Data Networks

Introduction to Networking

Dan Octavian Savu
&
Silvia Fressard-Batraneanu

CERN

ISOTDAQ 2015, Rio de Janeiro

Outline

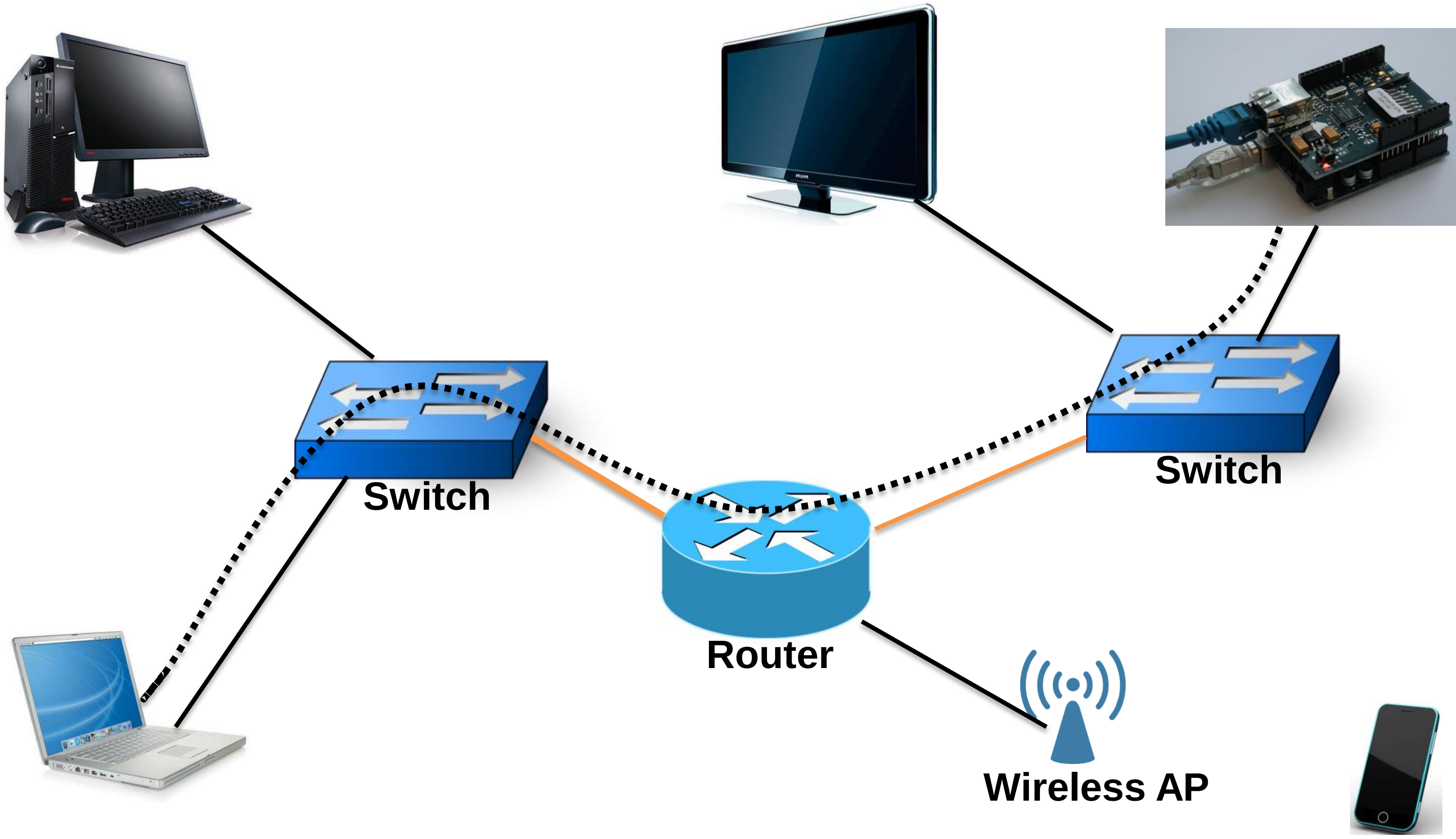
- Introduction
  - Networking basics
  - OSI reference model
- Technologies and protocols
  - Ethernet
  - IP (Internet Protocol)
  - TCP vs. UDP
  - Routing
- Network monitoring
- Software defined networking

What is a network ?

- A ***network*** is simply two or more computers connected together so they can exchange information. At the same time it can be a complex interconnected system of objects and people (Internet)

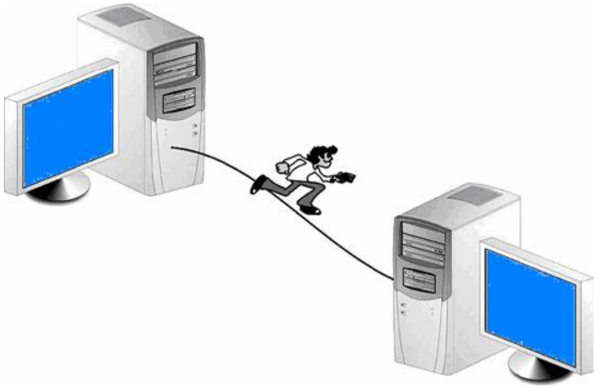
- ***End-host devices*** are hosts attached to a network
- A ***source host*** is the place where the data originally comes from
- A ***destination host*** is the place where the data is being sent to
- ***Networking devices*** are waypoints along paths for data to travel along
- **Links** are direct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

What is a network ?



Why do we need a network ?

- Sneaker Net
  - Inefficient data communication;
  - Many copies of the same file;
  - Reliability, scalability, flexibility… issues.



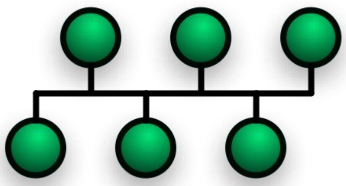
- (High speed) networks connecting all hosts help address slow transmission of information
- Interconnected datacenter servers help minimize redundant copy of files
- File sharing, resource sharing, communication & collaboration, group organization, remote access, data backups etc

Network types

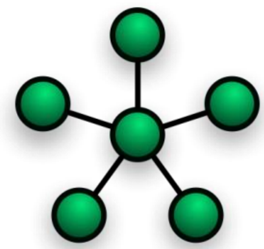
- Networks have different varieties to suit different purposes and needs
- **LAN** (small size, high speed, physical proximity)
- **WAN** (long distance, lower data transfer rates)
- **MAN** (metropolitan area network)
- **PAN** (immediate space around a person)
- **SAN** (connecting storage farms, high speed)
- **VPN** (private network extension across a shared or a public network)

Network structure

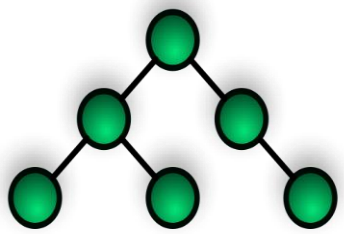
- The structure of a network is known as the ***topology***
  - **Physical** = The way the network is cabled
  - **Logical** = The way devices use the network to communicate



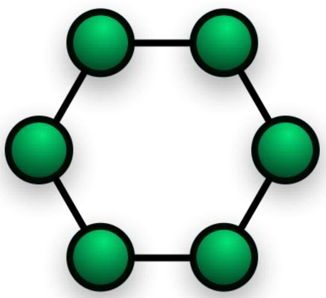
Bus Topology



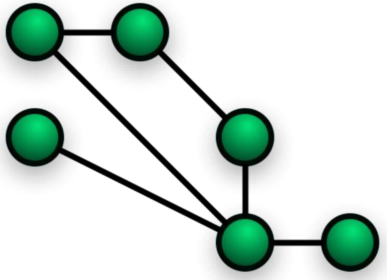
Star Topology



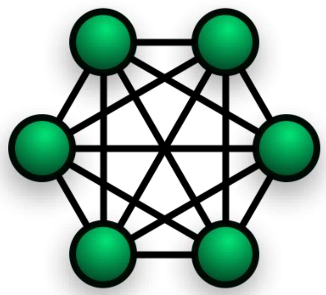
Hierarchical Topology



Ring Topology



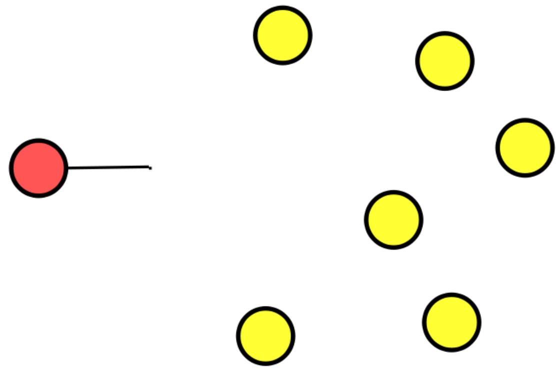
Partial Mesh Topology

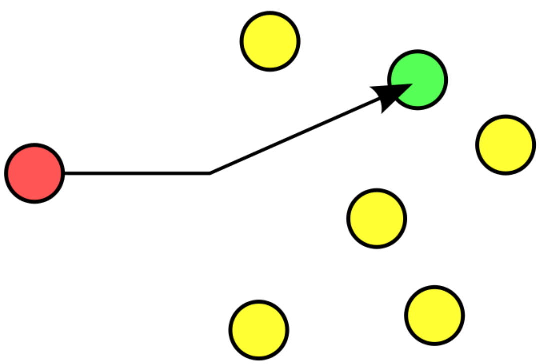


Fully Mesh Topology

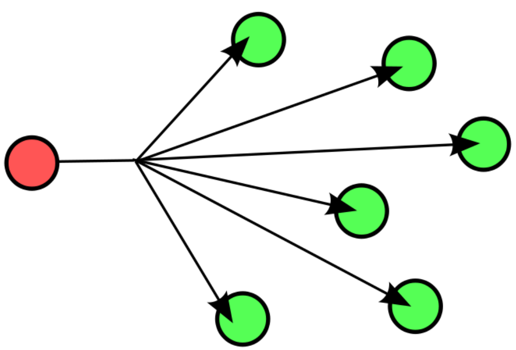
Network communication

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

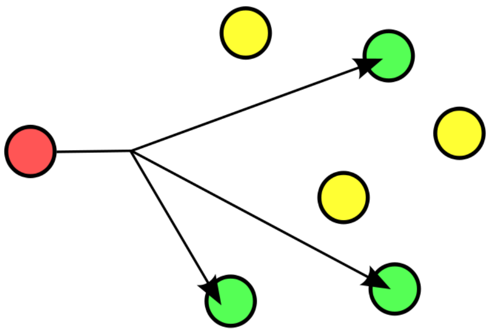




Unicast

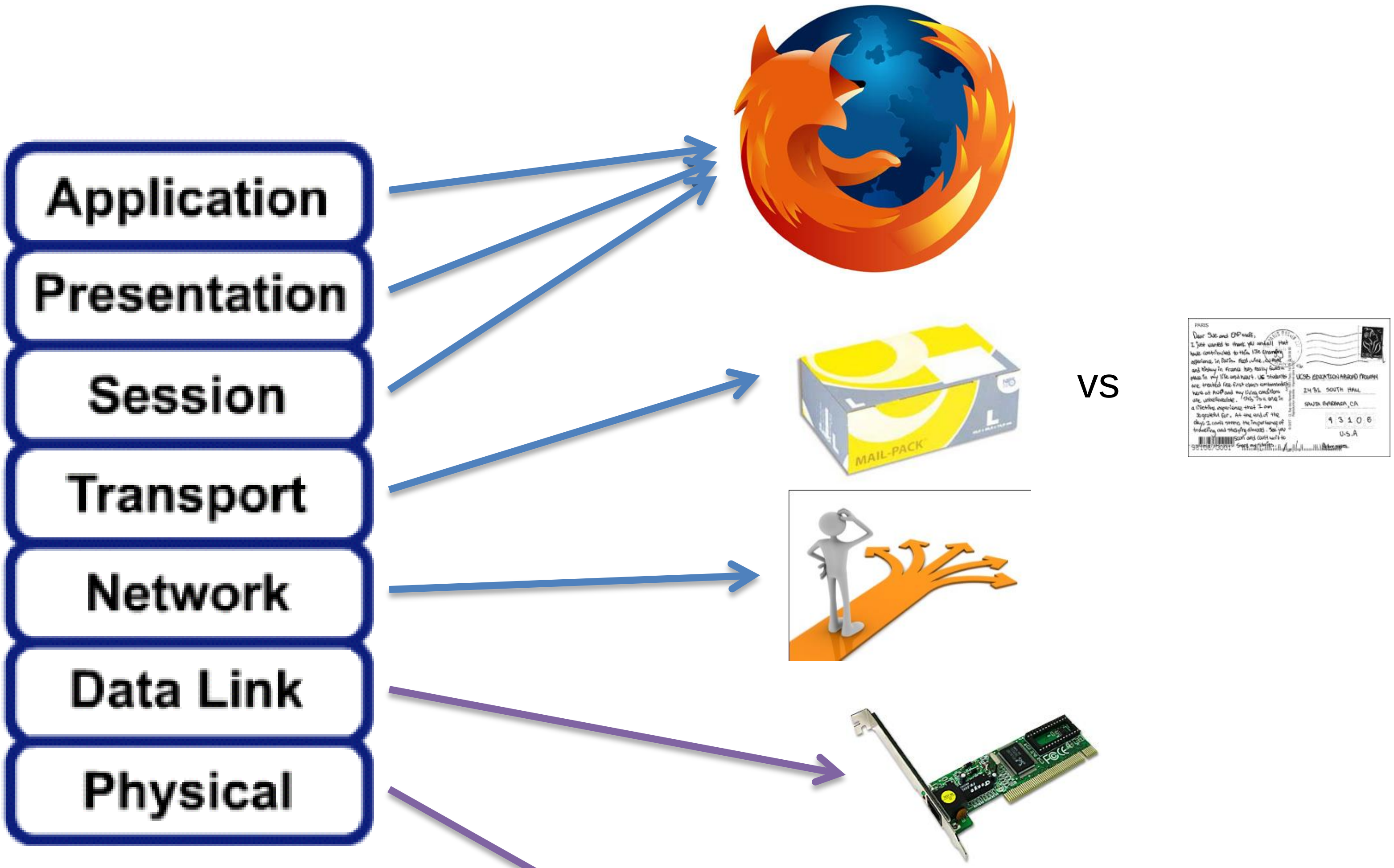


Broadcast



Multicast

OSI Model. Divide et impera.



