
KubeCon
CloudNativeCon
North America 2024

KubeCon CloudNativeCon
North America 2024

Solving the Kubernetes Networking API Rubik's Cube

Lior Lieberman
Site Reliability Engineer @ Google
Ingress2gateway Maintainer
LiorLieberman

Doug Smith
Principal Software Engineer @ Red Hat
K8s Network Plumbing Working Group & CNI
@dougbtv

Surya Seetharaman
Principal Software Engineer @ Red Hat
SIG-Network Contributor
@tssurya

Shane Utt
Senior Principal Software Engineer @ Red Hat
SIG Network Chair / Gateway API Maintainer
shaneutt

Subprojects & Working Groups

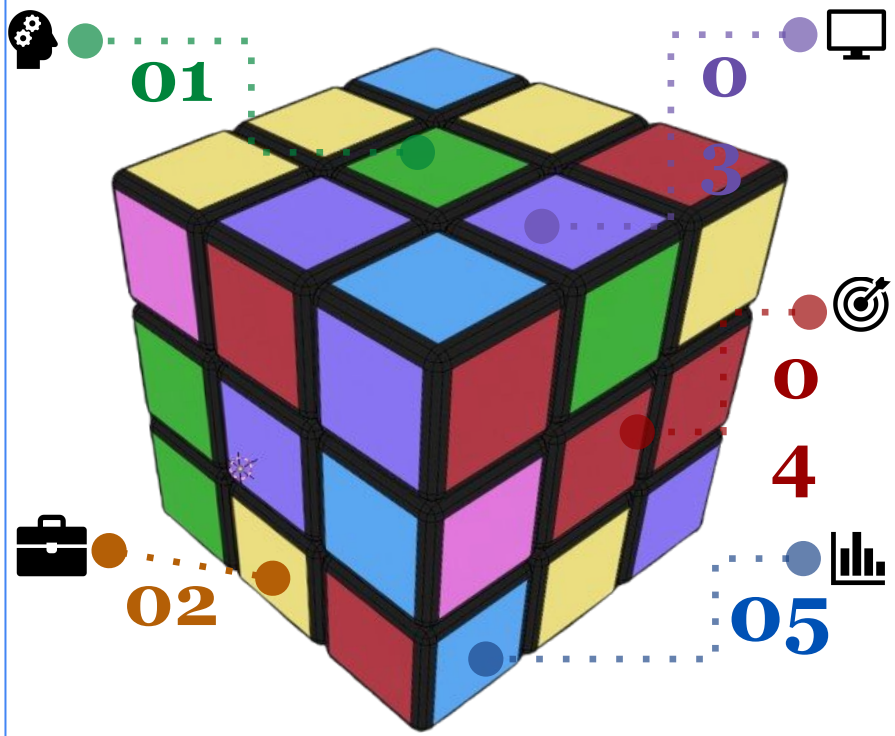
SIG-Network Subprojects

- Gateway API
- NetworkPolicy API
- Multi-Network

There's more!

Container Network Interface

Is a standard for configuring network interfaces for containers, and is how network providers plug their solutions into Kubernetes.



K8s Network Plumbing WG

Special Mention

WG Serving

LLM Instance Gateway
(Application Level)

Device Management WG

- Dynamic Resource Allocation (**DRA**)
  (Infrastructure Level)
- AI/ML + Networking

