Hardware Offloading Kernel Bridging/Routing for EVPN using TC

Rohith Basavaraja, Roopa Prabhu, Jul 2021

Sample Config..

#Bridge

```
auto bridge
iface bridge
    bridge-ports vni10 vni20 vniRED pf0vf0 pf0vf1 pf1vf0
    bridge-vids 10 20 30
    bridge-vlan-aware yes
    bridge-pvid 1
```

#Bridge Ports

```
auto pf0vf0
iface pf0vf0
     bridge-pvid 10

auto pf0vf1
iface pf0vf1
     bridge-pvid 10

auto pf1vf0
iface pf1vf0
     bridge-pvid 20
```

#SVIs

```
auto vlan10
iface vlan10
    address 10.1.10.1/24
    vrf RED
    vlan-raw-device bridge
    vlan-id 10

auto vlan20
iface vlan20
    address 10.1.20.1/24
    vrf RED
    vlan-raw-device bridge
    vlan-id 20

auto vlan4001
iface vlan4001
    vrf RED
    vlan-raw-device bridge
    vlan-id 4001
```

#L2 VNI

```
auto vni10
iface vni10
    bridge-access 10
    mstpctl-portbpdufilter yes
    mstpctl-bpduguard yes
    bridge-learning off
    bridge-arp-nd-suppress on
    vxlan-id 10
```

#L3 VNI

```
auto vniRED
iface vniRED
    bridge-access 4001
    vxlan-id 4001
    mstpctl-portbpdufilter yes
    mstpctl-bpduguard yes
    bridge-learning off
    bridge-arp-nd-suppress on
    mtu 9152
```

EVPN Bridging Entries

#FDB Entries

```
44:38:39:00:00:0a dev vni10 vlan 10 extern_learn master bridge
44:38:39:00:00:0a dev vni10 dst 10.10.10.1 self
7a:5c:43:c2:e7:15 dev pf0vf0 vlan 10 master bridge
ee:15:40:f1:9a:68 dev pf0vf1 vlan 10 master bridge
```

Resulting TC entries

Encap/Tx Direction

```
filter block 1 protocol ip pref 100 flower chain 0 handle 0x2be
  indev pf0vf0
  dst_mac 44:38:39:00:00:0a
  eth_type ipv4
  in_hw in_hw_count 1
        action order 1: tunnel_key  set
        src_ip 10.10.10.3
        dst_ip 10.10.10.1
        key_id 10
        dst_port 4789
        csum pipe
         index 17 ref 1 bind 1 installed 15 sec used 15 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0
        no_percpu

        action order 2: mirred (Egress Redirect to device vni10) stolen
        index 17 ref 1 bind 1 installed 15 sec used 0 sec
        Action statistics:
        Sent 1470 bytes 15 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 1470 bytes 15 pkt
        backlog 0b 0p requeues 0

filter block 1 protocol ip pref 100 flower chain 0 handle 0x2bd
  indev pf0vf1
  dst_mac 44:38:39:00:00:0a
  eth_type ipv4
  in_hw in_hw_count 1
        action order 1: tunnel_key  set
        src_ip 10.10.10.3
        dst_ip 10.10.10.1
        key_id 10
        dst_port 4789
        csum pipe
         index 16 ref 1 bind 1 installed 15 sec used 15 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0
        no_percpu

        action order 2: mirred (Egress Redirect to device vni10) stolen
        index 16 ref 1 bind 1 installed 15 sec used 0 sec
        Action statistics:
        Sent 1372 bytes 14 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 1372 bytes 14 pkt
        backlog 0b 0p requeues 0
        no_percpu
        no_percpu
```


NVIDIA.

