KUBERNETES
COMMUNITY DAYS CHINA 2022

使用 eBPF 代替 iptables 加速服务网格

DaoCloud 刘齐均

为何会有 **Merbridge** 项目？

- 社区一直在讨论，希望有一个基于 eBPF 的项目，为服务网格提供加速能力。
- 社区之前没有相关的开源产品。

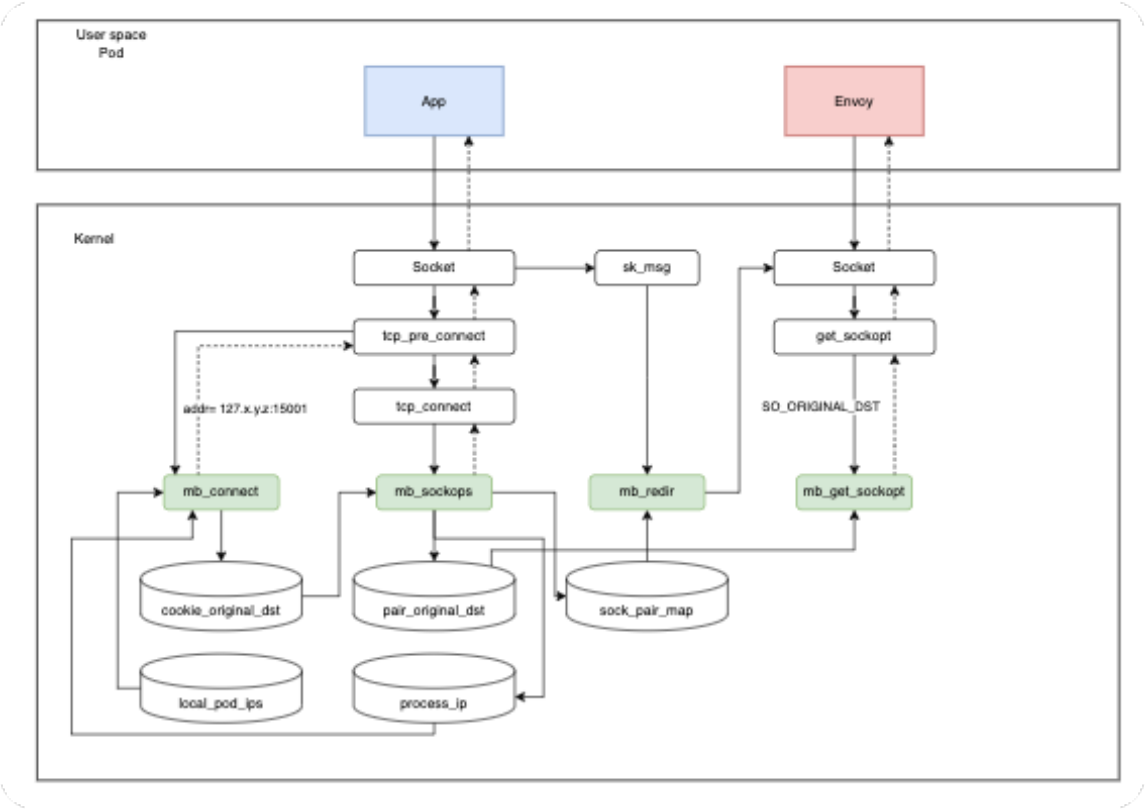
Merbridge 的目标是什么？

- 使用 eBPF 完全代替 iptables 实现应用加速。
- 不修改任何服务网格相关代码。
- 尽可能简单易用。
- 专注于服务网格领域的加速。

> 部分原理介绍

如何使用 eBPF 做流量拦截？

- 确定需要拦截的流量
  - 出口：通过检测当前 NS 是否存在 Sidecar 监听端口。
  - 入口：通过 tc 程序
- 怎么做拦截？
  - 修改目标地址为 127.0.0.1:15001
- 如何让 Sidecar 获取原始目标 IP 地址？
  - 通过 eBPF 修改 get_sockopt 函数处理 SO_ORIGINAL_DST.



四元组冲突问题

- Q：eBPF 是内核级别的，但是在网格场景中确实共享内核，在流量转发的路径中可能存在四元组冲突。

方案	缺点
修改 Destination IP，将其从 127.0.0.1 修改为 127.128.x.y, 其中 x, y 随着每次连接的建立而增加。	在 ipv6 下无法工作。
引入 socket cookie，为每个连接加上唯一标志，这样也可以避免冲突。	内核版本要求很高（5.15+）。
修改 sip，将 127.0.0.1 改为 Pod IP，这样也可以避免冲突。	需要能够识别 Pod IP。

如何在 eBPF 程序中获取 Pod IP？

- 1. 在 Pod 创建的时候通过 CNI 插件，在 Pod 的 NetNS 中监听一个特殊的端口。
- 2. 在为这个特殊的端口设置一个独立的 Mark。
- 3. 将这个独立的 Mark 和这个 Pod IP 存入一个 map。
- 4. eBPF 在运行的时候通过 lookup_tcp 可以获取到这个特殊端口的 socket。
- 5. 通过 socket 的 Mark 去 map 中查询当前 Pod 的 IP 地址即可。

> 能力介绍及 **RoadMap**

Merbridge 现有能力

- 在服务网格场景下，使用 eBPF 完整代替 iptables 的能力。
- 通过 sock redir 能力实现网路加速。
- 完全无侵入、无改造。
- 支持 Istio、Linkerd、Kuma 等主流的服务网格。
- 完整支持 Istio、Kuma 的所有能力，包括流量过滤规则等。
- 为 Kuma 网格带来 12% 的网络延迟降低。
- Ambient Mesh 模式支持（alpha），无惧任何 CNI 的影响。
- ……

RoadMap

- Ambient Mesh 支持（已 alpha）。
- 更低的版本要求（已规划）。
- 节点间加速。
- 双栈支持。
- ……

使用 eBPF 代替 iptables 加速服务网格

KUBERNETES
COMMUNITY DAYS CHINA 2022

> Demo

>

社区

网址: https://merbridge.io/zh/

项目地址：https://github.com/merbridge/merbridge

Slack 交流：
https://join.slack.com/t/merbridge/shared_invite/zt-11uc3z0w7-DMyv42eQ6s5YUxO5mZ5hwQ

KUBERNETES
COMMUNITY DAYS CHINA 2022

感谢观看

DaoCloud 刘齐均