📝 Meeting Notes

Envoy AI Gateway Community Meeting

GitHub - envoyproxy/ai-gateway | Website

📅**Every Thursday**
8:00 am US Pacific | 11:00 am US Eastern | 4:00pm London, UK | 5:00pm Europe Central

📞**Zoom link - Join here**

🤝**Code of conduct - View here**

🎥 **Recordings - View here** *or view in your personal dashboard on LFX*

Feel free to add topics to discuss to meeting agendas.

November 27, 2025

Status Future

Attendees

Please add your name to the attendee list.

- Host: TBD

Agenda

Feel free to add items to the agenda. Please add your name in brackets ***[name].***

- Standing Item **[Host]** Welcome
- ***Example:*** Discussion **[Name]**
-
- Standing Item AOB - Any Other Business

November 20, 2025

Status Future

Attendees

Please add your name to the attendee list.

- Host: TBD

Agenda

Feel free to add items to the agenda. Please add your name in brackets ***[name].***

- Standing Item **[Host]** Welcome

- ***Example:*** Discussion **[Name]**
-
- Standing Item AOB - Any Other Business
-

November 13, 2025

Status Skipped KubeCon Week

November 6, 2025

Status Next

Attendees

Please add your name to the attendee list.

- Host: TBD

Agenda

Feel free to add items to the agenda. Please add your name in brackets ***[name].***

- Standing Item **[Host]** Welcome
- ***Example:*** Discussion **[Name]**
-
- Standing Item AOB - Any Other Business

October 30, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Dan Sun (Bloomberg)
- Alexa Griffith (Bloomberg)
- Aaron Choo (Bloomberg)
- Gavrish Prabhu (Nutanix)
- Siddharth Shah (Nutanix)
- Hrushikesh Patil(Nutanix)
- Ayush Sawant (Nutanix)
- Ignasi Barrera (Tetrate) GH: nacx
- Sukumar Gaonkar (Bloomberg)
-

Agenda

- Welcome [Host]
- V0.4 Release Timeline & Preview [Erica]
    - What's the target date
    - Preview of Release notes https://deploy-preview-1456--envoy-ai-gateway.netlify.app/release-notes/v0.4
- Unified thinking api https://github.com/envoyproxy/ai-gateway/issues/1463 (yang/dan)
-
- AOB (Any Other Business)

October 23, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Aaron Choo (Bloomberg)
- Dan Sun (Bloomberg)
- Ignasi Barrera (Tetrate) GH: nacx
- Yuhong Du (Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Takeshi Yoned (Tetrate)
- Siddharth Shah (Nutanix)
- Gavrish Prabhu (Nutanix)
- Hrushikesh Patil(Nutanix)
- Yao Weng (Bloomberg)
- Xiaolin Lin (Bloomberg)
- Ayush Sawant (Nutanix)
- Sailesh Duddupudi (Nutanix)
- Manish Singh(Datum.net) GH:mksinghtx
- Camilo Aguilar (Redpanda)
- Alexa Griffith (Bloomberg)
- Brett Mertens (Datum.net)

Agenda

- ☑️Welcome [Host]
- ☑️Welcome new maintainers [Erica]
- ☑️MCP Proxy Demo with aigw CLI [Ignasi]
- ☑️https://github.com/envoyproxy/ai-gateway/pull/1280 [Hrushikesh]

- ☑️Per backend and route backend ref header mutation for ChatCompletion [Sukumar/Aaron]
    - https://github.com/envoyproxy/ai-gateway/pull/1414
- ☑️Message timeout issues with envoy 1.36 ? [dan/yuhong]
    - No longer honoring the extproc message timeout with 10s we set on aigw extproc
- AOB (Standing Item)

October 16, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Ross Morrow (CoactiveAI)
- Aaron Choo (Bloomberg)
- Ignasi Barrera (Tetrate) GH: nacx
- Sukumar Gaonkar (Bloomberg)
- Siddharth Shah (Nutanix)
- Hrushikesh Patil(Nutanix)
- Takeshi Yoneda (Tetrate): GH mathetake
- Gavrish Prabhu(Nutanix) GH: gavrissh
- Camilo Aguilar (Redpanda)
- Ayush Sawant (Nutanix)
- Manish Singh (Datum.net) GH:mksinghtx
- Sailesh Duddupudi (Nutanix)
- Xiaolin Lin (Bloomberg)
- Ravi Verma (IMESH)

Agenda

Please add items to the agenda :)

- Welcome [Host]
- Finalize 0.4 timeline [Erica]
    - Blocked on https://github.com/envoyproxy/gateway/issues/7248
- Welcome new maintainers [Erica]
- https://github.com/envoyproxy/ai-gateway/pull/1280 [Hrushikesh]
- Any PRs Waiting on Review? [Erica]
- Notable Improvements or New Capabilities [Erica]
    - **First-party Anthropic (api.anthropic.com) Support:** The team has added support for first-party Anthropic, which allows users to interact with the Anthropic API directly.
    - **Improved Metrics and Logging:** The team has made improvements to the metrics and logging system, including distinguishing the /messages endpoint metrics from the /chat/completions endpoint
    - Future TODO: https://github.com/envoyproxy/ai-gateway/issues/1372
- AOB (Standing Item)

October 9, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate) GH: mathetake
- Yuhong Du (Bloomberg)
- Ajay Nagar (Nutanix)
- Ashwini Vasanth (Nutanix)
- Aaron Choo (Bloomberg)

- Gavrish Prabhu (Nutanix)
- Ayush Sawant (Nutanix)
- Yao Weng (Bloomberg)

Agenda

- Welcome [Host]
- Change encoding/json to goccy/go-json to improve decode efficiency [Yuhong Du]
    - For large payload (larger than 1mb, preprocessing takes a long time)
    - Tested with 1-5 mb payload, switching to goccy/go-json could reduce preprocessing latency by 30-35%
    - goccy/go-json does not support omitzero, it uses omitempty instead
- Bug Reports [Erica]
    - https://github.com/envoyproxy/ai-gateway/issues?q=is%3Aissue+is%3Aopen+label%3Abug
- Flaky Tests [Erica]
    - Noticed we seem to have flaky tests, some efforts to improve
- ReRank API Support (Ayush)
- 0.4 timeline
    - Roughly End of October
- AOB (Standing Item)

October 2, 2025

Status Past

Attendees

- Host: Aaron Choo (Bloomberg)

- Yao Weng (Bloomberg)
- Yan Avlasov - Google
- Takeshi Yoneda (Tetrate)
- Siddharth Shah (Nutanix)
- Ayush Sawant (Nutanix)
- Alexa Griffith (Bloomberg)
- Gavrish Prabhu (Nutanix)
- Dan Sun (Bloomberg)

Agenda

- Welcome [Host]
- [Takeshi] MCP
    - Blog: https://aigateway.envoyproxy.io/blog/mcp-implementation
    - PR: https://github.com/envoyproxy/ai-gateway/pull/1260
    - Design Doc: https://github.com/nacx/ai-gateway/blob/58e17c829e967425a41c044aab18af6f7c02dae7/docs/proposals/005-mcp-gateway/proposal.md
- [Yan Avlasov] MCP support is being added to Envoy and will be used as a reference for k8s agent-gateway
- AOB (Standing Item)

September 25, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Ajay Nagar (Nutanix)
- Siddharth Shah (Nutanix)

- Ayush Sawant (Nutanix)
- Sailesh Duddupudi (Nutanix)
- Dan Sun (Bloomberg)
- Aaron Choo (Bloomberg)
- Yan Avlasov (Google)
- Gavrish Prabhu (Nutanix)

Agenda

- Welcome [Host]
- Prioritized Request Feature Issue request: https://github.com/envoyproxy/ai-gateway/issues/1228
- AOB (Standing Item)

September 18, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Siddharth Shah (Nutanix)
- Sailesh D (Nutanix)
- Aaron Choo (Bloomberg)
- Johnu George(Nutanix)
- Gavrish Prabhu (Nutanix)
- Takeshi Yoneda (Tetrate)
- Ajay Nagar (Nutanix)

Agenda

- Welcome [Host]
- Key Themes in Issues [Erica]
    - The new issues for the envoyproxy/ai-gateway project cover a few key themes:
        - **Usability and Ease of Use**: Issues around automatically setting default configurations and removing validation steps suggest a need to improve the developer experience and lower the barrier to entry.
        - **Compatibility and Interoperability**: Requests for OpenAI-compatible endpoints and handling of larger GRPC messages indicate a focus on ensuring the gateway works seamlessly with a variety of AI/ML tools and services.
        - **Configuration and Deployment**: The issue around migrating configuration from raw YAML to Helm values points to a need to streamline the deployment and management of the gateway.
- Pending Open PRs [Erica / All]
    - Any Blockers?
- Helm Restructure: https://github.com/envoyproxy/ai-gateway/issues/1186 [Sailesh D]
- AOB (Standing Item)
    - *Any other business?*
    - Adding Envoy AIGW to https://github.com/awslabs/ai-on-eks ?

