「Cloud Native Summit」专场

KubeSkoop: 容器网络问题的自动化诊断系统

王炳燊 2022年05月28日

About Me

- 王炳燊（溪恒）

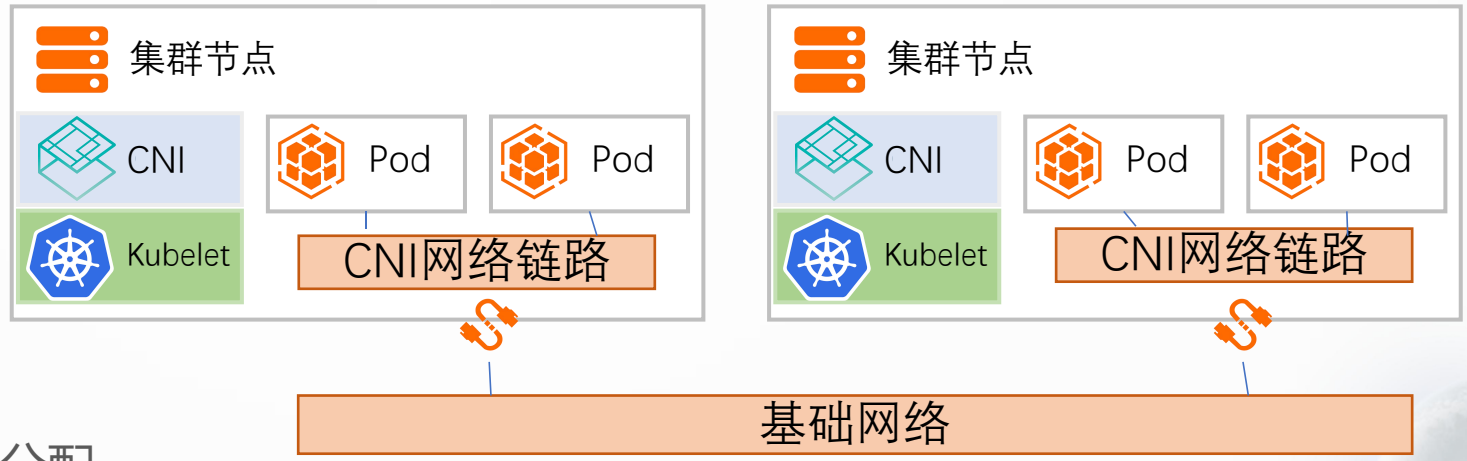
- @阿里云容器服务 ACK （Alibaba Cloud Container Service for Kubernetes）
  - 产品覆盖全球30+个地域，最大的集群超过10k的节点
  - 我负责其中容器网络方案设计和研发

- 长期聚焦在云原生开源技术栈
  - Terway – 阿里云云原生CNI插件
  - Raven – OpenYurt边缘社区云边互通
  - KubeSkoop – 容器网络问题诊断系统

Kubernetes 容器网络

网络连通性 - CNI

Kubernetes集群网络连通性
- 每个Pod有自己独立的网络空间
- 集群中Pod和Node节点可以互相通信
  - 同节点通信
  - 跨节点通信



CNI插件来实现的集群连通性要求
- 集群级别唯一的地址分配
- 集群纬度的网络打通

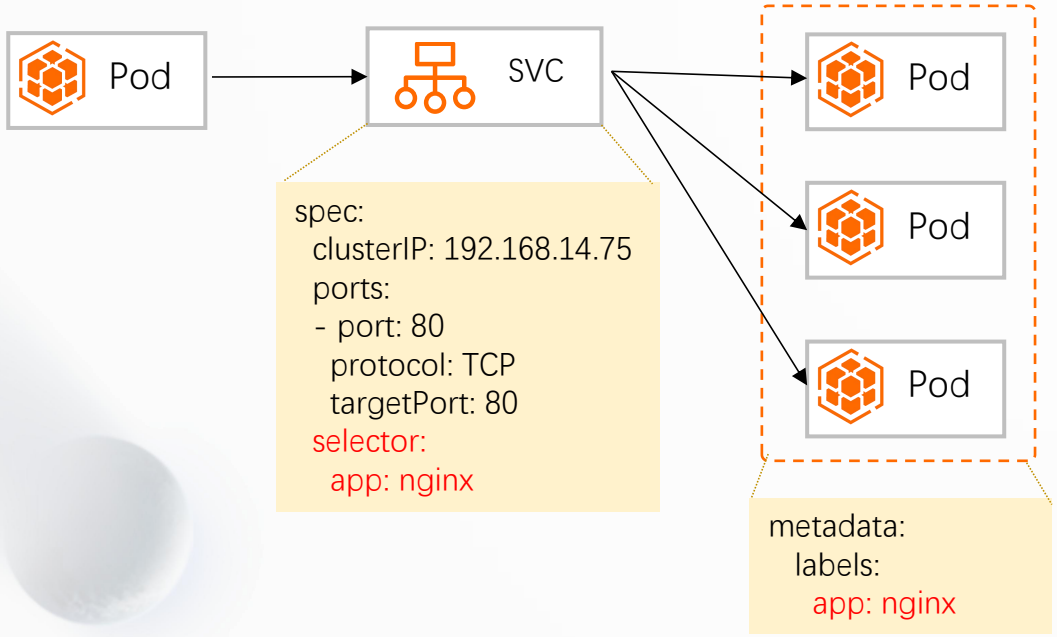
不同CNI插件的不同网络实现
- 地址分配： 每节点按段分配、全局分配
- 网络虚拟化实现： bridge、ipvlan、sriov等
- 跨节点的网络连通性实现： overlay、underlay、BGP等

