KubeCon CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024


KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024
Redefining Service Mesh
Leveraging eBPF to Optimize Istio Ambient Architecture and Performance

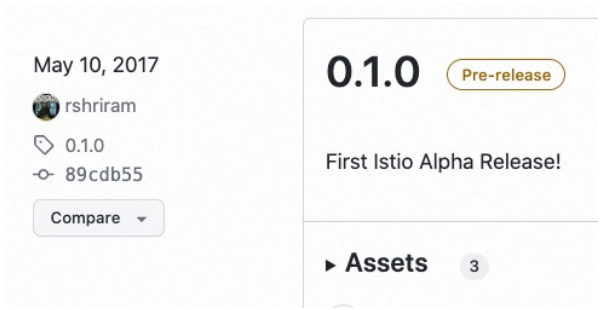
Speaker

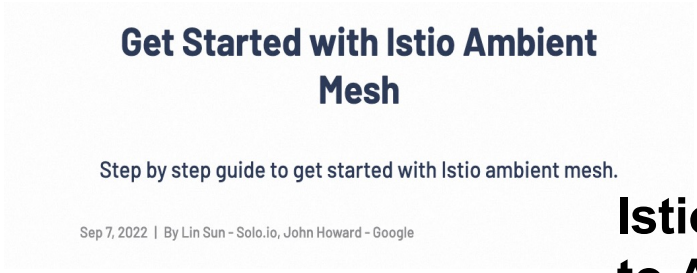
Yuxing Zeng

Technical Expert , Alibaba Cloud

Istio & Envoy member, has rich experiences in cloud native fields such as Kubernetes、 Networking、 Istio、 Envoy、 Nginx Ingress 、 CoreDNS, etc.

Istio Ambient 发展历程





2017

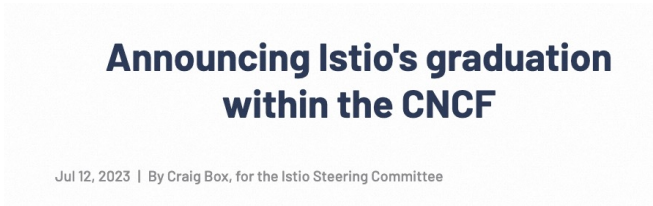
2020

Istio 1.5, Simplified Istio (istiod)

2022

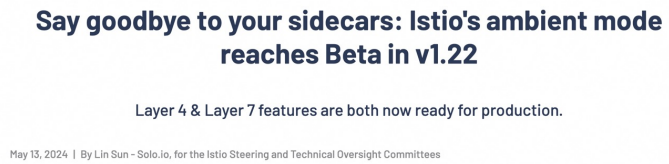
Istio 1.18, Ambient to Alpha

2023

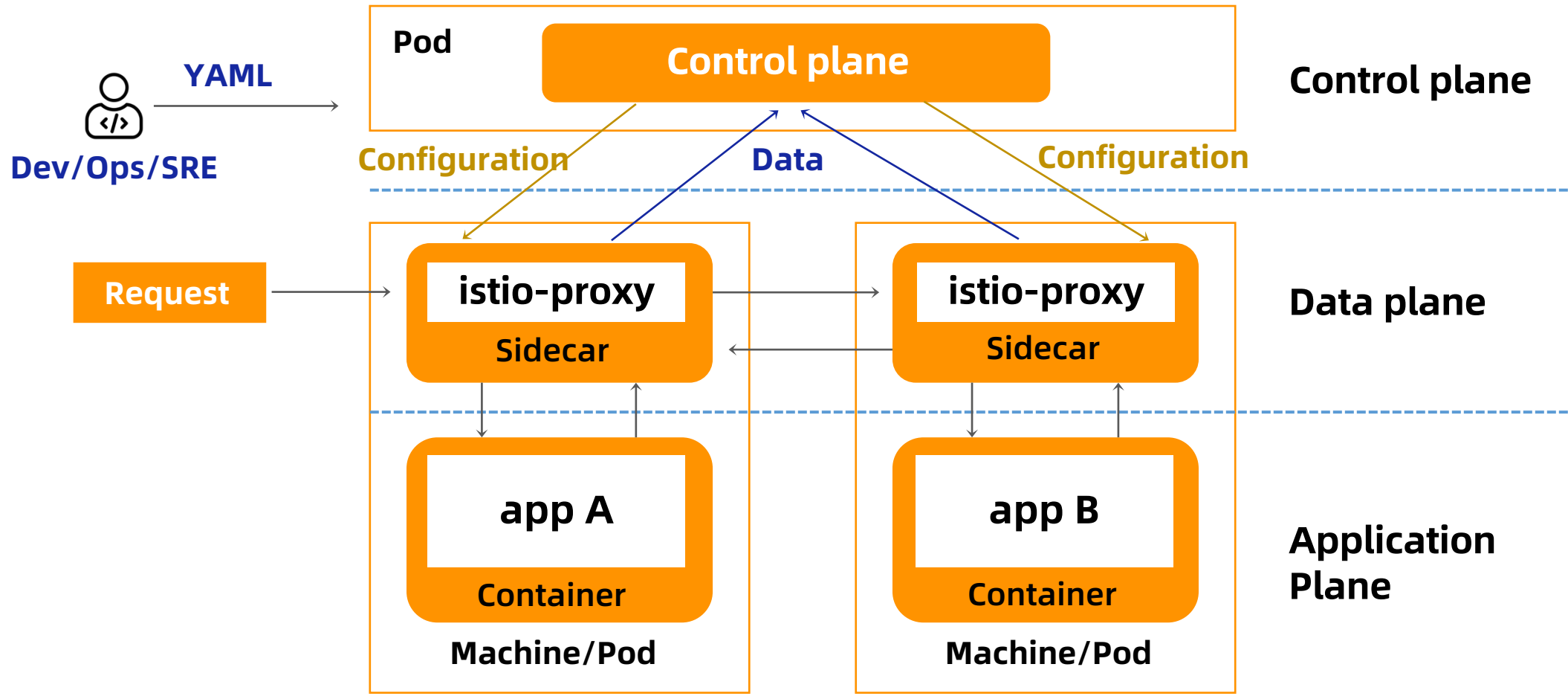


Istio 1.22, Ambient to Beta?

