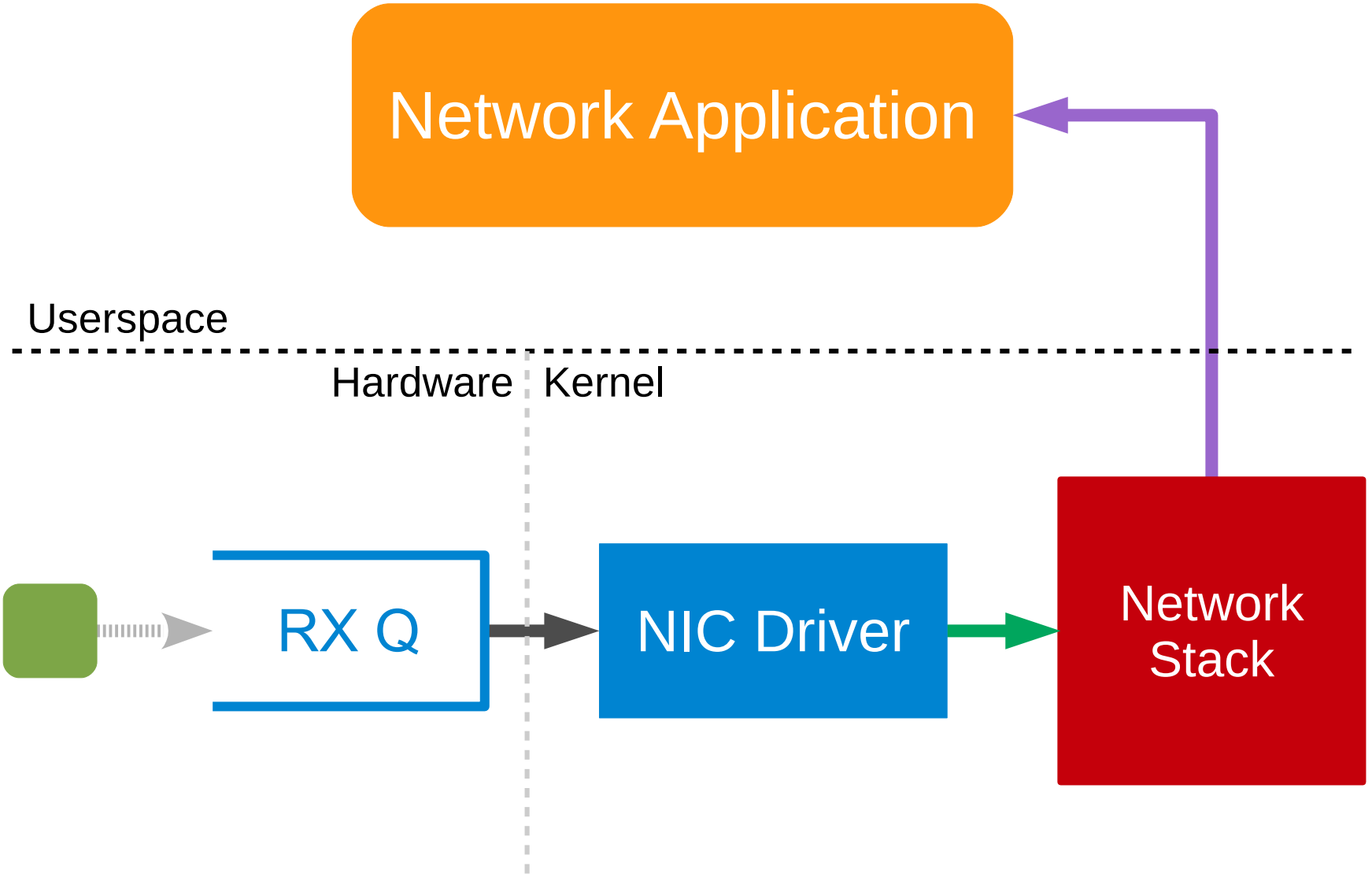
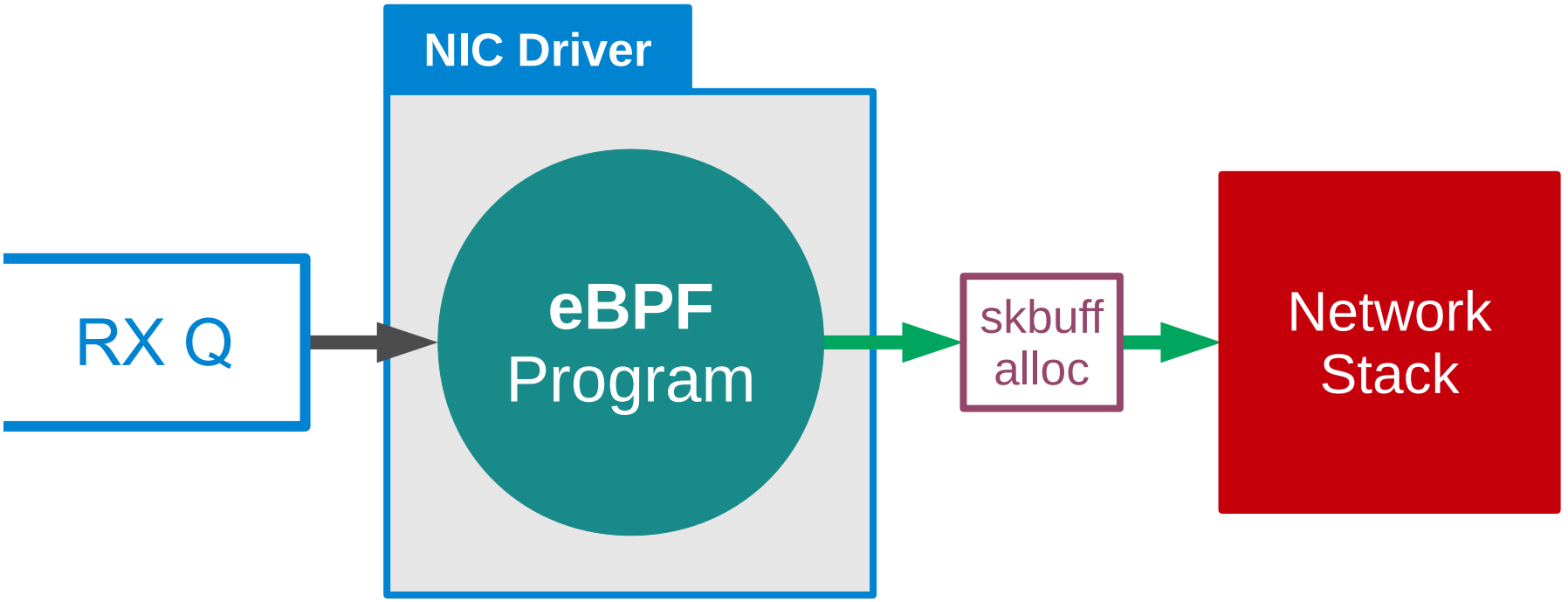
XDP: Networking from Kernel to Userspace