Kubernetes 容器网络

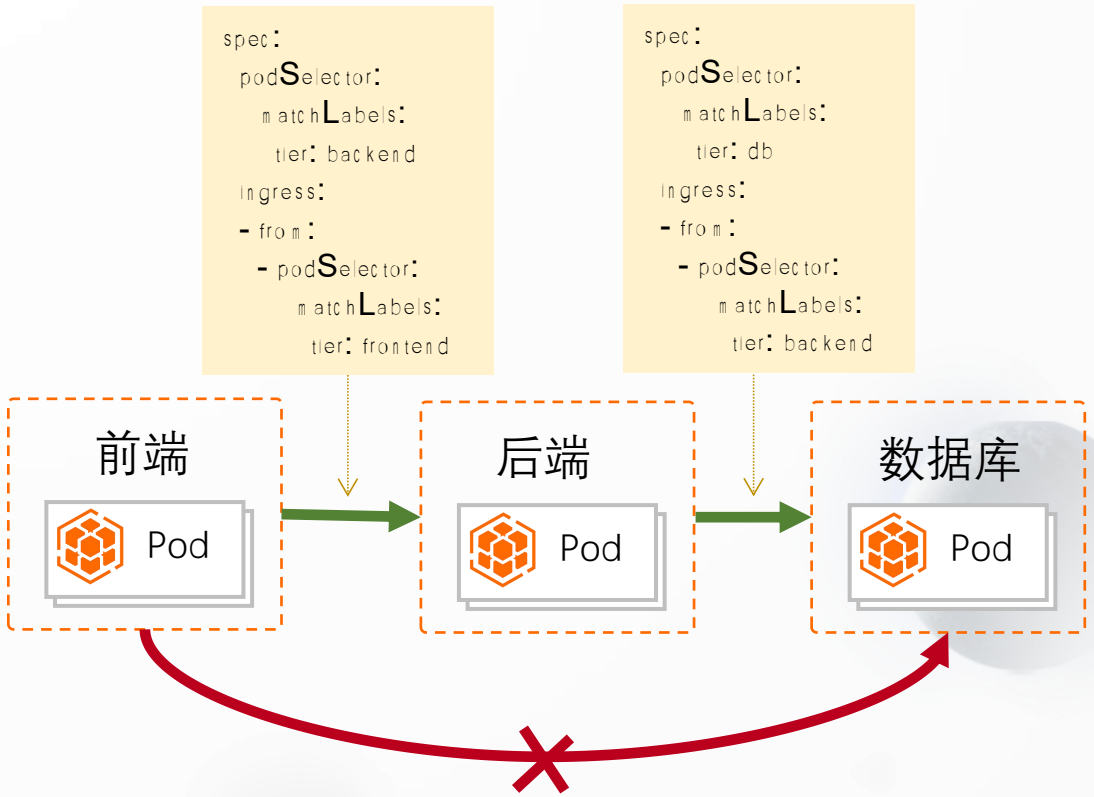
服务发现和网络策略

Kubernetes基于Label的服务发现和网络策略

Kubernetes服务发现 -- Service



Kubernetes网络策略 -- NetworkPolicy



Kubernetes 网络问题的复杂性

逻辑概念的复杂性

- Ingress/Service/NetworkPolicy 配置灵活
  - LabelSelector选择到不预期的Pod
  - 多个NetworkPolicy规则重叠
  - Service NAT的端口和实际监听的不符合
- ServiceMesh
  - 更复杂的7层网络策略
- 第三方CNI插件
  - 自定义的网络扩展

数据面实现的复杂度

- 数据平面复杂：
  - 数据面多层处理，ServiceMesh/KubeProxy/CNI
  - 多种CNI网络实现(Overlay/Underlay...)
- 协议栈链路复杂：
  - 链路长，涉及网卡驱动/netfilter/route/bridge等
  - 配置繁琐难懂
- 底层网络配置：
  - 每个云厂商配置不同
  - 安全组、路由表等配置复杂

