

ISOTDAQ 2016

WEIZMANN INSTITUTE FOR SCIENCE

NIKO NEUFELD, CERN EP

# PRACTICAL ASPECTS OF NETWORKS FOR DATA ACQUISITION



# THE DATA TORRENT

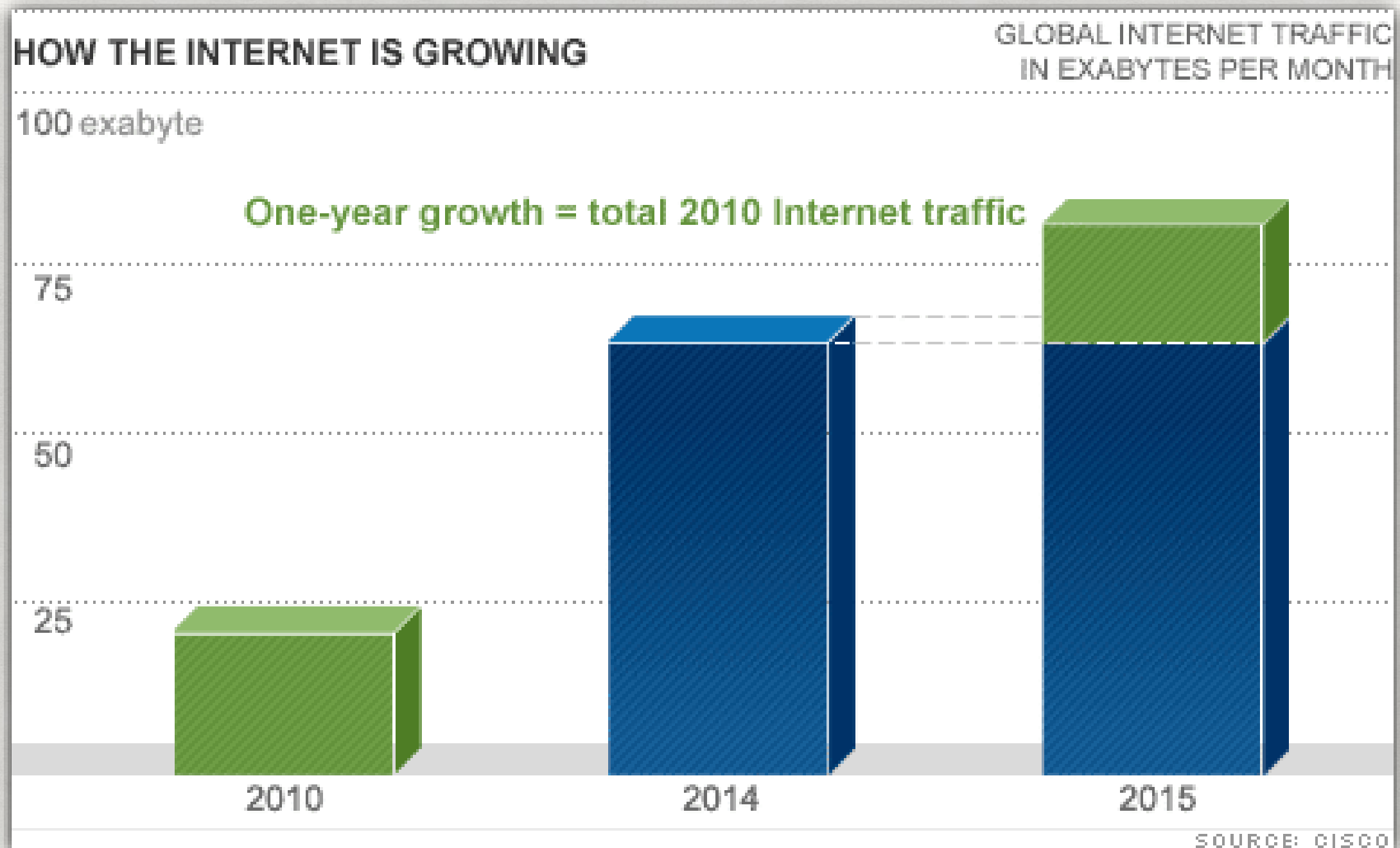
## EXPERIMENTS GENERATE “BIG” DATA

- LHC: (per experiment) about 100 GB/s(!) to the software trigger (HLT), about 1 GB/s to storage
- SKA: 68 Tbits/s (raw), 0.5 to 10 PB/day(!) of image data
- ESRF: > 8 GB/s for fast cameras, > 10 TB retained data (for a single beamline!)
- CTA: 4 GB/s to storage, 1.5 - 20 PB / year



# THE DATA TORRENT

## DATA GROWTH ON THE INTERNET





# THE INTERNET IS NO LONGER FOR ...

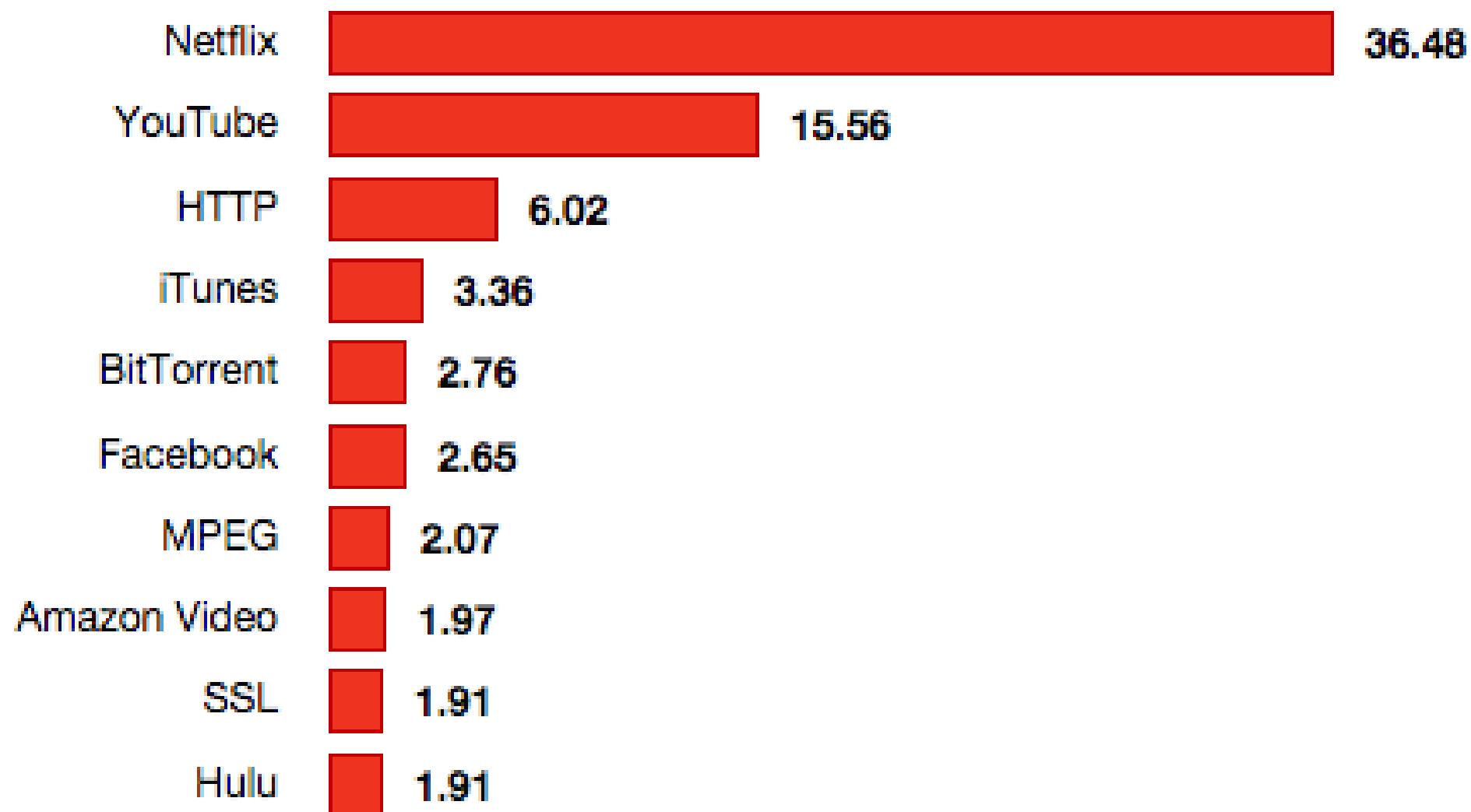


- It's for Video and mobile
- 15 bytes become 3.5 MB



# WHAT IS THE BANDWIDTH USED FOR?

## Top sites by percentage of downstream internet traffic in North America





HIGH DATA RATE  
REQUIRED?  
NETWORKS ARE THE WAY  
TO GO!



# NETWORKS

## GENERAL

- Networks connect multiple devices over (large) distances
- In a network devices are equal ("peers") unlike in bus systems (VME, USB, PCI), which have "masters" and "slaves"
- In a network devices communicate directly with each other
- Data and control use the same path (again unlike in many bus-systems)
- At the signaling level network technologies are normally (multiple) serial, many can use multiple physical layers (—> later) including optical and radio-transmitted



# NETWORKS

## TECHNOLOGIES

- The telephone network (now being replaced by IP)
- Ethernet (IEEE 802.3)
- ATM (the backbone for 2G cell-phones), RapidIO (3G and 4G)
- InfiniBand, OmniPath
- and many, many more
- Network technologies are sometimes functionally grouped
  - Cluster interconnect (OmniPath, InfiniBand) 1 m - 150 m (in a data-centre)
  - Local area network (LAN), up to ~ 10 km (in and between buildings)
  - Wide area network (ATM, SONET) > 50 km (between cities, countries...)



# WHAT MAKES A NETWORK?

- A network has two important aspects
  - A physical implementation: wires, optical fibres, devices, connectors, cables, radio-transmitters, etc...
  - A protocol defining how data are exchanged between nodes on the network



# NETWORKS

## PROTOCOL STACK

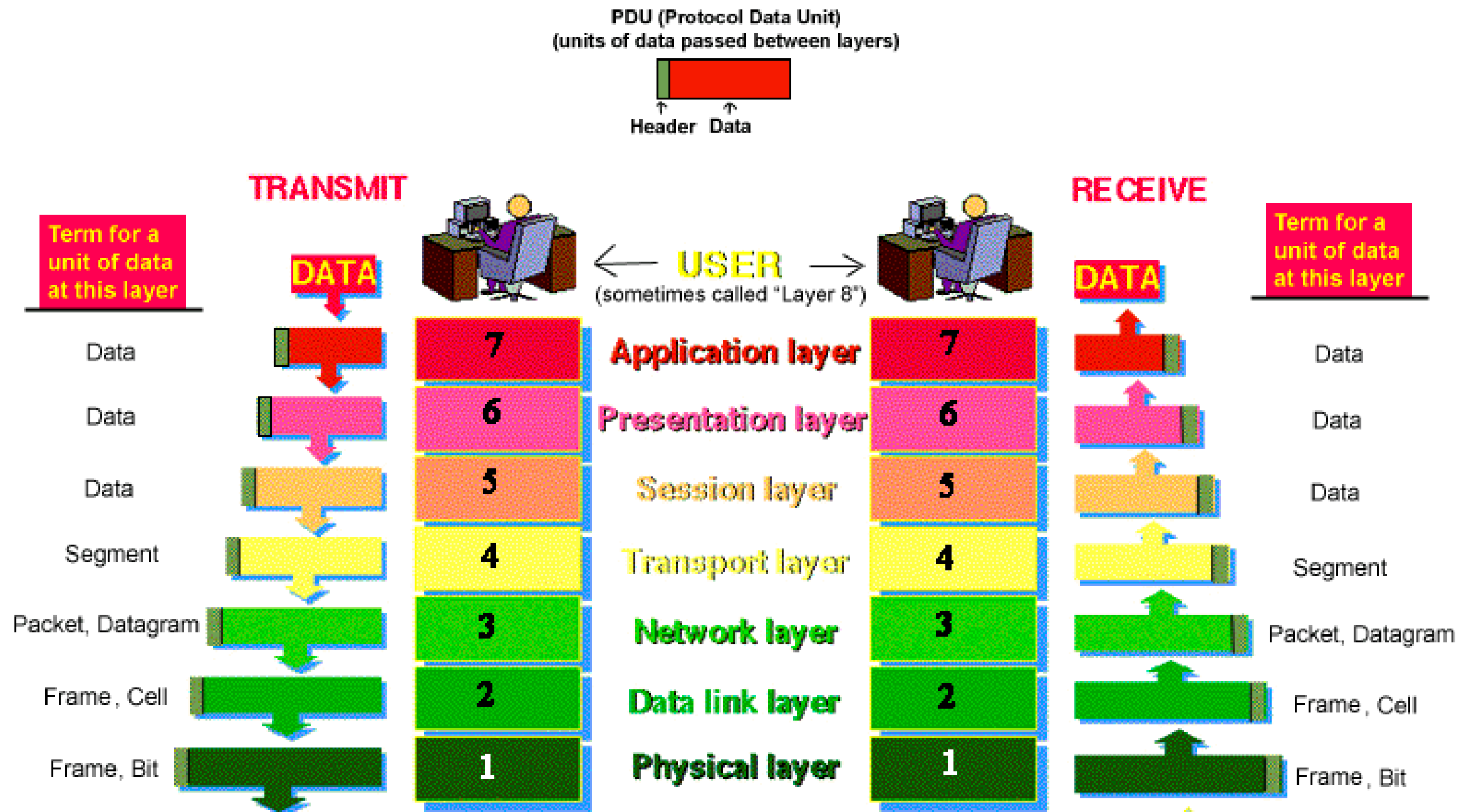
### OSI model

Layer	Name	Example protocols
7	Application Layer	HTTP, FTP, DNS, SNMP, Telnet
6	Presentation Layer	SSL, TLS
5	Session Layer	NetBIOS, PPTP
4	Transport Layer	TCP, UDP
3	Network Layer	IP, ARP, ICMP, IPSec
2	Data Link Layer	PPP, ATM, Ethernet
1	Physical Layer	Ethernet, USB, Bluetooth, IEEE802.11



