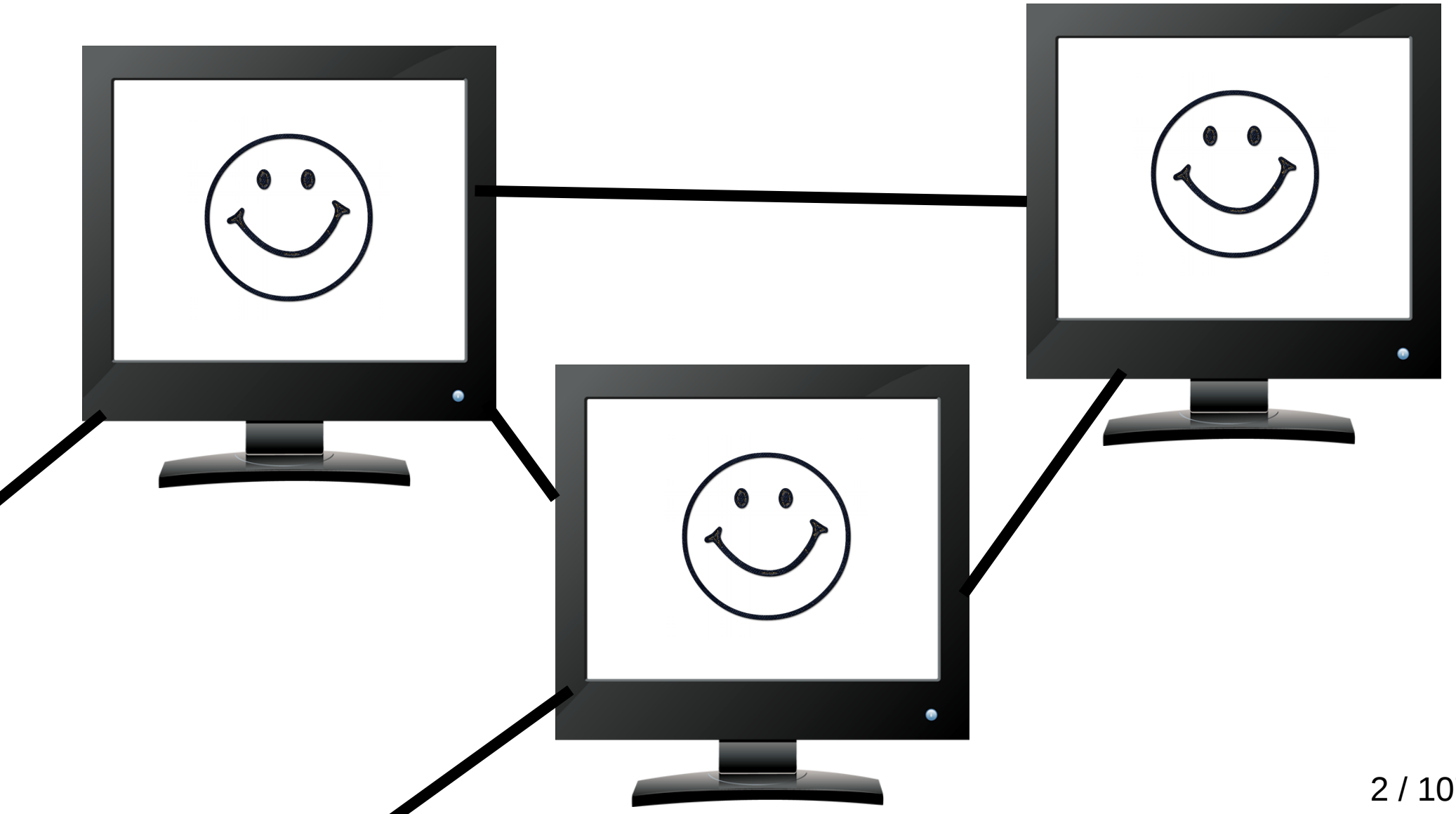
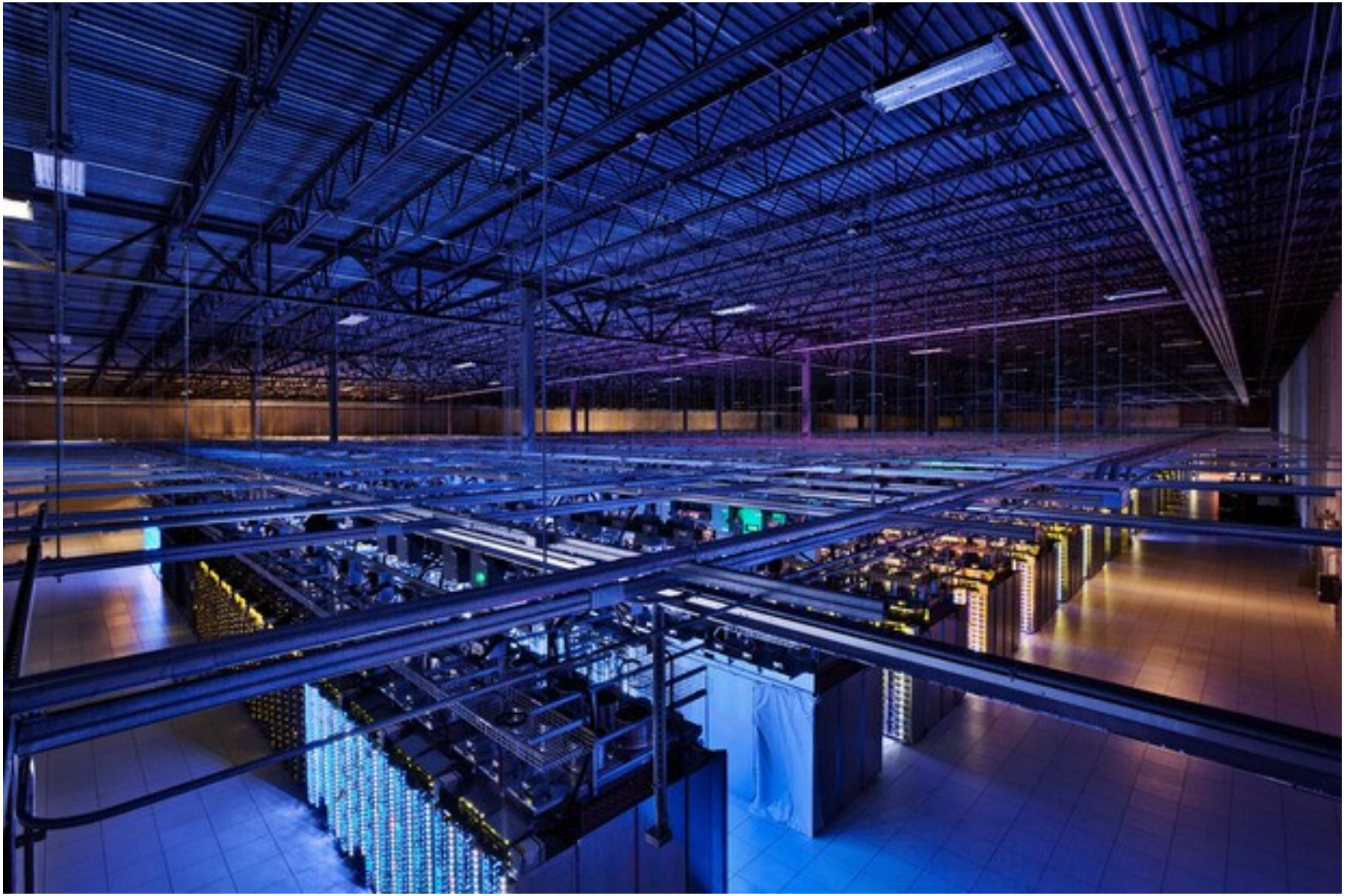
HPC


НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ
МИФИ

Remember?



Remember?

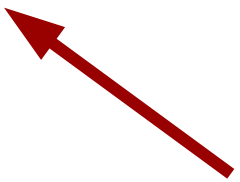


Основные характеристики вычислительного кластера

- Пиковая (теоретическая) производительность (Tflops)
- Linpack производительность (Tflops)
- BogoMIPS производительность?
- Пропускная способность Interconnect (Gbps)
- Задержки Interconnect (ns)

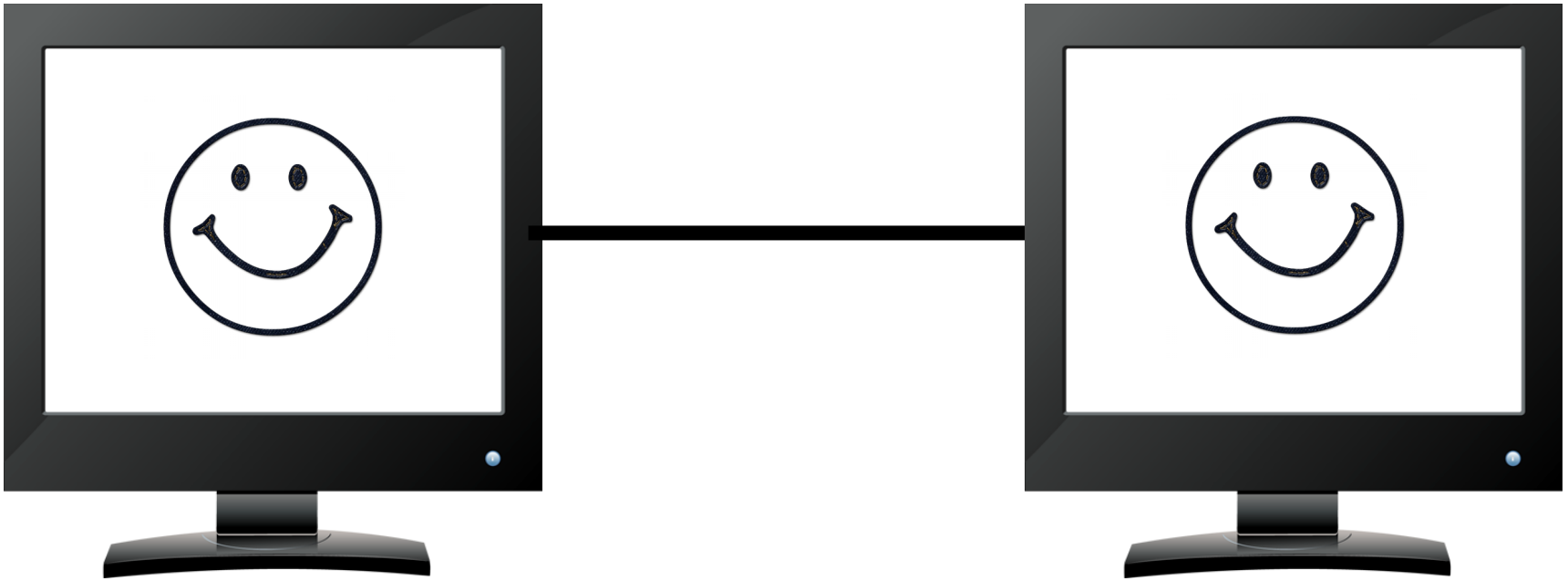
Основные характеристики вычислительного кластера

- Пиковая (теоретическая) производительность (Tflops)
- Linpack производительность (Tflops)
- BogoMIPS производительность?
- Пропускная способность Interconnect (Gbps)
- Задержки Interconnect (ns)



сеть

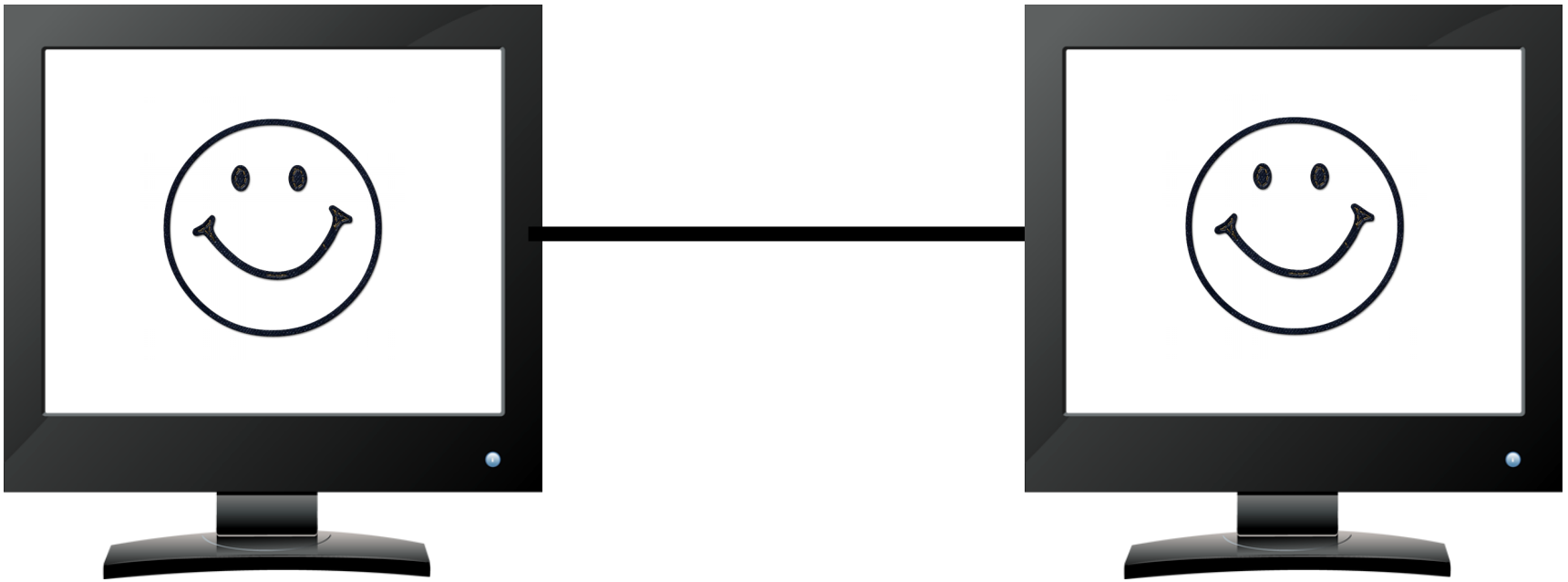
Что такое компьютерная сеть?



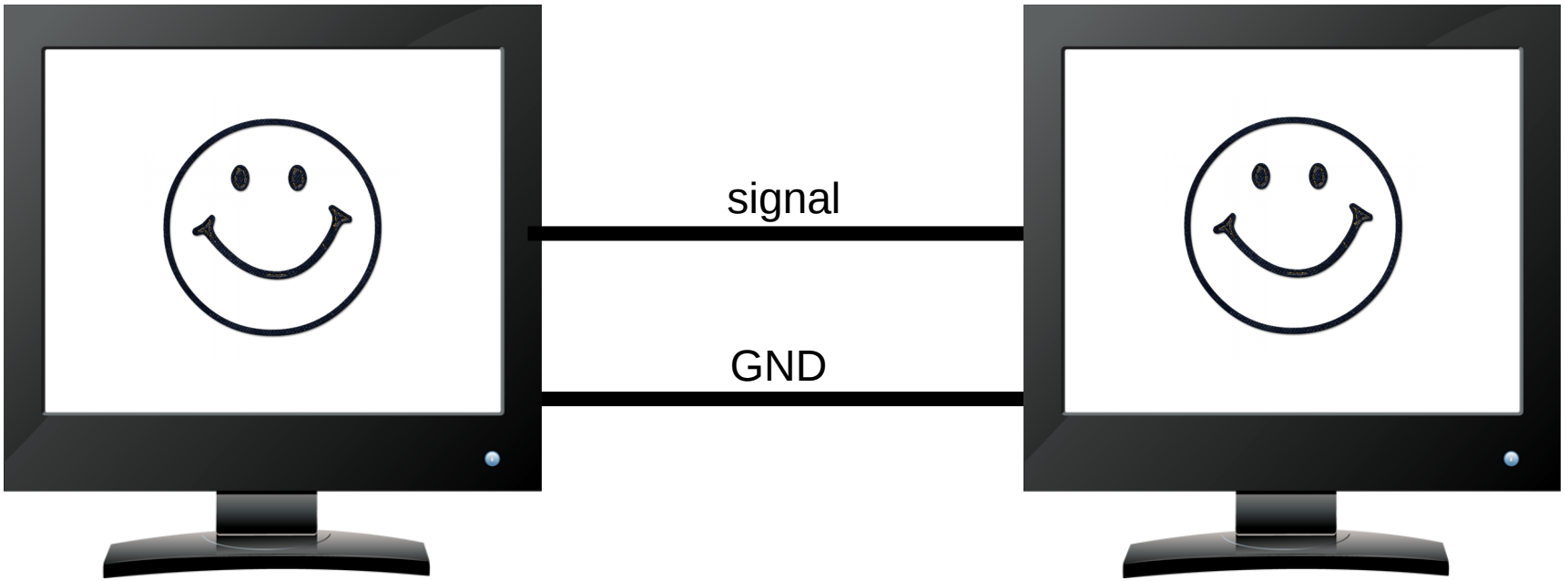
Что такое компьютерная сеть?

- Универсальность (любой компьютер)
- Масштабируемость
- Скорость
- Низкая стоимость