Kubernetes 网络问题传统定位手段

Kubernetes网络问题定位

定位流程长

- 在Pod、Node等多处同时抓包
- 定位丢包点
- 分析丢包点配置（iptables/内核参数/云厂商网络等）

大量定位时间开销

- 压测等手段复现问题
- 需要了解网络插件原理
- 大量的抓包和信息采集

人员经验要求高

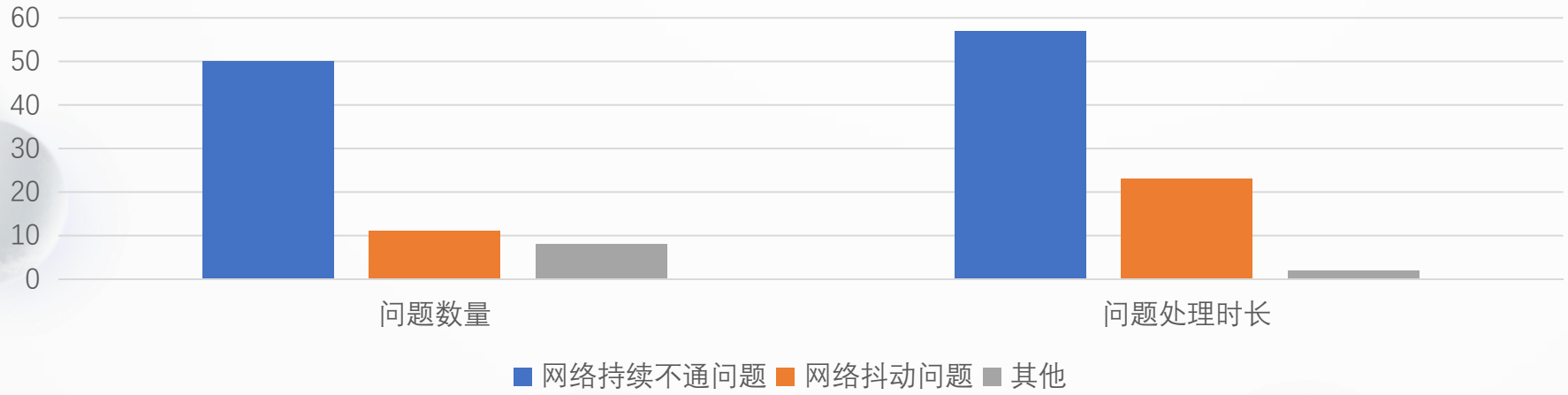
- 熟练的运维能力
- 精通Linux协议栈
- 精通对应云厂商的网络配置

KubeSkoop

Kubernetes网络问题类型和自动诊断系统

- 网络持续不通问题：
  持续的无法访问： ping不通、connect超时、DNS无法解析等
- 网络抖动问题：
  - 偶发的网络问题： 偶尔业务的超时，504，偶发Reset等
  - 网络性能问题： 网络性能低、QPS压不上去等
- 80% 都是可以依赖经验解决的已知问题，处理时间浪费在问题上报、信息收集和验证上

