

# Sudhakar Chundu

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

<b>Senior Cloud Architect - AI Infrastructure &amp; MLOps   Trackonomy Systems</b>	<b>10/2023 to Present</b>
<ul style="list-style-type: none"><li>Architected serverless GPU infrastructure reducing inference costs by 65% serving 5M+ daily predictions with 99.97% uptime</li><li>Deployed multi-platform inference across 15+ edge locations supporting NVIDIA CUDA and Apple MLX</li><li>Built AI monitoring system tracking inference latency, GPU utilization using Prometheus and Grafana</li><li>Established edge AI framework with K3s and NVIDIA Jetson devices supporting operations in 8 countries</li><li>Implemented GPU scheduling and load balancing across heterogeneous clusters</li><li>Engineered automated ML pipelines reducing deployment time from days to 15 minutes</li><li>Reduces cloud costs from \$10.8M to \$2.2M through GPU optimization</li><li>Developed Infra Applications managing tasks to reduce operational overhead and security compliance issues</li></ul>	
<b>Site Reliability Engineer - ML Infrastructure   Amazon Web Services (via Wipro)</b>	<b>02/2022 to 10/2023</b>
<ul style="list-style-type: none"><li>Delivered GPU cluster management with Kubernetes ensuring 99.95% availability for ML workloads</li><li>Built observability stack integrating Prometheus, Grafana with GPU metrics</li><li>Implemented MLOps practices: model versioning, automated retraining for 30+ models</li><li>Created automation pipelines reducing manual operations by 85%</li><li>Deployed GitOps-based releases supporting 50+ daily deployments with zero downtime</li></ul>	
<b>Cloud Infrastructure Engineer - Azure ML   Microsoft (via Wipro)</b>	<b>02/2020 to 02/2022</b>
<ul style="list-style-type: none"><li>Engineered AI infrastructure using Azure ML and AKS with GPU nodes</li><li>Implemented Infrastructure-as-Code improving deployment speed by 80%</li><li>Integrated observability for ML workloads with custom GPU metrics</li><li>Delivered compliance-as-code pipelines for various requirements</li></ul>	
<b>Cloud Architect - Healthcare AI Infrastructure   Harvard Pilgrim Health Care</b>	<b>06/2018 to 02/2020</b>
<ul style="list-style-type: none"><li>Led cloud modernization for HIPAA-regulated AI/ML applications to AWS, implementing GPU-enabled clusters</li><li>Designed ML pipeline with end-to-end encryption and audit logging</li><li>Implemented observability reducing false alerts by 60%</li></ul>	
<b>Senior Infrastructure Architect   Tata Consultancy Services (TCS)</b>	<b>05/2005 to 06/2018</b>
<ul style="list-style-type: none"><li>Designed infrastructure for global clients in government, telecom, and financial sectors</li><li>Deployed early container-first architecture with Kubernetes clusters</li><li>Built automation-first CI/CD practices improving application delivery</li></ul>	

## EDUCATION

## **SKILLS**

---

**AI/ML Infrastructure**

**GPU Computing**

**Edge AI**

**Multi-Platform**

**AI Monitoring**

**Container Orchestration**