



Networking Track
Vancouver, 14 November 2018

eBPF & Switch Abstractions

Nick Viljoen <nick.viljoen@netronome.com>

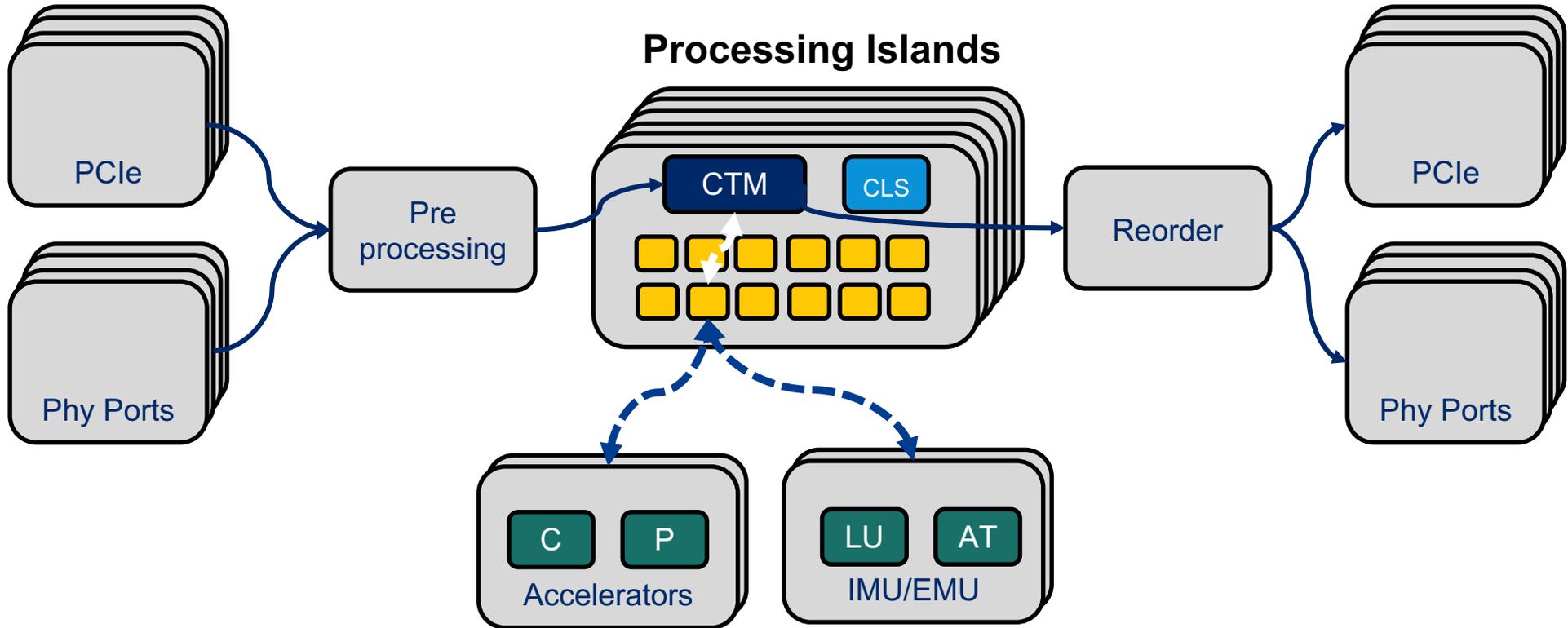
NETRONOME

- Background
- The Multi-Host NIC Abstraction
- The Switchdev Based Multi-Host NIC Abstraction
- Currently upstream (*as of 2 weeks ago*)
 - Boot
 - Setting Switchdev Mode
 - Loading Qdiscs
- Next Steps (*currently being upstreamed*)
 - Generalising Qdisc Offload
 - Adding clsact Qdiscs (u32, cls_bpf)
- Future Work

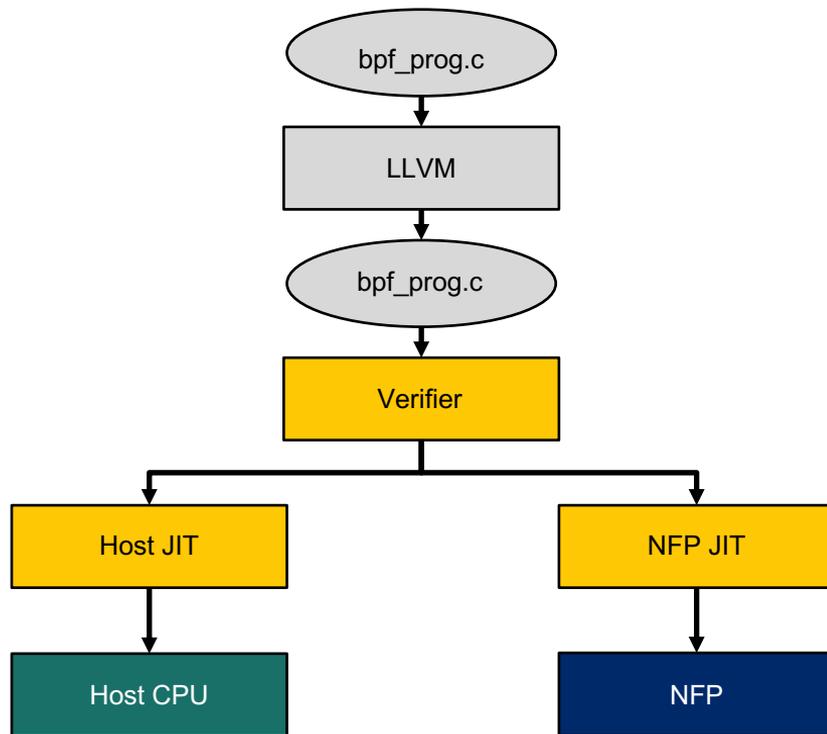
Background: HW, BPF JIT, NIC as a Switch

