

ASHISH KUMAR

Lead DevOps & MLOps Engineer | Cloud & AI Platform Architect

Phone: +91-9411301729 • **Email:** ak@aum.onl • **Portfolio:** ak.aum.onl • **Location:** Ramgarh, Jharkhand, India (Remote Available)

PROFESSIONAL SUMMARY

Accomplished **Cloud Platform & MLOps Architect** with **16+ years of total experience** driving scalable enterprise Linux systems, multi-cloud platforms (AWS, Azure, GCP), and enterprise Kubernetes orchestration. Over the **past 3+ years**, extensively focused on **AI/ML Infrastructure, MLOps, and LLMOps**—architecting GPU-accelerated Kubernetes clusters (EKS/GKE with NVIDIA GPU Operator & Karpenter), orchestrating distributed model training and fine-tuning pipelines (Kubeflow, Ray), deploying low-latency LLM inference architectures (Triton, vLLM, Ollama), and scaling vector databases (Milvus, Qdrant) for enterprise GenAI. Certified Kubernetes Administrator (**CKA**) and HashiCorp Certified Terraform Associate with a proven track record of reducing GPU spend by 35%+, achieving 99.99% system availability, and accelerating delivery cycles through AI-augmented workflows.

CORE TECHNICAL COMPETENCIES

- AI/MLOps & AI Infrastructure:** Kubeflow, Ray & Ray Serve, Triton Inference Server, vLLM, Ollama, MLflow, NVIDIA GPU Operator, Multi-Instance GPU (MIG), Distributed Training & Fine-Tuning Pipelines, Hugging Face, Model Packaging & Quantization.
- GenAI & Vector Databases:** Milvus, Qdrant, OpenSearch Vector Engine, Pinecone, Retrieval-Augmented Generation (RAG) Architecture, Semantic Prompt Caching (Redis), LLM Observability & Guardrails (Langfuse, OpenTelemetry).
- Cloud Platforms:** Amazon Web Services (AWS – 12 yrs), Google Cloud Platform (GCP – 7 yrs), Microsoft Azure (7 yrs), Hybrid Cloud & Bare-Metal (OpenStack, VMware vSphere/ESXi).
- Container Orchestration & GitOps:** Kubernetes / EKS / GKE / AKS (8 yrs – CKA Certified), Docker, Podman, Helm (7 yrs), ArgoCD (GitOps – 5 yrs), Kustomize, Service Mesh (Istio), OpenShift.
- IaC & Configuration Management:** Terraform (11 yrs – HashiCorp Certified), Terragrunt, Packer (11 yrs), Ansible & AWX (12 yrs), Puppet, Chef, Bash Scripting (15 yrs), Python Automation & AI Pipelines (8 yrs).
- CI/CD & DevSecOps:** Jenkins (10 yrs), GitOps, Screwdriver CI, GitLab CI, GitHub Actions, Snyk, SonarQube, Trivy, OWASP ZAP, HashiCorp Vault, AWS IAM / Least Privilege, CIS Benchmarks, PCI-DSS.
- Observability & SRE:** Prometheus, Grafana, Datadog, Splunk (5 yrs), New Relic, Wavefront, ELK Stack, Nagios/Icinga (14 yrs), Zabbix, Alertmanager, PagerDuty, SLO/SLI tracking, Anomaly Detection.
- Systems, Networks & Databases:** Enterprise Linux (RHEL, CentOS, Rocky, Ubuntu, Debian – 16 yrs), TCP/IP, DNS, IPtables / Security Groups (16 yrs), Cisco & Fortinet Firewalls, CDN; PostgreSQL, MySQL, YugabyteDB, MongoDB, DynamoDB, Redis.
- AI-Augmented Engineering Tools:** Cursor, GitHub Copilot, custom LLM incident triage bots, AI-assisted code generation & automated root cause analysis.