Resulting TC entries

Decap/Rx Direction

```
filter protocol ip pref 100 flower chain 0 handle 0x2d2
  dst_mac 7a:5c:43:c2:e7:15
  eth_type ipv4
  enc_dst_ip 10.10.10.3
  enc_key_id 10
  enc_dst_port 4789
  in_hw in_hw_count 2
        action order 1: tunnel_key  unset pipe
         index 8 ref 1 bind 1 installed 12 sec used 12 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: mirred (Egress Redirect to device pf0vf0) stolen
        index 15 ref 1 bind 1 installed 12 sec used 0 sec
        Action statistics:
        Sent 1176 bytes 12 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 1176 bytes 12 pkt
        backlog 0b 0p requeues 0
        no_percpu

filter protocol ip pref 100 flower chain 0 handle 0x2bc
  dst_mac ee:15:40:f1:9a:68
  eth_type ipv4
  enc_dst_ip 10.10.10.3
  enc_key_id 10
  enc_dst_port 4789
  in_hw in_hw_count 2
        action order 1: tunnel_key  unset pipe
         index 1 ref 1 bind 1 installed 15 sec used 15 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: mirred (Egress Redirect to device pf0vf1) stolen
        index 1 ref 1 bind 1 installed 15 sec used 0 sec
        Action statistics:
        Sent 1470 bytes 15 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 1470 bytes 15 pkt
        backlog 0b 0p requeues 0
        no_percpu
```


NVIDIA.

Issues with translating EVPN FDB entries to tc rules..

- Can’t share FDB entries across bridge members ports of same vlan

For eg: In TX (Encap) direction for FDB entries 44:38:39:00:00:0a, tc entries has to be replicated for all the bridge members belonging to same vlan

In Rx (Decap) direction we need to create separate decap entries for each of the local MACs

- Implicit tc rule generation is required.

For eg: When local MAC are learnt we need to implicitly create and install decap entries.

NVIDIA.

EVPN Route Entries

#EVPN Routes

```
10.1.10.101 via 10.10.10.1 dev vlan4001 table RED proto bgp metric 20 onlink
10.1.20.102 via 10.10.10.2 dev vlan4001 table RED proto bgp metric 20 onlink
```

#Neigh entries

```
10.1.10.201 dev vlan10 lladdr 7a:5c:43:c2:e7:15 REACHABLE
10.1.10.202 dev vlan10 lladdr ee:15:40:f1:9a:68 REACHABLE
10.1.20.203 dev vlan20 lladdr 7a:37:da:3c:d1:61 REACHABLE
```

Resulting TC entries

Encap/Tx Direction

```
filter block 1 protocol ip pref 50001 flower chain 0 handle 0x2bd
  indev pf0vf0
  eth_type ipv4
  dst_ip 10.1.10.101
  in_hw in_hw_count 1
        action order 1:  pedit action pipe keys 4
         index 5 ref 1 bind 1 installed 48 sec used 48 sec
         key #0  at eth+4: val 00003ec7 mask ffff0000
         key #1  at eth+8: val d53f1b0a mask 00000000
         key #2  at eth+0: val 44383900 mask 00000000
         key #3  at eth+4: val 00090000 mask 0000ffff
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: tunnel_key  set
        src_ip 10.10.10.3
        dst_ip 10.10.10.1
        key_id 4001
        dst_port 4789
        csum pipe
         index 5 ref 1 bind 1 installed 48 sec used 48 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0
        no_percpu

        action order 3: mirred (Egress Redirect to device vniRED) stolen
        index 5 ref 1 bind 1 installed 48 sec used 0 sec
        Action statistics:
        Sent 4704 bytes 48 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 4704 bytes 48 pkt
        backlog 0b 0p requeues 0
        no_percpu
```

```
filter block 1 protocol ip pref 50001 flower chain 0 handle 0x2be
  indev pf1vf0
  eth_type ipv4
  dst_ip 10.1.10.101
  in_hw in_hw_count 1
        action order 1:  pedit action pipe keys 4
         index 6 ref 1 bind 1 installed 48 sec used 48 sec
         key #0  at eth+4: val 00003ec7 mask ffff0000
         key #1  at eth+8: val d53f1b0a mask 00000000
         key #2  at eth+0: val 44383900 mask 00000000
         key #3  at eth+4: val 00090000 mask 0000ffff
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: tunnel_key  set
        src_ip 10.10.10.3
        dst_ip 10.10.10.1
        key_id 4001
        dst_port 4789
        csum pipe
         index 6 ref 1 bind 1 installed 48 sec used 48 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0
        no_percpu

        action order 3: mirred (Egress Redirect to device vniRED) stolen
        index 6 ref 1 bind 1 installed 48 sec used 0 sec
        Action statistics:
        Sent 4704 bytes 48 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 4704 bytes 48 pkt
        backlog 0b 0p requeues 0
        no_percpu
```