2024



Istio 数据面的演变过程 Sidecar -> Ambient



Sidecar 模式

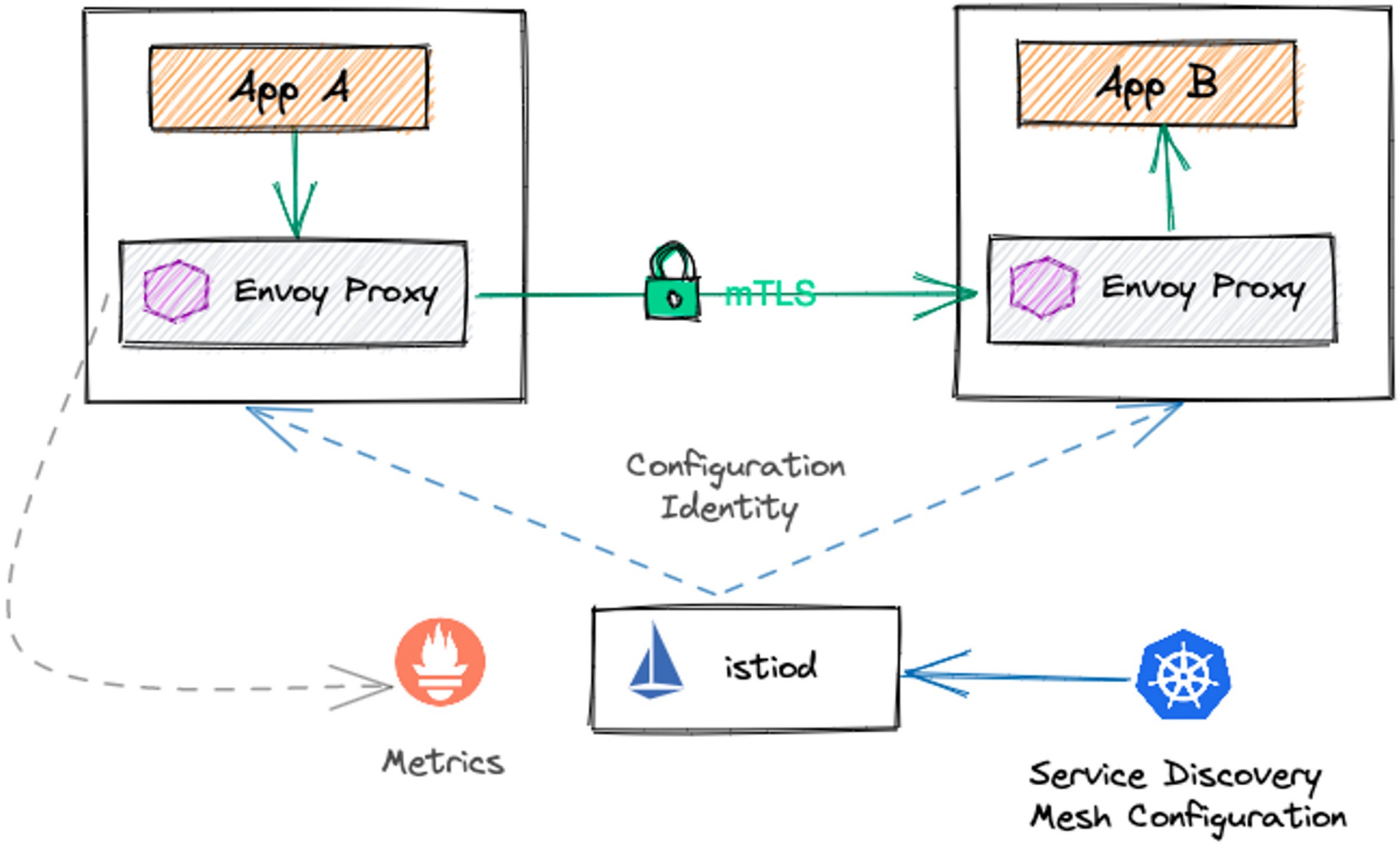
- 流量管理
- 安全
- 可观测性

Sidecar 模式的挑战

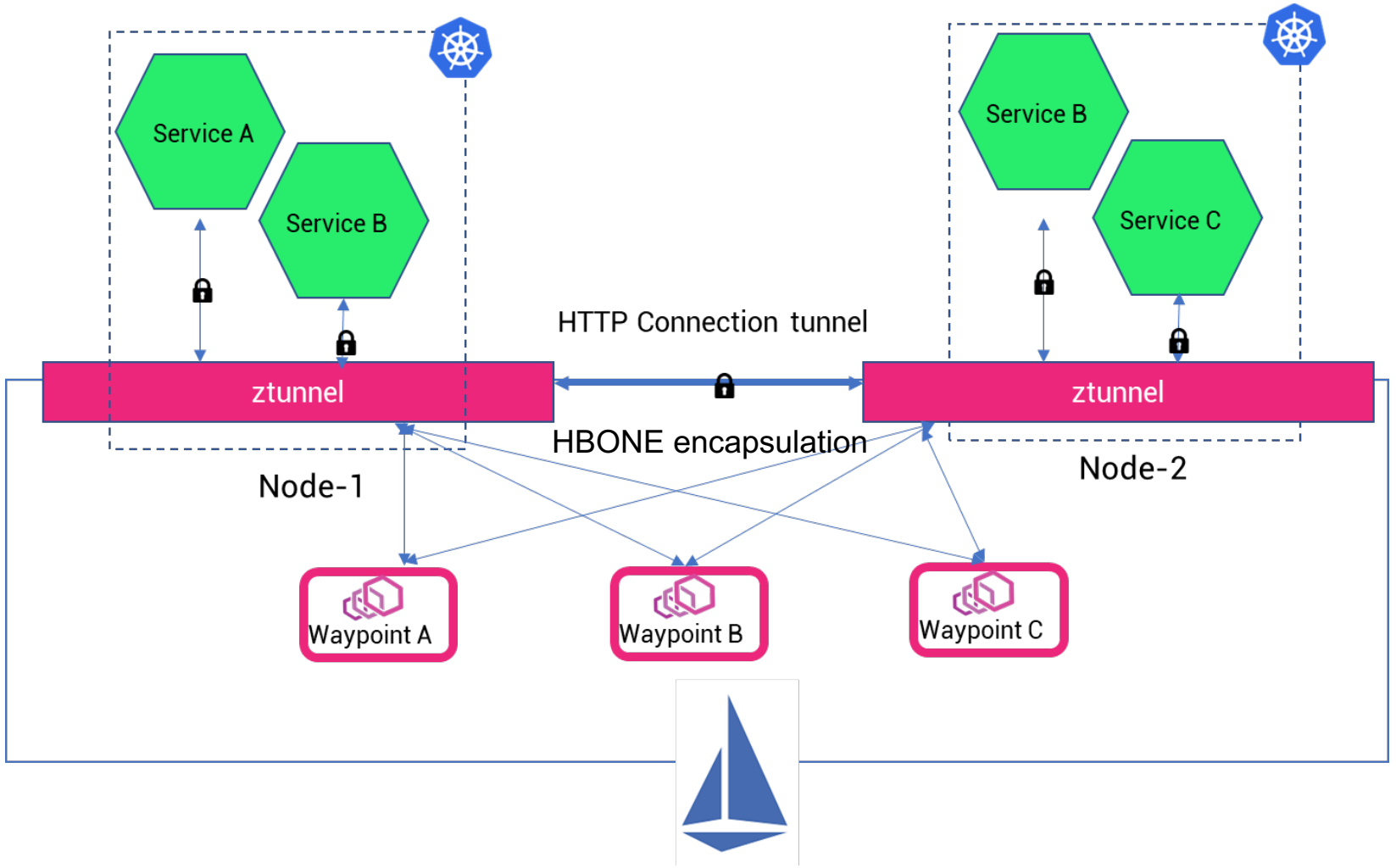
OPERATIONAL COMPLEXITY

OVERHEAD COST

PERFORMANCE



About Ambient



Istio Ambient: 一种新的数据平面模式

设计理念：将数据平面分层， 以此允许用户以更渐进增量的方式采用服务网格技术

7层高级处理：
功能丰富

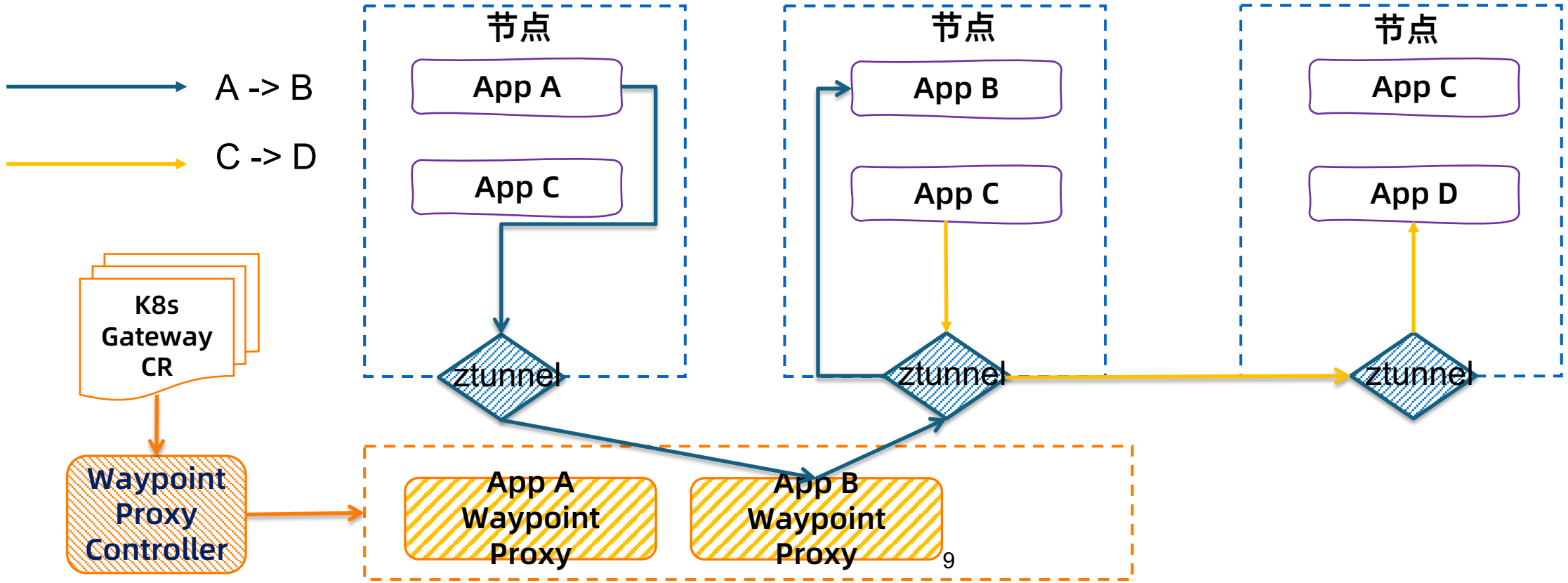
- 流量管理：HTTP路由、负载均衡、熔断、限流、故障容错、重试、超时等
- 安全：面向7层的精细化授权策略
- 可观测：HTTP监控指标、访问日志、链路追踪

