Kernel Networking Walkthrough

LinuxCon 2015, Seattle

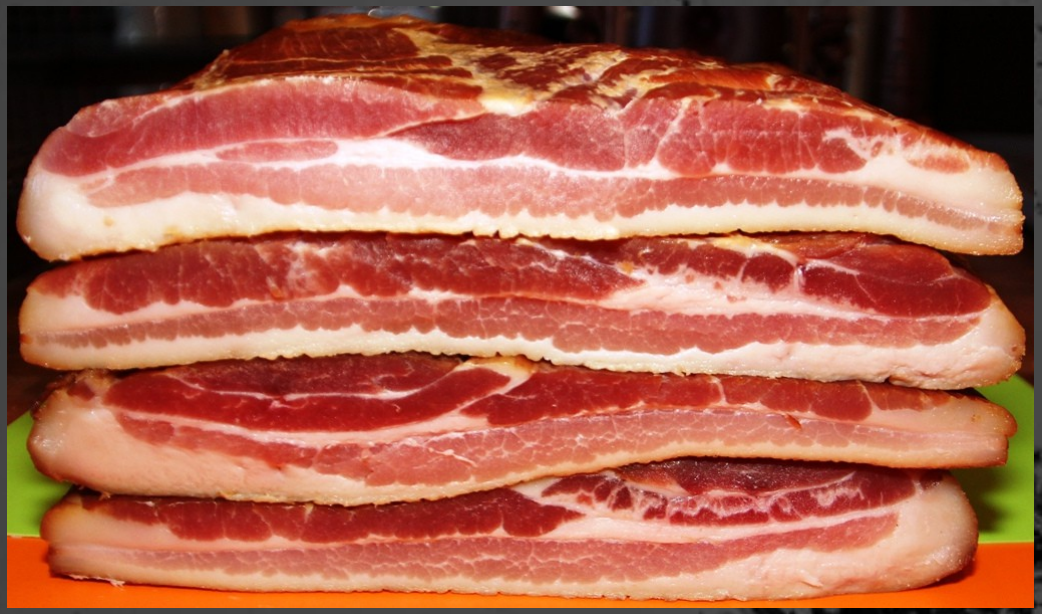
Thomas Graf
Kernel & Open vSwitch Team
Noiro Networks (Cisco)

Agenda

- **Getting packets from/to the NIC**
  - NAPI, Busy Polling, RSS, RPS, XPS, GRO, TSO
- **Packet processing**
  - RX Handler, IP Processing, TCP Processing, TCP Fast Open
- **Queuing from/to userspace**
  - Socket Buffers, Flow Control, TCP Small Queues
- **Q&A**

