Achieved 90%+ accuracy in classifying medical abstracts using NLP techniques. (Python, TensorFlow, Pandas).

Secure Video Conferencing Suite. Architected a cross-platform video app featuring end-to-end encryption and real-time signaling via Firestore. (Flutter, WebRTC).