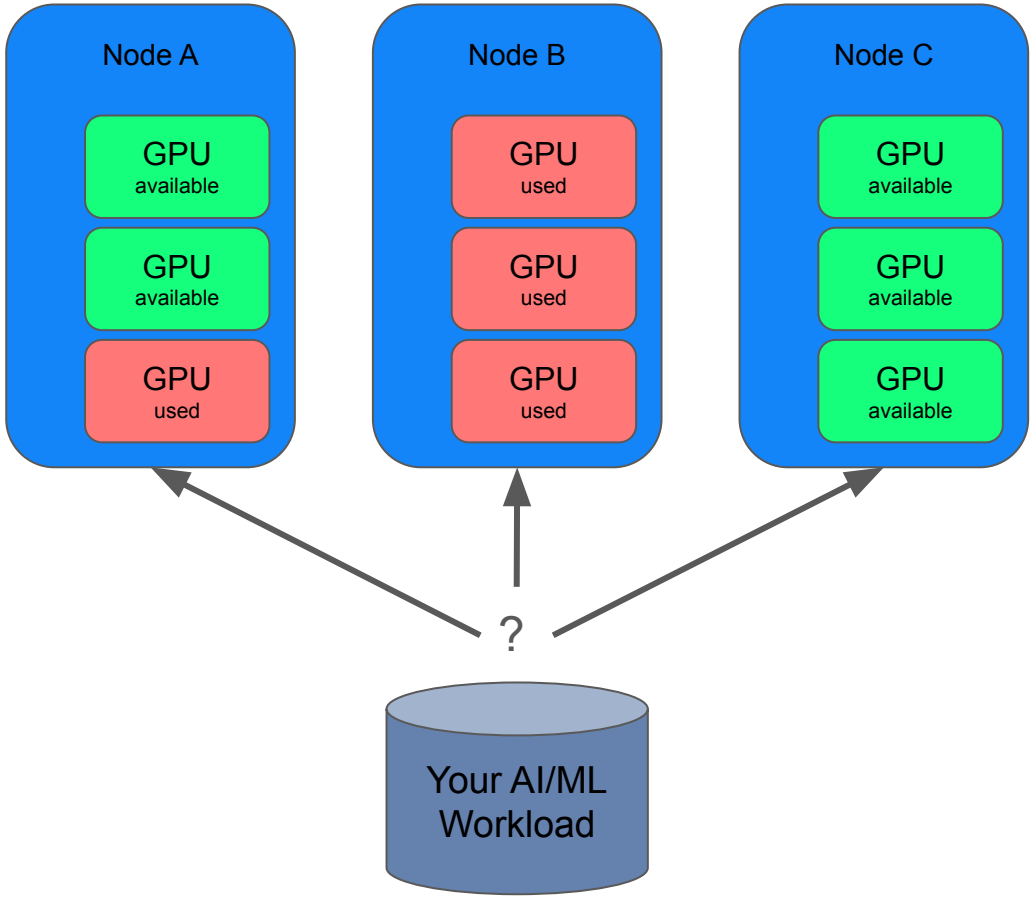
KubeCon

CloudNativeCon

North America 2024

AI + Networking at the Infrastructure level

DRA: What problem are we trying to solve?



Dynamic Resource Allocation (DRA) in Kubernetes

KubeCon CloudNativeCon North America 2024

What is DRA?

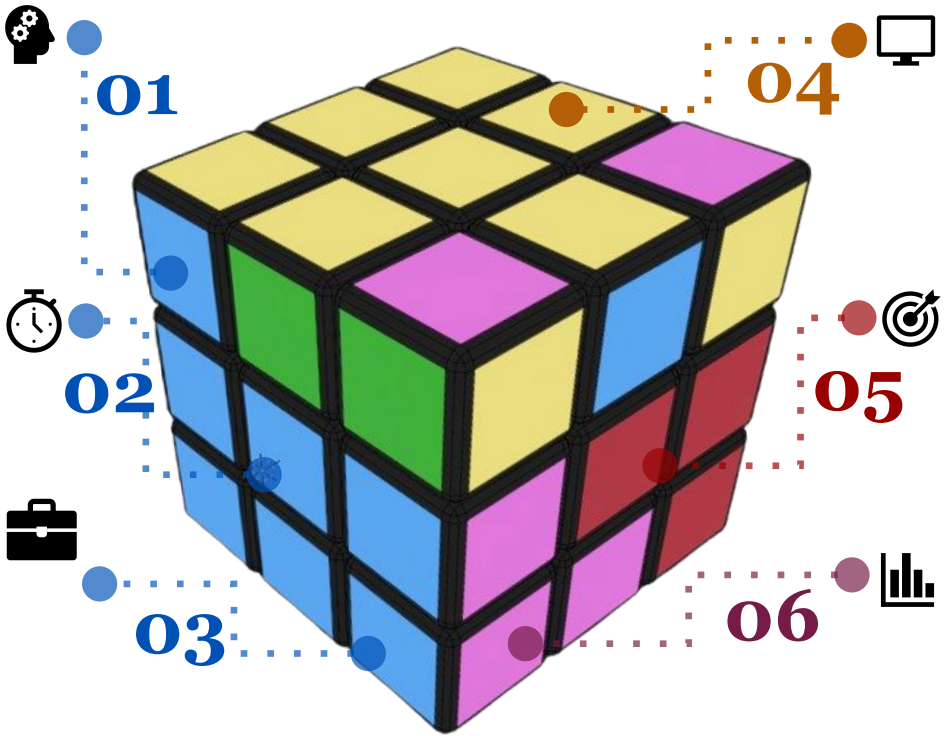
How to schedule workloads (pods) that rely on consumable hardware resources

