Kernel Networking Walkthrough

Thomas Graf – Principal Software Engineer
Networking Services
Red Hat
Feb 7, 2014

Agenda

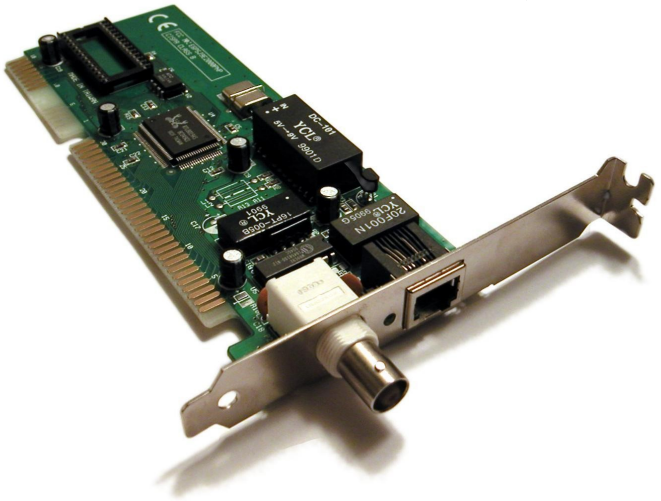
- How does a packet get in and out of the net stack?
  - NAPI, Busy Polling, RSS, RPS, XPS, GRO, TSO
- How does a packet get through the net stack?
  - RX Handler, IP Processing, TCP Processing, TCP Fast Open
- How to account for memory and do flow control?
  - Socket Buffers, Flow Control, TCP Small Queues
- Q&A

