AI and ML: the boring (yet critical) ops side

Rob Koch, Principal at Slalom Build
Milad Vafaeifard, Lead Software Engineer at Epam

TAG Contributor Strategy
DEAF & HARD OF HEARING
WORKING GROUP

Rob Koch

Principal at Slalom Build;
CNCF DHHWG Co-Chair

Connect with me on LinkedIn

Milad Vafaeifard

Lead Software Engineer at Epam;
CNCF DHHWG Member

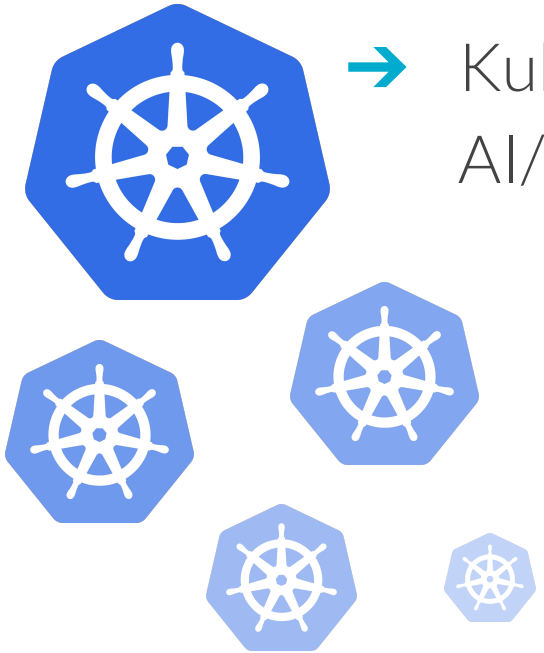
Connect with me on LinkedIn

Overview

Why the focus on Ops?

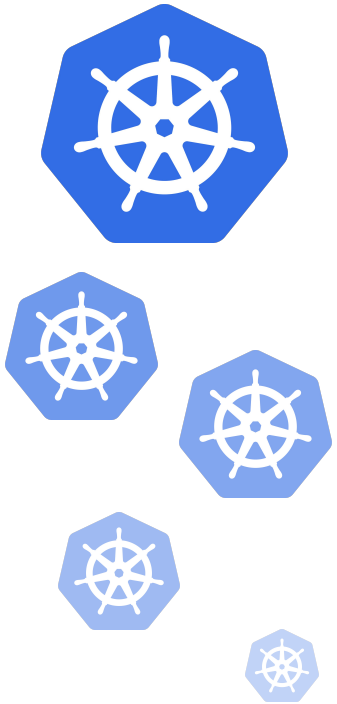
- **AI/ML**: Glamorous applications, but behind the scenes, they need strong infrastructure.
- **Critical aspects:** Compute resources, separation of data, reliability, observability.
- **Service meshes:** Simplify these challenges, allowing engineers to focus on AI/ML innovations.

What's Up with AI/ML in Kubernetes?



→ Kubernetes **orchestrates**
AI/ML workloads effectively

What's Up with AI/ML in Kubernetes?



Challenges with AI/ML workloads

- **Massive** compute and GPU requirements.
- **Large** datasets.
- **Isolation** of user data and processes.

How Does Kubernetes Make This Easy?

Strengths of Kubernetes

→ **Orchestration** & **scaling**

(using autoscalers/Karpenter)

How Does Kubernetes Make This Easy?

Strengths of Kubernetes

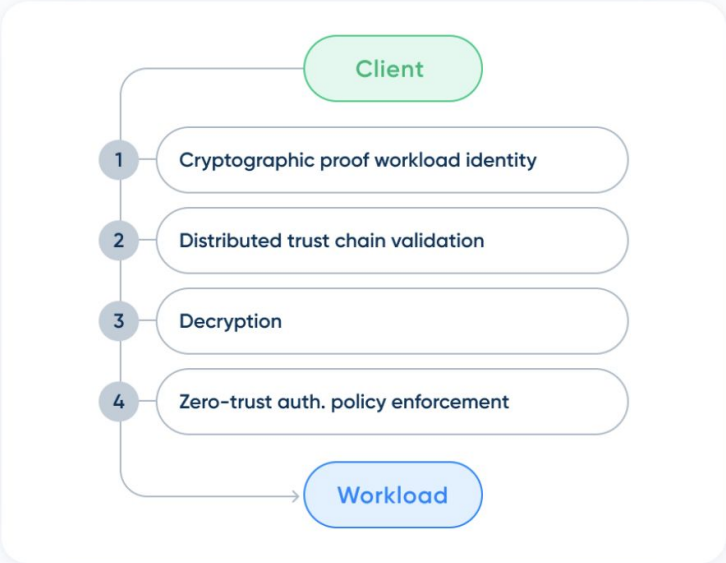
→ **Support** for GPUs and specialized hardware.

How Does Kubernetes Make This Easy?

Strengths of Kubernetes

→ **RBAC** (Role-Based Access Control) and strong **networking** features.