Why DRA?

We need a richer way to schedule our pods to interact with hardware, especially for AI/ML

ResourceClaim Status

ResourceClaim Status KEP merged



GPU Allocation

The main target for DRA, and structured parameters for DRA, for richer allocation, including partial allocation

DRA for Networking

Which often have hardware resources associated with your nodes (SR-IOV with RDMA/Infiniband),

DRA + CNI PoC

Currently in flight! Looking for an implementation that gives feedback to DRA for Networking

Multi Networking in Kubernetes Ecosystem

KubeCon CloudNativeCon North America 2024

MultiHoming for Pods

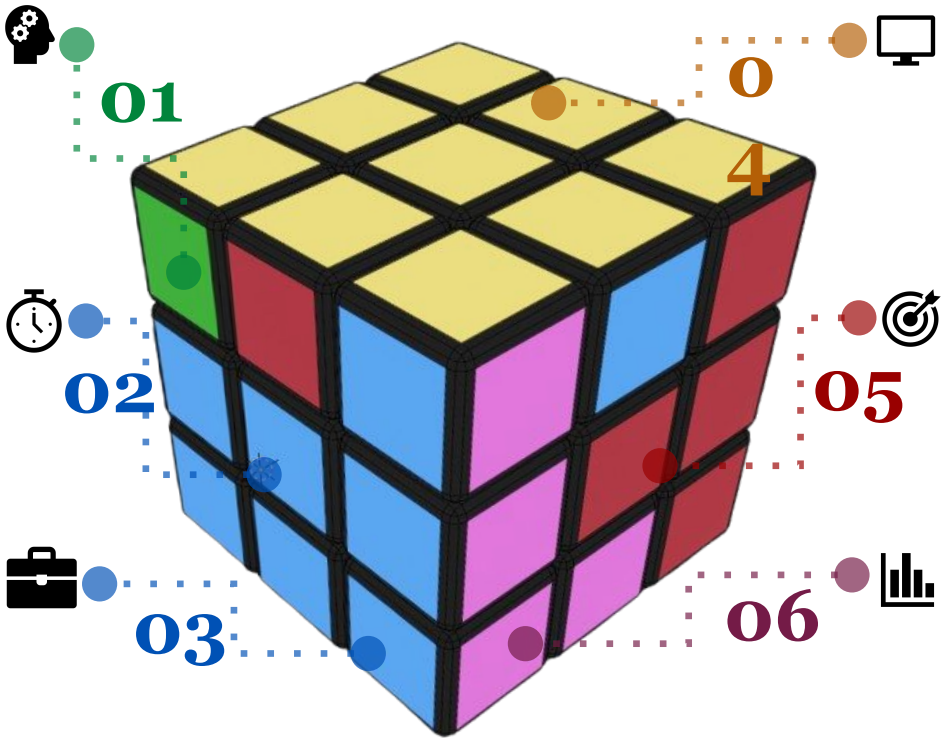
Ability to create additional networks for the additional interfaces on your pods

Multi Networks History

- sig-network (top bread slice)
- K8s NPWG (jam)
- Implementations of CNI (bottom bread slice)

Current Status

- sig-network-multi-network
  - KEP-3698 details



Use cases or thoughts from audience?