Resulting TC entries

Decap/Rx Direction

```
filter protocol ip pref 50001 flower chain 0 handle 0x2d7
  eth_type ipv4
  dst_ip 10.1.10.201
  enc_dst_ip 10.10.10.3
  enc_key_id 4001
  enc_dst_port 4789
  in_hw in_hw_count 2
        action order 1:  pedit action pipe keys 4
         index 15 ref 1 bind 1 installed 30 sec used 30 sec
         key #0  at eth+4: val 00003ec7 mask ffff0000
         key #1  at eth+8: val d53f1b0a mask 00000000
         key #2  at eth+0: val 7a5c43c2 mask 00000000
         key #3  at eth+4: val e7150000 mask 0000ffff
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: tunnel_key  unset pipe
         index 9 ref 1 bind 1 installed 30 sec used 30 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 3: mirred (Egress Redirect to device pf0vf0) stolen
        index 20 ref 1 bind 1 installed 30 sec used 30 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0
        no_percpu
```

```
filter protocol ip pref 50001 flower chain 0
filter protocol ip pref 50001 flower chain 0 handle 0x2c3
  eth_type ipv4
  dst_ip 10.1.20.203
  enc_dst_ip 10.10.10.3
  enc_key_id 4001
  enc_dst_port 4789
  in_hw in_hw_count 2
        action order 1:  pedit action pipe keys 4
         index 7 ref 1 bind 1 installed 31 sec used 31 sec
         key #0  at eth+4: val 00003ec7 mask ffff0000
         key #1  at eth+8: val d53f1b0a mask 00000000
         key #2  at eth+0: val 7a37da3c mask 00000000
         key #3  at eth+4: val d1610000 mask 0000ffff
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 2: tunnel_key  unset pipe
         index 2 ref 1 bind 1 installed 31 sec used 31 sec
        Action statistics:
        Sent 0 bytes 0 pkt (dropped 0, overlimits 0 requeues 0)
        backlog 0b 0p requeues 0

        action order 3: mirred (Egress Redirect to device pf1vf0) stolen
        index 8 ref 1 bind 1 installed 31 sec used 0 sec
        Action statistics:
        Sent 2156 bytes 22 pkt (dropped 0, overlimits 0 requeues 0)
        Sent software 0 bytes 0 pkt
        Sent hardware 2156 bytes 22 pkt
        backlog 0b 0p requeues 0
        no_percpu
```


NVIDIA.

Issues with translating EVPN routes to tc rules..

- Similar to bridge entries for l3 entries/routes tc entries (translation for routes) can’t be shared across interfaces belonging to same vrf.
- Again, similar to bridge entries, we need implicit tc rule generation is required. For eg: for every local routes we need to add decap tc rules.

NVIDIA.

