KubeCon CloudNativeCon
North America 2026

New Traffic, New Rules: Standardizing AI-Aware Networking in Kubernetes

Nir Rozenbaum, Keith Mattix



Agenda

- Speakers Intro
- What is WG AI Gateway?
- Proposals
  - Payload Processing
  - Backend
  - Endpoint Selector
  - Routability
  - Multi-Stage Route
- What's Next
- Get Involved

Speakers Intro

KubeCon CloudNativeCon North America 2026



Nir Rozenbaum
Senior Engineering Lead
NVIDIA



Keith Mattix
Senior Architect
Solo.io

What is WG AI Gateway?

Standardize AI traffic management at the Gateway layer

Standards development

Declarative APIs and standards for AI workload networking in Kubernetes

Leveraging existing standards

Reuse Gateway API building blocks whenever possible

Community collaboration

Foster discussion and build consensus on best practices for AI workloads

Extensible architecture

Composable & pluggable processing for AI-specific gateway extensions

KubeCon CloudNativeCon North America 2026

Payload Processing



Payload Processing – Why?

AI Inference Request

```
POST /v1/chat/completions
Authorization: Bearer eyJhbGci...
Content-Type: application/json

{
  "model": "llama-3-8b",
  "messages": [
    {"role": "user", "content": "Hi"}
  ],
  "stream": true
}
```

The routing key is in the payload

Example: the model lives in the JSON body

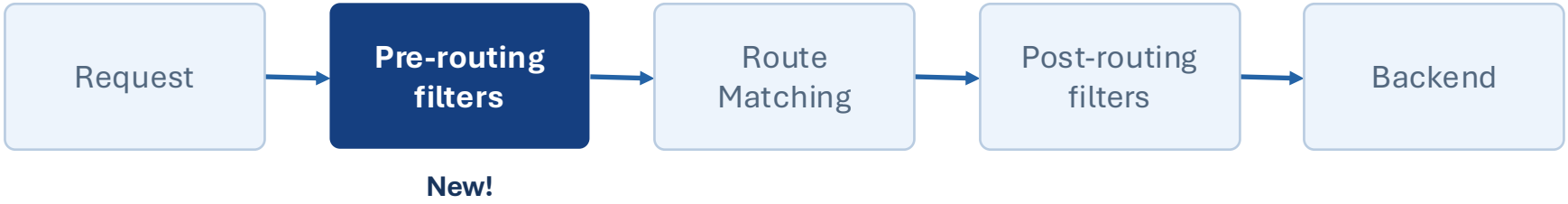
Gateway API can't see it

Matches use path, headers, query and method, and filters only run after a route is chosen

So everyone build their own

ext_proc services, Lua, Wasm, vendor policies

GEP-5224: Pre-Routing Filters



The gap

Every filter runs after route matching, so nothing can shape the request before a route is chosen.

The addition

An ordered list of filters on the Listener that runs before any route is matched.

The guarantees

Filters run in listed order, routes are matched once after the last one, and any filter can reject the request.

GEP-5293: Dynamic Headers

KubeCon CloudNativeCon North America 2026

Today: a fixed value

```
- type: RequestHeaderModifier
  requestHeaderModifier:
    set:
    - name: X-Model
      value: llama-3-8b
```

GEP-5293: a value from the request

```
- type: RequestHeaderModifier
  requestHeaderModifier:
    set:
    - name: X-Model
      valueFrom:
        json(request.body).model
```

Illustrative syntax

Values can come from:

- Body fields
- JWT claims
- Headers, query, cookies
- ...

Payload Processing – How?

Read the body → set X-Model header → match the route → InferencePool

Gateway: GEP-5224 + GEP-5293

```
kind: Gateway
spec:
  listeners:
  - name: http
    filters:
      requests:
      - type: RequestHeaderModifier
        requestHeaderModifier:
          set:
          - name: X-Model
            valueFrom:
              json(request.body).model
```

HTTPRoute: unchanged API

```
kind: HTTPRoute
spec:
  rules:
  - matches:
    - headers:
      - name: X-Model
        value: llama-3-8b
    backendRefs:
    - kind: InferencePool
      name: llama-3-8b
```

GEP-5224 gives the phase, GEP-5293 gives the dynamic value.