Audience Interaction Element

Multi Net aware K8s APIs

Support of Services, Gateways and NetworkPolicies on Multi Networks?

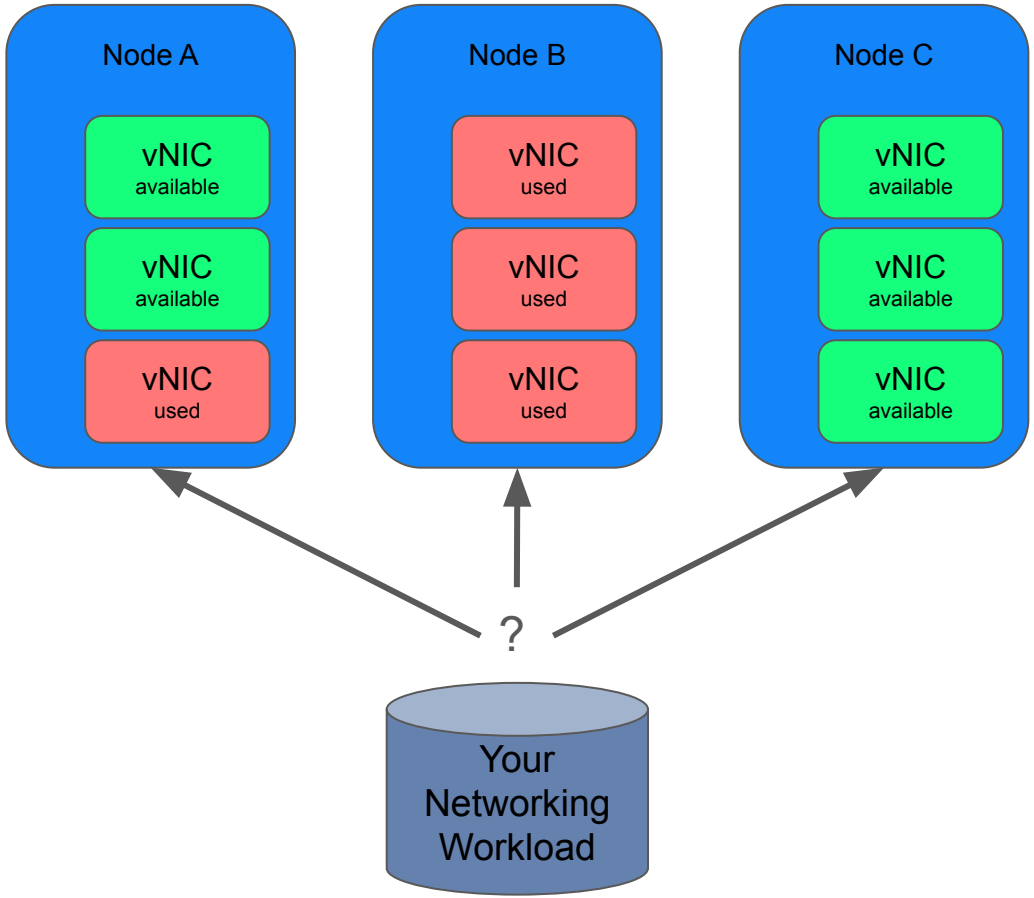
AI/ML Lossless Networks

Using multi-networks for achieving optimal, performant, secure network traffic for you LLM training

DRA: Dynamic Resource Allocation for Networking

How do DRA and Multi-Net help with AI use cases?

