Networking for DAQ

Silvia Fressard-Batraneanu

CERN

ISOTDAQ 2017, Amsterdam

**Courtesy of Dan Savu and Stefan Stancu*

Outline

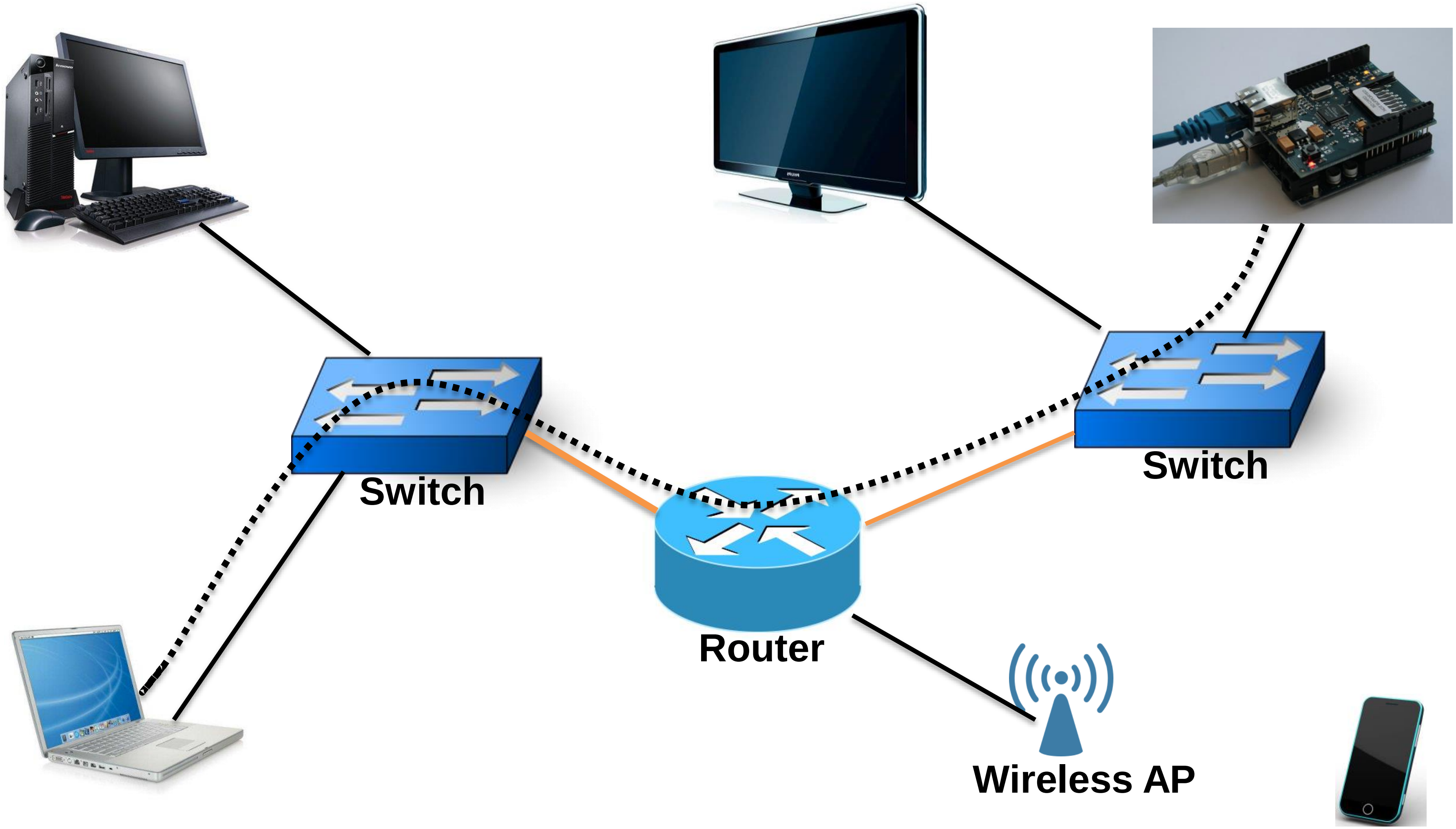
- Networking basics
- OSI reference model
- OSI Layer 1 + 2 : Ethernet and VLANs
- OSI Layer 3 : IP, ARP, routers and routing
- OSI Layer 4 : TCP and UDP protocols
- Data encapsulation efficiency
- Quality of Service
- Network monitoring
- Networks for DAQ: characteristics and optimizations

What is a network ?

- A ***network*** is simply two or more devices connected together so they can exchange information. At the same time it can be a complex interconnected system such as the Internet.

- ***End-host devices*** are devices attached to a network
- A ***source host*** is the place where the data originally comes from
- A ***destination host*** is the place where the data is being sent to
- ***Networking devices*** are waypoints along paths for data to travel along
- **Links** are direct data paths between adjacent devices
- A **route** is the path between any two network points

What is a network ?



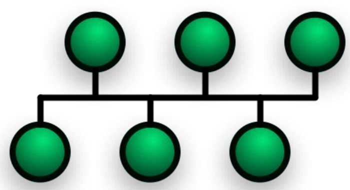
Network types

Networks come in many flavors to suit different purposes and needs

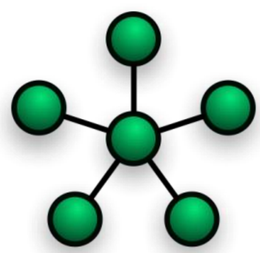
- **LAN** (small size, high speed, physical proximity)
- **WAN** (long distance, lower data transfer rates)
- **MAN** (metropolitan area network)
- **SAN** (connecting storage farms, lossless, high speed)
- **VPN** (private network extension across a shared or a public network)

Network structure

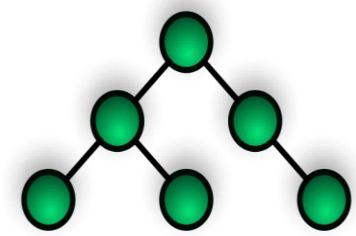
- The structure of a network is known as the ***topology***
  - **Physical** = The way the network is cabled
  - **Logical** = The way devices use the network to communicate



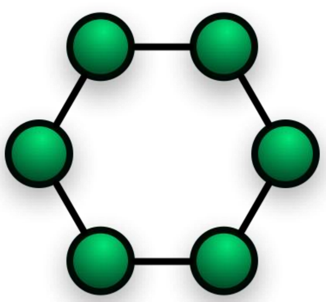
Bus Topology



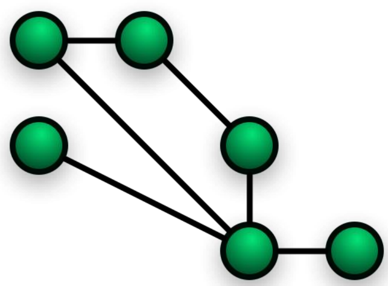
Star Topology



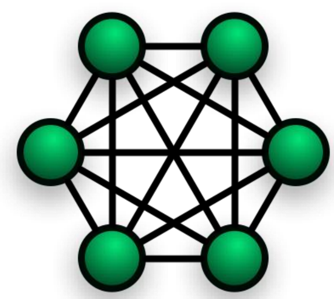
Hierarchical Topology



Ring Topology



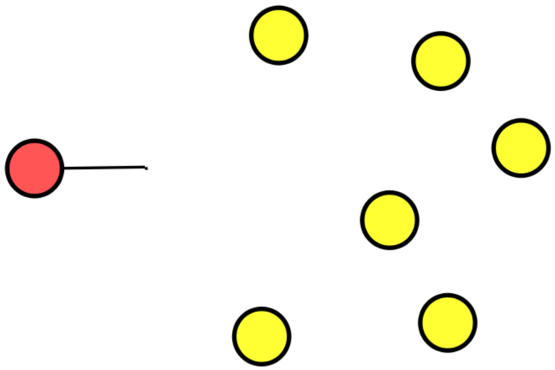
Partial Mesh Topology

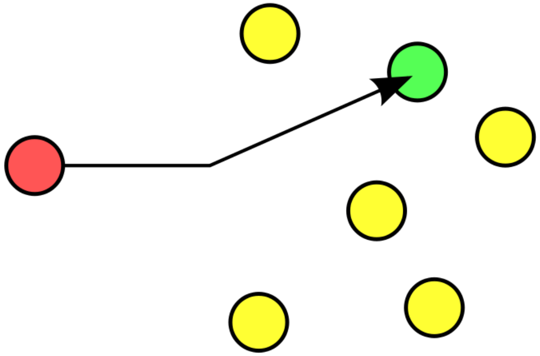


Fully Mesh Topology

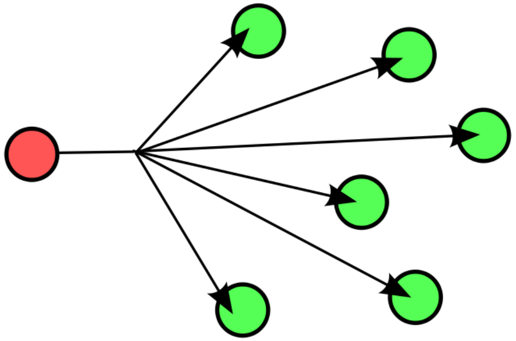
Network communication patterns

- One-to-one
- One-to-all
- One-to-many

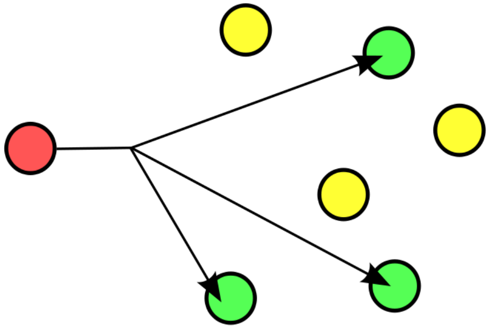




Unicast



Broadcast



Multicast

OSI Model. Divide et impera.

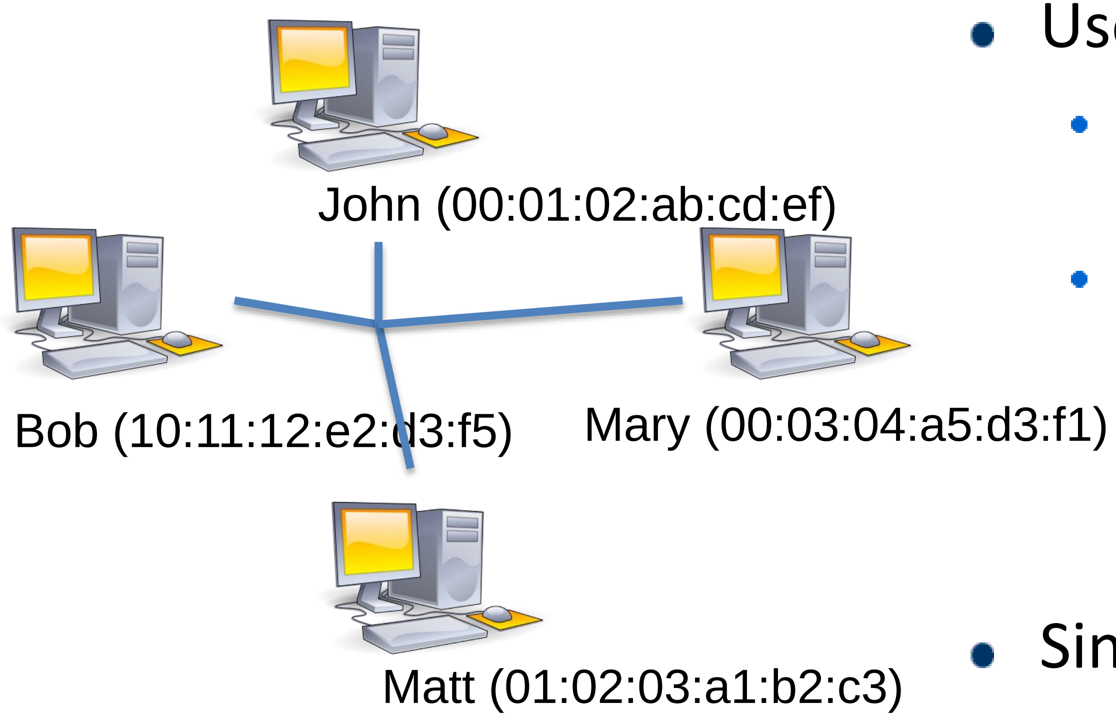
- Application
- Presentation
- Session
- Transport
- Network
- Data Link
- Physical

