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

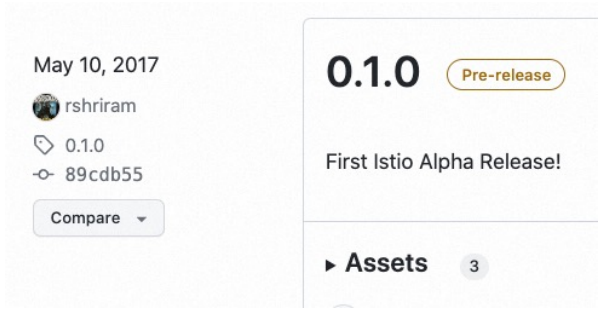
Speaker

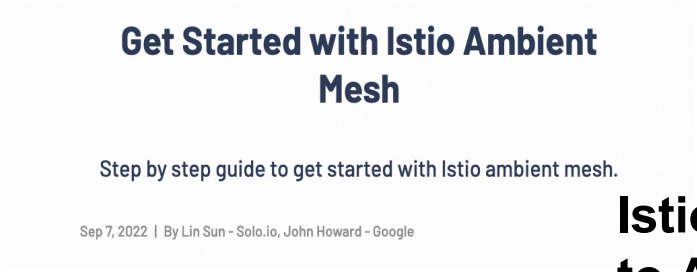
Yuxing Zeng

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 history





Istio 1.18, Ambient to Alpha

Istio 1.22, Ambient to Beta?

2017

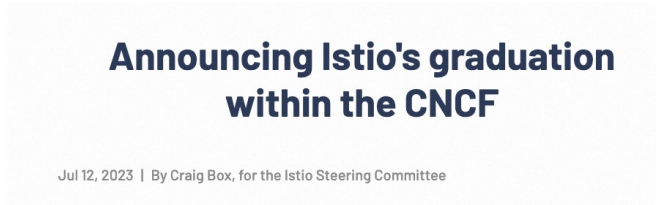
2020

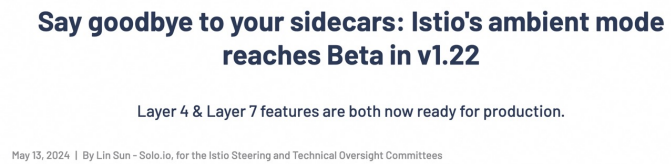
2022

2023

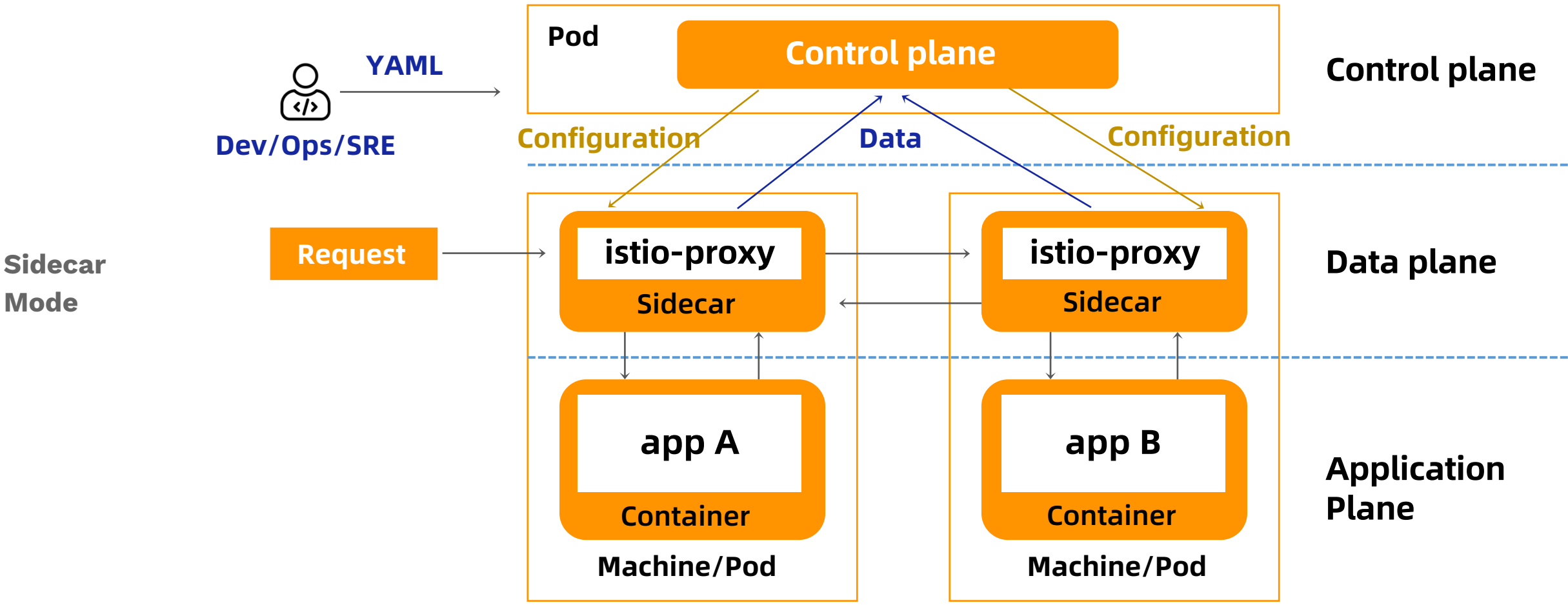
2024

Istio 1.5, Simplified Istio (istiod)