# HOW TO CALL THINGS

## COMMON TERMINOLOGY HELPS

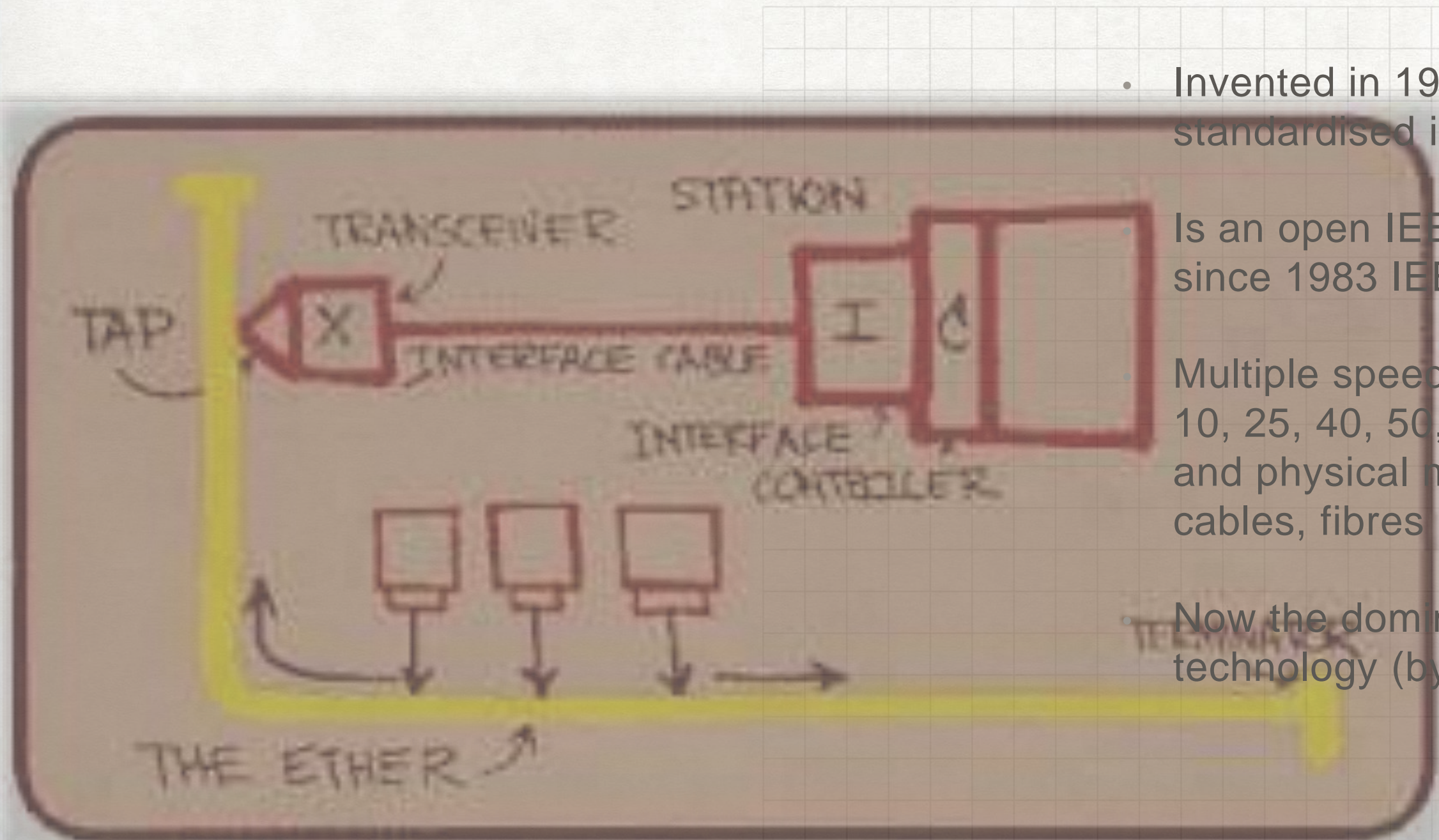


IN NETWORKING WE DO NOT TALK ABOUT BYTES BUT ABOUT OCTETS (8 BITS)



# ETHERNET

## ONE NETWORK TO RULE THEM ALL



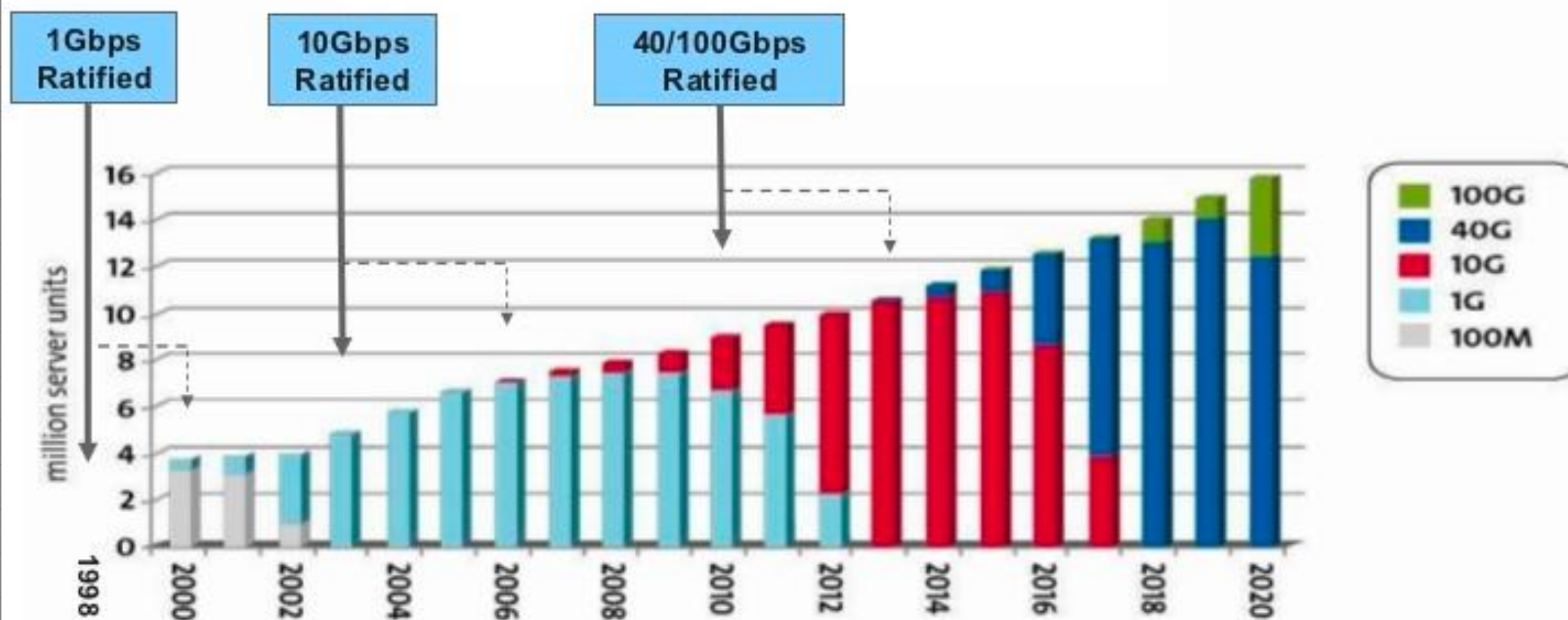
- Invented in 1973, standardised in 1980
- Is an open IEEE standard since 1983 IEEE 802.3
- Multiple speed grades (1, 10, 25, 40, 50, 100 Gb/s) and physical media: cables, fibres
- Now the dominant LAN technology (by far)



# THE RISE OF ETHERNET

## Historical and Predicted Port Delivery by Ethernet Speed

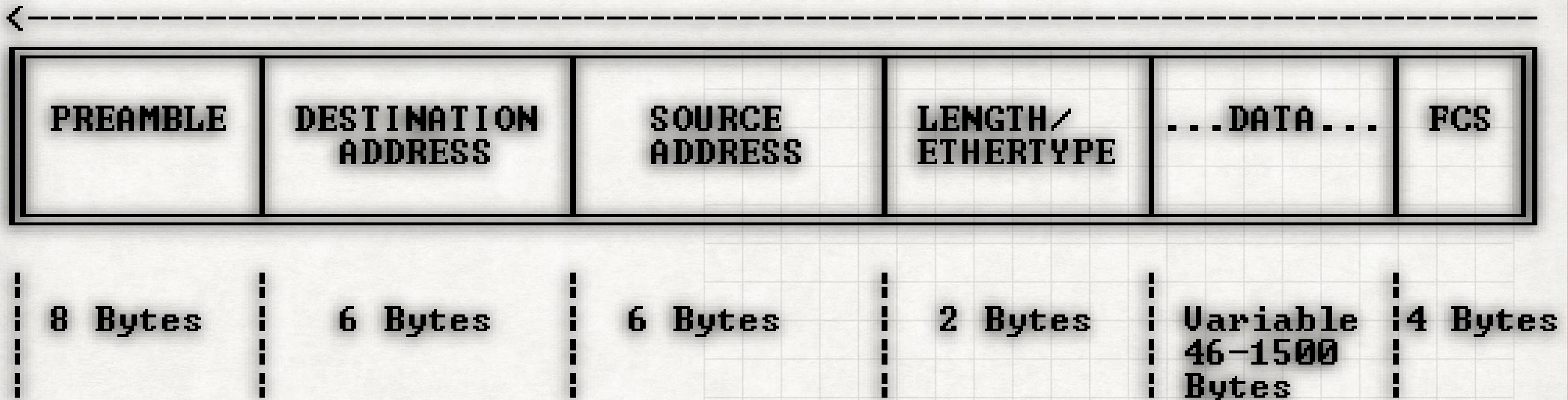
- Recent history suggests that standards ratification and infrastructure cabling lead actual port sales by ~3 years



Source: Intel/Broadcom (May 2007)



# ETHERNET



- 48 bit addresses, also called MAC addresses (Media Access Control)
- Payload: min 64 bytes, max 1518 bytes —> many but not all devices support “jumbo” frames up to 9000 bytes. The maximum number acceptable by a device is called the *Maximum Transmission Unit (MTU)*
- Source address usually “burnt” into a device (in practice read from an EEPROM or similar and set by software)
- You can override your source address (using `ethtool(8)` or `ip(8)` command)



# NETWORK PROTOCOLS

## IP

- Two major versions
  - IPv4 and IPv6 (focus on v4 for the rest of this)
- In IP devices carry a 32-bit address written as 4 decimals separated by dots
- IP groups several devices into a network / subnet using the net mask
  - Net mask
    - 8, 16, 24 bit wide net masks have special names (Class A, Class B and Class C)
- The network itself has an address (all unmasked bits 0) e.g. 137.158.0.0
- There is a broad-cast address (all unmasked bits 1) e.g. 137.158.255.255

137.158.128.0/**17**

(netmask **255.255.128.0**)

1111 1111	1111 1111	1 000 0000	0000 0000
1000 1001	1001 1110	1 000 0000	0000 0000

198.134.0.0/**16**

(netmask **255.255.0.0**)

1111 1111	1111 1111	0000 0000	0000 0000
1100 0110	1000 0110	0000 0000	0000 0000

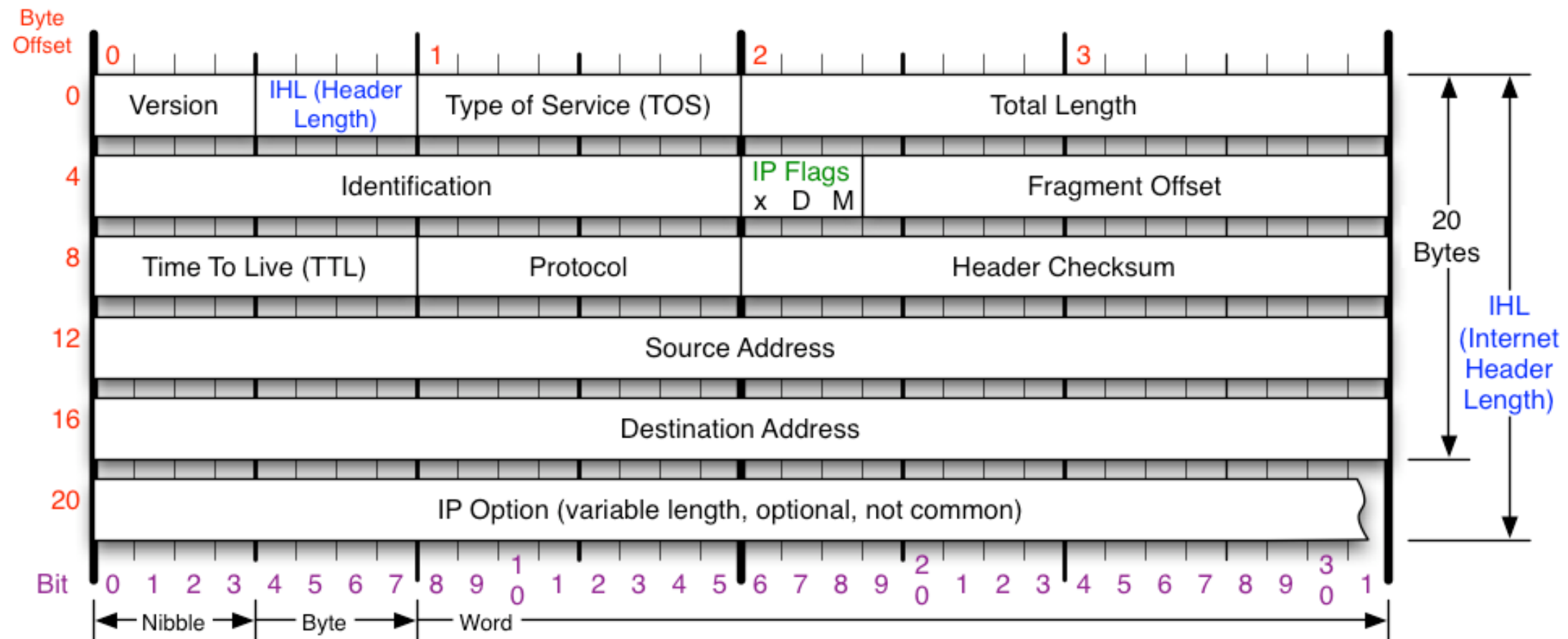
205.37.193.128/**26**

(netmask **255.255.255.192**)

1111 1111	1111 1111	1111 1111	11 00 0000
1100 1101	0010 0101	1100 0001	10 00 0000



# THE IPV4 HEADER



## Version

Version of IP Protocol. 4 and 6 are valid. This diagram represents version 4 structure only.

