Introduction to Linux Kernel TCP/IP procotol stack

雕梁
核心系统服务器平台组
diaoliang@taobao.com
simohayha.bobo@gmail.com
http://www.pagefault.info
2011/01/15

Agenda

Introduction

Networking code in the Linux kernel tree

L2 (Link Layer)

L3 (Network Layer)

L4 (Transport Layer)

Config and benchmark tools

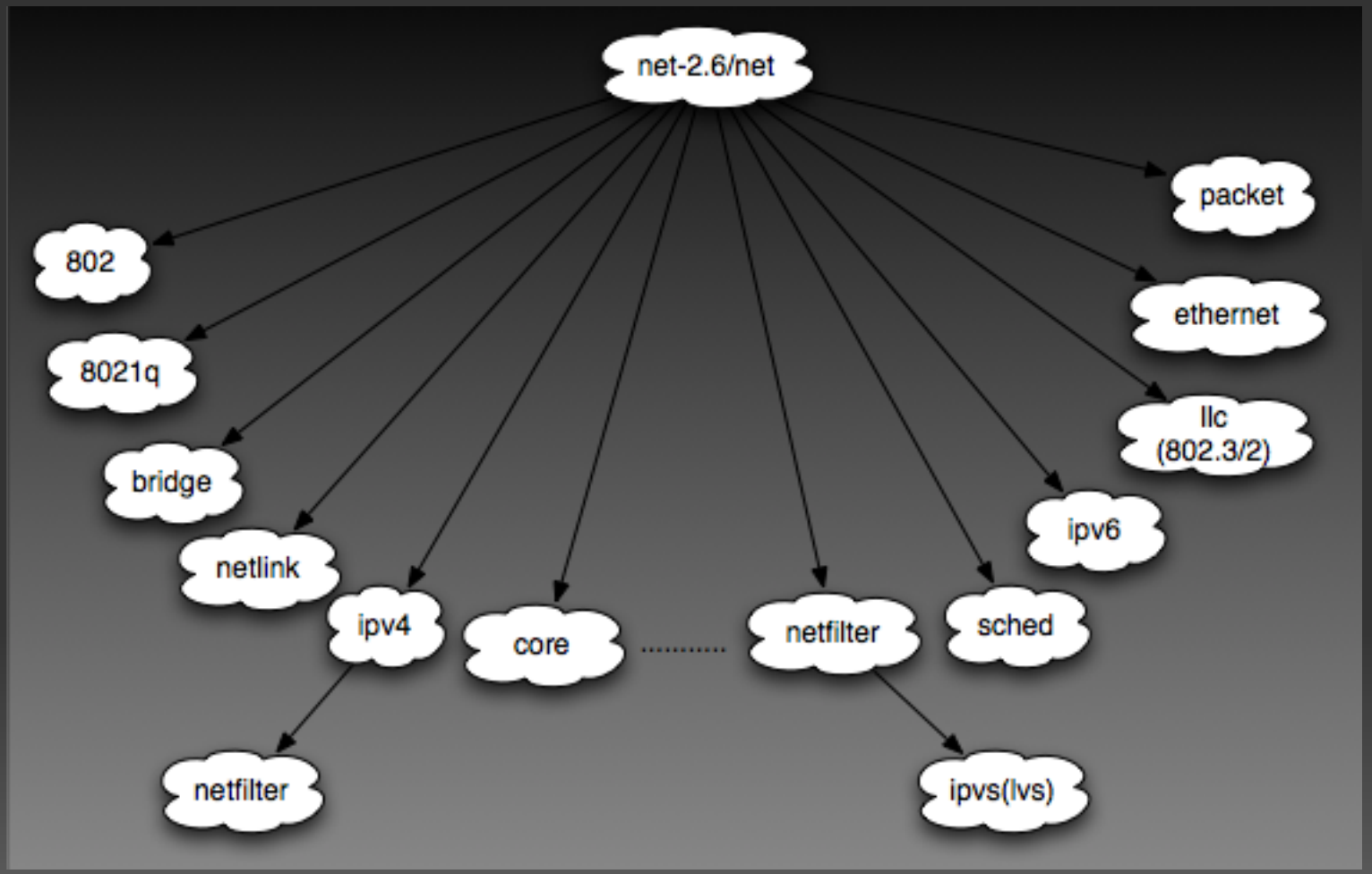
Resource

Introduction

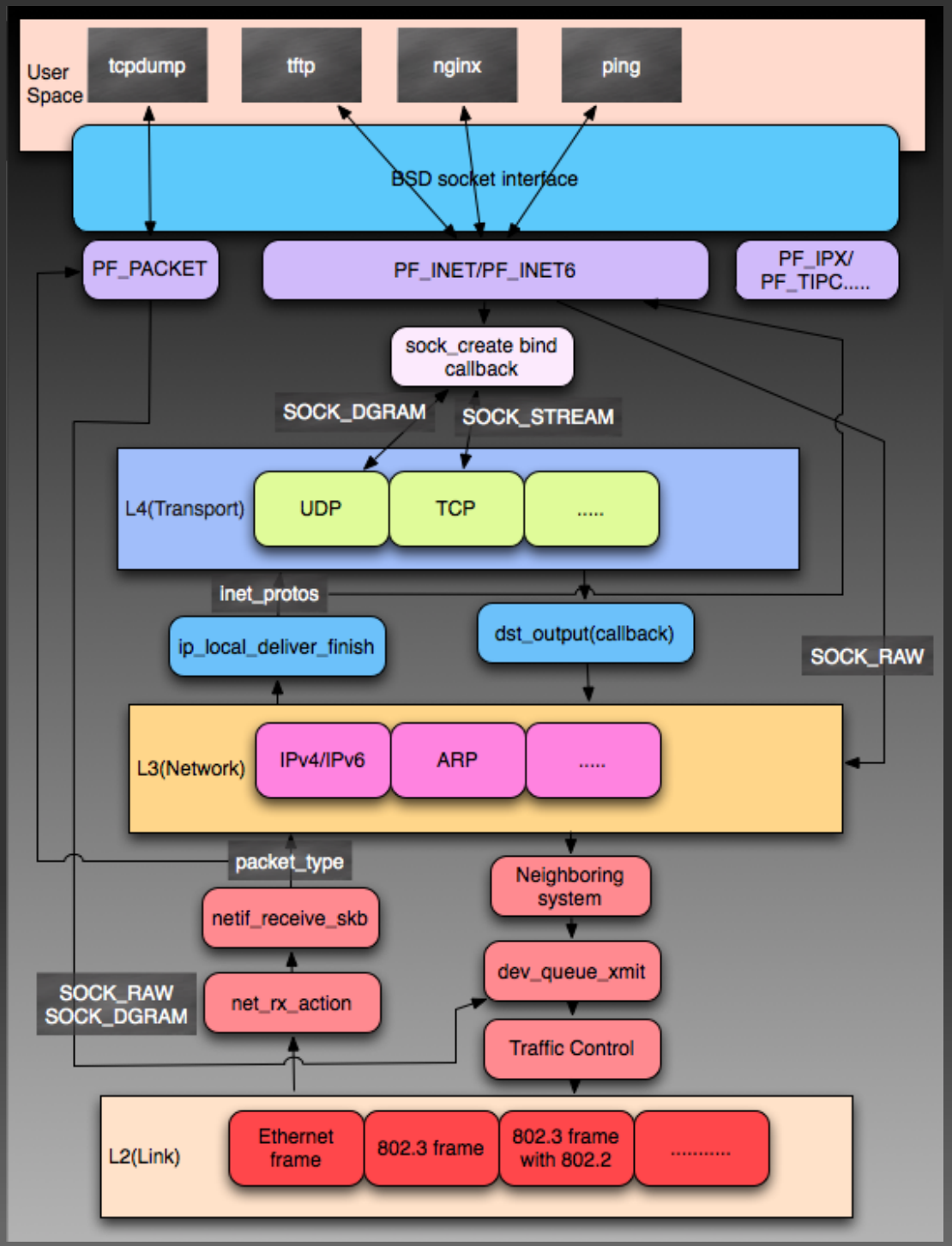
- Source
  - http://git.kernel.org/
  - net-next-2.6 and net-2.6
- Developer
  - Alan Cox, David Miller, Eric Dumazet, Patrick Mchardy etc.
- Traffic directions
  - input , forward and output
- Layer
  - L2(Link Layer)/L3(Network Layer)/L4(Transport Layer)
- Device interface
  - PCI/PCI-E

Networking code in the Linux kernel tree

Net-Kernel source tree



Big picture


User Space
tcpdump
tftp
nginx
ping
BSD socket interface
PF_PACKET
PF_INET/PF_INET6
PF_IPX/ PF_TIPC.....
sock_create bind callback
SOCK_DGRAM
SOCK_STREAM
L4(Transport)
UDP
TCP
.....
inet_protos
ip_local_deliver_finish
dst_output(callback)
SOCK_RAW
L3(Network)
IPv4/IPv6
ARP
.....
packet_type
Neighboring system
netif_receive_skb
dev_queue_xmit
SOCK_RAW SOCK_DGRAM
net_rx_action
Traffic Control
L2(Link)
Ethernet frame
802.3 frame
802.3 frame with 802.2
.....

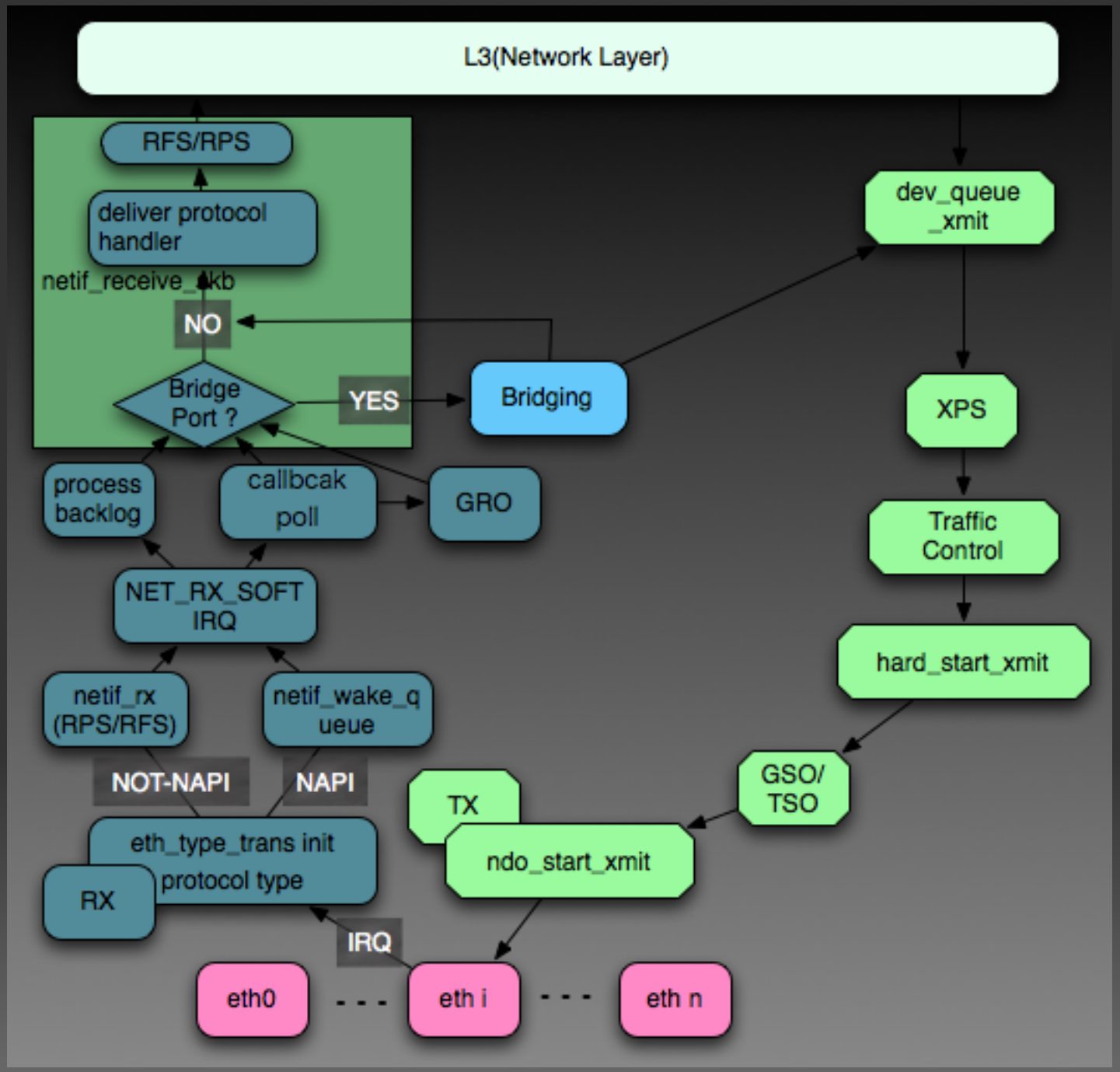
Link layer

- Frame type
  - 802.3/802.2/802.2-SNAP/Ethernet
- Input
  - Driver
    - NAPI
      - Poll + Interrupt
  - Soft interrupt
    - GRO
      - feed packet to network stack
    - RPS/RFS
      - make steer in SMP
  - Protocol handler
    - use eth_type_trans
    - Packet_type list

Link layer

- Output
  - Traffic Control
  - Soft interrupt
    - Transmit SKB
      - Scatter/Gather DMA
    - Free skb
    - XPS
      - multiqueue
      - avoid cache line bouncing
      - improve locality
- Bridge
  - Virtual device, must bind one or more real device
  - Spanning Tree Protocol

Link Layer bigmap



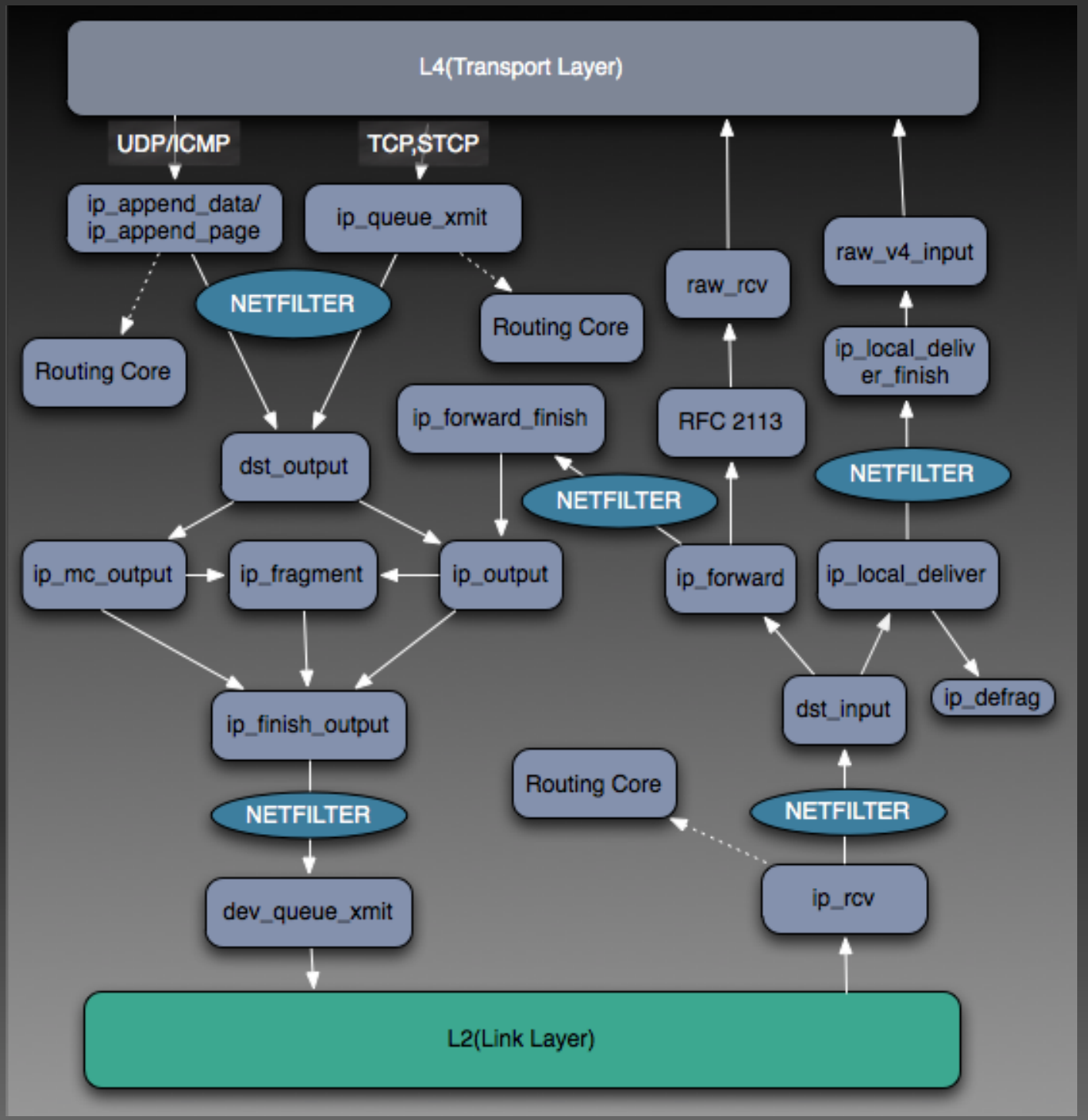
Network Layer(IP)

- Input
  - Protocol handler
    - net_protocol array
  - defragment
    - Hashtable
      - Each IP packet being defragmented save in a list
    - stored in kernel memory until they are totally processed
- Output
  - fragment
    - MTU
    - Scatter/Gather IO
    - udp
  - neighboring

Network Layer(IP)

- Forward
  - process ip option
  - igonore defragmentation
    - Router Alert option
- Route
  - Forwarding Information Base(routing table)
  - cache
- Netfilter
  - HOOK point
    - NF_IP_LOCAL_OUT/ NF_IP_LOCAL_IN etc..
- Management
  - Long-living IP peer information
    - AVL tree
  - IP statistics
    - per cpu data ipstats_mib
    - /proc/net/snmp

Network Layer Bigmap



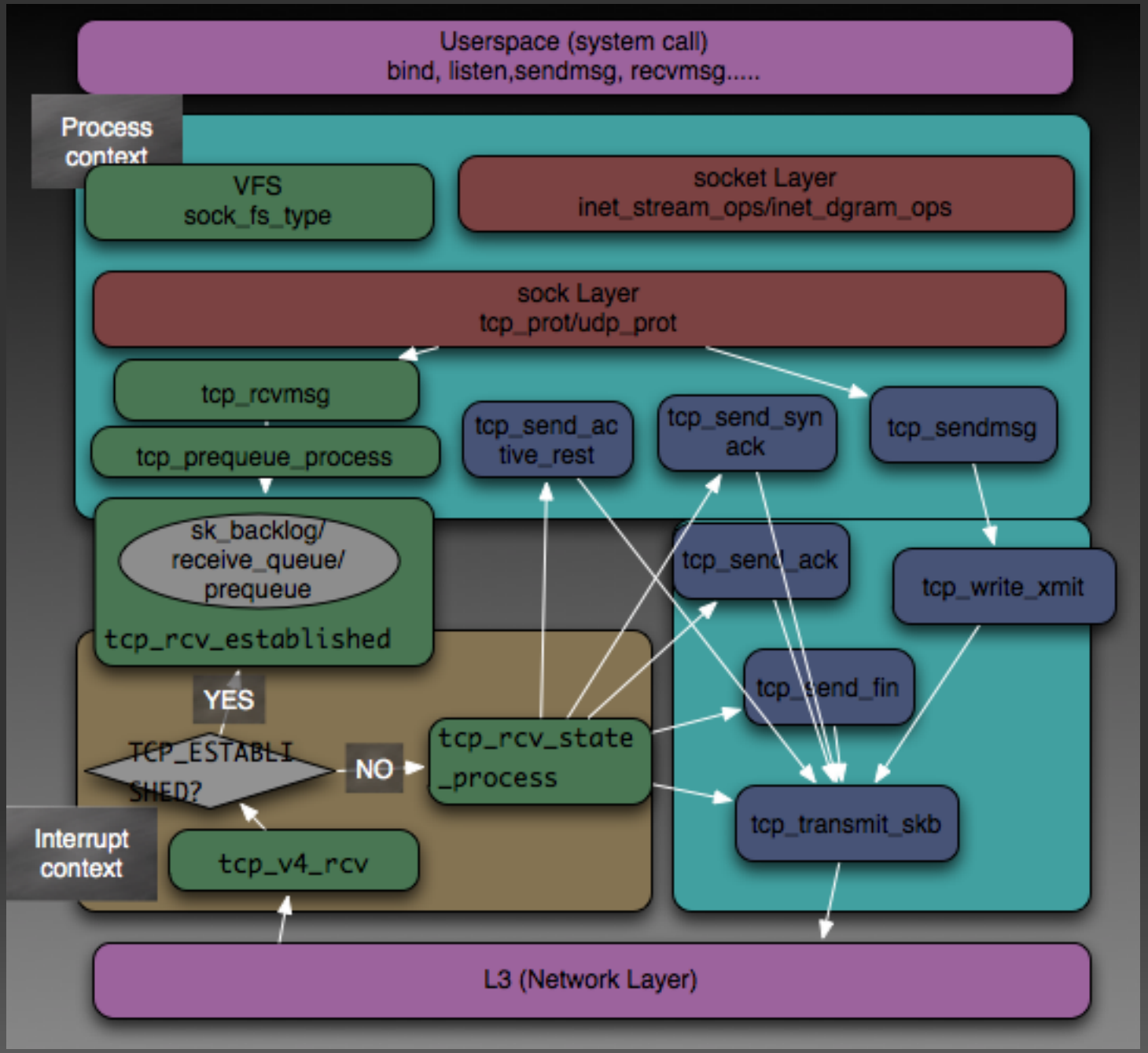
Transport Layer (tcp)

- Init
  - bind callback (sock_create)
  - Three handshrek
    - accept queue
    - syn table
    - create new socket fd and change state
- Manage socket
  - inet_ehash_bucket
    - TCP_ESTABLISHED <= sk->sk_state < TCP_CLOSE
  - inet_bind_hashbucket
    - local binding port info
  - listening_hash
    - socket in TCP_LISTEN state

Transport Layer (tcp)

- Output
  - Tcp push
  - Congestion control
    - state transition
    - congestion windows
    - packet count
- Input
  - fast path and slow path
  - Interrupt context/ Process context
  - sk_backlog/receive_queue/prequeue
- Tcp state transition
  - Kernel control
- Timer
  - Retransmit/keep-alive/time-wait etc

TCP Bigmap


Userspace (system call)
bind, listen,sendmsg, recvmsg.....
Process context
VFS
sock_fs_type
socket Layer
inet_stream_ops/inet_dgram_ops
sock Layer
tcp_prot/udp_prot
tcp_rcvmsg
tcp_prequeue_process
tcp_send_active_rest
tcp_send_synack
tcp_sendmsg
sk_backlog/
receive_queue/
prequeue
tcp_rcv_established
tcp_send_ack
tcp_write_xmit
tcp_send_fin
YES
TCP_ESTABLISHED?
NO
tcp_rcv_state_process
tcp_transmit_skb
Interrupt context
tcp_v4_rcv
L3 (Network Layer)

Config and Benchmark Tools

- Ethtool
  - offload fetures
- Benchmark and test tools
  - Netperf/pktgen
  - Mpstat/tcpstat

- Proc FileSystem
  - /proc/net
  - /proc/sys/net
    - ipv4
    - core
- Sys FileSystem
  - /sys/class/net/ethx

Resource

- http://kernelnewbies.org
-
- http://kernel.org
-
- http://www.kernelplanet.org
-
- https://lkml.org
-
- http://vger.kernel.org/vger-lists.html
-
- http://www.pagefault.info/?tag=kernel
-