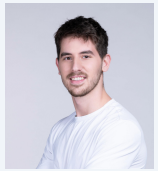
KubeCon CloudNativeCon North America 2026

#KubeCon #CloudNativeCon

Gateway API

Building the Future of Traffic



Lior Lieberman
@LiorLieberman
Staff Eng Lead
@ Google



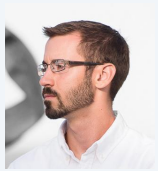
Norwin Schnyder
@snorwin
Sr. Cloud SWE
@ Airlock



Ricardo Katz
@rikatz
Principal SWE
@ Red Hat



Evaline Ju
@evaline-ju
Sr. SWE
@ IBM Research



Grant Spence
@gcs278
Sr. SWE
@ Red Hat

Agenda

KubeCon CloudNativeCon North America 2026

1. **What is Gateway API?**
   Ingress successor, the journey to v1.7, where the API stands today
2. **The Ecosystem**
   Gateway API, Inference Extension, Kube Agentic Networking, WG AI Networking
3. **Release v1.6 - L4 goes GA**
   TCPRoute & UDPRoute, OSI layer coverage, one Gateway for many cases, egress
4. **Agentic Networking**
   KAN v0.2, AccessPolicy, TelemetryPolicy
5. **Release v1.7 and beyond**
   Session persistence, pre-routing filters, payload processing, ClusterTrustBundle, HTTPRoute retries (Standard)
6. **Conformance**
   Conformance testing and updated requirements
7. **Community**
   Contributors & leads, how to get involved

What is Gateway API?

KubeCon CloudNativeCon North America 2026

“Three resources instead of one?” Yes, on purpose.

Role-oriented

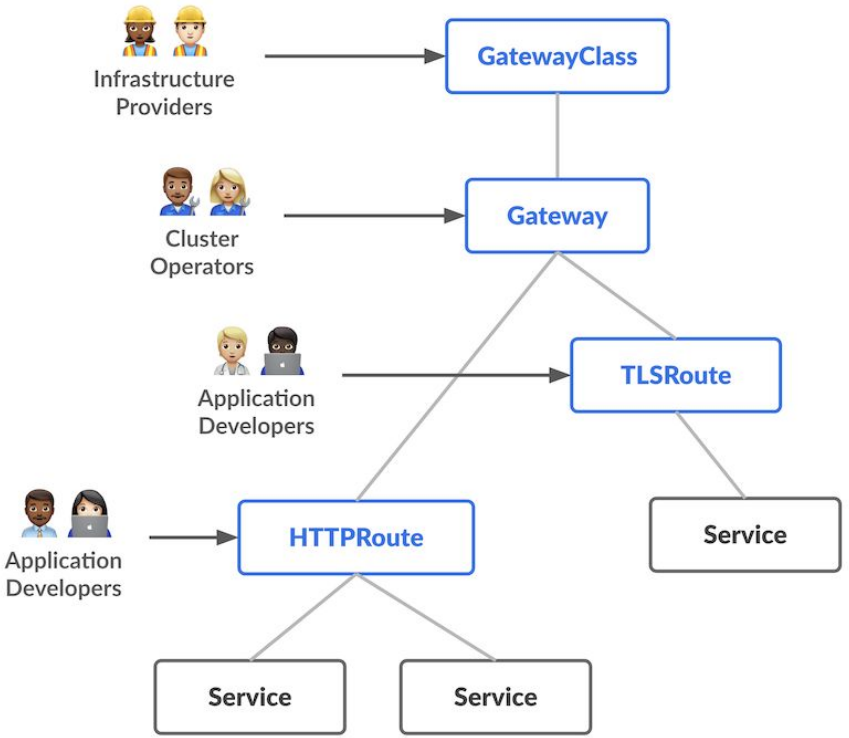
Infra, platform and app teams each own their layer

