Sudhakar Chundu

San Jose +1 513-666-0099 chundubabu@gmail.com Site Reliability Engineer (SRE) linkedin.com/in/schundu

SUMMARY

Site Reliability Engineer with over 20 years of experience in building AI/ML infrastructure and distributed systems. Expertise in automation, cloud services, and Kubernetes for efficient deployment and orchestration. Proven track record of delivering high performance with 99.95% SLAs and cost reductions up to $8M+ through performance tuning and optimization. Skilled in incident response and conducting blameless postmortems to enhance reliability.

EXPERIENCE

Senior Cloud Architect - AI Infrastructure & MLOps | Trackonomy Systems **10/2023 to 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, aligned with Site Reliability Engineering best practices
- Established edge AI framework using K3s and NVIDIA Jetson devices, supporting offline operations in 8 countries
- Implemented container orchestration with GPU scheduling, KEDA autoscaling, and load balancing across heterogeneous clusters (A100, V100, T4)
- Reduced cloud costs from $10.8M to $2.2M through GPU optimization and caching
- Developed Infra Applications reducing manual work and improving security compliance through automation

Site Reliability Engineer - ML Infrastructure | Amazon Web Services (via Wipro) **02/2022 to 10/2023**

- Managed GPU clusters at scale with Kubernetes and KEDA, ensuring 99.95% availability for ML inference workloads
- Developed observability stack with Prometheus, Grafana, and CloudWatch to monitor GPU metrics and model performance
- Established MLOps best practices for model versioning and blue-green deployments for 30+ models
- Automated pipelines in Go/Python reducing manual operations by 85%
- Executed GitOps-based deployments across 15+ regions, supporting 50+ daily deployments with zero downtime

Cloud Infrastructure Engineer - Azure ML | Microsoft (via Wipro) **02/2020 to 02/2022**

- Engineered scalable AI infrastructure using Azure ML and AKS, ensuring effective GPU utilization and hybrid connectivity via ExpressRoute
- Implemented infrastructure as code (IaC) pipelines with Terraform and Azure Bicep, enhancing deployment speed by 80%
- Integrated custom monitoring using Prometheus and Grafana for GPU and inference metrics
- Delivered compliance-as-code solutions for SOC 2, HIPAA, ISO 27001 requirements

Cloud Architect - Healthcare AI Infrastructure | Harvard Pilgrim Health Care **06/2018 to 02/2020**

- Led cloud modernization for HIPAA-compliant AI/ML workload deployment to AWS, leveraging GPU-enabled EKS clusters
- Designed secure ML pipelines with encryption, audit logging, and secure model serving
- Implemented observability using ELK Stack and Prometheus, reducing false alerts by 60%

Senior Infrastructure Architect | Tata Consultancy Services (TCS) **05/2005 to 06/2018**

- Developed enterprise infrastructure strategies for global clients across multiple industries
- Pioneered container-first architectures with production Kubernetes deployments
- Advanced CI/CD automation improving application delivery efficiency

EDUCATION

Bachelor of Engineering | Acharya Nagarjuna University

SKILLS

AI/ML Infrastructure

GPU Computing

Edge AI

Multi-Platform

AI Monitoring

Container Orchestration