## Header Length

Number of 32-bit words in TCP header, minimum value of 5. Multiply by 4 to get byte count.

## Protocol

IP Protocol ID. Including (but not limited to):

1 ICMP	17 UDP	57 SKIP
2 IGMP	47 GRE	88 EIGRP
6 TCP	50 ESP	89 OSPF
9 IGRP	51 AH	115 L2TP

## Total Length

Total length of IP datagram, or IP fragment if fragmented. Measured in Bytes.

## Fragment Offset

Fragment offset from start of IP datagram. Measured in 8 byte (2 words, 64 bits) increments. If IP datagram is fragmented, fragment size (Total Length) must be a multiple of 8 bytes.

## Header Checksum

Checksum of entire IP header

## IP Flags

x D M

x 0x80 reserved (evil bit)  
D 0x40 Do Not Fragment  
M 0x20 More Fragments follow

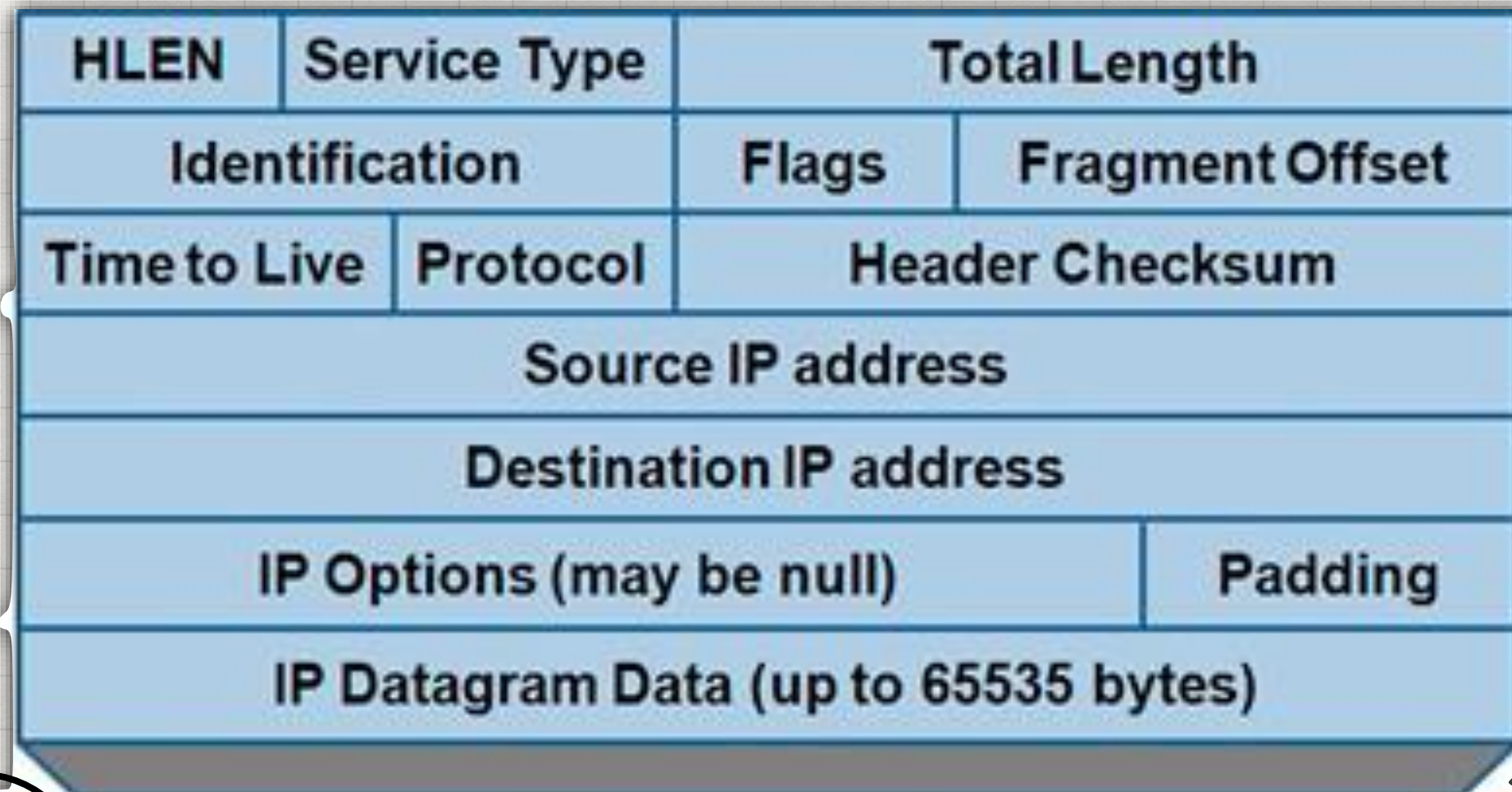
## RFC 791

Please refer to RFC 791 for the complete Internet Protocol (IP) Specification.



# NETWORK PROTOCOLS

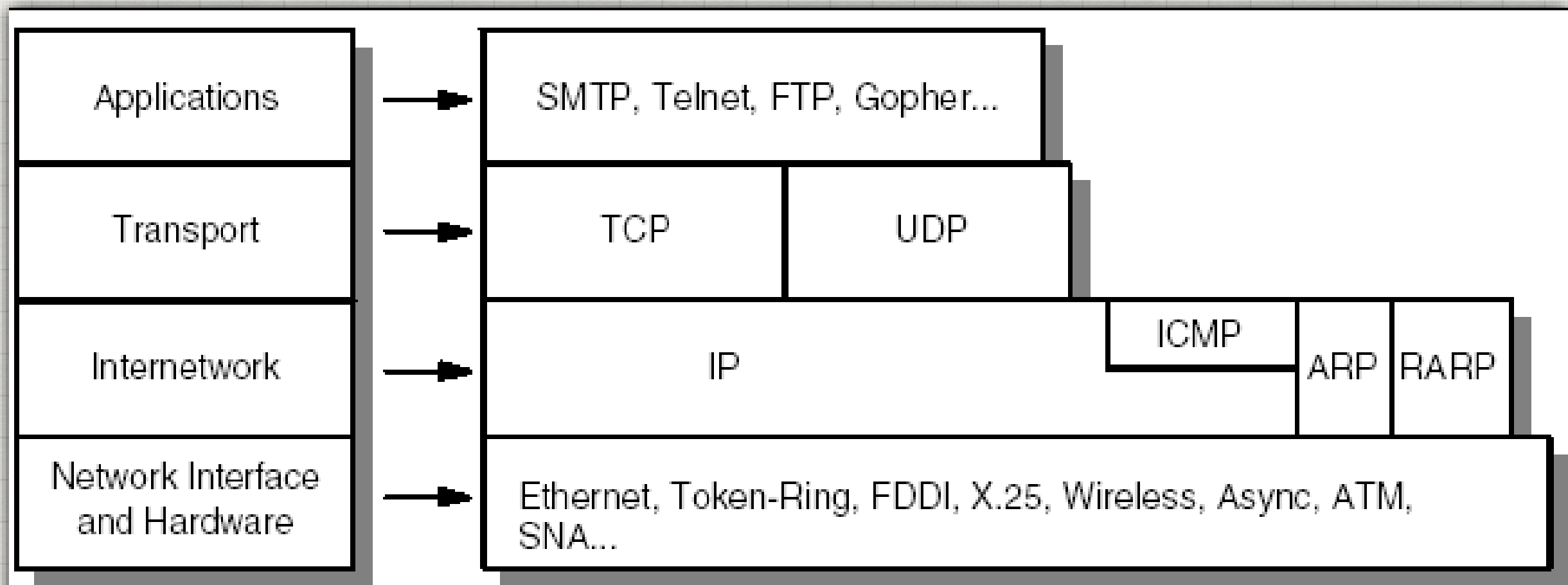
## IP OVER ETHERNET





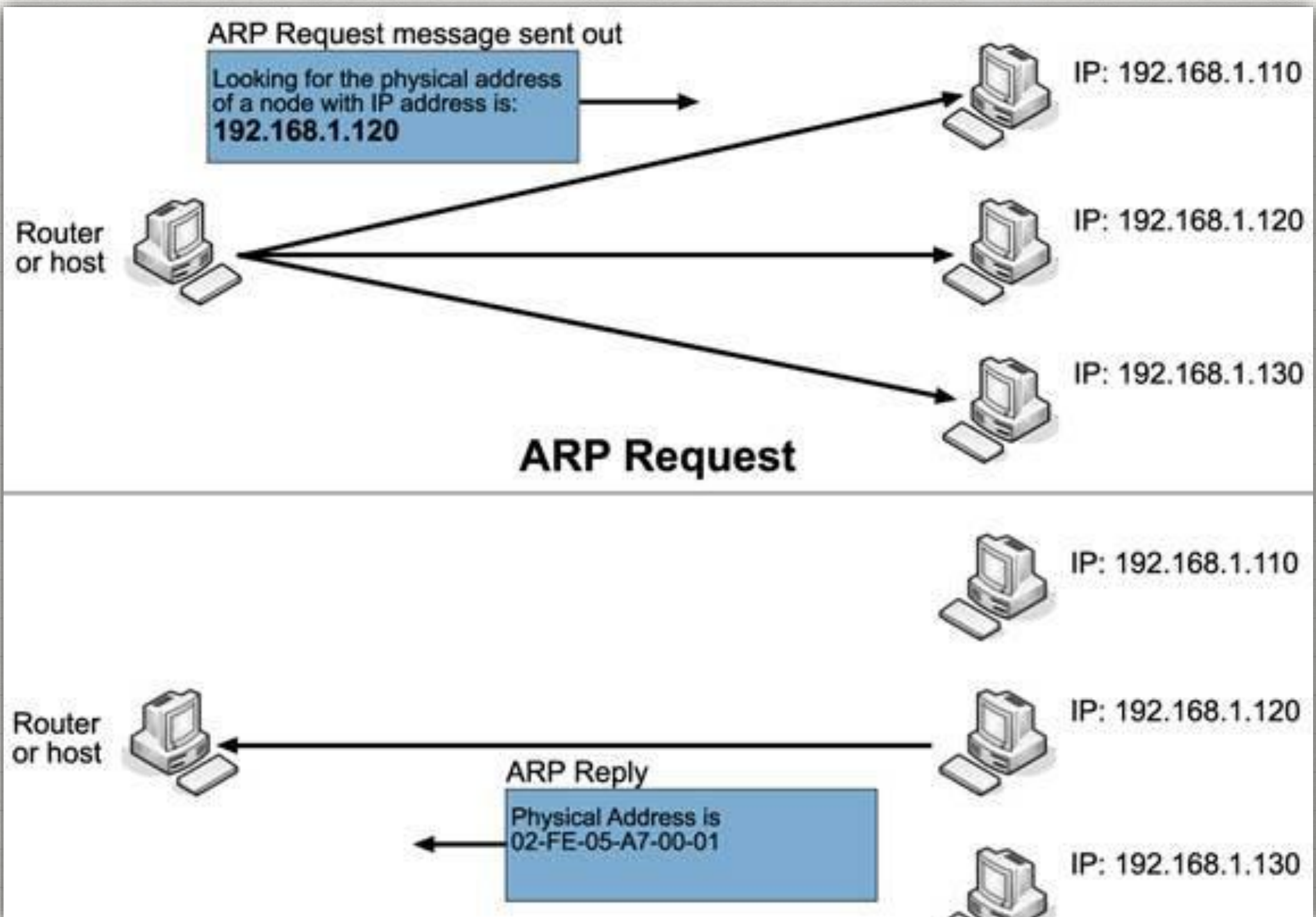
# THE INTERNET PROTOCOLS

## A SIMPLIFIED STACK





# RELATING IP AND ETHERNET: ARP





# ARP

- ARP requests are Ethernet broadcasts
- They are sent to and received by all members of a Virtual LAN (VLAN). A VLAN is an arbitrary selection of ethernet devices (hosts or switches). Broadcasts are only forwarded within a VLAN - VLANs are so-called “broadcast domains”.
- Because the protocol is relatively costly, ARP replies are cached
  - With all the usual problems of caches
- On Linux you can look at the cache using `ip neighbour show`
- You can also manipulate this table / sometimes it is useful to fix an Ethernet / IP relation



# NETWORK PROTOCOLS

## UDP

- Adds concept of “port” to IP
- Unreliable datagrams == messages
  - Can get lost
  - Can arrive out of sending order(!)
- Sent in one go
- Maximum size 64 kB

UDP Header – RFC 768



Common UDP Well-Known Ports

Port	Description
7	Echo
19	Chargen
37	Time
53	Domain
67	Bootps (DHCP)
68	Bootpc (DHCP)
69	Tftp
137	Netbios-ns

Port	Description
138	Netbios-dgm
161	Snmp
162	Snmp-trap
500	Isakmp
514	Syslog
520	Rip
33434	Traceroute

### Length

The number of bytes in the entire datagram, including the header; minimum value = 8

### Checksum

Calculated using a pseudo header that includes the IP source and destination addresses, protocol and UDP length, UDP header and data.

