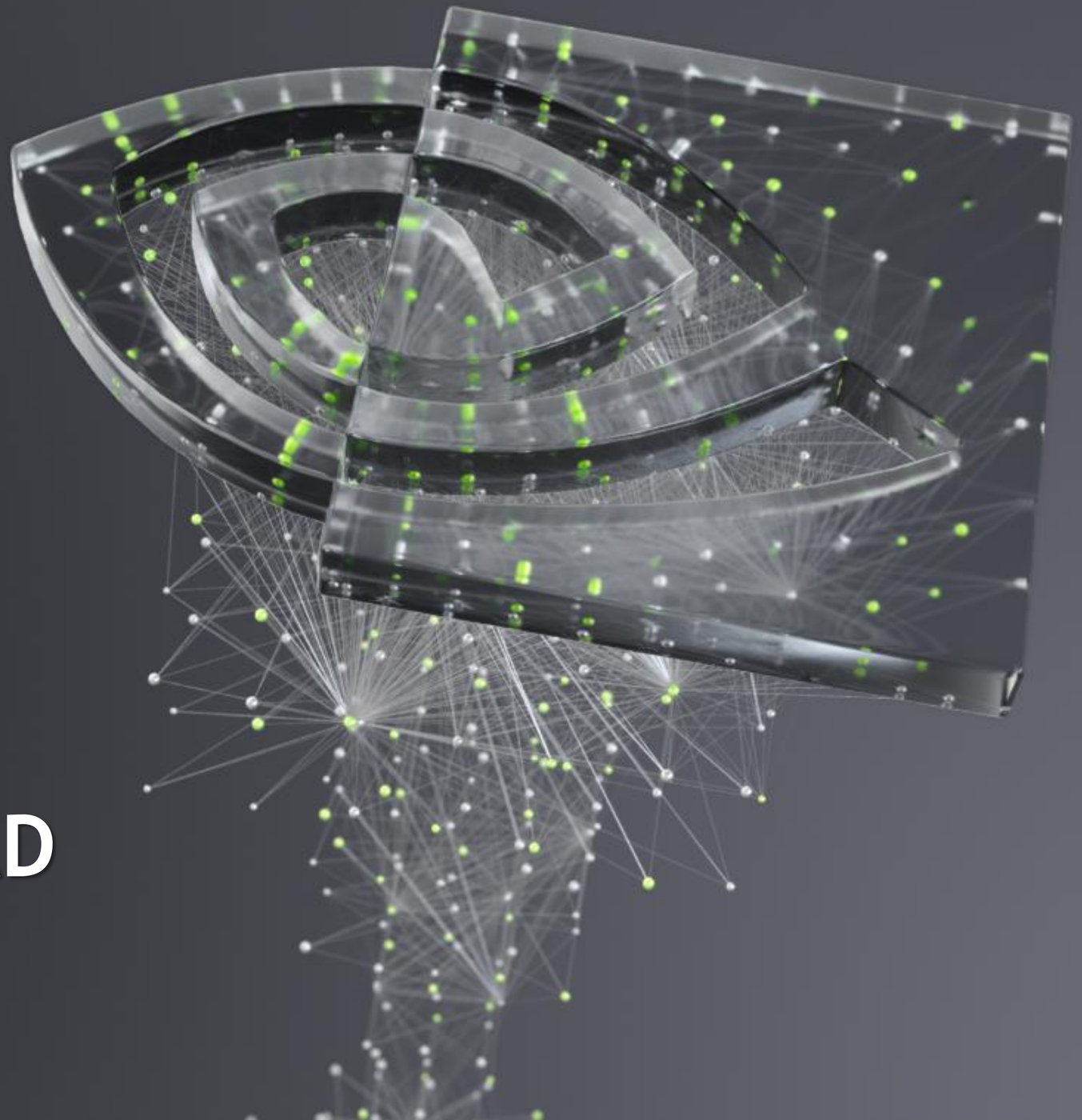


HIERARCHICAL QOS HARDWARE OFFLOAD

Yossi Kuperman, Maxim Mikityanskiy, 2020

AGENDA

Hierarchical Token Bucket

Brief description of HTB and its issues

HTB offload solution

Modifications to HTB to solve the issues and offload the logic

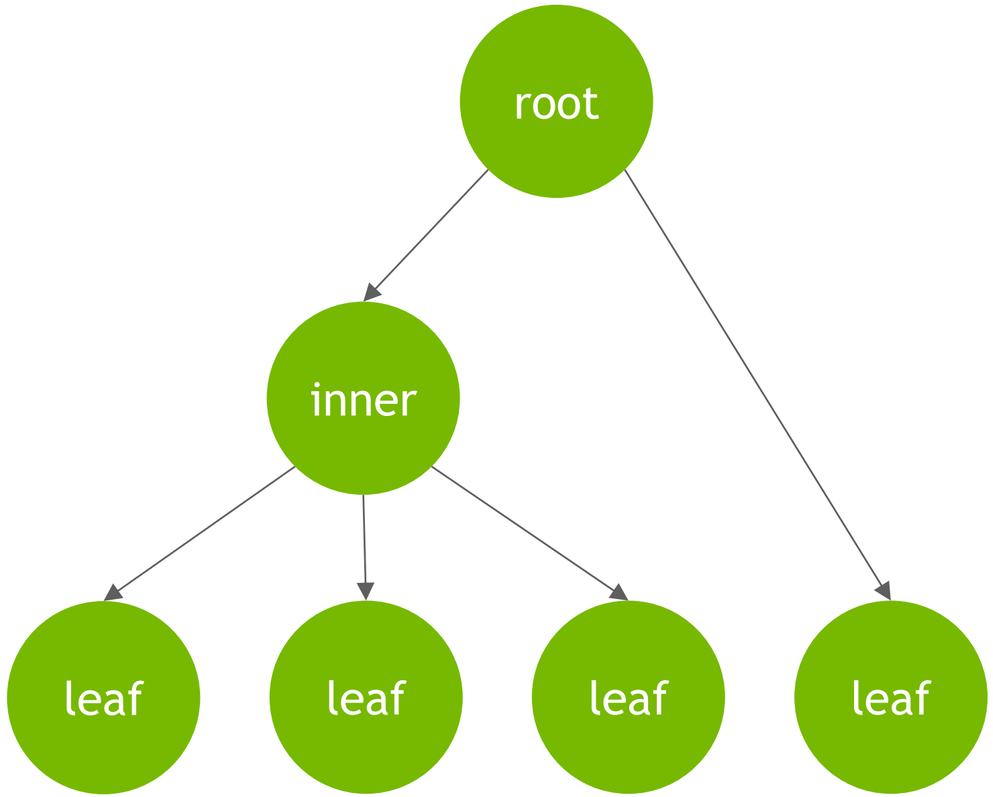
---

