Subeom Kwon

+82 10-5312-6290 | subeomkwon@gmail.com | linkedin.com/in/subeomkwon | github.com/subeom7
Seoul, South Korea

Summary

Software Engineer with hands-on expertise designing and operating end-to-end AI/ML inference pipelines in production environments. Specialized in infrastructure-agnostic AI deployment using Kubernetes and Docker, with proven experience serving 15 production AI models at scale for a medical AI platform. Reduced cloud costs by 90% through architecture migration and KEDA-based autoscaling. Certified Kubernetes Administrator (CKA) with experience in CI/CD, observability, and AWS cloud services.

Education

Virginia Tech — Blacksburg, VA, United States
Bachelor of Science in Computer Science — *May 2023*
- Graduated *Cum Laude*

Experience

JLK — Seoul, South Korea
Software Engineer (Alternative Military Service; Available from May 14, 2026) — *June 2024 – May 2026*
- Fulfilled mandatory military service as a Software Engineer at a **publicly traded medical AI company specializing in stroke analysis**, under the Industrial Technical Personnel program.
- **End-to-End AI/ML Pipeline:** Designed and operated a production-grade AI inference pipeline serving 15 medical AI models on an infrastructure-agnostic Kubernetes cluster from Redis Queue ingestion through GPU-based model execution. Replaced 24/7 cloud containers with a Scale-to Zero architecture, reducing monthly costs by 7M KRW ($5k).
- **Event-Driven Autoscaling:** Architected a dynamic scaling system where each AI model is isolated by a dedicated **Redis Queue**. Configured KEDA to trigger pod creation only upon request arrival, optimizing GPU utilization.
- **Resource Scheduling Strategy:** Solved GPU scarcity on a 5-node cluster by utilizing Kubernetes **PriorityClasses**. Designed a scheduling strategy that differentiates between 'Default' (High Priority) and 'Scale-out' (Low Priority) pods to prevent resource starvation.
- **Deployment & Performance Optimization:** Resolved critical Cold Start latency (5–10 min) caused by large CUDA images by engineering a **Jenkins** pipeline that triggers a DaemonSet to pre-pull images cluster-wide, enabling near-instant AI model rollouts.
- **Observability:** Built a monitoring stack using **Prometheus, Grafana, Node Exporter, and cAdvisor** to track node, pod, GPU, and container metrics for production AI workloads.

Projects

AWS S3 Migration for Medical Imaging Data | *S3, Terraform, AWS DataSync* — Jan 2025 – Feb 2025
- **Data Migration at Scale:** Migrated **250 million DICOM objects (60TB)** from Ncloud Object Storage to AWS S3 to support secure medical data operations and VANTA readiness for US market expansion.
- **Algorithmic Load Balancing:** Overcame AWS DataSync limitations and network bottlenecks for small-file transfers. Engineered a custom data sharding pipeline using a **Min-Heap greedy algorithm** to evenly balance the network payload based on **object count** across 8 parallel worker nodes.
- **Migration Orchestration & Concurrency:** Streamlined the end-to-end workflow by integrating Terraform and Python. Utilized **ThreadPoolExecutor** for concurrent S3 pagination, provisioned DataSync tasks via **Terraform**, and orchestrated parallel executions using **boto3**, reducing a week-long manual process to a few days.

Technical Skills

Languages: Python, Java
Cloud & DevOps: Kubernetes, Docker, KEDA, Helm, ArgoCD, Jenkins, DataSync, EC2, S3, CloudFront
Infrastructure: Linux, Redis, Nginx
AI/ML Infrastructure: GPU Inference Operations, CUDA-based Containers, AI Model Serving Infrastructure
Backend: FastAPI, Celery, Gunicorn, Uvicorn
Observability: Prometheus, Grafana, Node Exporter, cAdvisor
Version Control: GitLab

Certifications

CKA (Certified Kubernetes Administrator) | *The Linux Foundation* — Feb 2026