第二届 eBPF开发者大会

www.ebpftravel.com

基于eBPF实现混部场景下的网络QoS管理

中国·西安

混部场景下，业务QoS存在挑战，网络QoS是关键指标

业界现状

- 数据中心基础设施支出大，服务器占IDC成本大头
- 服务器资源利用率低，平均在15%
- 不同类型业务使用独立资源池(分开部署)

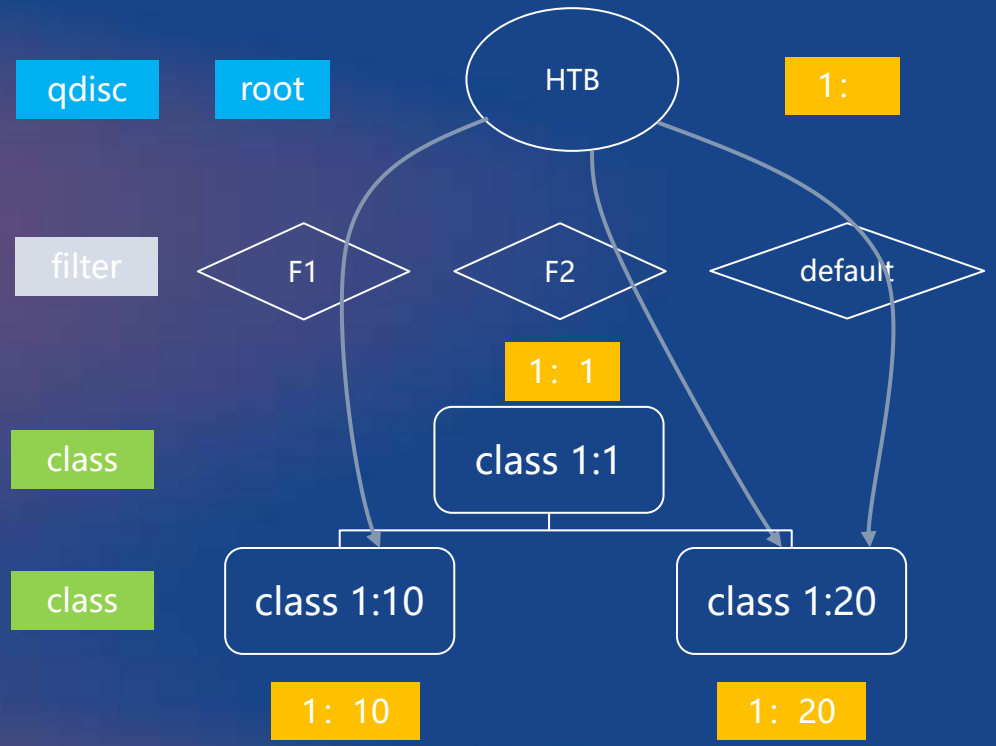
混部目标

- 在线业务和离线业务混合部署，降本增效

混部场景（鱼和熊掌）

- 离线业务填充在线业务波谷 → 利用率（鱼）
- 在线业务抢占离线任务资源 → 质量保障（熊掌）

linux TC简介



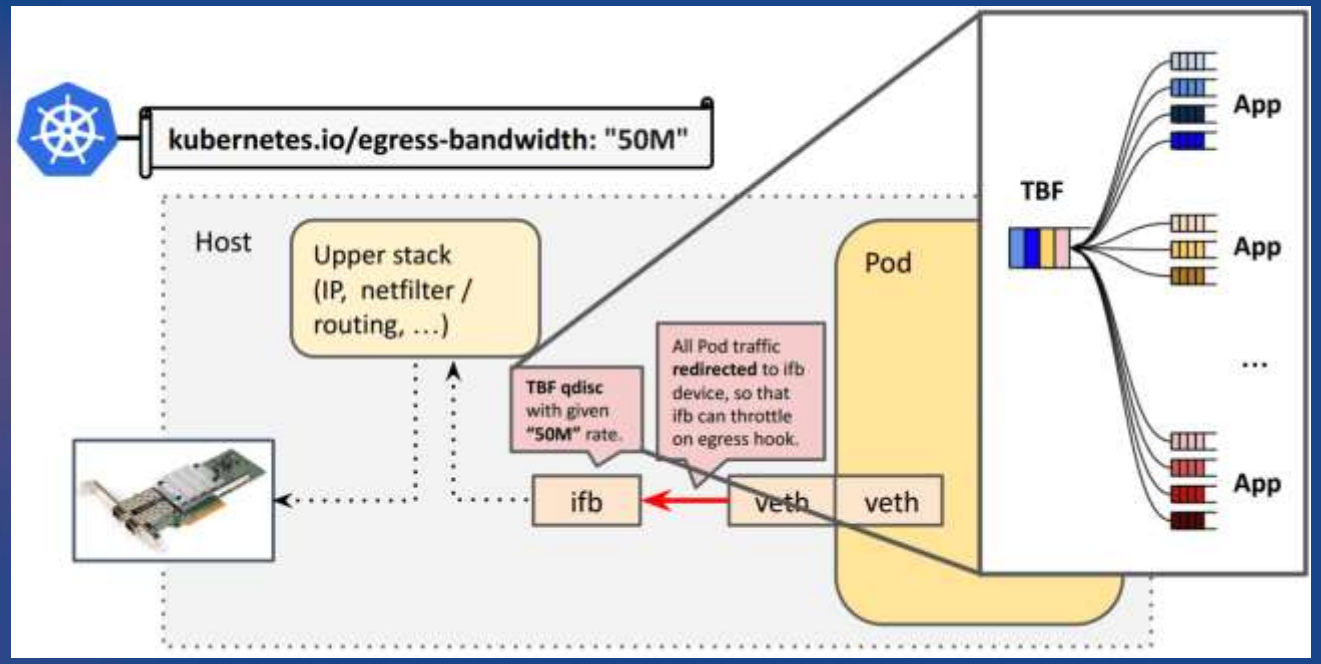
TC（traffic control）是Linux内核中的一个网络流量控制工具，它可以用来控制网络流量的带宽、延迟、丢包等参数，从而实现网络流量的优化和管理。

TC的基本功能：
shaping：出方向流量进行限速
scheduling：对报文进行调度
policing：入方向流量进行限速
dropping：超过预设值的流量都会被直接丢弃（ingress/egress）

TC的关键组件：
- qdisc 队列规则(queueing discipline)
  - classless qdisc：pfifo | TBF | ingress...
  - classful qdisc：HTB | prio | ...
- classes
  - 对于classful qdisc可以配置不同类别，并通过优先级实现优先处理某些队列流量
- filters
  - 通过filter来决定流量进入到哪个队列中
  - 类别： u32 | bpf | cgroup | ...

https://man7.org/linux/man-pages/man8/tc.8.html

k8s 带宽管理：基于TBF实现网络QoS限速




```
apiVersion: v1
kind: Pod
  metadata:
    name: iperf-slow
  annotations:
    kubernetes.io/ingress-bandwidth: 10M
    kubernetes.io/egress-bandwidth: 50M
```

