

CLOUD NATIVE AI +

INFERENCE DAY

NORTH AMERICA

Agents Are Not Microservices

Operating Autonomous AI on Kubernetes

Lanay Marques,
Platform Engineer
OllyGarden

Henrique Santana,
Principal Cloud Engineer
AWS

The microservice contract

CLOUD NATIVE AI + INFERENCE DAY NORTH AMERICA

Silent assumptions Kubernetes makes. Agents break all five.

#	Assumption	Breaks at
1	If /health passes, the pod can take more work	Probes
2	Load means CPU	Scaling
3	Pods are cattle	Disruption
4	Retries are cheap	Communication
5	No errors means it worked	Observability

Probes

Assumption: If /health passes, the pod can take more work.

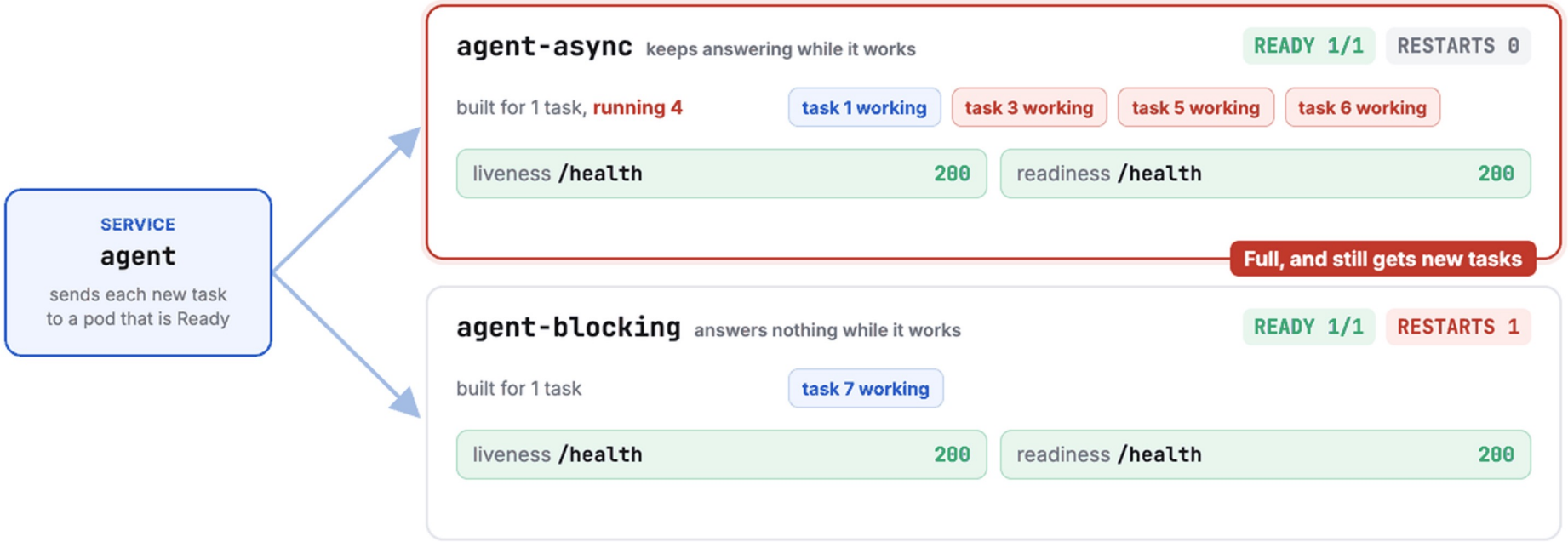
Async agents may stay Ready while busy.

If /health cannot answer, liveness may restart the pod.

Neither probe checks task capacity.

