

Daniel Mandragona

ML Performance Engineer | TPU Kernels · Parallelism & Sharding · Quantum Computing

 Sunnyvale, CA  danielmandragona@gmail.com  954-940-1207  dandragona.me  github.com/dandragona

EXPERIENCE

Google	Sep 2018 – Present
Machine Learning Engineer	Mountain View, CA
ML Performance Engineering — MaxText	Apr 2026 – Present

- Delivered a 25% speedup on an FFT Pallas TPU kernel for a customer engagement.
- Unlocked a 2× speedup on a large customer’s model via kernel fusion and parallelism strategies to remove exposed comms.
- Microbenchmarked Splash Attention to isolate bottlenecks, yielding a further 10% speedup on an already-tuned kernel.
- Cut Wan 2.1 397B diffusion step time 25% via Unified Sequence Parallelism (Ring Attention + DeepSpeed-Ulysses).
- Ran roofline analysis of Qwen 3.5 on next-gen TPUs to isolate compute- vs. memory-bound regimes and prioritize fixes.
- Identified and benchmarked mHC-lite optimization for DeepSeek V4, sidestepping the costly iterative Sinkhorn solve.

Quantum AI (20% project)	2025 – Present
---------------------------------	----------------

- Accelerated the Tesseract QEC decoder 2× on the d=23 surface code via various profiling-driven optimizations.
- Statistical modeling (peak-finding, Skew-Normal/GMM fits) to characterize Two-Level-System (TLS) features in qubit data.
- Researched adaptive TLS thresholding and built custom metrics to assess algorithm performance.
- Applied computer-vision techniques to identify stationary TLS features in heatmap data.

Certificate Authority Infrastructure — Google Trust Services	2019 – 2022, 2024 – 2026
---	--------------------------

- Fixed low-level CA bottlenecks: Go/C interop bugs, DB hotspot thrashing, network egress, DNS resolver latency.
- Designed load-balancing for atypical network constraints; migrated servers into secure, sensitive machine pools.
- Implemented ACME (RFC 8555), integrated performance testing frameworks, and hardened the system for public release.
- Discovered and [reported](#) a critical DoS vulnerability in Go/Crypto’s DSA implementation.
- On-call, rollouts, and outage triage; supported [WebTrust](#) compliance audits and Key Destruction Ceremonies.

Google Research — Perception & Play Abuse	2018 – 2019
--	-------------

- Built a MapReduce ML pipeline (billions of points) for image-saliency prediction on a tuned ResNet U-Net.
- Developed multimodal ML infrastructure for abuse detection in Google Play, decoupling feature generation from inference.

PRESENTATIONS

QEC: From Classical Errors to the Surface Code — Quantum AI, Google	Oct 2025
Classical and quantum error correction, from the Hamming code up to the surface code for QEC.	
Berry Phase & Chern Numbers — Master’s Presentation, TAMU	Nov 2024
Gauge invariance and numerical computation of Berry phase and Chern numbers for topological systems.	
Quantum Markov Chains — Quantum Algorithms, CSCE 640 – TAMU	Nov 2023
Szegedy’s quantization of Markov chains and its quadratic speedup in mixing time over the classical walk.	
Weyl Quantization Lecture — Topics in Physics for Mathematicians, MATH 689 – TAMU	Dec 2023
Presented the mathematical theory for converting classical phase-space $L^2(\mathbb{R}^{2n})$ -observables to be self-adjoint operators on a quantum Hilbert space.	
Visual Saliency Prediction — Perception Research Showcase, Google	Mar 2019
Presented the topic of image saliency and its motivations, and the ML infrastructure my team used for prediction and evaluation.	

PUBLICATIONS

A Chaotic Chemical Reactor With and Without Delay: Bifurcations, Competitive Modes, and Amplitude Death.
S. Roy Choudhury and Daniel Mandragona. <i>Int. J. Bifurc. Chaos</i> , 2019.

EDUCATION

M.S. Mathematics (GPA 3.93) · Texas A&M University, College Station, TX	Aug 2022 – Dec 2024
B.S. Mathematics, Computational Sciences (GPA 3.71) · University of Central Florida, Orlando, FL	Aug 2013 – May 2018