Based on RFC 768

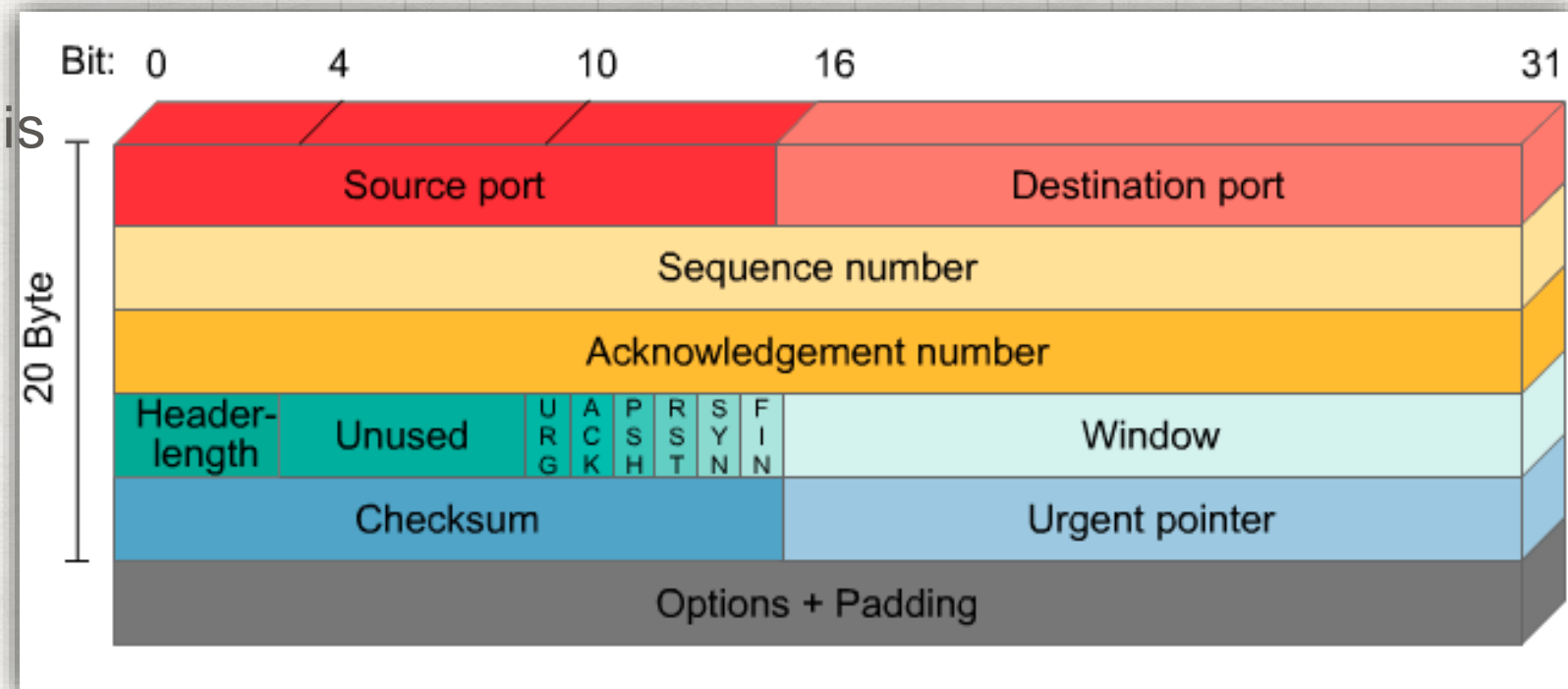
Source: [optimus5.com](http://optimus5.com)



# NETWORK PROTOCOLS

## TCP/IP

- TCP/IP is a child of Unix:
  - everything is a file and what is not a file looks like a file
  - sockets, byte-streams, there is no “maximum” size
  - TCP is a connected protocol
- TCP also has ports which are distinct from UDP ports
- TCP is reliable - packets do not get lost and arrive in sending order



Source: [www.kaderali.de](http://www.kaderali.de)

- Data are sent in chunks called “segments” the maximum segment size is abbreviated to MSS

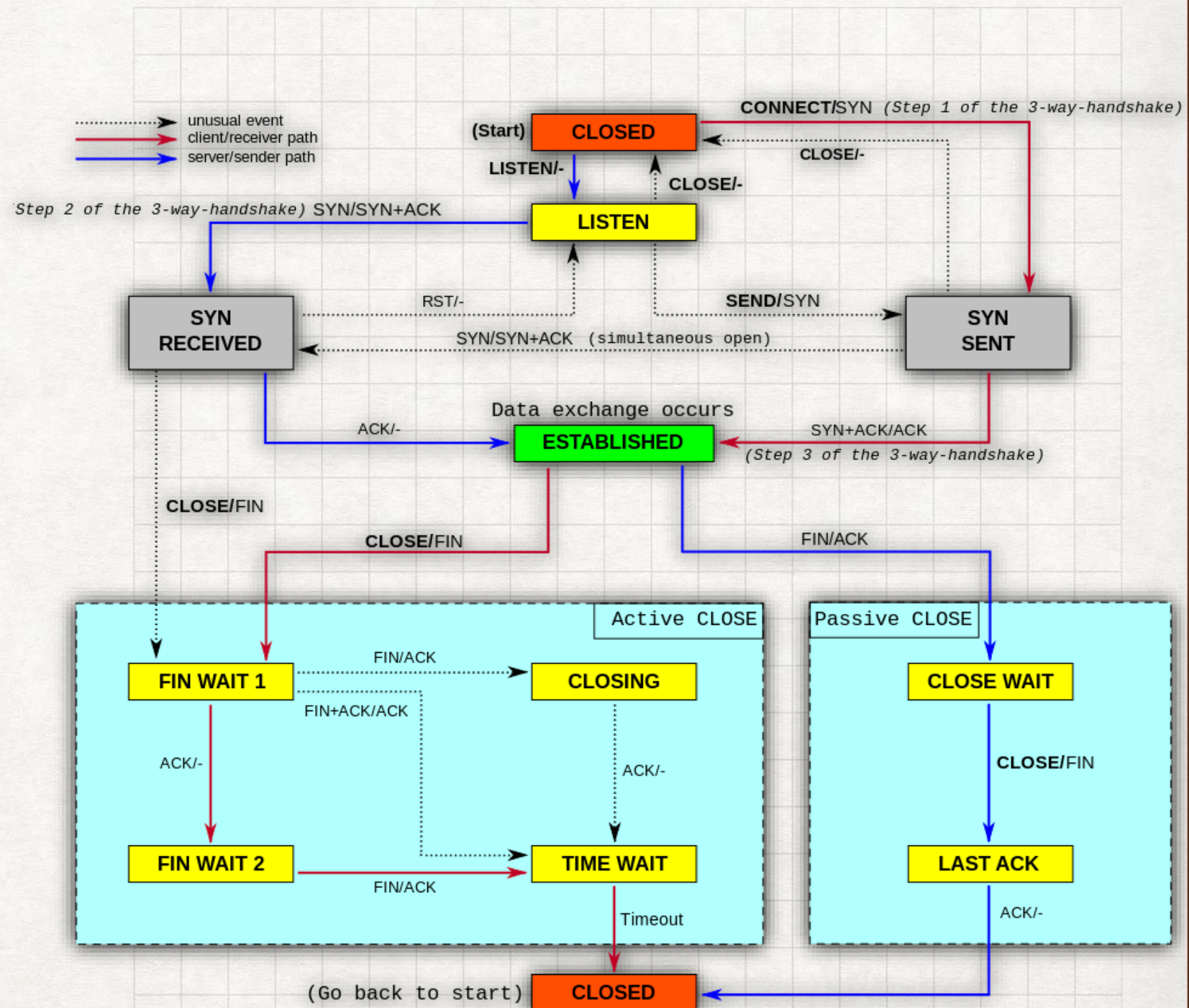


# TCP/IP

## STATE DIAGRAM (SIMPLIFIED)

## STATE DIAGRAM (SIMPLIFIED)

- This needs meditation and a good book (or wikipedia)
- Key points are that every attempt is made that sender and receiver are synchronised throughout the communication (handshakes)
- When used properly - also during shutdown, no data should be lost or corrupted (ever)
- But all this is not quick! For DAQ applications the connection needs to remain.



"Tcp state diagram fixed new" by Scil100. Licensed under CC BY-SA 3.0 via Commons

- [https://commons.wikimedia.org/wiki/File:Tcp\\_state\\_diagram\\_fixed\\_new.svg#/media/File:Tcp\\_state\\_diagram\\_fixed\\_new.svg](https://commons.wikimedia.org/wiki/File:Tcp_state_diagram_fixed_new.svg#/media/File:Tcp_state_diagram_fixed_new.svg)



# NETWORK PROTOCOLS

## SOME FEATURES OF TCP/IP 1)

- MSS and MTU path discovery
  - TCP uses an algorithm to determine the maximum allowed segment size which fits into the least common MTU along the path - this is even dynamic but requires ICMP (not in UDP)
- Nagle algorithm
  - Normally TCP waits a bit to collect small messages into a single segment to be close to the MSS.
  - This can be very bad for latency and can be switched off by setting the TCP\_NODELAY option (setsockopt)



# NETWORK PROTOCOLS

## MORE FEATURES OF TCP/IP

- TCP/IP was conceived to work well over long distance, unreliable lines
- There is a lot(!) of literature and algorithms to achieve good performance under such conditions (“slow start”, “window scaling”, “bandwidth delay product”, ...). Many socket options (tcp(7)) allow to influence this
- In DAQ (usually) a reliable LAN is used and none of this is relevant, as there should be no appreciable packet-loss. This should be verified of course using counters and tools such as wireshark, etc...



# RELIABLE DATAGRAMS

- In DAQ we normally want to send reliable datagrams
  - Can use UDP with homemade re-transmission
  - Or TCP with NO\_DELAY
- POSIX has a few more little pitfalls, e.g. read and recv can return fewer bytes than requested(!). You will see therefore code like this

```
int i = 0;

while (i < len) {
    int n = recv(socket, buf, len, 0);
    if (n < 0) {
        // error
        break;
    } else {
        i += n;
    }
}
```

- Lots of boring boilerplate code has been omitted
- Error-checking is of course (even more) vital in network programming



# PROTOCOLS

## RELIABLE MESSAGES: 0MQ

- Message libraries such as 0MQ or DIM take away some of the pain of dealing with sockets directly
- They encapsulate error-handling and some oddities of the socket semantics (like the boring while loop on recv) and the setup of the addresses
- Why reinvent the wheel? If you can use them!

```
# receiver
import zmq
context = zmq.Context()
socket = context.socket(zmq.REP)
socket.bind("tcp://127.0.0.1:5000")

while True:
    msg = socket.recv()
    print "Got", msg
    socket.send(msg)

# sender
import zmq
context = zmq.Context()
socket = context.socket(zmq.REQ)
socket.connect("tcp://127.0.0.1:5000")
for i in range(10):
    msg = "msg %s" % i
    socket.send(msg)
    print "Sending", msg
    msg_in = socket.recv()
```



# PROTOCOL OVERHEADS

## SOME NUMBERS AND WISDOM

- “In protocol design, perfection is achieved not when nothing can be added, but when nothing can be taken away”
- “Any technology can be **two out of the following three**: cheap, fast, reliable”
  - Every layer a adds a header:
    - Ethernet 14 octets
    - IP 20 octets
    - TCP 20 octets



In addition protocols add packet over heads  
(acknowledgments for TCP for example)



# A SIDENOTE ON DOCUMENTATION

## BEFORE GOOGLE THERE WAS MAN

- The man pages come in several sections
  - Section 7 contains all the protocols (man 7 ip, man 7 tcp)
  - Section 2 contains the sys-calls (man 2 setsockopt)
  - Section 8 contains the configuration tool (man 8 ip)
- It takes a bit of practice to read them, but understanding them means understanding sockets, protocols, etc...
- Wikipedia is usually an excellent complement

### NAME

`baby -- create new process from two parents`

### SYNOPSIS

`baby -sex [m|f] [-name name]`

### DESCRIPTION

`baby` is initiated when one parent process polls another server process through a socket connection in the BSD version or through pipes in the System V implementation. `baby` runs at low priority for approximately forty weeks and then terminates with a heavy system load. Most systems require constant monitoring when `baby` reaches its final stages of execution.

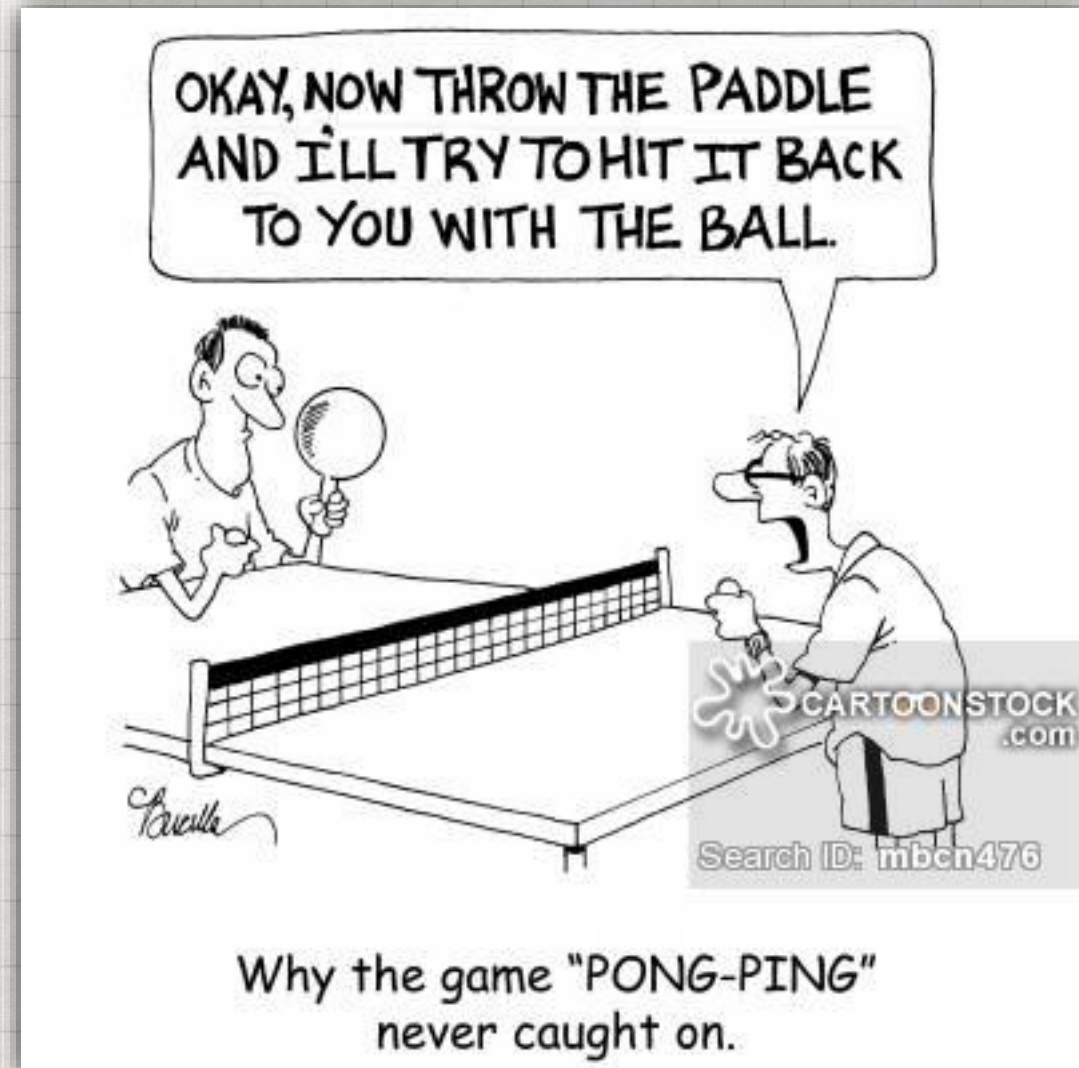
Older implementations of `baby` did not require both initiating processes to be present at the time of completion.



