Aditya Agashe

✉ aditya@agashe.me | ☎ 540.998.3989 | in linkedin.com/in/agasheaditya

Executive Summary

Staff Software Engineer focused on architecting distributed systems and enterprise AI infrastructure at multi-cloud scale. Led development of a centralized AI Gateway serving 100+ teams and processing 1B+ tokens/day, driving resiliency, governance, and cost optimization.

Professional Experience

Staff Software Engineer

NBCUniversal *Sept 2024 - Present*

AI Gateway – Multi-Cloud LLM Access & Governance

- Led the architecture and development of a centralized AI Gateway enabling 100+ teams to access LLMs across AWS, Azure, and GCP, scaling to 1B+ tokens/day.
- Designed platform-level resiliency and traffic controls (rate limiting, circuit breakers, multi-cloud failover) and implemented AI safety guardrails including prompt injection detection and content filtering.
- Established end-to-end observability and governance (cost, token usage, time-to-first-token), driving performance optimization and enterprise-wide AI spend control.
- Supported cross-team adoption of the AI Gateway through design reviews, architecture guidance, and tool evaluation; mentor engineers and participate in technical interviews.

AI Support Agent for On-Call Operations

- Architected and built a Slack-based AI agent serving as first-line support for engineering teams
- Leveraged historic support conversations and internal documentation to autonomously resolve repetitive operational queries
- Integrated with MCP servers to retrieve team-specific entities and contextual data for accurate responses
- Reduced on-call interruption load by handling routine questions, enabling engineers to focus on critical incidents

MCP Server Framework

- Architected a secure, OAuth2-enabled MCP server with RBAC to support multi-team access and governance
- Standardized deployment on AWS and Azure container platforms with built-in observability and health monitoring
- Established reusable server templates and best practices, enabling teams to rapidly build and deploy new MCP services

Staff Software Engineer

Extend, Inc *Jan 2022 - Sept 2024*

- Architected and delivered a RAG-based microservice using LangChain, Pinecone, and OpenAI Embeddings to power Extend's chatbot, enabling contextual, real-time policy and product support
- Led development of scalable serverless microservices on AWS serving millions of users daily, improving performance and system resilience
- Designed a high-throughput REST API handling 300+ requests/second, increasing reliability and reducing latency across critical transaction flows
- Built event-driven Node.js microservices integrating E-commerce platforms with Extend via Webhooks, Lambda, API Gateway, Kafka, and DynamoDB
- Delivered low-code integration tooling that accelerated merchant onboarding and expanded platform reach
- Enhanced front-end SDK performance and UX, increasing user engagement by 20%

Senior Software Engineer

Dassault Systemes *July 2015 - Dec 2021*

- Designed and developed web applications for Simulation Data Management in an Agile team, focusing on complex data visualization and interpretation with Java, JavaScript, and C++.
- Built responsive dashboard applications using ReactJS, Redux, and XState, significantly enhancing user experience and application performance.
- Developed back-end APIs and Java algorithms for the DS 3DExperience platform, improving data management, processing, and analytics capabilities.
- Prototyped a Web-Socket communication protocol for seamless integration between C++, Java, and JavaScript, boosting system interoperability and performance.

Technical Expertise

Architecture & Distributed Systems: Multi-cloud platform architecture (AWS, Azure, GCP) • High-throughput APIs • Event-driven microservices • Resiliency patterns • Reliability & performance optimization
AI & LLM Systems: Enterprise LLM platforms • RAG systems • Agent architectures • Vector search • AI safety & guardrails • Token/cost governance
Cloud & Infrastructure: AWS (Lambda, API Gateway, Step Functions, Bedrock, DynamoDB) • Azure (OpenAI, Logic Apps, Key Vault) • Containerized deployments • OAuth2/RBAC • Observability
Data & Messaging: Kafka • DynamoDB • PostgreSQL
Programming: Python • TypeScript/Node.js • Java • C++

Education

Master of Science in Computer Science Virginia Tech GPA: 3.61/4.00
Bachelor of Engineering in Computer Engineering University of Pune *First Class*