Jerry Duncan

Seattle, WA | jerry@jerryduncan.dev | linkedin.com/in/jerry-duncan-dev

EXPERIENCE

ByteDance | Seattle, WA — Q3 2023 – Present

Machine Learning Infrastructure Engineer (Platform Team)

- **Cut idle GPU hours by ∼30%** by building custom resource scheduler scaling on worker utilization instead of raw GPU utilization, reducing infra costs across production ML workloads
- **Sustained 2k+ QPS and hundreds of millions of tasks monthly** by building v2 async job queue (Redis, MongoDB, S3-equivalent), enabling dynamic workflows for TikTok ML features
- **Delivered 99.99% availability across US, EU, and ROW** by deploying multi-region clusters with disaster recovery and automated failover, ensuring uninterrupted TikTok ML workloads during data center outages
- **Maintained uninterrupted operation of mission-critical async job queue** by onboarding 6 engineers during turnover cycles, authoring runbooks, and mentoring successors into ownership

ByteDance | San Jose, CA (Remote) — 2022 – Q1 2024

Machine Learning Optimization Engineer

- **Cut training costs by ∼$1.5M/month (>6M in H1 2023)** by accelerating SDXL throughput 4.7× on 1k A100s; earned internal excellence award and partnered with NVIDIA NeMo on kernel-level optimizations
- **Saved ∼10M GPU-hours and cut training latency up to 4×** on A10 GPUs by optimizing diffusion model training for TikTok features (AI Self, AI Moji), powering 100M+ generations
- **Expanded compute capacity during GPU shortages** by evaluating 256 TPUs and migration tradeoffs, securing additional A100 GPUs through strategic resource exchange
- **Recognized as ML optimization expert** supporting ∼100 engineers; led deep dives on FSDP, DeepSpeed, and Megatron, and shaped PhD intern pipeline through ∼100 resume screens and ∼10 interviews

ByteDance | San Jose, CA (Remote) — Summer 2021

Software Engineering Intern — ML Systems

- **Enabled parameter-server scaling for ML researchers** by adding BytePS distributed backend support to an internal PyTorch-based training framework
- **Improved training efficiency and prevented costly misconfigured runs** by implementing profiling, config validation, and early stopping features
- **Extended internal PyTorch training framework**, later adopted by multiple research teams, strengthening department-wide ML infrastructure

SKILLS

Deep Expertise: Python, CUDA/NCCL, distributed training (FSDP, DeepSpeed, Megatron-LM), large-scale model optimization, GPU scheduling, multi-region infrastructure, Kafka, Redis
Working Knowledge: Go, C/C++, TPUs, Hive, MongoDB, RDS, RocketMQ, Spark, Docker, CI/CD, S3

EDUCATION

University of Tennessee, Knoxville — 2019–2021

M.S. & B.S. Computer Science, summa cum laude
Relevant coursework: Scalable & Resilient AI/ML Systems, Deep Learning, Reinforcement Learning

PATENTS

Improving Task Execution and Resource Management (U.S. Patent Application No. US 19/053098, pending)