- Many core, fully programmable network processor
 - 48-96 preprocessing cores
 - 54-120 programmable cores, 8 threads per core, MIMT
 - up to 4 PCIe
 - Up to 40 ports supported
 - ~17MB of on chip memory
 - 2-24GB of DRAM
 - Distributed & Transactional memory architecture
- Low power
 - ~ 10-35W (dependent on chip + frequency)

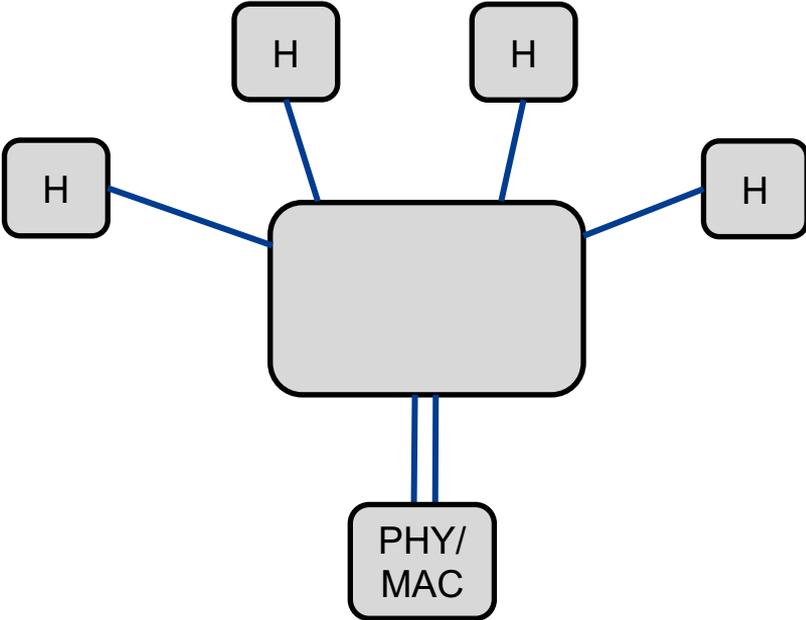
= Flow Processing Core (FPC)



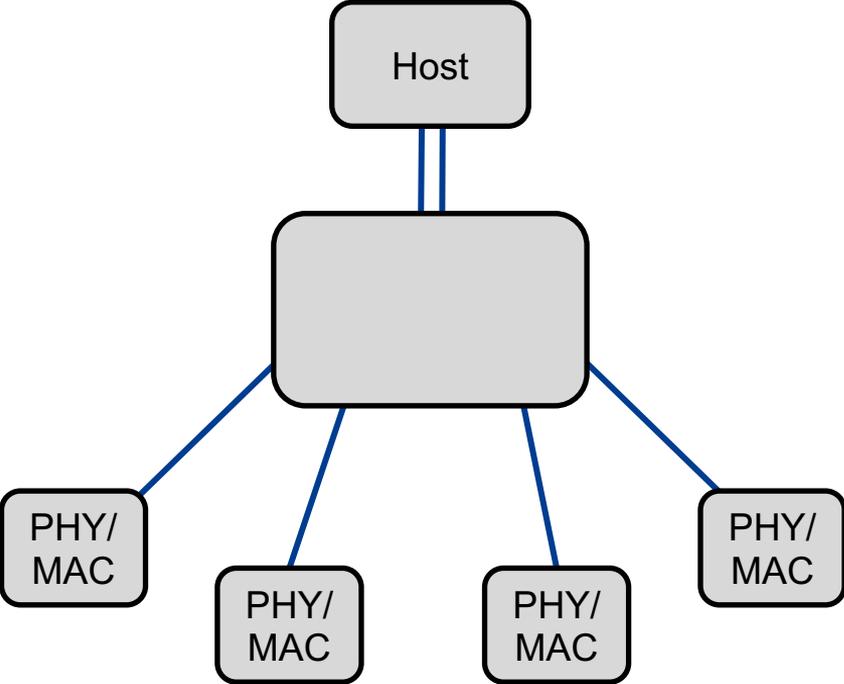
- Program is written in standard manner
- LLVM compiled as normal
- The nfp's jit is called like any other architectures jit
- This converts the BPF bytecode to NFP machine code
- Translation reuses the verifier infrastructure in kernel
- Defining the FPC datapath code using BPF