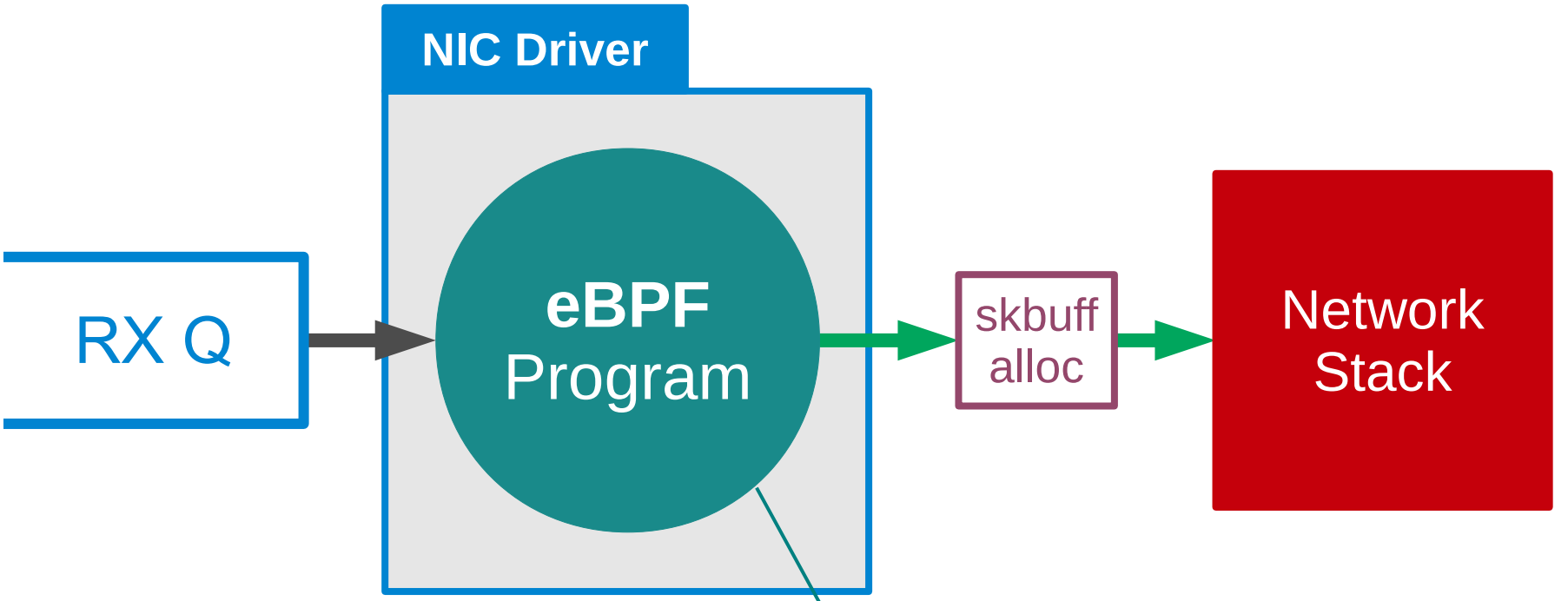
openSUSE Taiwan 2018

Gary Lin
Software Engineer, SUSE Labs
glin@suse.com

eXpress Data Path


Network Application
Userspace
Hardware
Kernel
RX Q
NIC Driver
Network Stack


NIC Driver
RX Q
eBPF
Program
skbuff
alloc
Network
Stack



- Loaded dynamically
- Verified with eBPF verifier
- Inspect and modify packets

XDP Actions

- **XDP_ABORTED**

  Indicate eBPF program error
- **XDP_DROP**

