



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Istio's Ambient Mesh: The Real Cost of Sidecar-less Tracing

Mofesola Babalola
Staff SRE

Hannah Olukoye
Engineering Manager



The Session Promise



KubeCon



CloudNativeCon

Europe 2026

1

Before/After Cluster Metrics

2

L4 vs L7 Latency Benchmarks

3

The 'Cognitive Load' of Sidecar-less Debugging

4

Solving the Observability Blind Spot

5

Feedback/Resources



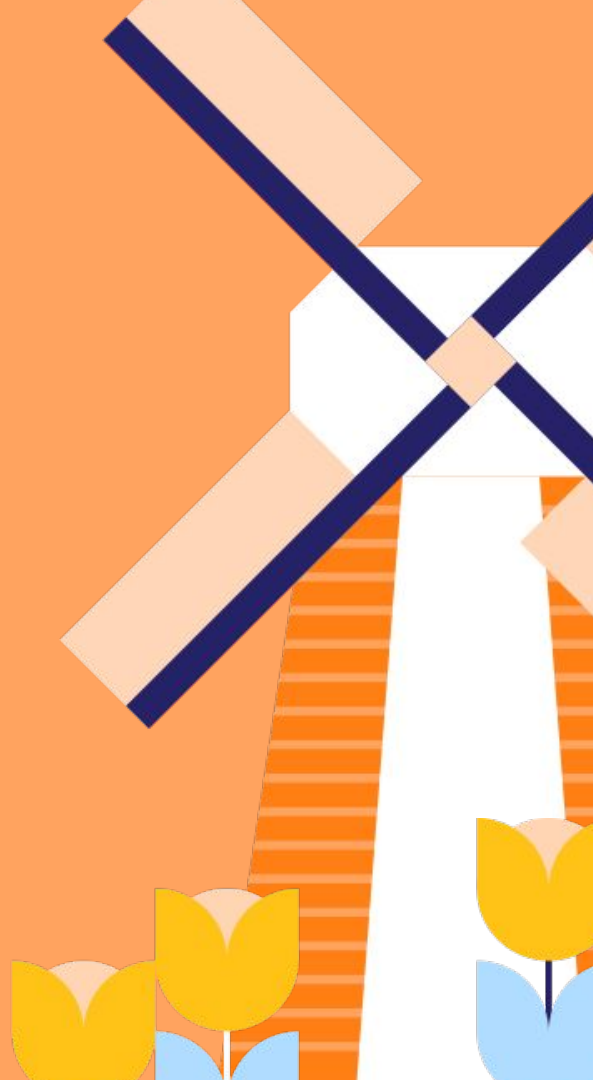
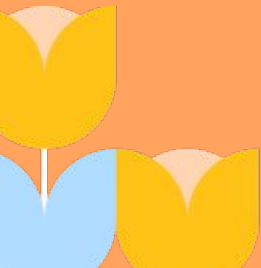
KubeCon



CloudNativeCon

Europe 2026

1 Before/After Cluster Metrics



The SRE's Problem: The 'Sidecar Tax'