带宽管理插件：配置pod annotation，并通过 TC TBF 实现限速。

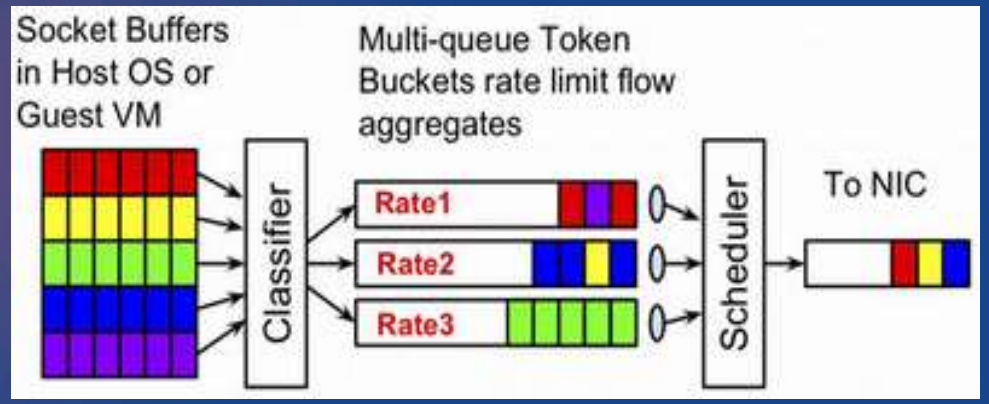
方案原理：

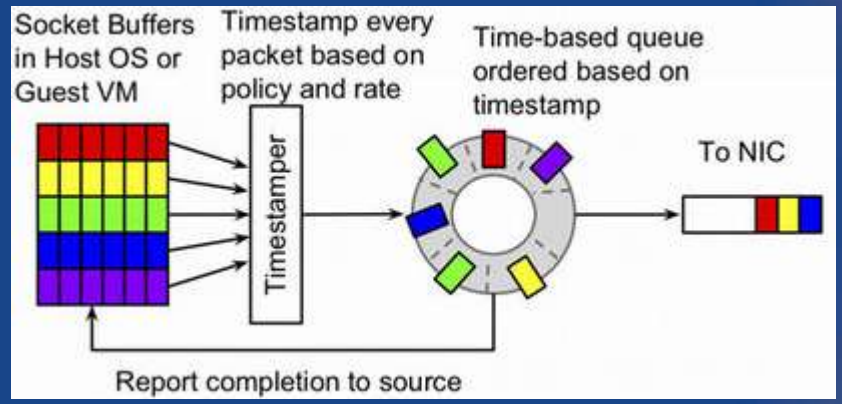
- Pod的egress对应veth host侧的ingress，ingress无法做流量整形，因此加了ifb网卡；
- Pod流量出来后重定向到ifb网卡，通过fib TBF qdisc实现限速；

方案问题：

- 锁竞争：TBF qdisc 所有 CPU 共享一个锁（qdisc root lock），因此存在锁竞争；流量越大锁开销越大；
- 缓冲区膨胀：多了一层ifb网卡排队，缓冲区增大；
- 时延增加：原来veth pair网卡对转发，现在多了个ifb；