4层基础处理：
低资源
高效率

- 流量管理：TCP路由
- 安全：面向4层的简单授权策略、双向TLS
- 可观测：TCP监控指标及日志

Istio 中业务应用程序和数据平面代理分离

Waypoint代理	• L7 组件完全独立于应用程序运行， 安全性更高；每个身份（Kubernetes 中的服务帐户）都有自己专用的 L7 代理，避免多租户 L7 代理模式引入的复杂度与不稳定性； • 通过K8s Gateway CRD定义触发启用；
ztunnel	将 L4处理下沉到 CNI级别，来自工作负载的流量被重定向到 ztunnel， 然后ztunnel识别工作负载并选择正确的证书来处理；
与Sidecar模式兼容	Sidecar模式仍然是网格的一等公民，可以与部署了 Sidecar 的工作负载进行本地通信；



Ambient L4 – Ztunnel

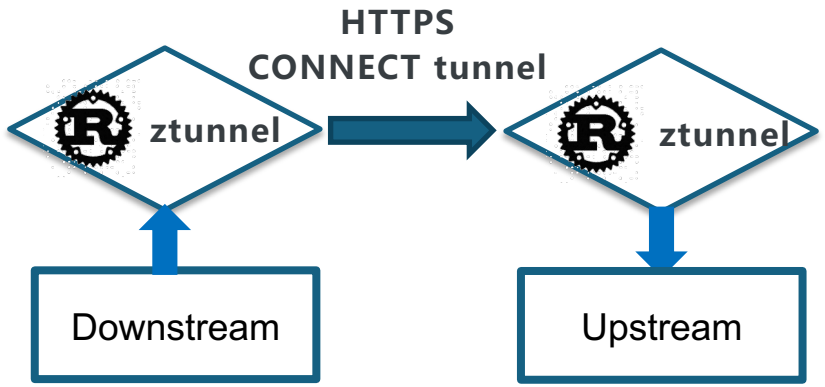
Ambient Mesh 下的流量路径

ztunnel提供L4流量代理能力

- ztunnel：用rust写成的新组件
- 同样通过XDS API从控制平面获取配置
- 只需要工作负载/服务ip、L4策略、工作负载证书配置，相对envoy配置更轻

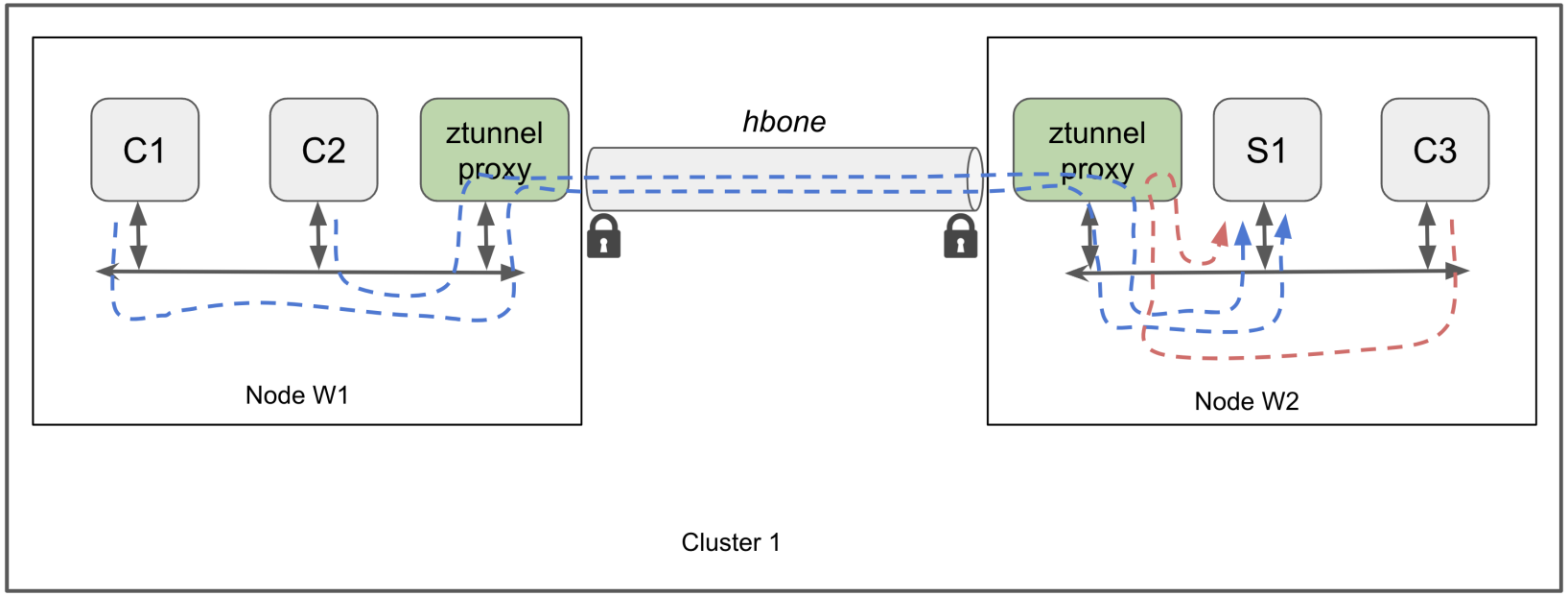
提供能力：

- 与其它节点的ztunnel通信时，使用HBONE协议，加密L4流量
- 保存Ambient Mesh中vip与pod id映射关系，实现简单的负载均衡
- 执行L4的授权策略



HBONE数据包组成

Ambient 已知问题



Ztunnel 的稳定性和性能：

- HA 问题： DamonSet 方式部署Ztunnel ，ztunnel pod 是单点，异常将导致整个节点业务流量受影响 https://github.com/istio/ztunnel/issues/40
- HBONE RTT 和 Ztunnel 连接池实现的性能问题
- 相关Issue: https://github.com/istio/istio/issues/48271 https://github.com/istio/istio/issues/48949

运维复杂度：

- HBONE 协议引入导致的运维的复杂度，而这些都是L4 、L7Envoy 应用层面需要感知的

Ambient 局限性- Upgrade

Upgrade with Helm

4 minute read page test

Follow this guide to upgrade and configure an ambient mode installation using Helm. This guide assumes you have already performed an ambient mode installation with Helm with a previous minor or patch version of Istio.

⚠ In contrast to sidecar mode, ambient mode supports moving application pods to an upgraded data plane without a mandatory restart or reschedule of running application pods. However, upgrading the data plane will briefly disrupt all workload traffic on the upgraded node and ambient mode does not currently support canary upgrades of the data plane.

Node cordoning and blue/green node pools are recommended to control blast radius of application pod traffic disruption during production upgrades. See your Kubernetes provider documentation for details.