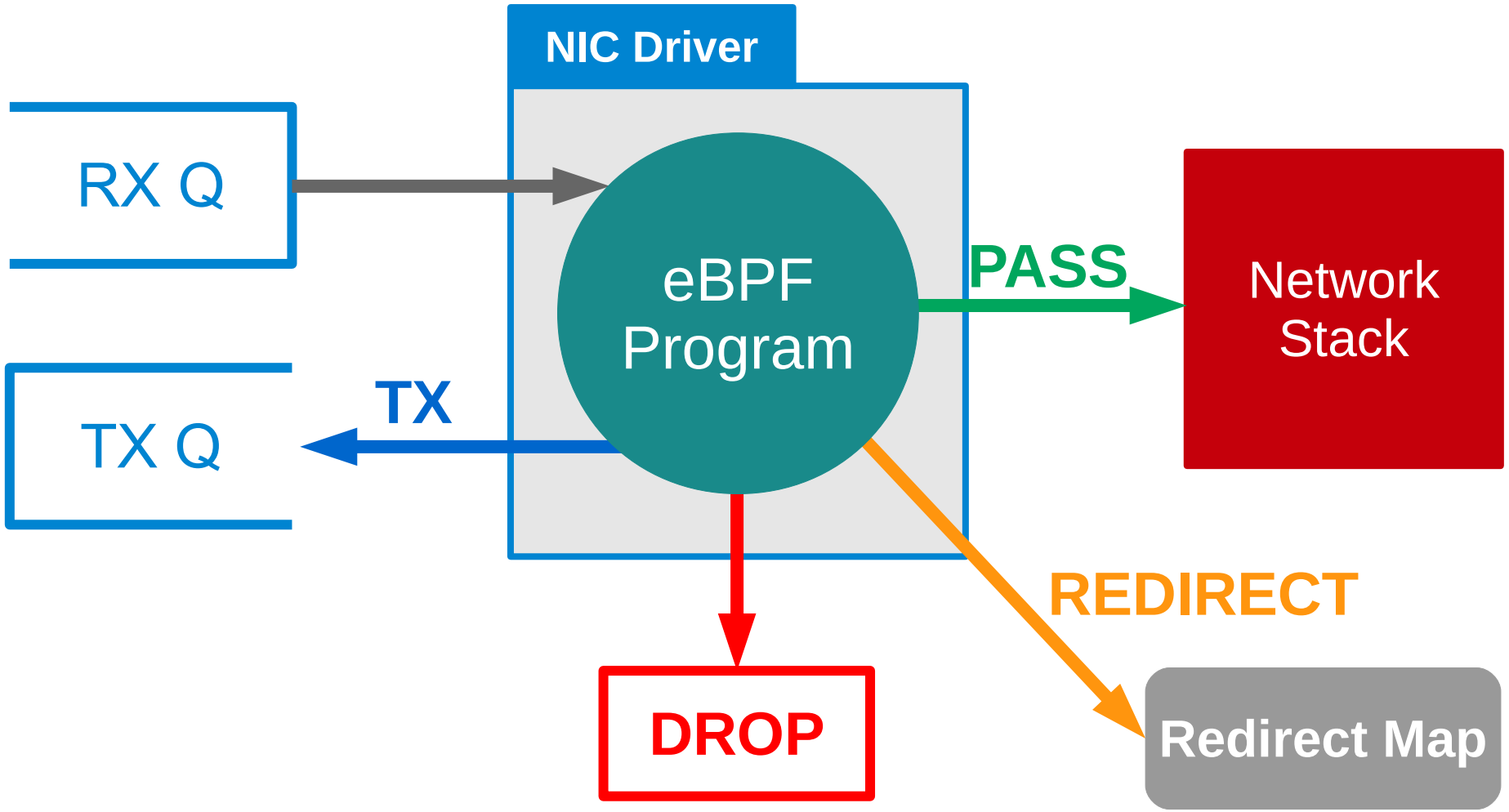
  Drop the packet
- **XDP_PASS**

  Pass the packet to the network stack
- **XDP_TX**

  Transmit the packet through the same NIC
- **XDP_REDIRECT**

  Redirect the packet to other NIC, CPU, or an AF_XDP socket

eXpress Data Path

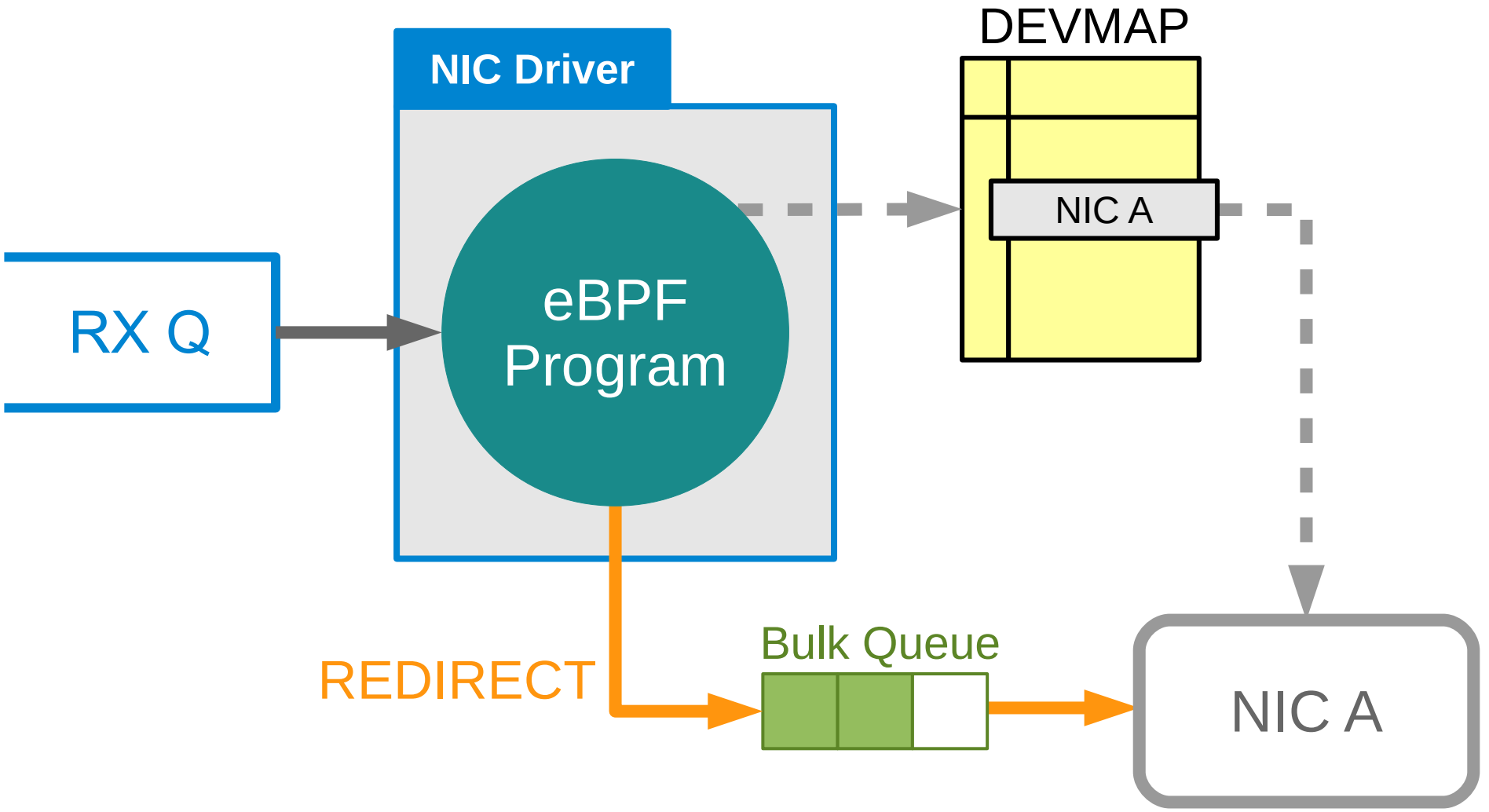


XDP Use Cases

- DDoS attack mitigation
- Load balancing
- Tunnelling: packet header handling
- Network sampling and monitoring
- And more...