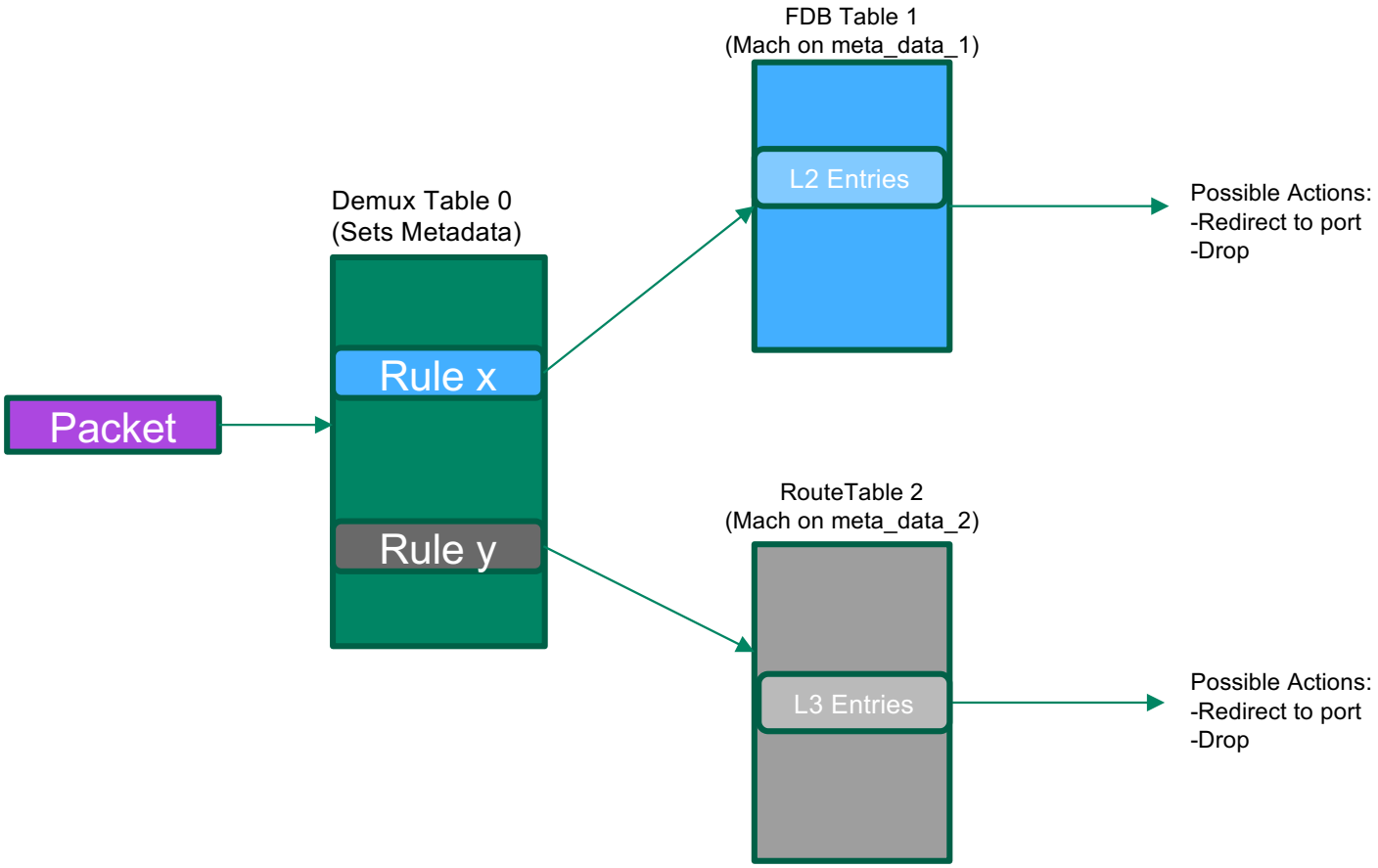
Possible Options to improve

- Add support for metadata. To have support for Action to set metadata and also to use as flow rule match pattern. This will also require splitting rules. So we need analyse trade-off between scale/simplicity Vs Latency.

- Removing the constraints that decap action should be part of terminal action.

NVIDIA.

Possible Options to improve with meta data



Possible Options to improve with meta data

Sample TC Rules
==========
L3
==

tc filter add block 1 pref 3 protocol 802.1Q flower indev dst_mac 00:05:00:01:00:0a pf0vf1 vlan_id 1004 action **set_meta_1 0xfad** action goto chain 10001