From Queues to Earliest Departure Time



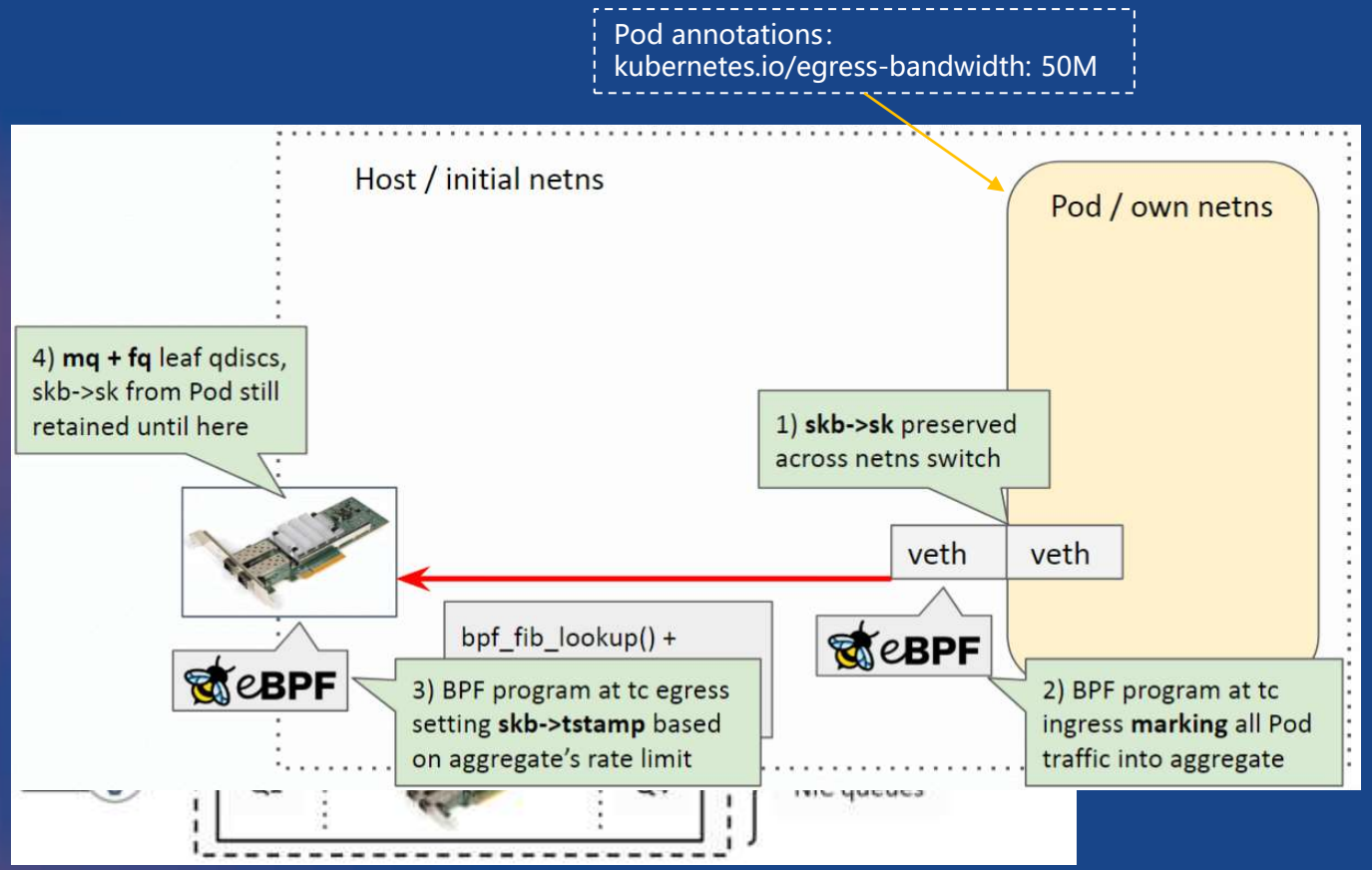


复杂、脆弱、级联队列
- 维护队列的CPU开销
- multi-cpu间共享队列的竞争开销

更好的CPU效率 & 更小的队列
- 每个skb可以设置最早离开时间(EDT)
- 时间轮调度

4.20: The TCP stack switched to Early Departure Time
From Queues to Earliest Departure Time:
https://documents.pub/document/oct-2018-david-wetherall-presenter-nandita-dukkipati-talks2018davidwetherall.html?page=12

cilium: eBPF + EDT实现Pod出方向带宽管理



cilium egress限速工作原理：

- Pod veth主机侧在tc ingress处挂载bpf prog，设置ctx->queue_mapping = aggregate
- 主机网卡侧在tc egress处挂载bpf prog，根据限速配置调整 skb->tstamp
- mq + fq按时间戳公平调度skb发包