September 11, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate)
- Xiaolin Lin (Bloomberg)
- Alexa Griffith (Bloomberg)
- Aaron Choo (Bloomberg)
- Siddharth Shah (Nutanix)
- Gavrish Prabhu (Nutanix)
- Johnu George(Nutanix)
- Ayush Sawant (Nutanix)
- Snehlata (Nutanix)
- Jay Prasad (Nutanix)

Agenda

- Welcome [Host]
- Issues to Address [All]
    - #1125: Usage diagram is misleading with regards to supported features
    - #1119: AIServiceBackend status is not correct when updating to an invalid BSP ref name
    - Support for OpenAI Responses API #980
- Notable Open PRs [All]
    - feat: add bedrock reasoning stream support #1173
    - refactor: moves filterapi under internal #1179
    - refactor: using openaigo SDK for non-stream response #1147 - Note this is a async discussion

- Discussion - AIGatewayRouteRuleBackendRef does not have Namespace field, Can we add it? [Gavrish]
    - We need to implement “ReferenceGrant” stuff in AIGW [Takeshi] https://gateway.envoyproxy.io/docs/api/gateway_api/r in eferencegrant/
- AOB (Standing Item)

September 4, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Ayush Sawant (Nutanix)
- Siddharth Shah (Nutanix)
- Ajay Nagar (Nutanix)
- Takeshi Yoneda (Tetrate)
- Gavrish Prabhu (Nutanix)

Agenda

- ✅Welcome [Host]
- ✅Adopters on Site [Erica]
- ✅Fix request duration metric recording issue: https://github.com/envoyproxy/ai-gateway/pull/1160 (Ayush)
- Recently Raised PRs needing Review [All]
    - https://github.com/envoyproxy/ai-gateway/pull/1138
    - https://github.com/envoyproxy/ai-gateway/pull/1160
- Recently Raised Issues worth Noting [All]
    - https://github.com/envoyproxy/ai-gateway/issues/1149

- AOB (Standing Item)

August 28, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Siddharth Shah (Nutanix)
- Gavrish Prabhu (Nutanix)
- Johnu George(Nutanix)
- Aaron Choo (Bloomberg)
- Ayush Sawant (Nutanix)
- Ajay Nagar (Nutanix)
- Snehlata (Nutanix)
- Takeshi (Tetrate)
- Ross Morrow (CoactiveAI)
- Ben Green (Hicap)
- Javier Cevallos (Hicap)
- Sukumar Gaonkar (Bloomberg)

Agenda

- Welcome [Host]
- Notable new issues raised (Erica / All)
- Notable new features merged (Erica / All)
- Prune Security Headers (Siddharth)
- Performance Numbers (Gavrish)
- AOB (Standing Item)

August 21, 2025

Status Past

Attendees

- Host: Aaron (Bloomberg)
- Xiaolin Lin(Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Siddharth Shah (Nutanix)
- Jay Prasad (Nutanix)
- Gavrish Prabhu (Nutanix)
- Alexa Griffith (Bloomberg)
- Yao Weng (Bloomberg)
- Johnu(Nutanix)

Agenda

- Welcome [Host]
- ***Refactor to using openaigo SDK in ResponseBody*** *(see: code changes) - This is getting larger and more complicated. Would appreciate any feedback before more time is put in.*
    - *Started from adding support for bedrock reasoning where refactor was contained to code changes there (see: code changes)*
        - *Reasoning content needs to be a custom field returned since openai doesn't include this. LiteLLM does something similar. I created custom structs with custom marshal (see: code)*
    - *Merged main then errors - because the* ***new span implementation*** *meant that every ResponseBody must use same struct, which led to*
    - *New PR to "quickly" (famous last words for refactor lol)* ***refactor for every ResponseBody to use openaigo SDK ChatCompletion struct***

    - *The new issue: Recently added field "Obfuscation", not in openai docs, not in openaigo SDK for ChatCompletion, but it is in the openai platform response. (see: responses)*
    - *What we can do now: custom struct similar to solution for reasoningContent for aws bedrock resp.*
        - *Created Issue in openai go repo*
        - *Pro: I favor using openai go because managing our custom structs is a pain. Seems like this openaigo repo will gain popularity and support. (i hope)*
        - *Con: now this becomes more confusing/complicated.*
- v0.3 release 🎉🎉🎉
- AOB (Standing Item)

August 14, 2025

Status Past

Attendees

- Host: Takeshi Yoneda or Dan Sun
- Alam Ahmed (cae)
- Ajay Nagar (Nutanix)
- Xiaolin lin (Bloomberg)
- Aaron Choo (Bloomberg)
- Gavrish Prabhu
- Siddharth Shah (Nutanix)

Agenda

- Welcome [Host]
- v0.3 release status

    - https://github.com/envoyproxy/ai-gateway/pull/1068
    - https://github.com/envoyproxy/ai-gateway/issues/1015
    - *https://github.com/envoyproxy/ai-gateway/issues/840*
- *Allow configuring default fallback @siddharth*
- AOB (Standing Item)

August 7, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate)
- Siddharth Shah (Nutanix)
- Jay Prasad (Nutanix)
- Xiaolin lin (bloomberg)
- Aaron Choo (bloomberg)
- Arif Setiawan (Tetrate)
- Mohammad Anwari (Tetrate) GH: mdamt
- Alam Zaib Ahmad (cae)
- Ajay Nagar (Nutanix)
- Yao Weng (Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Madhu CH

Agenda

- Welcome [Host]
- Adopter Story: Tetrate Agent Router Service [Mohammad Anwari]
    - *Bring your questions for Adopter Anwari*

- Review Draft v0.3 Release Notes [Erica]
    - https://deploy-preview-983--envoy-ai-gateway.netlify.app/release-notes/v0.3
- Prefix-based URL handling breaks extproc processing #1002 [Siddharth]
    - https://github.com/envoyproxy/ai-gateway/issues/1002
- Showcase VSCode Plugin POC [Alam]
- *Add items here*
-
- AOB (Standing Item)

July 31, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Yan Avlasov (Google)
- Sailesh D (Nutanix)
- Johnu(Nutanix)
- Siddharth Shah (Nutanix)
- Thomas Gschwendtner (Google)
- Suren Raju (Careem)
- Alam Ahmed(cae) #
- Jay Prasad (Nutanix)
- Ajay Nagar (Nutanix)
- Gavrish Prabhu (Nutanix)
- Dan Sun (Bloomberg)
- Aaron Choo (Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Xiaolin Lin (Bloomberg)

Agenda

- Welcome [Host]
- Adopter Story Preview: Tetrate Agent Router [Erica]
- V0.3 Release [Erica]
- Road to GA [Erica]
- [Yan Avlasov] Participation in k8s AI gateway WG https://github.com/kubernetes/community/pull/8521
-
- AOB (Standing Item)

July 24, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Alam Ahmad
- SIddharth Shah (Nutanix)
- Ayush Sawant (Nutanix)
- Ajay Nagar (Nutanix) GH: nagar-ajay
- Takeshi Yoneda (Tetrate) GH: mathetake
- Gavrish Prabhu (Nutanix)
- Curtis Maddalozzo (Bloomberg)
- Xiaolin Lin (Bloomberg)
- Aaron Choo (Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Jay Prasad (Nutanix)
- Boteng Yao (Google)

Agenda

- Welcome [Host]
- *Add items here*
- V0.3 Admin [Erica]
- support json-patch in request extra_body: Draft PR [Sukumar]
- Vertex Integration Status Check-in [Erica]
- AI Gateway questions [Johnu]
- AOB (Standing Item)

July 17, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
-
- *Vadim Dabravolski (Bloomberg)*
- Suren Vartanian (Bloomberg)
- Aaron Choo (Bloomberg)
- Dan Sun (Bloomberg)
- Curtis Maddalozzo *(Bloomberg)*
- Alexa Griffith (Bloomberg)
- Yan Avlasov (Google)
- Yao Weng (Bloomberg)
- Xiaolin Lin (Bloomberg)
- Sukumar Gaonkar (Bloomberg)
- Nagar-ajay (Nutanix)

Agenda

- Welcome [Host]
- *Native Anthropic API support*
    - *https://github.com/envoyproxy/ai-gateway/issues/847*
    - Have a pass-through option for
- Endpoint Picker support progress https://github.com/envoyproxy/ai-gateway/pull/823
    - Did multiple iterations, finally getting closer to the finish line
    - Only blocker: https://github.com/envoyproxy/gateway/pull/6524
- [Yan] Envoy buffering improvements: https://github.com/envoyproxy/envoy/pull/40256 and https://github.com/envoyproxy/envoy/pull/40254
- Support for retries across multiple clusters: https://github.com/envoyproxy/envoy/issues/40264
- MCP Support in 0.4 [Erica]
- AOB (Standing Item)

July 10, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate) GH: mathetake
- Sukumar Gaonkar (Bloomberg)
- *Curtis Maddalozzo (Bloomberg)*
- *Yao Weng (Bloomberg)*
- *Aaron Choo (Bloomberg)*
- *Dan Sun (Bloomberg)*

- Yan Avlasov (Google)
- Xiaolin Lin (Bloomberg)

Agenda

- Welcome [Host]
- New Features Merged Recently [Erica]
    - https://github.com/envoyproxy/ai-gateway/pull/793
    - https://github.com/envoyproxy/ai-gateway/pull/819 - a step in GCP Gemini support
- Inference Extension Work Progress Check-in [Erica]
    - Getting ready for review, expecting a PR in next few days soon
- Review for https://github.com/envoyproxy/ai-gateway/issues/844 [Sukumar]
- Problems
    - https://github.com/envoyproxy/ai-gateway/issues/840
- AOB (Standing Item)

July 3, 2025

Status Past

Attendees

- Host: Dan Sun
- Takeshi Yoneda (tetrate)
- Sukumar Gaonkar
- *Liping Tang (F5)*

Agenda

- Welcome [Host]
- *Content safety check (Liping Tang F5)*
- Kubecon JP talk
    - Access AI Models Anywhere: Scaling AI Traffic with Envoy AI Gateway
- V0.3 release features
    - GCP / Anthropic
    - Inference Gateway proposal
- Local development requires Envoy >=v1.34
    - https://github.com/envoyproxy/ai-gateway/blob/310d428f361360010d20fc9aa6e11a4ca5a0bea0/CONTRIBUTING.md
-
- AOB (Standing Item)

