KubeCon CloudNativeCon North America 2026

Verkada

Tackling Alert Fatigue:

How Verkada Triages 4,700 Alerts/Month at $0.33

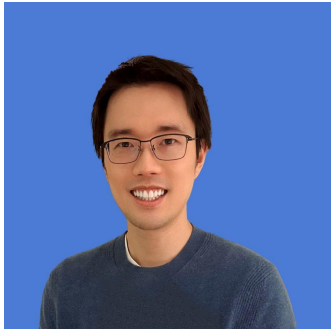
Rohan Sood and Adrian Ng, Verkada

10 November 2026, 3:40 PM, Salt Palace, Level 1, Room 151



Presenters

Verkada



Adrian Ng

Director of Engineering,
Platform Infrastructure



Rohan Sood

Senior Infrastructure
Engineer

Agenda

Verkada

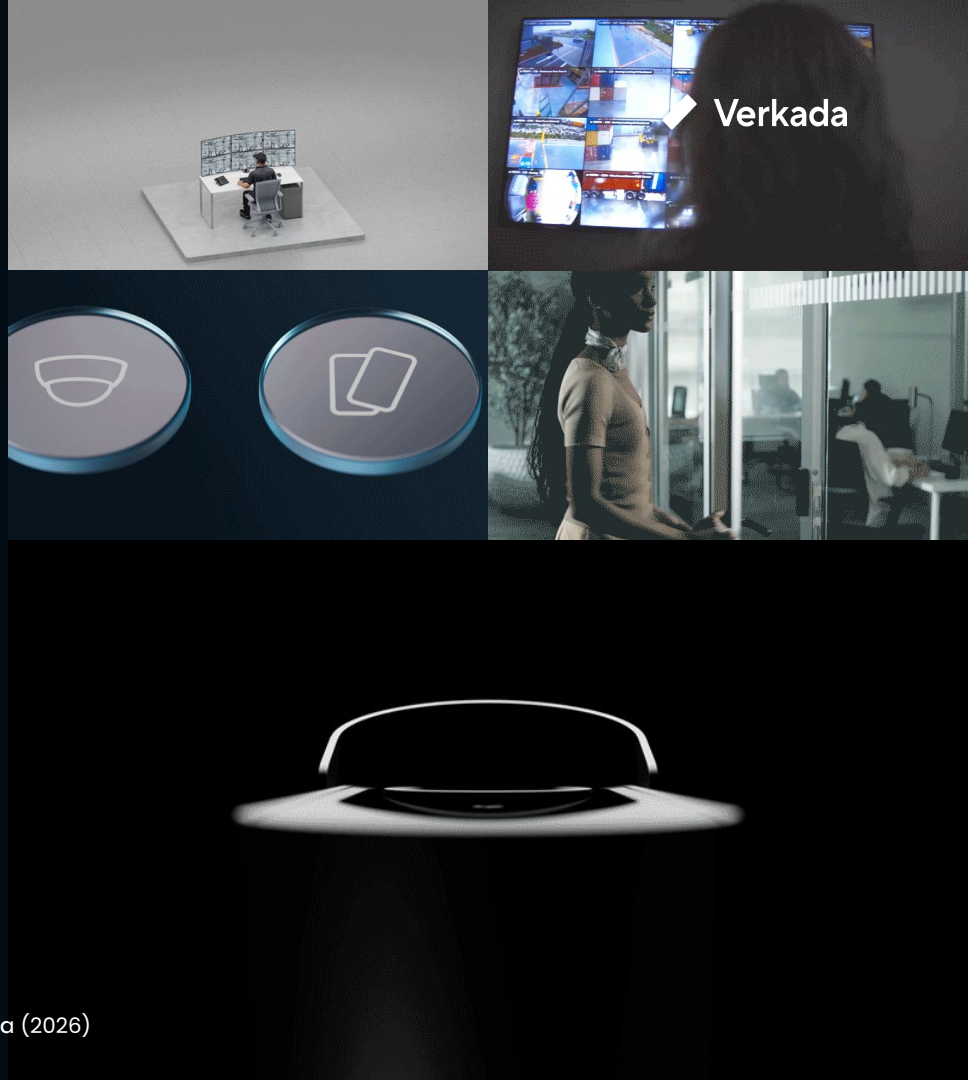
1	**The problem**	Our scale, alert fatigue
2	**How it works**	One alert, step by step
3	**High accuracy**	Rules and nightly evals
4	**Incident mode**	Then the results
5	**Lesson learned**	Where state lives