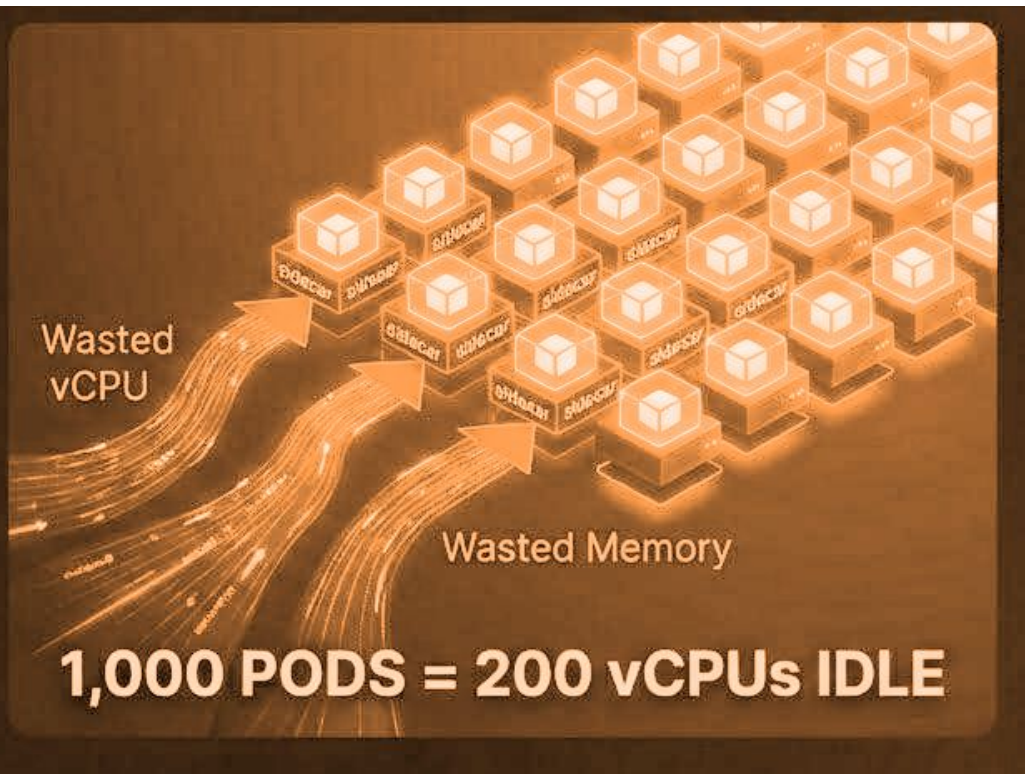


KubeCon



CloudNativeCon

Europe 2026



- **Resource Overhead:** ~0.20 vCPU & 60MB per pod
- **Scaling:** Linear growth of idle proxies
- **The Math:** 1,000 pods = 200 vCPUs doing 'nothing'

Operational Pain: The 'Restart Storm'



KubeCon



CloudNativeCon

Europe 2026

Problem

Envoy is coupled to the App Lifecycle

Impact

Security patches
=
100% Pod Rollout

Friction

Coordination overhead
between SRE and Dev teams



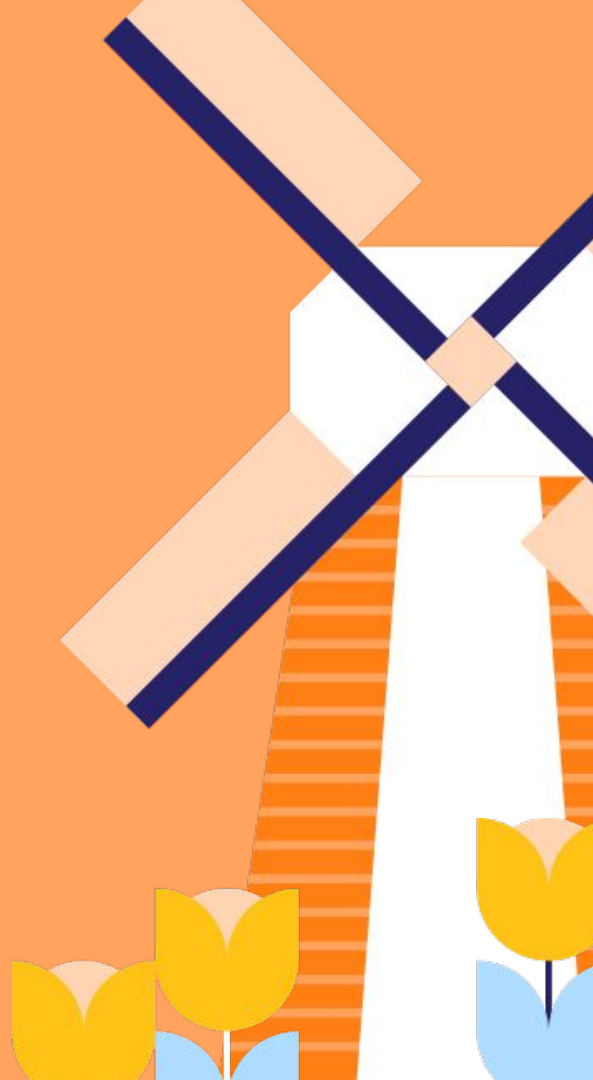
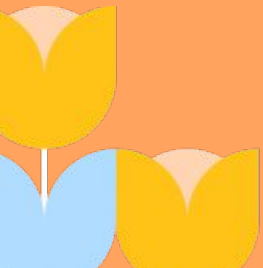
KubeCon