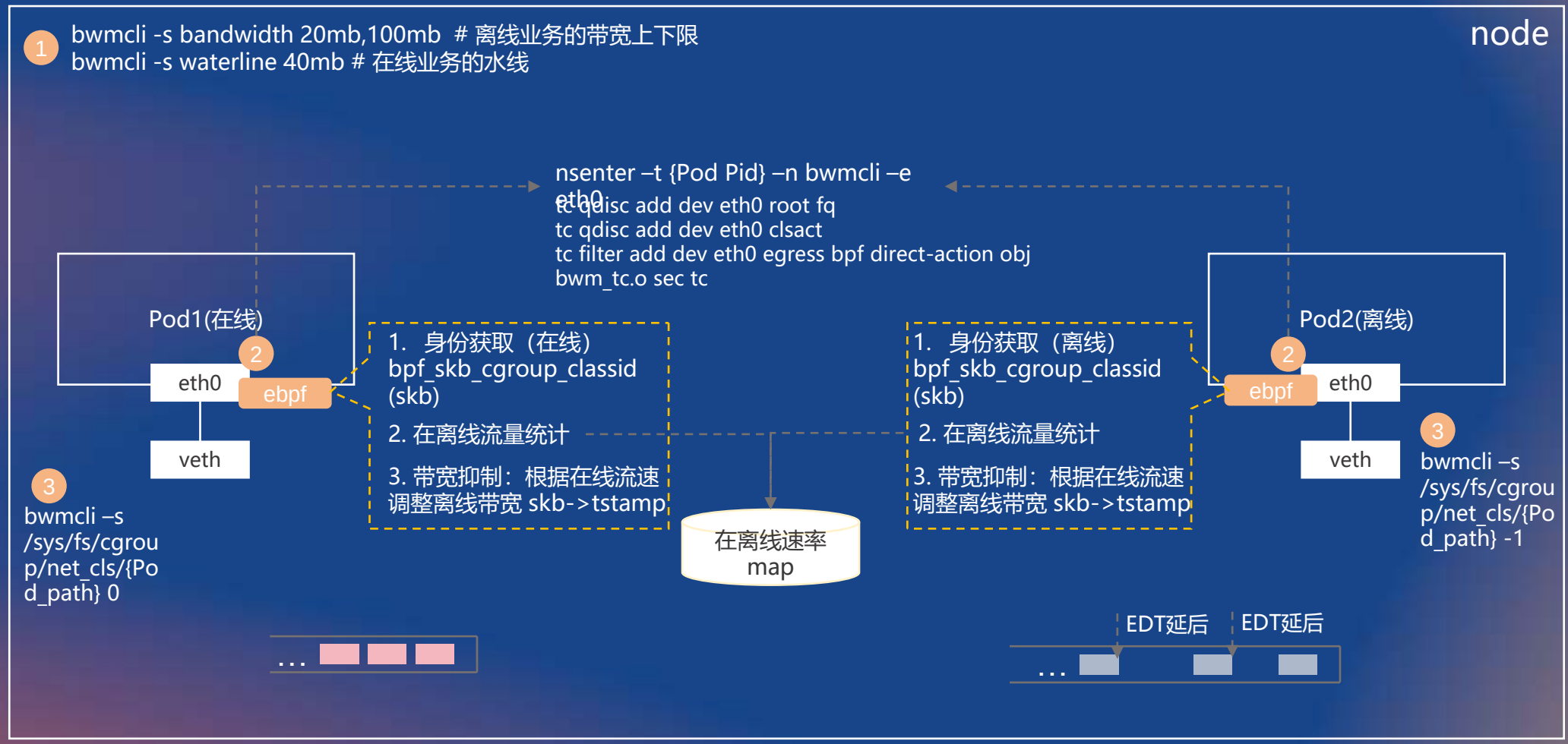
方案特点：

- 相比tc tbf等，实现免锁限速功能
- 多队列处理，避免缓存区膨胀

cilium启动使能带宽管理功能：

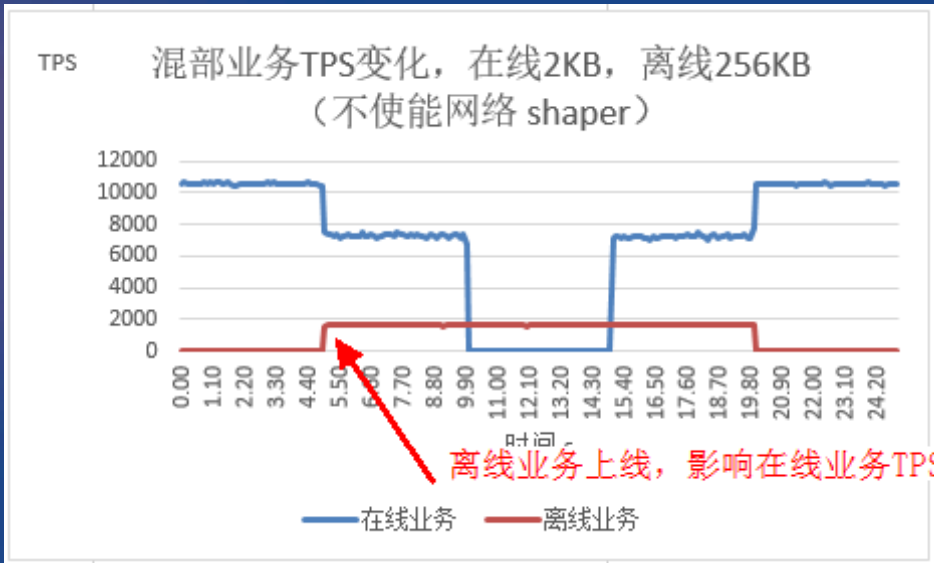
```
helm upgrade cilium cilium/cilium --version 1.13.4 \
--namespace kube-system \
--reuse-values \
--set bandwidthManager.enabled=true  # 仅支持出方向带宽管理
```