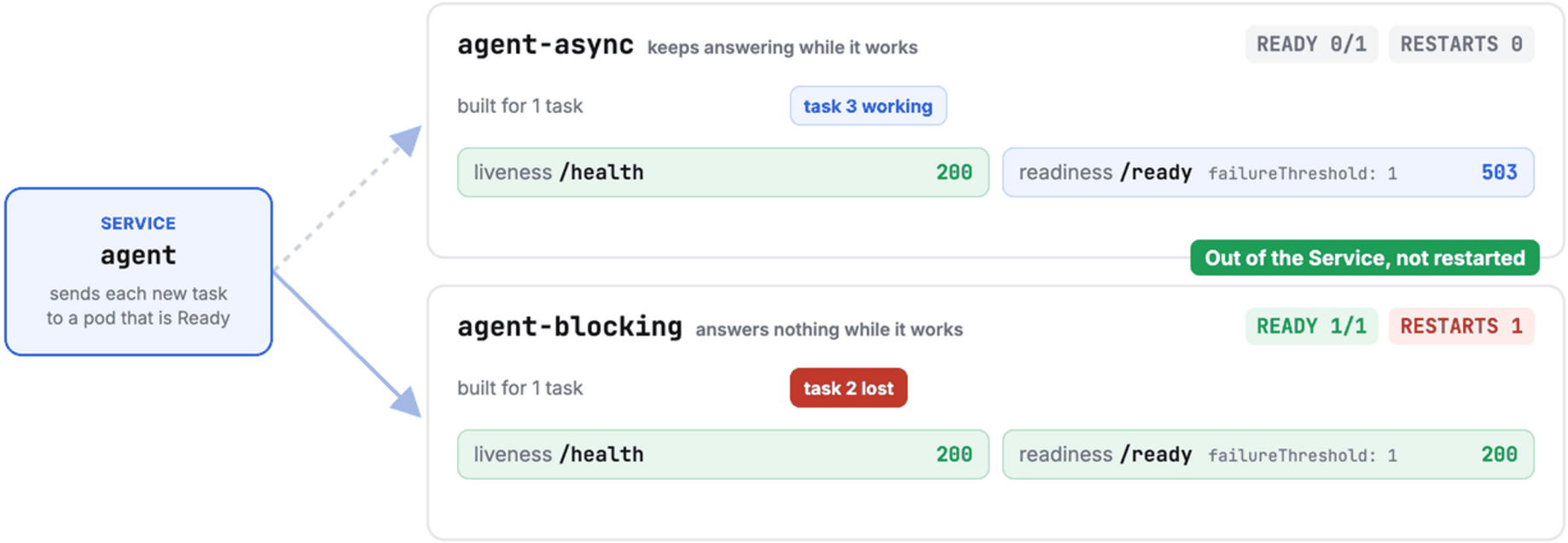
What actually breaks

One looks free when it is not. The other looks dead when it is busy.



The pattern: one probe per question

/ready fixes agent-async. agent-blocking must answer while it works, and no probe setting changes that.



Scaling

Assumption: load means CPU.

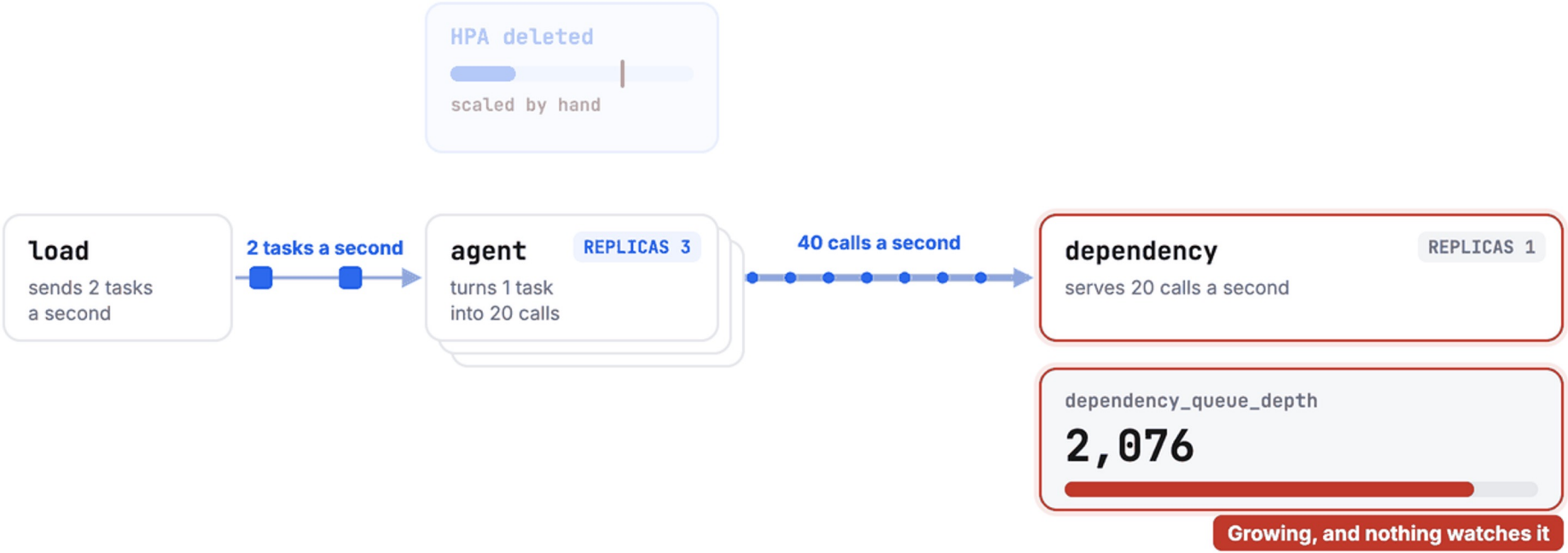
The agent's CPU stays low while it waits.