阿里云容器服务ACK 上常见网络问题种类和比例：



KubeSkoop

项目实现

K8s常见网络问题

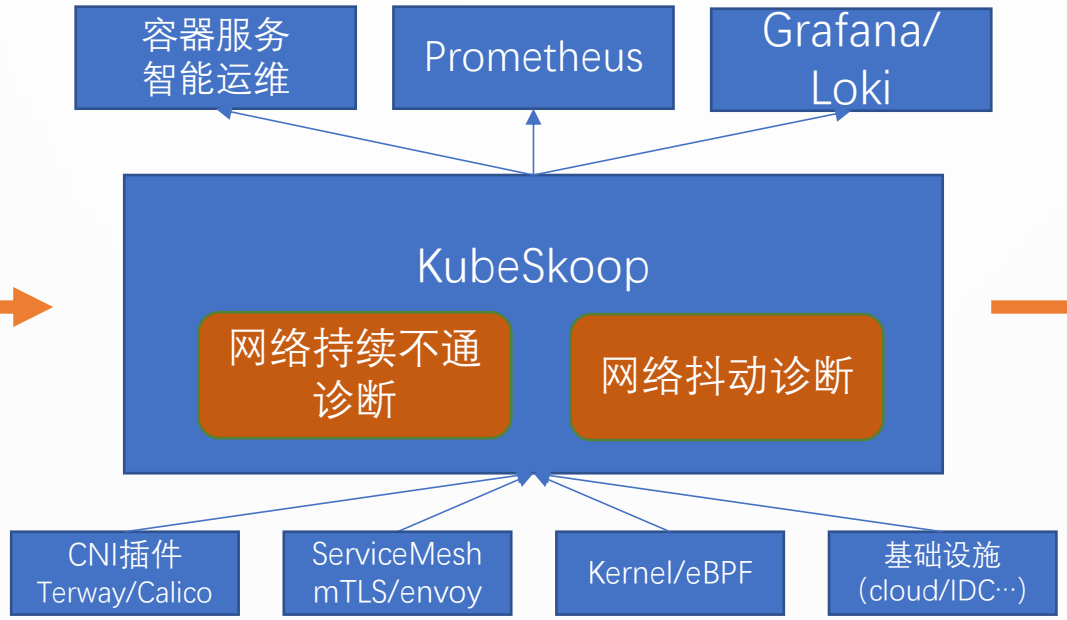
网络不通

- DNS解析异常
- SVC无法访问
- Pod无法联通
- Mesh路由错误

网络抖动

- 抖动无法复现
- 异常点定位
- 应用/平台定界
- 异常根因定位

使用KubeSkoop诊断



定位问题根因

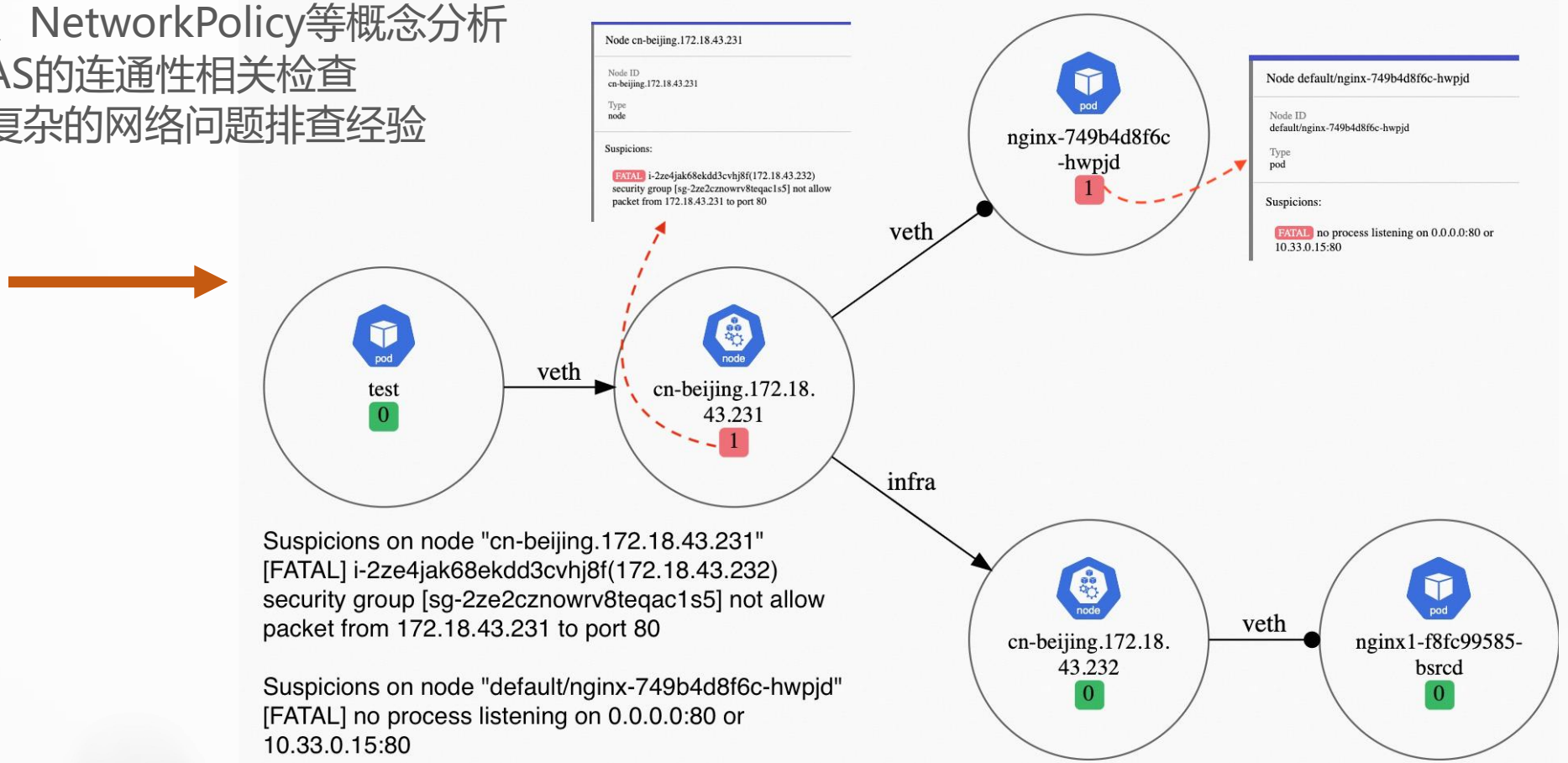
- 全链路一键诊断
- 网络栈延迟分析
- 网络异常事件识别回溯

KubeSkoop

通过KubeSkoop一键诊断网络持续不通问题

- 用户指定不通的来源目的，等待输出诊断的结果
- 自动构建网络访问链路，分析链路问题
- 包含Kubernetes Service、NetworkPolicy等概念分析
- 全面覆盖协议栈、底层IAAS的连通性相关检查
- 无需了解网络插件实现和复杂的网络问题排查经验