# NETWORK TOOLS TO KNOW

## ICMP PING

- Internet Control Message Protocol (ICMP) is a suite of utility messages to notify network devices about various problems and status of other devices
- The most well known is the Echo Request message which is used by the Ping command
- Watch out for the latency (round-trip time) and its stability



```
gw22:~>ping lab14
PING lab14.lbdaq.cern.ch (10.128.210.114) 56(84) bytes of data.
64 bytes from lab14.lbdaq.cern.ch (10.128.210.114): icmp_seq=1 ttl=60 time=1.11 ms
64 bytes from lab14.lbdaq.cern.ch (10.128.210.114): icmp_seq=2 ttl=60 time=1.17 ms
64 bytes from lab14.lbdaq.cern.ch (10.128.210.114): icmp_seq=3 ttl=60 time=1.20 ms
```



# NETWORK UTILITIES TO KNOW

## IPERF

- Standard tool
- URL:  
<https://iperf.fr/>
- Use only iperf v3
  - v2 has known bugs, in particular for bi-direction tests
- It can use multiple streams in parallel (important at high speeds)

```
<perf-3.1.1plus02:~/devel/iperf-3.1.1>./src/iperf3 -c localhost
Connecting to host localhost, port 5201
[ 4] local ::1 port 43530 connected to ::1 port 5201
[ ID] Interval            Transfer          Bandwidth        Retr   Cwnd
[ 4]  0.00-1.00    sec  1.79 GBytes    15.4 Gbits/sec     0    1.12 MBytes
[ 4]  1.00-2.00    sec  1.84 GBytes    15.8 Gbits/sec     0    1.12 MBytes
[ 4]  2.00-3.00    sec  1.82 GBytes    15.6 Gbits/sec     0    1.25 MBytes
[ 4]  3.00-4.00    sec  1.87 GBytes    16.0 Gbits/sec     0    1.37 MBytes
[ 4]  4.00-5.00    sec  1.83 GBytes    15.8 Gbits/sec     0    1.62 MBytes
[ 4]  5.00-6.00    sec  1.84 GBytes    15.8 Gbits/sec     0    1.75 MBytes
[ 4]  6.00-7.00    sec  1.77 GBytes    15.2 Gbits/sec     0    1.81 MBytes
[ 4]  7.00-8.00    sec  1.86 GBytes    15.9 Gbits/sec     0    1.87 MBytes
[ 4]  8.00-9.00    sec  1.86 GBytes    16.0 Gbits/sec     0    1.87 MBytes
[ 4]  9.00-10.00   sec  1.87 GBytes    16.0 Gbits/sec     0    1.87 MBytes
- - - - -
[ ID] Interval            Transfer          Bandwidth        Retr
[ 4]  0.00-10.00   sec  18.3 GBytes    15.8 Gbits/sec     0
[ 4]  0.00-10.00   sec  18.3 GBytes    15.8 Gbits/sec     0
iperf Done.
```

sender  
receiver



# NETWORK TOOLS TO KNOW

## IPTRAF

- Needs root privs (“sudo”) or CAP\_NET\_ADMIN (man capabilities)
- Good & quick overview on how much “stuff” is going on
- Detailed analysis needs other tools

### IPTraf

#### Statistics for eth0

	Total Packets	Total Bytes	Incoming Packets	Incoming Bytes	Outgoing Packets	Outgoing Bytes
<b>Total:</b>	987	174243	504	41811	483	132432
<b>IPv4:</b>	987	160419	504	34749	483	125670
<b>IPv6:</b>	0	0	0	0	0	0
<b>TCP:</b>	966	158586	494	33878	472	124708
<b>UDP:</b>	10	846	4	304	6	542
<b>ICMP:</b>	11	987	6	567	5	420
<b>Other IP:</b>	0	0	0	0	0	0
<b>Non-IP:</b>	0	0	0	0	0	0

<b>Total rates:</b>	66.6 kbits/sec	<b>Broadcast packets:</b>	0
	50.0 packets/sec	<b>Broadcast bytes:</b>	0

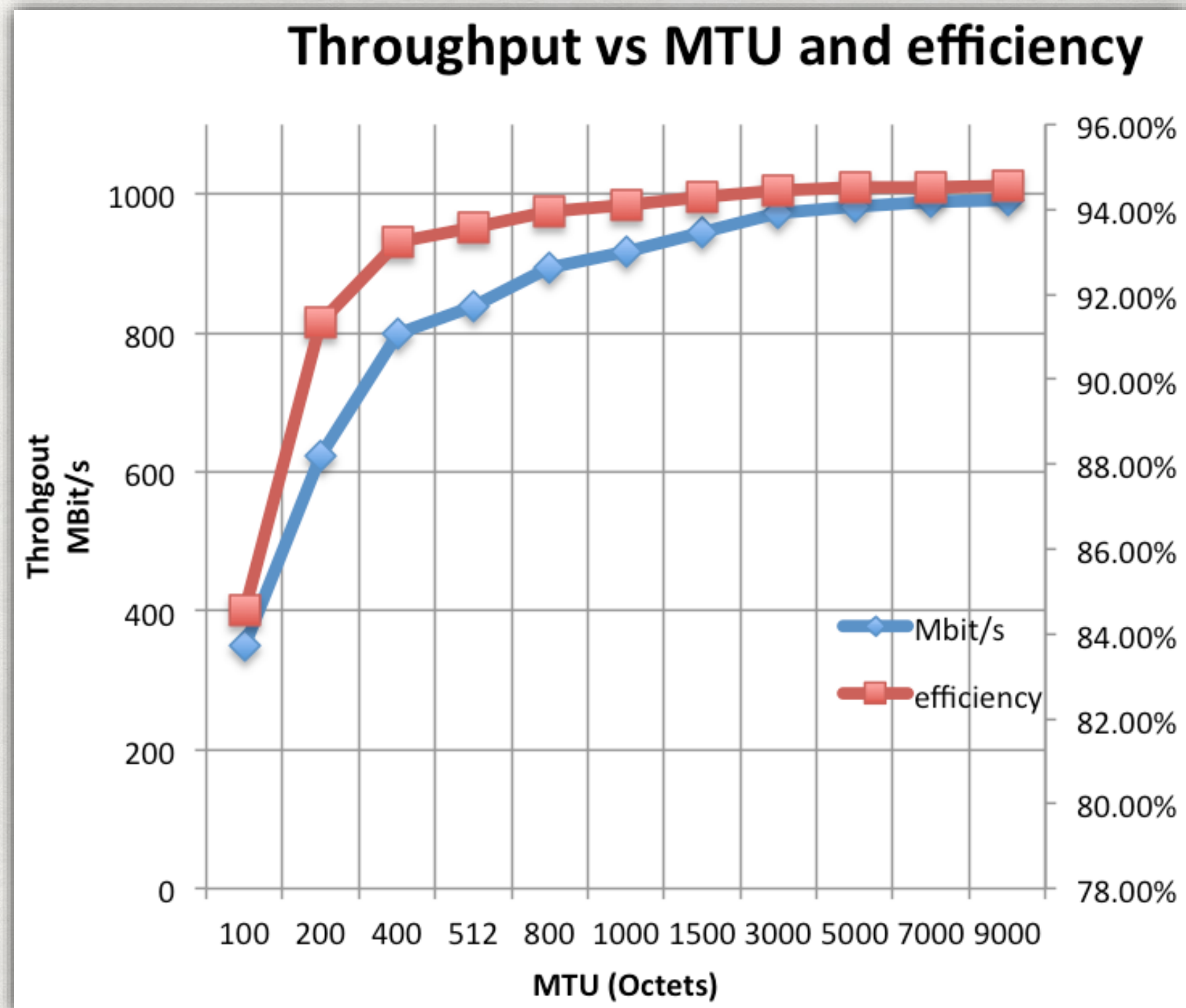
<b>Incoming rates:</b>	18.4 kbits/sec
	25.6 packets/sec

<b>Outgoing rates:</b>	48.2 kbits/sec	<b>IP checksum errors:</b>	0
	24.4 packets/sec		



# THE EFFECT OF THE MTU

- Measured on I350 network controller
- iperf3 TCP/IP no special tunings
- CPU Intel 2630v3 (8 core)
- Note that standard MTU 1500 does not give line-rate even though efficiency is already high. This effect is stronger on weaker CPUs
- $\text{eff} = (\text{MTU} - 54) / \text{MTU}$

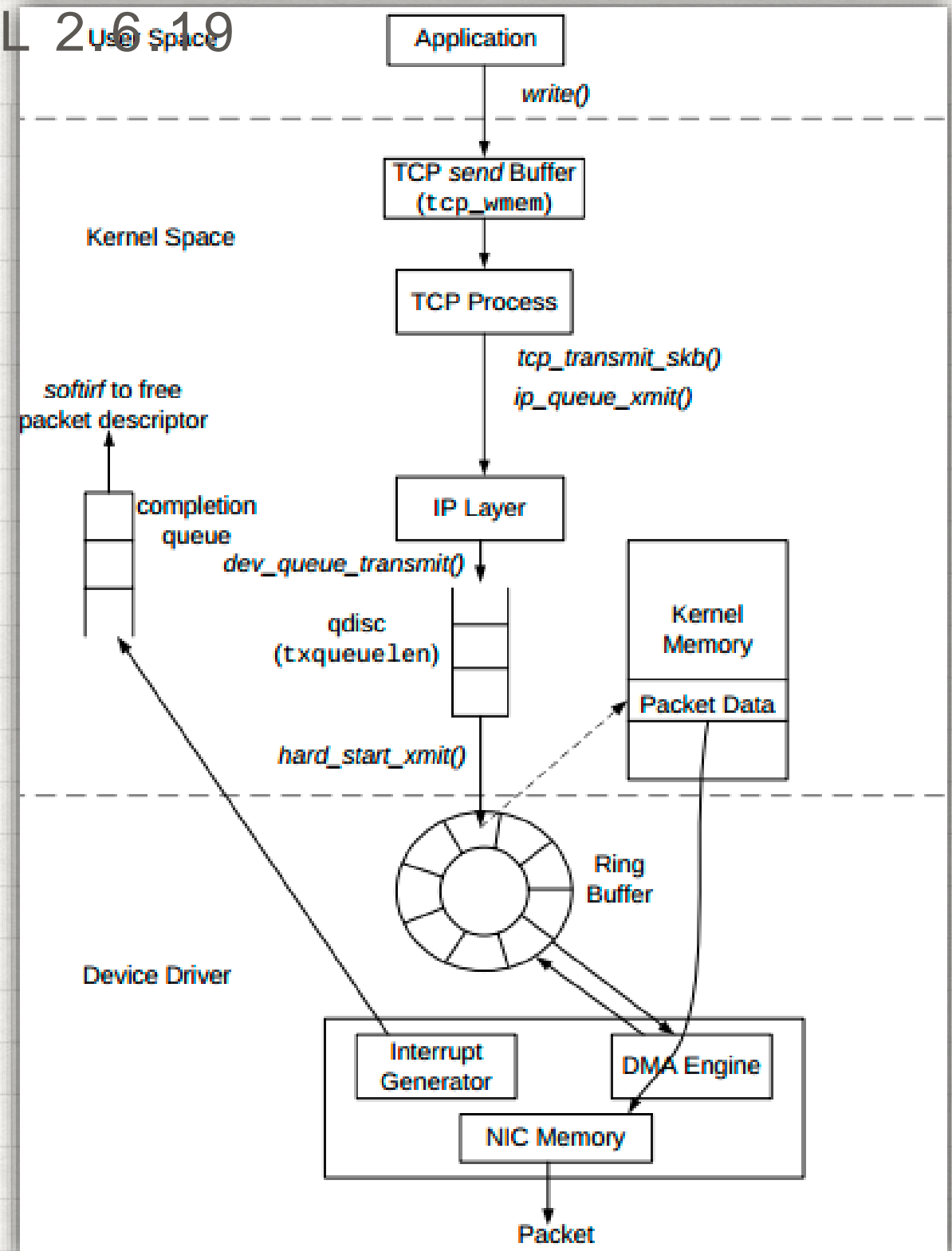




# THE LINUX NETWORK STACK

KERNEL 2.6.19

- The NIC DMA's the packets from/to buffers managed by the kernel `sk_buff`
- Protocol verification and handling done in the kernel
- Read/Write calls from user-space applications normally entail a copy from/to kernel space
- Standard Linux IP protocols do not support zero-copy
- Sizes of many of these buffers can be influenced by kernel parameters



