Mohammed Zafeeruddin

SDE II · Platform / DevOps / MLOps Engineer

Hyderabad, India · +91 95154 96400
mohammed.xafeer@gmail.com · zafeer.dev
github.com/Zafeeruddin · LinkedIn · x.com/itsZafeer

50+	**2,000+**	**55h → 5h**	**5d → 2h**	**SAR 300K**
k8s nodes operated	cameras on the platform	model training via PyTorch DDP	full platform deploy	yearly savings (Expro)

SUMMARY

Software engineer with 2+ years at an AI startup. I joined as an intern, converted to full time early and was promoted to SDE II in Apr 2026. I write the Python control-plane services behind **Baseer Builder**, our no-code computer-vision platform, and own its MLOps, DevOps and infrastructure. That covers Kubernetes across 50+ GPU nodes, CI/CD, secrets, networking, storage and observability, for deployments ranging from a 2,000-camera Saudi municipality to air-gapped government clusters. I also port speech models onto Qualcomm Cloud AI 100 accelerators (PyTorch → ONNX → compiled QPCs) and serve them behind OpenAI-compatible APIs.

TECHNICAL SKILLS

Languages	Python, Bash, TypeScript, SQL
Orchestration	Kubernetes (kubeadm, HA control plane, Jobs, StatefulSets, DaemonSets), Docker, containerd, Helm, Kubeflow Trainer
IaC & Delivery	Terraform, Ansible, Jenkins, Argo CD, GitHub Actions, Azure DevOps, GitLab CI, Harbor, HashiCorp Vault
Networking	NGINX, HAProxy, Keepalived (VIP), MetalLB, Ingress, VLANs, VPC design, WireGuard, Linux network namespaces
ML & Data	PyTorch DDP, MLflow, Kafka, MQTT, ClickHouse, PostgreSQL, Redis, InfluxDB, Qdrant, NFS/NAS
AI Accelerators	Qualcomm Cloud AI 100 (qaic-compile, QEfficient), ONNX export, vLLM-based inference image, LiteLLM
Cloud & Ops	AWS (EKS, EC2, VPC, ALB, API Gateway, RDS, S3, SQS, IAM), Alibaba (ECS, ACK, ACR), Cloudflare, Prometheus, Grafana

EXPERIENCE

Masterworks · Nabeh

May 2024 – Present · Hyderabad

SDE II (Apr 2026 – now) · **Software Engineer** (Mar 2025 – Mar 2026) · **Engineering Intern** (May 2024 – Feb 2025)

Baseer Builder: no-code computer-vision platform

product · MLOps & platform owner

The full CV lifecycle in one product: data, labelling, training, notebooks, real-time inference, camera and device fleets, and analytics.

- **Architected the microservice platform** and own the core logic of the training, notebook, face registration, camera and device services, all written in Python.
- **Model Training:** built an on-demand training service that runs as a Kubernetes Job for detection, classification and segmentation, with selectable architectures and weights. Added **PyTorch DDP** across pooled GPUs, which cut a production run from **55h to 5h** (−90%).
- Added pause, stop, resume and priority controls, live **MLflow** metrics, and two support services (logger and status) that stream progress over **Kafka**. Used Kubeflow Trainer so training persists across reboots.
- **Jupyter Notebooks:** per-user StatefulSets, each on its own subdomain via Ingress + MetalLB, with code on NFS-backed NAS. Idle notebooks get a warning and are then shut down automatically to free GPUs.
- **Face Registration:** an on-demand Kubernetes Job that supports multiple face models and stores embeddings in **Qdrant**.
- **Camera Service:** an RTSP → HLS pipeline on MediaMTX with an API for the backend to add or remove cameras, tuned to convert thousands of streams in parallel.
- **Device onboarding:** v1 used Ansible + Bash + MQTT to install Docker, NVIDIA drivers and the container toolkit, join the cluster and pre-pull images. **v2 (Deploy Manager)** replaced it to close security gaps. A persistent service now launches Ansible from Kubernetes Jobs to onboard single or bulk devices, **patch** to a pinned version, or **deboard** with full cleanup. It reports a 7-phase lifecycle to the backend over Kafka, with device health from kube-prometheus-stack.

On-prem Arabic speech AI on Qualcomm Cloud AI 100

model porting · inference serving

Speech-to-text for a Saudi government live-translation platform that has to run on-prem on Qualcomm accelerators, not on NVIDIA GPUs or a cloud API.

- **Cohere Transcribe:** ported the model (48-layer Conformer encoder, 8-layer decoder, 4 GB checkpoint) to Cloud AI 100 Ultra: PyTorch → fixed-shape **ONNX** → **qaic-compile** → QPCs. It matched the PyTorch reference token for token and ran about **17× real time**.
- Fixed a compiler failure: the ONNX export produced **1,688 tensor files (7.6 GB)** that qaic-compile could not map. Wrote a repack tool that merges them into one weights file.
- **Qwen3-ASR (Arabic, ~1.7B):** ported the audio encoder by hand and compiled a KV-cache decoder QPC for the same hardware.
- **Serving:** both models run on Qualcomm's vLLM-based Cloud AI inference image behind OpenAI-compatible **/v1/audio/transcriptions** APIs and a LiteLLM gateway. Each program owns one AI100 SoC, deployed as Kubernetes workloads with images in Harbor.

EPM: Eastern Province Municipality, Saudi Arabia

government · delivered on Baseer Builder

- Designed and operate the infrastructure for **200+ CV use cases on 2,000+ cameras**, sized for **50+ A100-class nodes**, as a dedicated environment alongside dev and prod.
- Moved every platform service (frontend, backend, Node-RED, agents, health service, databases, Redis) from Docker into Kubernetes. Full platform deploy went from **5 days to 2 hours**, with dynamic multi-GPU configuration.
- Migrated time-series storage from InfluxDB to **ClickHouse** and ran Redis and databases as StatefulSets on persistent storage.

