Gao Zhiyuan

+86 13262224615, +852 52140290
alapha23@gmail.com

Summary

Infrastructure & Cloud Support Engineer specialising in AWS availability, incident response, and performance tuning. 4+ years troubleshooting complex production issues across EC2, VPC, S3, EKS and serverless stacks for high-traffic enterprises. Adept at reproducing hard-to-trace errors, performing root-cause analysis, and automating fixes with Terraform/CloudFormation & Python/Bash. Proven record of cutting latency 60% → 10% and AWS spend $25M → $5M while ensuring secure, compliant operations. Thrive in fast-moving, multicultural teams and communicate technical findings clearly to engineers and business stakeholders.

Core Competencies

AWS Infrastructure: EC2, S3, VPC, IAM, CloudWatch, CloudFront, Lambda, RDS, EKS/ECS
Troubleshooting&Ops: Incident triage, MTTR reduction, log-driven RCA, cost optimisation
IaC&Automation: Terraform, AWS CDK/CloudFormation, Ansible, CI/CD (GitHub Actions)
Scripting&Languages: Python, Bash, Go (basic) – build diagnostic tooling & remediation scripts
Containers&OS: Docker, Kubernetes, Linux (Debian/Alpine/CentOS), Server administration
Observability: Prometheus/Grafana, CloudWatch Logs & Metrics, X-Ray, ELK stack
Soft Skills: Customer-facing support, cross-team collaboration, documentation

Work Experiences

Amazon Web Services
Associate Solutions Architect *Jun. 2024 - now*

1. Resolved cross-region latency spikes for a global e-commerce client by analysing CloudFront TCP reset patterns, designing& deploying an NGINX keep-alive solution; cut inconsistent RTT from 60% to 10% with zero downtime.

2. Automated EKS unmanaged→managed node-group migration with a Python user-data checker, slashing manual effort 80% and avoiding extended-support fees.

3. Improve Yolov8 Inference Throughput by 185%
- Using Pytorch to convert the model weight of Yolov8 for AWS Inf1 Instance.
- Gained 1.85x throughput increase whilst lowering the cost down to 25%.

Trilogy
Software Engineer *Aug. 2022 - Aug. 2023*

1. Led company-wide AWS cost-optimisation initiative: built automated right-sizing scripts (Python+AWS SDK), introduced multi-tenancy, resource-ownership tagging and anomaly detection → annual spend $25 M → $5M.

2. Deployed Terraform modules & Jenkins pipelines to standardise VPC & IAM baselines across 15 accounts, shrinking mean provisioning time from 3 days to 4 hours.

MISE
Devops Engineer (Contract) *Jan. 2022 - Sep. 2022*

Investigated and fixed Lambda cold-start latency spikes during traffic surges; tuned memory/timeout and introduced provisioned concurrency → p95 response time -40%

Hardened API Gateway + WAF configuration, blocking OWASP top-10 vectors; documented remediation steps for support runbooks.

MetaMUI-SovereignWallet
Python SDK Engineer *Jun. 2021 - Jan. 2022*
Ported Rust blockchain SDK to Python, writing integration tests to replicate consensus edge cases; accelerated developer onboarding 3×.

Projects

Seoul National University 2019
Project Name: Accomodate Input Spikes with AWS Lambda Functions
A hyrbid system of Dockers and AWS Lambda Functions, where stable workloads are running on docker and input rate spikes are offloaded to AWS Lambda Functions,
Implemented job scheduling/balancing algorithm to coordinate Lambda Functions' life cycle, data relay, and system-level dynamic scaling

Google Summer of Code
Apache Foundation *2019*
Engineered Apache Nemo to process single-stage batch data with AWS Lambda Functions

Education

National Cheng Kung University Taiwan
Engineering Science *Sep. 2015 - May. 2017*
Seoul National University Korea
B.S., Computer Science and Engineering *Sep. 2019 - Aug. 2024*

Tech Stack Snapshot AWS • Terraform • CloudFormation / CDK • Python&Bash • Linux &Windows Admin • Docker/Kubernetes • GitHub Actions • Prometheus/Grafana • Ansible • Jira / Agile Scrum

Languages: Chinese, English (TOEFL 100), Japanese (JLPT N1) and Korean