CloudNativeCon

Europe 2026

2 L4 vs L7 Latency Benchmarks



Enter Ambient Mode: Decoupling the Data Plane



KubeCon

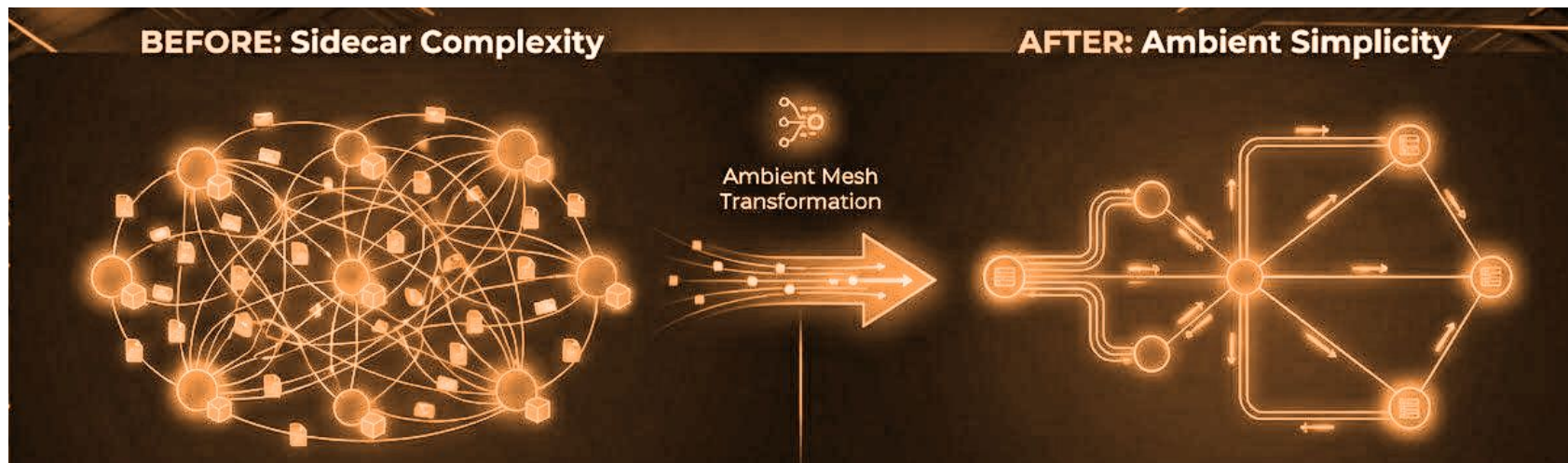


CloudNativeCon

Europe 2026

- ztunnel (L4): Shared, node-level agent
- Waypoint (L7): Selective, per-ServiceAccount proxy

The Goal: Mesh features without pod-level injection



ztunnel: The Rust-Powered 'Secure Pipe'



KubeCon



CloudNativeCon

Europe 2026

- **Function:** mTLS, L4 Telemetry, TCP Logging
- **Efficiency:** Purpose-built Rust (No Envoy overhead)
- **Footprint:** ~0.06 vCPU and ~12MB per node