Portable & expressive

Features live in the spec, not in annotations

One API across the stack

North-south, mesh, egress and AI traffic



The journey so far

KubeCon CloudNativeCon North America 2026

2019	2023	2024	Feb 2026	Jun 2026	Oct 2026
Project founded	**v1.0 GA** Gateway, HTTPRoute	**v1.1** Mesh & GRPCRoute GA	**v1.5** TLSRoute & ListenerSet GA	**v1.6** TCPRoute & UDPRoute GA	**v1.7** Retries GA + new Experimental

3
releases since KubeCon NA 2025

30+
implementations

65
GEPs and counting

Where the API stands today

HTTP & gRPC

- HTTPRoute **GA**
- GRPCRoute **GA**
- CORS **GA · v1.5**
- Retries **GA · v1.7**
- External auth **EXP**

L4 & TLS

- TLSRoute **GA · v1.5**
- TCPRoute **GA · v1.6**
- UDPRoute **GA · v1.6**
- ListenerSet **GA · v1.5**
- Pre-routing filters **EXP · v1.7**

Backends & Ops

- BackendTLSPolicy **GA**
- Backend (egress) **EXP · v1.6**
- TelemetryPolicy **EXP · v1.7**
- ClusterTrustBundle **EXP · v1.7**
- Address routability **EXP · v1.7**

GA = Standard channel **EXP** = Experimental channel

So many projects, what should I look at?

KubeCon CloudNativeCon North America 2026

🌐 Gateway API

Core Standard

Establishes the APIs for Gateway, routes, L4/L7 traffic

gateway-api.sigs.k8s.io

🤖 Inference Ext.

AI Workloads

Extend Gateway API backends for inference workloads

inference-extension.sigs.k8s.io

🧠 Agentic Network

Routing & Governance

Routing and Governance for AI cases, extending Gateway API

kube-agentic-networking.sigs.k8s.io

⚙️ WG AI Networking

Working Group

Captures proposals for AI cases that may land as concrete implementations

** Does not own code*

🔄 Synergy & Collaboration

All initiatives continuously collaborate and evolve together!

New on 1.6 - L4 cases

KubeCon CloudNativeCon North America 2026

🔌 TCPRoute & UDPRoute

Protocol + Port Focus

Designed for cases where traffic management relies strictly on Layer 4 attributes without needing higher-layer inspection.

⚡ Legacy & Passthrough

Zero Inspection Passing

Covers legacy architectural gaps or workloads requiring raw traffic to fully pass through on specific ports without inspection.

🎯 Port Ranges Roadmap

Future Expansion

Support for dynamic or contiguous port ranges is actively being defined for future iterations.

👏

Contributor Highlight

Special thanks to **Zac Nixon** from **AWS** for implementing and moving forward with this feature!

Extending the OSI layer coverage

KubeCon CloudNativeCon North America 2026

🌐 Application Routing (L7)

- **HTTPRoute** & **GRPCRoute**
- Advanced path/header routing & traffic splitting

🔒 TLS Routing (L4/L5 - SNI)

- **TLSRoute** (SNI hostname matching)
- TLS Termination & Passthrough (pre-L7)

⚡ Direct L4 Routing

- **TCPRoute** & **UDPRoute**
- Protocol + Port direct passthrough
- Multi-protocol listener sharing on same Gateway

Extending the OSI layer coverage

KubeCon CloudNativeCon North America 2026

🤖 AI Cases

🌐 Application Routing (L7)

- **HTTPRoute** & **GRPCRoute**
- Advanced path/header routing & traffic splitting

🔒 TLS Routing (L4/L5 - SNI)

- **TLSRoute** (SNI hostname matching)
- TLS Termination & Passthrough (pre-L7)

⚡ Direct L4 Routing

- **TCPRoute** & **UDPRoute**
- Protocol + Port direct passthrough
- Multi-protocol listener sharing on same Gateway

STANDARD (v1.6.0)

