



KubeCon



CloudNativeCon

India 2026

LLMs Behind Bars: Sandboxes at Scale

Prashanth Pai
CodeRabbit



martian

An Unbiased OSS Benchmark For Code Review Agents

Measure which tools actually catch bugs, improve code, and get their suggestions adopted by developers.

Modify the benchmark to fit your use case.

What matters to you? —

How important is it to avoid noise vs being thorough?

Less noise ← Balance → More Thorough

0 0.5 1 2 ∞

F score (manual input)

● Online Tracker
Real open source repos

● Offline Benchmark
50 hard-to-find bugs

Last month ▾

Share ↗

Code Review Leaderboard

FILTERED FOR: prioritize thoroughness | Across 6,530 scored PRs

Tool	F ₁ Score ↕	Precision ↕	Recall ↕	Total PRs
#1 Coderabbit	48.6%	65.9%	48.5%	69,422
#2 Gemini Code Assist	48.4%	73.4%	48.2%	26,741
#3 Claude	42.1%	62.1%	42.0%	9,247
#4 Qodo Code Review	40.7%	62.9%	40.6%	3,216
#5 Cubic Dev AI	38.6%	72.6%	38.4%	5,775
#6 Greptile Apps	37.5%	73.3%	37.4%	11,318
#7 GitHub Copilot	35.3%	63.9%	35.1%	87,353
#8 ChatGPT Codex Connector	33.8%	71.5%	33.6%	47,162
#9 Cursor	33.4%	69.9%	33.2%	8,598






150,000+
Monthly active users

12M+
Pull requests reviewed

2M+
Installs across repositories





-  [Code Reviews](#)
-  [Autofix](#)
-  [Review Chat](#)
-  [Pre-Merge Checks](#)
-  [Agent for Slack](#)



Sandbox?



KubeCon



CloudNativeCon

India 2026

Isolated environment designed for **AI agents** to **safely** run untrusted code

- Tool calls
- MCPs
- Run LLM-generated code
- Run user-generated code
- Coding agents
- Persist and restore state

Local Laptop
Claude/Codex
OpenClaw

Remote Servers
Background Agents



- **Exfil**
 - secrets, API keys, SSH keys, repo code, customer data
- **Access internal networks**
 - private services/APIs, metadata, infra, databases
- **Poison**
 - generated code, builds, dependencies, caches, artifacts
- **DoS or Cost \$\$**
 - crypto mining, fork bombs, runaway builds



