Networking Traffic Control

Cong Wang

@c0ngwang

Agenda

- Recent updates
- Upcoming updates
- Some challenges, especially locking

What Is New?

- pfifo_fast qdisc becomes lockless, not really
- TC filter chain and shared block
- Hardware offloading
- extack support

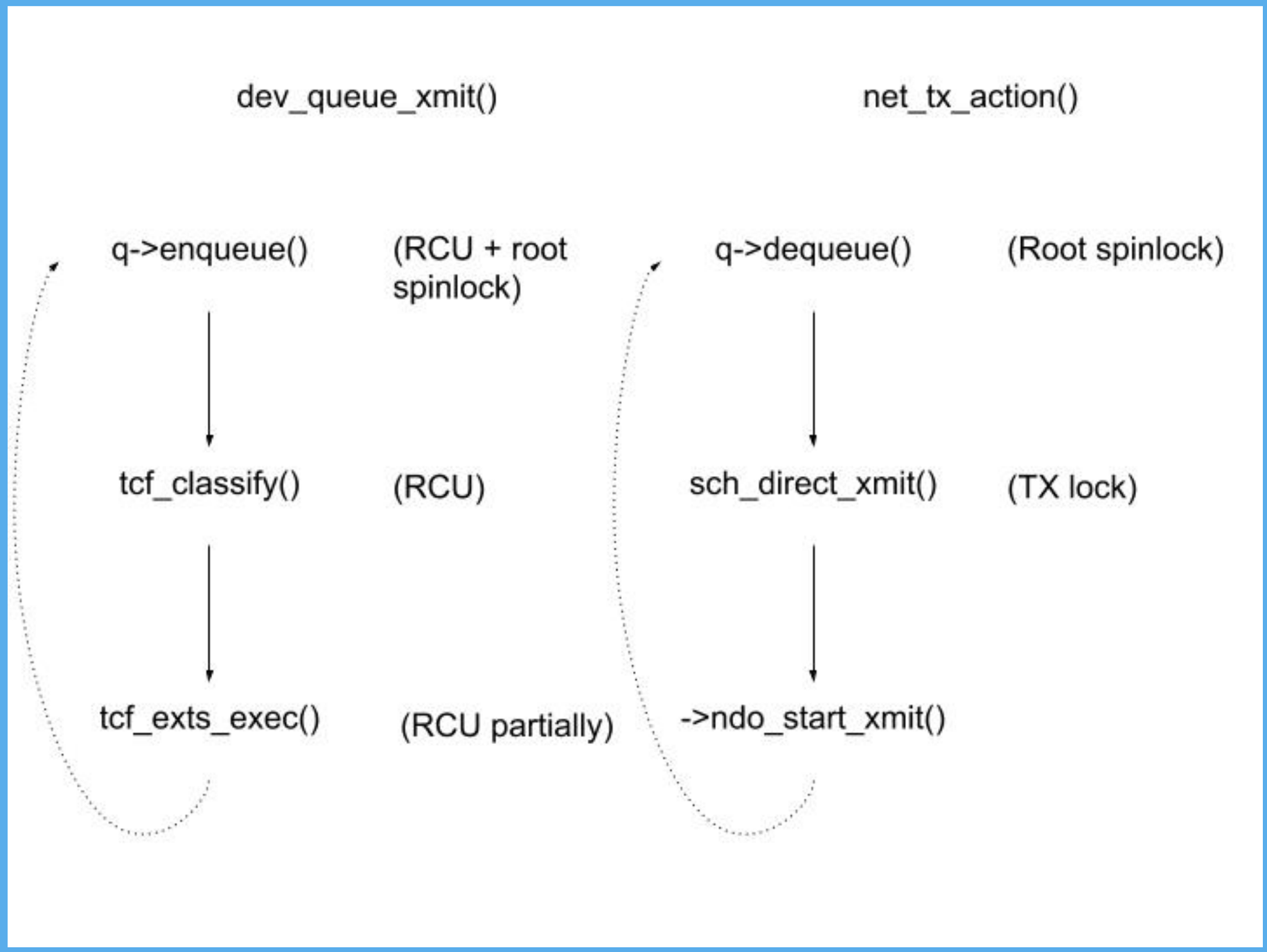
What's Next?

