AWS cost audit: Series B AI startup, $38.4K → $22.4K monthly.

Cloud Cost Optimization Audit · 2 weeks · $750 engagement fee

BEFORE	AFTER	ANNUALIZED SAVINGS	ROI
$38,400/mo	**$22,400/mo**	**$192,000**	**256×**
+14% MoM, pre-audit	42% reduction	at stable run rate	on the $750 audit

THE SETUP

50-engineer AI infrastructure startup, 18 months post Series B. Agent platform running on AWS us-east-1 and us-west-2: EKS for control plane, GPU inference on p4d.24xlarge and g5.12xlarge, OpenSearch for vector retrieval, RDS Postgres for application state. Monthly AWS run rate at $38,400 and climbing 14% month over month. CFO wanted a defensible answer before the next board meeting. Engineering didn't have the bandwidth to stop and look.

WHAT I FOUND

- **$4,800/mo** **Idle GPU capacity overnight.** Three p4d.24xlarge instances left running for dev experiments, 14 hours/day unused. Replaced with an on-demand scheduler and a shared inference endpoint.
- **$3,500/mo** **CloudWatch Logs over-ingestion.** Debug-level app logs retained 60 days across all environments. Dropped non-prod to 7-day retention, capped dev/staging at INFO.
- **$3,200/mo** **Inter-AZ traffic from the RAG pipeline.** 12M+ cross-zone calls/day to OpenSearch. Moved vector retrieval to same-AZ replicas, kept cross-AZ for failover only.
- **$2,600/mo** **OpenSearch oversized for actual load.** Cluster sized for peak QPS 8× above observed p99. Rightsized in place. No latency regression.
- **$1,900/mo** **Zero Savings Plans coverage.** Two years of stable baseline EKS + RDS usage, paying full on-demand. Applied a 1-year Compute Savings Plan at 100% of observed baseline.

WHAT I DID

Two weeks. Three calls. One shared Google Doc. Read the Cost and Usage Report for Q1 and Q2, cross-referenced against Compute Optimizer and Trusted Advisor, sampled traffic patterns from VPC Flow Logs, and spent a day with the eng lead walking through the RAG pipeline and inference layer. Every finding shipped with a runbook: the change, the owner, the rollback, and the 7-day verification check.

OUTCOME

MONTHLY SPEND	FIRST-YEAR SAVINGS	ENGAGEMENT COST
$22,400 from $38,400	**$192,000**	**$750** + optional retainer

> **What wasn't sold.** No managed service contract. No ongoing dashboard access. No "let us run your cloud." The team runs the changes. I'm available by the hour when they hit something sharp.