# AI Gateway Working Group - Meeting Notes

## Template

<Do not edit>

### Jan 1, 2025

Meeting Host: <HOST_NAME>
Agenda

- Placeholder

AIs

- Placeholder

</Do not edit>

## Meeting Notes

### Oct 9, 2025

Meeting Host: Nir Rozenbaum
Agenda

- Moving us to the CNCF / Kubernetes Calendar and setting up YouTube sync?

AIs

- Placeholder

### Sep 29, 2025

Meeting Host: Shane Utt

- recording: https://zoom.us/rec/share/g2ZGP5RGNTVGhazxIr7JSAlcTjm0b251zbOOurIxUqUdrLOy5qZOQnNNZ9nr18e0.Erk9lHtCaO06zbPS

- summary: https://zoom-lfx.platform.linuxfoundation.org/meeting/97629695442-1759147200000/summaries?password=9772756e-0c2f-4d64-ac22-a3cbeb33859c
- transcript: https://zoom.us/rec/play/HmjRkQrjDDFHKLXsiexdJWbzVX5iR4X3h6zPXSTdEVLU-eQv800CIZpiNSHSkthNmQP8nJHYCO0uoCII.nr2zyaI4cG7aWukW?eagerLoadZvaPages=sidemenu.billing.plan_management&accessLevel=meeting&canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fzoom.us%2Frec%2Fshare%2Fg2ZGP5RGNTVGhazxIr7JSAlcTjm0b251zbOOurIxUqUdrLOy5qZOQnNNZ9nr18e0.Erk9lHtCaO06zbPS

## Attendance

- Shane Utt
- Sanjeev Rampal
- David Breitgand
- David Martin
- Nir Rozenbaum
- Huamin Chen
- Roderick Kieley

## Agenda

- [sanjeez] Define architectures in terms of Gateways: are there going to be multi-Gateway solutions for “AI Gateway”?
    - [shane] we should discuss, and we need to come up with our definitions
- [DavidB] Front he featureset, the current scope is all prescheduling and postscheduling decisions. Is scheduling involved?
    - [shane] scheduling, as in the GIE, is specifically out of scope (covered by WG Serving and GIE, whom we can work with if we want to influence some changes in scheduling)
- [nir] We should consider the use case of combining realtime inference and batch processing (offline inference). As of today these are two different WGs: Serving and Batch, they are working on different projects on different technologies but under the same umbrella. There’s no standardization here, and they’re both doing inference, but there are some systems which will utilize both use cases.
    - [nir] when I think of an “AI gateway” I need a *standard way* to make an inference request. I need this to be standard regardless of whether its a realtime or batch request.
    - [nir] using prompt guards as an example: its relevant for both types of requests: an incoming request for realtime, vs batch, both could run the same prompt guard. I would like to re-use these building blocks regardless.
    - [shane] +1
- [shane] “Full Request Processing?” “Body-Based Processing?” “Entire Contents… based… stuff?”
    - waf example in gwapi
    - proposal for how we do this - get the specification where it needs to be

    - [Huamin] concerns about standardising when things change/new use cases arise
        - “Semantic Processing”?
    - What other use cases are there outside AI?
    - It’s about the *entire* contents. We need to focus on that.
- [Sanjeev] identifying ‘extension blocks’ and the logical model first
    - [Shane] I’m in favor of not getting too much into ext-proc vs wasm vs others. focus on use cases. preference to distance ourselves from implementation if we can. Open to suggestion.
    - [Sanjeev] Actually, I’m more worried about: what is the simplest logical model that we can think about? Agree that we don’t want to focus on super opinionated implementations for now, but what’s our basic model?
    - [Nir] we do want to show something working eventually & prove this.
- [Huamin] Focus on the greater Linux and CNCF landscape, (I have a project called semantic-router under pytorch, for instance)
    - [shane] +1
- [Nir] I was focusing on the phrase “the intersection of AI and Networking”. Reading the agenda today, I noticed “envoy” and “ext-proc” already being said, maybe we should define them at a higher level? Gateway level?
    - [shane] HUGE +1
    - [nir] I am reminded of a request from NGINX for the GIE: they asked for a ext-proc compatibility layer. But, why do we need that? Why not just make this a natural part of the Gateway? I don’t want different Gateways to have to go mimic ext-proc, but rather have them implement things within the gateways.
        - [shane] +1
- [sanjeev] Should we just start with Google docs?
    - [shane] *personally* I prefer we work through an iterative process via PRs, with those PRs being smaller and focused as much as possible.
    - [nir] keep in mind our repository is a staging ground, we want to move fast with fast iterations, with minimal process.

## AIs

- [shane] Actual word definitions “AI Gateway”, “Inference Gateway”, etc at the Kubernetes documentation level.
- [shane] For our eventual proposals, we need to define some topologies and architectures, a logical model that is our starting point for APIs and implementation details.
- [nir] Open issue for standardizing inference APIs for online inference and offline (batch) - https://github.com/kubernetes-sigs/wg-ai-gateway/issues/4