PROFESSIONAL EXPERIENCE

Lead DevOps & MLOps Engineer / Sr. Cloud & AI Infra Specialist | ProvaanTech Pvt Ltd May 2024 – Present | Remote, Pune

- Tech Stack:** AWS (EKS, EC2, S3, IAM, Lambda), GCP (GKE), NVIDIA GPU Operator, vLLM, Triton, Ray/Ray Serve, Kubeflow, Milvus, Terraform, Helm, ArgoCD, Ansible, Python, Prometheus, Grafana, Snyk.
- AI/ML Infrastructure & GPU Orchestration:** Architected multi-cloud GPU-accelerated Kubernetes clusters on AWS EKS and GCP GKE via Terraform & Helm, configuring NVIDIA GPU Operator and Multi-Instance GPU (MIG) slicing for A100/H100 & L4 workloads.
- Low-Latency LLM Serving:** Engineered production-grade model serving using vLLM, Triton Inference Server, and Ray Serve, achieving sub-50ms Time-to-First-Token (TTFT) and boosting inference throughput by 40%.
- GPU Cost Optimization:** Automated dynamic GPU autoscaling using Karpenter and Spot instance orchestration, slashing enterprise accelerator cloud compute expenses by 35%.
- Vector Database & RAG Platform:** Deployed and managed distributed Milvus and OpenSearch Vector Engine clusters on Kubernetes to support enterprise RAG pipelines with sub-second vector search indexing.
- GitOps & Continuous Delivery:** Standardized end-to-end GitOps deployment pipelines via ArgoCD, Helm, and Kustomize for automated model versioning, continuous canary rollouts, and zero-downtime microservice deployments.
- AI Observability & DevSecOps:** Built Prometheus and Grafana dashboards tracking GPU memory pressure, temperature, token throughput, and model drift; automated container and model security scanning with Snyk, Trivy, and SonarQube.
- AI-Assisted Incident Triage:** Developed Python LLM incident triage bots and Slack integrations for automated anomaly analysis, reducing Mean Time to Resolution (MTTR) by 45%.

Production Engineer II (IC Level 3) – Cloud & AI/Data Platform Ops | Yahoo Software India Apr 2023 – Apr 2024 | Remote, Bengaluru

- Tech Stack:** Linux (CentOS/RHEL), AWS, OpenStack Bare-Metal, Kubernetes, Helm, ArgoCD, Screwdriver CI, Ansible AWX, Terraform, Packer, Python, Splunk, Nagios, MySQL, Traffic Server (CDN).
- Large-Scale High-Throughput Reliability:** Owned L3 production reliability engineering for distributed data platforms and media services handling multi-million daily requests across hybrid cloud environments (AWS & OpenStack bare-metal).
- HPC & Linux Kernel Tuning:** Engineered OS-level network stack, memory, and kernel optimizations on bare-metal and virtualized hosts, significantly reducing latency and packet drops for distributed data streaming and batch jobs.
- Cloud Migration & IaC Automation:** Led migration of legacy bare-metal workloads to AWS using automated Terraform modules, Packer automated golden image pipelines, and Ansible AWX configuration management.
- CI/CD & Automated Patching:** Built automated build/release pipelines with Screwdriver CI and ArgoCD on Kubernetes; automated zero-downtime security patching across 3,000+ Linux instances using Ansible AWX.
- Incident Automation:** Developed custom Python telemetry correlation scripts and Slack bot integrations with Splunk and Nagios for real-time automated outage diagnosis and health checks.