eBPF: The Magic Under the Hood

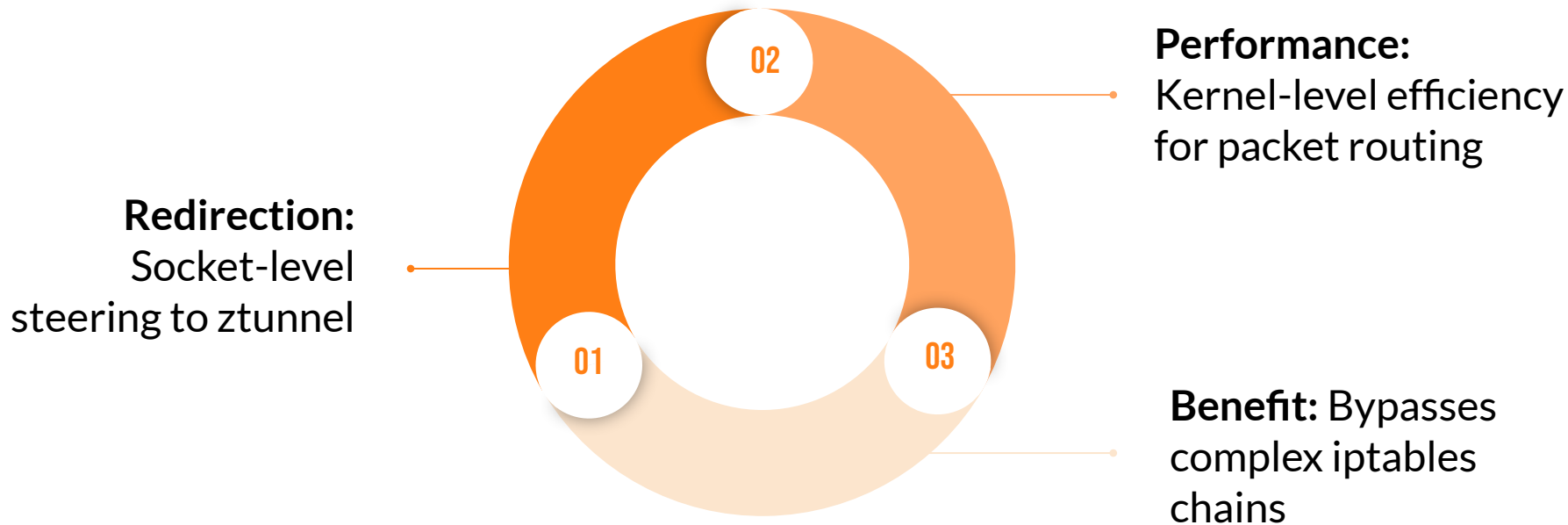


KubeCon



CloudNativeCon

Europe 2026



HBONE: The 'Blind' Tunnel Protocol



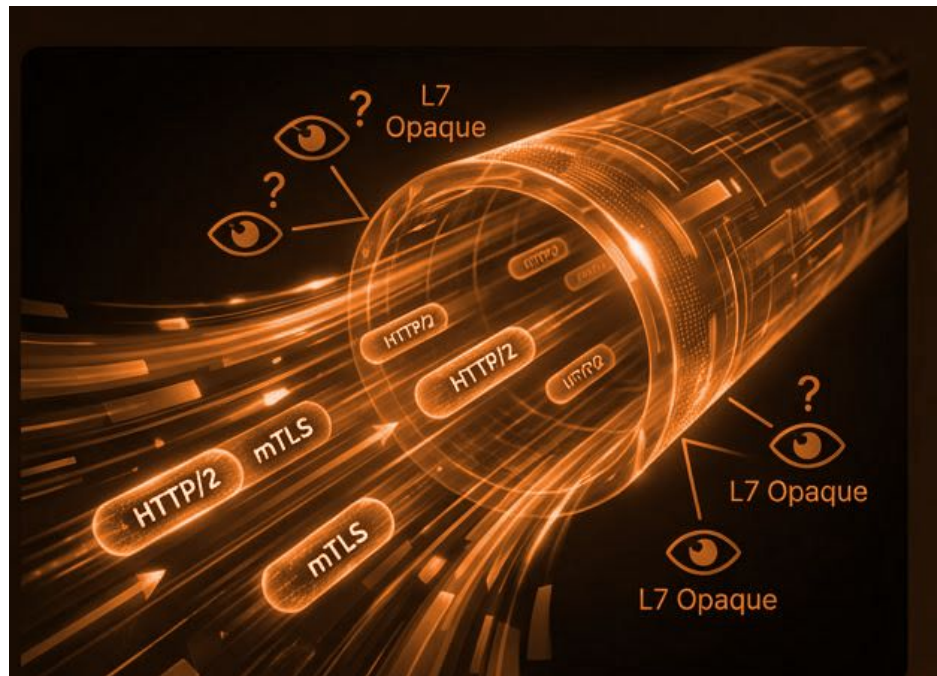
KubeCon



CloudNativeCon

Europe 2026

- **Tech:** HTTP-Based Overlay Network
- **Mechanism:** mTLS via HTTP/2 CONNECT tunnels
- **Limitation:** Encapsulates traffic; L7 is opaque to ztunnel



The Tracing Conundrum: L7 Blindness



KubeCon



CloudNativeCon

Europe 2026

- ❑ **Sidecar:** Sees L7 headers by default (MITM)
- ❑ **Ambient L4:** No header inspection = No trace spans
- ❑ **The Gap:** Broken service graphs in ztunnel-only mode

The Waypoint Requirement



KubeCon



CloudNativeCon

Europe 2026

- ❖ **Requirement:** L7 observability requires an Envoy Waypoint
- ❖ **Opt-in:** Only pay for L7 where you need Tracing/AuthZ
- ❖ **Incremental:** You don't have to boil the ocean

Scaling Waypoints: Production YAML (Part 1)



