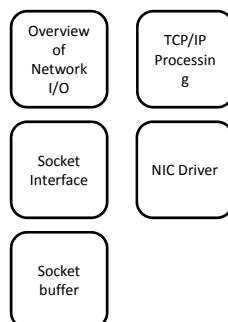# Operating System Design and Implementation
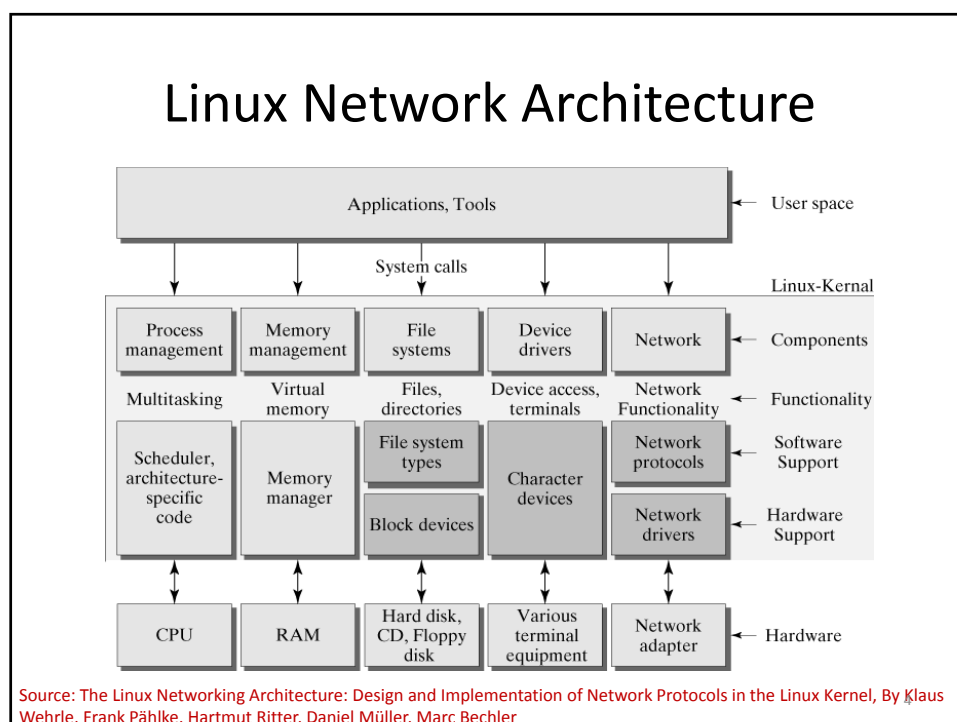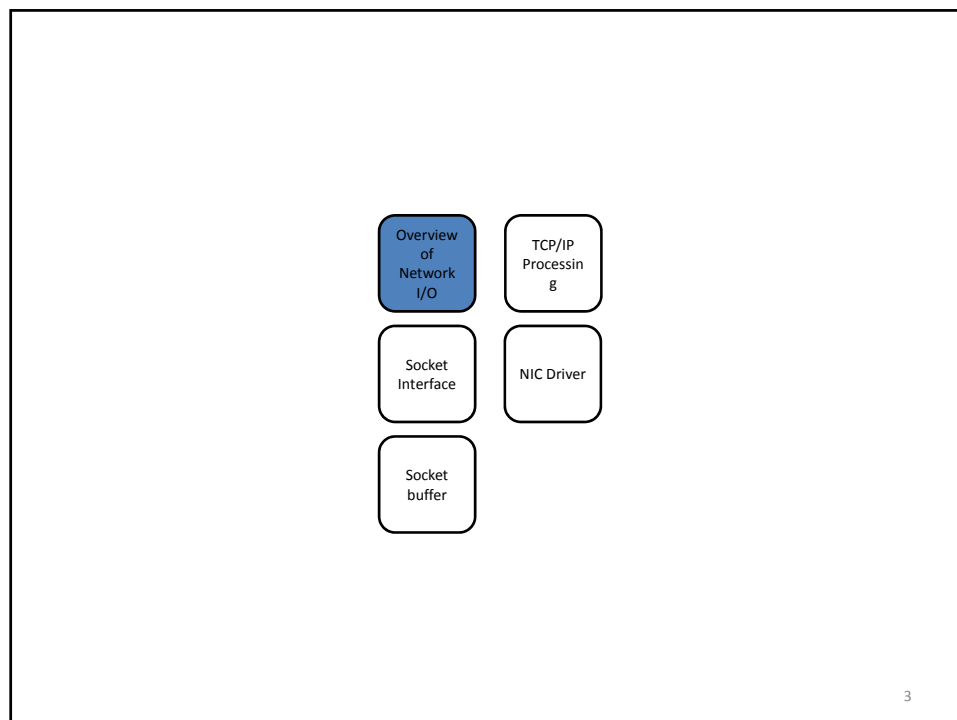## *Network Protocol and NIC Driver*

Shiao-Li Tsao
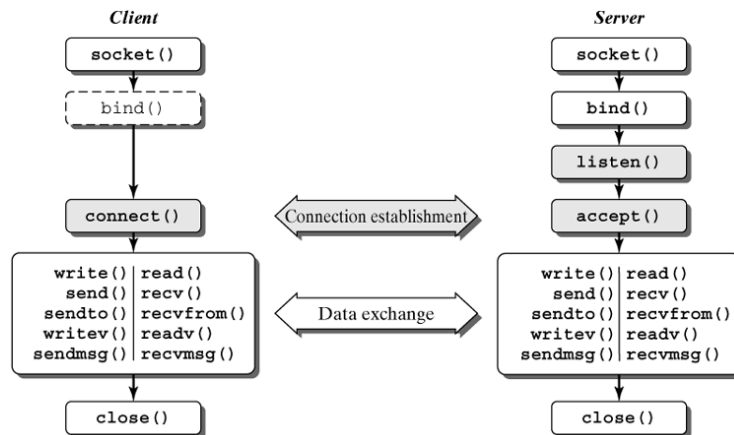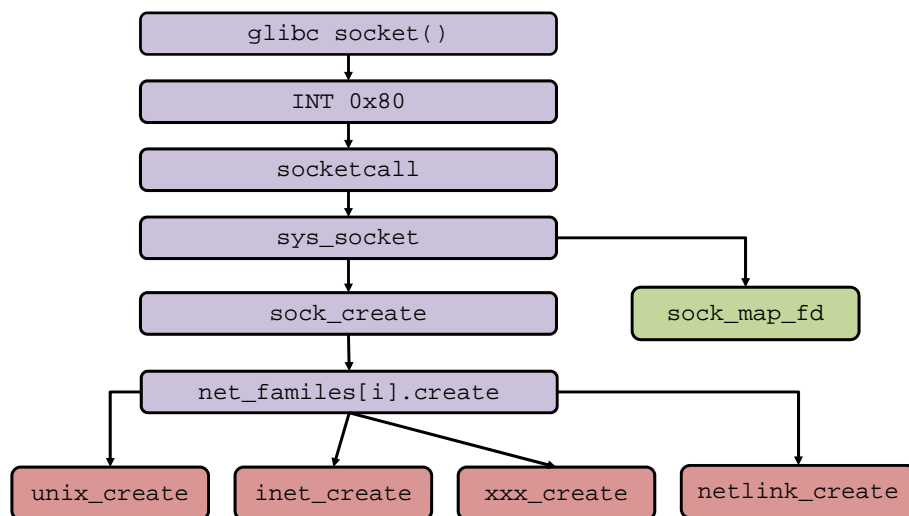
1

Overview of Network I/O

TCP/IP Processing

Socket Interface

NIC Driver

Socket buffer

2

Overview of Network I/O

TCP/IP Processing

Socket Interface

NIC Driver

Socket buffer

3

Linux Network Architecture

| Applications, Tools | | | | | ← User space |

System calls

Linux-Kernal

| Process management | Memory management | File systems | Device drivers | Network | ← Components |

| Multitasking | Virtual memory | Files, directories | Device access, terminals | Network Functionality | ← Functionality |

| Scheduler, architecture-specific code | Memory manager | File system types | Character devices | Network protocols | ← Software Support |
| | | Block devices | | Network drivers | ← Hardware Support |

| CPU | RAM | Hard disk, CD, Floppy disk | Various terminal equipment | Network adapter | ← Hardware |

Source: The Linux Networking Architecture: Design and Implementation of Network Protocols in the Linux Kernel, By Klaus Wehrle, Frank Pählke, Hartmut Ritter, Daniel Müller, Marc Bechler
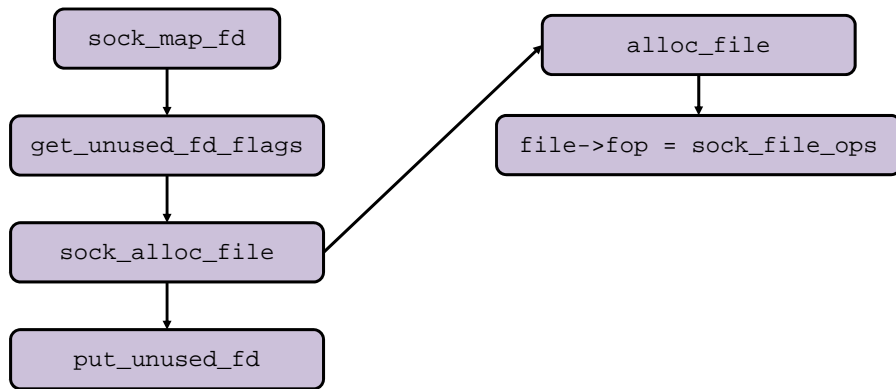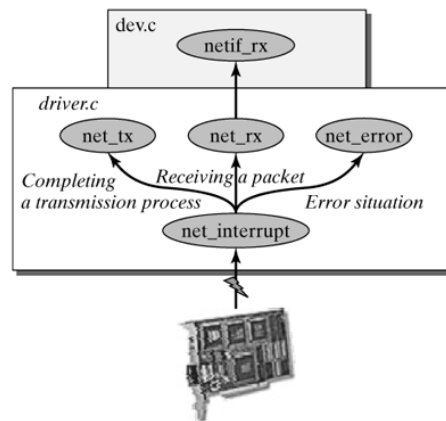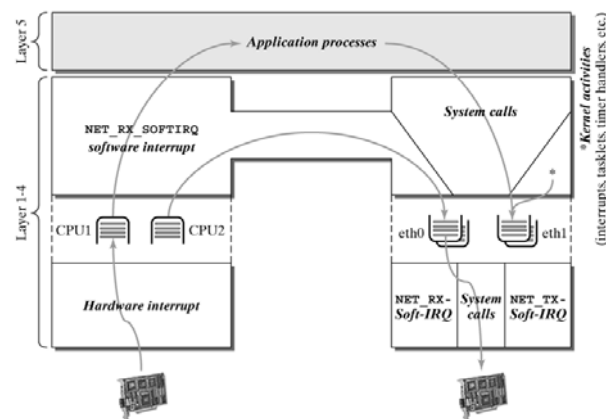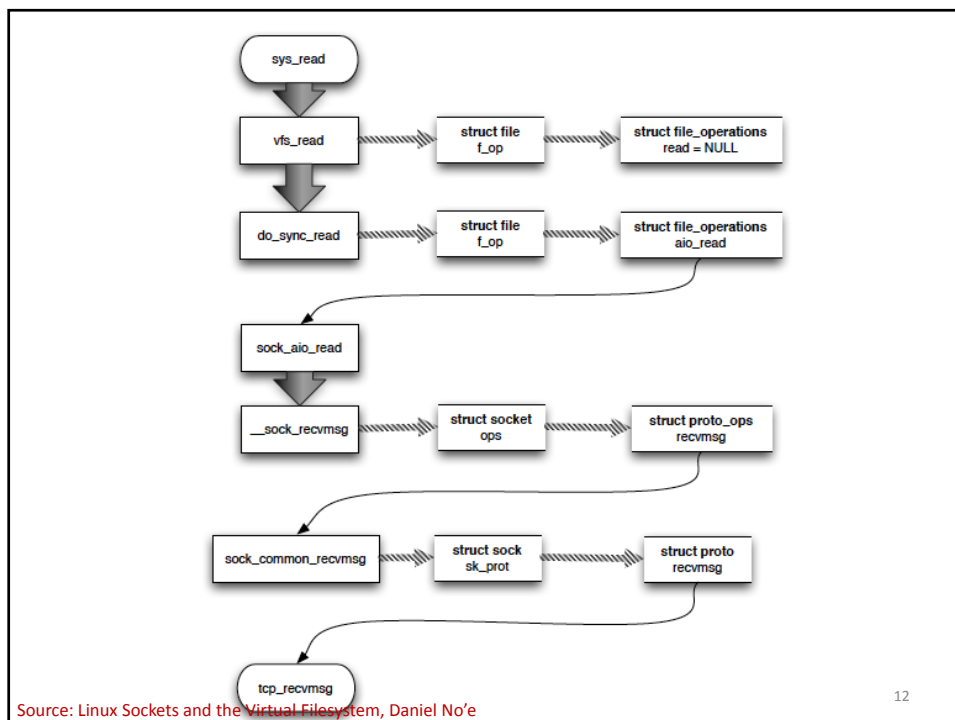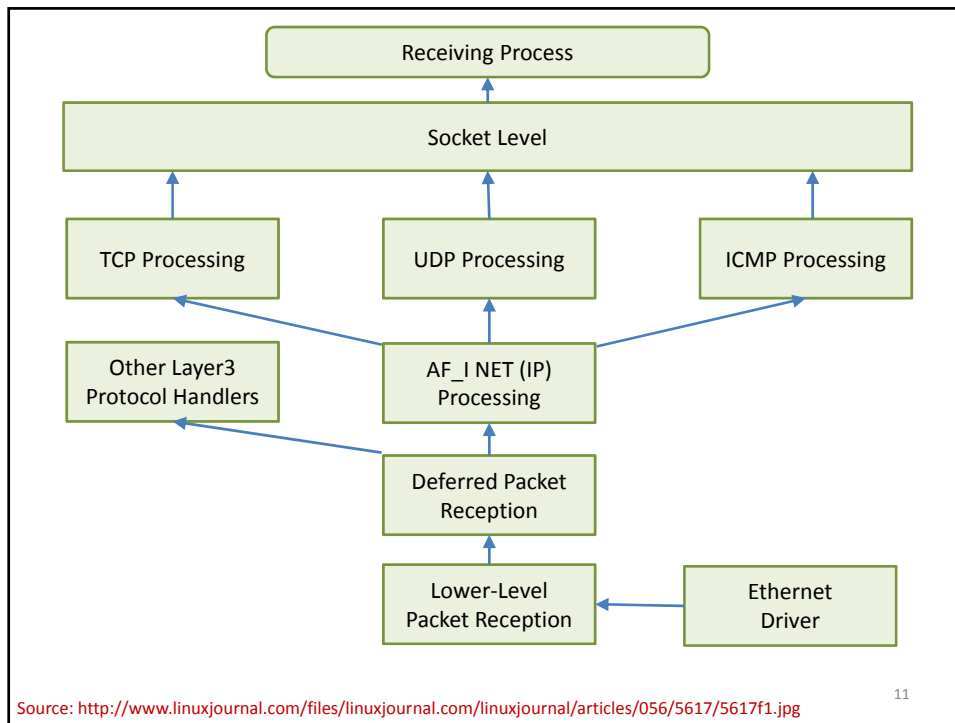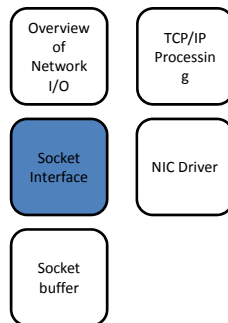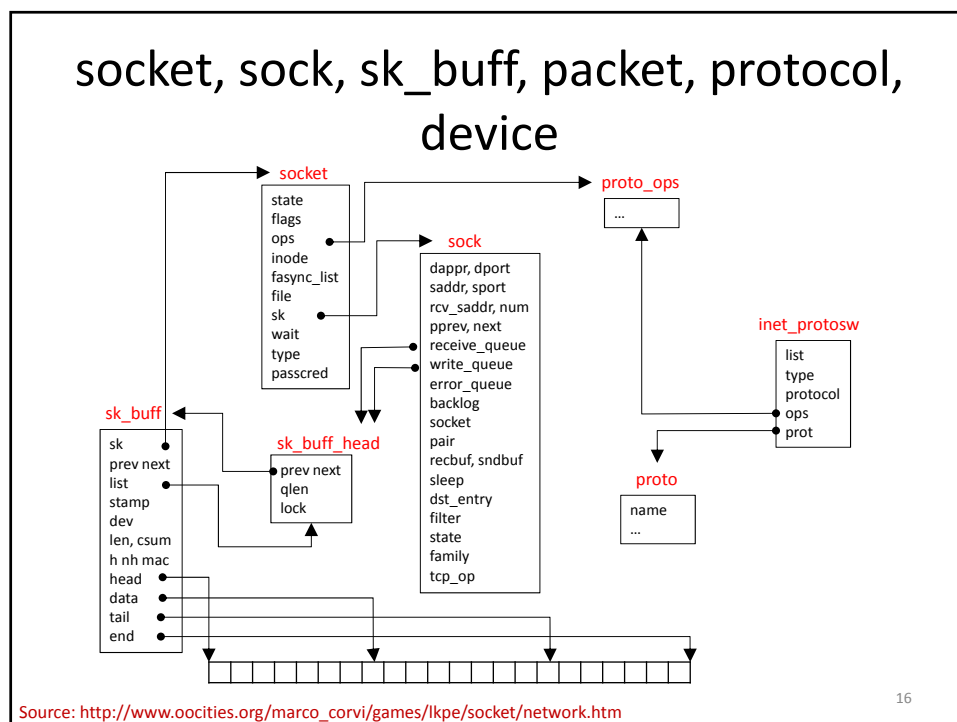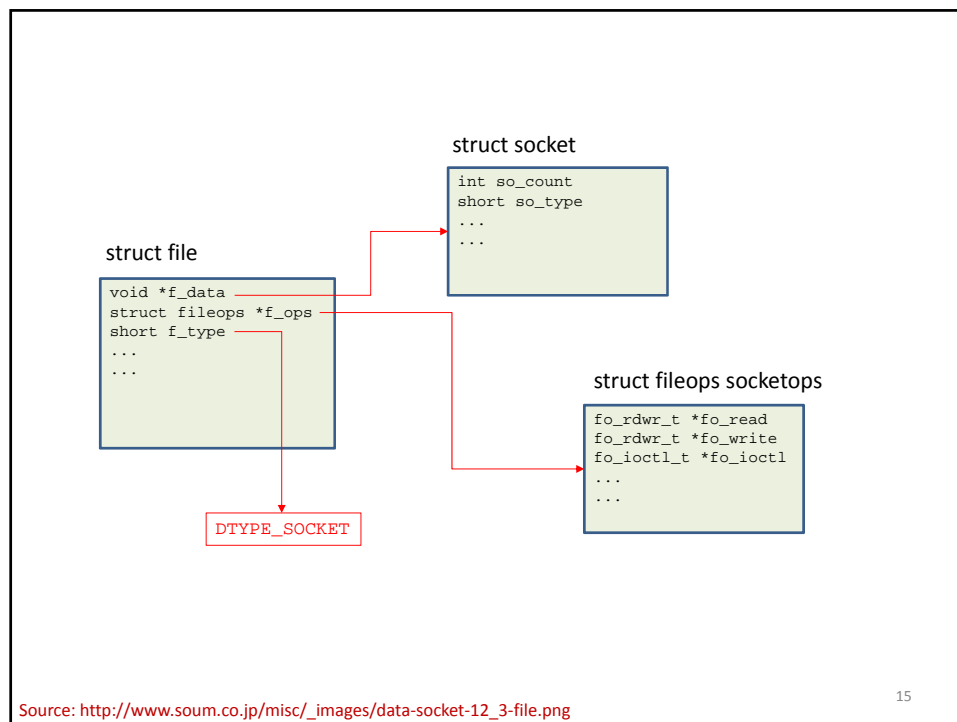
TCP/IP Program

socket()

socket() (Cont.)

```
sock_map_fd
```

```
get_unused_fd_flags
```

```
sock_alloc_file
```

```
put_unused_fd
```

```
alloc_file
```

```
file->fop = sock_file_ops
```

7

---

write(socket, data, length)	read(socket, data, length)
sys_write()	sys_read()
	sock_readv_writev()	
sock_sendmsg	sock_recvmsg
sock->ops-sendmsg	sock->ops-recvmsg
inet_sendmsg	sock_common_recvmsg
tcp_sendmsg ... tcp_transmit_skb	tcp_rcvmsg
ip_queue_xmit	ip_rcv
dev_queue_xmit	netif_rx
NIC TX	NIC ISR

8

Network interface and interrupt handler

- Interrupt (top half)

Network interface and interrupt handler

- Top halves and bottom halves

國立交通大學資訊工程學系 曹孝櫟老師

Receiving Process → Socket Level → TCP Processing / UDP Processing / ICMP Processing → Other Layer3 Protocol Handlers / AF_I NET (IP) Processing → Deferred Packet Reception → Lower-Level Packet Reception ← Ethernet Driver

11

sys_read → vfs_read — struct file f_op — struct file_operations read = NULL

do_sync_read — struct file f_op — struct file_operations aio_read

sock_aio_read

__sock_recvmsg — struct socket ops — struct proto_ops recvmsg

sock_common_recvmsg — struct sock sk_prot — struct proto recvmsg

tcp_recvmsg

12

國立交通大學資訊工程學系 曹孝櫟老師

Overview
of
Network
I/O

TCP/IP
Processin
g

Socket
Interface

NIC Driver

Socket
buffer

13

Socket Interface

- Linux kernel has only one socket related system call
  - sys_socketcall(int call, unsigned long *args)
  - Kernel distribute the function based on "call" parameter

14

struct socket

```
int so_count
short so_type
...
...
```
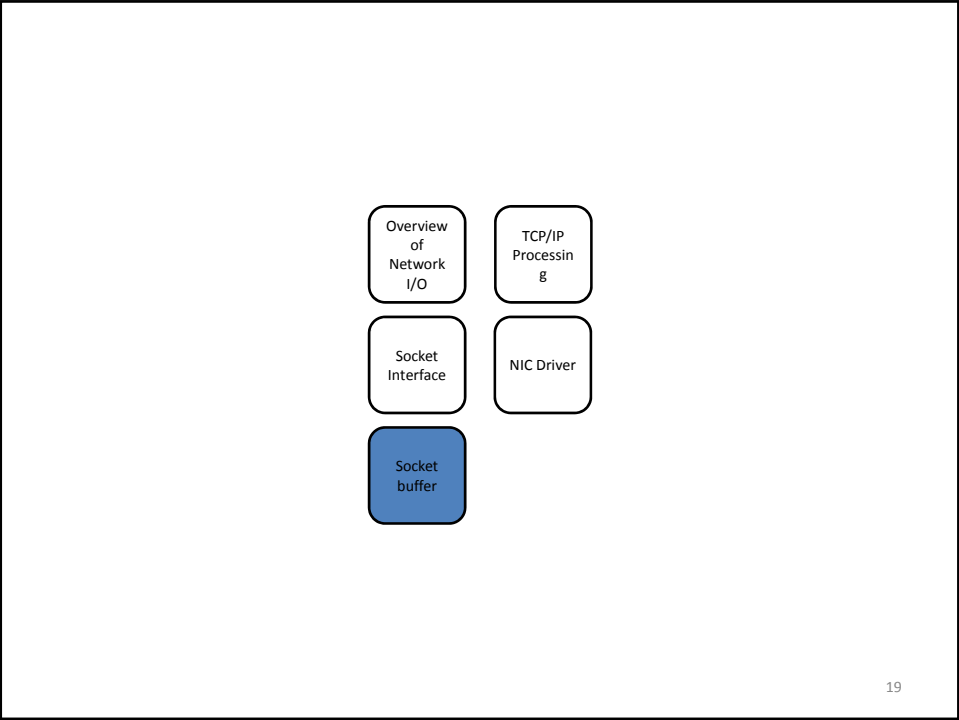
struct file

```
void *f_data
struct fileops *f_ops
short f_type
...
...
```

struct fileops socketops

```
fo_rdwr_t *fo_read
fo_rdwr_t *fo_write
fo_ioctl_t *fo_ioctl
...
...
```

`DTYPE_SOCKET`

socket, sock, sk_buff, packet, protocol, device

socket

```
state
flags
ops
inode
fasync_list
file
sk
wait
type
passcred
```

proto_ops

```
...
```

sock

```
dappr, dport
saddr, sport
rcv_saddr, num
pprev, next
receive_queue
write_queue
error_queue
backlog
socket
pair
recbuf, sndbuf
sleep
dst_entry
filter
state
family
tcp_op
```

inet_protosw

```
list
type
protocol
ops
prot
```

sk_buff

```
sk
prev next
list
stamp
dev
len, csum
h nh mac
head
data
tail
end
```

sk_buff_head

```
prev next
qlen
lock
```

proto

```
name
...
```

UDP Example

struct socket   struct sock

| ops | | prot |
| |
| sk |

udp_prot   inet_dgram_ops

"UDP"   PF_INET,
udp_close   inet_release
udp_connect   inet_bind
udp_disconnect   …
…

17

Kernel initialization

- Figure 5-1 in "Understanding Linux Network Internals"

18

Overview
of
Network
I/O

TCP/IP
Processing

Socket
Interface

NIC Driver

Socket
buffer

Network Protocol Layers

Source: The Linux Networking Architecture: Design and Implementation of Network Protocols in the Linux Kernel, By Klaus Wehrle, Frank Pählke, Hartmut Ritter, Daniel Müller, Marc Bechler
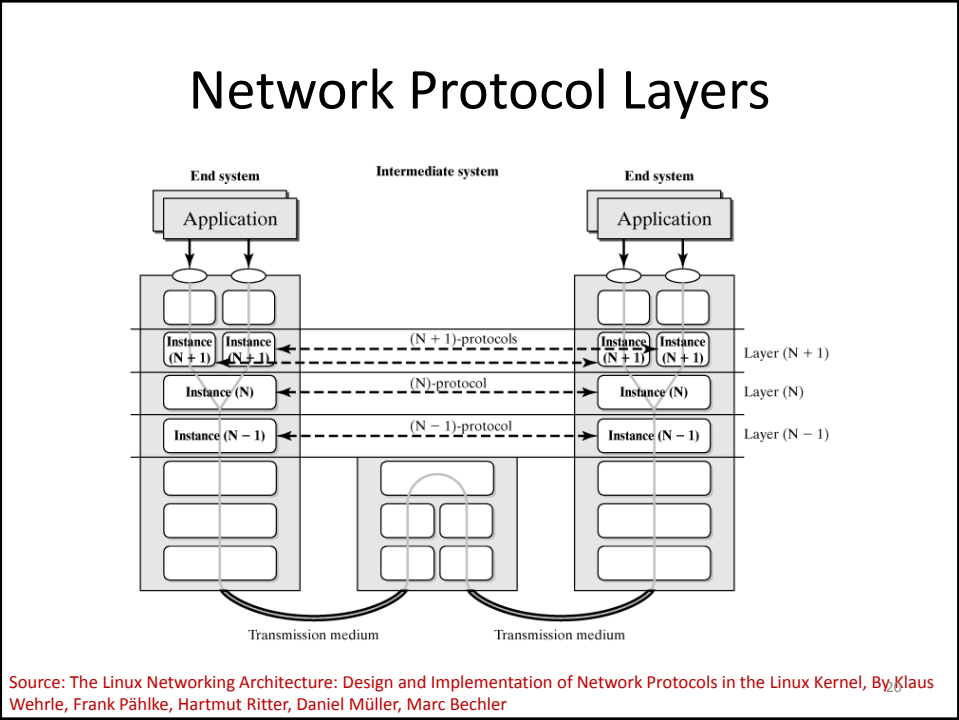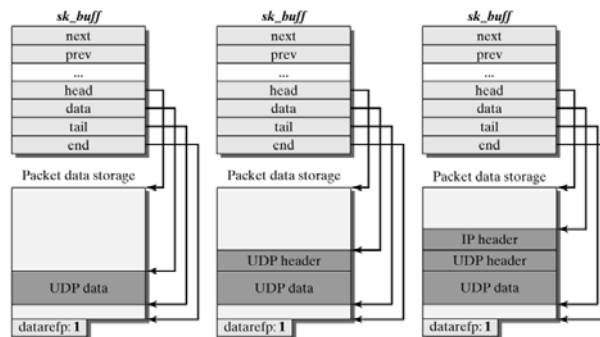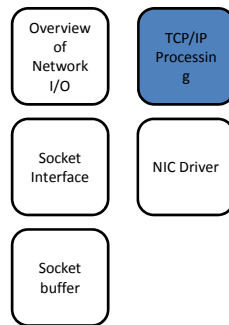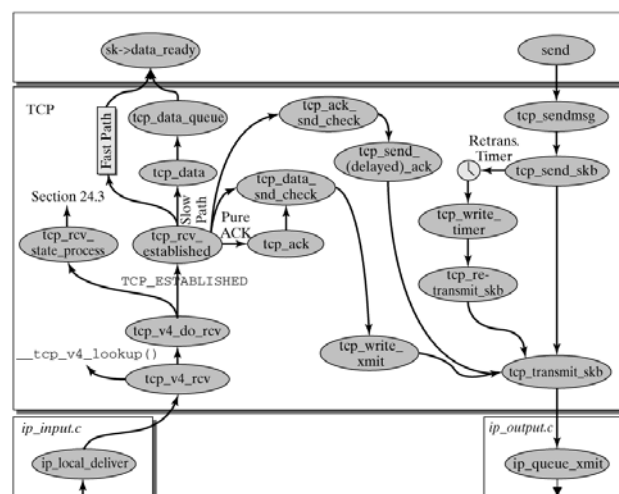
Internet Protocol Layers

SDU vs, PDU

Ethernet Framing without Segmentation or Checksum Offloading
Linux Ethernet Tx / Rx Framing No Segmentation or Checksum Offloading for TCP/IP.

Socket Buffer

Copy Packet

Copy sk_buff (skb_clone)

Packet Queue

Buffer management

- Changes to the packet buffers across the protocol hierarchy

29

TCP Protocol

Source: The Linux Networking Architecture: Design and Implementation of Network Protocols in the Linux Kernel, By Klaus Wehrle, Frank Pählke, Hartmut Ritter, Daniel Müller, Marc Bechler
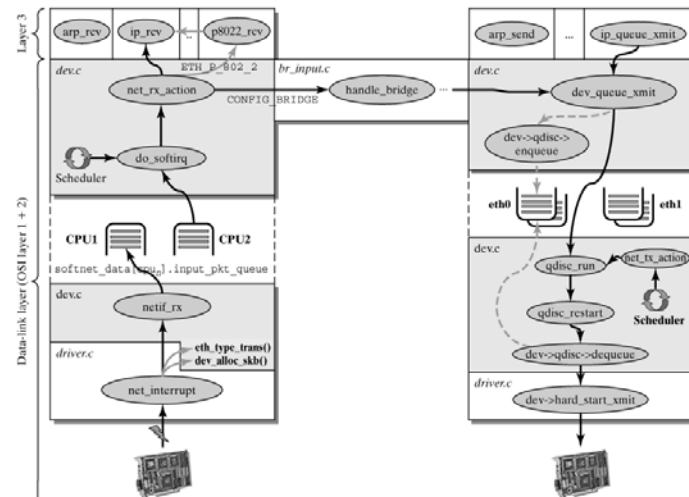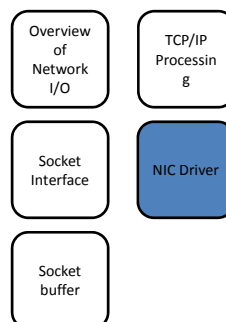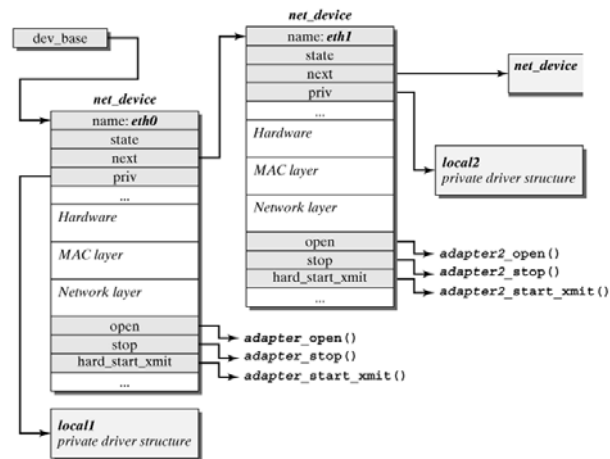
IP Protocol
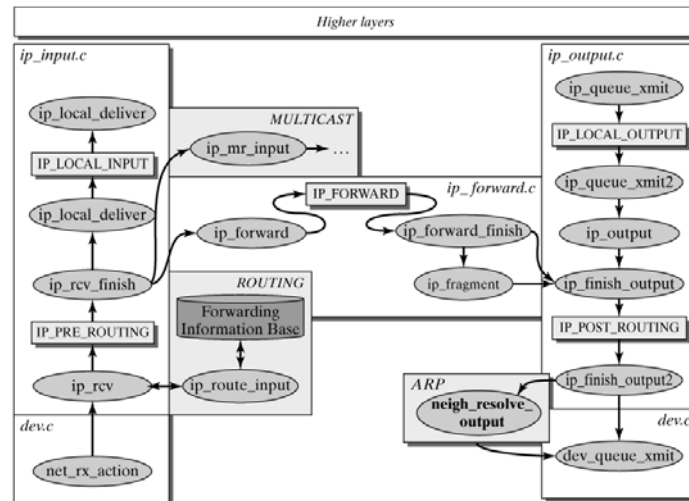
inet_protos
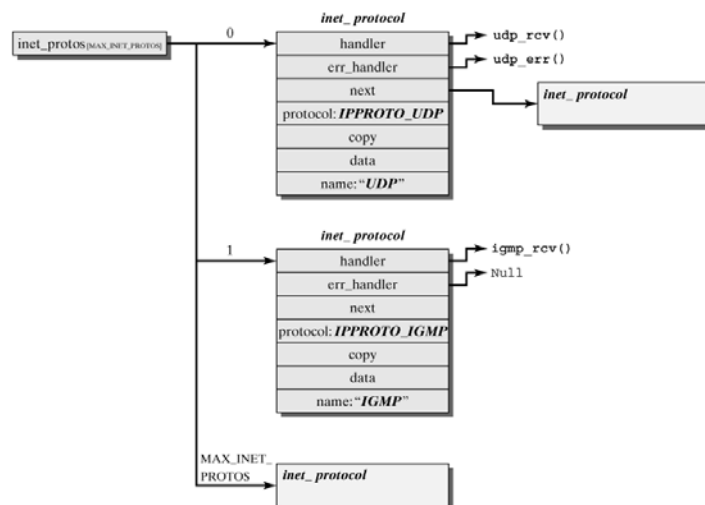
NIC driver and protocol stack

Source: The Linux Networking Architecture: Design and Implementation of Network Protocols in the Linux Kernel, By Klaus Wehrle, Frank Pählke, Hartmut Ritter, Daniel Müller, Marc Bechler
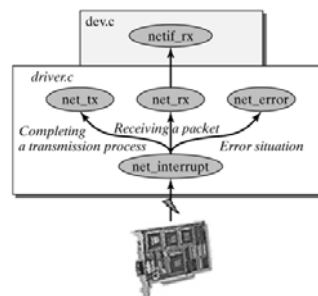

34

NIC driver and protocol stack

Network interface and interrupt handler

- Interrupt (top half)

Network interface and interrupt handler

- Top halves and bottom halves

NIC driver and protocol stack
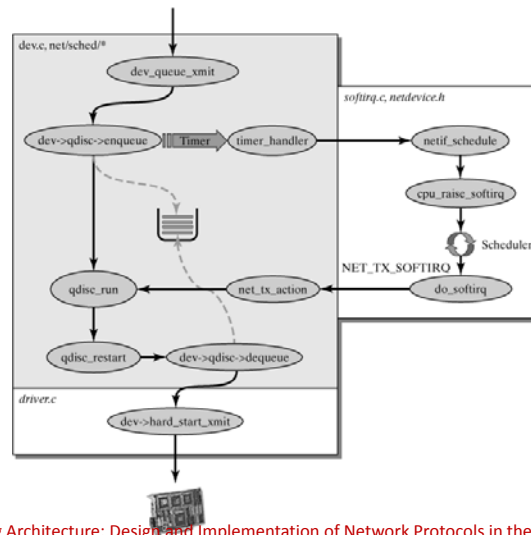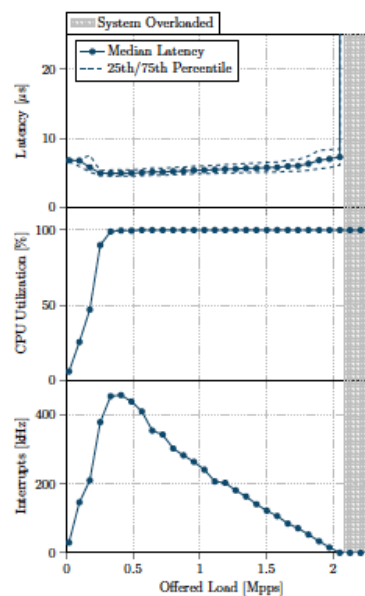
NIC driver and protocol stack

- TX

Source: The Linux Networking Architecture: Design and Implementation of Network Protocols in the Linux Kernel, By Klaus Wehrle, Frank Pählke, Hartmut Ritter, Daniel Müller, Marc Bechler

NAPI

Source: Optimizing Latency and CPU Load in Packet Processing Systems by Paul Emmerich et al.

Source: "Understanding Linux Network Internals," By Christian Benvenuti

41

Source: "Understanding Linux Network Internals," By Christian Benvenuti

42