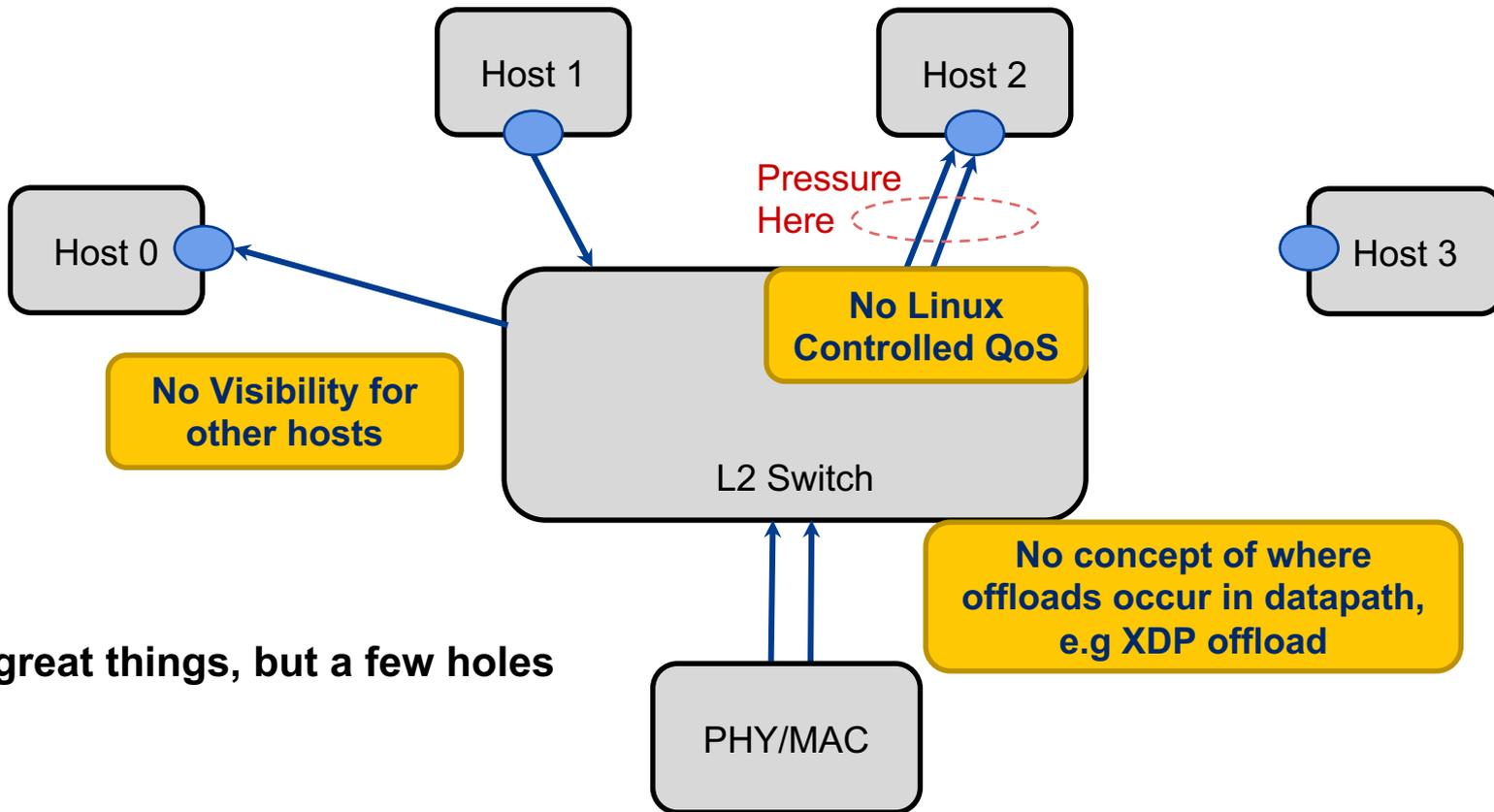
Multi-Host



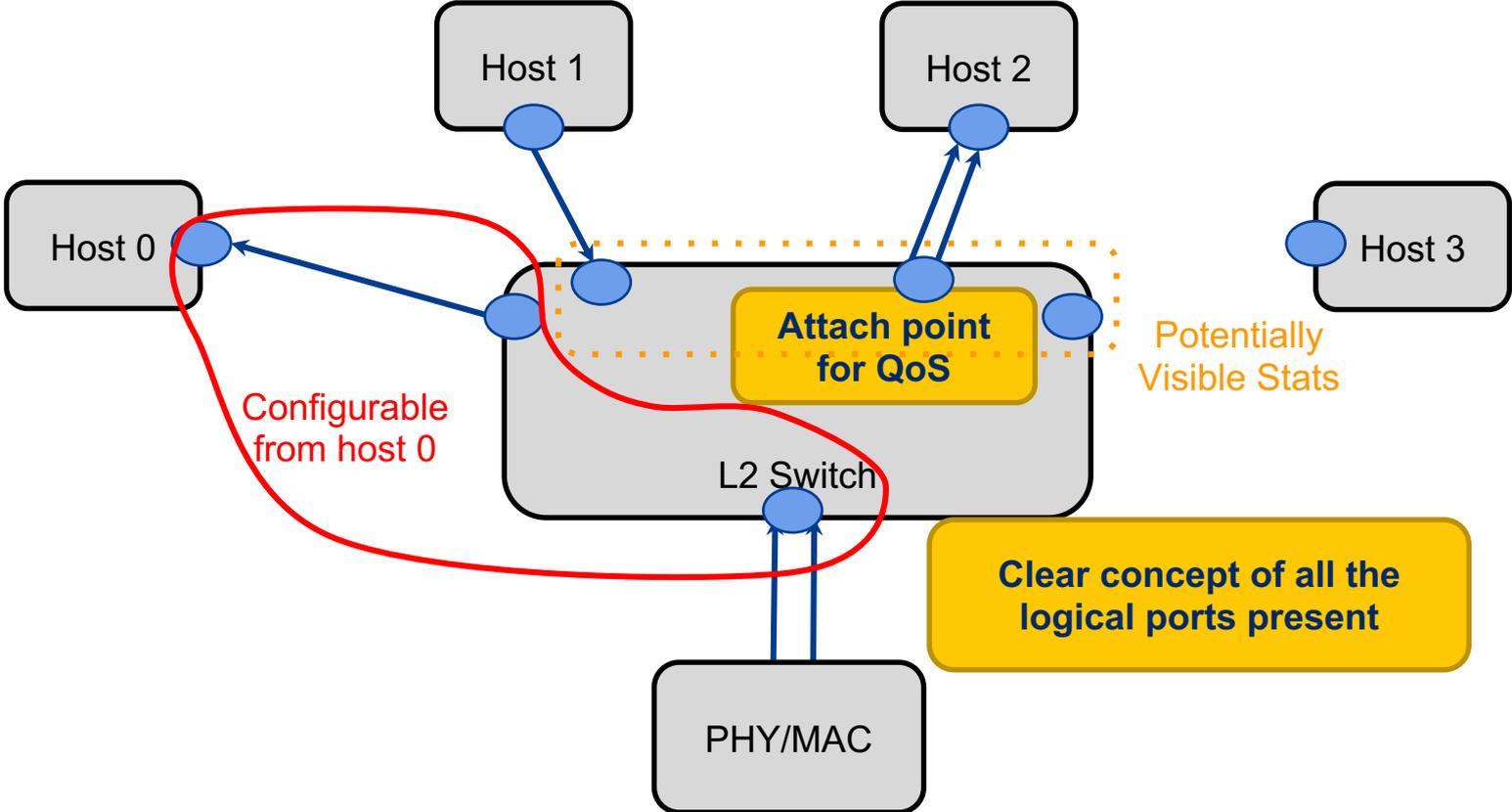
“Multi-Homed”



