Muhammad Awais Khan

+92-3365206052 | ssc.awaiskhan.2490@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Systems-focused **Cloud Platform & MLOps Engineer** bridging the gap between low-level GPU optimizations and scalable infrastructure. Expert in **container orchestration**, eliminating serverless **cold starts**, and automating robust **CI/CD pipelines**. Proven track record of architecting high-performance systems and managing end-to-end deployment workflows to drive operational efficiency in production GenAI environments.

TECHNICAL SKILLS

- **Infrastructure & Cloud: Kubernetes (K8s)**, **Docker**, AWS, Google Cloud Platform (GCP), Cloud Run, Serverless Infrastructure.
- **CI/CD & Automation: GitHub Actions**, Git, Linux/Bash Scripting, Automated Testing Pipelines.
- **MLOps & Optimization:** RunPod, **CUDA** (Custom Kernels), TensorRT, Triton Inference Server, Memory Caching (vmtouch), Model Drift Detection.
- **Languages & Frameworks:** Python, C++, PyTorch, FastAPI.

EXPERIENCE

AI Engineer | *Funsol Technologies · Islamabad, Pakistan* | *July 2025 – Present*

- **RunPod Bottleneck:** Engineered a high-performance storage layer on RunPod by migrating model weights out of **Docker** layers into persistent volume mounts, slashing serverless inference **cold starts** from 80s to <1s.
- **Throughput:** Optimized production image-to-video inference pipelines via Sage Attention and vmtouch memory caching, doubling system throughput and significantly reducing GPU compute costs.
- **Hardware:** Authored custom **CUDA** kernels from scratch in C++ to optimize memory tiling and global memory access, maximizing hardware utilization on high-end GPUs.
- **Infrastructure Architecture:** Architected 5 multimodal generative AI pipelines on serverless GPU infrastructure, managing full integration with production web application backend.
- **CI/CD Automation:** Engineered automated **CI/CD** workflows using **Docker** and **GitHub Actions** for zero-downtime ML model deployments, enabling continuous iteration across production inference endpoints.
- **Pipeline Reliability:** Designed and deployed end-to-end serverless inference pipeline for the ACE-Step 1.5 generative audio model, containerized with **Docker** and orchestrated to RunPod via **GitHub Actions**.

INFRASTRUCTURE & MLOPS PROJECTS

Automated ML Retraining & Deployment System (ML-Auto-Retraining-System)

- Architected an end-to-end self-healing MLOps pipeline featuring automated model drift detection and continuous integration.
- Configured **GitHub Actions CI/CD** workflows to automatically trigger cloud-based GPU retraining and rolling deployments upon detecting performance degradation, achieving zero-downtime updates.

Knowledge Nest Production Deployment (knowledge-nest-deployment)

- Containerized a complex RAG-based knowledge retrieval engine using multi-stage **Docker** builds, reducing image sizes by 60% for rapid container provisioning.
- Deployed scalable containerized infrastructure to cloud environments (GCP Cloud Run / AWS), implementing robust health checks and environment variable management for production reliability.

CUDA Neural Network Inference Acceleration

- Implemented custom **CUDA** kernels from scratch, optimizing shared memory tiling and coalesced global memory access patterns to maximize memory bandwidth utilization.

- Leveraged GPU Tensor Core operations to accelerate matrix multiply-accumulate workloads, achieving measurable throughput improvements over baseline implementations.

EDUCATION

FAST National University of Computer and Emerging Sciences | *2022 – 2026*
BS Computer Science | Islamabad, Pakistan

LEADERSHIP & CERTIFICATIONS

- **Head of Photography**, Google Developer Groups on Campus (GDGoC)
- **Certifications:** Introduction to TensorFlow for AI, ML, and Deep Learning (DeepLearning.AI / Coursera)