source: <http://www.ece.virginia.edu/cheetah/documents/papers/TCPlinux.pdf>



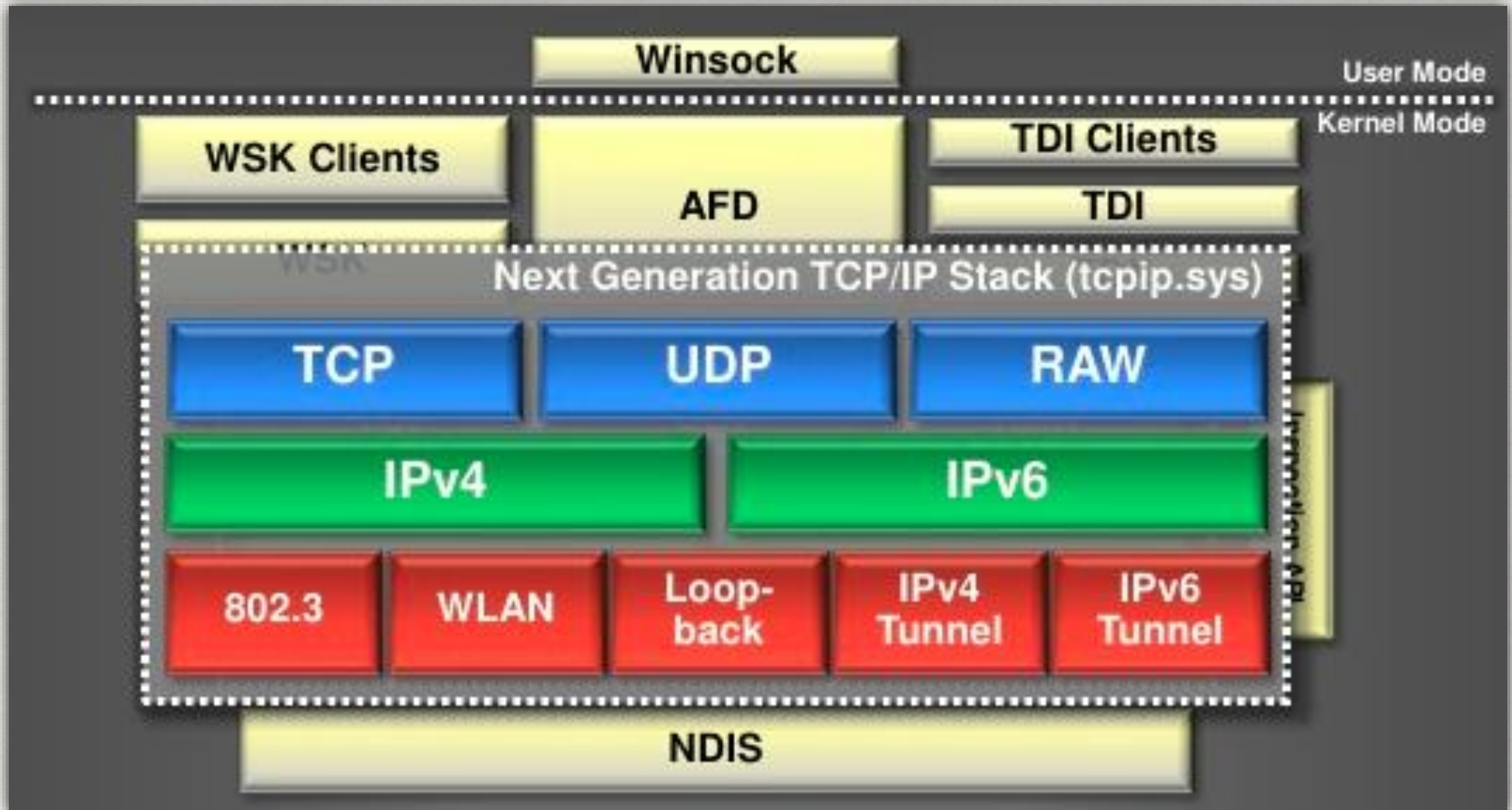
# PERFORMANCE

## LINUX TUNING - KERNEL PARAMETERS

- `net.core.rmem_max`: upper limit for buffer in sockets for reading
  - increase if you loose a lot of messages (UDP) or performance is bad (TCP) - note you also need to adapt the TCP options
- `net.core.wmem_max`: same for writing
- `net.core.netdev_max_backlog`: how many packets can be waiting to be handled
  - increase if you loose messages, can be caused by competing applications or hardware (IRQ sharing!)
- To change use
  - `sysctl -w <parametername> <value>`
- To read use
  - `sysctl <parametername>`
- Documentation in every kernel source tree in the doc folder or <https://www.kernel.org/doc/Documentation/kernel-parameters.txt>



# THE WINDOWS NETWORK STACK

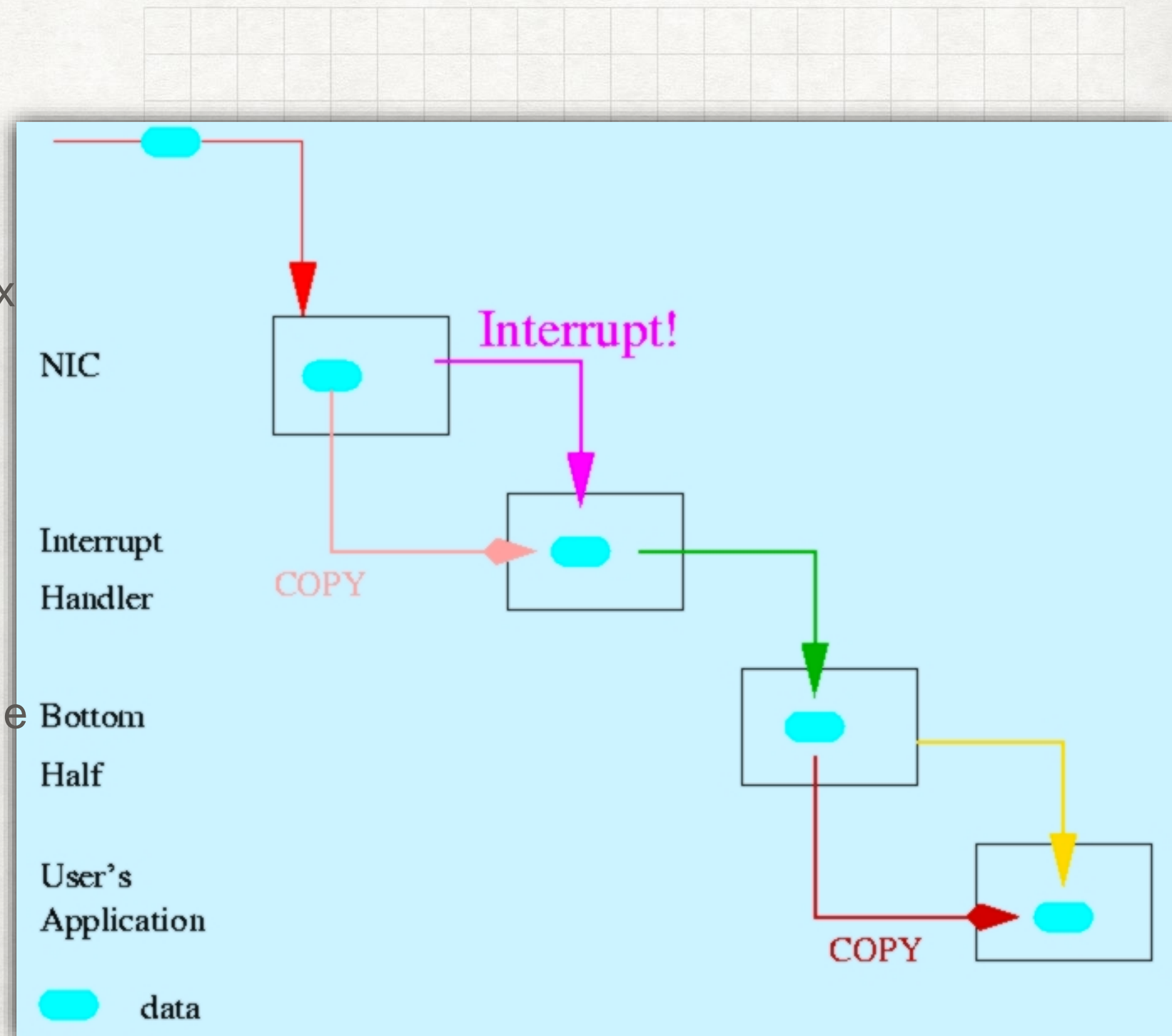


- The good news: winsock is an exact copy of the Berkeley socket interface



# NETWORK ADAPTERS & DRIVERS

- Main manufacturers are today: Intel, Broadcom, in the high end you find also Chelsio, Solarflare, Mellanox
- Watch out for difference between “desktop” and “server” adapter
  - desktop adapters can have very small on-chip memory and be vulnerable to packet loss

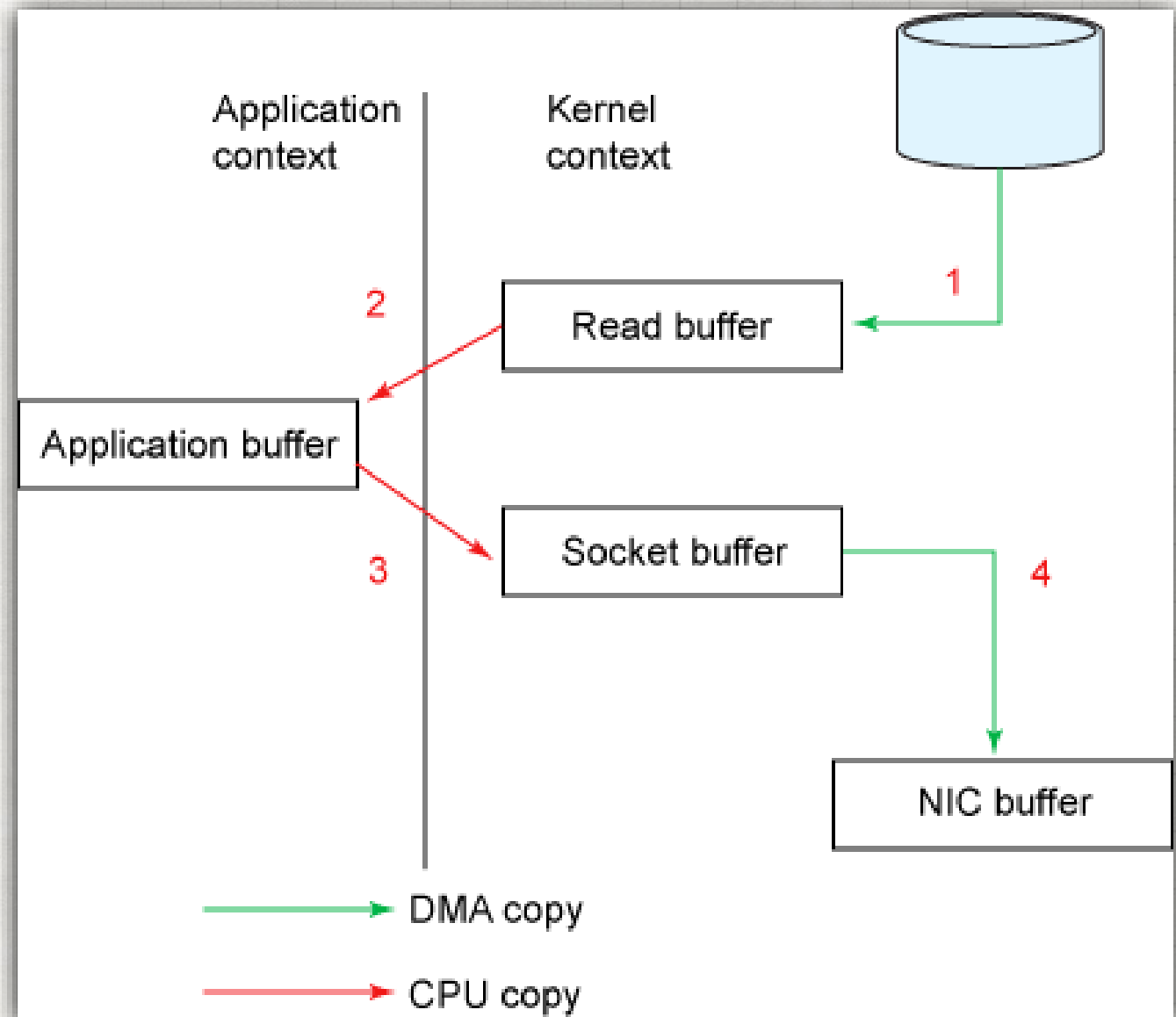




# ZERO-COPY

## THE PROBLEM

- Buffering is crucial in networking because you never know when data will arrive
- But (shared) buffers are risky from a security and stability point of view
- This is avoided by copying to private buffers
- Copy operations however need CPU cycles and memory bandwidth