Istio Data Plane Mode: Sidecar –> Ambient

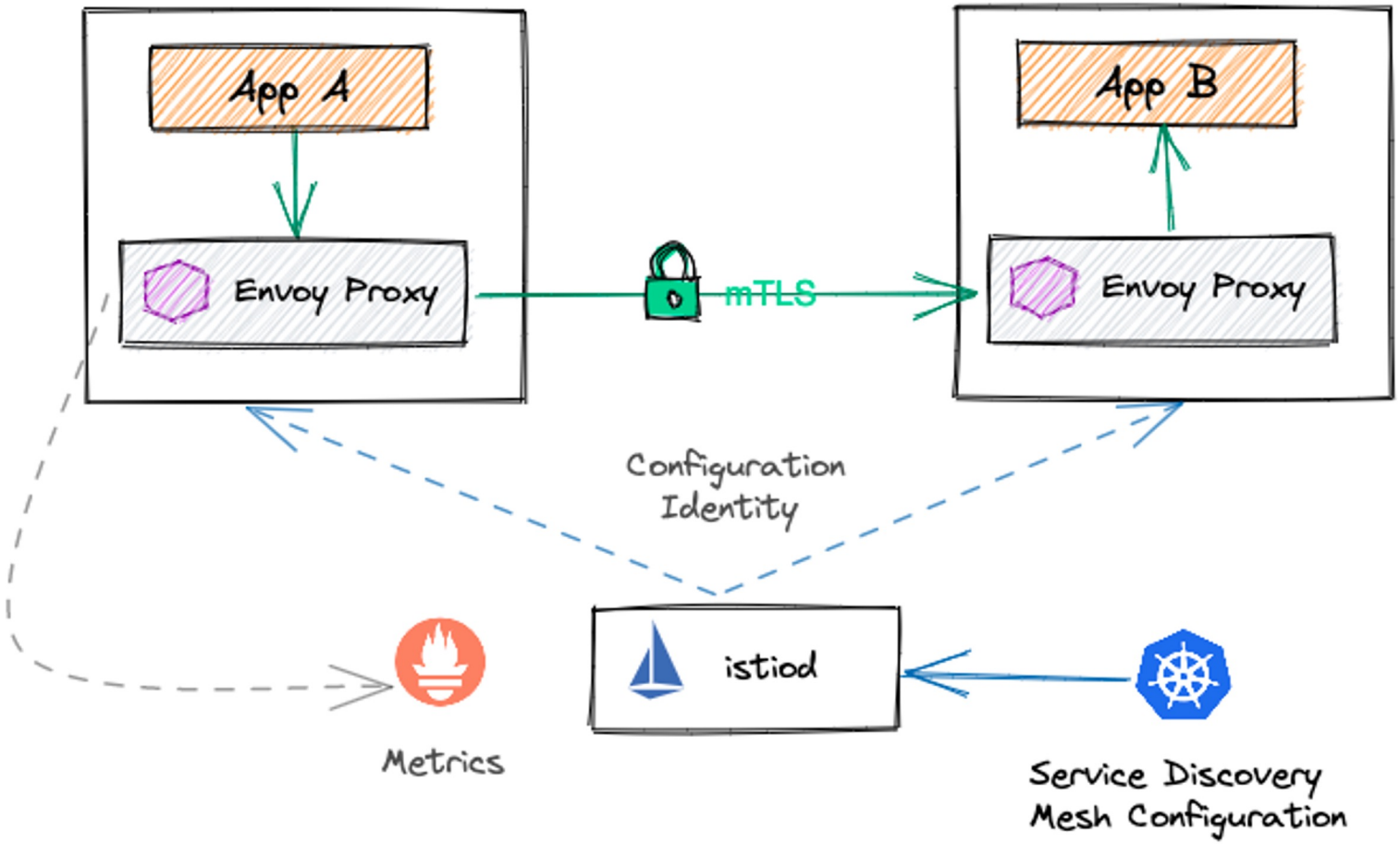


Challenges with Sidecar Mode

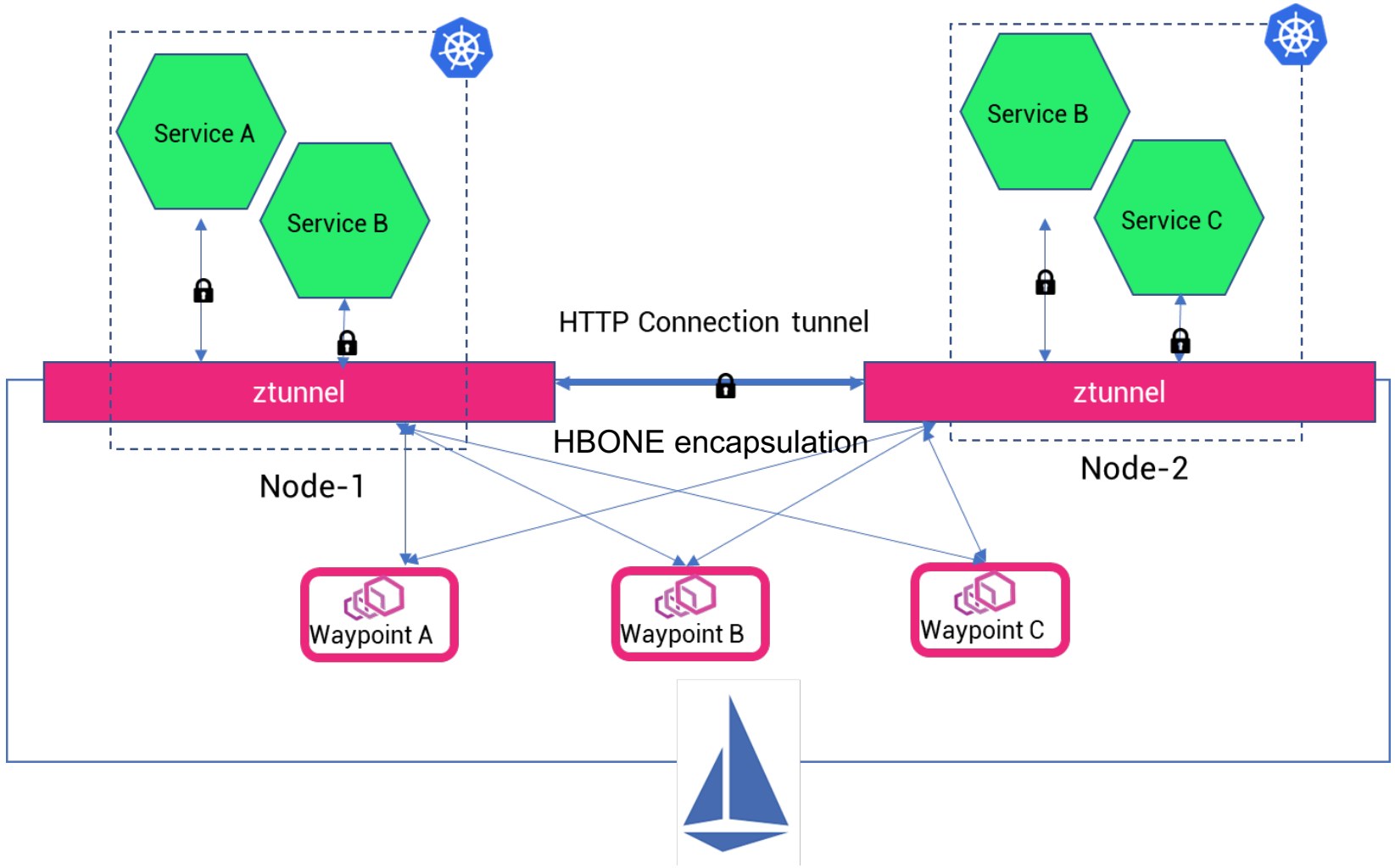
OPERATIONAL COMPLEXITY

OVERHEAD COST

PERFORMANCE



About Ambient



Goal:

- Simplify Operations
- Cost Reduction
- Improve Performance

Istio Ambient Mode

Design Concept: Layer the data plane to allow users to adopt service mesh techno logies in a more incremental manner.