```
~# kubeskoop \
    --src 10.33.0.13  \
    --dst 192.168.14.75 \
    --dport 80
```



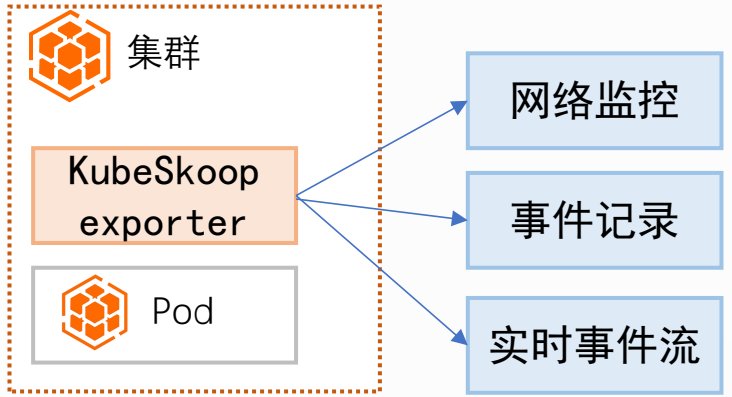
完整使用文档：https://kubeskoop.io/docs/guide/diagnose/intro

KubeSkoop

通过KubeSkoop诊断网络抖动场景

基于eBPF的容器网络异常监控

- 云原生部署，与Prometheus等可观测体系对接
- Pod级别的网络监控能力
- 精简、低开销的内核异常监控
- 覆盖驱动、netfilter、TCP等完整协议栈，几十种异常场景识别
- 支持网络问题的Metrics和事件回溯



业务异常时间点 21：25 左右
通过回溯KubeSkoop监控
业务Pod重传↑，Virtnet指令延时突增
为底层虚拟化执行超时导致软中断卡住而延迟重传
然后通过事件分析到调用栈和根因