- Allocation of AI/ML pods for training and inference which rely on network communication at scale




KubeCon
CloudNativeCon
North America 2024
AI + Networking at the Application level

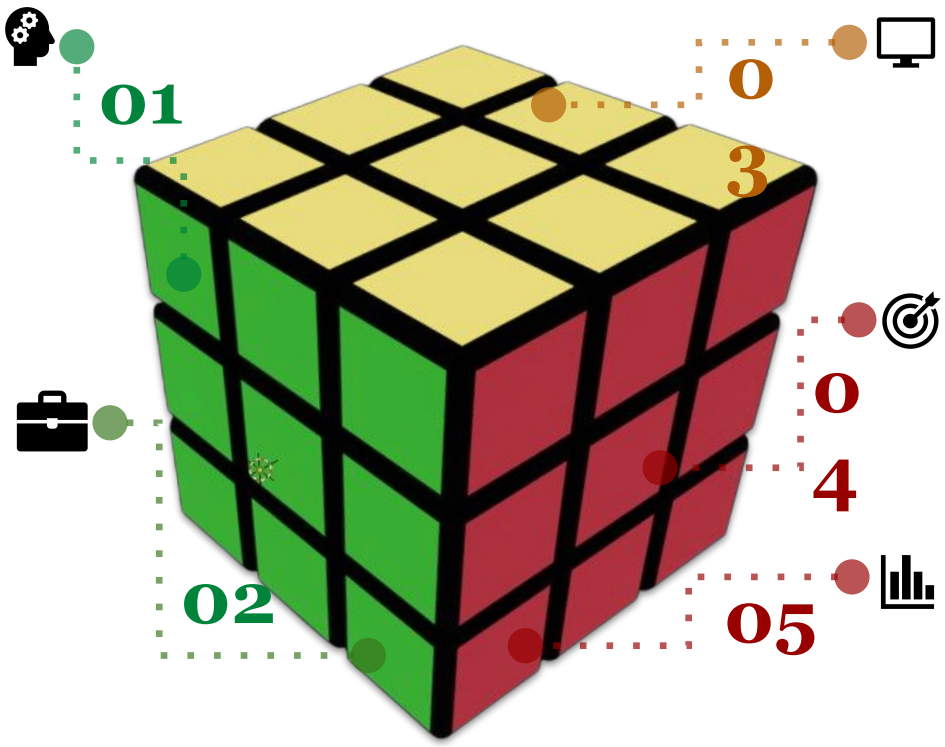
Gateway API and LLM Instance Gateway

What is Gateway API?

- Next Generation of **Ingress** API
- Provides Routing APIs
- Variety of Use Cases

Over 25 API Implementations!