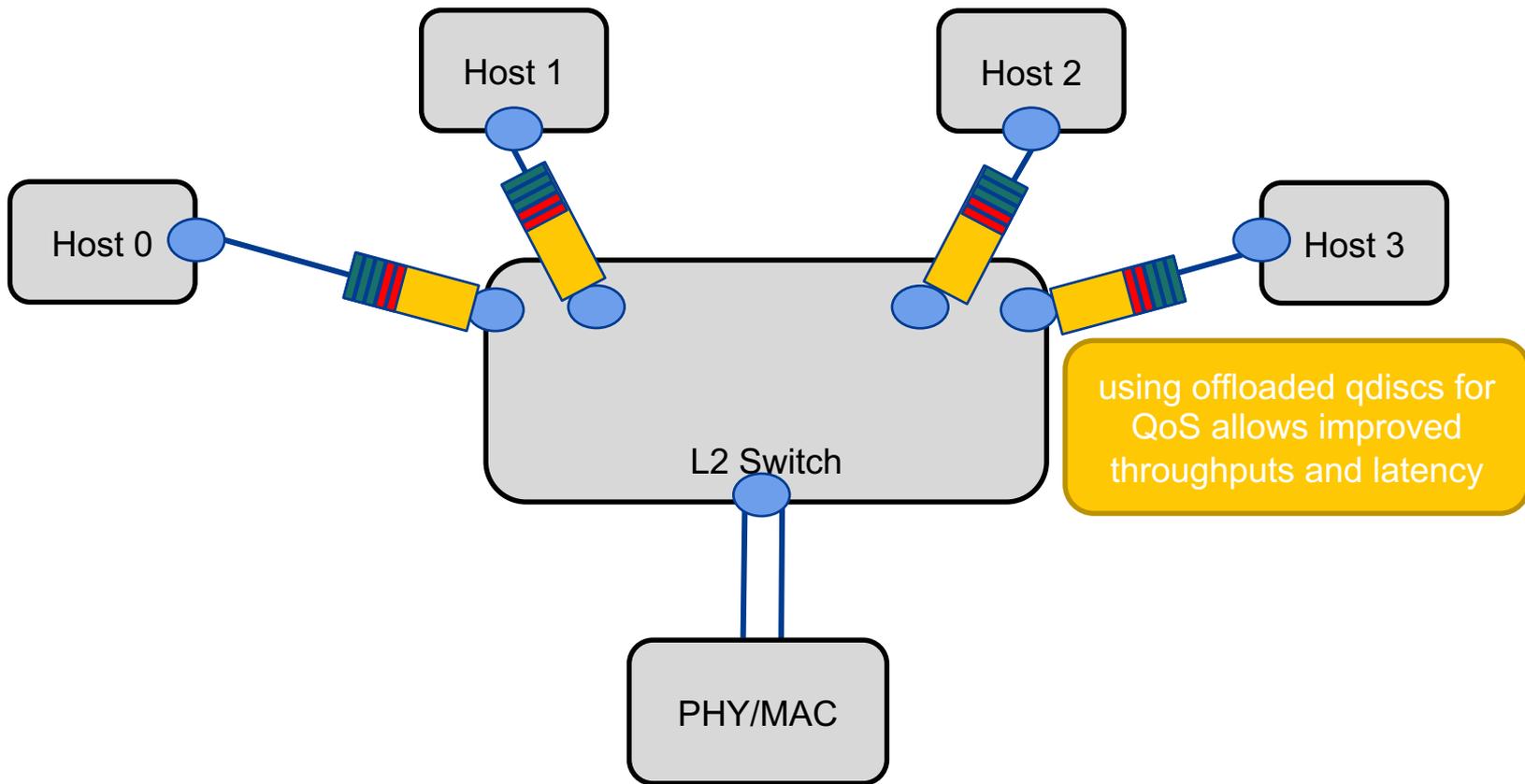
Multi-Host NIC Architecture



Many great things, but a few holes



Current Work: Switchdev and Simple Qdisc Offload



- Initialisation (.probe)
 - App Initialisation
 - vNIC Allocation
- Entering Switchdev Mode (.devlink_eswitch_set)
 - Spawning Representors
- Qdisc Setup (.ndo_tc_setup)
 - Attaching Qdisc to struct nfp_abm_link

Kernel

pci_epf_core.c

probe

Driver

nfp_net_main.c

nfp_net_pci_probe()