监控回溯

事件记录

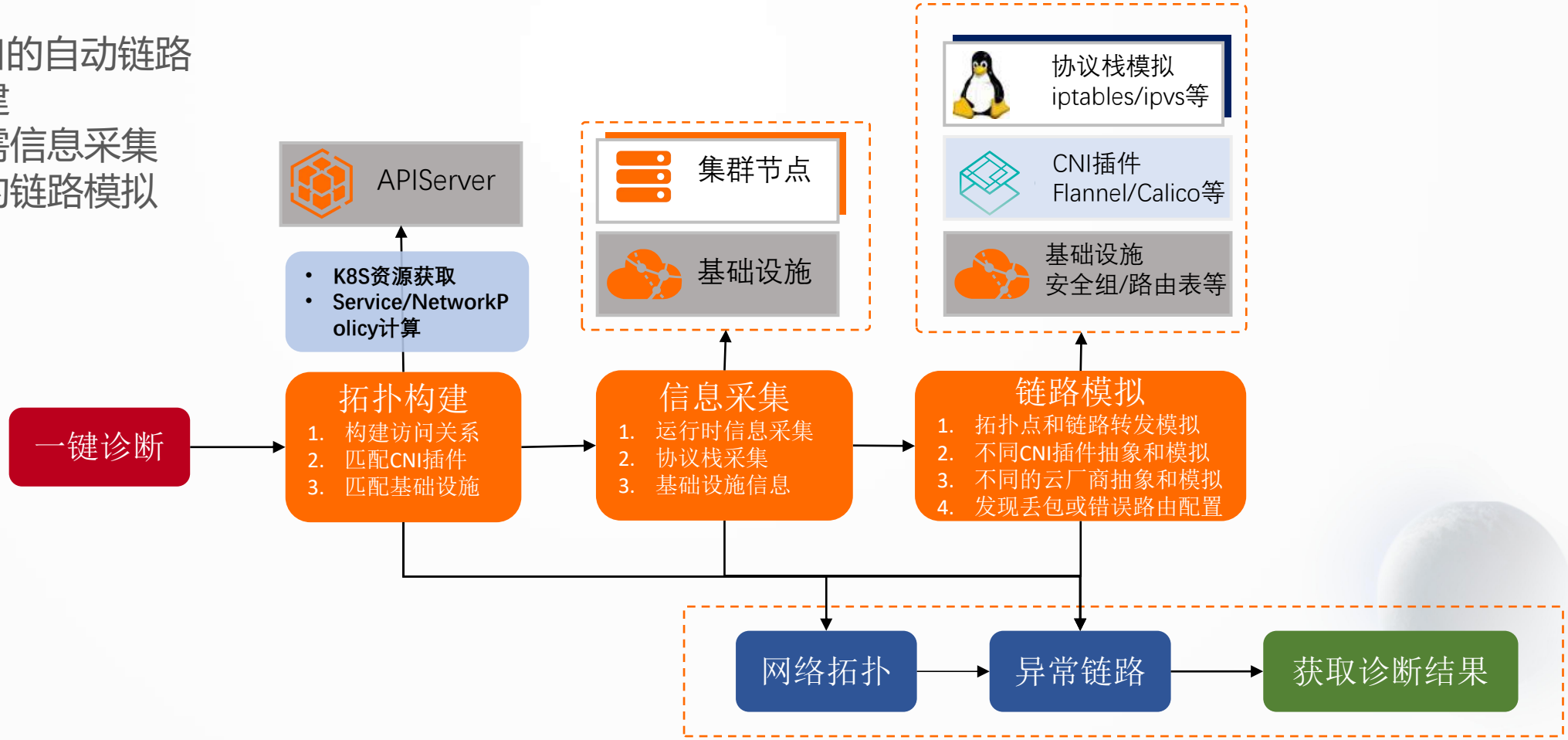
```
INFO  March  9 21:21:05 VIRTCMDEXCUTE nginx-ingress-controller-5c7fb5594-jwhnw protocol=TCP saddr=100.121.88.193 sport=27057 daddr=10.33.0.11 dport=443  stacktrace:
virtnet_send_command+0x1
virtnet_set_rx_mode+0x251
__dev_change_flags+0x9c  dev_change_flags+0x21
do_setlink+0x257  __rtnl_newlink+0x600
rtnl_newlink+0x44  rtnetlink_rcv_msg+0x119
netlink_rcv_skb+0x4e  netlink_unicast+0x1d7
netlink_sendmsg+0x240  sock_sendmsg+0x5f
____sys_sendmsg+0x232  ___sys_sendmsg+0x75
__sys_sendmsg+0x49  do_syscall_64+0x30
entry_SYSCALL_64_after_hwframe+0x61
```

