Ivan Morrow

GitHub | Ivanmorrow14@gmail.com |309-756-6188 | LinkedIn

Professional Summary

Platform and Site Reliability Engineer with 5 years of experience building and optimizing scalable, high-performance infrastructure across cloud and on-prem environments. Proven track record supporting mission-critical workloads by driving improvements in system reliability, observability, and incident response. Skilled in investigating and resolving complex production issues in both cloud-native and low-latency distributed systems. Proficient in Python, Kubernetes, Cloud Engineering, and CI/CD tooling, with a proven track record of implementing robust monitoring and alerting systems, driving operational excellence, and removing bottlenecks within software systems with iterative improvements. Committed to providing best-in-class infrastructure and support for development teams.

Technical Skills

Programming Languages: Python (6 yrs.), SQL (5 yrs.), Bash (1 yr.), Rust (2 yrs.), Typescript (2 yrs.), C++ (2 yrs.)
Cloud Technologies: GCP (2 yrs.), Kubernetes (1 yr.), Docker (1 yr.), Terraform (1 yr.), Helm (1 yr.), ArgoCD (1 yr.)
Tools: Splunk (5 yrs.), Grafana (4 yrs.), Airflow (2 yrs.), Gitlab (1 yr.), Gitpod (1 yr.)
ML Frameworks: PyTorch (1 yr.), TensorFlow (1 yr.)
Operating Systems: Linux (5 yrs.), Windows (5 yrs.)

Professional Experience

Belvedere Trading – Chicago, IL (2020 - Present)

Cloud DevOps Platform Engineer May 2024 – Present

- Architect and manage cloud infrastructure using Infrastructure as Code, ensuring reliability, scalability, and security.
- Built a full-stack CI/CD observability pipeline by scraping GitLab Runner metrics via pod monitoring and exporting detailed pipeline stats from the GitLab API to BigQuery, powering Grafana dashboards used to monitor reliability and SLAs.
- Used observability insights from GitLab pipeline and runner metrics to identify and address key CI/CD bottlenecks and failure patterns.
- Identified image pull failures as a key source of instability in CI/CD cluster and implemented a solution by migrating Docker images to Google Artifact Registry co-located in same region as GKE cluster and configuring proxy repositories to locally cache third-party containers, eliminating pull errors and reducing job failure rate from ~20% to 0.4%.
- Provide first class support to internal engineers on our cloud development platform troubleshooting SDLC challenges, optimizing CI/CD pipelines, and driving continuous improvements in the software development lifecycle.
- Developed robust data pipelines in Python to process and publish terabytes of data daily, leveraging GCP, BigQuery, and Airflow.
- Built reusable, internal tools using Python and Bash to streamline cloud operations, improve developer productivity, and enhance system observability.

Product Specialist - Electronic Trading January 2023 – October 2024

- Acted as a technical liaison between production support and development teams, resolving high-impact trading system issues and driving cross-functional collaboration to improve system reliability and developer velocity.
- Developed real-time health monitoring dashboards to track market data latency and throughput enabling proactive incident detection and minimizing downtime.
- Deployed observability solutions (metrics, logs, alerts) to improve visibility into production systems, reducing mean time to detection and accelerating root cause analysis.
- Tuned and re-architected market data services to reduce latency from 100–600ms to 5–15ms, significantly improving trading performance in latency-sensitive environments.
- Developed automation and monitoring tools in Python to streamline operational workflows, monitor critical services, and enhance overall production system stability.
- Collaborated with engineering teams on system design and rollout strategies to enhance maintainability and fault tolerance.
- Architected and managed service rollout strategies across on-prem infrastructure by analyzing CPU core usage, evaluating service-specific core requirements, and optimizing host selection to ensure efficient resource utilization and performance at scale.

Production Engineer - Trading System September 2020 – January 2023

- Led overnight Level 1 production support (4 pm - 2 am) for a distributed trading system, diagnosing and resolving critical incidents under high-pressure conditions.
- Built a Python-based service to ingest email alerts and surface them in the team's internal monitoring tool, centralizing incident visibility and reducing mean time to resolution by ensuring no critical alerts were overlooked.
- Maintained and optimized a globally distributed trading system with hundreds of microservices across multiple data centers, ensuring high availability and low-latency performance.
- Built automated monitoring and alerting tools in Python to improve system reliability, reduce incident response time, and prevent recurring issues.

Coe College – Cedar Rapids, IA June 2018 – Aug 2018

Machine Learning - Undergraduate Research Assistant

- Designed and implemented a neural network using TensorFlow to detect and localize neutron emissions based on scintillation patterns in a handheld detection device.
- Built a model that analyzed light emissions from scintillating glass to infer neutron directionality, enabling real-time user guidance toward radioactive sources (e.g., illicit nuclear material).
- Collaborated with a multidisciplinary research team and co-authored a peer-reviewed publication detailing the model's development and results..

Projects

K8s Based Distributed LLM Inference Pipeline

- GitHub Repo
- Designed and deployed a scalable LLM inference API on GKE using FastAPI and Hugging Face Transformers.
- Configured Helm charts for infrastructure-as-code deployment and simplified cluster updates.
- Optimized model inference performance by experimenting with OpenVINO quantization and benchmarking response latency (~367ms on GPT-2).
- Integrated HuggingFace and PyTorch.

Quantized Flower Classification Model - Edge TPU

- GitHub Repo
- Fine-tuned and quantized a MobileNetv2 model enhanced with Attention for efficient deployment on a Coral Edge TPU.
- Model trained to classify flowers based on images.
- Optimized the entire pipeline: parallel data augmentation with TPUs, compiled and deployed final 8-bit model to Edge TPU.
- Demonstrated low-latency, on-device inference (<5 ms per inference).
- Practiced hardware-aware design.

Education

Coe College, Cedar Rapids, IA
BA – (May 2020)
Majors: Computer Science and Physics

Colorado State University, Fort Collins, CO
Graduate Certificate in Computer Systems Engineering – (Current)
GPA: 4.0

Relevant Courses

- CS545 - Machine Learning (Fall '24)
- ECE528 - Embedded Systems and Machine Learning (Fall '24)
- Fast.ai (Winter '25)
- Coursera - Machine Learning Specialization (Winter '24)