Agent Harness <=> Sandbox



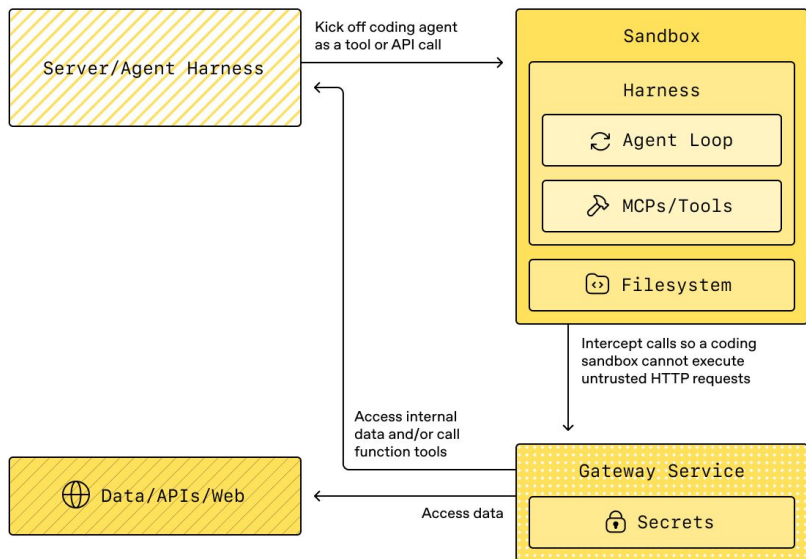
KubeCon



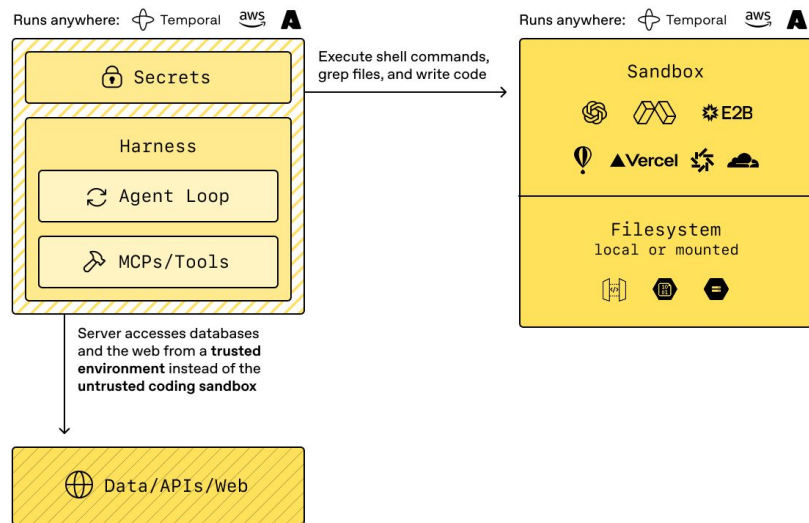
CloudNativeCon

India 2026

Harness in compute



Harness separate from compute



Sandboxes @ CodeRabbit



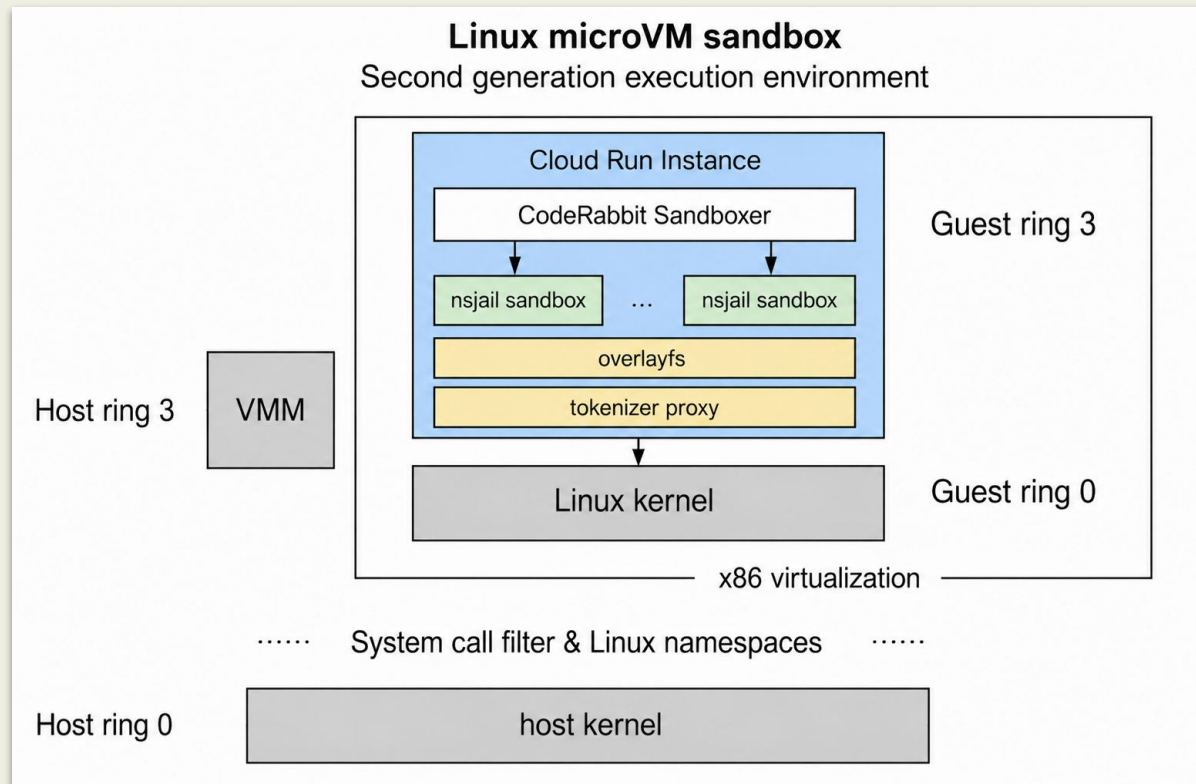
KubeCon



CloudNativeCon

India 2026

- Agent harness outside sandbox
- Agent streams cmds to run in sandbox
- Locked-down tools, packages
- Multi-Tenant