Current status

Known challenges and status of development and submission

HTB

Hierarchical Token Bucket



Shaping occurs in leaf nodes

Child nodes borrow tokens from parents

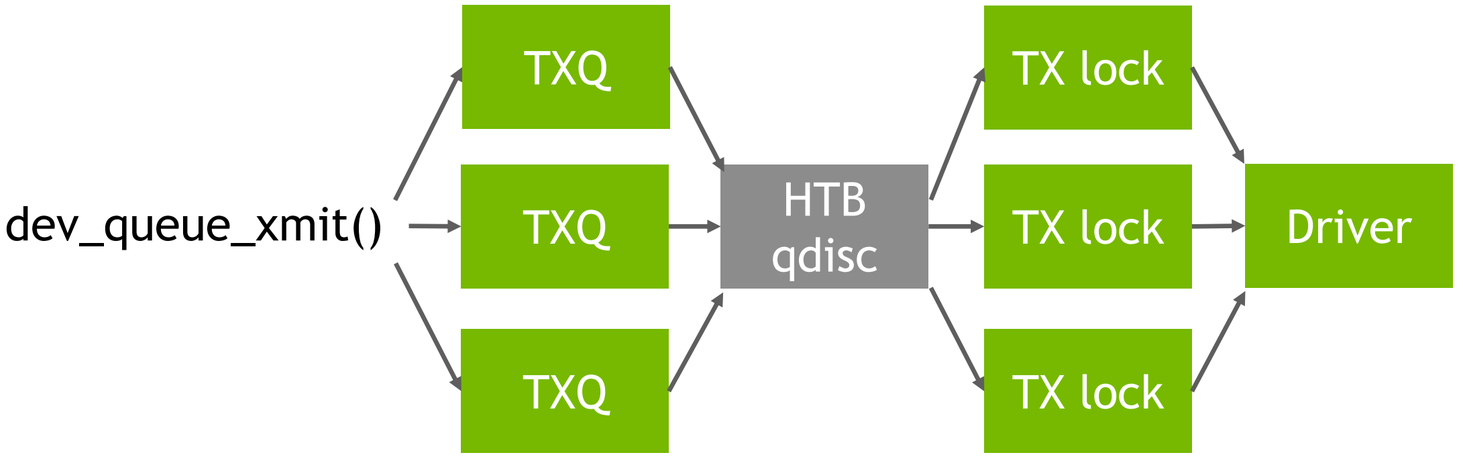
Classification:

```
# tc filter add dev eth0 parent 1:0
protocol ip flower dst_port 80 classid
1:10
```