vs

1011101010000111100101010100011

All **P**eople **S**eem **T**o
Need **D**ata **P**rocessing

Ethernet

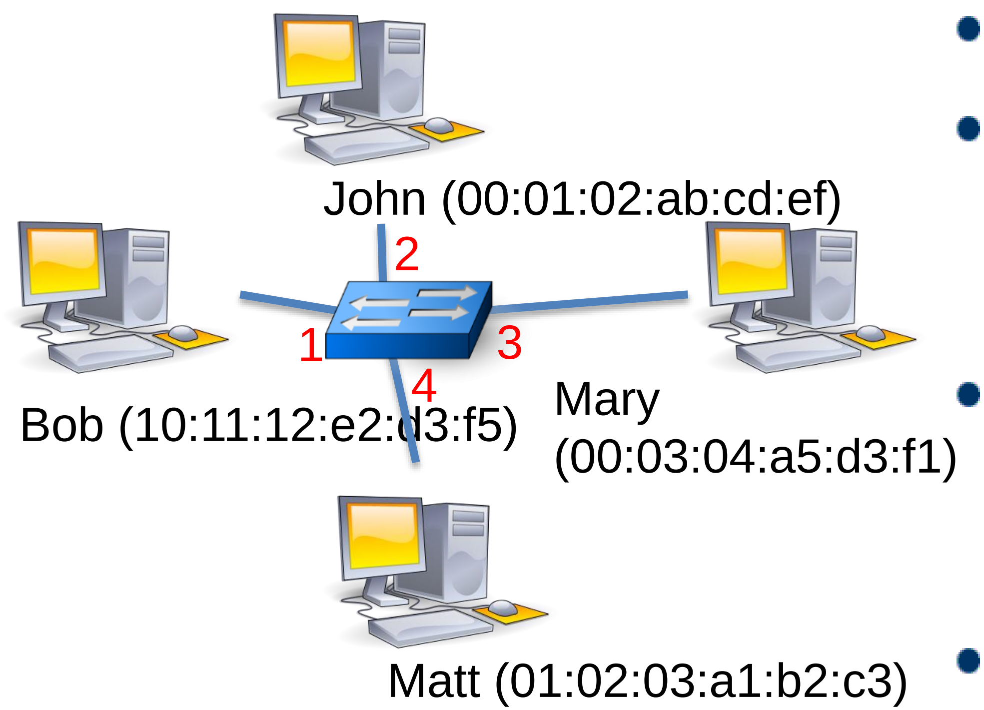


8	6	6	2	46 ~ 1500 bytes	4
Pre-amble	Dest.	Source	Type/ Length	Data	Frame check

Basic Ethernet frame

- Used for the first two layers
  - Defines wiring and signaling standards (Layer1)
  - Defines a flat addressing schema with local visibility, called **MAC** (Layer 2) -> *MAC address on 48 bits, usually in hex format*
- Single broadcast domain
- Frame based technology

Ethernet. Switch

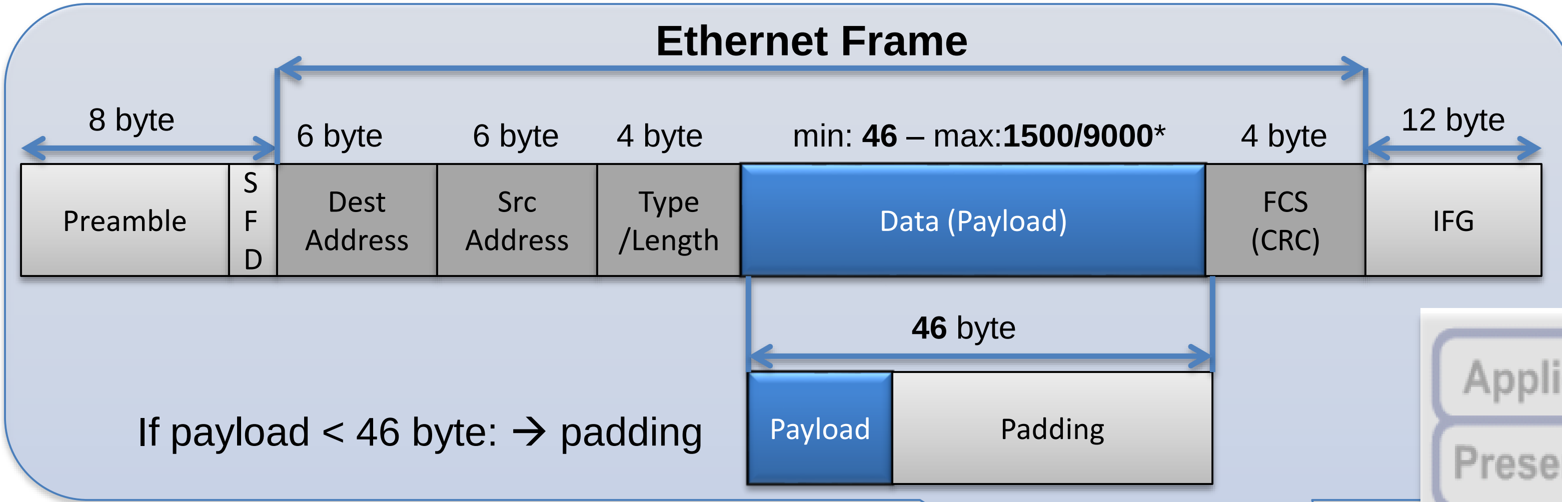


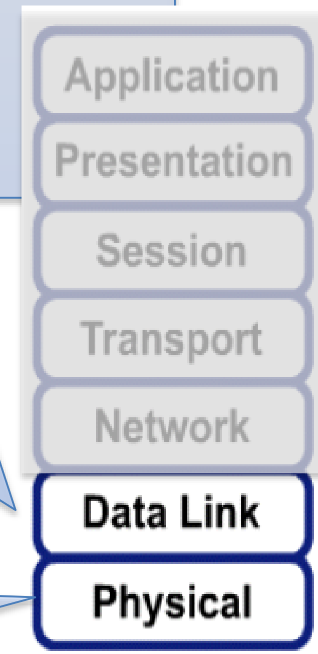
- Layer 2 device
- *Switches* frames to the correct segment using MAC addresses
- *Learns* the MAC addresses by storing them in a dedicated table
- *Floods* the network before learning

MAC address table

Port	MAC Address
2	00:01:02:ab:cd:ef(John)
4	01:02:03:a1:b2:c3(Matt)
....	.....

Ethernet





All flavors of media and speeds:

- … even slower but this is now history
- 100 Mbit/s: copper (UTP), fiber
- 1 Gbit/s: copper (UTP), fiber
- 10 Gbit/s: fiber, copper (twinax, UTP)
- 40 Gbit/s: fiber
- 100 Gbit/s: fiber

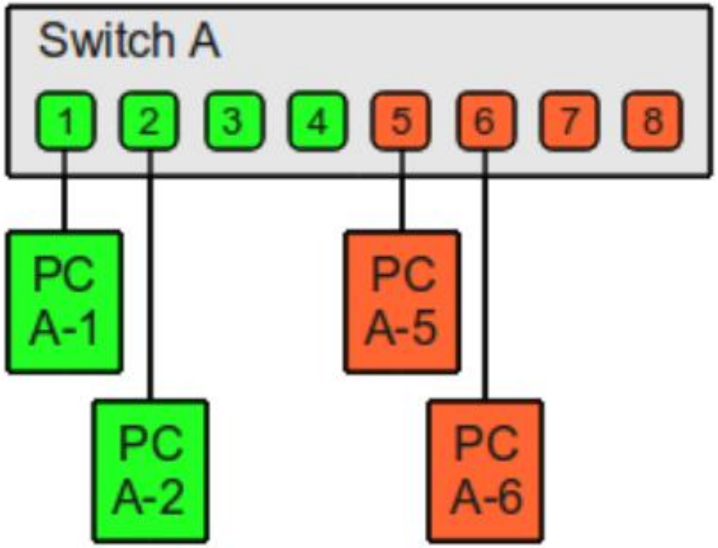
Ethernet Standards

The Evolution of Ethernet Standards to Meet Higher Speeds				
Date	**IEEE Std.**	**Name**	**Data Rate**	**Type of Cabling**
1990	802.3i	10BASE-T	10 Mb/s	Category 3 cabling
1995	802.3u	100BASE-TX	100 Mb/s*	Category 5 cabling
1998	802.3z 802.3z	1000BASE-SX 1000BASE-LX/EX	1 Gb/s	Multimode fiber Single mode fiber
1999	802.3ab	1000BASE-T	1 Gb/s*	Category 5e or higher Category
2003	802.3ae 802.3ae	10GBASE-SR 10GBASE-LR/ER	10 Gb/s	Laser-Optimized MMF Single mode fiber
2006	802.3an	10GBASE-T	10 Gb/s*	Category 6A cabling
2015	802.3bq	40GBASE-T	40 Gb/s*	Category 8 (Class I & II) Cabling
2010 2015	802.3ba 802.3ba 802.3bm	40GBASE-SR4/LR4 100GBASE-SR10/LR4/ER4 100GBASE-SR4	40 Gb/s 100 Gb/s 100 Gb/s	Laser-Optimized MMF or SMF Laser-Optimized MMF or SMF Laser-Optimized MMF
2016	SG	Under development	400 Gb/s	Laser-Optimized MMF or SMF
Note: *with auto negotiation				

Virtual LAN(VLAN)

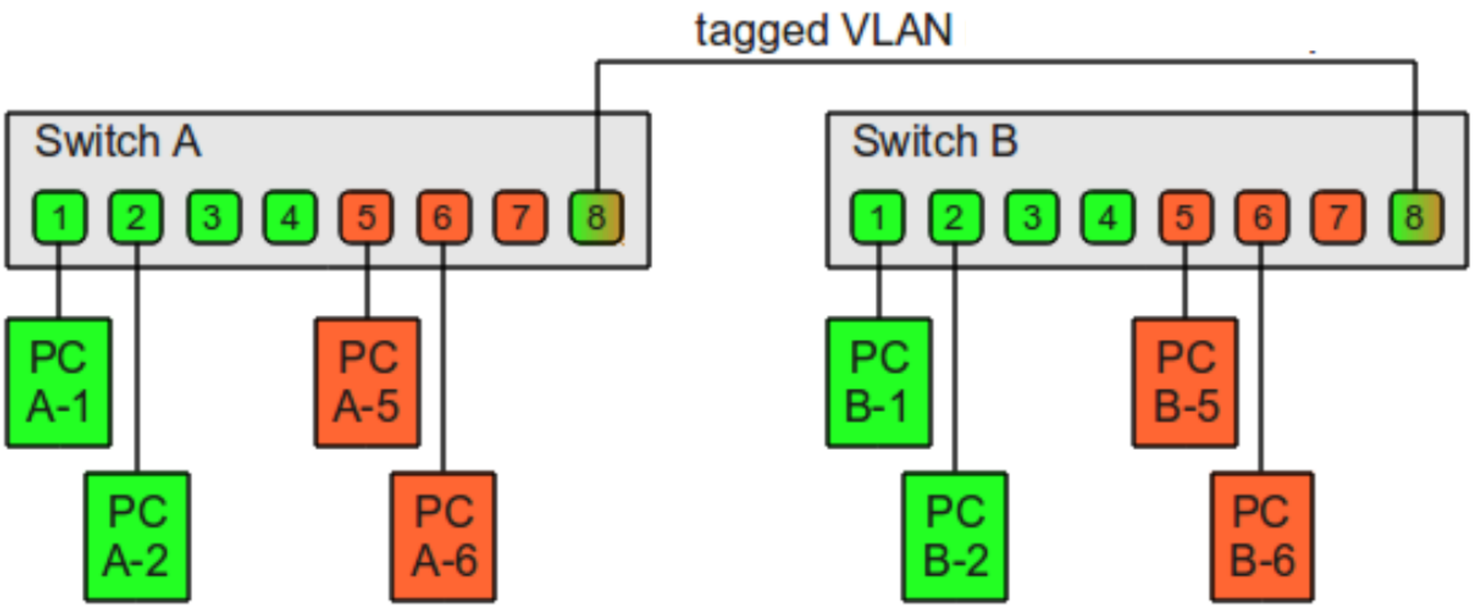
Layer 2 feature which allows the creation of more than one network on a switch and thus:

- Logically grouping host
- Reducing the broadcast domain
- Improving security
- Reducing costs
- Simplifying design and administration



Virtual LAN(VLAN)

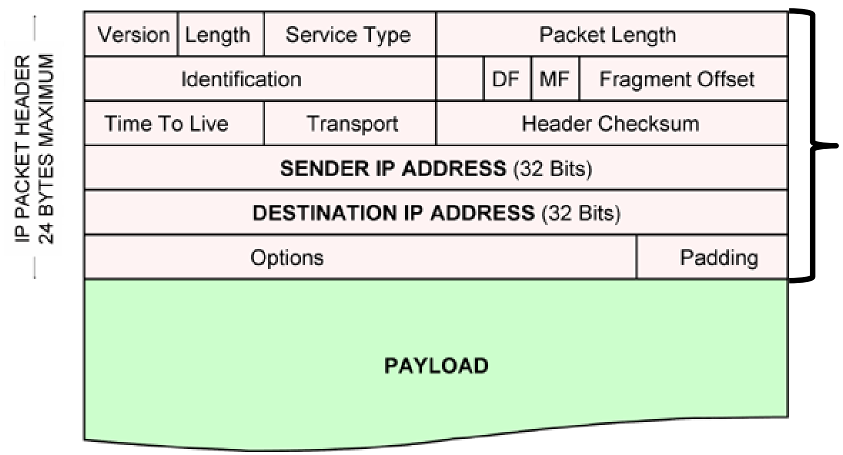
VLANs can span over multiple switches



Ports 1-7 – untagged
Ports 8 – tagged(allowing traffic from both VLANs)

IP protocol

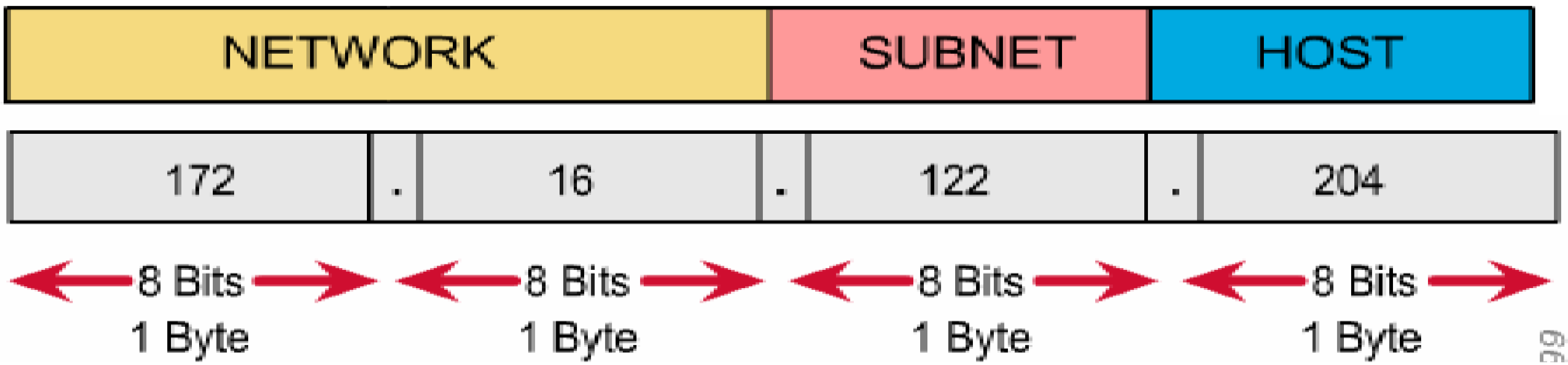
- Connectionless, best effort Layer 3 protocol
- Designed to be encapsulated into layer 2 protocols(such as Ethernet)
- Defines a hierarchical (logical) addressing schema capable of connecting all the hosts in the world
- Routes packets towards destination using best available path with the help of routing protocols



IP Header (20 bytes without options)

IP Addressing

- 32 bit address space (IPv4)
- Hierarchical addressing (similar to postal addressing)
- Global visibility



10101100	00011111	01111010	11001100	
		AND(bitwise)		
11111111	11111111	11111111	00000000	Subnet mask (255.255.255.0)
10101100	00011111	01111010	00000000	Subnet address (172.16.122.0)

Address Resolution Protocol(ARP)

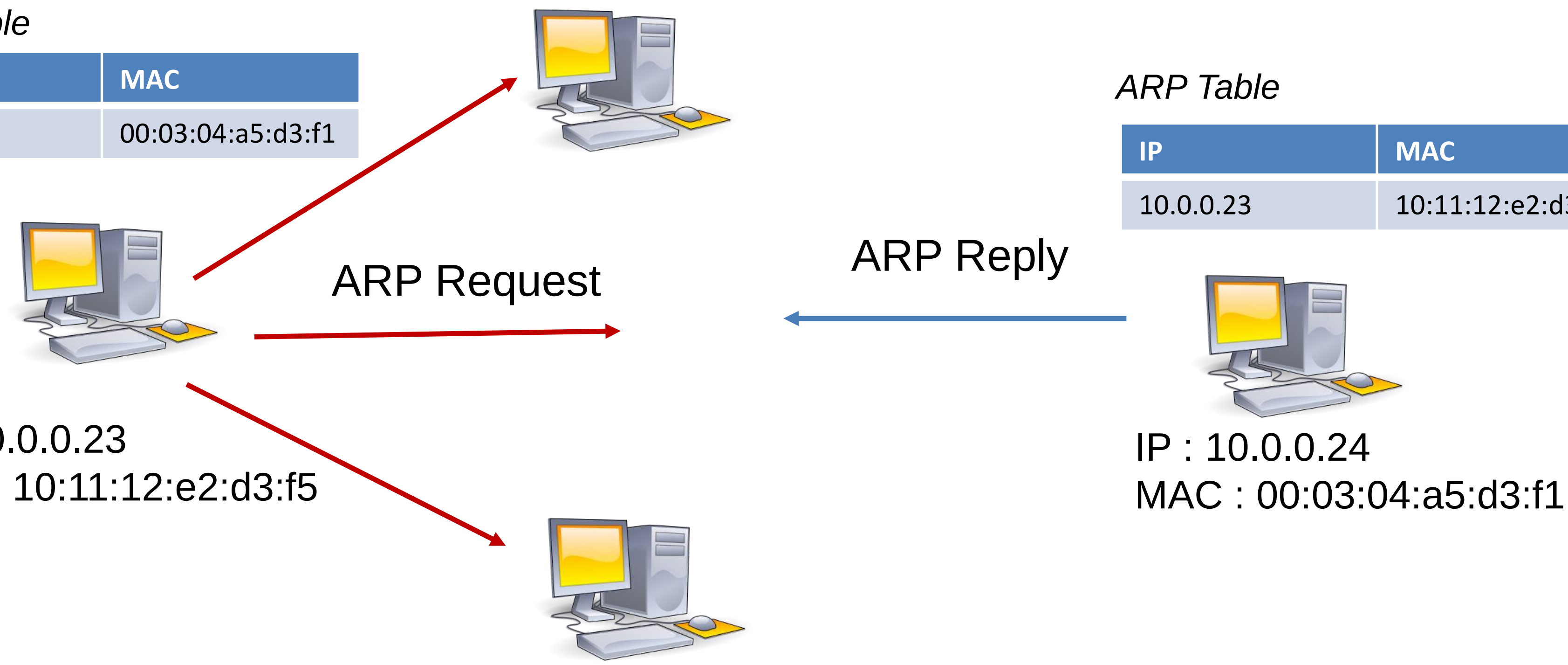
Used to map IP addresses with Ethernet MAC addresses

ARP Table

IP	MAC
10.0.0.24	00:03:04:a5:d3:f1

ARP Table

IP	MAC
10.0.0.23	10:11:12:e2:d3:f5



ARP Request(broadcast) : For the host with IP address 10.0.0.24, please reply with your MAC address to IP 10.0.0.23.

ARP Reply(unicast): I have 10.0.0.24 and I have MAC 00:03:04:a5:d3:f1.

Routers

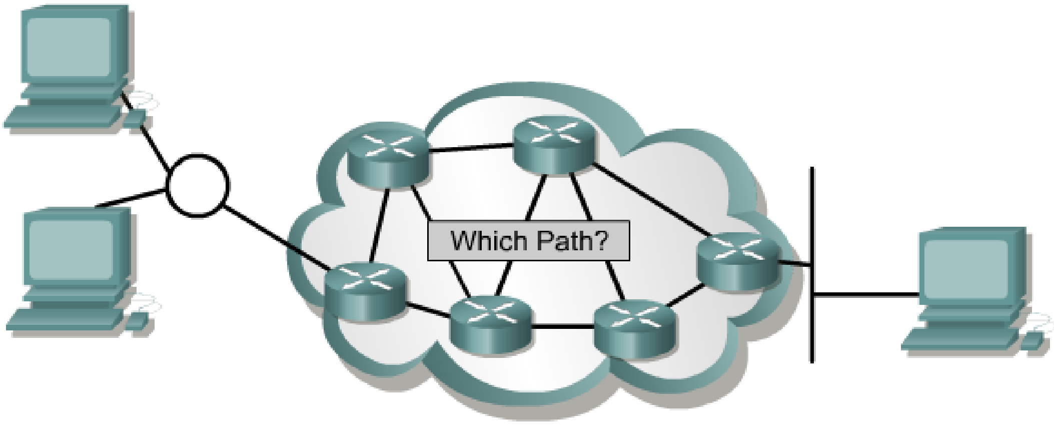
- Layer 3 networking devices
- **Connect** together **separate networks**, sometimes of various networking technologies(Ethernet, ATM, Fiber Channel, etc)
- Make path determination decision based upon logical addresses (such as IP). The process is called **routing**.
- Routing and switching are similar concepts, but are in different layers:
  - Routing occurs in Layer 3, uses IP
    - Maintains routing tables (IP network addresses)
    - Maintains ARP tables (IP to MAC mappings)
  - Switching occurs in layer 2, uses MAC
    - Maintains switching tables (MAC address to port mappings)

Routing

The **process of selecting paths** in a network along which to send network traffic, based upon logical addresses (such as IP).