- Bring your own implementation - no default



LLM Instance Gateway

Sponsored by WG Serving

Tailored AI Routing

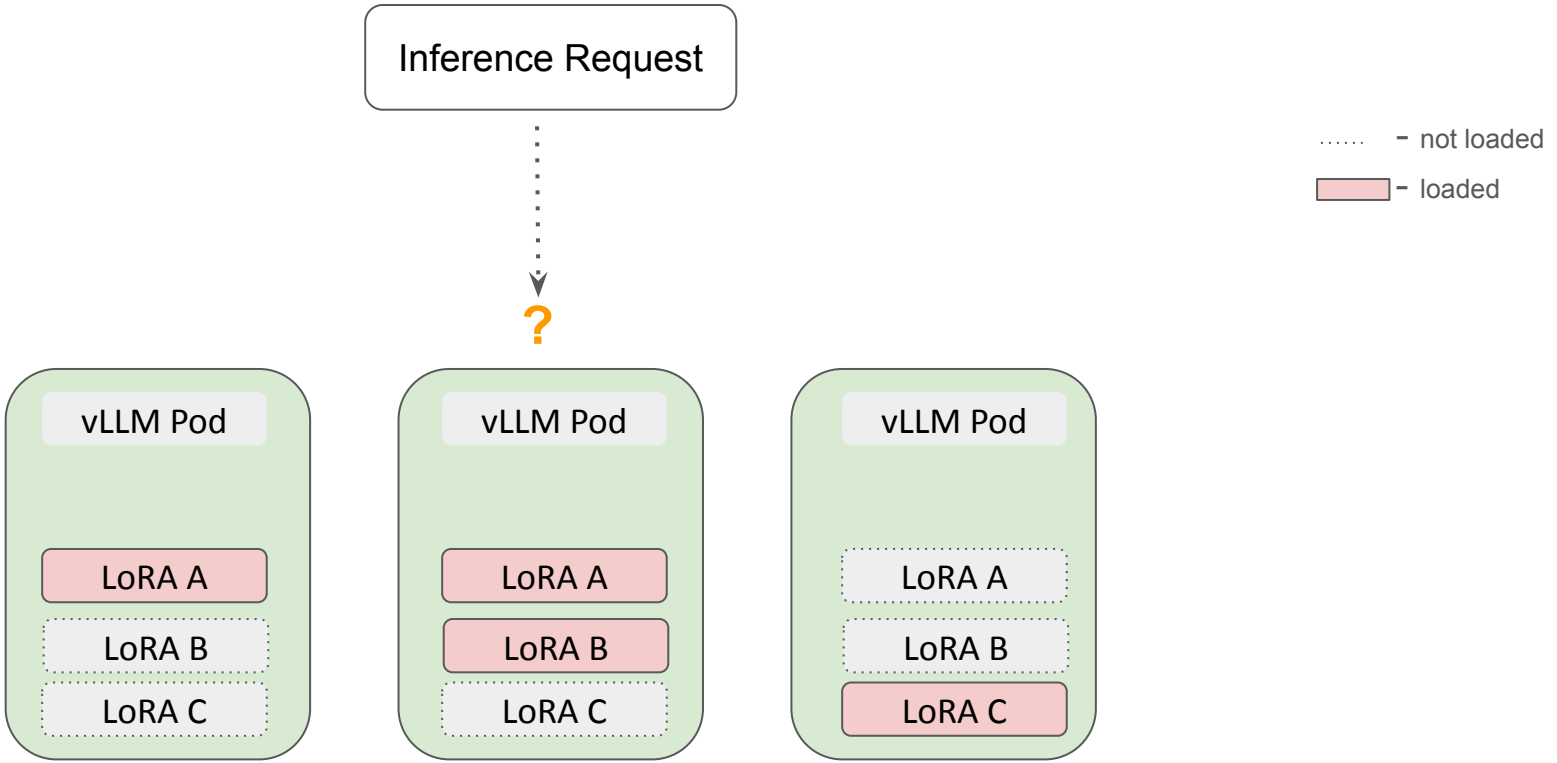
Semantic Routing
Configurable Operational Objectives
More Efficient Resource Utilization

Building with Gateway API

LLM Instance Gateway - LoRA

LoRA (Low Rank Adaptation) - layers of lightweight parameters that can be added to a larger model to specialize it and improve its performance for specific tasks. By enabling the larger model adaptation to those specific tasks, the adapters speed up and improve task-specific responses, and reduce the need to retrain new models for specific tasks.

