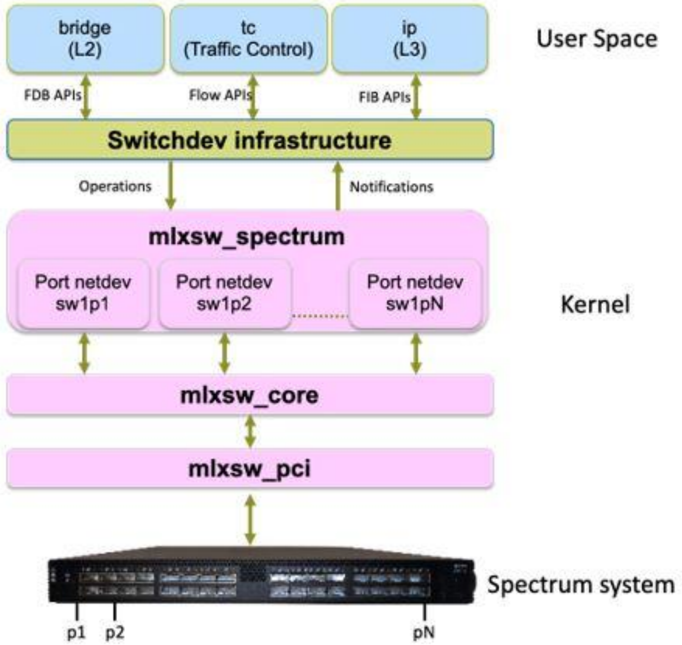
# Mlxtoolkit:
## A Tool to Ease Life With Switchdev

Alexander Zubkov
Qrator Labs CZ s.r.o.
green@qrator.net

# Switchdev

- in-kernel infrastructure
  - since 2014
- map dataplane to Linux (offload)
  - bridging
  - routing
  - filtering



Courtesy of Mellanox Technologies
https://blog.mellanox.com/2018/12/mellanox-spectrum-linux-switch-powered-by-switchdev/

# Linux toolkit

- iproute2
  - ip
  - bridge
  - devlink
  - tc
- sysctl
- ethtool
- …

# Switch configuration

route

ip

vrf

vlan

bridge

bond

port (interface)

- startup
  - big script
  - somewhat manageable
- runtime changes
  - complex
  - human errors
- not the worst part

# tc filter show

```
filter protocol ip pref 1 flower chain 101
filter protocol ip pref 1 flower chain 101 handle 0x1
  eth_type ipv4
  ip_proto icmp
  in_hw in_hw_count 1
    action order 1: gact action pass
     random type none pass val 0
     index 1 ref 1 bind 1
    used_hw_stats immediate

filter protocol ip pref 2 flower chain 101
filter protocol ip pref 2 flower chain 101 handle 0x1
  eth_type ipv4
  ip_proto tcp
  in_hw in_hw_count 1
    action order 1: gact action pass
     random type none pass val 0
     index 2 ref 1 bind 1
    used_hw_stats immediate

...
```

# mlxtoolkit

- https://gitlab.com/qratorlabs/mlxtoolkit
- MIT license
- Perl
- mlxrtr
  - interfaces
  - routing
- mlxacl
  - ACL
  - shared acl
  - chain per vlan

# mlxtoolkit

## mlxrtr

```
[port 1]
split 4
[bond srv1]
slave port1/0, port1/1
[bond srv2]
slave port1/2, port1/3
[vlan 10]
native port2
vrf ext
ip 192.0.2.2/31
[vlan 20]
tag bond srv1, bond srv2
vrf int
ip 198.51.100.1/24
[loopback 10]
vrf ext
ip 192.0.2.1/32
```

```
[vrf ext]
table 100
route 0.0.0.0/0 via 198.51.100.2 dev vlan20
[vrf int]
table 200
route 0.0.0.0/0 via 192.0.2.3 dev vlan10
route 203.0.113.0/24 via 198.51.100.2 dev vlan20
```

## mlxacl

```
[vlan10]
ip_proto icmp dst_ip 192.0.2.2 action pass
src_ip 203.0.113.0/24 action drop
dst_ip 203.0.113.0/24 action goto [ex1]
dst_ip 203.0.113.0/24 action drop
action pass
[ex1]
ip_proto icmp action pass
ip_proto tcp action pass
action drop
```

