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)

- **Owned design and global rollout of async inference platform** on Kubernetes, sustaining 2k+ QPS and 100M+ monthly tasks; introduced a Mon/Thu release cadence with canary rollouts to reduce outage risk from Friday releases and improve on-call stability
- **Directed GPU utilization strategy** via a scheduling service that cut idle time ∼30% across clusters, reclaiming capacity for latency-critical workloads and significantly lowering cost-to-serve
- **Drove reliability roadmap** by implementing global failover for inference dispatch, achieving 99.99% availability and ensuring uninterrupted delivery for mission-critical TikTok ML features
- **Built and scaled team capability** by onboarding 5 engineers, transferring ownership of high-availability serving systems, and establishing team practices that sustained long-term reliability across the team

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

Machine Learning Optimization Engineer

- **Led optimization strategy for diffusion models**, delivering up to 4× faster train/infer and saving ∼10M GPU-hours; enabled TikTok generative features used in 100M+ creations
- **Directed large-scale performance work on SDXL in collaboration with NVIDIA NeMo**, boosting throughput 4.7× on 1k+ A100s; earned an excellence award and drove multi-million $ monthly savings
- **Expanded compute capacity and efficiency at scale**, improving a 12k-GPU production run for MegaScale LLM MoE by 10% and evaluating TPU migration across 256 devices to maintain continuity during global GPU shortages
- **Recognized as a domain expert in ML optimization**, advising 100+ engineers on best practices for distributed training (FSDP, DeepSpeed, Megatron) and inference (ONNX, TensorRT), and owning the PhD intern pipeline, including resume screening, interviews, and final hiring decisions

ByteDance | San Jose, CA (Remote) Summer 2021

Software Engineering Intern — ML Systems

- **Enabled parameter-server scaling in PyTorch-based training framework** by extending it with BytePS distributed backend support, broadening capacity for ML researchers
- **Improved training reliability and efficiency** by adding profiling, config validation, and early stopping, reducing misconfigurations and wasted compute

TECHNICAL SKILLS

Programming/Systems: Python, C/C++, Go; Docker, Kubernetes; AWS, GCP, S3
Distributed/Infra: Kafka, Redis, RocketMQ; multi-region fault tolerance; async systems
ML Serving & Infrastructure: PyTorch internals, CUDA, NCCL, custom kernels, GPU scheduling; ONNX Runtime, TensorRT; quantization, distillation, compression; observability/monitoring, resiliency, canary deployments

EDUCATION

University of Tennessee, Knoxville 2016 – 2021

B.S. Computer Science, summa cum laude, 2019
M.S. Computer Science, summa cum laude, 2021

PATENTS

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