Together: body-based routing, portable and in-process.

Payload Processing – Still Open

Current focus

* In-process pre-routing filters
* Payload processing as dynamic filters

Future

External processing:
Calling out to an external service, such as ext_proc, remains an open question.

KubeCon CloudNativeCon North America 2026

Backend



The problem: Service is over-employed

The Service resource has (at least) 2 jobs:

Frontend: How is it discovered?

Stable DNS name • ClusterIP • Producer oriented

Backend: How do we connect?

Endpoints • TLS • Wire Protocol • Consumer Oriented

1 resource: 2 concerns and 2 personas

Gateways / proxies need backend info to make connections – especially with AI

Why the existing model breaks down

External destinations

An external LLM or MCP server requires an ExternalName Service just to satisfy backendRef (and ExternalName has various security risks)

Scattered settings

BackendTLSPolicy couples backend (client) configuration with producer (server-side) ownership, leading to sprawl and complex debugging flows

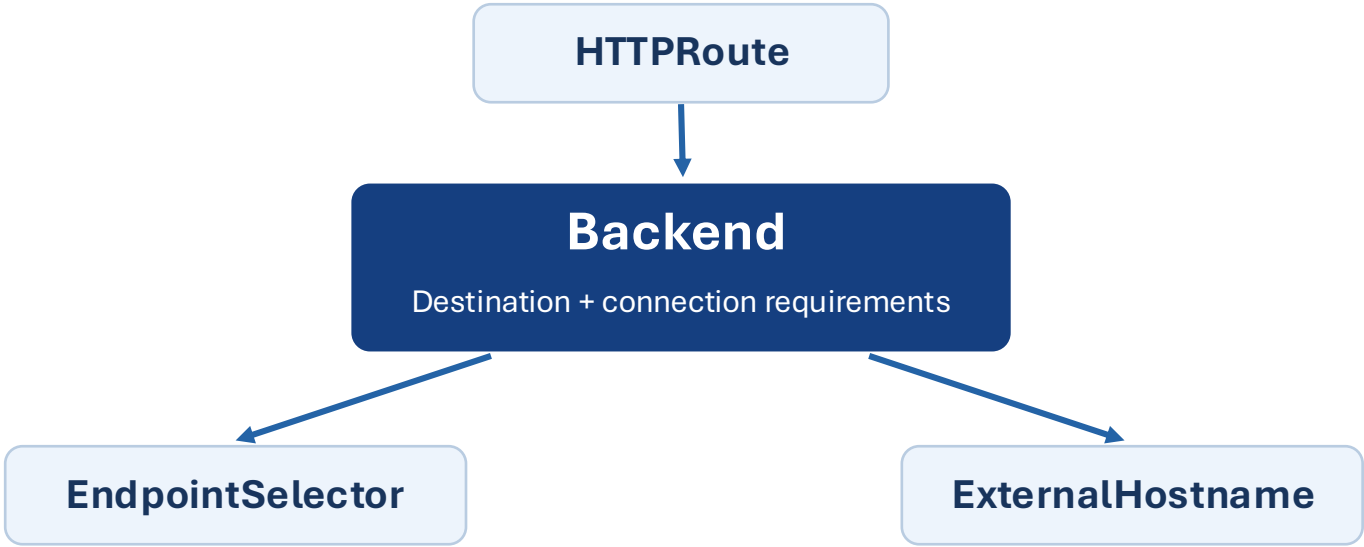
Cost/Benefit Mismatch

Trying to get a bag of endpoints without a ClusterIP (e.g. to prevent kube-proxy load balancing) means controllers have to duplicate all the work of endpointslice-controller. Often not worth it

AI egress exposed a generic Kubernetes networking gap—not an AI-only problem.

Backend: a consumer-owned destination

BACKEND



One abstraction: internal endpoints or external FQDNs.