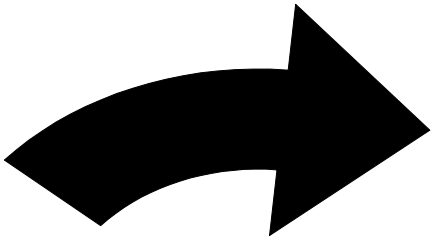
Touring the Network Stack

Expectation

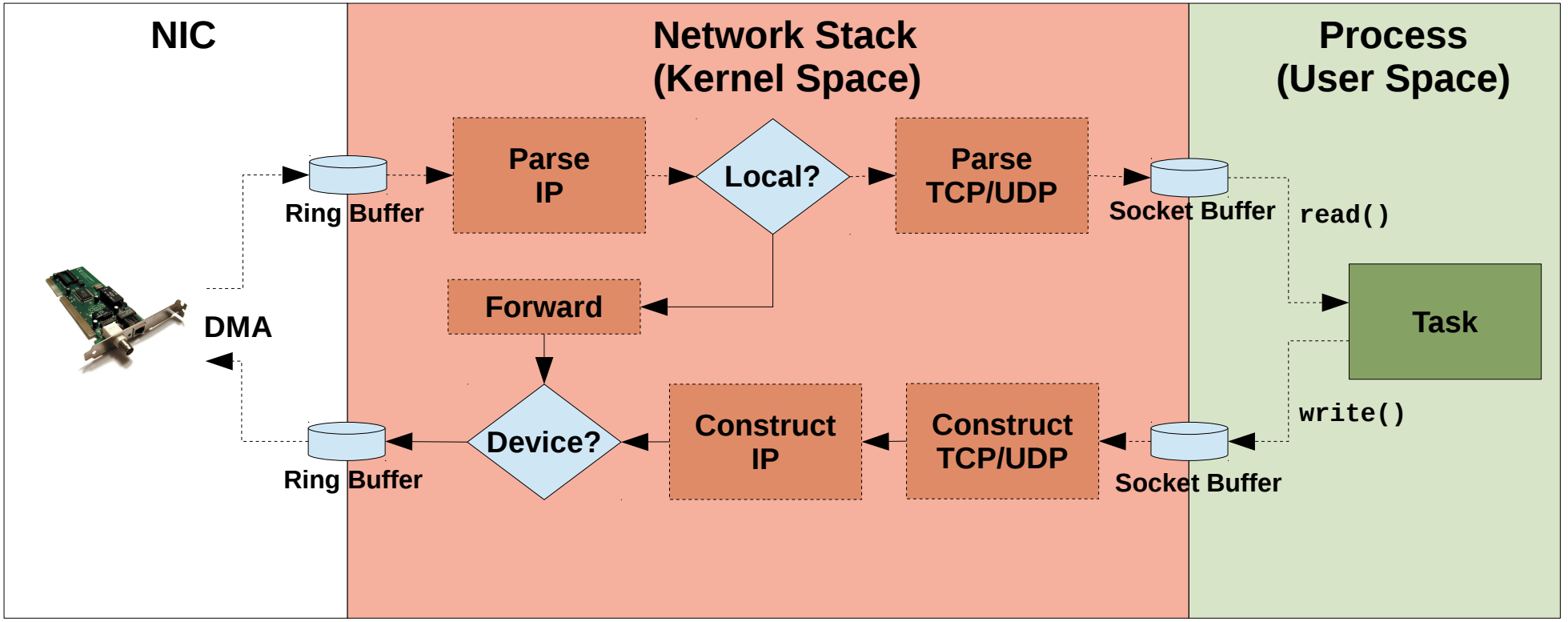
Reality

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

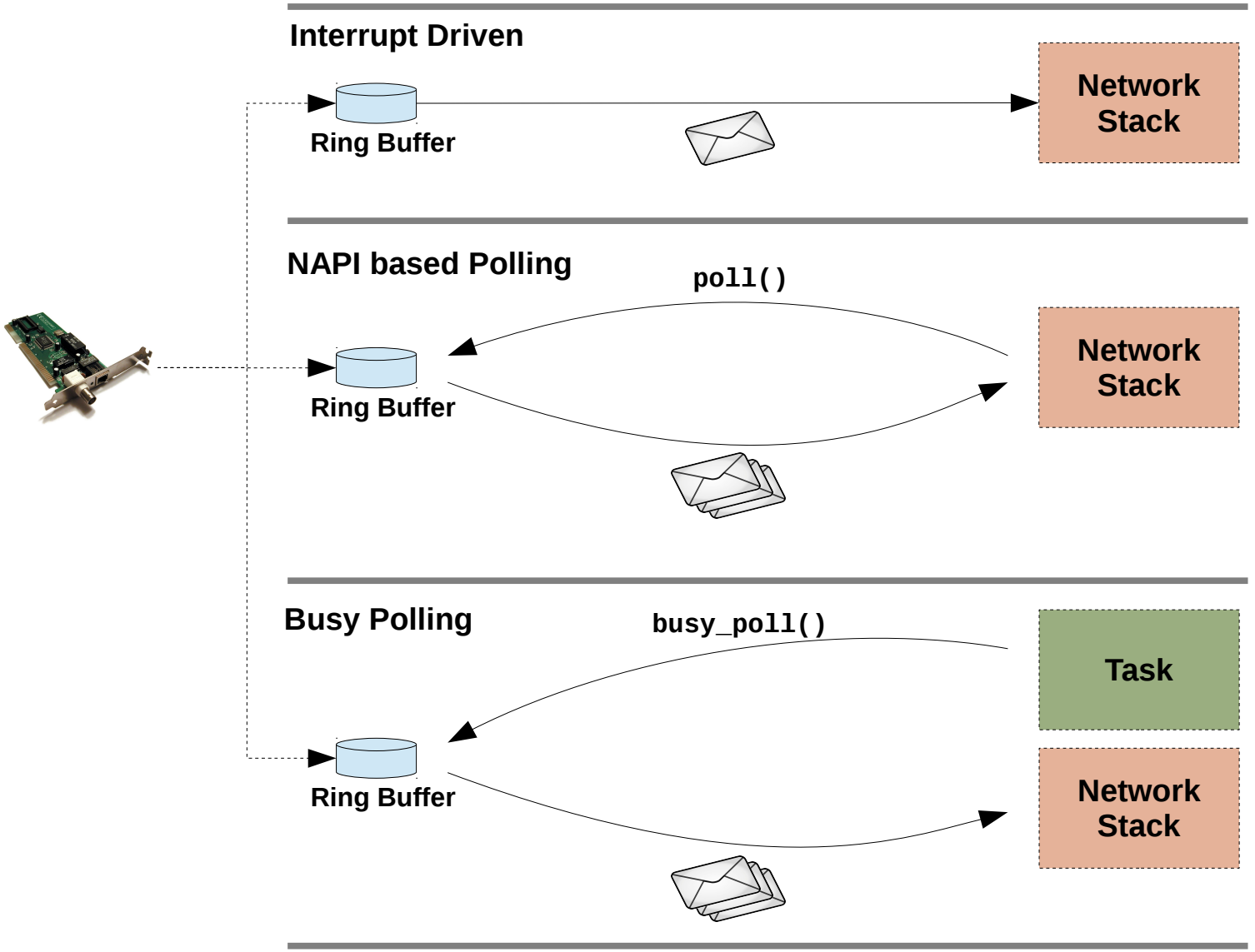




Receive & Transmit Process

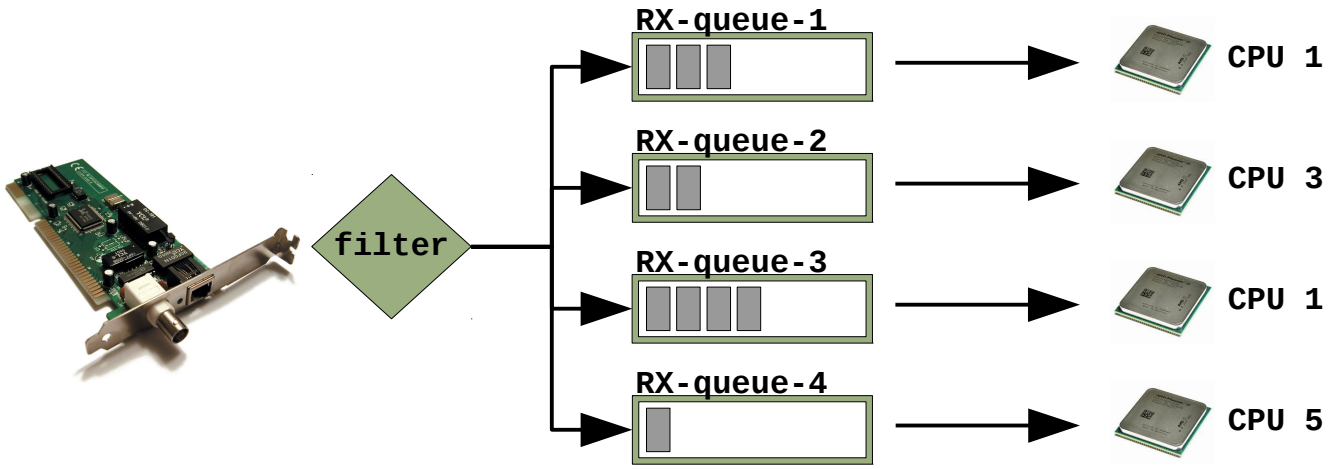


The 3 ways into the Network Stack



RSS – Receive Side Scaling

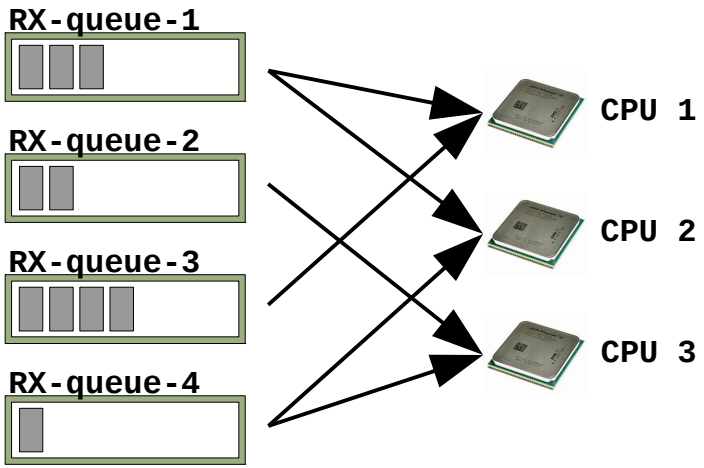
- NIC distributes packets across multiple RX queues allowing for parallel processing.
- Separate IRQ per RX queue, thus selects CPU to run hardware interrupt handler on.



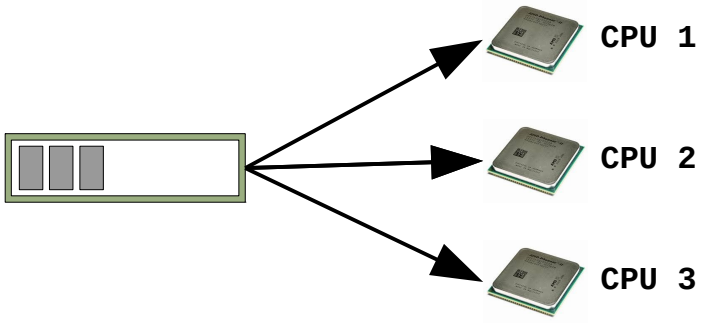
RPS – Receive Packet Steering

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

... redo queue - CPU mapping

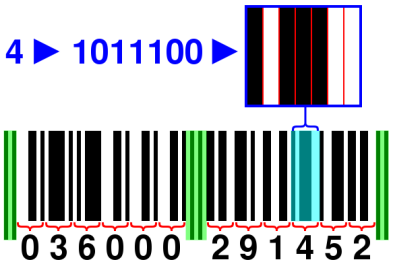


... distribute single queue to multiple CPUs



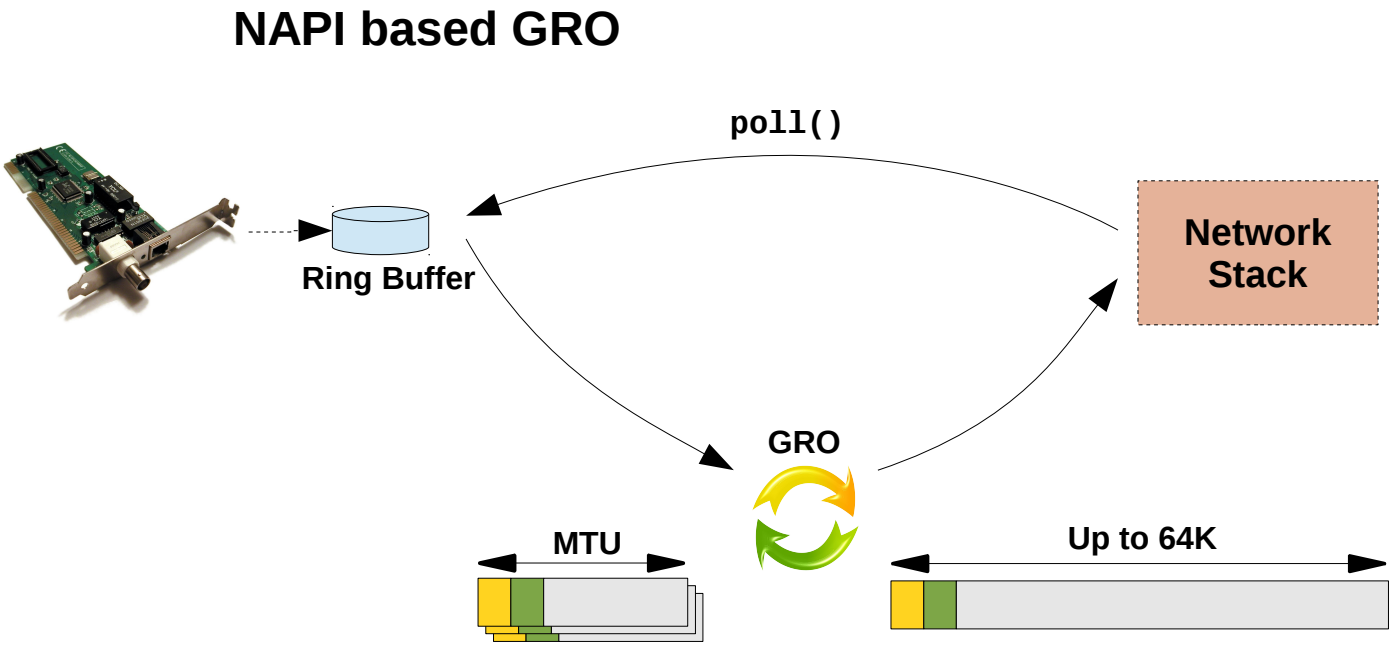
Hardware Offload

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

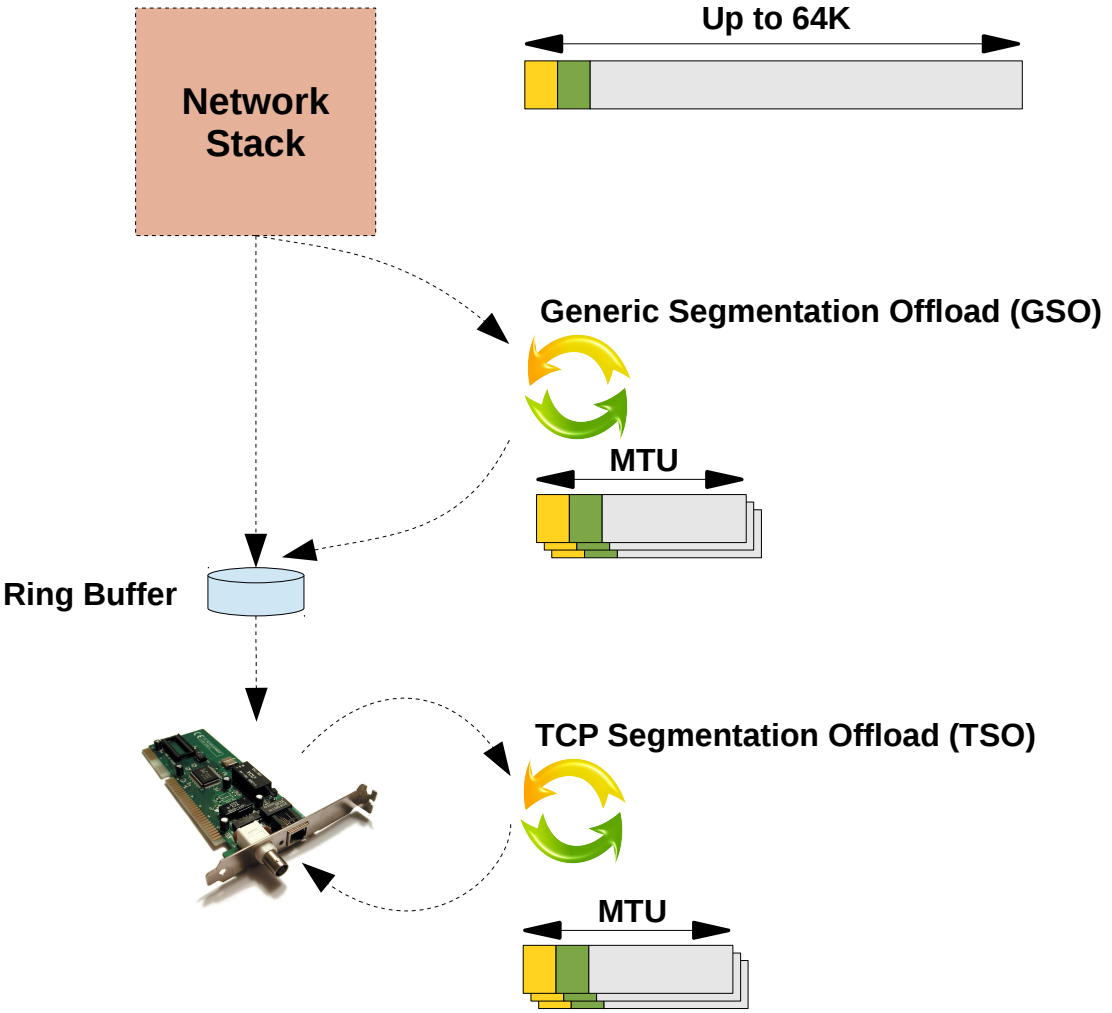


- Virtual LAN filtering and tag stripping
  - Strip 802.1Q header and store VLAN ID in network packet meta data.
  - Filter out unsubscribed VLANs.

Generic Receive Offload



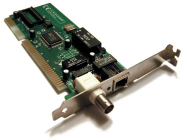
Segmentation Offload



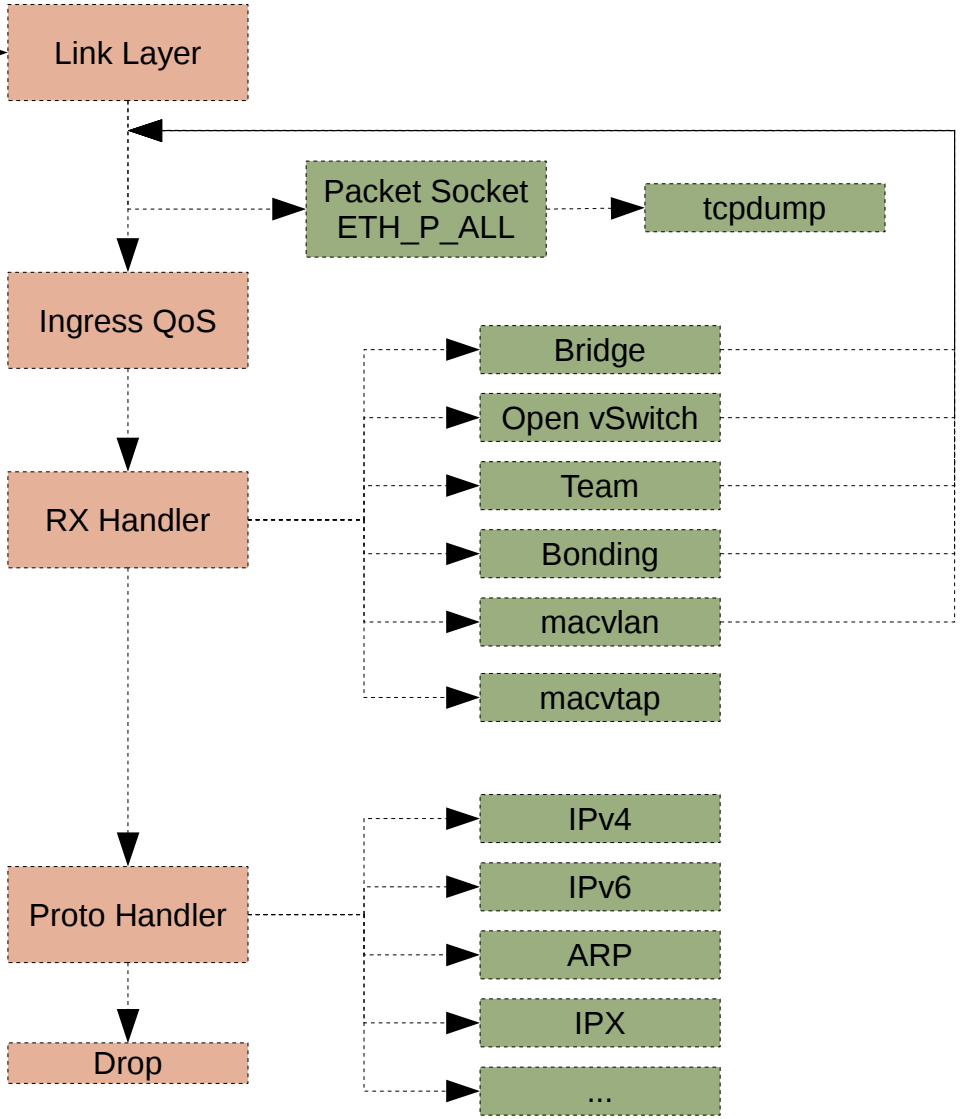
How does a packet get through the Network Stack?

(c) Karen Sagovac

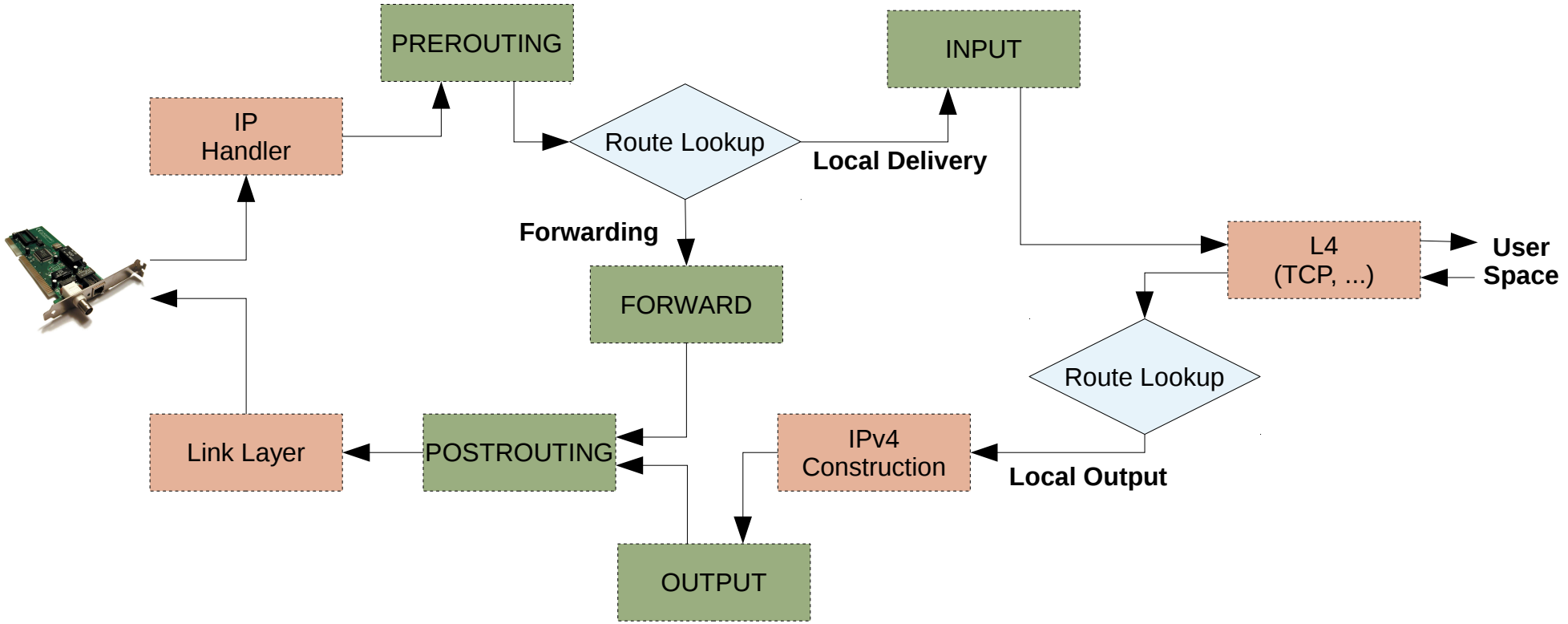
Packet Processing



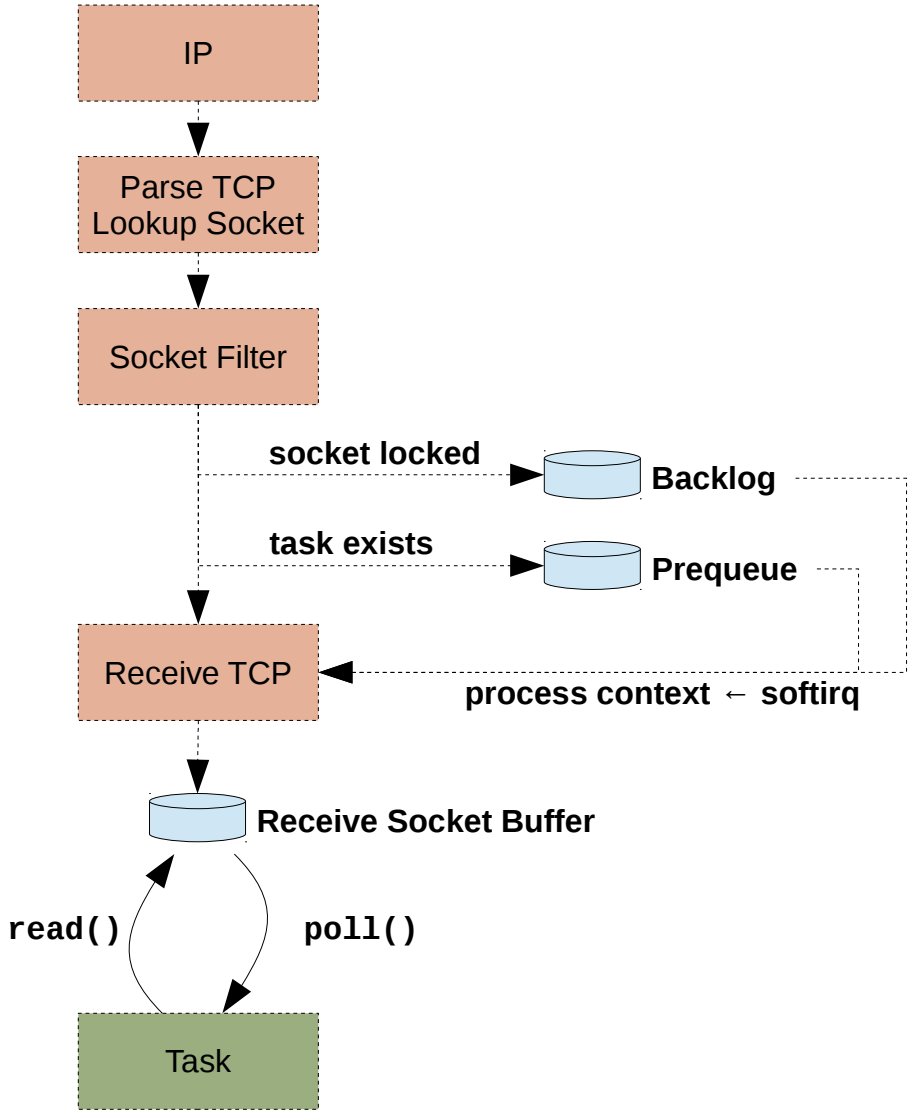
Feast of the hungry chicks