DevOps & Cloud Infrastructure Lead Kloud9 Technologies India		Feb 2022 – Oct 2022 Remote, Bengaluru
Tech Stack: AWS, Azure, GCP, Linux, Kubernetes (EKS/AKS), Terraform, Packer, Ansible, Helm, Jenkins, GitOps, Python, Bash, New Relic, Splunk, MongoDB, DynamoDB, AWS Lambda, Dynamics 365.		
Engineering Leadership: Led a team of 11 DevOps and SRE engineers across multi-cloud infrastructure deliverables; conducted agile sprint planning, architectural reviews, and technical mentorship. Multi-Cloud IaC Standardization: Architected reusable Terraform and Packer automation frameworks across AWS, GCP, and Azure, reducing infrastructure provisioning turnaround from days to under 30 minutes. Serverless & Data Pipeline Architecture: Engineered serverless microservices using AWS Lambda, API Gateway, and DynamoDB for event-driven workflows and customer analytics pipelines in enterprise e-commerce platforms (Shopify, Magento). Enterprise Systems & Observability: Managed high-availability deployment of Microsoft Dynamics 365; standardized CI/CD pipelines via Jenkins & GitOps; implemented observability in New Relic & Splunk, maintaining 99.95% uptime.		
Sr. Technical Support Engineer (Kubernetes & Cloud Infra) YugaByte India		Mar 2021 – Nov 2021 Remote, Bengaluru
Tech Stack: YugabyteDB, Linux, AWS, GCP, Azure, Kubernetes, Helm, Terraform, Packer, Ansible, Python, Git, Zendesk, PagerDuty.		
- Automated deployment of distributed YugabyteDB clusters across AWS and GCP using Terraform modules, Packer, and Ansible. - Engineered and optimized Kubernetes Helm charts for distributed database orchestration, fault tolerance, and multi-region replication. - Troubleshoot critical customer issues spanning Raft consensus, query performance, and storage subsystem bottlenecks. - Integrated YugabyteDB with enterprise client workloads and high-throughput e-commerce platforms.		
Site Reliability Engineer (Observability & Telemetry) VMware Software India		Jul 2020 – Mar 2021 Remote, Pune
Tech Stack: AWS, Terraform, Linux, Python, AWS Lambda, Wavefront (Tanzu Observability), logz.io, Slack API.		
- Spearheaded enterprise observability migration from logz.io to VMware Wavefront across multi-account AWS environments. - Built serverless log aggregation and metrics harvesting pipelines utilizing Python and AWS Lambda. - Automated incident alerting pipelines from Wavefront directly to Slack channels, reducing alert latency by 60%.		
Technical Lead (Cloud Operations & SRE) elfnfochips		Feb 2018 – Jan 2020 Pune
Tech Stack: Linux, AWS, GCP, Azure, Docker, Kubernetes (EKS & Azure VMs), Terraform, Packer, Ansible, Jenkins, Zabbix.		
- Led a team of 16 DevOps/SRE engineers; architected multi-cloud infrastructure on AWS, Azure, and GCP; achieved 25% AWS cost savings. - Automated IAM key rotation and Route53 DNS modifications with Lambda; containerized monolithic workloads onto Kubernetes (EKS & Azure VMs).		
Senior Linux Administrator & DevOps Engineer Rave Technologies India		Jul 2016 – Jan 2018 Mumbai
Tech Stack: Enterprise Linux, AWS, Puppet, Bamboo, Jenkins, Git, Docker, MySQL, Nagios, Terraform, NFS, Samba.		
- Administered 2,000+ Linux enterprise servers using Puppet and Bash/Python; built automated CI/CD pipelines via Jenkins and Bamboo. - Successfully executed lift-and-shift migration of infrastructure from VMware vSphere to AWS EC2 with automated scaling and HIDS security.		
Early Career Foundation (2011 – 2016)		
ControlCase International	Sr. Executive (Infra & SecOps) – Implemented PCI-DSS directives, SIEM, Nagios, Cisco/Fortinet firewalls; automated vSphere-to-AWS migration with Terraform.	Feb 2014 – Mar 2016
CDAC (former NCST), Mumbai	Project Engineer – Implemented BOSS Linux in mission-critical infra of INDIAN NAVY ; deployed network print servers; conducted Linux training for naval officers.	Jul 2011 – Jan 2014

CERTIFICATIONS

Certified Kubernetes Administrator (CKA)
The Linux Foundation | Credential ID: CKA-2000-008142-0100

HashiCorp Certified: Terraform Associate
HashiCorp

CNSS (Networking & System Security)
CDAC Hyderabad

EDUCATION

Bachelor of Technology (B.Tech)
IIMT Engineering College, Meerut | UPTU Lucknow (2007 – 2011)

Higher Secondary Certificate (HSC – 10+2, PCM)
RTC Ormanjhi, Jharkhand Academic Council, Ranchi

Secondary School Certificate (SSC – 10th)
RTC High School, CBSE Board