Platform, DevOps & Infrastructure

company-wide

- **Kubernetes admin** for an on-prem HA cluster (HAProxy) of 50+ nodes across dev, prod and EPM.
- **CI/CD:** two Jenkins pipelines detect changed services, lint, build and push images, and bump tags in an infra repo that **Argo CD** syncs to three environments. Also built GitHub Actions pipelines with SemVer and multi-environment deploys, with config in **Vault** as the single source of truth.
- **Harbor registry** on NAS with backups, NGINX in front, and per-team accounts and quotas. Used by **8 teams**; saves about **$1K/mo** and reduced image vulnerabilities.
- **HA NGINX:** removed a single point of failure with Keepalived virtual IPs across machines and GitHub-synced config. Built a dual-network setup that fails over automatically when one network drops.
- **Security & network:** key-only SSH, per-team restricted users, AD-backed machine login, every exposed service behind NGINX, an ISP-agnostic office network split into VLANs, and a DNS registrar move saving about **$500/mo**.
- **Observability:** Prometheus + Grafana for every service in the cluster, per-camera uptime and device health, with alerts routed to the team. Multi-stage Docker images throughout; a compose-based setup took install time from 2h to 5min.

CLIENT DELIVERY

Saudi Ministry of Economy and Planning: AmplifAI (RAG) & Mudawin (meeting bot)

air-gapped on-prem Kubernetes

Built and deployed both production apps in a cluster with no internet egress, a private registry, internal DNS and SAML/ADFS, all operated over VPN.

- Dockerized both platforms and wrote their manifests for dev, staging and production, wiring each to databases, block storage and Ingress.
- **Unlocked horizontal scaling:** recording bots (one pod per Teams/Meet/Zoom/Webex meeting) were all pinned to **one node** by a shared ReadWriteOnce volume and a required podAffinity. I moved recording to node-local ephemeral storage and streamed finished multi-GB recordings back as a tar over the **k8s exec channel**, with no temp files. Bots now spread across **4 workers**, with right-sized requests and a one-env-var rollback.
- **Prevented an outage:** the checked-in manifest would have deleted **22 live env vars**, including every Graph, Teams, Webex and SQL credential. I shipped a targeted strategic-merge patch instead and reconciled the repo to zero drift.

Expro: Saudi government expenditure entity

Alibaba ECS/ACR · Azure DevOps · Vault

- Designed a lean dev and production topology that saved **SAR 300K/year**. Built approval-gated, rollback-capable Azure DevOps pipelines across **4 branch-mapped environments** with PR checks that block merges.
- Made **Vault** the only source of config: every deploy re-renders env from Vault, staging gets its own secret tree under a read-only policy, token renewal is automated, and a refresh pipeline recreates containers without a rebuild.
- Hardened access with per-machine, key-only deploy accounts without sudo, and locked down world-writable deploy directories.
- Fixed a Docker bridge /16 that blackholed the office subnet, registry pull failures on build agents, deploy stages that ran after failed builds, an env file emptied when Vault was unreachable, and a volume orphaned by a compose rename.

Monsha'at: Saudi government entity

Alibaba ACK + ECS

- Deployed AI services on ECS and platform services on ACK, connected across **multiple VPCs**.
- Replicated dev, QA and pre-prod clusters and built GitLab CI/CD for the AI service.

Humain: storage migration

NVMe · ClickHouse

- Migrated staging `/serve` and ClickHouse data to a **3.5 TB NVMe** with UUID mounts and boot ordering.
- Verified checksums, kept rollback copies, and validated Docker, Kubernetes and ClickHouse after cutover.

Computer-vision PoCs (internship)

May 2024 – Feb 2025

- **Dubai Airport:** turnaround monitoring on live cameras across 3 gates, with a Selenium + SMTP watchdog that detects expired streams, re-logs in, restarts them and emails the team.
- **Ajdan, King Salman Military Base, Gold Chain:** crowd counts, unauthorized-access alerts and multi-camera staff presence. I trained the models and ran inference on live streams.

PROJECTS

Causeway : private-network camera access & recording

Python 3.12 · Next.js 16 · WebRTC · WireGuard · S3

- View and record cameras behind VPNs, jump hosts and SSH tunnels. Each customer runs in its own **Linux network namespace** (Fortinet, GlobalProtect, WireGuard), so a dropped tunnel can't leak traffic.
- 8-gate diagnostics from VPN to stream handshake, WebRTC viewing with LL-HLS fallback, and crash-safe MPEG-TS recording assembled into MP4 on S3. **298 automated tests**.

kuiqctl : persistent kubeadm for roaming Linux hosts

Bash · Python · kubeadm · containerd · Calico · systemd

- Open-source CLI for a single-node cluster that stays Ready across Wi-Fi and DHCP changes (stable loopback identity, mDNS, a systemd address watcher), with preflight gates and lifecycle locks on destructive operations.

CloudPad: on-demand dev sandboxes

K8s · FastAPI · React · AWS

- Chat to build a site and get a live URL and a browser VS Code. Two pods share one volume: one serves the code, one writes it.

AWS EKS document platform

EKS · RDS · S3 · SQS

- Private EKS on a multi-AZ VPC behind API Gateway → VPC Link → internal ALB. Workers read SQS jobs; Pod Identity replaces static keys.

EDUCATION

B.E. Computer Science & Engineering, Osmania University, Hyderabad · GPA 8.4 / 10 — 2021 – 2025