Inference Extension Meeting Notes

Project Repo: https://github.com/kubernetes-sigs/gateway-api-inference-extension

Recordings: https://www.youtube.com/@kubernetes-wg-serving

Template

<Do not edit>

Jan 1, 2025

Meeting Host: <HOST_NAME>

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).

AIs

- Placeholder

Maintainer Lawnmowing (aka: Did you groom the backlog last week)

- Last week's maintainer:
- Next week's maintainer:

</Do not edit>

Meeting Notes

Current Zoom link:
https://zoom.us/j/96271651417?pwd=NViXawg6lMsRjgXbu2YmW8DxWqbjta.1

Nov 6, 2025

Meeting Host: Daneyon Hansen

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts)
- [robscott] [SIG-NETWORK] KubeCon Break Room Sessions
    - Also at maintainer summit: https://sched.co/28aCp
- [kellen] I used the zoom link on our main README and it brought me to the correct meeting (same one used in meeting description)
- [kallner] feat: Extend the text based configuration to include feature flags and the SaturationDetector's configuration
- [nirro] https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1818
- [kallner] Istio 1.28.0 has shipped with InferencePool/v1 support
- [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1502

Oct 30, 2025

Meeting Host: Daneyon Hansen

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts)
- [danehans] Should we cancel the 11/13 meeting due to KubeCon?
- [kellen] [Public]Multiple Base Models hosted on a single Endpoint Picker
- [kellen] v1.2 date?
    - My thinking is near Thanksgiving
- [Abdullah / Shimi] [Public Doc] Serving Online Batch via Inference Gateway
- [zetxqx] some updates on the [Public] Inference pool level model name redirect and traffic splitting ,mainly naming and discussion on EPP and BBR

- [srampal] Zoom link still incorrect on community pages, need to update everywhere with the right current zoom link (added now to this doc above .. okay?)

AIs

- ~~[danehans] Ask serving working group lead to cancel 11/13, Thanksgiving, Xmas, NYE meetings.~~ Complete (xref)
- ~~[kellen] Update the meeting info in the project readme.~~

Oct 23, 2025

Meeting Host: Nir Rozenbaum

Agenda

- [zetxqx] Inference pool level model name redirect and traffic splitting https://docs.google.com/document/d/12yR_nAWM-Tg2ZmgGYX1h-dlUNi0AqYoACUjNEIipl0M/edit?tab=t.0
- [nir] multi pool management through bbr proposal: [Public] Evolving BBR for Multi-Pool scenarios
- [davidbr] Issue https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1750
  One cannot leave context non-specified when configuring an EnvoyFilter with istio 1.28. It was intentionally left open to cover both sidecars and waypoints. Using ANY allows to deploy, but the filter does not seem to work
  - [kellen] have we tried this with GKE or kGW?
  - [davidbr] I only tried with Istio yet
  - [davidbr: UPDATE]: kgateway path works OK
- [nir] conformance open issues are starting to pile up:
  Examples: #1693, #1680, #1670, #1655, …

Oct 16, 2025

Meeting Host: Abdullah Gharaibeh

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).
- [robscott] v1.1 release for multi-cluster (InferencePoolImport)
- [danehans] PSA- I'm in the process of cutting the v1.0.2 patch release. I need https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1735 tagged.
- [ahg-g] [Public] EPP As a Standalone Request Scheduler

- [nir] plan the release plan 🙂
    - [danehans] Here is the first stab at a simple release planning process- The release manager is responsible for creating the milestone, the release planning issue, linking the two, communicating the release plan, etc.
- [nir] InferenceObjective - is it really optional?
    - [kellen] This is a bug, we should just fix imo
- [srampal] Initial feedback/ discussion on InferencePool extension for BBR
    - WIP: refer to Tab 2 of this doc
- [nir] endpoint-served conformance tests + implementation in gateways: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1670

AIs

- Placeholder

Oct 9, 2025

Meeting Host: Kellen Swain
Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).
- [bobzetian] https://github.com/kubernetes-sigs/gateway-api-inference-extension/discussions/1695
- [nir] we migrated to formal kubernetes youtube channel, we have a new playlist link: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1692
- [robscott] [SIG-NETWORK] Agentic Networking for Kubernetes
- [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1693
- [kellen] Evolving the EPP Framework
- [pier] Multi-modal support
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1617
- [srampal] please add the latest zoom link for community calls to the top of this agenda doc. There was some confusion this week as some of us had an outdated zoom link apparently & couldnt join.

AIs

- Placeholder

Oct 2, 2025

Meeting Host: Daneyon Hansen

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).
- [danehans] Looking for reviews of this PR that adds multi-cluster API types.
- [yangli] Looking for discussion/review on this discussion
- [srampal] Some GIE open topics on InferencePool, body routing as started here

AIs

- [nir] make sure community calls are being uploaded. Ping Yuan tang and @youtube-admins.
  Update from Nir: pinged youtube admins. With the help from Sig Contribex, I'm going to migrate the recordings to the general Kubernetes youtube channel and create a playlist for our project.
  (Youtube admins don't have access to wg-serving youtube channel).
-

Sep 25, 2025

Meeting Host: Nir Rozenbaum

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).
- [danehans] Final review discussion of the Multi-Cluster Inference Pooling (MCIP) proposal.
- [syw14] Plan to make change this week to support multi-port for inferencePool (from maxItem: 1 to 8): detailed requirements from https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1336
  - [shmuelk] See https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1519. I have code for this. Waiting on Istio fix.
  - [danehans] Adding support for multiple InferencePool targetPorts is being tracked by this kgateway issue: https://github.com/kgateway-dev/kgateway/issues/12238.
  - Consensus is to add multiple targetPorts to InferencePool in the next GIE release.

