

Prafulla Anurag

Lead Software Engineer

Lead platform engineer with 7 years building distributed systems at scale — multi-region cell architectures, Kubernetes control planes, and AI inference infrastructure. Currently leading AI Infrastructure at Razorpay; previously built Microsoft Search backend.

Bengaluru, India • anuragprafulla@gmail.com • linkedin.com/in/anuragprafulla

EXPERIENCE

Razorpay Software Limited, Bengaluru

February 2022 – Present

Lead Software Engineer Nov 2024 – Present • **Senior Software Engineer** Feb 2022 – Oct 2024

AI Infrastructure

- Built and operate Razorpay's LLM gateway on top of LiteLLM (self-hosted) — a multi-tenant proxy fronting commercial LLM APIs and Fireworks-hosted OSS models with per-user/team/app API keys, provider failover, streaming responses, and per-tenant cost attribution. Serves ~100 teams at 400K+ requests/day (50 RPS peak); drove a 20% reduction in LLM spend.

Developer Platform

- Led a team of 4 engineers building the control plane for Razorpay's pre-production infrastructure — a multi-tenant Kubernetes platform serving 1,000+ engineering consumers across 100+ services.
- Designed and implemented a Kubernetes-native control plane (custom operators + CRDs) provisioning 5,000+ ephemeral environments per week on AWS EKS. Cut provisioning time from 10min → 2min by replacing per-environment Helm installs with a CRD that forks a warm baseline deployment, eliminating bootstrap overhead.
- Built a workflow orchestration engine on Argo Workflows (Go) that executes 2,000+ end-to-end test pipelines per week across the service fleet.

Cell-Based Architecture

- Designed the IaC abstraction for Razorpay's cell-based architecture using Terragrunt stacks — composable, region-isolated deployment units where each cell is a self-contained instance of the full service fleet for one country. Bounded the failure domain of a regional outage to that country's traffic alone.
- Bootstrapped the first cell end-to-end: brought up the initial service set on the new abstraction and established the patterns adopted across all subsequent cells. Foundation now runs 10+ cells hosting 100+ services, including Razorpay's Singapore and US deployments.

Microsoft India Development Center, Bengaluru

June 2019 – February 2022

Software Engineer 2 Oct 2021 – Feb 2022 • **Software Engineer** Jun 2019 – Sep 2021

- Built the extensibility framework for Microsoft Search's query layer (C# / .NET) — a multi-tenant system serving all M365 enterprise customers — replacing per-source bespoke connectors with a declarative configuration model that enabled new data source integrations without engineering involvement.
- Designed and built the Answers Platform: a low-latency federation layer returning "intent of answer" responses within an 850ms p90 SLA, followed by an async callback for the resolved answer. Powered instant search experiences across third-party content in Microsoft Search.
- Owned the Substrate Content Domain project in the Query layer — managed query-time configuration for heterogeneous content types across the M365 Substrate (Exchange, SharePoint, OneDrive, Teams).

SKILLS

Languages: Go, Python, C#, C++

Cloud: AWS (EKS, VPC, IAM, S3), Terraform, Terragrunt, Docker

Data & Messaging: Kafka, MySQL, Postgres, Redis

Kubernetes: Custom operators, CRDs, Helm, Argo Workflows

AI Infrastructure: LiteLLM, AWS Bedrock, Fireworks

CI/CD: GitHub Actions, Spinnaker

EDUCATION

Indian Institute of Technology, Roorkee — B.Tech in Electrical Engineering | CGPA: 8.34/10

July 2015 – May 2019

TALKS & WRITING

- Declarative GitOps Workflows for E2E Testing — PlatformCon 2022

AWARDS & RECOGNITION

- Microsoft E+D India, Be a Leader Award 2020
- KVPY Fellow with AIR 108