TCPRoute / UDPRoute examples

KubeCon CloudNativeCon North America 2026

```yaml
apiVersion: gateway.networking.k8s.io/v1
kind: Gateway
metadata:
  name: my-tcp-gateway
spec:
  listeners:
  - name: foo
    protocol: TCP
    port: 8080
    allowedRoutes:
      kinds:
      - kind: TCPRoute
```

```yaml
apiVersion: gateway.networking.k8s.io/v1
kind: TCPRoute
metadata:
  name: tcp-app-1
spec:
  parentRefs:
  - name: my-tcp-gateway
    sectionName: foo
  rules:
  - backendRefs:
    - name: my-foo-service
      port: 6000
```

One Gateway, multiple cases

KubeCon CloudNativeCon North America 2026

Gateway Resource

```
listeners:
  - name: http, protocol: HTTP, port: 80
    allowedRoutes: {kinds: [{kind: HTTPRoute}]}
  - name: tls, protocol: TLS, port: 8443
    allowedRoutes: {kinds: [{kind: TLSRoute}]}
  - name: tcp, protocol: TCP, port: 10000
    allowedRoutes: {kinds: [{kind: TCPRoute}]}
  - name: udp, protocol: UDP, port: 10000
    allowedRoutes: {kinds: [{kind: UDPRoute}]}
```

🌐 HTTP Listener (Port 80)

- **Protocol:** HTTP • **Port:** 80
- **Allowed Route:** HTTPRoute

🔒 TLS Listener (Port 8443)

- **Protocol:** TLS • **Port:** 8443
- **Allowed Route:** TLSRoute (SNI)

⚡ TCP & UDP Listeners (Port 10000)

- **Protocols:** TCP / UDP • **Port:** 10000
- **Allowed Routes:** TCPRoute / UDPRoute

Egress Gateways

KubeCon CloudNativeCon North America 2026

🌐 External Resources

Allowing workloads to consume external resources:

- External providers like OpenAI or Anthropic
- External REST & Web APIs

🛠️ Traffic Modification

Using HTTPRoute to modify traffic:

- Injecting headers (like auth headers)
- Centralized outbound policy enforcement

🛡️ Routability & Security

Restricting Gateway exposure:

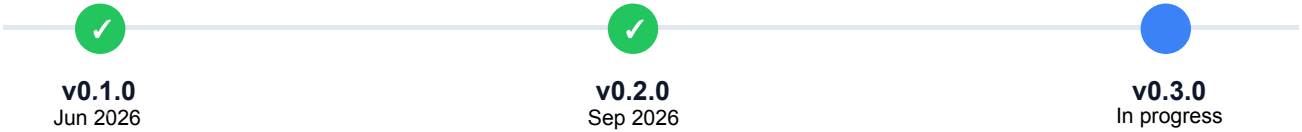
- Limit Gateway routability
- Avoid usage as an open "public proxy"

📌 Resources & Specifications

GEP-5093: for routability | **GEP-4894:** for Backend
Demo: Link

Kube Agentic Networking

Gateway API subproject



WHAT YOU GET TODAY

- Control which agents can call which MCP tools
- Identity-aware policies (SPIFFE, ServiceAccount)
- External auth delegation for custom logic
- Distributed tracing for agent traffic

WHAT'S NEW

- TelemetryPolicy API merged in Gateway API
- CLI to inspect MCP tools on backends
- Conformance test suite

EXPERIMENTAL

AccessPolicy

TODAY - V1ALPHA1

- Control which agents can call which MCP tools
- Delegate auth decisions to your identity provider
- Write custom authorization rules in CEL

PROPOSED AND DISCUSSED CHANGES

- Simplify MCP matching with July spec
- JWT claims / broader identity types
- Session-scoped grants in CEL

NEW agentic-net-tool list-tools — inspect MCP tools on your backends before writing policy

ALIGNING WITH JULY MCP SPEC

MCP now exposes tool info via standard HTTP headers

TODAY – V1ALPHA1

```
authorization:
  type: Inline
  mcp:
    methods:
      - name: tools/call
        params:
          - lookup-order
```

PROPOSED – V1ALPHA2

```
authorization:
  type: Inline
  headers:
    - name: Mcp-Name
      value: lookup-order
```

EXPERIMENTAL (v1.7.0)

Telemetry Policy (GEP-4768)