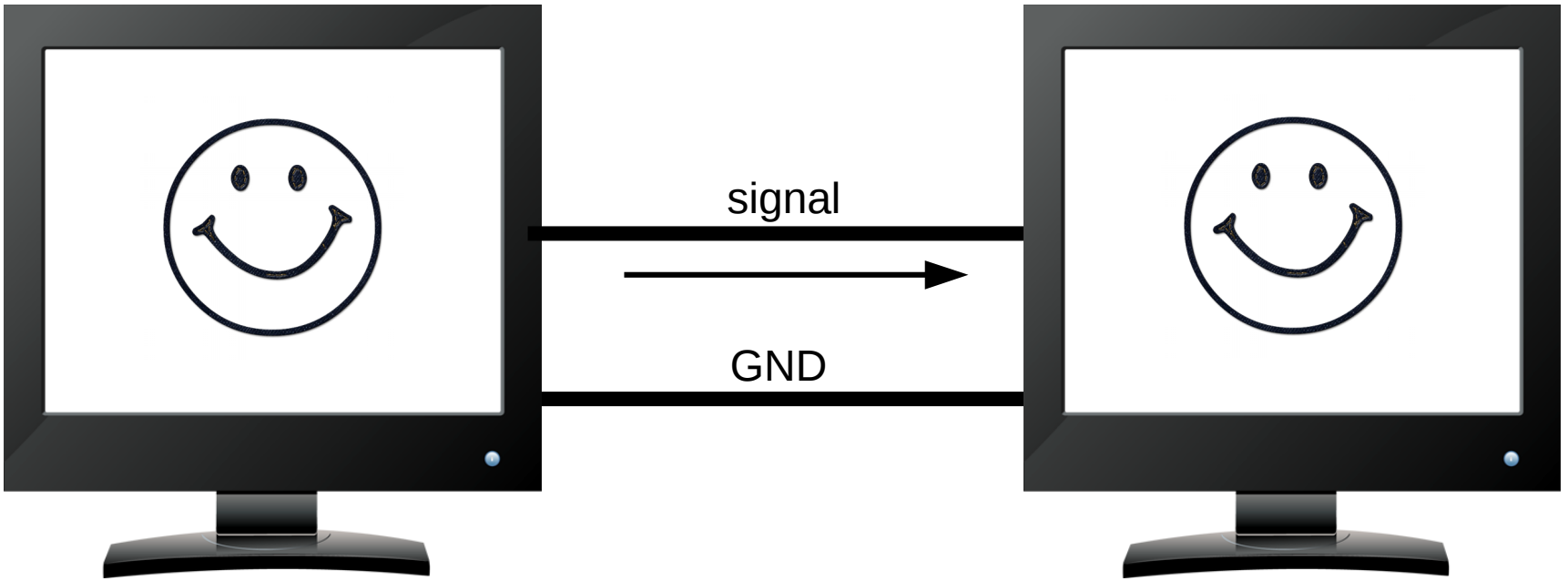
Ethernet



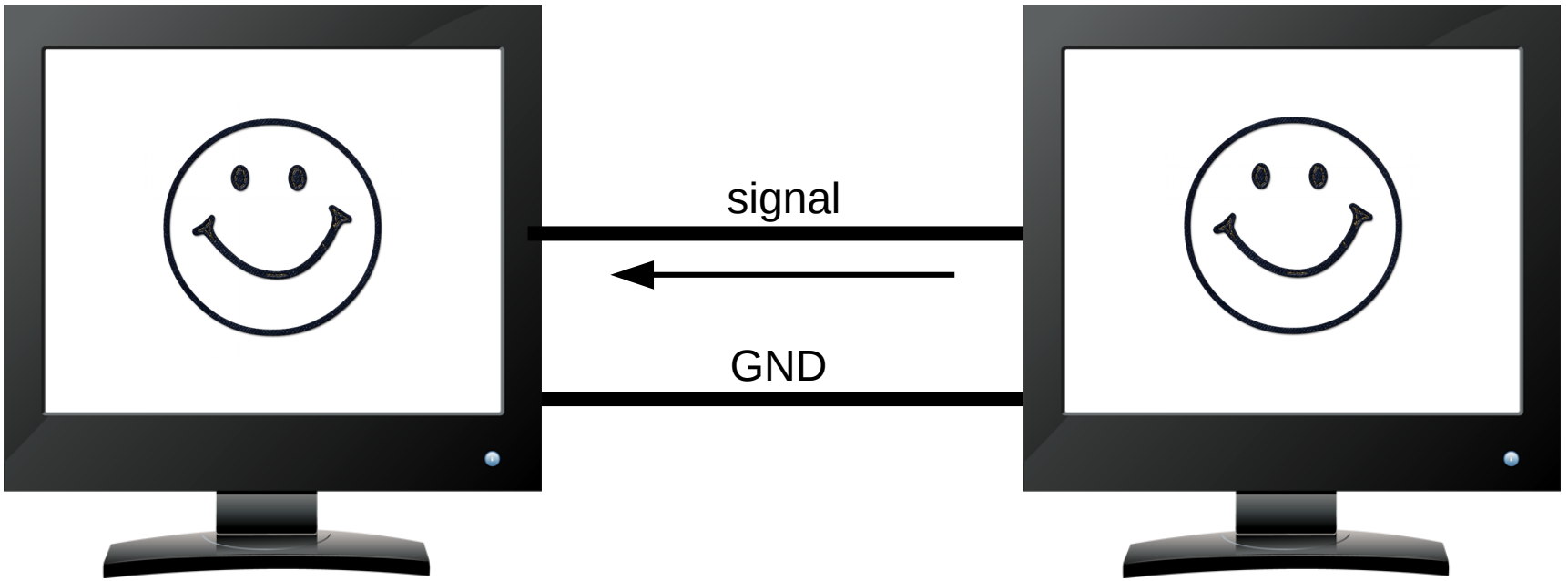
Ethernet



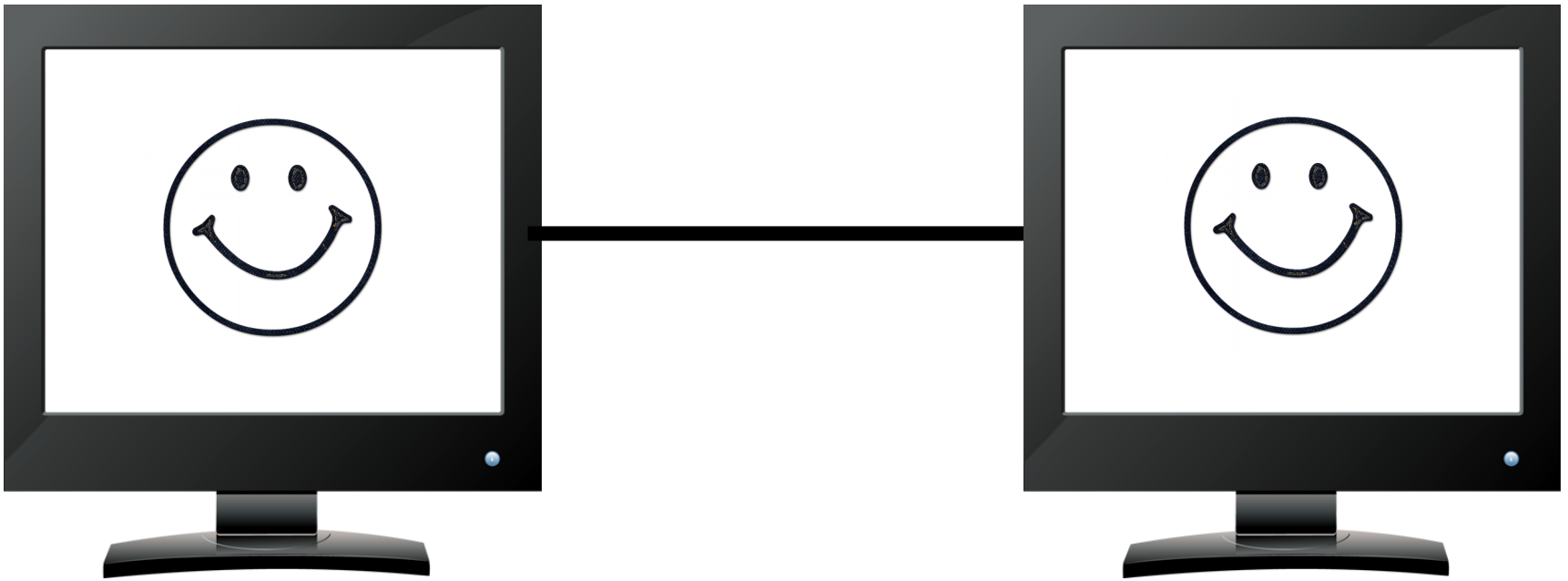
Ethernet



Ethernet

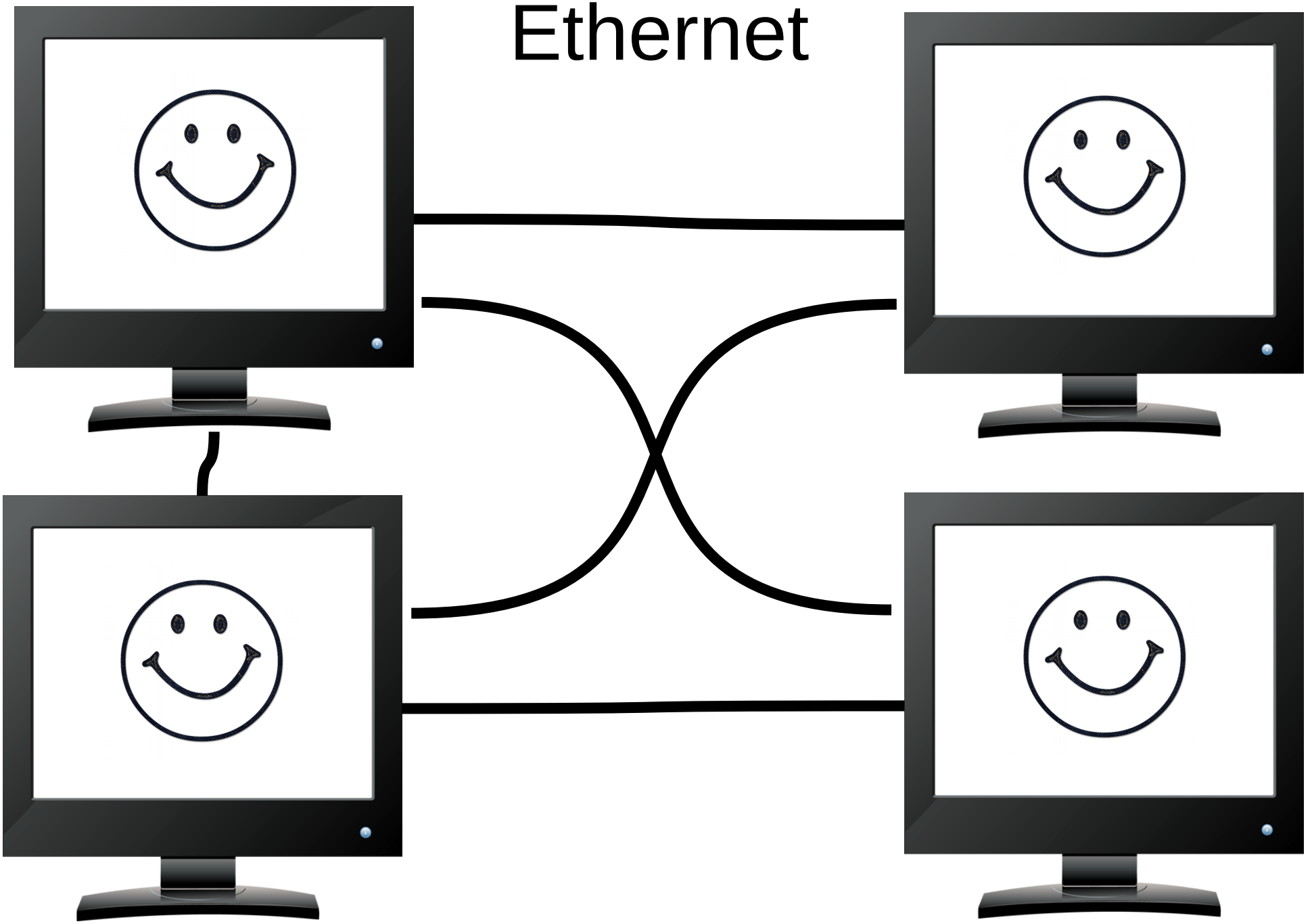


Ethernet



Ethernet

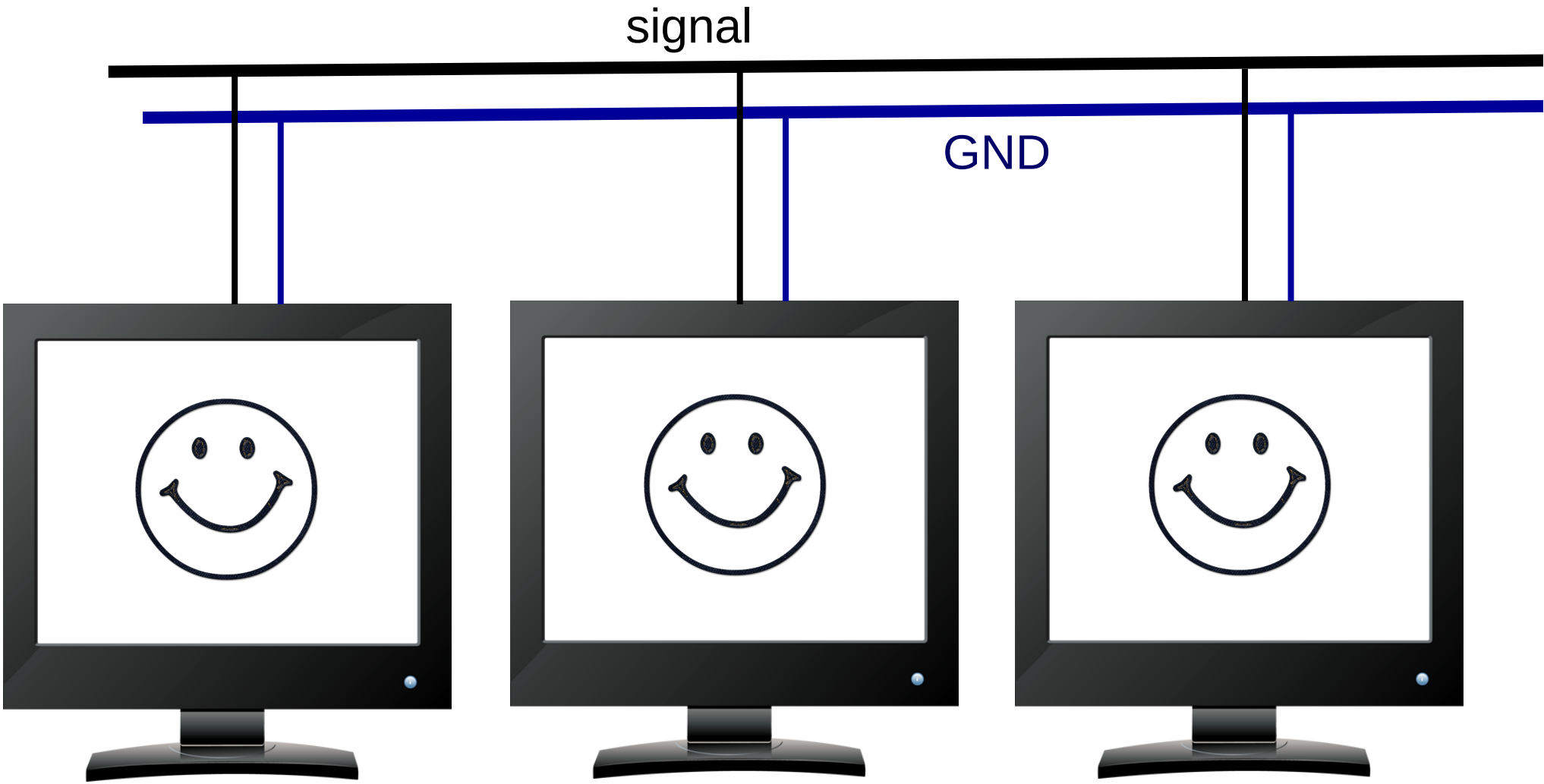
Ethernet



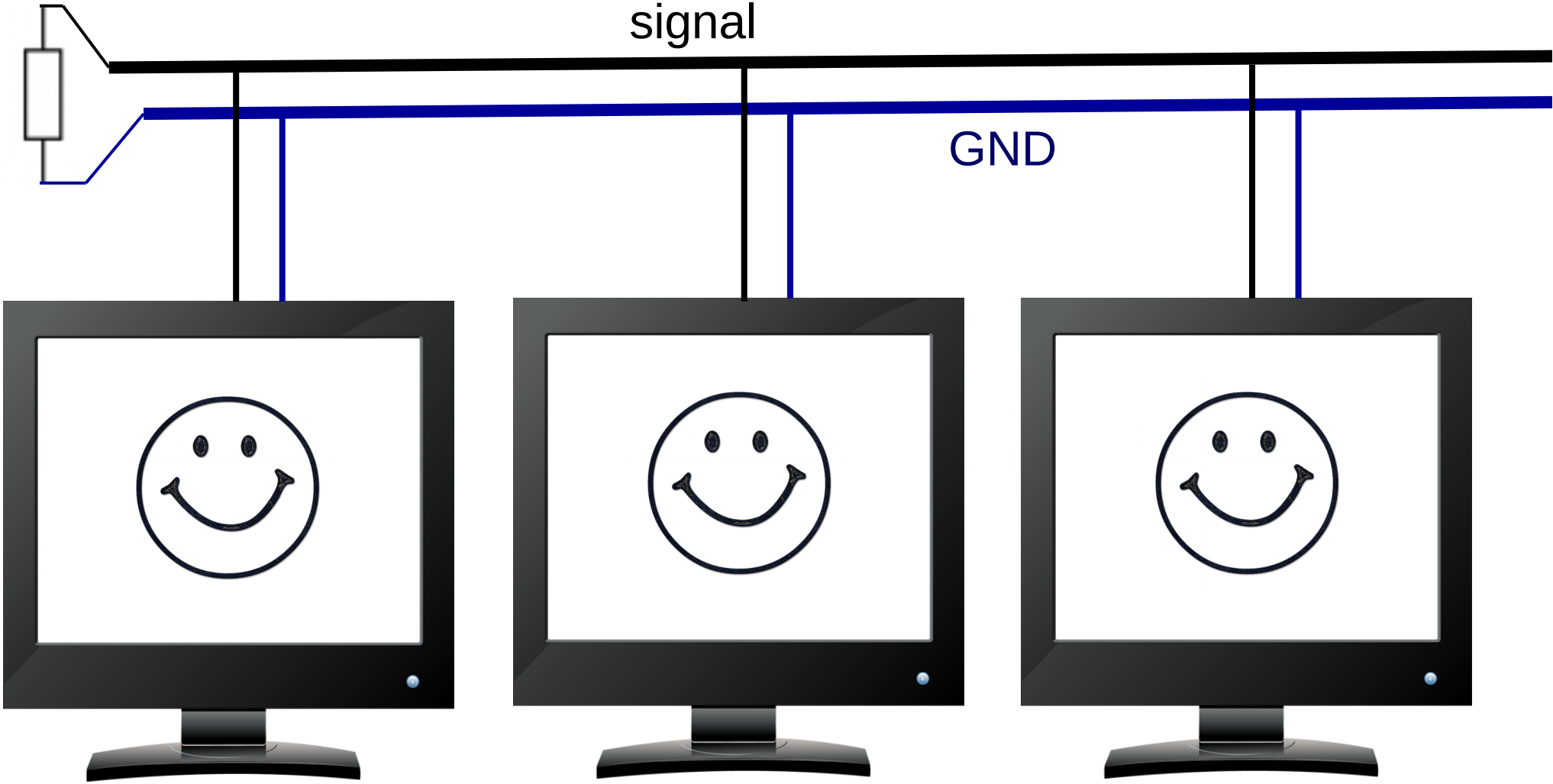
Ethernet

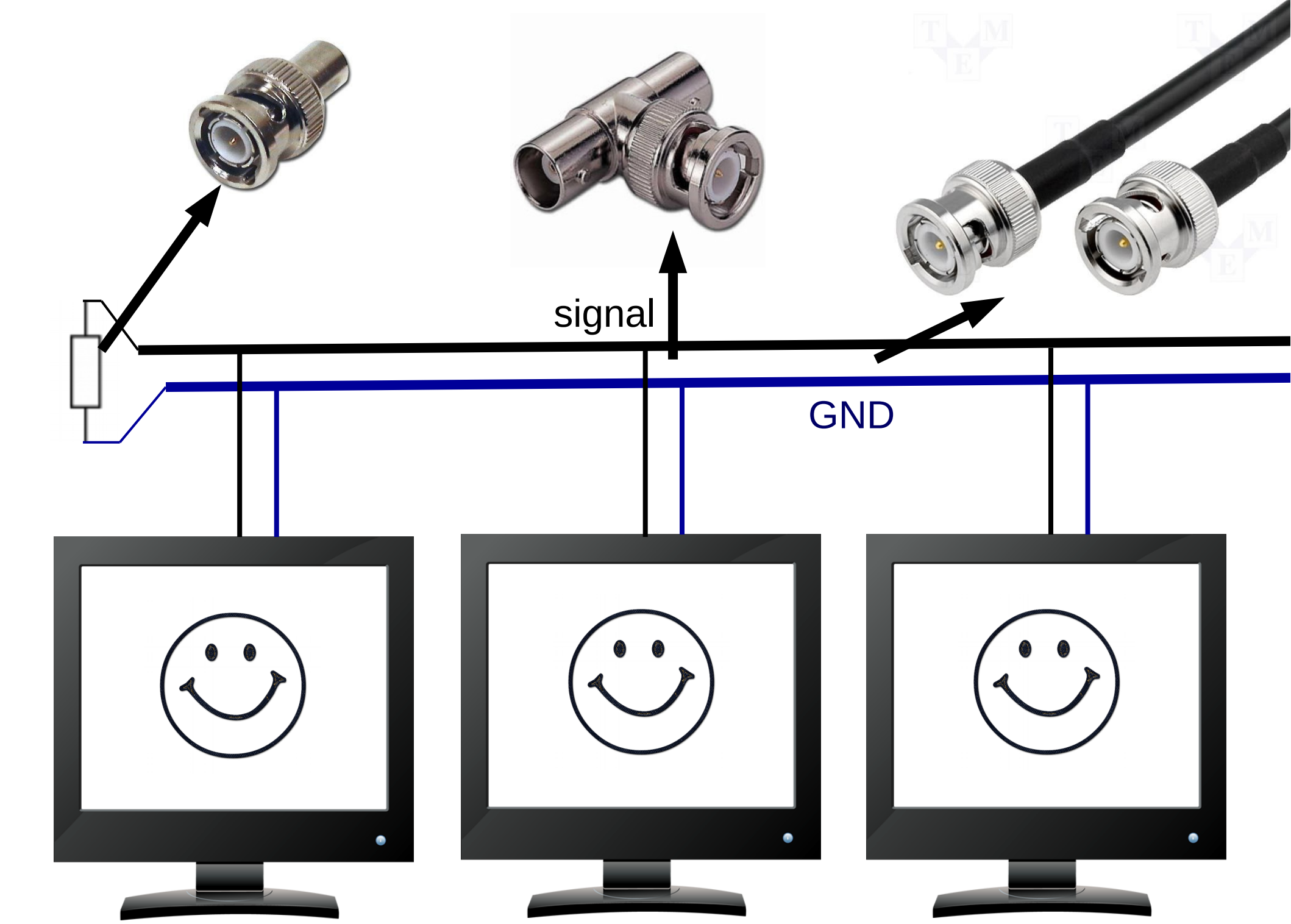
Ethernet

Ethernet

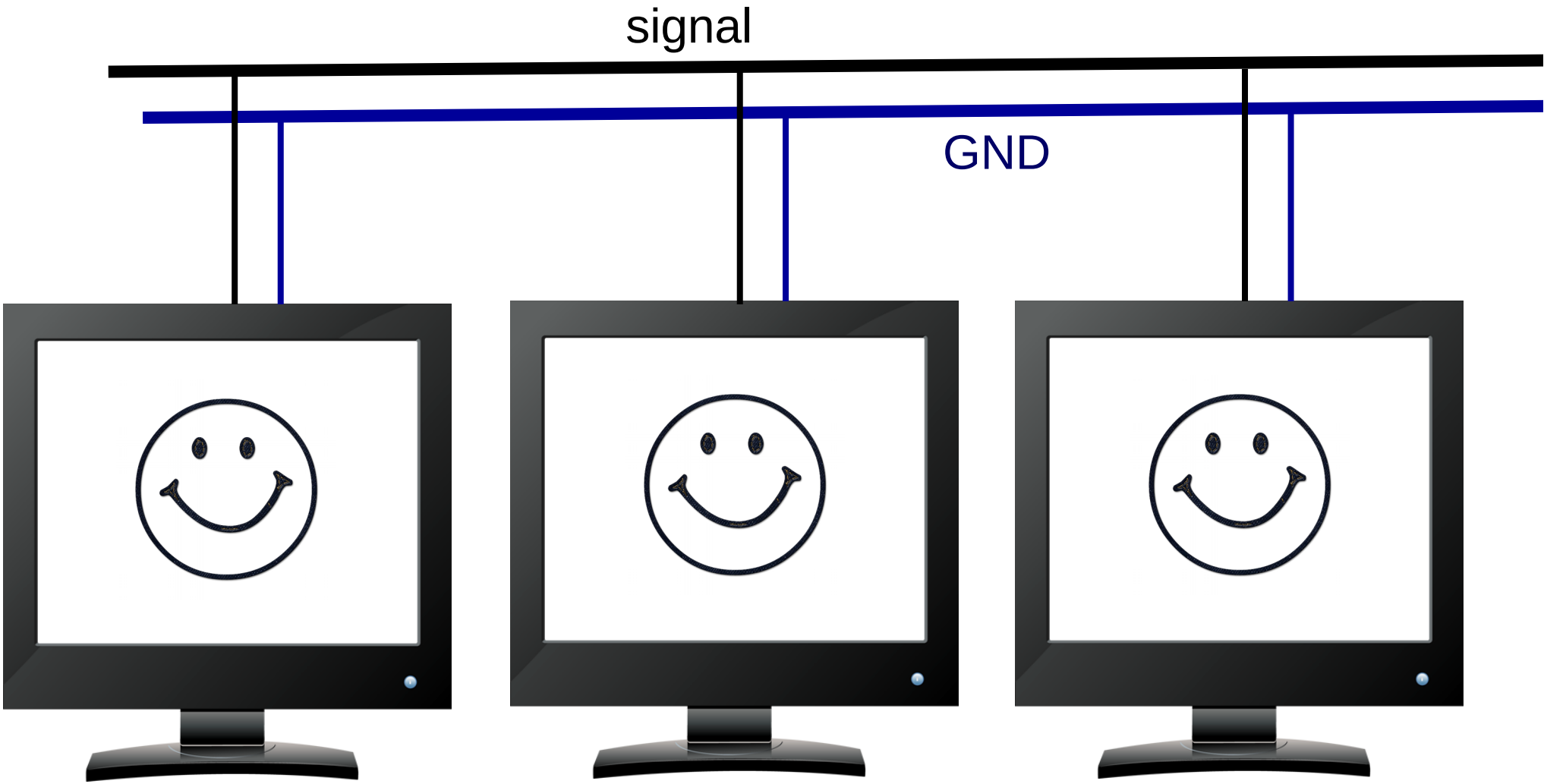


Ethernet

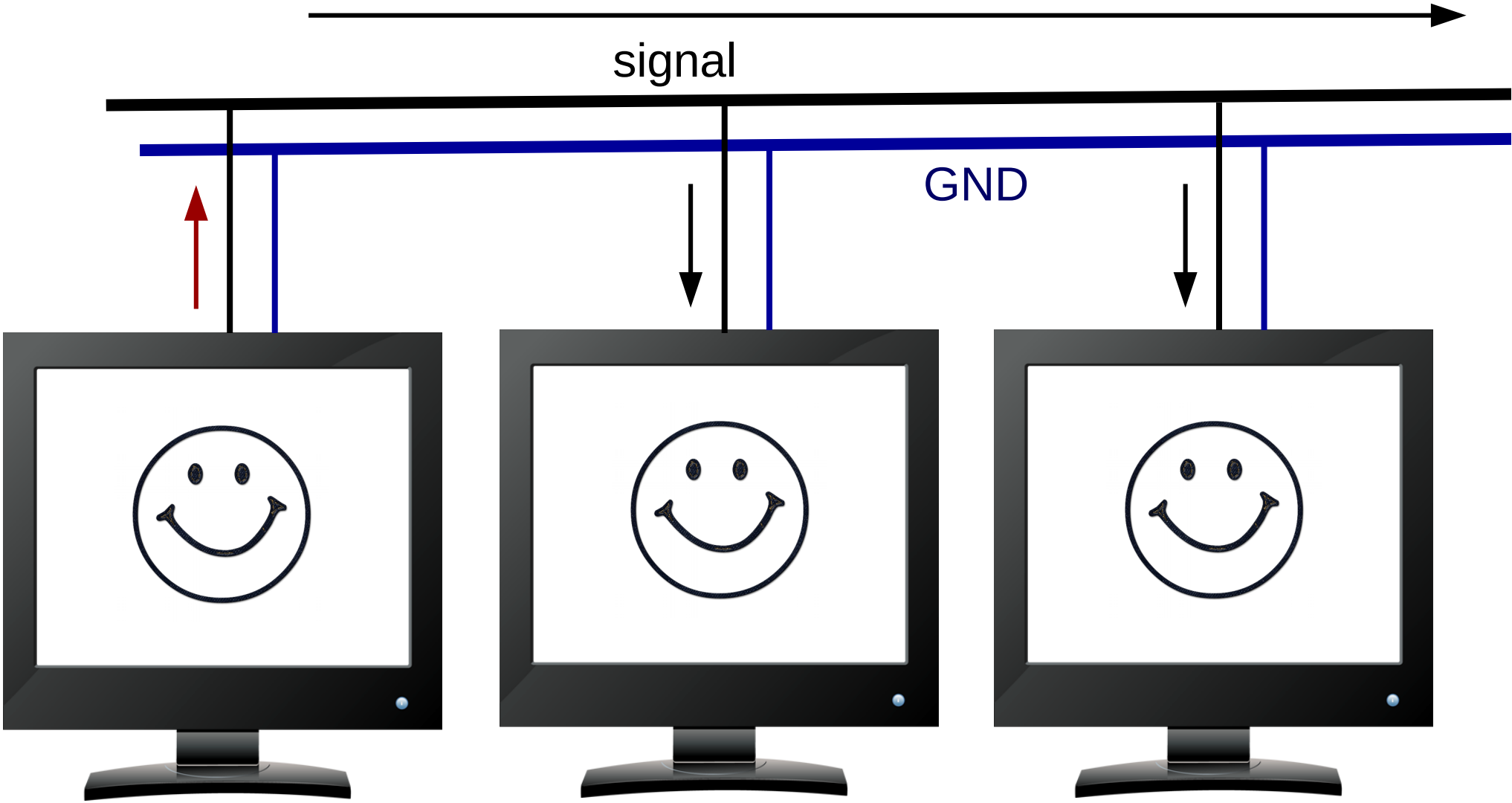



signal
GND

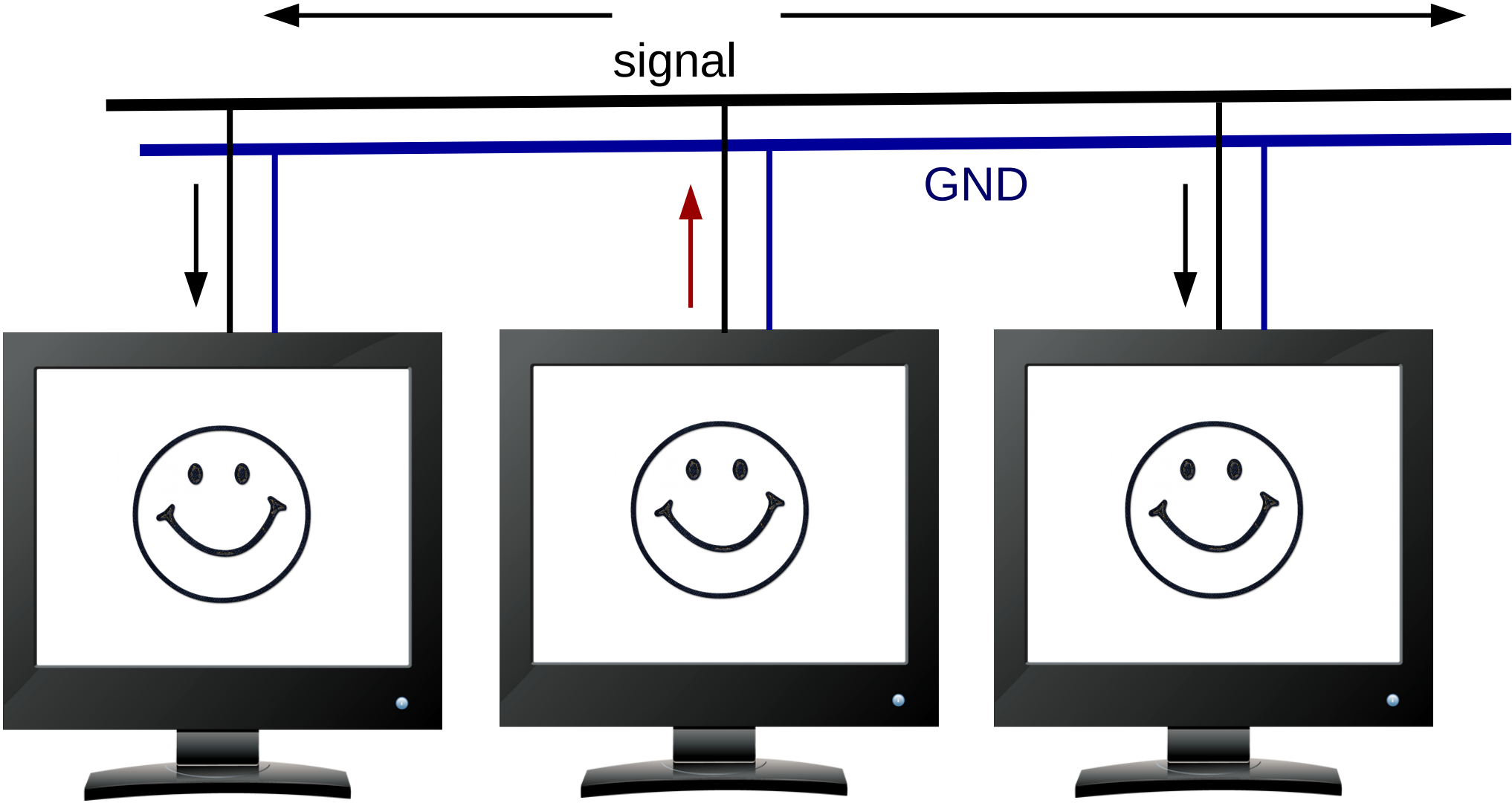
Ethernet



Ethernet

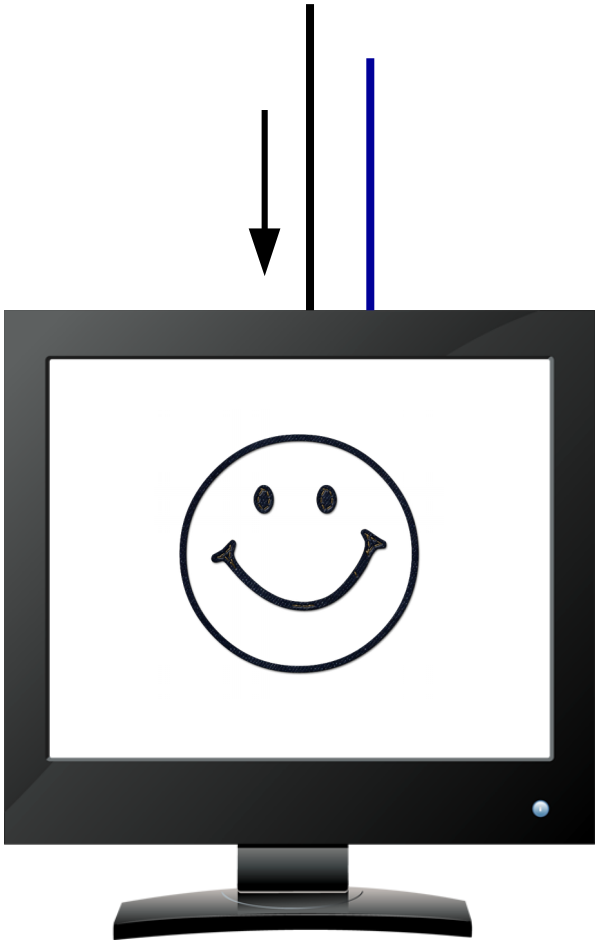


Ethernet



Ethernet

- Мне ли пришёл данный кадр?
- Кому слать ответ?



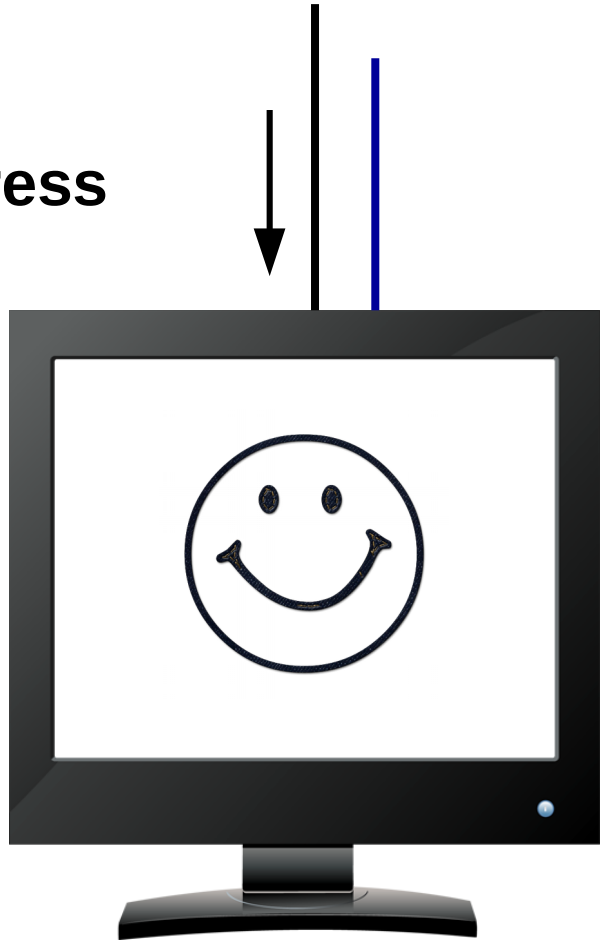
Ethernet

- Мне ли пришёл данный кадр?
- Кому слать ответ?

Destination MAC address

Source MAC address

Media Access Control

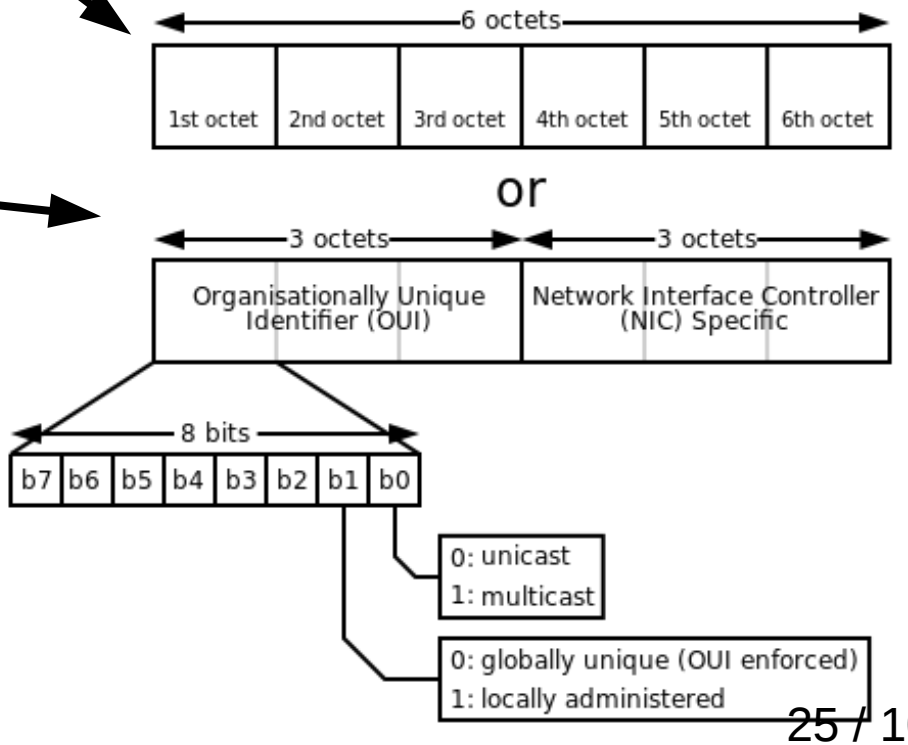


Ethernet

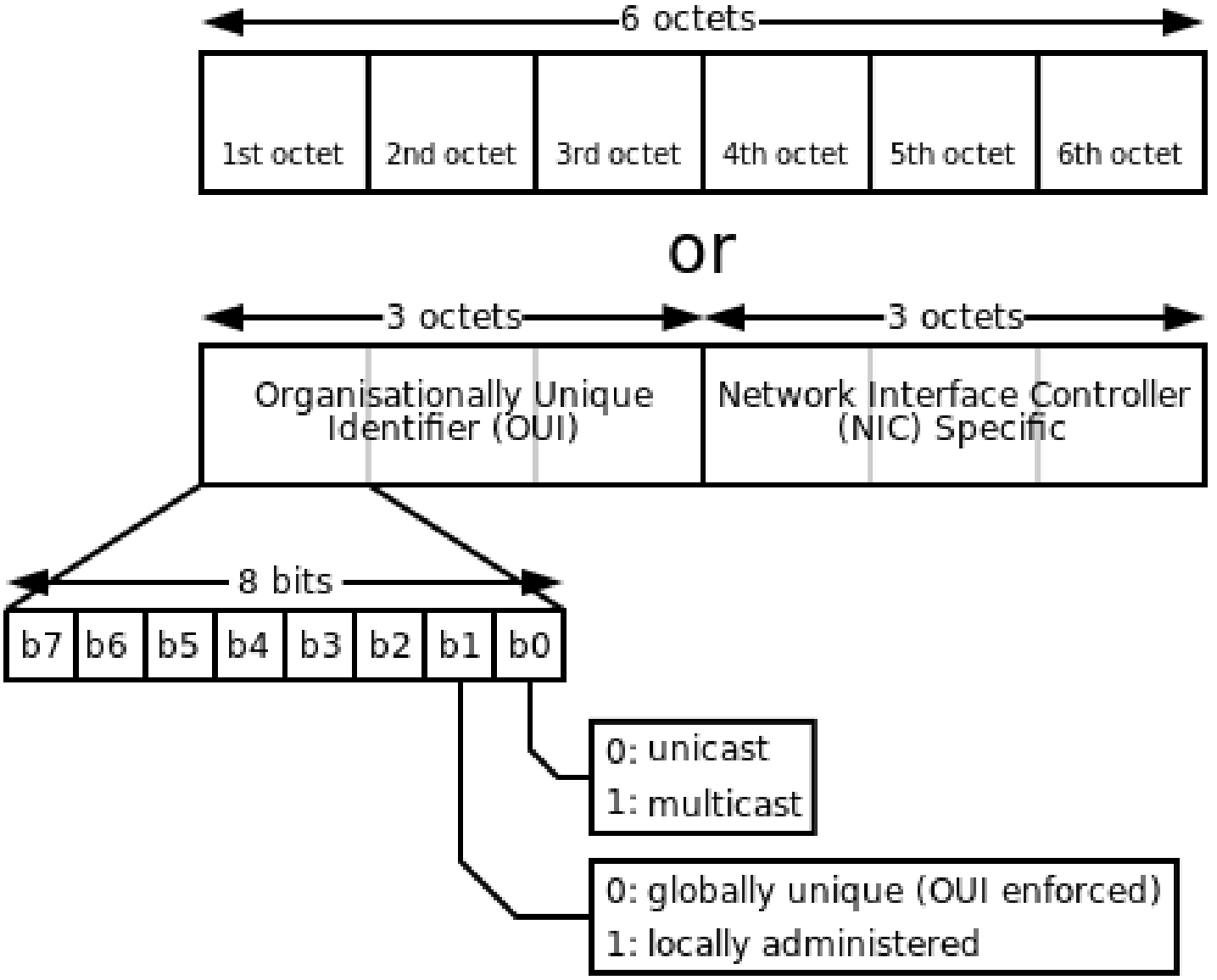
- Мне ли пришёл данный кадр?
- Кому слать ответ?

Destination MAC address

Source MAC address

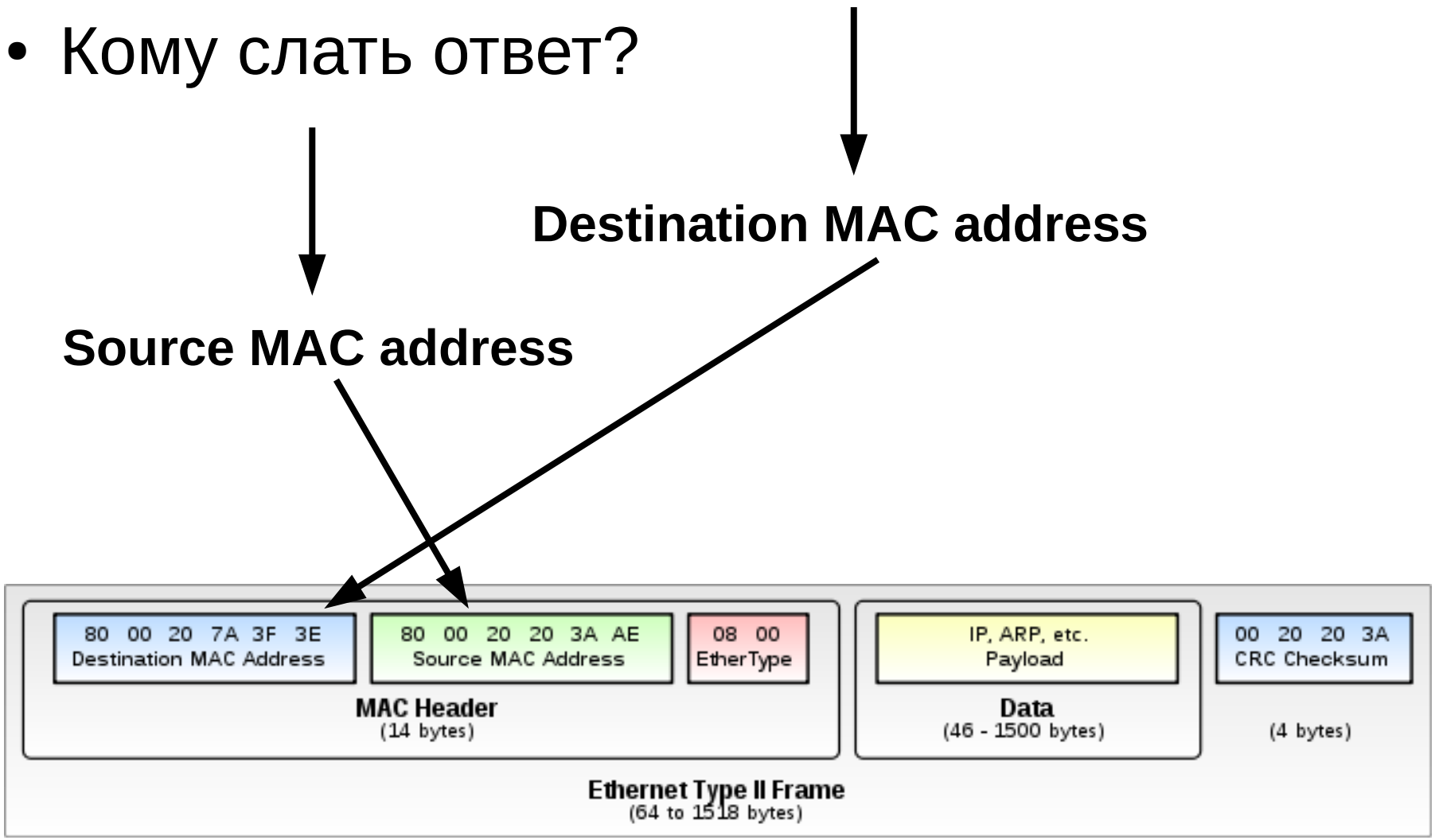


Ethernet

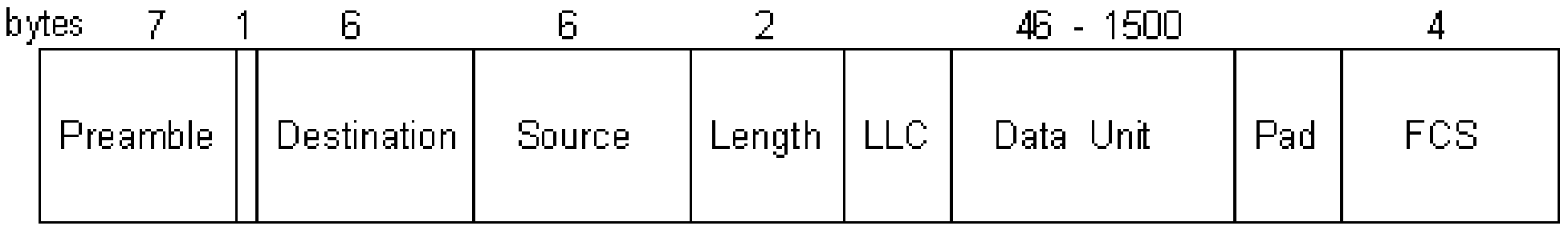

6 octets
1st octet
2nd octet
3rd octet
4th octet
5th octet
6th octet
or
3 octets
3 octets
Organisationally Unique Identifier (OUI)
Network Interface Controller (NIC) Specific
8 bits
b7
b6
b5
b4
b3
b2
b1
b0
0: unicast
1: multicast
0: globally unique (OUI enforced)
1: locally administered

Ethernet

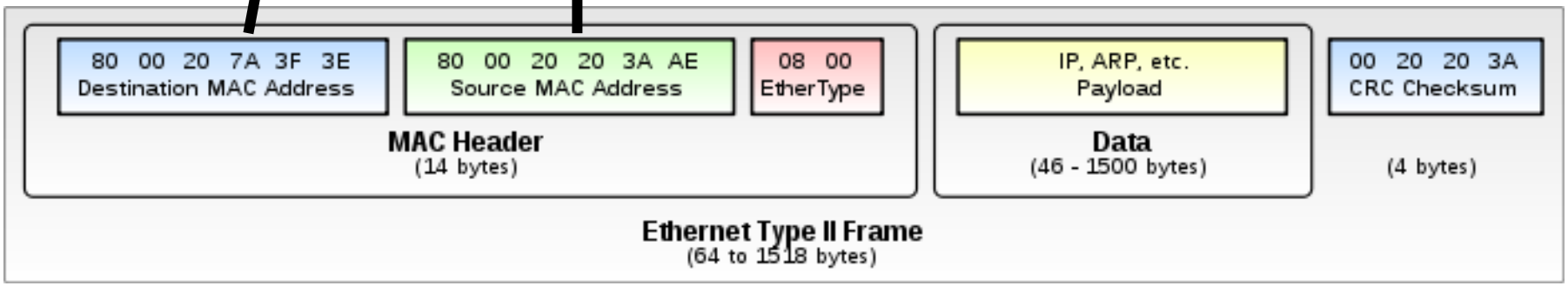
- Мне ли пришёл данный кадр?
- Кому слать ответ?



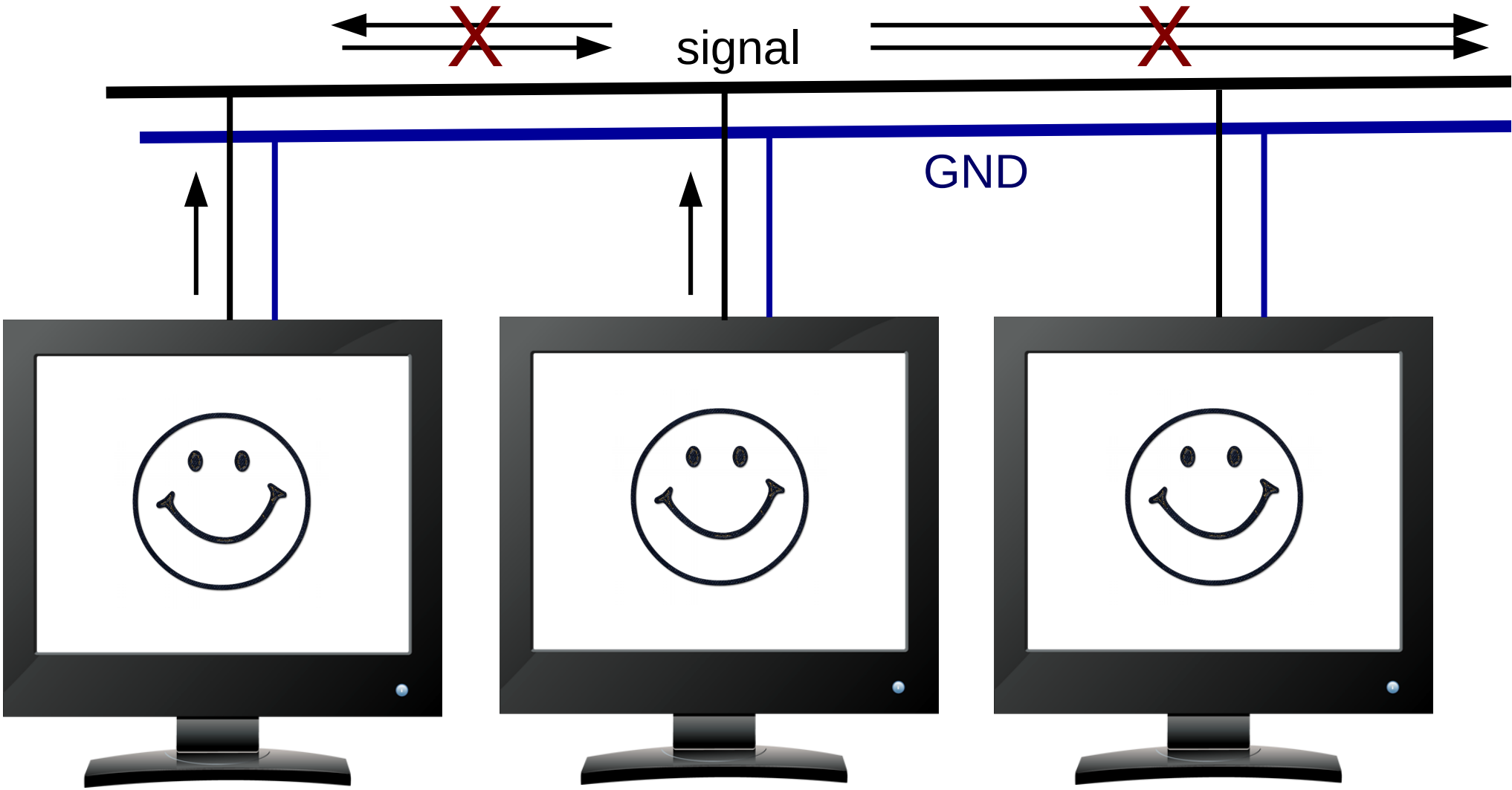
Ethernet



IEEE 802.3 Frame



Ethernet



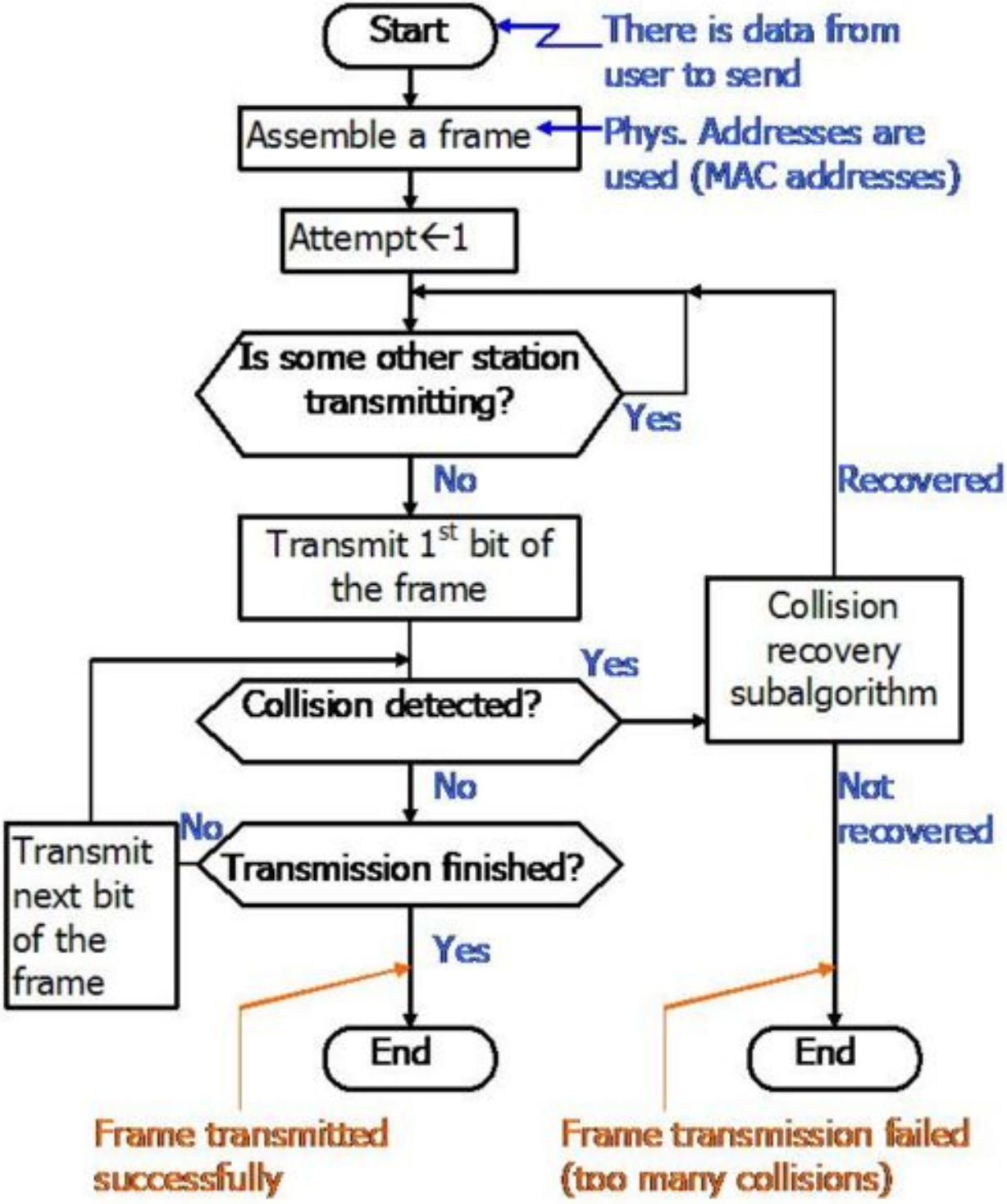
Ethernet

Большая сеть

Ethernet

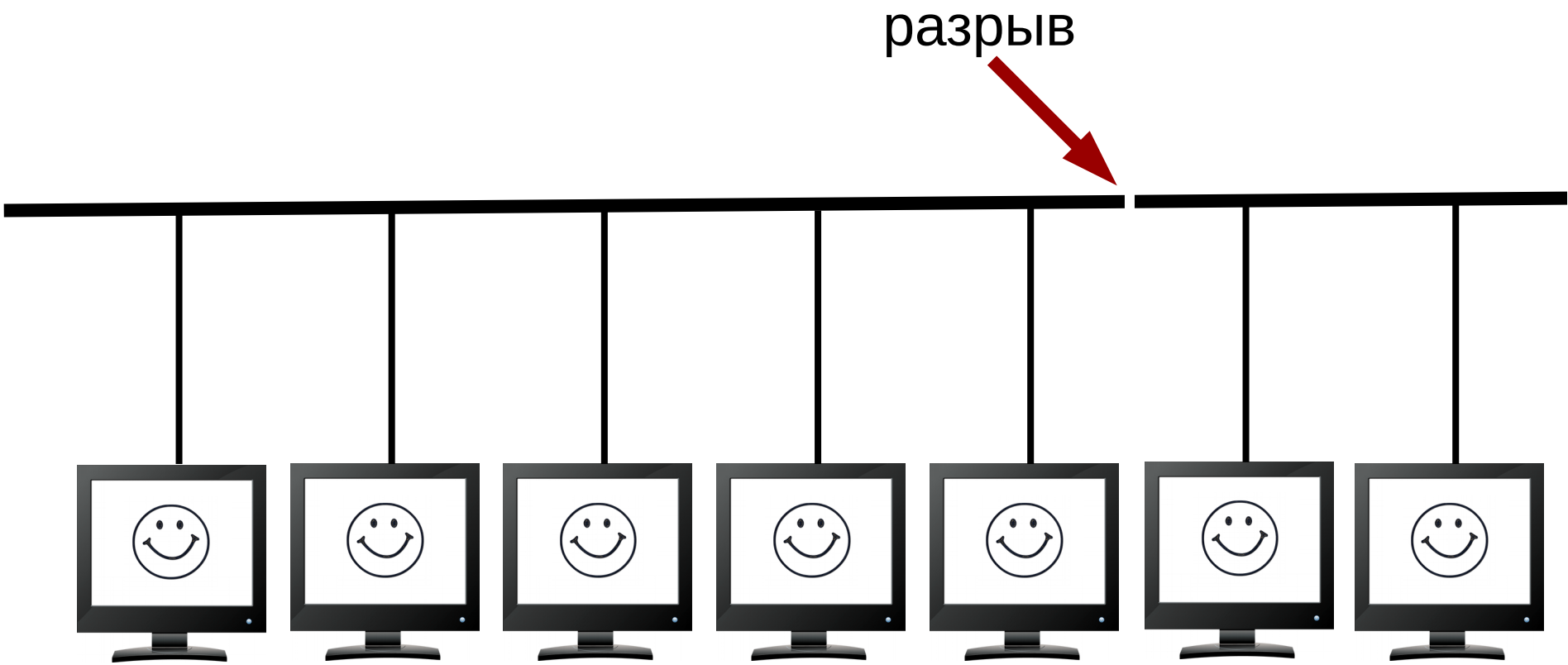
Ethernet

общая шина, коллизии


Start
There is data from user to send
Assemble a frame
Phys. Addresses are used (MAC addresses)
Attempt←1
Is some other station transmitting?
Yes
No
Recovered
Transmit 1st bit of the frame
Collision recovery subalgorithm
Yes
Collision detected?
No
Not recovered
No
Transmit next bit of the frame
Transmission finished?
Yes
End
End
Frame transmitted successfully
Frame transmission failed (too many collisions)

