Thiago Nazario - Principal Cloud Security Architect | Hardened by Design™ Methodology

https://github.com/thiagonazario

ARCHITECTURAL AUDIT REPORT

Project: Ghost Protocol – FinOps-First AI Infrastructure

Date: February 9, 2026

Status: Hardened & Optimized

Architect: Thiago Nazário

---

1. EXECUTIVE SUMMARY

This document provides a technical audit of the **Ghost Protocol** infrastructure. The architecture was designed to host private Large Language Models (LLMs) using a **Hardened-by-Design** methodology. The primary goal is to achieve maximum security (Zero-Trust) while maintaining a **Zero-Waste** financial policy through advanced FinOps automation.

2. CORE ARCHITECTURAL PILLARS

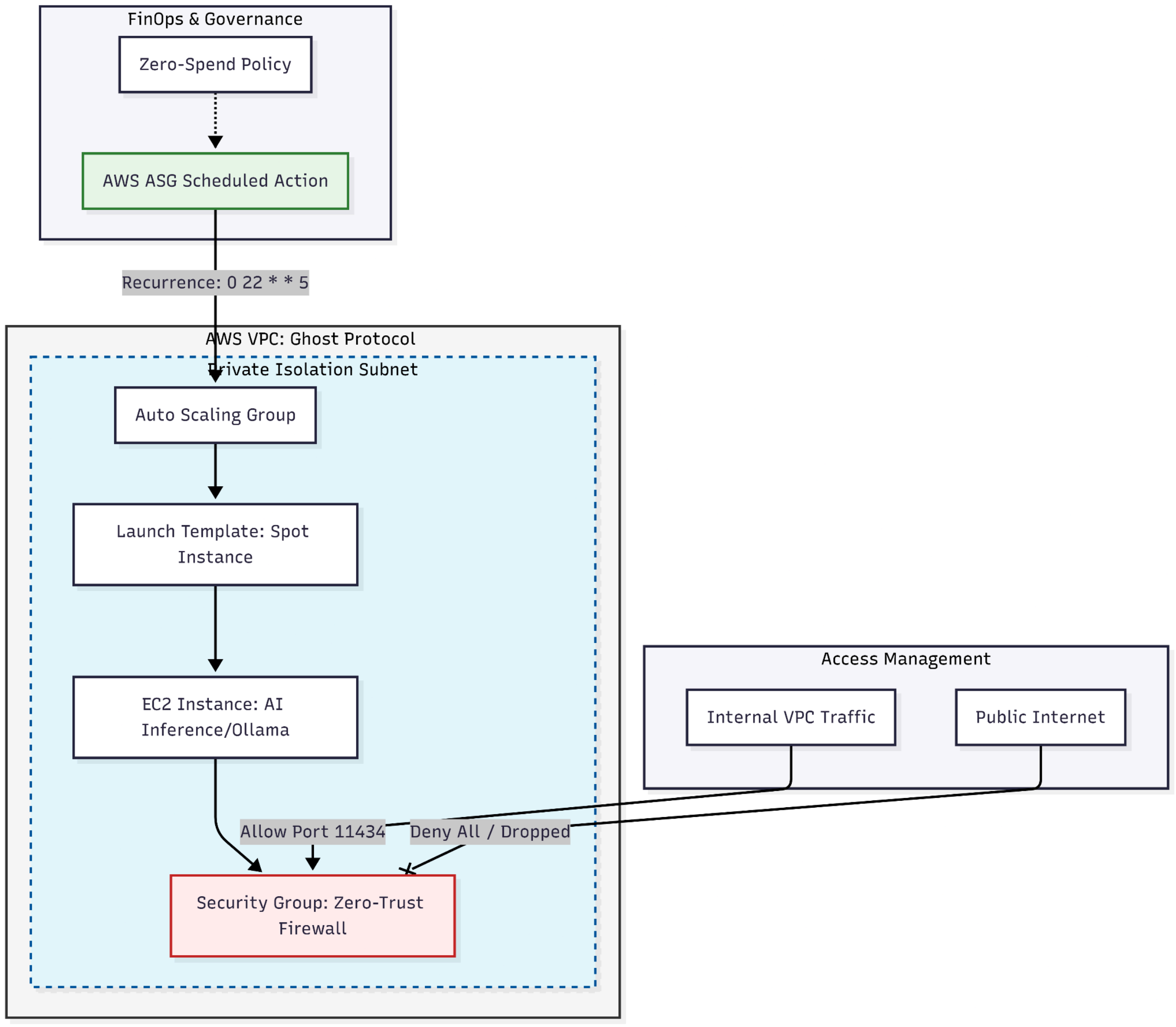
2.1 Hardened-by-Design Networking

- **Isolation Strategy:** The workload is deployed within a strictly private VPC subnet.
- **Zero-Trust Firewall:** Security Groups are configured with a "Deny-by-Default" stance. Ingress is only permitted on Port 11434 (Ollama) from verified internal CIDR ranges.
- **Attack Surface Reduction:** By eliminating Public IPs and NAT Gateway dependencies where possible, the infrastructure remains invisible to external scanners.

2.2 FinOps & Cost Engineering

- **Spot Instance Utilization:** The compute layer leverages AWS EC2 Spot Instances, providing a **70-80% cost reduction** compared to standard On-Demand pricing.
- **Automated Lifecycle (The Kill-Switch):** Implementation of aws.autoscaling.Schedule to scale the infrastructure to **zero** during non-operational hours (Weekends).
- **Infrastructure as Code (IaC):** 100% of the stack is managed via Pulumi, ensuring total cost transparency and preventing "zombie resources."

[PAGE 2 - VISUAL EVIDENCE]



---

3. TECHNICAL STACK & GOVERNANCE

- **Cloud Provider:** AWS (Amazon Web Services)
- **IaC Engine:** Pulumi (Python SDK)
- **Compute:** EC2 Auto Scaling Groups (Spot)
- **Compliance:** DevSecOps Standards / SOC2-ready patterns

4. AUDIT CONCLUSION

The Ghost Protocol infrastructure is an elite-tier solution for sovereign AI operations. It demonstrates that high-performance AI inference can be delivered with fiscal responsibility and military-grade security.