AIs

- [nir] make sure community calls are being uploaded. Ping Yuan tang and @youtube-admins.
  Update from Nir: pinged youtube admins. With the help from Sig Contribex, I'm going to migrate the recordings to the general Kubernetes youtube channel and create a playlist for our project.
  (Youtube admins don't have access to wg-serving youtube channel).
  Once we do that we should update the link at the header of this doc. In progress.
- [nir] create tracker issue for implementing endpoint served metadata https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1670

Sep 18, 2025

Meeting Host: Daneyon Hansen

Agenda

- Review action items from the previous meeting and verify that the cloud recording is set. Verify the recording from the last meeting (before the meeting starts) (xref).
- [danehans] Final review discussion of the Multi-Cluster Inference Pooling (MCIP) proposal.
  - Namespace/name sameness between exported an InferencePool and InferencePoolImport- Should the ref'd EPP also require ns/name sameness (xref comment)?
    - Potential Istio friction- When do we need to make the cluster decision. No concept of cluster name. First choose cluster then EPP? Keith to doc the Istio use case in more details and coordinate with John.
  - Should the list of exported clusters be represented in spec instead of status to support observedGeneration?
    - Should export controller details be removed (impl detail).
    - Do not follow MCS for spec/status design (been debated in the community for yrs).
    - Doc that status condition obsGen is unused.
    - Freshness? Timestamp potentially, Rob to discuss with sig-arch.
  - Thoughts on multi-implementation multi-cluster InferencePools (xref).
- [danehans] Version the quickstart (xref).
- [shmuelk] Multiple TargetPorts in an InferencePool
  - https://github.com/istio/istio/issues/57638

AIs

- Placeholder

Sep 11, 2025

Meeting Host: Kellen Swain

Agenda

- [ryan] multi-cluster inference extention: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1374
- [kellen] post-v1 focus areas
    - Flow Control
        - Etai: Can we extend the queuing operation to track external state and events (e.g., wait on LoRA loaded state)?
    - SLO prediction
    - CI overhaul
        - More benchmarks
        - Automate scale testing
        - RC validation should be pretty automated
    - Pluggable system expansion
    - InferenceObjective API expansion
        - Including solving traffic splitting
        - Tackling how we express LoRA in a Pool
    - Potential BBR expansion
    - Ensuring InferencePools expose the proper data/metrics to operate with a larger cluster
    - Expand documentation on which scorers/algos to use
        - How do scorers in llm-d interact with the scorers in IGW?
            - Look into what we are allowed to document wrt llm-d in IGW documentation
    - Nebulous: how much IGW depends on vLLM?
        - Is the goal of IGW to become independent of vLLM
            - Open an issue to at least track the level of dependence on vLLM
                - Example sg-lang issue: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1141

AIs

- Placeholder

Sep 4, 2025

Meeting Host: Nir Rozenbaum

Agenda

- [nir] release updates
    - [danehans] Today is the scheduled date for the v1 milestone. This date needs to be pushed out if we do not release today.
- [nir] Saylor to introduce himself. NGINX are looking into implementing support for GIE.
  https://github.com/nginx/nginx-gateway-fabric/issues/3644 - epic issue (high level)
  https://github.com/nginx/nginx-gateway-fabric/pull/3800 - detailed design
- [nir] epp protocol - should we use headers instead of envoy metadata?
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1394
- [robscott] revisit multi-cluster:
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1374

AIs

- Placeholder

Aug 28, 2025

Meeting Host: Daneyon Hansen

Agenda

- [danehans] Review AIs from last meeting
  [danehans] v1.0.0-rc.1 has been released.
- [danehans] v1.0 milestone issue burndown and update the milestone target date.
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1484
- [srampal] Initial discussion on BBR extension
- [kellen] how do we want to handle releasing the API for GW maintainers

AIs

- Placeholder
- Release cadence discussion -
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/discussions/1495

Aug 21, 2025

Meeting Host: Kellen Swain

Agenda

- InferencePool v1 api review
  - Xiyue Yu Optional non-pointer usage https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1173#discussion_r2277950132
  - danehans: Refactors v1 InferencePool status
- [rob] Viability of Model Name Redirect and Traffic splitting for Gateways
  - Does/can request header modification happen *before* EPP callout?
  - [Kellen] continue this conversation in slack
- [kellen] Release cadence post-v1.0
  - Feature gating
  - Better auto for release
  - Release channels
  - Comms in open
  - Road map
  - [shmuel/nir] tags on main instead?
    - Shmuel - tags on main are essentially a sha
      - Some large features may be half-baked if we cut quickly
  - [nir] 2 weeks too quick - 6 weeks is too long
    - 3 weeks might be right?
    - [Cong] Consider a regular/frequent release cadence (ideally with automation) where we don't need to plan, just pick the latest and best release candidate that pass all the qualification; (e.g., GKE releases every week (mostly) with automated qualification)
- [kellen] v1.0.0-rc cut
  - Do we wait to resolve the traffic splitting issue?
- [nir] poor performance in bursty workloads - **(NEXT WEEK)** https://github.com/llm-d/llm-d-inference-scheduler/issues/298
- [ryan]- **(NEXT WEEK)** https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1374

Action Items

- [kfswain] spin up discussions in IGW repo for:
  - Release cadence
  - Traffic splitting feasibility
  - V1.0.0-rc timeline

Aug 14, 2025

Meeting Host: Daneyon Hansen

Agenda

- [xiyue] Decision of https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1336? Should we change TargetPortNumber int32 to become TargetPorts []Port(length == 1) to allow for possible future extension
- [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1372
- [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1373
- [srampal] Intro discussion on potential future topic - BBR & LLM routing
- [ricardo] https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1366 some quick discussion about KAL and why we are doing it
- [robscott/bexxmodd] multi-cluster follow up https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1374

Action Items

- [daneyon] will split up 2 and 3 from https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1336
- [ricardo] create a makefile target for golangci-lint + kal https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1380
    - The linter is in, no make target PR in yet; under review
-

Aug 7, 2025

Meeting Host: Nir Rozenbaum

Agenda

- [xiyue] EPP now supports either v1 or v1a2 InferencePool(see https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1277). But for the default, do we want to default to v1a2 InferencePool or v1 InferencePool?
- V1.0 API Review
    - [xiyue]

https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1173#discussion_r2214003795
    - [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1173#discussion_r2215474675
    - [robscott] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1273
- [maroon] llm-d kv cache precise scorer - upstream to GIE open questions
- [maroon] new extension point to support DP use case
  llm-d-inference-scheduler Data Parallelism
- [robscott] Follow up on [SIG-NETWORK] Multi-Cluster Inference Gateways
- [liu-cong] approximate prefix cache scorer user feedback + enhancements to make it more user friendly and more precise (issue #971, #972, #973, #1304). WIll come up with a plan.