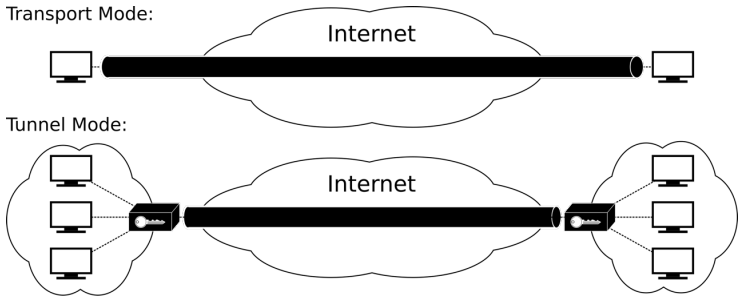
升级ztunnel 会导致对应节点上的工作负载流量中断

对于生产环境来说，这个是致命的

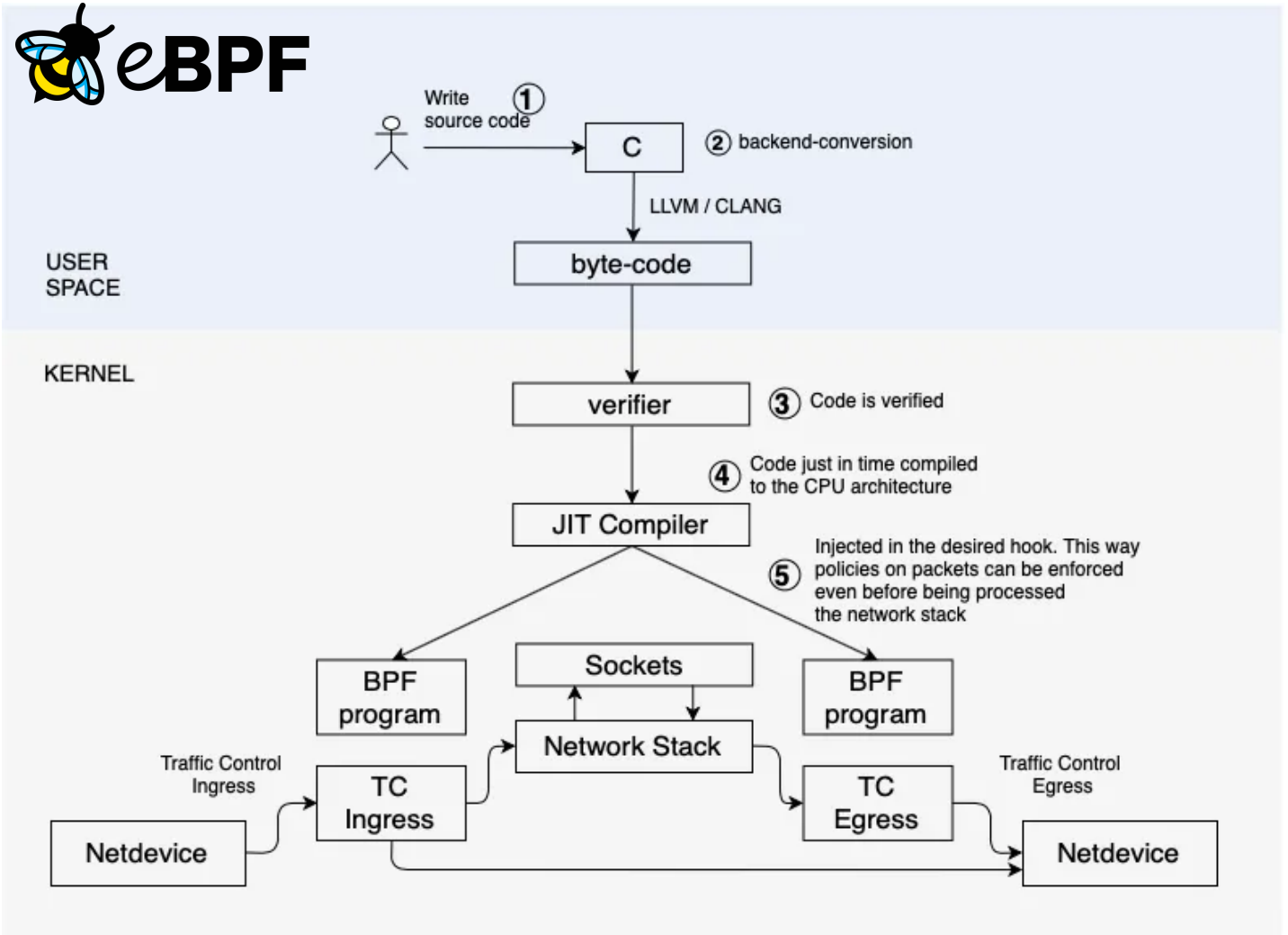
About Issue:
https://github.com/istio/istio/issues/51126

Ambient L4-Ztunnel 的替代实现 eBPF + IPSec

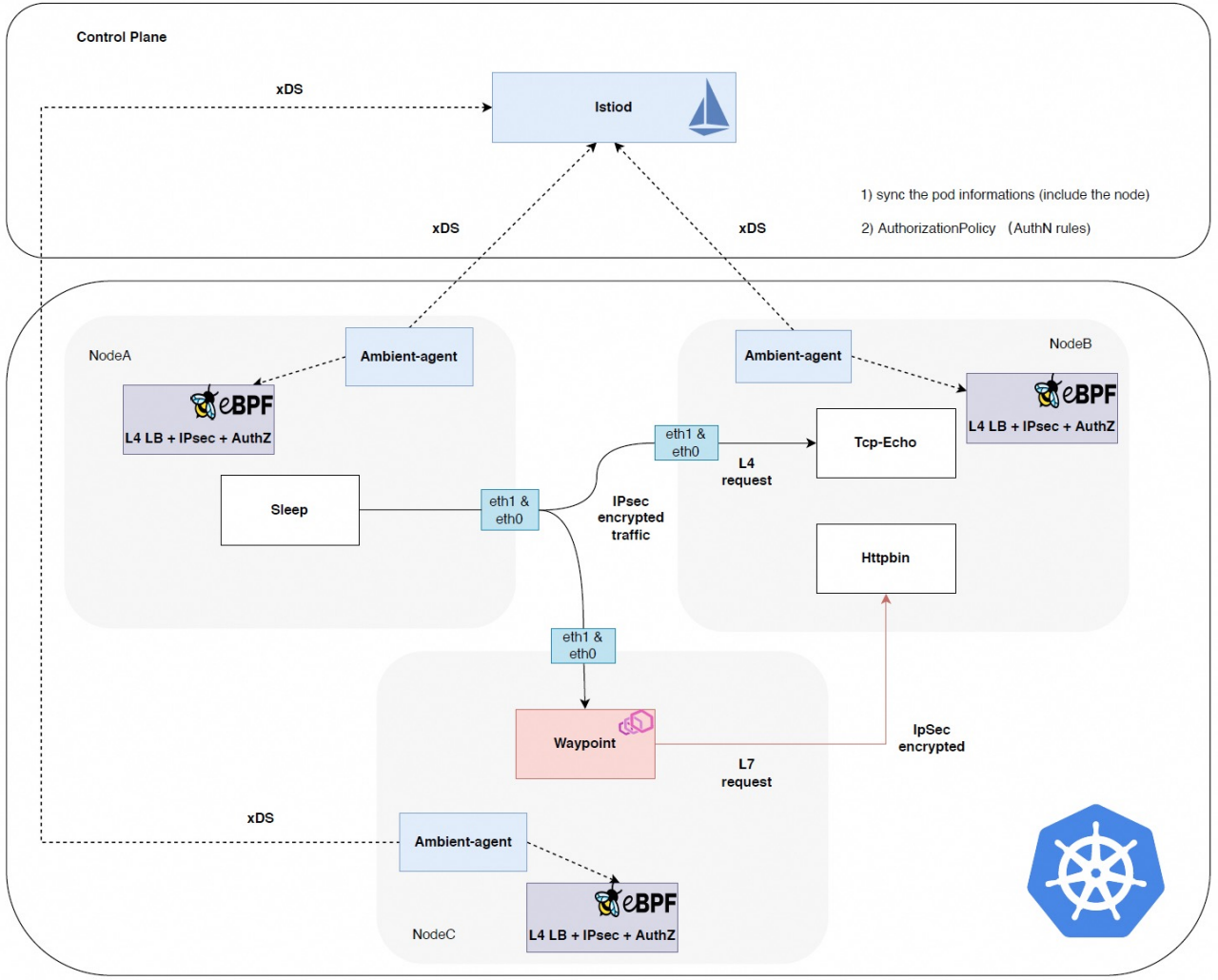
- eBPF LB 替代ztunnel LB 功能
- IPSec/WireGuard 替代 mTLS(HBONE)，IPSec/WireGuard 是一个Linux kernel 自带的功能