完整指标参考：https://kubeskoop.io/docs/guide/exporter/exporter-description

KubeSkoop

网络不通诊断能力设计

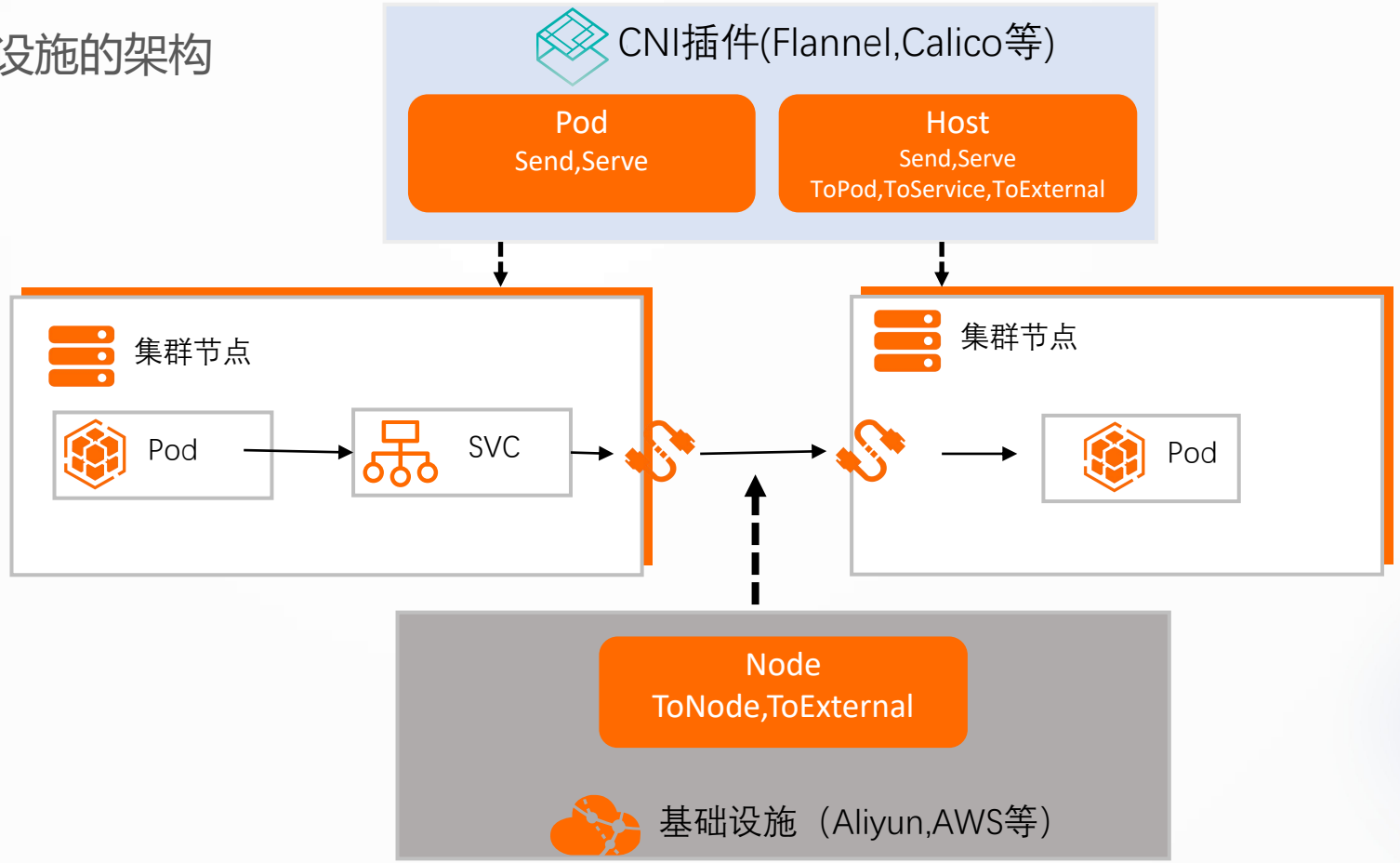
- 根据传入IP端口的自动链路分析和拓扑构建
- 根据拓扑的按需信息采集
- 全面链路覆盖的链路模拟
  - 协议栈
  - CNI插件
  - 基础设施



