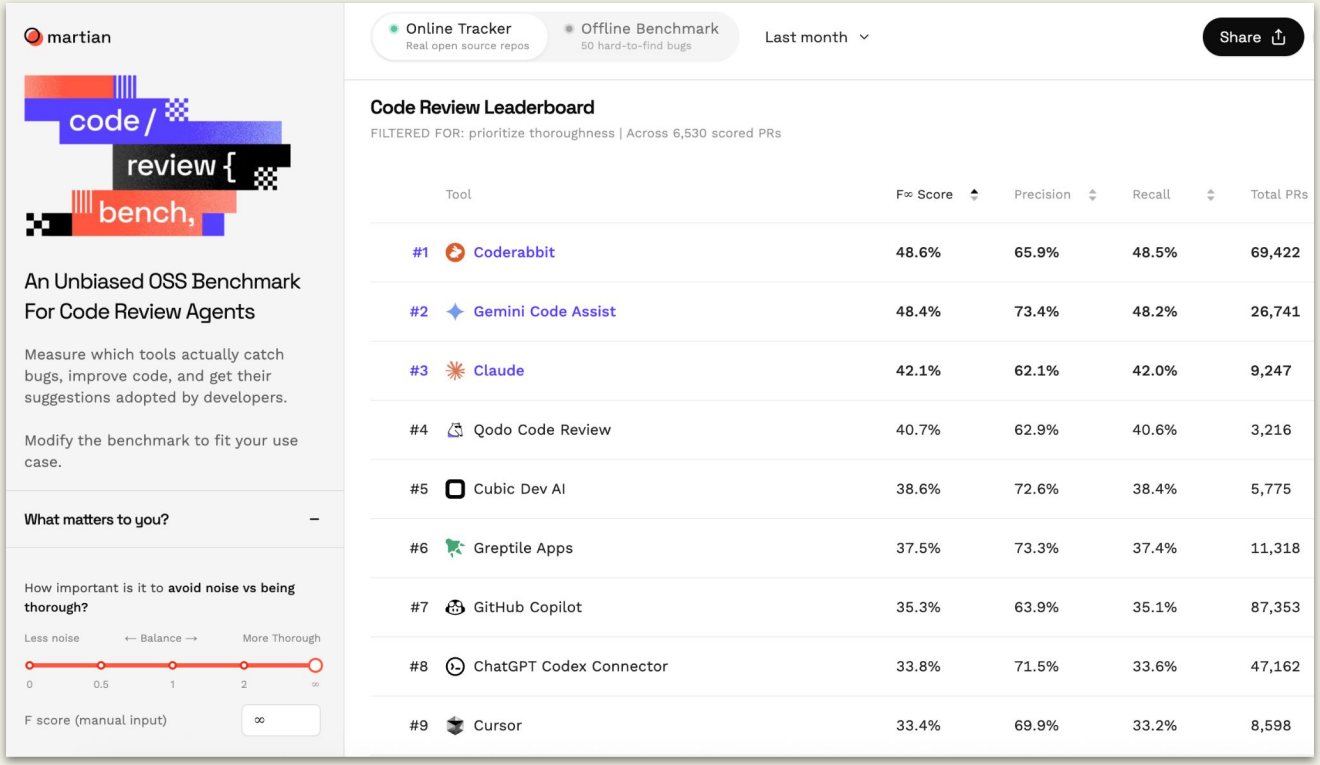
KubeCon CloudNativeCon India 2026

LLMs Behind Bars: Sandboxes at Scale

Prashanth Pai
CodeRabbit

CodeRabbit

KubeCon CloudNativeCon India 2026



https://codereview.withmartian.com/?beta=10

150,000+
Monthly active users

12M+
Pull requests reviewed

2M+
Installs across repositories

Demo

- 🔗 Code Reviews
- 🔗 Autofix
- 🔗 Review Chat
- 🔗 Pre-Merge Checks
- 🔗 Agent for Slack

Sandbox?

