
KubeCon
CloudNativeCon
India 2026
#KubeCon #CloudNativeCon
Zero-GPU Autopilot: Orchestrating Kagent and Kgateway for Private, Self-Healing Clusters
- ASHOK M , DigitalOcean
- DILLIBABU SAMPATH , Wells Fargo

# Problem Statement



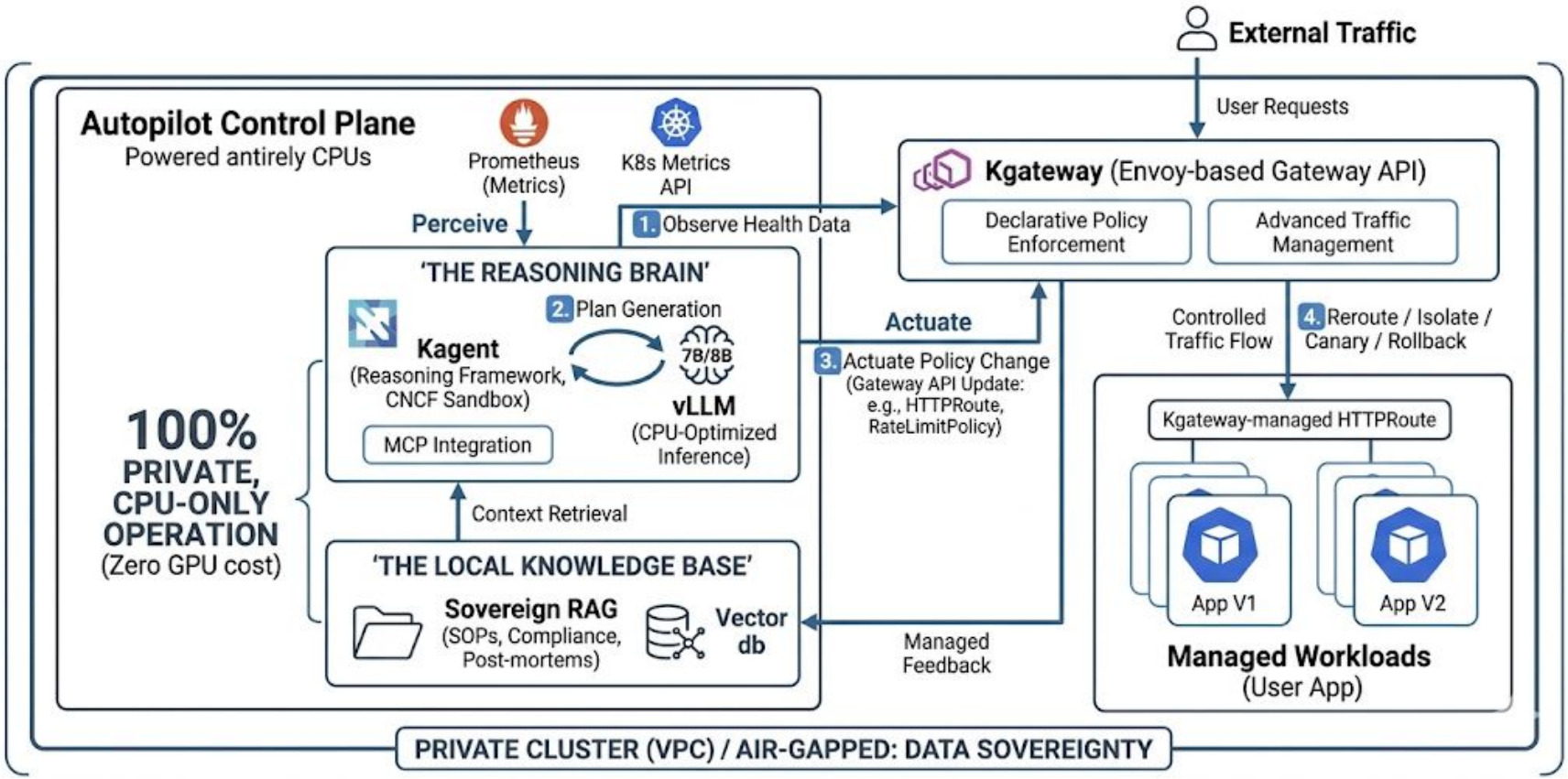
- MTTR Crisis
- Traditional Failures
- GPU Cost
- AI Based SRE Agent Complexity for Simple tasks
- Tuning AI for Company related best practices
- Data Privacy and Data Sovereignty