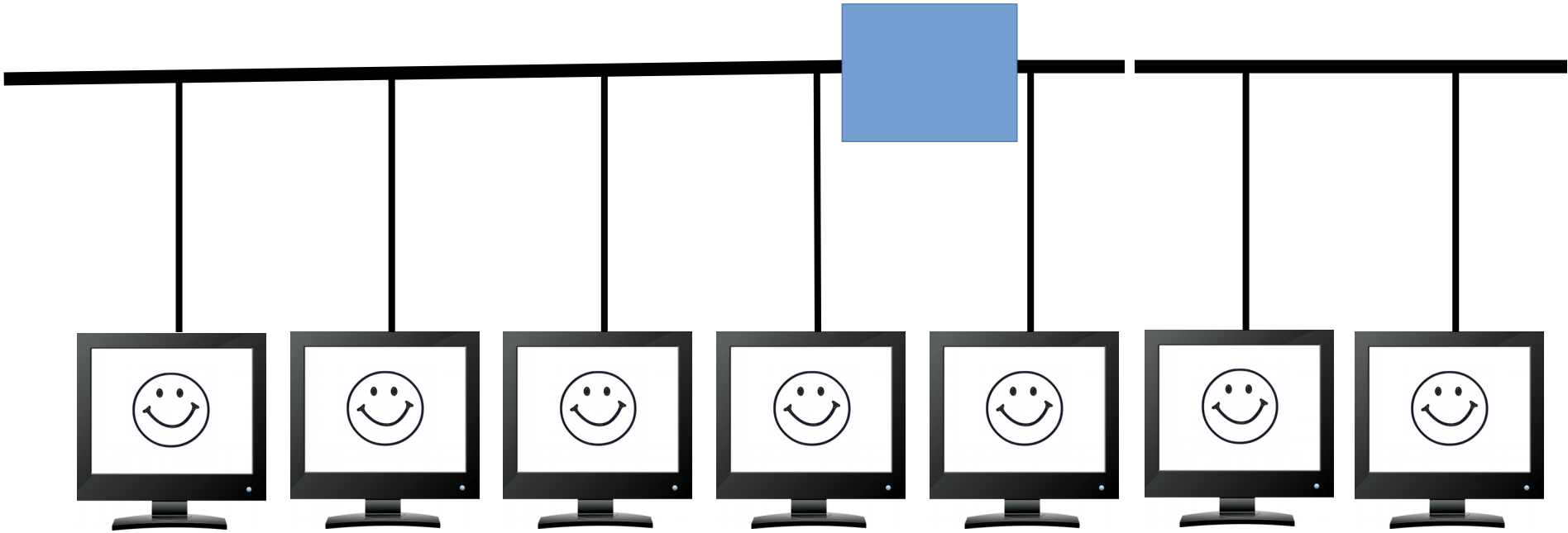
Ethernet

Ethernet



Ethernet

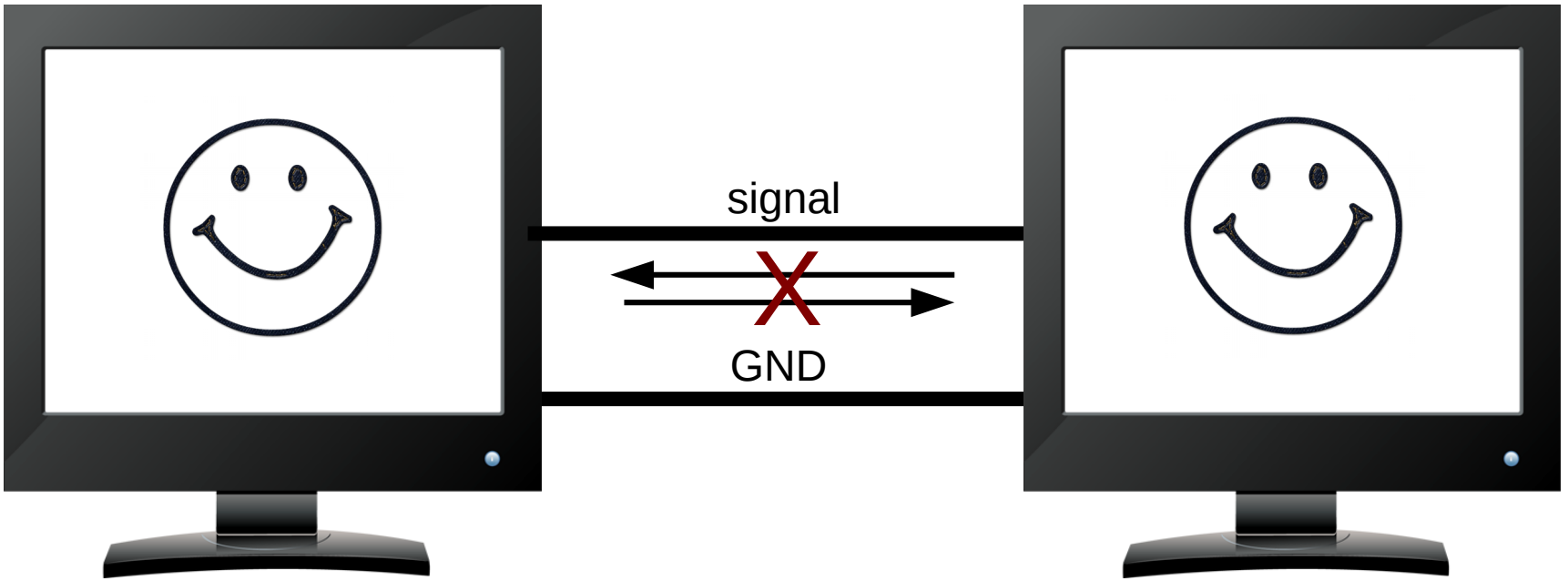
repeater



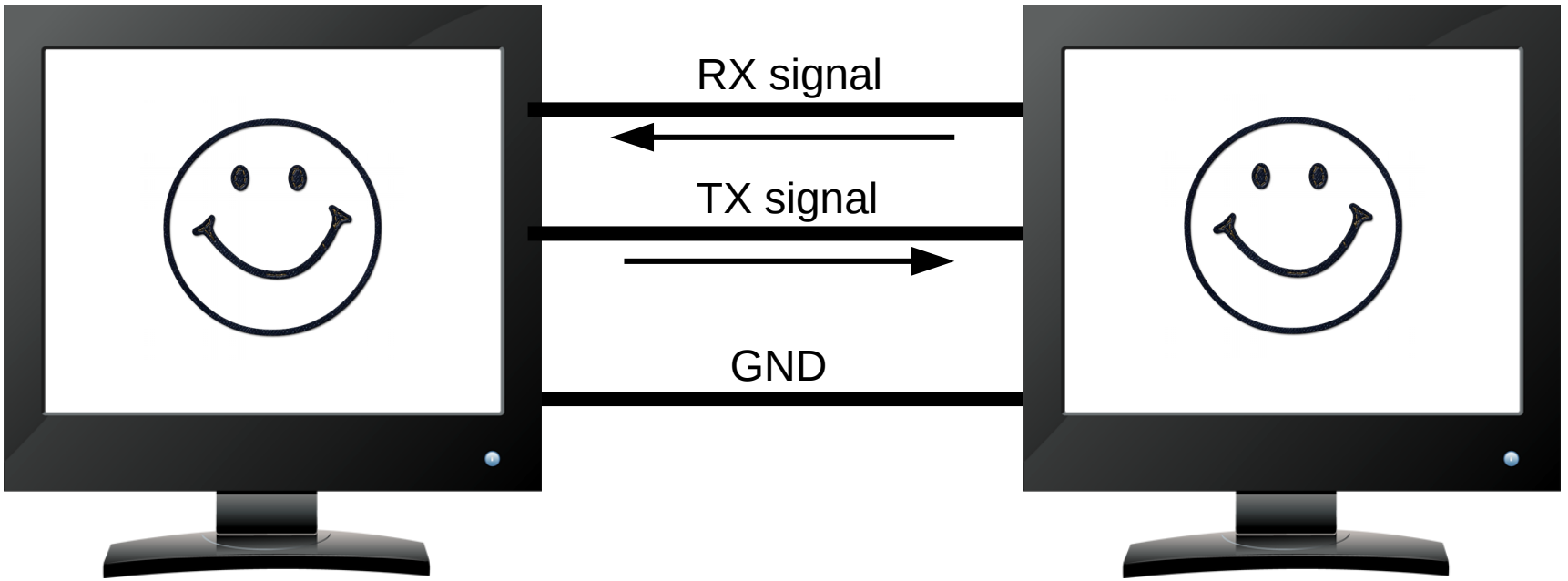
Ethernet

- Плохая масштабируемость
- Низкая надёжность
- Коллизии
- Низкая скорость

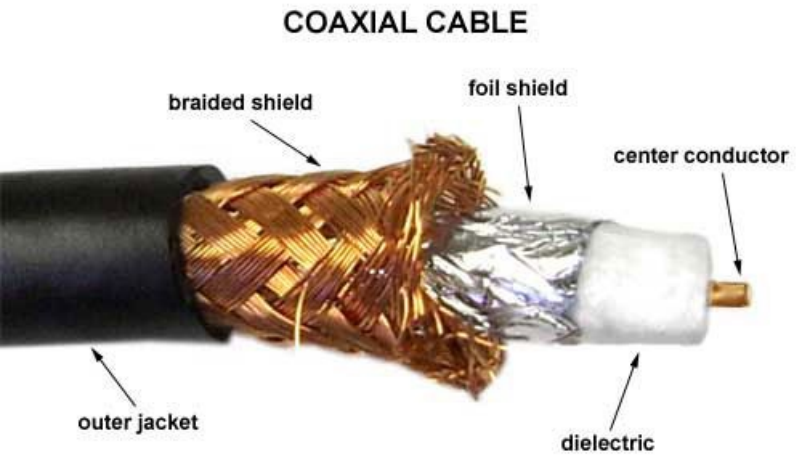
Ethernet



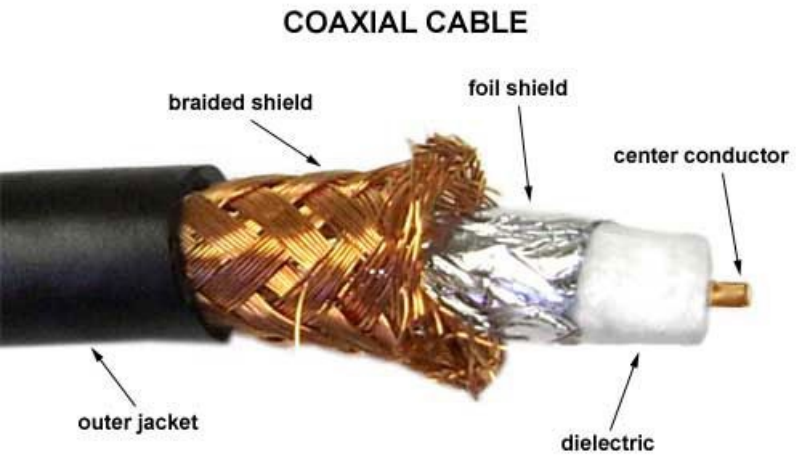
Ethernet



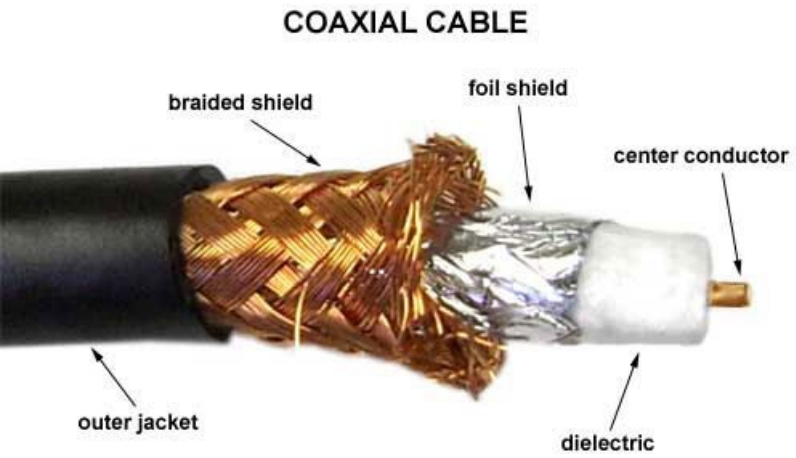
Ethernet


COAXIAL CABLE
braided shield
foil shield
center conductor
outer jacket
dielectric

Ethernet

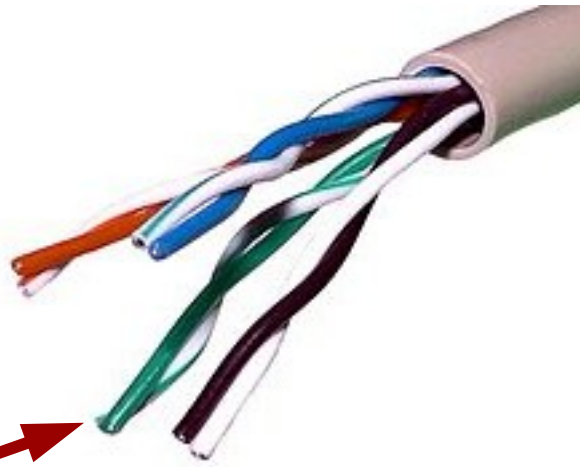

COAXIAL CABLE
braided shield
foil shield
center conductor
outer jacket
dielectric

Ethernet

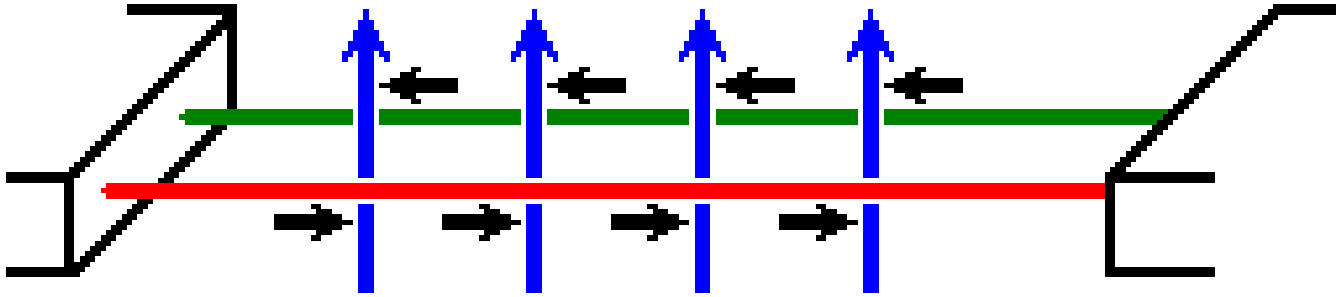


RX

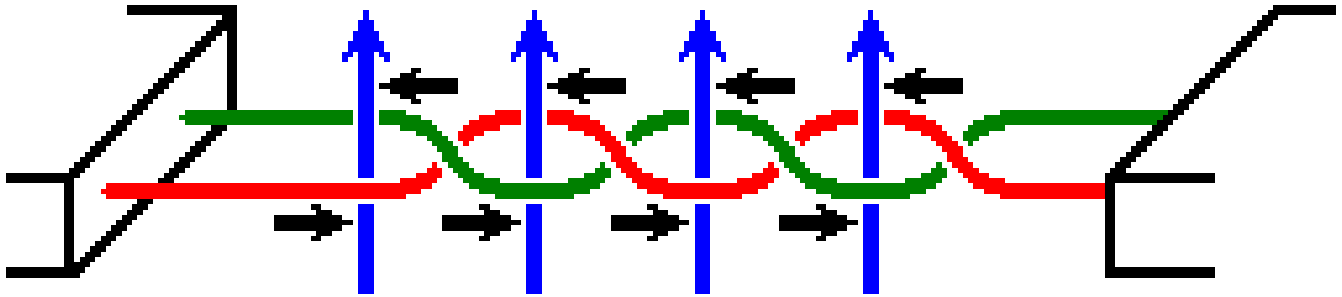
TX



Ethernet



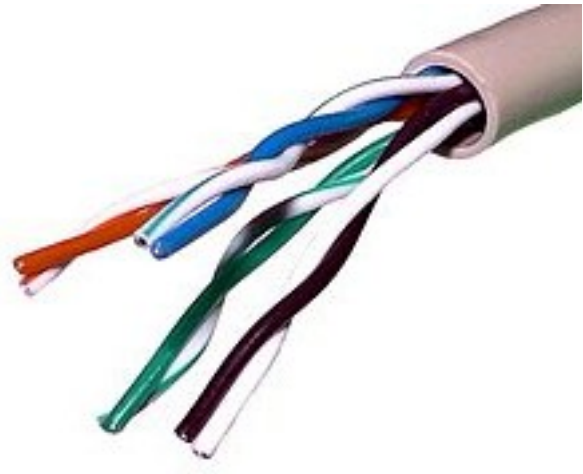
Straight cable



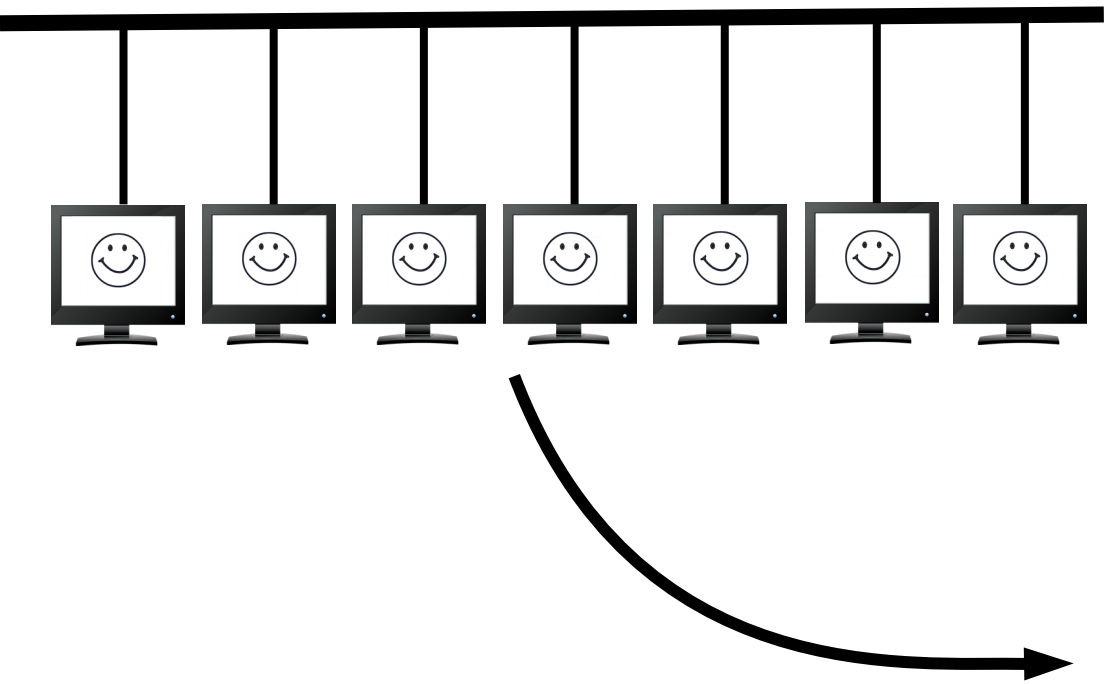
Twisted pair cable

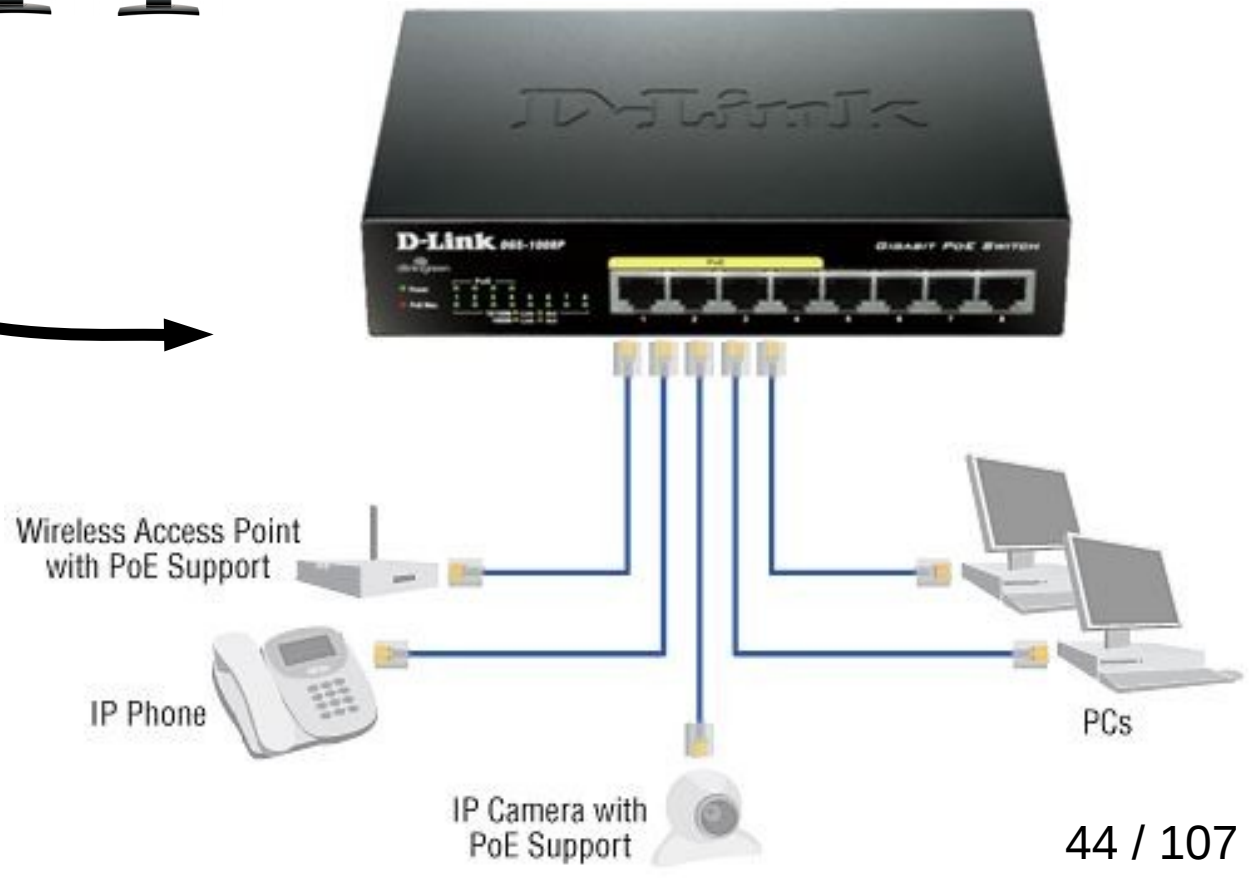
Magnetic field

Induced noise current



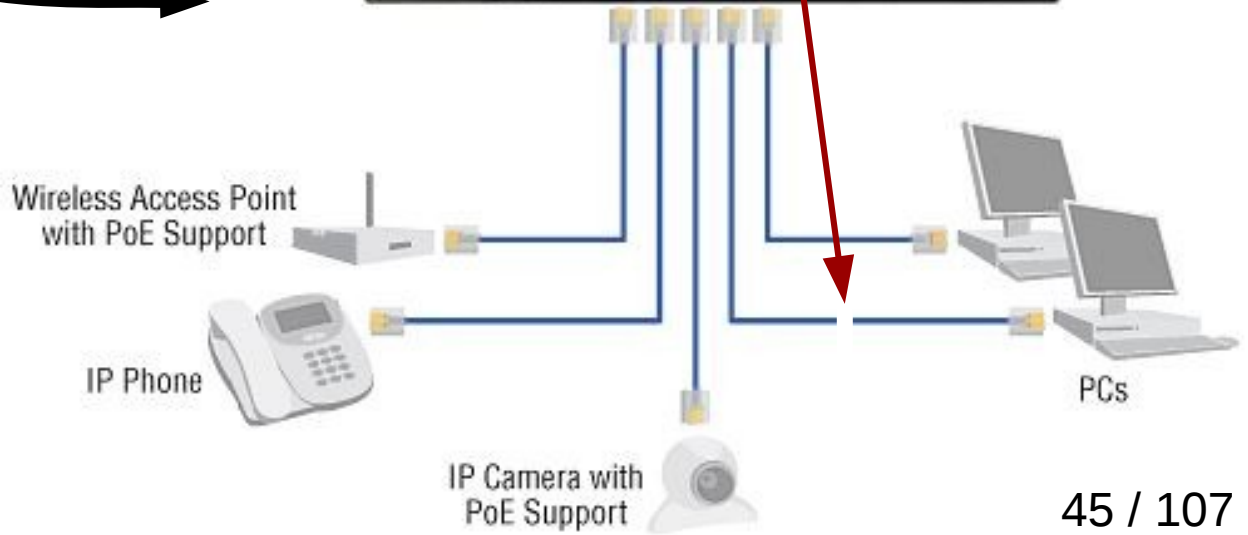
Ethernet




Wireless Access Point
with PoE Support
IP Phone
IP Camera with
PoE Support
PCs

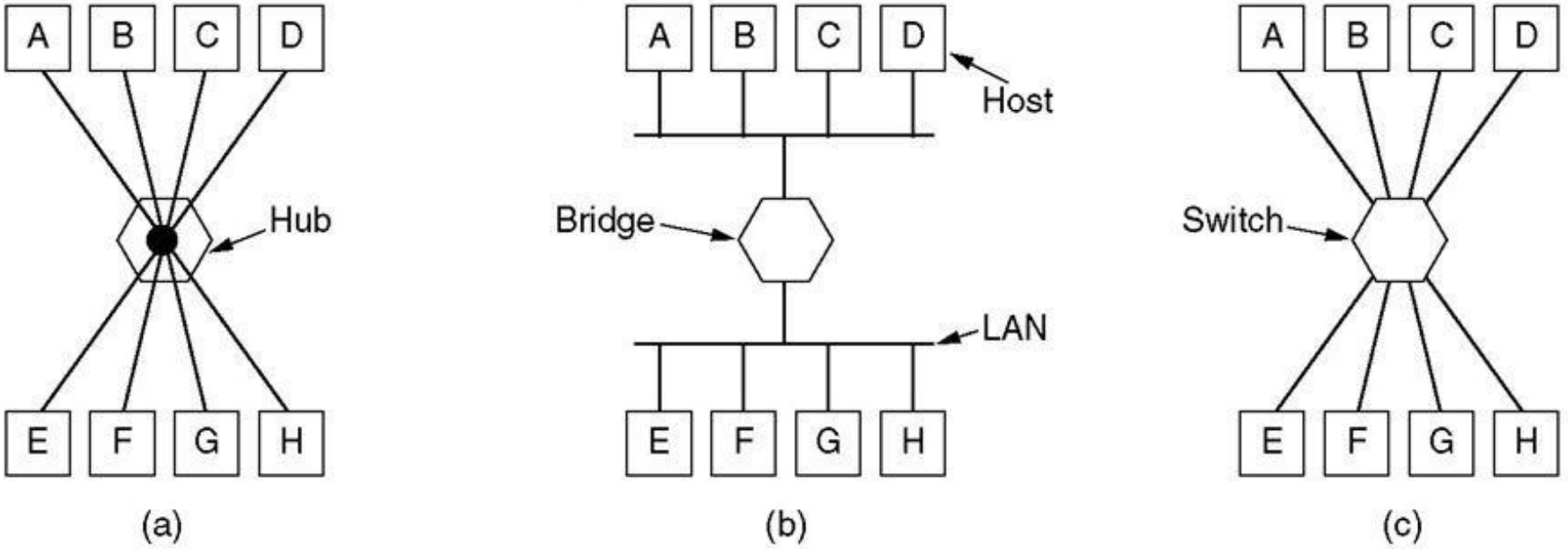
Ethernet

Разрыв, но сеть работоспособна



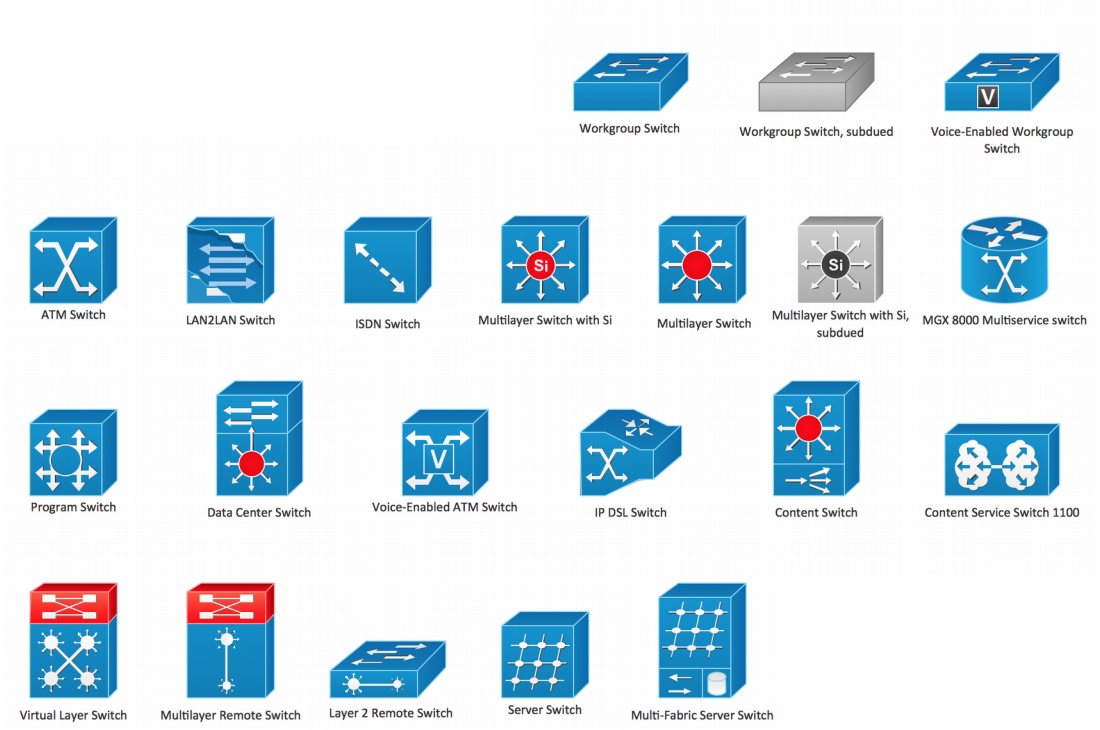
Ethernet

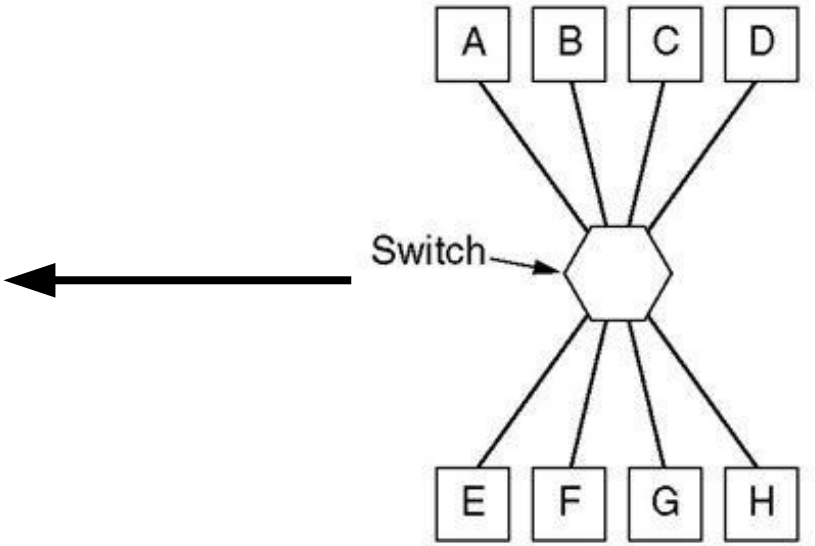
Ethernet



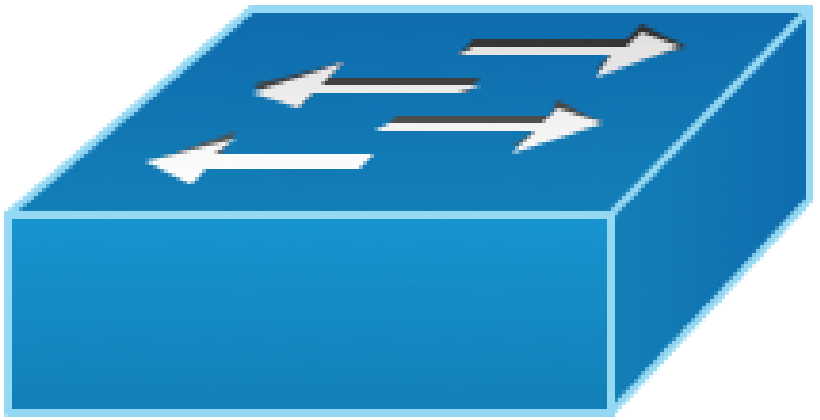
(a) Концентратор (b) Мост (c) Коммутатор