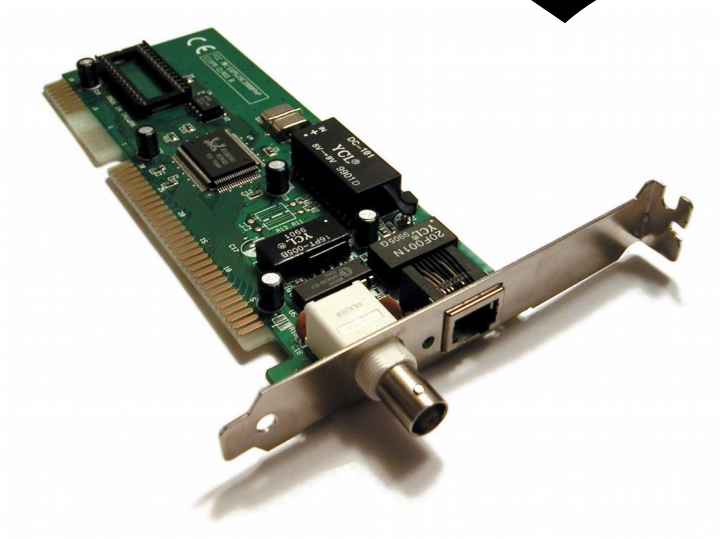
Touring the Network Stack

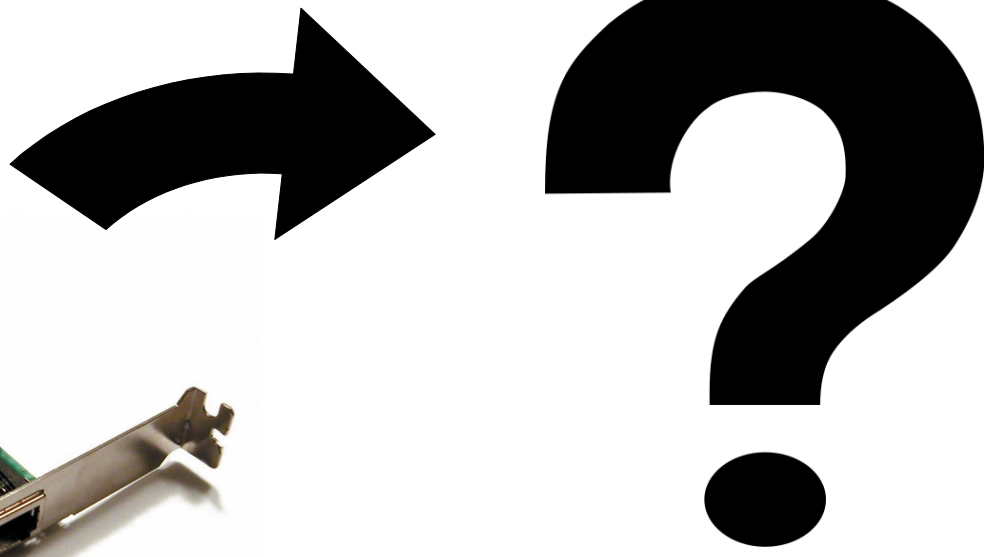
Expectation



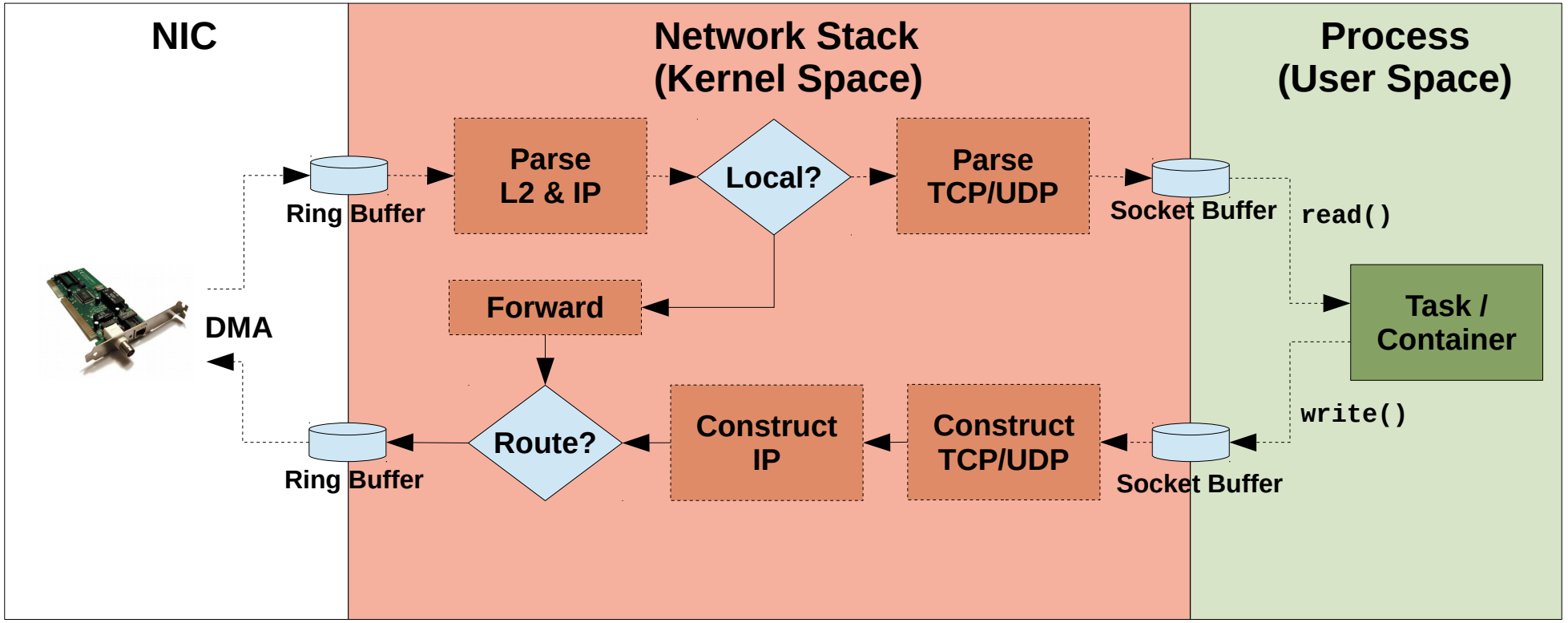
Reality

How does a packet get in and out of the Network Stack?