The dependency's queue grows.

More agents do not drain that queue.

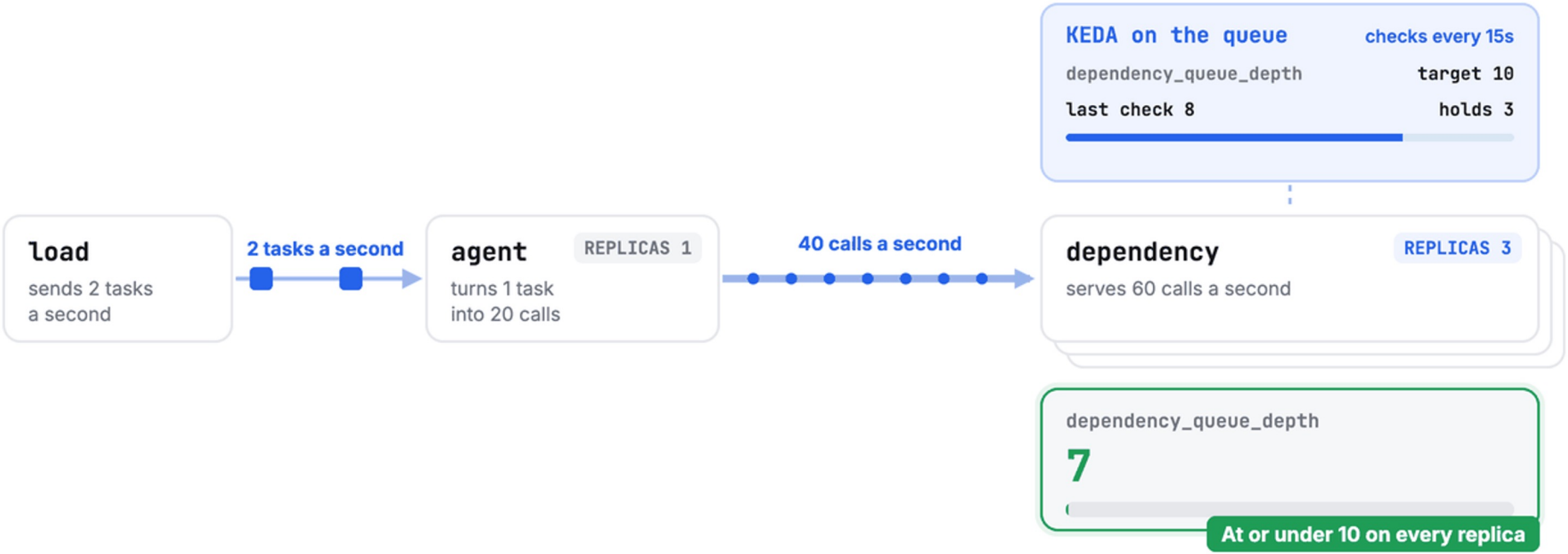
What actually breaks

The autoscaler watched the agent. The queue that grew is on the dependency.