Backend is not a replacement for Service

Rather, it's a fit-for-purpose subset

KEEP SERVICE

ClusterIP, DNS, and general-purpose in-cluster discovery remain Service responsibilities.

ADD BACKEND

Gateway-native endpoint selection and connection metadata can evolve independently.

Kubernetes Foundation: EndpointSelector

Extended Capabilities: ExternalHostname, TLS, MCP…

Our goal is to be backwards compatible but forward-looking

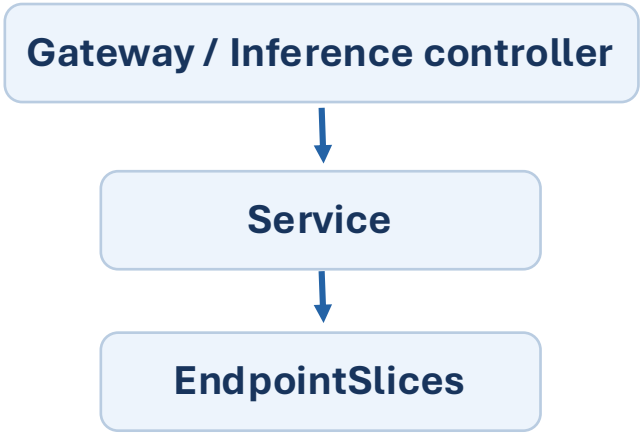
KubeCon CloudNativeCon North America 2026

Endpoint Selector



Endpoint discovery - Service dependency

ENDPOINTSELECTOR



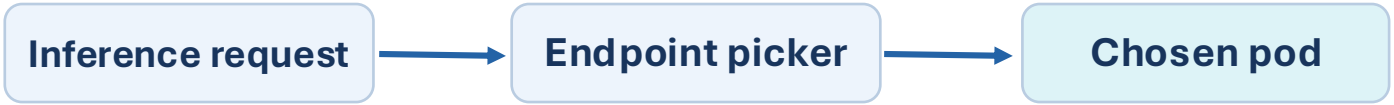
The forced choice for implementations

Create a Service and accept a VIP/DNS frontend, or maintain EndpointSlices yourself as pods churn.

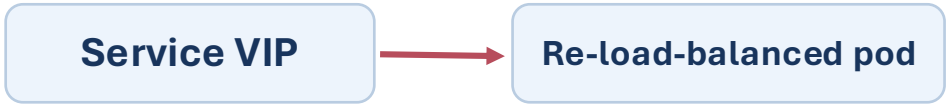
Endpoint discovery and frontend exposure are separate responsibilities.

Why inference routing makes this urgent

ENDPOINTSELECTOR



Correct: picker chooses the target based on model-serving state.

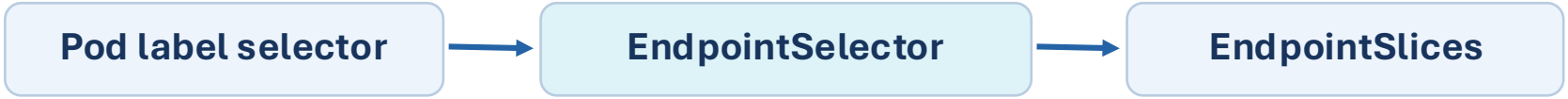


Risk: a second load-balancing decision defeats the picker.

Intelligent routing needs direct, reliably maintained endpoint sets.

Reuse the Kubernetes machinery

ENDPOINTSELECTOR



Kubernetes reconciles

Tracks pod additions, removals, and readiness changes.

Controllers consume

Gateway, inference, mesh, and other systems reuse the endpoint set.

No user-authored Service; no duplicate endpoint-controller logic.

Two APIs, two layers of responsibility

ENDPOINTSELECTOR

Gateway API: Backend

What destination should this gateway use, and with what TLS/protocol settings?

Kubernetes: EndpointSelector

Which pods match a selector, and what EndpointSlices represent them?

Gateway API defines connection intent; Kubernetes supplies endpoint discovery.