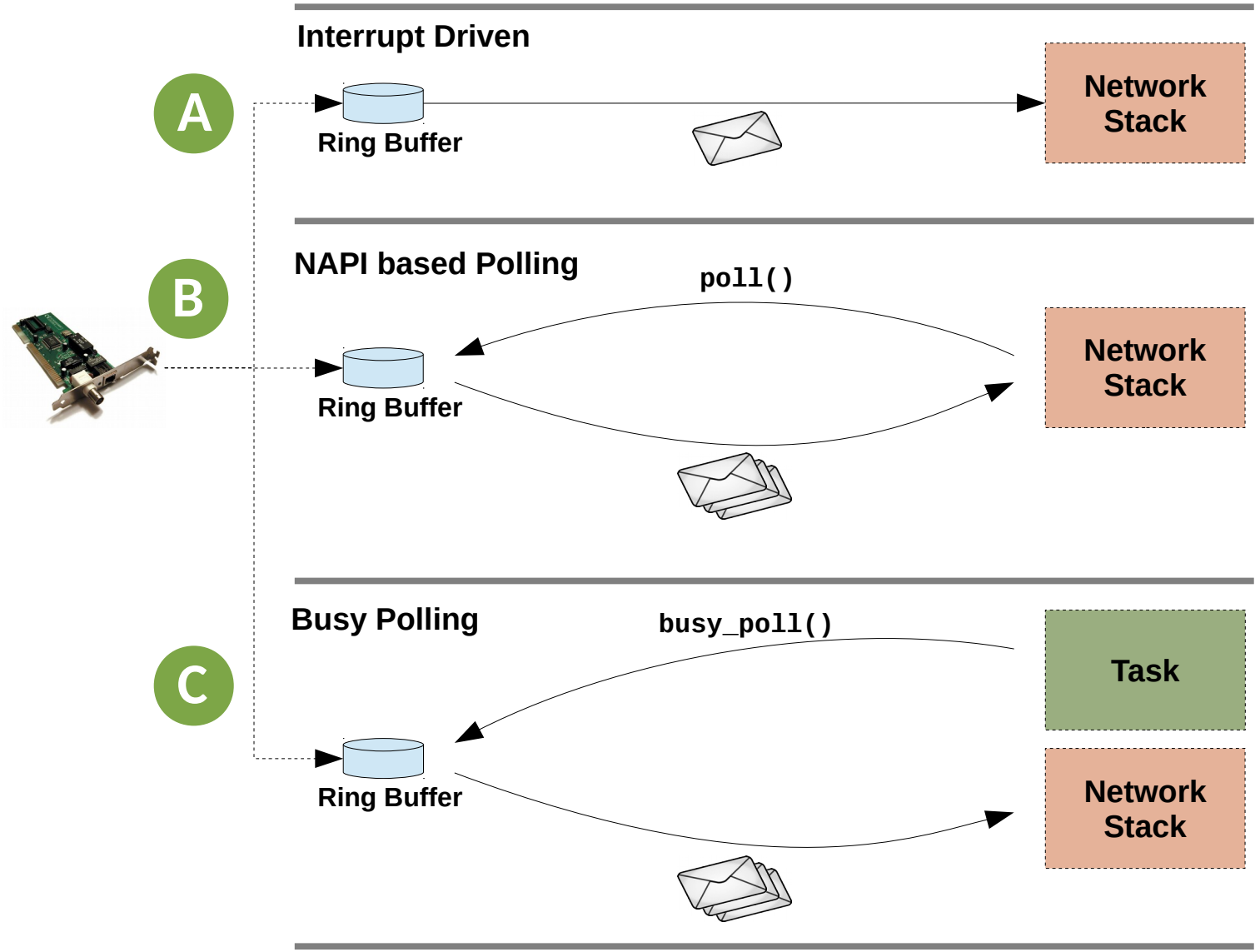




Receive & Transmit Process

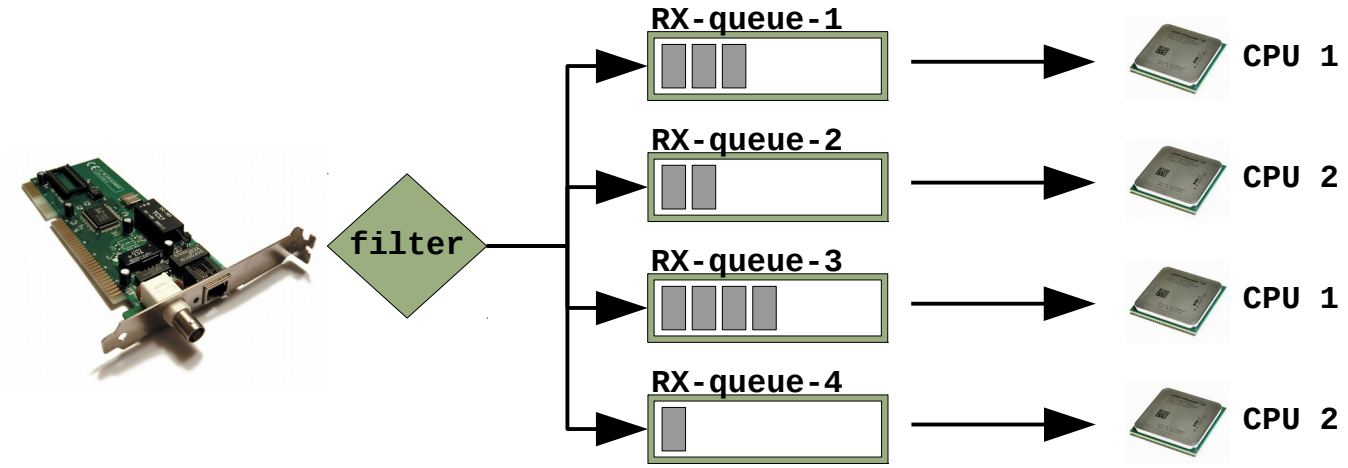


The 3 ways into the Network Stack



RSS – Receive Side Scaling

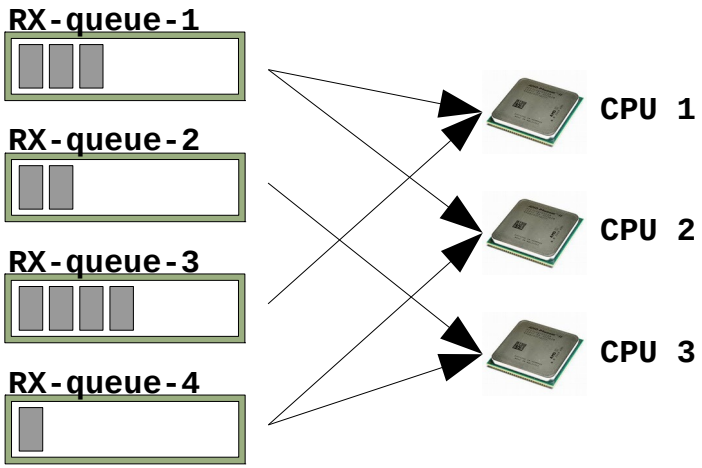
- NIC distributes packets across multiple RX queues allowing for parallel processing.
- Separate IRQ per RX queue, thus selects CPU to run hardware interrupt handler on.