HTB DRAWBACKS

Single HTB instance, single lock, not aware of multi-queue netdevs

1. Contention by flow classification
2. Contention by handling packets



SOLUTION FOR CLASSIFICATION

Flow classification still takes place in software

Classification takes place at the clsact hook

HTB skips classification if priority “points” to a class

For example, replace:

```
# tc filter add dev eth0 parent 1:0 protocol ip flower dst_port 80 classid 1:10
```

with an equivalent filter using skbedit action:

```
# tc filter add dev eth0 egress protocol ip flower dst_port 80 action skbedit priority 1:10
```

Thread-safe and lock-free classification

REMOVING THE LOCK CONTENTION

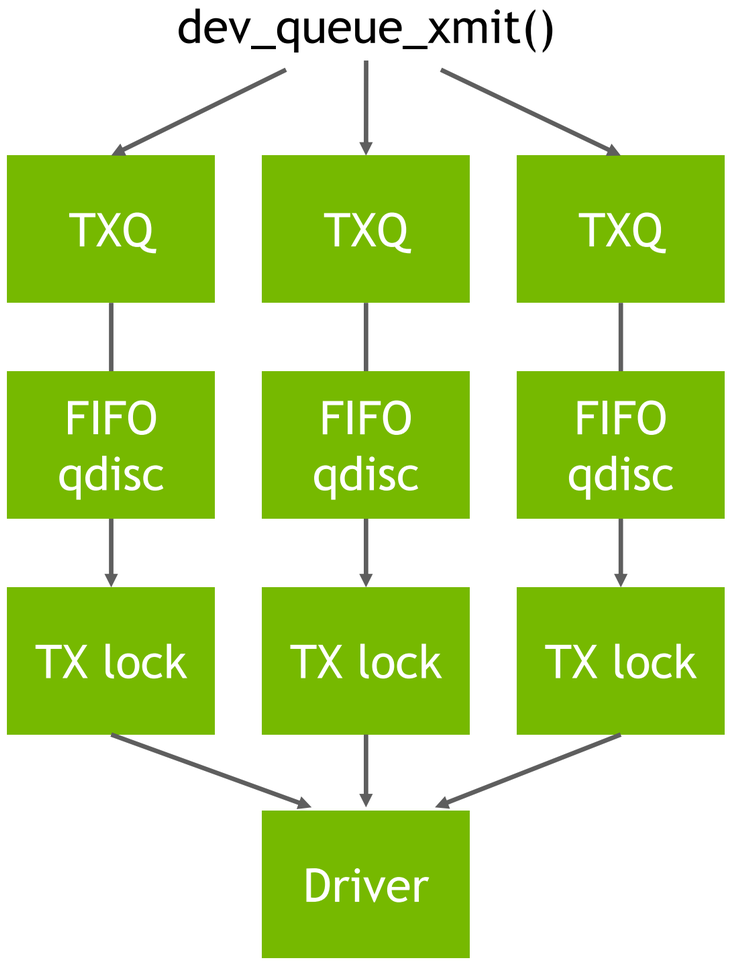
HTB will present itself as mq/mqprio does

