Sudhakar Chundu

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

SUMMARY

Accomplished Site Reliability Engineer specializing in multi-cloud infrastructure and GPU computing with over 13 years of experience. Expertise in designing scalable AI and ML platforms, optimizing distributed systems, and enabling automation through Kubernetes and NeoCloud. Proven success in reducing costs by 73% and maintaining 99.97% uptime through advanced observability and automation strategies.

EXPERIENCE

Distinguished Cloud AI Architect (HPC Platform Lead) | Trackonomy Systems **10/2023 to 12/2025**

- Designed, deployed, and operated Slurm-based GPU compute platform — 65 GPUs across 8 nodes using Slurm, Slinky (Slurm-on-Kubernetes), and NVIDIA BCM; achieved 99.97% uptime serving 12+ enterprise clients for AI inference and training workloads
- Integrated Slurm with Kubernetes orchestration layer; implemented hybrid scheduling, unified authentication, and telemetry pipelines using vcluster for multi-tenant GPU workload isolation
- Optimized cluster utilization and scheduling — developed fair-share policies, QoS configurations, and preemption strategies for GPU and CPU workloads across enterprise clients
- Built platform features improving developer experience — Python-based job submission APIs, automated environment setup, GPU utilization tracking dashboards, and internal tooling for job scheduling optimization
- Implemented observability pipelines using Prometheus, Grafana, and custom exporters; unified 5 monitoring platforms into Datadog for consolidated APM, logs, metrics, and distributed tracing
- Supported full job lifecycle — from container image builds and environment configuration to debugging and performance tuning of Slurm jobs; implemented automated incident remediation
- Hands-on redesign of multi-cloud architecture (Azure, AWS) reducing services from 27→9 (67%); migrated VMs to AKS/EKS with Helm and GitOps workflows
- Built real-time IoT stream processing using Apache Flink; implemented Databricks platform (Jobs, Compute, Unity Catalog) for ML training and analytics workflows
- Reduced infrastructure costs 73% ($10M→$2.7M) through GPU utilization optimization, fair-share scheduling, and vendor consolidation; cut deployment time 95% via CI/CD automation
- Led SRE L3 operations — root cause analysis (RCA), incident management, blameless postmortems, SLI/SLO definition, error budget tracking, on-call rotations, and runbook development for production GPU clusters
- Architected data pipelines with Apache Kafka for real-time event streaming; implemented Delta Lake for ACID transactions and data versioning on cloud storage (S3, Azure Blob)

Cloud & SRE Architect | OSDU Data Platform **08/2018 to 10/2023**

- Hands-on architecture of scalable GPU/ML pipelines: sensor data → Kafka ingestion → validation → feature engineering → model training → Databricks BI; achieved 99.99% availability
- Designed EKS/Fargate clusters for ML workloads; created 50+ Terraform modules (VPC, EKS, RDS, DynamoDB, S3, SageMaker, IAM); implemented self-healing auto-scaling reducing downtime 75%
- Built CI/CD pipelines for 10+ microservices using GitHub Actions/ArgoCD; integrated MLflow for model versioning and experiment tracking; reduced deployment time 80%
- Implemented multi-tenant environments using AWS Organizations, SCPs, and Terraform workspaces for isolated ML training across 4+ enterprise clients
- Built Prometheus/Grafana/ELK observability stack improving issue resolution 50%; production troubleshooting with TBBT methodology; DR with RPO<1hr, RTO<4hrs
- Performed SRE L3 duties — conducted RCA for critical incidents, led blameless postmortems, defined SLIs/SLOs for platform services, managed error budgets, maintained on-call rotations, authored operational runbooks
- Implemented Apache Kafka for real-time data ingestion; built Delta Lake layers on S3 for data lakehouse architecture with ACID compliance and schema evolution

Multi Cloud Architect | Tata Consultancy Services **05/2007 to 08/2018**

- 5 years hands-on Linux/Unix systems administration — WebSphere, WebLogic middleware, kernel tuning, performance optimization, high-availability configurations across CNA, PWC, Verizon
- Migrated legacy systems to AWS, Azure, OpenStack; implemented CI/CD with Jenkins/Ansible; managed large-scale distributed deployments across multi-cloud environments
- L3 support and operations — incident management, root cause analysis, performance troubleshooting, capacity planning, on-call support, network and storage troubleshooting for enterprise applications

EDUCATION

Bachelor of Engineering | Acharya Nagarjuna University

Engineering bachelor’s program focusing on foundational principles applicable to reliability engineering and infrastructure optimization.

SKILLS

HPC & GPU Computing: Slurm, Slinky (Slurm-on-Kubernetes), NVIDIA BCM, vcluster, CUDA, GPU scheduling & orchestration, fair-share policies, QoS, preemption, job lifecycle management

SRE L3 Operations: Root cause analysis (RCA), incident management, blameless postmortems, SLI/SLO/SLA management, error budgets, capacity planning, on-call rotations, runbook development, chaos engineering

System Design: Scalability (horizontal/vertical), high availability (HA), fault tolerance, load balancing, caching strategies, database sharding, CAP theorem, microservices, event-driven architecture, API design, rate limiting, circuit breakers, distributed transactions, service mesh, data replication, disaster recovery

Linux & Systems: Unix/Linux internals, systemd, kernel tuning, environment modules, performance tuning, troubleshooting compute/storage/network layers

Container Orchestration: Kubernetes (EKS, AKS), Docker, Helm, container image builds, hybrid scheduling (Slurm + K8s integration)

Cloud Platforms: Azure (AKS, VMs, Event Hub), AWS (EKS, EC2, Fargate, SageMaker), OpenStack, multi-cloud GPU deployments