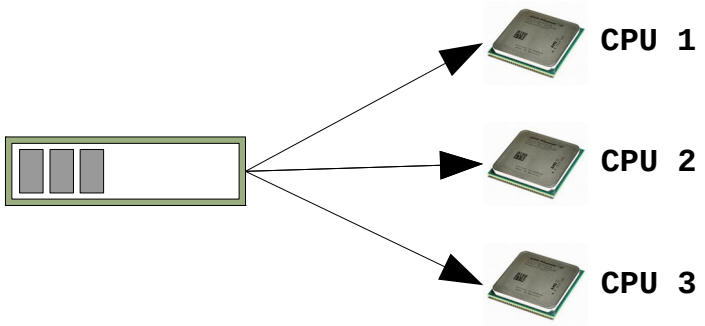
RPS – Receive Packet Steering

- Software filter to select CPU # for processing
- Use it to …

… redo queue - CPU mapping

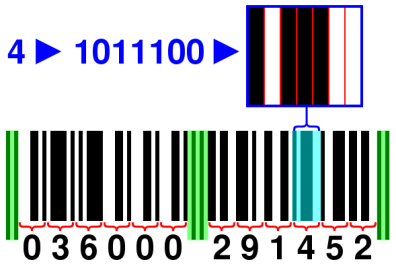


… distribute single queue to multiple CPUs



Hardware Offload

- RX/TX Checksumming
  - Perform CPU intensive checksumming in hardware.

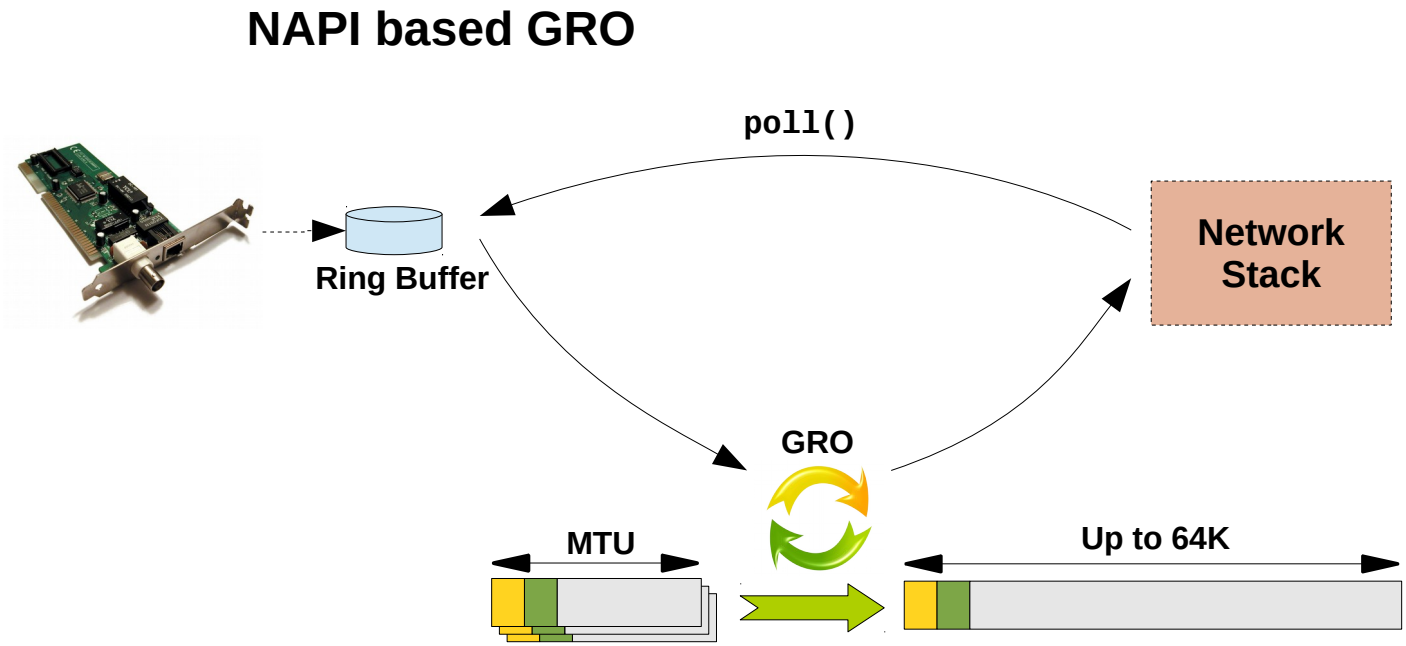


- Virtual LAN filtering and tag stripping
  - Strip 802.1Q header and store VLAN ID in network packet meta data.
  - Filter out unsubscribed VLANs.