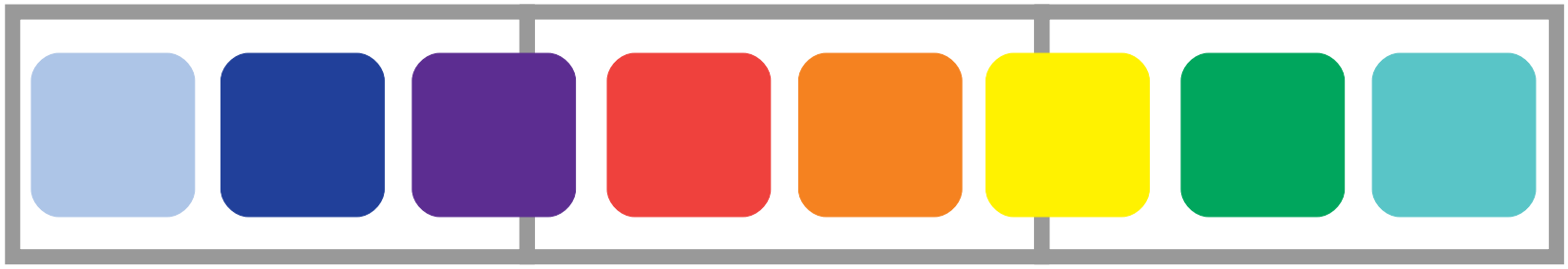
XDP REDIRECT – DEVMAP

XDP REDIRECT – DEVMAP



Memory Model Change

Conventional RX Buffer



Page 1 Page 2 Page 3

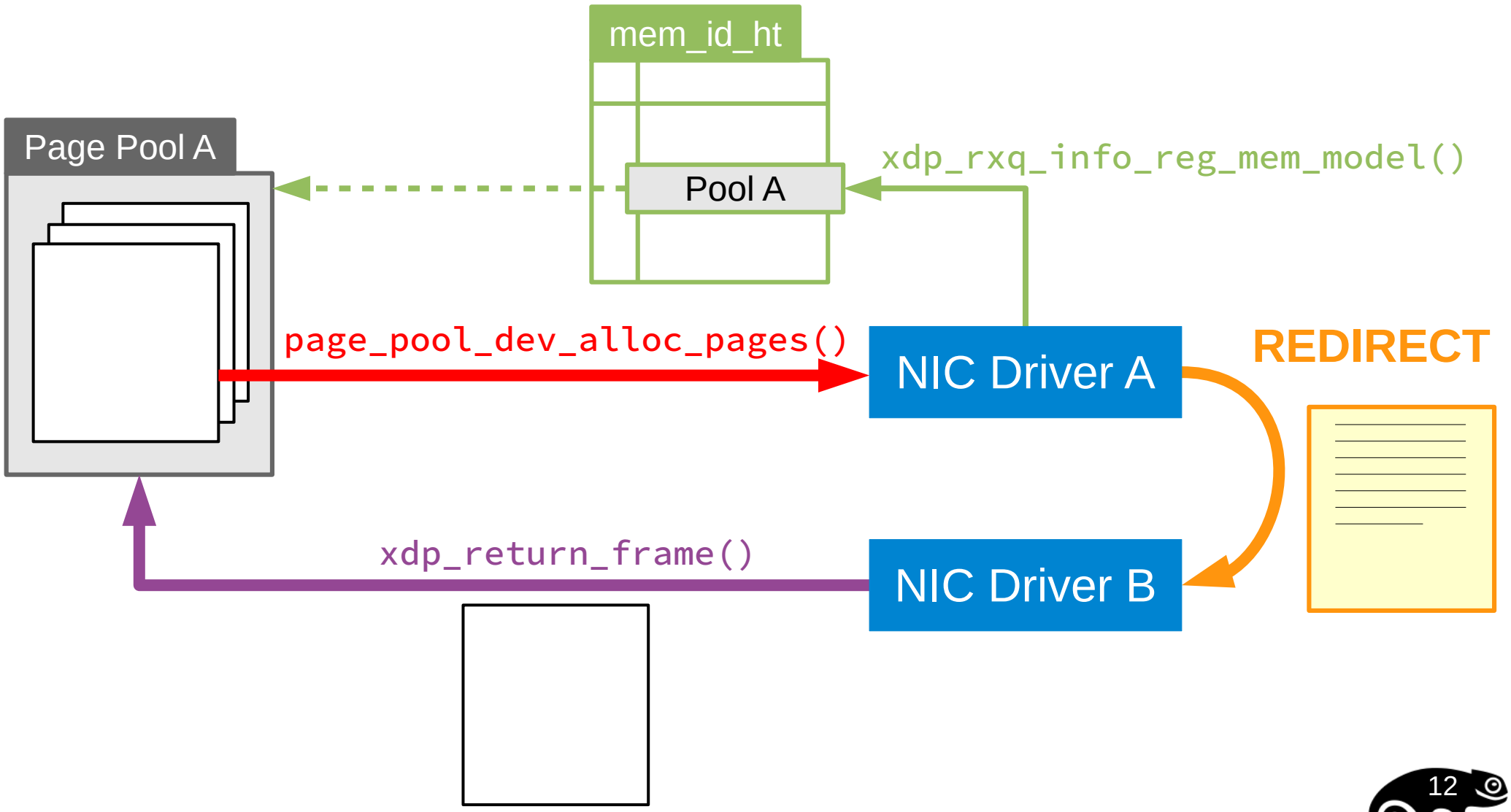
XDP

One Packet Per Page

Page 1 Page 2 Page 3

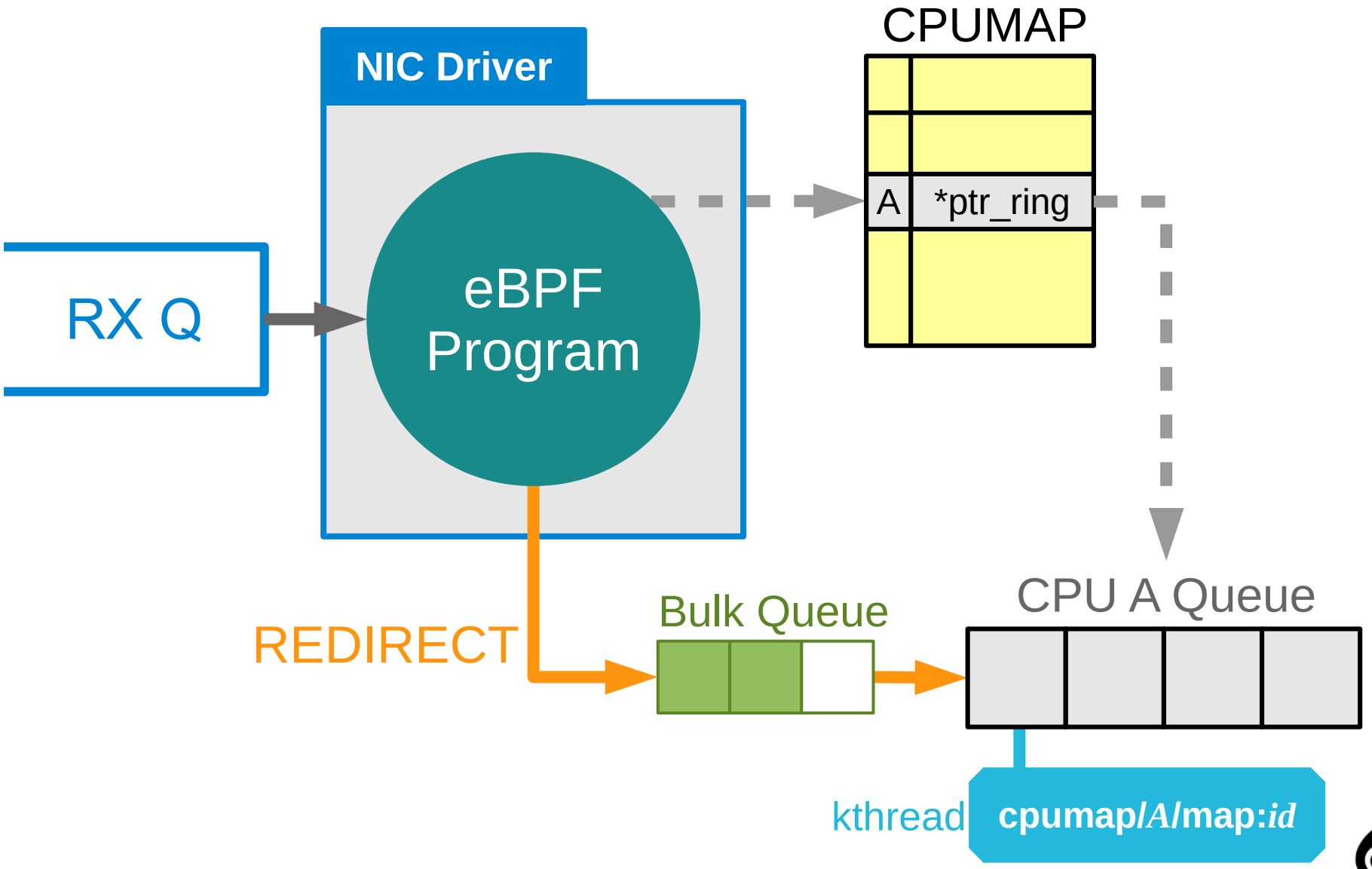
NOTE: Some drivers such as ixgbe adopt a different memory model.