L7 features

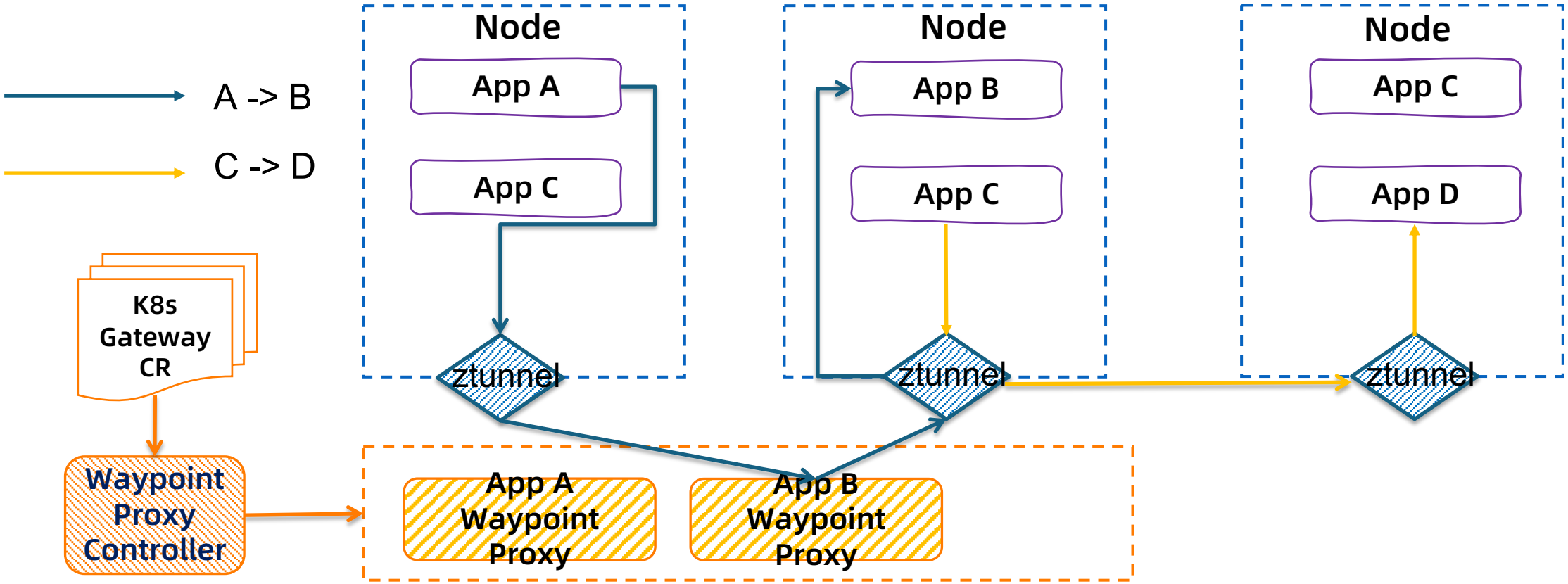
- Traffic management：HTTP Route、Load Balance、Circuit Breakers、Rate limit、Retry、Timeout、Fault injection etc.
- Security：authorization policy for the L7
- Observability：Trace 、metrics 、logs

L4 features

- Traffic management：TCP Route
- Security： authorization policy for the L4
- Observability

Separation of Business Applications and Data Plane Proxies in Istio

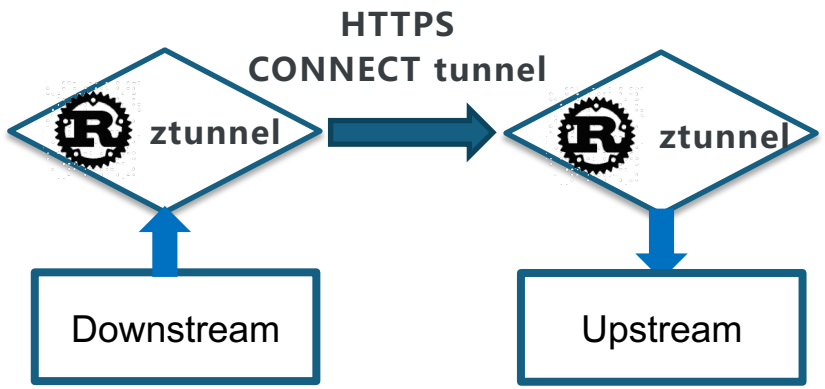
Waypoint	• run completely independently of the application, enhancing security; each identity (service account in Kubernetes) has its own dedicated L7 proxy, avoiding the complexity and instability introduced by a multi-tenant L7 proxy model
ztunnel	• The traffic from the workload is redirected to ztunnel, which then identifies the workload and selec ts the appropriate certificate for processing.



Ambient L4 — Ztunnel

ztunnel traffic management

- ztunnel : Rust based
- Using XDS API Sync the configure
- Configures: Workload/Service ip、L4 AuthN policy、Workload certs

