

JAYAPRAKASH ALAWALA (JP)

Enterprise Technology Leader — Cloud Migration & Modernization at Scale | AI Infrastructure, GenAI Inference & Agentic Platforms | Industry Thought Leader

Bangalore, India | +91 9900622490 | alawalajp@gmail.com Web: cloud-tech-consulting.com | LinkedIn: linkedin.com/in/jpalawala | YouTube: youtube.com/@ctcwithjp

Experience Summary

- **Principal Containers Specialist Solution Architect (L7)** — Amazon Web Services (AWS) Internet Services Private Ltd, **September 2019 – Present**.
- **DevOps Technical Leader II (Grade 11)** — Cisco India Pvt Ltd, **Jan 2013 – Oct 2018**.
- **Application Engineer Tech Lead** — Texas Instruments India Pvt Ltd, **July 2001 – Dec 2012**.

Executive Summary

Technology leader operating at the intersection of the two transformations enterprises are funding today: **cloud migration & modernization at scale**, and **becoming AI-first**. Principal Solutions Architect (L7) at AWS with 20+ years' experience, trusted advisor to **500+ enterprise customers** — India's largest private banks, OTT, fintech and global SI/ISV platforms.

As a business leader, I grew Application Modernization **582% over three years**, landed **105% of revenue target (+34.2% YoY)**, and founded a partner-led delivery programme that compressed enterprise container migrations from **6–18 months to 6–10 weeks** across 5 partners and 47 customers. **As a technologist**, I architect GenAI inference on both **NVIDIA GPU and AWS silicon (Trainium/Inferentia)** — cutting **GPU inference cost ~60%** and **time-to-first-token up to 5×** — and have solved cross-region GPU scarcity as a reusable industry pattern. **As an industry voice**, I shape the AWS container roadmap through **29 product feature requests** and an upstream Kubernetes proposal, author **13 AWS blogs** and workshops **delivered 168 times** across the field, speak at **AWS re:Invent 2025**, reach **~700 subscribers** on my technical YouTube channel, and grew a technical community from **87 to 187 members** while mentoring **12+ engineers** to senior and principal level.

Technical Expertise

- **Cloud Migration & Modernization:** AWS Transform (agentic migration & modernization), Trianz Concierto Agentic Platform (Agentic Assess / Migrate / Modernize), AWS MGN, discovery & assessment, TCO and business-case modelling
- **Containers & Cloud Orchestration:** AWS, Kubernetes, Amazon EKS/ECS, ROSA (OpenShift on AWS), Fargate, ECR, Docker, API Gateway, Lambda
- **GenAI / AI Infrastructure:** Amazon Bedrock & Bedrock AgentCore, LLM inference (vLLM, KubeRay, Karpenter, S3/EFS/FSx Lustre), NVIDIA GPU orchestration, AWS Inferentia/Trainium, RAG & agentic architectures, Amazon Q, Kiro
- **Agentic AI & GenAI Platform Stack:** LangGraph/LangChain, LiteLLM & Kong AI Gateway, Langfuse & Arize (observability), Ray/Ray Serve, Ragas (LLM-as-a-Judge), Milvus & Redis (vector/RAG), MCP tool servers, Bedrock Guardrails, KEDA, FastAPI, Open WebUI, FlexAI, Strands SDK
- **Platform Engineering, Observability & FinOps:** Prometheus, Grafana, CloudWatch Container Insights, ADOT/OpenTelemetry, X-Ray, Fault Injection Simulator, KubeVela/OAM; Karpenter, Cluster Autoscaler, HPA/VPA, EC2 Spot, Graviton, Fargate Spot
- **Security, CI/CD & IaC:** IAM, Kubernetes RBAC, Secrets Manager, Falco, OPA/Gatekeeper, Kyverno, CIS/PCI-DSS/HIPAA; ArgoCD, CodePipeline, Jenkins, Blue-Green/Canary; Terraform, AWS CDK, CloudFormation, Crossplane, KRO, Boto3/Python
- **Networking & Service Mesh:** Istio, Linkerd, VPC design, ALB/NLB, ECS Service Connect, Cloud Map, Route 53, K8s Gateway API, VPC Lattice
- **Partner/ISV Ecosystem & Languages:** Kong, Nirmata/Kyverno, Arize, Anyscale · Python, Go, Bash · Amazon Linux 2023, Bottlerocket · DRM (VGDRM, Widevine, PlayReady)