June 26, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- *Add your name*

Agenda

- Welcome [Host]
- [Yan Avlasov]: Buffering for large requests + body rewrites: https://github.com/envoyproxy/envoy/issues/40028
- Highlight discussions on API Design [Erica]
    - https://github.com/envoyproxy/ai-gateway/issues/764
    - https://github.com/envoyproxy/ai-gateway/issues/675

- 0.3 release plan
    - https://github.com/envoyproxy/ai-gateway/milestone/3
-
- AOB (Standing Item)

June 19, 2025

Status Skipped

CANCELLED IN OBSERVANCE OF JUNETEENTH

June 12, 2025

Status Next

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- *Add your name*
- Sukumar Gaonkar - Bloomberg
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- *Alam Zaib Ahmad*
- *Yuhong Du* (Bloomberg)
- Dan Sun (Bloomberg)
- Alexa Griffith (Bloomberg)
- Curtis Maddalozzo (Bloomberg)

Agenda

- Welcome [Host]
- v0.2 Release Review [Erica]
- AIGateway Resources Conversation [Dan/Erica]
- v0.3 Call for Features [Erica]
    - Review open issues https://github.com/envoyproxy/ai-gateway/issues
    - New feature ideas?
- Problems [All]
- *Add items here*
-
- AOB (Standing Item)

June 5, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Yao Weng (Bloomberg)
- Curtis Maddalozzo (Bloomberg)
- Aaron Choo (Bloomberg)
- *Sukumar Gaonkar (Bloomberg)*
- *Alam zaib ahmad (cae)*
- Alexa Griffith (Bloomberg)
- Dan Sun (Bloomberg)
- Takeshi (Tetrate)

Agenda

- Welcome [Host]
- v0.2 Release [Erica]
    - Any Blockers
        - Bug found: https://github.com/envoyproxy/ai-gateway/issues/682
            - Alexa will raise a PR for the fix later today
                - Takeshi to review
    - Review Release Notes https://deploy-preview-679--envoy-ai-gateway.netlify.app/release-notes/v0.2/
- Discuss how to handle backend specific parameters in request/response transformation [Sukumar]
- Kubernetes AI gateway SIG
    - [PUBLIC] Kubernetes AI Gateway WG Proposal
- *Add items here*
- v0.3 Call for Feature Issues [Erica]
- AOB (Standing Item)

May 29, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Rob Scott (Google) GH: robscott
- Takeshi Yoneda (tetrate) GH: mathetake
- Curtis Maddalozzo (Bloomberg) GH: cmaddalozzo
- *Add your name*

Agenda

- Welcome
- 0.2 Release [Erica]
    - Remaining items:
        - Reinstate Inference extension implementation [Takesh]
        - Azure OIDC [aaron]
        - Docs for fallback, retry and LB
    - Target date: Latest 7 June
- *Implement cascading retry by priority [Curtis/Yao]*
    - *Support host and priority predicates in backendTrafficPolicy retry config*
        - *https://github.com/envoyproxy/gateway/issues/6181*
    - *Patch priority in envoy ai gateway extension server*
        - *Ideally it should be defined on HTTPRoute, there was an issue Arko created 2 years ago*
        - *https://github.com/kubernetes-sigs/gateway-api/discussions/2304*
- AOB (Standing Item)

May 22, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Yan Avlasov (Google)
- Dan Sun (Bloomberg)
- Yao Weng (Bloomberg)
- Curtis Maddalozzo (Bloomberg)
- Alexa Griffith (Bloomberg)
- Aaron Choo (Bloomberg)
- James Cummins (Microsoft)
- Takeshi Yoneda (Tetrate)
- Anjali Taneja (Bloomberg)
- Shashank Goel

Agenda

- Welcome
- 0.2 Release [Erica]
    - New Features
    - Breaking Changes
- Refactor to ExtProc as Sidecar [Takeshi]
    - https://github.com/envoyproxy/ai-gateway/pull/599
    - https://github.com/envoyproxy/ai-gateway/pull/629

- Model fallback [Curtis/Yao]
- AOB

May 15, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- Ignasi Barrera (Tetrate) GH: nacx
- Shashank Goel
- Ajay Nagar
- Dan Sun (Bloomberg)

Agenda

- Welcome
- 0.2 Release [Erica]
- GCP Integration [Dan]
- OpenAI API Evolution [Erica]
- MCP Protocol Support Discussion from Previous meeting: [Proposal: Envoy Support for Model Context Protocol (MCP) · Issue #39174 · envoyproxy/envoy · GitHub]
- Trivia Game Demo [Erica]
- KubeCon NA Proposals

May 8, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Yan Avlasov (Google)
- Dan Sun (Bloomberg)
- Alam Zaib Ahmed (CAE)

Agenda

- Welcome
- Failover Progress [Dan]
- CNCF Live AI GW [Erica]
- Semantic Caching Feature Question [Erica]

May 1, 2025

Status Past

Attendees

- Host: Erica Hugherbg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate) GH mathetake
- Yan Avlasov (Google)
- Alam Ahmed (CAE)

- Ellice Kwak (Tetrate)
- John Landa (Tetrate) GH johnlanda
- Sneha Prasad
- Dan Sun(bloomberg)

Agenda

- Welcome
    - Introductions
- Discuss Fallback design doc

  Envoy AI Gateway Model Fallback Design Proposal **Dan/Yao**
    - Update on POC for Fallback Upstream Auth/Translator - Takeshi
- AOB
    - Envoy in the Era of AI: Building the Future of Traffic Handling Together
    - https://github.com/i-am-bee and ACP - Erica

April 24, 2025

Status Past

Attendees

- Host: Erica Hugherbg (Tetrate) GH: missberg
- Xiaolin Lin (Bloomberg LP) GH: xiaoliln593
- Yao Weng (Bloomberg LP)
- Gavrish Prabhu (Nutanix)
- Ajay Nagar (Nutanix)
- Takeshi Yoneda (Tetrate) GH mathetake

Agenda

- Welcome
- Review Action Items from Last Meeting
    - LLM Metrics-based routing
        - Draft configuration/API for external endpoint picker by next meeting. **Dan / Siva**
        - Pending PR for Envoy load balancing extension. **Yan**
    - 0.2 Release Status **Takeshi**
    - Support MCP and A2A Protocols
        - An Envoy issue for required data plane features to support MCP. **Yan**
            - **https://github.com/envoyproxy/envoy/issues/39174**
        - An AI Gateway issue to gauge community interest and guide config/resource design. **Yan**
    - Blog Posts
        - Reference Architectures (Dan to set date)
        - AI Gateway + KServe (Dan/Siva)
        - Inference Extension & Endpoint Picker (Erica + Yan + Takeshi; target: end of April)
        - Upstream Auth in AI Gateway (Xiaolin + Aaron; target: mid-May)
- Discuss Fallback design doc
  Envoy AI Gateway Model Fallback Design Proposal **Dan/Yao**
- Adding more Providers for Translation (Erica)
- KubeCon NA ATL (Erica)
- Any Other Business

April 17, 2025

Status Past

Attendees

- Host: Erica Hugherbg (Tetrate) GH: missberg
- John Landa (Tetrate) GH: johnlanda
- Xiaolin Lin (Bloomberg) GH: xiaolin593
- Ignasi Barrera (Tetrate) GH: nacx
- Takeshi Yoneda (Tetrate) GH: mathetake
- Sivanantham Chinnaiyan(Ideas2IT Technologies)
- Yan Avlasov (Google)

Agenda

- Welcome
- LLM Metrics-based routing using inference extension API (Dan/Siva)
    - Metrics-Based Load Balancing for Envoy AI Gateway Inference Exte…
- 0.2 Release Planning (Erica)
    - Before or After EG 1.4
- Support MCP and A2A protocols (Yan)
- Blog Posts (Erica)
    - Reference Architectures - AIGW in Action how it is used (Dan to decide date)
        - AIGW + Kserve
    - Inference Extension blog post - more Envoy broad topic - End of April
    - Upstream auth in AIGW post - Mid May

Key Decisions & Takeaways

1. Welcome

- Welcomed Siva from Ideas2IT and KServe community as a new contributor.
- Quick introductions from all attendees helped align context across organizations.

2. LLM Metrics-Based Routing Using Inference Extension API (Dan/Siva)

Key Decisions:

- The team will not cache metrics inside Envoy or the load balancer due to freshness and performance needs.
- Metrics-based routing will be implemented via an external endpoint picker to preserve flexibility and extensibility.
- Reference implementation from Kubernetes Gateway API inference extension will be leveraged, with the potential to offer a Envoy AI Gateway-specific implementation later.

Takeaways:

- Endpoint picker will support multiple scoring filters (e.g., capacity, cost, KV cache).
- The picker may run as a sidecar or separate service with fast, frequent metric scraping.
- Adding this functionality improves latency and hardware efficiency significantly.

Next Steps:

- Dan and Siva to draft configuration/API for external endpoint picker by next meeting.
- Yan to finalize and submit pending PR for Envoy load balancing extension.

3. 0.2 Release Planning (Erica)

Key Decision:

- Delay 0.2 release until after Envoy Gateway 1.4 is released (in 2 weeks) to align versions.

Takeaway:

- No blockers to waiting. Users can still test features already merged on main.

4. Support MCP and A2A Protocols (Yan)

Key Decisions:

- The group supports exploring MCP support in AI Gateway, recognizing the value of a complete model-serving ecosystem.
- A2A (agent-to-agent) support acknowledged as a potential future need but not a priority now.
- MCP functionality would likely need a dedicated route type (e.g., `MCPGatewayRoute`), separate from inference routes.

Takeaways:

- MCP is relevant for serving context or tools to models, especially in self-hosted environments.
- This expands AI Gateway's applicability for advanced agent-based workloads.

Next Steps:

- Yan to create:
    - An Envoy issue for required data plane features to support MCP.
    - An AI Gateway issue to gauge community interest and guide config/resource design.

