eBPF for Switches

Using eBPF combined with Switchdev
Nick Viljoen (DRAFT)
Current SoA
Multi-Host NIC

L2 Switch

Host 0

Host 1

Host 2

Host 3

Key

Network device

PHY/MAC
Proposed Architecture

Host 0

Host 1

Host 2

Host 3

L2 Switch

PHY/MAC

DRAFT
Problems

- Debugging
- Consistency of Throughput/Latency
  - Qdisc Offload
- Offloads
  - To a fully programmable datapath
Debugging
Visibility (Lack of)
Proposed Architecture

Host 0

Host 1

Host 2

Host 3

L2 Switch

PHY/MAC

Configurable from host 0

Visible stats (read only mode From host 0)
Offloads-Conceptual architecture
Qdisc Offload

Host 0

Host 1

Host 2

Host 3

Qdisc (RED)

L2 Switch

PHY/MAC
Fully Programmable Datapath

Qdisc (RED)

Programmable RSS redirect()

Host 0

Host 1

Host 2

Host 3

XDP

PHY/MAC
Implementation

- Initialisation
- Switchdev Mode
- Qdisc Offload
- Next Steps
- Future Work
Initialisation

- **nfp_app.c**
  - ctrl.c
    - nfp_abm_ctrl_qm_disable
    - nfp_abm_ctrl_read_params
  - main.c
    - nfp_abm_vnic_alloc
    - nfp_abm_init
- **nfp_net_main.c**
  - nfp_net_pf_app_init
  - nfp_net_pf_alloc_vnics
  - nfp_net_pci_probe()
- **pci_epf_core.c**
  - probe
  - nfp_net_pci_probe()
Qdisc offload

Diagram:

- nfp_app.c
  - main.c
    - ctrl.c
      - nfp_mbox_cmd()
  - nfp_abm_setup_tc_red
  - nfp_abm_setup_tc_mq
  - struct nfp_qdisc_red
  - struct nfp_abm
  - struct nfp_abm_link
  - struct nfp_repr

Diagrams:
- tc qdisc
  - tc qdisc add (mq/red)
  - tc qdisc replace
  - tc qdisc del
  - tc qdisc show
  - tc -s qdisc show
- sch_red.c/sch_mq.c
- nfo_setup_tc
- Qdisc offload
- DRAFT
Addition of Classification

DRAFT

main.c
- ctrl.c
  - nfp_mbox_cmd

app->type->setup_tc

struct nfp_abm
- struct nfp_abm_link
- struct nfp_abm_link
- struct nfp_qdisc_clsact
  - nfp_abm_u32_knode_replace

nfp_app.c
- app->type->setup_tc

cls_u32.c

tc filter
- tc filter add (u32)
- tc filter replace
- tc filter del
- tc filter show

tc qdisc
- tc qdisc add (mq/red)
- tc qdisc replace
- tc qdisc del
- tc qdisc show
- tc -s qdisc show

sch_ingress.c/sch_red.c/sch_mq.c

ndo_setup_tc

Addition of Classification
Addition of BPF

- prog.o
- verifier.c
- offload.c
- nfp_app.c
- ctrl.c
- nfp_abm
  - nfp_abm_setup_tc_red
  - nfp_abm_setup_tc_mq
- nfp_abm_link
- nfp_abm_repr
- tc filter
  - tc filter add (u32/bpf)
  - tc filter replace
  - tc filter del
  - tc filter show
- tc qdisc
  - tc qdisc add (mq/red)
  - tc qdisc replace
  - tc qdisc del
  - tc qdisc show
- tc -s qdisc show
- sch_ingress.c/sch_red.c/sch_mq.c
- cls_bpf.c/cls_u32.c
- nfp_mbox_cmd()
- sch_ingress.c
- nfp_abm_setup_cls_block
- cls_bpf.c/cls_u32.c
- sch_red.c/sch_mq.c
- sch_mq.c
- sch_red.c
- sched.c
- nfp_bpf_jit
- nfp_bpf_jit_prepare
- nfp_bpf_jit
- nfp_bpf_translate
- nfp_bpf_map_alloc
- offload.c
- nfp_bpf_verifier_prep
- nfp_bpf_translate
- nfp_bpf_map_alloc
- nfp_bpf_ctrl_alloc_map
- nfp_bpf_jit
- nfp_bpf_jit_prepare
- nfp_bpf_jit
- nfp_bpf_translate
- nfp_bpf_map_alloc
- nfp_bpf_ctrl_alloc_map
- nfp_bpf_jit
- nfp_bpf_jit_prepare
- nfp_bpf_jit
- nfp_bpf_translate
- nfp_bpf_map_alloc
- nfp_bpf_ctrl_alloc_map
- nfp_mbox_cmd()
Summary

● The combination of switchdev and eBPF could be used to almost completely define the working of a multihost NIC
● Could be applied to other switches too
● Philosophical question of how far heterogeneous processing should go
  ○ However many things can be done before this point is reached
  ○ Cross applicable helpers for XDP
    ■ E.g Checksums