Who we are

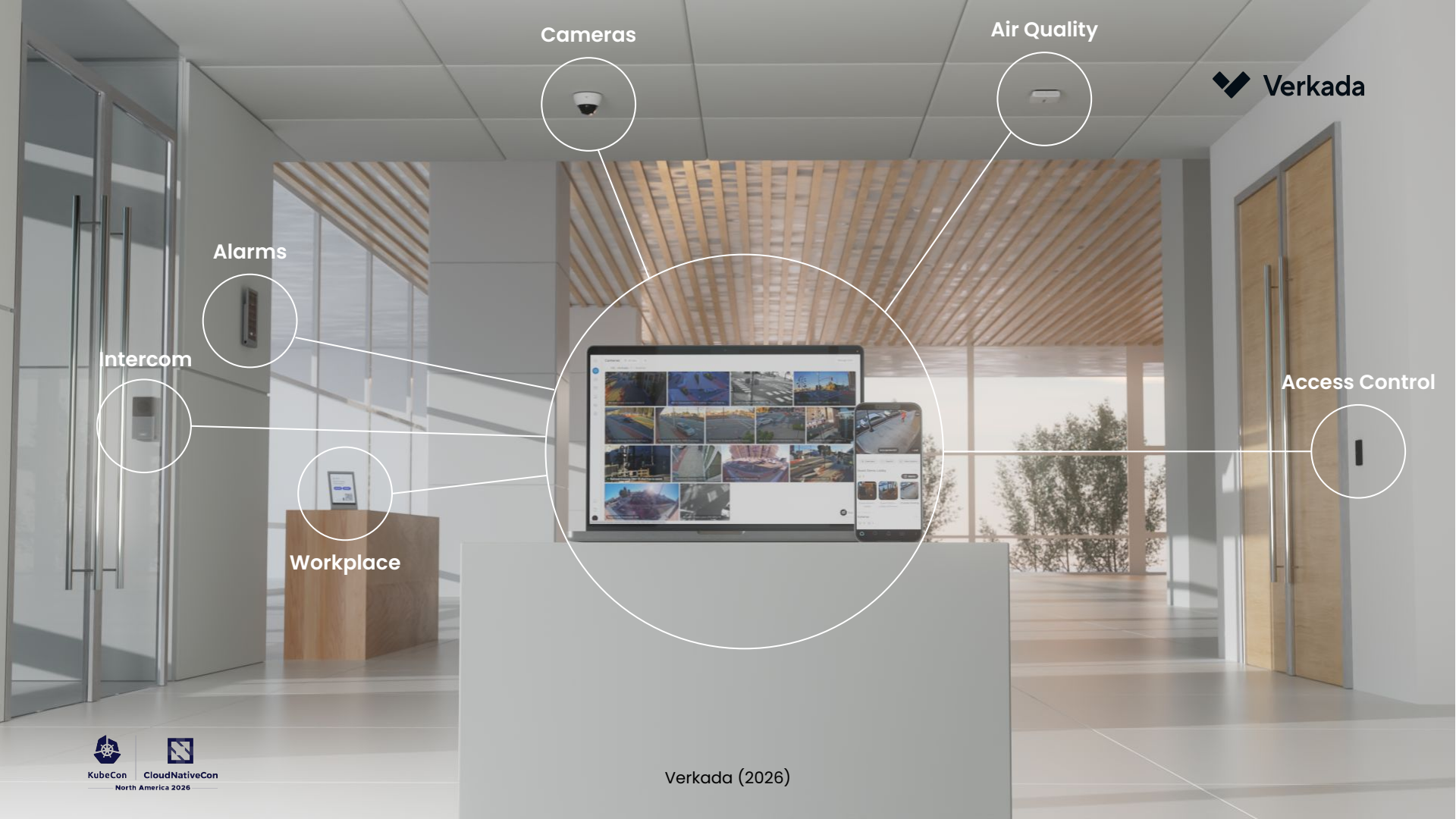
Purpose

Protect people and places in a privacy-sensitive way.

We've built an integrated platform for cloud-managed physical security.

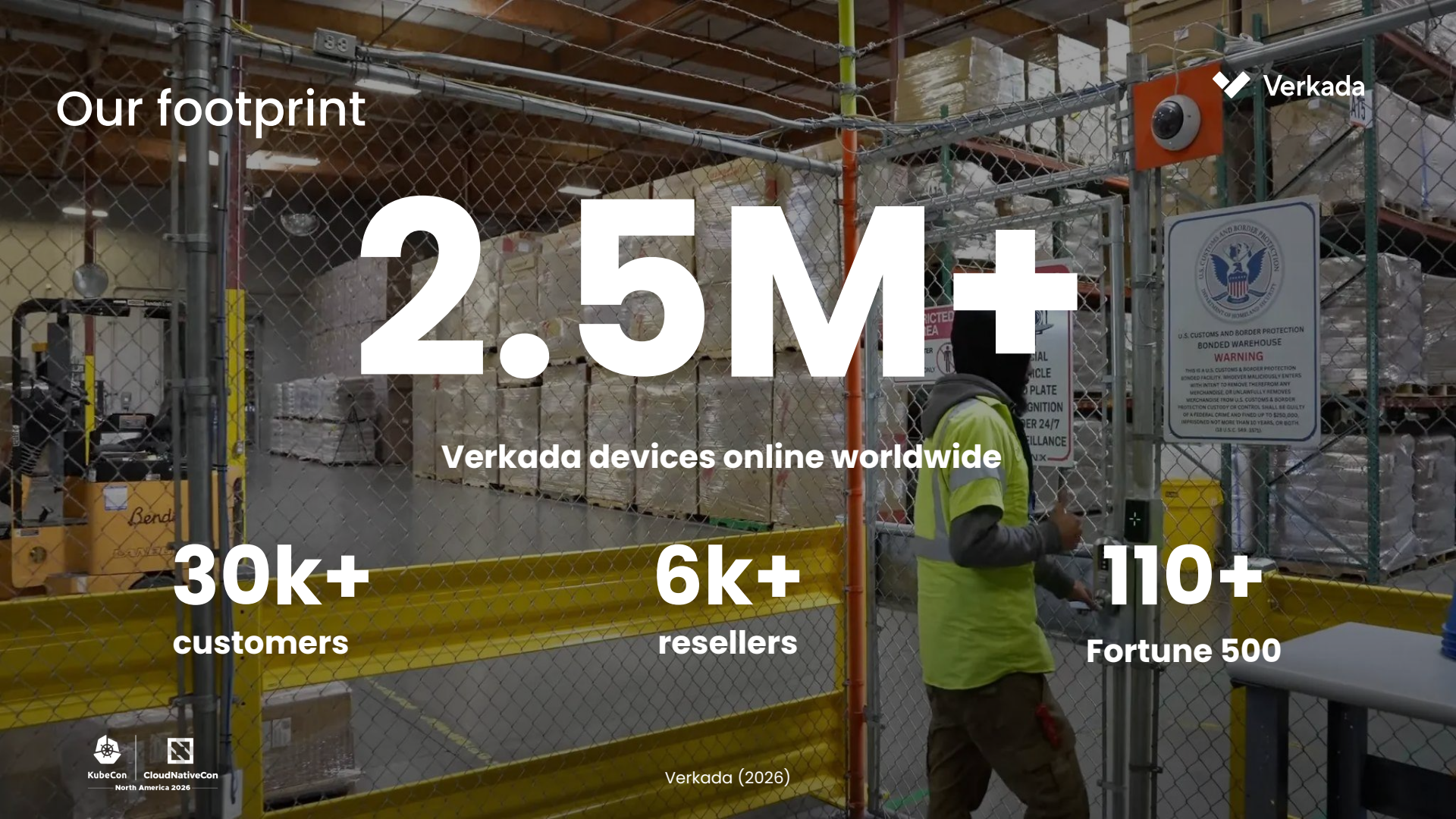


Verkada (2026)