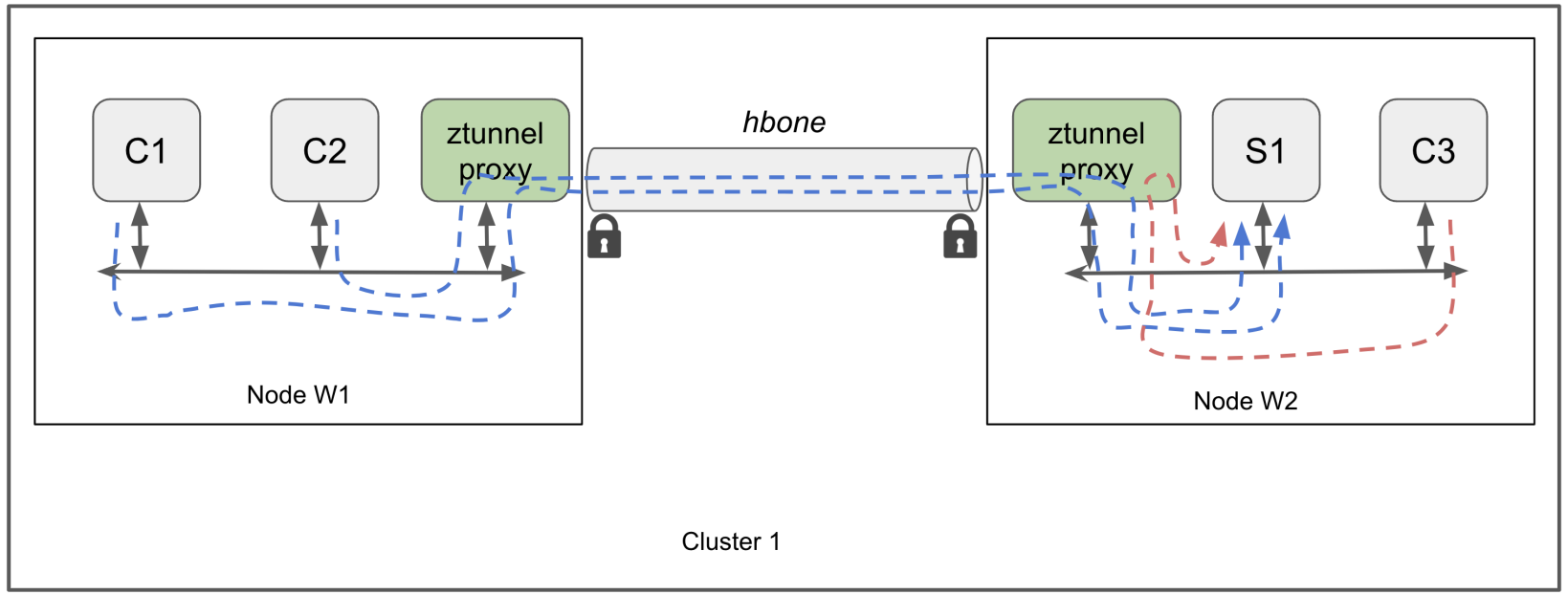


Main Features :

- AuthZ & AuthN, Using HBONE
- L4 Load balance

HBONE数据包组成

Ambient Known Issues



Ztunnel HA & Performance：

- HA Issue： Using DamonSet , Only one ztunnel pod per node, it is a single point, and if the pod fails or exception, it will affect the mesh traffic on the entire node https://github.com/istio/ztunnel/issues/40
- HBONE Traffic RTT and Ztunnel Performance, about Issues: https://github.com/istio/istio/issues/48271 https://github.com/istio/istio/issues/48949

Operations and maintenance are complex ：

- The introduction of the HBONE protocol adds complexity to operations, which is something that L4 and L7 Envoy need to be aware of at the application layer.

Ambient Limitation — Upgrading will disrupt all workload traffic

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.

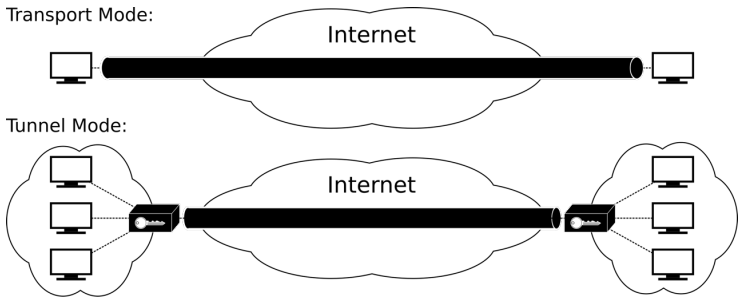
Node cordoning and blue/green node pools are recommended to control blast radius of application pod traffic disruption during production upgrades. See your Kubernetes provider documentation for details.

For the production environment, this is critical.

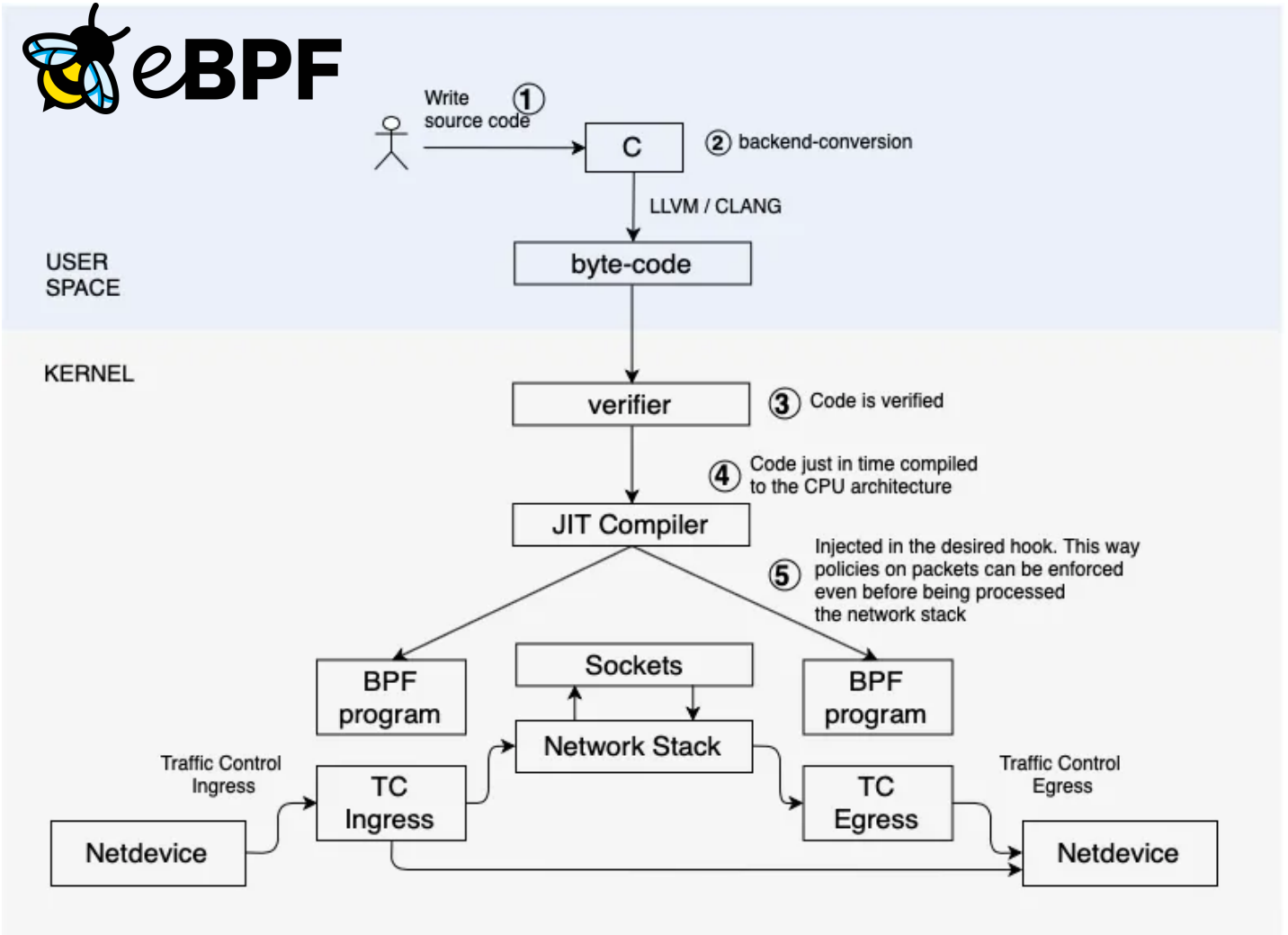
About Issue:
https://github.com/istio/istio/issues/51126