Latency



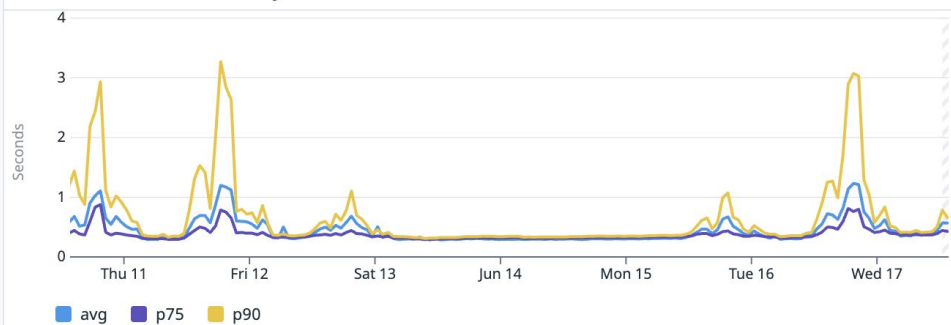
KubeCon



CloudNativeCon

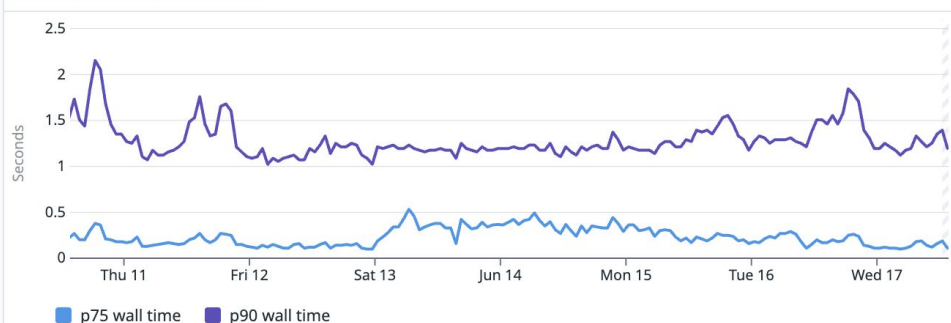
India 2026

Sandbox Creation Latency



~700ms
sandbox startup

Command Run Duration



50+
third-party tools

~250ms
most cmd runs



Scale



KubeCon

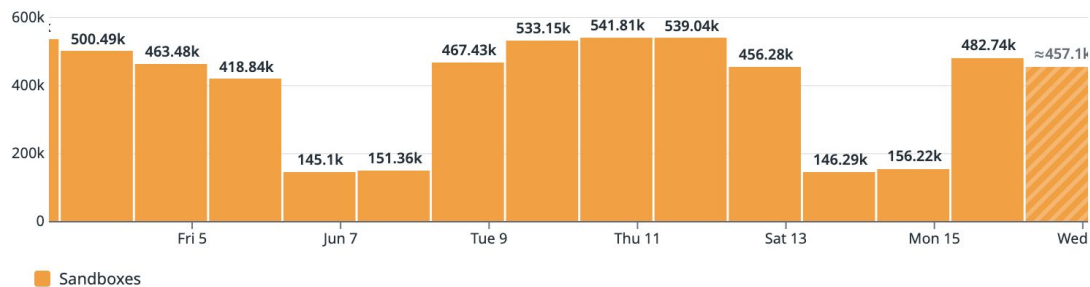


CloudNativeCon

India 2026

Sandboxes Spawed

2w

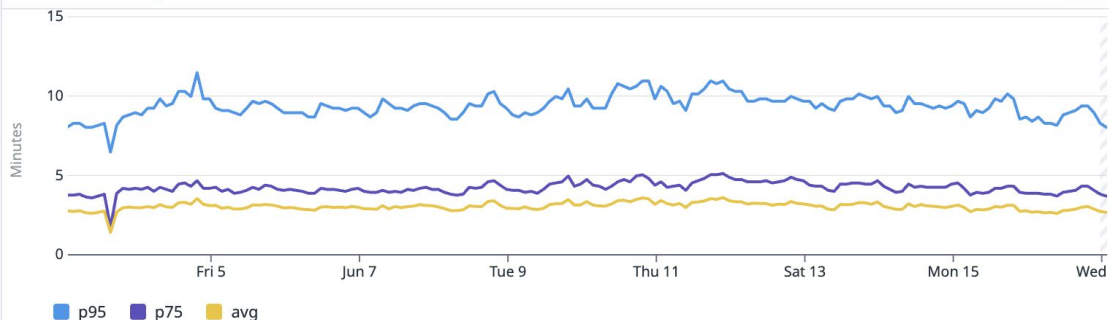


500k
sandboxes/day

