

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

- Set team direction (OKRs, prioritization) and spearheaded the architecture of a production-scale LLM inference framework, adopted by 150 client teams and supporting over 2500 QPS across diverse models.
- Enabled the launch of critical features like **AI Overviews** by increasing offline TPU duty cycle to >70%—a crucial launch requirement that saved O(10k)s chips—through novel traffic optimization and hardware fungibility.
- Influenced core infrastructure by collaborating with Google's Flume and LLM Deployment teams by proposing and getting buy-in for Google-wide optimization; also managed support for over 180 client engineers.

Senior Software Engineer, Search Infrastructure

Google

📅 2020 – 2023

📍 Mountain View, CA

- Co-designed and drove development of APIMirror, a service enabling products like Gemini Extensions to securely interact with third-party APIs.
- Led a team to build 3PDB, a scalable C++ platform ingesting 30 terabytes of structured data daily from thousands of third-party providers, now a Google-wide standard.

Software Engineer, Data Ingestion Infrastructure

Google

📅 2015 – 2020

📍 Mountain View, CA

- Built and scaled numerous large-scale C++ distributed systems, including the first backend for the public Google Indexing API and the serving infrastructure for Google News.

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 data-flows using openflow," in *2015 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, IEEE, 2015, pp. 52–58.

ABOUT ME

"A Staff Engineer combining a decade of systems expertise with a long-standing passion for machine learning. Two years ago, I formally pivoted into a challenging ML infrastructure role, successfully applying years of self-directed learning on ML models and deployment bottlenecks. My goal is to tackle core ML challenges by building highly efficient, next-generation platforms."

TECHNICAL SKILLS

ML & Data Infrastructure

Large-Scale Distributed Systems

ML Workload Optimization

LLM Inference & Serving

API Design

PERSONAL PROJECTS

JeopardyGPT 🔗

- Built a complex, multiplayer web game using 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
Received last year for impact and collaboration on ML infrastructure.

INTERNSHIPS

Software Engineer Intern

Facebook

📅 2014

📍 Menlo Park, CA

Software Engineer Intern

Directi

📅 2013

📍 Mumbai, India