# mlxrtr (init)

```
sysctl -w ...
...
ip rule del pref 0
ip rule add pref 30000 table local
...
devlink port split pci/0000:01:00.0/25 count 4
ip link set dev enp1s0np1s0 down
ip link set dev enp1s0np1s0 name port1-0
ip link set dev port1-0 up
...
ip link set dev enp1s0np10 name port10
ip link set dev enp1s0np11 name port11
...
tc qdisc add dev port1-0 ingress_block 100 ingress
tc qdisc add dev port1-1 ingress_block 100 ingress
```

sysctl

ip rule

port split

rename

prepare for acl

# mlxrtr (init)

```
ip link add name bond_srv1 type bond lacp_rate fast min_links 1 \
    mode 802.3ad xmit_hash_policy 'layer3+4'
ip link set dev bond_srv1 down
...
ip link add name loop10 type dummy
ip link set dev loop10 down
ip link add name switch type bridge vlan_filtering 1
ip link set dev switch down
ip link add name vrf-ext type vrf table 100
ip link set dev vrf-ext down
...
ip link set dev port1-0 down
ip link set dev port1-0 master bond_srv1
```

add bond

add “loopback”

add switch

add vrf

attach to bond

# mlxrtr (init)

```
ip link set dev port2 master switch
ip link set dev port2 down
...
ip link set dev loop10 master vrf-ext
ip link set dev loop10 down
ip link add link switch name vlan10 type vlan id 10
ip link set dev vlan10 down
...
ip link set dev vlan10 master vrf-ext
ip link set dev vlan10 down
...
bridge vlan del vid 1 dev port2
bridge vlan add vid 10 dev port2 pvid untagged
bridge vlan add vid 10 dev switch self
```

attach to switch

add L3 vlan

attach to vrf

map vlans

# mlxrtr (init)

```
ip link set dev port1-0 up
...
ip -4 address add 192.0.2.1/32 dev loop10
...
ip -4 route replace 0.0.0.0/0 metric 0 table 200 \
     proto static nexthop via 192.0.2.3 dev vlan10 weight 1
ip -4 route replace blackhole 0.0.0.0/0 metric 4278198272 table 200 \
     proto static
ip -4 route replace 203.0.113.0/24 metric 0 table 200 \
     proto static nexthop via 198.51.100.2 dev vlan20 weight 1
```