- Create simple qdisc (FIFO) per TX queue
- Only when offload mode is set

HTB serves as the root qdisc

- Aggregate statistics and report to user
- Delegate the requests to the driver

HTB code is no longer part of the data-path



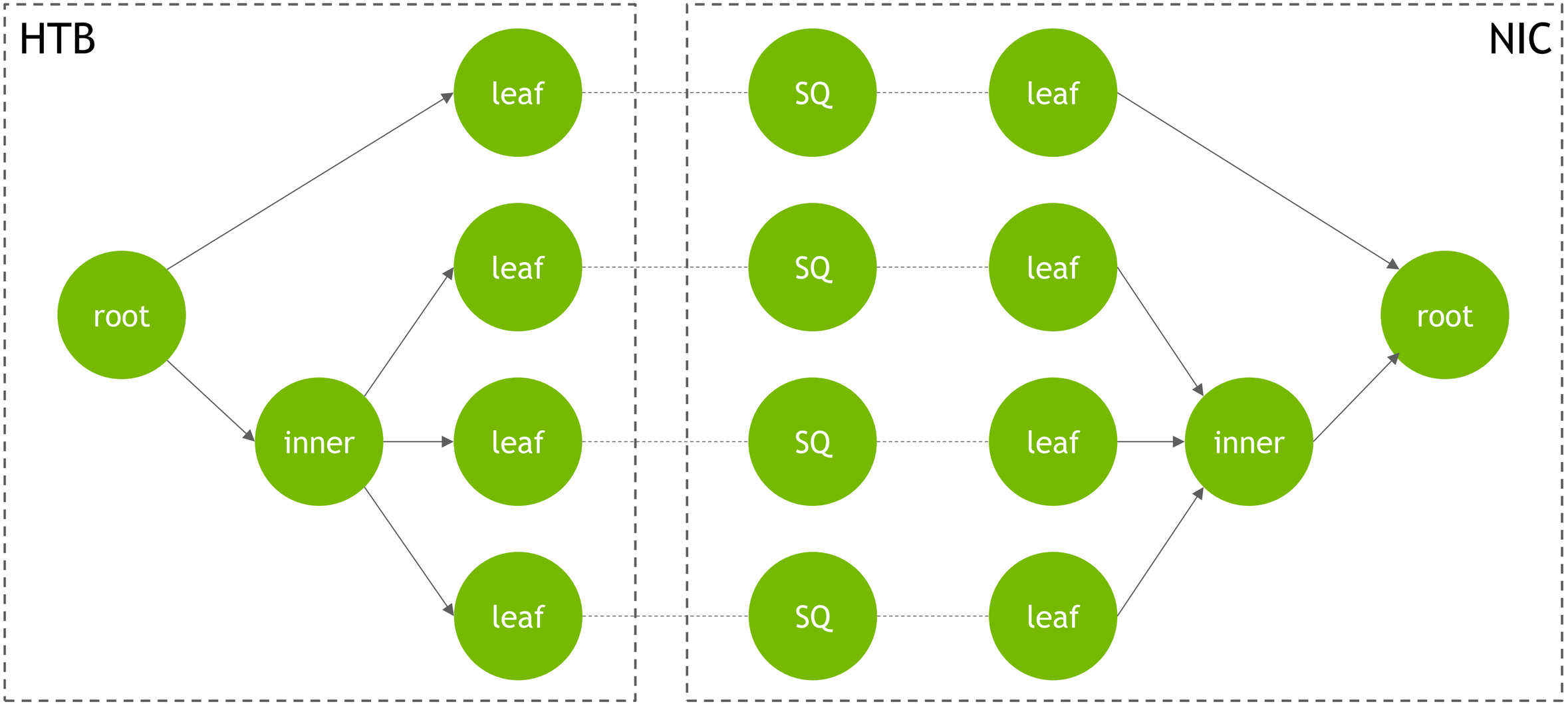
HARDWARE OFFLOAD

HTB uses ndo_setup_tc to provide the QoS tree structure to the driver, which recreates it in the NIC

All streams don’t have to fight for a single lock anymore

1. HTB registers as a multi-queue qdisc (like mq) and creates qdiscs per queue
2. Each leaf class is backed by a hardware queue
3. Clsact happens before ndo_select_queue, so the driver can pick a queue corresponding to the class
4. Rate limiting is performed by the hardware

HARDWARE OFFLOAD



PACKET FLOW

1. clsact sets skb->priority to a leaf class ID
2. ndo_select_queue looks at skb->priority and picks the TX queue
3. The SKB is enqueued into the per-queue qdisc of that TX queue
4. The SKB is dequeued from the per-queue qdisc
5. The driver puts the SKB into the hardware Send Queue
6. The NIC does the shaping and transmits the packet

HARDWARE OFFLOAD ADVANTAGES

No contention on a single lock: different traffic classes don’t interfere with each other, which allows for better throughput