5. Blog Posts (Erica)

Key Decisions:

- Multiple blog posts scoped out to increase visibility and education:

    1. Reference Architectures (Dan to set date)
    2. AI Gateway + KServe (Dan/Siva)
    3. Inference Extension & Endpoint Picker (Erica + Yan + Takeshi; target: end of April)
    4. Upstream Auth in AI Gateway (Xiaolin + Aaron; target: mid-May)

Takeaways:

- Posts should balance real-world usage with generic architecture guidance.
- Erica will lead writing efforts and help structure collaboration.

Next Steps:

- Contributors to begin drafting according to agreed timelines.
- Siva / Dan to share link to existing KServe integration docs/presentation for reuse.

April 10, 2025

Status Past

Attendees

- Host: Dan Sun
- Erica Hughberg [Tetrate]
- Yan Avlasov [Google]
- John Landa [Tetrate]
- Ajay Nagar [Nutanix]

- Xiaolin Lin [Bloomberg]

Agenda

- Welcome [Host]
- Discuss 0.2 release [Suggested topic by Erica]
- Kubecon recap?
    - From envoy kiosk
        - no body based routing
        - Egress use cases, 1st anthropic
- Inference extension follow up
    - Added to https://gateway-api-inference-extension.sigs.k8s.io/implementations/
    - Current implementation is not bind to the HTTPRoute
- KServe/Envoy AI gateway integration
    - Manage LLM as a service https://kserve.github.io/website/latest/admin/ai-gateway_integration/#create-an-inferenceservice

---

❌ April 3, 2025

Status Skipped

Skipping because of KubeCon Europe

🗓️ March 27, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate)
- Melissa Salazar(Independent)
- Takeshi Yoneda(Tetrate)
- Yan Avlasov (Google)
- Andres Guedez (Google)

Agenda

- Welcome [Host]
- Inference Extension Implementation Update [Takeshi]
    - https://github.com/envoyproxy/ai-gateway/pull/493
- CLI Tool intro [Takeshi]
    - https://aigateway.envoyproxy.io/docs/cli/
- Azure Integration Update - Docs progress? [Erica]
- Local chat sample app [Erica]
- AOB

Notes

Shmuel Kallner - I will not be attending today's call. I can report that we will be contributing our improved vLLM Simulator to the Open Source. The question right now is to where. There are multiple projects that are interested in such a capability. We are looking into the possibility of contributing this code into a repo under the vLLM umbrella. A decision hasn't been made yet and apparently won't be made until after KubeCon Europe.

Key Decisions & Takeaways

1. Inference Extension Implementation Update [Takeshi]

- **Implementation Status**:
    - The implementation is complete and ready for review; it's a foundational pull request to set the groundwork.
- **Next Steps**:
    - Takeshi will address current PR comments and aim to merge the foundational PR soon.
    - Clarify handling of ingress and egress use cases separately.
    - Discuss further in London, including intelligent routing usecase for Bloomberg (load balancing based on metrics rather than simple Round Robin).

2. CLI Tool Intro [Takeshi]

- **Overview**:
    - A new experimental CLI was created to debug and visualize Envoy AI Gateway configurations.
    - Similar to the existing Envoy Gateway CLI, it helps users understand how to translate Gateway resources into Envoy configuration.
- **Use cases**:
    - Primarily for debugging, but also useful for local development/testing of AI applications.
- **Next Steps**:
    - The community is encouraged to test and provide feedback.

3. Azure Integration Update - Docs Progress [Erica/Yao]

- **Current Status**:
    - Documentation is still in progress. Bloomberg team is working on it actively.

- **Next Steps**:
    - Erica offered to help draft documentation if the Bloomberg team can provide rough notes or preliminary drafts.
    - Yao will follow up internally to provide initial content.

4. Local Chat Sample App [Erica]

- **Demo Overview**:
    - Erica built a small chat demo app using React, demonstrating interaction with multiple AI backends (e.g., OpenAI, local Ollama, Mistral) via Envoy AI Gateway.
- **Purpose**:
    - Intended as a fun and interactive way to showcase AI Gateway capabilities locally.
- **Next Steps**:
    - Erica plans to share the source code publicly on GitHub for others to experiment or use as a teaching/demo resource.

5. Any Other Business

- **Community Collaboration**:
    - A London meeting (at KubeCon) is planned to explore specific use cases, notably Bloomberg's cloud-provider fallback and intelligent routing scenarios.
    - Wednesday was tentatively agreed upon as the best day for in-person collaboration.

- **Envoy Proxy PR Update [Yan]**:
    - Yan's Envoy Proxy PR is still open due to ongoing technical discussions, but there are no fundamental blockers. We aim to resolve and merge it soon.

- **Housekeeping & Docs [David Xia]**:

    - David previously suggested Markdown linting improvements for docs. No recent action has been taken, but it's recognized as a good housekeeping task to revisit soon.
    - David suggested maintaining a gallery of sample/demo apps on the project website to engage the community.

Action Items

- Takeshi to finalize PR reviews for Inference Extension implementation.
- Erica to coordinate London meetup.
- Yao to provide rough notes for Azure integration docs.
- Erica to publish chat sample app in her GitHub.
  Continue discussions on advanced routing/fallback use cases in upcoming London meeting.

🗓️ March 20, 2025

Status Past

Attendees

- Host: Erica Hughberg (Tetrate) GH: missberg
- Takeshi Yoneda (Tetrate) GH: mathetake
- Xiaolin Lin (Bloomberg) GH: xiaolin593
- Ajay Nagar (Nutanix) GH: nagar-ajay
- Dan Sun (Bloomberg): GH: yuzisun
- Aaron Choo (Bloomberg) GH: aabchoo
- Gavrish Prabhu (Nutanix) GH: gavrissh

Agenda

- Welcome [Host]
- Review action items from last meeting [Erica]
- Review Inference Extension Proposal Implementation [Takeshi]
    - https://github.com/envo`yproxy/ai-gateway/pull/492/files?short_path=8fabb64#diff-8fabb64fdce6b69f5aae92c2071d924484accf18a08631c49f57f1ad15257ca1
    - https://github.com/envoyproxy/envoy/pull/38757
    - PoC (general direction in envoyproxy/ai-gateway repo): https://github.com/envoyproxy/ai-gateway/pull/493
    - Model Server Protocol Spec: [PUBLIC] Model Server Protocol for LLM Instance Gateway
- Update on recently merged PRs [Erica]
- Envoy ai gateway hybrid mode [dsun]
    - Sample diagram: https://docs.google.com/document/d/10e1sfsF-3G3Du5nBHGmLjXw5GVMqqCvFDqp_O65B0_w/edit?tab=t.4dc3y8mwy92u
- What is the envoy-ai-gateway release cadence?[Xiaolin]
    - https://github.com/envoyproxy/ai-gateway/blob/main/RELEASES.md

Key Decisions & Takeaways

1. Welcome [Host]

- General introductions completed.
- Reminded attendees to add their names to the attendee list.

2. Review Action Items from Last Meeting [Erica, Takeshi, Yan, Aaron, Dan]

- The OpenAI schema extension proposal has not started yet.
- Multi-level failover discussions ongoing; dependency identified on an upstream PR.
- Inference extension proposal discussed further in agenda item 4.

- Load balancing extension PR open and under active discussion; consensus on its usefulness achieved, technical details remain.
- Encouragement to promote EnvoyCon Europe actively.

3. Update on Recently Merged PRs [Erica, Xiaolin]

- Azure OpenAI integration (upstream authorization) merged.
- Identified a limit issue when hosting 16+ models; an issue was created, and a workaround suggested.

4. Envoy AI Gateway Hybrid Mode [Dan, Erica, Andres, Yan, Takeshi]

- Agreement on implementing inference extension initially as part of AI Service Backend (secondary citizen).
- Future native support (HTTP route) would likely move to Envoy Gateway itself based on community feedback and usage.
- Hybrid mode essential for advanced load balancing and integration with KServe.
- Discussions around the metrics standardization ongoing; current initial implementation scrapes Prometheus metrics directly.
- Decision: Separate the implementation of pluggable infrastructure and the specific algorithms for endpoint selection.

5. Envoy AI Gateway Release Cadence [Xiaolin, Erica, Dan, Takeshi]

- Established release cadence aligns with Envoy Proxy quarterly releases.
- Significant feature enhancements could prompt additional releases between quarterly cycles.
- Patch releases for critical bug fixes as needed.
- Azure integration likely to be included in an upcoming release, pending successful testing and documentation.

Other Business

- IBM's interest in AI Gateway as both frontend and in-cluster proxy noted.
- IBM team has extended the upstream test server with enhanced metrics, potential PR contribution suggested.

Action Items

- **Takeshi - Inference Extension Implementation**
    - Continue building out the inference extension PR. Post updates or questions as needed.
- **Xiaolin / Dan - Azure Upstream Auth Documentation**
    - Add documentation for the new Azure integration in Envoy AI Gateway. This is necessary before cutting a release that includes it.
- **Shmuel - Draft PR for Extended Fake LLM Server PR**
    - Open a (draft) PR showing how they extended the test server with added metrics and features so others can evaluate merging it upstream.
- **Dan - KServe + Envoy AI Gateway Integration**
    - Continue work on KServe PR to auto-generate AI Gateway resources; share progress at upcoming meetings.
- **Maintainers - 0.2 Release Planning**
    - Plan a release cut for 0.2 with key new feature being Azure Integration (Translation & Auth)

🗓️ March 13, 2025

Status Past

Attendees

- Host: Erica (Tetrate) GH: missberg
- Melissa (Independent)
- Dan Sun (bloomberg)
- Gavrish Prabhu (Nutanix) GH: gavrissh

- Huamin Chen (Red Hat)
- Yan Avlasov (Google)
- Ajay Nagar (Nutanix) GH: nagar-ajay
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- Ignasi Barrera (Tetrate) GH: nacx
- Andres Guedez (Google) AndresGuedez
- Takeshi Yoneda (Tetrate)
- Enock Kasaadha (Independent)

Agenda

- Welcome [Host]
- OpenAI schema extension (aaron/dsun)
    - vLLM supports a handful of guided generations that would be great to have as part of the ai-gateway
- Backend failover priority for more than 2 levels (dsun/yao)
    - https://github.com/envoyproxy/gateway/issues/5442
    - OSS Endpoint Picker: https://github.com/kubernetes-sigs/gateway-api-inference-extension/
        - Reference implementation: see pkg/epp/...
- Store metrics like TTFT/TPOT in metadata for downstream consumption
- Reference Architecture Diagrams (Erica)
    - Showcase how Envoy AI Gateway fits into a system
    - KServe integration doc https://github.com/kserve/website/pull/454
- Inference Extension to Gateway API (Erica)
    - Takeshi, Yan, collab on design doc?