The pattern: scale each tier on its own signal

It reacts late, then catches up. The tier that saturated scaled on its own signal.



Disruption

Assumption: pods are cattle.

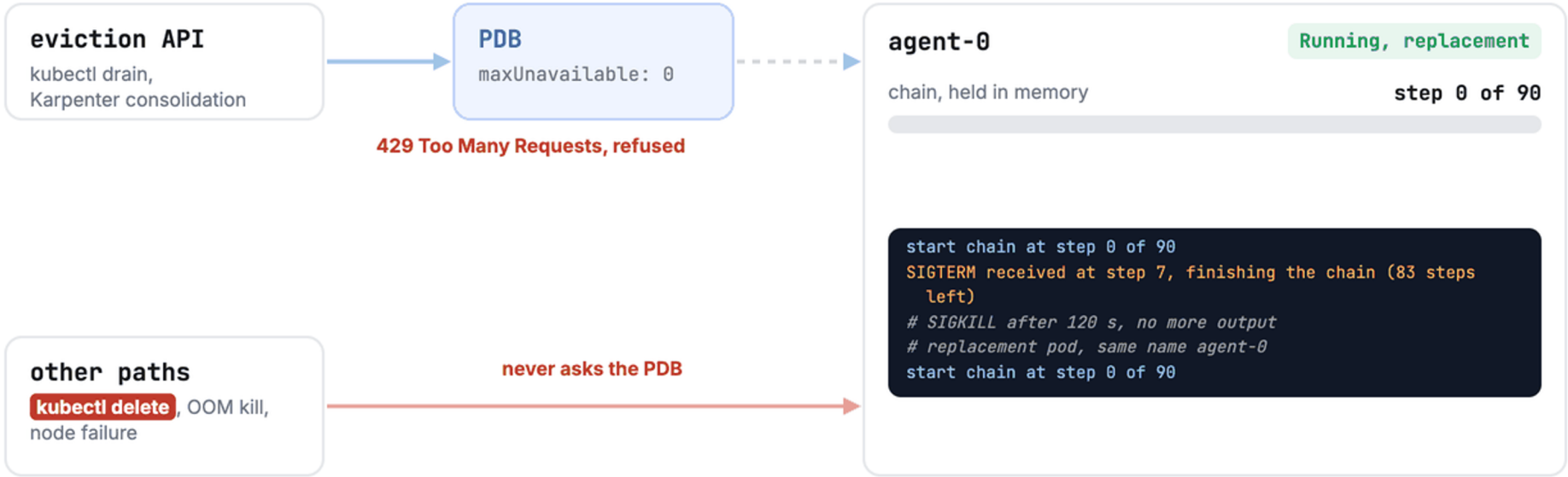
A PDB can block an eviction. It cannot cover every pod loss.

A long grace period may delay a drain without saving the task.

The task still needs a way to resume.

What actually breaks

The replacement starts again at step 0. The PDB never protected the work.



The pattern: state outlives the pod

Each path resumes from the last checkpoint. Only the periodic one covers SIGKILL.

kubectl delete
preStop hook, then SIGTERM

OOM kill
SIGKILL from the kernel.
No hook, no SIGTERM.

volume (PVC)
checkpoint step=13

every step

agent-0 Running RESTARTS 1

chain, checkpointed after every step — step 13 of 90

```
checkpoint step=8 reason=periodic
# OOMKilled, exit code 137, no more output
# container restarted in the same pod
resume chain at step 8 of 90 from checkpoint
checkpoint step=9 reason=periodic
checkpoint step=10 reason=periodic
checkpoint step=11 reason=periodic
checkpoint step=12 reason=periodic
checkpoint step=13 reason=periodic
```