# Our Approach



- 100% Private Operations: Not a single log, metric, or prompt ever leaves the secure VPC boundary.
- Zero External AI APIs: Completely decoupled from costly third-party closed-source LLMs.
- Zero Dedicated GPUs Required: Capitalizes on optimized, CPU-only local model serving.
- Deterministic Policies: AI actions are governed by declarative Kubernetes Gateway API rules.
- Kubernetes based local agent for solving problems in K8s Cluster.
- Kubernetes based gateway for handling the Network traffic and its related issues

# Zero-GPU Autonomous SRE Architecture: closed-loop self-healing



# Why vLLM?



- **vLLM + Intel Backends**
  - Direct integration with OpenVINO & IPEX
  - Native CPU execution → up to 10× faster
- **Paged Attention**
  - Solves KV cache fragmentation (60–80% waste)
  - Uses RAM "pages" like OS virtual memory
  - Enables high-concurrency agent sessions on CPU
  - Prevents OOM during parallel alerts

# Kagent: Sovereign "Reasoning Brain"



- CNCF Sandbox Framework: A native Kubernetes reasoning engine built to understand cluster states—not a generic chatbot.

**1. MCP (Model Context Protocol)**

- Standardized bridge between LLMs and the Kubernetes API.
- Replaces risky raw shell scripts with safe, structured API tool execution.

**2. Context Compaction**

- Filters out complex cluster event "noise" into compact datasets.
- Fits massive telemetry logs into the smaller context windows of {7B} or {8B} CPU-optimized models.

**3. Bounded Agentic Flow**

- Multi-step deterministic plan generation.
- Structured, logical execution: Inspect Metrics ->Match Local SOP->Patch Route.

# Kgateway: The Policy Guardrail



- Envoy-Based Gateway API: Modern, policy-driven traffic routing (built by Solo.io) to manage, secure, and insulate the cluster's data plane.

**1. Declarative Actuation**

- Operates strictly through standard declarative resources like HTTPRoute or TCPRoute.
- Replaces unsafe network "hacks" with structured, auditable GitOps-friendly configs.

**2. The "Safety Sandbox"**

- Hard, human-defined policy boundaries (e.g., global Rate Limit policies).
- Instantly rejects any agentic action that violates human-configured cluster limits.

**3. Network-Layer Mitigation**

- Envoy-native traffic shifting and canarying running entirely on standard host CPUs.
- Resolves cascading errors instantly at the L7 layer with zero costly, slow pod restarts.

# Closed Loop

KubeCon CloudNativeCon India 2026

**1. Intent-Based Automation**

*Old Way:* Manual, panic-driven imperative commands (e.g., kubectl delete pod).

*Sovereign Way:* Prometheus Alert -> Kagent RAG Triage -> Kgateway HTTPRoute Patch.

**2. Separation of Concerns**

*Kagent (Intention):* Proposes logical mitigation steps (e.g., "Isolate failing v2 pods").

*Kgateway (Enforcement)*: Evaluates and enforces the hard boundaries (e.g., "Deny if weight < 50%").

**3. The "Zero-GPU" Synergy**

*Kagent:* Outputs ultra-compact, structured JSON tool calls—no expensive, lengthy creative text generation.

*Kgateway:* Handles high-concurrency traffic shifting natively via high-performance C++ (Envoy) on CPU.


# Demo Scenario

KubeCon CloudNativeCon India 2026

# Q & A

# Reference

**vLLM**

vLLM Docs – https://vllm.ai/docs

GitHub – https://github.com/vllm-project/vllm

**Kagent**

Kagent Docs – https://kagent.dev/docs

GitHub – https://github.com/kagent-dev/kagent

**Kgateway**

Kgateway Docs – https://docs.kgateway.io

GitHub – https://github.com/kgateway-dev/kgateway

⚠️ **Disclaimer: All slides and information are sourced from open-source materials and do not include any proprietary content.**