Peaks at 33k
sandboxes/hour

Sandbox Lifecycle Duration

2w



Lifecycle duration
3-10 mins



Autonomous Agents



KubeCon

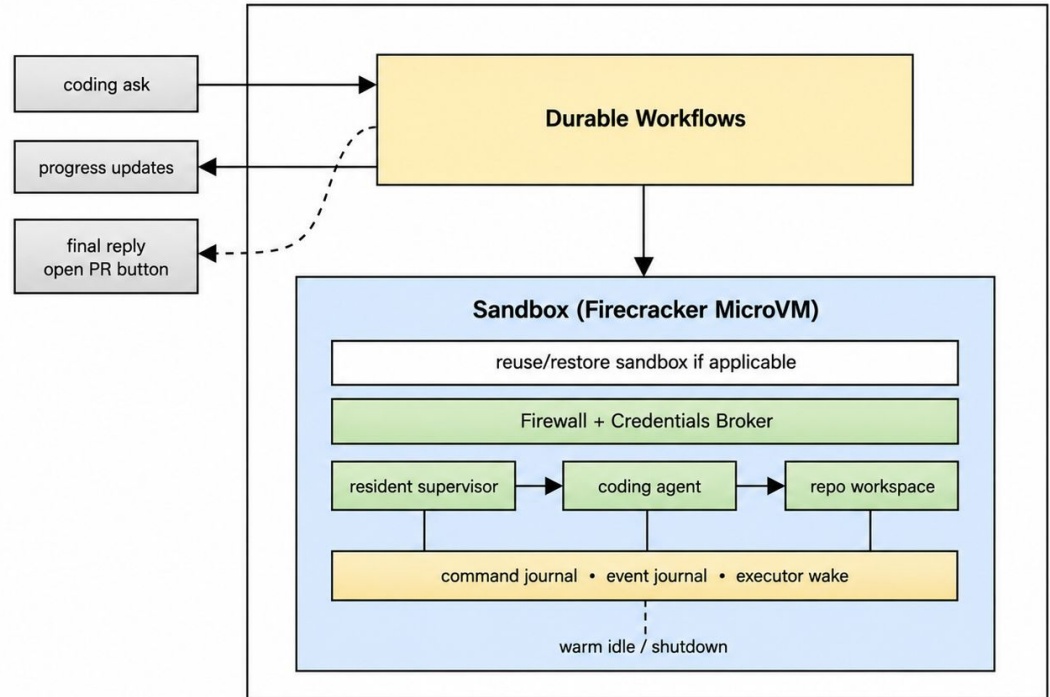


CloudNativeCon

India 2026

- Agent harness inside sandbox
- Stronger Isolation
- Dependency Installation
Flexibility + Snapshots
- Single Tenant per VM
- Operationally complex if self hosted

CodeRabbit Agent Lifecycle



What makes up a Sandbox?