KubeCon CloudNativeCon North America 2026

Verkada (2026)


Verkada
Our footprint
2.5M+
Verkada devices online worldwide
30k+
customers
6k+
resellers
110+
Fortune 500
U.S. CUSTOMS AND BORDER PROTECTION
BONDED WAREHOUSE
WARNING
KubeCon
CloudNativeCon
North America 2026
Verkada (2026)

Our scale

Verkada

50K

People/vehicle detection ingested per second

~1M

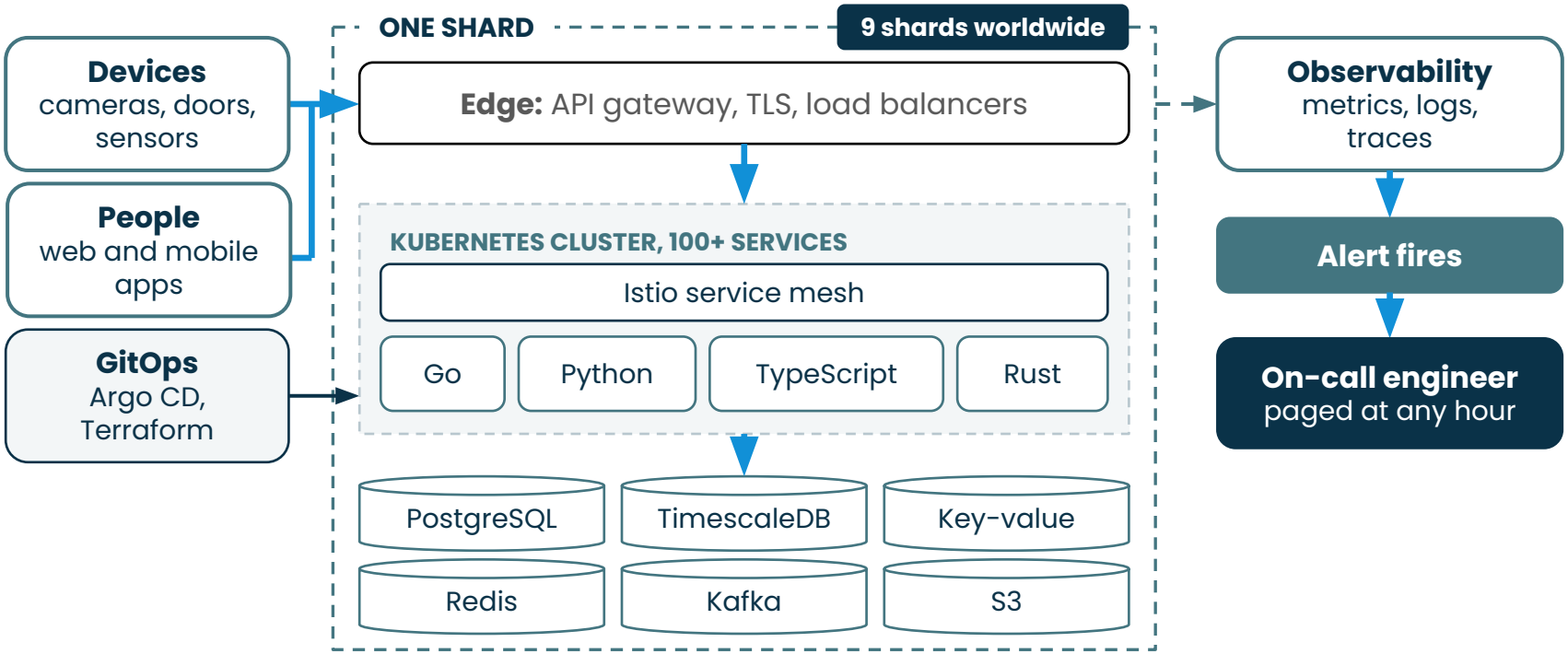
HTTP requests to our network gateway per second

3.2G

video streamed from cloud per second

Verkada shard architecture

Verkada



Simplified view; every shard runs the same stack

The problem

Verkada

Alert fatigue 1

Fix time also a function of oncall domain knowledge

Alert growth 2

Every feature and customer adds alerts.

Alerts keep changing 3

Definitions and priorities evolve over time.

10+

alerts on a busy day, for one on-call

KubeCon CloudNativeCon North America 2026

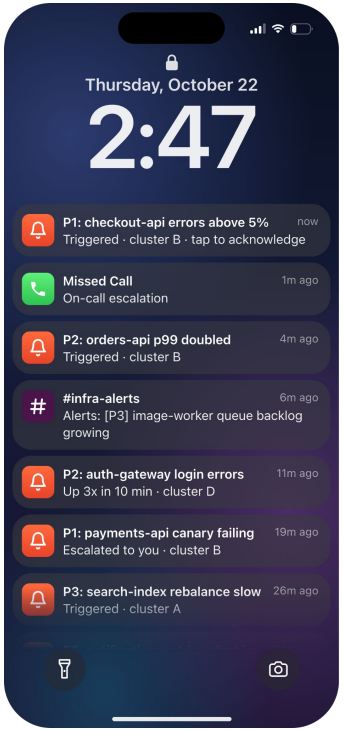
Source: alert volume per channel, October 2026

Rohan’s own story: On-call at 2 a.m.