1011101010000111100101010100011

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

OSI Model. Divide et impera.

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

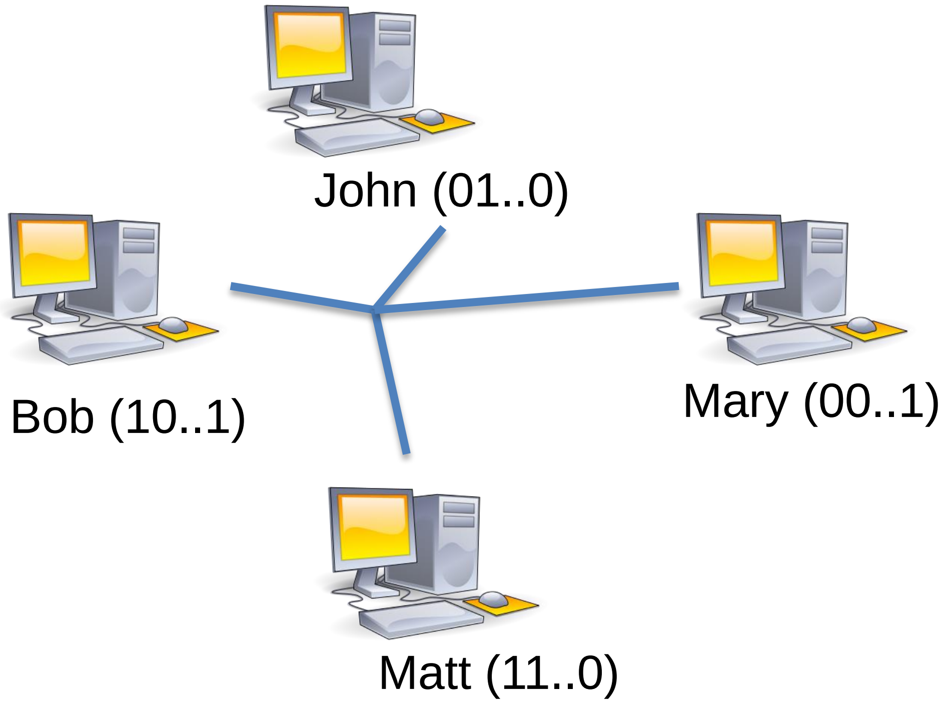
Why layers in OSI ?

- Simplifies understanding of networking
- Breaks networking tasks into smaller, manageable, chunks
- Allows for platform independence
- Provides a standard for networking manufactures
- Easier to determine the correct networking protocol required to connect
- Problem investigation is easier and debugging time is shortened

Outline

- Introduction
  - Networking basics
  - OSI reference model
- Technologies and protocols
  - Ethernet
  - IP (Internet Protocol)
  - TCP vs. UDP
  - Routing
- Network monitoring
- Software defined networking

Ethernet

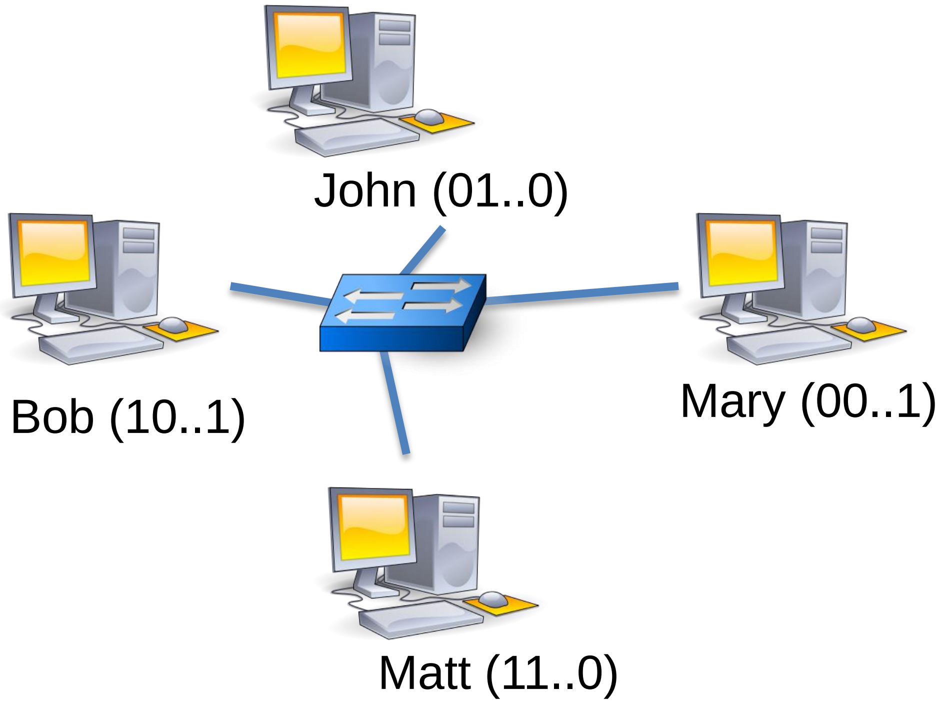


- Small to medium size group of computers
- Frame based technology
- Defines wiring and signaling standards (Layer 1)
- Defines a flat addressing scheme with local visibility, called **MAC** (Layer 2)

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

Basic Ethernet frame

Ethernet (switch)



- Analyses incoming frames and switches them to correct segment using MAC addresses (a process called switching)
- Simultaneous data transmissions without medium sharing
- Layer 2 device

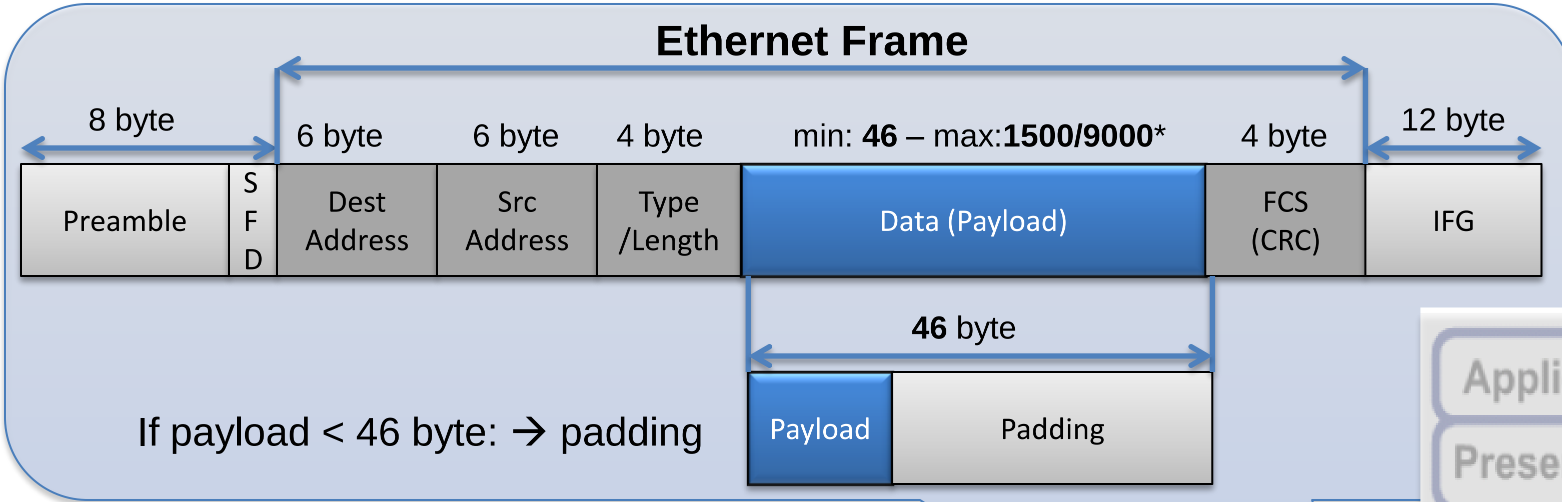
8	6	6	2	46 ~ 1500 bytes	4
Pre-amble	Dest.	Source	Type/ Length	Data	Frame check

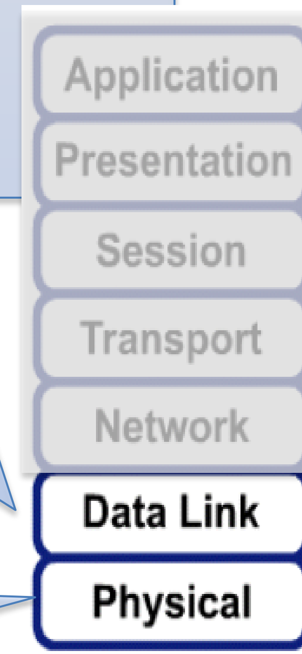
Basic Ethernet frame

Ethernet (reliable since 1973)

- Created at Xerox in 1973, released as an open standard in the early 80s
- Later modified to comply with the OSI model, ratified as IEEE 802.3 in 1985
- Ethernet has evolved significantly since then:
  - Proved flexible as a technology, able to upgrade to new media and faster data transmission speeds.
  - 10Gig Ethernet ratified as IEEE 802.3ae
  - Optical fiber has joined copper as media of choice for the IEEE 802.3 family
- Flexibility came through the simplicity of Ethernet's structure
- Ease of installation and maintenance

Ethernet





All flavors of media and speeds:

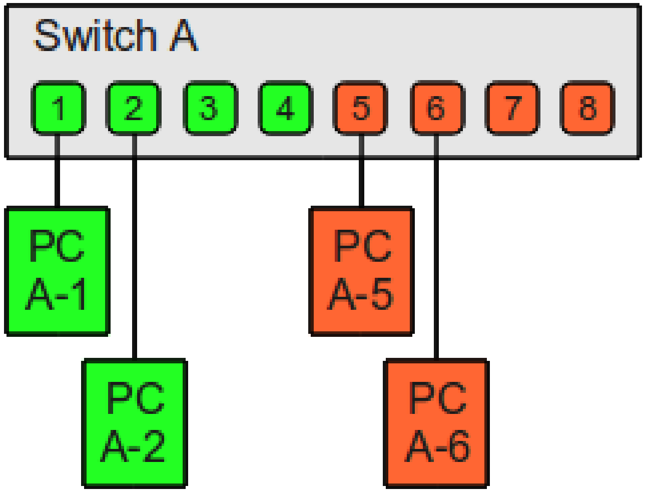
- … even slower but this is now history
- 100 Mbit/s: copper (UTP), fiber
- 1 Gbit/s: copper (UTP), fiber
- 10 Gbit/s: fiber, copper (twinax, UTP)
- 40 Gbit/s: fiber
- 100 Gbit/s: fiber

Ethernet Standards

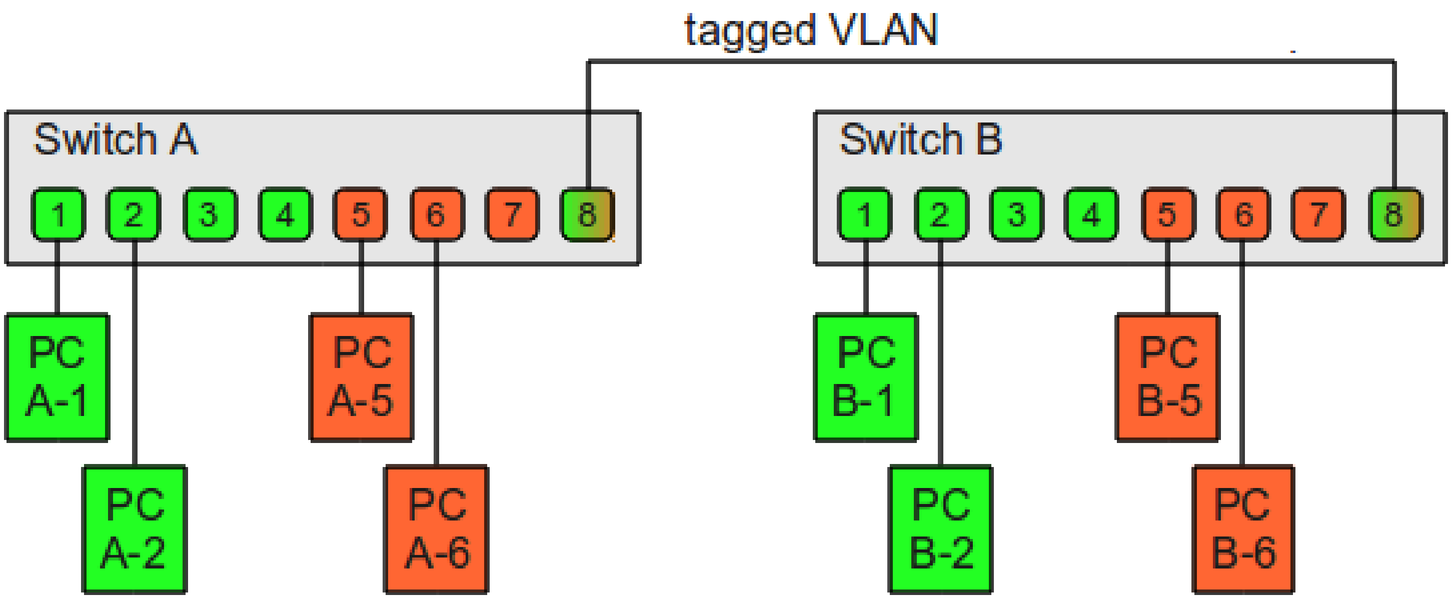
The Evolution of Ethernet Standards to Meet Higher Speeds				
Date	**IEEE Std.**	**Name**	**Data Rate**	**Type of Cabling**
1990	802.3i	10BASE-T	10 Mb/s	Category 3 cabling
1995	802.3u	100BASE-TX	100 Mb/s*	Category 5 cabling
1998	802.3z	1000BASE-SX	1 Gb/s	Multimode fiber
	802.3z	1000BASE-LX/EX		Single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