L4 alternative implement: using eBPF + IPSec

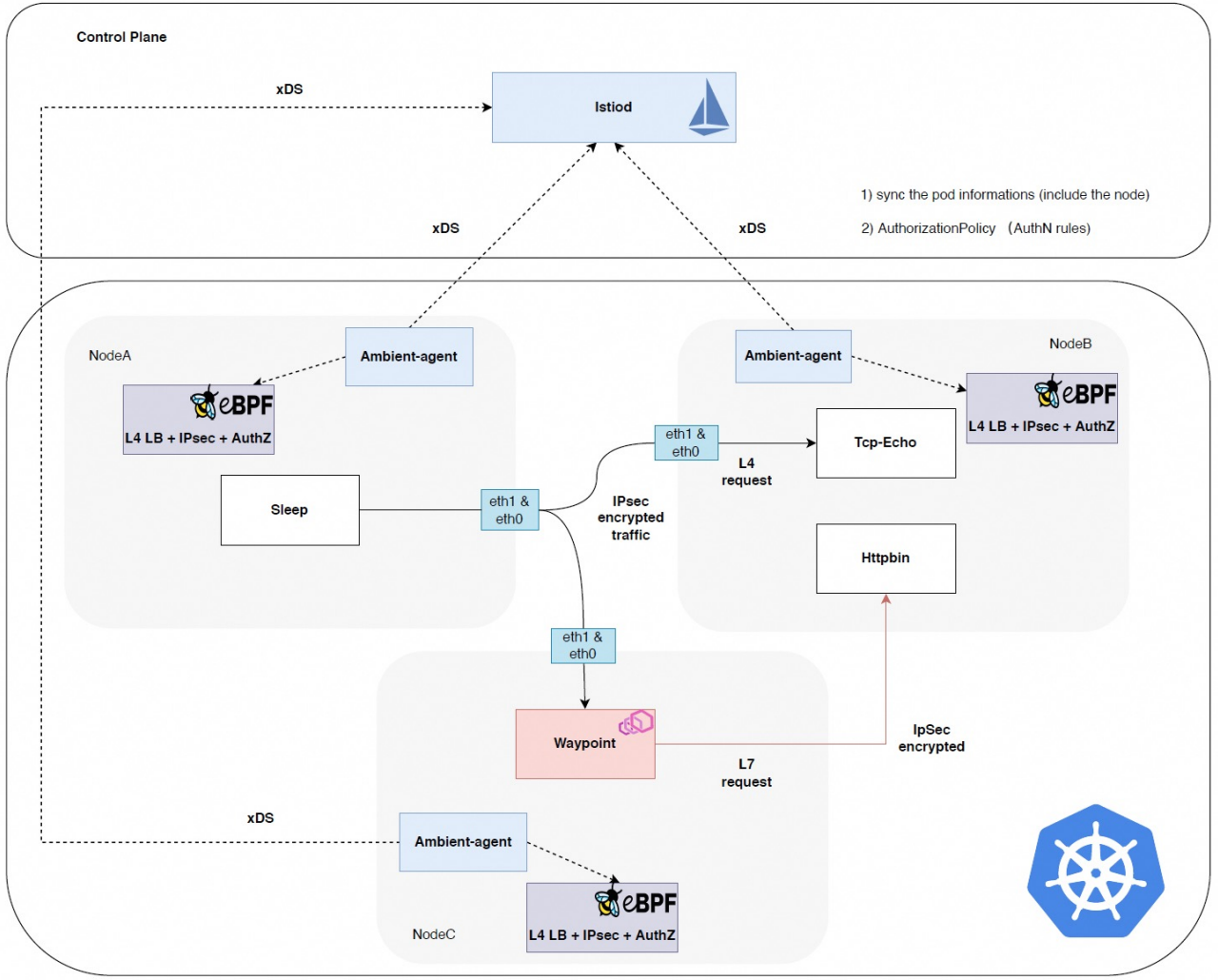
- eBPF LB replace ztunnel LB
- IPSec/WireGuard replcace mTLS(HBONE) , IPSec/WireGuard is a feature included by default in the Linux Kernel