KubeSkoop

网络不通诊断能力设计

- 方便扩展的CNI插件和基础设施的架构
- 已支持Flannel, Calico CNI
- 已支持阿里云基础设施



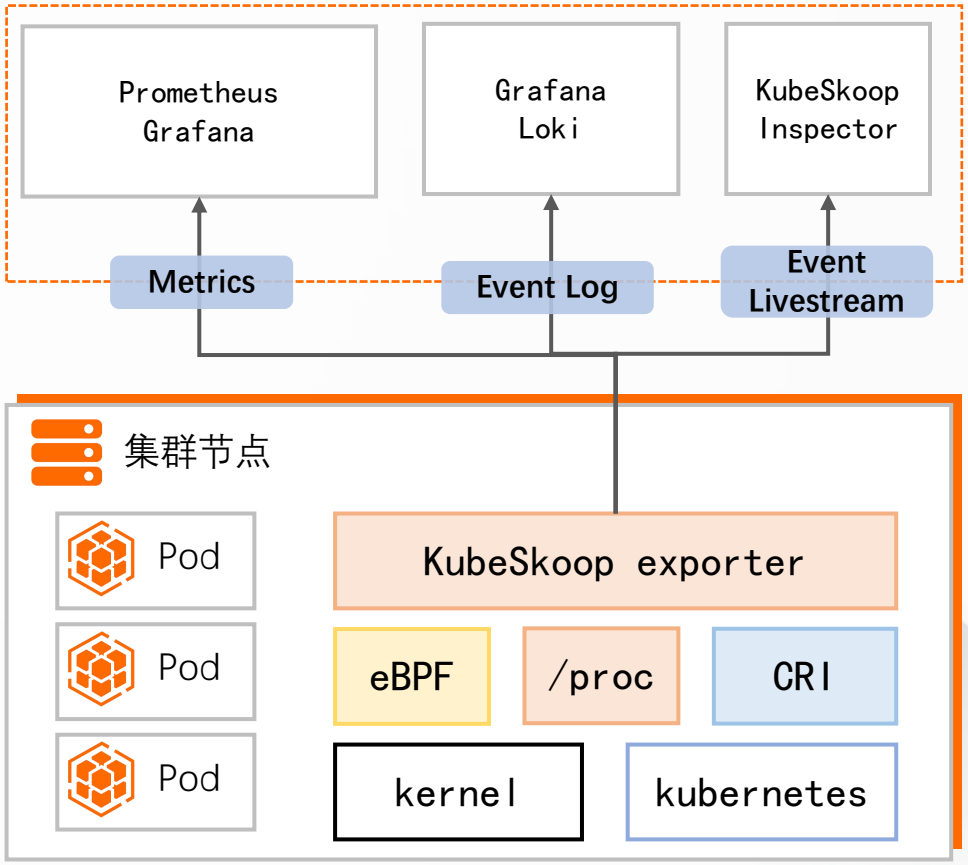
KubeSkoop

网络抖动的诊断设计

深度容器网络监控采集：
- 通过eBPF采集协议栈关键点
- 采集procfs下内核透出信息用于回溯
- 采用CRI接口关联采集点和Pod

容器指标和异常事件预处理：
- 网络异常Metrics过滤，减少开销
- 多指标聚合生成异常Event

网络Metrics和Event展示：
- 通过Prometheus+Grafana存储和回溯异常时间点指标
- Grafana Loki记录异常事件
- KubeSkoop Inspector查看实时异常事件流

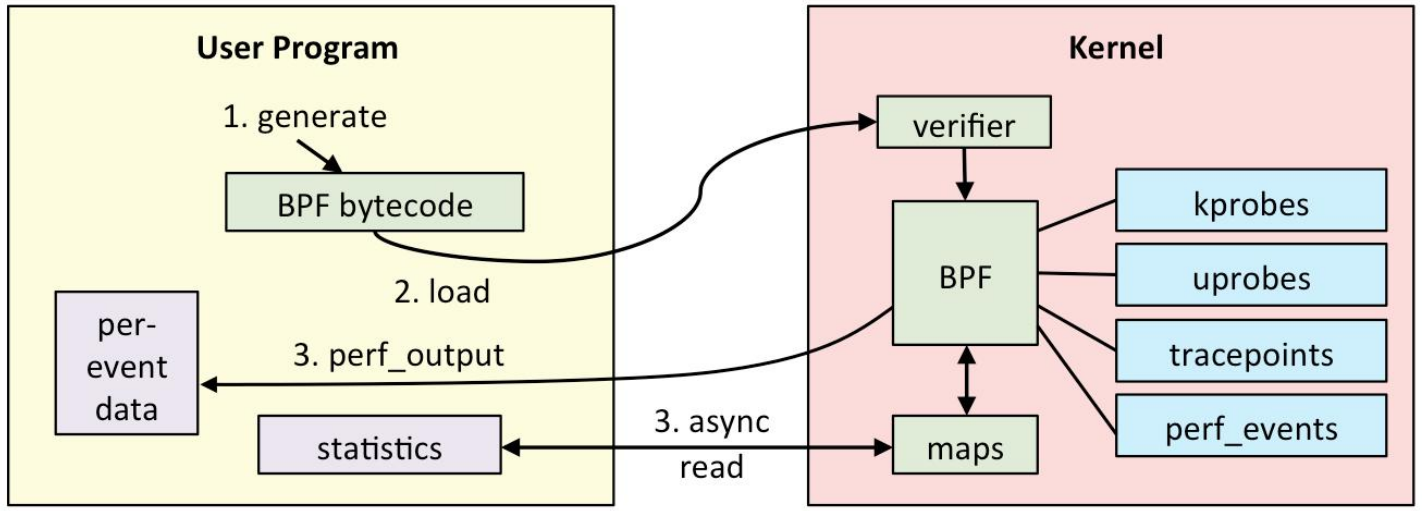


KubeSkoop

网络抖动的诊断设计

采用eBPF作为KubeSkoop主要的数据采集来源

- eBPF可以在内核中动态注入代码运行
- 内核中执行效率高
- 自带eBPF校验，避免宕机
- 可以通过map和perf_event和用户态通信



KubeSkoop如何使用eBPF

- 采用CO-RE方式减少编译开销，提升内核兼容性
- 减少在关键路径上的程序注入
- 尽量在eBPF程序中过滤异常数据，以减少内存开销
- 默认注入低开销程序，根据需求可动态插拔eBPF采集模块和修改过滤参数

```
# cat config.yaml
---
tracers:
- name: latency
  contexts:
    filters:
    - source_port: 6379
    - dest_port: 6379
    plugins:
    - 'ipvs‘
    - 'iptables'
```

eBPF采集配置

KubeSkoop

未来规划

- 增加更多云厂商和网络插件的支持
- 支持模拟发包和追踪以定位未知问题点，缩小排查范围
- 提供KubeSkoop Analysis 工具，智能分析KubeSkoop的指标和事件，降低问题理解门槛
- 不限于网络诊断，增加存储、性能等诊断
- 应用层感知，对7层协议(http,redis等)的感知和处理

GOTC

THANKS

项目官网： kubeskoop.io

钉钉扫码加入用户群

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE