Howard Nguyen

howardanguyen@gmail.com phoxelua www.howardanguyen.com howardanguyen

Senior software engineer with 10+ years of experience scaling high performance distributed systems and AI/ML infrastructure

Experience

Pinterest *C++, Python, real-time inference*

SENIOR SOFTWARE ENGINEER (AI/ML SERVING) *2021 - Present*

- Architected the company's paved path for ML inference and GenAI on Kubernetes by synchronizing the ORKs of 6+ XFN teams over 2+ years. This included shipping a prototype CRD with automated CICD, topology based model deployment, user defined partitioning, monitoring, sidecar injection, quota management, and custom UI
- Led a multi-team initiative to deliver remote inference and distributed GPU model partitioning, resulting in $3M of annual infra savings, 10X more model capacity, and 3% boost in engagement. This enabled the launch of the largest model in Pinterest history
- Acted as the sole technical lead for the team which included reviewing *every* design doc, postmortem, and client request. Also served as the final point of escalation for all blockers and difficult debugging work
- Acted as the interim engineering manager, reporting directly to the VP of Engineering, to set yearly OKRs, track individual project progress, organize morale events, present technical brown bags, review performance packets, and provide career mentorship
- Led an org-wide strategic objective to secure S3 data isolation between dev and prod workflows
- Led a multi-team project to deliver the second generation of model and feature monitoring at Pinterest
- Assumed ownership of model deployment service from a partner team, saving $2M in annual infra, increasing success rate from 76% to 98%, decreasing E2E latency by 40%, and deprecating 20,000 lines of code
- Managed teammate's projects simultaneously, ensuring the delivery of the company's first GenAI use case, feature trimming ($1.3M infra savings), migration onto Weights & Biases, hardware benchmarking, model composition, and U24/AMI upgrade
- Architected and built a self-service cluster management and dark traffic configuration system that would be adopted company-wide

Meta (Instagram) *C++, Python, stream processing*

SENIOR SOFTWARE ENGINEER (ML PLATFORM) *2018 - 2021*

- Led initiative to track demand within Instagram Machine Learning (IGML). This included implementing distributed tracing for client cost attribution, aggregating metrics across shared services, and creating a standard system for regression detection and admission control. Also delivered teamwide keynote and workshops on managing resources and solving future regressions
- Led initiative to map the data lineage of IGML's ecosystem into a knowledge graph then delivered a stakeholder keynote on how to leverage it for better dependency management, privacy control, and cost-benefit analysis
- Rearchitected IGML client logging and real-time processing services into a more efficient layer that reduced host footprint by 25%
- Cut real-time feature storage footprint by 20% by deleting features with low model significance
- Optimized feature storage serialization and batched keys based on access to reduce footprint by 30% and cache latency by 15%
- Migrated ML features to new framework, reducing database QPS by 75%, fleet size by 10%, and increasing engagement by 0.7%

Modsy *Django, MySQL, AWS*

LEAD BACK-END ENGINEER *2017 - 2018*

- Architected and scaled a custom, in-house 3D render pipeline that increased throughput by 3X
- Saved customer success 20 hours/week by creating an custom order management system with automated refunding
- Improved query times by 50% and unified data access patterns by creating a data access layer for products and 3D render jobs
- Increased unit test coverage from 58% to 80% and removed 20,000 lines of code by standardizing best coding practices

Captricity *Django, PostgreSQL, Celery*

FULL-STACK ENGINEER *2015 - 2017*

- Optimized job throughput by using Celery to asynchronously batch Amazon Human Intelligence Tasks (HITs) based on perceived effort, task type, and completion time. Further reduced job turnaround-time by building real-time repricing of undesirable HITs.

Skills

Proficient with: Python, PostgreSQL/MySQL, Docker, Kubernetes, Tensorflow Serving, AI/agentic coding, Django, Flask, AWS EC2/S3/EBS

Experience with: C++, Java, Go, MATLAB, Ruby on Rails, Kafka, Memcached, Redis, Celery, MongoDB, Presto, Hive, RocksDB, Flink

Familiar with: Nvidia Dynamo/Triton, PyTorch, MLflow, Weights & Biases, HTML/CSS, JavaScript/TypeScript, Angular, React, Elixir

Education

University of California, Berkeley *3.5 GPA*

B.S. IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE *Graduated 2015*

Coursework: Artificial Intelligence, Algorithms, Computer Architecture, Databases, Data Structures, Machine Learning, Graphics