# Adrián Schmidt Staff ML Engineer

Santa Clara, CA · adrian.schmidt@gmail.com · +1 408-555-0119 · linkedin.com/in/adrianschmidt

## PROFILE SUMMARY

- **Staff ML Engineer** with **11 years** leading **large-scale training and inference** programs across **GPU, autonomous-vehicle, and foundation-model** workloads, specializing in **distributed training, inference-runtime performance, and ML platform governance**.
- Hands-on across primary language (**Python**), systems language (**C++** with CUDA basics), training framework (**PyTorch** DDP and FSDP), inference runtime (**TensorRT-LLM**, **vLLM**, **Triton**), distributed compute (**Ray** and **NCCL**), and orchestration (**Kubernetes** with the NVIDIA GPU operator).
- Deep expertise in **multi-node H100 training**, **FSDP migration**, **continuous batching and paged attention**, and **KV-cache tuning**, using **RFC-driven design reviews** and **SLO-backed rollouts** to ship **inference that gets cheaper every quarter**.
- Partners with **Research, Product, Hardware, and Finance** on inference-cost roadmaps and capacity planning, briefing executives on **6-quarter cost-of-inference projections** and owning the on-call rotation for 4 high-stakes services.
- Tech lead of **7 ML engineers** who set up the team's **RFC governance** (18 RFCs shipped), the inference-runtime onboarding curriculum, and the **Inference Platform forum** the wider org now runs on.

## TECHNICAL SKILLS

| | |
|---|---|
| **Languages:** | Python, C++ (CUDA basics), Bash, Rust (basics) |
| **Distributed Training:** | PyTorch DDP and FSDP, Megatron-LM, DeepSpeed, NCCL, Horovod, multi-node H100 clusters |
| **Inference Runtime:** | TensorRT-LLM, vLLM, Triton Inference Server, continuous batching, paged attention, KV-cache tuning, quantization |
| **Orchestration:** | Kubernetes (GKE), NVIDIA GPU operator, Ray, Slurm, Argo Workflows |
| **ML Platform:** | MLflow, Weights & Biases, model registry, RFC governance, on-call playbooks |
| **Observability & Cost:** | Prometheus, Grafana, DCGM GPU metrics, inference SLO dashboards, cost-of-inference reporting |
| **Leadership:** | tech lead of 7, hiring loops, architecture reviews, executive briefings |

## EDUCATION

**ETH Zurich** — M.S. in Computer Science — Zurich, Switzerland · Sep 2012 - Jun 2014

**Universität Hamburg** — B.Sc. in Computer Science — Hamburg, Germany · Sep 2009 - Jul 2012

## WORK EXPERIENCE

### NVIDIA — Staff ML Engineer — Santa Clara, CA · Mar 2021 - Present

- Tech lead for the **inference-runtime team** of 7 engineers, owning the **Triton and TensorRT-LLM** stack that serves **6 production models at 22k QPS** across internal and partner workloads.
- Led the **FSDP migration** from legacy DDP across 4 training programs, unlocking **multi-node H100** training at trillion-parameter scale and cutting per-step time by **34%**.
- Drove an inference-optimization program on 3 flagship models with **paged attention, continuous batching, and KV-cache tuning** in **TensorRT-LLM**, lifting throughput **52%** and cutting p99 latency **38%**.
- Set up the team's **RFC governance** and chair the bi-weekly Inference Platform forum, shepherding **18 RFCs** through review, including the org-wide **vLLM** evaluation and adoption plan.
- Own the on-call rotation for **4 high-stakes inference services**, from runbooks and post-incident reviews to **SLO error-budget enforcement**, holding **99.95% availability** for 6 straight quarters.

### Cruise — Senior ML Engineer — San Francisco, CA · Jul 2014 - Feb 2021

- Owned the **perception model training platform**, building **Horovod** distributed training pipelines for 12 perception models across LiDAR, camera, and radar modalities.
- Built the on-vehicle inference runtime on a custom **GPU operator**, serving **8 perception models** at **p99 latency under 25 ms** on the compute budget of a single drive unit.
- Led the **Kubernetes** training-cluster migration across 3 GPU pools, unlocking **4x training throughput** and cutting per-experiment cost by **42%**.