LLM Instance Gateway - The Problem

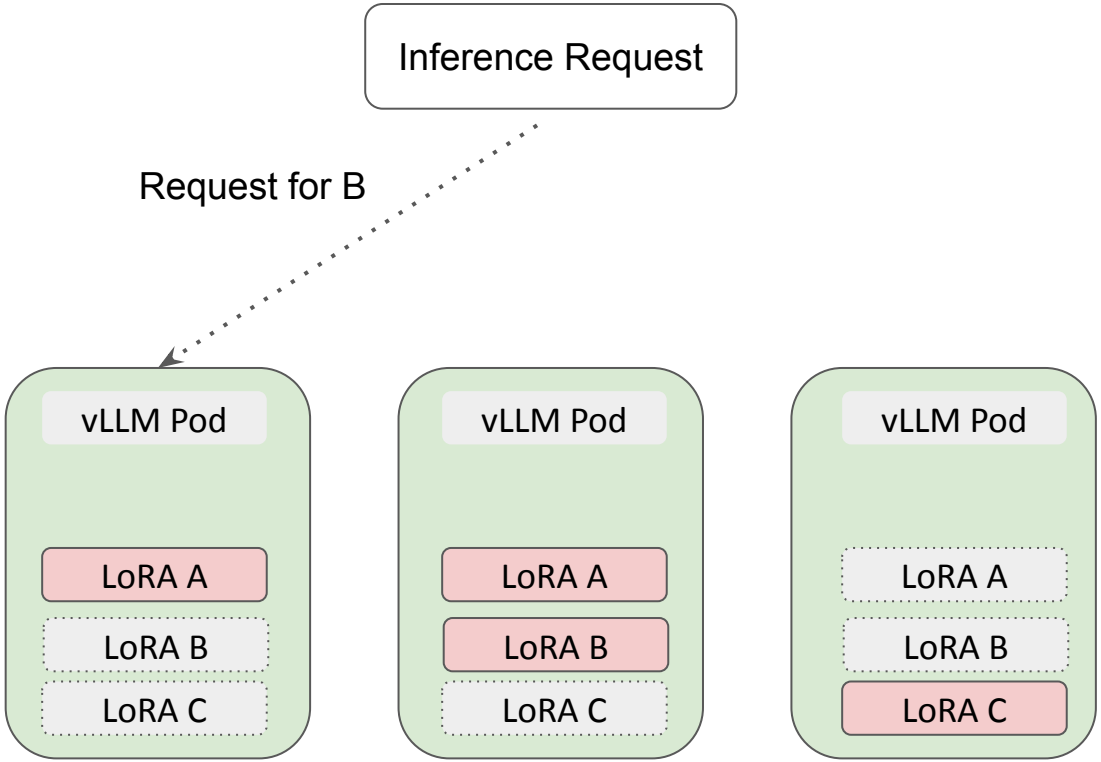


Drawing is based on the llm-instance-gateway proposal - https://github.com/kubernetes-sigs/wg-serving/tree/main/proposals/012-llm-instance-gateway#overview

