Aditya Jha

Cloud & AI Solutions Architect | Pre-Sales | Program Management | AWS Certified (Cloud & AI Practitioner)

Bengaluru, India | jha.aditya17@gmail.com | linkedin.com/in/adityarjha | github.com/adityarjha17 | medium.com/@jha.aditya17

SUMMARY

Cloud and AI solutions architect, from production Kubernetes on AWS to LLM inference sizing and GPU clusters. Turns client requirements into sized, costed designs, writes the proposal and sees it through delivery, for universities, enterprises, and defence and government tenders. Pre-sales and program management experience; AWS Certified Cloud and AI Practitioner.

EXPERIENCE

MiPhi Semiconductors | *Solutions Architect (Pre-Sales) & Program Manager* Feb 2026 - Present

- Closed a ₹2 Cr+ Kubernetes GPU cluster order with an education customer, via a channel partner, within 7 months of joining. Delivered the BoQ, power and cooling sizing, and licensing, SLA and security responses.
- Lead solutions architect for the team's AI deals across 10+ enterprise, education and public-sector engagements, including an 18,000-user on-prem coding-agent proposal to replace a cloud LLM API bill.
- Appointed project manager of a 5-person agentic AI product team (scrums, Jira, weekly management updates).
- Program manager for a multi-party server program with a government R&D body and two industry partners.
- Audited a GPU fleet capacity model, fixed 4 modelling errors, rebuilt it across 40 model and precision combinations.
- Led a 12-deliverable go-to-market package for an on-prem agentic AI appliance, from sales deck to TCO model and user manual.

Osfin.ai | *DevOps & Platform Engineer* Jun 2025 - Jan 2026

- Built and secured production Amazon EKS clusters with modular Terraform (IRSA, managed node groups, workload isolation), cutting cluster provisioning time by 40%.
- Fixed CI/CD and Helm pipeline failures across environments, cutting build failures by 30%.
- Automated multi-cloud operations on AWS, OCI and Azure with Jenkins, n8n and Python, cutting compute cost by 60% and saving 6+ engineer-hours a week.

AMD India | *Co-op Intern, GenAI* Jul 2024 - Jun 2025

- Fine-tuned Indic-language LLMs with INT8, INT4 and GPTQ quantisation for about 20% better inference efficiency on low-resource GPUs.
- Configured and deployed an end-to-end fine-tuning suite with automated inference, evaluation and benchmarking, orchestrated on Kubernetes.

Corizo | *Trainee* May 2023 - Jul 2023

- Designed a highly available multi-tier AWS architecture (EC2, RDS, S3, auto scaling, load balancing) with Lambda automation and CloudWatch alerting.

EDUCATION

Kalinga Institute of Industrial Technology (KIIT), Bhubaneswar, Odisha 2021 - 2025

B.Tech, Computer Science and Engineering | CGPA 9.31 / 10

SKILLS

Cloud & Platform: AWS (EKS, VPC, IAM, Lambda), Kubernetes, Terraform, Helm, ArgoCD, Jenkins, Docker, Prometheus, Grafana, Azure, OCI

AI & LLM Infrastructure: LLM inference sizing, Fine-tuning, vLLM, KV-cache offloading, Quantisation, RAG, Agentic AI, NVIDIA B200 / RTX PRO 6000

Solution Architecture: Requirements to architecture, BoM / BoQ, TCO modelling, Concurrency modelling, RFQ & tender responses, Proposals

Delivery: Program management, Scrum, Jira, Multi-party coordination, Status reporting

CERTIFICATIONS

- AWS Certified AI Practitioner, Amazon Web Services | Oct 2025
- AWS Certified Cloud Practitioner, Amazon Web Services | Sep 2025
- McKinsey.org Forward Program, McKinsey.org | Jul 2025
- Career Essentials in GitHub Professional Certificate, GitHub, LinkedIn Learning | May 2025
- Real World AI Applications with Gemini and Imagen, Google Cloud | Apr 2025
- Prompt Design in Vertex AI, Google Cloud | Apr 2025

HIGHLIGHTS

- Represented MiPhi Semiconductors at Convergence India, UPITS and Bengaluru Tech Summit