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

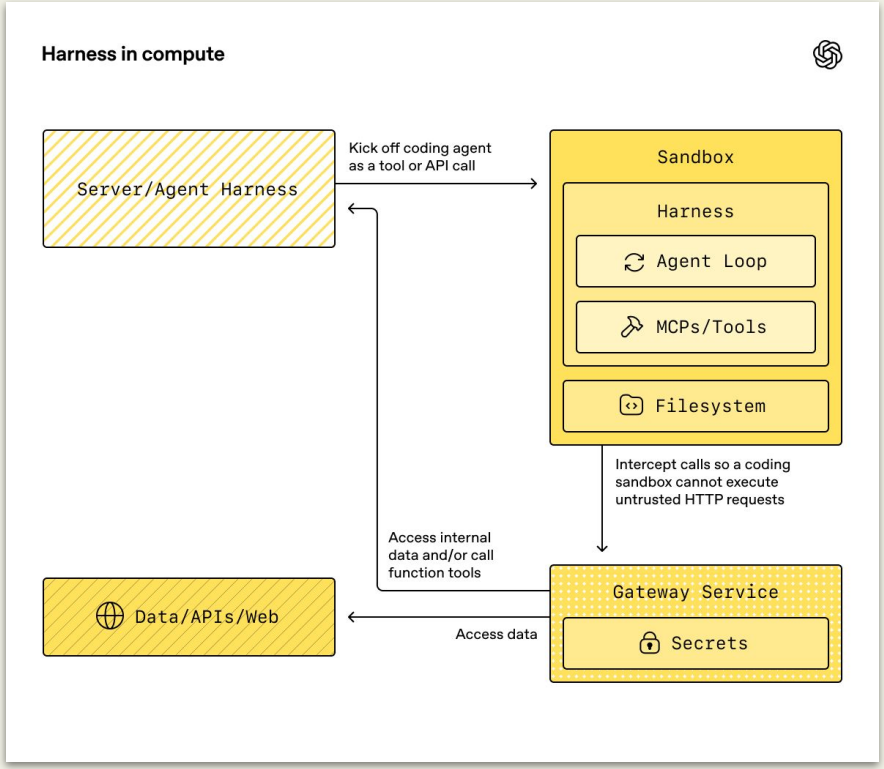
Local Laptop
Claude/Codex
OpenClaw

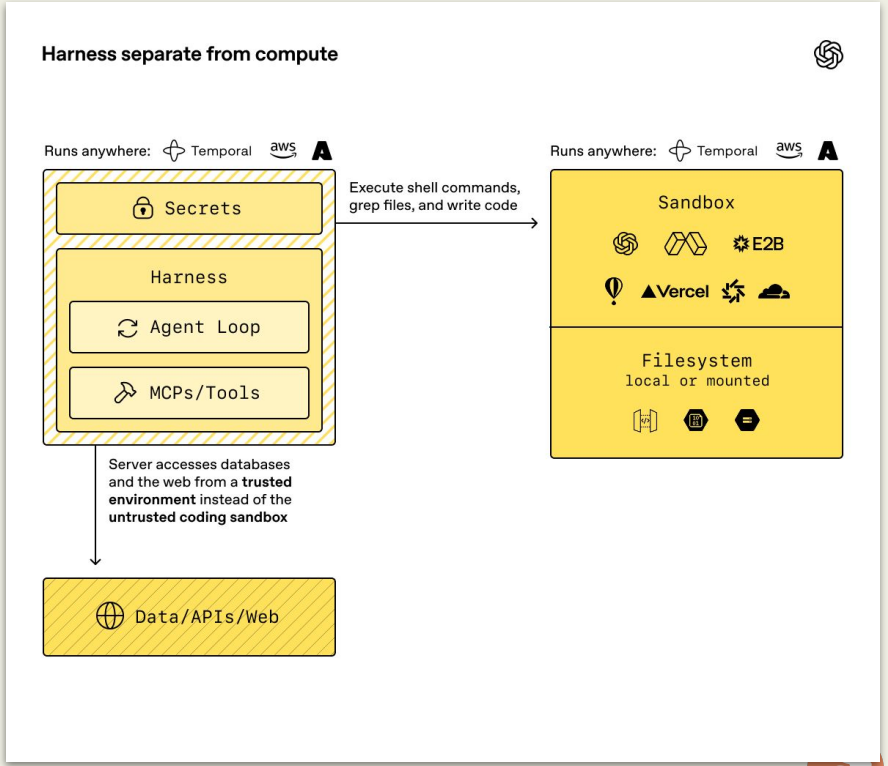
Remote Servers
Background Agents

Threat