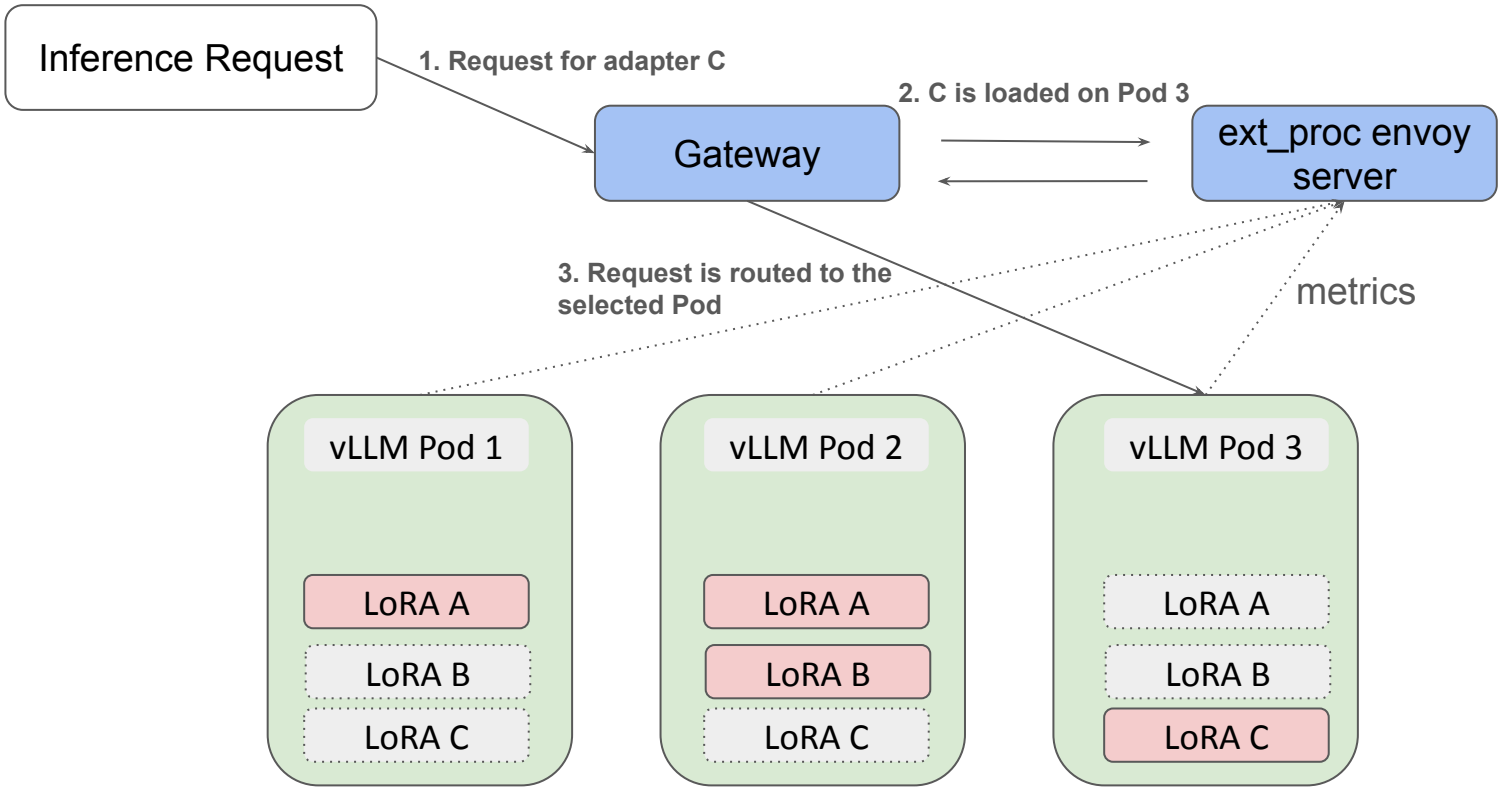
LLM Instance Gateway - The Problem

Request that lands on a backend without the requested adapter leads to:

- **Reduced Throughput**
- **Increased Latency**
- **Inefficient Hardware Utilization**



LLM Instance Gateway - Intelligent Routing



LLM Instance Gateway - Learn More


KubeCon
CloudNativeCon
North America 2024
What’s Next?

Looking to the future

- We'd like to see native Kubernetes APIs be the element that users interact with
  - And only the specific level APIs when you need to, like DRA
- *Everyone* interested in K8s networking technology should dig in and make sure that $upcoming_k8s_net_tech fits your needs!
  - Is it intuitive to you? It should be!
  - If not give your feedback, read KEPs, file a GitHub issues, make PRs
- Almost everything here is **new** and experimental
  - You can help shape it!
  - The more people involved, the more well-rounded the solution will end up
  - Want to know how you can get involved?

DRA KEPs

Get Involved!

- Subscribe to mailing lists
  - https://groups.google.com/g/kubernetes-sig-network
  - https://groups.google.com/a/kubernetes.io/g/wg-serving
  - https://groups.google.com/a/kubernetes.io/g/wg-device-management
- Join channels on https://kubernetes.slack.com
  - #sig-network, #sig-network-gateway-api, #wg-serving, #wg-device-management
- Meetings
  - https://github.com/kubernetes/community/blob/master/sig-network/README.md#meetings
  - https://github.com/kubernetes/community/blob/master/wg-serving/README.md#meetings
  - https://github.com/kubernetes/community/tree/master/wg-device-management#meetings
- Overwhelmed?
  - CNCF Mentoring (GSoC, LFX)

https://kubernetes.slack.com

https://contribute.cncf.io/contributors/

KubeCon | CloudNativeCon

North America 2024

Questions?

Leave feedback for our talk->

KubeCon | CloudNativeCon

North America 2024

Orange -> Red

Pink

Green

Blue

White -> Violet

yellow