- A few new qdisc's
- RCU-completeness for TC actions (WIP)
- Review TC tree lock and RTNL lock

@c0ngwang

Rcu Completeness

- Qdisc layer still has spinlocks on fast path
- Filter layer is already RCU-complete
- Action layer, copy is still missing...
- Completely get rid of TC action spinlocks


dev_queue_xmit()
net_tx_action()
q->enqueue()
(RCU + root spinlock)
q->dequeue()
(Root spinlock)
tcf_classify()
(RCU)
sch_direct_xmit()
(TX lock)
tcf_exts_exec()
(RCU partially)
->ndo_start_xmit()

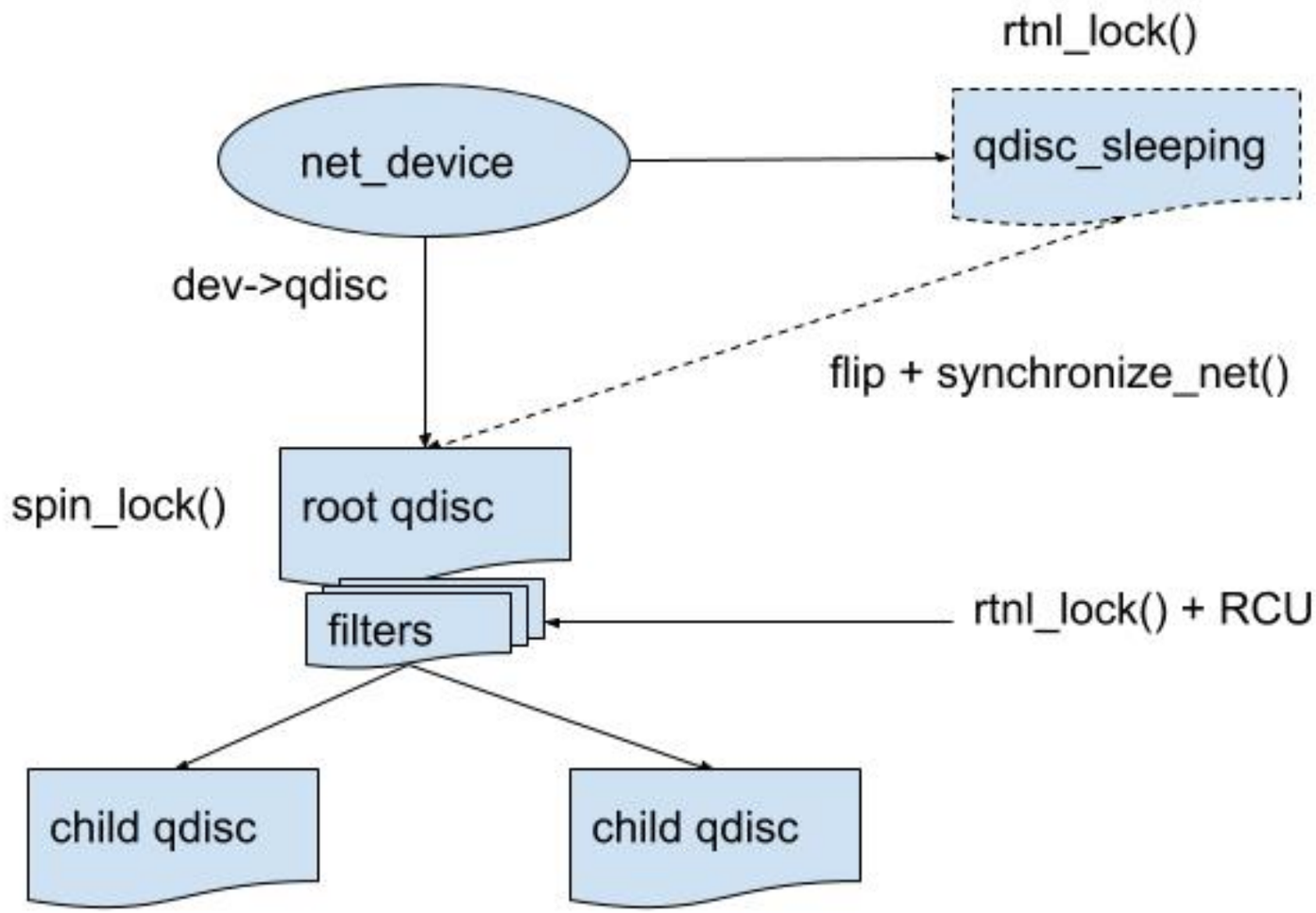
Lockless Qdisc?