KubeCon CloudNativeCon North America 2026

WHY IT MATTERS

- First Gateway API native observability resource
- Agent workflows span many services — you need traces to debug them
- Vendor-neutral: same policy across Gateway implementations

CURRENT SCOPE

- Sampling rate and parent-based sampling
- OTel collector as a backend destination
- Custom span attributes (literal, header, attribute)

```yaml
apiVersion: gateway.networking.x-k8s.io/v1alpha1
kind: XTelemetryPolicy
spec:
  targetRefs:
    - kind: Gateway
      name: agent-gateway
  tracing:
    mode: Enabled
    samplingRate:
      numerator: 10
      denominator: 100
    provider:
      backendRef:
        name: otel-collector
        port: 4317
    attributes:
      - name: env
        sourceType: Literal
        literalValue: prod
```

EXPERIMENTAL (v1.7.0)

Session Persistence (GEP-1619)

EVOLUTION

- **Leaner & Predictable:** Dropped idleTimeout; explicit cookie.path
- **Standardized API:** Clean Cookie vs. Header discriminators
- **Moved to Backend Resource:** Unified attachment point

WHY MOVE ATTACHMENT TO BACKEND?

- **Natural Ownership:** Defined directly on backends
- **Conflict Reduction:** One place, no precedence conflicts
- **Clear Discoverability:** Direct configuration inspection

```yaml
apiVersion: gateway.networking.x-k8s.io/v1alpha1
kind: XBackend
metadata:
  name: my-app-backend
spec:
  type: EndpointSelector
  port: ...
  endpointSelector:
    matchLabels:
      app: my-app
  sessionPersistence: # GEP-1619
    type: Cookie
    absoluteTimeout: "1h"
    cookie:
      name: app-session
      lifetimeType: Permanent
      path: "/"
```

WIP

Payload Processing

KubeCon CloudNativeCon North America 2026

GOAL

Inspect, transform, and guardrail full HTTP request and response message bodies at the proxy edge.

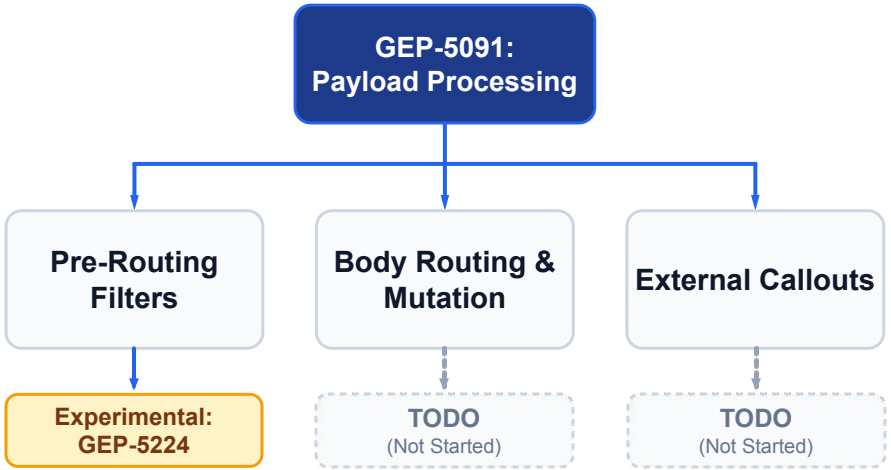
USE CASES

- **Body Routing:** Extract model attributes into headers
- **Guardrails:** Block prompt injections and jailbreaks
- **PII Compliance:** Scrub sensitive data before GPUs

KEY DIMENSIONS

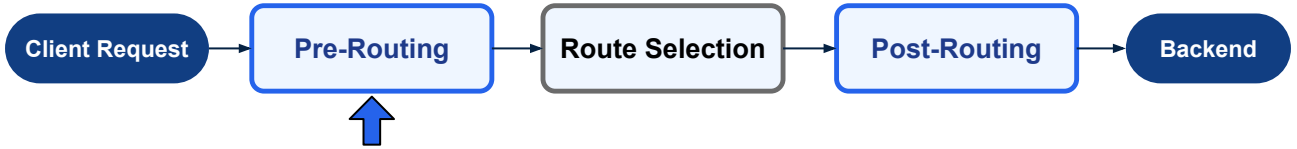
- **WHEN:** PreRouting or PostRouting
- **DIRECTION:** Request or Response
- **HOW:** InProcess (CEL) or ExtProcess (gRPC)
- **ACTION:** Inspect or Mutate or Short-Circuit