Virtual LAN (VLAN)

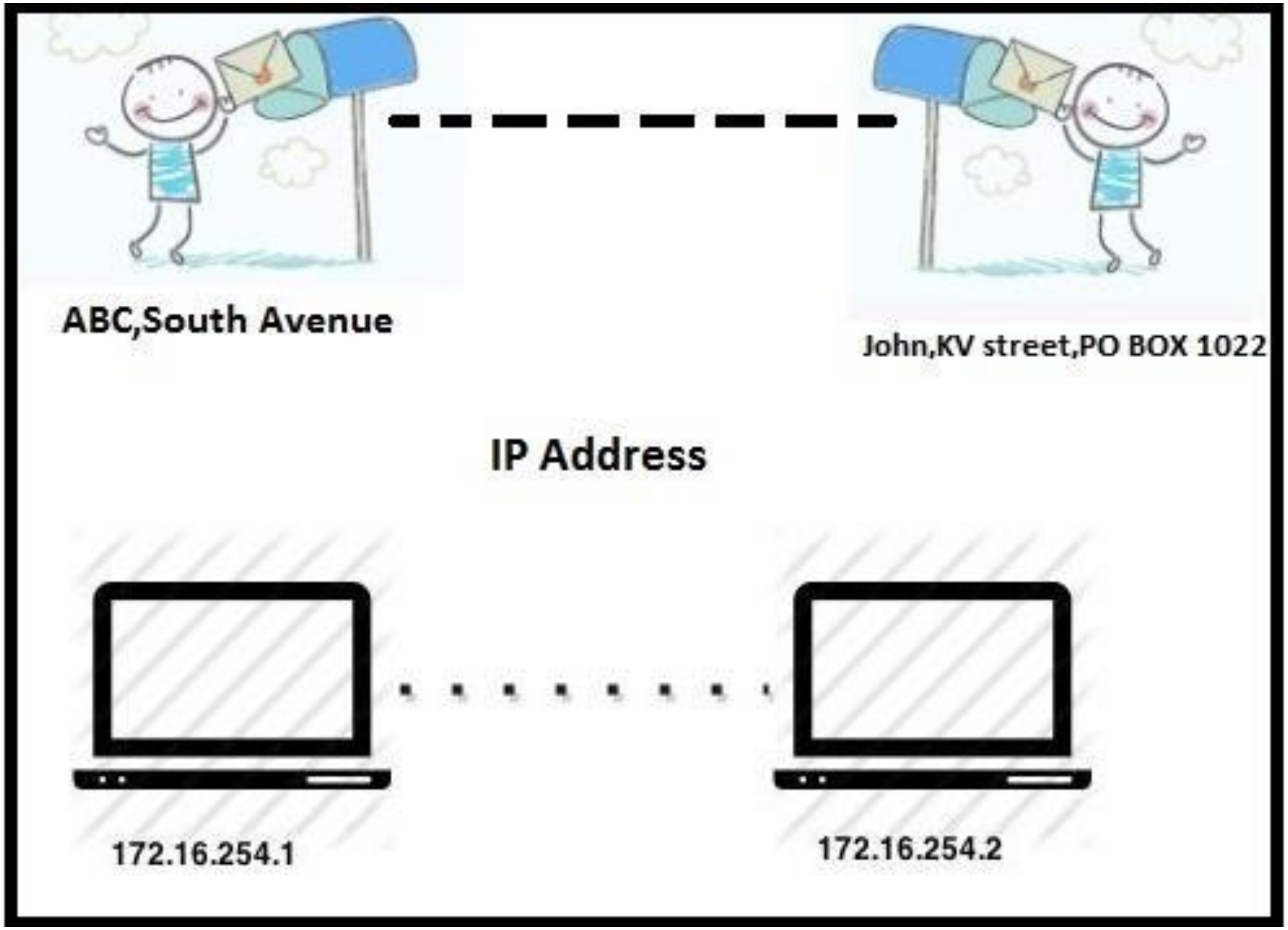
- OSI Layer 2
- Logical grouping of hosts
- Simplifies network design and administration



Virtual LAN (VLAN)



IP (Internet Protocol)



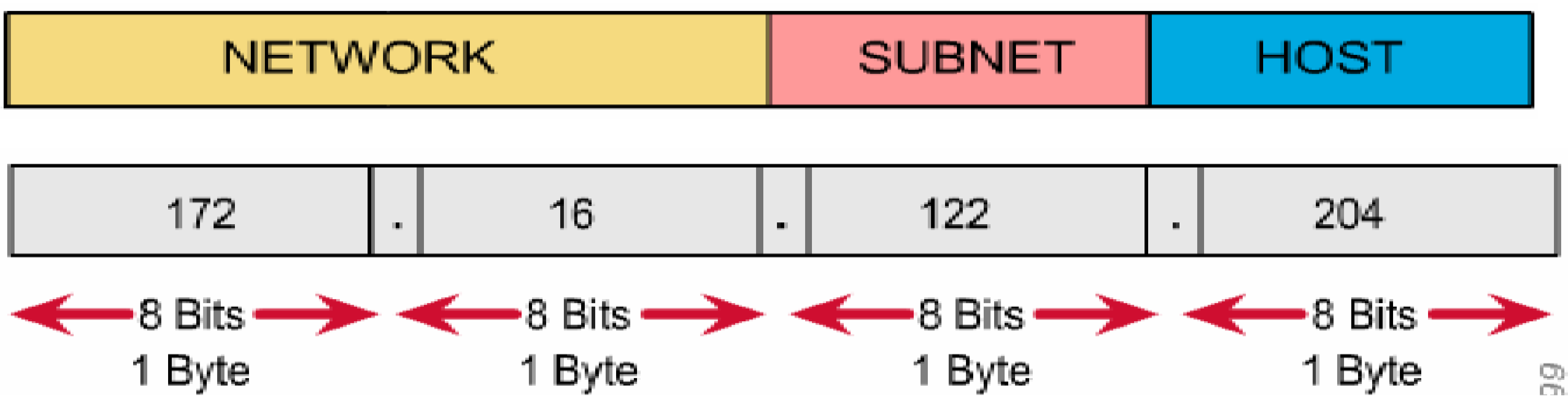
IP. (Un)reliable since 1974.

- Connectionless, best effort protocol
- Designed to be encapsulated into layer 2 protocols , such as Ethernet
- Initially created by Vint Cerf and Bob Kahn in 1974
- IPv4 described in RFC 791 (1981) hyperlink

- Defines a hierarchical (logical) addressing scheme capable of connecting all the hosts in the world (Layer 3)
- Routes packets towards destination using best available path, with the help of routing protocols (Layer 3)

IP Addressing

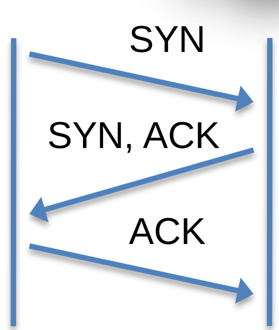
- 32bit address space (IPv4)
- Hierarchical addressing (similar to postal addressing)
- Global visibility
- ARP (Address Resolution Protocol) used to map an IP address with an Ethernet MAC address (layer 2, local visibility)

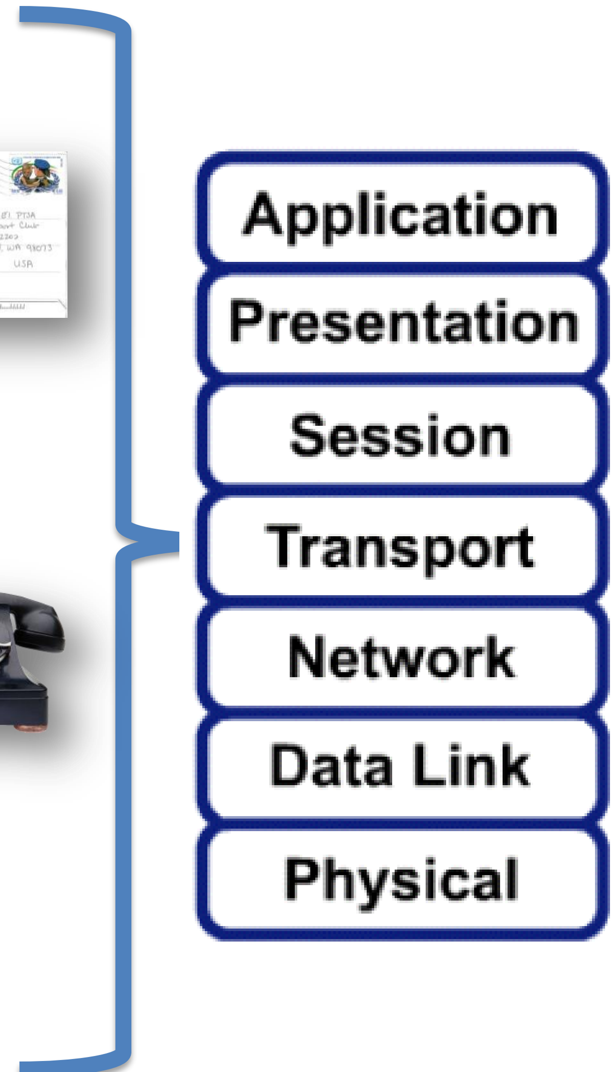


Major Transport Protocols: TCP and UDP

- Unreliable Datagram Protocol (UDP)
  - Unreliable but simple
  - Connectionless
  - RFC 768
    - http://tools.ietf.org/html/rfc768

- Transport Control Protocol (TCP)
  - Connection oriented protocol
  - Flow control
  - Lossless
  - RFC 793
    - http://tools.ietf.org/html/rfc793





TCP Packet Header

Bits

0 8 16 31

<table>
<tr><td colspan="3">Source Port</td><td colspan="2">Destination Port</td></tr>
<tr><td colspan="5">Sequence Number</td></tr>
<tr><td colspan="5">Acknowledgment Number</td></tr>
<tr><td>Data Offset</td><td>Reserved</td><td>Code</td><td colspan="2">Window</td></tr>
<tr><td colspan="3">Checksum</td><td colspan="2">Urgent Pointer</td></tr>
<tr><td colspan="4">Options</td><td>Padding</td></tr>
<tr><td colspan="5">Data</td></tr>
</table>

Size of TCP header without options: **20 bytes**

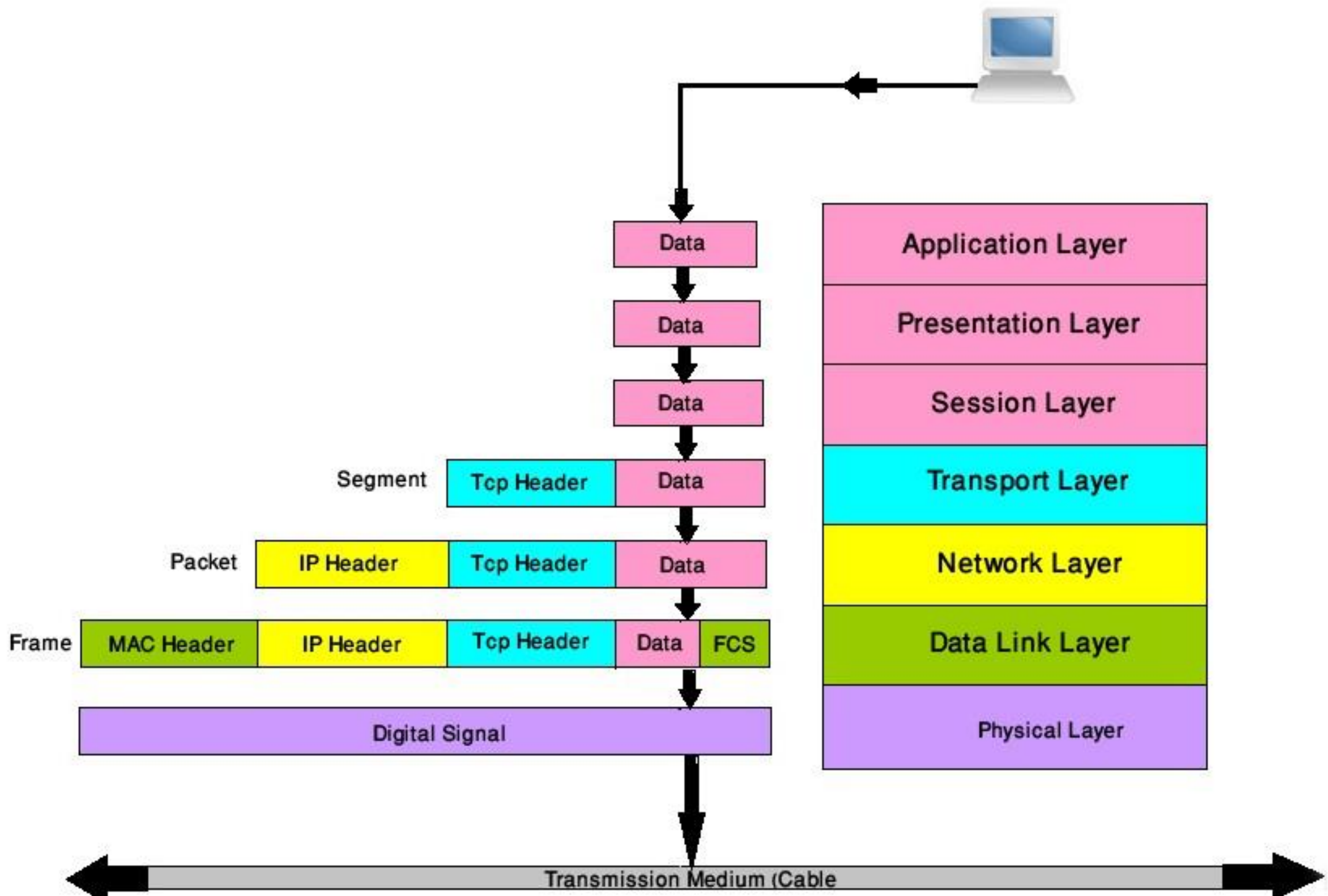
UDP Packet Header

Bits

0	16	31
SOURCE PORT NUMBER	DESTINATION PORT NUMBER	
LENGTH	CHECKSUM	

Size of UDP header: **8 bytes**

Data Encapsulation & Decapsulation



Link Aggregation & QoS

- Link Aggregation (LAG)
  - Combining several links in parallel
  - Increased throughput
  - Redundancy
- Quality of Service (QoS)
  - Ability to provide different priority to different applications, users or data flows
  - Guarantee a certain level of performance to a data flow
  - E.g.: required bit rate, delay, packet drop probability

Routers

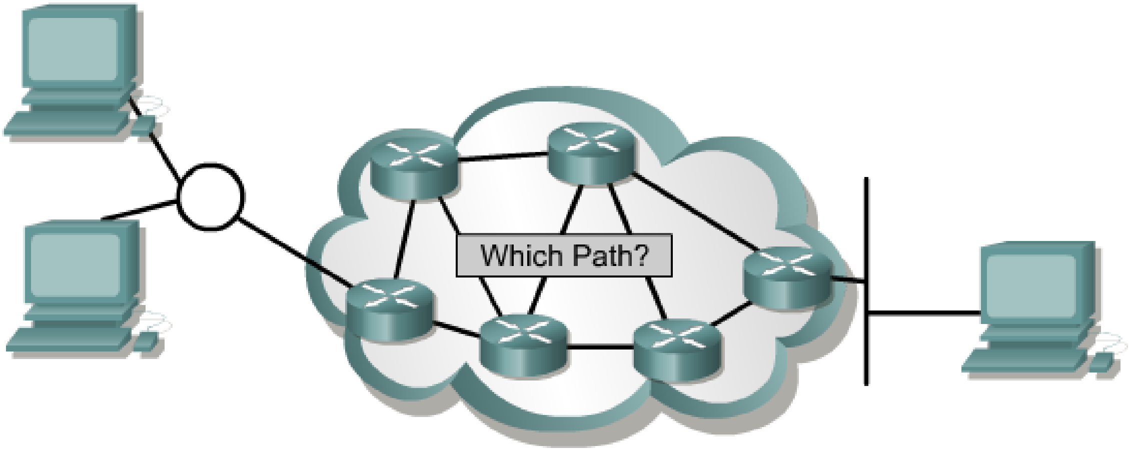
- **Connect** together **separate networks**, sometime of various networking technologies (ex: Ethernet and DSL)
- Make path determination decision based upon logical addresses (such as IP). The process is called **routing**.
- Layer 3 networking devices
- Routing and switching are similar concepts, 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 addresses)

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
  - Link State



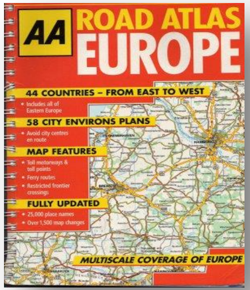
Routing. Dynamic routing

Distance Vector Protocols

- Each router tell its neighbors about its view over the network
- Routes are advertised as a vector of distance and direction.
- Routers do not have knowledge of the entire path to a destination

Link State Protocols

- Each router tells the world about its neighbors
- Routes are computed based on the network connectivity map (topological database)
- Routers have knowledge of the entire path to a destination

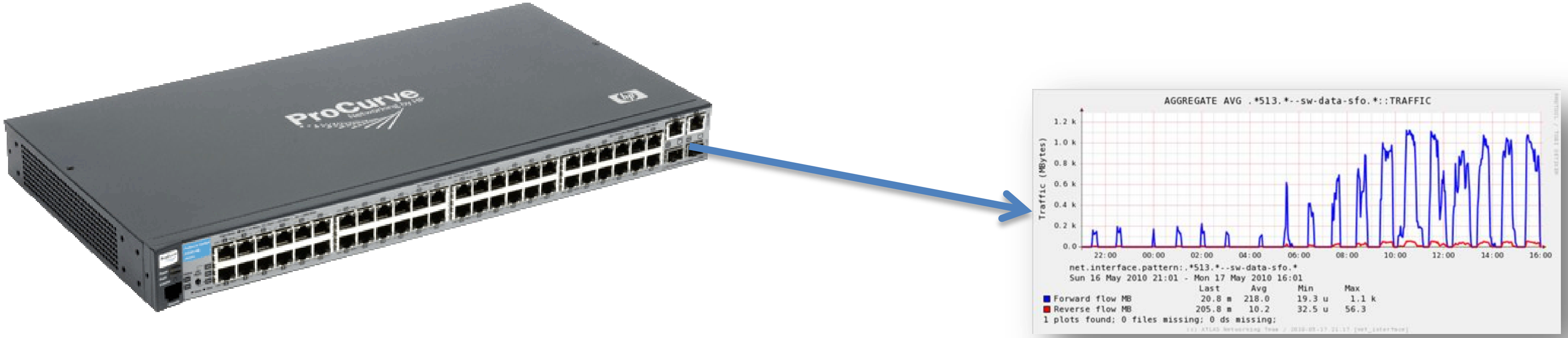


Outline

- Introduction
  - Networking basics
  - OSI reference model
- Technologies and protocols
  - Ethernet
  - IP (Internet Protocol)
  - TCP vs. UDP
  - Routing
- Network monitoring
- Software defined networking

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);



Network Monitoring. sFlow & NetFlow

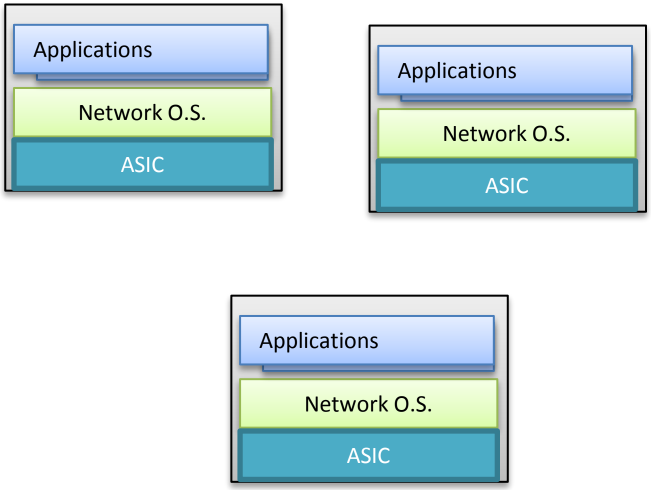
- Network monitoring technology to gather flow-related statistics;
- Can track the source and destination for packets that passes through an interface;

- sFlow compute statistics based on a sampling mechanism;
- NetFlow keeps a record for every flow. If needed, it can also use sampling.

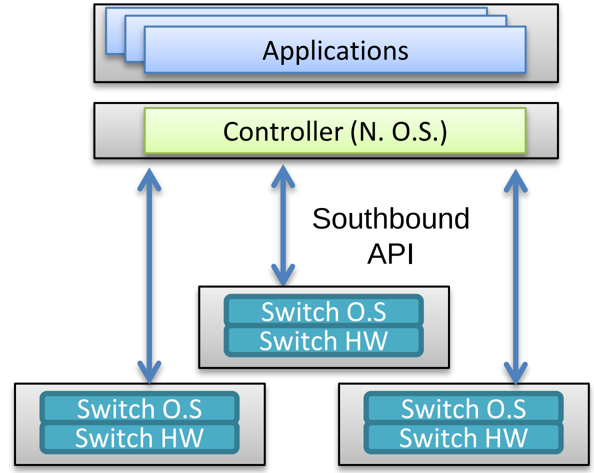
Outline

- Introduction
  - Networking basics
  - OSI reference model
- Technologies and protocols
  - Ethernet
  - IP (Internet Protocol)
  - TCP vs. UDP
  - Routing
- Network monitoring
- Software defined networking

Software Defined Networking



- **Distributed protocols**
  - Each switch has a brain
  - Hard to achieve optimal solution
- **Network configured indirectly**
  - Configure protocols
  - Hope protocols converge



- **Global view of the network**
  - Applications can achieve optimal performance
- **Southbound API gives fine grained control over switch**
  - Network configured directly
  - Allows automation
  - Allows definition of new interfaces

The show must go on …

Data Networks

Networking for Data Acquisition

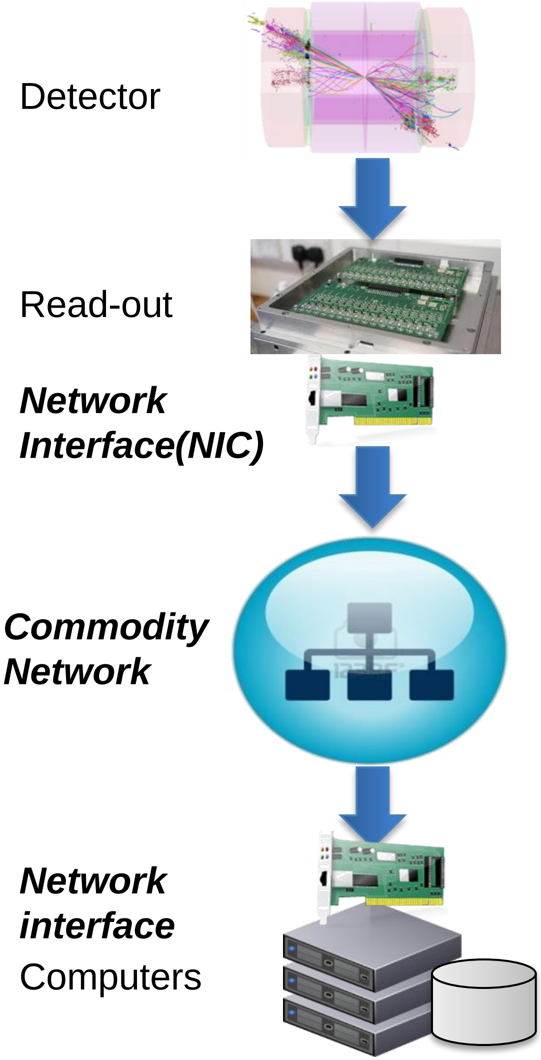
Acknowledgements
Stefan Stancu
Eukeni Pozo
Wainer Vandelli

Dan Octavian Savu
&
Silvia Fressard-Batraneanu

CERN

ISOTDAQ 2015, Rio de Janeiro

Data Acquisition uses networks

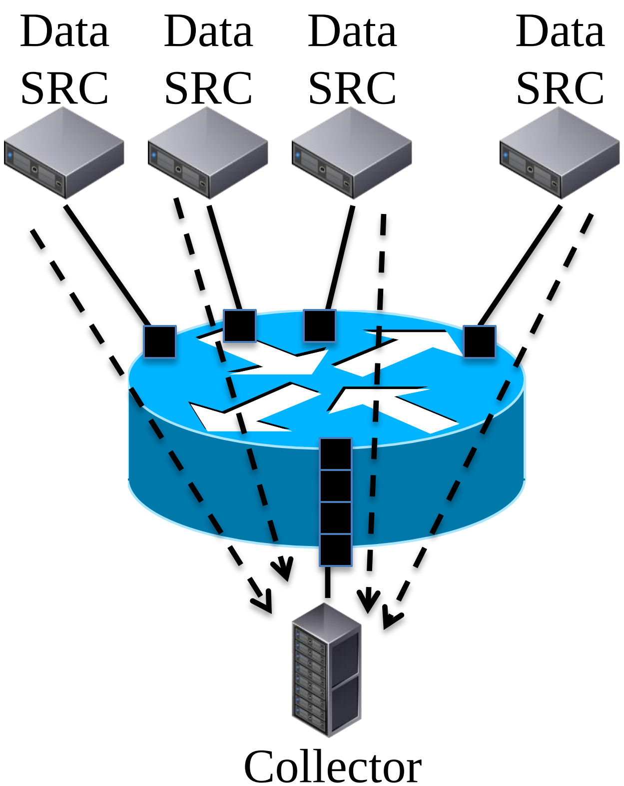


- Detector
  - Measure physical phenomena
- Read-Out
  - Digitize and perform basic processing
  - Possibly data buffers
  - **Interface to network**
- **Commodity Network**
  - Connect all read-outs to analysis computers
  - Allows computers to collect data from all sources
- Computer(s)
  - **Interface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