- It is a ring buffer!
- No tree lock, but consumer and producer lock
- Decouple enqueue and dequeue
- FIFO is easy, it is class-less, filter-less

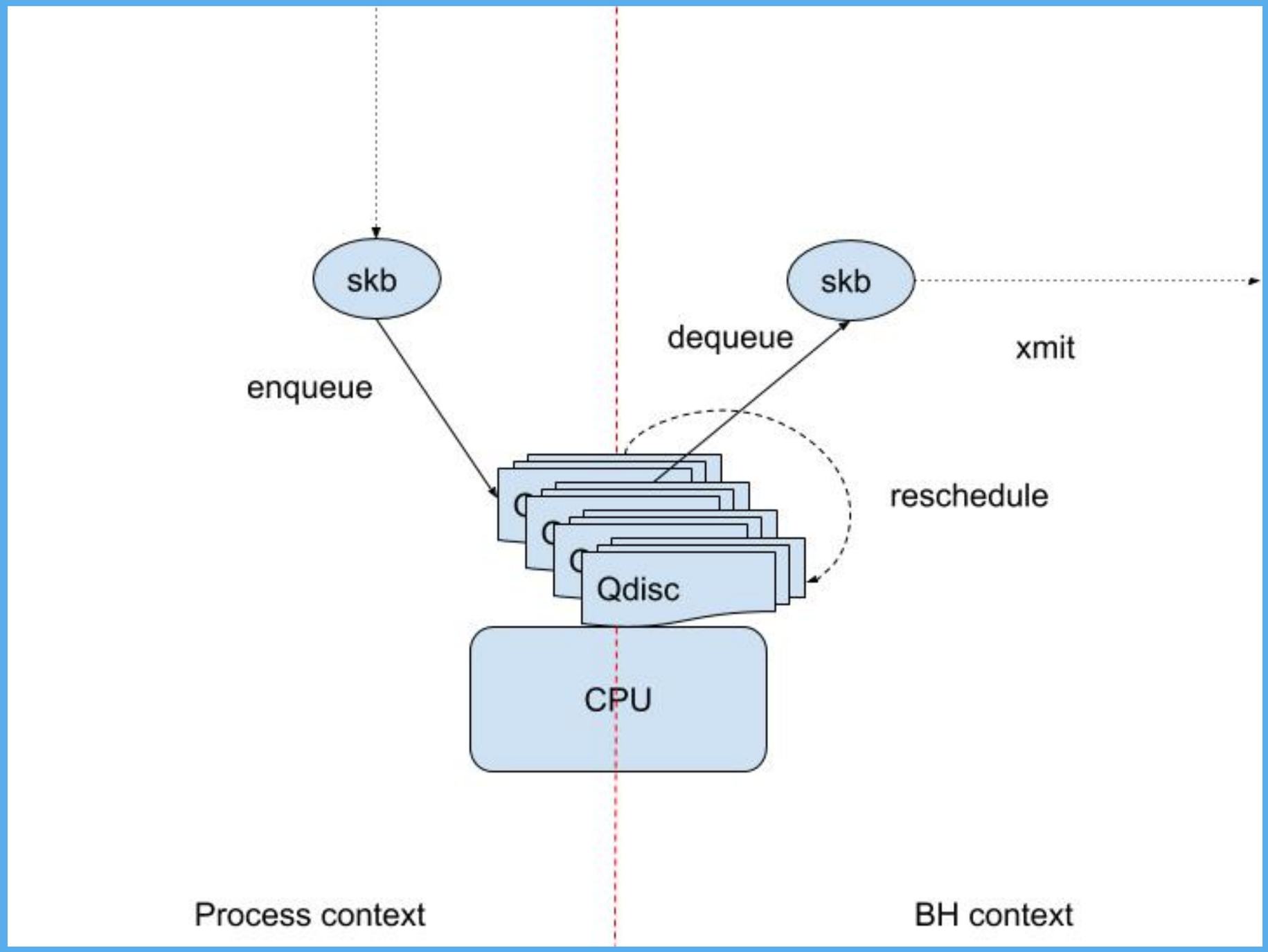
Really Lockless?

- skbs enqueued and dequeued on fast path
- TC is hierarchical by design, always locks root
- BH context vs process context
- Different internal data structures

	qdisc	filter and embedded action	standalone action
add	dev->qdisc_sleeping	dev->qdisc	idr
delete	dev->qdisc_sleeping	dev->qdisc	idr
dump	dev->qdisc	dev->qdisc	idr
sync with fast path	synchronize_net()	call_rcu()	Filter's call_rcu()