Verkada

3x

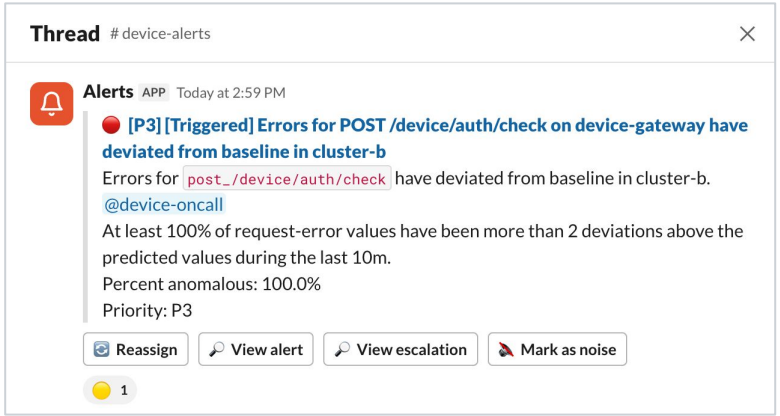
more pages I acknowledged, 2025 to 2026

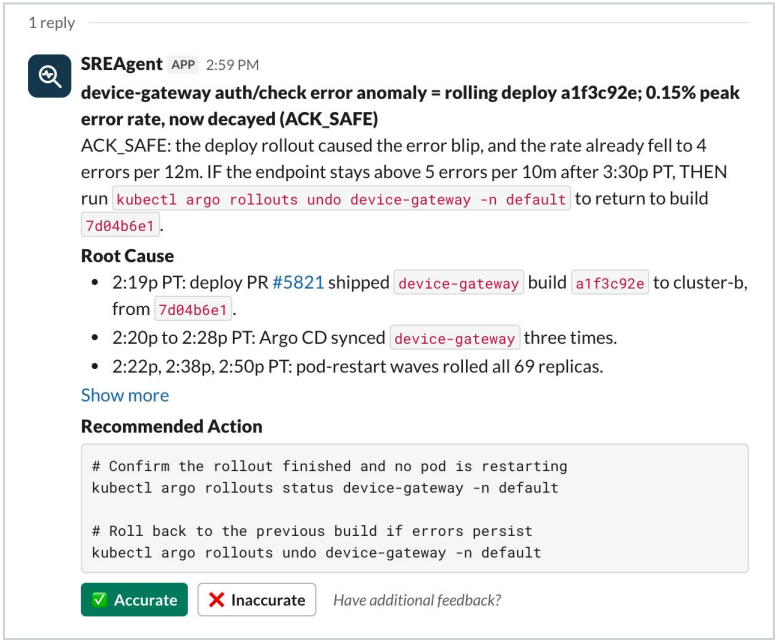


Source: Paging system, 2025 to 2026; phone illustrative

SREAgent – user experience

Verkada





Recreation of a real alert and reply, Nov 2026; names and details changed

Auto-triage 4,700 alerts a month

Verkada

4,712

triages per month

14,238 over 92 days, Jul-Sep 2026