KubeCon



CloudNativeCon

India 2026



Containers are legos

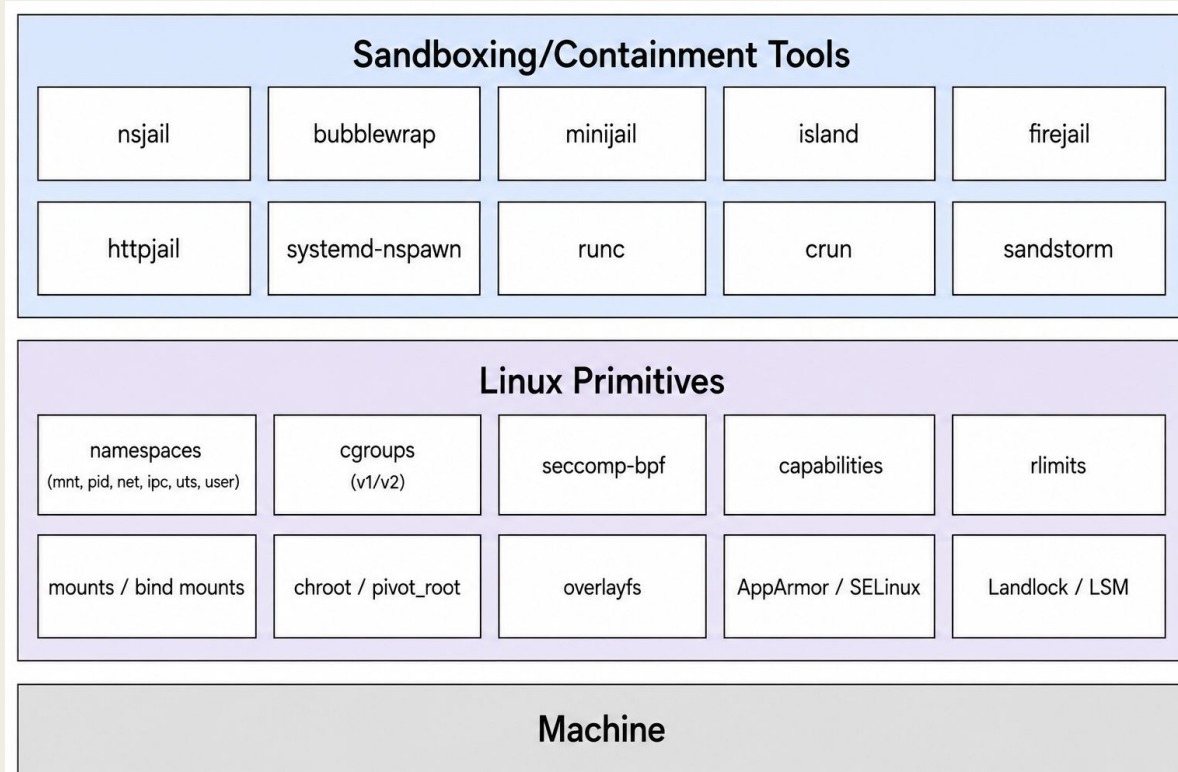


KubeCon



CloudNativeCon

India 2026



MicroVMs



KubeCon



CloudNativeCon

India 2026



Untrusted Workload

AI-generated code • CI scripts • User repos • Dependency hooks • Tests

No secrets

Network policy

CPU / memory / time limits

No persistent data



Container Runtime inside Guest

Podman • Docker-compatible UX • OCI images • No host Docker socket

Rootless where possible

Read-only mounts

Drop capabilities

Seccomp / AppArmor



Ephemeral Guest OS

Dedicated kernel • Disposable roots • No host credentials

Ephemeral image

Minimal packages

No persistent data

No host access