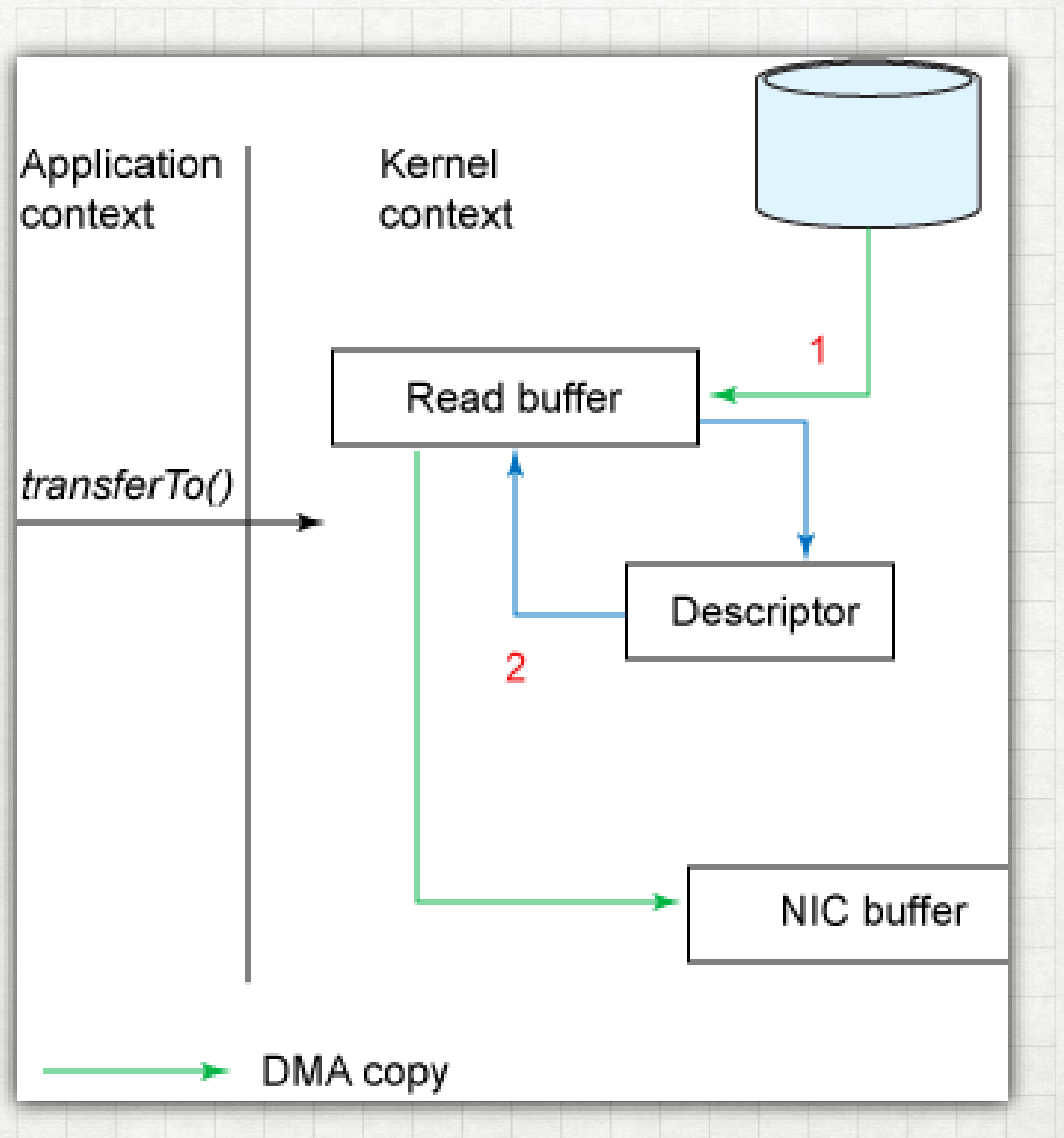
Traditional copy-model in a file-server (source IBM)



# ZERO-COPY

## A SOLUTION

- The key is to let the DMA engines talk to each other
- For this application has to own the buffer (and lock the pages in RAM)
- Kernel safety functions are bypassed
- This is partially possible in Linux for fileserver using the sendfile system call
- In general this is a problem only for data-transfers  $> 10$  Gbit/s and (on Linux at least) requires not using the standard TCP/IP stack





# CONNECTING MANY DEVICES

## SWITCHES

- Switches normally operate at the Ethernet level (Layer 2)
- There are managed switches, which allow configuration and monitoring and unmanaged switches which are fully automatic and not configurable
- Monitoring is usually done via SNMP (a UDP based protocol)
  - From linux use snmpwalk
- For DAQ buffer-sizes should be configured as large as possible
  - Often space can be made available by reducing the number of traffic classes (used for priority)
- Configuration is using a web-interface or a Command Line Interface (CLI) which is usually a variant of the CLI invented by Cisco (if you know one, you will be annoyed but not stopped by any other)



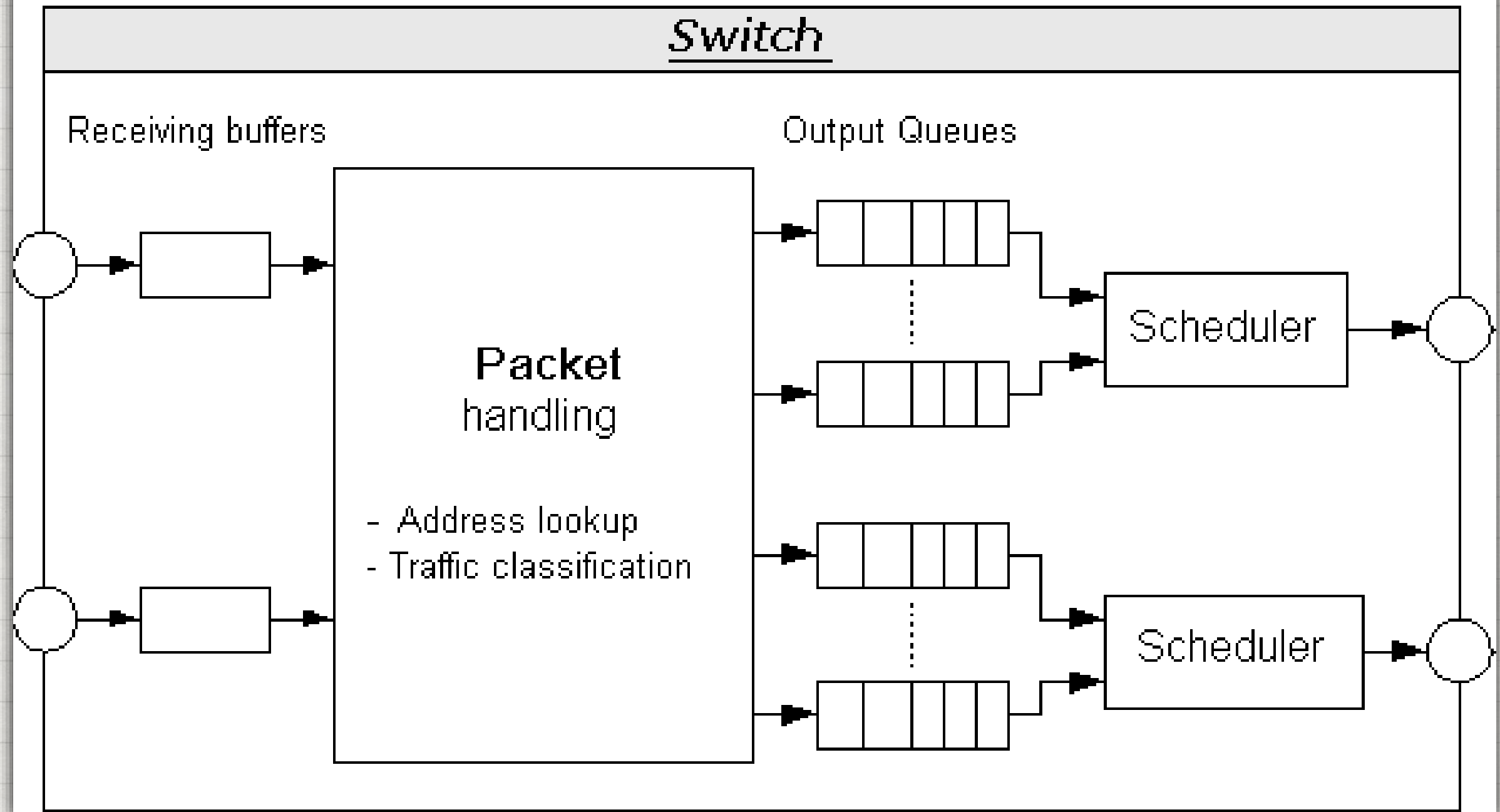
# NETWORK TOOLS TO KNOW

## ETHTOOL

- ethtool allows you to manipulate the Ethernet layer and your NIC
- It shows very useful information , eg:
  - #link status  
sudo ethtool eth1 | grep Link  
Link detected: yes
  - #link statistics  
sudo ethtool eth1 -S  
# loads of numbers defined by 802.3 (watch out for errors)  
Even a small amount of errors usually means a serious problem (cable, optics, etc...)
- It can be used to set important parameters, eg:
  - # set number of DMA descriptors for receiving to 4000  
sudo ethtool eth1 -G rx 4000



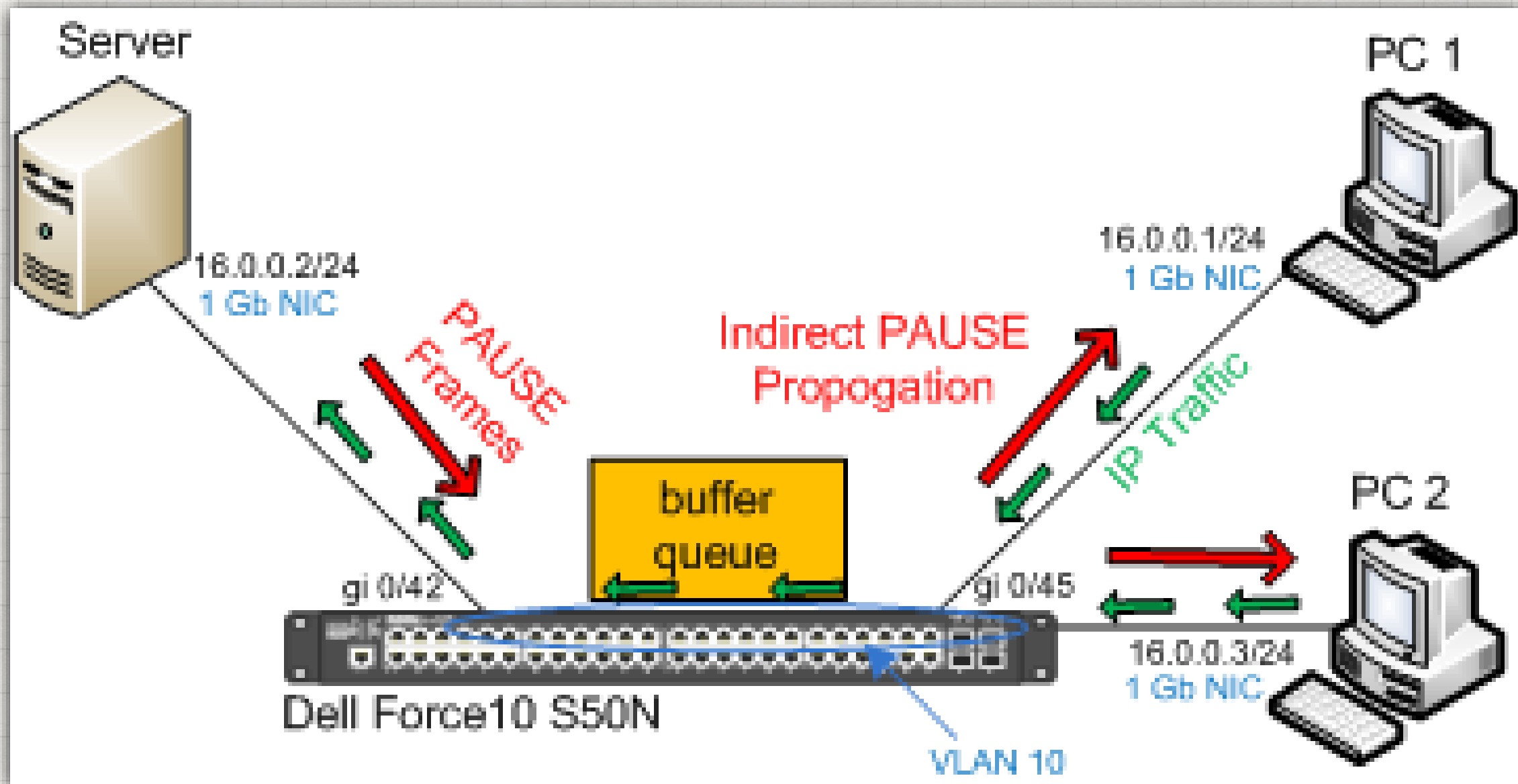
# ETHERNET SWITCH INSIDE





# ETHERNET FLOW CONTROL

## AND IT'S PROBLEMS





# CONNECTING NETWORKS

## ROUTERS

- In DAQ we tend to use private LANs
- For larger or multistage systems, it is almost always better to use layer 3 (IP). A layer-3 switch is called a router.
  - IP addresses are easier to manage, they have symbolic names, there is a (simple) loop protection
- Routers offer many, many features, most of which are not needed in a LAN for DAQ
- A static setup eases debugging, the failover features of internet routing protocols are usually too slow to be of big help in DAQ



# NETWORK TOOLS TO KNOW

## TRACEROUTE

- traceroute (tracert on M\$-Windows)
- Checks the connectivity between nodes. It shows all (many) of the intermediate devices.
  - If you loose packets any of them are a candidate
- It can show you if things go the way you think they would

```
hltb1001:~>traceroute -n www.cern.ch
traceroute to www.cern.ch (188.184.9.235), 30 hops max, 60 byte packets
 1  10.130.122.254  1.061 ms  1.067 ms  1.099 ms
 2  172.16.1.1  0.357 ms  0.407 ms  172.16.1.5  0.447 ms
 3  137.138.18.241  2.460 ms  2.456 ms  2.626 ms
 4  172.24.2.145  0.361 ms  0.387 ms  0.381 ms
 5  172.24.3.17  0.457 ms  172.24.3.81  0.454 ms  0.493 ms
 6  172.24.3.178  0.517 ms  172.24.3.158  0.404 ms  172.24.3.178  5.135 ms
 7  188.184.9.235  0.564 ms  0.572 ms  0.548 ms
```



# NETWORK TOOLS TO KNOW

## LIBPCAP

- Filtering lets you focus on what's interesting
- It is much more efficient to filter at the libpcap ("capture") level than using the display filter in Wireshark
- This is shared by many tools: dumpcap, tcpdump, etc..

