

OPEN TO WORK · REMOTE

Felix Tran

Platform & AI Infrastructure Engineer

Platform engineering, SRE and AI infrastructure. I build the layer between the model and the people using it — privately, on infrastructure you control, at a cost per answer you can defend.

EU citizen — no work permit required in the EU **Remote-first** from Staffanstorp, Sweden (CET)

Open to **remote and hybrid roles**

8+ years cloud and platform engineering	**5 years** production Kubernetes	**EU citizen** no work permit required in the EU	**Remote-first** CET — overlaps EU and UK hours

Profile

I have spent eight years building cloud platforms where failure is a compliance event, not a nuisance. The Swedish approach applies: architecture should be quiet, reliable and boring. GCP and Azure as the ground, Kubernetes as the substrate, Terraform and ArgoCD as the delivery model.

Most AI projects do not fail for lack of GPUs. They fail in the layer between the model and the people using it — the inference stack, the context budget, the agent loop, and the evaluation that never happened. That is the layer I build: private inference, quantisation and KV-cache tuning, agent harnesses with permissions, audit trails and per-run token budgets.

Roles I am a good fit for

- Platform Engineer
- Site Reliability Engineer
- DevOps Engineer
- Kubernetes / Cloud Platform Engineer
- MLOps / AI Infrastructure Engineer
- Cloud Architect (GCP / Azure)

Working remotely

Distributed work is mostly a documentation problem. The infrastructure is the easy part; what decides whether a remote engagement lands is written context — decisions in git, runbooks next to the cluster, evals in CI, and handovers that survive a timezone gap rather than a Friday call. That is how I work alongside client teams, and it is what I leave behind when the engagement ends.

The constraint that travels with the work is data. Regulated teams — fintech, health, public sector — cannot send their documents to a hosted API, and working remotely does not change that. So the stack I bring is portable: open-weight models on GPUs you control, retrieval over your own sources, and an audit trail a compliance review will accept, whether the team sits in one building or six.

At a glance

BASED
Staffanstorp, Sweden

RIGHT TO WORK
EU citizen — no permit required in the EU

TIME ZONE
CET — overlaps EU and UK hours

OPEN TO
Remote and hybrid roles, EU-wide

felix@cloudape.ai

LinkedIn profile

Download CV (PDF)

Skills

CLOUD AND INFRASTRUCTURE AS CODE
Google Cloud Platform · Microsoft Azure · Terraform · Cloud Build · Azure DevOps

KUBERNETES AND DELIVERY
AKS · GKE · Multi-tenant clusters · ArgoCD · GitOps · GitHub Actions · GitLab CI · Jenkins · Docker

AI INFERENCE
vLLM · Ollama · llama.cpp · MTPLX · GGUF / AWQ / GPTQ quantisation · Batching and KV-cache sizing · Speculative decoding · Model routing

AGENTS AND RETRIEVAL
DeepSeek Harness · Open WebUI · MCP and tool calling · Qdrant · RAG

OBSERVABILITY AND OPERATIONS
LGTM (Loki, Grafana, Tempo, Mimir) · Splunk · SLOs · PostgreSQL · Nginx

LANGUAGES AND AUTOMATION
Python · Go · Bash · GitHub Enterprise API automation

Languages

ENGLISH
Full professional — working language

SWEDISH
Native

Experience

Founder & CEO

Cloud Ape AB

May 2024 – present · Staffanstorp, Sweden · home office

- Runs open-weight models privately on Apple Silicon, single-GPU workstations, rented cloud GPUs and Kubernetes — Ollama, llama.cpp, vLLM and MTPLX.
- Tunes inference for cost and latency through quantisation (GGUF/AWQ/GPTQ), batch sizing, KV-cache budgeting, speculative decoding and model routing.
- Builds local agents on DeepSeek Harness, Open WebUI, tool calling and MCP servers, with permissions, audit logging and per-run token budgets.
- Delivers engagements as code — model configs, eval suites, runbooks and dashboards — handed to the client's own team.

Senior AI Platform Engineer VIA CLOUD APE

Riksantikvarieämbetet (Swedish National Heritage Board)

Aug 2026 – present · Sweden · public sector

- Builds RAG pipelines on Qdrant and provisions secure, local autonomous coding agents on open-weight models.
- Designs declarative, multi-tenant Kubernetes environments delivered through ArgoCD GitOps.
- Runs the LGTM stack for platform health and GPU utilisation telemetry.

Platform Engineer VIA CLOUD APE

Inter IKEA Group

May 2024 – Apr 2026 · Malmö, Sweden

- Managed GitHub Enterprise infrastructure integrated with Azure, including API automation.
- Built CI/CD pipelines with GitHub Actions and Docker containerisation.
- Implemented observability with Splunk and ran JFrog Artifactory for artefact management.
- Automated security, access control and repository management at scale.

Industrial Automation and DevOps VIA CLOUD APE

Pågen

Sep 2025 – Dec 2025 · Malmö, Sweden

- Designed a proof of concept for remote management and configuration of Raspberry Pi devices with Ansible AWX.
- Integrated Azure DevOps Pipelines with Sonatype Nexus for automated Docker image builds and versioning.
- Delivered a secure on-prem solution for sensitive production data in the bakery industry.

Site Reliability Engineer

Accessy AB

Jan 2023 – May 2024 · Malmö, Sweden

- Designed and operated Azure Kubernetes infrastructure as Terraform IaC.
- Established GitLab CI/CD pipelines and Docker-based deployment automation.
- Owned monitoring, PostgreSQL operations and Nginx ingress, plus security hardening and high-availability operation.

Cloud Developer

Qlik

Jul 2021 – Jan 2023 · Lund, Sweden

- Built cloud-native microservices on Kubernetes in Go with MongoDB.
- Implemented API integrations and event-driven architectures, with Grafana monitoring and Splunk observability.

Education

Blekinge Tekniska Högskola

BSc, Computer Science

2018 – 2020

EC Utbildning

Higher Vocational Education Diploma, IT Support

2018 – 2019

Certifications

- Google Professional Cloud Developer
- Google IT Automation with Python Professional Certificate
- GitHub Foundations
- Jenkins Engineer
- Microsoft Certified: Azure Fundamentals

More Google Cloud, Microsoft, ITIL and Scrum credentials — badges at cloudape.ai/certifications.

Cloud Consultant

Forefront Consulting

Jan 2021 – Jun 2021 · Malmö, Sweden

- Architected GCP solutions with Terraform and Cloud Build CI/CD.
- Integrated Google Cloud AutoML and AI Platform for enterprise clients.

Cloud Consultant

Softhouse Consulting

Feb 2020 – Dec 2020 · Malmö, Sweden

- Built Jenkins and GitLab CI delivery pipelines, and Terraform-provisioned cloud environments.
- Managed artefacts in JFrog Artifactory and automated operations in Bash and Python.

Cloud Consultant SOGETI

Sep 2019 – Feb 2020 · Malmö, Sweden

IT Support Specialist TIETO

Jan 2018 – Mar 2019 · Malmö, Sweden

Easy Asia Trading AB, Malmö — CEO (2012–2017), Operations Manager (2009–2012), System Operator (2005–2008). Twelve years in operations and systems before moving into cloud engineering.