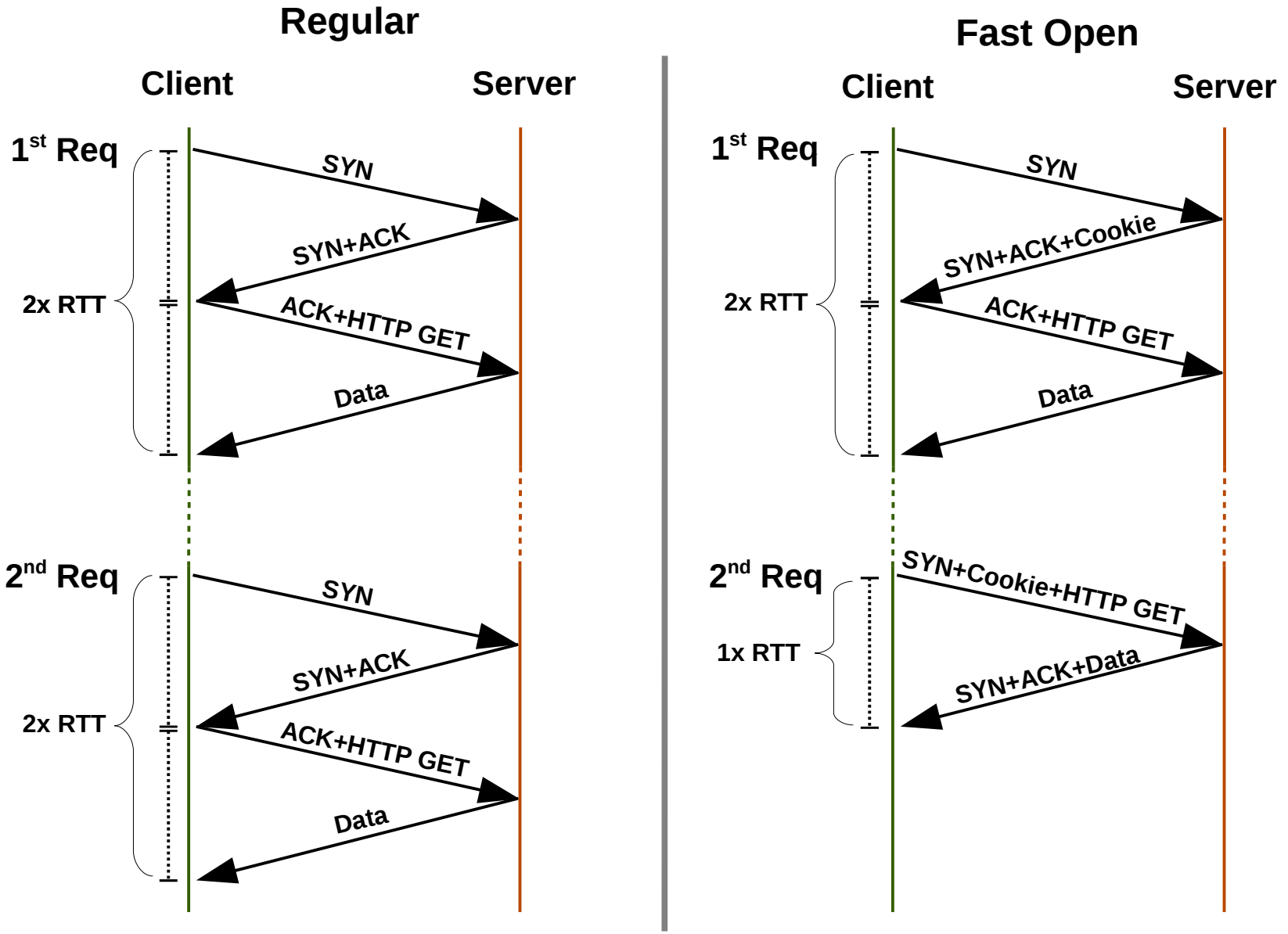
IP Processing



TCP Processing



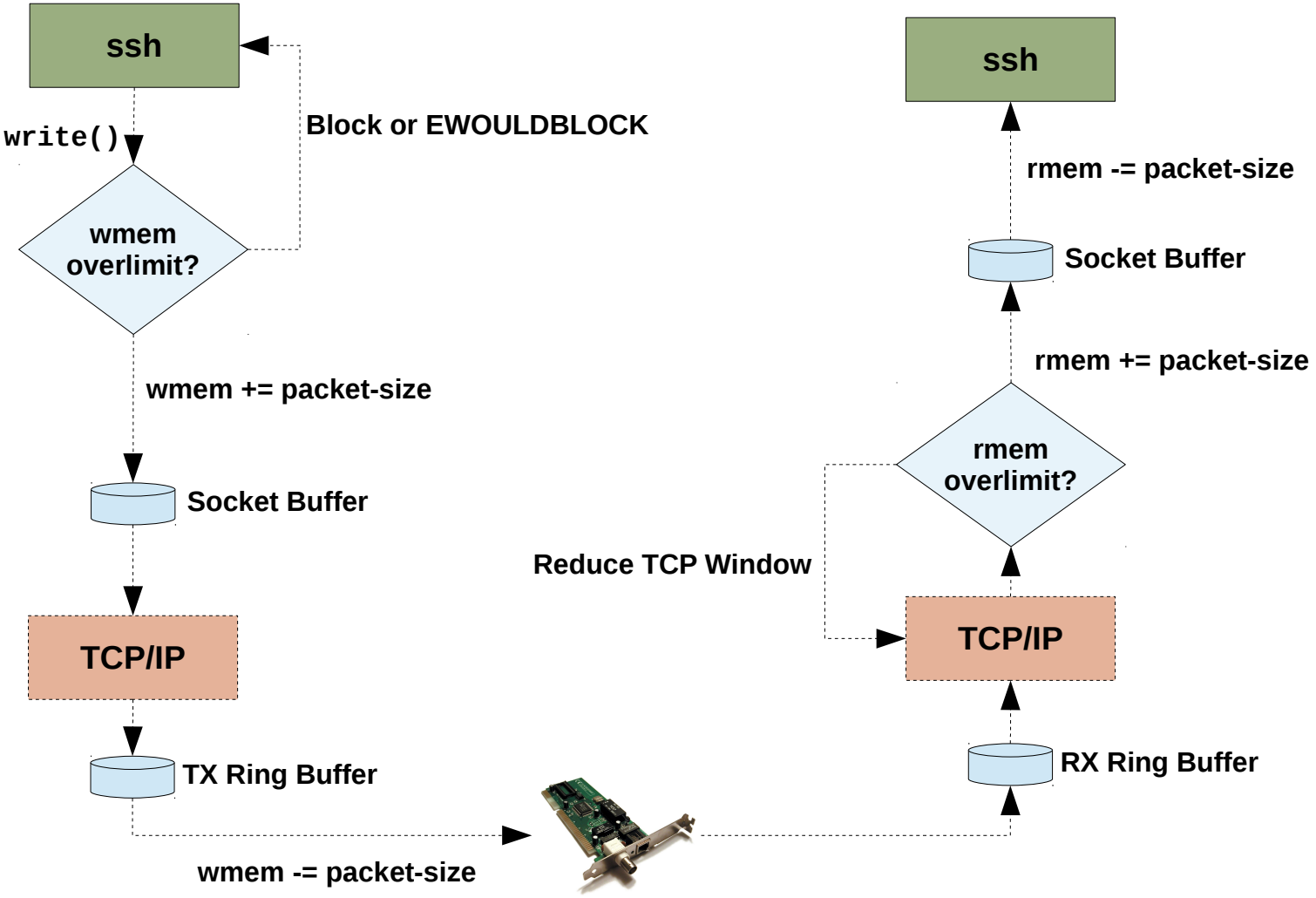
TCP Fast Open
(net.ipv4.tcp_fastopen)