nfp_app.c

nfp_net_pf_app_init
nfp_net_pf_alloc_vnics

App
Abstraction

main.c

nfp_abm_init

struct nfp_abm

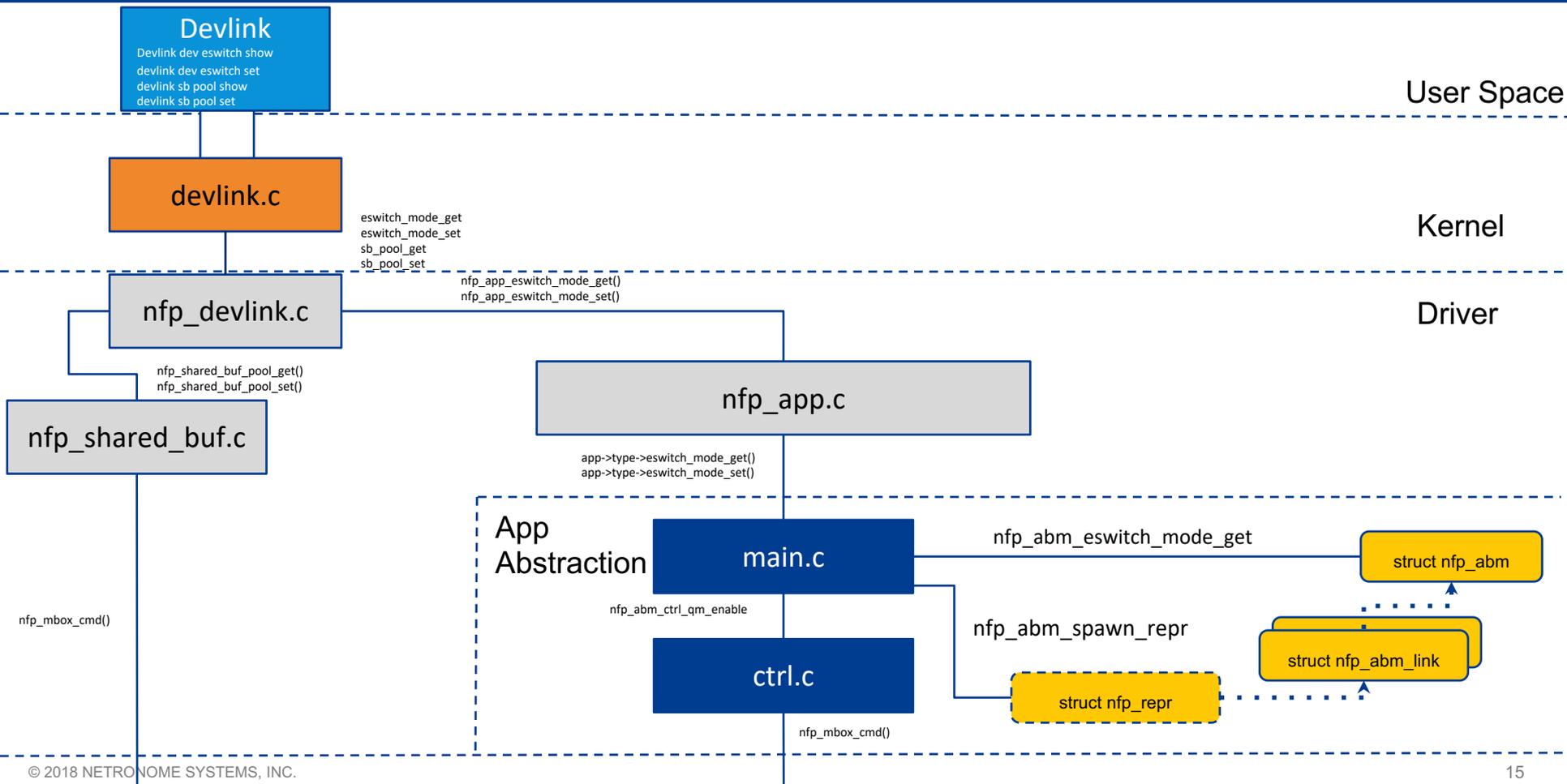
nfp_abm_ctrl_qm_disable
nfp_abm_ctrl_read_params

ctrl.c

nfp_abm_vnic_alloc

struct nfp_abm_link

nfp_mbox_cmd()



User Space

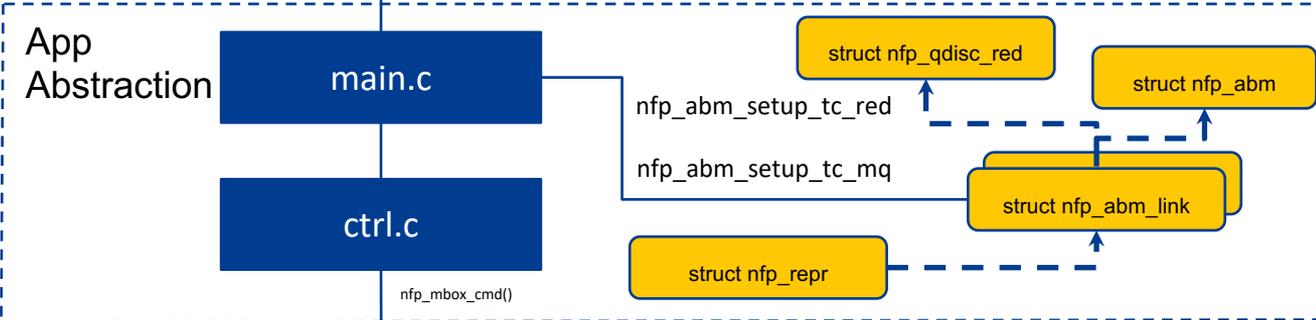
```
tc qdisc
tc qdisc add (mq/red)
tc qdisc replace
tc qdisc del
tc qdisc show
tc -s qdisc show
```

Kernel

```
sch_red.c/sch_mq.c
```

Driver

```
nfp_app.c
```



Next Steps: Extending the Egress Representor Architecture

- Generalising Qdisc Offload
 - Structure changes
 - nfp_abm_link
 - nfp_qdisc
- The clsact Qdisc
 - Motivation
 - Architecture

Before

```
struct nfp_abm_link {  
    struct nfp_abm *abm;  
    struct nfp_net *vnic;  
    unsigned int id;  
    unsigned int queue_base;  
    unsigned int total_queues;  
    u32 parent;  
    unsigned int num_qdiscs;  
    struct nfp_red_qdisc *qdiscs;  
};
```

```
struct nfp_red_qdisc {  
    u32 handle;  
    struct nfp_alink_stats stats;  
    struct nfp_alink_xstats xstats;  
};
```