KubeCon



CloudNativeCon

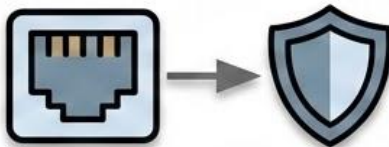
Europe 2026

Concurrency



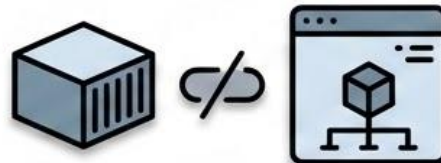
2 (Optimized for
250m CPU limits)

Protocol



HBONE
(Listener on Port
15008)

Lifecycle



Independent of the
application pods

Scaling Waypoints: Production YAML (Part 2)



KubeCon



CloudNativeCon

Europe 2026

HPA Targets



60% CPU and 70%
Memory targets

Stabilization Window



300s scale-down
window to prevent
flapping

Spread Constraints



TopologySpreadCons
traints for High
Availability

Benchmarking Methodology



KubeCon



CloudNativeCon

Europe 2026

Environment: EKS (m6i.xlarge),
1,000 Pod Cluster

Tools: Fortio for load,
Prometheus/Grafana for tracking

Target: Stable 2,000 Requests Per
Second (RPS)

Resource Consumption: The 76% Memory Win



KubeCon



CloudNativeCon

Europe 2026

Resource Performance at 2,000 Requests Per Second

Data Plane Component	CPU Usage (milli)	Memory Footprint (MiB)
Traditional Sidecar (Envoy)	638m	288Mi
Ambient L7 (Waypoint)	370m	68Mi
Ambient L4 (ztunnel)	60m	12Mi

Latency: The 'Extra Hop' Reality



KubeCon



CloudNativeCon

Europe 2026

Latency Breakdown (Sidecar vs. Ambient)

Architecture	Average	99th Percentile (p99)	99.9th Percentile (p99.9)	The Stability
Traditional Sidecar	1.60 ms	2.44 ms	25.81 ms	High volatility due to pod resource contention.
Ambient L4 (Rust)	0.52 ms	0.99 ms	1.63 ms	15x more stable than sidecars.
Ambient L7 (Waypoint)	1.71 ms	2.75 ms	10.00 ms	60% reduction in tail latency spikes.

Tail Latency at Scale (p99)

Sidecar



p99.9: 25.81ms



Ambient



p99.9: 10.00ms

(Stable tail due to decoupled infrastructure)



✓ **Winner: Resource Isolation.** By moving the proxy out of the app pod, we eliminate the jitter caused by shared CPU limits.

The 'Real Cost' Summary: Financials



KubeCon



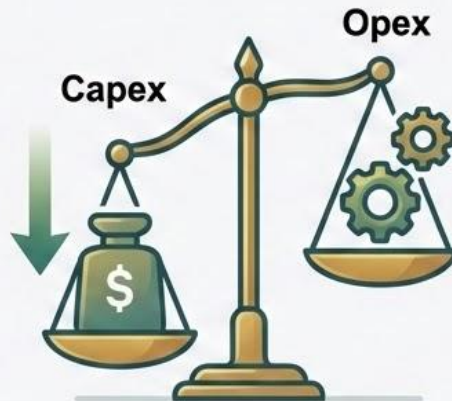
CloudNativeCon

Europe 2026

Projected Savings:
\$1.8M - \$2.2M
annually



Projected Savings:
\$1.8M - \$2.2M annually



Capex vs Opex: Lower compute costs vs New Ops training



The Big Win: Reclaiming
'Ghost' capacity

The Hidden Cost: Cross-AZ Charges



KubeCon



CloudNativeCon

Europe 2026

- ★ Danger: Traffic hopping across Availability Zones to a Waypoint
- ★ Multiplier: AZ egress costs can double your networking bill
- ★ Mitigation: Zone-aware routing and local replicas

Security Gain: Identity Isolation



KubeCon



CloudNativeCon

Europe 2026

- **Sidecar:** Keys are inside the pod (Vulnerable to RCE exfiltration)
- **Ambient:** Keys are in ztunnel (Protected by node boundary)
- **Logic:** Compromised app != Compromised mesh identity