Fast path vs. slow path


skb
skb
dequeue
xmit
enqueue
reschedule
Qdisc
CPU
Process context
BH context

@c0ngwang

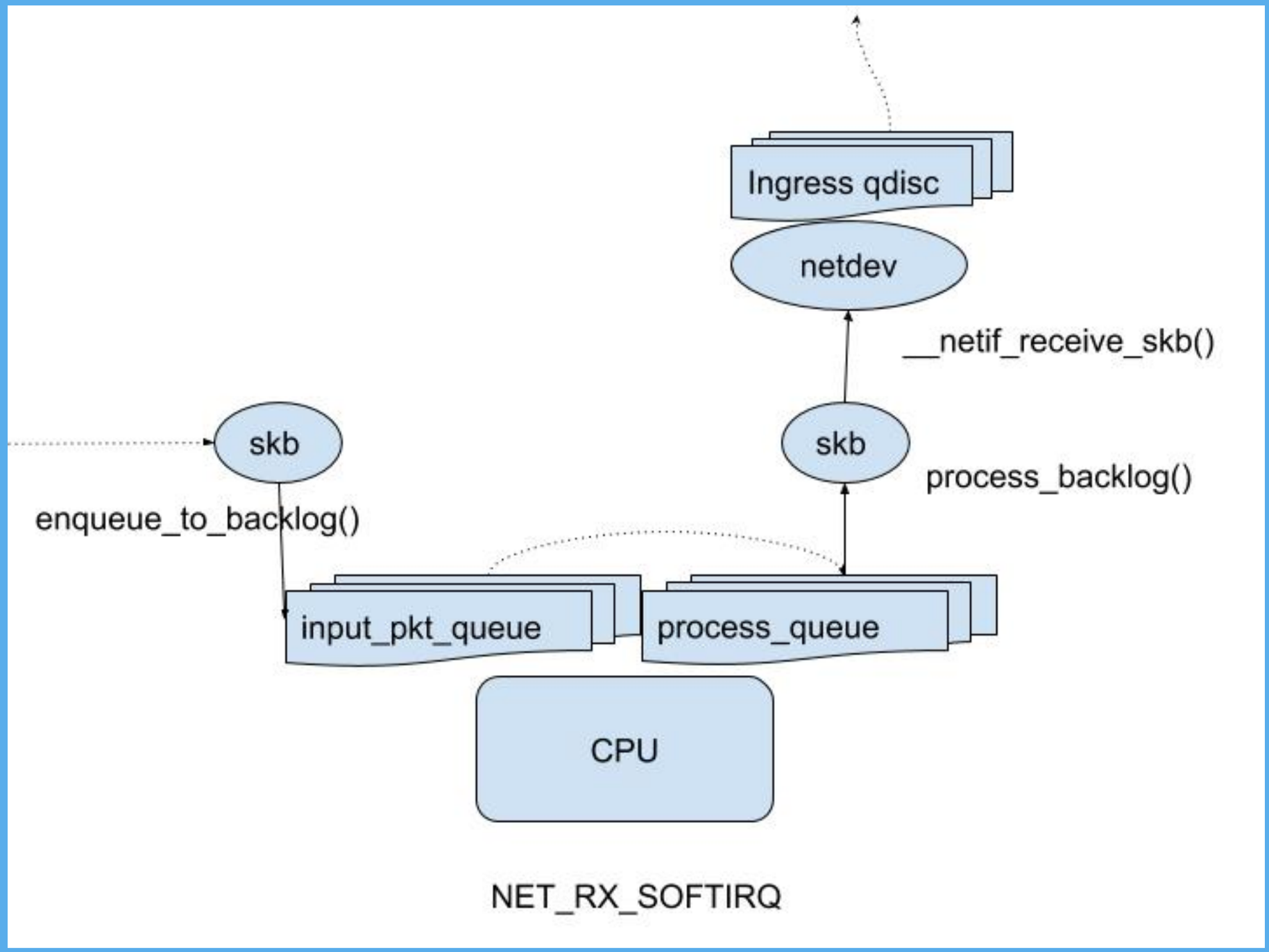
Idea: Locking

- dev->qdisc_sleeping can be eliminated?
- Break down RTNL?
- Tree lock can break down to each impl./layer?
- Lockless queue?

Idea: Tx Backpressure

- Process can’t retrieve TX status from BH
- Queued != Transmitted
- TCP small queue leverages skb destructor
- Not every qdisc respects TCP ECN

@c0ngwang


Ingress qdisc
netdev
__netif_receive_skb()
skb
process_backlog()
skb
enqueue_to_backlog()
input_pkt_queue
process_queue
CPU
NET_RX_SOFTIRQ

Idea: Ingress Shaping

- No queueing on RX side: RX->TX is ugly
- Hard to decouple enqueue and dequeue
- Throttle Rx queue?
- No feedback to sender: TCP ECN? Throttle TCP ACK?

Idea: DEV->TX_QUEUE_LEN

- Modern qdiscs are not just one single queue
- Multiple hardware queues
- pfifo_fast resize
- sch->limit

Thank You!

@c0ngwang