KubeCon CloudNativeCon North America 2026

Routability



An address needs a reachability contract

ROUTABILITY

Internal-only gateways

Knative and platform teams need in-cluster addresses without provider-specific annotations.

Egress gateways

Workloads need an address to send outbound traffic through a gateway.

Multi-address gateways

One Gateway can expose addresses with different reachability scopes.

An IP address or hostname alone does not communicate where it is reachable.

Routability belongs to each address

ROUTABILITY



spec.addresses requests a scope; status.addresses reports the provisioned scope.

What “Cluster” does—and does not—mean

ROUTABILITY

GUARANTEED

For IPAddress: a Kubernetes Service ClusterIP in the ServiceCIDR, reachable from inside the cluster.

NOT GUARANTEED

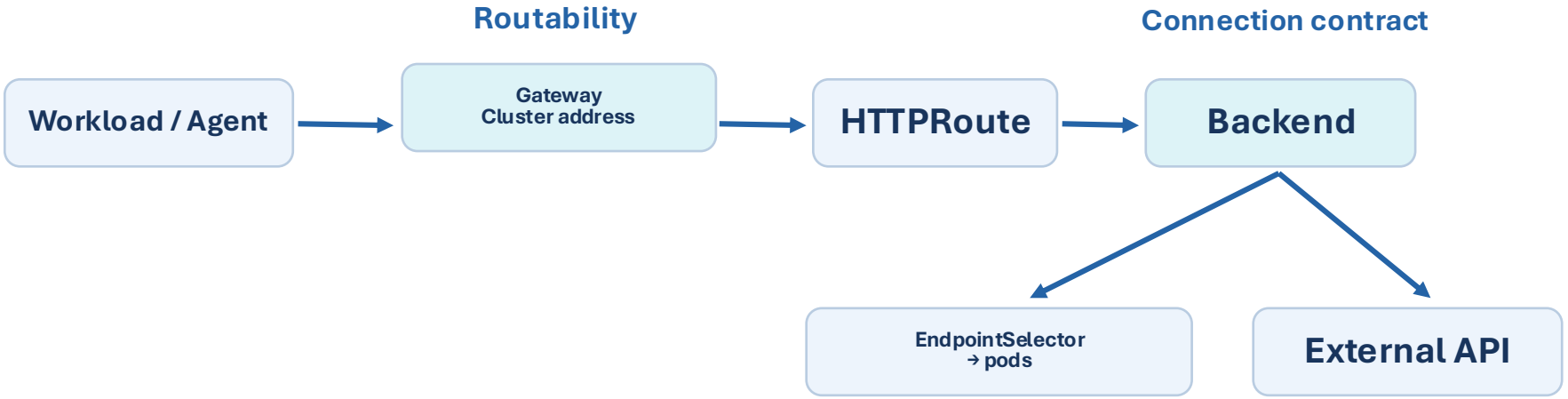
Not a network-isolation policy. An administrator may make the address routable beyond the cluster.

Reachability description ≠ access-control enforcement.

Putting the pieces together

KubeCon CloudNativeCon North America 2026

ARCHITECTURE



Reach the Gateway → choose the destination → discover endpoints or connect externally.

KubeCon CloudNativeCon North America 2026

Multi-Stage Route



Multi-Stage-Route – Why?

Today: one request, one backend call



AI workloads: one request, several backend calls



Fragmented

Different distributed inference projects orchestrate prefill and decode differently