- **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

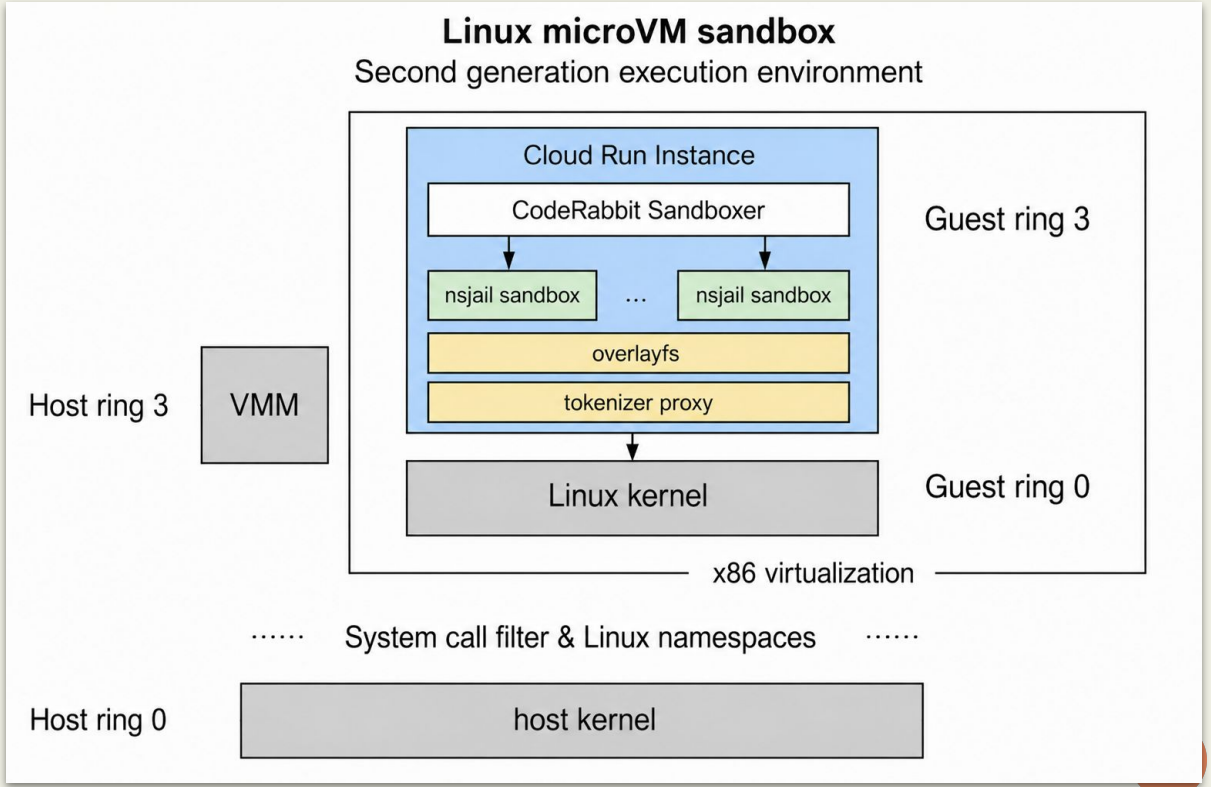




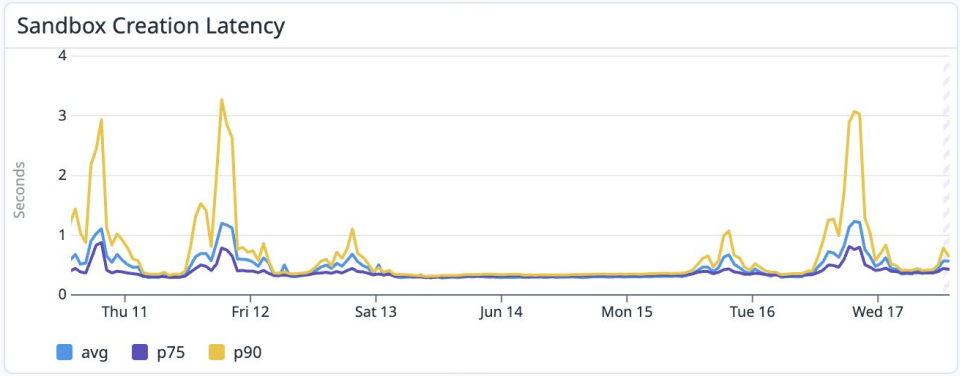
https://developers.openai.com/api/docs/guides/agents/sandboxes