$0.33

per investigation

about $1,555 a month at this volume

139 s

median time to triage

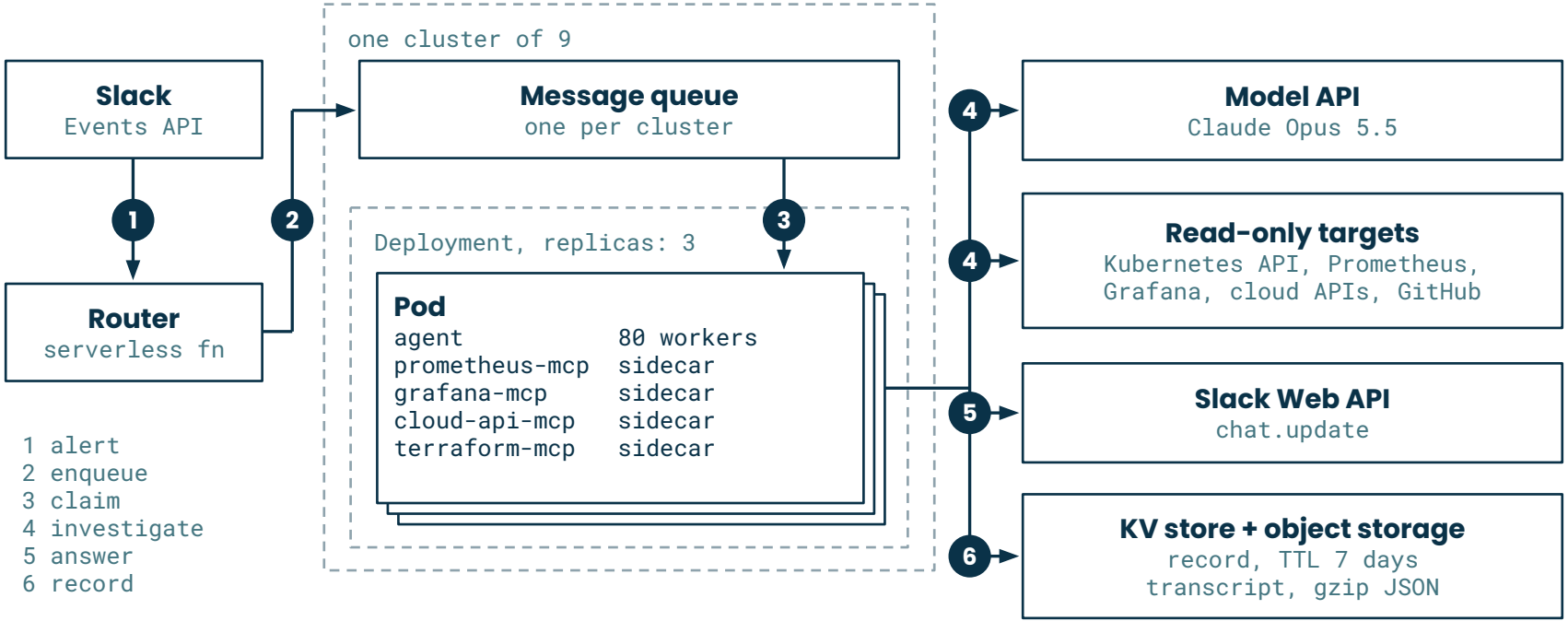
October 2026

KubeCon CloudNativeCon North America 2026

Source: Triage records, Jul to Sep 2026; median latency, October 2026

How it works

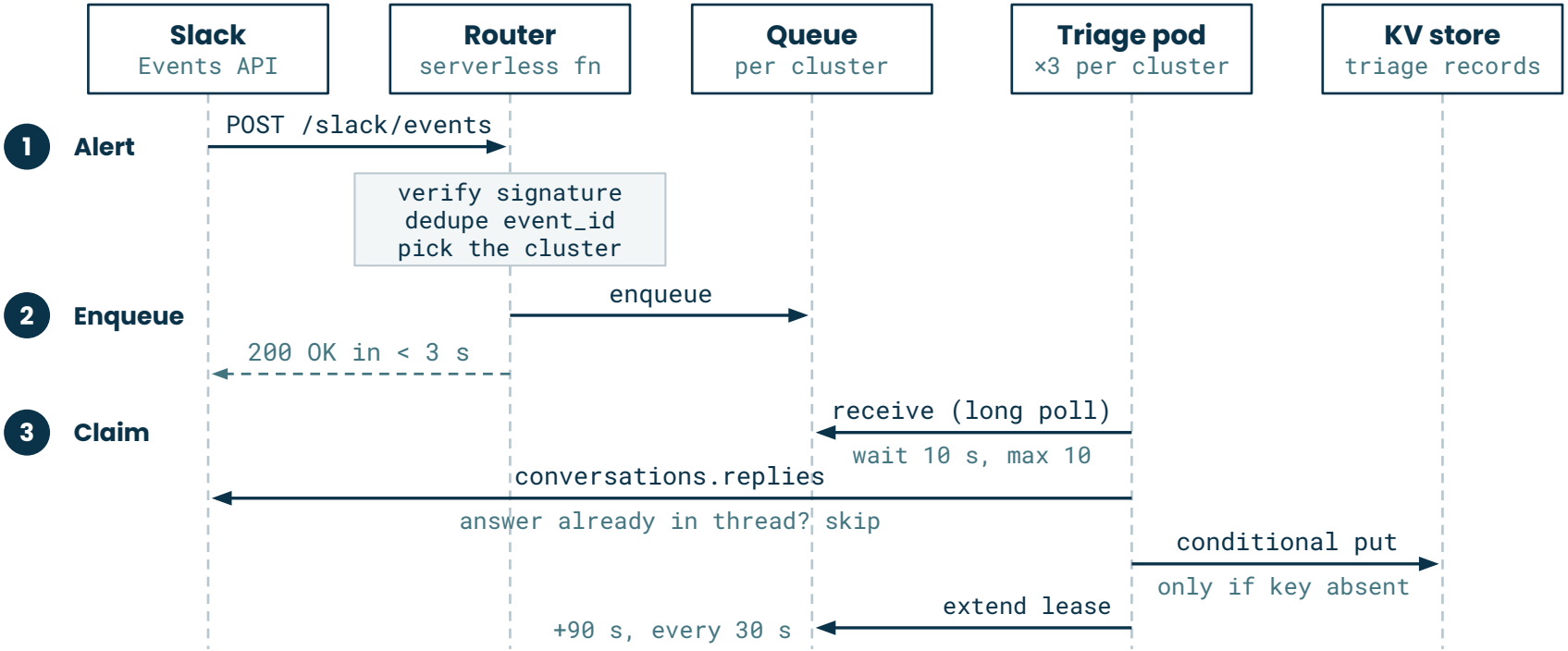
Verkada



Source: Production architecture; nine production clusters, three triage pods each