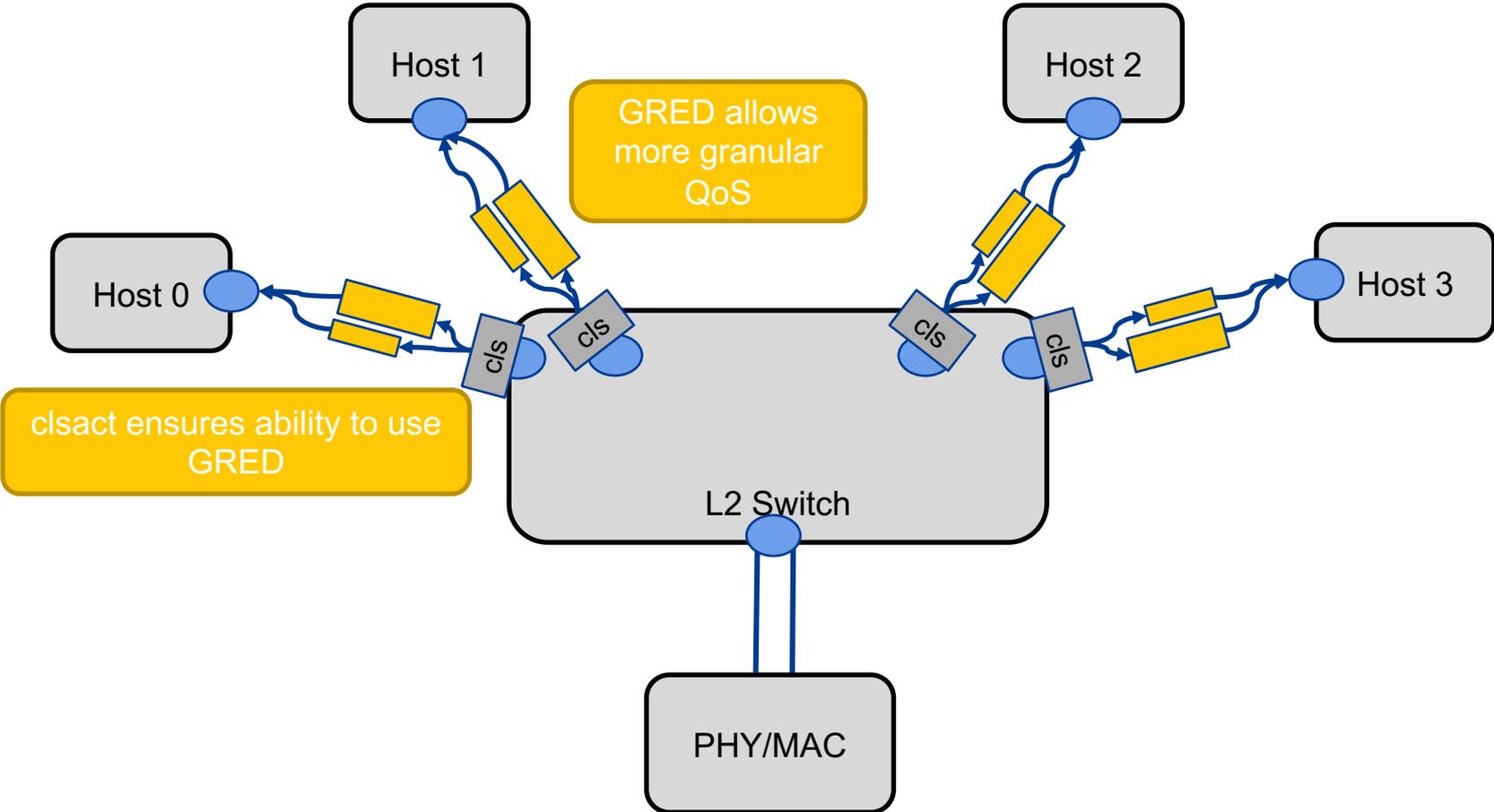
After

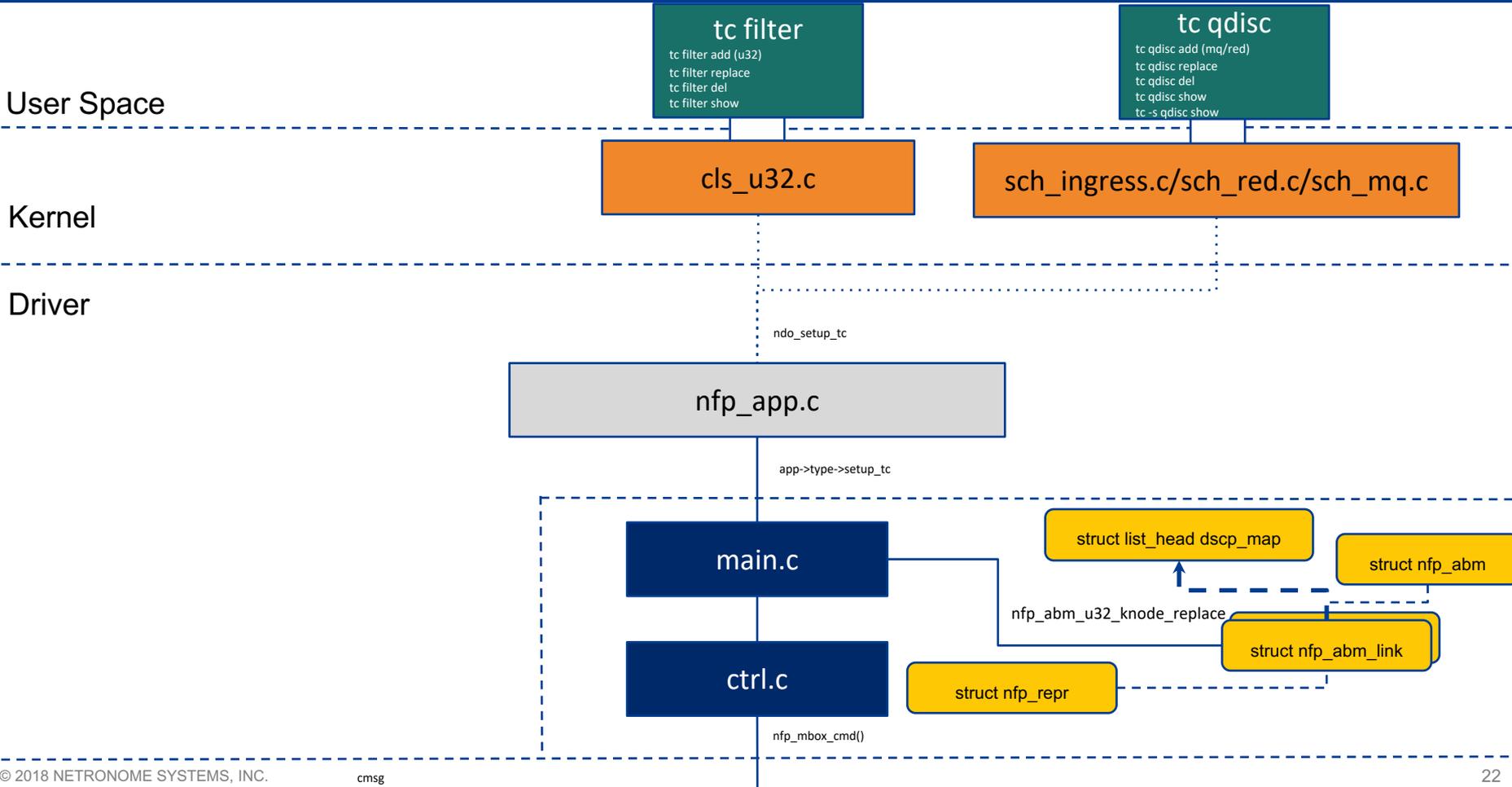
```
struct nfp_abm_link {
    struct nfp_abm *abm;
    struct nfp_net *vnic;
    unsigned int id;
    unsigned int queue_base;
    unsigned int total_queues;
    struct nfp_qdisc *root_qdisc;
    struct radix_tree_root qdiscs;
};
```

```
struct nfp_qdisc {
    struct net_device *netdev;
    enum nfp_qdisc_type type;