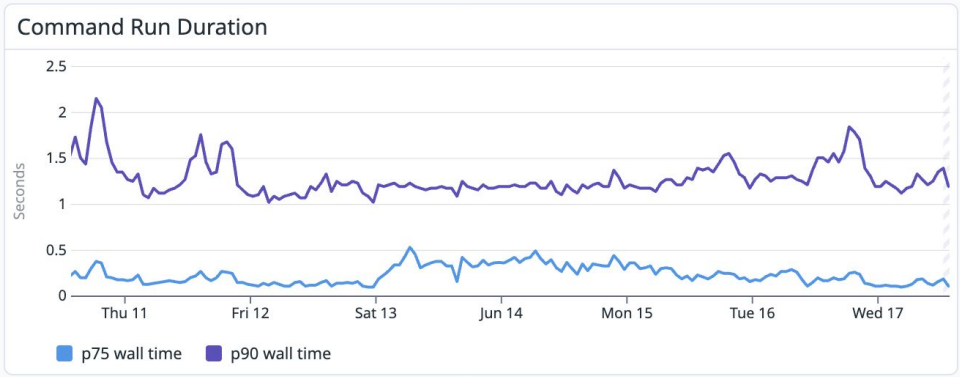
Sandboxes @ CodeRabbit

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



Latency



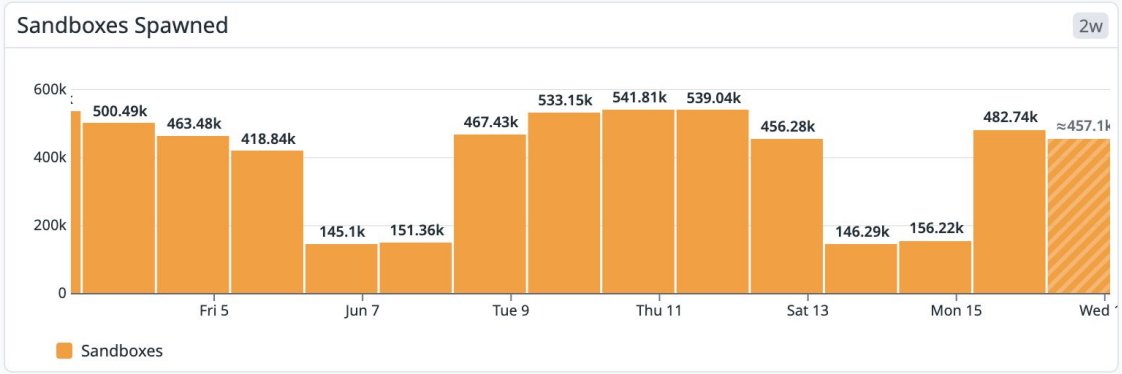


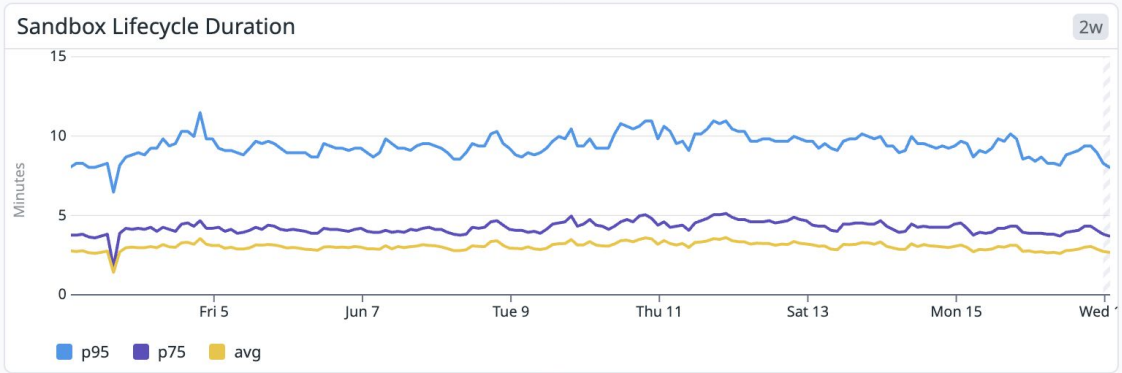
~700ms
sandbox startup

50+
third-party tools

~250ms
most cmd runs

Scale