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

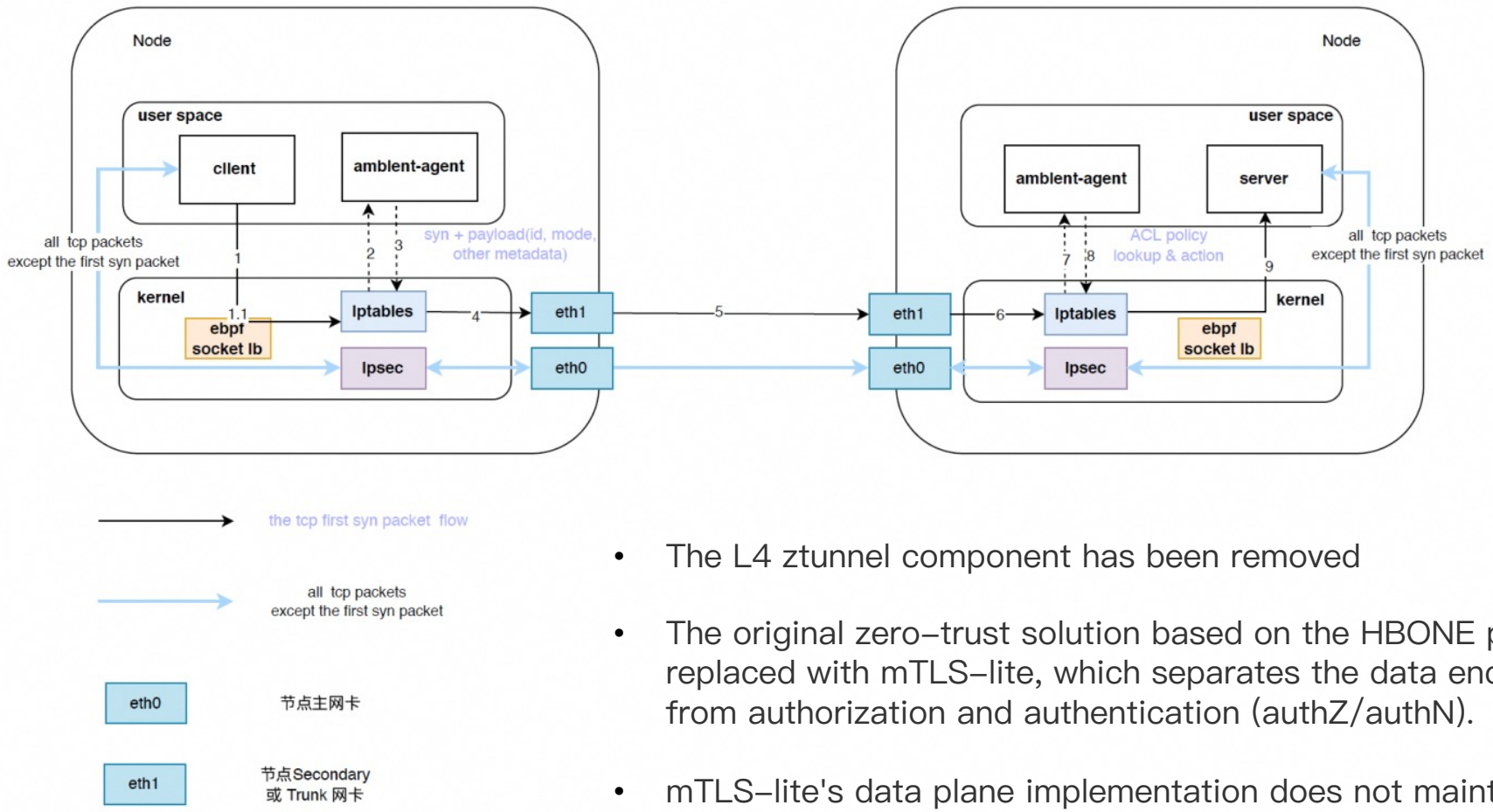


Ambient optimization – Mocket Project



- using eBPF L4 LB
- IPSec + AuthZ replace HBONE + AuthZ
- Observability

Mocket Core



- The L4 ztunnel component has been removed
- The original zero-trust solution based on the HBONE protocol has been replaced with mTLS-lite, which separates the data encryption channel from authorization and authentication (authZ/authN).
- mTLS-lite's data plane implementation does not maintain TCP connections, and the ambient-agent upgrade is seamless for users.

Istiod Support mocket gatewayClass


```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
```