- Segmentation Offload

Generic Receive Offload

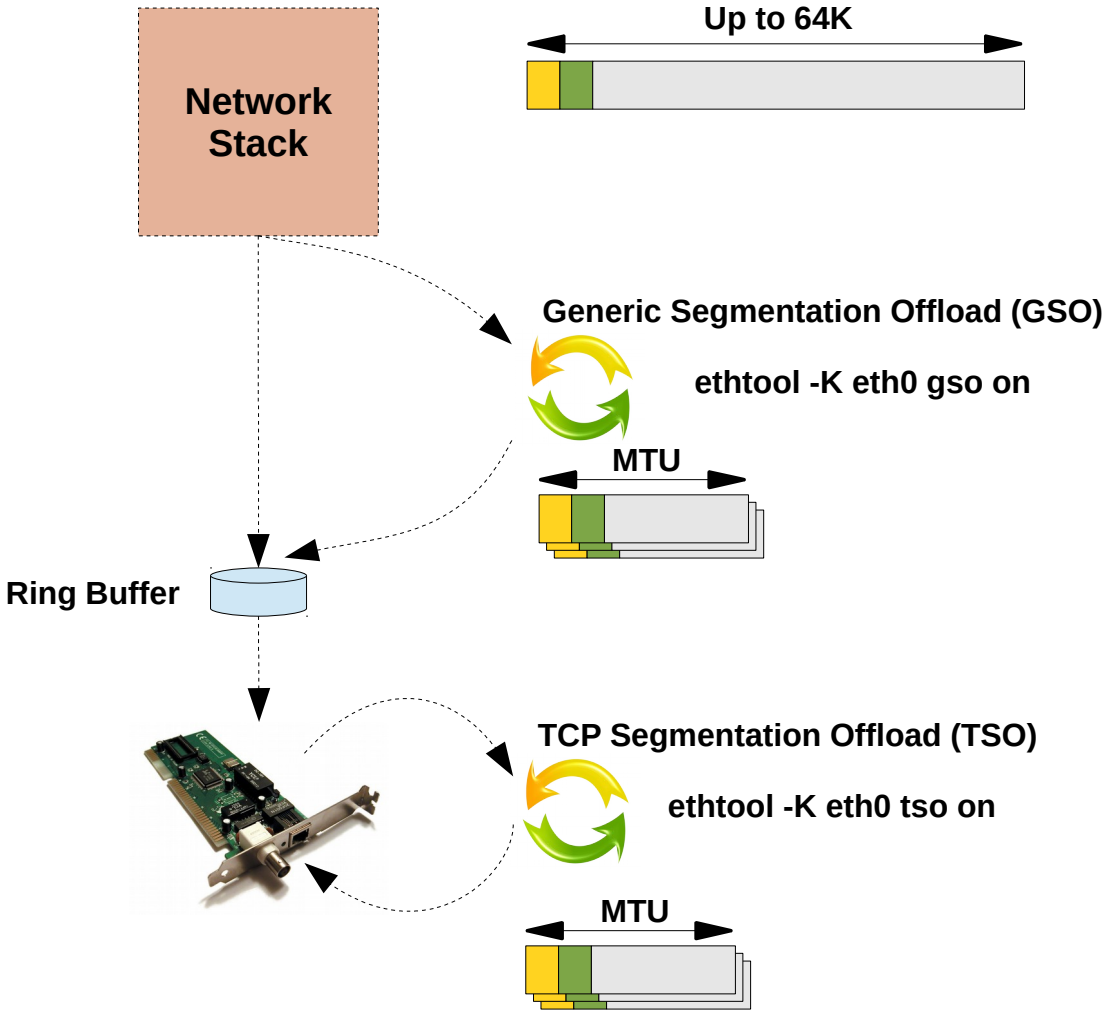
(ethtool -K eth0 gro on)



It's more effective to process 1x64K bytes packet instead of 40x1500 bytes packets.

Segmentation Offload

(ethtool -K eth0 tso on)
(ethtool -K eth0 gso on)



How does a packet get through the Network Stack?