From alert to pod

Verkada

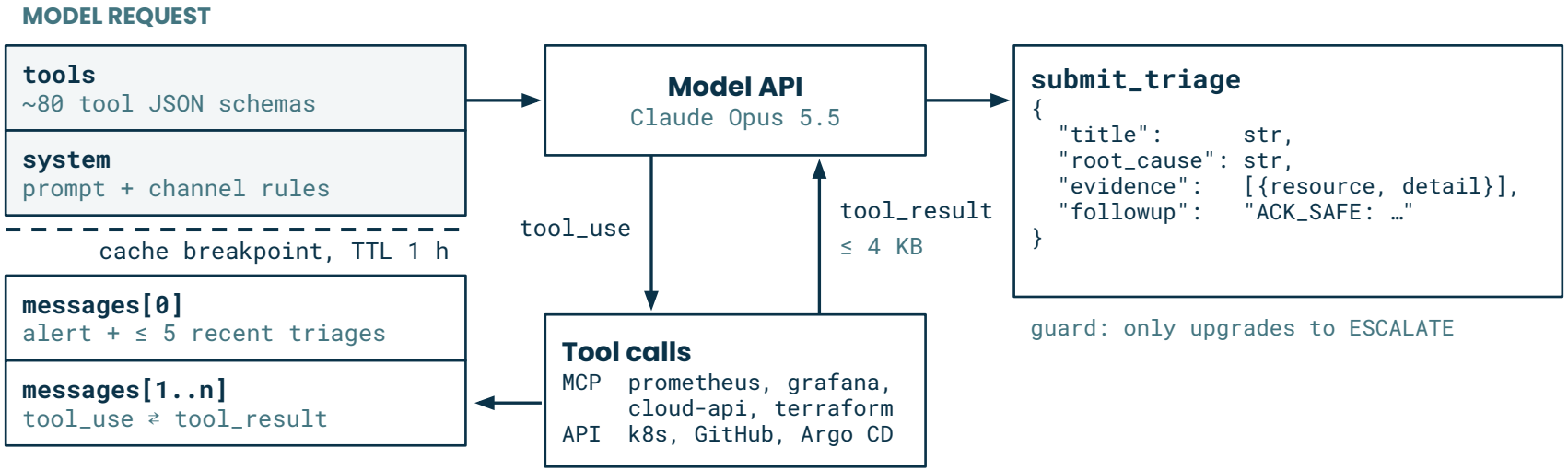


KubeCon CloudNativeCon North America 2026

Source: Router and queue consumer code, November 2026

The investigation loop

Verkada



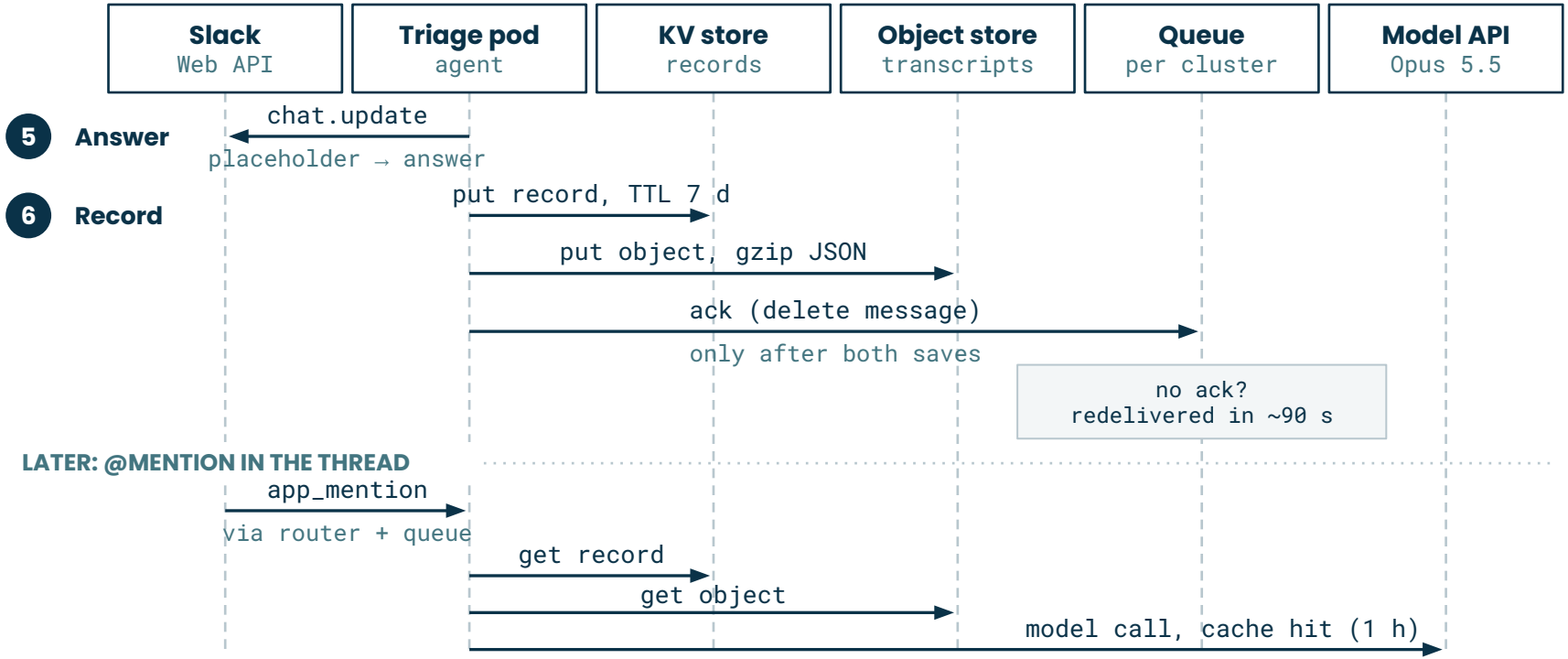

```
max_turns = 15   ·   1 turn left: nudge to submit   ·   cost cap hit: one forced submit_triage
tool calls: 3 tries with jitter   ·   circuit breaker: 3 failures → open 60 s
```

KubeCon CloudNativeCon North America 2026

Source: Triage engine and agent config

Answer, save, resume

Verkada



Source: Queue consumer, triage store and session store code, November 2026

Enabling teams to write triage logic

Verkada

One Markdown file per alert channel

Scope, strategy, and when to escalate

No code change to retune

```
Covers the device-provisioning services.

INVESTIGATION STRATEGY
1. Recent deploys are the first signal. Compare
   rollout times with the alert.
2. The main service is database-heavy. Check
   pod logs for pool exhaustion first.
3. The thin loader fails on downstream calls.
   Do not dig into its own pods.

Answer ACK_SAFE: bounded blast radius.
Answer ESCALATE: wide or sustained impact.
```

A wrong answer becomes one new line, written by the team that owns the alert.