MicroVM / VMM Layer

Firecracker • crosvm • Cloud Hypervisor • libkrun • QEMU-microvm

Minimal device model

Seccomp filter

Jailer / Sandbox

Small TCB



Host OS & KVM

Bare metal • Hardened kernel • Minimal services • Worker isolation

KVM hypervisor

IOMMU

cgroups

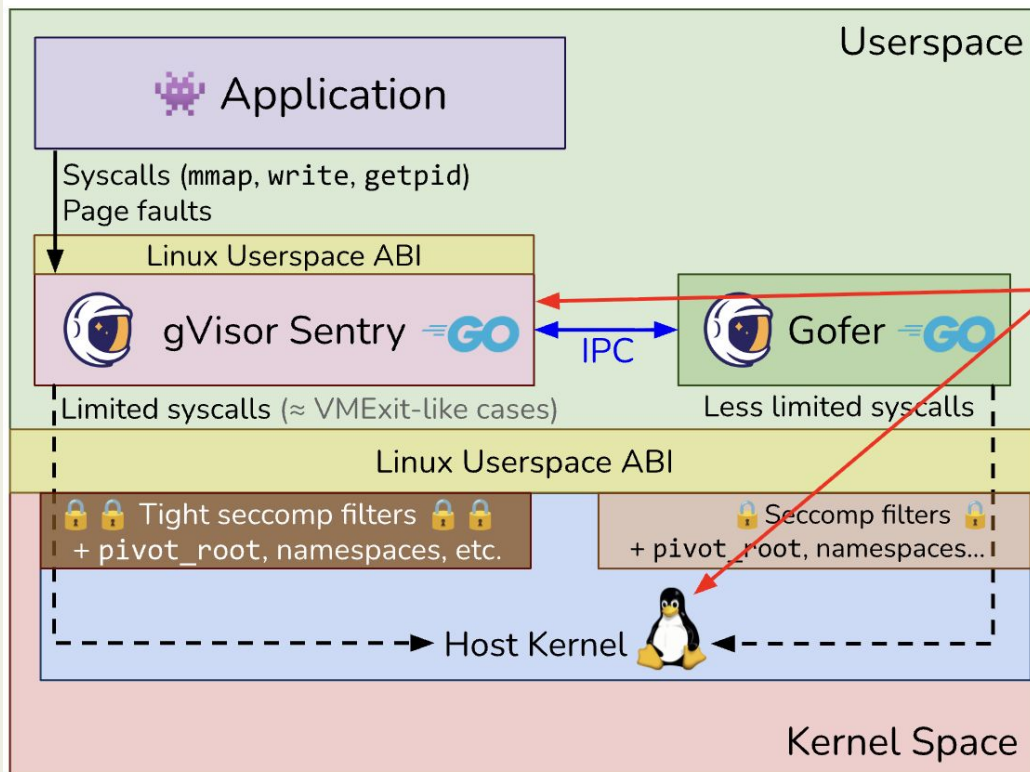
Audit logs

Security updates



Hardware virtualization: Intel VT-x / AMD-V





- 🛡️ Many more layers of defense.
- 🐧 **Distinct** kernel implementation
- 💡 Need to break both kernels!
- 🏋️ No hardware virt. required
- 📦 No workload-specific tailoring
- ✨ Fewer kernel-invasive parts



k8s: Kata Containers

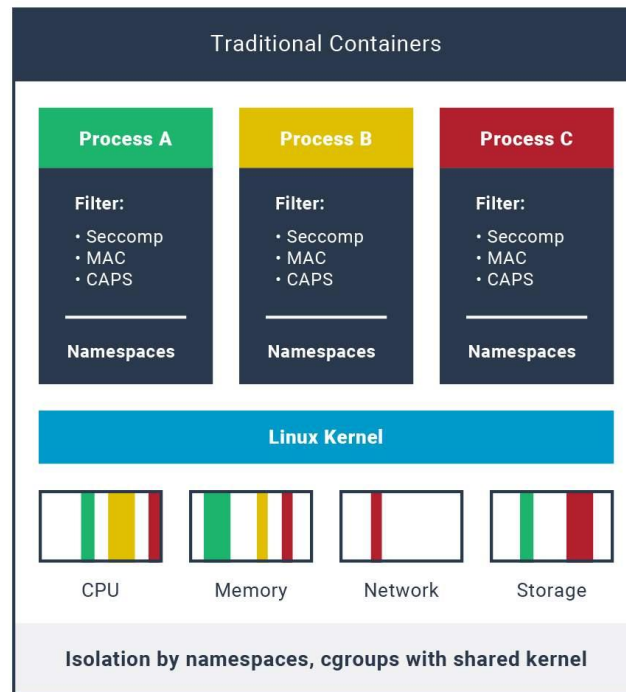
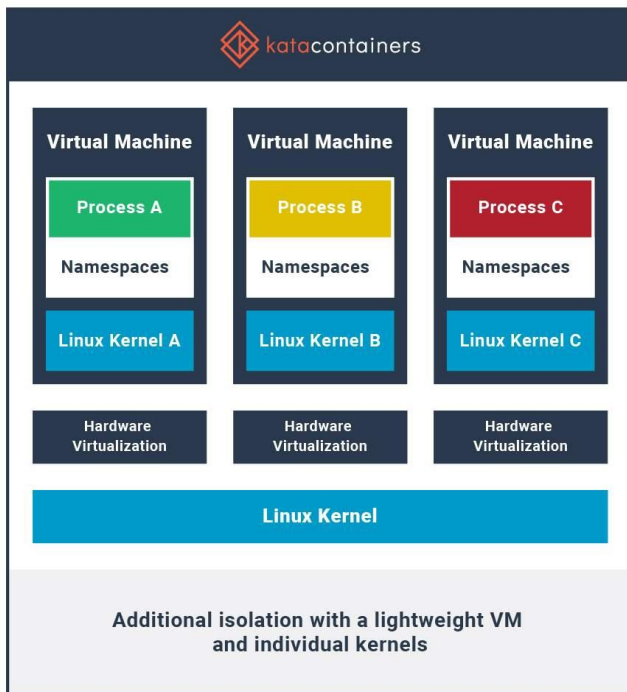