Outline

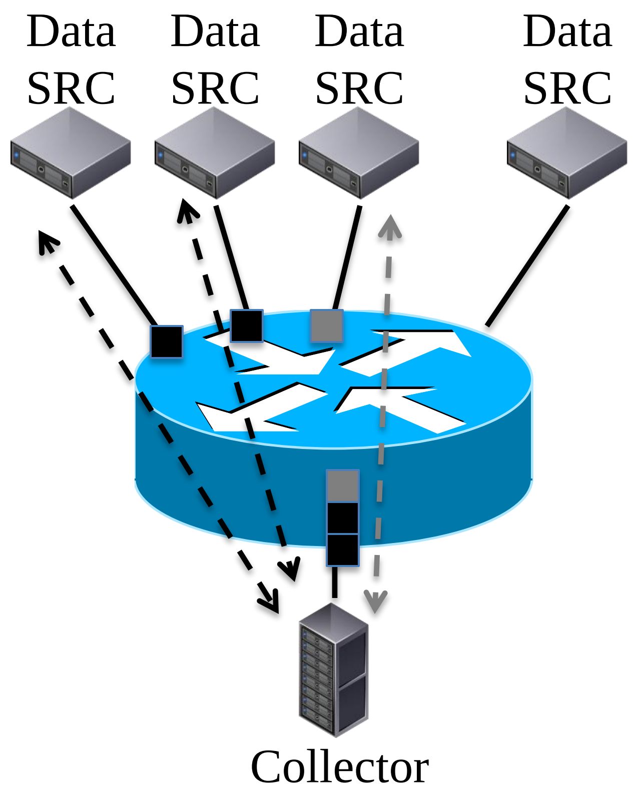
- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

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
- Security
- Performance
  - *High throughput AND low latency*
  - *Substantial tuning*
    - *Data flow software*
    - *Network itself*
  - *Advanced performance monitoring*
- Low cost

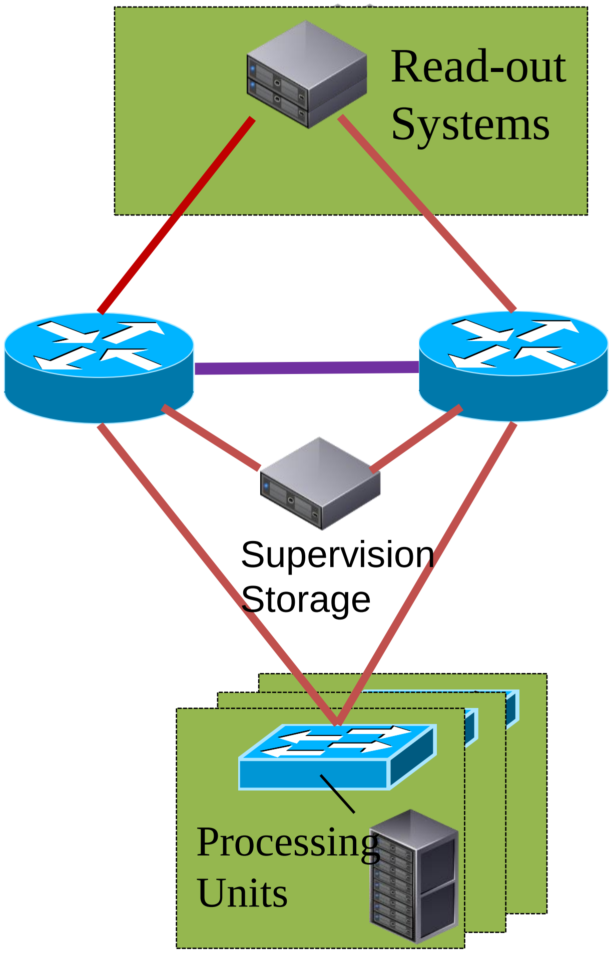
LHC DAQ networks characteristics

- Private local networks
- Flat network topology
- Congestion hot-spots
- Packet loss caused by
  - HW failures
  - Transmission errors
  - Congestion manifested by discarded packets
- 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



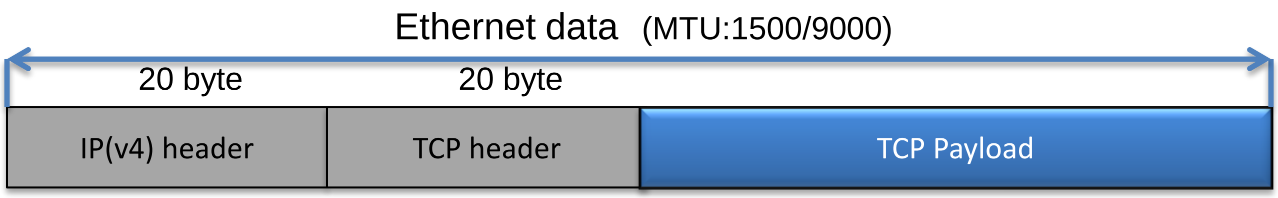
ATLAS DAQ Network

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

Outline

- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

TCP: Fragmentation



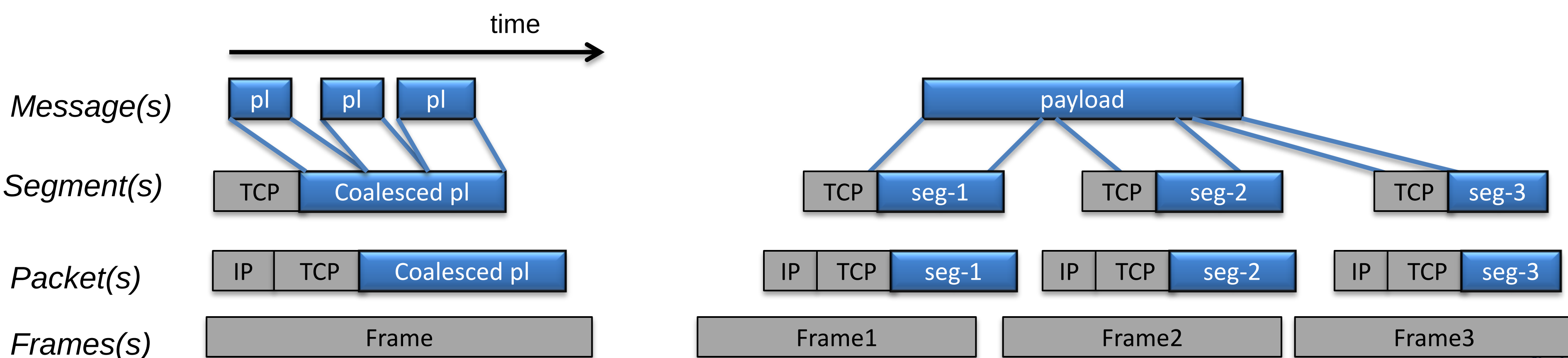
Buffers data for a short while before sending it

Knows the MTU size

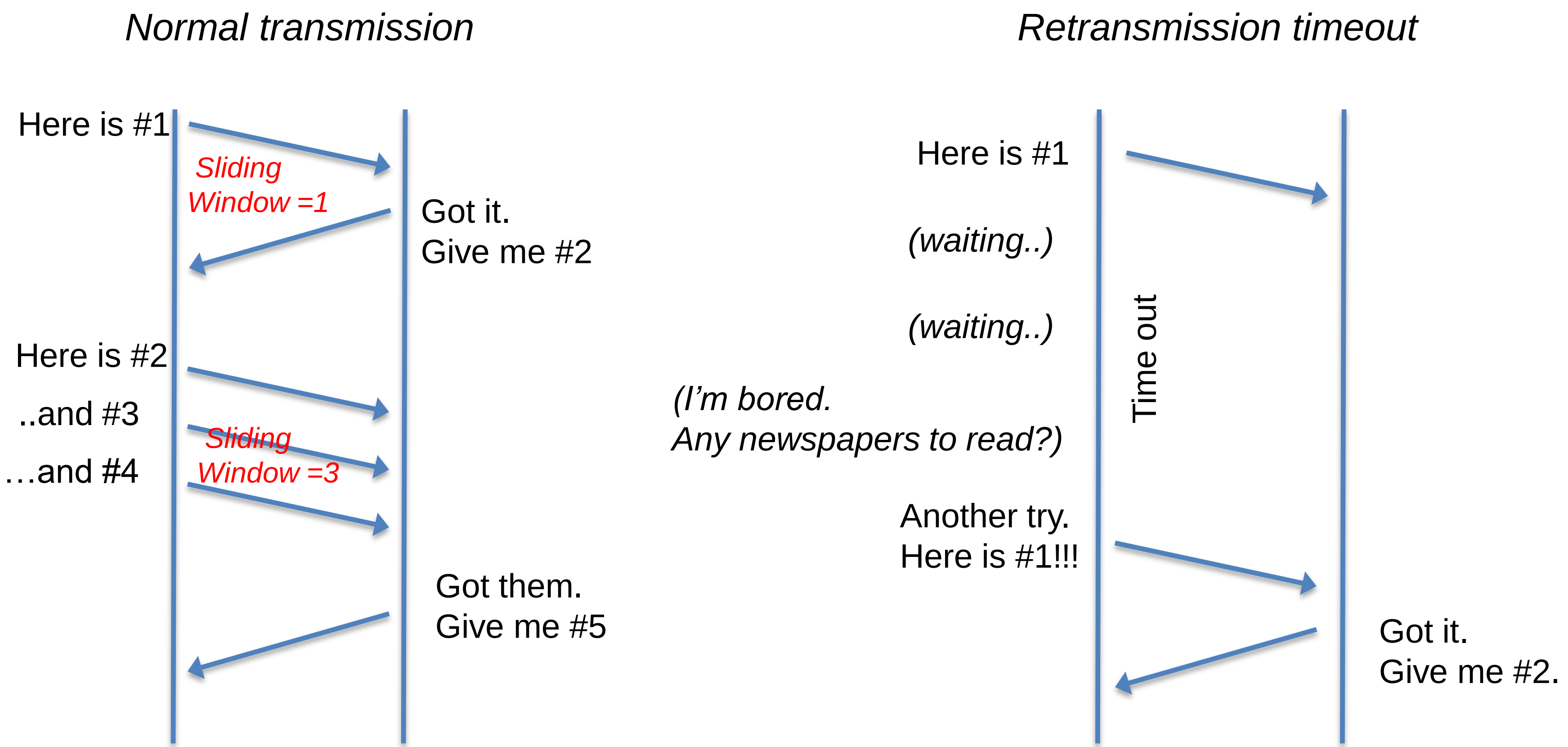
Coalesces or segments data depending on payload size

Payload < MTU
May coalesce using the nagle algorithm

Payload > MTU
Does segmentation
No IP fragmentation

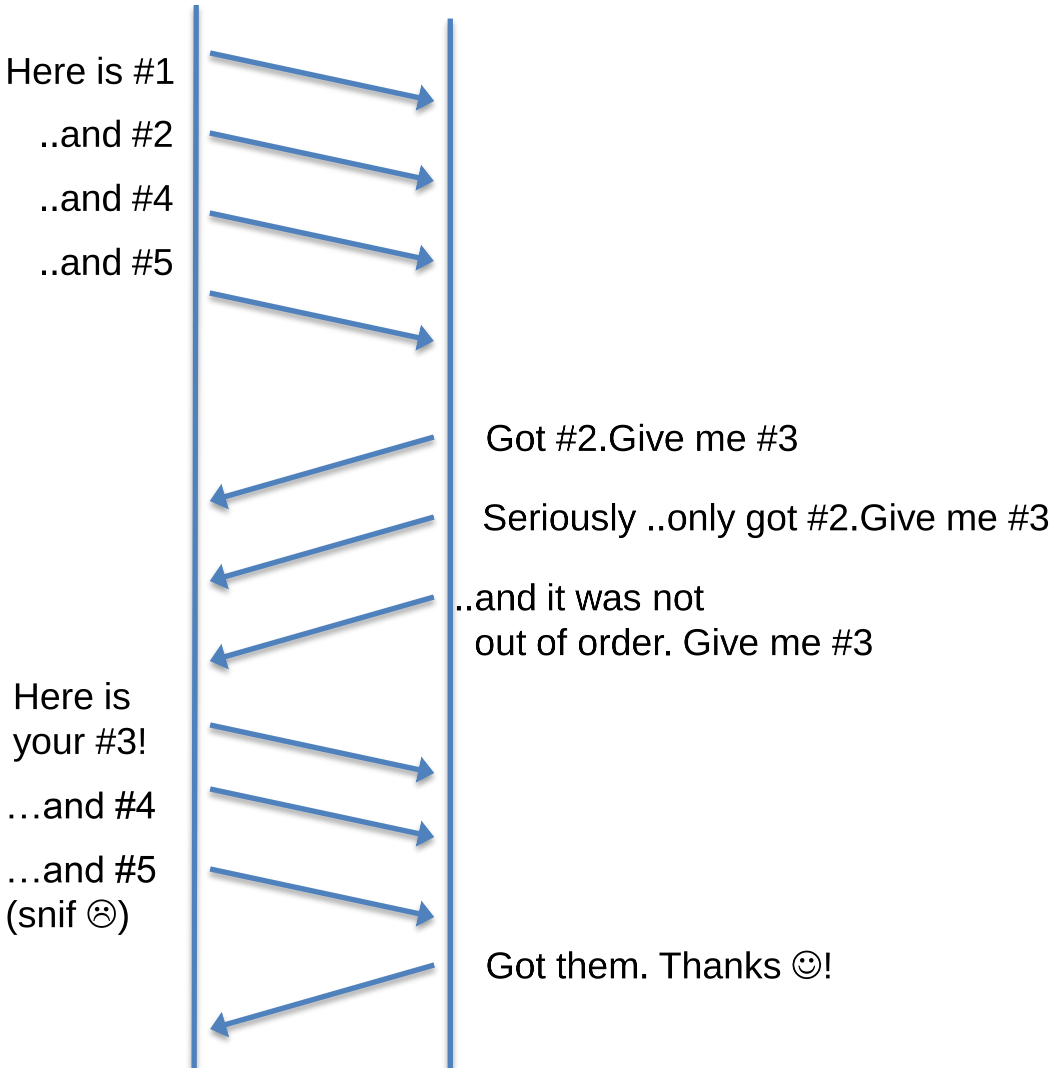


TCP: Reliable transmission(1)