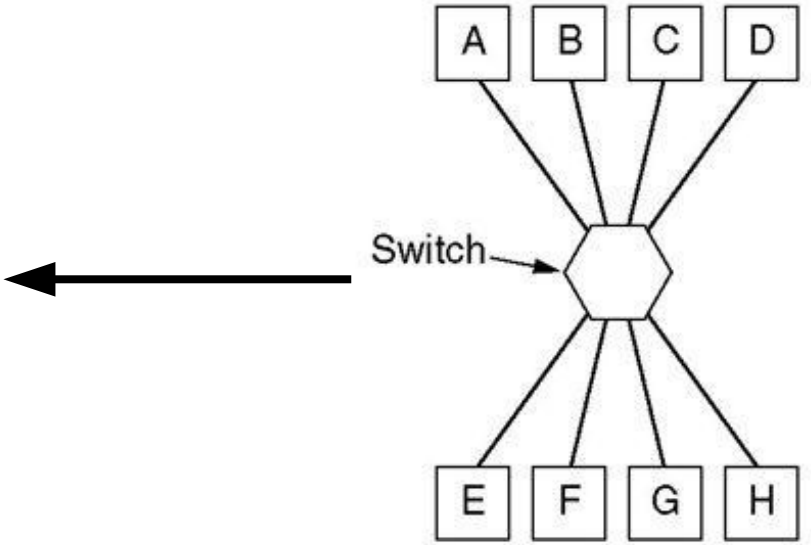
Ethernet

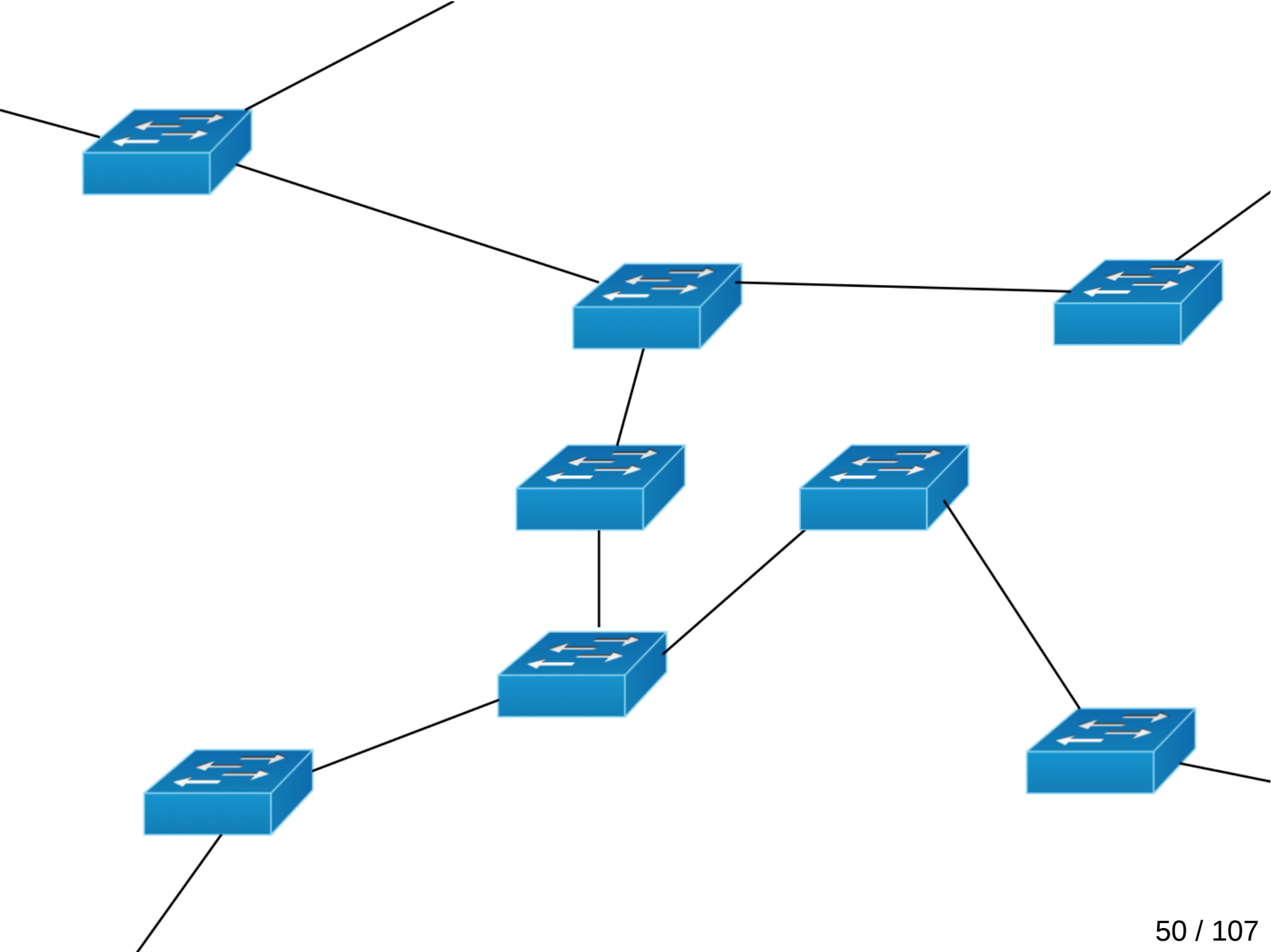

Workgroup Switch
Workgroup Switch, subdued
Voice-Enabled Workgroup Switch
ATM Switch
LAN2LAN Switch
ISDN Switch
Multilayer Switch with Si
Multilayer Switch
Multilayer Switch with Si, subdued
MGX 8000 Multiservice switch
Program Switch
Data Center Switch
Voice-Enabled ATM Switch
IP DSL Switch
Content Switch
Content Service Switch 1100
Virtual Layer Switch
Multilayer Remote Switch
Layer 2 Remote Switch
Server Switch
Multi-Fabric Server Switch