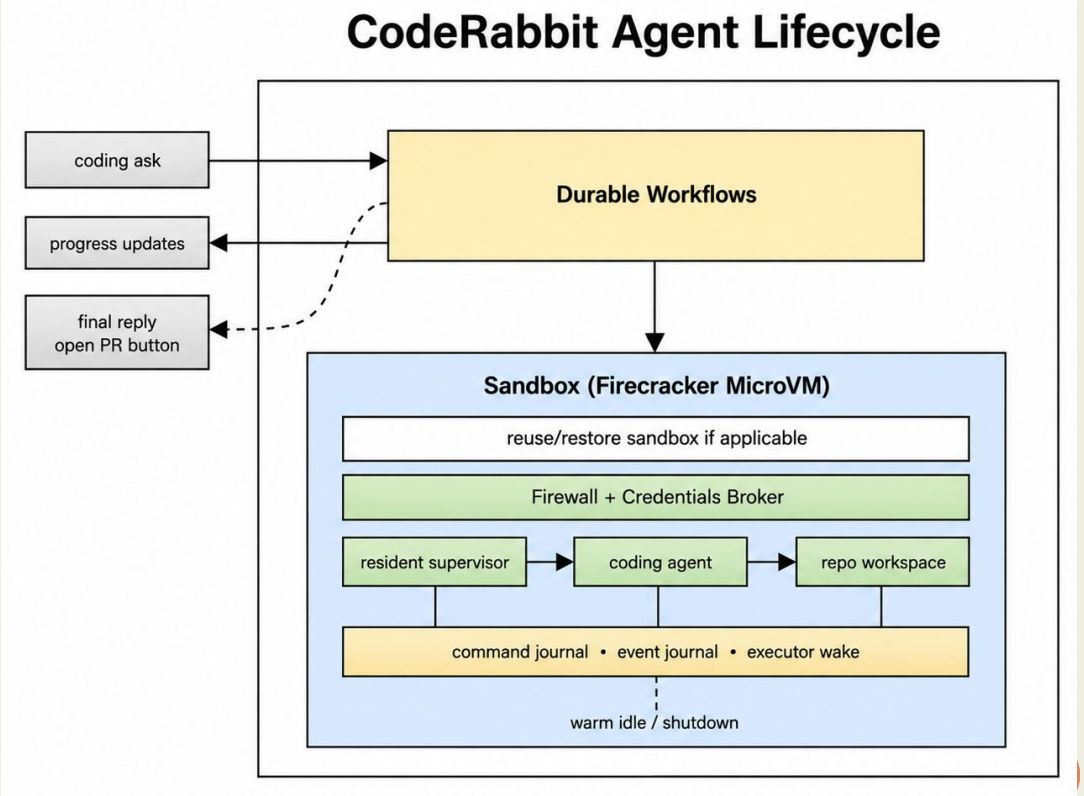
500k
sandboxes/day

Peaks at 33k
sandboxes/hour

Lifecycle duration
3–10 mins

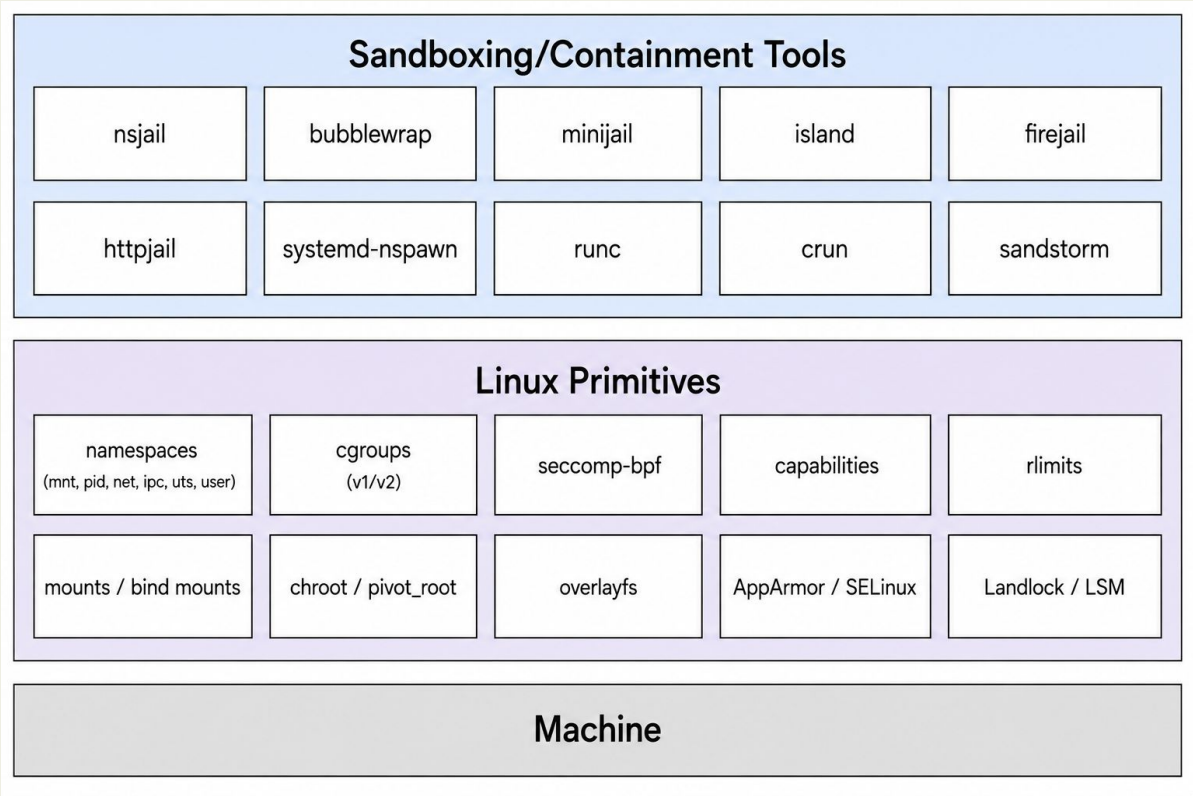
Autonomous Agents

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

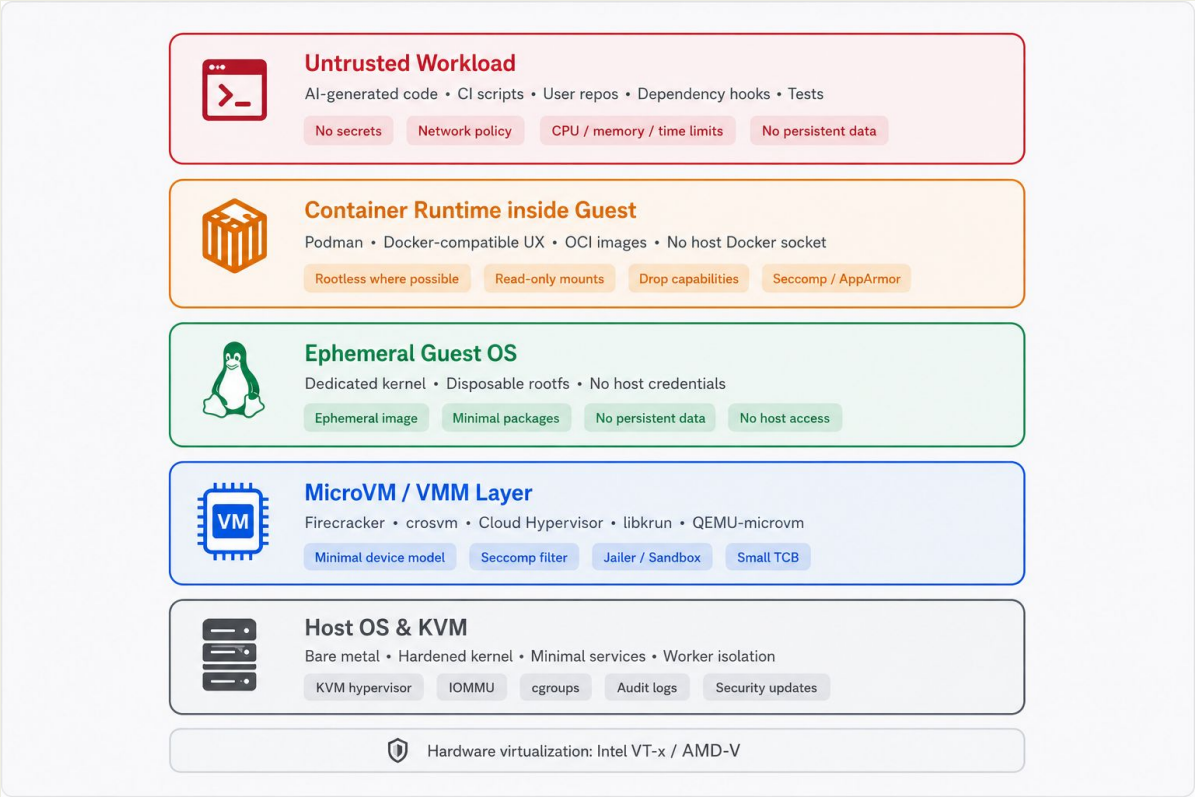


What makes up a Sandbox?