Meeting Notes

Jul 31, 2025

Meeting Host: Kellen Swain

Agenda

- [xiyue] InferencePool v1 CRD introduction and InferencePool alpha CRD deletion. See context in https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1201

InferencePool API Version	Group	API Types	Generated Client	CRD
v1alpha2	inference.networking.x-k8s.io	Yes	Yes	No
v1	inference.networking.k8s.io	Yes(use `type InferencePool` `v1.InferencePool` syntax)	Yes	Yes

- [robscott] [SIG-NETWORK] Multi-Cluster Inference Gateways
- [nir] enhancements for config api - suggestion
- [danehans] Plan to workaround fail-open conformance test issues/1266.
    - TBD based on timeline for Envoy fix.
-

Meeting Notes

Jul 24, 2025

Meeting Host: Nir Rozenbaum

Agenda

- Review action items from the previous meeting and verify the cloud recording is set in . Verify the recording from the previous meeting (before the meeting starts).
- [nir] quickstart broken in main - need to verify this is fixed asap
- [robscott] Follow up on [SIG-NETWORK] Multi-Cluster Inference Gateways
- [kellen] Proposal for how we handle the InferenceModel successor(s) :https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1199
    - Open question(s):
        - Should these CRDs allow multi-pool reference?
        - Is the phased approach _mostly_ acceptable to this audience?
- [abdullah] Model Name Redirect and Traffic splitting
- [andresguedez] https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1226

AIs

- Placeholder

Jul 17, 2025

Meeting Host: Daneyon Hansen

Agenda

- [xiyue] move x-k8s InferencePool to apix directory and add v1 InferencePool to api/v1 directory and change epp to use v1 InferencePool only, see details in https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1116 and https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1118
- [xiyue] InferencePool v1 CRD introduction and InferencePool alpha CRD deletion. See context in https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1150

InferencePool API Version	Group	API Types	Generated Client	CRD
v1alpha2	inference.networking.x-k8s.io	Yes	Yes	No
v1	inference.networking.k8s.io	Yes(use `type InferencePool` `v1.InferencePool` syntax)	Yes	Yes

    - Encourage users and controllers to exclusively focus on the v1 API going forward, thus we omit the v1alpha2 CRD from the release. However, this PR leaves controllers with the option to support both v1a2 and v1 concurrently.
- [robscott] SIG-Net API Review
    - [SIG-NETWORK] InferencePool GA Review
        - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1173
- Config API, need good defaults to simplify the experience
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1169
- [robscott] [SIG-NETWORK] Multi-Cluster Inference Gateways
- [nir] release 0.6 - which capabilities are we targeting? target date? https://github.com/kubernetes-sigs/gateway-api-inference-extension/milestone/9
-

Jul 10, 2025

Meeting Host: Kellen

Agenda

- [shane] “AI Gateway” Working Group Proposal
    - https://groups.google.com/a/kubernetes.io/g/dev/c/XC_8qAyk8W0
    - https://groups.google.com/g/kubernetes-sig-network/c/j50pypPlLSk
    - https://docs.google.com/document/d/10WTdHYW5x2rw6BTgDzW7X-5QNesAh205MuoaUe5-IQg
- [nir] config api and removal of plugins env vars and initialization through code. https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/1130
- [bobzetian] how to add inference-gateway-extension annotation version
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1134
- [bobzetian] conformance tests for v0.5.0 release:
    - one minor PR is needed: 1133(add a consts version)
    - successful reports:
        - gke-gateway:https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1005

        - istio https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1102
- [danehans] Kgateway conformance status update- All but 1 conformance test passing, ~1 week ETA for a conformance profile PR.

Jul 3, 2025

- [shane] CANCELLED: some of us joined today, but there were only a few (probably due to US holiday) so we decided we might as well just wait until next week.

Jun 26, 2025

Agenda

- Review action items from the previous meeting and verify cloud recording.
- **Note**: Next week is a long holiday weekend for the US, no meeting next week
- [danehans] In case you missed the news, v0.4.0 was released. Check out our v0.5.0 milestone to see what's on the roadmap.
- [robscott] Conformance test updates
    - InferencePool ResolvedRefs: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1070
- [ahg-g] Revisiting The InferenceModel API

AIs

- [Kellen] perf benchmark dashboard for v0.5
- [daneyon] create a backlog + v0.6 milestones
-

Jun 19, 2025

Agenda

- Review action items from the previous meeting and verify cloud recording.
- [danehans] Confirm meeting recording to the cloud.
- [kellen] https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1007
    - How do we want to handle the match?
- [danehans] v0.4 release status (scheduled for today)
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1020 is needed to fix a docs/release bug.
    - Has anyone completed a perf benchmark of v0.4.0-rc.1?
- [danehans kellen] Add Nir as a maintainer.

AIs

- [danehans] Update template to include recording check
- [danehans] Cut v0.4 release after https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/1020 merges and is backported

Jun 12, 2025

Agenda

- Review action items from the previous meeting.
- [liu-cong] v0.4 should be ready to go
    - Cutting the first RC today (06/12), will soak for a week to give time for benchmarking comparison, determining conformance test progress, etc.
    - Will decide next week what we would like to do with the release
- [elevran] Review new data layer extensibility proposal
  Pluggable/Extensible Data Layer - Design proposal
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/703
- [kellen] Fairness
    - Revisiting The InferenceModel API &
      Proposal: New SLO Parameters in InferenceSchedulerObjective don't directly address fairness, but fairness is potentially the most important design consideration in a multitenant pool.
    - what is the 'primary key' of a workload?
        - Whatever InferenceObjective you match with
- [nirrozenbaum] e2e tests pre submit job:
  https://github.com/kubernetes/test-infra/pull/34974

AIs

- Llm-d benchmarking & pd meetings are in conflict with this one
- Cut a GA/v1.0 milestone
- Review conformance testing implementation progress:
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/977

Jun 5, 2025

Agenda

- Review action items from the previous meeting.
- [danehans] PSA: Inference Extension k8s blog post.
- [Cong] v0.4 release check
    - Chase status update