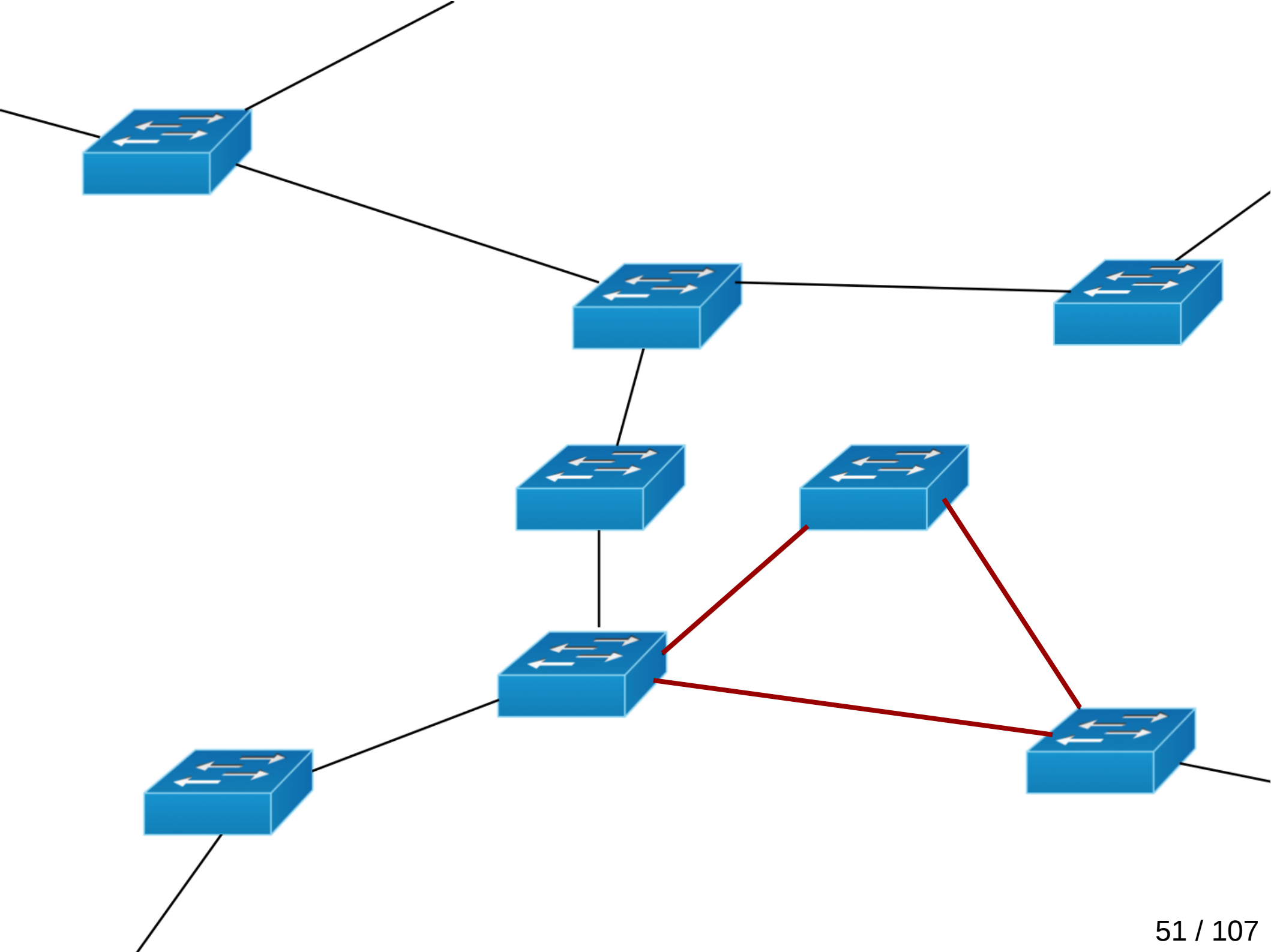

A
B
C
D
Switch
E
F
G
H

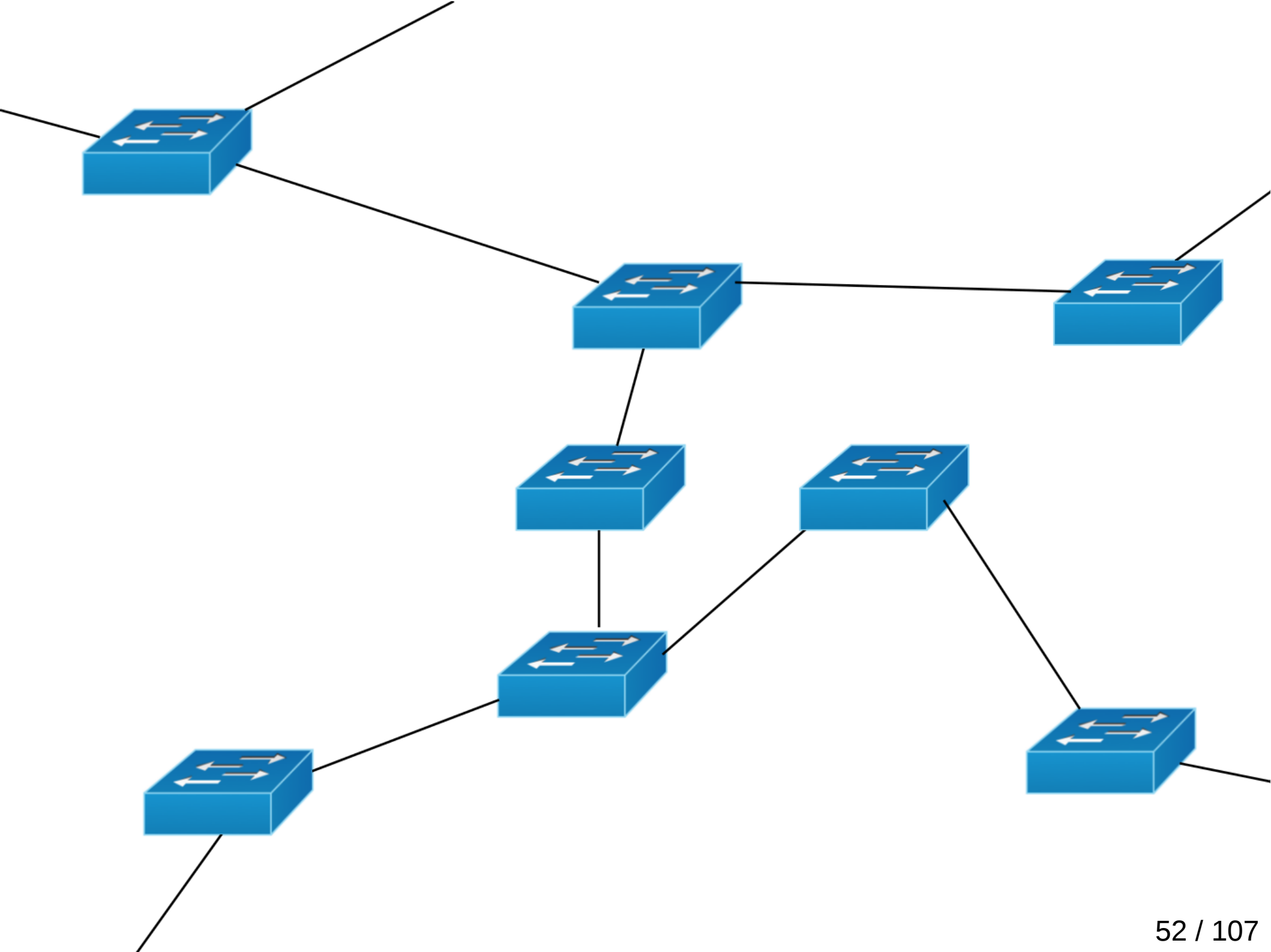
Ethernet







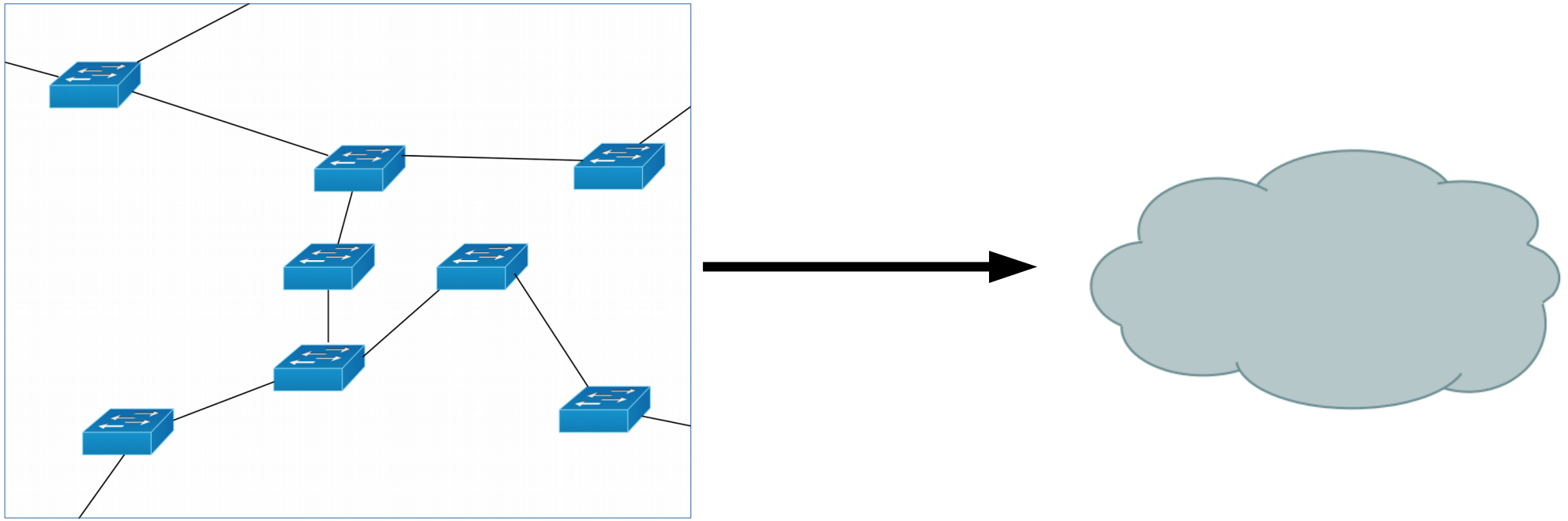


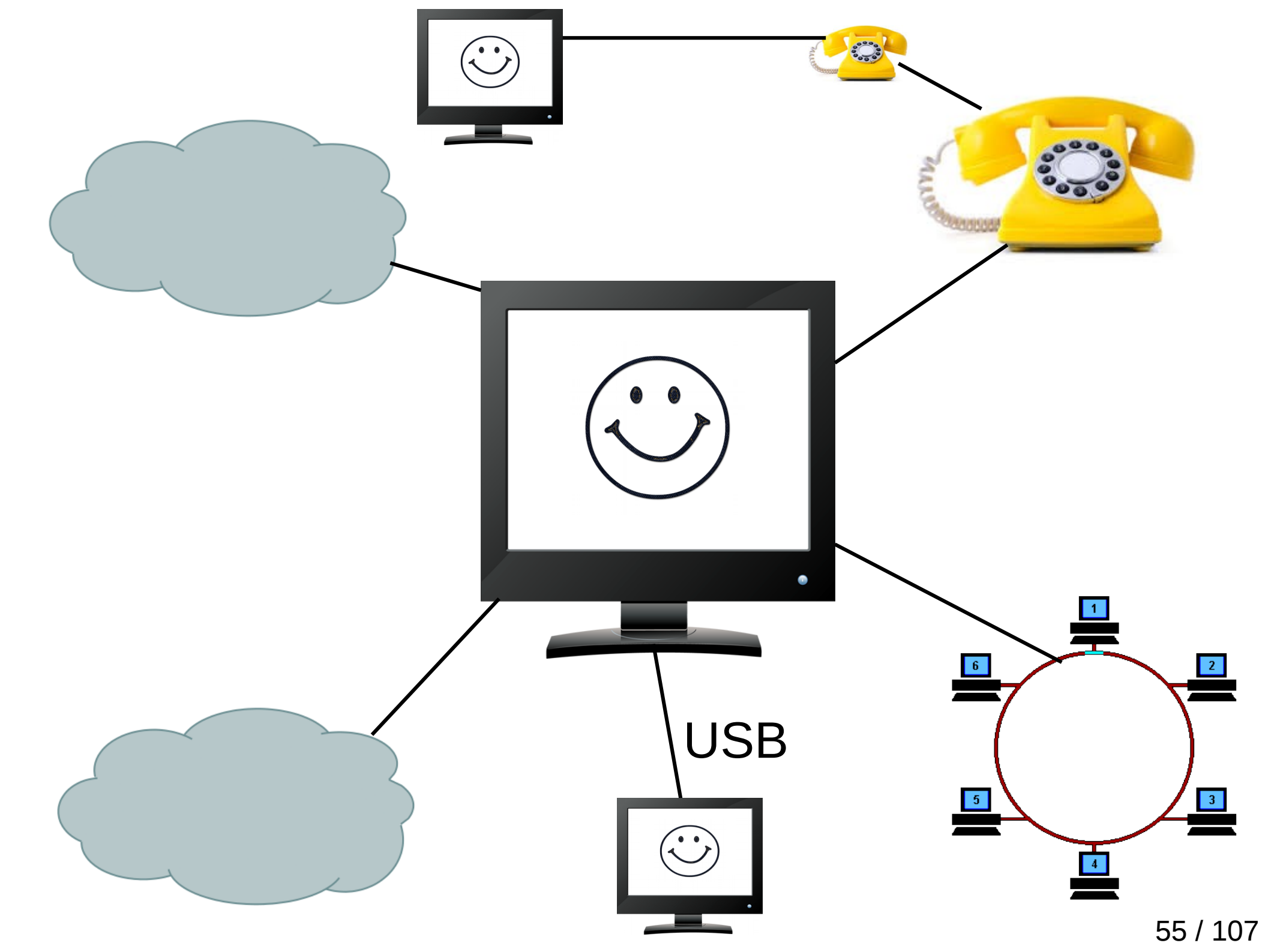


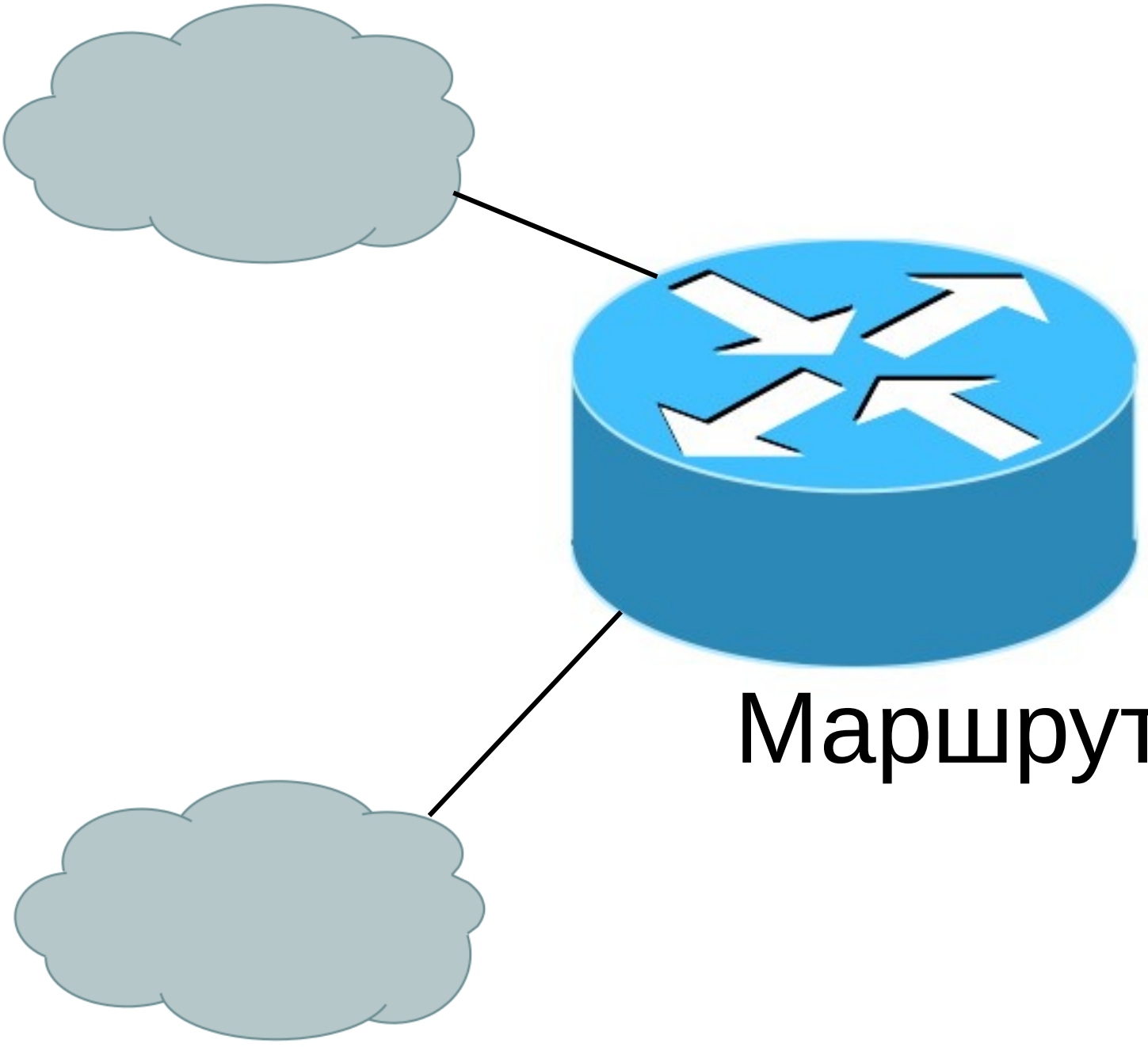
Ethernet

- Не универсально
- Плохо масштабируемо

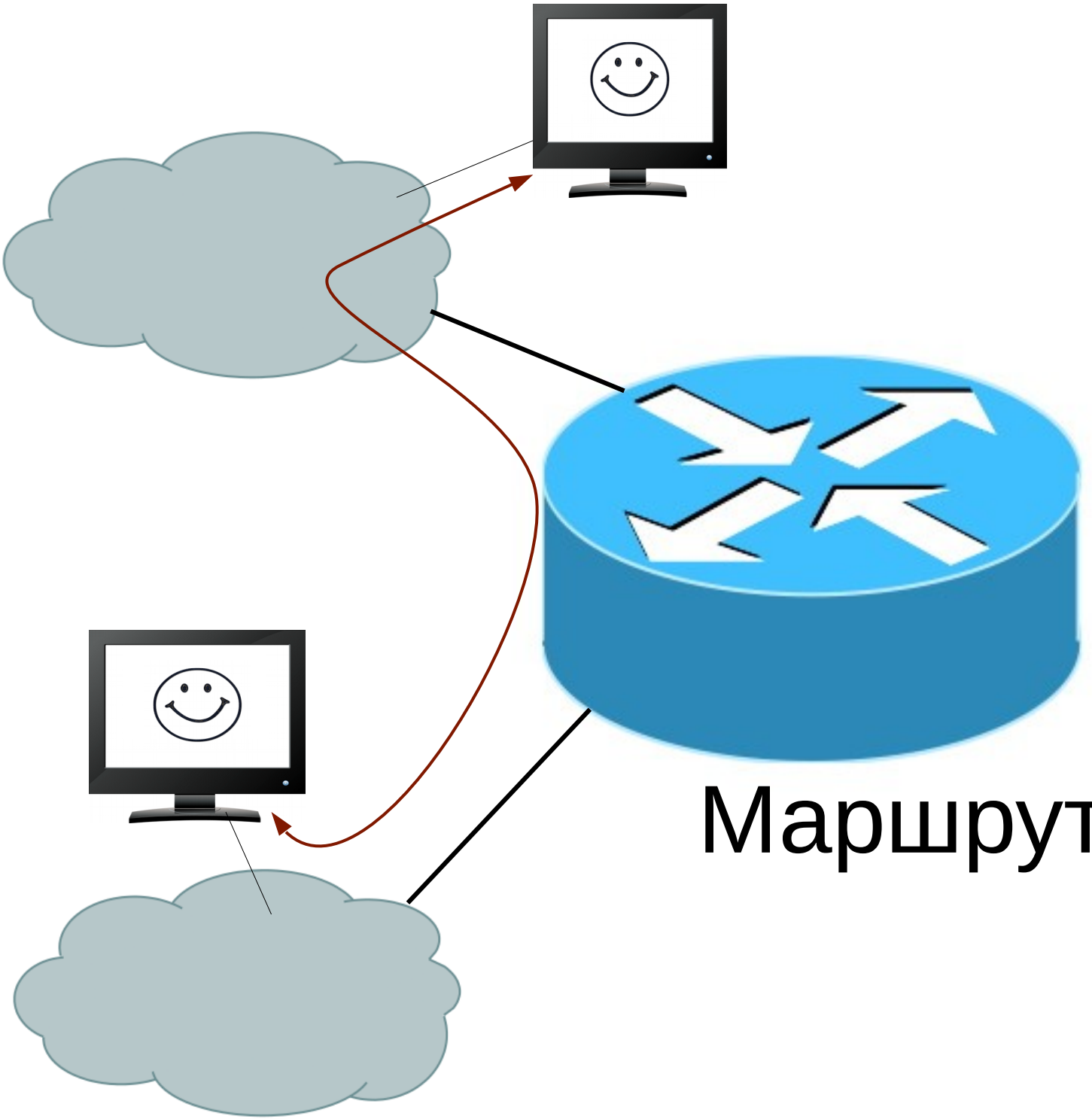
Ethernet



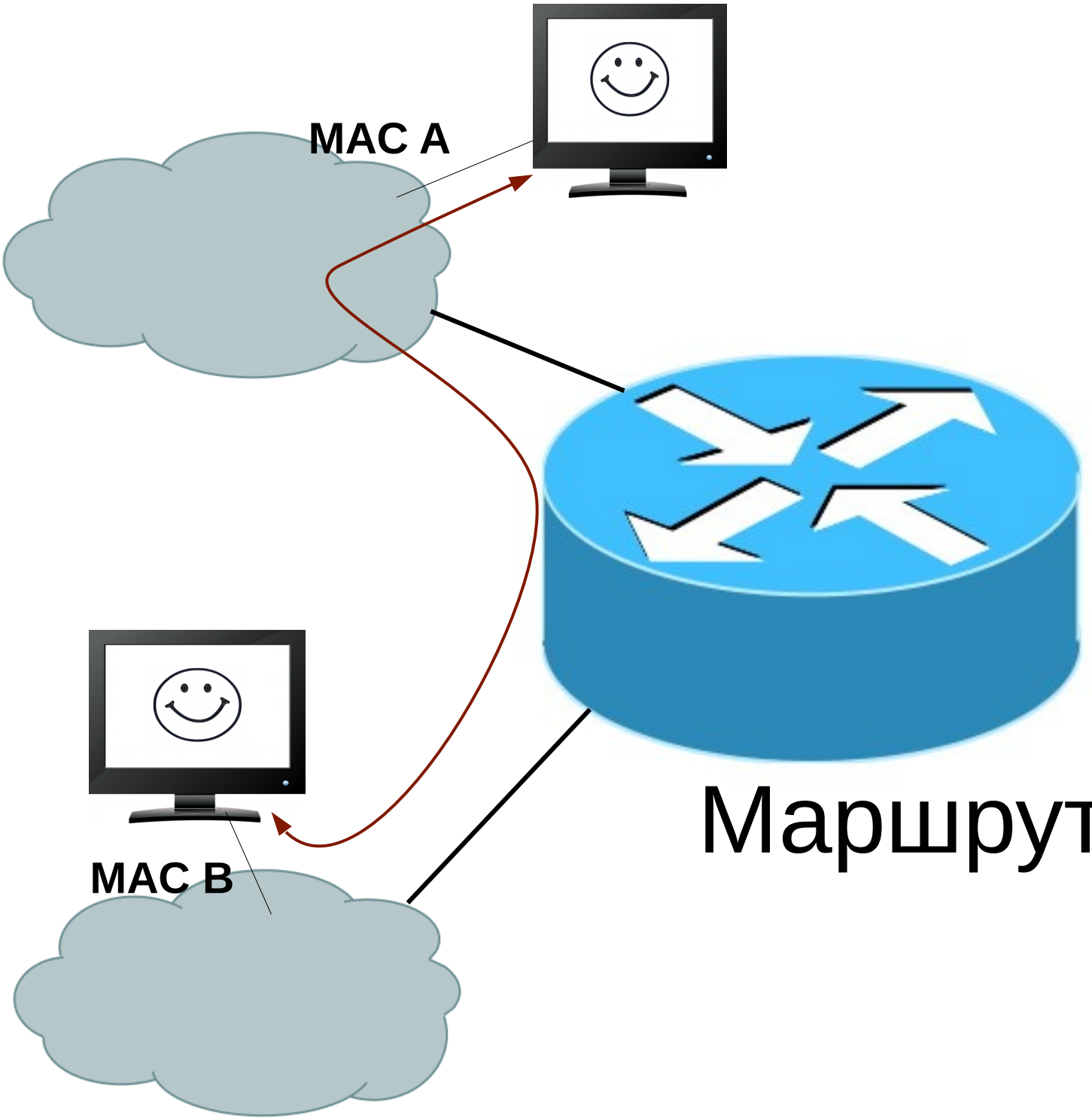

USB
1
2
3
4
5
6



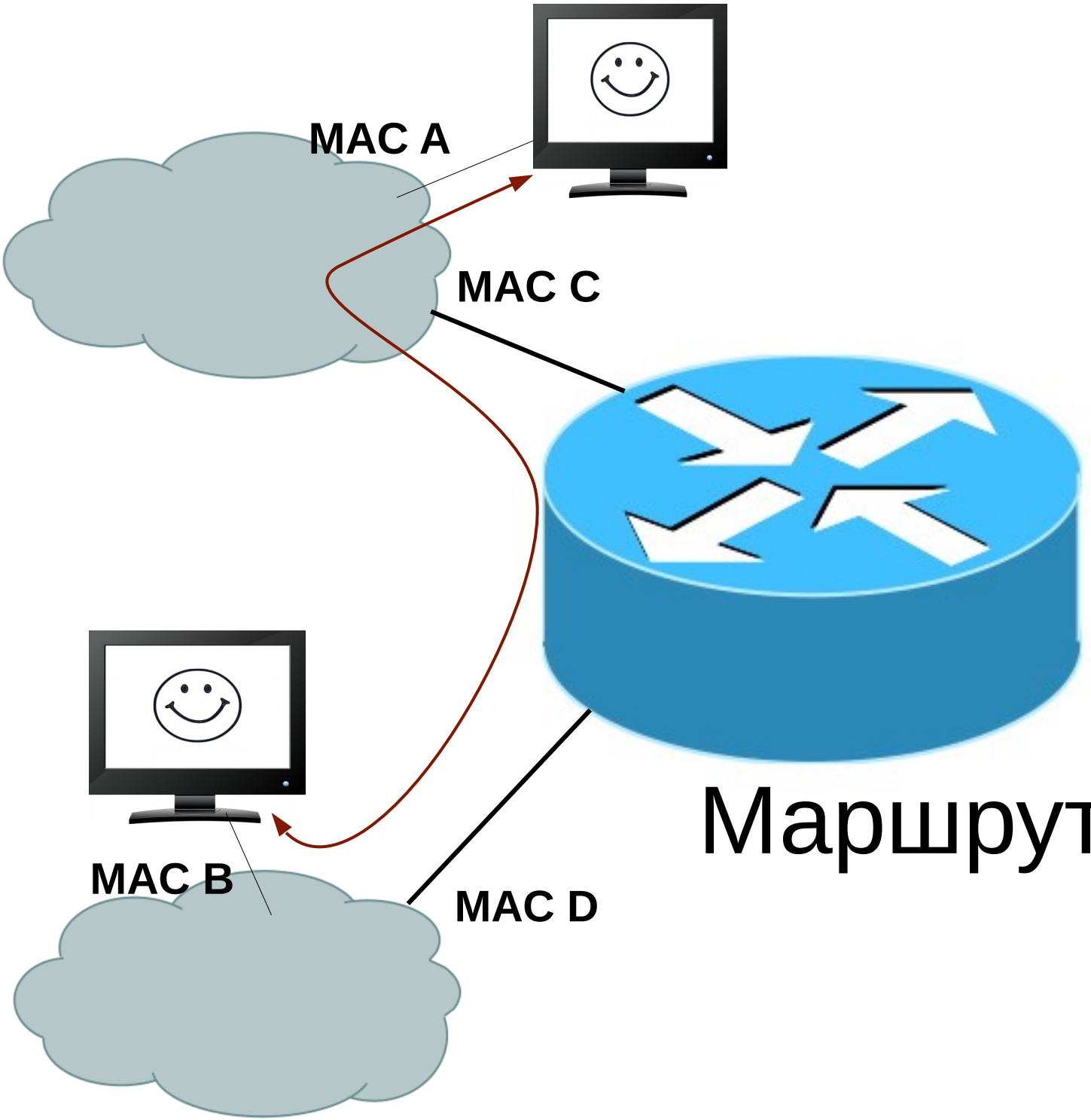
Маршрутизатор



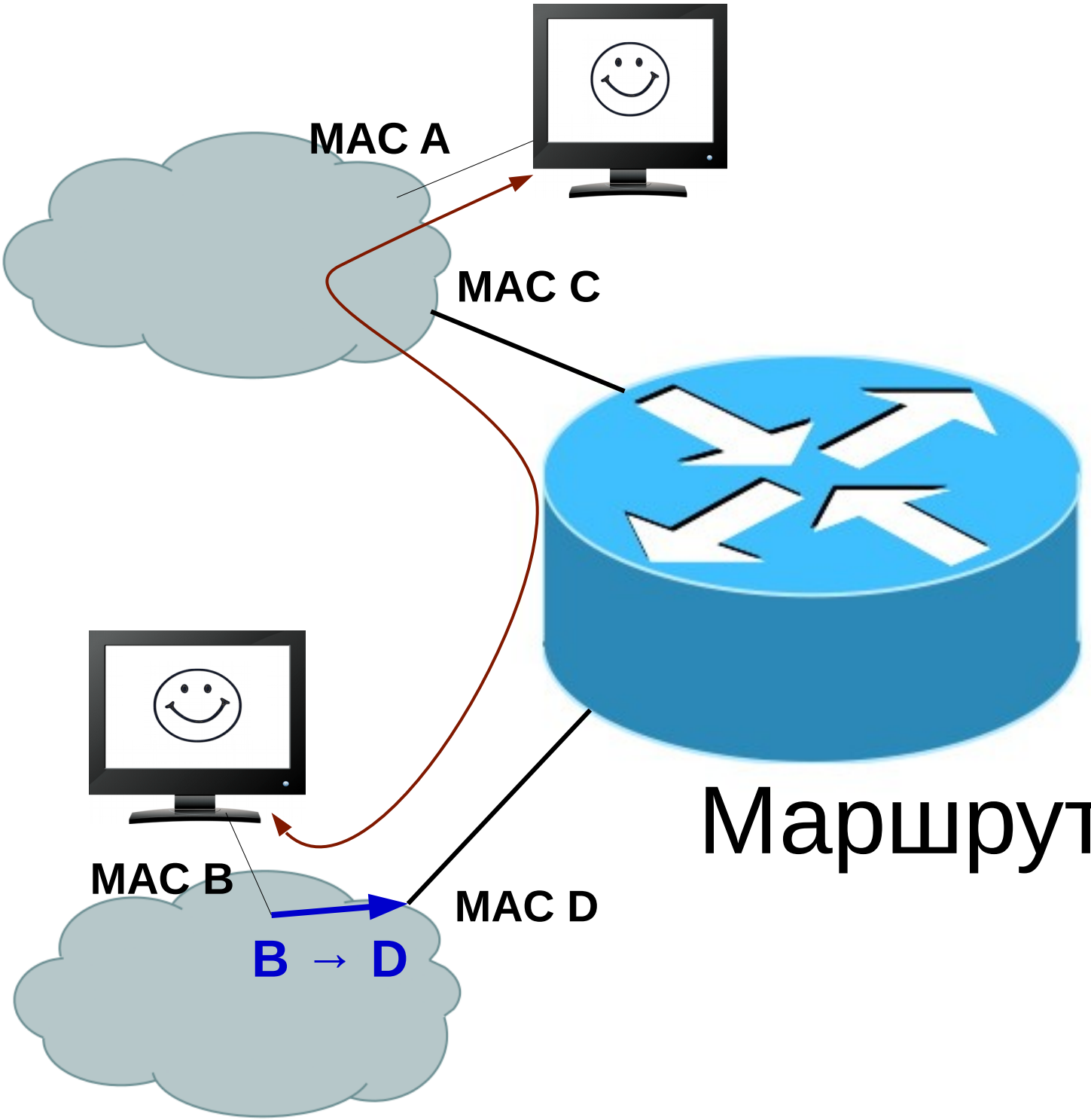
Маршрутизатор



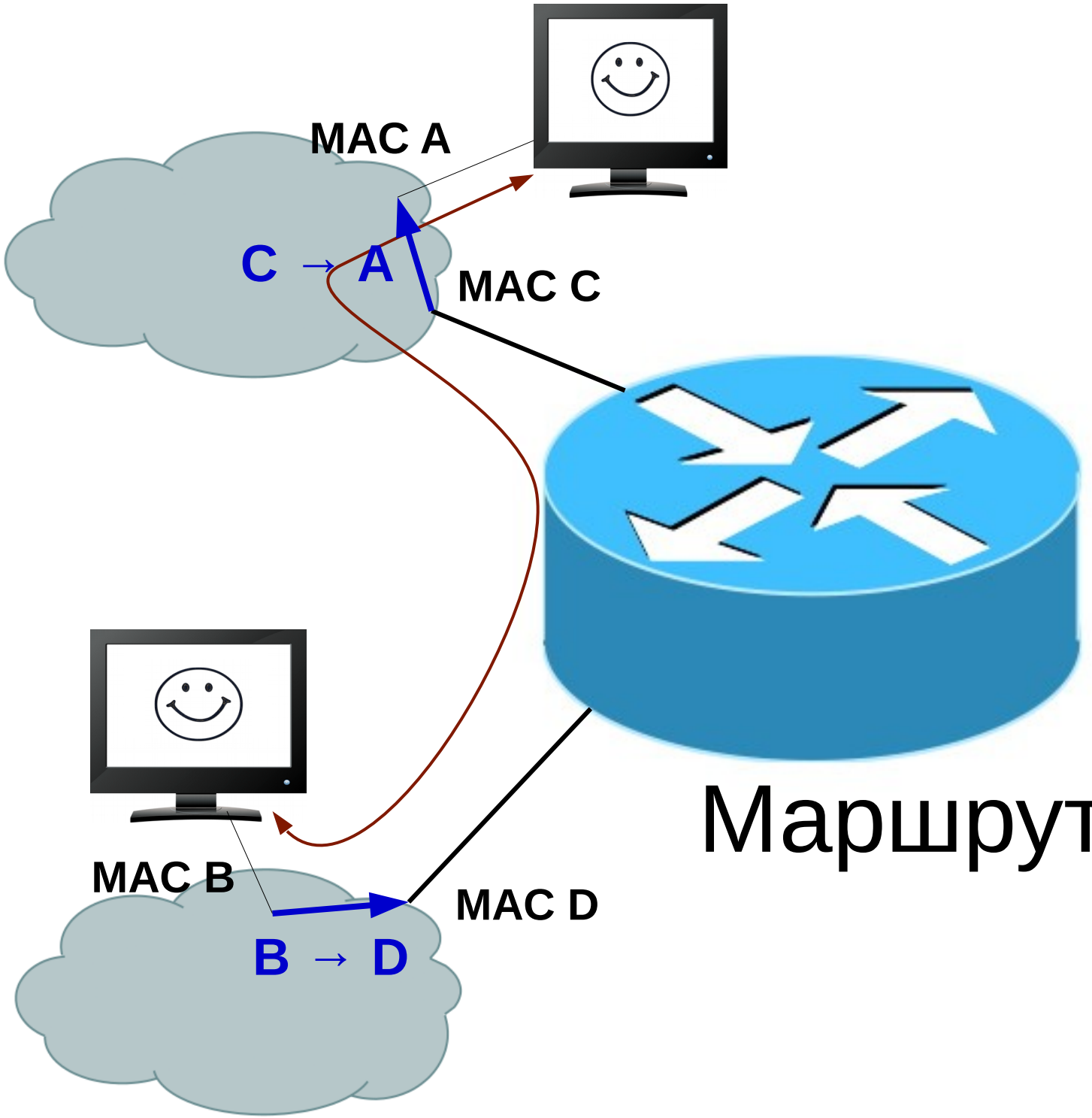
Маршрутизатор



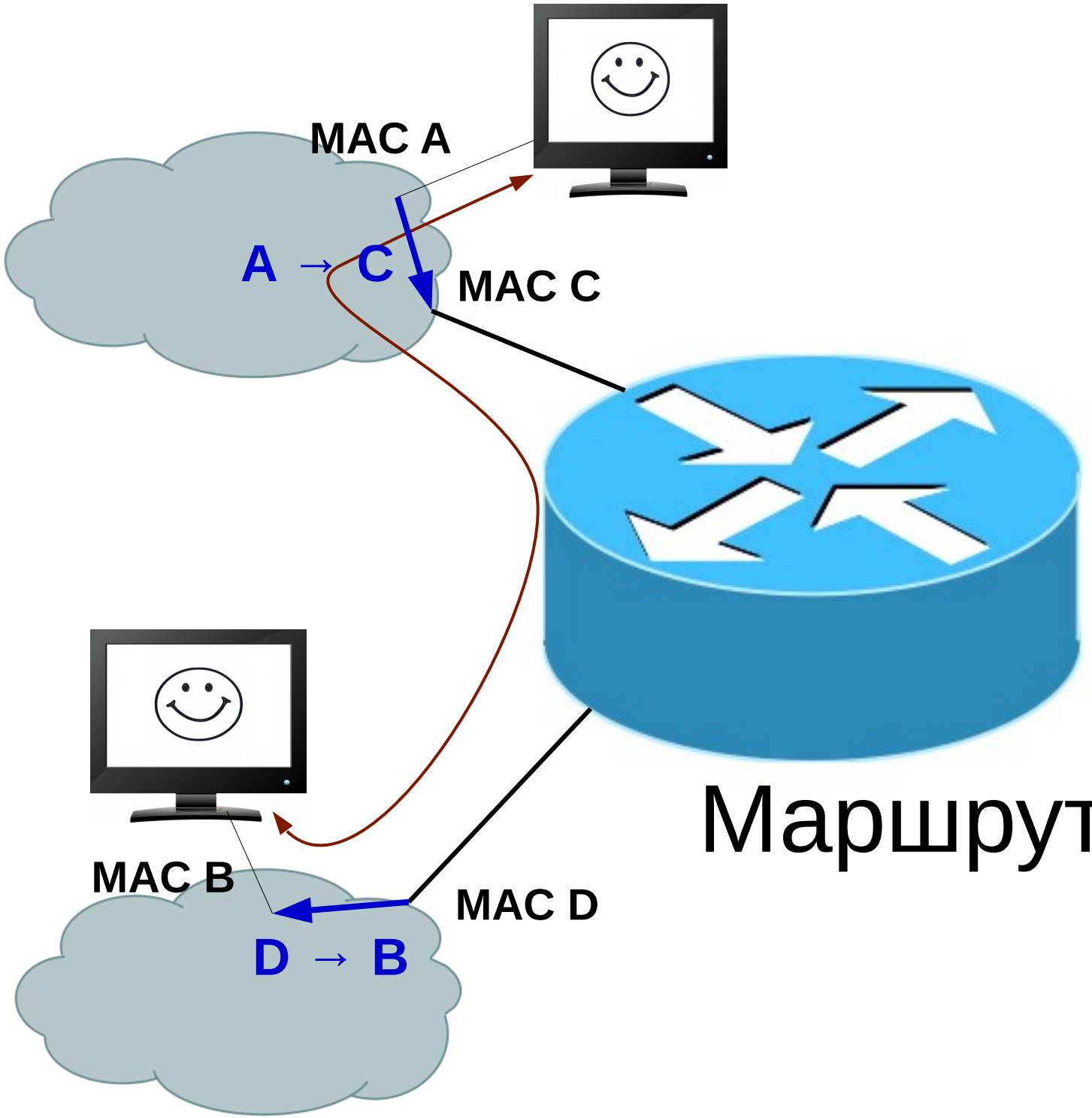
Маршрутизатор



Маршрутизатор

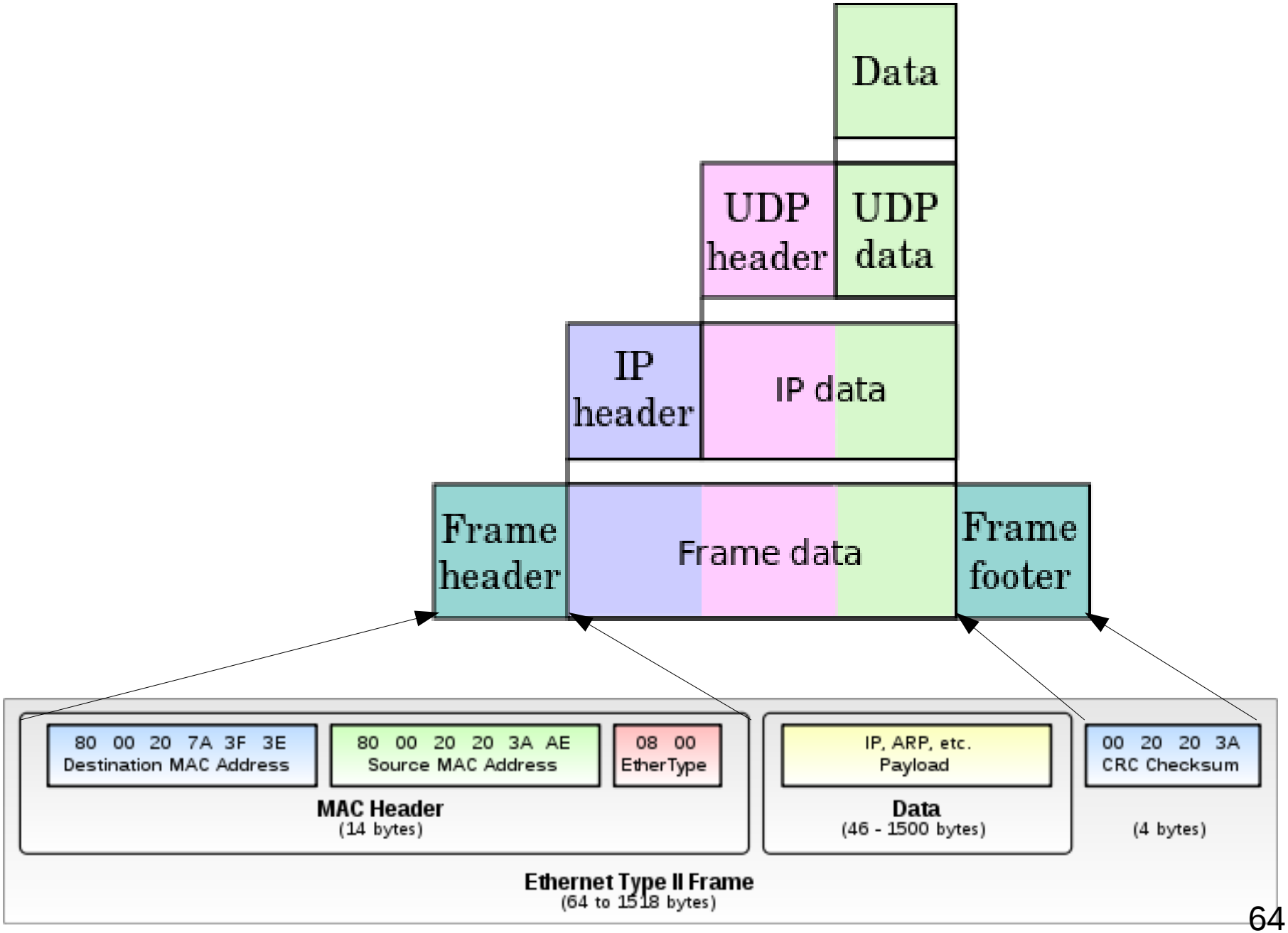


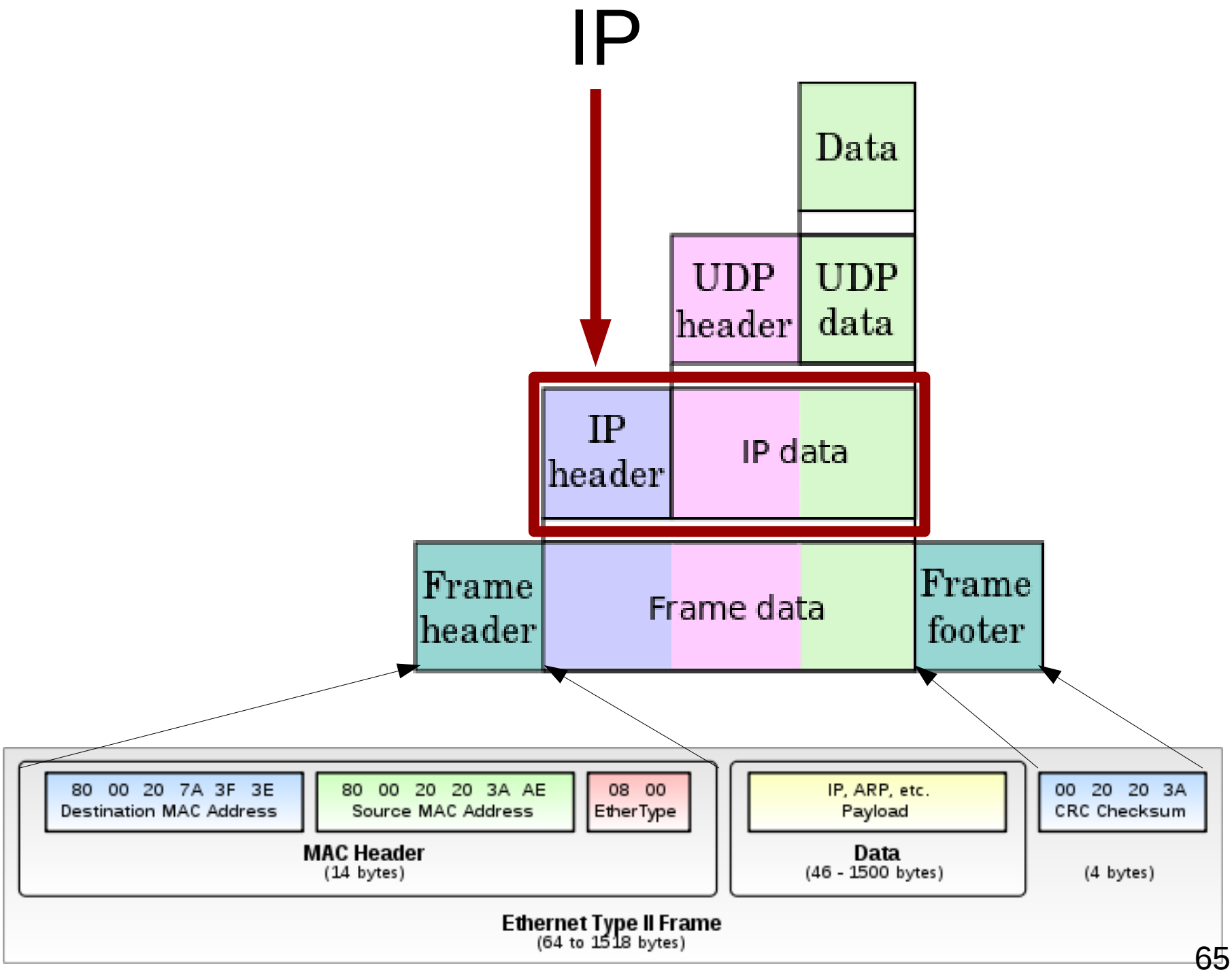
Маршрутизатор



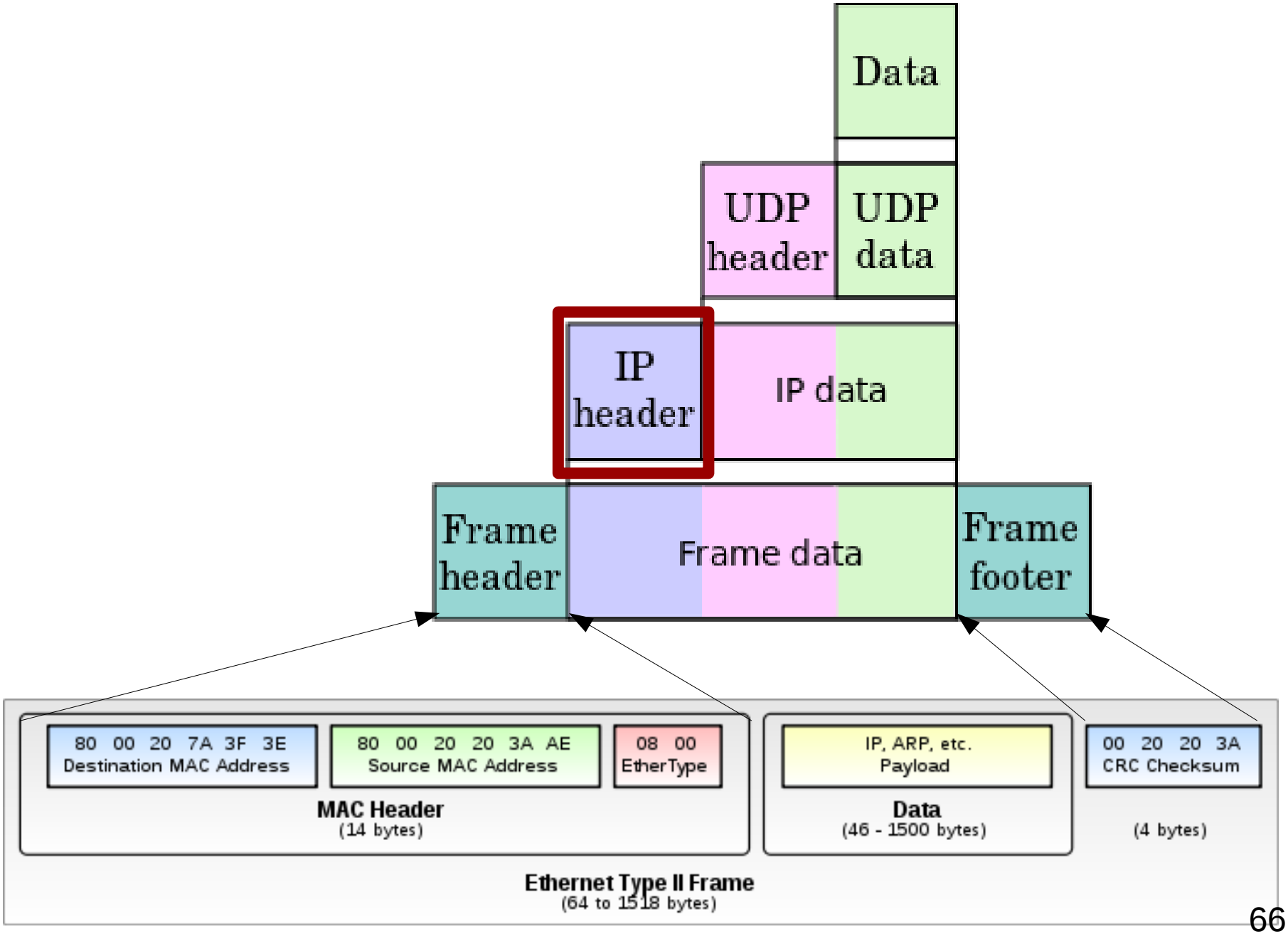
Маршрутизатор

IP


Data
UDP header
UDP data
IP header
IP data
Frame header
Frame data
Frame footer
80 00 20 7A 3F 3E
Destination MAC Address
80 00 20 20 3A AE
Source MAC Address
08 00
EtherType
MAC Header
(14 bytes)
IP, ARP, etc.
Payload
Data
(46 - 1500 bytes)
00 20 20 3A
CRC Checksum
(4 bytes)
Ethernet Type II Frame
(64 to 1518 bytes)

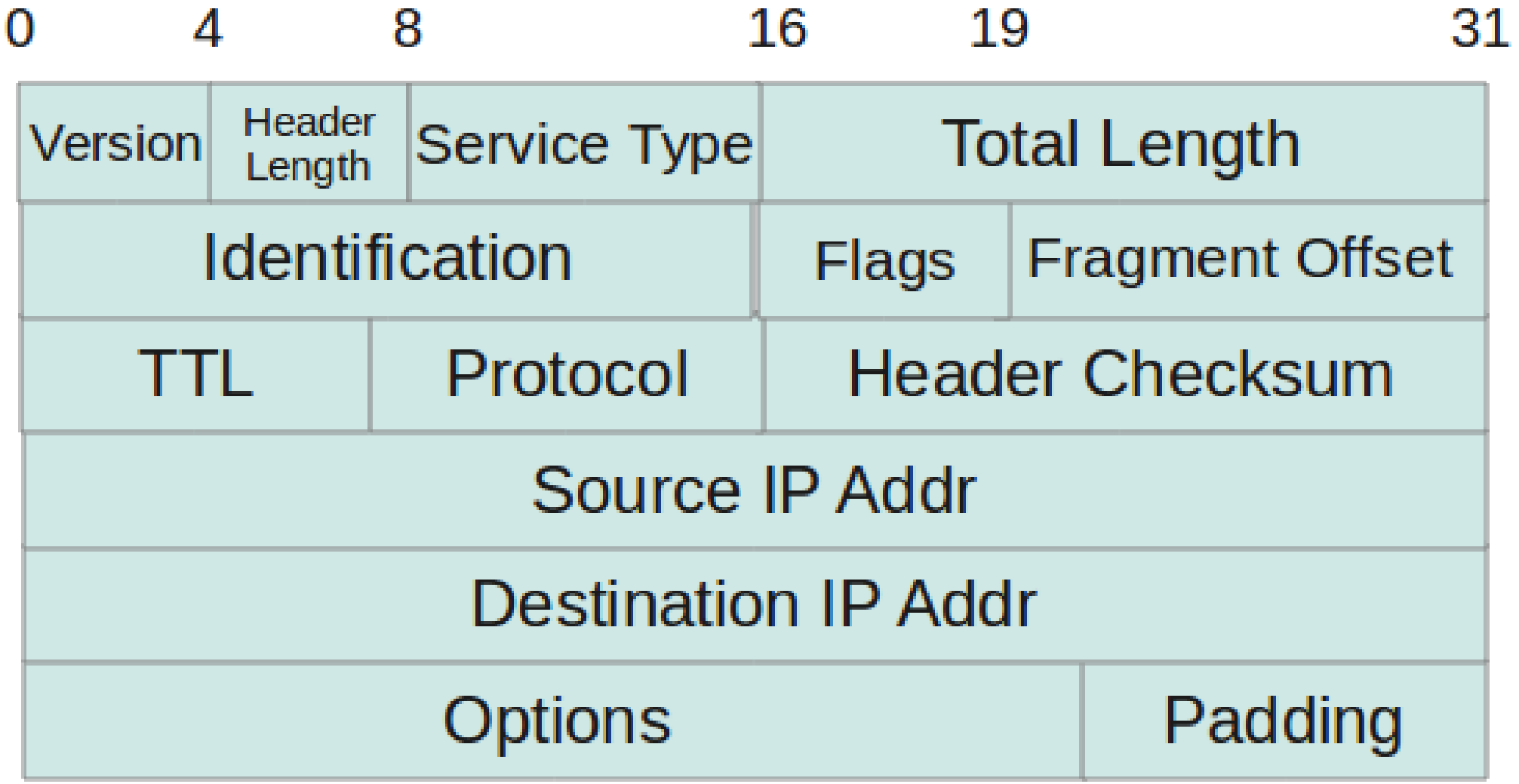

IP
Data
UDP header
UDP data
IP header
IP data
Frame header
Frame data
Frame footer
80 00 20 7A 3F 3E
Destination MAC Address
80 00 20 20 3A AE
Source MAC Address
08 00
EtherType
MAC Header
(14 bytes)
IP, ARP, etc.
Payload
Data
(46 - 1500 bytes)
00 20 20 3A
CRC Checksum
(4 bytes)
Ethernet Type II Frame
(64 to 1518 bytes)

IP

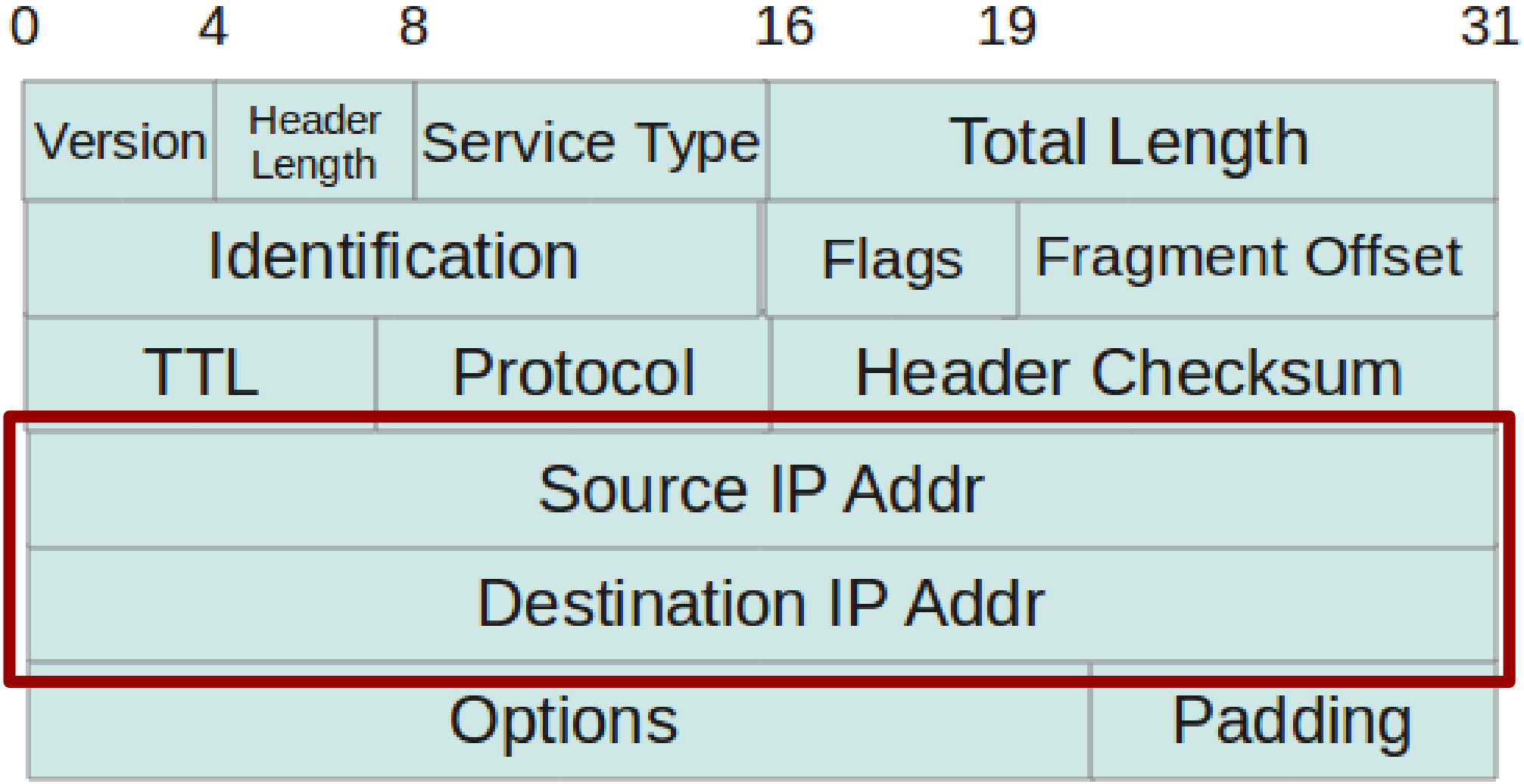

Data
UDP header
UDP data
IP header
IP data
Frame header
Frame data
Frame footer
80 00 20 7A 3F 3E
Destination MAC Address
80 00 20 20 3A AE
Source MAC Address
08 00
EtherType
MAC Header
(14 bytes)
IP, ARP, etc.
Payload
Data
(46 - 1500 bytes)
00 20 20 3A
CRC Checksum
(4 bytes)
Ethernet Type II Frame
(64 to 1518 bytes)

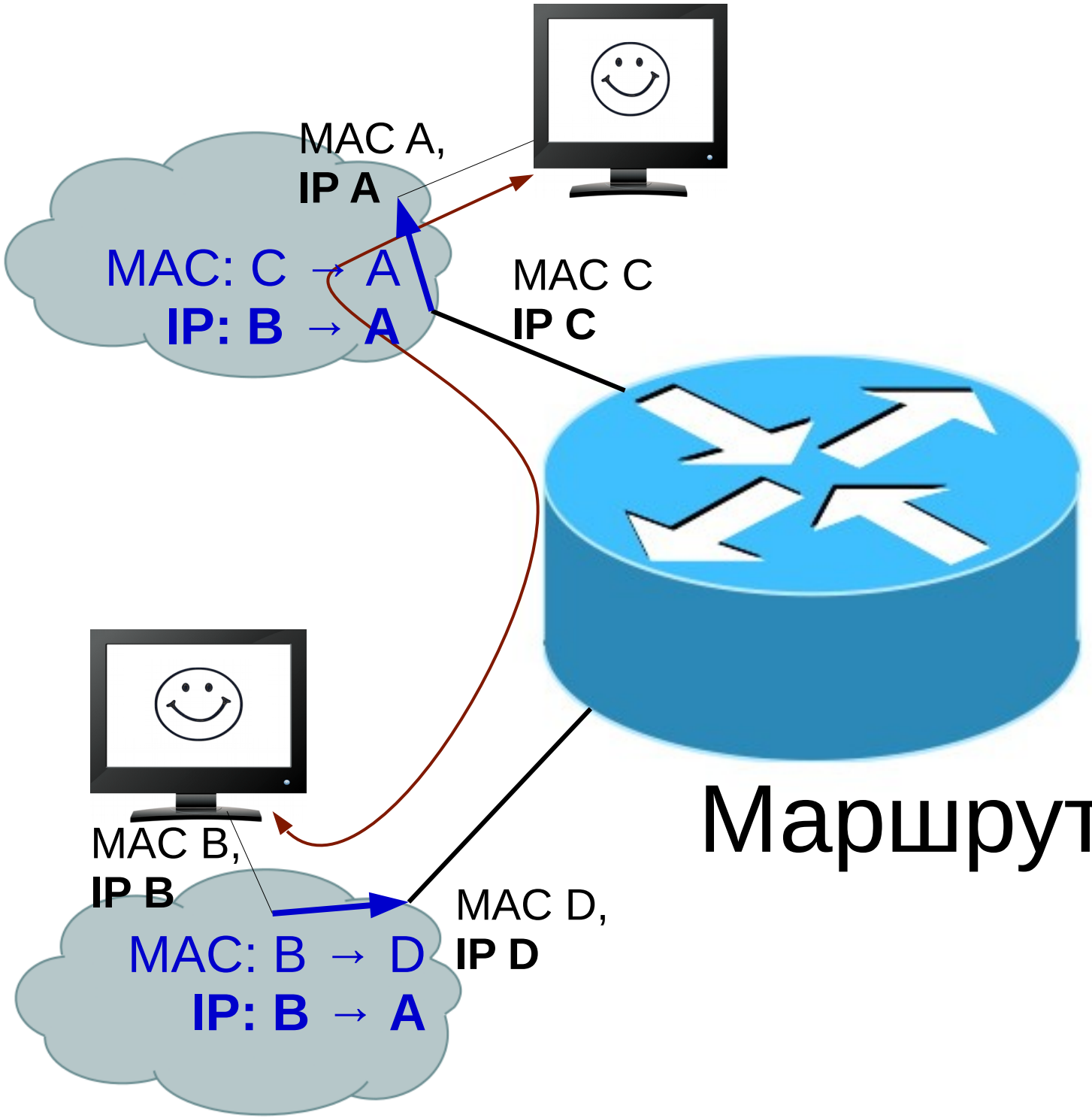

IP
header

IP

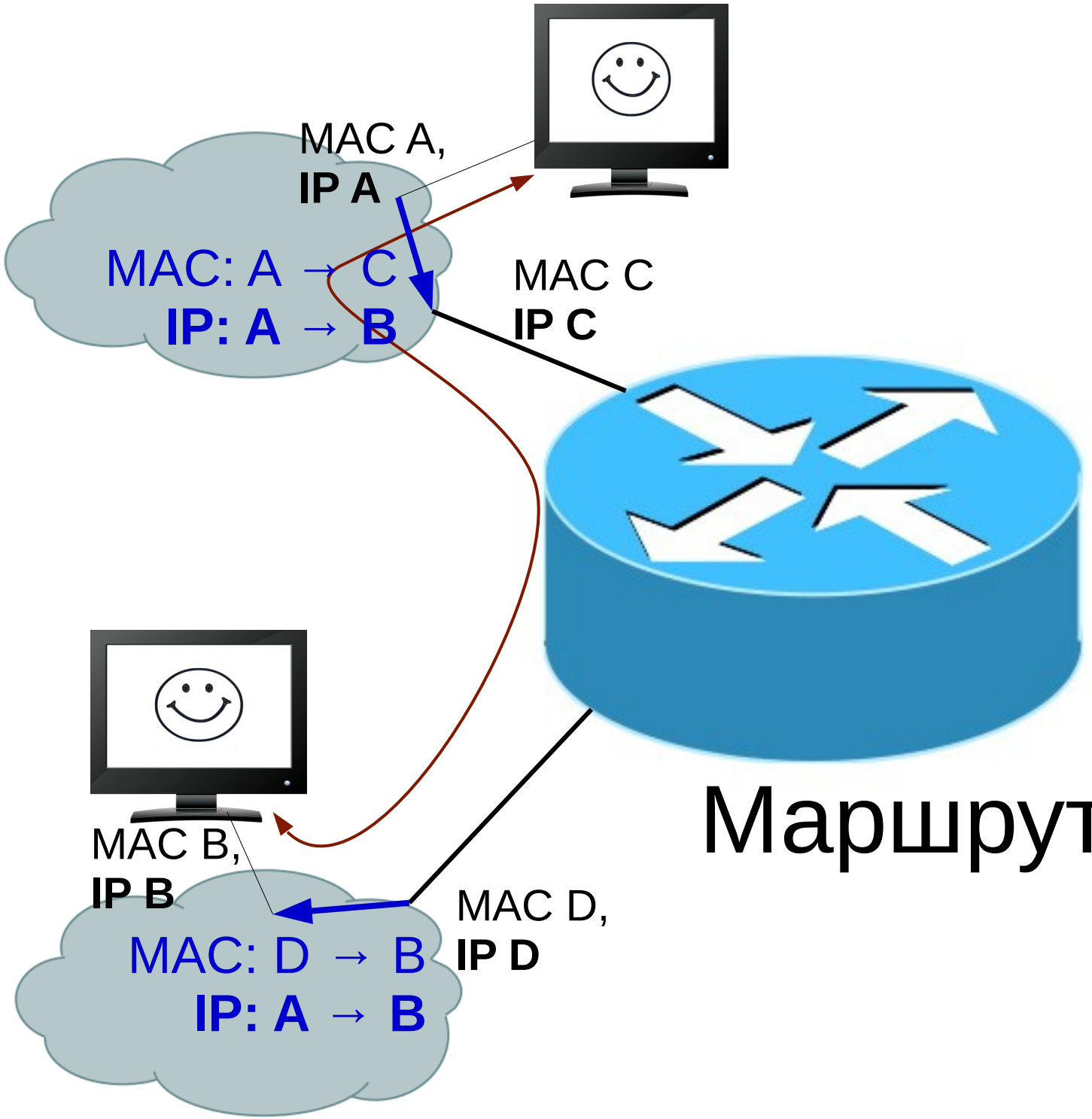

0
4
8
16
19
31
Version
Header Length
Service Type
Total Length
Identification
Flags
Fragment Offset
TTL
Protocol
Header Checksum
Source IP Addr
Destination IP Addr
Options
Padding

IP


0
4
8
16
19
31
Version
Header Length
Service Type
Total Length
Identification
Flags
Fragment Offset
TTL
Protocol
Header Checksum
Source IP Addr
Destination IP Addr
Options
Padding



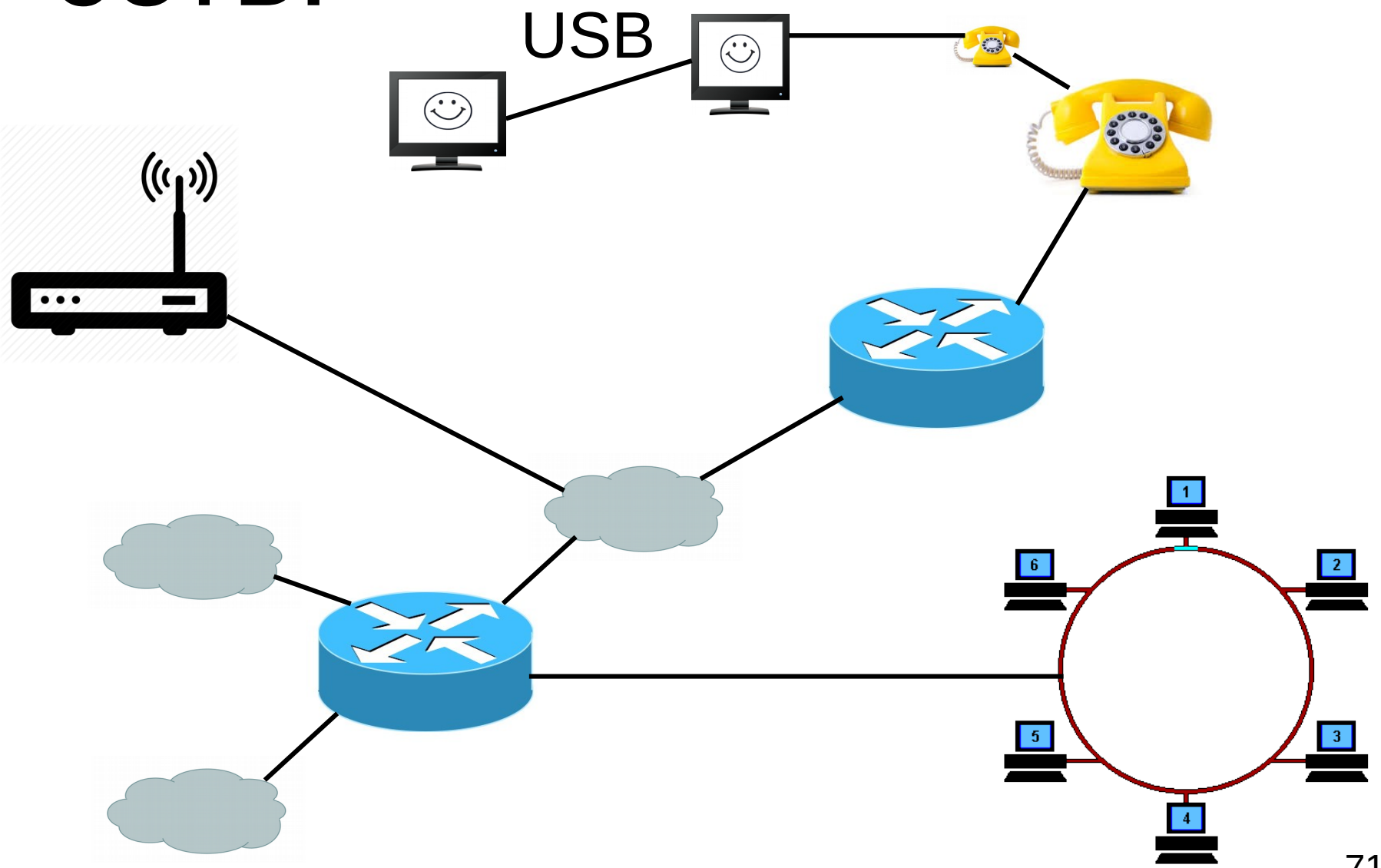
Маршрутизатор



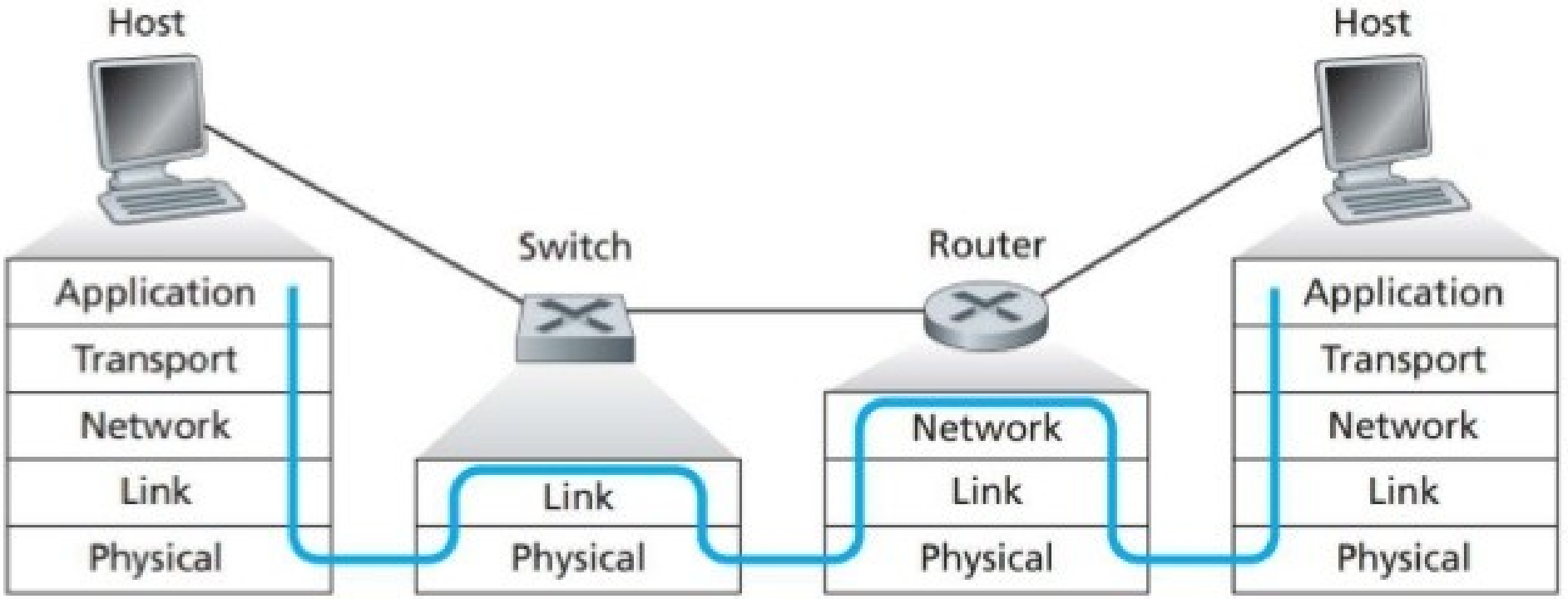
Маршрутизатор

IP

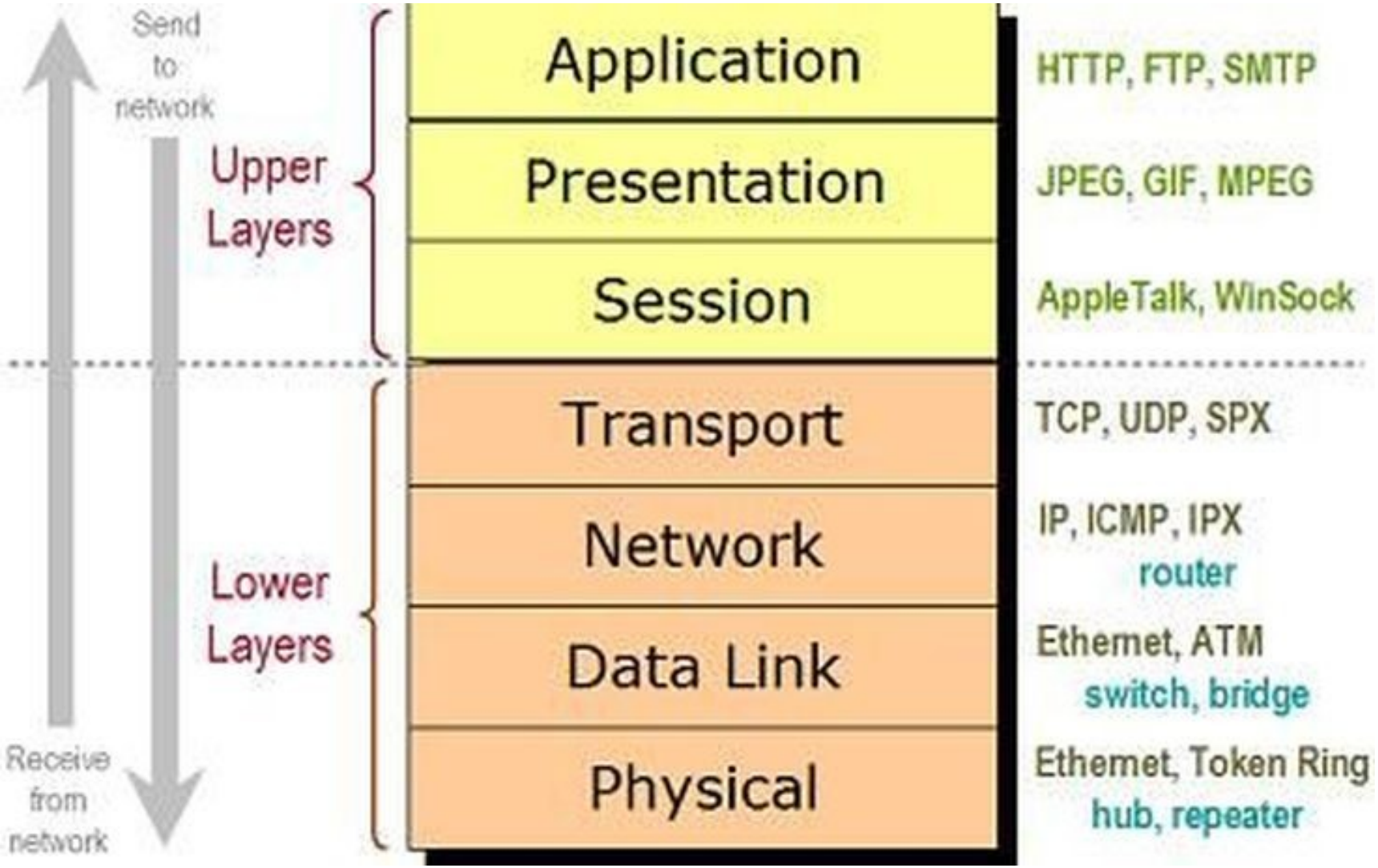
IP-сеть:



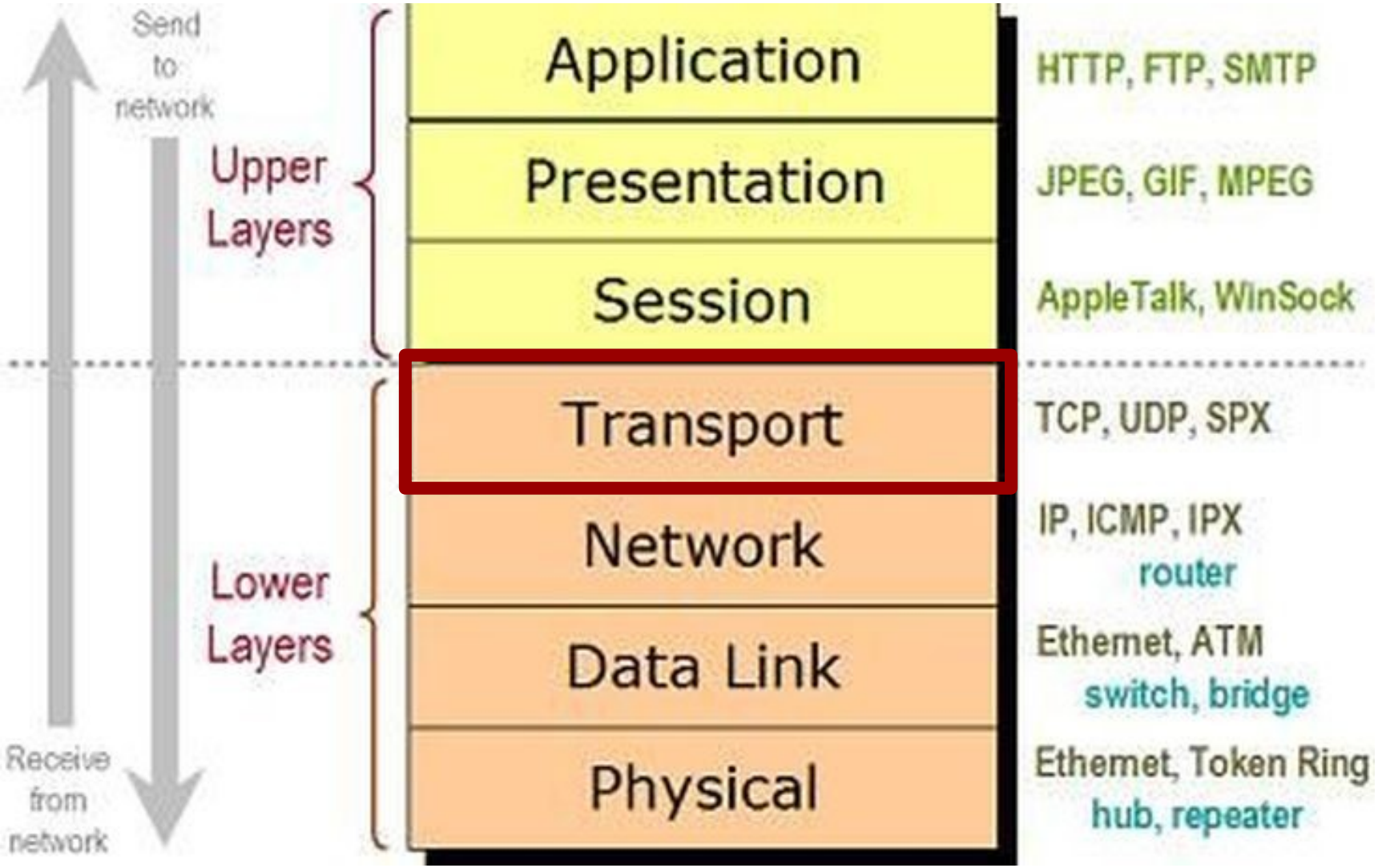
IP

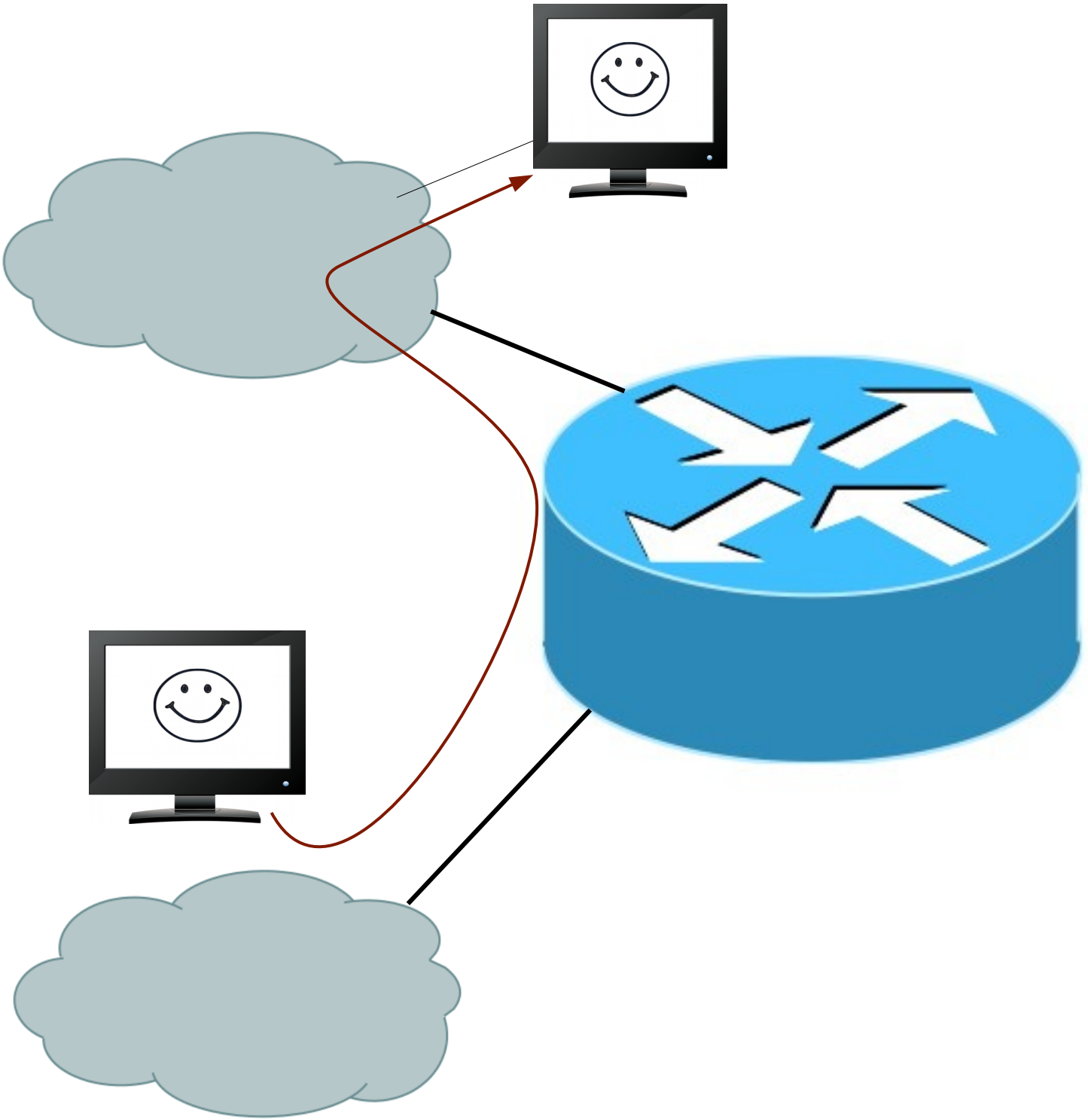

Host
Host
Switch
Router
Application
Transport
Network
Link
Physical
Link
Physical
Network
Link
Physical
Application
Transport
Network
Link
Physical

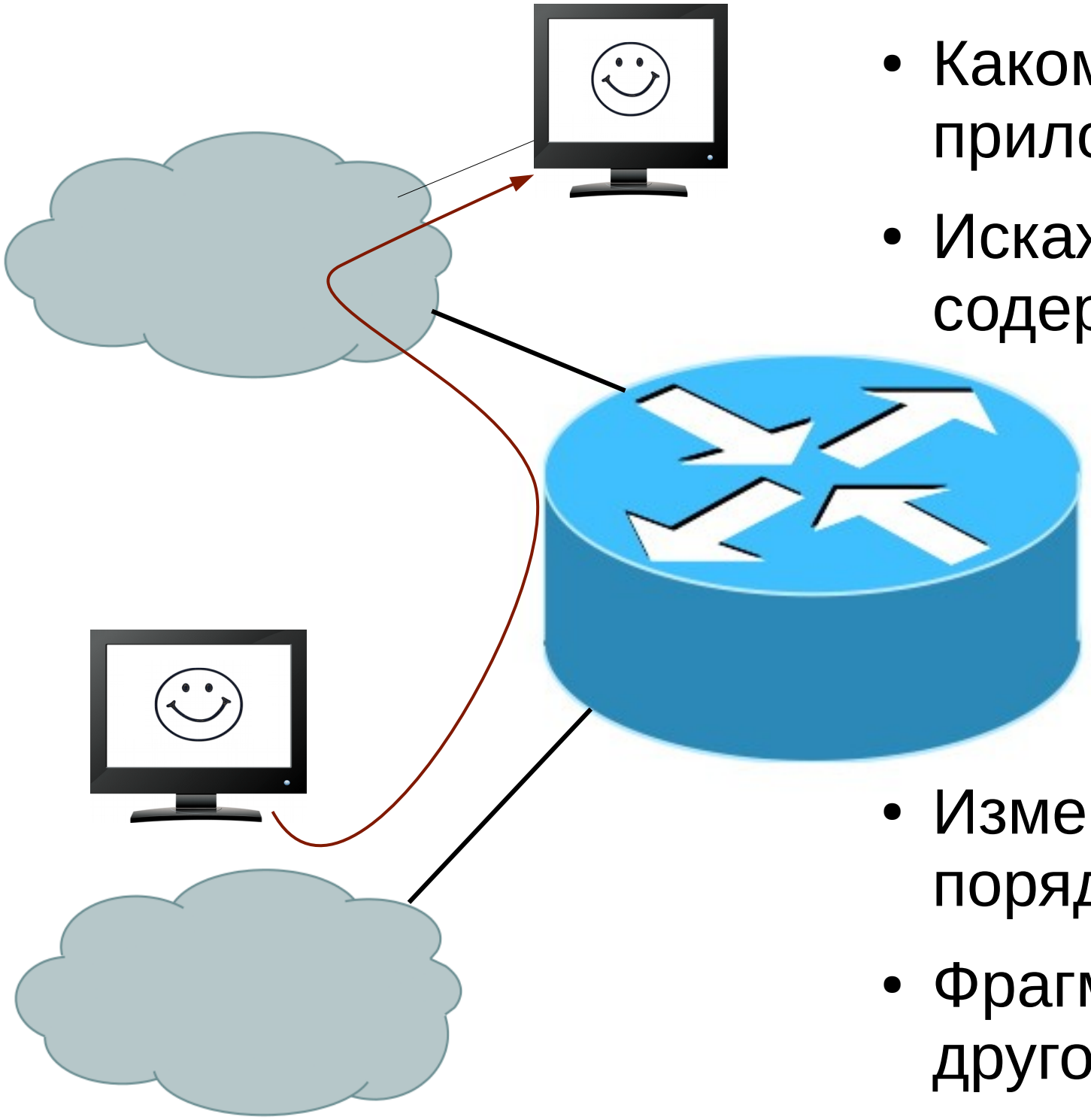
OSI


Send to network
Upper Layers
Lower Layers
Receive from network
Application
Presentation
Session
Transport
Network
Data Link
Physical
HTTP, FTP, SMTP
JPEG, GIF, MPEG
AppleTalk, WinSock
TCP, UDP, SPX
IP, ICMP, IPX
router
Ethernet, ATM
switch, bridge
Ethernet, Token Ring
hub, repeater

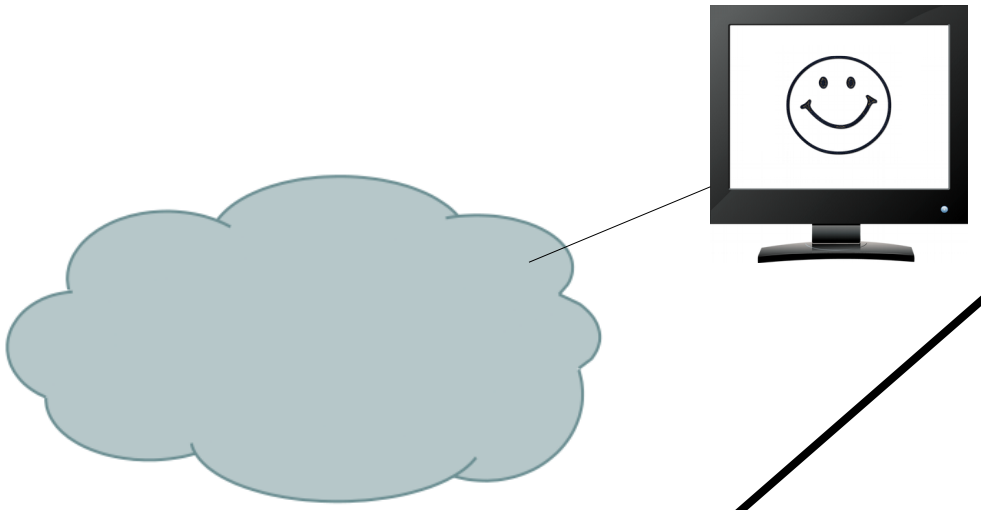
OSI


Send to network
Upper Layers
Lower Layers
Receive from network
Application
Presentation
Session
Transport
Network
Data Link
Physical
HTTP, FTP, SMTP
JPEG, GIF, MPEG
AppleTalk, WinSock
TCP, UDP, SPX
IP, ICMP, IPX
router
Ethernet, ATM
switch, bridge
Ethernet, Token Ring
hub, repeater





- Какому приложению?
- Искажение содержимого?
- Изменение порядка пакетов?
- Фрагментация и другое…



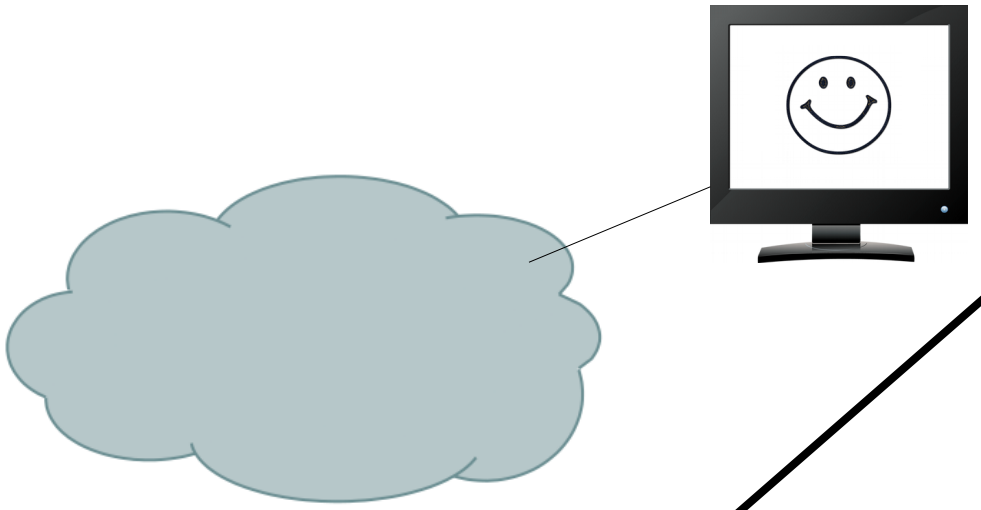
- Какому приложению?
- Искажение содержимого?

контрольная сумма

порт

порядковый номер

- Изменение порядка пакетов?
- Фрагментация и другое…



- Какому приложению?
- Искажение содержимого?

порт

контрольная сумма

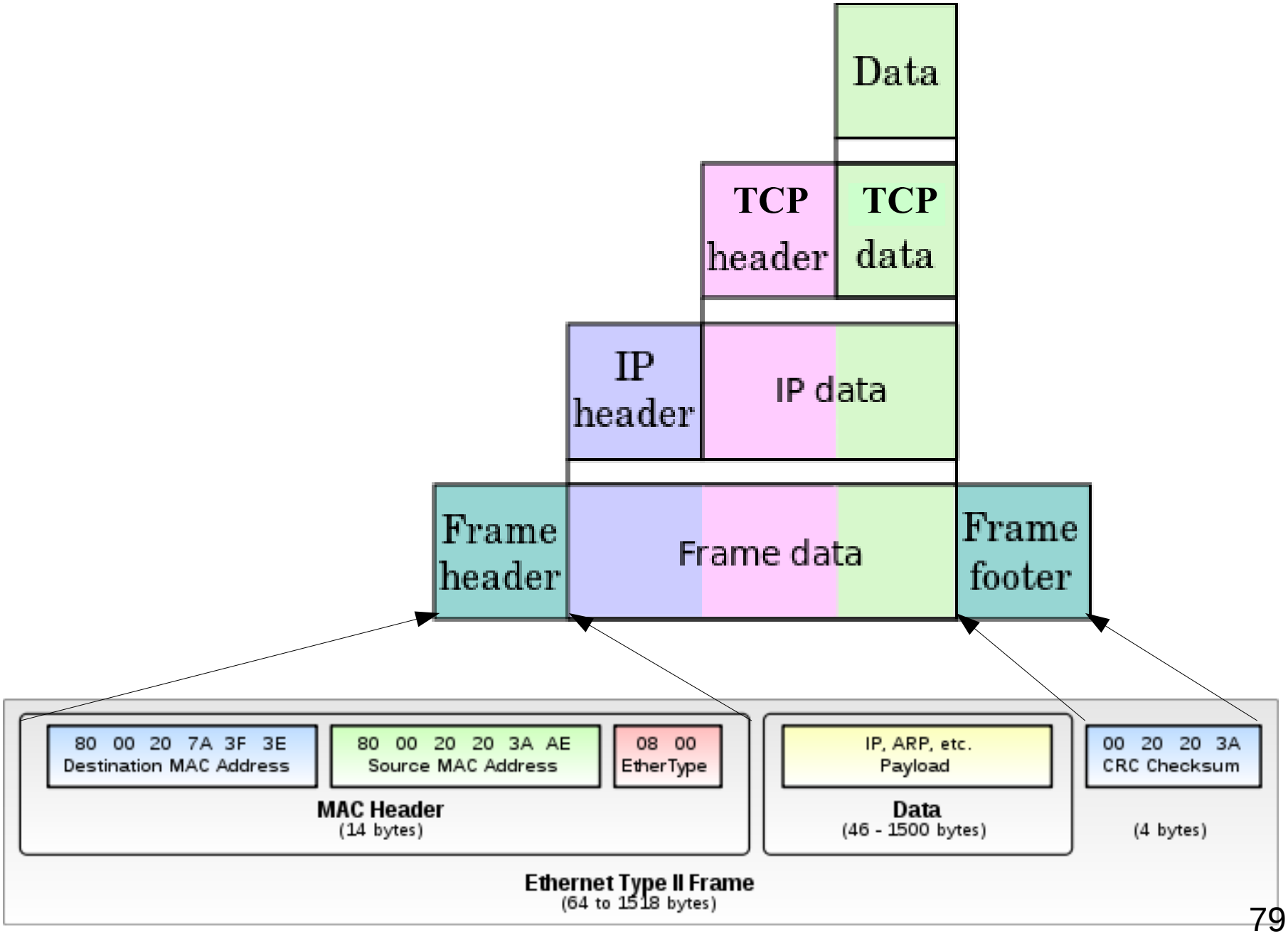
порядковый номер

- Изменение порядка пакетов?

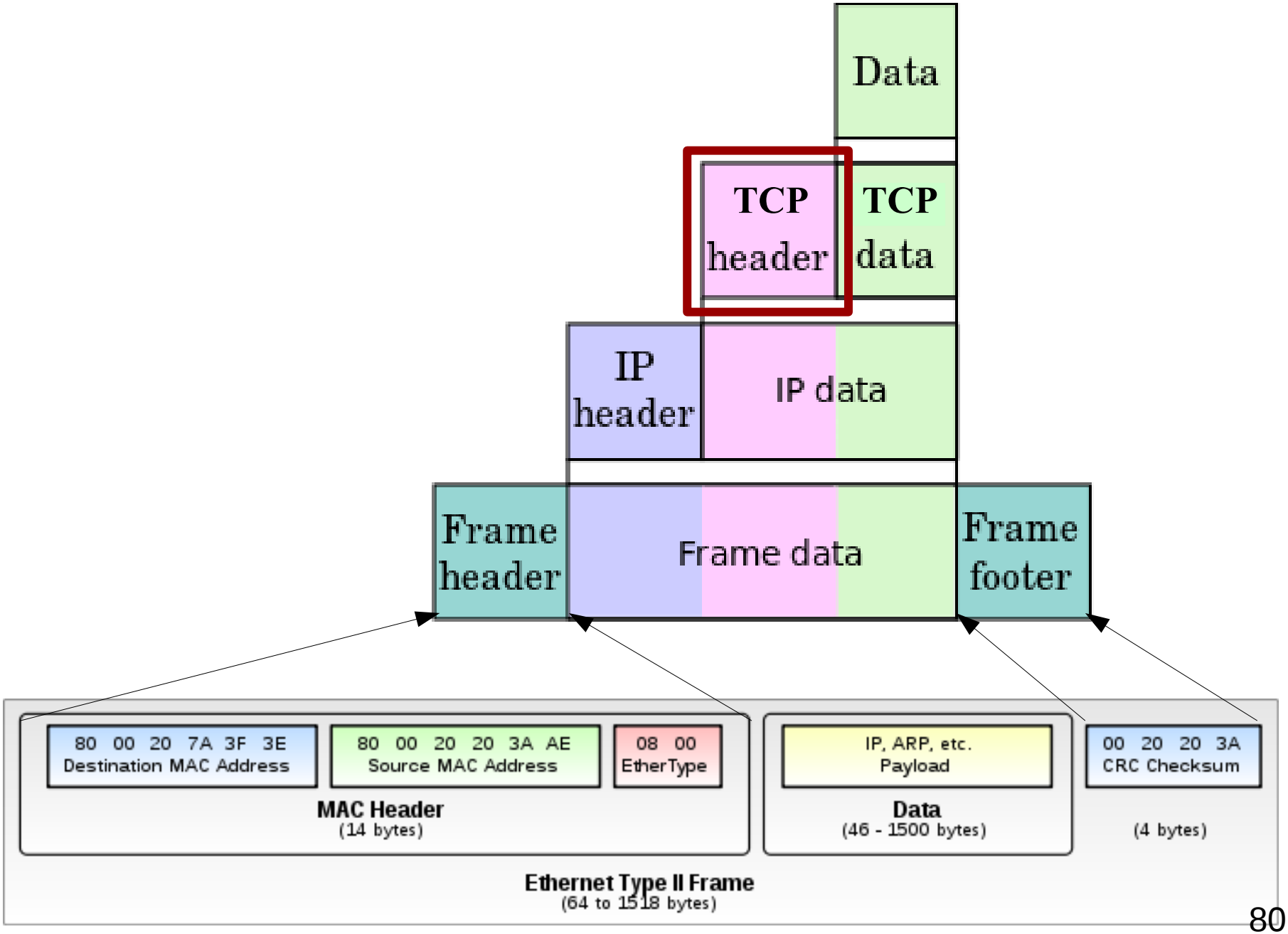
И прочее…

- Фрагментация и другое…

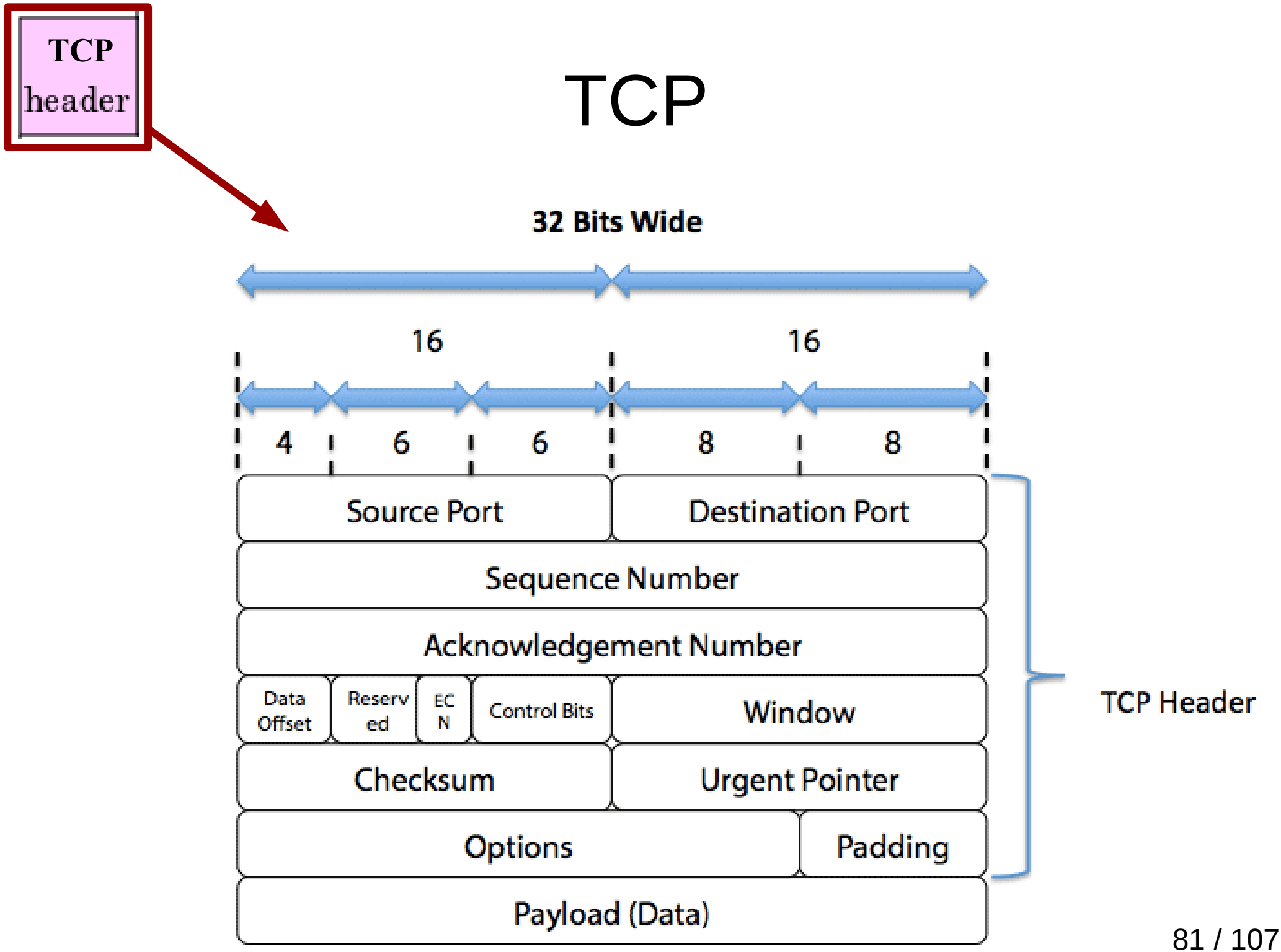
TCP


Data
TCP header
TCP data
IP header
IP data
Frame header
Frame data
Frame footer
80 00 20 7A 3F 3E
Destination MAC Address
80 00 20 20 3A AE
Source MAC Address
08 00
EtherType
MAC Header
(14 bytes)
IP, ARP, etc.
Payload
Data
(46 - 1500 bytes)
00 20 20 3A
CRC Checksum
(4 bytes)
Ethernet Type II Frame
(64 to 1518 bytes)

TCP


Data
TCP header
TCP data
IP header
IP data
Frame header
Frame data
Frame footer
80 00 20 7A 3F 3E
Destination MAC Address
80 00 20 20 3A AE
Source MAC Address
08 00
EtherType
MAC Header
(14 bytes)
IP, ARP, etc.
Payload
Data
(46 - 1500 bytes)
00 20 20 3A
CRC Checksum
(4 bytes)
Ethernet Type II Frame
(64 to 1518 bytes)

TCP

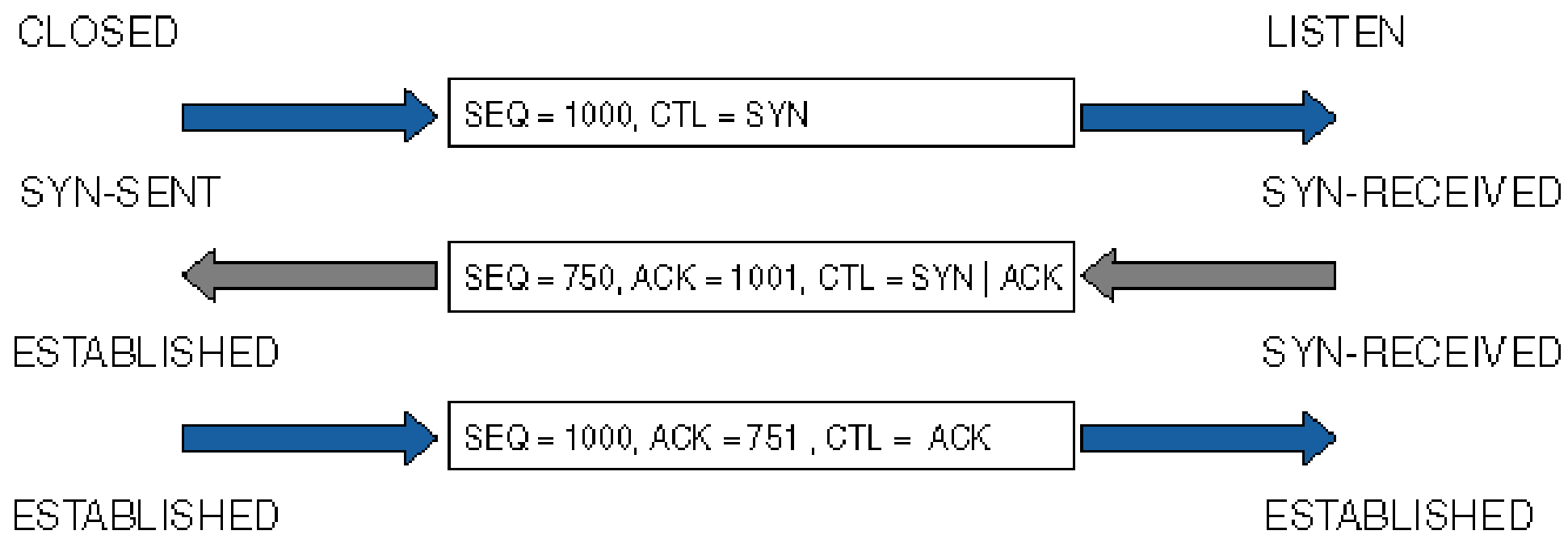

TCP header
32 Bits Wide
16
16
4
6
6
8
8
Source Port
Destination Port
Sequence Number
Acknowledgement Number
Data Offset
Reserv ed
EC N
Control Bits
Window
Checksum
Urgent Pointer
Options
Padding
Payload (Data)
TCP Header