XDP Page Pool



XDP REDIRECT – CPUMAP

XDP REDIRECT – CPUMAP

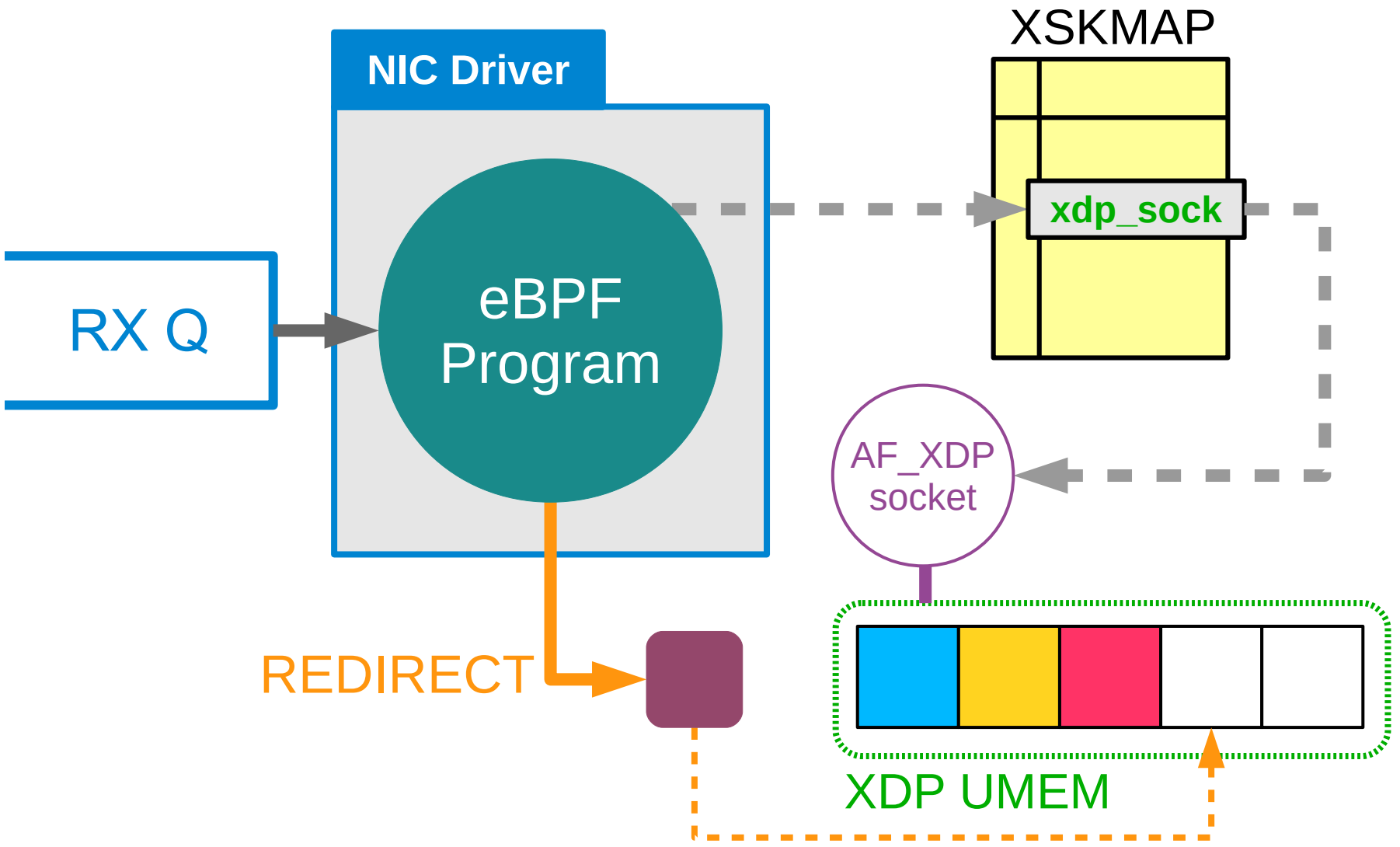


Looks like RFS/RPS?

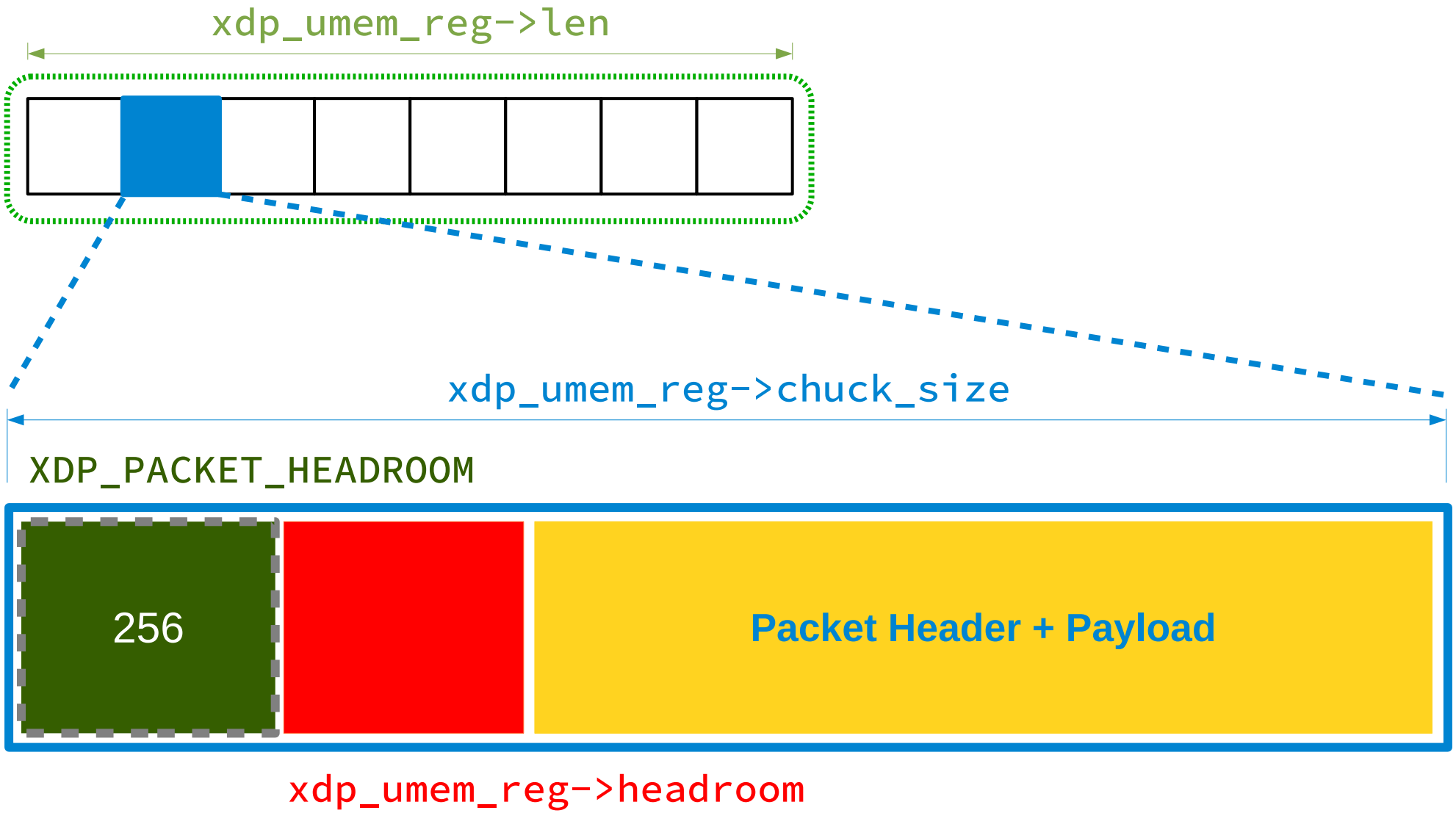
CPUMAP is more customizable!