KubeCon



CloudNativeCon

India 2026



k8s: Agent Sandbox

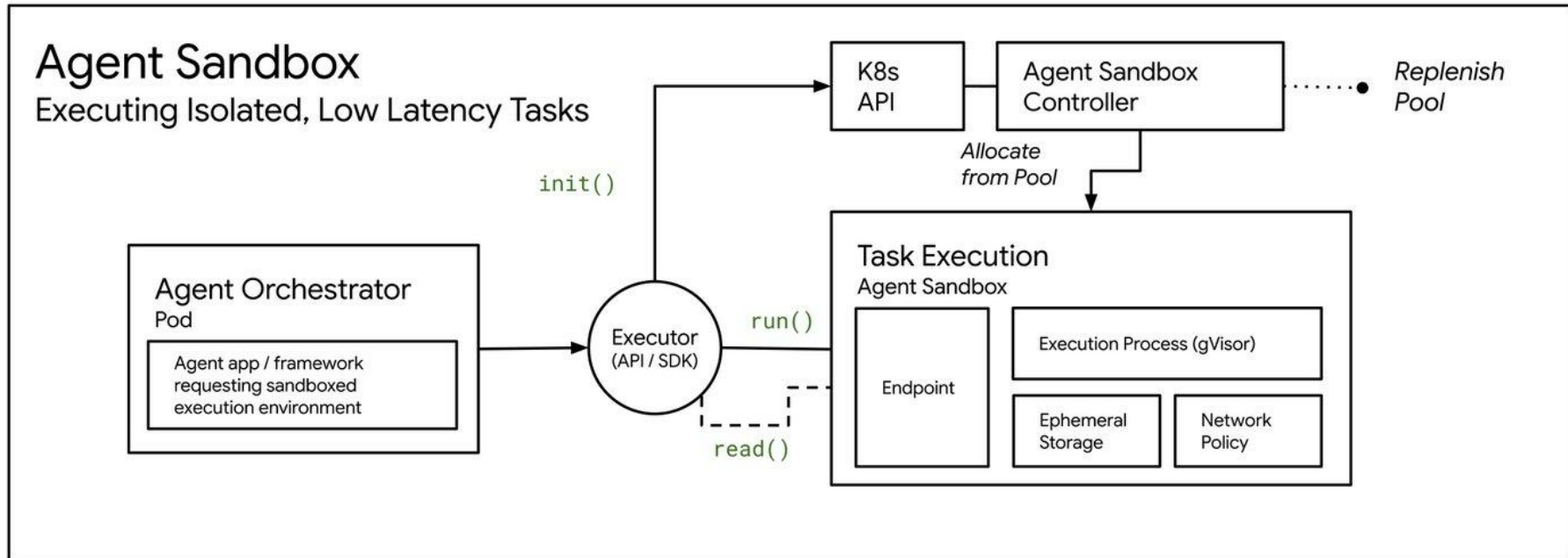


KubeCon



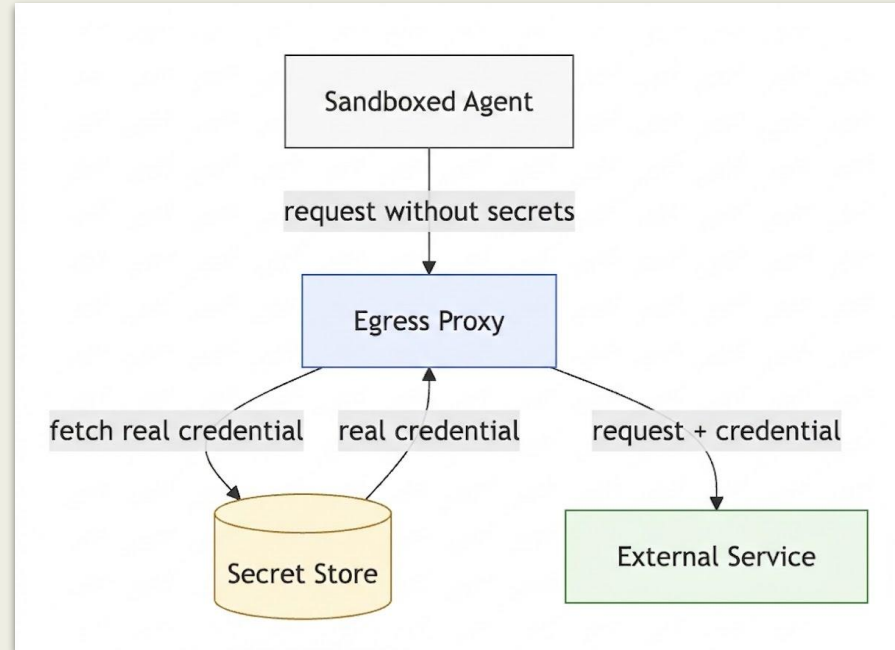
CloudNativeCon

India 2026



Proxy: Broker Secrets

- Inject fake credentials into the sandbox
- Egress Proxy needs access to real creds
- Define rules per sandbox
- Example: Envoy + Credential Injector



