Andrew Garmon

andrew.garmon@gmail.com | (863) 838-9932 | linkedin.com/in/andrewgarmon

Experience

Software Engineer III, Site Reliability (Core ML Infrastructure), Google | Pittsburgh, PA

2024 – Present

- Designed an end-to-end telemetry system for Google's core ML training platform (XManager) and ML workloads, created a multi-quarter roadmap, and implemented the design with a team of engineers, enabling holistic observability via custom data pipelines, distributed tracing, and end-to-end SLIs that accurately measure user-perceived reliability and latency.
- Partnered with the Google DeepMind AI platform team to deploy Gemini shared inference servers, developing load tests to benchmark serving throughput and scalability and inform capacity planning for shared TPU resources.
- Drove a fundamental redesign of the platform's traffic management system by authoring a proposal that differentiated between quota, rate-limiting, and throttling to improve stability and ensure fair resource allocation under high load.
- Enhanced resource efficiency for ML Training services on Borg by tuning performance, optimizing horizontal and vertical autoscaling configurations and resolving conflicts between autoscaling and capacity management systems.
- Established and led the team's incident management practice, authoring postmortems after outages to drive improvements in alerting and triage automation. Designed and delivered drills to train 30+ engineers on incident response protocols.
- Pioneered the migration of service security policies to an Infrastructure as Code (IaC) model, automating the configuration of access controls and emergency bypass procedures to reduce toil, mitigate outage risk, and improve security posture.
- Leveraged CI/CD pipelines to improve developer velocity by configuring automated release builds and rollout schedules, establishing automated presubmit and integration testing against release candidates.

Software Engineer II, Site Reliability (Core ML Infrastructure), Google – Pittsburgh, PA

2022 – 2024

- Rapidly mastered on-call responsibilities for Gemini base model training infrastructure, ensuring the successful landing of massive ML training workloads by demonstrating expertise in infrastructure and model architecture to mitigate outages.
- Mitigated data corruption outages in ML training services by independently designing large-scale database recovery tests, driving the resolution of critical gaps in backup testing, environment isolation, and data recovery processes.

Leading Petty Officer, US Navy – Norfolk, VA

2016 – 2022

- Led and mentored 3 teams of 40 technicians through a complex, multi-year systems overhaul, directed integration testing for 11 advanced RADAR, weapons, and distributed data systems, and spearheaded a 5-year critical infrastructure integrity project, ensuring 100% system isolation across 3,000 compartments and saving the US Navy $2.1MM in projected costs.

Education

University of Florida

2022

B.S. in Computer Science, *Cum Laude*

Skills

Languages & Operating Systems: Python, C++, Go, Bash, SQL, Linux Environment

Cloud & Infrastructure:

- **Distributed Systems:** Microservices Architecture, API Design, RPC, Protocol Buffers
- **Container Orchestration, IaC, and Automation:** Kubernetes, Borg, Docker, Infrastructure as Code (IaC), Performance Tuning & Optimization, Capacity Planning, Traffic Management (Rate Limiting, Throttling)
- **Networking:** Load Balancing Policies, HTTP Traffic Analysis, DNS/TCP Fundamentals

Observability & Telemetry:

- SLI/SLO Design & Implementation, Distributed Telemetry Design (Metrics, Tracing, Logging), Monitoring & Alerting
- *Experience with systems conceptually analogous to Prometheus, Grafana, and the ELK Stack.*

CI/CD & Build Systems:

- CI/CD Pipelines, Automated Testing Frameworks, Build Systems (Bazel), Version Control (Git, Mercurial)

ML Systems & Operations (MLOps):

- Debugging large-scale ML workload issues with performance, data integrity, and training metrics.
- Familiarity with ML stack, including frameworks (TensorFlow, JAX), compilers (XLA), and hardware accelerators (TPUs).
- Strong conceptual understanding of ML/LLM fundamentals, model architecture, and the training lifecycle.

Volunteering

Panelist and Mentor, Google Veterans Network | Pittsburgh, PA

2023 – Present

- Served as a technical panelist for VetsInTech, sharing expertise on AI/ML careers with an audience of military veterans.
- Participated in Google VetNet career panels and resume reviews to guide veterans seeking roles in tech.