    - Prefix cache aware scheduling and scheduler framework are important enough for a release
    - Will cut scope if needed to ensure the 0.4 release out in Mid-June timeframe
- [Cong] Short-term scheduler v2 enhancements targeting graduation at v0.5 release
    - contribution/collaboration welcome
- [robscott, bobzetian] HTTPRoute.BackendRef.Port discussion
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/discussions/918
- [robscott] FYI [SIG-NETWORK] InferencePool GA Pre-Review
- [danehans] Conformance EPP shim #893
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/922
- [elevran] heads up - new data layer extensibility proposal
  Pluggable/Extensible Data Layer - Design proposal
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/703
-

AIs

- [Cong] Convert the enhancements to issues; benchmark automation script [this requires more effort and I currently don't have an ETA].
- [nirrozenbaum] push a PR for CI pre-submit job that runs e2e tests and uses llm-d simulator

May 29, 2025

Agenda

- Review action items from the previous meeting.
- [nirrozenbaum] discuss InferenceModel CRD
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/discussions/872
    - Related: Revisiting The InferenceModel API
- [nirrozenbaum] Istio bug - not returning response body
  https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/888
- [robscott] What should happen when HTTPRoute references an InferencePool using a different port?
    - A: HTTPRoute status set to ResolvedRefs: False
        - Clarify the expected `ResolvedRefs` status for an HTTPRoute when its `backendRef.port` mismatches the port on the referenced `InferencePool`. Full context #887 & #886
    - B: Requests get 503
    - C: Requests get 404
- [robscott] GA for InferencePool

    - Target timeline: August
    - Graduation criteria: https://gateway-api.sigs.k8s.io/concepts/versioning/#graduation-criteria
    - Aiming for conformance test completion in June
    - Aiming for 3 conformant implementations in early July
- Revisiting The InferenceModel API

AIs

- ~~[danehans] Create an issue regarding meeting recordings being updated to YouTube.~~
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/932

May 22, 2025

Agenda

- Review action items from the previous meeting.
- [danehans][5-10m] Updates on KubeCon CFPs (xref).
- [anyone] llm-d announcement recap and add'l discussion?
- [kfswain] v0.4 drive to completion
    - v0.5 scope/ drive to GA

AIs

- ~~[danehans] Create milestone for InferencePool GA~~
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/milestone/6

May 15, 2025

Agenda

- [danehans] [10-15m] KubeCon NA CFP topics and speakers (May 27 deadline)
    - How can we all coordinate to have a well orchestrated set of topics
        - We should aim towards multi-speaker/multi-org talks
- [Huamin Chen] [10m]
    - Demo of latest update on Semantic Processor
    - Follow up the "where to live" discussion
        - Pluggable architecture will let a user add their own extensions
            - Currently EPP/IGW is focused on scheduling & routing; less focus around classification, semantics
- [shane] [5m] discussions about getting other proxy implementations on board
    - revisit proxy_wasm discussion from Kubecon Salt Lake
        - Rob: ++ long-wished for integration
        - Shane: we should create an issue

- [shane] [5-10m] where are we at with performance?
    - Is anyone doing some intense perf/scale testing with the ext-proc implementation?
    - Is Golang GC going to hit us pretty hard?
        - [huamin] in general, Rust is going to be better than Golang for performance: https://medium.com/@dmytro.misik/go-vs-rust-web-service-performance-7fb10bbf9a9f
        - [abdullah] So far in our setups the backend resources for model serving overshadows anything that we need to do with the epp.
        - [cong] we did some benchmarking that showed the epp doing well under 1000s qps without struggling too bad. Prompt size though might hurt us eventually though we'll need to keep tabs.
        - [cong] maybe we should test more extremes however.
        - [kellen] let's find the breaking point for the epp, let's push it until it falls over
        - [nir] ibm+redhat have a vllm simulator that's going to be open sourced soon, it may be able to help here
            - [kellen] ++
- [robscott] Multicluster + Inference
    - [robscott] Would others be interested in multicluster?
    - [nir] Is the intent for us to implement multicluster? Or work with others?
        - [robscott] implementation details are per implementation
        - [robscott] will cut issue to gauge further interest
- [robscott] June 5 API review at SIG-Network meeting
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/830 needs feedback.
- [xiyue] Conformance Test for Isio and Kgateway
    - Sina Chavoshi is currently working on the implementation of Conformance Test Technical Design for Gateway API Inference Extension, will get it done in early June
    - Does Istio and Kgateway have plans to dedicate time to pass the conformance tests and submit a conformance report before the end of June or before GA?
        - [danehans] Yes, I am currently adding InferencePool status support. I plan on having Kgateway fully conformant in the next minor release.
    - [bugbug] https://docs.google.com/document/d/1F82G_o3ENM94Imx5pYKG0x-m508_FSJyfbu1d8J8CCk/
        - [robscott] When is Istio's next release?
            - [shane] would love to see this(shared doc) before KubeCon ATL

AIs

- Include my thoughts for what the CFP topics could be
- [danehans] Create an issue to track kcon CFPs and link to slack channel.

    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/841
- [shane] Creation of issue discussing how we can work with other proxies (proxy_wasm for example) & just create a simple gRPC shim
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/840

May 8, 2025

Agenda

- [Abdullah] (15m) Pool sharing scenarios (xref)
    - [Clayton] we should pick the best possible use case to argue against sharing in the pool vs not
        - We can then extract that to a principle to help guide
- [shane] chat about EPP HA deployment · Issue #692
- [ricky] prefix aware scorer
    - Related: WIP Implementation & Proposal

AIs

- Placeholder

May 1, 2025

Agenda

- Review action items from the previous meeting.
- [shane/ahg/robscott/kfswain] discuss enhancement proposals
    - Context: https://kubernetes.slack.com/archives/C08E3RZMT2P/p1743178535027789
    - Process state machine:
        - Initial idea
        - Idea accepted; needs design
        - Design drafted
        - Design reviewed/accepted
        - Work scoped/planned
- [Huamin] https://docs.google.com/document/d/1bTW397hgcARaGTAKq8QwmcbYLeHzJbxu2t6OZxu0FTg/edit?usp=sharing
    - Scoping: where does semantic caching ‘live’
        - How/Can we avoid making EPP monolithic
        - E2E latency
            - not an issue with semantic caching, but a problem in general as we expand features/extensions