TCP

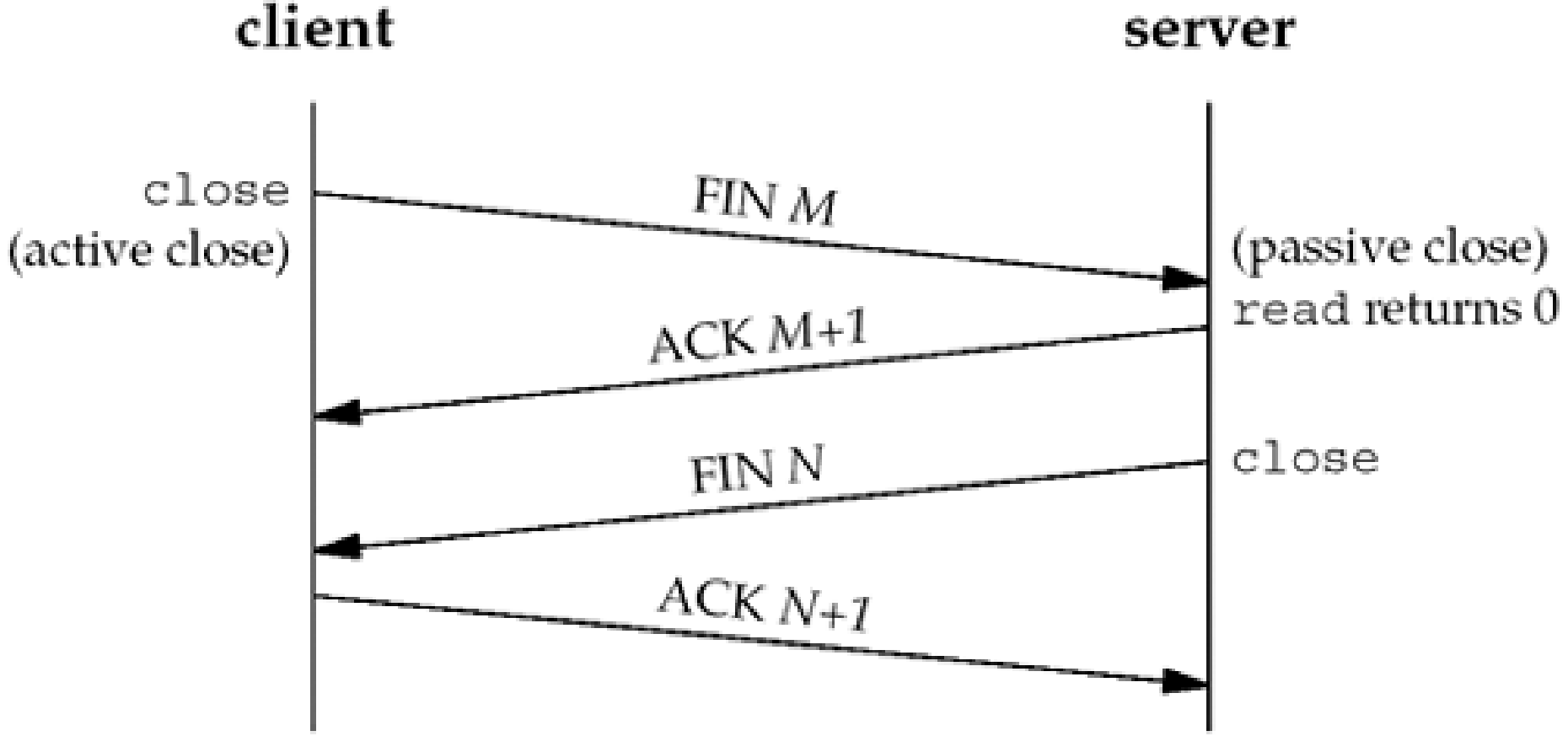
TCP State

TCP Packet

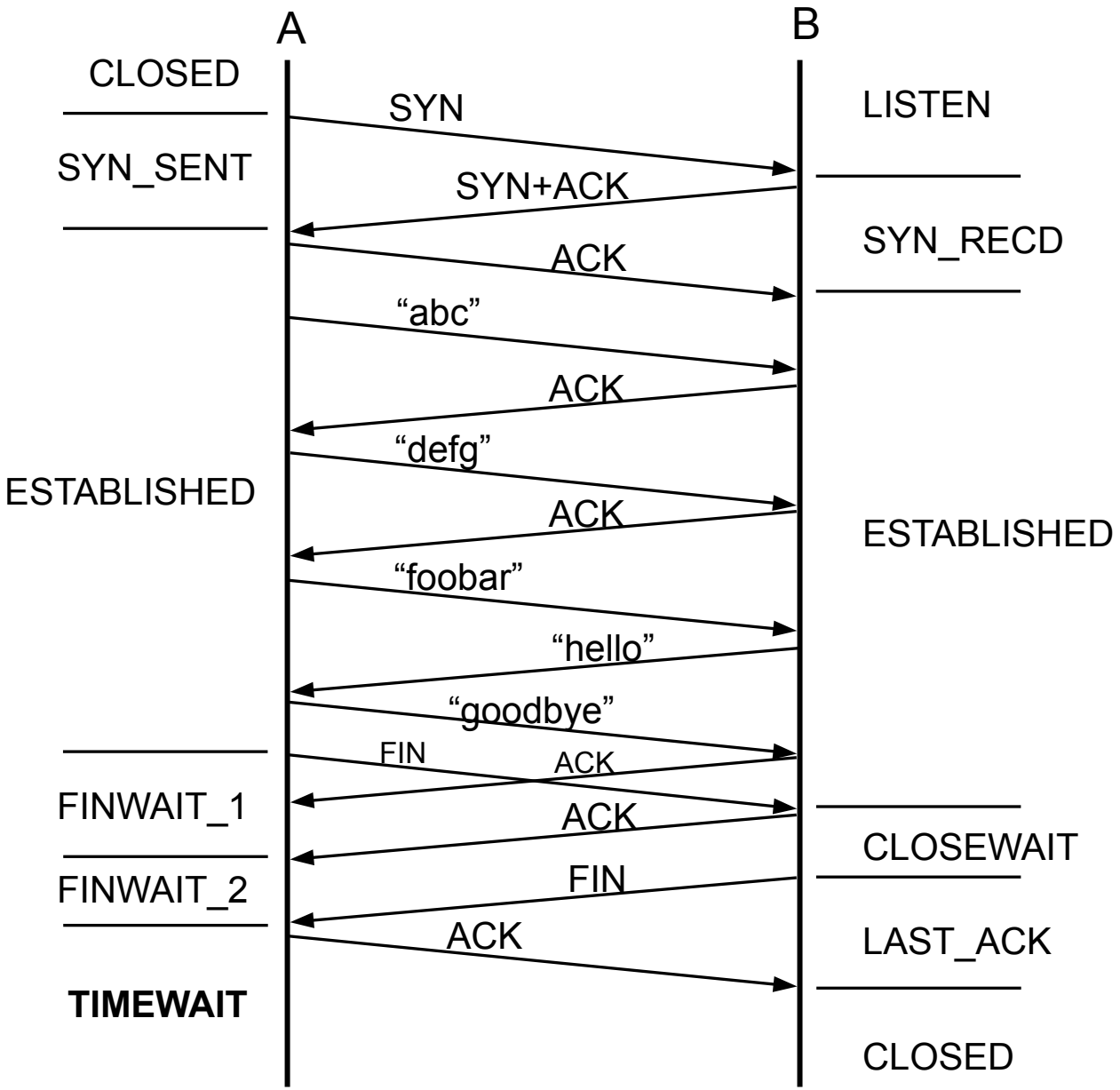
TCP State



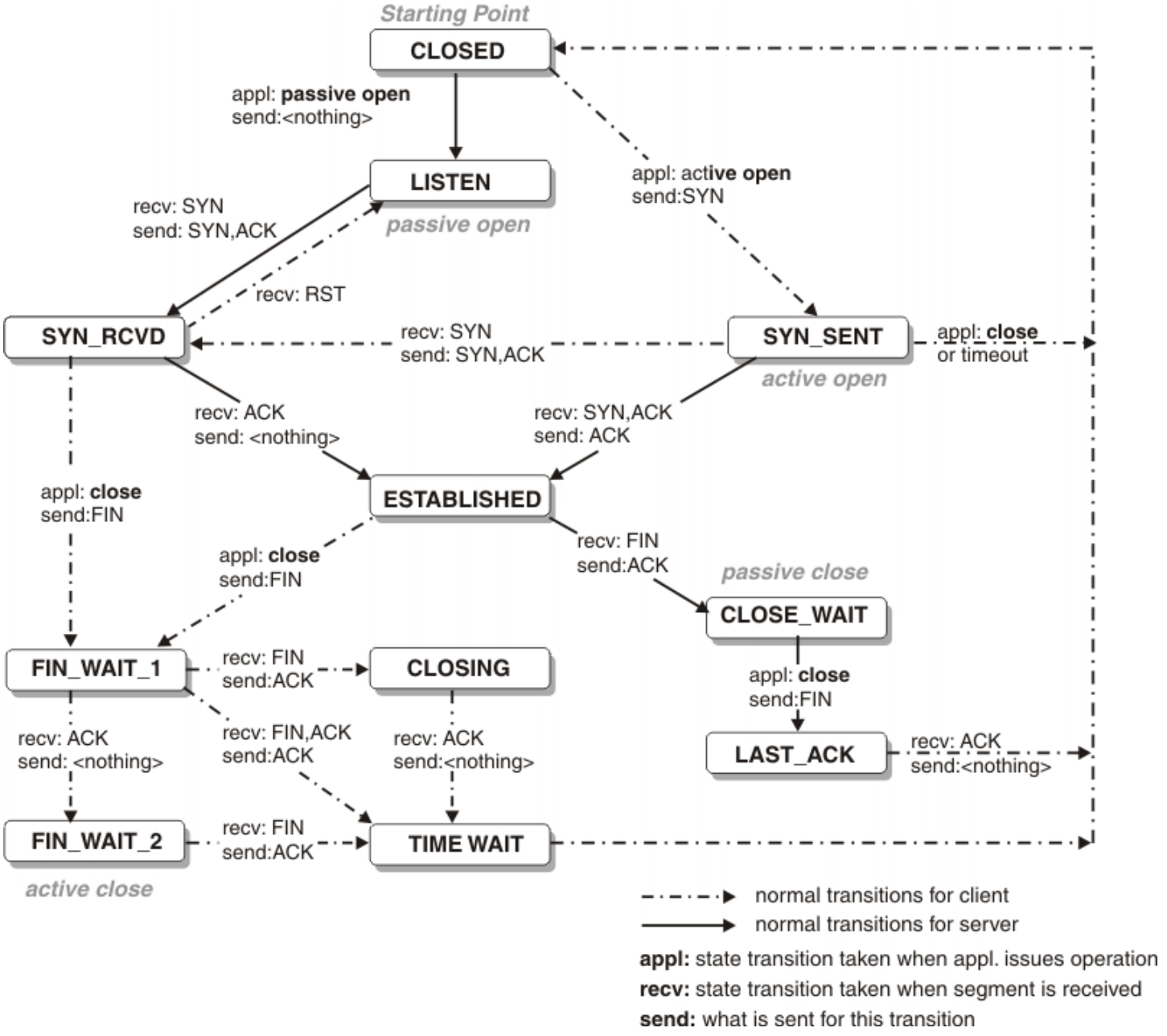
TCP


client
server
close
(active close)
FIN M
(passive close)
read returns 0
ACK M+1
FIN N
close
ACK N+1

TCP


A
B
CLOSED
SYN
LISTEN
SYN_SENT
SYN+ACK
SYN_RECD
ACK
“abc”
ACK
“defg”
ESTABLISHED
ACK
ESTABLISHED
“foobar”
“hello”
“goodbye”
FIN
ACK
FINWAIT_1
ACK
CLOSEWAIT
FINWAIT_2
FIN
ACK
LAST_ACK
TIMEWAIT
CLOSED

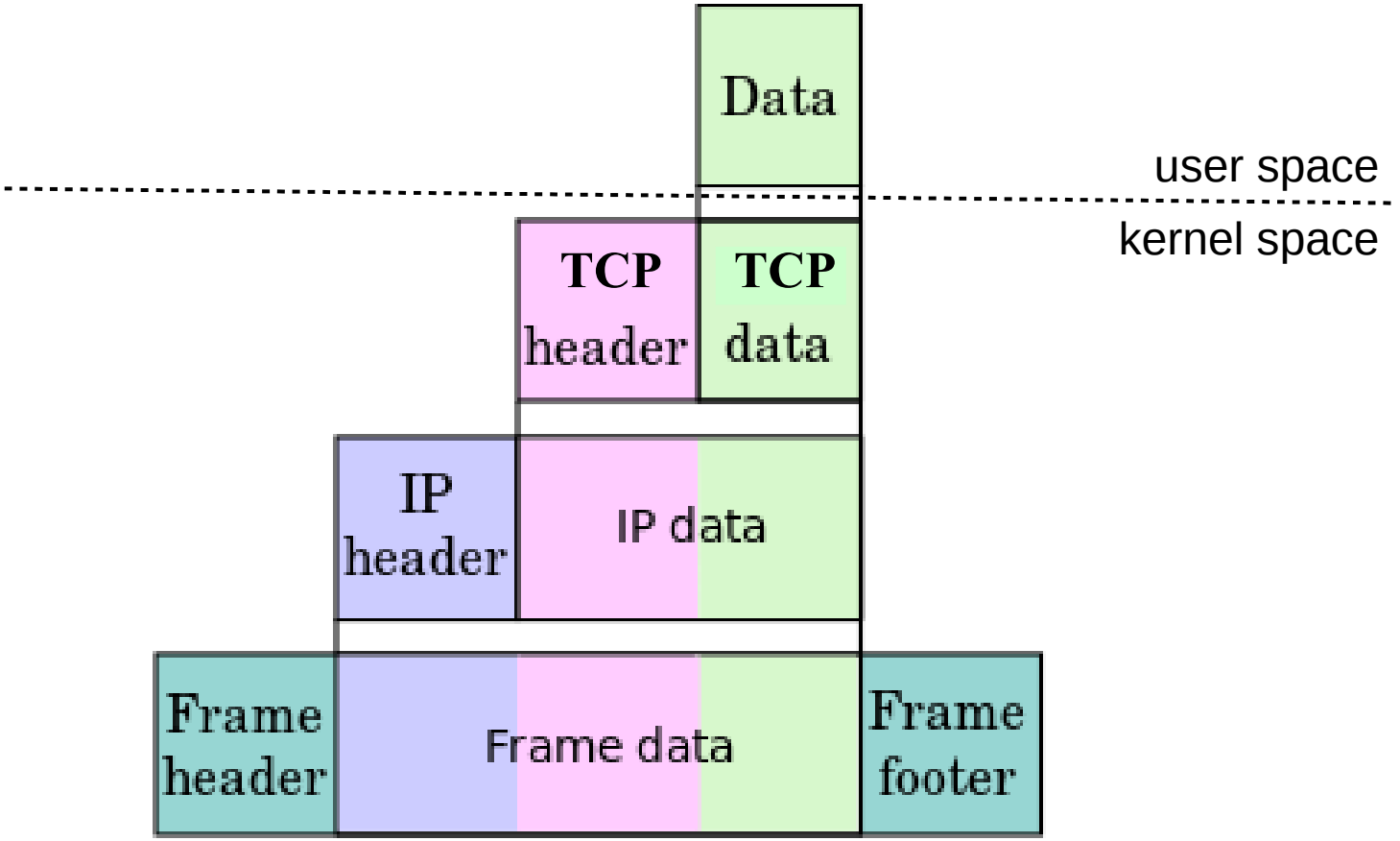
TCP


Starting Point
CLOSED
appl: passive open
send:<nothing>
LISTEN
passive open
appl: active open
send:SYN
recv: SYN
send: SYN,ACK
recv: RST
SYN_RCVD
recv: SYN
send: SYN,ACK
SYN_SENT
appl: close
or timeout
active open
recv: ACK
send: <nothing>
recv: SYN,ACK
send: ACK
ESTABLISHED
appl: close
send:FIN
recv: FIN
send:ACK
passive close
CLOSE_WAIT
appl: close
send:FIN
FIN_WAIT_1
recv: FIN
send:ACK
CLOSING
recv: ACK
send: <nothing>
recv: FIN,ACK
send:ACK
recv: ACK
send:<nothing>
LAST_ACK
recv: ACK
send:<nothing>
FIN_WAIT_2
recv: FIN
send:ACK
TIME WAIT
active close
normal transitions for client
normal transitions for server
appl: state transition taken when appl. issues operation
recv: state transition taken when segment is received
send: what is sent for this transition

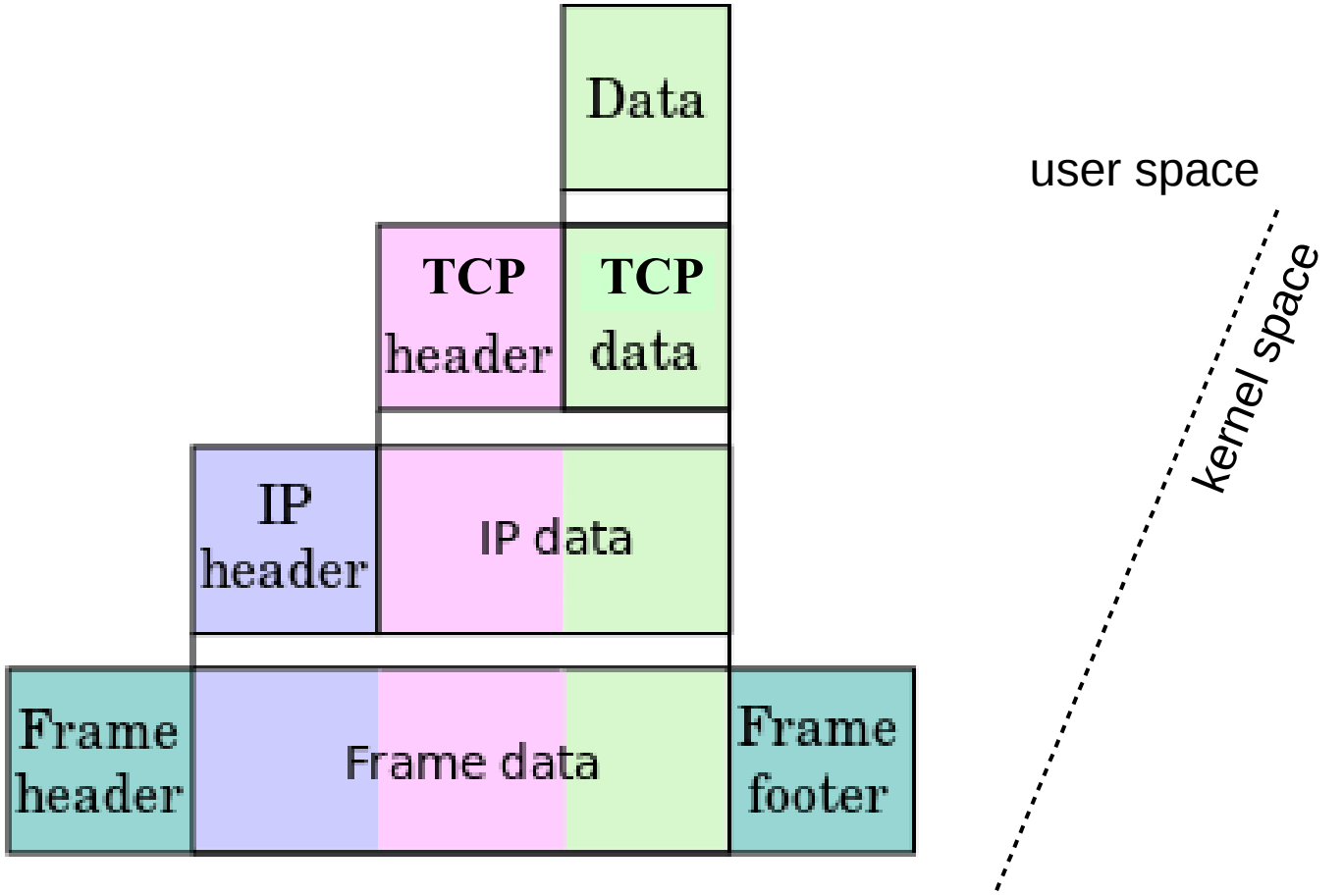
TCP vs UDP

TCP	UDP
Keeps track of lost packets. Makes sure that lost packets are re-sent	Doesn't keep track of lost packets
Adds sequence numbers to packets and reorders any packets that arrive in the wrong order	Doesn't care about packet arrival order
Slower, because of all added additional functionality	Faster, because it lacks any extra features
Requires more computer resources, because the OS needs to keep track of ongoing communication sessions and manage them on a much deeper level	Requires less computer resources
Examples of programs and services that use TCP: - HTTP - HTTPS - FTP - Many computer games	Examples of programs and services that use UDP: - DNS - IP telephony - DHCP - Many computer games

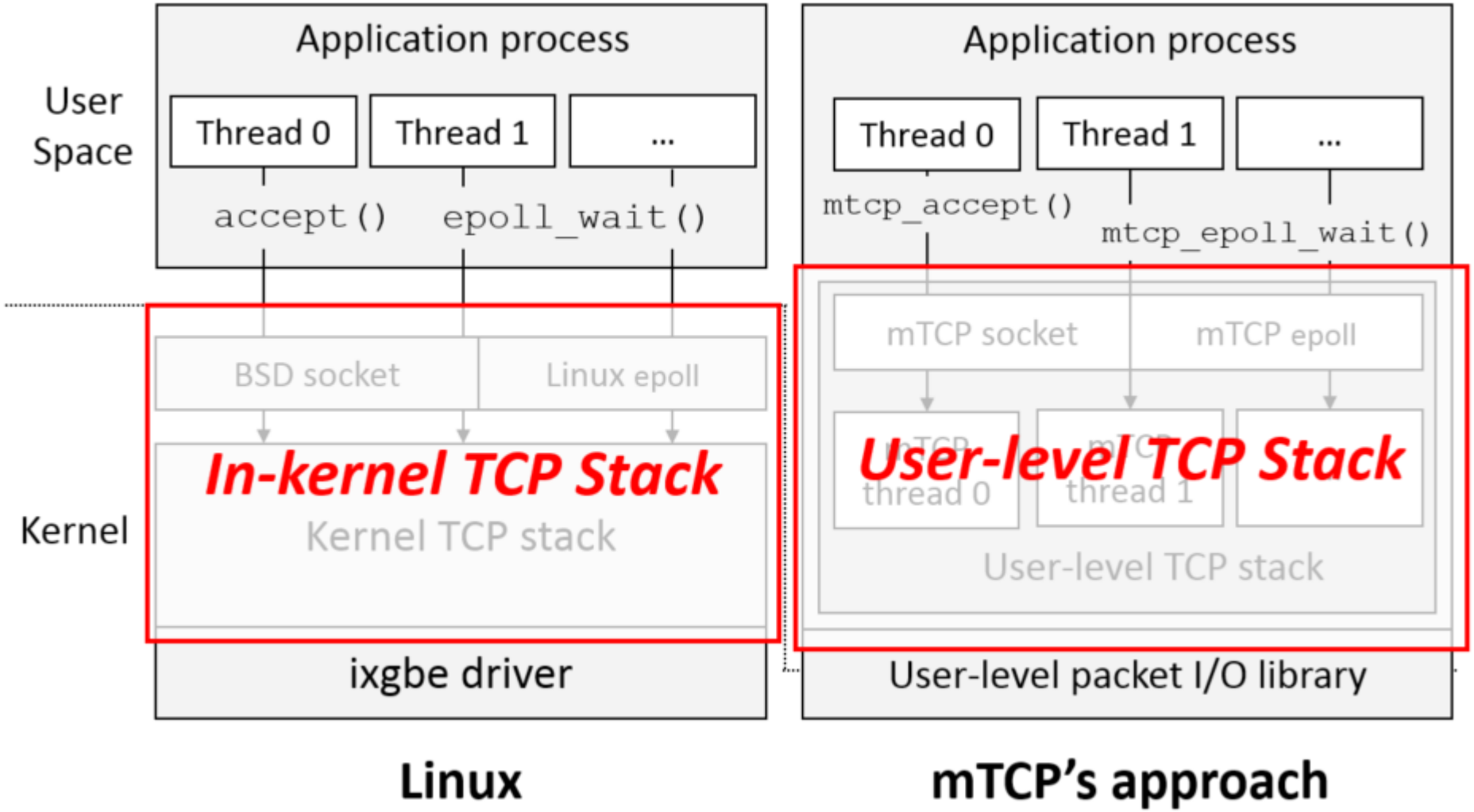
User TCP/IP


Data
user space
kernel space
TCP header
TCP data
IP header
IP data
Frame header
Frame data
Frame footer

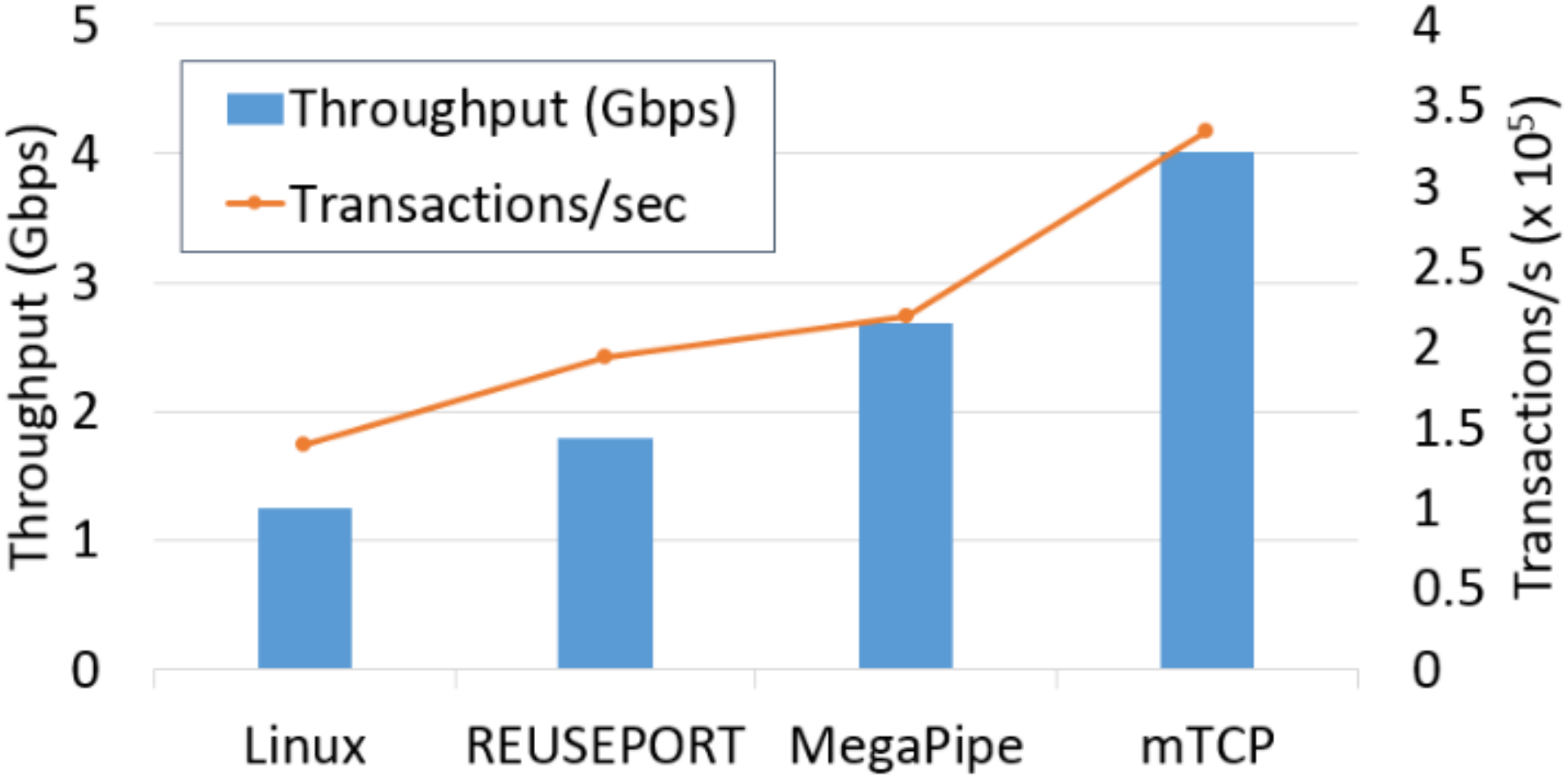
User TCP/IP


Data
TCP header
TCP data
IP header
IP data
Frame header
Frame data
Frame footer
user space
kernel space

User TCP/IP

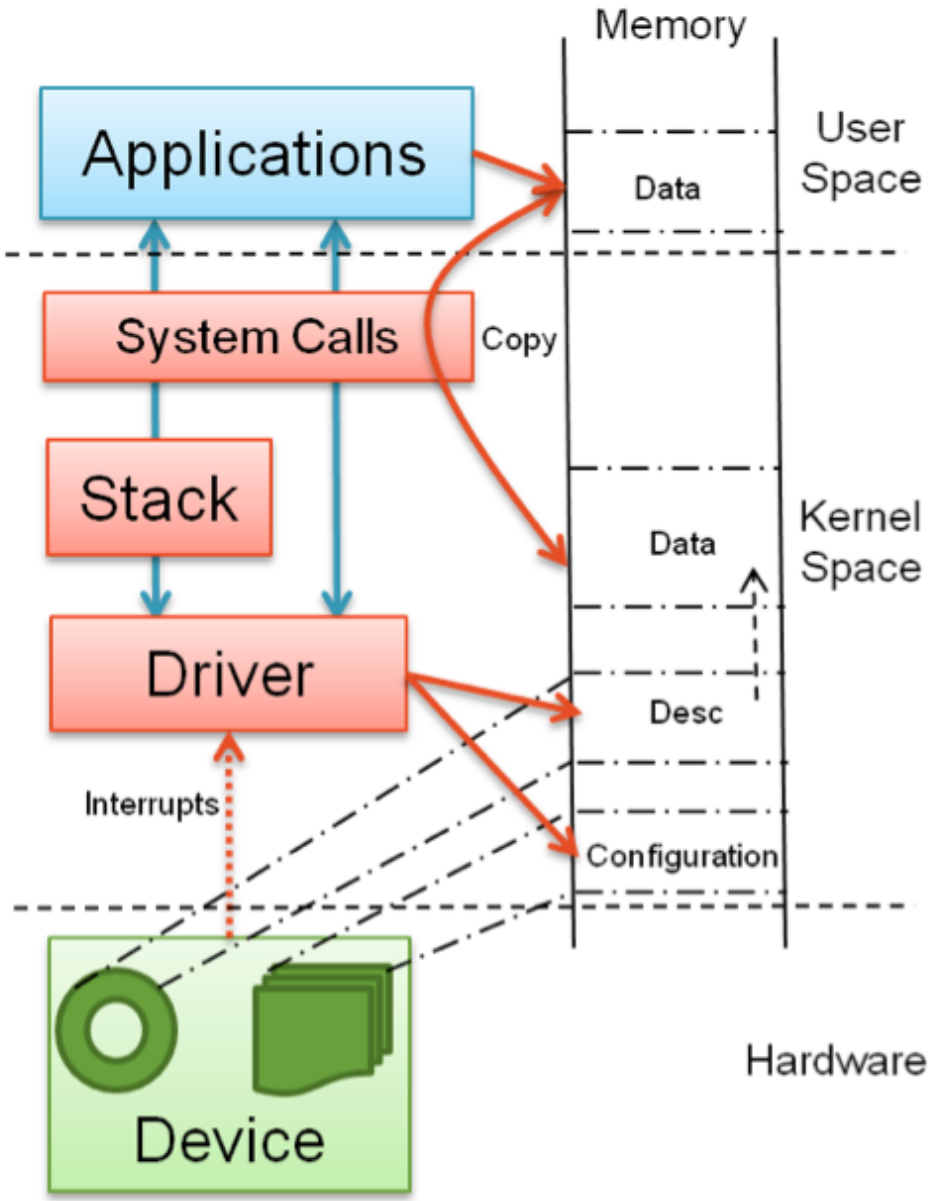

User Space
Kernel
Application process
Thread 0
Thread 1
...
accept()
epoll_wait()
BSD socket
Linux epoll
In-kernel TCP Stack
Kernel TCP stack
ixgbe driver
Linux
Application process
Thread 0
Thread 1
...
mtcp_accept()
mtcp_epoll_wait()
mTCP socket
mTCP epoll
thread 0
thread 1
User-level TCP Stack
User-level TCP stack
User-level packet I/O library
mTCP's approach

User TCP/IP


Throughput (Gbps)
Transactions/sec
Throughput (Gbps)
Transactions/s (x 10^5)
0
1
2
3
4
5
0
0.5
1
1.5
2
2.5
3
3.5
4
Linux
REUSEPORT
MegaPipe
mTCP

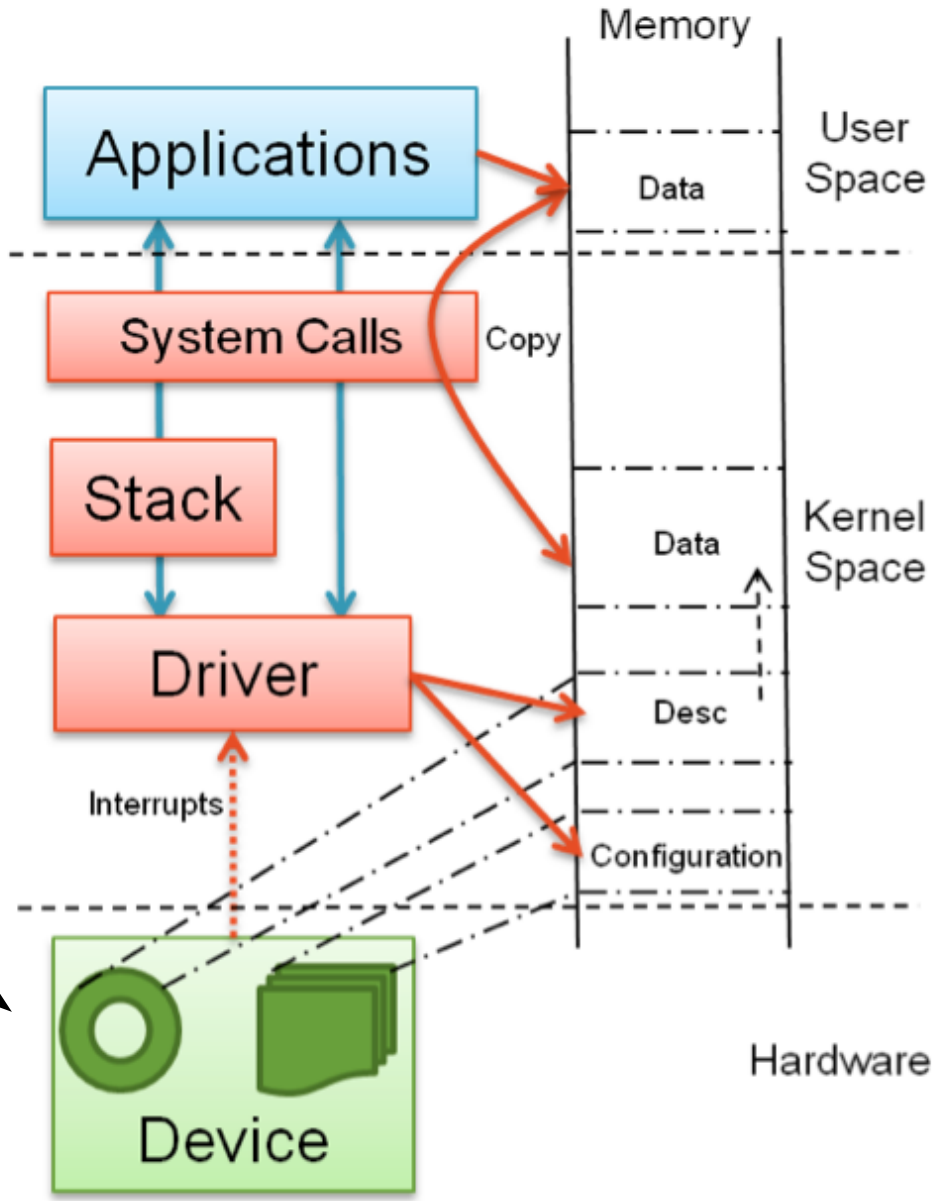
Infiniband & RDMA

TCP/IP over Ethernet:



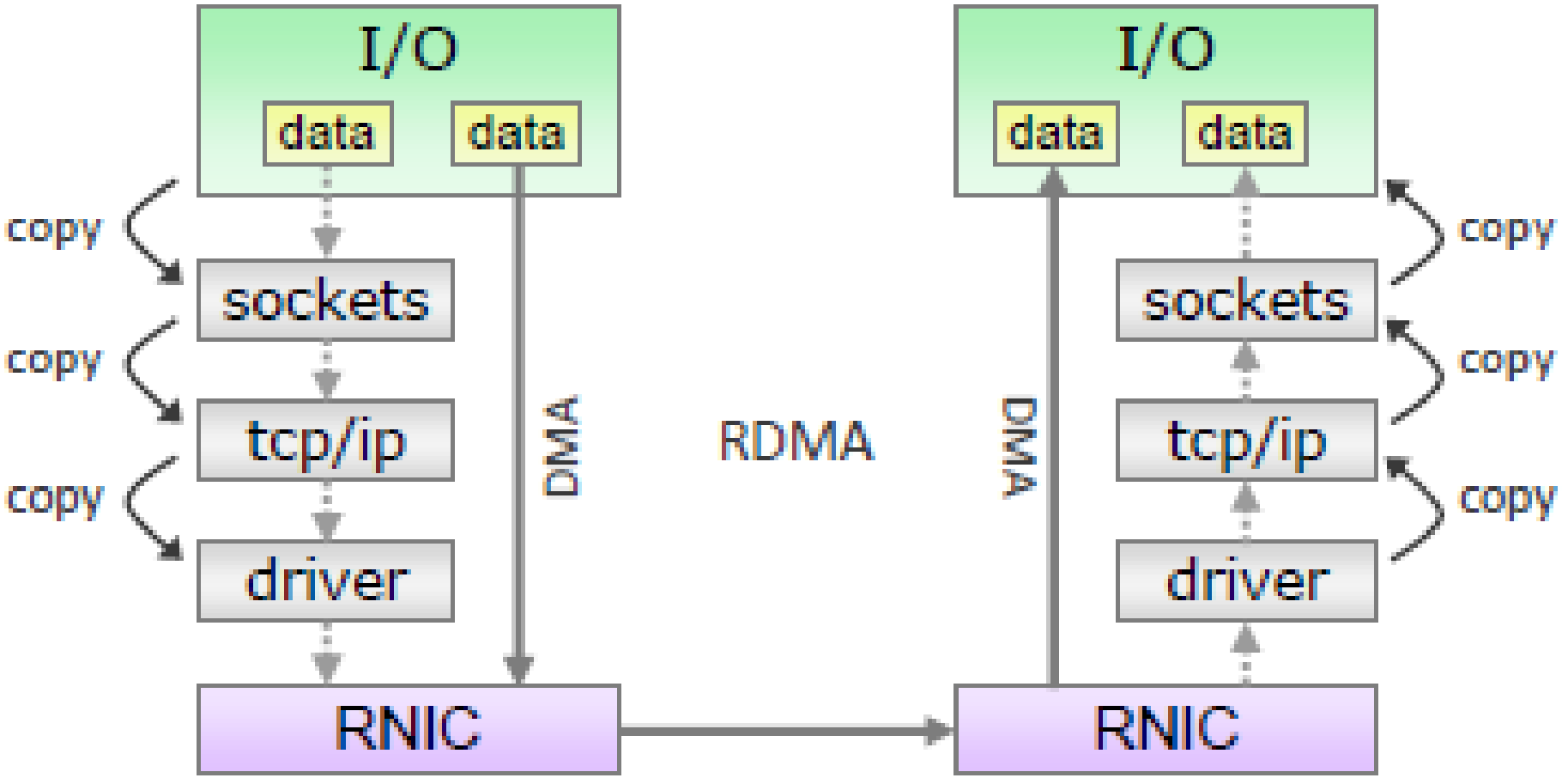
Infiniband & RDMA

TCP/IP over Ethernet:

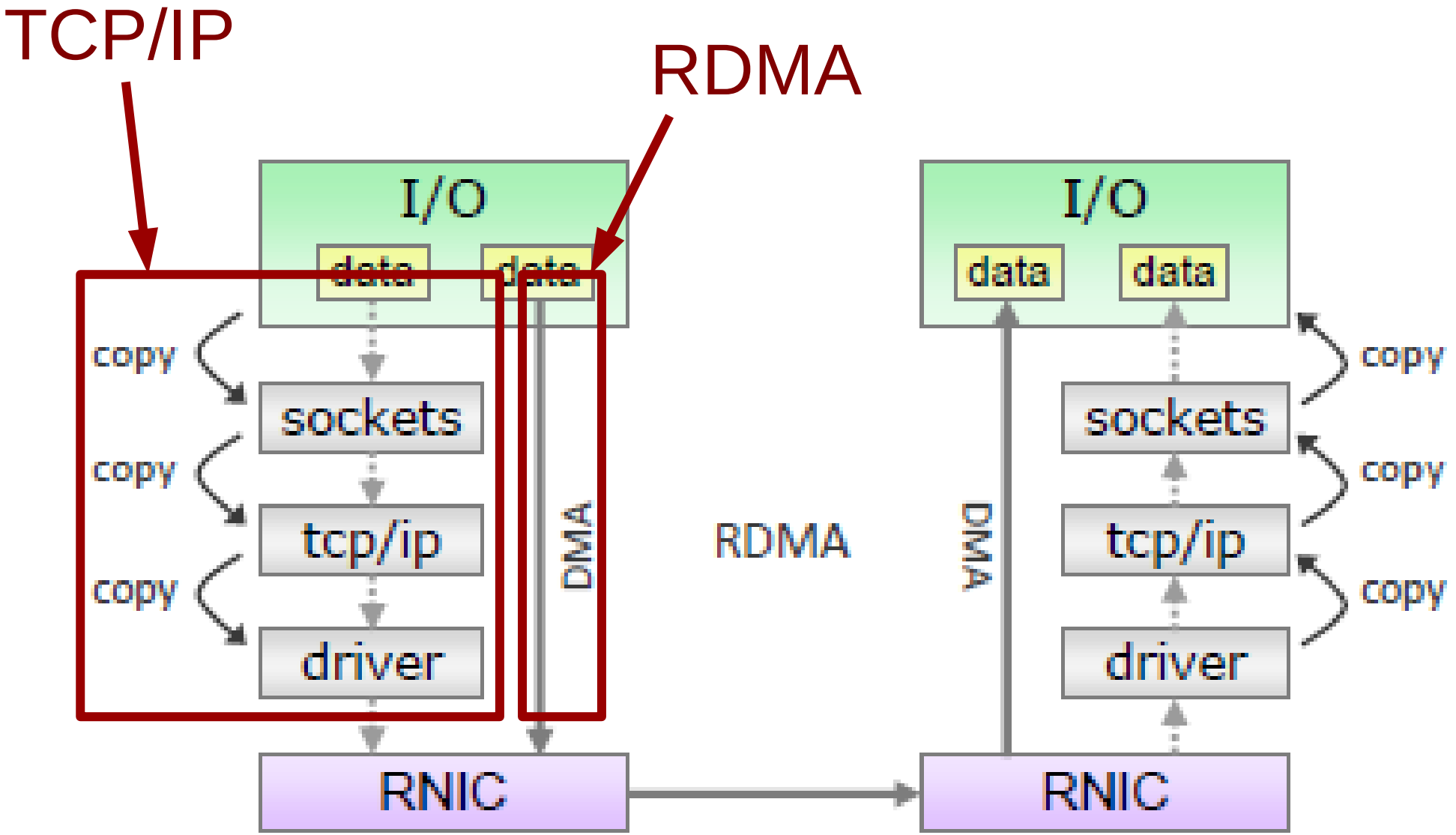


внутри тоже буферы

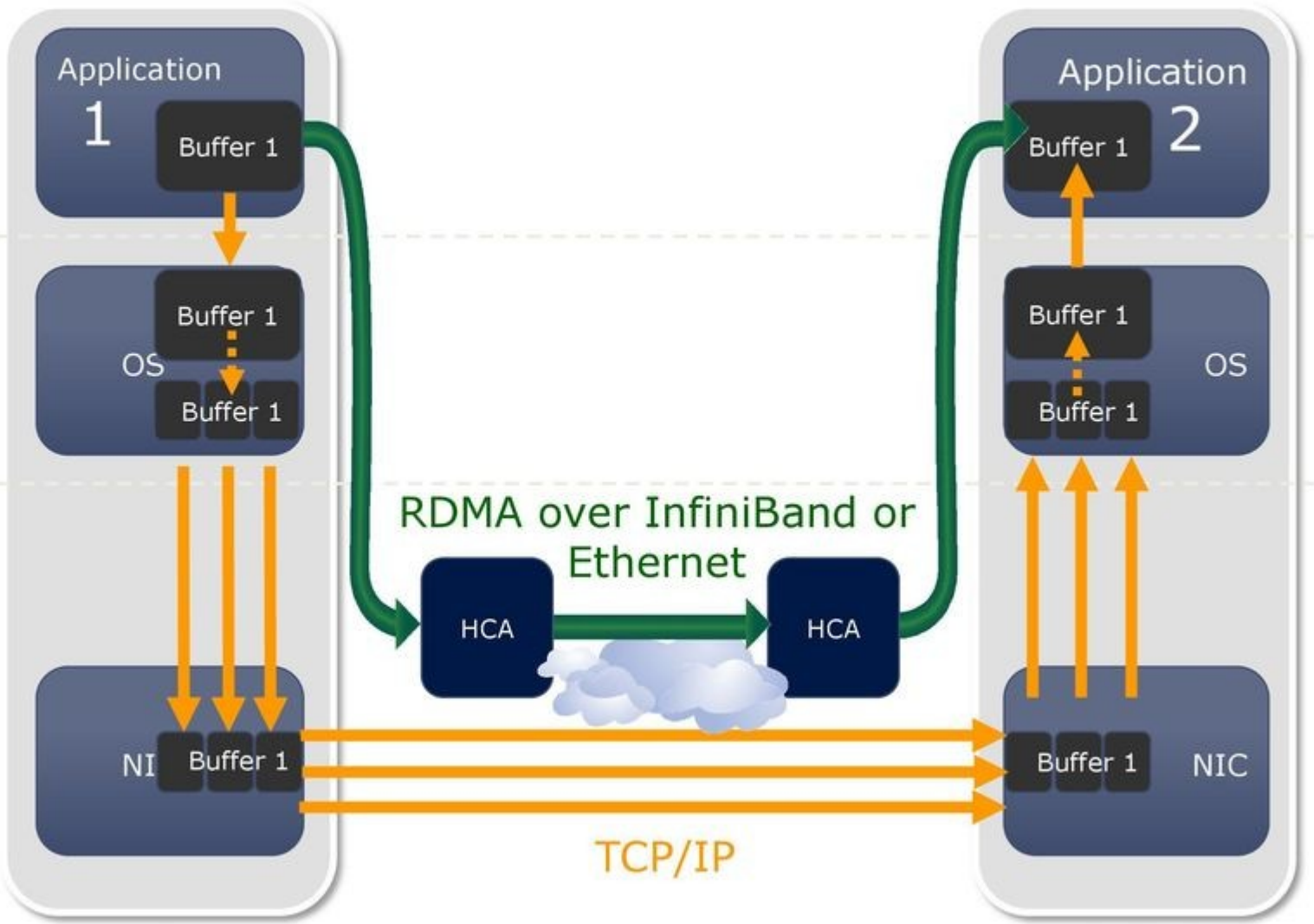
Infiniband & RDMA



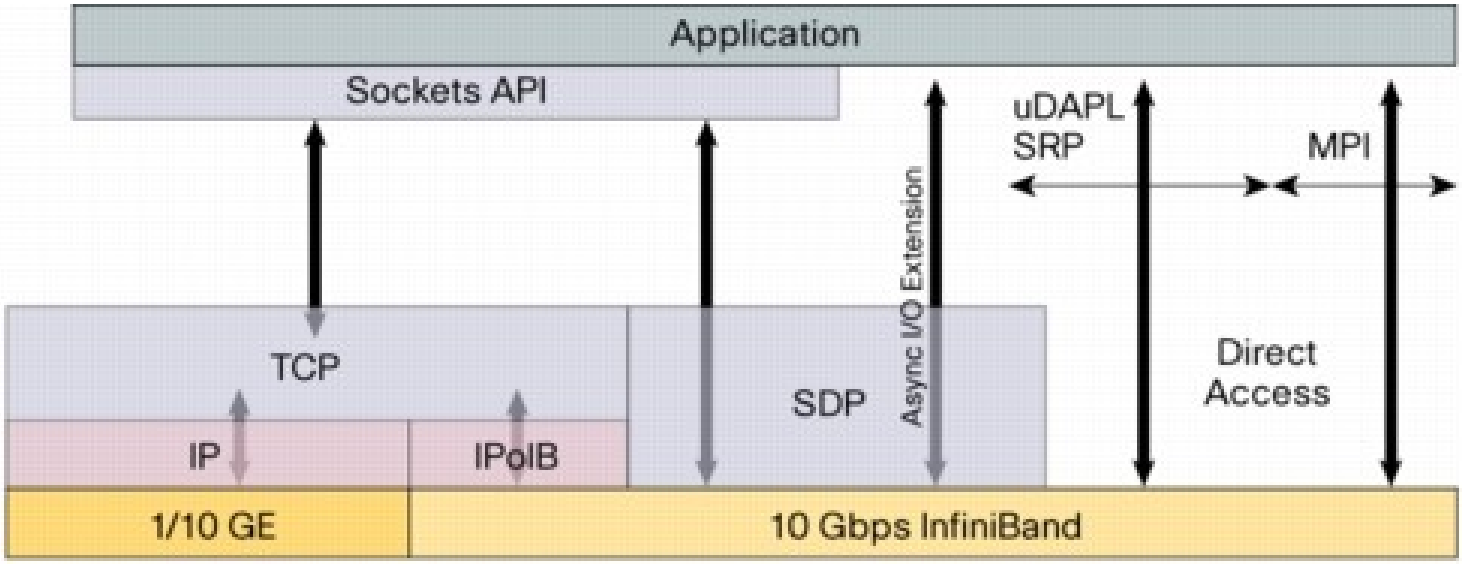
Infiniband & RDMA



Infiniband & RDMA



Infiniband & RDMA



Throughput	1 Gbps	3.9 Gbps	4.1 Gbps	4.5 Gbps	7.9 Gbps	8 Gbps	8 Gbs
Latency	40 usec	40 usec	20 usec	15 usec	15 usec	8 usec	4.5 usec
CPU Utilization	30%	80%	85%	25%	4%	1%	>1%

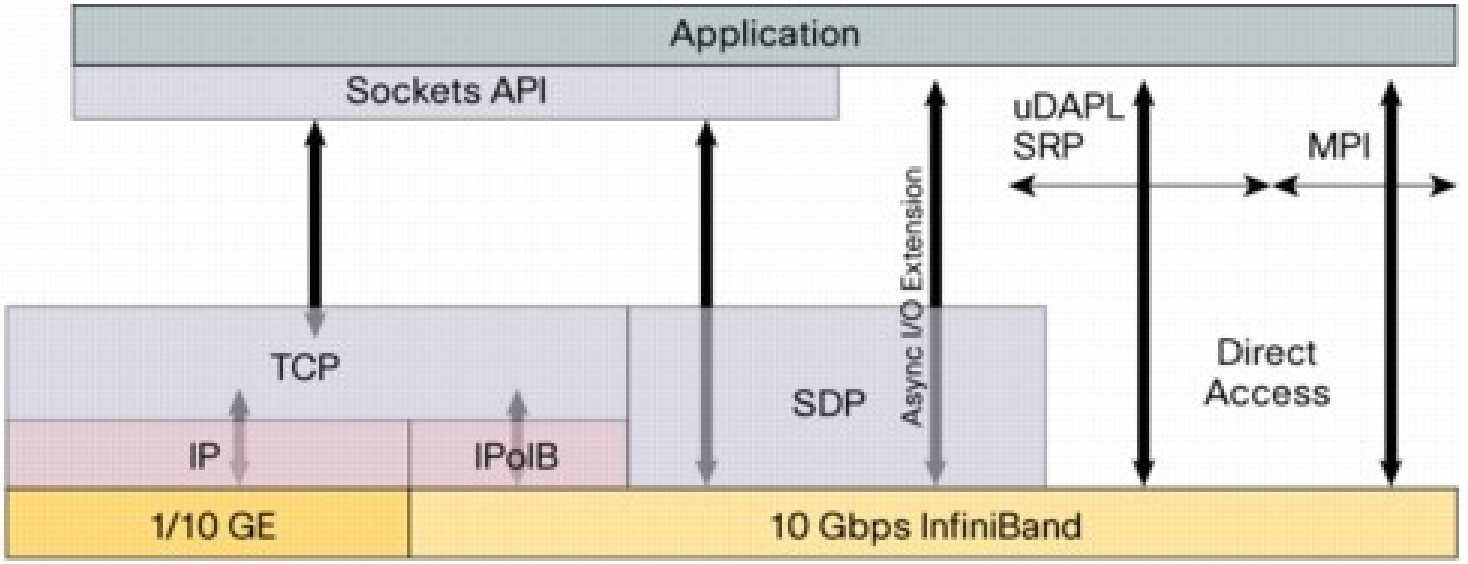
* данные за 2006-ой год

MPI: Message Passing Interface

SRP: SCSI Remote Protocol

uDAPL: User-level Direct Access Programming Language

Infiniband & RDMA



Throughput	1 Gbps	3.9 Gbps	4.1 Gbps	4.5 Gbps	7.9 Gbps	8 Gbps	8 Gbs
Latency	40 usec	40 usec	20 usec	15 usec	15 usec	8 usec	4.5 usec
CPU Utilization	30%	80%	85%	25%	4%	1%	>1%

MPI: Message Passing Interface

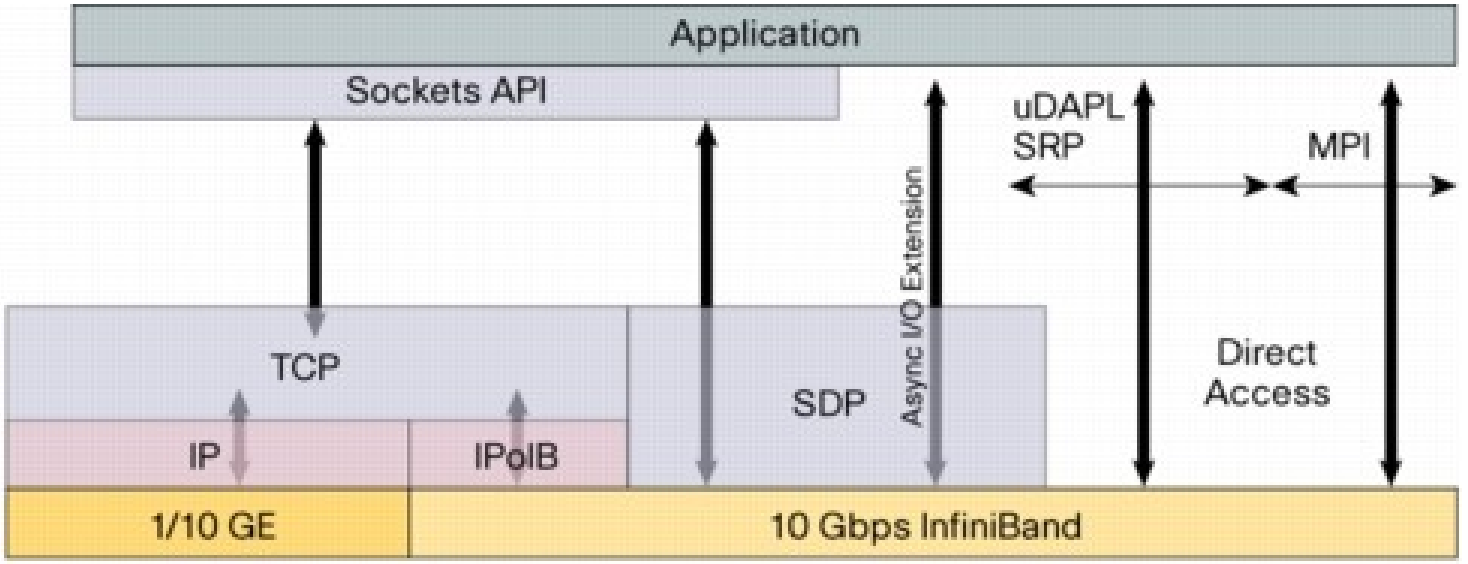
SRP: SCSI Remote Protocol

uDAPL: User-level Direct Access Programming Language

FDR: 700ns

EDR: 500ns

Infiniband & RDMA



Throughput	1 Gbps	3.9 Gbps	4.1 Gbps	4.5 Gbps	7.9 Gbps	8 Gbps	8 Gbs
Latency	40 usec	40 usec	20 usec	15 usec	15 usec	8 usec	4.5 usec
CPU Utilization	30%	80%	85%	25%	4%	1%	>1%

MPI: Message Passing Interface

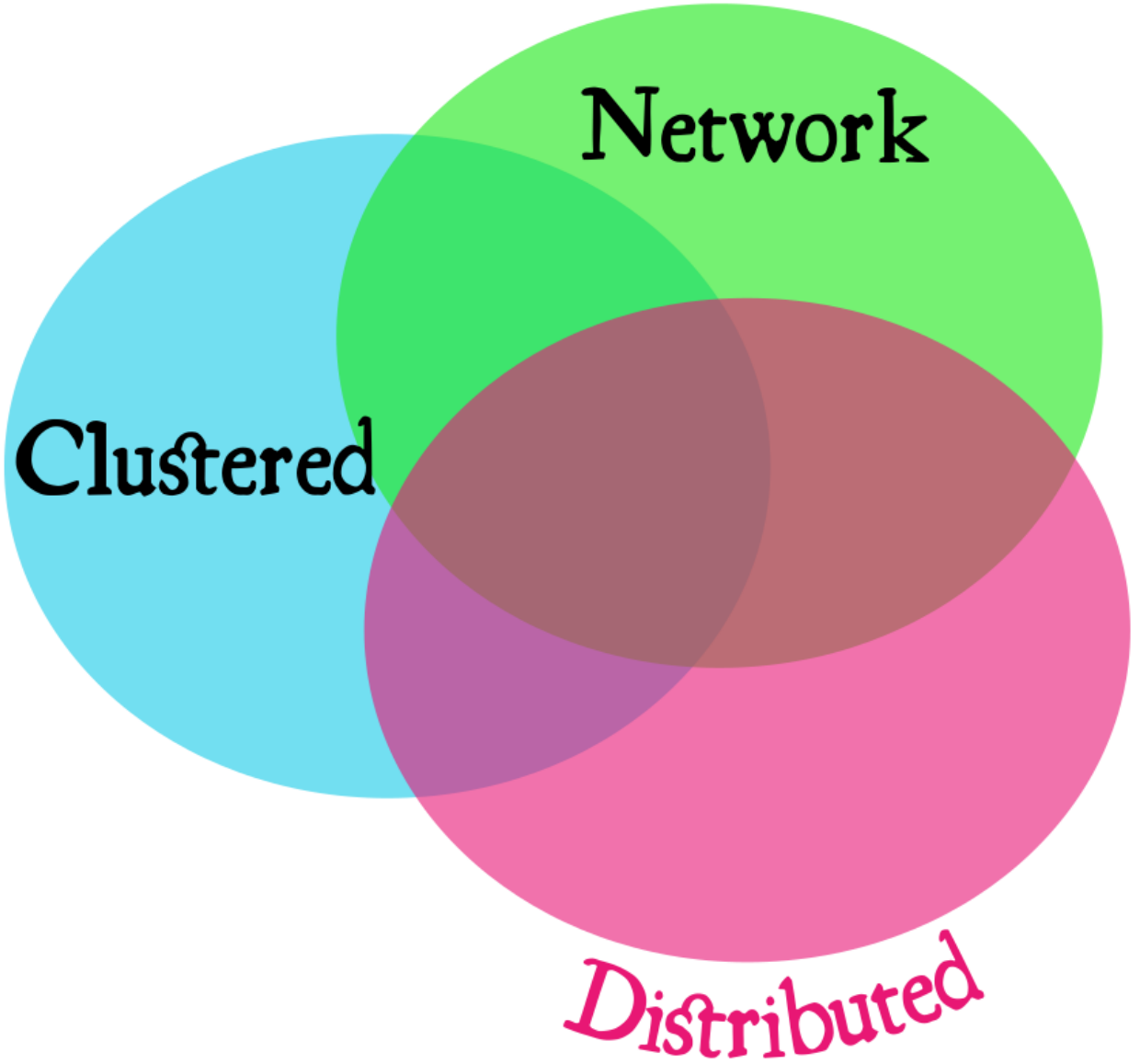
SRP: SCSI Remote Protocol

uDAPL: User-level Direct Access Programming Language

Latencies

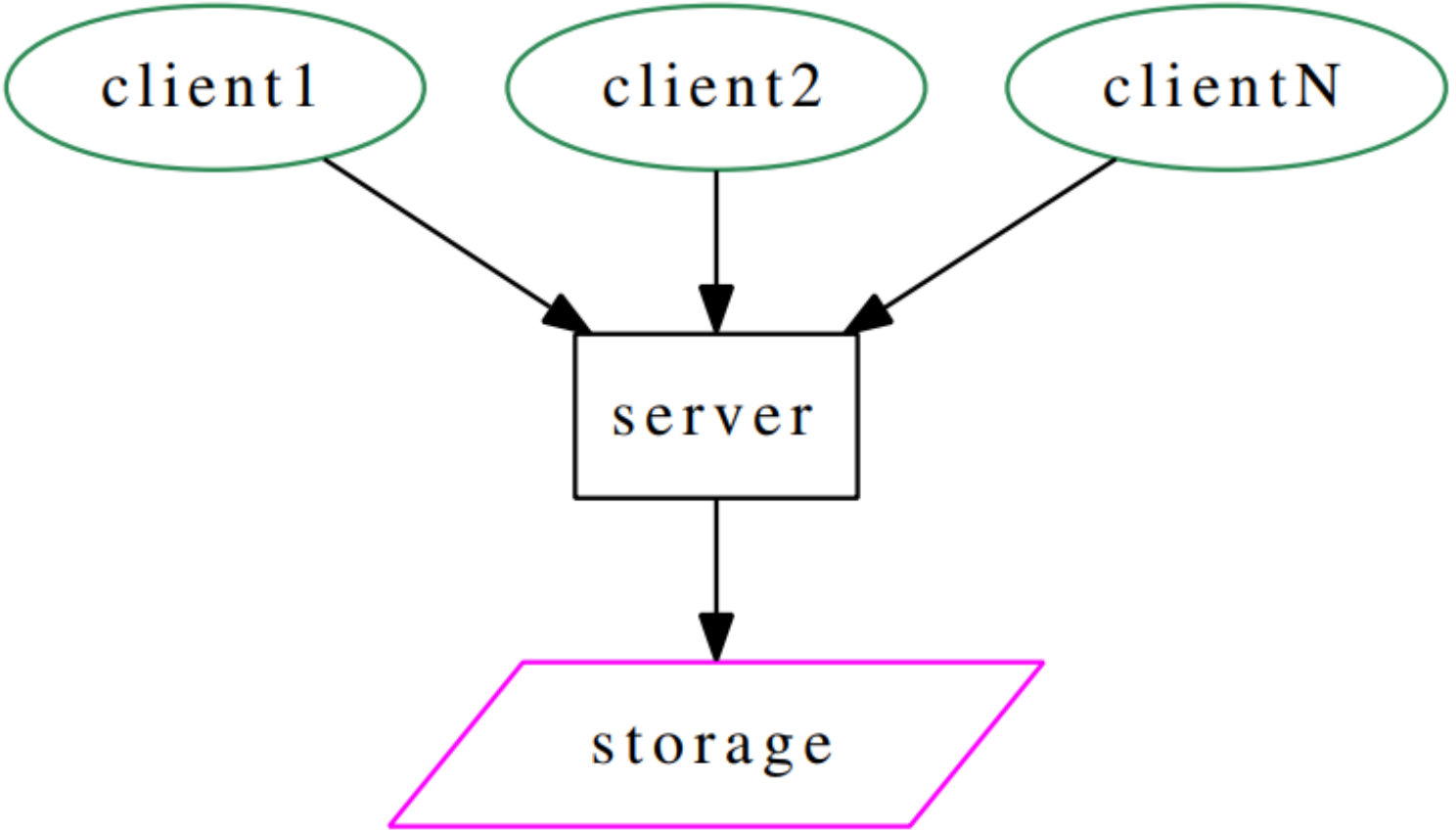
Event	Latency	Scaled
1 CPU cycle	0.3 ns	1 s
Level 1 cache access	0.9 ns	3 s
Level 2 cache access	2.8 ns	9 s
Level 3 cache access	12.9 ns	43 s
Main memory access (DRAM, from CPU)	120 ns	6 min
Solid-state disk I/O (flash memory)	50–150 µs	2–6 days
Rotational disk I/O	1–10 ms	1–12 months
Internet: San Francisco to New York	40 ms	4 years
Internet: San Francisco to United Kingdom	81 ms	8 years
Internet: San Francisco to Australia	183 ms	19 years
TCP packet retransmit	1–3 s	105–317 years
OS virtualization system reboot	4 s	423 years
SCSI command time-out	30 s	3 millennia
Hardware (HW) virtualization system reboot	40 s	4 millennia
Physical system reboot	5 m	32 millennia

File systems in HPC



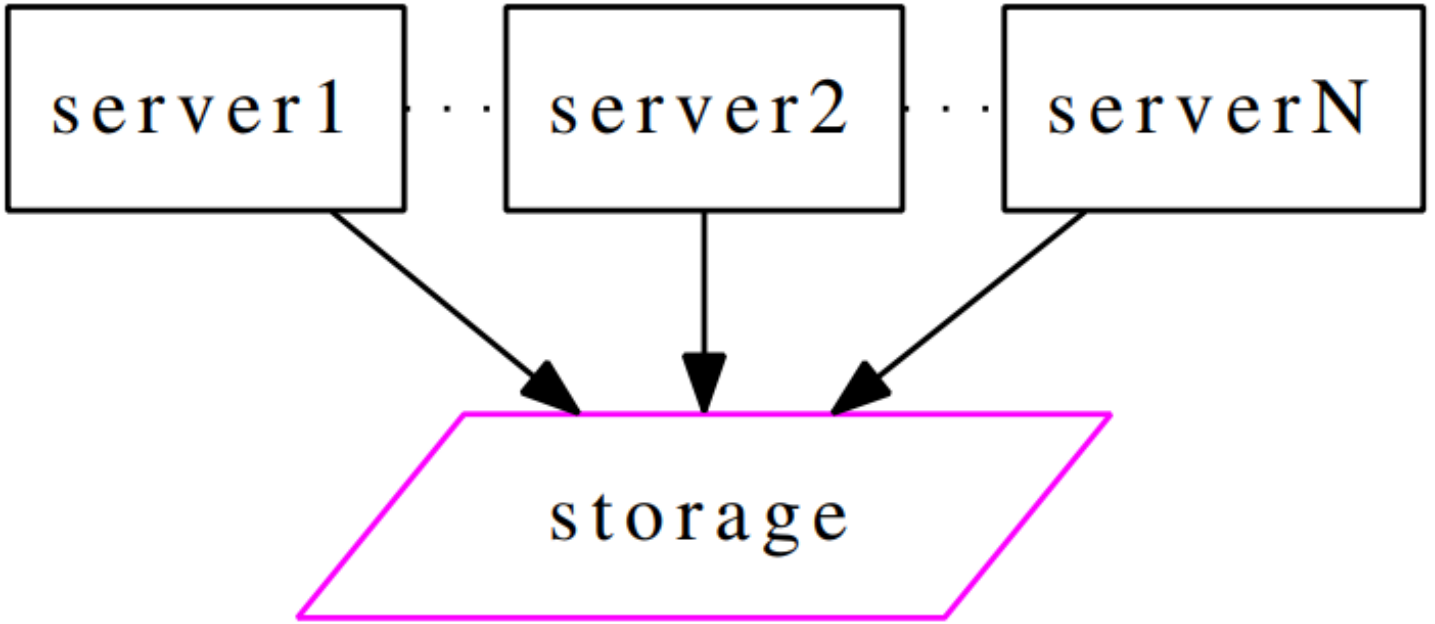
File systems in HPC

Network file system:



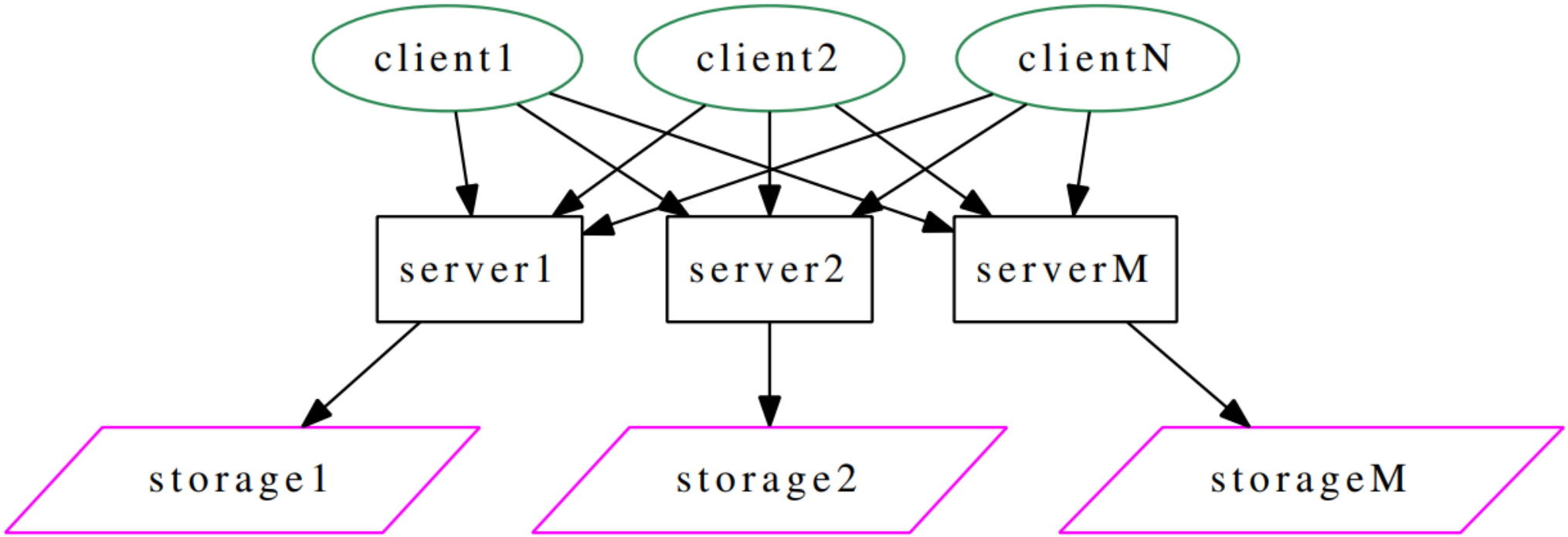
File systems in HPC

Clustered file system:

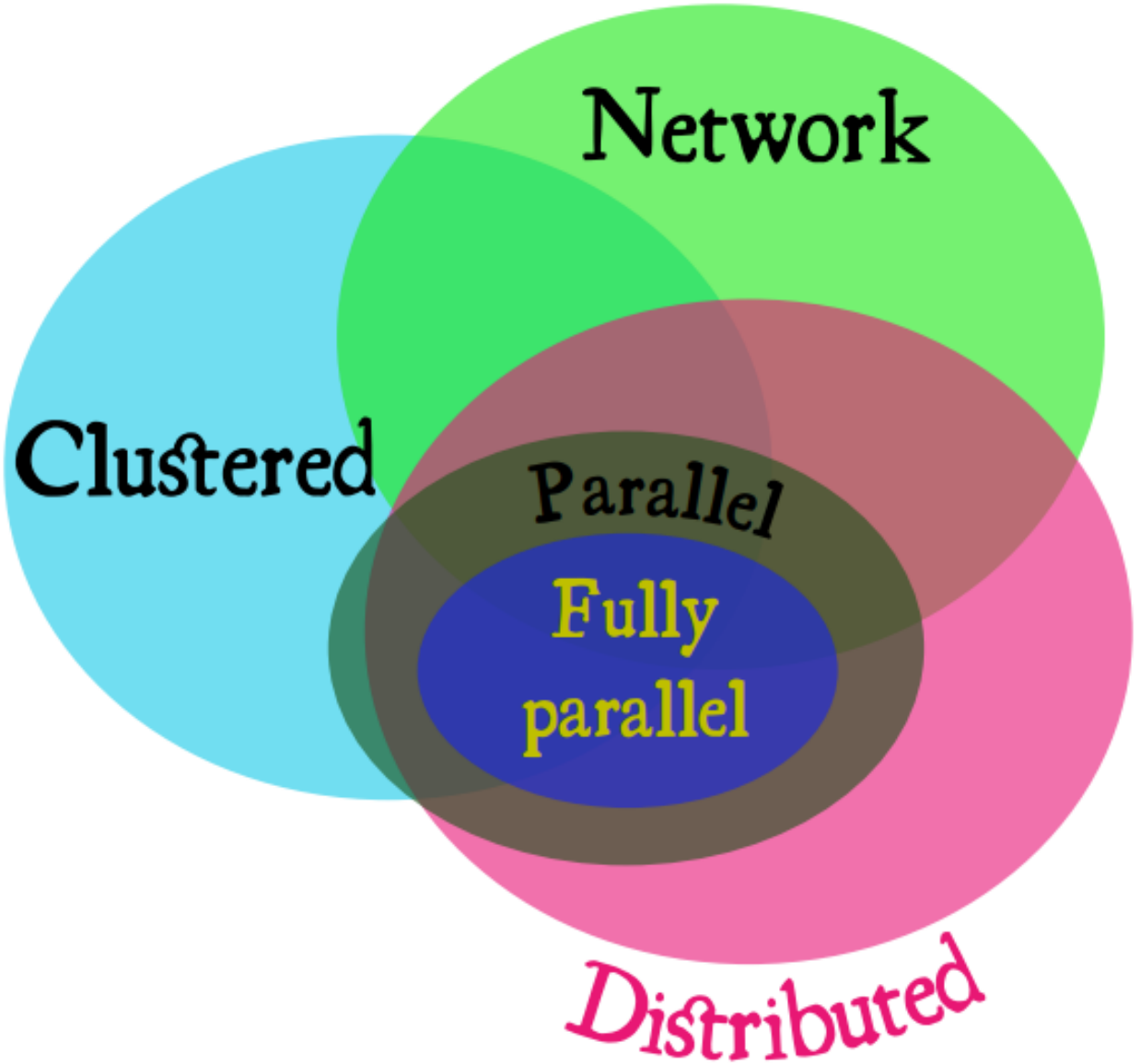


File systems in HPC

Distributed file system:



File systems in HPC



File systems in HPC

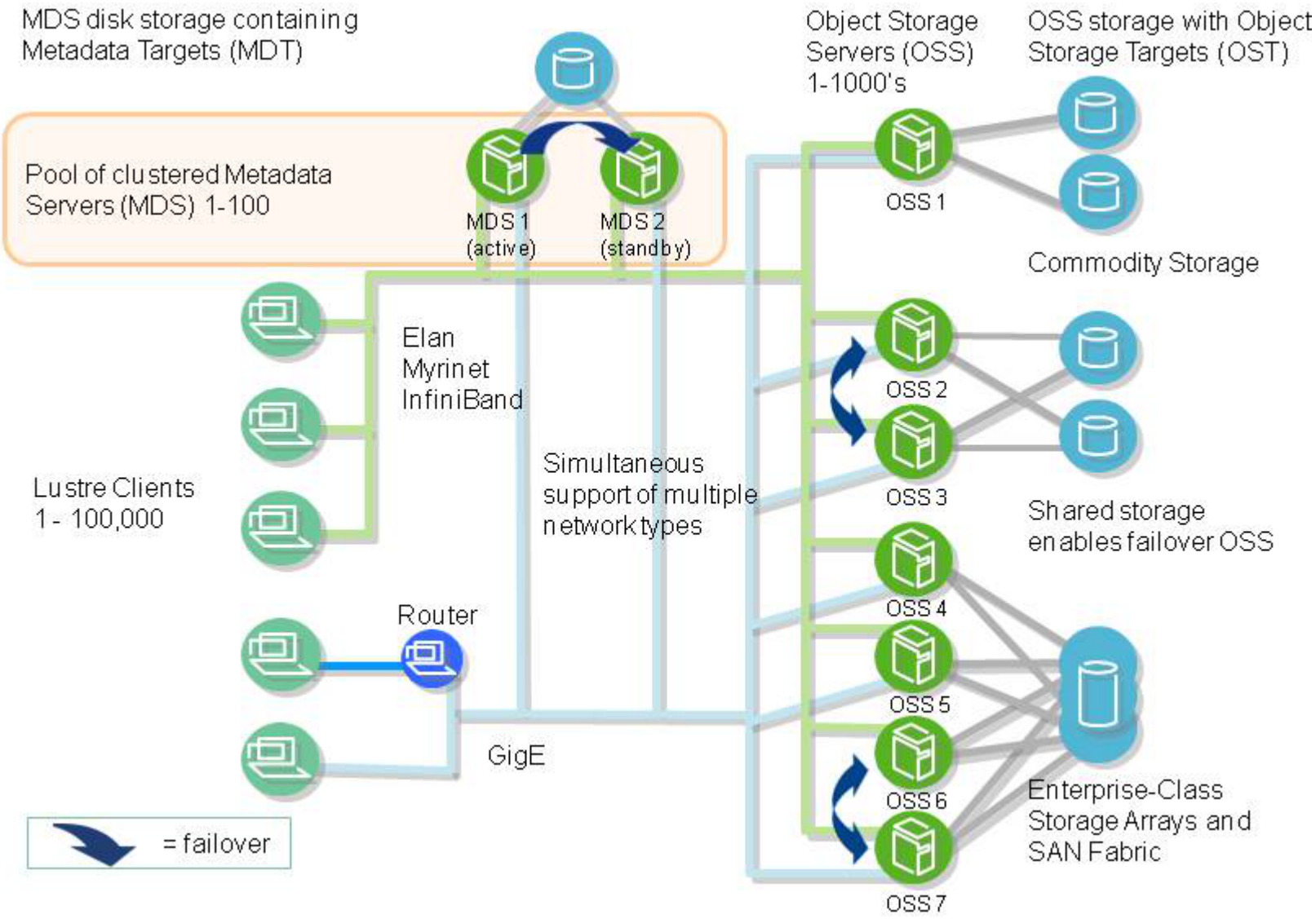
Parallel file system:

- Параллельный доступ ко всем серверам
- Параллельный доступ к данным
- Распределение нагрузки

Fully parallel file system:

+ Параллельный доступ к метаданным

File systems in HPC



Q&A