Certifications

- **AWS Certified AI Practitioner (AIF-C01)** — 2025
- **AWS Certified Solutions Architect – Professional (SAP-C02)** — 2024
- **AWS Certified Solutions Architect – Associate (SAA-C03)** — 2020
- **Kong AI Gateway Operations** — Kong ([Credly](https://credly.com/certifications/kong-ai-gateway-operations))
- **Kong Gateway Operations** — Kong ([Credly](https://credly.com/certifications/kong-gateway-operations))
- **Kong Gateway Foundations** — Kong ([Credly](https://credly.com/certifications/kong-gateway-foundations))
- **AWS AI-Driven Development Lifecycle (AI-DLC) Ambassador — Foundational (L100)** — AWS ([Credly](https://credly.com/certifications/aws-ai-driven-development-lifecycle-ambassador-foundational))
- **AWS Knowledge: AI-Driven Development Lifecycle Foundations** — AWS ([Credly](https://credly.com/certifications/aws-knowledge-ai-driven-development-lifecycle-foundations))

AWS Experience — Sept 2019 – Present

Role: Principal Container Specialist Solutions Architect (L7) · **Focus:** AI infrastructure, GenAI platforms, Kubernetes/EKS · **Region:** India / APJ

AI-First Enterprise Transformation — Flagship Wins

- Anchored the **cloud-native, microservices-based container architecture** on **Amazon EKS** for a **top India private-sector bank's 3-year CRM modernization** — **120M customers, 45K concurrent users, 9,400 branches, 100TB+ data, 350M+ records** — a competitive win reclaiming the workload to AWS.
- Single-threaded containers authority on **India's largest AWS account** — a **national OTT streaming platform** — home of the **72M-viewer world-record streaming concurrency on Amazon EKS**; authored the definitive account-consolidation strategy that proved AWS scalability at world-record concurrency.
- **Lead EKS architect and single-threaded owner** for a top India private bank's internet-banking (NetBanking) platform — owning end-to-end design, scale, networking, and security — and architected its **disaster-recovery strategy at near-zero RPO and ~12-minute RTO**, setting a new resiliency high-bar for mission-critical banking workloads on AWS.

AI Infrastructure — GPU & AWS-Silicon Inference at Scale

- **Sole author of the KV-Cache Offloading section** of the flagship, globally-used **NVIDIA GPU GenAI-on-EKS workshop** (vLLM + LMCache + Amazon ElastiCache) — delivering **up to 5× faster time-to-first-token and ~60% GPU-cost reduction**, with a companion AWS Containers blog.
- **Designed and built a cross-region GPU-scarcity autoscaling pattern** that solves the field's #1 GenAI-batch blocker — overflow activates only on a genuine GPU capacity signal, drains work cross-region and scales back to zero on recovery, with no job loss. Proven live, **endorsed by AWS Principal GTM Compute leadership**, region- and accelerator-agnostic, and extended to a national financial-markets platform for on-premises-to-AWS burst.
- **Authored the complete Ray and Anyscale chapter** for AWS's GenAI accelerator workshop — an end-to-end GPU path covering LLM serving and autoscaling, multi-model GPU packing, fault tolerance, batch inference, distributed fine-tuning and operations — plus the open-source KubeRay counterpart, so customers can adopt self-hosted or managed.
- **Advised the Amazon EKS product-management team on GPU capacity fungibility and Dynamic Resource Allocation (DRA)** — static versus just-in-time GPU provisioning and reservation strategy to share one GPU pool across training and inference — guidance that shaped the recommended path for a large FSI customer. GPU capacity practice (Capacity Blocks, on-demand reservations, checkpointing, DRA, lazy image loading) is part of the best-practice content I teach.
- **AWS silicon (Trainium/Inferentia/Neuron) depth**: root-caused a severe GenAI model-loading bottleneck for a large India FSI bank — proving the delay was storage weight-loading rather than model compilation — driving AWS's move to high-throughput shared storage for LLM inference; authored the **Llama-model storage price/performance benchmark on AWS Neuron**; and surfaced a model-architecture coverage gap plus a performance regression to the Neuron service team.