The 3-Way GEP Split



EXPERIMENTAL (v1.7.0)

Pre-Routing Filters (GEP-5224)




```yaml
listeners:
- name: https
  port: 443
  protocol: HTTPS
  tls: ...
  filters: # GEP-5224
    requests:
    - type: ExternalAuth
      externalAuth:
        protocol: HTTP
        backendRef:
          name: global-authz
          port: 9000
```

KEY ARCHITECTURAL EVOLUTION

- **Listener-Level Execution:** Runs logic before route selection
- **Strict Ordering:** Resolves filter ambiguity at edge processing
- **Initial API:** Only ExternalAuth and ExtensionRef

WHY RUN PRE-ROUTING?

- **Early Edge Rejection:** Block/auth requests before route matching
- **Global Enforcement:** Apply uniform policies across listeners
- **Feed-Forward Metadata:** Inject context for downstream routing

EXPERIMENTAL (v1.7.0)

Support for ClusterTrustBundle (GEP-5075)

KubeCon CloudNativeCon North America 2026

ClusterTrustBundle (GA in Kubernetes 1.37):
cluster-scoped resource for distributing X.509 trust anchors (root CAs) across the cluster.

HOW IT'S USED?

```yaml
caCertificateRefs:
  - group: certificates.k8s.io
    kind: ClusterTrustBundle
    name: example.com:my-signer:v1
```

- Supported for TLS validation (`caCertificateRefs`) in `Backend`, `BackendTLSPolicy` and `Gateway`
- Referenced by name only — no `namespace` field, no `ReferenceGrant`

```yaml
apiVersion: certificates.k8s.io/v1
kind: ClusterTrustBundle
metadata:
  name: example.com:my-signer:v1
spec:
  signerName: example.com/my-signer
  trustBundle: |
    -----BEGIN CERTIFICATE-----
    ...
    -----END CERTIFICATE----
```

STANDARD (v1.7.0)

HTTPRoute Retries (GEP-1731)

KEY PRINCIPLES

- **Mask transient failures:** retry on connection failures and status codes
- **Safe by default:** no retry once the request may have been processed (reset after send)
- **Bounded:** count-based, backoff between attempts, capped by the route's timeouts

THE RETRY POLICY (`rules[].retry`)

`codes` - HTTP status codes that trigger a retry

`attempts` - maximum number of retries

`backoff` - base interval for duration between retries

Connection failures are retried whenever `retry` is set

```yaml
apiVersion: gateway.networking.k8s.io/v1
kind: HTTPRoute
metadata:
  name: retry-example
spec:
  parentRefs:
  - name: example-gateway
  rules:
  - backendRefs:
    - name: backend-svc
      port: 8080
    retry:
      codes:
      - 500
      - 503
      - 504
      attempts: 2
      backoff: 100ms
```

KubeCon CloudNativeCon
North America 2026

Conformance



Conformance Testing

KubeCon CloudNativeCon North America 2026

WHAT IS CONFORMANCE

- **Test suite** shipped with every release that checks an implementation against the API spec
- **Guarantees API semantics** and **portability** across implementations
- **Publicly available, reproducible reports** that users can use to verify implementations

78

Features

170

Test scenarios
(147 Gateway / 23 Mesh)

Mesh Profile

HTTPRoute

Organization	Project	Version	Mode	Extended Features	Core	Mesh Cluster IP Matching	Mesh Consumer Route	Mesh HTTPRoute Backend Request Header Modification	Mesh HTTPRoute Named Route Rule
Buoyant	Buoyant Enterprise for Linkerd	enterprise-2.20.2	default	7/9 features	✅	✅	✅	✅	✅
cilium	cilium	v1.20.0	default	8/9 features	✅	✅	❌	✅	✅
istio	istio	1.31	default	8/9 features	✅	❌	✅	✅	✅

→ check out the Controller Matching Wizard (kudos to @bexxmodd)

Conformance Requirements

KubeCon CloudNativeCon North America 2026

Conformance Levels

Conformant

Partial Conformance
deprecated · removed in v1.8

~~Stale~~
removed in v1.6

Gateway API v1.7 updates the conformance requirements for implementations.