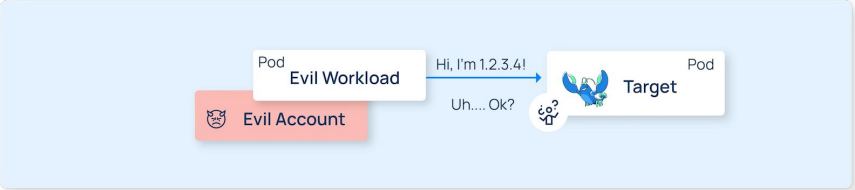
How Does Kubernetes NOT Make It Easy?



Kubernetes **does not** automatically provide

➔ Zero trust security

How Does Kubernetes NOT Make It Easy?



Kubernetes **does not** automatically provide

➔ Workload authentication and secure communications

How Does Kubernetes NOT Make It Easy?

Kubernetes **does not** automatically provide

- ➔ Dedicated hardware (like GPUs) can be tricky to manage

Why Linkerd?

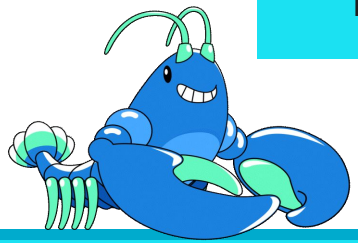
Operationally Simple

Micro proxy means no unnecessary overhead

Most Secure Service Mesh

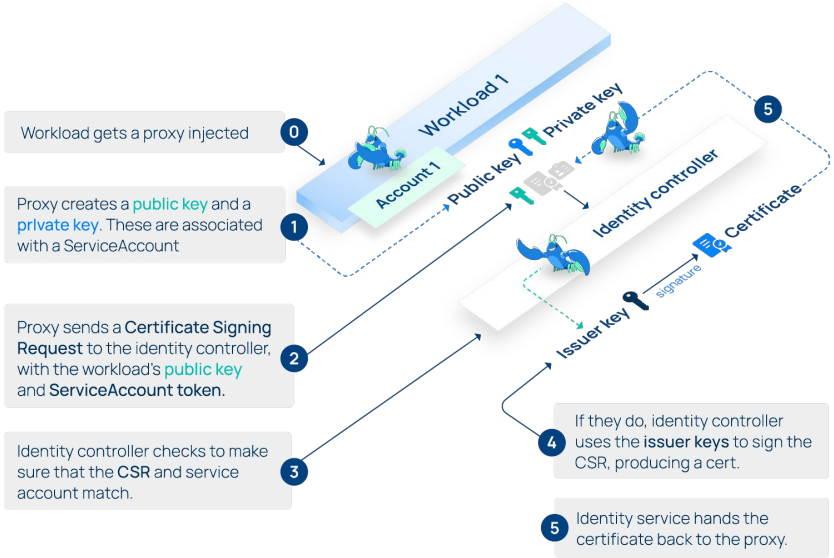
Written in Rust

Google's Chromium project found that 70% of serious security bugs are due to memory safety problems. Linkerd avoids these problems by using Rust.



Meshes to the Rescue

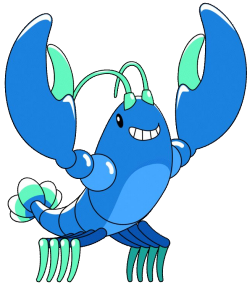
Service meshes like **Linkerd** provide



→ **Zero trust**: secure communications, workload authentication

Meshes to the Rescue

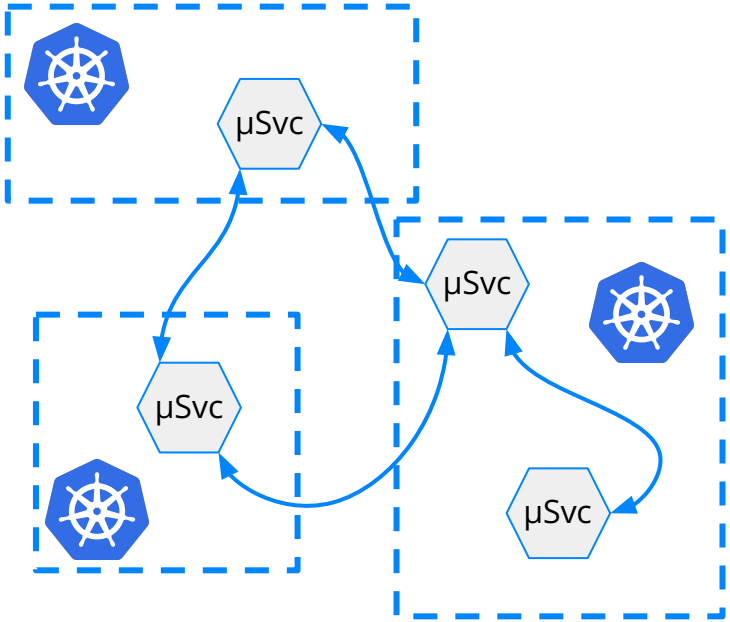
Service meshes like **Linkerd** provide



→ **Observability**: critical to understanding failures or performance issues.