https://en.wikipedia.org/wiki/IPsec

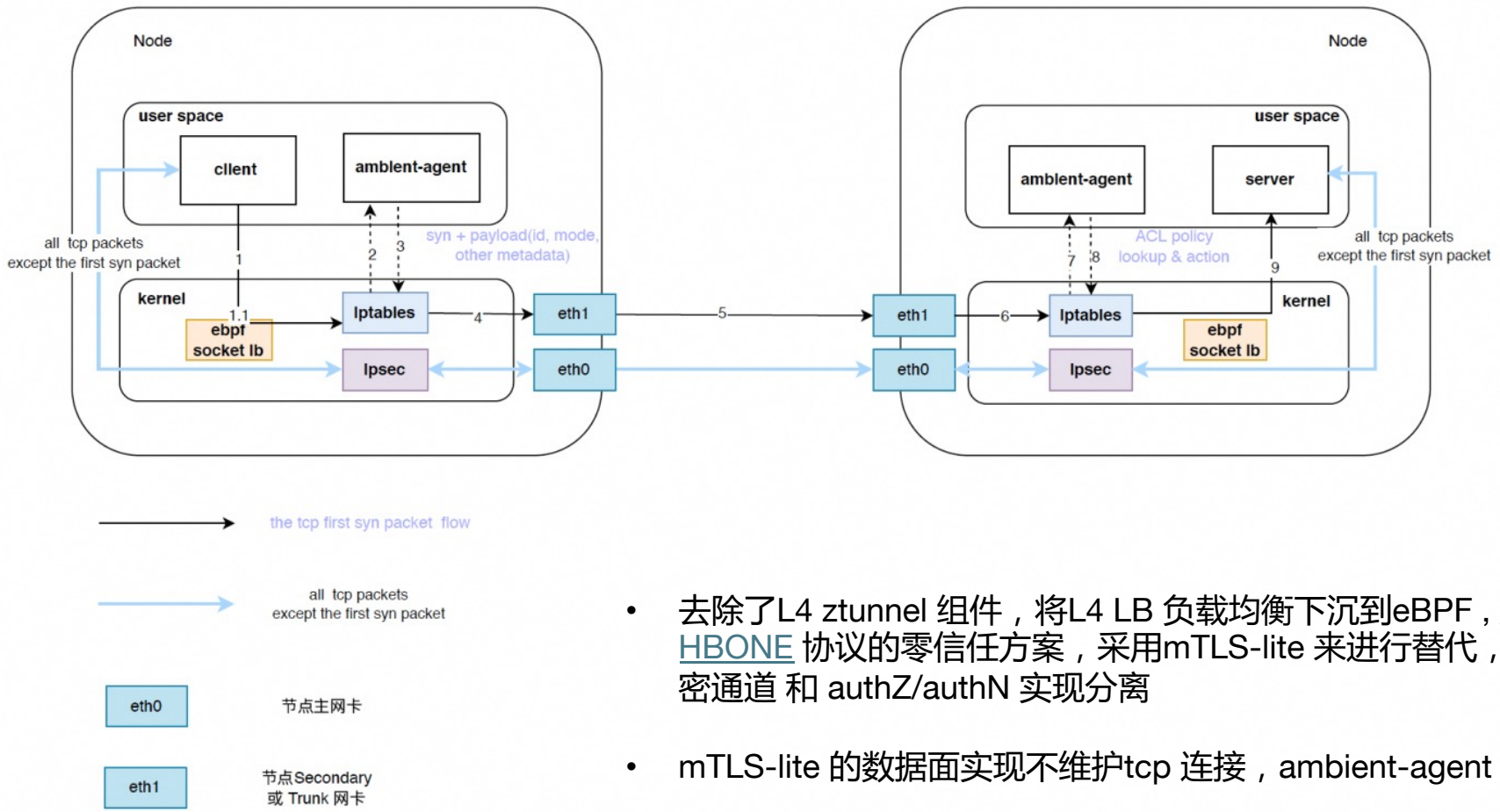


Ambient 优化-Mocket 方案



- eBPF L4 LB 来替代 Ztunnel 负载均衡能力
- IPSec + AuthZ 来替代HBONE，同时覆盖支持AuthorizationPolicy 相关功能
- Observability

Mocket 核心原理



- 去除了L4 ztunnel 组件，将L4 LB 负载均衡下沉到eBPF，原有的基于 HBONE 协议的零信任方案，采用mTLS-lite 来进行替代，mTLS-lite 将数据加密通道 和 authZ/authN 实现分离
- mTLS-lite 的数据面实现不维护tcp 连接，ambient-agent 升级对用户无感知。

Mocket 下对应Istiod 的适配修改


```yaml
apiVersion: gateway.networking.k8s.io/v1beta1
kind: Gateway
metadata:
  annotations:
    istio.io/for-service-account: test
  name: test
  namespace: default
spec:
  gatewayClassName: mocket-waypoint
  listeners:
    - name: serviceA
      port: 16001
      protocol: mTLS-lite
```

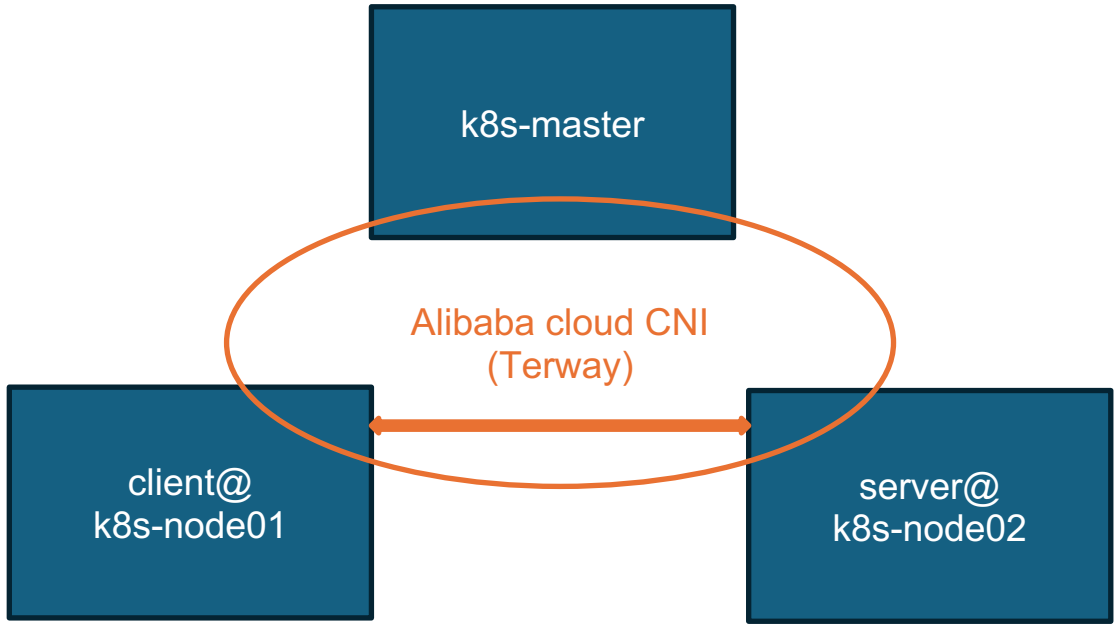
对应适配修改Istiod以支持mTLS-lite

KubeCon CloudNativeCon
THE LINUX FOUNDATION OPEN SOURCE SUMMIT
AI_dev Open Source GenAI & ML Summit
China 2024