A conformant implementation:

NEW MUST be from a *publicly available* version, i.e., tagged release, hosted platform version or evaluation build (no commit hashes)

NEW MUST have *all* profiles pass within a report, with no skipped tests (1-5 skipped tests → Partial Conformance)

MUST have ≥ 1 Conformant report if several are submitted

MUST target Gateway API version n or n-1 (n-2 → Partial Conformance)

Prerelease Reports NEW

- Prove that an implementation will support a forthcoming feature (serve as evidence for graduating it from Experimental to Standard)
- Less strict requirements
- Not used to generate the Implementations List or feature matrices

KubeCon CloudNativeCon
North America 2026

Who Makes Gateway API Possible?

& How to Get Involved



Thank you to our contributors!

KubeCon CloudNativeCon North America 2026

Since KubeCon NA 2025 (Atlanta)

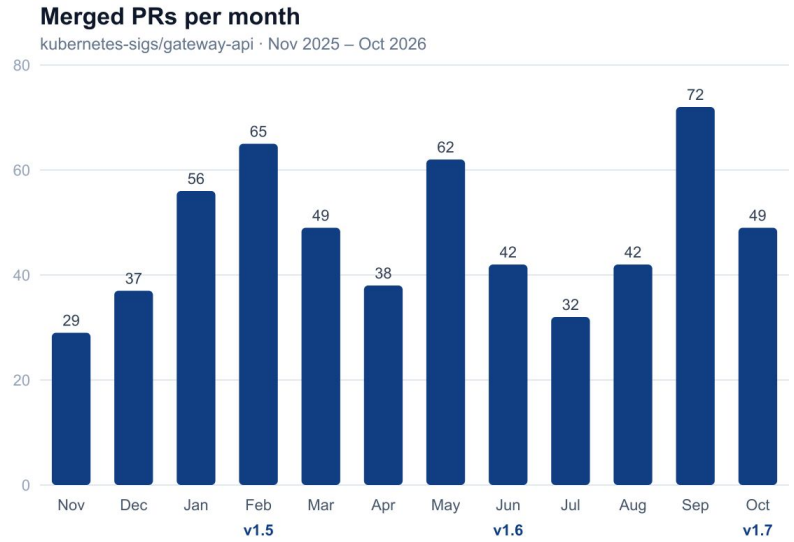
550+
PRs merged

120+
active contributors

80+
first-time contributors

3
releases: v1.5 → v1.7

370+ contributors all-time



Thank you to our leads!

KubeCon CloudNativeCon North America 2026

Gateway API maintainers



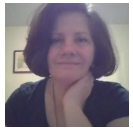


Mesh (GAMMA) leads






Conformance & GEP reviewers




Kube Agentic Networking




Inference Extension





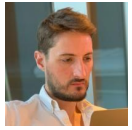


Emeritus









Get involved

KubeCon CloudNativeCon North America 2026

1 **Try Experimental, tell us what breaks**

Backend, TelemetryPolicy, pre-routing filters, ClusterTrustBundle

2 **Own a GEP**

Help take features to GA: design, implementation and conformance tests

3 **Implementers: submit conformance reports**

The v1.7 conformance requirements are live

Into AI traffic? Inference Extension · Kube Agentic Networking · WG AI Gateway

gateway-api.sigs.k8s.io #sig-network-gateway-api kubernetes-sigs/gateway-api

Feedback

KubeCon CloudNativeCon North America 2026



Gateway API Project Update: - Building the Future of Traffic