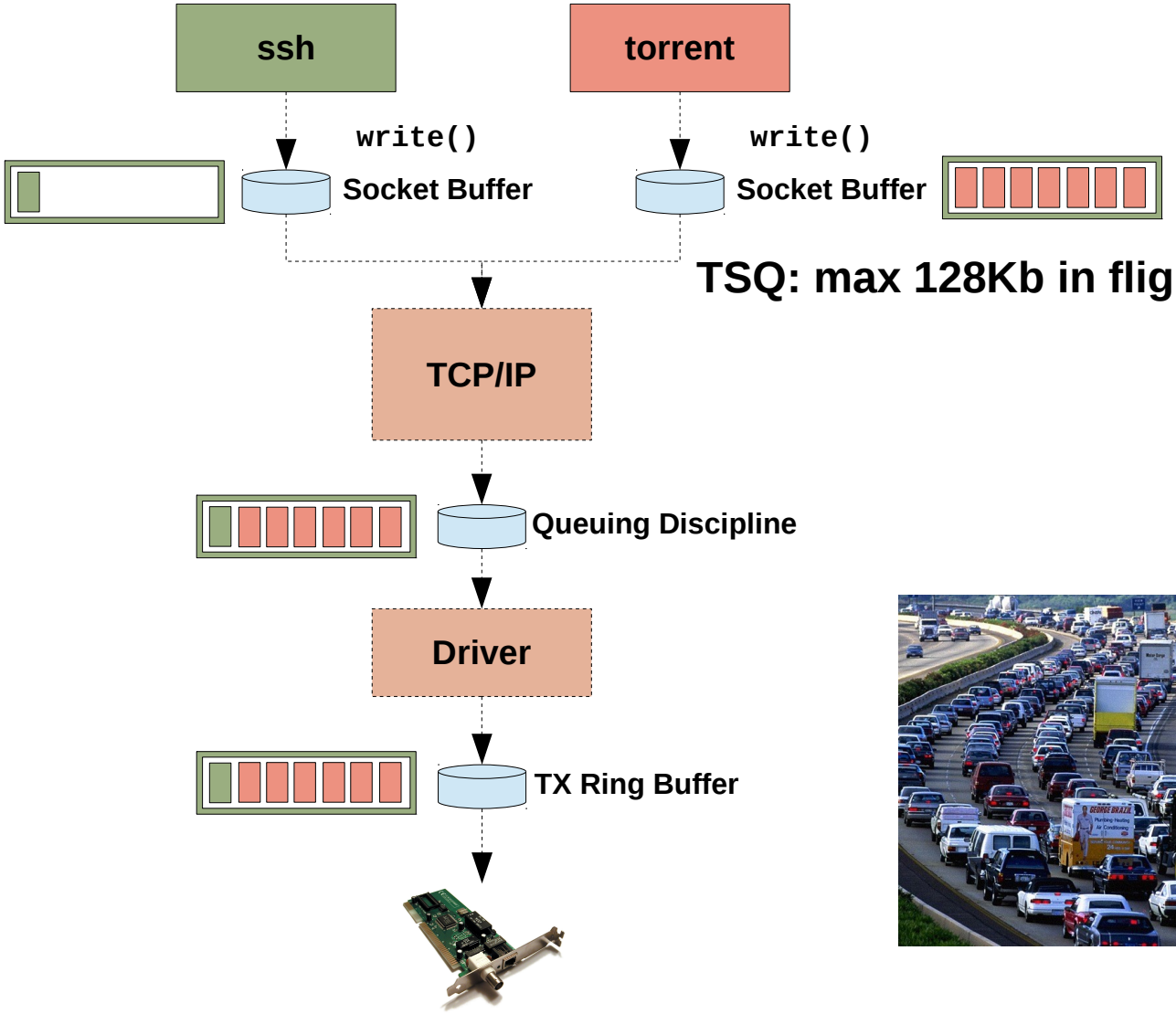
Memory Accounting & Flow Control

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

Feedback Page

- http://devconf.cz/f/1

Coming Up Next:

NetworkManager for Enterprise

Dan Williams