Key Decisions & Takeaways

1. **OpenAI Schema Extension**

- The proposal is to extend the AI Gateway schema to include additional guided generation features (e.g., guided JSON, regex, choice, grammar).
- The approach will maintain backward compatibility with the OpenAI schema while supporting multiple providers.
- A formal OpenAPI spec for AI Gateway's supported schema will be created for clarity and documentation.
- A design document will be initiated to outline the approach.

2. **Backend Failover Priority for More than Two Levels**
   - Current Envoy Gateway supports only two failover levels (active-passive mode), which is insufficient for multi-region failover use cases.
   - The goal is to support failover across multiple regions and between provisioned vs. on-demand endpoints.
   - Envoy Gateway needs to expose priority fields for failover levels.
   - Further discussion on Slack will determine the best approach, potentially leveraging existing load-balancing strategies or extending capabilities.
3. **Reference Architecture Diagrams**
   - The AI Gateway team will create reference architecture diagrams for:
     - Generic cloud-native usage.
     - Cloud-provider-specific setups (GCP, AWS, etc.).
     - Integration with other popular industry solutions (e.g., KServe).
   - Collaboration with the CNCF AI Working Group to create a **white paper** on how AI Gateway fits into GenAI infrastructure.
   - Contributors interested in helping with reference architectures should reach out.
4. **Inference Extension to Gateway API**
   - Takeshi will take the lead on implementation and collaborate with Yan and others from Google.
   - A design document will be created to align with the Kubernetes inference API proposal.
   - Yan will open-source a new load-balancing extension to support this.
   - The team will coordinate via Slack and GitHub issues to track progress.

5. **AI Gateway at KubeCon**
   The team plans to **meet up on Monday** before the conference officially starts.
   - Takeshi, Dan, and Erica are running an **Envoy AI Gateway workshop** on Monday afternoon.
   - The inference API and related work should ideally have a preview/demo at the workshop.

Action Items

[Aaron, Dan] Start a design doc for OpenAI Schema Extension proposal.
[Yao, Dan] Investigate best approaches for multi-level failover and share findings in Slack.
[Erica, Huamin] Kick off collaboration with CNCF AI Working Group on a **white paper** for AI Gateway in GenAI infrastructure.
[Takeshi] Lead the implementation of the **Inference Extension to Gateway API** and raise a design document.
[Yan] Open-source the **load balancing extension** and coordinate with Takeshi for integration.
[All interested contributors] Reach out to Erica to help with **reference architectures** for AI Gateway documentation.
[Everyone] Help promote **EnvoyCon Europe** and **KubeCon** sessions on social media.

🗓️ March 6, 2025

Status Past

Attendees

- Host: Erica (Tetrate) GH: missberg

- Huamin Chen (Red Hat), GH: rootfs
- Xiaolin Lin (Bloomberg), GH: xiaolin593
- Shmuel Kallner (IBM), GH: shmuelk
- Takeshi Yoneda (Tetrate), mathetake
- Dan Sun (Bloomberg), yuzisun
- Yao Weng (Bloomberg), yaoweng04
- Aaron Choo (Bloomberg), aabchoo

Agenda

- Welcome [Host]
- KubeCon Europe - London - [Erica]
    - Workshop on 31st March - Takeshi/Erica/Dan/Yan(TBD)
    - In person meeting?
        - Who will be at KubeCon?
    - All talks related to envoy ai gateway?
        -
- Provider Integration Poll
    - https://github.com/envoyproxy/ai-gateway/discussions/386
- Feature Progress Updates
    - Initial metrics:
        - https://opentelemetry.io/docs/specs/semconv/attributes-registry/gen-ai/
        - https://github.com/envoyproxy/ai-gateway/pull/459
    - Azure stuff
        - https://github.com/envoyproxy/ai-gateway/pull/424
        - https://github.com/envoyproxy/ai-gateway/pull/445
- Look at: https://github.com/NVIDIA-AI-Blueprints/llm-router
    - Can this be complimentary?
    - [Huamin] Per a discussion in last meeting, this could happen in an extproc chain
- *Please add agenda points*

🗓️ Feb 27, 2025

Status Past

Attendees

- Host: Erica (Tetrate) GH: missberg
- Aaron Choo (Bloomberg)
- Huamin Chen (Red Hat)
- Eric Mariasis
- Takeshi Yoneda (Tetrate)
- Xiaolin Lin (Bloomberg)
- Yan Avlasov (Google)
- Yao Weng (Bloomberg)
- Dan Sun (Bloomberg)
- Shmuel Kallner (IBM)
- Arshardh Ifthikar (WSO2)

Agenda

- Welcome [Host]
- Release celebration! 🎉 - Erica
- What's next? - Erica
    - https://github.com/envoyproxy/ai-gateway/discussions/386
    - https://github.com/envoyproxy/ai-gateway/issues/423
    - Resilience features
- (Huamin) Update on https://github.com/envoyproxy/ai-gateway/pull/433
    - An initial gRPC proto is defined
    - Extproc with semantic cache using the gRPC proto is here. Note, this PoC only supports non-streaming OpenAI schema based requests.
    - The demo can be found here. There are three panes in the terminal, the top one is the test client using openai schema. The middle pane is the semantic processor with semantic cache (using FAISS). The bottom

pane is the extproc. https://github.com/envoyproxy/ai-gateway/commit/a622ffcd366915b774115f1d4ab72b83a70cb653The tests were run twice: the first time without cache took about 7s, the 2nd run with cache hit took 0.5s
    - Comments and suggestions are welcome
- (Yan) proposal for external endpoint picking protocol: https://github.com/kubernetes-sigs/gateway-api-inference-extension/tree/main/docs/proposals/004-endpoint-picker-protocol

---

❌ Feb 20, 2025

Status Skipped

Attendees

- Host: [TBA]

Agenda

- Welcome [Host]
-

🗓️ Feb 13, 2025

Status Past

Attendees

- Erica Hughberg (Tetrate)
- Huamin Chen (Red Hat)
- Takeshi Yoneda (Tetrate)
- Yan Avlasov (Google)
- Eric Mariasis
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- David Xia (Spotify)
- Dan Sun (Bloomberg)

Agenda

- Welcome [Erica]
- 0.1 Release Status [Takeshi]
    - Site DNS (aigateway.envoyproxy.io)
    - OIDC: https://github.com/envoyproxy/ai-gateway/pull/306
    - More docs!
- Target repository for APIs
    - Who is doing work?
- Inference Gateway API - how to get there? [Erica]
    - v0.1.0 alpha release: https://github.com/kubernetes-sigs/gateway-api-inference-extension/releases/tag/v0.1.0
- Quick demo of Envoy AI Gateway with vLLM services on EKS [Huamin]
    - https://github.com/rootfs/ansible-receipes/blob/main/curl-vllm.sh

- Discuss how we can expose HTTP route features such as timeout, failover [dsun]
    - https://github.com/envoyproxy/ai-gateway/issues/34
- Prompt Caching Mechanism Discussion

Meeting Summary

Release Status (Takeshi)

- **OIDC PR (#306)**: Near completion (90%). Aaron is working on the final test coverage.
- **Site DNS (aigateway.envoyproxy.io)**: CNCF action required for setup. Follow-up needed.
- **Documentation Improvements**: Additional docs were added; more are in progress.
- **Release Timeline**: Targeting completion in the next two weeks.

Action Items:

- [Erica] Follow up with CNCF on DNS setup.
- [Takeshi/Aaron] Finalize OIDC test coverage and implementation.
- [Team] Continue adding documentation.

Inference Gateway API - Implementation Discussion (Erica, Yan, Andres)

- **v0.1.0 Alpha Release**: Kubernetes SIG Gateway API Inference extension is formally released.
- **Target Repository & Implementation**:
    - Short-term: Implement in Envoy AI Gateway for fast iteration.
    - Long-term: Migrate into Envoy Gateway for broader adoption.
    - There is no active implementation work yet.
- **Concerns:**
    - Where features should be placed (AI Gateway vs. Envoy Gateway).

    - Potential duplication of work.

Action Items:

- [Yan] Create an issue detailing the implementation plan and API considerations.
- [Andres/Erica] Validate API alignment with Envoy Gateway maintainers.
- [Team] Discuss further on Slack to coordinate contributions.

Demo: Envoy AI Gateway with vLLM Services on EKS (Huamin)

- **Setup:**
    - Uses Ansible to deploy EKS with small vLLM models.
    - Demonstrates traffic routing and inference handling.
    - It helps visualize resource requirements.
- **Proposal:**
    - Share ‘recipes’ for easy setup and comparison of configurations.
    - Potential inclusion in the documentation.

Action Items:

- [Huamin] Refine and share deployment recipes.
- [Erica] Explore the integration of deployment recipes into documentation.