    /***/
    struct nfp_qdisc **children;

    /***/

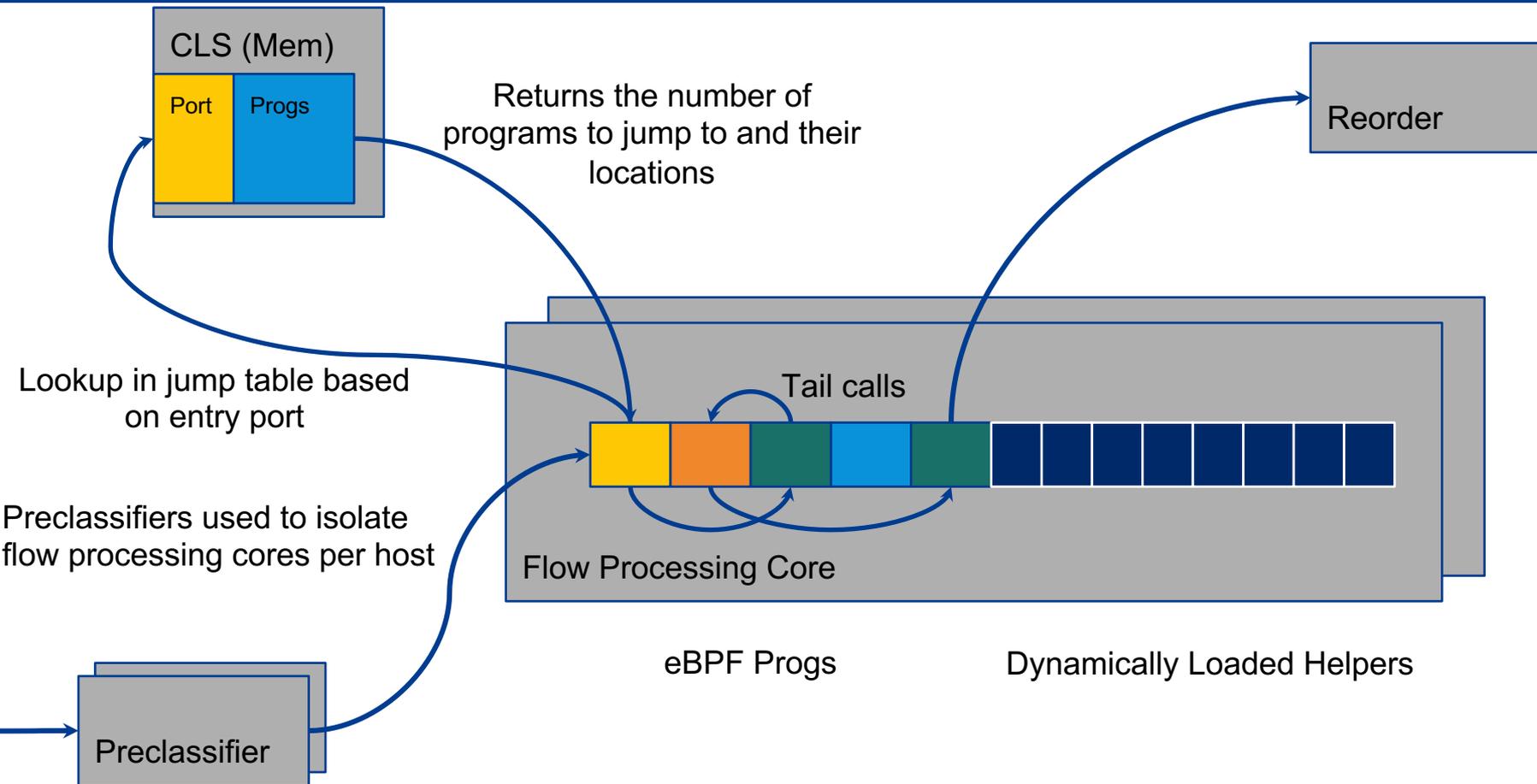
    union {
        /* TC_SETUP_QDISC_PRIQ */
        struct {
            const u32 *map;
        } prio;
        /* TC_SETUP_QDISC_RED */
        struct {
            u32 threshold;
            struct nfp_alink_xstats stats;
            struct nfp_alink_xstats prev_stats;
        } red;
    };
};
```