(c) Karen Sagovac

Packet Processing

Link Layer

Packet Socket
ETH_P_ALL

tcpdump

Ingress QoS

RX Handler

Bridge

Open vSwitch

Team

Bonding

macvlan

macvtap

Proto Handler

IPv4

IPv6

ARP

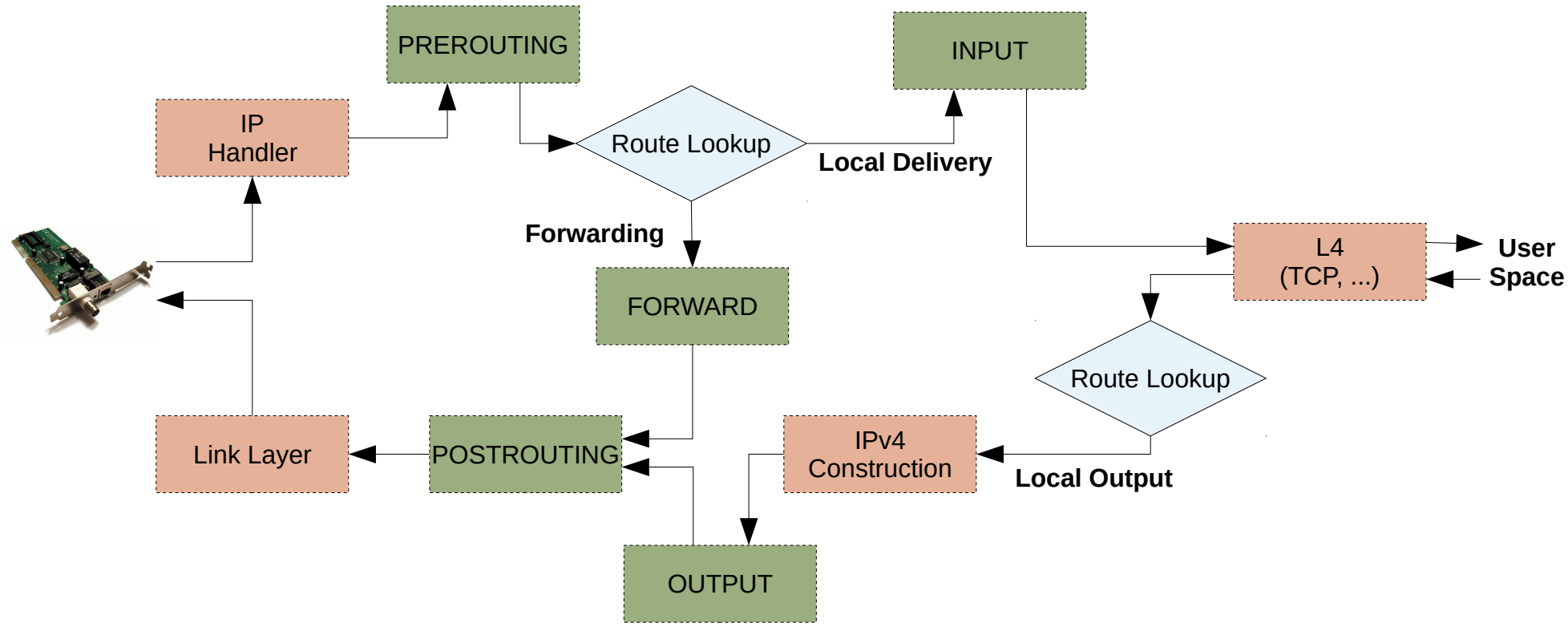
IPX

...

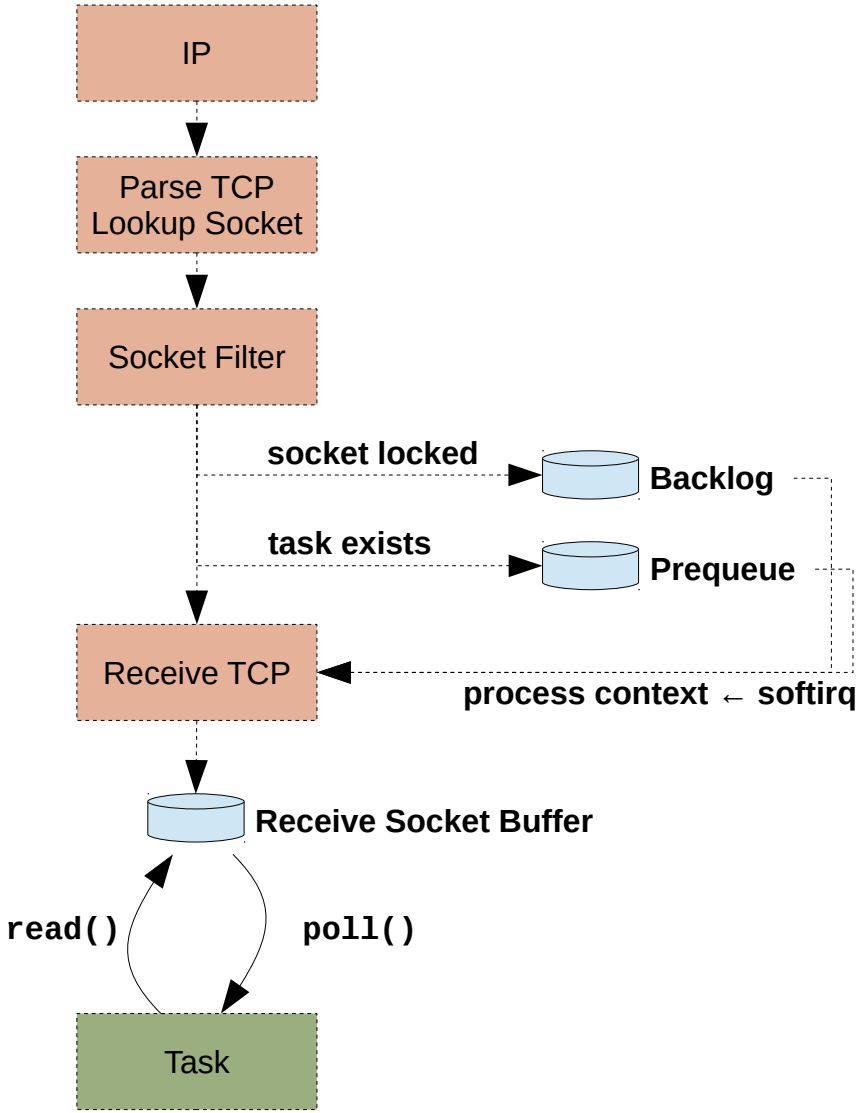
Drop

The Feast!