KubeCon



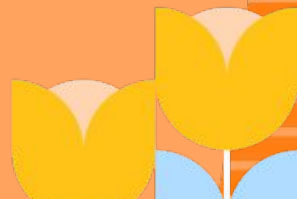
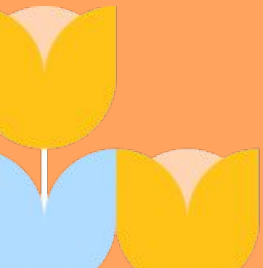
CloudNativeCon

Europe 2026

03

The 'Cognitive Load' of Sidecar-less

Debugging



Operational Cost: The 'Cognitive Load'



KubeCon



CloudNativeCon

Europe 2026

- ★ **Change:** You cannot 'exec' into a pod to debug the mesh
- ★ **New Tooling:** `istioctl ztunnel-config / waypoint-config`
- ★ **Skillset:** SREs must understand eBPF and node-level logs

Debugging Workflow in Ambient





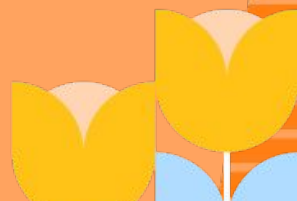
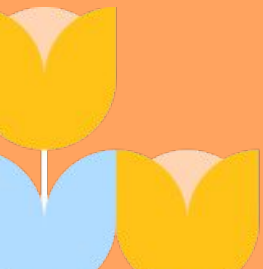
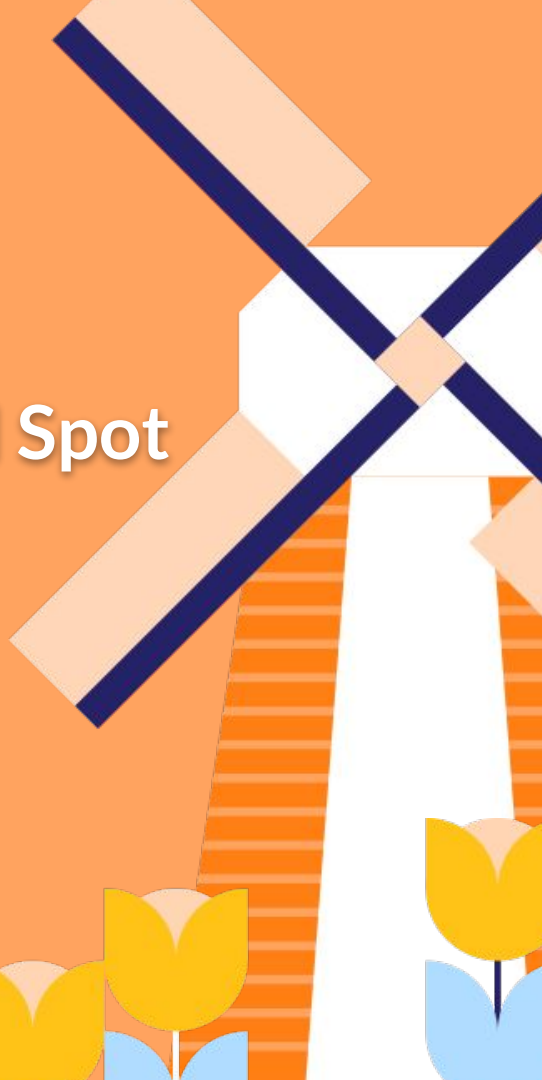
KubeCon



CloudNativeCon

Europe 2026

04 Solving the Observability Blind Spot



Observability Strategy: Correlating Spans



KubeCon



CloudNativeCon

Europe 2026

- ❑ **Solution:** Combine ztunnel metrics with Waypoint traces
- ❑ **Result:** Rich, context-aware telemetry without universal overhead
- ❑ **Pro-tip:** Use selective tracing for the 'High Value' paths

Zero-Downtime Upgrades: The Holy Grail



KubeCon



CloudNativeCon

Europe 2026

- ★ Sidecar: Patching Envoy = Rollout the App
- ★ Ambient: Patch ztunnel/Waypoint = App stays running
- ★ Outcome: Decoupled infrastructure and product lifecycles

Final Recommendations

Use ztunnel cluster-wide for mTLS floor

Deploy Waypoints surgically for L7 observability

Watch your AZ layout and use Topology Spread



KubeCon



CloudNativeCon

Europe 2026

Feedback & Resources



Mofesola Babalola

<https://www.linkedin.com/in/mofesola/>

Hannah Olukoye

<https://www.linkedin.com/in/hannaholukoye/>

<https://github.com/mofesola/istio-kubecon-demo>