Proxy: Tokenize Secrets



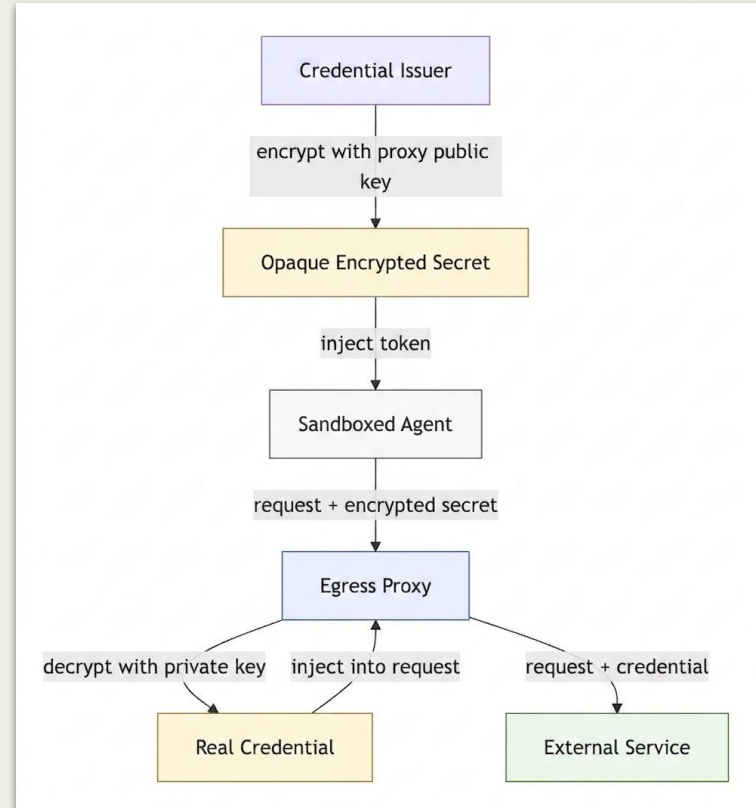
KubeCon



CloudNativeCon

India 2026

- Sandbox sees opaque encrypted token payload
- Egress Proxy becomes stateless
- Behaviour encoded in the payload
- Con: format validation may fail



Proxying Tools



KubeCon



CloudNativeCon

India 2026

- Envoy Proxy + credential_injector
- Mitmproxy (TLS-terminating)
- Squid: caching proxy with access control lists
- LiteLLM: gateway + rate limiting
- Tokenizer proxy: superfly/tokenizer

Redirection

- Base URLs
- HTTP_PROXY
- proxychain
- iptables
- httpjail



Or... just use MCP or Custom Tools



KubeCon



CloudNativeCon

India 2026

- Tool calls a service/proxy which injects the credentials
- Pros:
 - No TLS interception needed
 - Credentials out: agent only sees the tool interface
- Cons:
 - Host/maintain MCP servers
 - MCP Authorization still needed



Takeaways



KubeCon



CloudNativeCon

India 2026

- **Compute**

- MicroVMs > gVisor > containers/jails
- Cold start is mostly irrelevant; use snapshots
- k8s: Kata Containers, AgentSandbox

- **Ideal sandbox pattern: container-in-a-VM**

- Container for DX, VM for isolation

- **Protecting Secrets**

- Assume exfil is always possible
- Broker/tokenize short-lived secrets/tokens
- k8s: Envoy + credential_injector_filter

Sandbox Providers

Vercel

E2B

Modal

Daytona

CloudFlare

North Flank

Fly.io Sprites

Docker Sandboxes





KubeCon



CloudNativeCon

India 2026



LLMs Behind Bars: Sandboxes at Scale for AI on a Short Leash

Feedback

Thank
You

