

Ganesh Ravichandran

SF Bay Area, CA | ganeshravichandran.com | github.com/TheBatmanofButler | the.ganesh.ravichandran@gmail.com

EDUCATION AND SKILLS

Columbia University, B.A. in Computer Science

Sep. 2014 – Feb. 2018

Math coursework: Calculus III, Linear Algebra, Discrete Mathematics

Post-baccalaureate math coursework:

- | | |
|--|-----------------------|
| - Real Analysis I, II (UC Santa Cruz) | Jan. 2026 – Jun. 2026 |
| - Ordinary Differential Equations for Scientists and Engineers (UMass Amherst) | May 2026 – Jun. 2026 |
| - Multivariable Calculus (UMass Amherst) | Jun. 2026 – Aug. 2026 |
| - Numerical Analysis (UC Berkeley) | Jun. 2026 – Aug. 2026 |

Technical skills:

Proficient: Python, NumPy, TypeScript/JavaScript, Kubernetes, Docker, AWS

Familiar: PyTorch, JAX, CUDA, C++, MATLAB

EXPERIENCE

Independent ML Performance and Systems Research

Oct. 2024 – Present

Conducted self-directed research into GPU architectures, Transformer optimization, and distributed training methodologies to advance understanding of ML performance bottlenecks and scalability challenges. Published in-depth technical articles on optimization techniques at thebatmanofbutler.substack.com.

Key projects include:

- Designing and assembling a **custom GPU cluster** with four NVIDIA GTX 1070 Ti GPUs with full x16 PCIe bandwidth and AMD Threadripper 2920X, optimizing for parallel training workloads.
- Implementing and **training GPT-2 locally**, leveraging PyTorch Memory Profiler for **performance analysis** and implementing activation recomputation to **reduce memory overhead during training**.
- Studying current literature on computational efficiency in deep learning, including data parallelism, ZeRO memory optimization, quantization techniques, and speculative decoding, to build **foundational knowledge in ML performance optimization**.

Senior Software Engineer at SOOT (AI copilot for visual media)

Feb. 2021 – Oct. 2024

First full-time engineering hire. Architected and maintained critical infrastructure for ML-driven applications. Built scalable systems handling hundreds of thousands of files per month and optimized distributed computing resources. Mentored junior engineers and drove technical architecture decisions.

Engineered high-performance ML pipeline architecture for production workloads:

- Built an ML-powered file ingestion pipeline leveraging CLIP embeddings, UMAP dimensionality reduction, FAISS vector search, and Stable Diffusion image generation. **Processed over 50,000 image uploads per week with 99.9% reliability.**
- Maintained and optimized a distributed architecture of Kubernetes worker pods and message queues to **promote throughput, minimize request latency, and handle expensive, concurrent ML workloads**. Implemented intelligent UMAP model updates with incremental learning capabilities to **maintain model quality at scale**.

Optimized distributed computing infrastructure for scale and performance:

- Designed and developed a highly-scalable static file serving pipeline using CloudFront, Lambda@Edge, and S3 to handle distributed data access at scale. Integrated multiple parts of our microservices architecture, including CloudFront, Lambda@Edge, S3, and custom RBAC logic. **Scaled system to handle over 300,000 file requests per week with sub-100ms latency.**
- Implemented horizontal autoscaling for Kubernetes clusters based on queue depth metrics from SQS,

optimizing resource utilization. **Achieved 2x increase in processing capacity while reducing infrastructure costs.**

- Engineered multi-resolution image serving pipeline by leveraging texture atlas compression techniques to **reduce bandwidth usage 4-5x compared to JPEG**. Supported both low-latency previews and lazy-loaded high-quality originals to **serve 10,000+ images per client per session**.
- Developed a local filesystem synchronization engine to automatically update uploaded user data based on local file changes, with the ability to **continuously watch and update over 500,000 files per user device**. This quality-of-life feature **increased average uploaded user data by 40%**.

Architected fault-tolerant systems with comprehensive observability monitoring:

- Promoted fault-tolerance through granular retry mechanisms and partial failure handling. Implemented checkpointing to preserve successful operations during batch processing, **reducing data reprocessing by 85% during failure scenarios**.
- Architected a comprehensive observability stack using OpenTelemetry and Zipkin for distributed tracing across microservices. **Improved system observability coverage by 80% and reduced mean time to detection by 25%**.
- Securely configured an Electron app update server and CI/CD pipeline with automated code signing to continuously deliver new features to ensure reliable, continuous delivery of new features to users.

Designed high-throughput API endpoints and authentication infrastructure:

- Extended GraphQL API functionality to support multiple highly scalable requests, including the implementation of a streaming architecture to **enable users to bulk download all of their account data, typically over 100,000 images**.
- Implemented authentication and authorization infrastructure integrating multiple identity providers, creating reusable libraries for service-to-service communication. Substantially reduced dev time dedicated to new auth provider integration.
- Collaborated with the design team to develop Vue components that enhanced user communication and transparency during system failures, resulting in a **90% increase in error messaging coverage**.

Software Engineer and Technical Lead at LiveStories

Sep. 2019 – Feb. 2021

Full-stack software engineer. Led development on the MVP during mid-pandemic product pivot.

Engineered core features and accelerated product performance:

- **Developed 4/7 of the MVP's critical features**, including authentication, location-based search, and chart editing.
- Rearchitected the search function for the data library and **reduced load time by 40%**.
- Shipped an email campaign launch platform for government clients to reach businesses based on their eligibility for a given aid program, used by more than half of our clients.
- Built tabular React UIs for government clients to analyze aid eligibility for over 10,000 businesses.
- Developed and maintained UIs to empower struggling businesses to apply to more than 130 municipal and federal aid programs and gain access to **over \$600,000 in available pandemic relief**.

Orchestrated MVP Launch after strategic product redirection during COVID:

- **Managed three engineers across three continents** to develop and launch an MVP to replace the legacy app.
- Advised CEO on design/development strategy and cut sprint planning meeting lengths in half.
- Facilitated transition for existing users by integrating new MVP into the existing product UX.