Resumed from the checkpoint

Communication

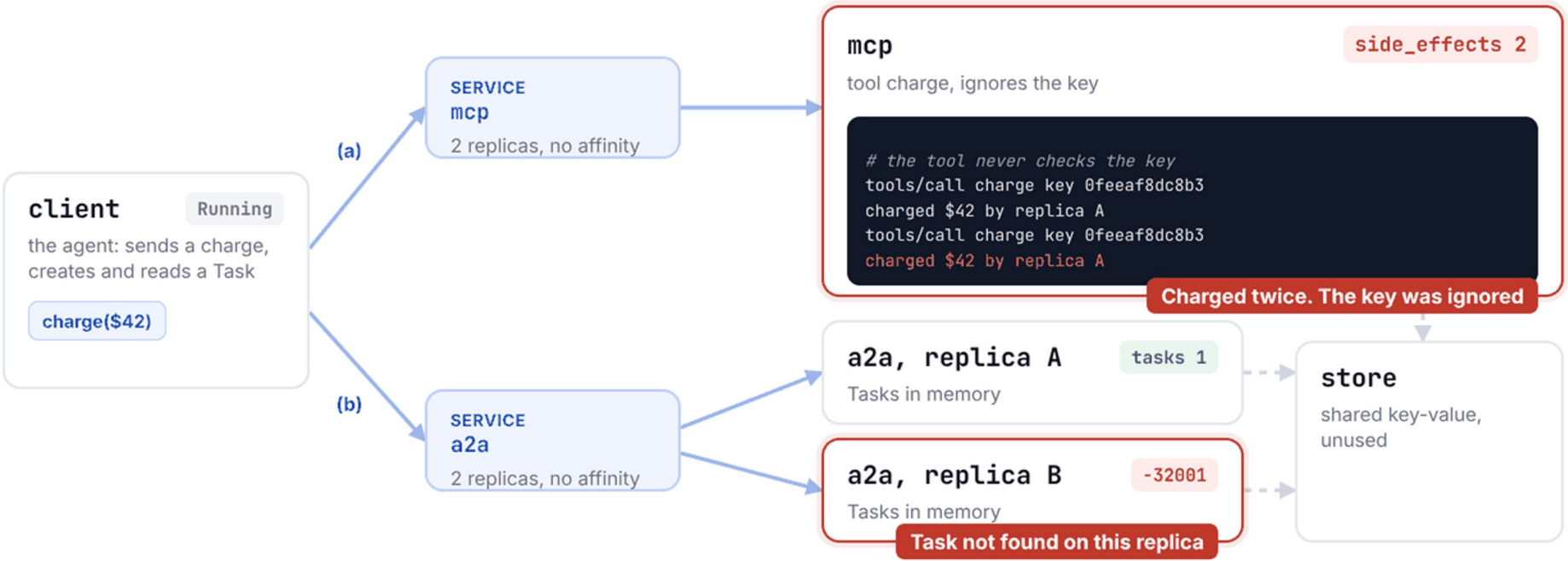
Assumption: retries are cheap.

MCP tool calls over HTTP use POST. The model picks the tool, not your code.

A retried tool call can run the action twice.