- [rob] GA in/by August?
    - InferencePool
    - Interface between Gateways and Endpoint Picker(s)
- [Cong] Prefix cache aware scheduling WIP PR, feedbacks welcome: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/768/files

AIs

- How do we want to migrate existing proposals?
- Define the GEEP process
- Delineating between geep and just regular old issue?
- Custom labels:
    - https://github.com/kubernetes/test-infra/blob/master/config/prow/plugins.yaml
- Placeholder

Apr 24, 2025

Agenda

- [Luke] [PUBLIC] EPP Flow Controller for Priority, Fairness, and Queuing
- [Kellen] EPP Multi-tenancy discussion
    - Should we support it?
    - Is it an anti-pattern?
- [Chavoshi] Conformance tests test case details
  [Public] Conformance Test Technical Design for Gateway API Inference Extension
- [demo - 15 min] semantic caching
    - Related [Public] Semantic Processing for Inference Extension

- [x] ~~dancyonhansen@gmail.com - Create an issue for an ops guide (spawned from Multi-tenancy discussion)~~

Apr 17, 2025

Agenda

- Review action items from the previous meeting.

- [Cong] Prefix caching POC + Scheduler plugins https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/602
- [Luke] Queueing: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/674
- [chavoshi] [Public] Conformance Test Technical Design for Gateway API Inference …
- [nir] if one wants to work with multiple base models, what is our recommended setup?
- [Huamin] [Public] Semantic Processing for Inference Extension
    - [Kellen] Quick one-liner summary for context: this is a proposal for how we approach handling semantic caching in IGW

Apr 10, 2025

Agenda

- Review action items from the previous meeting.
- [kfswain/shane] discuss current API design, and the potential for declarative routing rules
- [nir] SessionAffinity. If a pod was serving a request, it makes sense that the same pod will serve subsequent requests of the same client on a best effort basis (using session header). https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/675
- [kellen] EPP Refactor - Proposal on how to restructure the EPP deployable to make the code more extensible, maintainable, and approachable
- [Cong] Prefix caching POC + Scheduler plugins https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/602
- [rob] https://github.com/kubernetes/website/pull/49898
- [Luke] Queueing: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/674

AIs

- Create a discussion about “declarative routing rules”
    - Kellen and I are opting to move forward with a proposal, instead of the discussion, but avoid implementation details for the first pass.

Apr 3, 2025

<No Meeting; Kubecon week>

Mar 27, 2025

Agenda

- Review action items from the previous meeting.
- Cutting v0.3.0 this week
    - No new features, just quite a bit of polish for Kubecon
    - Tracking issue
- [robscott] KubeCon
    - KubeCon - Gateway API - Break Room Sessions
- [nir rozenbaum] ranking pods
    - [nir] pluggable way to specify which algorithms you want your endpoint picker to use, out of a “registered list”? Not ready for a full design, but want to see if this is something others need or see as something needed in the future.
        - [kellen] makes sense, in the short term fork the epp and make your own so we can get a POC, demo it, and then merge back in
        - [rob] the API is currently very barebones, the extensions are sort of a closed box
            - [rob] so q: what belongs as a toplevel concept, and what’s specific to an extension?
            - [rob] any extension could support multiple algorithms with more or dynamic sequencing, so on first glance this seems to make sense, we just need to figure out the how
            - [nir] the reference implementation is missing some things we need to get started, in particular some “registered list” seems like a low-hanging fruit to help build more dynamic extensions.
            - [nir] everyone is going to have potentially very different pod selection algorithms, so I would like to see this API specify this since it sounds universal
                - [rob] ++
                - [rob] We need to figure out better what’s going to be CORE and what’s going to be IMPLEMENTATION-SPECIFIC
                - [rob] however we have some leeway and time while people are just building the “closed box” implementations for now to figure that out.
                - [shane] should we consider more “prototyping in upstream”?
        - [kellen] Clayton mentions here the need to distill our routing algo to a scoring mechanism, which would help generalize algos & also stack rank

        - [Clayton] Should our focus after Kubecon be:
            - What features can we add with no config
            - What config will everyone need
        - [Clayton] dimensions of IGW extending
            - Forking EPP and implementing unblocks now
                - Gives time for a more general solution to mature
            - Implementing in EPP
            - Implementing in an extension CRD
            - Implementing in the API proper
        - [Kellen] Next steps:
            - Our scheduling algorithm needs to be more pluggable
- [shane] (unfortunately have to drop at the half hour, but if you feel like talking about this anyway, go for it)
    - Has anyone seen anything that Dynamo is doing so far that seems like an obvious gap for us yet? Still too early?
    - Separately: have any implementers run into any significant issues with Envoy? Are there areas here where Envoy struggles?

AIs

- [Kellen] Shape up https://github.com/kfswain/go-py-interface to be able to be discussed
- [Nir] Come up with an initial design to support pluggable algorithms. The first version might support a set of algorithms Inference extension supply. We should be able to support more than one algo with the ability to configure the order the algorithms are executed.

Mar 20, 2025

Agenda

- Review action items from the previous meeting.
- [Abdullah] https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/496
- [Nir] do we want to add a pre-submit CI job that runs e2e? After the latest changes in Makefile and e2e it's possible to run cpu based e2e tests (if GPUs are available even better).
    - [Kellen] to keep me honest: I want to move the image build to its own CI job so that it doesnt need to be rerun on a retest
    - [Nir] not sure on resourcing we will get
        - [Kellen] Can we check with #sig-testing on what kind of resourcing limits we have
        - [Kellen] [External] Standardizing Large Model Server Metrics in Kubernetes
    - [Abdullah] Only intel CPUs work for the E2E test

- [danehans] Blog post status update.
    - Benchmarking feedback
- [Kellen] Cross-collab meeting w/AIBrix
    - https://github.com/vllm-project/aibrix/issues/861#issuecomment-2730681254
    - In-person convos @ Kubecon
        - Any other AIBrix personnel?
    - AIBrix technical problems
        - How to make app-based router and proxy-based router work together?
            - Have to reimplement the async engine
        - Proxy based
            - Lacks programmability
            - How to make algo more pluggable
- [robscott] KubeCon
    - KubeCon - Gateway API - Break Room Sessions
    - wg-serving meet up?