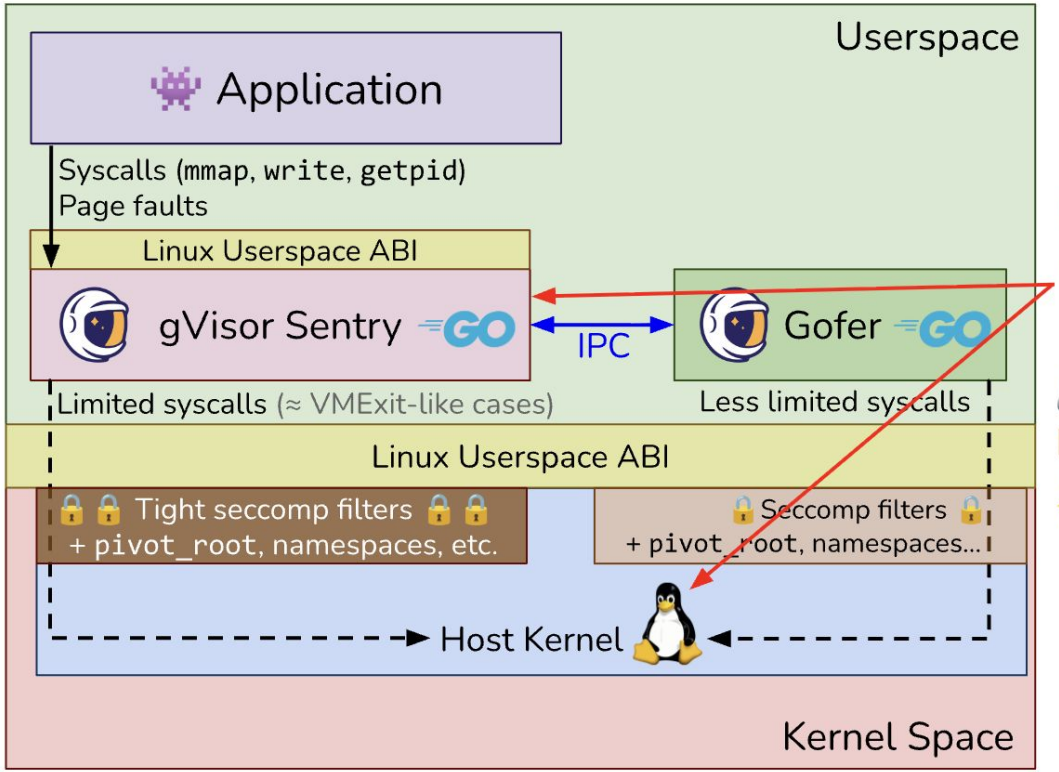
Containers are legos



MicroVMs

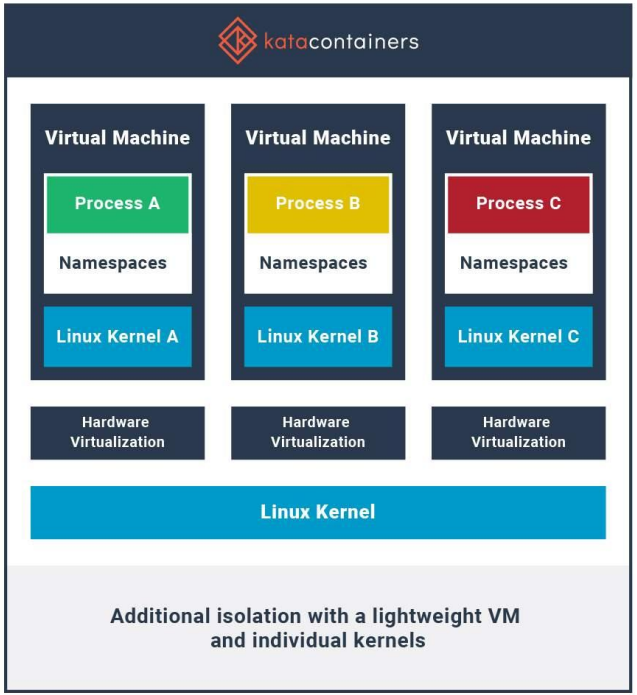

Untrusted Workload
AI-generated code • CI scripts • User repos • Dependency hooks • Tests
No secrets
Network policy
CPU / memory / time limits
No persistent data
Container Runtime inside Guest
Podman • Docker-compatible UX • OCI images • No host Docker socket
Rootless where possible
Read-only mounts
Drop capabilities
Seccomp / AppArmor
Ephemeral Guest OS