Comparison of Performance

Testbed on Alibaba cloud ECS instances


```
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              8
On-line CPU(s) list: 0-7
Thread(s) per core:  2
Core(s) per socket:  4
Socket(s):           1
NUMA node(s):        1
Vendor ID:           GenuineIntel
BIOS Vendor ID:      Alibaba Cloud
CPU family:          6
Model:               143
Model name:          Intel(R) Xeon(R) Platinum 8475B
BIOS Model name:     pc-i440fx-2.1
Stepping:            8
CPU MHz:             3200.062
CPU max MHz:         3800.0000
CPU min MHz:         800.0000
BogoMIPS:            5400.00
Hypervisor vendor:   KVM
Virtualization type: full
L1d cache:           48K
L1i cache:           32K
L2 cache:            2048K
L3 cache:            99840K
NUMA node0 CPU(s):   0-7
```



Testbed on Alibaba cloud ECS instances


```
https://help.aliyun.com/document_detail/416274.html
Last login: Tue Dec  5 08:58:30 2023 from 192.102.204.53
[root@k8s-master ~]# lscpu
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              8
On-line CPU(s) list: 0-7
Thread(s) per core:  2
Core(s) per socket:  4
Socket(s):           1
NUMA node(s):        1
Vendor ID:           GenuineIntel
BIOS Vendor ID:      Alibaba Cloud
CPU family:          6
Model:               143
Model name:          Intel(R) Xeon(R) Platinum 8475B
BIOS Model name:     pc-i440fx-2.1
Stepping:            8
CPU MHz:             3200.000
CPU max MHz:         3800.0000
CPU min MHz:         800.0000
BogoMIPS:            5400.00
Hypervisor vendor:   KVM
Virtualization type: full
L1d cache:           48K
L1i cache:           32K
L2 cache:            2048K
L3 cache:            99840K
NUMA node0 CPU(s):   0-7
Flags:               fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat
 pse36 clflush mmx fxsr sse sse2 ss ht syscall nx pdpe1gb rdtscp lm constant_tsc rep_g
ood nopl xtopology nonstop_tsc cpuid aperfmperf tsc_known_freq pni pclmulqdq monitor s
sse3 fma cx16 pdcm pcid sse4_1 sse4_2 x2apic movbe popcnt aes xsave avx f16c rdrand hy
pervisor lahf_lm abm 3dnowprefetch cpuid_fault invpcid_single ibrs_enhanced fsgsbase t
sc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm avx512f avx512dq rdseed adx smap av
x512ifma clflushopt clwb avx512cd sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsa
ves avx_vnni avx512_bf16 wbnoinvd ida arat avx512vbmi umip pku ospke waitpkg avx512_vb
mi2 gfni vaes vpclmulqdq avx512_vnni avx512_bitalg avx512_vpopcntdq rdpid bus_lock_det
ect cldemote movdiri movdir64b fsrm uintr md_clear serialize tsxldtrk amx_bf16 avx512_
fp16 amx_tile amx_int8 arch_capabilities
[root@k8s-master ~]# free
               total        used        free      shared  buff/cache   available
Mem:       32072068     3324236    14894720        6040    13853112    28281332
Swap:             0           0           0
[root@k8s-master ~]# uname -r
5.10.134-15.al8.x86_64
[root@k8s-master ~]# lsb_release -a
LSB Version:    :core-4.1-amd64:core-4.1-noarch
Distributor ID: AlibabaCloud
Description:    Alibaba Cloud Linux release 3 (Soaring Falcon)
Release:        3
Codename:       SoaringFalcon
[root@k8s-master ~]#
```