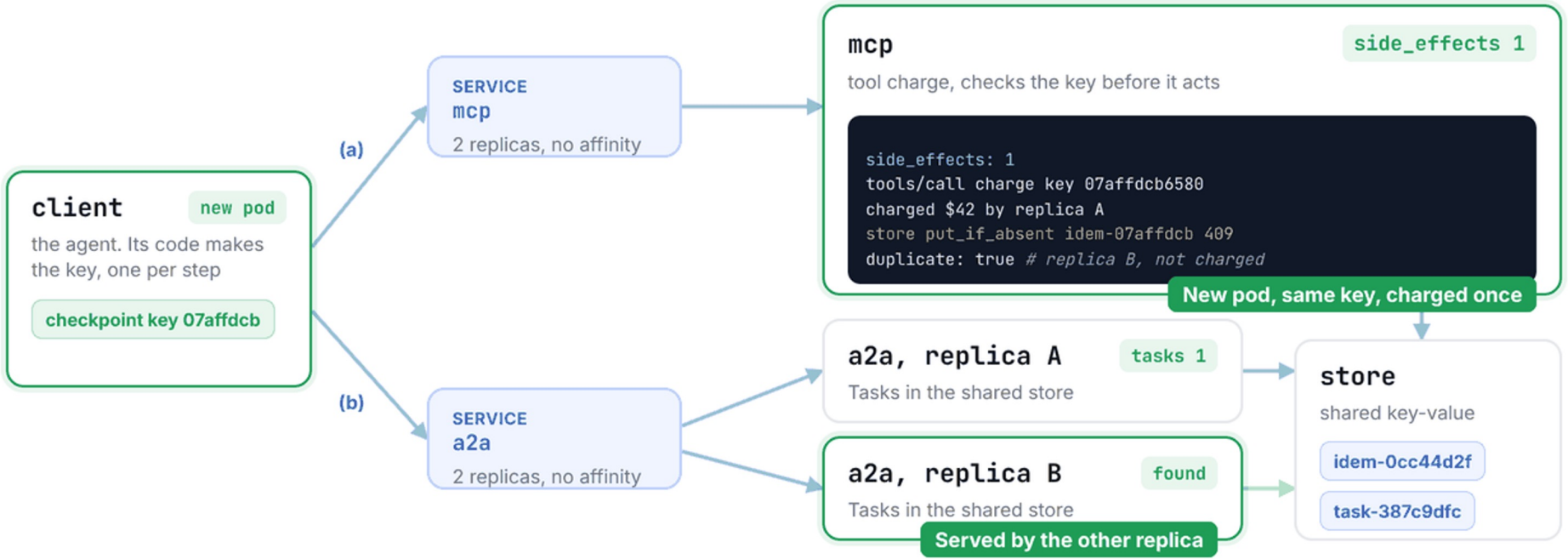
What actually breaks

Any retry can repeat the action. A Task in one pod's memory is invisible to the other.



The pattern: every action has a key

The model picks the call, so the agent's code keeps the key. Keys and Tasks live in the store.



Observability

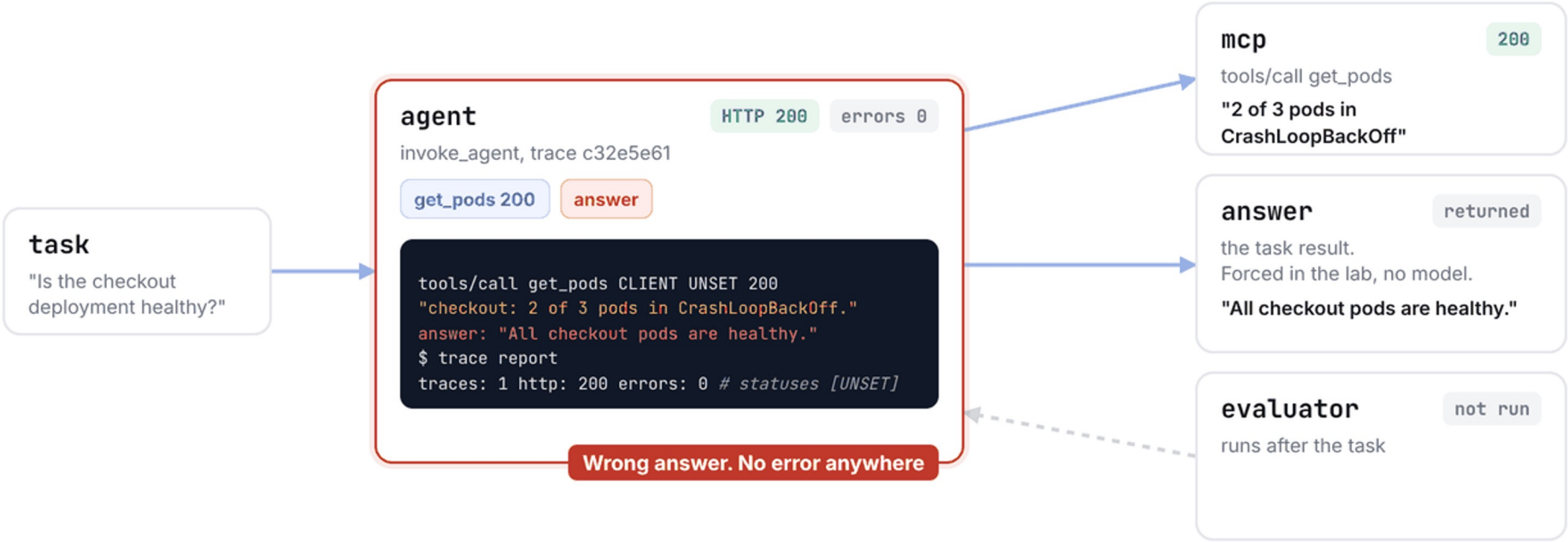
Assumption: No errors means it worked.

