Gao Zhiyuan

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

Summary

Cloud and AI engineer with extensive experience in designing and operating large-scale cloud infrastructures (AWS, Kubernetes) and developing AI solutions (RAG and multi-agent). Proven track record in optimizing systems for performance and cost, including AI workload acceleration (1.85x throughput) and enterprise-scale cost reduction ($25M → $5M). Skilled in Python, IaC, and AWS systems with a focus on automation and reliability. Fluent in English, Japanese (JLPT N1), Chinese, Korean, and experienced in cross-functional, multicultural teams.

Work Experiences

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

1. Global Network Infrastructure Optimization Project
- Poizon is a global E-Commerce platform which faces large traffic per day and latencies is a critical issue.
- Monitored and analyzed network latency between the USA and Singapore, identifying bottlenecks within internal CloudFront components causing TCP reconnections.
- Developed and implemented a PoP warm-up strategy and an NGINX-based solution to ensure persistent TCP connections.
- Reduced cross-region latency inconsistency from 60% to under 10%.

2. RAG-based Chatbot for AWS Knowledge
- A chatbot using Retrieval Augmented Generation, to answer technical questions specifically focused on AWS services
- Involving Multi-agent orchestration and MCP, rewrote agent loop with langchain and Strands SDK.
- Evaluate the RAG in terms of hallucination, self-knowledge dependence, Context Utilization and etc.

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%.

4. EKS Node Group Migration
- Customer faces an urgent smooth EKS version upgrade demand to avoid extended support costs.
- To migration from unmanaged node group to managed node group in production, I tackled challenges in subnet public-ip auto-assigning, bridged internet connections with VPC endpoints
- Created a user data script and troubleshooted cloud-init logs to make sure the new unmanaged node groups can acquire EIPs, then pod evictions to complete the migration.

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

1. AWS Cost Optimization project
- Lowered AWS Service Costsfrom 25 million USD annually down to 5 million USD in a team of 3.
- Designed non-intrusive, sustainable plans to guarantee resource sharing, user resource awareness, automated anomaly detection and permission rules.

2. Social Platform for Education
A social platform that enables 4-12 grade students to memorize curricular knowledge through making posts with the help of AI and memorization techniques.
The tech stack includes Next.js, Prisma/MySQL, GPT-3, Stable-Diffusion, Docker and AWS Devops.

3. Data Platform based on Azure AD
Construct a graph-based data platform and use it to calculate company employee interaction intensity. Extract data from MS 365 through Azure Graph API, engineer the data into graph relationships and maintain the in the graph database.

MISE
Devops Engineer *Jan. 2022 - Sep. 2022*
Addressed latency spikes in an AWS Lambda-based startup application during peak traffic hours.
Diagnosed and resolved performance bottlenecks by analyzing AWS Lambda, API Gateway, and MySQL RDS.
Streamlined deployment processes using Terraform Infrastructure as Code (IaC).

MetaMUI-SovereignWallet
Python SDK Engineer *Jun. 2021 - Jan. 2022*
Migrate Rust SDK to Python SDK for MetaMUI blockchain built with Substrate framework

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*

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

Skills: Languages: Python, Java, Typescript, C/C++, NextJS
DevOps: AWS, Redis, K8S, Docker
Others: AI Agent / RAG development, Serverless development, QEMU, React, Selenium