Capture only traffic to or from IP address 172.18.5.4:

```
host 172.18.5.4
```

Capture traffic to or from a range of IP addresses:

```
net 192.168.0.0/24
```

or

```
net 192.168.0.0 mask 255.255.255.0
```

Capture traffic from a range of IP addresses:

```
src net 192.168.0.0/24
```

or

```
src net 192.168.0.0 mask 255.255.255.0
```

Capture traffic to a range of IP addresses:

```
dst net 192.168.0.0/24
```

or

```
dst net 192.168.0.0 mask 255.255.255.0
```

Capture only DNS (port 53) traffic:

```
port 53
```

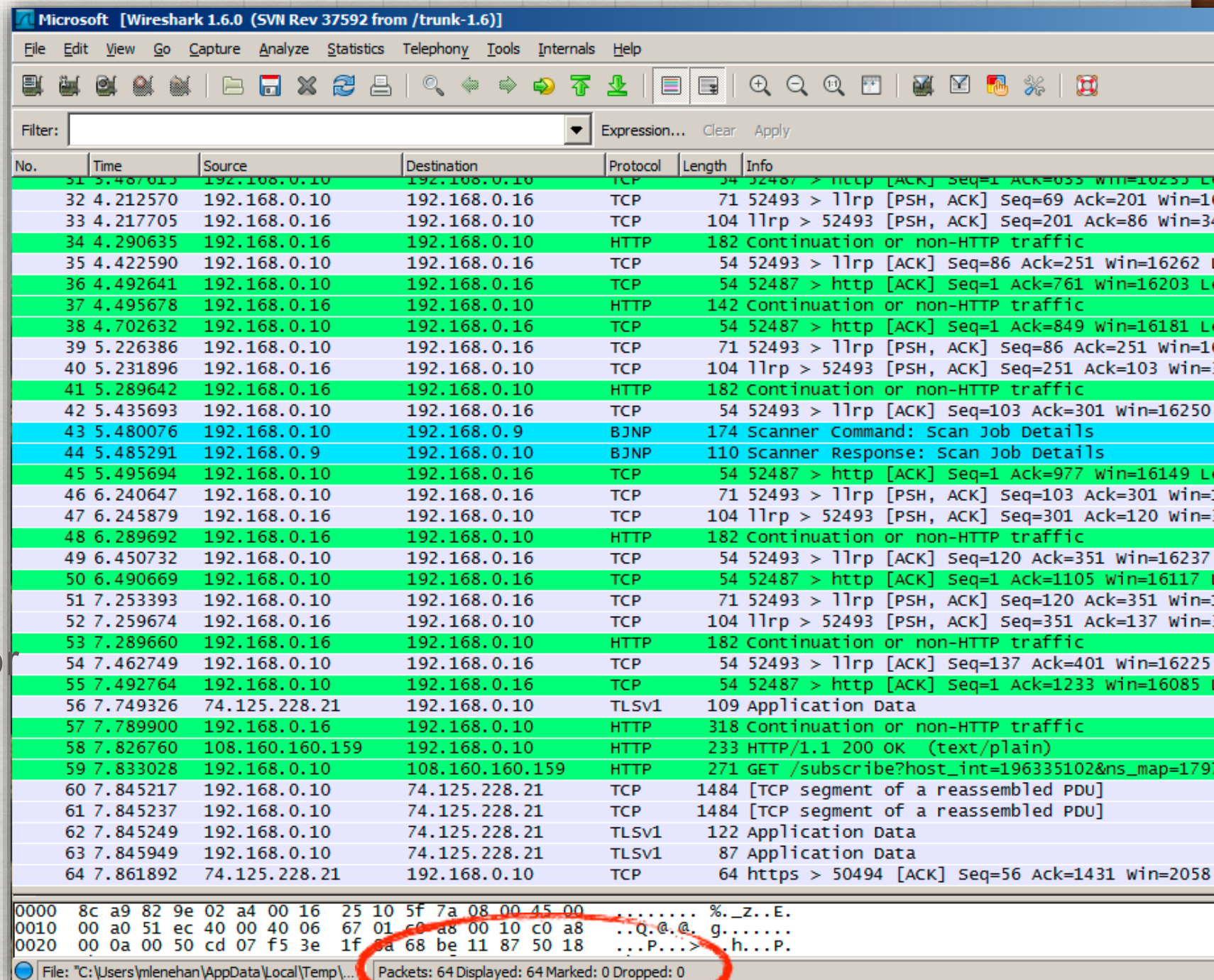
From the Wireshark wiki



# NETWORK TOOLS TO KNOW

## WIRESHARK

- A super-powerful tool
- Comes in a gui and a console version (tshark)
- Is built on top of libpcap
- It understands virtually all known network protocols
- Packet processing at high speed is tricky, watch out for dropped packets —> not seen does not mean not received





# INTERPRETING THE HEX DUMP

## NETWORK BYTE ORDER

	Low address				High address			
Address	0	1	2	3	4	5	6	7
Little-endian	Byte 0	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6	Byte 7
Big-endian	Byte 7	Byte 6	Byte 5	Byte 4	Byte 3	Byte 2	Byte 1	Byte 0
Memory content	0x11	0x22	0x33	0x44	0x55	0x66	0x77	0x88
64 bit value on Little-endian				64 bit value on Big-endian				
0x8877665544332211				0x1122334455667788				

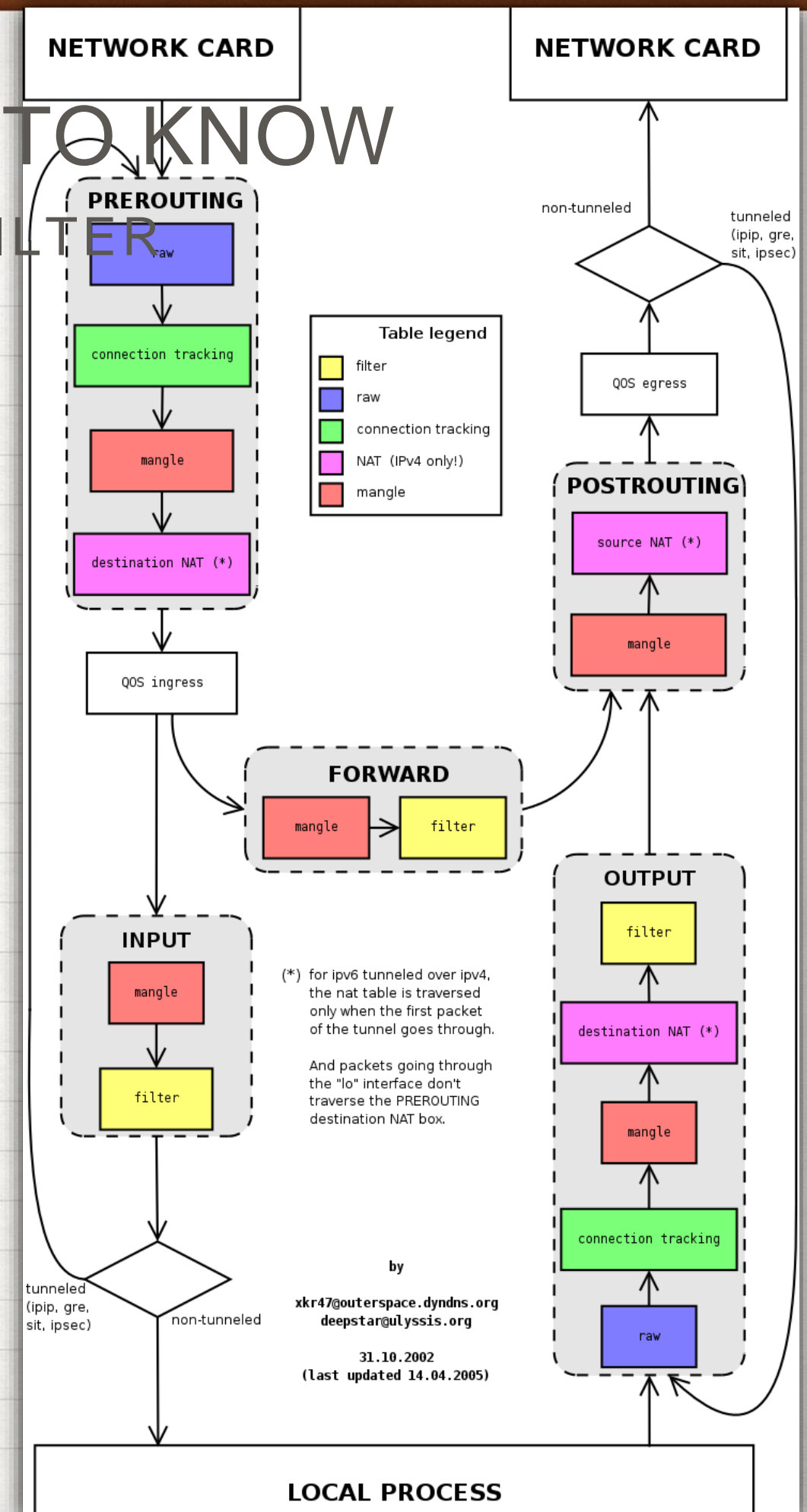
- Historically network byte order is big-endian
- In the 21st century almost all processors are little-endian
- POSIX has portable conversion functions (`man htonl`)
  - `htonl`, `ntohl`, `htons`, `ntohs` h ... host, n ... network



# NETWORK TOOLS TO KNOW

## IPTABLES / NETFILTER

- “Firewall” used for counting
- Monitoring independent of running application
- No copy from kernel space —> efficient





# WIFI FOR DAQ?

## NOT A GOOD IDEA

- WiFi is a shared medium (like the original Ethernet)
- Bandwidth is erratic
- Stability is poor
  - Check latencies with ping or throughput with iperf
- Your mileage may vary of course, if you're needs are far below the effective speed

Standard	eff. speed
802.11 b	2 Mbit/s
802.11 g	20 Mbit/s
802.11 n	40 Mbit/s
802.11 ac	60 Mbit/s

source: [www.speedguide.net](http://www.speedguide.net)



source: [www.dumbblog.com](http://www.dumbblog.com)



# FURTHER READING

- Best is simply to try out
- man-pages and wikipedia have practically all info which is needed
- For book-worms there is “Unix Network Programming, Volume 1: The Sockets Networking API (3rd Edition)”





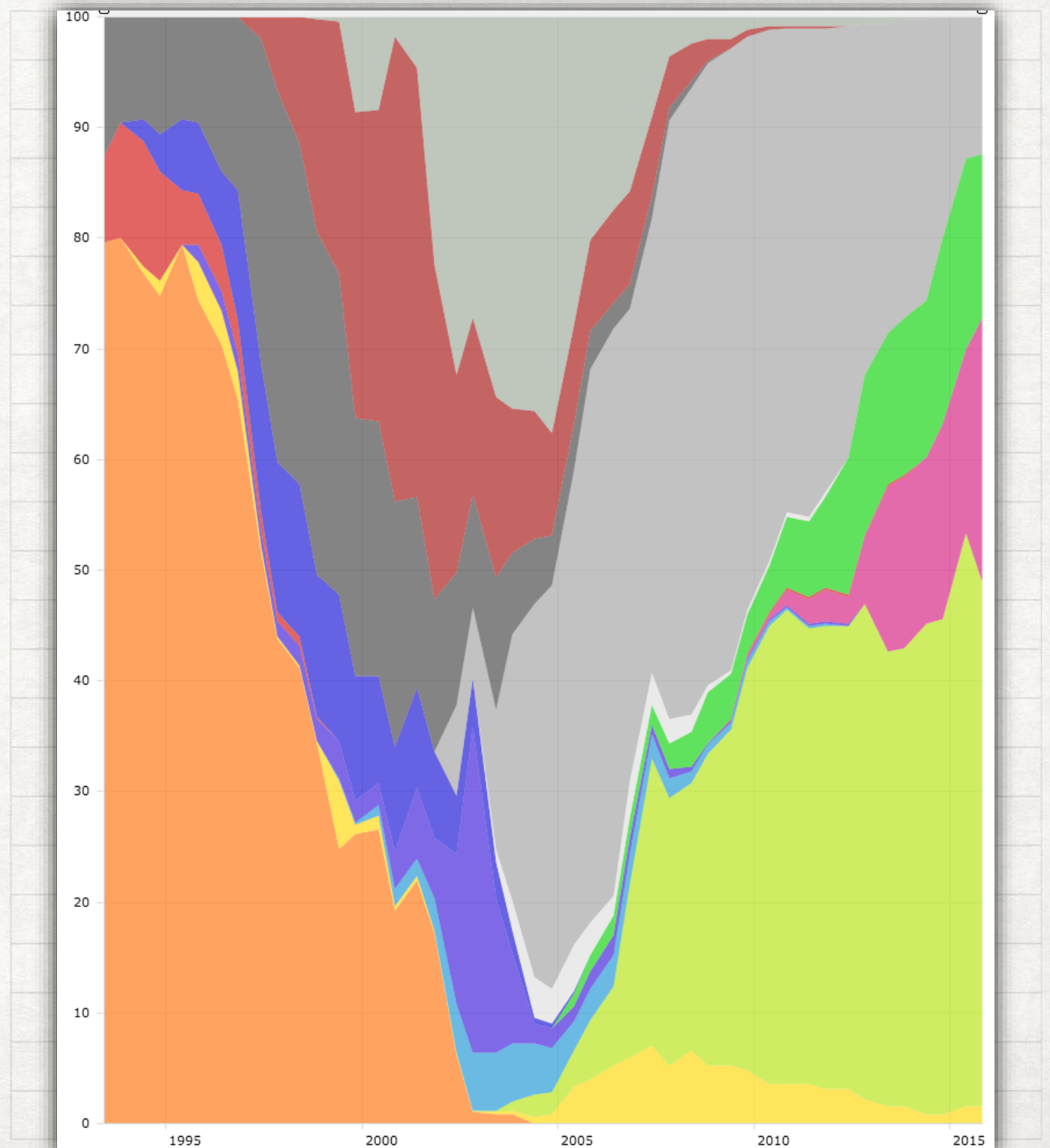
# APPENDIX



# THE CASE FOR CLUSTER-INTERCONNECTS

## COST, COST AND COST

- Per unit of bandwidth InfiniBand and OmniPath are more cost-effective than Ethernet (at least at the top-speeds)
- They tend also to use less CPU power than the TCP/IP stack (only relevant at speeds > 10 Gbit/s)
- Using them is \*much\* more difficult in practice - believe me - I've been there :-)





# THE OFED

## ANOTHER STACK