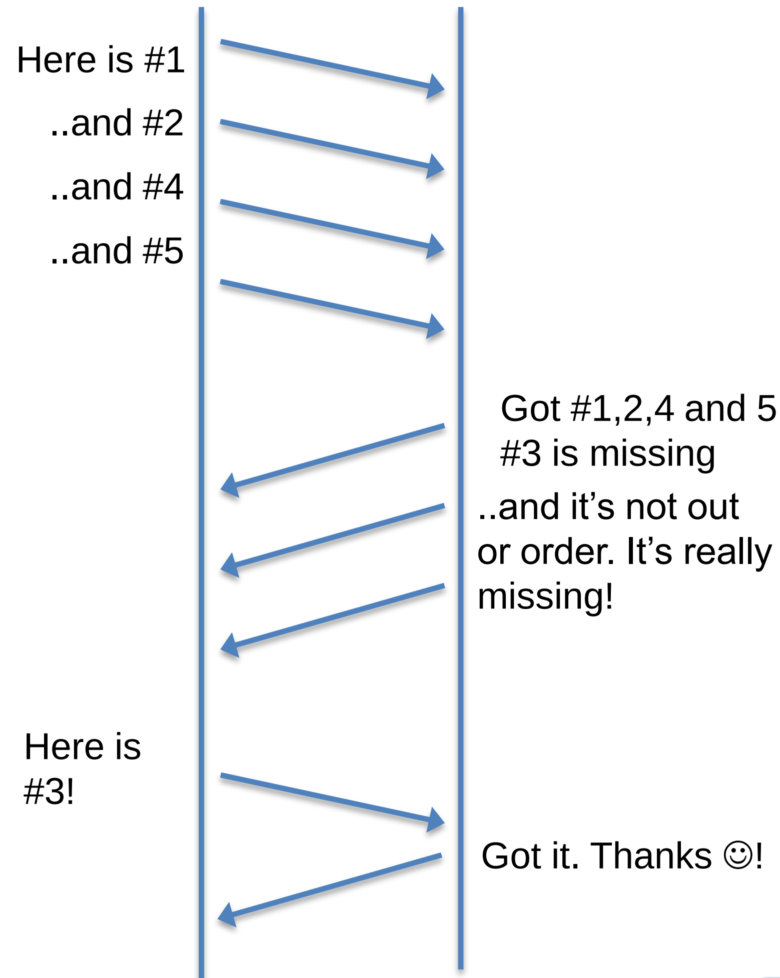


TCP: Reliable transmission(2)