HTTP Route Features (Timeout, Failover) (Dan, Takeshi)

- **Failover Mechanism:**
    - Goal: Failover between AI model providers (AWS Bedrock, OpenAI, etc.).
    - Initial approach: Use the existing Envoy Gateway failover policy.
    - Issues to solve:
        - Model-awareness in failover scenarios.
        - Handling authorization header transformations for different providers.
        - Need for active/passive backend selection.

Action Items:

- [Dan] Test existing failover mechanism in AI Gateway and report gaps.
- [Takeshi] Investigate backend traffic policy configuration improvements.
- [Team] Explore enhancements for model-aware failover.

Prompt Caching Mechanism Discussion (Dan, Yao)

- **Goal:** Reduce AI inference costs by caching processed prompts.
- **Challenge:**
  - Different providers (AWS Bedrock, Anthropic, etc.) have distinct caching mechanisms.
  - OpenAI does not support native caching.
  - Standardizing API for caching across providers.
- **Next Steps:**
  - Research industry efforts on standardized caching APIs.
  - Collaborate with NVIDIA and others on API unification.

Action Items:

- [Dan/Yao] Investigate existing caching standards and share findings.
- [Dan] Define a design proposal for caching API in AI Gateway.

🗓️ Feb 6, 2025

Attendees

- Erica Hughberg (Tetrate)
- Takeshi Yoneda (Tetrate)
- Huamin Chen (Red Hat)
- Darius Emrani (Scorecard.io)
- Melissa Salazar (Individual)
- Arshardh Ifthikar (WSO2)

- Dan Sun (Bloomberg)
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- Eric Mariasis

Agenda

- Welcome [Erica]
- 0.1 Release Status [Takeshi]
    - Blocker: AWS OIDC impl (Aaron)
    - Bunch of pending docs
    - Any other blockers?
- [Huamin]Prometheus and Grafana
    - https://github.com/envoyproxy/ai-gateway/commit/15541e0aacc66251a80d31bbd31dcf7226809185
    - Added backend, request, and token tracking. Any more metrics?
        - First token latency under streaming mode and time to first token (TTFT) and inter-token latency (ITL)
    - Sample Grafana dashboard https://github.com/rootfs/ai-gateway/blob/dev/docs/grafana-dashboard.json, with a screenshot
- Contributor guidance - Erica
- [Eric Mariasis] API Reference docs, discuss here or with Erica 1-1 about desired look and feel
- *[Darius Emrani] Model management reference/capabilities lookup*
    - *Mode: chat vs completion*
    - *Supports: Vision, audio, tool use etc*
    - *Max Input/output tokens*

Meeting Summary

🗓️ Jan 30, 2025

Attendees

- Huamin Chen (Red Hat) with coding cat Mellow
- Andres Guedez (Google)
- Takeshi Yoneda (Tetrate)
- Erica Hughberg (Tetrate)
- Noy Itzikowitz (Independent)
- Sébastien Han (Red Hat)
- Ignasi Barrera (Tetrate)
- Eric Mariasis
- Yao Weng (Bloomberg)
- Dan Sun (Bloomberg)
- Aaron Choo (Bloomberg)
- Guy Stone (Spotify)
- David Xia (Spotify)
- Ron Haberman (Red Hat)
- Arshardh Ifthikar (WSO2)
- Ikenna Chifo (Spotify)
- Melissa Salazar (Independent)
- Arko Dasgupta (Tetrate)
- Rob Scott (Google)
- Ron Haberman (Red Hat)

Agenda

- Welcome [Erica]
- 0.1 Release Status [Takeshi]

    - Milestone: https://github.com/envoyproxy/ai-gateway/milestone/1
    - Docs Site
        - ("Architecture/Design" doc stuff by Dan)
        - Maybe a Ratelimit tutorial?
        - DeepSeek example ?;)
    - Features
        - AWS+OIDC
    - Bugs
        - None (Hopefully)
    - When are we cutting RC?
    - Staged site: https://envoy-ai-gateway.netlify.app
        - aigateway.envoyproxy.io has been requested, waiting on it to be set up
- [Huamin] Semantic processor PoC and feedback
    - Auto LLM Model Selection and Routing PoC
    - Semantic processor service interface discussion
- [Andres] Envoy Extensions for Inference in K8s: Gateway Inference APIs
- *Add your items to the agenda*

Meeting Summary

0.1 Release Status [Takeshi]

- The milestone for the 0.1 release is mostly complete, with the remaining work focused on documentation and final touches.
- AWS OIDC Integration:
    - Aaron is testing the integration.
    - Plans to raise a draft PR by the end of the day.
    - Further testing is required before merging.
- Release Candidate (RC) Plan:
    - The target is to cut an RC by Monday
    - If issues arise, they should be raised ASAP.
    - A draft PR should be created to allow for early reviews.

- Documentation & Site Status:
    - Docs staged at https://envoy-ai-gateway.netlify.app.
    - aigateway.envoyproxy.io domain requested; awaiting CNCF setup.
    - More documentation and guides will be added (Melissa and Erica will collaborate on this).

[Huamin] Semantic Processor PoC and Feedback

- Auto LLM Model Selection and Routing PoC
    - Proposal to integrate semantic processing for intelligent LLM model selection.
    - The idea is to avoid hardcoding models and dynamically select the best model based on query complexity, latency, and coherence.
    - It uses a semantic processor (Python-based NLP service) to evaluate queries and route them to the most suitable model.
    - Demo showed:
        - Simple questions routed to GPT-3.5.
        - Complex queries are sent to a higher-capability model like GPT-4.
- Discussion on API Design for Semantic Processing
    - Need a straightforward API for adding external processors.
    - Potential use cases beyond model selection include caching and safety filtering (e.g., PII prevention).
    - Agreement on adding an experimental feature set allows users to try these capabilities before stabilizing the API.

[Andres] Envoy Extensions for Inference in Kubernetes

- Inference-specific APIs in Kubernetes Gateway
    - Introduces Inference Pool and Inference Model resources to manage self-hosted models.
    - Targets users hosting models on Kubernetes and optimizes tail latency and fairness in serving models.
-
-

- Implementation Considerations
    - Uses external processing (ExtProc) for endpoint selection.
    - Relies on Orca protocol for inline metric reporting.
    - Needs access to both HTTP headers and body payloads to parse AI-specific request details.
- Integration into Envoy AI Gateway
    - Andres and Rob (Google) are willing to help contribute to this integration.
    - Needs further discussion on:
        - Helm chart integration.
        - CRD installation approach.
        - Feedback mechanisms for real users.

Agreed Action Items

1. **0.1 Release Finalization**
    - Aaron: Raise PR for AWS OIDC integration.
    - Takeshi & team: Ensure all required docs are updated.
    - RC target: Cut by Monday.
2. **Experimental Feature Support for Semantic Processing**
    - Huamin: Create a GitHub issue to discuss the API approach for semantic processors.
    - Team: Establish a way to enable experimental external processors in the project.
        i. https://github.com/envoyproxy/ai-gateway/issues/233
3. **Inference API Integration with Envoy AI Gateway**
    - Andres, Rob: Support integration with AI Gateway. Share implementation information.
    - Arko, Erica, Takeshi, Dan: Discuss Helm chart and installation details.
    - Users and Contributors: Provide feedback on whether this feature fits their use case.
4. **Community Feedback & Contributions**
    - Spotify (David & team): Evaluate 0.1 release and provide feedback.
    - Further discussions on ensuring smooth API integration.

🗓️ Jan 23, 2025

Attendees

- Erica Hughberg (Tetrate)
- Takeshi Yoneda (Tetrate)
- Huamin Chen (Red Hat)
- Harvey Tuch (Google)
- Krishan Wjesena(Wso2)
- Melissa Salazar(Independent)
- Yan Avlasov (Google)
- Yao Weng (Bloomberg)
- Dan Sun(Bloomberg)

Agenda

- Welcome [Erica]
- Status of 0.1 Release Progress [Takeshi]
    - MUST:
        - PR waiting for the review by Dan(Bloomberg)
            - https://github.com/envoyproxy/ai-gateway/pull/151
            - https://github.com/envoyproxy/ai-gateway/pull/153
            - https://github.com/envoyproxy/ai-gateway/pull/156
            - https://github.com/envoyproxy/ai-gateway/pull/158
        - Documentation site `erica.hughberg@tetrate.io`
        - Rate limit end to end tests `takeshi@tetrate.io`
        - Auth API (OIDC?) impl: Aaron (Bloomberg)
        - Polishing the router code: Yao(Bloomberg)
            - https://github.com/envoyproxy/ai-gateway/issues/53

    - NICE TO HAVE:
        - https://github.com/envoyproxy/ai-gateway/issues/128
        - https://github.com/envoyproxy/ai-gateway/issues/134(Siva)
        -
- Survey Plans [Erica]
- Python Service proposal [Huamin] Envoy AI Gateway Extension Proposal
- Inference Instance Gateway Features Needed Discussion [Erica]
- *Add your items to the agenda*

🗓️ Jan 16, 2025

Attendees

- Erica Hughberg (Tetrate)
- Takeshi (Tetrate)
- Krishan (Wso2)
- Hussain (Google)
- Arshardh (WSO2)
- Mona Borham(Independent)

Agenda

- Current status of project
    - 90% Done
        - Need Backend Security Feature
            - PR raised by aaron
        - Control plane API?

            - Rate Limit Request Cost: https://github.com/envoyproxy/ai-gateway/pull/103
        - Minimum Doc
            - Concepts, Getting Started
        - Erica to speak with Phlax on Website publishing
        - Extensibility of translator
            - How to create translators?
                - Maybe provide a way to add for users to add their own transformers
- Dynamic Modules: https://github.com/envoyproxy/ai-gateway/issues/90
    - What is it?
    - How can it benefit AI GW project?
- Release cycle conversation from last week
    - https://github.com/envoyproxy/ai-gateway/pull/99
- Roadmapping/Feature needs
    - How to ensure we are building valuable features?