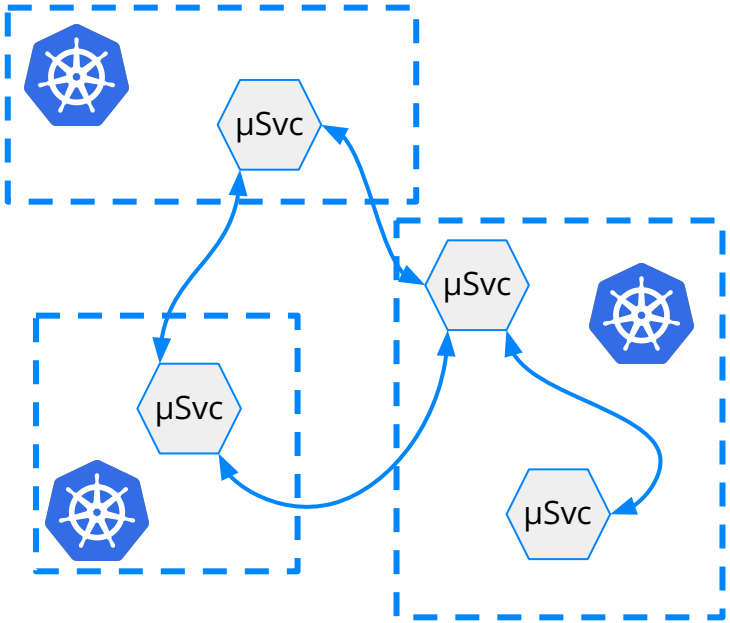
Meshes to the Rescue

Service meshes like **Linkerd** provide



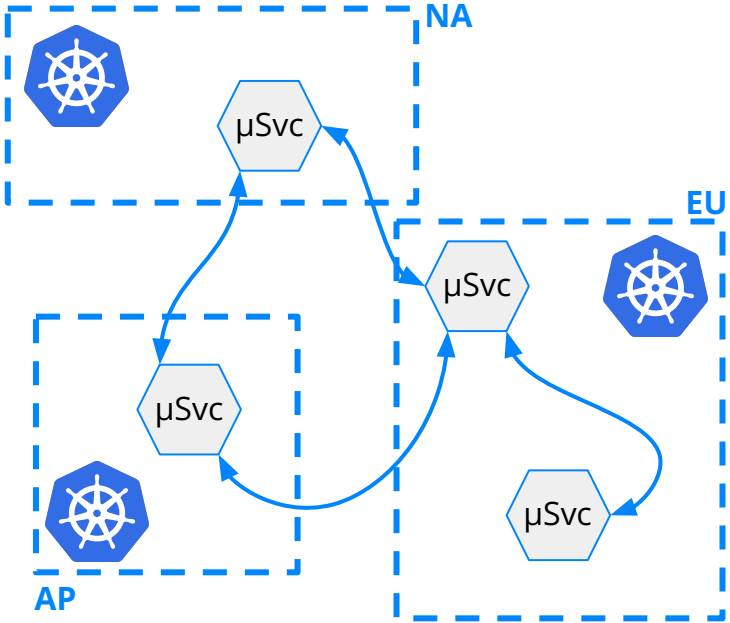
→ **Multi-cluster support**
for better hardware and data isolation.

Why Multicluster Matters



- ➔ Clusters **isolate** workloads and manage dedicated hardware **efficiently**
  - GPU clustering, **sensitive** datasets, **compliance** (GDPR, etc.)

Why Multicluster Matters



- Operational **partitioning**
  - Dev/test clusters vs. **production**
  - Geographical distribution for **compliance** or **cost savings**

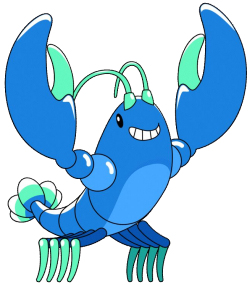
The Solution: Linkerd + Kubernetes

→ **Hands-on workshop:** Using Linkerd for cross-cluster communications

- Multicluster allows for multi-tenancy, IPv6 support, GPU workloads
- Starting point for deploying, managing, and debugging ML applications across clusters

BUOYANT SERVICE MESH ACADEMY

Benefits to AI/ML Ecosystem



- **Reliable** infrastructure for **critical** workloads.
- **Simplified** security, reliability, and observability.
- Key takeaway: Service meshes provide a **simplified** path to scaling and **securing** ML applications.

Conclusion

- ➔ Why *Ops Matter: Infrastructure is the **foundation** of AI/ML success
- ➔ Lean on service meshes to simplify the operational side and speed up AI/ML **innovation**

Recommended Reading

What is a Service Mesh?

buoyant.io/service-mesh-manifesto

mTLS Guide

buoyant.io/mtls-guide

Linkerd: Up & Running

oreilly.com/library/view/linkerd-up-and/9781098142308

Linkerd vs Istio

buoyant.io/linkerd-vs-istio

Why Linkerd doesn't use Envoy

linkerd.io/2020/12/03/why-linkerd-doesnt-use-envoy

Some pictures of DHH WG

Rob Koch

Principal at Slalom Build;

CNCF DHHWG Co-Chair

Connect with me on LinkedIn

Milad Vafaeifard

Lead Software Engineer at Epam;

CNCF DHHWG Member

Connect with me on LinkedIn