AIs

- [Kellen] to keep me honest: I want to move the image build to its own CI job so that it doesnt need to be rerun on a retest
- [Kellen] Work on sim server fleet (make issues)
    - [Shmuel] Follow up next week on simulated model server
- [Cong] Follow up on benchmarking graphics to display a more compelling case & have a user able to do that themselves
- [Kellen] Schedule AIBrix + IGW meeting post kubecon + design proposals to make IGW more pluggable/composable

Mar 13, 2025

Agenda

- Review action items from the previous meeting.
- [danehans] Provide additional context for https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/489.
- [danehans] Any updates on Kubecon- W/Th planned on having an IGW meetup?
- [Kellen] IGW conformance test suite
    - [Rob] May have someone that will have a design doc for this
        - Follow up next week to check progress
        - [Kellen] Make an issue on this to track
- [Kellen] intend to cut v0.2.0 this afternoon
- [robscott] Blog post: https://github.com/kubernetes/website/pull/49898

- Subsetting and fallbacks, the proposal is ready: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/445

AIs

- ~~[danehans] Create inf ext and vLLM community about CPU-based image.~~
- [kellen] Talk to Nir and remove CPU-based image and deployment refs, manifests, etc.
    - vLLM has a supported CPU-based image https://github.com/vllm-project/vllm/issues/14756 (Thanks Nir!!)
- [ceposta] Review https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/480, test, provide feedback, etc.

Mar 6, 2025

Agenda

- Review action items from the previous meeting.
- v0.2 RC out
- V0.3.0 in April
    - Targeting prefix cache aware routing
- Support for fallbacks and subsetting: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/445
- Kubecon
    - W/Th planned on having a IGW meetup

AIs

- [Nir] implement a CI test using the Qwen cpu based model

Feb 27, 2025 - Recording

Agenda

- [danehans] Version-controlled docs are needed.
    - Issue: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/407
    - I created this gist for a v0.1.0 quickstart guide.
    - The script used to manage the quickstart for a release is broken due to the guide being moved to `site-src`.
- [nirrozenbaum] cpu based vllm image usage in CI gating tests.
    - [Clayton] is an E2E test using CPU serving going to validate what IGW is doing (leveraging GPU/TPU characteristics)

        - Should we have a simulator fleet of model servers to validate IGW behaviors?
            - This would be in addition to our CPU-serving test
            - Potentially funding an accelerator based CI pool?
            - [Kellen] Follow up on getting an image for sim-model server
        - Follow up with vLLM to have a standard, accepted CPU based serving image
            - Folks who might want this
                - Envoy AI GW
                - Kserve
- [danehans] Any updates regarding “[clayton] Which controller writes status for InferenceModel?“ from the last meeting?
    - Remove **EndpointPickerNotHealthy** condition from InferencePool Status: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/385
- [robscott] v0.2 timeline
    - If target is Mar 10, we should have RC next week
    - Review fields wrt:
        - Validation
        - Immutability
- [robscott] KubeCon
    - Maintainer Summit: https://groups.google.com/a/kubernetes.io/g/dev/c/EEsRJt7Qbfl/m/cBT5pmGOAQAJ?utm_medium=email&utm_source=footer
    - Other working sessions: KubeCon NA 24 - Gateway API Break Room Sessions
    - Talks:
        - AI Day: https://sched.co/1u5fc
        - KubeCon: https://sched.co/1txC7
- [Nir] (if time) Run through cpu/qwen/CI setup

AIs

-
- [ahg-g] address https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/385

Feb 20, 2025 - Recording

Agenda

- [robscott] InferencePool should have nested status similar to HTTPRoute
    - https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/379
- [clayton] Which controller writes status for InferenceModel?

    - [clayton] Want to simplify/minimize work required to create alternative EPPs
    - [rob] Needs to be whichever component is actually implementing InferenceModel API (EPP right now). [danehans] +1 here. We should consider adding a ResolvedRefs condition to InferencePool status to ensure resources can be resolved using inferencepool.spec.selector.
    - [clayton] We should split control plane with data plane
    -
- Review action items from the previous meeting.
    - [Kellen] Created https://github.com/kubernetes-sigs/gateway-api/issues/3625
    - [Kellen] Issue Triage happened on the 19th
- [Clayton] Doing a refresh of README to more clearly explain what gateway extension actually is, and separate it from the idea of AI Gateway
    - Working on a diagram, want to share and get feedback
- [Abdullah] Tracking issue for the 0.2.0 release: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/362
    - Call for help: many are “good first issues”, please feel free to pick up any issue you are interested in, or reach out if you are not sure and want suggestions.
    - Target release date for 0.2.0; the suggestion is the first week of March for RC and one week later for the canonical release. I want to get suggestions.

AIs

- [robscott] Doc the domain of the gw controller (include in an arch doc).
    - [kellen] Ensure we have an architecture doc (principals, conventions, etc.)

Feb 13, 2025 - Recording

Agenda

- [Cong] Benchmark automation tool demo.
    - [danehans] I commented here about doc’ing perf testing.
- [Daneyon] Review action items from the previous meeting.
- [Kellen] Take Extension Protocol to GWAPI
    - [shane] as usual I have to leave at the halfway point! That said, I’m very interested in this: I want to try and find allies in the Gateway API community because my hope is that if there’s a lot of interest in the protocol (e.g. Gateway API implementations that want standardization for how you declare LB algorithms) we will actually be able to use their help to *accelerate* the GIE project by effectively having more people working on things needed here. We should be on the lookout for GWAPI implementations that could be allies here.
- [Daneyon] Does anyone have experience using the Nvidia Jetson Orin Nano for a dev env.
- [Rob] Envoy Original DST + Fail Open

AIs

- [danehans] Create a PR to track the need for a consolidated approach to benchmark tooling/docs. See: https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/332
- [Rob] Create issue regarding Orig DST cluster type
- [Kellen] Schedule a maintainer meeting to triage issues.
- [Kellen] Create issue in GWAPI discussing a general approach to GW extensions