Duplicated

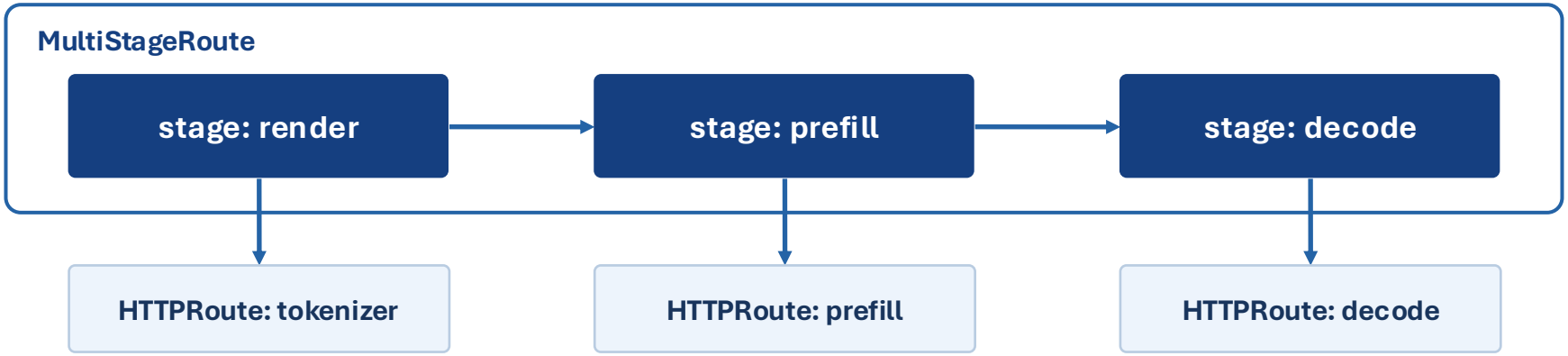
Retries, timeouts and failure handling are rebuilt outside the gateway

Opaque

Operators cannot follow one request across its stages

GEP-5250: Multi-Stage-Route

KubeCon CloudNativeCon North America 2026



Composition

Which stages run, in what order, and how each stage's request is built

Interaction

Each referenced route keeps its own backend, timeouts and retries

Reuse

A route works standalone or as a stage, unchanged

GEP-5250: Multi-Stage-Route

KubeCon CloudNativeCon North America 2026

MultiStageRoute

```
kind: MultiStageRoute
spec:
  parentRefs: [{name: ai-gateway}]
  matches:
  - path: {value: /v1/chat/completions}
  stages:
  - name: render
    routeRefs: [{name: tokenizer}]
  - name: decode-probe
    routeRefs: [{name: decode}]
    onResponse: [...]
  - name: prefill
    routeRefs: [{name: prefill}]
    filters: [...]
  - name: decode
    routeRefs: [{name: decode}]
    filters: [...]
```

HTTPRoute: prefill (unchanged)

```
kind: HTTPRoute
metadata: {name: prefill}
spec:
  rules:
  - filters:
    - type: RequestBodyModifier
      requestBodyModifier:
        set:
        - {name: /max_tokens, value: '1'}
    backendRefs: [{kind: InferencePool}]
    timeouts: {request: 60s}
```

Stages reference routes. The route owns the interaction: prefill means max_tokens=1, with or without a pipeline.

Between Stages: Payload Processing

Between stages

```
- name: decode-probe
  routeRefs: [{name: decode}]
  filters:
  - type: RequestBodyModifier
    requestBodyModifier:
      set:
      - name: /token_ids
        valueFrom: json(
          stage("render").response.body,
          "$.token_ids")
  onResponse:
  - match: {status: 200}
    action: Complete
  - match: {status: 412}
    action: Continue
```

stage()

A filter can read an earlier stage's response

Same filter style

RequestBodyModifier mirrors the dynamic RequestHeaderModifier

onResponse

Per request: Complete returns this response, Continue moves on

KubeCon CloudNativeCon
North America 2026

What's Next



What's Next

- Bring current proposals to GA
- Extend and close gaps, e.g., External Processing (in Payload Processing)
- New Proposals
- Consolidate AI Aware Networking efforts (AI Gateway WG, KAN, etc).

KubeCon CloudNativeCon North America 2026

Get Involved



Get Involved

KubeCon CloudNativeCon North America 2026

- Visit the working group GitHub repository
- Join the weekly meeting on Thursdays at 2PM EST
- Connect with us on Slack (#wg-ai-gateway)
  (visit https://slack.k8s.io/ for an invitation)
- Join the WG AI Gateway mailing list

Thank You!

Rate this session