Dedicated kernel • Disposable rootfs • No host credentials
Ephemeral image
Minimal packages
No persistent data
No host access
VM
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

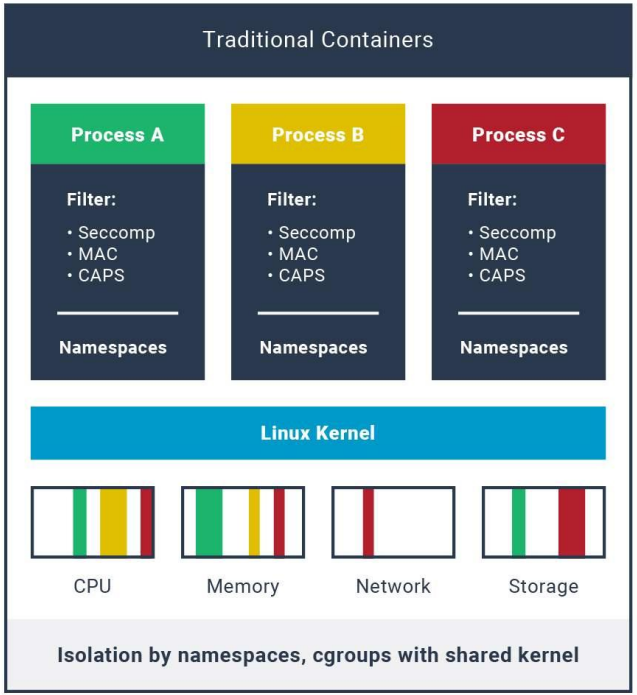
gVisor



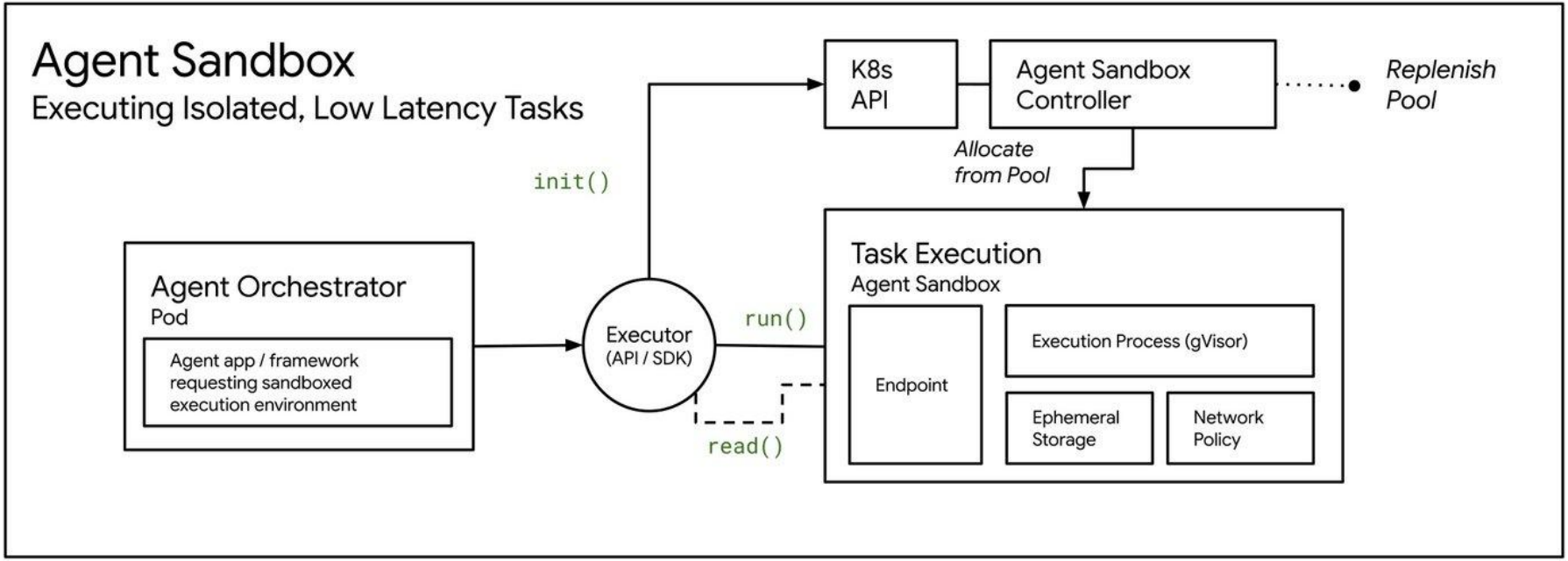
- Many more layers of defense.
- **Distinct** kernel implementation
  - Need to break both kernels!
- No hardware virt. required
- No workload-specific tailoring
- Fewer kernel-invasive parts