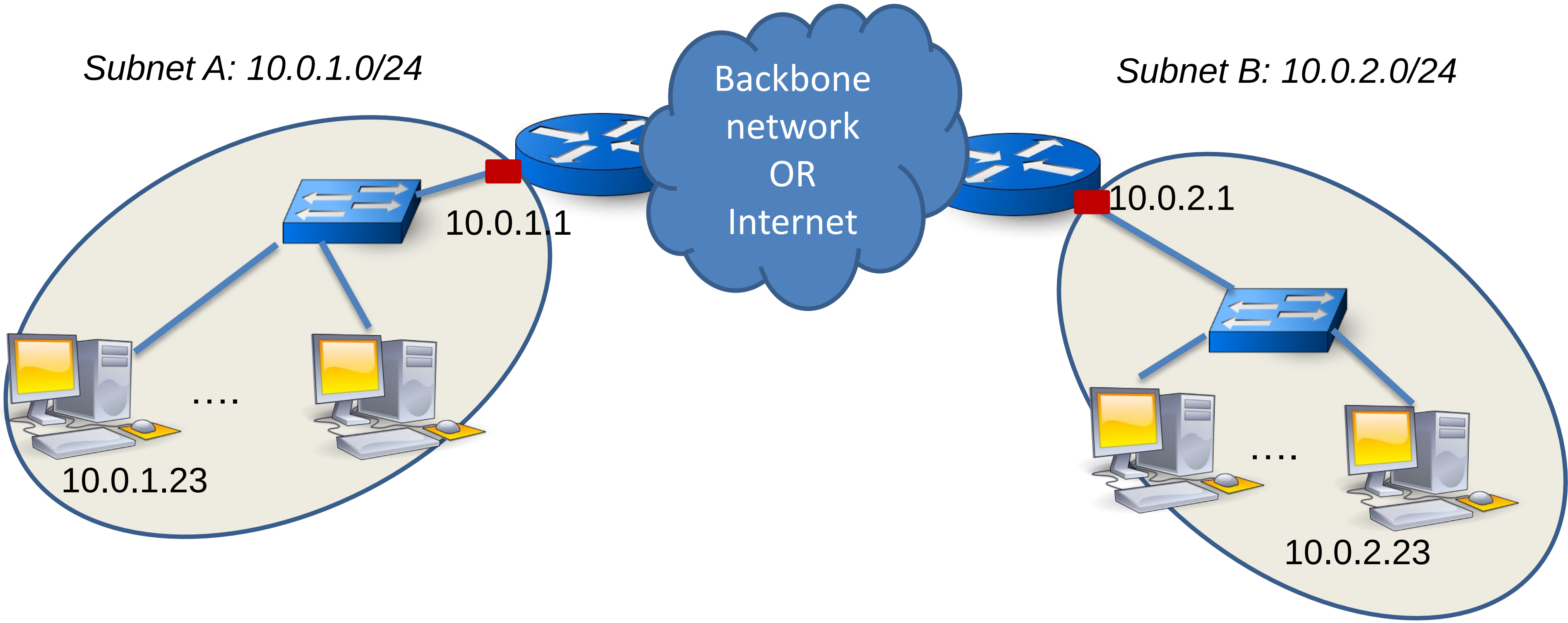
A **routing protocol** allows one router to share information with other routers regarding known network paths as well as its proximity

- Static routing
- Dynamic routing
  - Distance Vector(RIP, IGRP)
  - Link State(OSPF, IS-IS)



IP inter-network communication

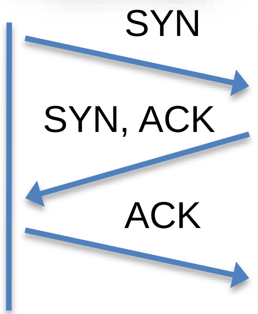
Default Gateway – The subnet exit point where packets need to be sent on their way to a different subnet

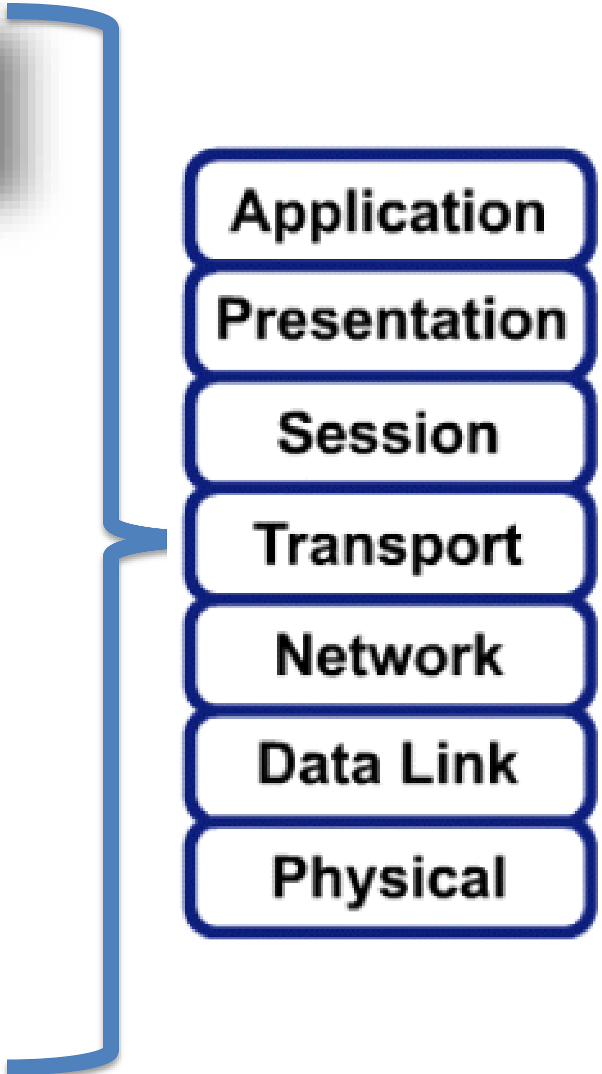


Major transport protocols: TCP and UDP

- **U**nreliable **D**atagram **P**rotocol
  - Connectionless
  - Simple/lightweight
  - Unreliable
  - RFC 768
    - http://tools.ietf.org/html/rfc768

- **T**ransport **C**ontrol **P**rotocol
  - Connection oriented
  - Heavyweight
  - Lossless
  - Congestion and flow control
  - RFC 793
    - http://tools.ietf.org/html/rfc793





TCP and UDP headers

TCP header

Bits

<table>
<tr><th>0</th><th>8</th><th colspan="2">16</th><th colspan="2">31</th></tr>
<tr><td colspan="3">Source Port</td><td colspan="3">Destination Port</td></tr>
<tr><td colspan="6">Sequence Number</td></tr>
<tr><td colspan="6">Acknowledgment Number</td></tr>
<tr><td>Data Offset</td><td>Reserved</td><td>Code</td><td colspan="3">Window</td></tr>
<tr><td colspan="3">Checksum</td><td colspan="3">Urgent Pointer</td></tr>
<tr><td colspan="5">Options</td><td>Padding</td></tr>
</table>

20 bytes without options

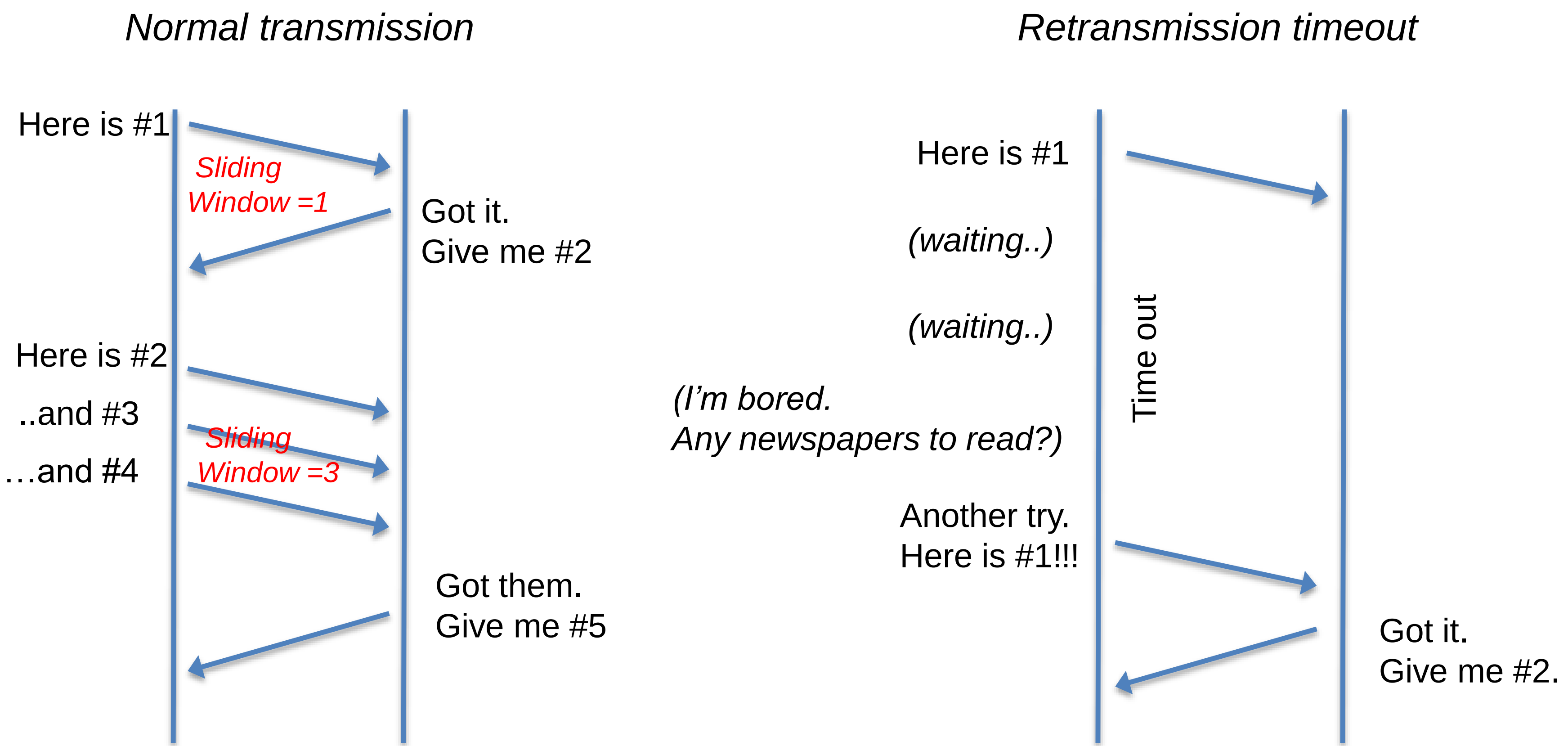
UDP header

Bits

<table>
<tr><th>0</th><th>16   31</th></tr>
<tr><td>SOURCE PORT NUMBER</td><td>DESTINATION PORT NUMBER</td></tr>
<tr><td>LENGTH</td><td>CHECKSUM</td></tr>
</table>

8bytes

TCP: How does it work?

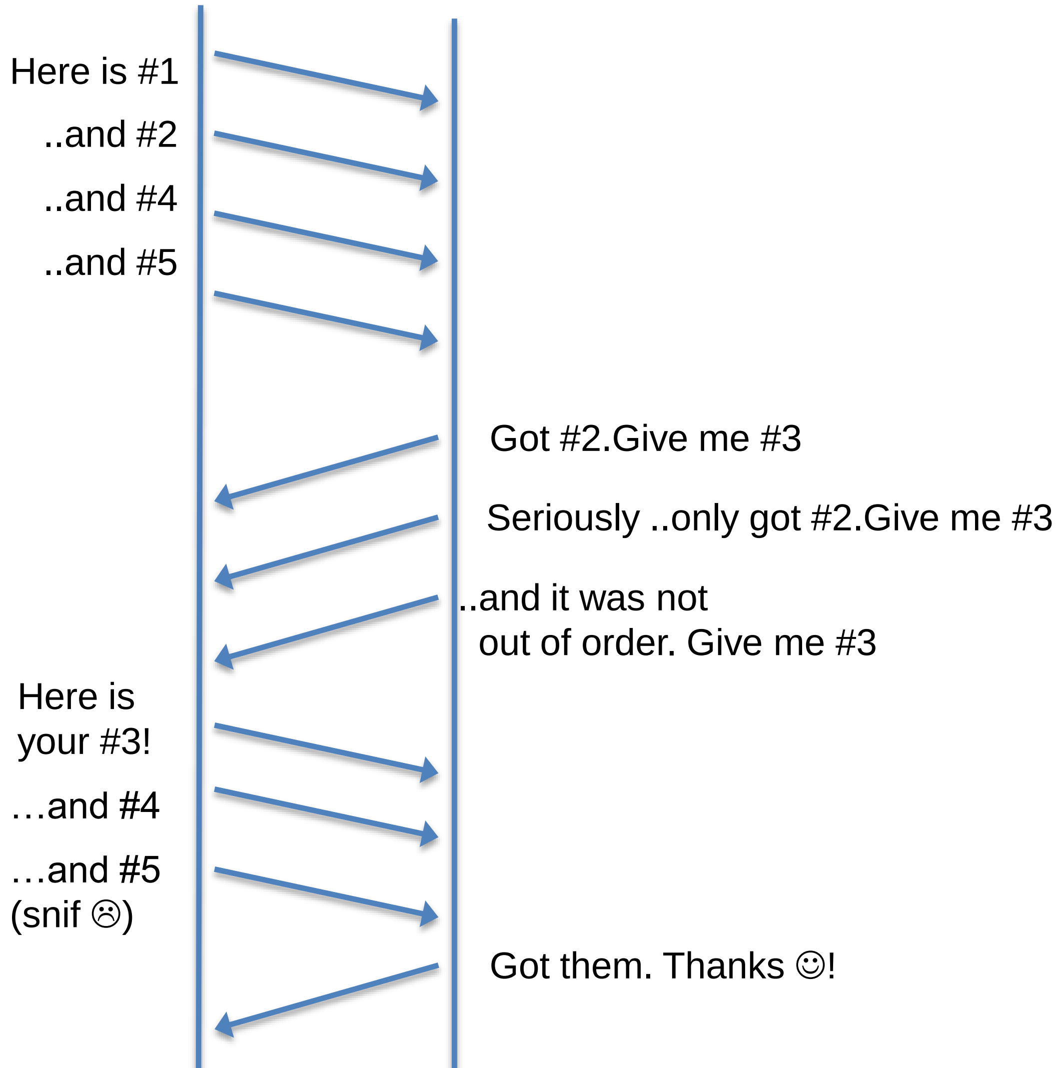


The *sliding window* is variable and depends on:

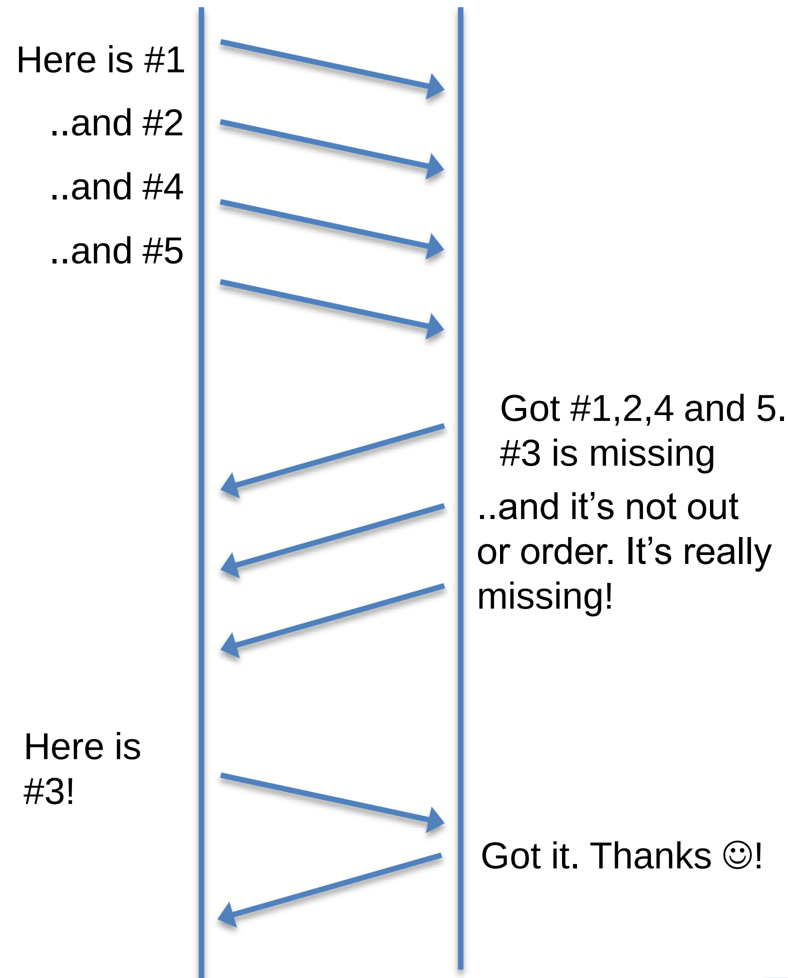
- the congestion control algorithm
- flow control parameters
- traffic congestion conditions
  (see next slides)

TCP: How does it work?

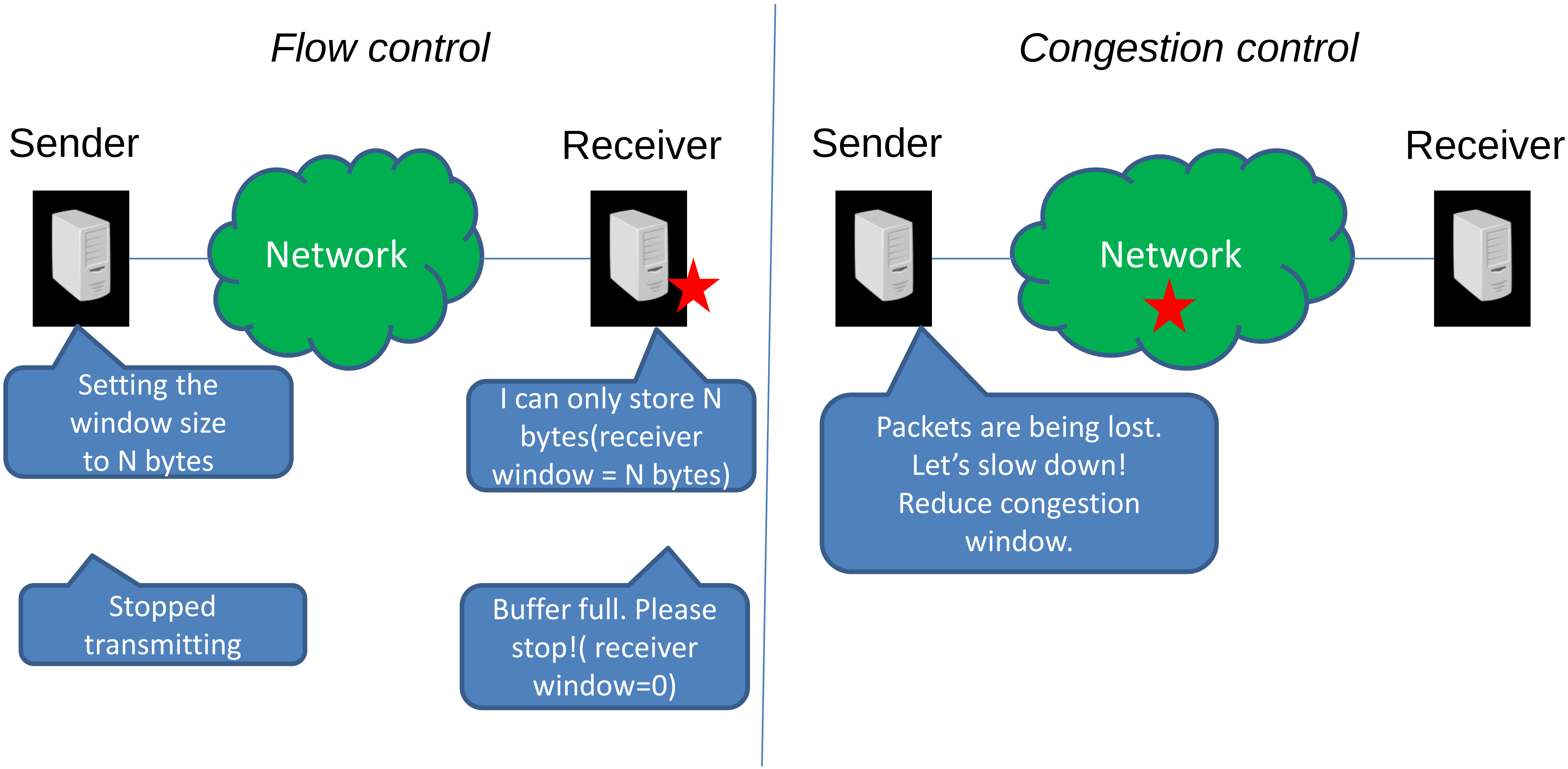
Cumulative acknowledgement



Selective acknowledgement

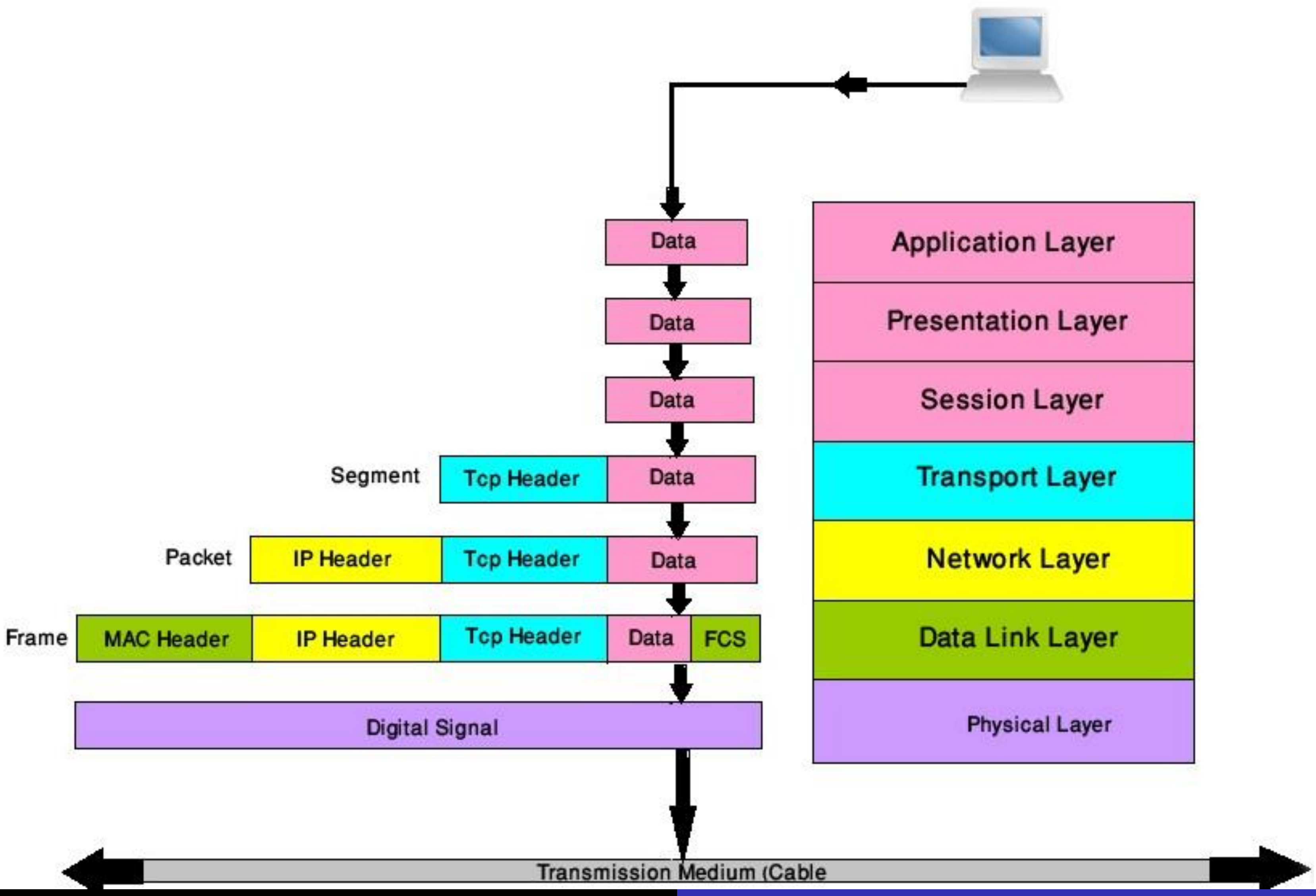


TCP: Flow vs congestion control



(Sliding) window size = min(Receiver Window Size, Congestion Window Size)

