Corigine
芯启源

Hardware Offload of QoS

Simon Horman — OVS+OVN 2022 Fall Conference

Agenda

- Backgrounder: rate-limiting
- Modeling OpenFlow Metering with Policer Action
- Open Issues

Backgrounder: Rate-Limiting

Open vSwitch Rate Limiting

Corigine

- Oldest QoS feature in OVS?
- Limit applied to packets received by an OVS port
  - bits/s and burst
  - packets/s and burst (since OVS v2.16 and Kernel v5.13)
- Over-limit packets are dropped
- Configured via OVSDB

```
ovs-vsctl set interface $dev ingress_policing_rate=1000
ovs-vsctl set interface $dev ingress_policing_burst=100

ovs-vsctl set interface $dev ingress_policing_kpkts_rate=2000
ovs-vsctl set interface $dev ingress_policing_kpkts_burst=400
```

TC Policer

- For kernel datapath TC police action is used to implement OVS rate limiting
- Also provides hardware offload facility

```
tc filter add dev $dev ingress protocol all \
    matchall \
    action police pkts_rate 2000 pkts_burst 400 \
    action police rate 1Mbit burst 125Kb
```

Open Flow Meters meet the TC Police Action

Open Flow Meters

Corigine

- n-band

- Each band has a target rate
  - bytes/s or packets/s
- Drop or DSCP mark packets

- Are independent of flows
  - May be used by zero or more flows via meter action
- Are independent of ports
  - May be used in flows attached to different ports

Meter Example

```
ovs-ofctl -O openflow13 add-meter br0  \
    meter=99,kbps,burst,band=type=drop,rate=1000,burst_size=100
for dev in $DEV1 $DEV2; do
    ovs-ofctl -O openflow13 add-flow br0 \
        in_port=$dev,ip,action=meter:99,output=$DEV3 \
    ovs-ofctl -O openflow13 add-flow br0 \
        in_port=$dev,ipv6,action=meter:99,output=$DEV3 \
done
```

Meters and Police Action

Corigine

Open Flow Meters

- n-band
- Each band has a target rate
  - bits/s or packets/s
- Drop or DSCP mark packets
- Are independent of flows
  - May be used by zero or more flows via meter action
- Are independent of ports
  - May be used in flows attached to different ports

TC Police Action

- 1-band
  - multiple bands via multiple actions
- Each instance has a target rate
  - bits/s or packets/s
- May drop packets
  - Or mark with assistance from other actions
- May be independent of filters
  - May be used by zero or more flows using index of action instance
- May be independent of netdevs
  - May be used by filters attached to different ports

Wow, that's a pretty good match!

TC Action Independent of Filters

Corigine

```
tc action add action police index 99 rate 1mbit burst 100k
for dev in $DEV1 $DEV2; do
    tc filter add dev $dev ingress protocol ip \
        flower ip_proto tcp action police index 99
    tc filter add dev $dev ingress protocol ipv6 \
        flower ip_proto tcp action police index 99
done
```

Hardware Offload

Corigine

Offload via Flow Rule API

- New tc_setup_type
  - TC_SETUP_ACT
  - First class citizen along with U32, Flower, Block, ...
- Hook into cls_api and act_api for action:
  - Creation/Modification
  - Deletion
  - Statistics

Status of Meters via TC

Corigine

Initial upstreaming complete

- Kernel
  - cls_api and act_api (v5.17)
  - Flow rule (v5.17)
  - NFP driver (v5.18)
- OVS
  - Nvidia implementation included in v3.0

… But there are some issues remaining

Open Issues

Open Issues

- Software Datapath and Offload Policy
- Statistics
- Police Action Eviction

Software Datapath and Offload Policy

Corigine

Problem

- Offload flags (skip_sw/skip_hw) of TC actions and rules must match
- TC police action instances are created with no flags set
- Rules are created with flags of offload policy
    - Default is not set
    - But other options are possible, f.e. offload=true, tc-policy=skip_hw
    - Which create a missmatch
- If there is a missmatch then rules will not be added to TC
    - Flows will not be offloaded to TC datapath

Proposal

- Create TC police action instances with offload flags of offload policy

Ref: [PATCH v2] netdev-linux: Allow meter to work in tc software datapath when tc-policy is specified
https://mail.openvswitch.org/pipermail/ovs-dev/2022-October/398721.html

Statistics - Problem

Corigine

Requirement

- Count the packets flowing through a meter on a per-flow basis
  - In Open Flow: Meter != Meter Action

Flows with meters offloaded to hardware have incorrect statistics

- In v3.0.0 flow statistics were based on the police action
  - But this may be shared between flows (overcounting)
- In v3.0.1 flow statistics are based on action following police action
  - Not incremented for packets dropped by police action (undercounting)

Statistics - Proposals

- On ovs-dev ML
  - Record statistics from dummy action placed before police action
  - e.g. gact with PASS as control

- Alternate proposal briefly discussed at TC Workshop at Netdev 0x16
  - Record statistics of TC rule (flow) itself

Police Action Eviction

Problem

- Stale police action instances may be left in TC datapath

Proposals

- Revalidate (all) police action instances
  - Patch posted
  - But this is expensive
- Track instances that couldn't be deleted and only revalidate those
- Track deletion of actions that use meter

Ref: [PATCH v2] dpif-netlink: add revalidator for offload of meters
https://mail.openvswitch.org/pipermail/ovs-dev/2022-October/398366.html


THANKS