Feb 6, 2025 - Recording

Agenda

- v0.1 released
- [robscott] Body-to-header translator for model params
- [danehans] Update on AIs from last week's meeting.
- [Kellen] Criticality convo
    - (this can be handled here, but wanted to call it out as its been brought up a decent amount)
    - You want to prioritize workloads not models
        - That is our current APIs intent (perhaps we need to work on wording, spec)
- [Jiaxin] CPU image for testing purpose. https://github.com/kubernetes-sigs/gateway-api-inference-extension/issues/259#issuecomment-2635485066
- [Kellen] Envoy Gateway Integration
-

Jan 30, 2025 - Recording

Agenda

- [danehans] Updates on topics from the last meeting, e.g. discussion tab, repo generalization, etc.
- [danehans] Create a presubmit CI job that runs e2e test (xref).
    -
    - We could try to run on CPUs if we can't find GPUs
- [danehans] Config proposal status update (xref).
- [robscott] v0.1 release
    - Steps required
    - Target date
    - Release candidate first?
    - Changelog?

        - Reference: https://github.com/kubernetes-sigs/gateway-api/blob/main/CHANGELOG/0.x-CHANGELOG.md#v010
            - https://github.com/kubernetes-sigs/gateway-api/blob/main/RELEASE.md (not much automation)
            - Consider creating a GH template to guide the release process.
    - Goal- Make integrations as easy as possible.
- [Kellen] Would love to hear from Joe about his multi-LoRA usecase and how we can work together on this!
    - Super exciting, happy to have ya

AIs

- [danehans] Document the release process. [update] @ahg-g and I completed this and Kellen confirmed with the v0.1.0 release.
- [danehans] Create v0.1.0.rc-1 and v0.1.0 milestones: https://github.com/kubernetes-sigs/gateway-api-inference-extension/milestones
- [ahg-g] Ping sig-net leads, i.e. Shane, for final review of config API PR.
- [Cong] Use `qwen/Qwen2.5-1.5B-Instruct` as backup, base model is small and there're lots of public adapters for e2e CI job.
- [Jeffwan] work on the vLLM CPU version for e2e CI job.
    - [danehans] I have this running in Docker but I'm still working through the details to run as a k8s deployment (xref).

Jan 23, 2025 - Recording

- [Kellen] Want to enable to Discussions tab on our GH repo
    - Would be great to have a place that is just discussions, not just issues
    - Didn't do this before simply because our active contributors weren't admins
    - Objections?
        - None, just would be good to provide some guidelines
- [Kellen] Integrating GIE with GW implementations
    - When should we start the generalization effort for this extension?
- [ahg] Gateway Inference Extension "Contracts"
- [Cong] Initial model server protocol proposal: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/164

Jan 16, 2025

Agenda

- Intros for new (and returning) faces!
    - We have a reasonable batch of newer folks, would be great to introduce everyone

- InferencePool config proposal: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/162
- Configurable scheduling filter PR: https://github.com/kubernetes-sigs/gateway-api-inference-extension/pull/169
    - What configurability do we need? for dev, for operator?
- [Kellen] V0.1 discussion
    - Just if we have time, there are still some hanging items

Jan 9, 2025 - Recording

Agenda

- V0.1 API review with sig-net leads happened on Tues:
    - PR containing the review
        - Criticality field
            - Make it optional (*) so if we create an orthogonal, we can oneof it with the Criticality field
            - Do we have a *concrete* usecase that requires more than the 3 current fields
                - If not, we should stick with the simple, limited field
    - Planning a release in ~ a week
    - Slides: [SIG-NETWORK] Gateway API Inference Extension v0.1 Review
- InferencePool config proposal
    - PR
    - Doc
- [robscott] Header choice for extensions that select endpoint
    - X-gateway-destination-endpoint?
    - Note: we should have a new folder in the docs dir to help clarify what is extension protocol and what is inference

AI's

- ☑ ~~Create Issue to discuss the Header Key used for Endpoint Selection~~
- ☑ ~~Lengthen this meeting to be 1hr~~

Dec 5, 2024 - Recording

- [robscott] [SIG-NETWORK] Renaming Instance Gateway + APIs
- [ahg] [Public] Revisiting the Latency Objective

Nov 21, 2024 - Recording

- Looking for real world usecases to benchmark against

    - Data would be great
- Project updates:
    - API implementation
        - Nearly done within repo; will create a quickstart guide tomorrow
        - Example implementation within Envoy Gateway planned, will begin implementation soon
    - Benchmarking
        - Link here:
        - [Abdullah] **Q**: has the community noticed latency impact on merged vs unmerged LoRA adapter inference?
            - How much?
            -
- [robscott, shaneutt] https://github.com/kubernetes-sigs/llm-instance-gateway/issues/41 follow-up from the SIG Network meeting
    - https://docs.google.com/document/d/1_w77-zG_Xj0zYvEMfQZTQ-wPP4kXkpGD8smVtW_qqWM/edit?tab=t.0#

Oct 31, 2024

- No agenda. Happy Halloween!

Oct 24, 2024 - Recording

- [Cong] Model server protocol
  [PUBLIC] Model Server Protocol for LLM Inference Gateway

Oct 17, 2024 - Recording

- Envoy Gateway is on-board to be the referential implementation
    - Thread - Envoy GW issue
    - Open to have contributors here if interested
    - Open Issue on our side:
        - Issue Link - I-GW issue
- CRD implementation PR opened
    - PR
- LoRA adapters orchestration
    - Proposing to start with a simple initial solution based on a sidecar
    - Jiaxin share a doc next week
        - Code at the end of Oct
- Model server protocol
    - Hope to have PR open this week (by next Th for sure)

Open Questions

- [Mattia] - Is there the desire to have multiple GWs support I-GW?
    - [Abdullah] Please take a look at the API Proposal PR, it is already merged, but you can still comment on it and we can patch the proposal https://github.com/kubernetes-sigs/llm-instance-gateway/pull/5
    - [Kellen] Yes, absolutely. Envoy Gateway implementation is simply referential, and any others would be very open to be supported. Very similar to our model server strategy (starting with vLLM, and expanding as we define our protocol)
        - Caveat: We use ext-proc heavily, and the given GW needs to support that