IP Processing

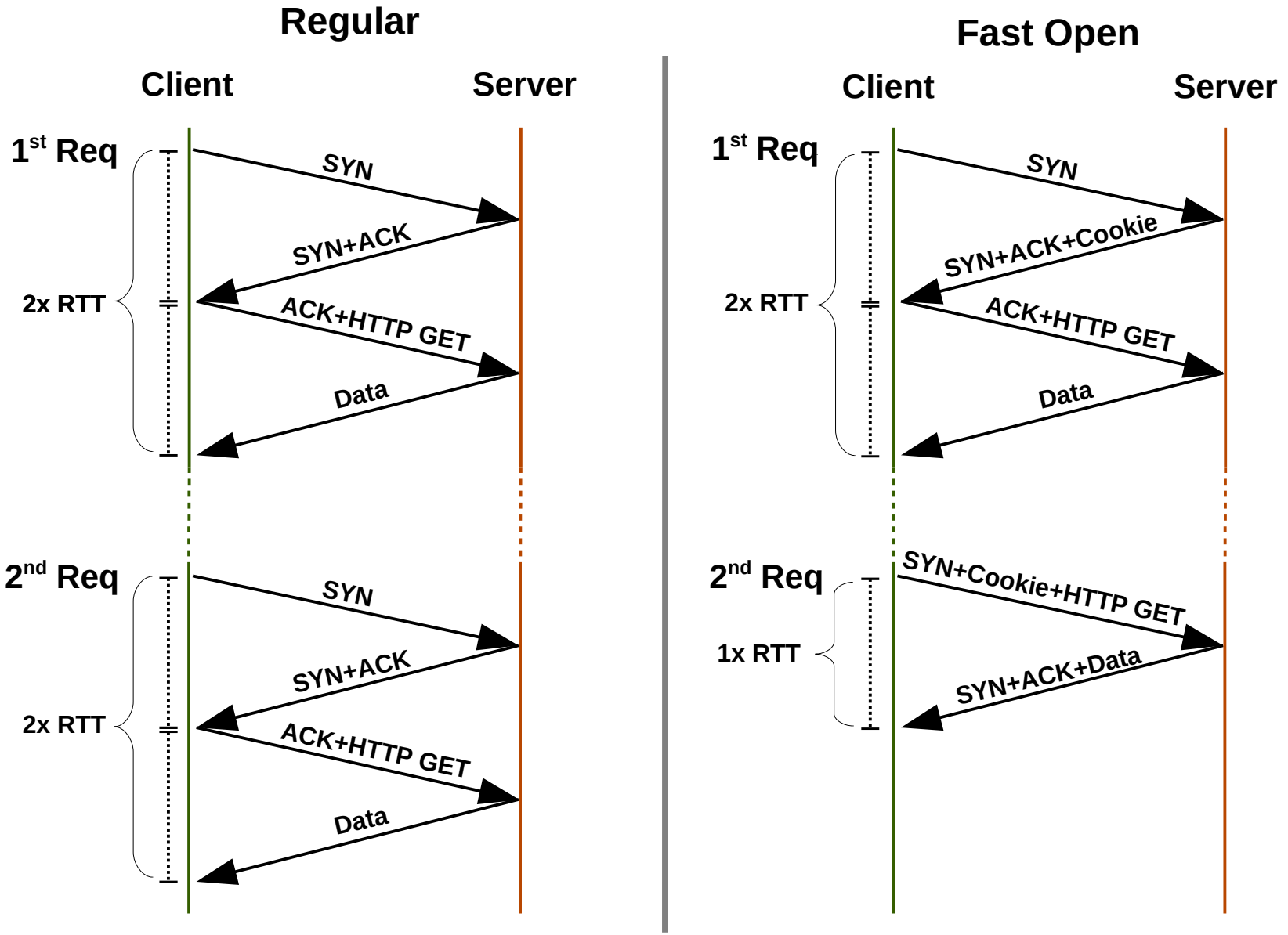


TCP Processing



TCP Fast Open

(net.ipv4.tcp_fastopen)