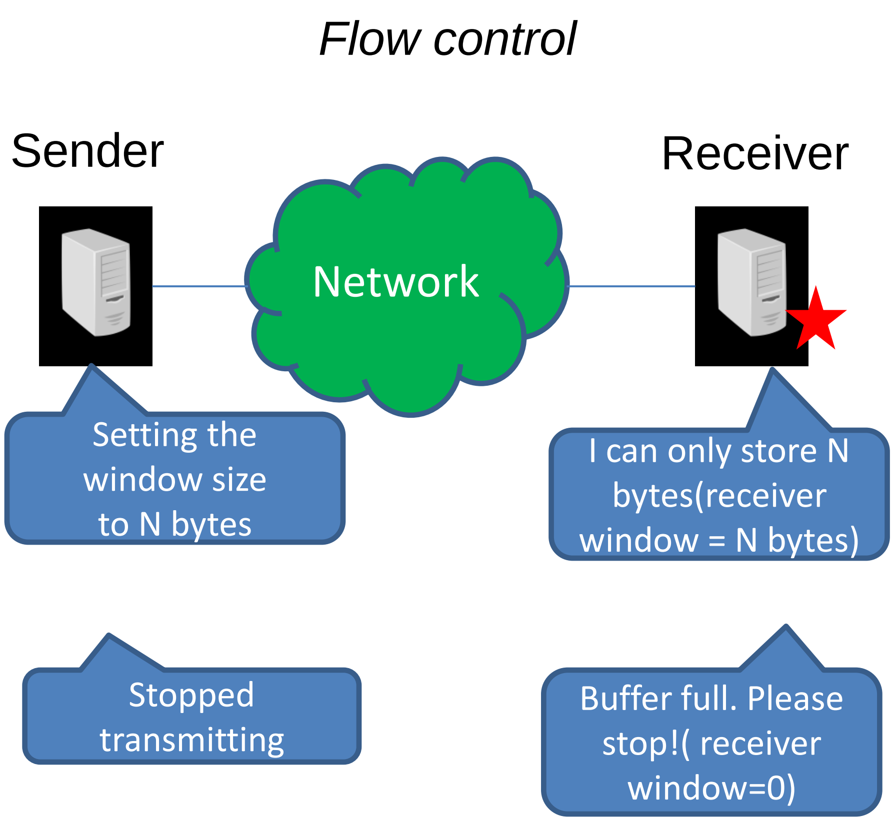
Cumulative acknowledgement

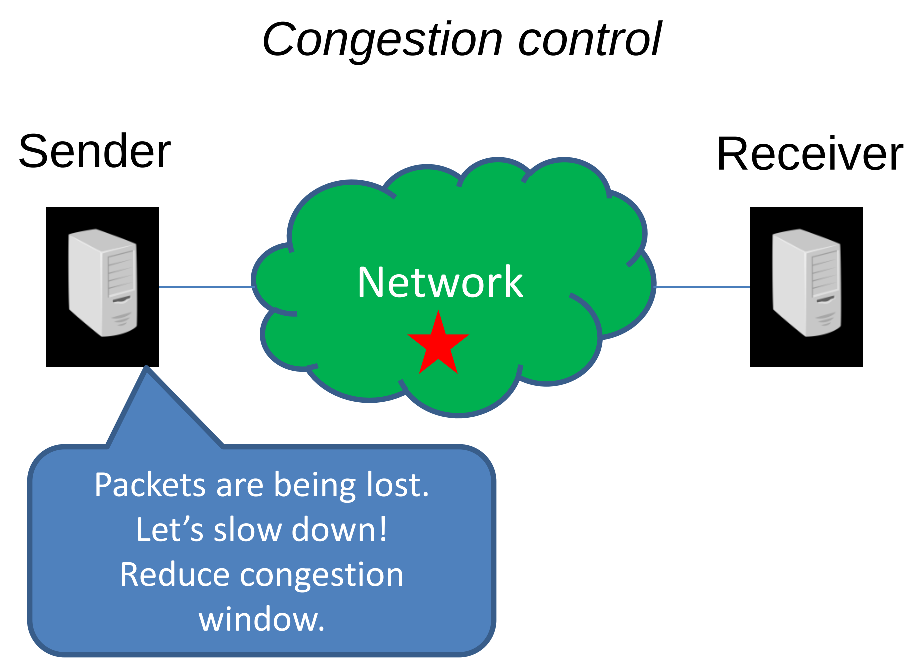


Selective acknowledgement

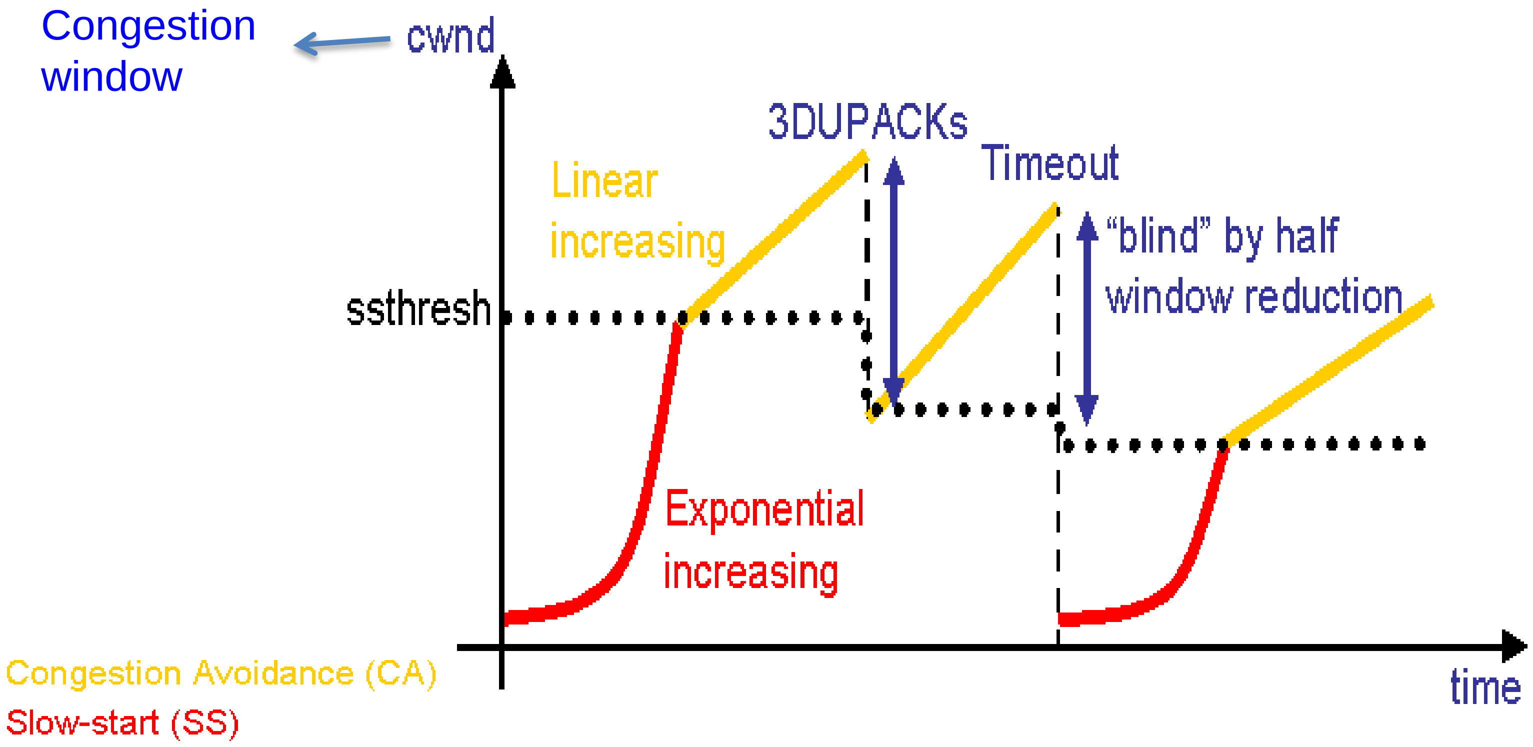


TCP: Flow vs congestion control



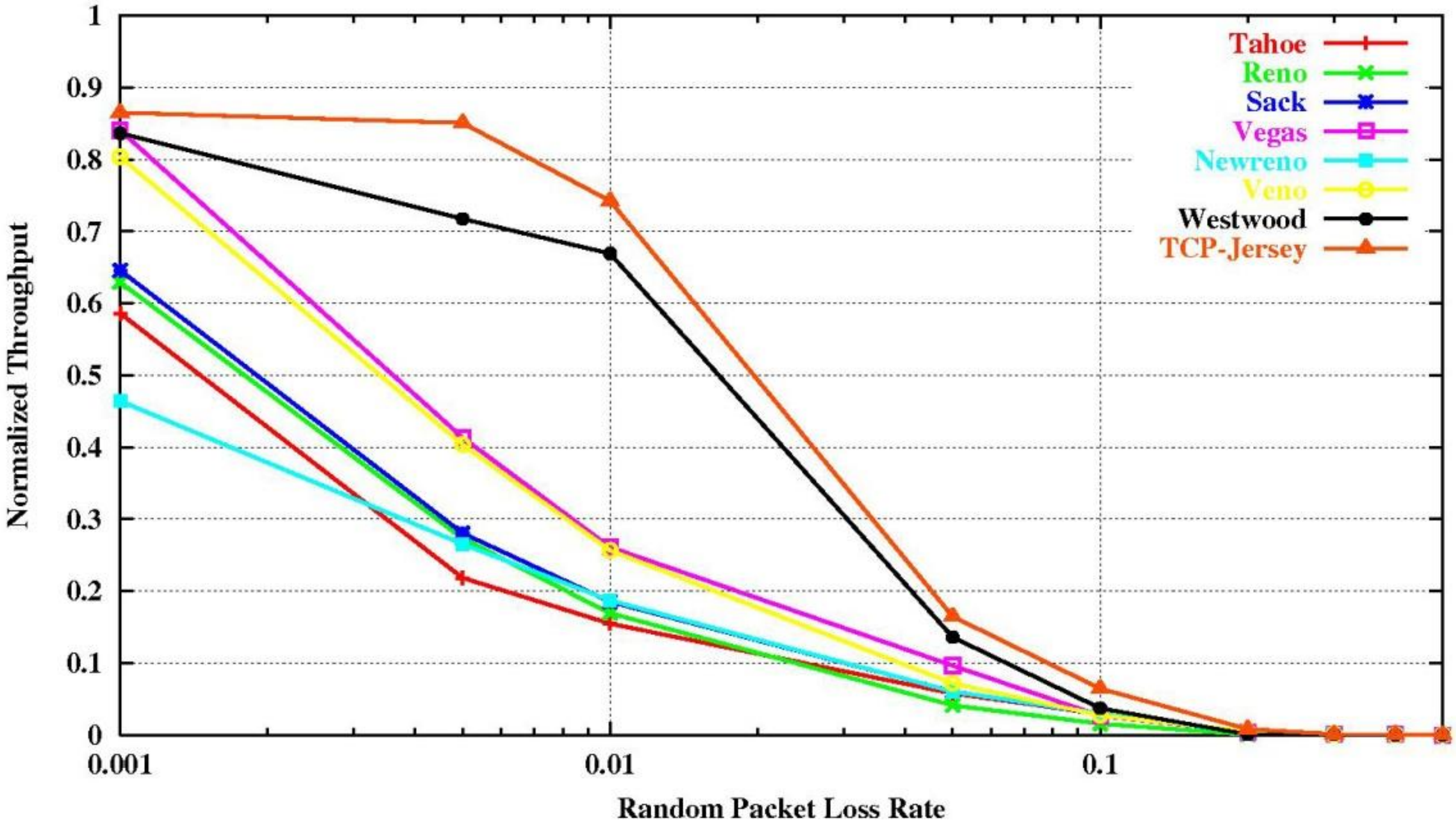


TCP: Congestion control example



TCP Westwood

TCP Performance with packet loss

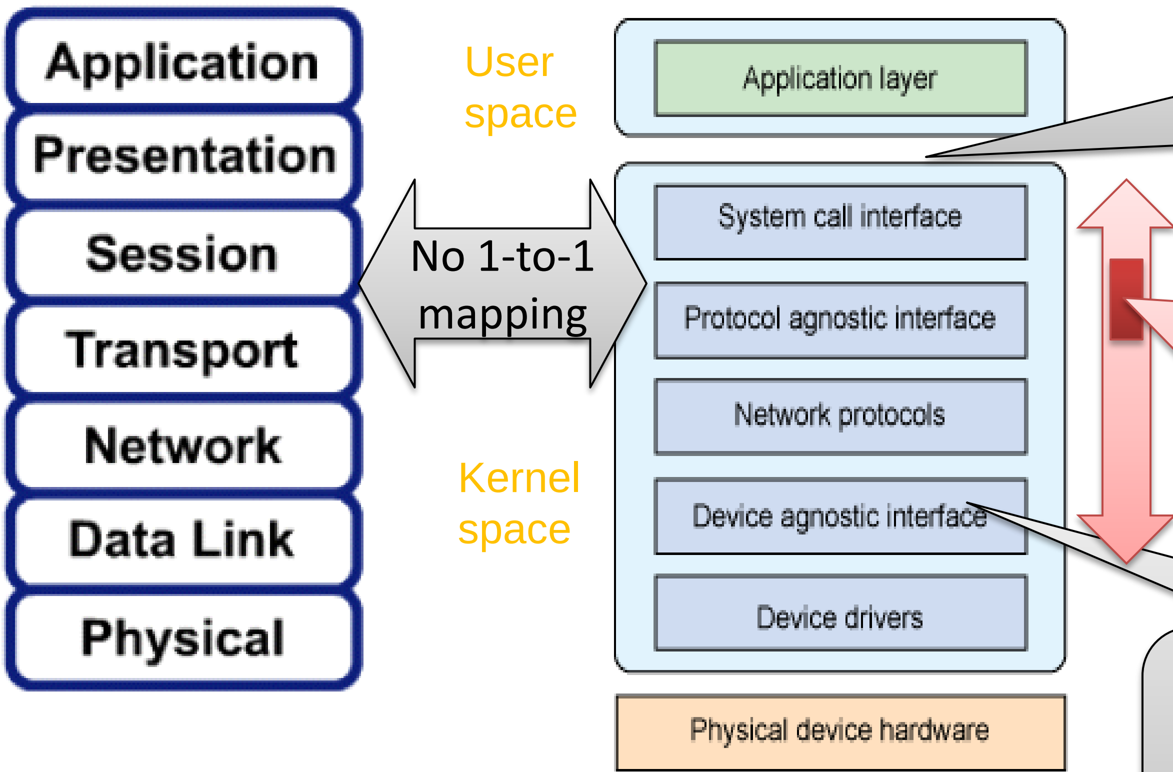


Outline

- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

From theory to practice : Linux N/W stack

OSI model | **Linux network stack**



System calls

- Application access to the network stack

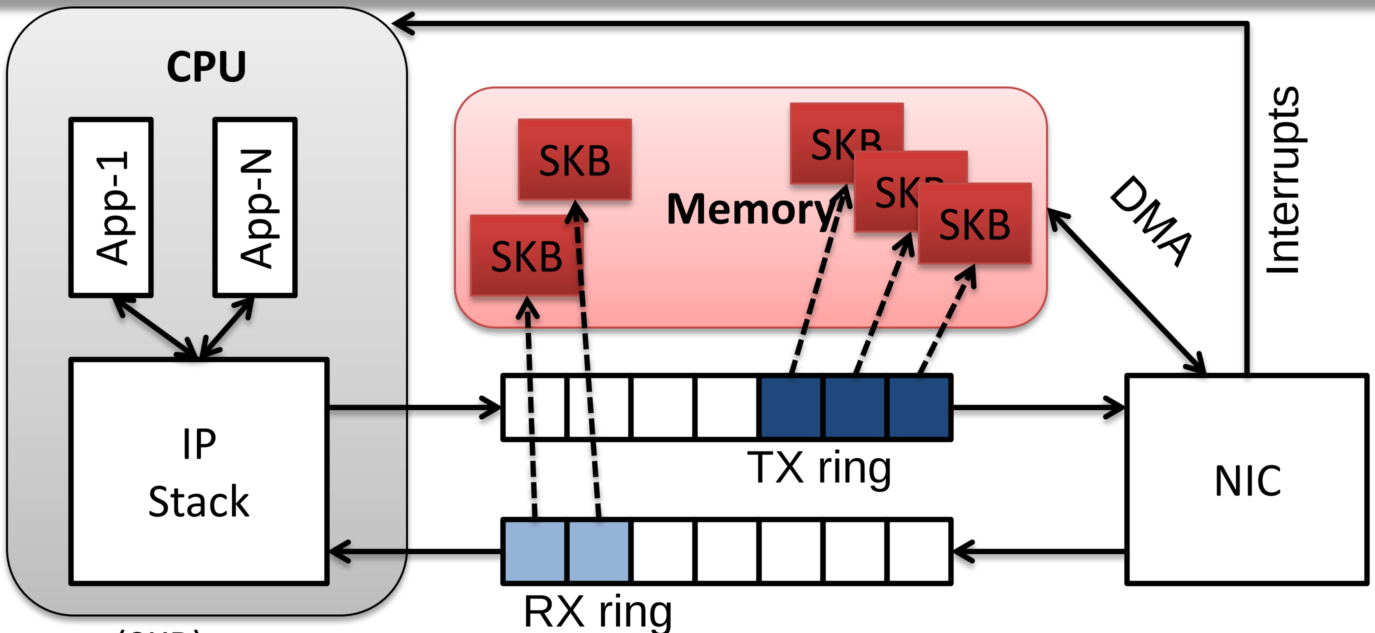
Socket buffer (SKB)

- Unique data structure used to move data between layers

Layering provides:

- a common interface to many transport protocols
- a common interface to many devices

Kernel – NIC interaction



Send

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

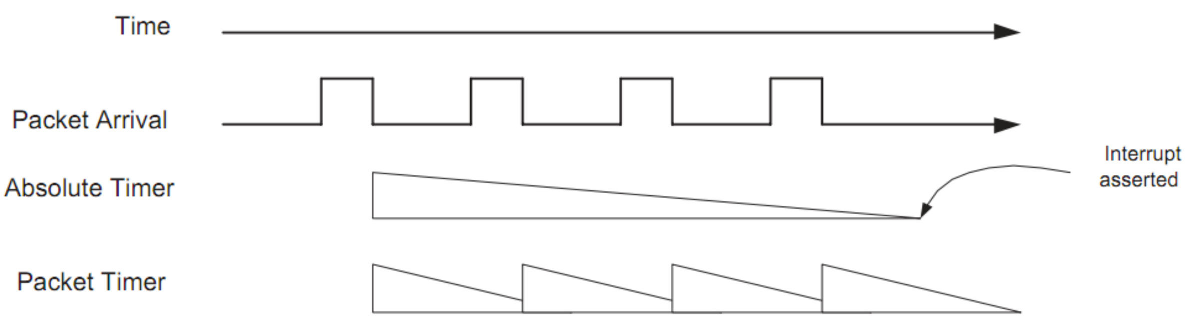
Receive

- 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



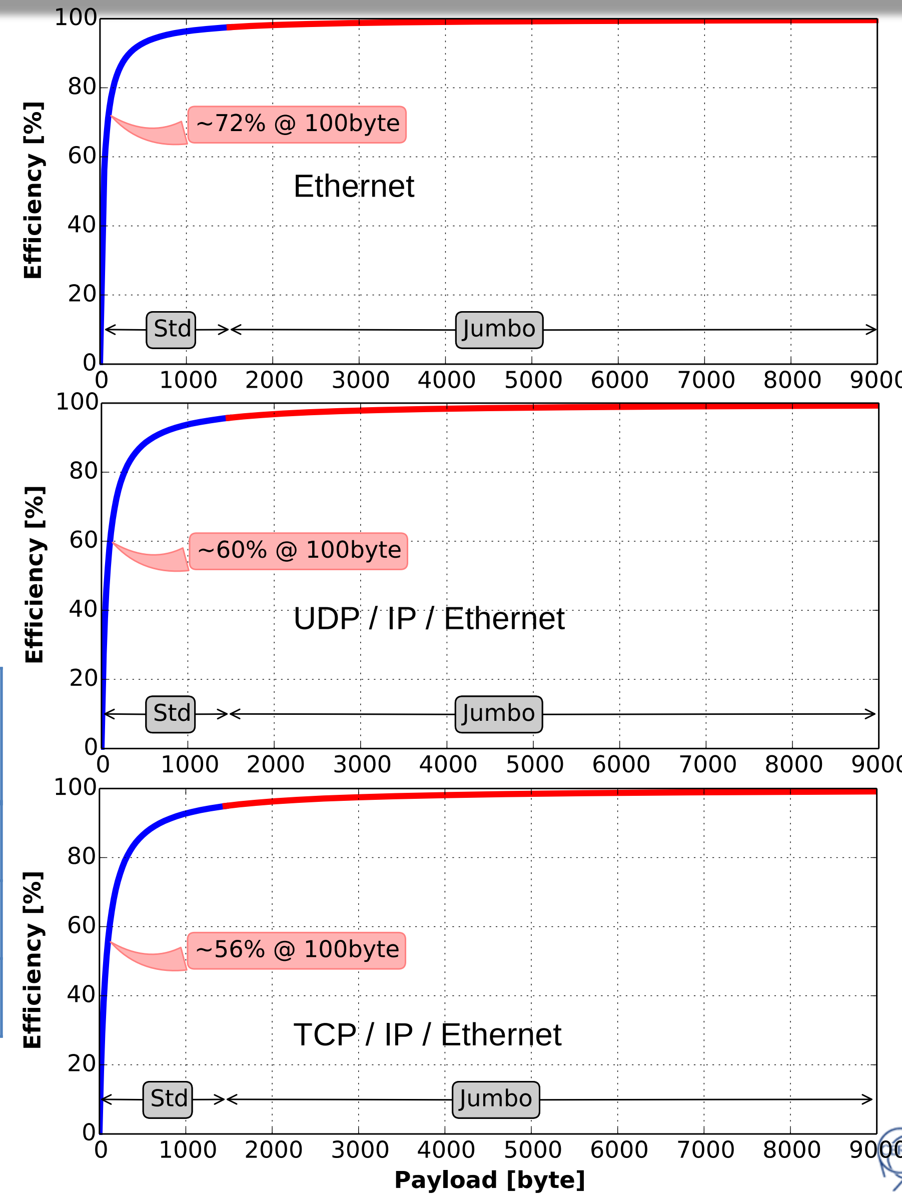
Precautions

- 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

Encapsulation – Efficiency

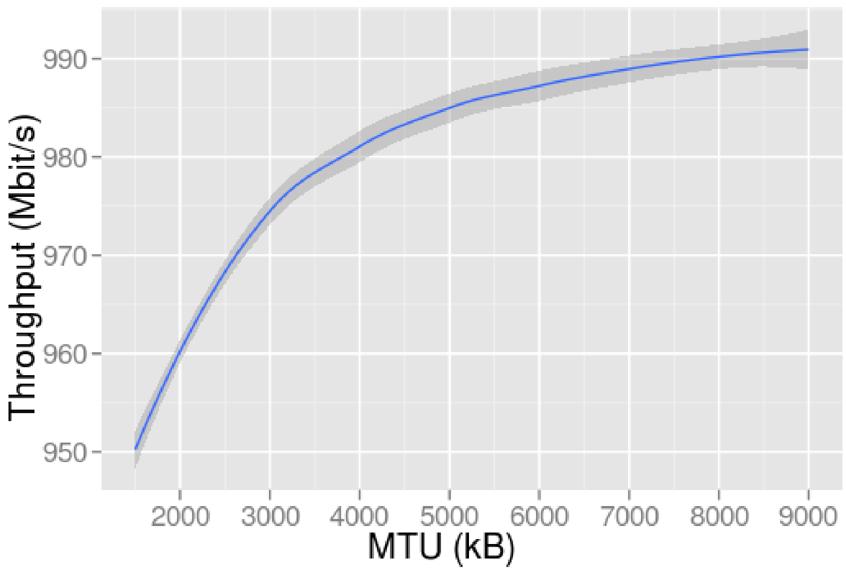
$$Efficiency = \frac{Payload}{Payload + Overhead}$$