XDP REDIRECT – XSKMAP

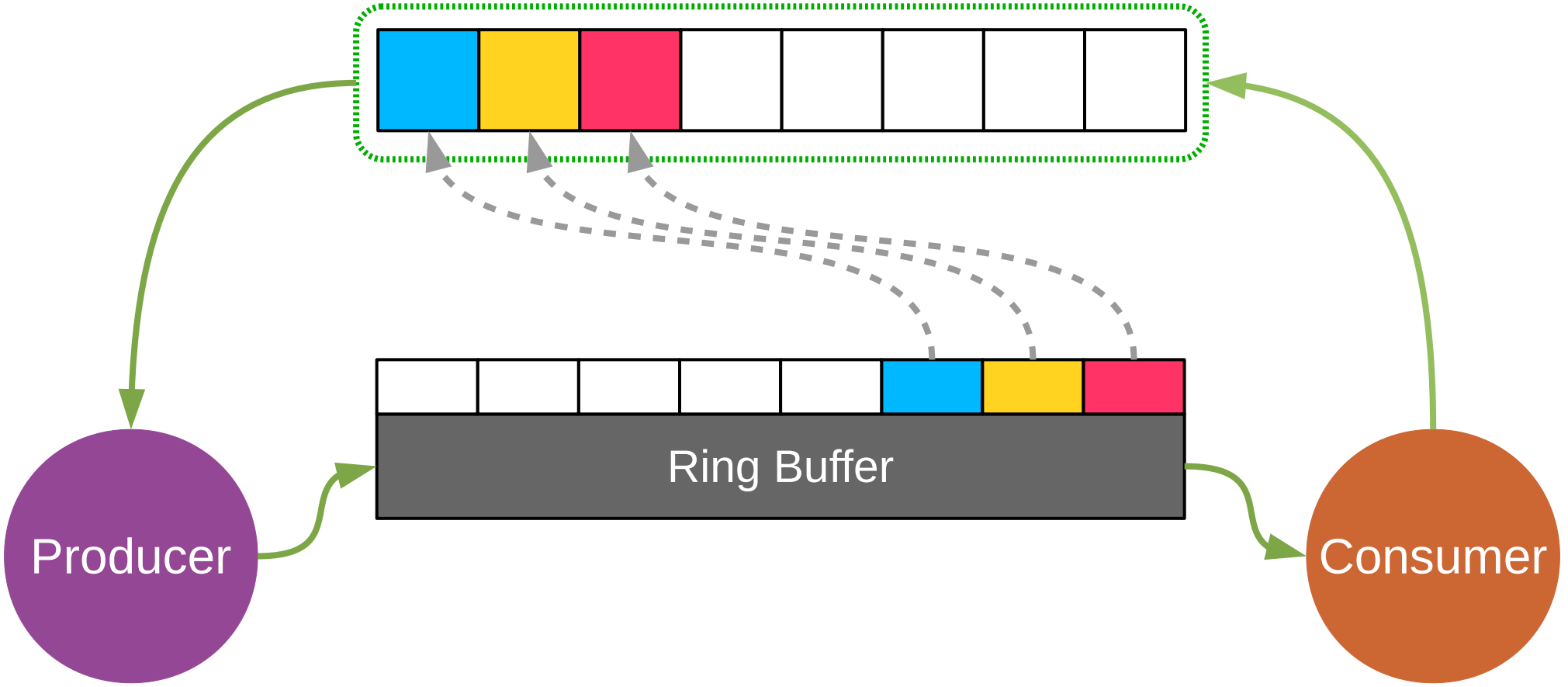
XDP REDIRECT – XSKMAP



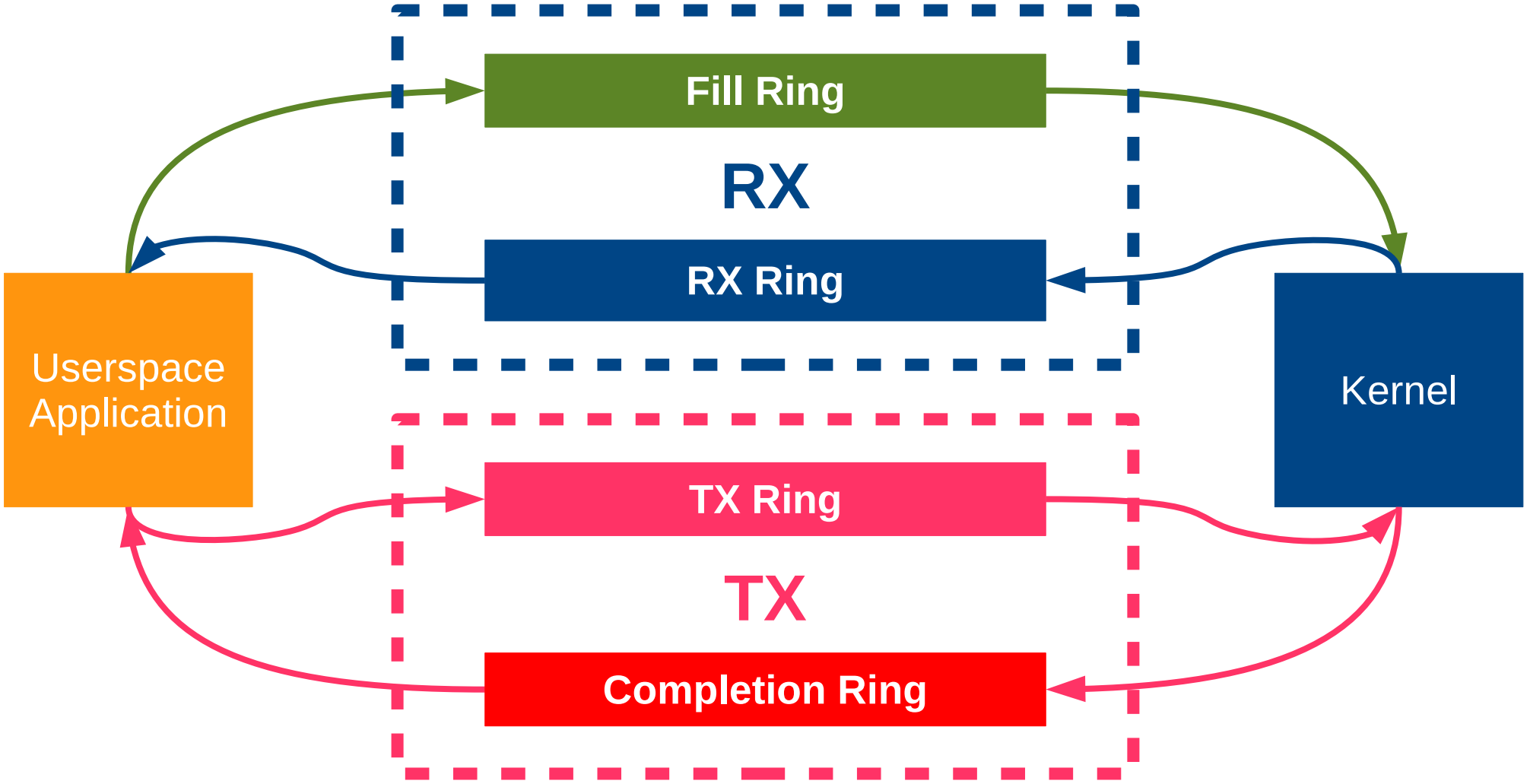
UMEM Chunk


xdp_umem_reg->len
xdp_umem_reg->chuck_size
XDP_PACKET_HEADROOM
256
Packet Header + Payload
xdp_umem_reg->headroom