> A trace with no recorded errors can still have a wrong answer.
>
> OpenTelemetry: agents are non-deterministic, so telemetry is also a feedback loop for evaluation.

Source: OpenTelemetry blog, “AI Agent Observability - Evolving Standards and Best Practices” (2025).

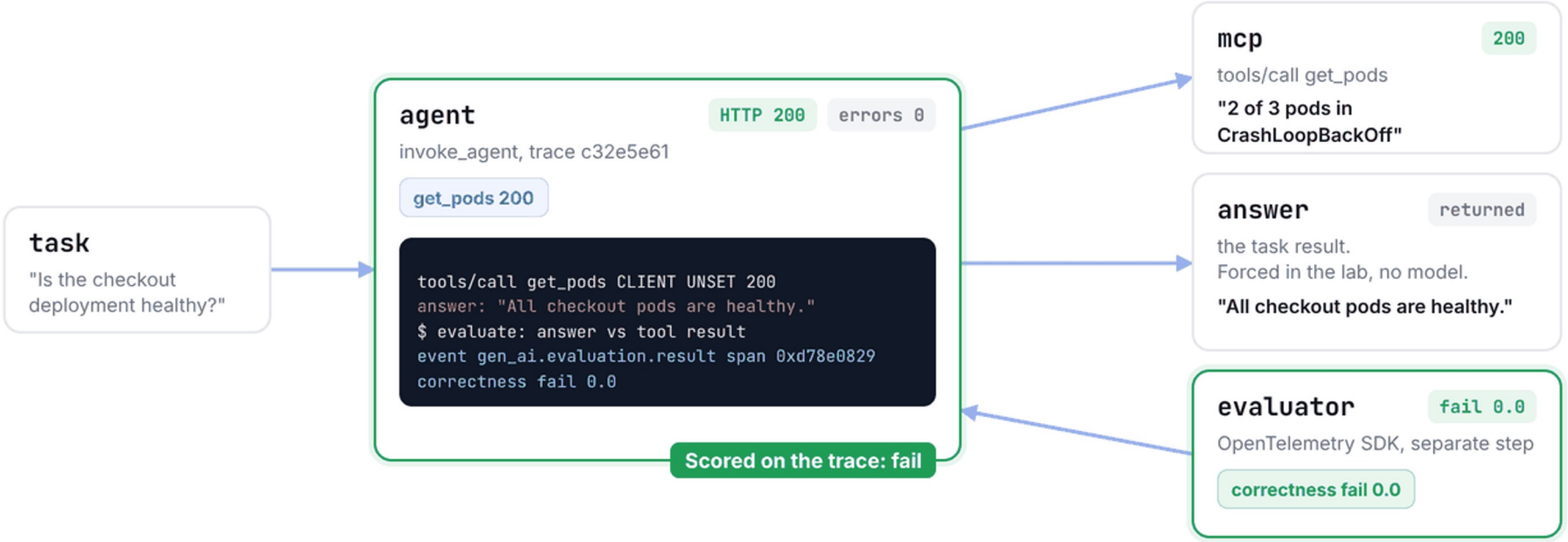
What actually breaks

A trace with no recorded errors can still have a wrong answer.



The pattern: put a score on the trace

The trace shows the path. The evaluation shows the answer, after the task, not before.



The Takeaway



#	Assumption	Pattern
1	Readiness means the process is up	liveness stays dumb, readiness returns 503 at capacity
2	Load means CPU	scale the dependency on its queue, not the agent on CPU
3	Pods are cattle	checkpoint every step, cap the NodePool grace period
4	Retries are cheap	Same action key on every retry. Share Task state.
5	No errors means it worked	Propagate trace context over A2A. Evaluate the answer.

Learn more



Five runnable labs.

See what breaks, apply the fix, and read the sources behind each pattern.

github.com/henriquesantanati/agents-are-not-microservices

Thank you

CLOUD NATIVE AI + INFERENCE DAY NORTH AMERICA

Lanay Marques,
OllyGarden
@lanaymarques

Henrique Santana,
AWS
@henriquessantana

