MIGUEL MERLIN

NYC Metro Area — mmerlin@stevens.edu — github.com/miguel-merlin — linkedin.com/in/miguel-angel-merlin-arriola

EDUCATION

Stevens Institute of Technology, Hoboken, NJ — Enrolled: September 2022 — Expected: May 2026
Computer Science and Math B.S. — Overall GPA: 3.91 — Major GPA 3.91

EXPERIENCE

Amazon.com (Alexa AI) — Bellevue, WA
Software Development Engineer Intern — June 2025 - August 2025

- Designed and implemented a Java-based Inference Throttling mechanism to maintain Alexa+ GPU cluster utilization near the optimal 80% target, deferring autoscaling triggers until necessary
- Developed a token-aware throttling strategy that estimates LLM token consumption (input and projected output tokens) to selectively defer high-cost requests when cluster load approaches saturation, ensuring latency guarantees for concurrent traffic.
- Integrated a priority-aware queuing system that classifies and schedules requests based on criteria such as traffic type (e.g., production vs. test) and estimated token usage, optimizing resource allocation under load.

Amazon.com (AGI) — Bellevue, WA
Software Development Engineer Intern — June 2024 - August 2024

- Contributed to the design and implementation of a high-throughput, distributed inference system in Java that underpins Alexa+, Amazon's next-generation LLM-powered voice assistant
- Designed and implemented a Prompt Construction Framework in Java that dynamically synthesizes LLM prompts by leveraging user intent, conversational context, and system metadata, significantly enhancing the relevance and coherence of Alexa+ model outputs.
- Engineered a reactive, non-blocking orchestration layer in the Alexa backend using JavaRx, enabling early-stage computation with partial upstream input and reducing end-to-end latency by 13%

Amazon.com (Alexa AI) — Bellevue, WA
Software Development Engineer Intern — May 2023 - August 2023

- Designed and implemented a high-performance, Java-based predicate evaluation engine that efficiently matches incoming JSON events against a dynamic set of rule-based JSON predicates.
- Developed an Experimentation Framework in Go to dynamically divert traffic from underperforming ML models based on real-time eligibility criteria, resulting in a 10% uplift in perceived customer satisfaction through intelligent routing.

Stevens Managed Investment Fund (SMIF) — Hoboken, NJ
Quantitative Developer — January 2024 - Present

- Engineered a Synthetic Limit Order Book (SLOB) framework to simulate market microstructure and evaluate execution strategies under varied liquidity regimes. The system enables data generation for high-frequency trading (HFT) models and supports robust strategy backtesting for alpha optimization and latency-sensitive execution.
- Architected the initial HFT system infrastructure, implementing a low-latency trading engine in C++ with Python bindings for orchestration, logging, and analytics. Integrated the system with Kubernetes for containerized deployment.
- Built a C++ portfolio optimization engine leveraging convex optimization techniques. The tool supports multi-asset backtesting via a simulated market environment and provides a Python interface for analysts to design and validate dynamic asset allocation strategies.

Stevens Blueprint — Hoboken, NJ
Technology Vice President — May 2023 - Present

- Developed a CDK-based Infrastructure Pipeline Builder that abstracts the creation of cloud resources and CI/CD workflows for Blueprint projects, enabling rapid and consistent deployment across multiple NPO applications.
- Implemented modular CDK constructs to provision complete cloud architectures—including Lambda functions, S3 buckets, DynamoDB tables, API Gateway, and EC2-backed Spring Boot services—supporting both serverless and containerized stacks, with full environment-specific configuration support and seamless deployment of React frontends via S3 + CloudFront.

SKILLS

- **Python**
- **Java**
- **JavaScript**
- **C++**
- **Go**
- **Machine Learning**