Source: 15 channel prompts, November 2026; excerpt sanitized

Nightly evaluation to continuously improve correctness

Verkada

RUNNING NIGHTLY

Replay eval

- Replays confirmed triages
- Same recorded tool results
- Model judge, median of 3 votes
- Fails below 90% accuracy

LIVE

Live A/B

- Real alerts, copied to a candidate
- Both answer the same alert
- Reviewers pick the better answer
- Candidate never posts in the thread

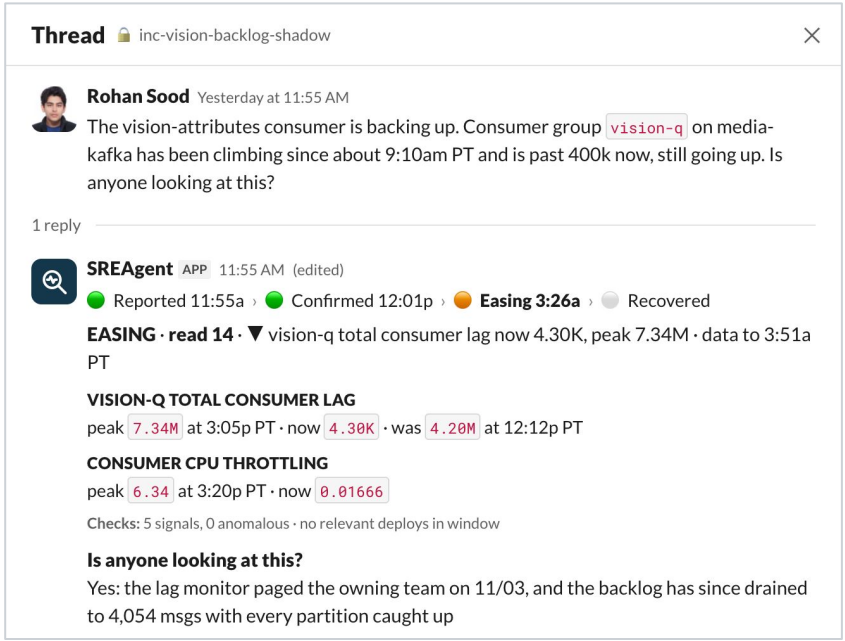
101 ratings since April: too few to compare two prompts

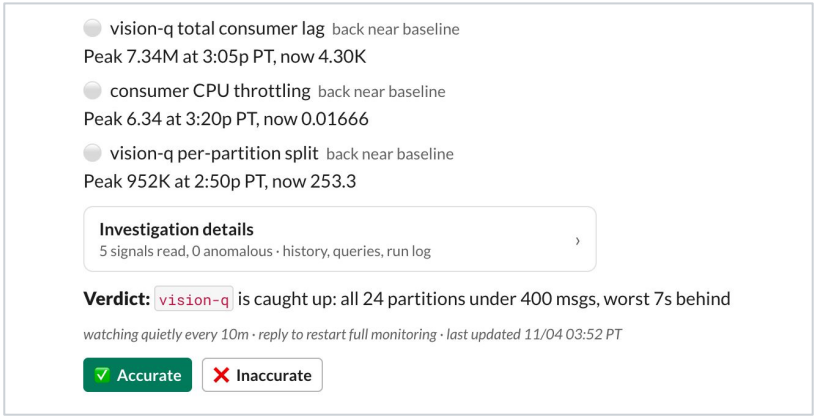
KubeCon CloudNativeCon North America 2026

Source: Eval harness, nightly jobs and live A/B, November 2026

Incident mode

Verkada





KubeCon CloudNativeCon North America 2026

Recreation of a real shadow-mode card, Nov 2026; names and details changed

Results

Verkada

~90,000
tool calls in October

~190 h
agent investigation, October

15
alert channels covered

69 of 101
ratings say accurate

WHY IT WAS SAFE TO SHIP

Read-only
Worst case: a wrong comment

$0.33 each
About $1,555 a month

Markdown rules
Teams retune with a pull request

In the open
Every answer stays in the thread

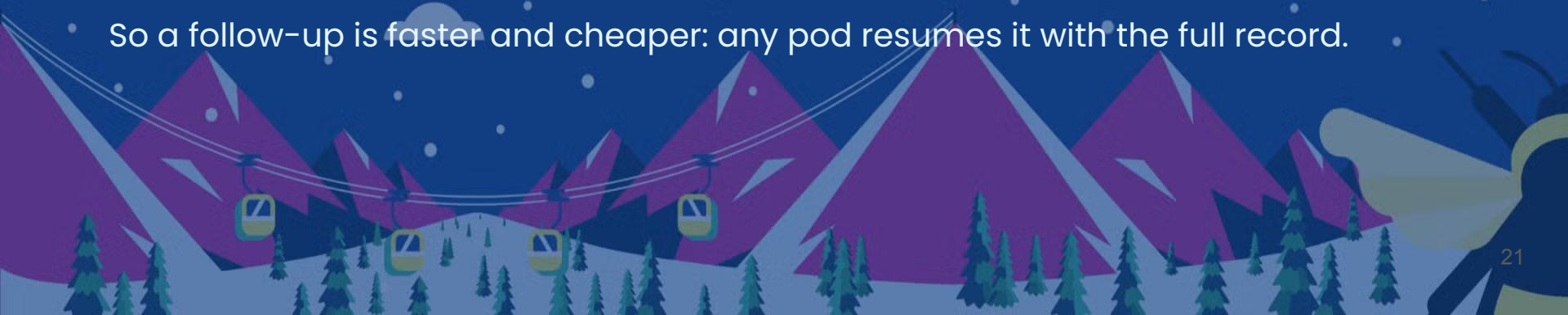
Source: Agent metrics, October 2026; feedback since April 2026

KubeCon CloudNativeCon North America 2026

LESSON LEARNED

Save every triage where any pod can load it

So a follow-up is faster and cheaper: any pod resumes it with the full record.