Rate limiting logic is offloaded to the NIC, reducing CPU load

KNOWN CHALLENGES

Qdiscs of leaf classes are applied before HTB logic, when offloaded

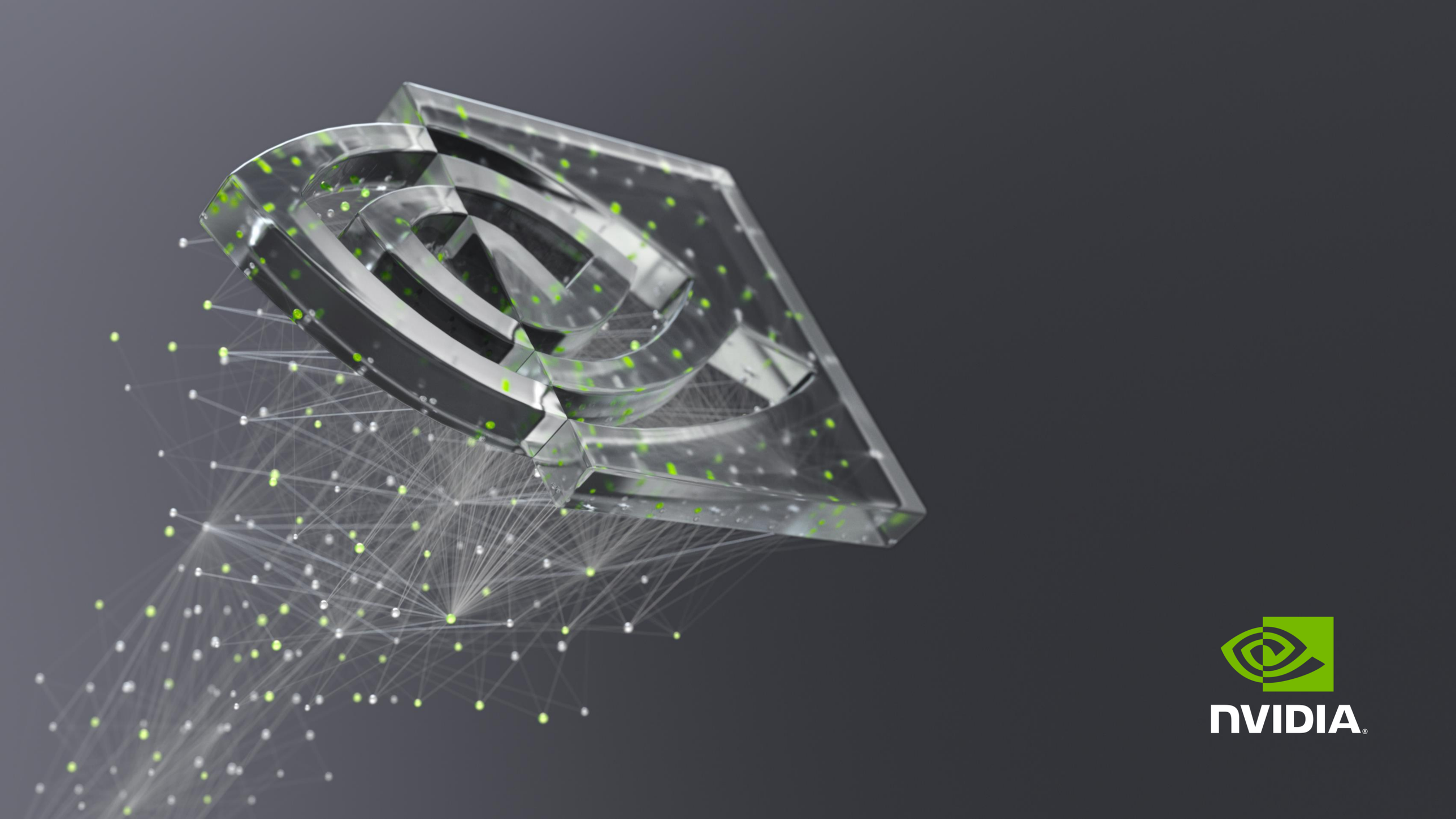
QoS TX queues have to be preallocated on alloc_etherdev_mqs

- Hardware queues are created and destroyed on demand
- real_num_tx_queues is changed by the driver when leaf classes change
- Deleting a leaf class may lead to gaps in TX queue numeration

CURRENT STATUS

PoC patches for mlx5 (using sysfs for configuration)

RFC was posted to netdev mailing list, showing the HTB offload interface


NVIDIA