Agentic AI Platforms & Scaled AI Enablement

- **Co-owner of AWS's flagship agentic-AI platform workshop on Amazon EKS** — authored the managed-SaaS track from scratch (Kubernetes-native **AI gateway + LLM observability + agent SDK**); delivered in **Singapore and London**, and the basis of a joint go-to-market with an AI-compute ISV.
- **Built the entire Token Economics (AI cost-governance) workshop with partners Kong and Arize** — per-tenant cost attribution, budget enforcement, cost-aware routing and semantic caching — delivered in Singapore and at a European tech summit.
- Authored **9 agent-ready guidance playbooks** for AWS's internal specialist AI agent and engineered the **first repeatable server-side A/B evaluation harness** proving a playbook measurably improves agent guidance with no regressions — recognised by the Worldwide Tech Leader for Containers.
- Shipped **7 public agent-skills (3 EKS + 4 ECS, including GPU/ML on ECS)** to AWS's open agent-skills library — customer-facing, opinionated golden-path guidance.
- **Agentic migration-platform evaluation (Trianz Concierto)**: completed a hands-on technical review of the **Concierto Agentic Assess / Migrate** platform — running discovery, assessment and TCO to completion, working through the landing-zone and network-migration agents, building migration plans and move groups end to end, and **benchmarking its cost model against AWS Transform on an identical 500-VM input** — delivering **36 findings framed as product feature requests** (cost defensibility, governance and workflow clarity, plus differentiation opportunities), which Trianz escalated to their product team.

Business Impact & Scale

- Created and executed a 3-phase Application Modernization Webinar Series (**3,500+ participants across 700+ customers**), driving major modernization pipeline and positioning AWS as category leader in containers.
- Led India Application Modernization to **105% of revenue target (+34.2% YoY)** and **+62.4% YoY pipeline growth** in 2025; the 1:Many AppMod program reached **2,000+ customers across 40+ sessions**.
- Drove sustained business growth 2022 → 2025 — a **582% overall increase** — through YoY gains of 324% (2023), 22% (2024), and 34.2% (2025).
- Founded the **Partner Packages program (5 SI/ISV partners, 47 customers)**, compressing container deployment from **6–18 months to 6–10 weeks** (>100% efficiency gain); drove multi-city EKS design workshops (CSAT 4.77/5).

Containers/EKS Enablement & Field Leadership

- Created, own, and maintain the **AWS EKS Security Immersion Day (since 2023)** — **delivered 168 times, generating ~USD \$54.3M in annualized opportunity revenue**.
- As **APJ Lead of the Container Ambassador Program**, grew the Container Technical Field Community from **87 → 187 members (+69% YoY 2024, +28% YTD 2025)**, multiplying field capacity across APJ.
- Drove **4 published AWS case studies** including **India's first EMR/Spot (a major music-streaming platform, 60% cost cut)** and **first ECS/Spot (a leading fintech, 50–70% cost cut)**.

- Founded a **6-language localization cohort (8 native-speaker reviewers across APJ/EMEA/AMER)** for the Platform Engineering on EKS workshop; French content shipped for AWS Summit Paris.
- Mentored **12+ engineers to senior/principal roles**; promotion assessor; founder of a specialist community cohort; 16 hiring interviews.

Enterprise Platform Advisory — Migration, Modernization & Multi-Tenancy

- Trusted advisor with **500+ enterprise customer conversations** across on-premises → AWS migration, multi-cloud re-architecture, hybrid cloud, DevOps/CI-CD, GenAI and security; single-threaded containers authority across **40+ enterprise accounts** spanning banking, fintech, OTT, SaaS and digital-native leaders.
- **Competitive multi-cloud wins**: led the container architecture for enterprise SaaS and AI platforms migrating **from other hyperscalers onto AWS**, including a multi-tenant enterprise-AI SaaS platform that achieved **30–40% lower cost per tenant** while displacing the incumbent.
- **Multi-tenant SaaS at extreme fleet scale**: designed and validated the network and isolation architecture for a data-platform ISV scaling to **1,000+ clusters per region** — custom networking with reusable address space, interconnect topology, a capacity-ceiling density model and a tenant security-group design — selecting the option that was **dramatically cheaper per month** than the alternatives.
- **Platform fleet operations**: root-caused IP-address exhaustion for a global consulting firm’s multi-tenant delivery platform, proving it was a provisioning-ordering race rather than a tuning problem, and delivered a **validated reference configuration plus a live reproduce-and-fix proof of concept** — opening a path to a 50-cluster-per-account estate with no platform refactor. For an XXL fintech, built the fleet-management, upgrade-at-scale and async-canary guidance for a custom-hardened-AMI estate.
- Unblocked FSI compliance and security gates for multiple banks and fintechs — establishing CIS-hardened, bank-grade EKS security baselines for regulated production workloads.