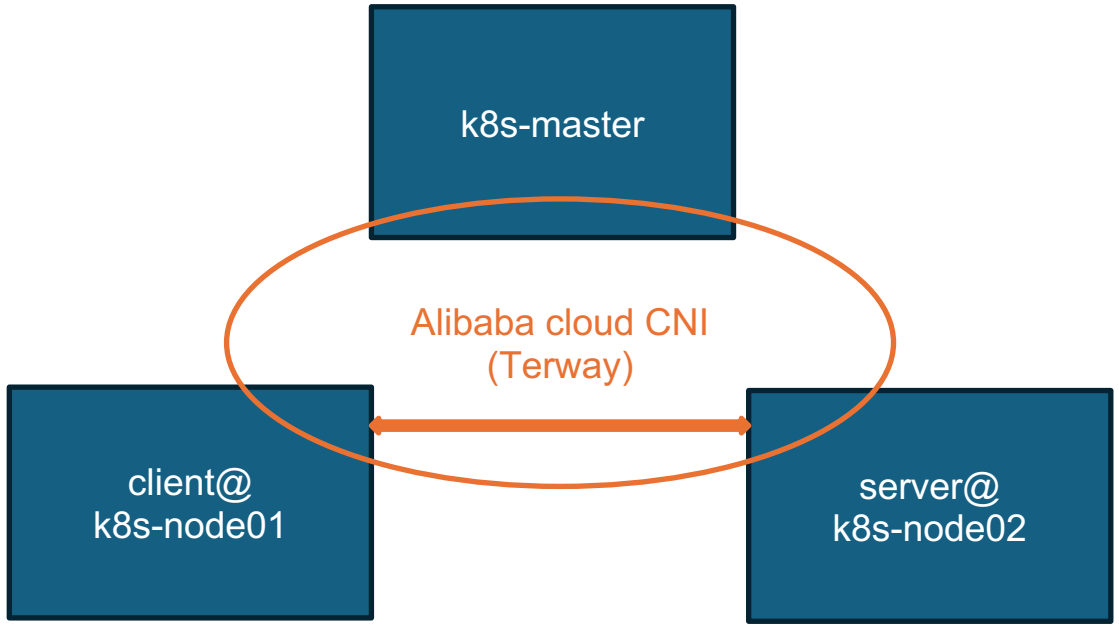
Modify Istiod to support mTLS–lite

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 ~]#
```

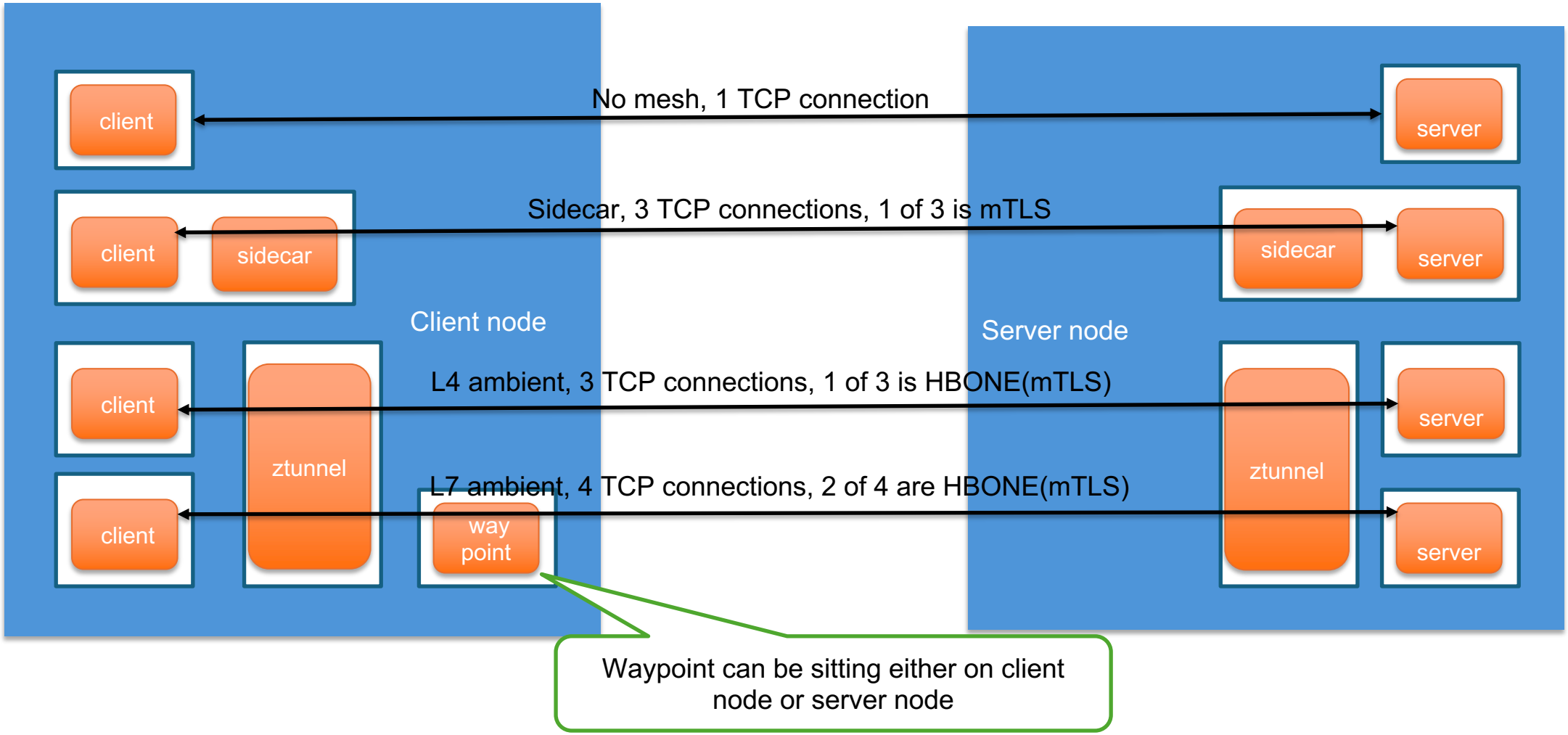
```
[root@k8s-node01 ~]#
[root@k8s-node01 ~]#
[root@k8s-node01 ~]# lscpu
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              8
On-line CPU(s) list: 0-7
Thread(s) per core:  2
Core(s) per socket:  4
Socket(s):           1
NUMA node(s):        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 ~]#
```

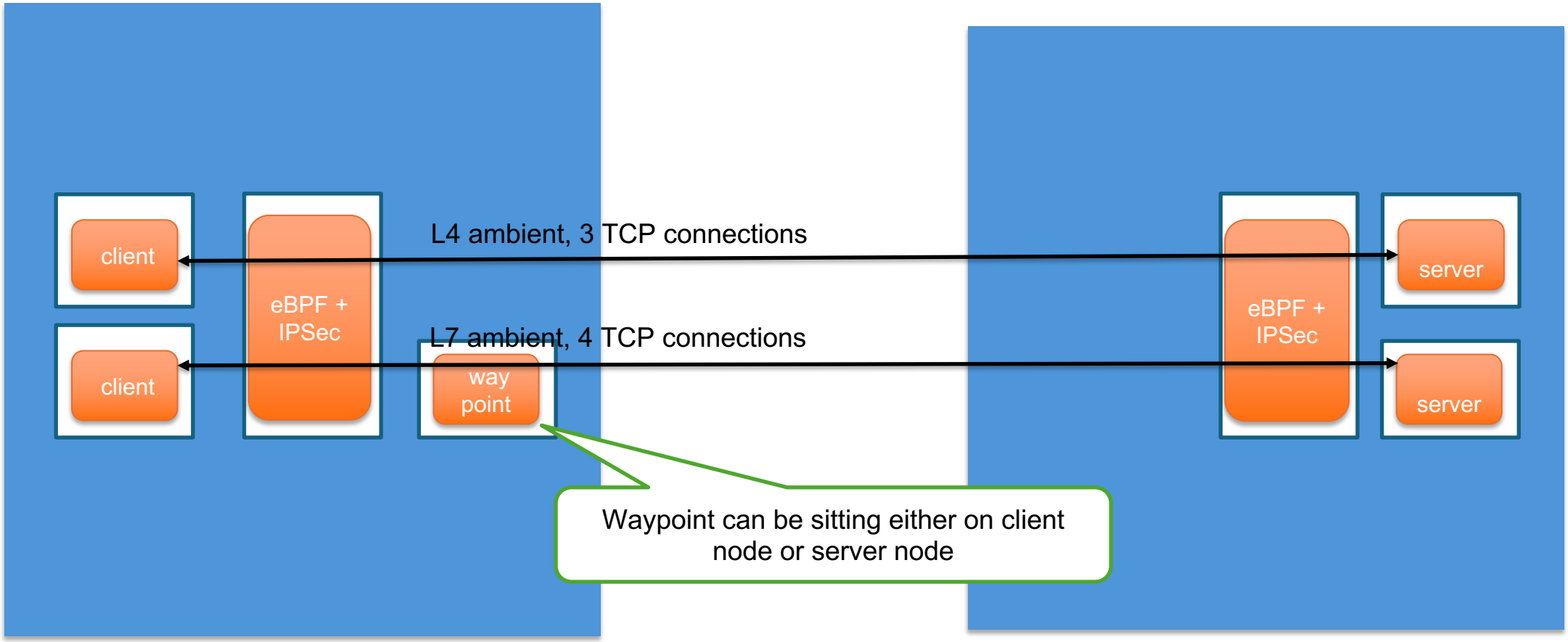
SW BOM

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	

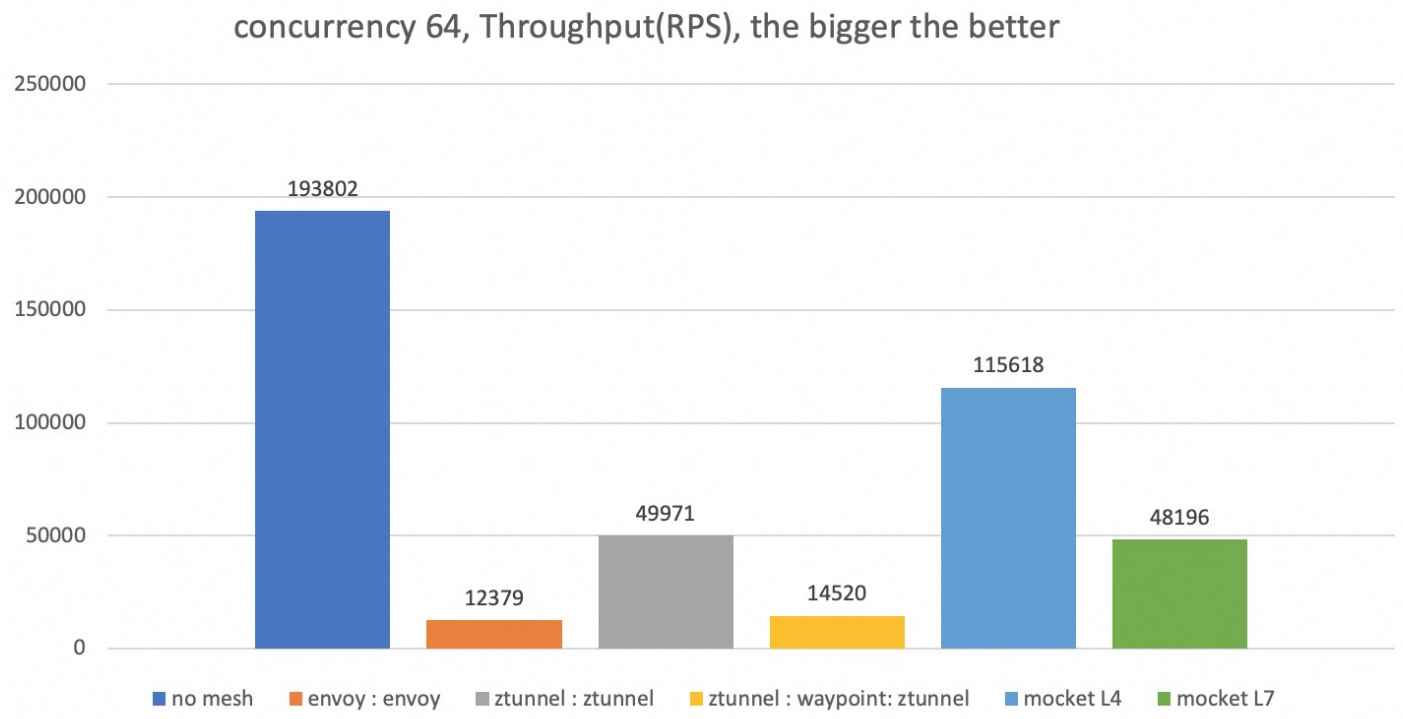
