

# Pulkit Goel

Tech Lead | ML Infrastructure & Large-Scale Systems

@pulkitg10@gmail.com  
Mountain View, CA  
linkedin.com/in/pulkitg10  
+16506600118

## EXPERIENCE

Tech Lead/Staff SWE, Search ML Infrastructure

Google

📅 2023 – Present

📍 Mountain View, CA

- Drove strategic direction and architected a production-scale LLM inference framework, adopted by 150+ client teams and supporting over 2500 QPS, significantly accelerating product development and scaling ML capabilities across Google Search.
- Directly enabled the launch of critical features like AI Overviews and numerous new experiments/launches for Google Search by increasing offline TPU duty cycle to >70%—a crucial launch requirement that minted significant new capacity equivalent to tens of thousands of chips (O(10,000s)) from existing resources—through novel traffic optimization and hardware fungibility, despite no new TPU grants since 2024.
- Championed and successfully launched multiple Google-wide infrastructure optimizations, partnering with Flume & LLM Deployment teams to define OKRs and steer progress, resulting in lower cost-per-million-tokens and higher duty cycle.
- Served as a lead subject matter expert, providing dedicated technical support to over 200 engineers, which accelerated successful platform adoption and unblocked critical development efforts for numerous client teams.

Senior Software Engineer, Search Infrastructure

Google

📅 2020 – 2023

📍 Mountain View, CA

- Co-designed and drove development of APIMirror, a secure gateway that dramatically reduced integration time with 3P APIs for products like Gemini Extensions, now connecting 20 3Ps.
- Led a team to build 3PDB, a scalable C++ platform ingesting 30 terabytes of structured data daily from thousands of 3P providers, now a Google-wide standard.

Software Engineer, Search Data Ingestion Infrastructure

Google

📅 2015 – 2020

📍 Mountain View, CA

- Architected and scaled a central C++ backend for feed serving (powering Google News, Blogger, etc.), handling over 100,000 QPS at its peak.
- Built the first backend for the public Google Indexing API, enabling partners to push data into Google's index with extremely low latency.

Software Engineer Intern, Product Quality

Facebook

📅 2014

📍 Menlo Park, CA

## ABOUT ME

*"Staff Engineer combining a decade of systems expertise with a passion for ML. I build highly efficient, next-generation platforms to tackle core ML challenges and deployment bottlenecks."*

## TECHNICAL SKILLS

ML & Data Infrastructure

Large-Scale Distributed Systems

ML Workload Optimization

LLM Inference & Serving

API Design

## PERSONAL PROJECTS

JeopardyGPT 🔄

- Architected & built a real-time, multiplayer web game using Angular, Firebase/Firestore, and a custom Colyseus.js backend, with game logic powered by LLMs.

## EDUCATION

B.Tech (Hons) & MS by Research, CS  
International Institute of Information Technology (IIIT)

📅 2010 – 2015

📍 Hyderabad, India

## HONORS & AWARDS

🏆 ACM-ICPC World Finalist 🔄  
2014 & 2015

👥 23 Peer Bonuses  
Awarded in 1 year, for impact and collaboration on ML infrastructure.

## PUBLICATIONS

- P. Goel, "Load balancing on the cloud," M.S. thesis, International Institute of Information Technology, Hyderabad, 2017.
- P. Goel, K. Rishabh, and V. Varma, "An alternate load distribution scheme in dhds," in 2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom), IEEE, 2017, pp. 218–222.
- P. Goel, L. Srinivasan, and V. Varma, "An adaptive routing scheme for heterogeneous dataflows using openflow," in 2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), IEEE, 2015, pp. 52–58.