Data Encapsulation & Decapsulation

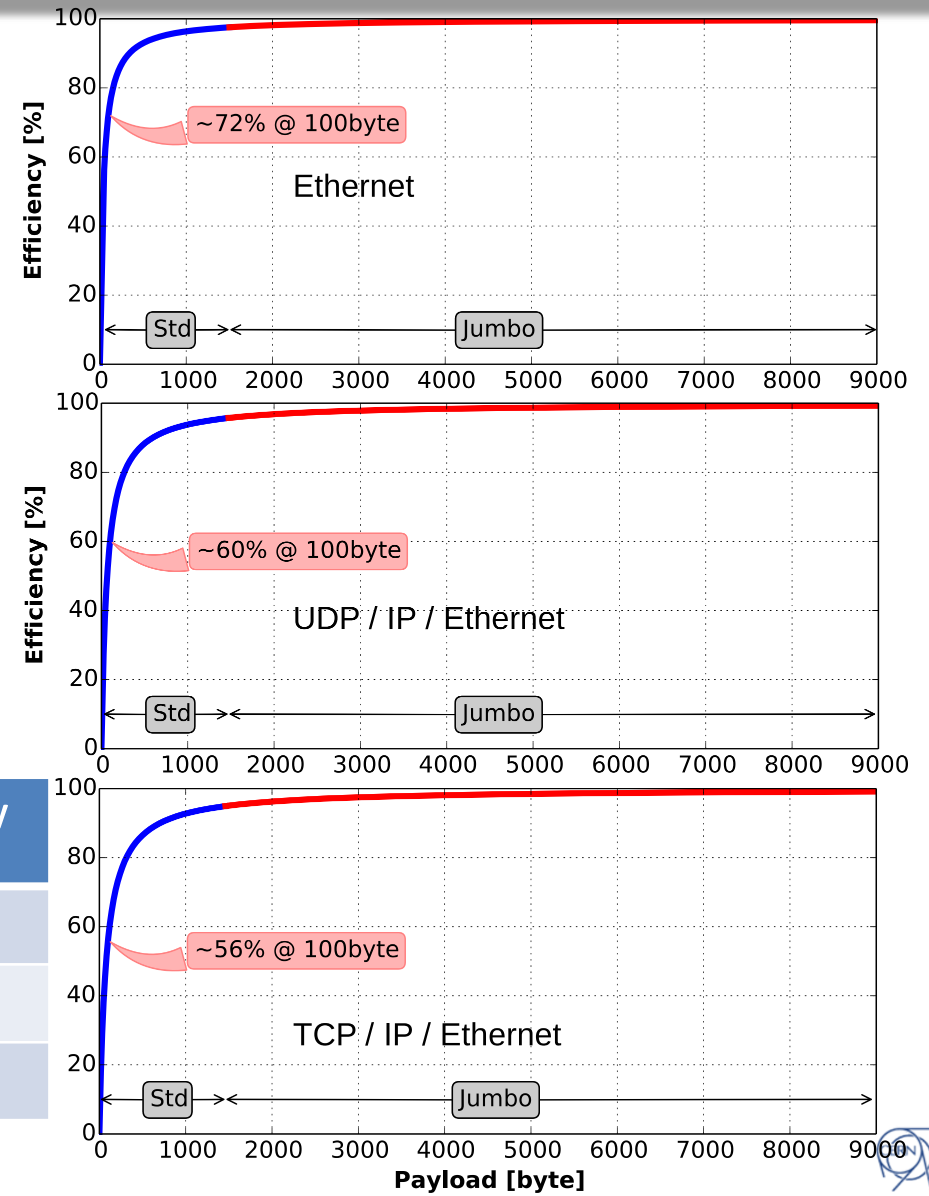


Encapsulation – Efficiency

$$Efficiency = \frac{Payload}{Payload + Overhead}$$

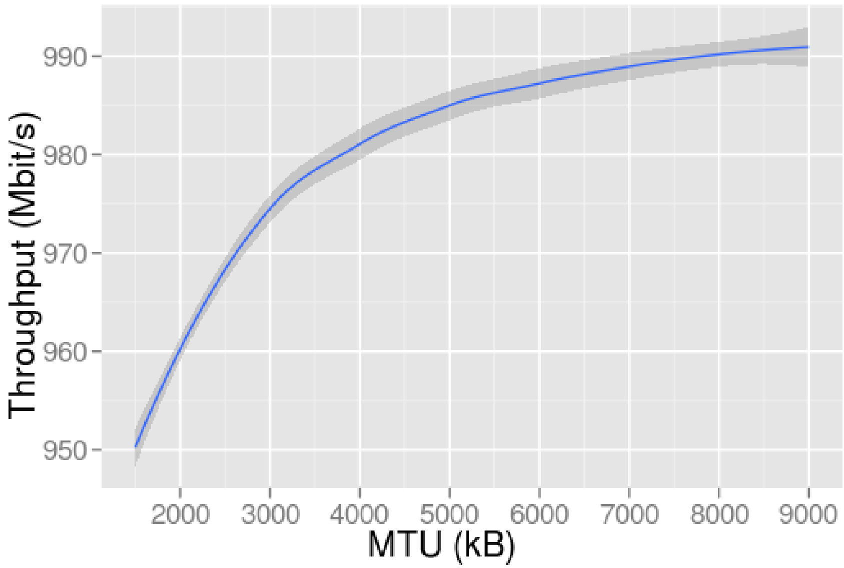
*Goodput = Efficiency * Throughput*

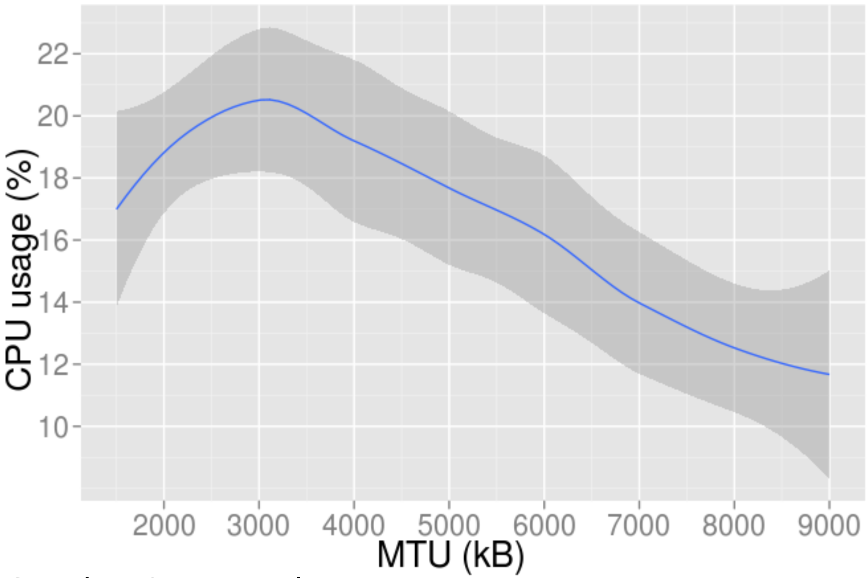
Encapsulation	Over head	Efficiency (1420 b)	Efficiency (100 b)	Efficiency (1 byte)
Ethernet	40b	97.2%	72%	1.2%
UDP/IP/Eth	68b	95.4%	60%	1.2%
TCP/IP/Eth	80b	94.6%	>56%	>1.2%



Jumbo Frames

- **Improve goodput**
  - 94% @ 1500 MTU
  - 99% @ 9000 MTU
- **Reduce the frame rate**
  - Lower interrupt rate
  - Less data dis/re – assembling for the CPU





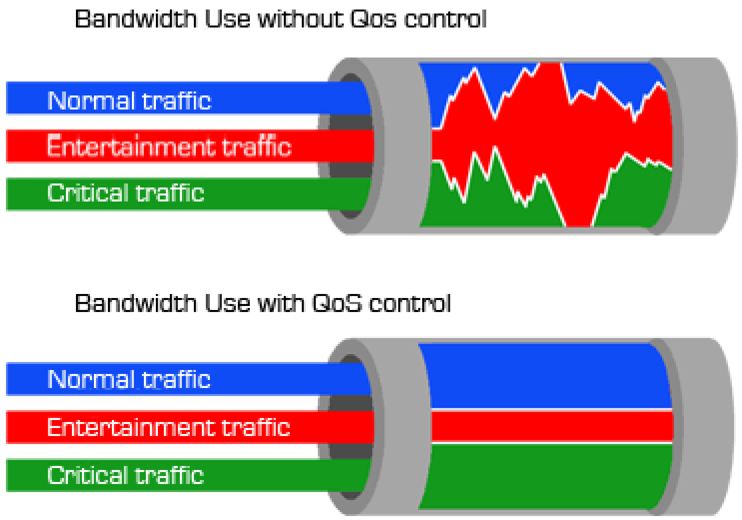
Tests performed on a Broadcom NIC and an 8 core Intel Xeon processor

Quality of Service (QoS)

- Enables traffic prioritization
  - Particularly useful for critical control messages
- Layer 4 – based on SRC/DEST port
- Layer 3
  - SRC/DEST IP Address

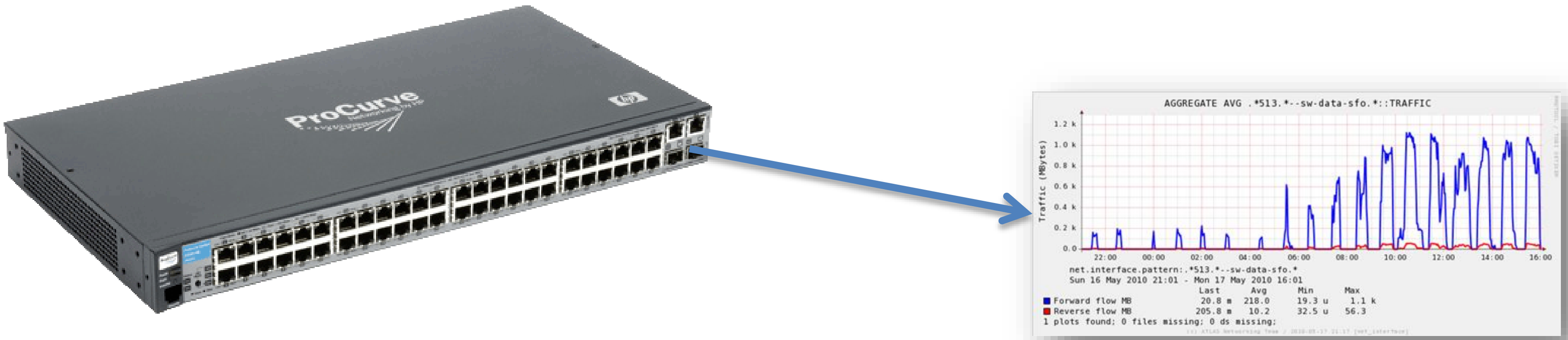
  OR
  - DiffServices (DSCP)
    - Sets the priority in a dedicated IP header field
    - Can be set at the application level
- Layer 2
  - SRC/DEST MAC Address

  OR
  - VLANs
    - Define overlapping(tagged) VLANs
    - Send traffic on a specific VLAN
    - Configure network devices to prioritize VLANs

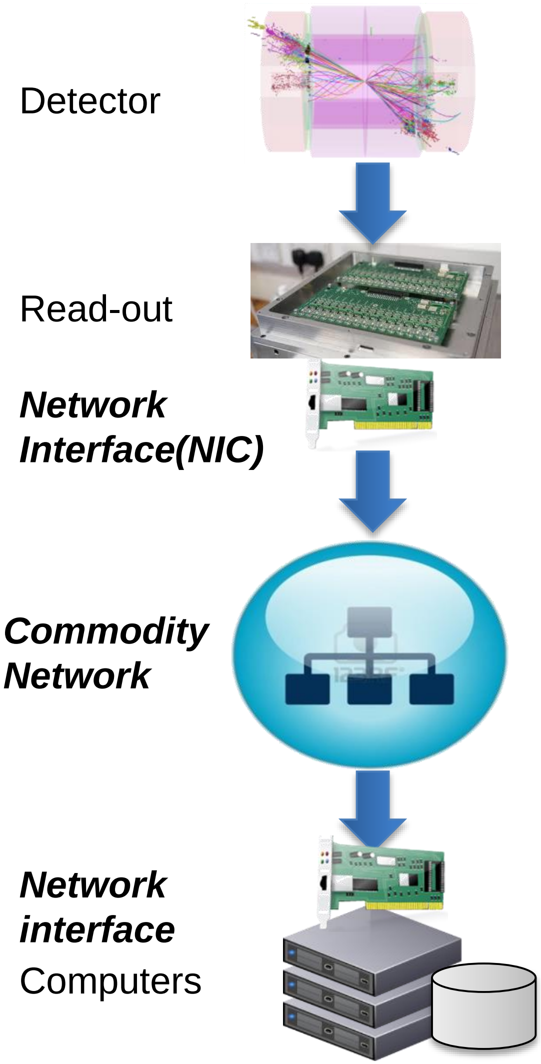


Network Monitoring. SNMP

- A standard protocol for managing devices on IP networks (switches, routers, computers etc);
- Exposes management data in the form of variables on the managed systems. These variables are then queried;
- Used to gather device-based or port-based statistics (traffic volume, errors, packets, discards, temperature etc);