k8s: Kata Containers



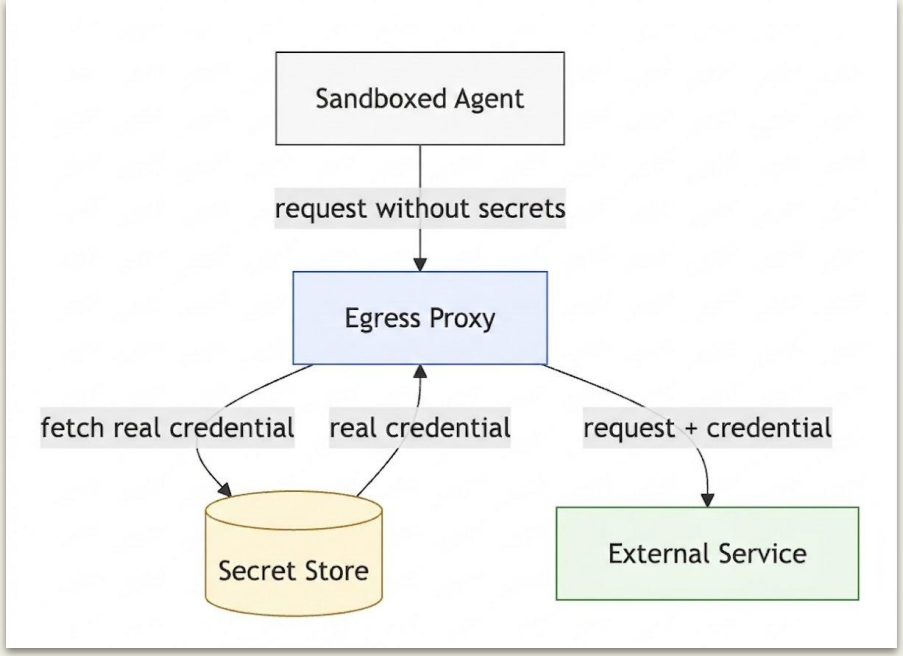


k8s: Agent Sandbox



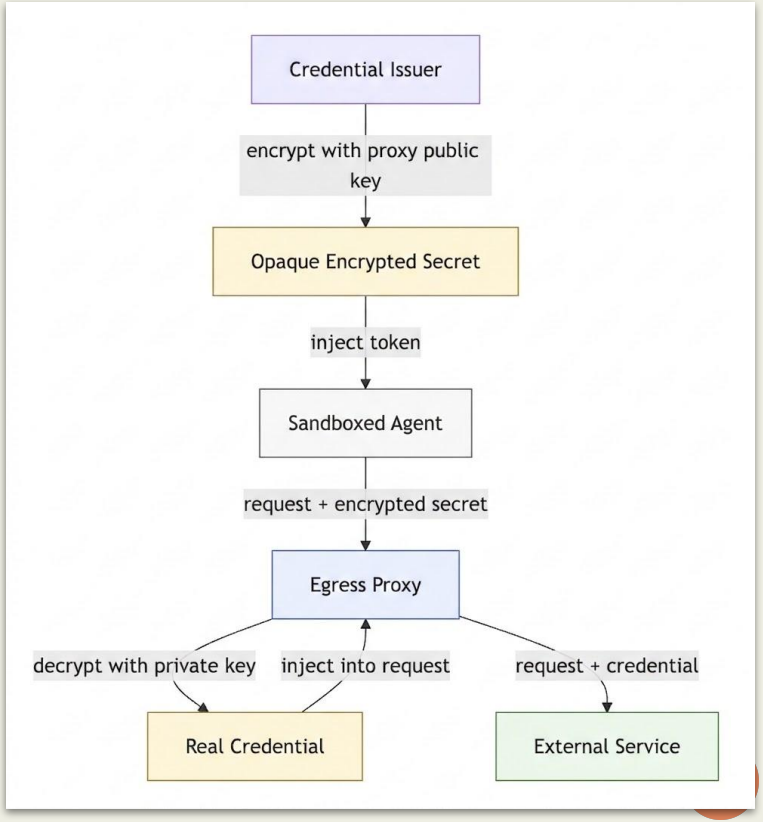
Proxy: Broker Secrets

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



Proxy: Tokenize Secrets

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



Proxying Tools

- 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

- 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

- **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

Feedback

Thank You