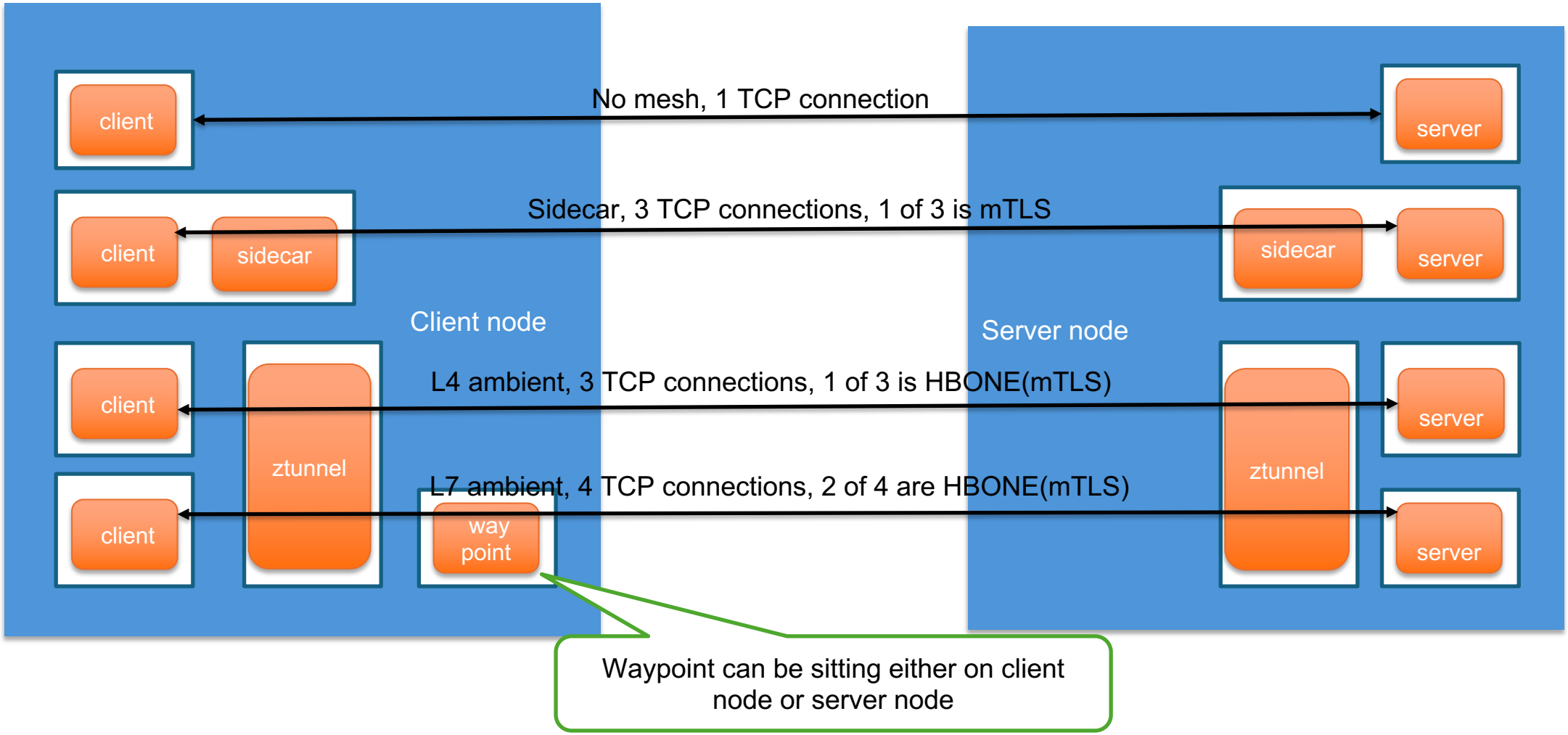
```
[root@k8s-node01 ~]#
[root@k8s-node01 ~]#
[root@k8s-node01 ~]# lscpu
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              8
On-line CPU(s) list: 0-7
Thread(s) per core:  2
Core(s) per socket:  4
Socket(s):           1
NUMA node(s):        1
Vendor ID:           GenuineIntel
BIOS Vendor ID:      Alibaba Cloud
CPU family:          6
Model:               143
Model name:          Intel(R) Xeon(R) Platinum 8475B
BIOS Model name:     pc-i440fx-2.1
Stepping:            8
CPU MHz:             3200.000
CPU max MHz:         3800.0000
CPU min MHz:         800.0000
BogoMIPS:            5400.00
Hypervisor vendor:   KVM
Virtualization type: full
L1d cache:           48K
L1i cache:           32K
L2 cache:            2048K
L3 cache:            99840K
NUMA node0 CPU(s):   0-7
Flags:               fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat
 pse36 clflush mmx fxsr sse sse2 ss ht syscall nx pdpe1gb rdtscp lm constant_tsc rep_g
ood nopl xtopology nonstop_tsc cpuid aperfmperf tsc_known_freq pni pclmulqdq monitor s
sse3 fma cx16 pdcm pcid sse4_1 sse4_2 x2apic movbe popcnt aes xsave avx f16c rdrand hy
pervisor lahf_lm abm 3dnowprefetch cpuid_fault invpcid_single ibrs_enhanced fsgsbase t
sc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm avx512f avx512dq rdseed adx smap av
x512ifma clflushopt clwb avx512cd sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsa
ves avx_vnni avx512_bf16 wbnoinvd ida arat avx512vbmi umip pku ospke waitpkg avx512_vb
mi2 gfni vaes vpclmulqdq avx512_vnni avx512_bitalg avx512_vpopcntdq rdpid bus_lock_det
ect cldemote movdiri movdir64b fsrm uintr md_clear serialize tsxldtrk amx_bf16 avx512_
fp16 amx_tile amx_int8 arch_capabilities
[root@k8s-node01 ~]# free
               total        used        free      shared  buff/cache   available
Mem:       32072064     2107232    19940256        4160    10024576    29500212
Swap:             0           0           0
[root@k8s-node01 ~]# uname -r
5.10.134-15.al8.x86_64
[root@k8s-node01 ~]# lsb_release -a
LSB Version:    :core-4.1-amd64:core-4.1-noarch
Distributor ID: AlibabaCloud
Description:    Alibaba Cloud Linux release 3 (Soaring Falcon)
Release:        3
Codename:       SoaringFalcon
[root@k8s-node01 ~]#
```

```
https://help.aliyun.com/document_detail/416274.html
Last login: Mon Dec  4 14:56:32 2023 from 10.1.0.205
[root@k8s-node02 ~]# lscpu
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              8
On-line CPU(s) list: 0-7
Thread(s) per core:  2
Core(s) per socket:  4
Socket(s):           1
NUMA node(s):        1
Vendor ID:           GenuineIntel
BIOS Vendor ID:      Alibaba Cloud
CPU family:          6
Model:               143
Model name:          Intel(R) Xeon(R) Platinum 8475B
BIOS Model name:     pc-i440fx-2.1
Stepping:            8
CPU MHz:             3200.000
CPU max MHz:         3800.0000
CPU min MHz:         800.0000
BogoMIPS:            5400.00
Hypervisor vendor:   KVM
Virtualization type: full
L1d cache:           48K
L1i cache:           32K
L2 cache:            2048K
L3 cache:            99840K
NUMA node0 CPU(s):   0-7
Flags:               fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat
 pse36 clflush mmx fxsr sse sse2 ss ht syscall nx pdpe1gb rdtscp lm constant_tsc rep_g
ood nopl xtopology nonstop_tsc cpuid aperfmperf tsc_known_freq pni pclmulqdq monitor s
sse3 fma cx16 pdcm pcid sse4_1 sse4_2 x2apic movbe popcnt aes xsave avx f16c rdrand hy
pervisor lahf_lm abm 3dnowprefetch cpuid_fault invpcid_single ibrs_enhanced fsgsbase t
sc_adjust bmi1 hle avx2 smep bmi2 erms invpcid rtm avx512f avx512dq rdseed adx smap av
x512ifma clflushopt clwb avx512cd sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsa
ves avx_vnni avx512_bf16 wbnoinvd ida arat avx512vbmi umip pku ospke waitpkg avx512_vb
mi2 gfni vaes vpclmulqdq avx512_vnni avx512_bitalg avx512_vpopcntdq rdpid bus_lock_det
ect cldemote movdiri movdir64b fsrm uintr md_clear serialize tsxldtrk amx_bf16 avx512_
fp16 amx_tile amx_int8 arch_capabilities
[root@k8s-node02 ~]# free
               total        used        free      shared  buff/cache   available
Mem:       32072068     2717872    16686192        5076    12668004    28888660
Swap:             0           0           0
[root@k8s-node02 ~]# uname -r
5.10.134-15.al8.x86_64
[root@k8s-node02 ~]# lsb_release -a
LSB Version:    :core-4.1-amd64:core-4.1-noarch
Distributor ID: AlibabaCloud
Description:    Alibaba Cloud Linux release 3 (Soaring Falcon)
Release:        3
Codename:       SoaringFalcon
[root@k8s-node02 ~]#
```

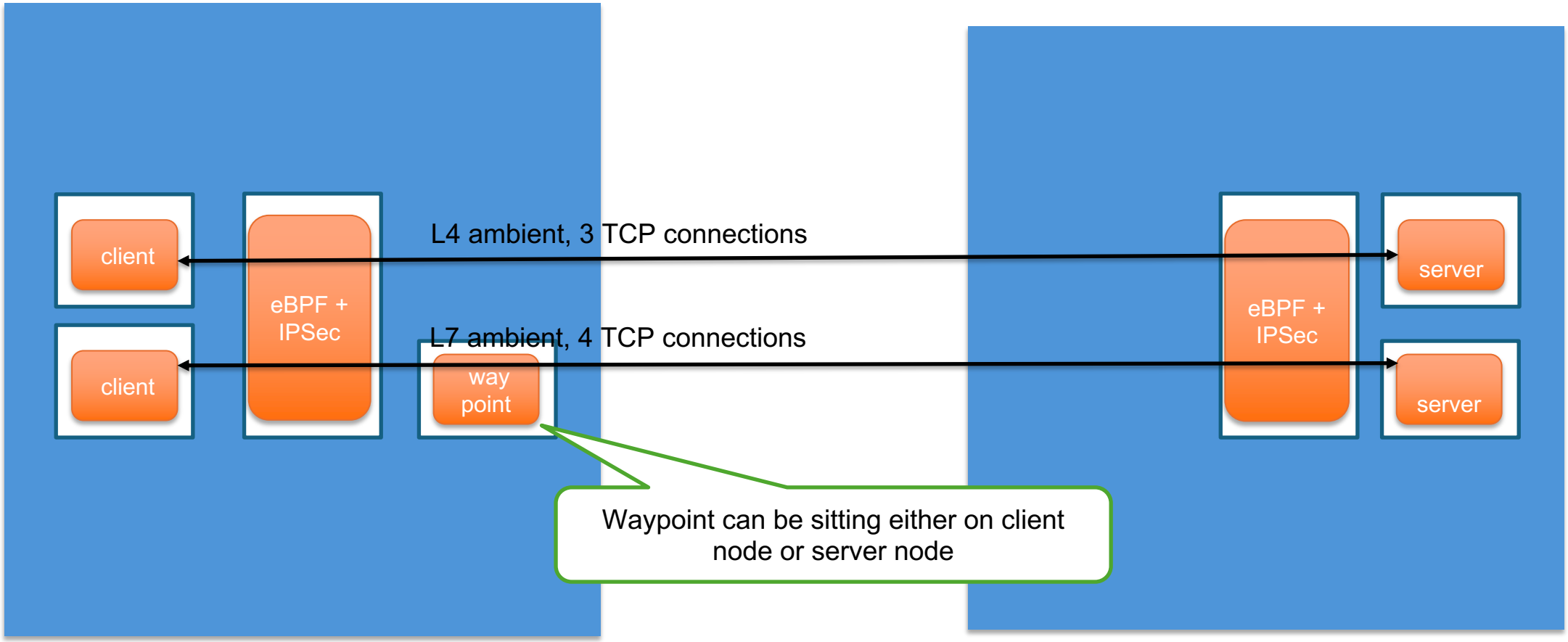
SW BOM

component	version	justification
Kubernetes	v1.28.2	GitCommit:89a4ea3e1e4ddd7f757228609035983e0387b2f
Runc	1.1.9	v1.1.9-0-gccaecfc
Containerd	1.6.24	61f9fd88f79f081d64d6fa3bb1a0dc71ec870523
Cri-dockerd	0.3.7	
Calico	v3.26.3	
Istio	1.20.0	ambient running in ebpf redirection mode
Fortio	1.17.0	
Mocket	0.0.10	

