



# Sudhakar Chundu

## AI Infrastructure Engineer

chundubabu@gmail.com  
San Jose - +1 513-666-0099  
[linkedin.com/in/schundu](https://linkedin.com/in/schundu)

Strategic engineering leader with 20+ years of experience in building robust AI/ML infrastructures, specializing in GPU orchestration and model serving platforms. Expert in serverless GPU computing, edge AI deployment, and ensuring multi-platform inference (NVIDIA, MLX). Demonstrated ability to deliver 99.95%+ SLAs for ML workloads while achieving cost reductions of over \$8M through intelligent optimization.

---

## EXPERIENCE

### Senior Cloud Architect - AI Infrastructure & MLOps

*Trackonomy Systems*

10/2023 - Present

- Architected serverless GPU infrastructure using Google Cloud Run and AWS Lambda containers, reducing inference costs by 65% while serving 5M+ daily predictions at 99.97% uptime
- Deployed multi-platform inference pipeline supporting NVIDIA CUDA, Apple MLX, and CPU fallback across 15+ edge locations with intelligent workload routing
- Built real-time AI monitoring system tracking inference latency (p50/p95/p99), GPU utilization, model drift, and throughput using Prometheus, Grafana, and custom Python exporters
- Established edge AI framework for low-connectivity environments using K3s and NVIDIA Jetson devices with offline model sync, supporting operations in 8 countries
- Implemented container orchestration for AI workloads with GPU scheduling, KEDA autoscaling, and load balancing across heterogeneous clusters (A100, V100, T4)
- Engineered automated ML pipelines using ArgoCD and GitOps, reducing deployment time from days to 15 minutes

### Site Reliability Engineer - ML Infrastructure

*Amazon Web Services (via Wipro)*

02/2022 - 10/2023

- Delivered GPU cluster management at scale with Kubernetes, Istio, and KEDA, ensuring 99.95% availability for ML inference workloads
- Built custom observability stack for AI workloads integrating Prometheus, Grafana, CloudWatch with GPU metrics and model performance dashboards
- Implemented MLOps best practices: model versioning, automated retraining, blue-green deployments for 30+ production models
- Created automation pipelines in Go/Python reducing manual operations by 85%, including GPU provisioning and model deployment

## **Cloud Infrastructure Engineer - Azure ML**

02/2020 - 02/2022

*Microsoft (via Wipro)*

- Engineered scalable AI infrastructure using Azure ML, AKS with GPU node pools, ExpressRoute for hybrid connectivity
- Implemented IaC for ML pipelines using Terraform and Azure Bicep, improving deployment speed by 80%
- Integrated Azure Monitor, Prometheus, Grafana for ML workloads with custom GPU and inference metrics
- Delivered compliance-as-code pipelines for SOC 2, HIPAA, ISO 27001 requirements

## **Cloud Architect - Healthcare AI Infrastructure**

06/2018 - 02/2020

*Harvard Pilgrim Health Care*

- Led cloud modernization for HIPAA-regulated AI/ML applications to AWS, implementing GPU-enabled EKS clusters
- Designed HIPAA-compliant ML pipeline with end-to-end encryption, audit logging, secure model serving
- Implemented observability using ELK, Prometheus, Grafana, reducing false alerts by 60%

## **Senior Infrastructure Architect**

05/2005 - 06/2018

*Tata Consultancy Services (TCS)*

- Designed enterprise infrastructure for global clients (government, telecom, financial services)
- Deployed early container-first architecture with production Kubernetes clusters and SDN
- Built automation-first CI/CD practices enabling application delivery at scale

## **EDUCATION**

### **Bachelor of Engineering**

*Acharya Nagarjuna University*

Studied core engineering principles with a focus on AI/ML technologies and cloud infrastructure.

## **SKILLS**

**AI/ML Infrastructure**

**GPU Computing**

**Edge AI**

**Multi-Platform**

**AI Monitoring**

**Container  
Orchestration**