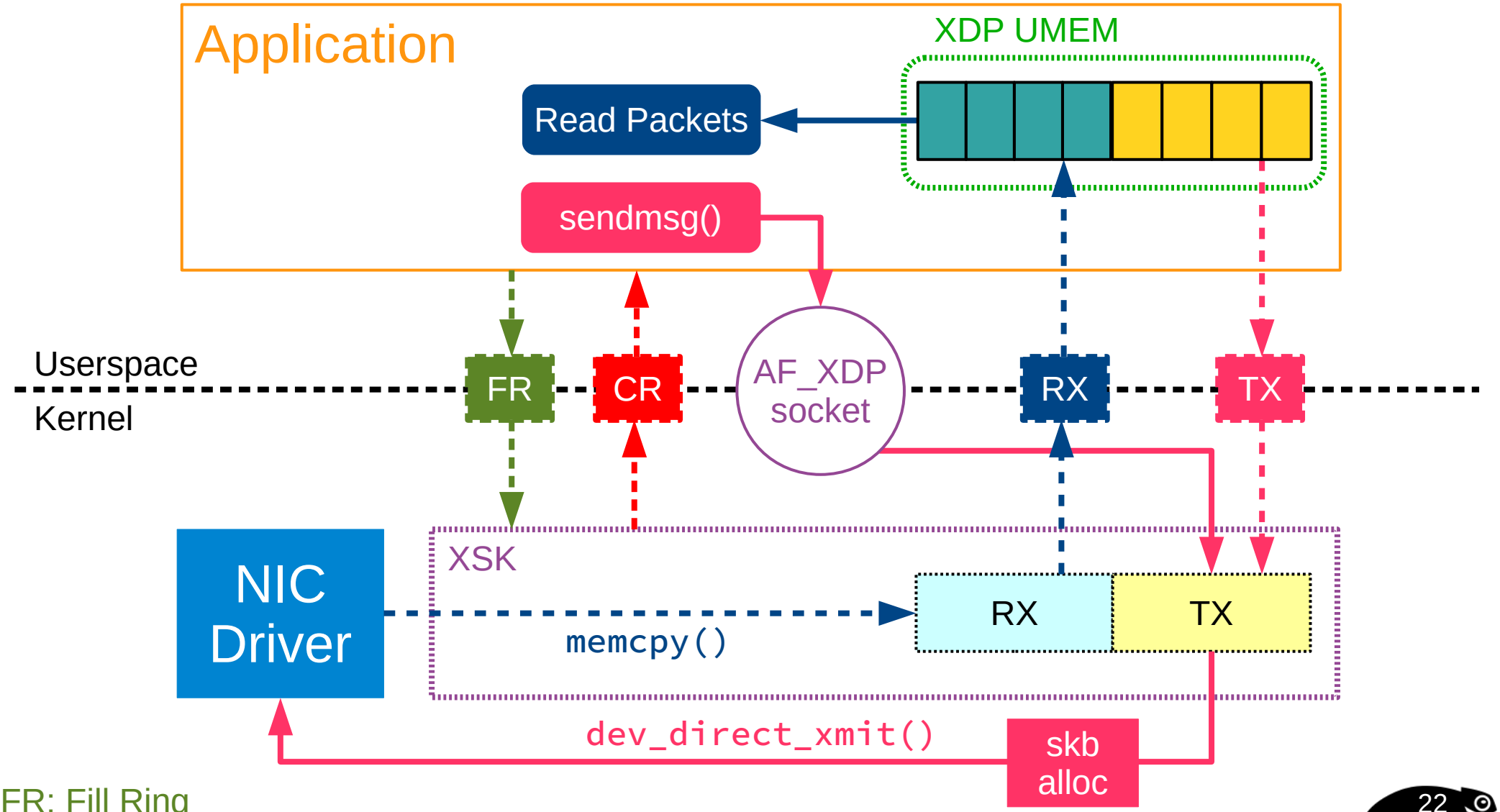
UMEM



UMEM Ring Buffers

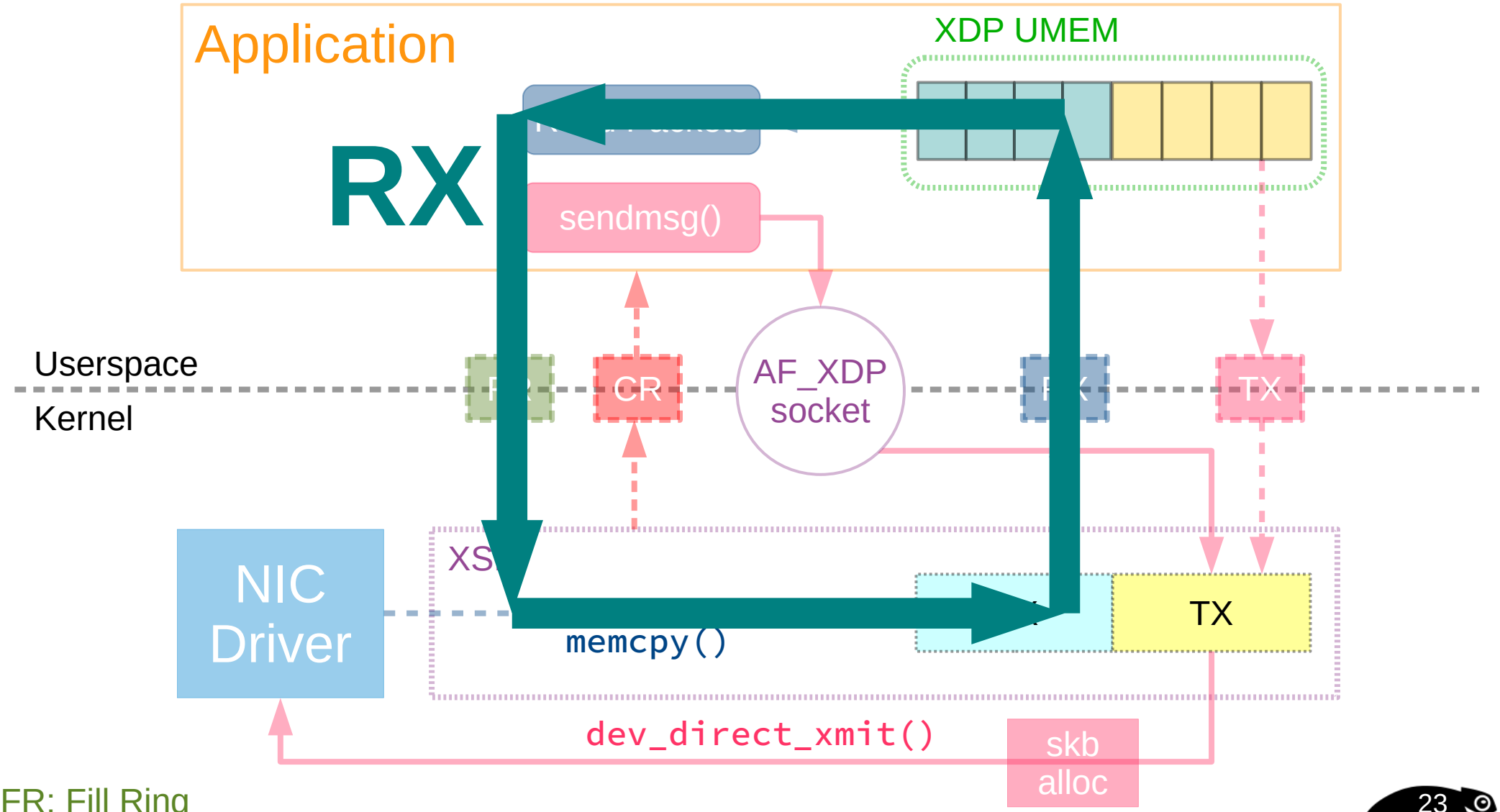


AF_XDP



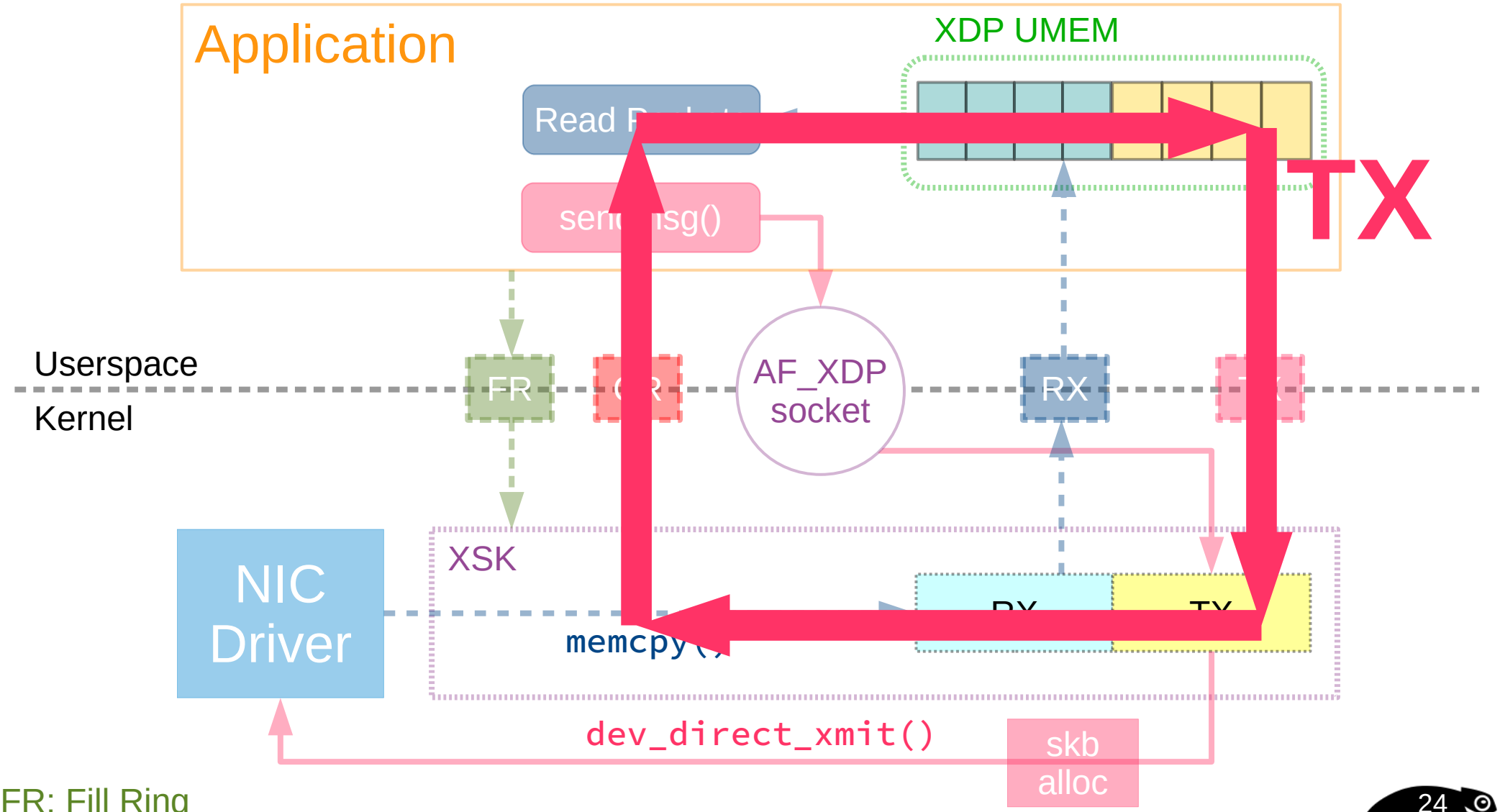
FR: Fill Ring
CR: Completion Ring

AF_XDP



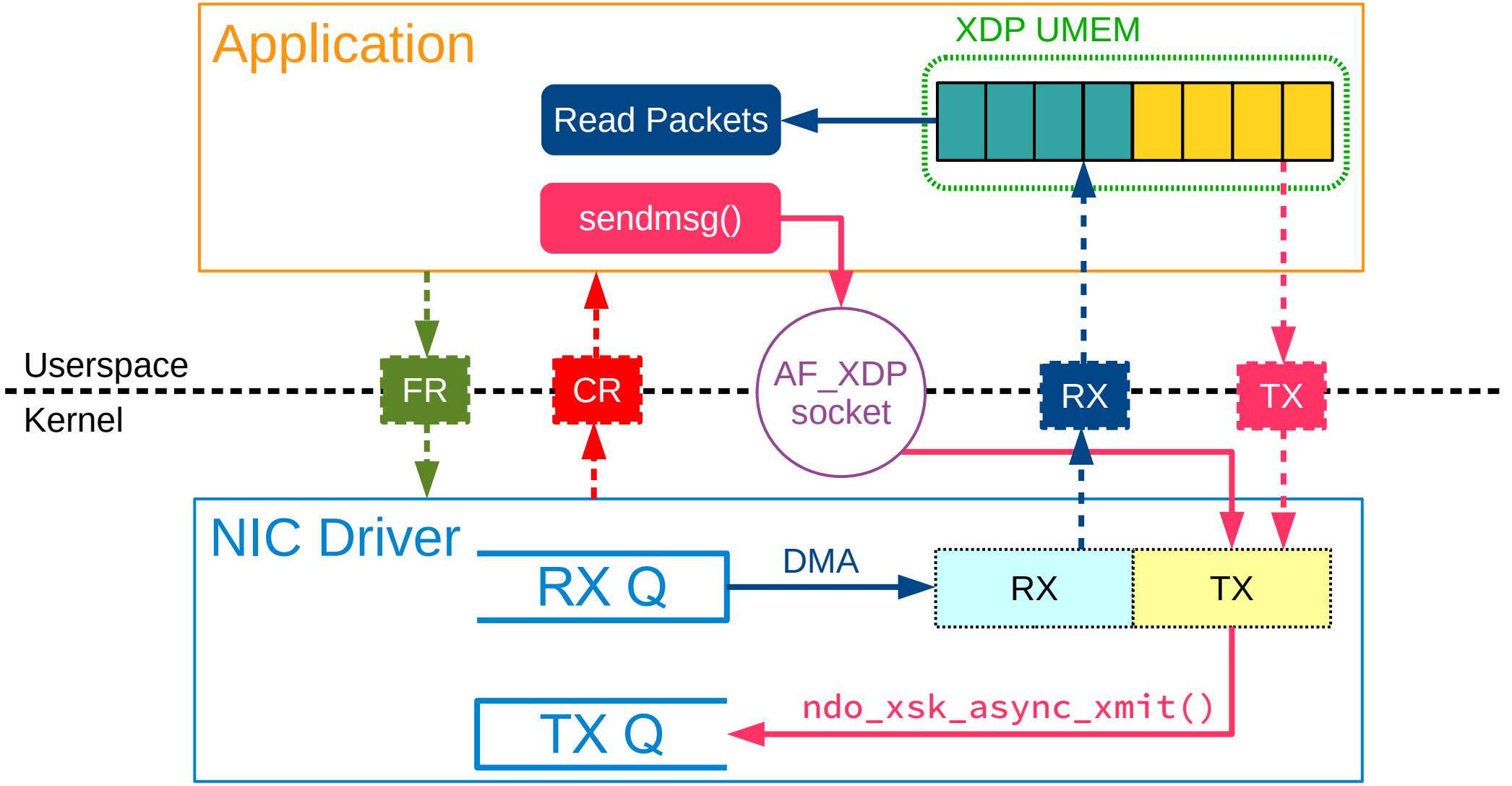
FR: Fill Ring
CR: Completion Ring

AF_XDP



FR: Fill Ring
CR: Completion Ring

AF_XDP (Zero-Copy)



FR: Fill Ring
CR: Completion Ring

Summary

- Processing packets earlier with XDP
- Newly-added XDP REDIRECT
  - DEVMAP, CPUMAP, and XSKMAP
  - Limited driver support (ixgbe, i40e, mlx5, virtio_net)
- Official kernel by-pass with AF_XDP
  - A replacement of DPDK?
  - Incoming Zero-Copy support (i40e & ixgbe in bpf-next)
  - Possible scenario: virtio_net + AF_XDP Zero-Copy

Question?

Thank You!


SUSE
We adapt. You succeed.

References

- **Linux kernel source v4.18:**
  - kernel/bpf/, net/xdp/, net/core/filter.c, and net/core/page_pool.c
- **Kernel git: bpf-next**
  - https://git.kernel.org/pub/scm/linux/kernel/git/bpf/bpf-next.git/
- **BPF and XDP Reference Guide**
  - https://cilium.readthedocs.io/en/stable/bpf/
- **Accelerating networking with AF_XDP**
  - https://lwn.net/Articles/750845/
- **BPF Features by Linux Kernel Version**
  - https://github.com/iovisor/bcc/blob/master/docs/kernel-versions.md
- **FOSDEM 2018: Fast Packet Processing in Linux with AF_XDP**
  - https://archive.fosdem.org/2018/schedule/event/af_xdp/