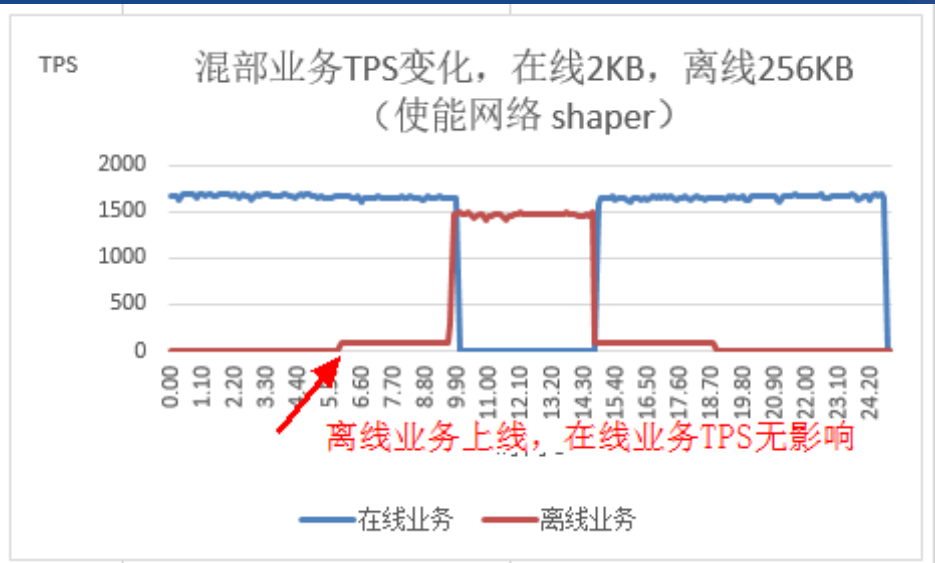
bwm：基于eBPF + EDT实现带宽抢占



- tc ebpf免锁流控
- node内共享在离线带宽水线
- EDT限速，实时带宽保障

技术效果：混部下高优先级应用带宽抢占时延<100ms





Volcano出口带宽保障

Volcano是基于k8s的批处理平台，可提供高性能任务调度引擎、高性能异构芯片管理、高性能任务运行管理等通用计算能力。

1. CCE控制台， 节点管理节点池中添加高级配置标签：volcano.sh/colocation=true

2. 插件中心 开启 在离线业务混部参数：

https://support.huaweicloud.com/usermanual-cce/cce_10_0701.html

3. 开启出口带宽保障，并配置带宽保障参数

```
kubectl edit configmap -nkube-system volcano-agent-configuration
...
data:
  colocation-config: |
    {
        "globalConfig":{
            "cpuBurstConfig":{
                "enable":true
            },
            "networkQosConfig":{
                "enable":true,
                "onlineBandwidthWatermarkPercent":80,
                "offlineLowBandwidthPercent":10,
                "offlineHighBandwidthPercent":40
            },
```

4. 业务部署时配置为离线作业，默认在线作业

```
kind: Deployment
apiVersion: apps/v1
spec:
  replicas: 4
  template:
    metadata:
      annotations:
        volcano.sh/qos-level: "-1"      # 离线作业标签
```

bwm后续演进计划


2024.05
2024.06
2024.09
2024.12
多优先级带宽管理
主机进程的带宽管理
低优先级流量反压
集成进Kmesh
流量治理引擎

欢迎关注社区

bwm：网络带宽管理
https://gitee.com/openeuler/oncn-bwm

Kmesh：内核级流量治理引擎
https://github.com/kmesh-net/kmesh