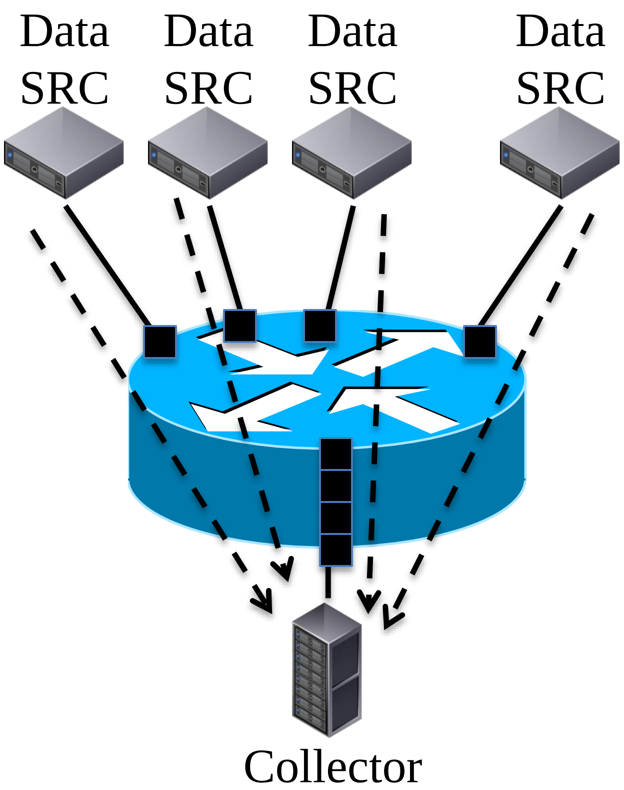


Data Acquisition uses networks



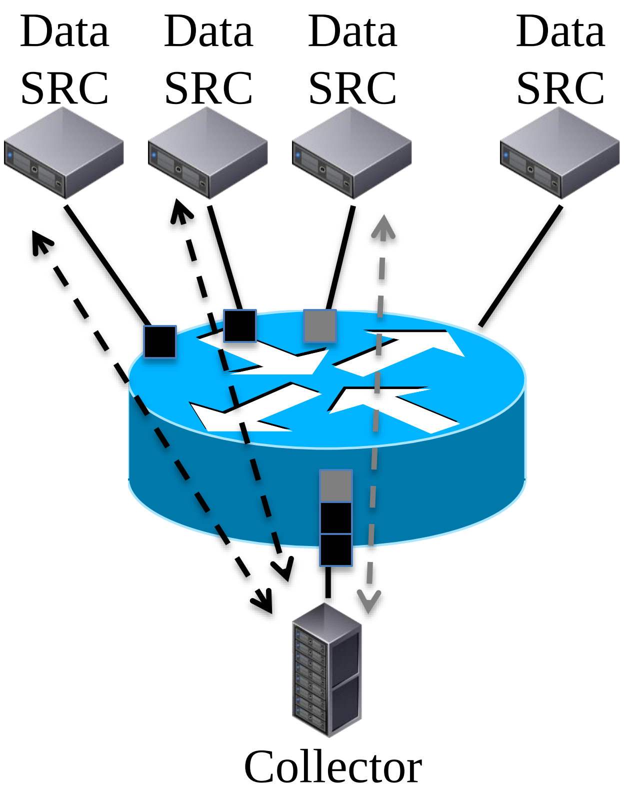
- Detector
  - Measure physical phenomena
- Read-Out
  - Digitize and perform basic processing
  - Possibly data buffers
  - **Interface to network**
- **Commodity Network**
  - Connect all read-outs to analysis computers
  - Allows computers to collect data from all sources
- Computer(s)
  - **Interface to network**
  - Collect data from all sources
  - Analyze and filter data
  - Store data

DAQ – push design



- Data SRCs simultaneously send data to a collector
- Fan-in effect on the switch
  - Packets need to be buffered before being sent to the Collector
  - The more sources, the worse
- Advantages:
  - Simple design of the data sources
- Disadvantages
  - Rely on network buffers for not losing data
  - Collector must cope with the rate

DAQ – pull design



- Data SRC buffer data and provide it on request
- Controlled fan-in effect on the switch
  - Collector can limit the number of outstanding requests
  - Not affected by the number of sources
- Advantages:
  - Better control of network traffic
  - Collector asks as much as it can handle
  - Collector can slow down in case of loss detection
- Disadvantages
  - Data sources complexity:
    - Buffering
    - Request-reply protocol implementation

LHC DAQ networks requirements

- High availability/Fault tolerance
  - Ideally, redundancy at every level
  - Advanced health monitoring
- Performance
  - *High throughput AND low latency*
  - *Substantial tuning*
    - *Data flow software*
    - *Network itself*
  - *Advanced performance monitoring*
- Security
- Low cost

LHC DAQ networks characteristics

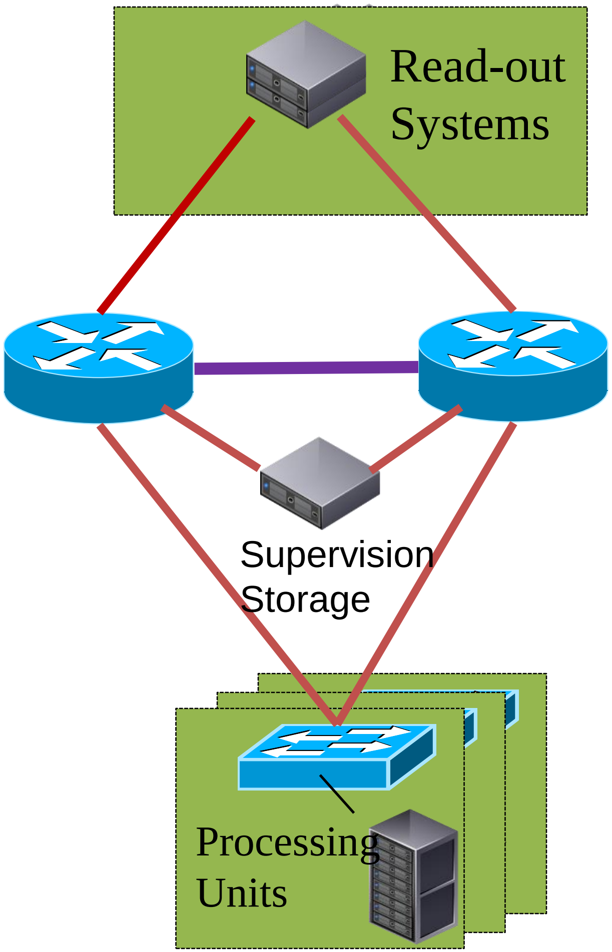
- Private local networks
- Congestion hot-spots
- Packet loss caused by
  - HW failures
  - Transmission errors
  - Buffer overflows
- Network latency and event building time

  much smaller than the TCP timeout

The golden rule:

Minimize packet loss and TCP retransmissions!

DAQ Network for a large experiment



ATLAS DAQ Network

- Pull architecture
- LHC DAQ systems use O(1000) nodes
  - too large for a single network device
- Typical multi-layer architecture
  - Aggregation layer
  - Core layer
- Simple, reliable and fast
  - Routing
  - Link aggregation

DAQ networks optimization : examples

- Hosts
  - Mainly at reception
    - more resource consuming side mainly because it has to reorder packets
  - Provide large kernel buffers and large socket buffer for the application
    - possible to increase the TCP window size
  - Interrupt(IRQ) tuning; especially important for link speed > 1Gbit/s
    - Use interrupt coalescing(an interrupt for more than one frame)
    - Use affinity to spread IRQs on all CPU cores
- Network devices
  - Enable jumbo frames on all ports to improve goodput and performance
  - Maximize network device buffers
    - Packet loss has a big impact on performance
  - Use QoS for critical control messages

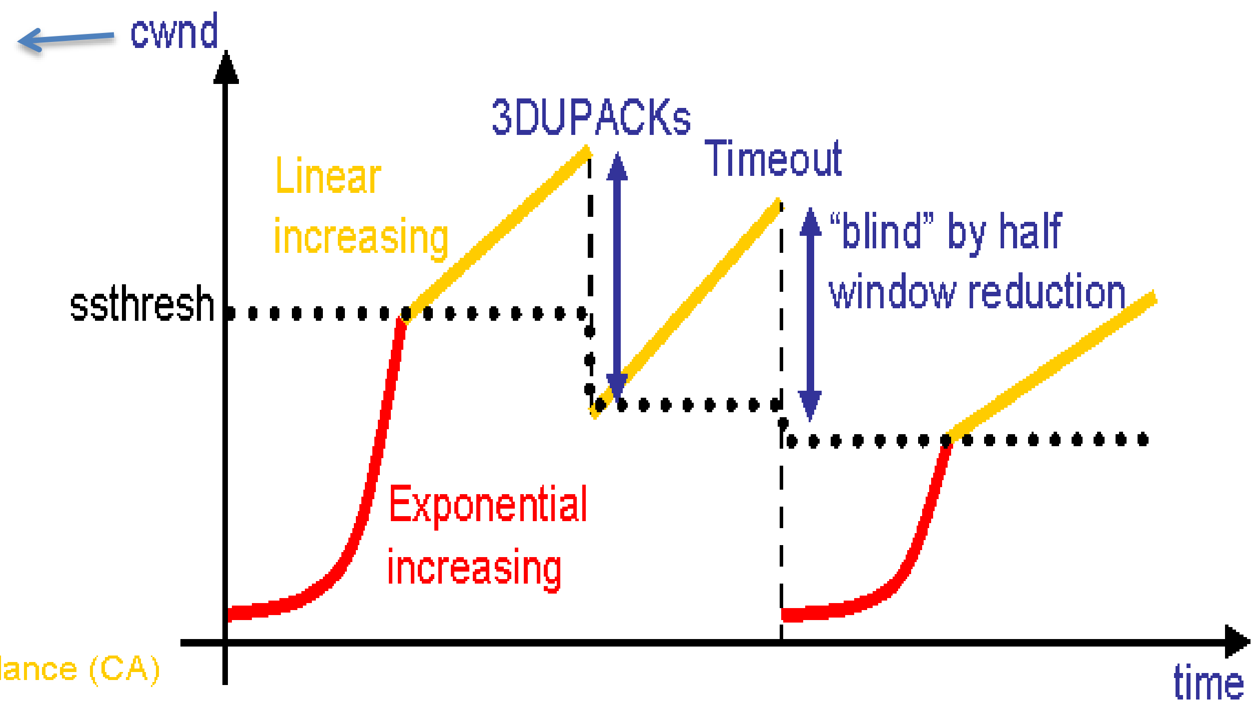
Conclusions

- Networking basics
  - OSI Model is your friend!
  - Ethernet is used to connect hosts together
  - IP is used to connect networks together
  - TCP is used to reliably transport data
  - Efficient data encapsulation can play an important role
  - QoS is used to prioritize traffic
  - SNMP is used to monitor the network

- Networks for DAQ
  - Performance is the key requirement
  - Very often, substantial tuning is needed to obtain an efficient large-scale DAQ system
  - Only scratched the surface by presenting Ethernet-based networks
    - There are other very interesting HPC technologies (Infiniband, Omnipath, etc)