🗓️ Jan 9, 2025

Attendees

- Erica Hughberg (Tetrate)
- Dan Sun (Bloomberg)
- Aaron Choo (Bloomberg)
- Arshardh Ifthikar (WSO2)
- Krishan Wijesena (WSO2)
- Yao Weng (Bloomberg)

Agenda

- Rate limiting
    - How to specify and configure the cost calculator such as input_token * weight + output_token
- EG routing issue for the routing calculation (Yao)
    - Tried two listeners but the host header is overwritten by the first listener when connection is established
- AI gateway release cycle
    - On top of regular official releases, can we do daily or weekly releases with named date tags ?
        - Use commit number as tag
        - Can be tricky if the feature has upstream dependency for envoy or gateway
- Running AI Gateway along Istio?

Note from Takeshi

- The v0.1.0 implementation update:
    - ExtProc implementation is done (Thanks for the help Dan!) for the currently merged API + RateLimit
    - Controller implementation is almost done with pending work-in-progress env tests
    - Remaining tasks from implementation pov
        - Finish controller env tests
            - https://github.com/envoyproxy/ai-gateway/pull/75 is the first one and the subsequent PRs will complete it
        - BackendSecurityPolicy implementation: assigned to Aaron
        - Real end to end test with EG under k8s envs
            - I will take care of this next week.
        - Rate Limit API
            - Even though it will be supported by EG API to achieve the “usage based rate limit”, we still need to provide a way to configure the metadata {namespace and key} in AI

Gateway CRD, which say we call BackendTrafficPolicy
https://github.com/envoyproxy/gateway/pull/4957
        - The design will take into account the “cost calculator” as I see in the agenda.
    - I need help from someone on
        - https://github.com/envoyproxy/ai-gateway/issues/53
        - https://github.com/envoyproxy/ai-gateway/issues/70
- I am looking to have a way to enable users to insert custom logics into extproc - basically it will allow them to build their own custom extproc image with some inserted custom logics inside the extproc implementation. So the former is to expose the main function of extproc and the latter is to have a public package with some Go interfaces to insert, say, a custom Router, custom cost calculator, and custom transformer (https://github.com/envoyproxy/ai-gateway/issues/37) etc. This is something Tetrate wants as well as many people hopefully. I will raise the issue tomorrow morning (in tokyo).
    - I envision this will be the extension point also used to implement semantics caching, prompt guard etc.

🗓️ Jan 2, 2025

Attendees

- Erica Hughberg (Tetrate)
- Dan Sun (Bloomberg)
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- Arshardh Ifthikar (WSO2)

Agenda

- [Erica] Welcome
- 0.1 release tracking
    - Had to move AWS Request signing into ext proc
    - Prioritize Backend Security Policy API
    - Traffic Policy - check Envoy Gateway PR see if it is in
    - Bedrock Conversion
        - AWS SDK
        - Need more comprehensive testing
    - Content
        - Docs for installing and trying out
        - Contribution guide
        - [non-blocking] CNCF Blog for 0.1 release
    - Moving Rate limiting - check with Takeshi what is left
- KServe and EG and Istio
- Feature Request
    - Integrate with Google Gemini 2.0
        - Seems to have OpenAI compatible interface
        - Needs to figure out

🗓️ Dec 26, 2024

Attendees

- Erica Hughberg (Tetrate)

Note

Only Erica joined, assuming everyone is having a great holiday break :)
Dropped off meeting at 11:17 eastern.

Please message in Envoy Slack #envoy-ai-gateway channel for async communication.

See you all next week 🎉
Happy new year!

December 19th

Attendees

- Erica Hughberg (Tetrate)
- Dan Sun (Bloomberg)
- Aaron Choo (Bloomberg)
- Sunil Ravipati(Yottasecure)
- Krishan Wijesena(Wso2)
- Hussain Chinoy (Google)

Agenda

- [Erica] Welcome
- [Erica] Review Action items from last meeting
    - Dan is going to publish design doc
    - https://envoy-ai-gateway-site-demo.netlify.app/ - website
    - Consider creating a BackendSecurityPolicy - that can be applied to a Backend Resource
        - Created an issue on envoy gateway
- [Erica] Discussed Envoy Proxy new features
    - Dynamic Modules - might enable us to build features into AI GW
- [Erica] Discussion - Content Security with a Prompt Guard

    - https://huggingface.co/meta-llama/Prompt-Guard-86M
    - Discussed the urgency of this, believe that for internal to model traffic doesn't have this as a hard requirement, however when app teams expose GenAI functionality in apps to end users, this become critical
- [Krishan] Semantic Caching
    - Will share design when it's more ready, Erica suggest that Krishan connect with Takeshi who might have ideas
- [Krishan] Offering to help with Envoy AI Gateway website
    - Erica is working on getting the initial setup merged
    - Discussed URL - likely be something like envoyproxy.io/ai-gateway and to set up a redirect from envoyaigateway.io

---

🗓️ December 12th

Attendees

- Erica Hughberg (Tetrate)
- Dan Sun (Bloomberg)
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)
- Krishan Wijesena (wso2)
- Faisal Afzal (HashiCorp)
- Sanjeewa Malalgoda(WSO2)

Agenda

- [Erica] Welcome
- [Erica] Review Action items from last meeting
    - Recap work to get ready for v0.1
    - Discussion: https://github.com/envoyproxy/ai-gateway/issues/42

    - API Definition to be done by end of year
        - Add to docs when ready
    - LLMBackendTrafficPolicy
- [All] Review GitHub
    - Issues
        - https://github.com/envoyproxy/ai-gateway/issues/31
        - https://github.com/envoyproxy/ai-gateway/issues/30
    - PRs
        - Site PR - has some items from Takeshi to address
        - LLMBackendTrafficPolicy - Yao to review comments and address
- *Add items to discuss*

🗓️ December 5th - Roadmap Focus

Attendees

- Erica Hughberg (Tetrate)
- Takeshi Yoneda (Tetrate)
- Krishan Wijesena (WSO2)
- Dan Sun (Bloomberg)
- Yao Weng (Bloomberg)
- Aaron Choo (Bloomberg)

Agenda

- [Erica] Welcome
- [Erica] Review Action items from last meeting
- [All] Roadmap items
    - Assign lead ownership of items
        - Semantic Caching - Krishan

        - Prompt Templating - Krishan
        - Metrics Exporting - Yao
        - Usage Limiting based on Tokens - Takeshi
        - Token per second (TPS) - For monitoring - Yao
        - Failover Logic between Providers and Models - Dan/Yao
        - Add LLM Providers - Erica to create issues for Providers to be added
            - GCP
- Road to 0.1.0
    - MVP of POC
    - Runnable on K8s
        - Make sure we have docs so people can learn how to run it
    - Erica to create a Milestone in GitHub
    - Target date: Jan 31, 2025
- AOB
    - Website

🗓️ Nov 27th - Feature Deepdive

Note: Looking to have this on a Wednesday due to the US Thanksgiving break
Having this meeting on Google Meet due to issue with moving Zoom meeting

To join the video meeting, click this link: https://meet.google.com/dyo-wcvd-pwp
Otherwise, to join by phone, dial +1 385-645-7711 and enter this PIN: 282 327 100#
To view more phone numbers, click this link:
https://tel.meet/dyo-wcvd-pwp?hs=5

Attendees

- Erica Hughberg (Tetrate)
- Takeshi (Tetrate)
- Arshardh Ifthikar (WSO2)
- Aaron Choo (Bloomberg)
- Yao Weng (Bloomberg)
- Sanjeewa Malalgoda (WSO2)
- Krishan Wijesena (WSO2)

Agenda

- [Erica] Welcome
- [Erica] Review Action items from last meeting
- Feature Deepdive
    - [Takeshi] POC Features Showcase
    - [All] New Features/Opportunities
        - Semantic Caching
            - Related topic: Example of Not a semantic cache, a KV cache https://github.com/LMCache/LMCache
        - Prompt Enricher
            - Templating - standardize, structure prompts for consistent input to the model
            - Decorating - enhance or modify a prompt by adding contextual information
        - Prompt Guard
            - Semantic - Leverage LLM
            - Keyword based prompt protection
        - Intelligent failover between providers [Important]
            - Important Feature to ensure resilience
            - Help to manage costs as well(switch between expensive and cheap models)
        - Metrics exporting - there is LLM for OTel
            - Raise issue for this in repo

        - Rate/Usage limiting
            - Token per second (TPS) - (LLM Specific) - Both client to gateway, and gateway to upstream target
            - Usage limiting - Based on business usage
            - Backend Protection - Protect backend usage based on utilization or cost. Maybe leverage TPS
        - Supporting Providers for AuthZ and Transformation
            - OpenAI (pass open ai key)
            - Cohere
            - Azure (pass access key) - has sso
            - Anthropic
            - Mistral/Llama2
            - Vertex/Gemini (api token)
            - AWS Bedrock (need to provide access/secret key)
- [Dan] POC Code Merge

Meeting Summary

Key Discussion Points

1. **Introductions and Collaboration Goals:**
    - New attendees introduced themselves, including their backgrounds in API gateways, AI gateways, and related technologies.
    - Emphasis on fostering collaboration to address GenAI traffic handling challenges.
    - Focus on establishing a foundation for open-source collaboration on AI gateways.
2. **Review of Action Items from the Last Meeting:**
    - Goals document finalized and merged into the repository with a "living" approach for future evolution.
    - A glossary to standardize terminology is being developed for consistency in communication.