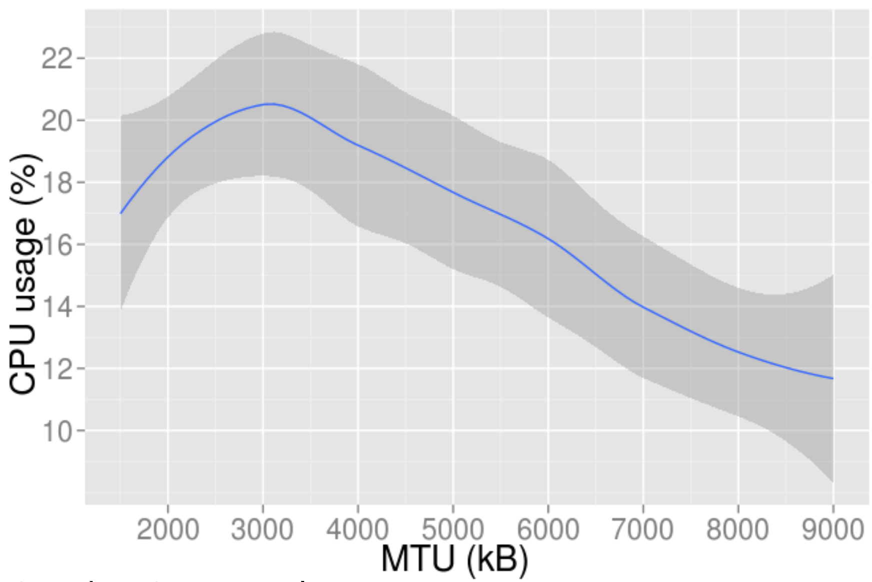
Encapsulation	Efficiency (100 byte)	Efficiency (1 byte)
Ethernet	72%	1.2%
UDP/IP/Eth	60%	1.2%
TCP/IP/Eth	> 56%	> 1.2%



Jumbo Frames

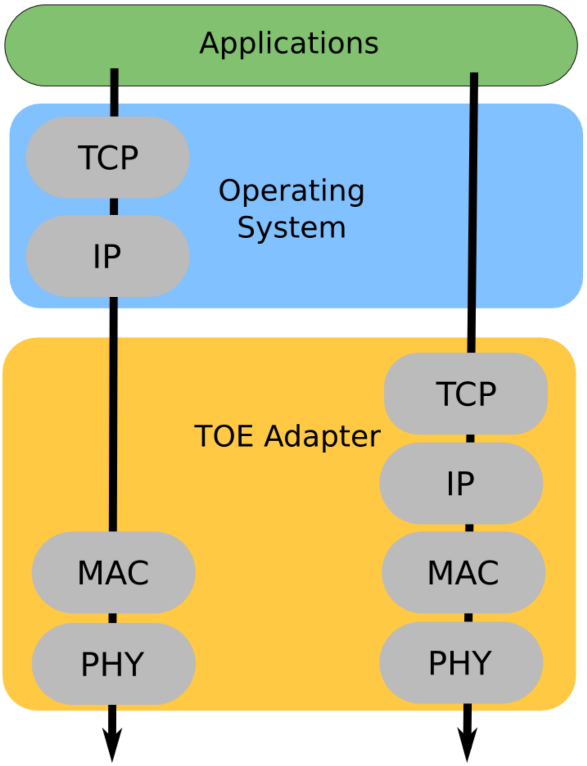
- **Improve max throughput**
  - 94% @ 1500 MTU
  - 99% @ 9000 MTU
- **Reduce the frame rate**
  - Lower interrupt rate
  - Less data dis/re – assembling for the CPU





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

NIC Offloading



TOE: export processing to hardware controllers

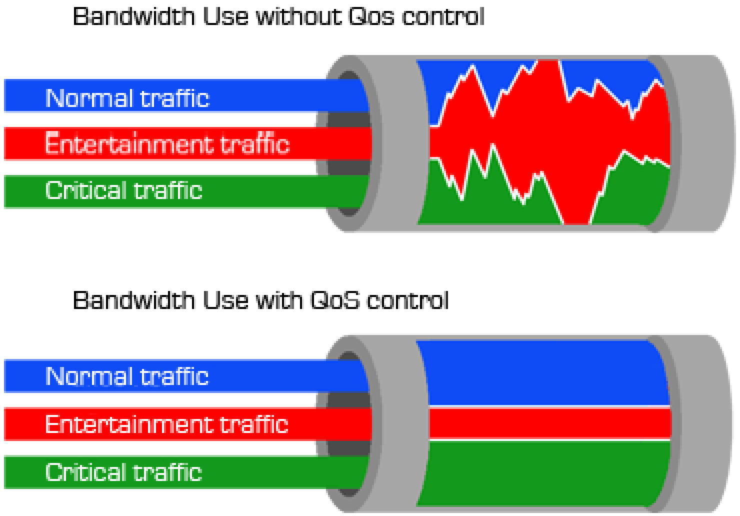
- Packet processing consists of numerous task -> CPU intensive for high bandwidth
- TCP Offload Engine: TCP/IP stack processed by the network device
  - Checksum computing
  - Transport protocol segmentation
- A TOE capable device will offer the OS a much larger MTU (SKB size)
  - TSO = TCP Segmentation Offload (send)
    - the NIC takes care of segmenting the large SKB
  - LRO = Large Receive Offload (receive)
    - the NIC assembles data from multiple frames/segments into a large SKB

Outline

- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

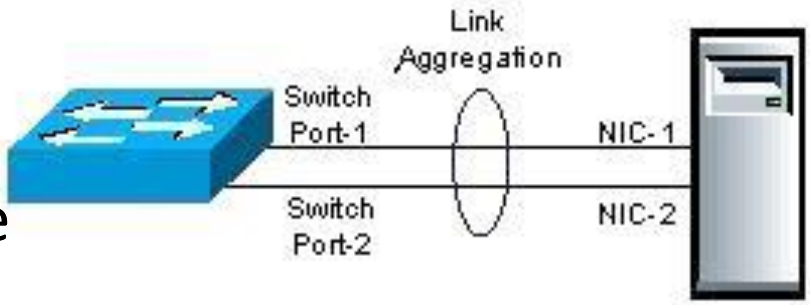
QoS

- Could be useful for DAQ
- Layer 3 - DiffServices (DSCP)
  - Set the priority in the ToS field of the IP header
  - Can be done at the application level
- Layer 2 –VLANs
  - Define overlapping(tagged) VLANs
  - Send traffic on a specific VLAN
  - Configure N/W devices to prioritize VLANs



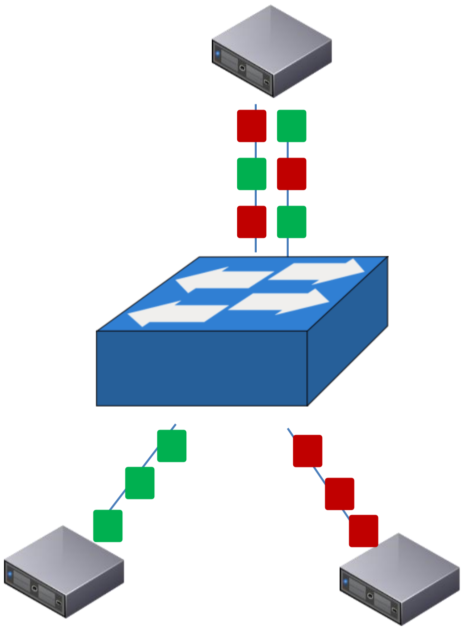
Link Aggregation - basics

- Combining two or more links to
  - increase throughput
  - provide redundancy
- In Linux is called bonding
- One master and one or more slaves
- Master MAC address becomes bond MAC address
- Can be static or dynamic (LACP)



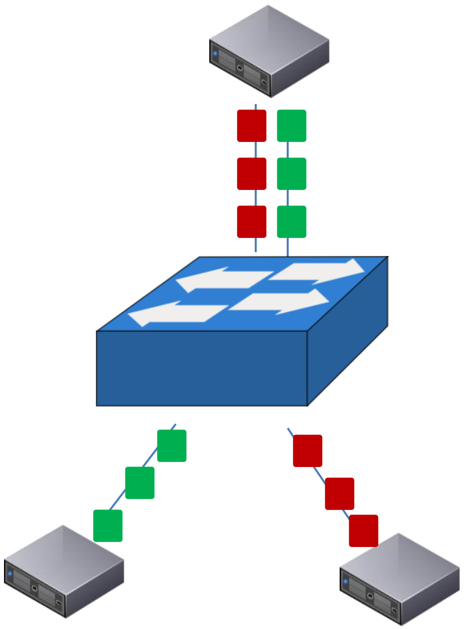
Link Aggregation – load balancing policies

Round robin



High throughput
Out of order packets

Hash-based



High throughput for many conversations
No out of order packets

Outline

- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

Basic optimizations for DAQ

- Hosts
  - Mainly for reception side
    - *Reception* is the more resource consuming side mainly because it has to reorder packets
  - Provide large kernel buffers and large socket buffer for the application (sysctl!)
  - Tune IRQ moderation
  - If possible, enable jumbo frames
  - Be aware of the specifics of the TCP congestion mechanism on your system
    - If needed tune them

- Network devices
  - Enable jumbo frames on all ports to improve bandwidth
  - Increase buffers
    - Packet loss has a big impact on performance (see previous TCP slides)
    - Pay attention to the delay they introduce
  - Use virtual output queueing to avoid head-of-line blocking

- Both
  - Evaluate the performance of your link aggregation groups
    - If needed, change the load balancing policy
  - Use QoS to prioritize critical traffic that risks of getting lost in the congestion spots

Outline

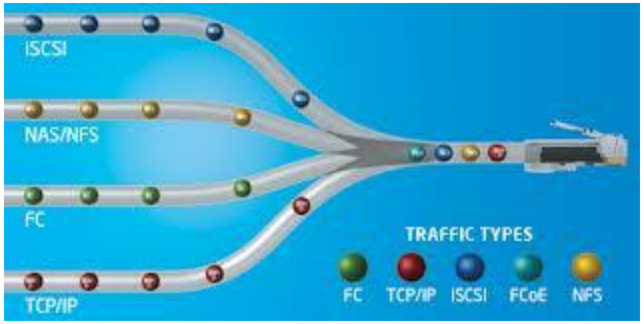
- DAQ networks for large experiments
- TCP protocol characteristics
- Linux networking characteristics and optimizations
- QoS and link aggregation specifics
- Optimization summary
- Network technology choices

Network Technology Choice

- **Ethernet** is a de-facto standard
  - OSI Layer-2
  - Largely used in the industry
  - Many providers
  - Its evolution (Converged Enhanced Ethernet)
    - Brings more reliability without the TCP complexity
    - Makes Ethernet usable for storage area networks

- **Infiniband**
  - One main provider
  - One step ahead regarding link speeds
  - Steep learning curve, not so much know-how around
  - Lower cost per port, proprietary connectivity more expensive
  - Key advantages related to performance and packet loss
  - More and more used in HPC
- Myrinet, FiberChannel ..

Converged (Enhanced) Ethernet

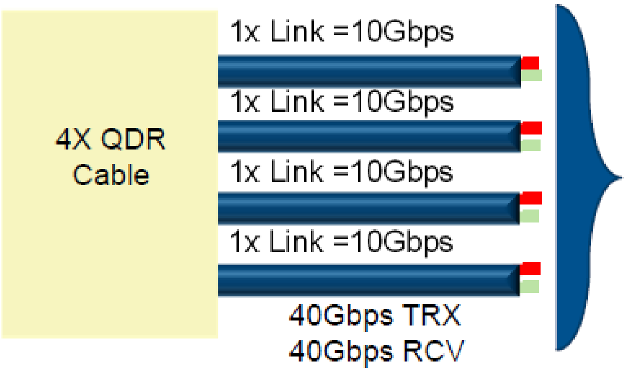


- Also known as "Lossless Ethernet"
- Aims to eliminate loss due to queue overflow and to be able to allocate bandwidth on links for selected traffic
- Combines a number of optional Ethernet standards into one umbrella:
  - Priority based flow control(PFC): Link level flow control for each Class of Service (CoS)
  - Enhanced Transmission Selection(ETS): Bandwidth assignment to each CoS
  - End-to-end Congestion notification(ECN): Per flow congestion control to supplement per link flow control
  - Data Center Bridging eXchange (DCBX): Exchange protocol used for conveying capabilities between neighbours

Infiniband

- *High speed*
  - *Uses multiple differential pairs(x4, x8,x12)*

    SDR - 2.5Gb/s per lane (10Gb/s for 4x)
    DDR - 5Gb/s per lane (20Gb/s for 4x)
    QDR - 10Gb/s per lane (40Gb/s for 4x)
    FDR - 14Gb/s per lane (56Gb/s for 4x)
    EDR (EDR) - 25Gb/s per lane (100Gb/s for 4x)

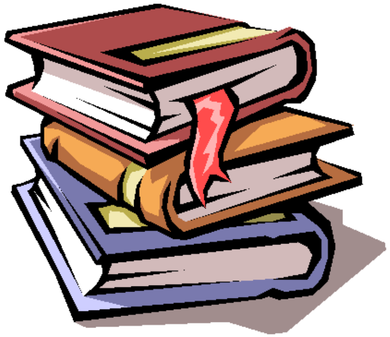


- *Low latency* **–** OSI layers 2-4 implemented in hardware
- *Low CPU Utilization with RDMA* (Remote Direct Memory Access)
  - communication bypasses the OS
- *Absolute credit based flow control*
  - assures NO packet loss within fabric even in the presence of congestion
  - receiver guarantees that enough space is allocated for N data blocks
- *Reliable transport protocols* for other packet loss

INFINIBAND™

References

- **Wikipedia**
- **IETF RFCs**
- **« man » pages**
- **Conference proceedings and journals**



A few noticeable:

1. Wenji Wu. The Performance Analysis of Linux Networking Packet Receiving
   http://lss.fnal.gov/archive/2006/pub/fermilab-pub-06-406-cd.pdf

2. Sequence diagrams for TCP/IP stack and many protocols
   http://www.eventhelix.com/RealtimeMantra/Networking/

3. 10 Gigabit Ethernet Association
   http://www.10gea.org/tcp-ip-offload-engine-toe.htm

4. Binary Increase Congestion Control for Fast, Long Distance Networks
   http://netsrv.csc.ncsu.edu/export/bitcp.pdf

5. Designing Cloud and Grid Computing Systems with InfiniBand and High Speed Ethernet:
   http://www.ics.uci.edu/~ccgrid11/files/ccgrid11-ib-hse_last.pdf

6. Unix Network Programming, Volume 1: The Sockets Networking API (3rd Edition) , Addison-Wesley Professional