Bryan Yue - Software Engineer

yue_bryan123@gmail.com | 425-749-2741 | https://bryanyue.com | https://linkedin.com/in/bryanyue322 | https://github.com/byue

Summary

Software Engineer with 7 years of experience developing production-grade distributed systems and cloud-native backend platforms at Amazon, Microsoft, and Bloomberg. Specializes in system design and performance optimization across high-throughput APIs, event-driven workers, and cloud-native microservices on AWS and Azure. Drives cross-functional roadmap planning, mentors engineers through complex technical trade-off discussions, and delivers reliable, observable production systems at scale.

Skills

- Languages: Java, C++, Python, C#, TypeScript, Clojure, Ruby, SQL, Bash
- Frameworks: Spring, .NET Core, FastAPI, React
- Storage: AWS DynamoDB, PostgreSQL, AWS Elasticache, Redis, Memcached, AWS S3
- Compute: AWS ECS, AWS EC2, AWS Lambda, AWS Step Functions
- Messaging and Streaming: Kafka Streams, AWS SQS, AWS SNS, AWS Kinesis, RabbitMQ, WebSocket, REST API, gRPC
- Deployment and Operations: CI/CD, GitHub Actions, Azure DevOps Pipelines, Docker, Docker Compose, AWS CloudFormation, AWS CDK, AWS AppConfig
- Observability: Prometheus, Grafana, Loki, AWS CloudWatch, AWS OpenSearch
- AI and Machine Learning: Azure OpenAI, AWS Bedrock, PyTorch, Pandas, Numpy

Work Experience

Amazon | Machine Learning Engineer | Jan 2026 - Present

- Designed an ads search query personalization feature using ML-derived brand preference signals, increasing CTR from 3% to 8.3% and CVR from 28.9% to 56.6%.
- Optimized real-time ads relevance model inference to p99 30ms via quantization, model selection, and compilation techniques, enabling low-latency ranking at production scale.

Amazon | Software Development Engineer | Nov 2021 - Jan 2026

- Architected AWS Shield annual backend roadmap and led peer scope planning with data-driven prioritization, mentoring teammates through technical design and code review standards for high-impact DDoS mitigation initiatives.
- Engineered near real-time mitigation automation in Java, Ruby, and Clojure for AWS EC2 network defenses using distributed event-driven pipelines, improving DDoS mitigation efficacy from 10% to 80% across global production traffic at scale.
- Designed and launched a centralized cross-team service for known-offender IP list management and deployment, reducing on-call MTTR by 30 minutes, leading incident response coordination, and driving blameless postmortem actions to harden mitigation workflows across orgs.

Microsoft | Software Engineer | Oct 2019 - Nov 2021

- Designed AzureStack bulk VM goalstate API in C# and caching strategy that reduced dataplane backpressure by 50x and cut VM provisioning p99 latency by 5 seconds, improving availability for high-concurrency VM provisioning scenarios by 11%.
- Migrated VM agent protocol to reduce infrastructure traffic by 8x while maintaining deployment safety and backward compatibility across heterogeneous VM environments.
- Ported Instance Metadata API capabilities from Azure to Azure Stack in C++, enabling container orchestration and VM provisioning workflows in resource-constrained on-premises environments.

Bloomberg | Software Engineer | Sep 2018 - Sep 2019

- Led migration of petabytes of historical market data into SamayDB, a high-throughput time-series database serving billions of daily queries, executing within a strict 6-hour maintenance window while meeting throughput and availability SLAs.
- Designed and prototyped a stock-insights ranking API integrating data science models, enabling higher-signal analytics for downstream consumers and accelerating internal tooling workflows.

Kernel Labs | Machine Learning Intern | Mar 2018 - Jun 2018

- Implemented a BLSTM speech separation model with L2 regularization and dropout, achieving 82% test set accuracy on multi-source audio decomposition.
- Built an audiobook web scraper to construct labeled train, validation, and test datasets, and visualized feature distributions using pandas and MFCC plots.

Projects

TradeStrike

- Architected dockerized services for near real-time stock volume and price spike alerting, achieving sub-1-second spike-to-notification p99 latency across a push-based event pipeline with aggressive Redis caching.
- Computed configurable rolling window average and standard deviation statistics over WebSocket streams via Kafka stream workers for real-time anomaly signal generation.
- Instrumented a full observability stack with Prometheus, Grafana, and Loki to monitor pipeline throughput, worker p99 latency, and alert delivery health.

Reelify

- Built a text-to-video Python API (FastAPI) backed by self-hosted multimodal LLM under 8-12GB VRAM constraints.
- Optimized model inference in Pytorch using quantization and keyframe interpolation, reducing frame render p99 latency by ~40%.

Education

University of Washington | Bachelor of Science in Computer Science | 2014 - 2018 | GPA: 3.89 | Phi Beta Kappa Honor Society