Comparing groups



Comparing groups



Performance Comparison of Mocket and Ambient – RPS

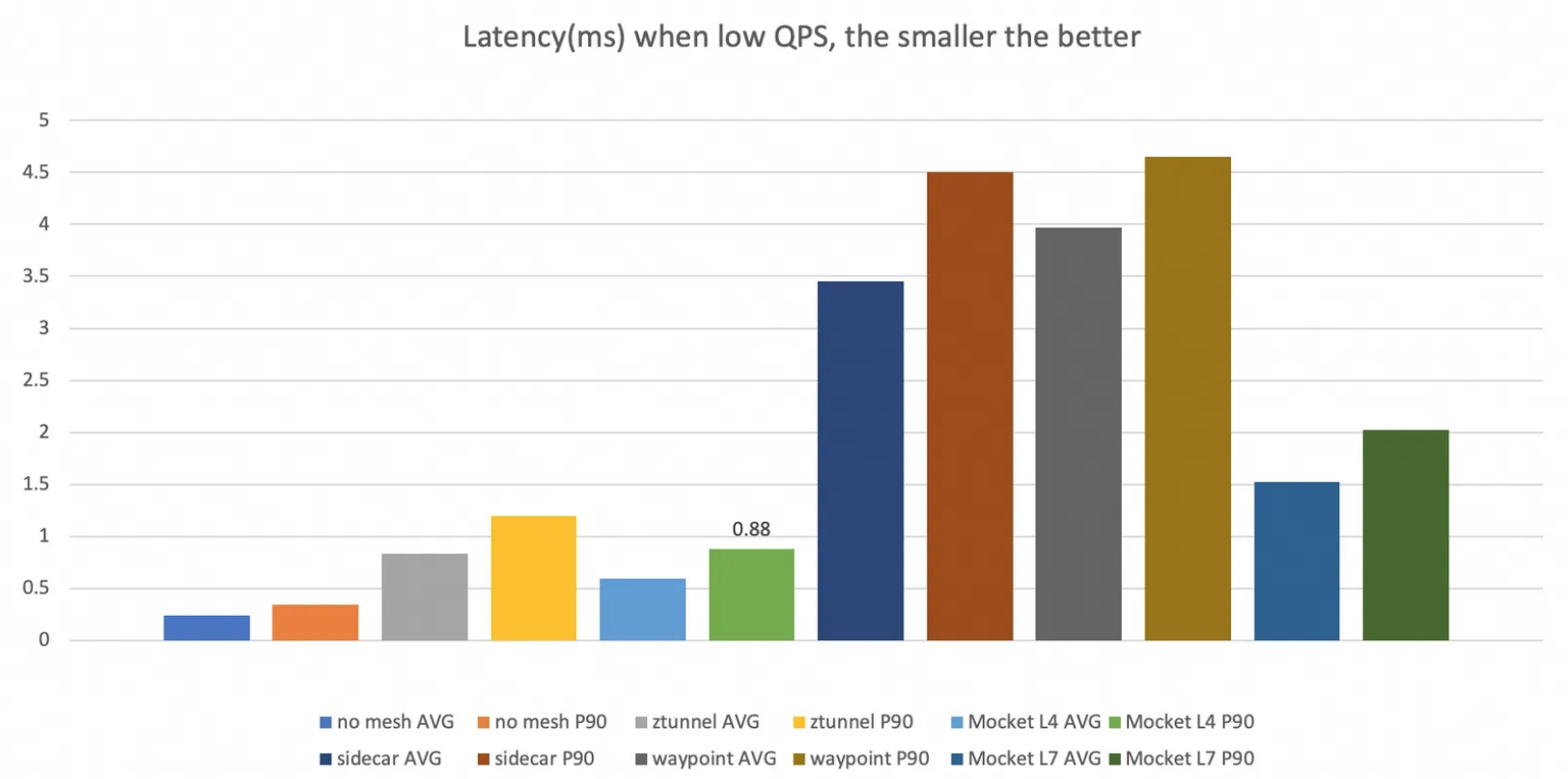


concurrency 64 ，1024 echo size，command：

```
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 comparison shows that **Layer 4 improved by 130% and Layer 7 improved by 230%** after optimization.

Performance Comparison of Mocket and Ambient – Latency



Test： Using a fixed （14,000） QPS, without reaching the service processing bottleneck

The average latency at L4 and L7 has **decreased by nearly 50% – 60%** compared to before optimization.

Q&A

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