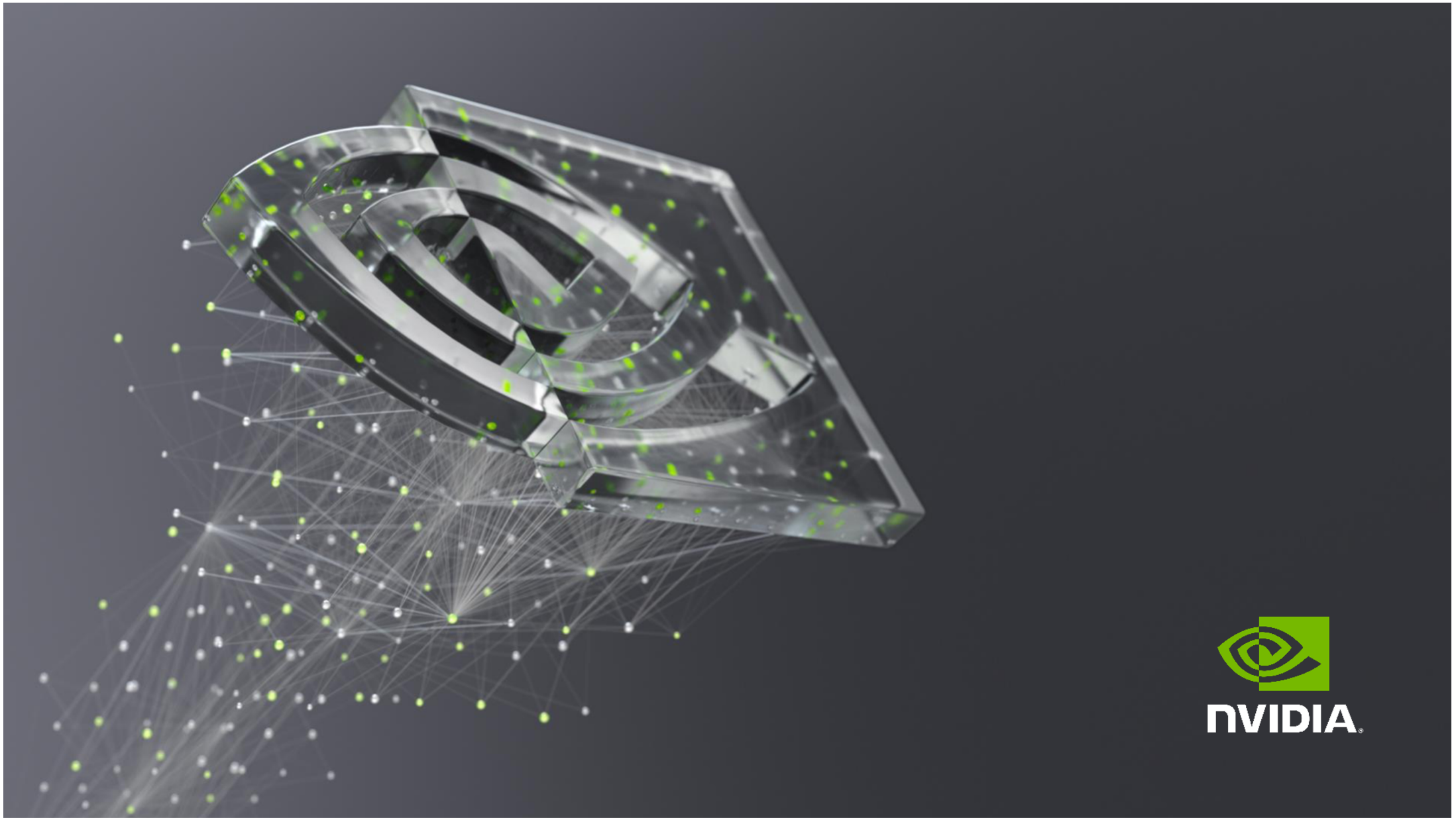
tc filter add block 1 chain 10001 pref 500 protocol ip flower **meta_1 0xfad** dst_ip 10.1.10.101 action pedit ex munge eth src set 3a:16:1c:9f:18:17 pipe action pedit ex munge eth dst set 00:02:00:00:00:2d pipe action tunnel_key set id 4001 src_ip 10.10.10.3 dst_ip 10.10.10.1 dst_port 4789 pipe action mirred egress redirect dev vniRED

L2
===

tc filter add block 1 pref 6 protocol 802.1Q flower indev pf0vf1 vlan_id 1002 action **set_meta_2 0x568** goto chain 1004

tc filter add block 1 chain 1002 pref 101 flower **meta_2 0x568** dst_mac 00:03:00:00:00:09 action vlan pop pipe action tunnel_key set id 1002 src_ip 27.0.0.18 dst_ip 27.0.0.13 dst_port 4789 pipe action mirred egress redirect dev vni10

NVIDIA.


NVIDIA