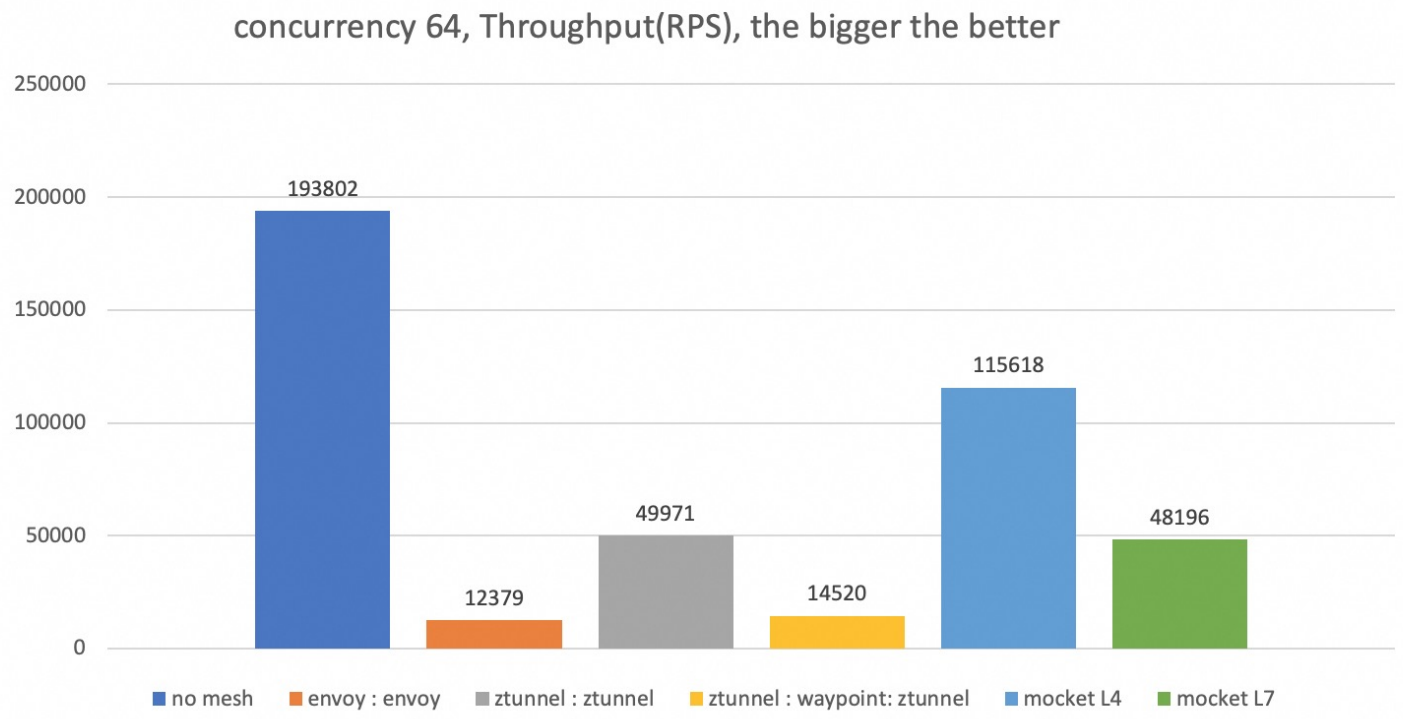
Comparing groups



Comparing groups



Mocket 和 Ambient 性能对比-RPS

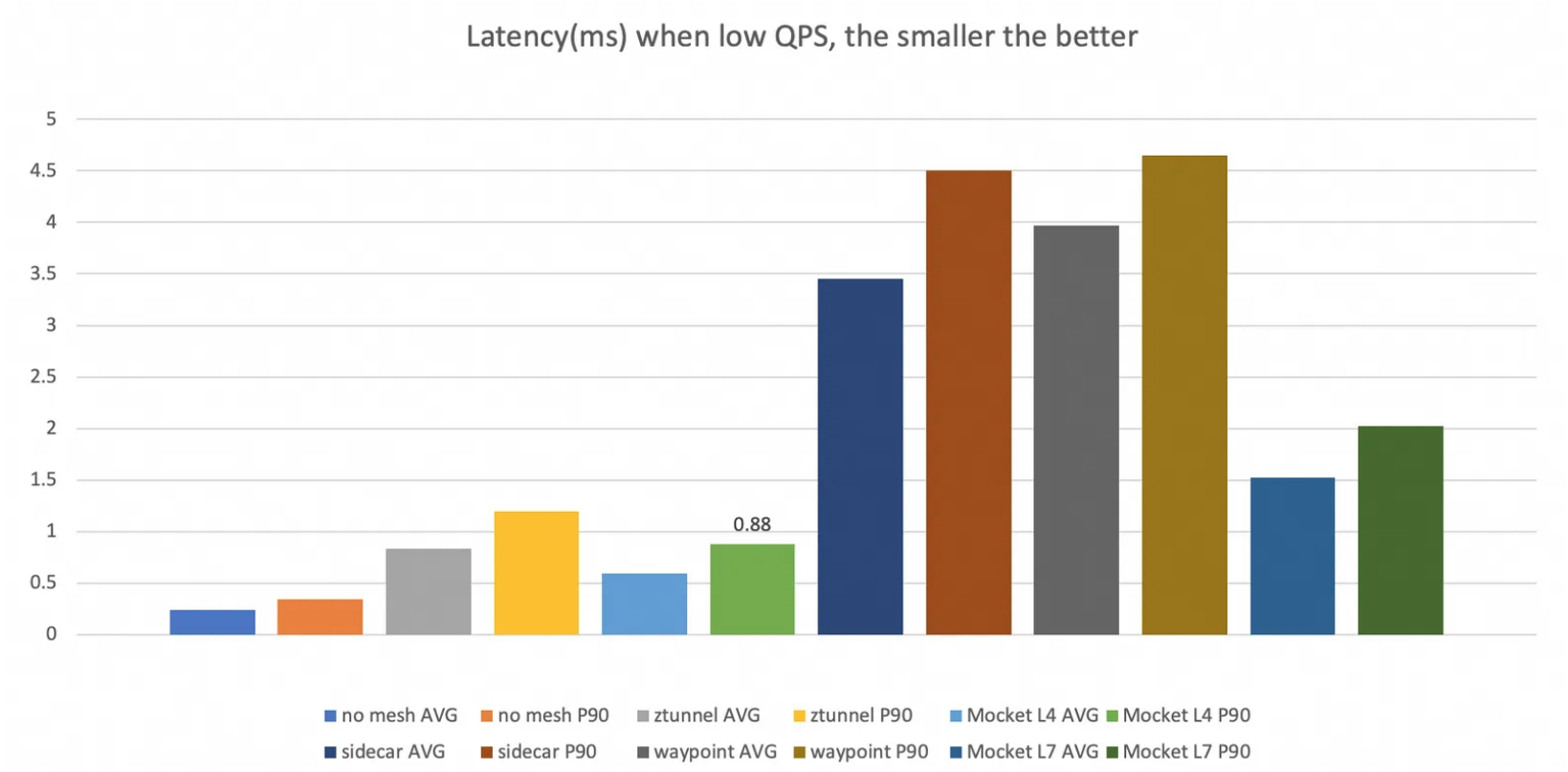


64 并发，1024 echo size ,压测命令：

fortio load -c 64 -qps -1 -t 30s -a -r 0.00005 -httpbufferkb=64 -labels ng-perf-test-xx http://fortioserver:8080/echo\?size\=1024

QPS 对比优化前，**L4 提升了130%，L7 提升了230%** 。

Mocket 和 Ambient 性能对比-Latency



测试： 采用固定QPS, 未达到服务处理瓶颈(14000 QPS)

Latency 对比优化前，**L4 、L7 AVG latency 降低接近50% - 60%。**

Q&A

Please scan the QR Code above
to leave feedback on this session