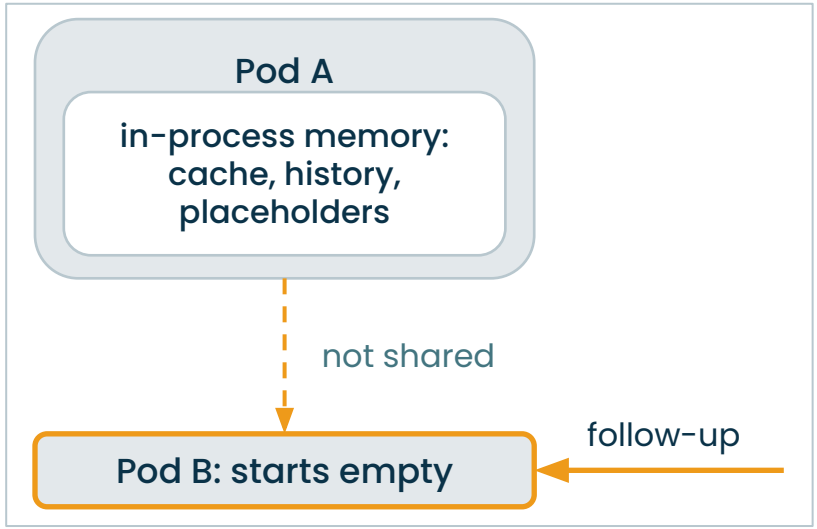


Saved triages make follow-ups faster

Verkada

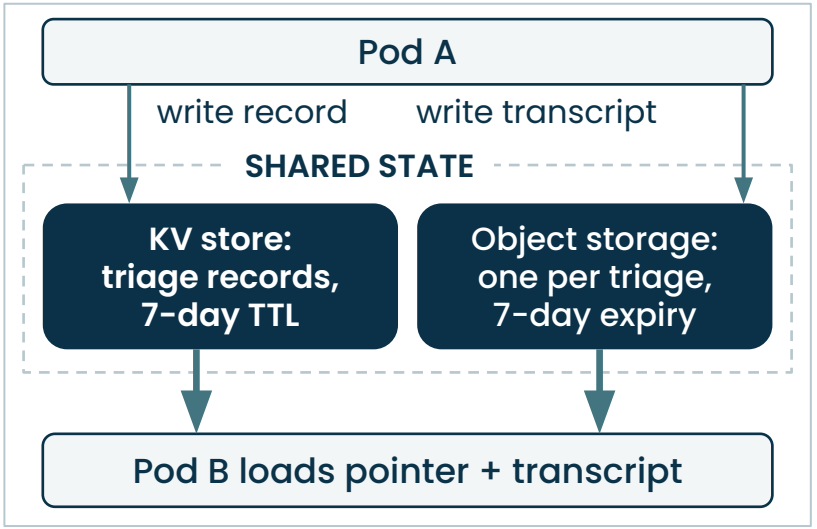
Faster and cheaper: resume from the saved record, pay for only the new turn.

BEFORE



Follow-ups started from nothing

AFTER



Any pod loads and resumes

KubeCon CloudNativeCon North America 2026

Source: SREAgent design doc, 2026

What each triage saves

Verkada

KV record per alert, 7-day TTL

Key: cluster + alert timestamp
Claim: conditional write, one pod wins
Summary: severity, root cause, verdict
Cost and tool-call counts
Feedback: Accurate or Inaccurate
Pointer: transcript key, version, size
Hash of the tool schemas

Transcript blob: 7-day expiry

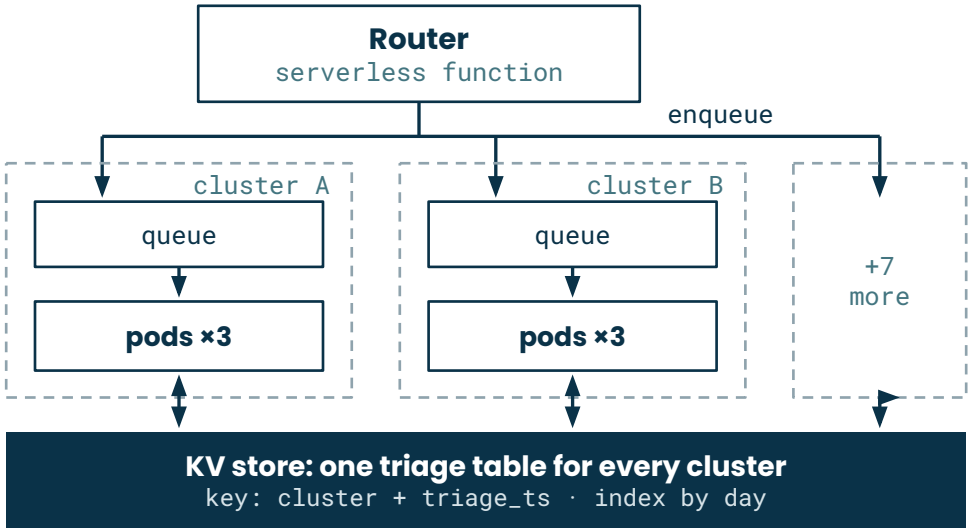
Full conversation, gzipped
Tool results with capture times
New version per follow-up
Loaded by any pod that resumes

Follow-ups take a compare-and-set lock, stale after 120 s, and bump the version with a conditional update.

Source: Store and follow-up handler; design doc

All clusters share one triage history

Verkada



- The router reads the cluster from the alert: marker, name, cloud account, region
- Any pod can read the fleet's recent triages, as reference, not as the answer
- Follow-ups return to the cluster that ran the triage

What it is not: no agent calls another cluster's tools. Each cluster keeps its own read-only access.

Source: Router cluster matching and shared triage store; prompt rules for prior triages

Takeaways

Verkada

1. Triage where the alert fires
2. Queue first, so restarts lose nothing
3. Checked answers, team-owned rules
4. Nightly evals, not ratings
5. Save every triage where any pod can load it

Rohan Sood and Adrian Ng, Verkada

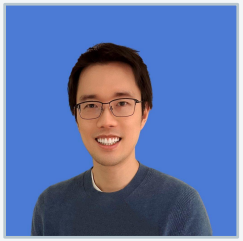
Tell us what you would ship first.

Rate this talk

sfeedback.com/IS5dS3

Connect with us, and we are hiring

Verkada



Adrian Ng

Director of Engineering, Platform Infrastructure

linkedin.com/in/cmuadrian



Rohan Sood

Senior Infrastructure Engineer

linkedin.com/in/rohansood10

We are hiring

verkada.com/careers

Scan for open roles