Srinivas Ritvik Chemudupati

(404) 996-9975 | ritvik.chemudupati@gmail.com | linkedin.com/in/ritvikcs | ritvikchemudupati.com

Summary

Platform engineer with 3 years at Deloitte building production Kubernetes infrastructure on AWS — deploying GPU-accelerated inference pipelines, designing self-service CI/CD platforms, operating multi-cluster environments at scale, and hands-on production incident response. Proven impact: 20% infrastructure cost reduction (∼$600K annually) supporting 75+ data scientists. Graduated May 2026 with MS in Computer Science (AI) from Georgia Tech.

Experience

Platform Engineer July 2022 – July 2025
Deloitte *Hyderabad, India*

- Reduced **Kubernetes** infrastructure costs by 20% (∼$600K annually) by consolidating from 7–8 to 3–4 clusters per account, rightsizing pods, tightening autoscaling bounds, culling idle notebooks, and removing orphaned **EBS** volumes and **ECR** images across multiple AWS accounts using **AWS Cost Explorer** for spend analysis.
- Deployed **NVIDIA Morpheus** on **AWS EKS** as sole owner of a production GPU inference platform serving 3 teams — configuring GPU scheduling via taints, tolerations, and node selectors (g5 over p3), **IRSA** for pod-level AWS credential injection, and **Lambda**-to-EKS authentication via **aws-auth** ConfigMap. Built **Ansible** automation for idempotent redeployment across clusters.
- Built self-service CI/CD platform for ML model deployments on **KServe** — eliminating manual platform engineer involvement via **GitHub OIDC** federation with AWS for keyless **ECR** push and **ArgoCD Image Updater** for automatic rollouts. Platform scaled to 75+ data scientists across teams, reaching VPC IP exhaustion under load.
- Operated **Kubeflow** on **AWS EKS** and contributed to recovery of a full production outage — manually reconstructing cluster state without `terraform apply` to avoid state mismatch, restoring AZ-specific **EBS** volumes by known notebook size, creating **Kubernetes** Jobs for AZ-targeted reattachment, and recovering data via Python **S3** sync scripts.
- Provisioned **EKS** infrastructure using **Terraform** — node groups, GPU (p3/g5) and CPU (m5) instance types, node labels for workload isolation, and cluster configuration across staging and production environments.
- Deployed **Robust Intelligence** AI Firewall in a hybrid architecture — resolving production networking failures including AWS hairpin NAT limitations, ALB annotations, and security group rules to connect the vendor control plane to on-cluster services.
- Implemented observability using **Prometheus** and **Grafana** across multiple **Kubernetes** clusters, building custom dashboards from pod logs and metrics for resource utilization and pipeline health.

Education

Georgia Institute of Technology Atlanta, GA
Master of Science in Computer Science – Specialization: Artificial Intelligence *Graduated May 2026*

Birla Institute of Technology and Science, Pilani Hyderabad, India
Bachelor of Engineering in Computer Science *2018 – 2022*

Projects

Bare-Metal Kubernetes Platform | *Kubernetes, k3s, ArgoCD, Prometheus, Grafana, Arch Linux* Ongoing

- Designing and deploying a self-managed bare-metal Kubernetes cluster using **k3s** on Arch Linux, replicating on-premises deployment patterns without cloud abstractions.
- Configuring **GitOps** workflows with **ArgoCD** and observability with **Prometheus** and **Grafana**, extending production operational patterns to edge and on-premises conditions.

Technical Skills

Languages: Python, Bash, SQL, C++
Cloud Platforms: AWS (EKS, EC2, S3, IAM, Lambda, ECR, SageMaker), Azure
Container Orchestration: Kubernetes, Docker, Helm, Kustomize, Knative
CI/CD & GitOps: GitHub Actions, ArgoCD, ArgoCD Image Updater, Git
ML Platforms: Kubeflow, KServe, NVIDIA Morpheus, MLflow, Hopsworks, PyTorch
Infrastructure as Code: Terraform, Ansible
Monitoring & Observability: Prometheus, Grafana, CloudWatch