SUPPLEMENT

The Honest Service Mesh Adoption Framework

Envoy response-flag triage table and adopt/defer decision thresholds

Ritvik Pandya · Engineering Lead, distributed systems & cloud-native infrastructure · linkedin.com/in/ritvik-pandya

This page and the next are the two reference artifacts the talk's abstract promises, in a form you can print or paste into a runbook. Views are my own; nothing here comes from a production system.

1 Envoy response-flag triage table

In a mesh, a 503 with an empty application log usually means the proxy failed, not the app. Envoy records why in a **response flag**, written right after the HTTP status in the proxy's access log. Read the flag first: it tells you which config object to open.

FLAG	CODE	ENVOY MEANING	FIRST THING TO CHECK IN ISTIO
UH	503	No healthy upstream hosts	Nothing ready behind the Service: selector vs. pod labels, readiness, DestinationRule subset labels. Or outlier detection has ejected every endpoint.
UF	503	Upstream connection failure	mTLS mismatch: a DestinationRule sending plaintext (`tls.mode: DISABLE`) to a STRICT workload, or `ISTIO_MUTUAL` to a pod with no sidecar.
UO	503	Upstream overflow (circuit breaking)	`connectionPool` limits and outlier settings in the DestinationRule.
NR	404	No route configured	VirtualService host, Service port name (`http` / `grpc` prefix), or gateway selector.
UC	503	Upstream connection termination	App idle timeout shorter than the proxy's. If it started right after a TLS or policy change, treat it as UF.
URX	varies	Retry limit reached	Upstream health: outlier detection and readiness probes. Retries hid the failure until they ran out.
UT	504	Upstream request timeout	VirtualService `timeout` vs. real upstream latency. Attempts × `perTryTimeout` must fit inside the route timeout.
UAEX	403*	Denied by external authorization	The ext_authz provider behind a CUSTOM AuthorizationPolicy. Native ALLOW/DENY denials carry no flag: look for `rbac_access_denied` in the response details.

* Typical; the authorization service sets the code. URX keeps the code of the last attempt.

WHERE TO LOOK

1. Make sure access logs are on. Not every install profile enables them: set `meshConfig.accessLogFile: /dev/stdout` or enable them with the Telemetry API.

2. Read the proxy, not the app. `kubectl logs <pod> -c istio-proxy`. In Istio's default format the flag is the token after the HTTP status.

3. Read the next two fields. Response code details say who set the code and why; the upstream transport failure reason names TLS handshake errors.

4. Confirm against live config. `istioctl proxy-config cluster | endpoint | route <pod>`

Ambient mode: ztunnel is not Envoy and logs L4 errors in its own format (`kubectl logs -n istio-system ds/ztunnel`). Response flags apply at waypoints and sidecars.

2 Adopt / defer decision thresholds

Not “is a service mesh good technology?” (it is) but “does it earn its cost at my scale, with my team, this year?” Run the gates in order and stop at the first that says defer.



1 GATE

Ownership

Does a named team own the control plane: upgrades, incidents and budget? If not, defer. Ambient mode makes this matter more: it removes the resource cost and leaves staffing as the binding constraint.

2 CHECK

Single concern

Need one item from the bundle? Price the narrower option first, and rule it out on the record.

IF YOU ONLY NEED	CONSIDER INSTEAD	WHEN IT WINS
Encrypted service-to-service traffic	**Cilium mutual auth (eBPF)**	Already on Cilium CNI, no L7 rules
Workload identity only	**SPIFFE / SPIRE**	Identity without intercepting traffic
A gRPC-only fleet	**Proxyless gRPC via xDS**	One protocol, no data-plane proxy
Retries, timeouts, breakers	**Libraries or ingress config**	One or two languages, low churn
Just TLS at the edge	**Gateway API / ingress controller**	North-south only, no east-west policy

3 COUNT

Adopt signals: at least two of four

Ownership from gate 1 plus at least two of these, and a mesh is likely net positive.

SIGNAL	COUNTS WHEN
Fleet size	Per-service resilience code has already drifted apart
Compliance need	mTLS and segmentation are a hard requirement with an audit attached
Traffic control	Canary and traffic shifting are recurring needs, not one-offs
Polyglot fleet	Several languages, so library-based resilience means N implementations

4 DEFER

Blockers: defer even if gate 3 passes

Unmeasured latency-critical path. Sub-millisecond budgets where the real hop cost hasn't been benchmarked.

Mid-migration. Already moving platforms or CNIs; a mesh compounds change risk.

5 SIZE

Data plane

DATA PLANE	SERVICE-COUNT GUIDE	WHY
Sidecar	**~20+ services**	Per-pod cost must be amortized; upgrades touch every pod.
Ambient	**Closer to ~10**	No per-pod tax; service count stops being the binding constraint.

Start with the ambient L4 overlay (ztunnel) and add waypoints only where L7 policy is real. If nearly every service needs L7, waypoints spread and the savings shrink.

BEFORE YOU COMMIT

Benchmark your baseline. Pooled HTTP/2 can make the latency cost net negative against an HTTP/1.1 baseline. Measure your protocol.

Size it. Sidecar memory × pod count is the number that surprises people. Compute it before the design review.

Budget telemetry. Sampling and metric scope go in the same proposal as compute. Cardinality, not volume, breaks the budget.

Gate mesh config like a deploy. Same review path, same rollback plan. A policy edit can cause an outage.