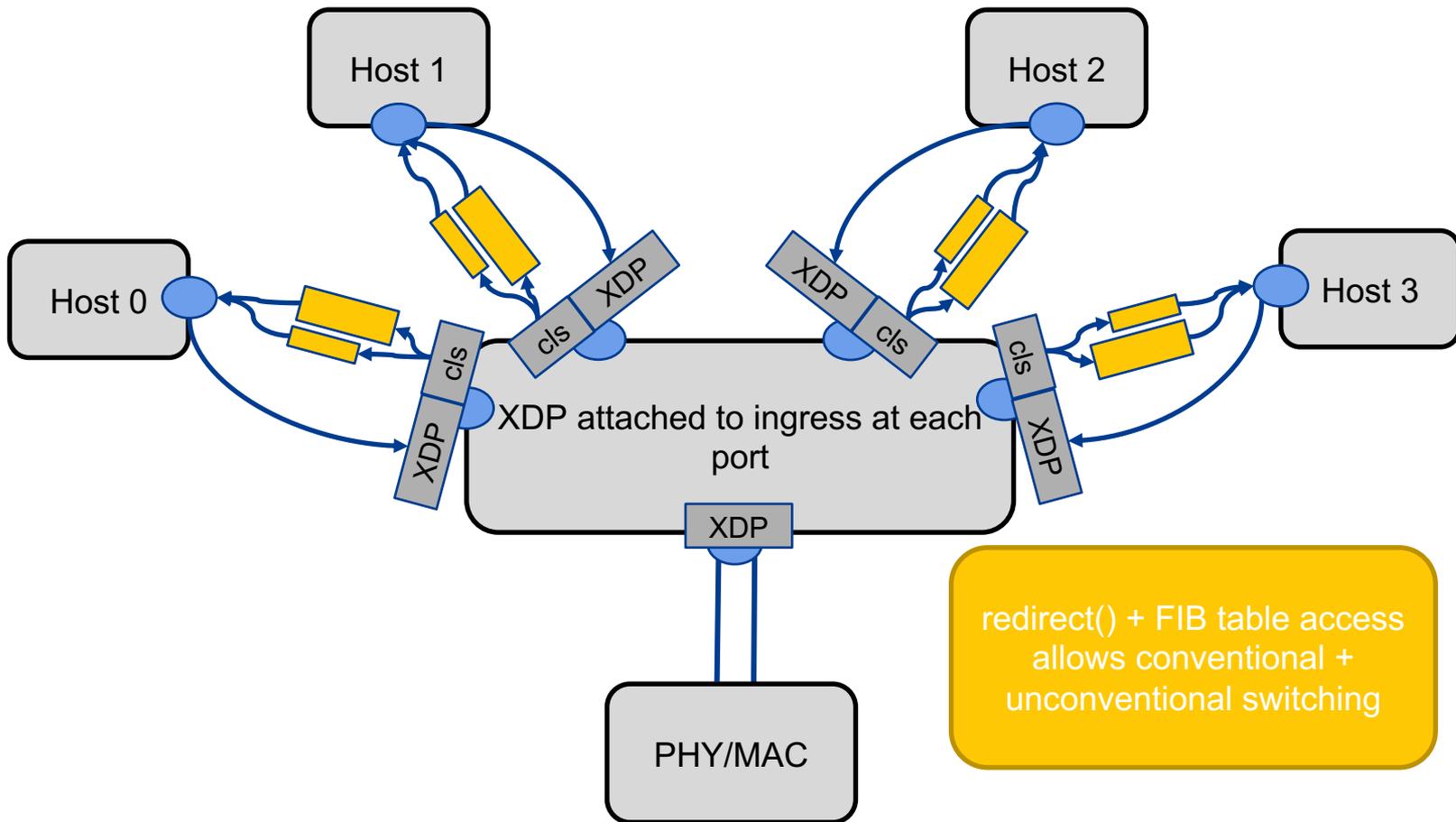


Future Work: Multi-host BPF Offload

- Firmware and BPF JIT
 - Flow Processing Core Datapath
- cls_bpf and Switchdev
 - Architecture
- XDP for Multihost Systems
 - Problems
 - Proposed Abstraction



- Challenges that have to be solved
 - XDP is an RX exclusive hook
 - Heterogenous architecture support is nascent
 - Security
- However more and more of the potential problems are falling away
 - e.g David Ahern's recent work on exposing the FIB table



- Proposing a fully flexible datapath for a multi-host NIC
- Achieved through a combination of switchdev, qdisc offload, cls_bpf and XDP
- Work in progress
 - Switchdev architecture and qdisc offload has been upstreamed
 - Next is simple clsact support
 - Followed by cls_bpf & XDP
- Provides potential for BPF defined pipelines in heterogenous architectures
 - See Jakub's talk at the microconference for more!

- Jakub Kicinski
- Jiong Wang
- Quentin Monnet
- David Beckett
- Edwin Peer
- Johan Moraal
- Mary Pham

Thank you!



Discussion
Questions/Comments