Oct 10, 2024 - Recording

Agenda

- Gateway integrations
    - Relevant PR: https://github.com/kubernetes-sigs/llm-instance-gateway/pull/18
    - Relevant issue: https://github.com/kubernetes-sigs/llm-instance-gateway/issues/19

Oct 3, 2024 - Recording

Agenda

- Algorithm doc
    - [Public] Inference Gateway Scheduling Algorithm
- Naming discussion v9001
    - UseCase -> **LLMService**
    - BackendPool -> **LLMServerPool**
    - OSS in alignment

Notes

- SLO metric
    - Jiaxin's team uses
        - TTFT
        - TPOT
        - Queuing time
- Output length prediction (using simple data analysis) is hard.
    - Especially so since it needs to be done per use case

Sep 26, 2024 - Recording

Agenda

- Naming of the 2 CRD objects
    - Trying to clarify the forward compatibility of these objects
    - Have a doc that we hope to share next week

Sep 19, 2024 - Recording

Agenda

- [Yuan] Include rationales for building default open implementation on top of Envoy in proposal
- [Evan] Did I hear there was an implicit assumption of a one-one mapping between objective types (latency/throughput) and backendpools? If so, what are the implications of this for the many-many (?) relationships allowed between pools and use cases?
    - [Answer] MVP will only focus on Latency objectives at this time, so we punt on this for now
    - [Answer] But in the future, we would bubble up an error for that usecase on the specific BEP if there is a mismatch
- [Kellen] BackendPool CRD will be explicit.
    - We do not want to duplicate Service capabilities when identifying backends.
- [Kellen] LLMUseCaseSet naming
    - https://github.com/kubernetes-sigs/llm-instance-gateway/pull/5
    - LLMRoute? (by the time you are at the BP, there isn't routing to be done)
        - HTTPRoute manages many of the routing decisions already, duplication of names may be confusing
    - ModelGroup is reasonably generic
        - Add enum as we add non-LLM usecases
        - 'Group' sidesteps the 'Set' debate
        - 'Model' keeps this generic enough to handle the future like StableDiffusion

Notes

- Using a service with BP
    - What if a request is sent directly to the service and circumvents the BP?
        - OOS for at least MVP, also nothing we can really do to stop this, other than potential extensions to model servers to emit metrics when non, gateway requests are sent.
        - Document these decisions and how a user might foot shoot should they not follow best practices
    - Multiple service support necessary

Sep 12, 2024 - Recording

Recording: https://youtu.be/siRavi_o1aA

Agenda

- API Proposal discussion
    - 'Changelog'
        - Proposing BackendPool to be implicit for MVP
        - Allowing a single objective metric for MVP
            - Name: DesiredAveragePerOutputTokenLatencyAtP95OverMultipleRequests
            - Specifying an objective gives higher priority over non-specified
        - Adding a glossary section (in-progress) to define:
            - Priority
            - Fairness
            - Lora Affinity
            - Latency Based Routing

Notes:

- If we do not provide a BackendPool object, and we attach no use cases. Where would the Error or Status percolate within K8s?
    - HTTPRoute is a valid option
- Time to first token is also another metric, users care about that because streaming can fake the perceived real E2E latency
- Axioms to clarify:
    - The Instance Gateway will attempt to minimize queuing at the model server level
    - The instance gateway will default to maximizing usage, and change behavior only in resource-constraint scenarios
- Questions:
    - Can a useCase reference multiple BEPs? Vice Versa?
    - How might I deploy this solution?
    - Are all the controllers necessary for this solution going to be provided by instance Gateway?
    - For MVP is a single objective is acceptable?
    - Clarify Latency/Throughput based routing
        - Define why we find them orthogonal/diametrically opposed
        - Call out clearly that we are focusing on Latency based routing *first*
        - [Evan] Did I hear there was an implicit assumption of a one-one mapping between objective types (latency/throughput) and backendpools? If so, what are the implications of this for the many-many (?) relationships allowed between pools and use cases?

Sep 5, 2024 - Recording

Recording: https://youtu.be/dEL5wAJSHdo

Agenda

- Execution
    - [5m] Administrative [Kellen]
        - Best time slot for this meeting?
            - Poll
            - EU participation
        - Do we want a slack channel dedicated to LLM-IG?
            - Clayton: suggest starting with slack channel until we become disruptive
    - [5m] Transferring and approving initial scope into a proposal doc in the repo
        - Ensure we all align on the top level goals
        - Ensure we have non-goals described well
        - Clayton / Yuan: Suggest we do it in the wg-serving repo - sub-project scope in proposals/, PR number in the name <PR_number>_proposal_details.md
    - [5m] Repo structure discussion - what expectations for proposals, docs, etc
        - Unblock execution on code
        - Use proposals to resolve design issues and have proposals that cover the core design
        - PoC PR - Please kick tires on dis
- Details
    - [15m] Roadmap review - Clayton to share a proposed roadmap of features and discuss them (at high level)
        - What features and objectives for MVP does everyone have?
        - Follow ups
            - Evan: What about traffic shadowing?
                - Clayton: good point, we have heard it's important and would hope the architecture either allows it via HTTPRoute or folks can get it to work somehow by MVP
            - Eduardo: Can we put the gateway in front of NIM servers?
                - Clayton: I expect so (for the LLM use case) even without LoRA, let's talk and take requirements
            - Mathis: Is there a long term plan to support a gateway to traditional models or multi-modal?
                - Clayton: believe we should
            - Clayton: need to better define the rollout workflows teams and users want, need feedback from consumers (assumption is that backends need to be incrementally and safely rolled out, and lora

adapters need to be progressively rolled out, both via traffic splitting)
    - API design
        - Discuss the current design and show difference between prototype and current state
        - Naming discussion
    - Architecture design
        - get proposals created as google docs and begin iterating towards a merged proposal
        - Include rationales for building default open implementation on top of Envoy in proposal
    - Principles discussion:
        - Continuously and progressively deployable
        - Address major architecture changes or alternatives when we hit the first fork
        - Backwards- and forwards-compatibility mindset
        - Generally shy away from smaller meetings, better to document in PRs, Google Docs, larger meetings(such as this one)
- Next meeting
    - Poll for time
    - Comment on PRs or discuss in wg-serving chat