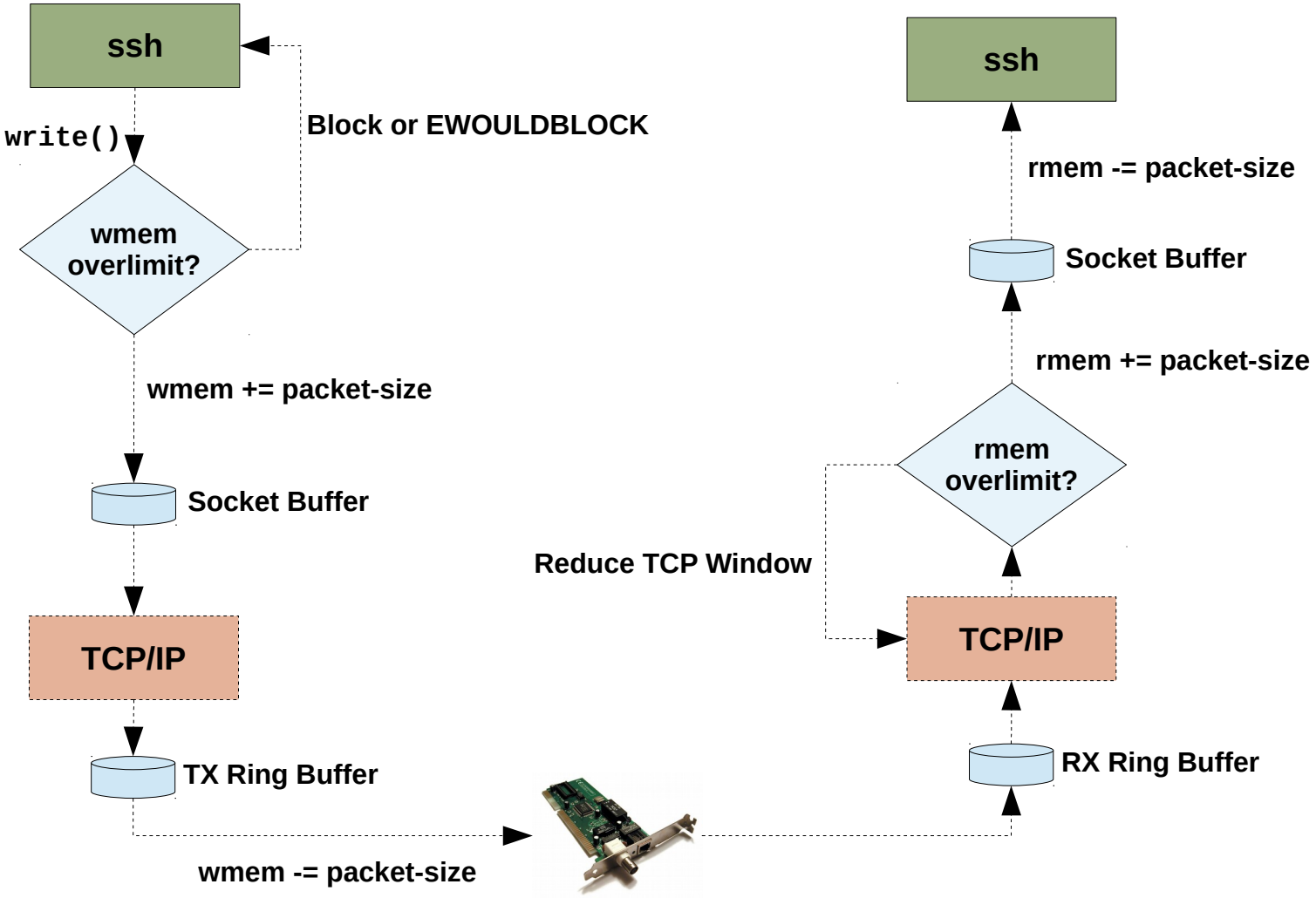


Memory Accounting & Flow Control

A Stack of Wheat ready for transport

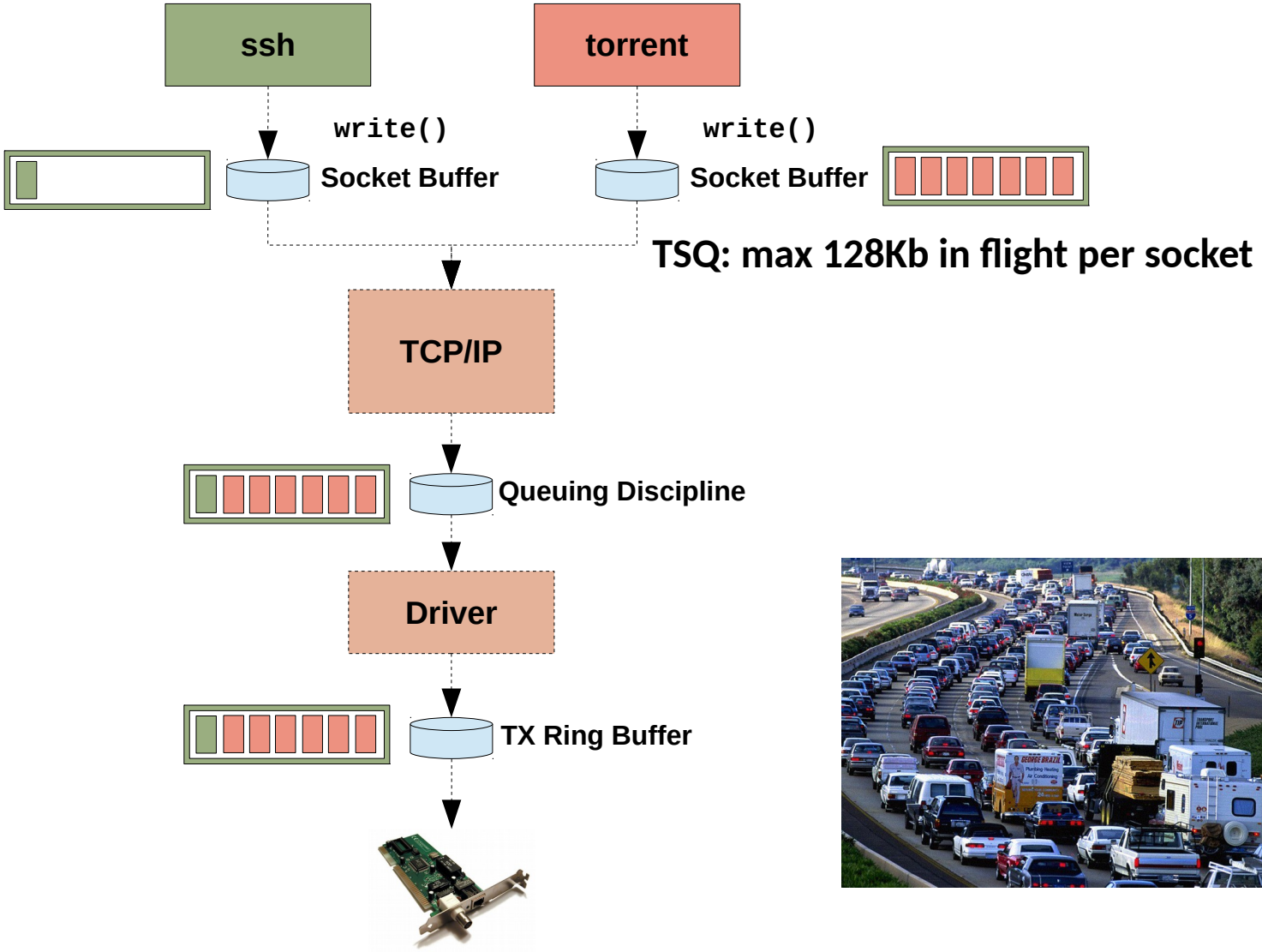
Socket Buffers & Flow Control

(net.ipv4.tcp_{r|w}mem)



TCP Small Queues

(net.ipv4.tcp_limit_output_bytes)



Q&A

Contact:

- **E-Mail: tgraf@suug.ch**
- **Twitter: @tgraf__**