Technical Innovation

- Engineered a **custom Kubernetes pod scheduler** that lifted a **major India insurtech’s EC2 Spot adoption from 23% to 46%** — winning the **Invent & Simplify Award (APJ, Q1 2021)**; published as an AWS blog + upstream Kubernetes Enhancement Proposal.
- Re-architected a **US healthcare-technology leader’s event-processing platform (6B events/month)** onto ECS + EC2 Spot, cutting cost **73%** while supporting **10× load at 0.3× cost**.
- Built a native-AWS **EKS Spot ↔ On-Demand failover solution** adopted in production by a global financial-software firm, and pioneered **stateful workloads on EC2 Spot (60% savings)** — together contributing to a **+71% YoY rise in India EC2 Spot adoption**.

Product & Thought Leadership

- Shaped the **Amazon EKS/ECS/Fargate roadmap through 29 Product Feature Requests & upstream contributions** — materially influencing container security benchmarks, native Kubernetes Network Policy, and a regional Amazon Managed Prometheus launch; co-submitted an upstream Kubernetes Enhancement Proposal.
- Authored **13 AWS Containers/Builder blogs** and **4 published case studies**; published 1 AWS Solutions Library guidance (EKS external SSO); co-created 2 EC2 Spot Game Days.
- Delivered **4 sessions at AWS re:Invent 2025 (CSAT up to 5/5)**; also KubeCon, AWS Summits, Tech Summits, Developer Days.
- **Open-source**: core contributor to AWS EKS security tooling and agentic-AI evaluation frameworks; author of published **aws - samples** reference solutions for EKS Spot resilience and access governance.

Honors & Recognition

- **15 AWS awards/recognitions (2020–2026)**: MVP (Q3 2020), Invent & Simplify (APJ, Q1 2021), APJ Compute Summit Award, WWSO India Excellence — Customer Obsession (5×), Unified Solutions, SA Champion, Trailblazer Deal, Outstanding Service Team, **Silver TFC Member — Containers (H1 2026)**.

Cisco Experience — Jan 2013 – Oct 2018

Role: DevOps Technical Leader II (Grade 11), Cisco Systems India · **Team Size:** 5 · **Clients:** global telecom & pay-TV operators

- Led a 5-engineer DevOps team owning the **Ejabberd** (Erlang/XMPP) messaging cluster — CI/CD, Salt/Ansible/Python IaC, and upgrade/rollback automation that **cut downtime from 2 hours to 15 minutes**.
- Architected a production **Prometheus observability stack** (custom Python metrics aggregation across 50+ deployments) — **cutting MTTR 40%** for a 500K+ household streaming platform.
- Led **multi-DRM implementation** (Google Widevine, Microsoft PlayReady, **Cisco VGDRM**) and security-layer hardening — secure streaming to **500K+ households** across STB/Android/iOS for a large-scale streaming business.
- Designed end-to-end DRM infrastructure (broadcast head-end system, security gateways, key-management) as resident **DRM SME across APAC**.
- Containerized streaming back-office microservices on Kubernetes/OpenShift (**6 months → 8 weeks**) and built a “One-Click” soak/load-test framework (500K-device simulation) for infra supporting **10M+ daily active users**.
- Automated tenant lifecycle + cost optimization — **35% infra-cost reduction at 92% utilization**.
- **9 Cisco awards** (Make Innovation Happen, Win Together, Connect Everything, ...).

Texas Instruments Experience — July 2001 – Dec 2012

Role: Application Engineer Tech Lead · **Team Size:** 3 · **Clients:** OEMs

- Led applications engineering for **OMAP mobile application-processor SoCs** across 6 major projects — embedded multimedia and computational photography for Motorola, LG, Google, Nokia, NEC, Dell and **50+ global partners**; accelerator-based performance optimisation to 1080p30; demos at CES/MWC.
- **Applications Lead (2009–2012):** managed 3-engineer teams on flagship smartphone/tablet reference platforms; **trained 200+ partners** across Korea/India/China and established an APAC developer-network support framework for 20+ partners.

References

1. **Rajesh Singh** — Head of App Modernization Business, Microsoft / Azure India
2. **Guru Balasubramaniam** — Head of Customer Engineering (Digital Natives), Google Cloud India

(Contact details available on request.)

Education

- **B.E. Computer Science & Engineering**, 85% — Osmania University College of Engineering (Autonomous), Hyderabad, 1997–2001. *EAMCET State Rank 153; State Rank 26 (Intermediate); State Rank 12 (X standard).*

Personal

- **Location:** Bangalore, India · **Languages:** English, Telugu, Hindi · **Passport:** valid until 2031