3. **Proof of Concept (PoC) Features:**
    - Takeshi highlighted three key PoC features:
        - Transformation: Translating OpenAI to AWS schemas or others.
        - Token-Based Rate Limiting: Controlling usage by tokens consumed in request/response cycles.
        - Upstream Authorization: Simplified authentication and authorization for various providers.
    - Agreement that these features address critical industry needs.
4. **Feature Prioritization and Roadmap:**
    - Semantic caching was recognized as useful for latency and cost but complex to implement.
    - Prompt enrichment (templating and decorating) and semantic prompt guards were considered "nice to have" but not critical for initial phases.
    - Intelligent failover between providers and metrics exporting were deemed high-priority features.
    - Discussions about rate limiting, including token-per-second policies, emphasized the need for dynamic and backend-protective mechanisms.
5. **Unified API Specification:**
    - Importance of creating or aligning with a unified API specification for GenAI traffic.
    - Mention of ongoing efforts in the AI community, including potential hosting of a spec under the LF AI & Data Foundation.
6. **Use Case Insights:**
    - Shared experiences with token-based rate limiting, semantic caching, and other features.
    - Importance of features like burst control and token-per-second metrics to manage resource usage dynamically.
7. **Metrics Exporting and Cost Control:**
    - Need for standardized metrics to monitor and manage usage.
    - Mention of existing discussions in the OpenTelemetry community for metric standardization.

Decisions

1. **Feature Focus for Immediate Development:**
    - Prioritize intelligent failover, upstream/backend protection, and upstream authorization features.
    - Semantic caching and prompt-related features to be addressed in later phases.
2. **Unified API Alignment:**
    - Move towards implementing a unified API spec once it is published, leveraging community standards.
3. **Collaborative Issue Management:**
    - Use GitHub issues for async discussions and prioritization of features and control plane configurations.

Action Items

- [ ] **Create GitHub Issues** **Erica Hughberg**
    - [ ] Add issues for prioritized features, including semantic caching, intelligent failover, token-per-second policies, and metrics exporting.
    - [ ] Raise an issue for control plane API configurations.
- [x] ~~**Connect with Unified API Efforts** **dansun1981@gmail.com**~~
    - [x] ~~Initiate collaboration threads with groups working on the unified API specification.~~
- [ ] **Glossary Development** **Erica Hughberg**
    - [ ] Finalize and integrate the glossary into the repository for better accessibility.
- [x] ~~**PoC Polishing:** **Takeshi Yoneda**~~
    - [x] ~~Review and refine PoC features before merging~~ **~~Takeshi Yoneda~~** **~~/ ALL~~**
    - [x] ~~Create design doc for Control Plane API~~ **~~Takeshi Yoneda~~**
- [ ] **Follow Up on Metrics Standardization** **yweng14@bloomberg.net**
    - [ ] Explore alignment with OpenTelemetry discussions for metric exporting standards.

🗓️ Nov 21st - Introduction Meeting (take 2)

Attendees

- Erica Hughberg (Tetrate)
- Takeshi Yoneda (Tetrate)
- Dan Sun (Bloomberg)
- Arshardh Ifthikar (WSO2)
- Aaron Choo (Bloomberg)
- Krishan Wijesena (WSO2)
- Yao (Bloomberg)

Agenda

- [Erica] Short summary of Background
    - Cloud Native LLM Gateway
    - #envoy-ai-gateway on Envoy Slack
    - ai-gateway GitHub in Envoy Proxy org
- [All] Recap learnings from KubeCon
    - Aaron Choo is sharing his learnings from KubeCon:
      GenAI and the gateways
- [Erica] Review & Finalize Goals
    - https://docs.google.com/document/d/19UGS1NJcfyTqkimvvPFfdPT4kemKKzAYcu4PtHqa4eo/edit?tab=t.0
- *Add agenda points here*

Meeting Summary

Key Discussion Points:

1. **Introductions and Context:**
    - New contributors from WSO2 (Arshad and Krishan) shared their experience in API management, Envoy Proxy, and LLM traffic handling.
    - Overview of Envoy AI Gateway POC and its goals to simplify integration with GenAI services.
2. **Learnings from KubeCon:**
    - Presentations on AI gateway features from Kong and Solo.io, including request/response transformations, rate-limiting, semantic caching, and unified API specifications.
    - Noted gaps between current POC features and those offered by other vendors.
    - Identified the need for a standardized glossary for GenAI-related terms to reduce language confusion.
3. **Features for Consideration:**
    - Semantic caching for cost and performance optimization.
    - Unified API for managing multiple LLM providers (already part of the current POC for AWS Bedrock and OpenAI).
    - Advanced failover mechanisms and intelligent traffic routing based on query context.
4. **Project Goals and Alignment:**
    - Draft project goals emphasize seamless communication between applications and GenAI services via Envoy Gateway.
    - Focus on reducing complexity for developers and providing secure, scalable GenAI traffic handling.
    - Discussion on extending support for instance-level gateways in addition to broader use cases.

Action Items:

☑ ~~**Goals Document Review:**~~

- [x] ~~All participants to review and provide feedback on the draft goals document by next week.~~

- [x] ~~**Agenda for Next Meeting:**~~
  - [x] ~~Dive deeper into specific features (e.g., caching, RAG implementation, failover strategies).~~
  - [x] ~~Schedule the next meeting for Wednesday due to Thanksgiving.~~
- [x] ~~**Resources to Share:**~~
  - [x] ~~Aaron to finalize and share the KubeCon presentation summarizing AI gateway features.~~
  - [x] ~~Erica to share:~~
    - [x] ~~Glossary for standardizing terms.~~
      ~~DRAFT - GenAI Traffic Handling Glossary~~
    - [x] ~~Recording of the Envoy AI Gateway POC demo.~~
      ~~https://share.descript.com/view/Xp87gpK3PCy~~
- [ ] **Community Collaboration:**
  - [ ] Draft doc for Roadmap to collaborate on

Next Meeting:

- Tentatively scheduled for Wednesday next week.
- Agenda to focus on feature deep dives and project roadmap.

Acknowledgments:

- Thanks to Aaron for presenting KubeCon learnings and to WSO2 contributors for sharing their expertise and aligning goals with the community.

Nov 14th - Canceled due to KubeCon NA

Please visit the Envoy Kiosk at KubeCon, and don't miss the co-located Envoy Con event.

🗓️ Nov 7th - Introduction Meeting

Due to zoom scheduling issues, the meeting was held on Google Meet.

Recording: 🎬 ikr-jhqa-eha (2024-11-07 11:13 GMT-5)

Attendees

- Erica Hughberg (Tetrate)
- Takeshi Yoneda (Tetrate)
- Aaron Choo (Bloomberg)
- Yao Weng (Bloomberg)
- Dan Sun (Bloomberg)
- Arko Dasgupta (Tetrate)

Agenda

- [Arko] LFX Meeting time incorrect, Arko will raise a ticket with CNCF
- [Erica H] Welcome & Background
    - Cloud Native LLM Gateway
    - #envoy-ai-gateway on Envoy Slack
    - ai-gateway GitHub in Envoy Proxy org
- [Erica H] Connect at KubeCon
    - Start a thread in Slack to find a good time to meet
    - [Arko] LLM Routing Session focused on Ingress - maybe we should meet
        - KubeCon NA 24 - Gateway API Break Room Sessions
    - [Dan] Look to see if we can arrange a meeting with Terry from RedHat

- [Erica H] First draft Goals Doc
  - [DRAFT] Envoy AI Gateway GOALS
- [Erica H] Introductions of Attendees
  - Erica
  - Dan
  - Takeshi
  - Aaron
  - Yao

Meeting Summary

1. **Introduction**
   - This was the inaugural Envoy AI Gateway community meeting. Erica announced that a second introductory meeting will occur after KubeCon, focusing on resolving time zone and scheduling challenges.
2. **Background**
   - The project was initiated by a document Dan Sun created on the concept of a "Cloud Native LLM Gateway," which inspired community interest in leveraging Envoy for AI gateway needs, specifically around managing traffic between clients and AI services (both self-hosted and cloud-based models).
   - Dan noted Bloomberg's need for a unified gateway to facilitate access to closed-source AI models, citing the benefits of open-source contributions for tackling complex, cross-organizational challenges. This aligns with the approach Bloomberg took when developing other open-source platforms.
   - Community Involvement
     i. The project aims to gather community input on common challenges and solutions for AI Gateway functionality, inviting both contributions and feature requests from diverse organizations. Erica emphasized that as AI gateway needs evolve, solving these challenges together will benefit the industry.
3. **KubeCon Coordination**

- Several attendees will be present at KubeCon, where they will use Slack to coordinate informal meetups. Erica encouraged attendees to visit the Envoy kiosk and suggested setting up a community room meetup.
- Arko highlighted an LLM routing session focused on ingress, with plans to explore egress discussions, potentially converging on a shared API.
- Dan will reach out to Terry at Red Hat to arrange a potential meeting with maintainers at KubeCon to discuss project synergies and align on terminology to avoid overlap and confusion.

4. **First draft Goals Doc**
    - Erica has drafted and will share a goals document for community feedback. She emphasized the importance of keeping the goals concise and focused to enable future scalability.
5. **Introductions**
    - Attendees introduced themselves and shared their roles in the project. Highlights included Takeshi's experience contributing to Envoy and Arko's role as a maintainer of the Envoy Gateway project.
6. **Closing Remarks**
    - Erica thanked everyone for joining and reiterated that future meetings would be better organized. She encouraged anyone interested in contributing to join the discussions on Slack and GitHub. The meeting adjourned with a reminder to connect at KubeCon and continue the discussion in the next community call post-KubeCon.

Resources

Hybrid GenAI Gateway

Hybrid GenAI Gateway on Kubernetes

Both an Inference Instance Gateway and an Egress Gateway