TCP: Congestion control example

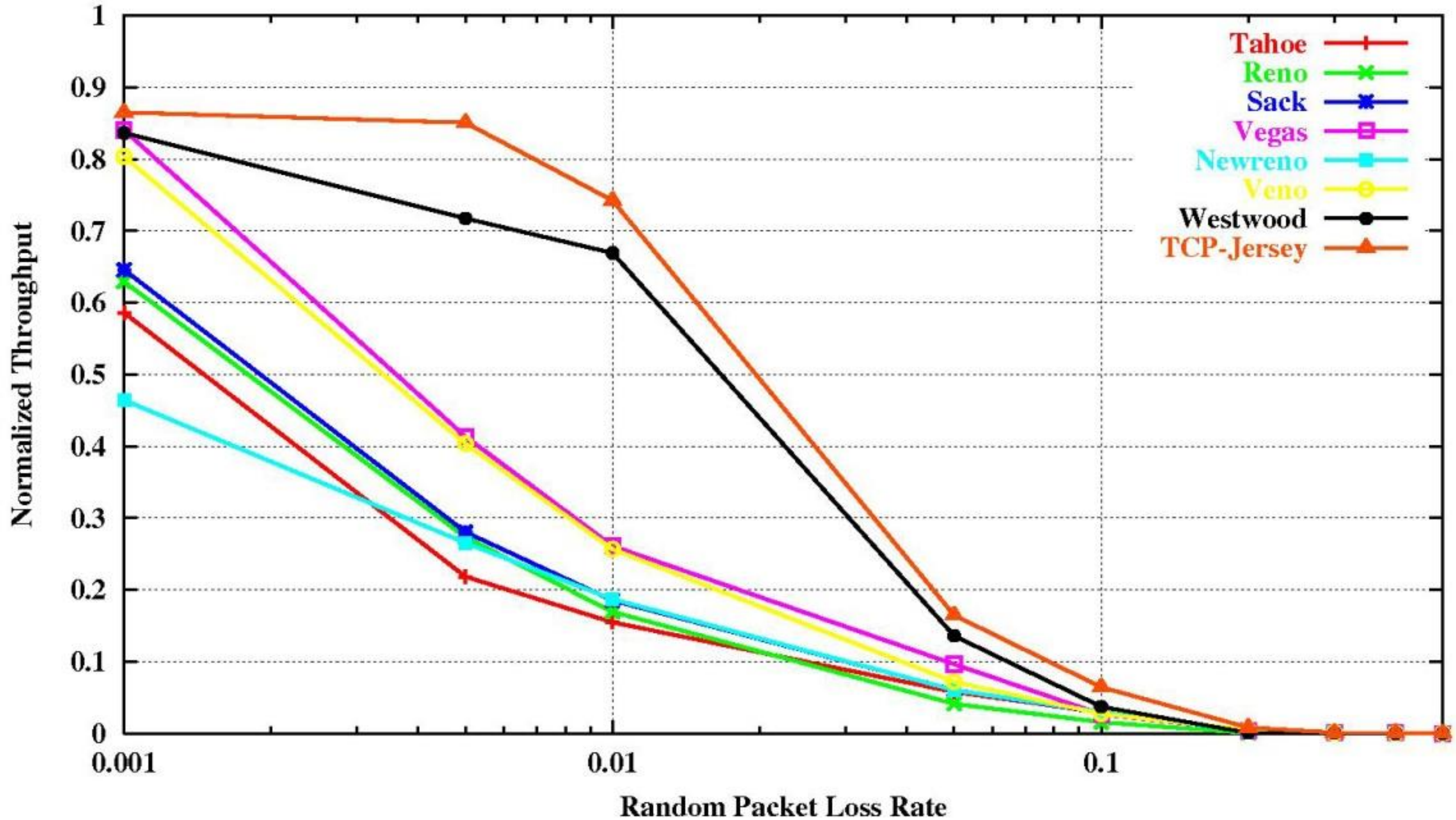
Congestion window



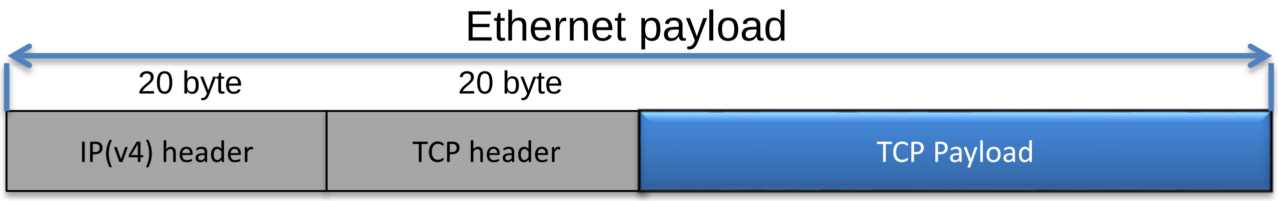
Congestion Avoidance (CA)
Slow-start (SS)

TCP Westwood

TCP Performance with Packet loss



TCP/IP (over Ethernet)



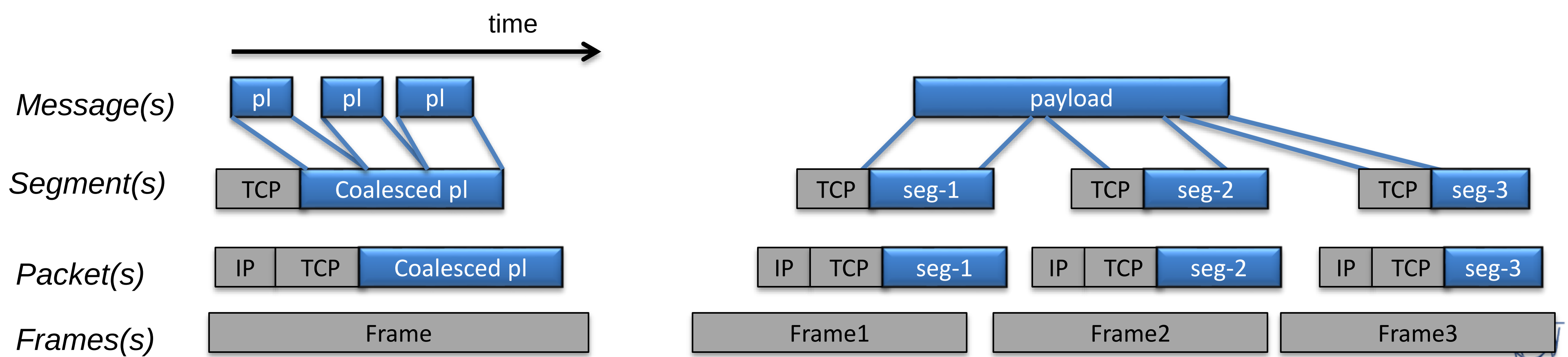
Data is buffered for a short before sending it

The sender knows the maximum transmission unit(MTU) size (typically 1500 bytes)

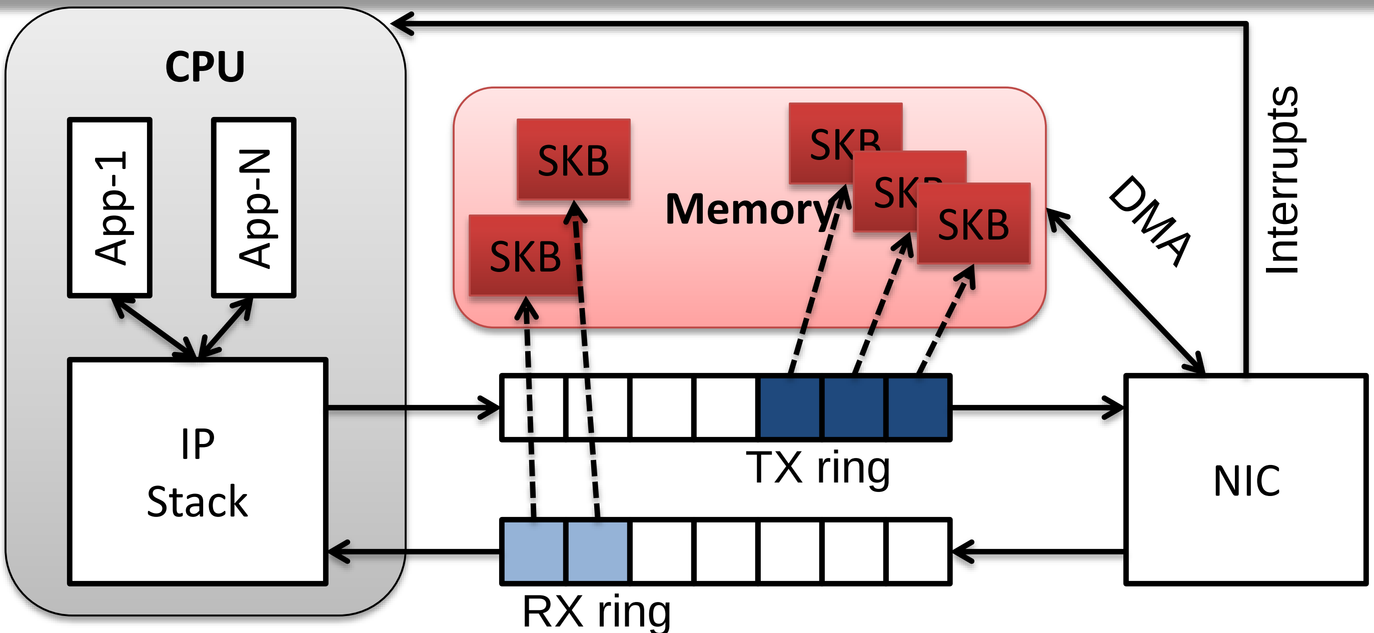
Coalesces or segments data depending on payload size

<u>Payload < MTU</u>
May coalesce

<u>Payload > MTU</u>
Does segmentation



Kernel – NIC interaction



Send

- Data in memory (SKB)
- Descriptor to TX ring
- NIC fetches data via DMA
- NIC **interrupts** when finished sending

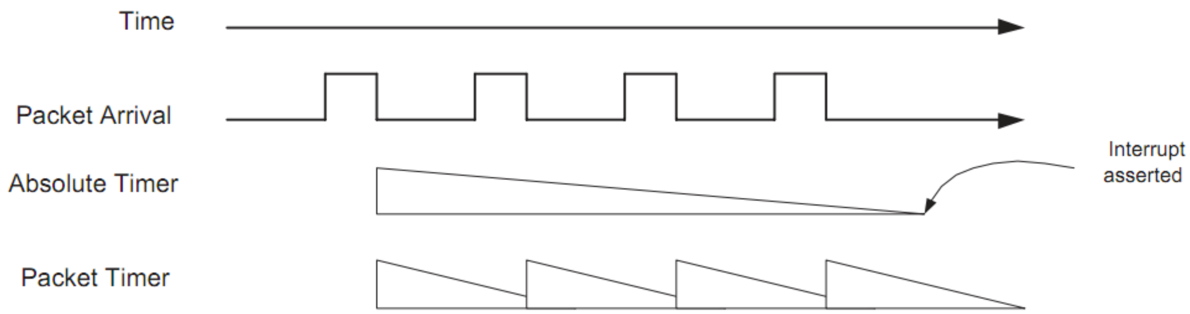
Receive

- NIC puts data in memory (SKB) via DMA
- NIC puts descriptor in RX ring
- NIC **interrupts**
- CPU fetches the SKB and frees up the RX ring descriptor

Interrupt coalescing

- Hardware interrupt has a cost
  - Context switch of a CPU
    - Saving and loading registers and memory maps, updating various tables and list
  - Happens every time an Ethernet frame is received
    - 1538 bytes -> 12304 bits -> 1 frame every 1.23 µs @ 10 GbE

- Lower the rate with *interrupt coalescing*
  - 1 interrupt for several frames



Precautions

- Do not add too much latency in case of low traffic
- Careful with the ring buffer size
  - Packets are discarded if the buffer is full