

Abishek K C

Birmingham, AL (open to relocation) | abhishekkc701@gmail.com | 334-372-7920

abishekkc.com | linkedin.com/in/abishekkc701 | github.com/abishekkc701

Technical Skills

Languages & Fundamentals: Python, Go, Bash/Shell, SQL, C++, REST APIs, gRPC, Protocol Buffers, Asynchronous I/O (asyncio), Thread Synchronization (threading.Lock)

Cloud & Infrastructure: Google Cloud Platform (GCP), Kubernetes (GKE), Docker, Redis Streams, Google Cloud Pub/Sub, Cloud TPU, NVIDIA GPUs (L4, T4), KEDA, Terraform (IaC)

Observability & Reliability: OpenTelemetry (OTel), Datadog, Prometheus, Grafana, Jaeger, LangSmith, SLIs/SLOs, Error Budget Governance, Distributed Tracing, Alert Noise Suppression

Systems, Delivery & AI Tooling: Helm, ArgoCD (GitOps), CI/CD Automation, Claude Code Skills, Machine-Readable JSON Runbooks, Schema Validation (Pydantic), Chaos Engineering

Experience

Site Reliability Engineer – Motion Industries

Dec 2023 – Present

Contract via Aurum Data Solutions | Birmingham, AL (Remote / Hybrid)

- **Architected a fault-tolerant distributed inference pipeline in Python and Redis Streams**, implementing dedicated GPU worker sub-processes, thread-safe synchronization (threading.Lock) to eliminate race conditions, deterministic in-order sequence reassembly, and automated XCLAIM dead-letter queue consumers bounded at 100k entries to prevent out-of-memory (OOM) faults across 500-item batches.
- **Scaled GKE agent workloads to absorb a 10x enterprise rollout (20 to 200 branches) and drained persistent 500+ Pub/Sub message backlogs to near-zero** by engineering asynchronous worker concurrency (asyncio) to parallelize I/O-bound LangChain chains, PostgreSQL connection-pooled queries with KEDA autoscaling.
- **Established the company's first AI governance framework and automated test harnesses on GKE to manage multi-GB fine-tuned LLMs**; engineered CI/CD verification services that benchmarked model accuracy against evaluation suites, promoting compliant weights while auto-purging failed container images, cloud artifacts, and ephemeral GPU pods to prevent cluster storage and VRAM exhaustion.
- **Prevented capacity-induced outages and protected core serving SLOs through global accelerator shortages** by partnering with Google TAMs to engineer multi-architecture scheduling across preemptible NVIDIA L4/T4 GPU spot instances and migrating production Vertex AI endpoints to Cloud TPU-backed serving.
- **Accelerated production troubleshooting and Root Cause Analysis (RCA), cutting MTTR by ~65% (~100 mins)** across GCP/GKE microservices by deploying distributed tracing (OpenTelemetry, Datadog, LangSmith) and designing AI-native, machine-readable operational artifacts (schema-validated JSON runbooks) and Claude Code skills to drive automated, AI-assisted diagnostics during live incidents.
- **Delivered a core PI Planning reliability milestone by codifying Terraform-managed SLOs (7-day and 30-day windows) and deploying an AI alert-summarization pipeline**; cut alert fatigue by ~40% across standardized Microsoft Teams channels by replacing static threshold alerts with multi-window error-budget burn rates that extracted and summarized root-cause failure keywords.
- **Hardened downstream ERP and database integrations against cascading failures** by implementing backend resilience patterns in Python—including connection pooling, circuit breakers, and exponential backoff with jitter—eliminating distributed timeout propagation across synchronous microservice dependencies.
- **Engineered a production Python deal-scoring microservice that ranks enterprise accounts by profitability and renewal risk**, developing modular service interfaces, Microsoft Fabric data pipelines, Entra ID OAuth/RBAC security, and strict schema validation with transactional, idempotent writeback guards to guarantee financial data integrity.
- **Eliminated operational toil and elevated code quality by engineering custom CLI developer tooling and Claude Code skills** that ingested repository schemas to auto-generate boundary-case unit and integration test suites, refactor legacy backend components, and enforce architectural linting standards.

Projects

ML Infrastructure & Reliability System

Python, FastAPI, Kubernetes

- **Architected an autoscaling ML inference service in Python and FastAPI on GKE**, sustaining sub-120ms p95 latency under simulated 5x traffic surges with end-to-end telemetry (Prometheus, Grafana, Jaeger).

Education

Troy University – B.S. in Computer Science

May 2026

Relevant Coursework: Distributed Systems, Operating Systems, Computer Networks, Data Structures & Algorithms

Certifications

AWS Certified Solutions Architect – Associate (2024)

AWS Certified Cloud Practitioner (2024)