link up

add ip

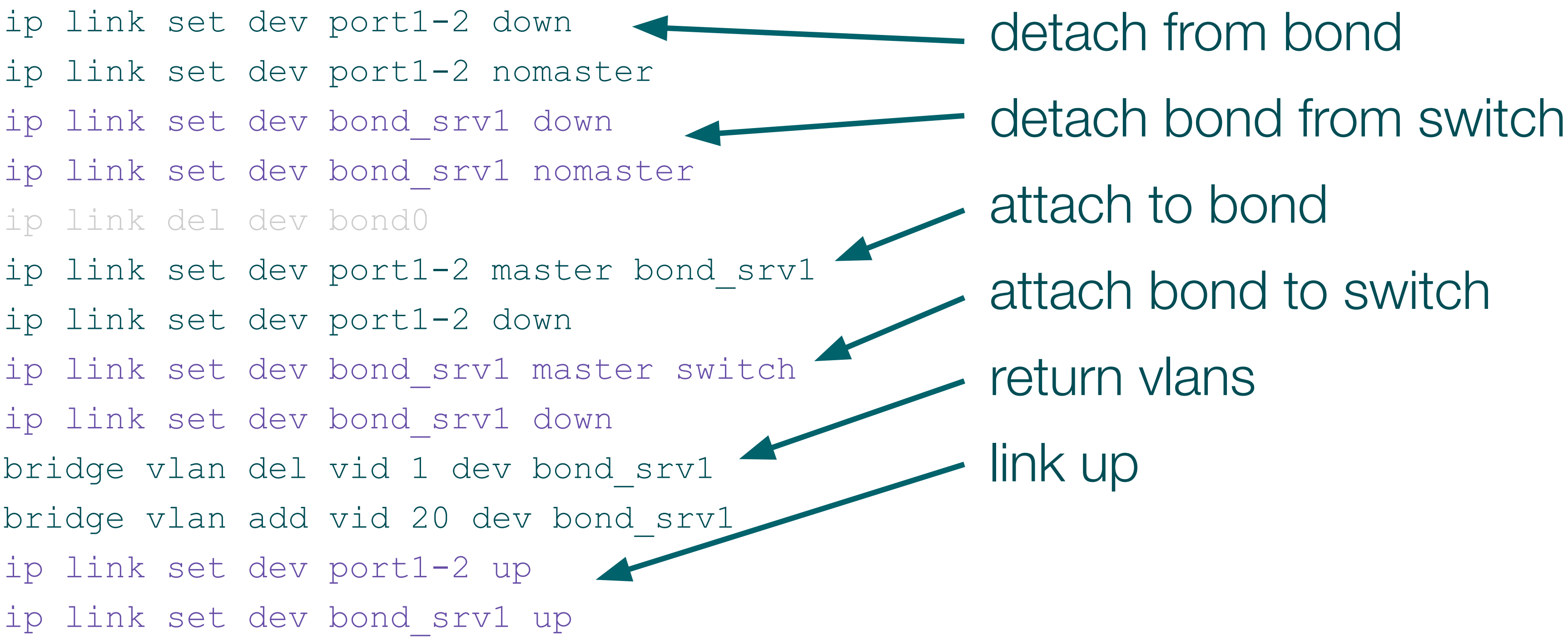
add routes

# mlxrtr (change)

## move port between bonds

```
 [port 1]
 split 4
 [bond srv1]
-slave port1/0, port1/1
+slave port1/0, port1/1, port1/2
 [bond srv2]
-slave port1/2, port1/3
+slave port1/3
 [vlan 10]
 native port2
 vrf ext
```

# mlxrtr (change)



```
ip link set dev port1-2 down
ip link set dev port1-2 nomaster
ip link set dev bond_srv1 down
ip link set dev bond_srv1 nomaster
ip link del dev bond0
ip link set dev port1-2 master bond_srv1
ip link set dev port1-2 down
ip link set dev bond_srv1 master switch
ip link set dev bond_srv1 down
bridge vlan del vid 1 dev bond_srv1
bridge vlan add vid 20 dev bond_srv1
ip link set dev port1-2 up
ip link set dev bond_srv1 up
```

detach from bond

detach bond from switch

attach to bond

attach bond to switch

return vlans

link up

# mlxacl (init)

```
tc filter add block 100 ...
protocol ip chain 101 pref 1 flower ip_proto icmp action pass
protocol ip chain 101 pref 2 flower ip_proto tcp action pass
protocol ip chain 101 pref 3 flower action drop
protocol ipv6 chain 101 pref 4 flower action drop
protocol ip chain 100 pref 1 flower ip_proto icmp \
    dst_ip 192.0.2.2 action pass
protocol ip chain 100 pref 2 flower src_ip 203.0.113.0/24 action drop
protocol ip chain 100 pref 3 flower dst_ip 203.0.113.0/24 action goto \
    chain 101
protocol ip chain 100 pref 4 flower dst_ip 203.0.113.0/24 action drop
protocol ip chain 100 pref 5 flower action pass
protocol ipv6 chain 100 pref 6 flower action drop
protocol 802.1q chain 0 pref 1000 flower vlan_id 10 action goto chain 100
protocol 802.1q chain 0 pref 1001 flower action pass
```

fill chains

match vlans

# mlxacl (change)

## delete one rule

```
 [vlan10]
-ip_proto icmp dst_ip 192.0.2.2 action pass
 src_ip 203.0.113.0/24 action drop
 dst_ip 203.0.113.0/24 action goto [ex1]
 dst_ip 203.0.113.0/24 action drop
```

# mlxacl (change)

```
tc filter add block 100 protocol ip chain 102 pref 1 \
    flower src_ip 203.0.113.0/24 action drop
tc filter add block 100 protocol ip chain 102 pref 2 \
    flower dst_ip 203.0.113.0/24 action goto chain 101
tc filter add block 100 protocol ip chain 102 pref 3 \
    flower dst_ip 203.0.113.0/24 action drop
tc filter add block 100 protocol ip chain 102 pref 4 flower action pass
tc filter add block 100 protocol ipv6 chain 102 pref 5 flower action drop
tc filter add block 100 protocol 802.1q chain 0 pref 1002 \
    flower vlan_id 10 action goto chain 102
tc filter add block 100 protocol 802.1q chain 0 pref 1003 \
    flower action pass
tc filter del block 100 chain 0 pref 1000
tc filter del block 100 chain 0 pref 1001
tc filter del block 100 chain 100
```

new chain

new vlan match

rearrange