

# Sudhakar Chundu

📍 San Jose 📞 +1 513-666-0099 ✉ [chundubabu@gmail.com](mailto:chundubabu@gmail.com) 💻 Senior Site Reliability Engineer 🌐 [linkedin.com/in/schundu](https://www.linkedin.com/in/schundu)

## SUMMARY

Strategic engineering leader with 20+ years building AI/ML infrastructure and optimizing GPU clusters. Expertise in Kubernetes operations and Infrastructure-as-Code for high-performance networking. Proven record in reducing costs by \$8M+ while maintaining 99.95%+ SLAs. Skilled in implementing observability stacks and CI/CD pipelines, enhancing DevOps efficiency.

## EXPERIENCE

Senior Cloud Architect - AI Infrastructure & MLOps | Trackonomy Systems

10/2023 to Present

- Architected serverless GPU infrastructure reducing inference costs by 65% serving 5M+ daily predictions with 99.97% uptime
- Deployed multi-platform inference across 15+ edge locations supporting NVIDIA CUDA and Apple MLX
- Built AI monitoring system tracking inference latency, GPU utilization using Prometheus and Grafana
- Established edge AI framework with K3s and NVIDIA Jetson devices supporting operations in 8 countries
- Implemented GPU scheduling and load balancing across heterogeneous clusters
- Engineered automated ML pipelines reducing deployment time from days to 15 minutes
- Reduces cloud costs from \$10.8M to \$2.2M through GPU optimization
- Developed Infra Applications managing tasks to reduce operational overhead and security compliance issues

Site Reliability Engineer - ML Infrastructure | Amazon Web Services (via Wipro)

02/2022 to 10/2023

- Delivered GPU cluster management with Kubernetes ensuring 99.95% availability for ML workloads
- Built observability stack integrating Prometheus, Grafana with GPU metrics
- Implemented MLOps practices: model versioning, automated retraining for 30+ models
- Created automation pipelines reducing manual operations by 85%
- Deployed GitOps-based releases supporting 50+ daily deployments with zero downtime

Cloud Infrastructure Engineer - Azure ML | Microsoft (via Wipro)

02/2020 to 02/2022

- Engineered AI infrastructure using Azure ML and AKS with GPU nodes
- Implemented Infrastructure-as-Code improving deployment speed by 80%
- Integrated observability for ML workloads with custom GPU metrics
- Delivered compliance-as-code pipelines for various requirements

Cloud Architect - Healthcare AI Infrastructure | Harvard Pilgrim Health Care

06/2018 to 02/2020

- Led cloud modernization for HIPAA-regulated AI/ML applications to AWS, implementing GPU-enabled clusters
- Designed ML pipeline with end-to-end encryption and audit logging
- Implemented observability reducing false alerts by 60%

Senior Infrastructure Architect | Tata Consultancy Services (TCS)

05/2005 to 06/2018

- Designed infrastructure for global clients in government, telecom, and financial sectors
- Deployed early container-first architecture with Kubernetes clusters
- Built automation-first CI/CD practices improving application delivery

## EDUCATION

SKILLS

AI/ML Infrastructure

GPU Computing

Edge AI

Multi-Platform

AI Monitoring

Container Orchestration