

Gao Zhiyuan

+86 13262224615
alapha23@gmail.com

Work Experiences

Amazon Web Services

Associate Solutions Architect

Jun. 2024 - now

1. Global Network Infrastructure Optimization Project

- Monitored and analyzed network latency between the USA and Singapore, identifying bottlenecks within individual 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
- Evaluate the responses in terms of hallucination, self-knowledge dependence, Context Utilization and etc.
- The tech stack includes AWS Lambda, RDS, Opensearch, Bedrock/Claude, Sagemaker, Rerank.

3. Multimodal Content Guardrailing

- Employed a model unlearning project to forget certain concepts during inference time
- Moderate Content through AWS Nova, Bedrock Guardrails, and the model unlearning project

4. Client Security Incident Response

- Managed customer security incidents through advanced attack behavior analysis, device fingerprint profiling, and cloud attack path mapping.
- Enhanced client cloud security by designing and implementing robust security architectures and strategies.

Trilogy

Software Engineer

Aug. 2022 - Aug. 2023

1. AWS Cost Optimization project

- Lowered AWS Service Costs from 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.

4. Automate Company ERP operations on NetSuite

In charge of project ideation, development and delivery and was solely responsible for communication with the clients. Tech stack involves Selenium, Chrome Plugin, Python and Javascript

MISE

Devops Engineer

Jun. 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).

Projects

Seoul National University

Research Intern

City Energy Lab

Sep. 2023 - Now.

RAG-based Chatbot for Urban Planning

- Built a RAG chatbot with FAISS, parent-child chunking, keyword-extraction for domain-specific knowledge

- Designed a chatbot with NextJS, Typescript, Prisma, FAISS and OpenAI API

Seoul National University

2019

Project Name: Accomodate Input Spikes with AWS Lambda Functions

A hybrid 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

Chinese Academy of Science

PLCT Lab

QEMU Researcher

2020

One patch to QEMU upstream to emulate Nuclei RISC-V SoCs with customised interrupt controllers and registers,

Delivered an oral presentation at CRVA 2020

Education

National Cheng Kung University

Taiwan

B.S., Political Science

Sep. 2015 - May. 2017

Seoul National University

Korea

B.S., Computer Science and Engineering

Sep. 2019 - Aug. 2024

Skills: Languages: Python, Java, Typescript, C/C++

Backend Development: NextJS, Typescript, Prisma, MySQL

DevOps: AWS, Redis, K8S, Docker

Others: AI Agent / RAG development, Serverless development, QEMU, React, Selenium