Networking for Data Acquisition Systems

Fabrice Le Goff - 14/02/2018 - ISOTDAQ

Outline

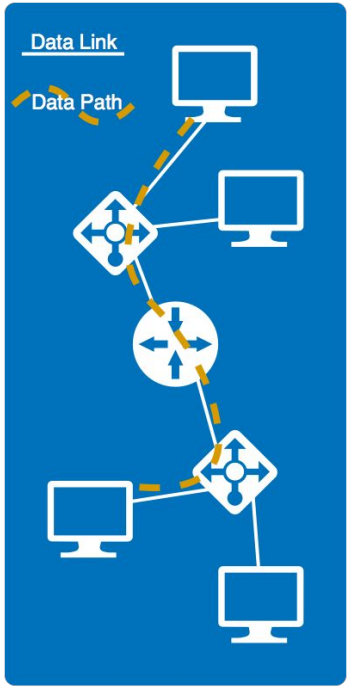
- Generalities
- The OSI Model
- Ethernet and Local Area Networks
- IP and Routing
- TCP, UDP and Transport
- Efficiency
- Networking for Data Acquisition Systems

Generalities

Definitions

A network is the interconnection of computing devices, called **nodes**, via **data links** to exchange information.

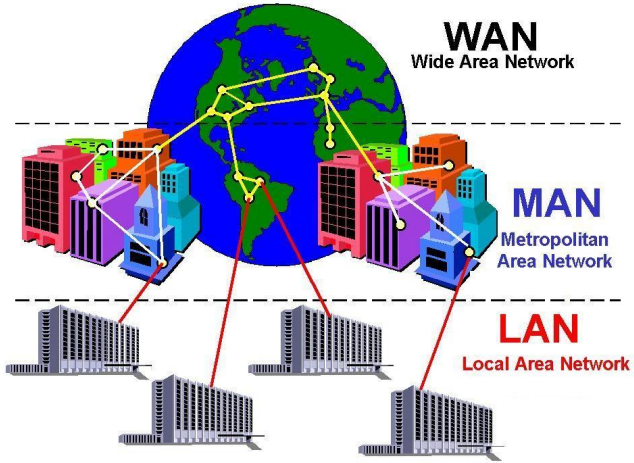
- Nodes
  - **Source**: emitting data
  - **Destination**: receiving data
  - Intermediate: transferring data (router, switch, bridge, etc.)
- Information is exchanged in the form of blocks, called **packets**
  - Except for some point-to-point connections that exchange continuous bit streams
  - Packet (L3) or frame (L2), datagram (L4), segment (L4) depending on the protocol
- A **data path**, or **route**, is a succession of nodes and links connecting a source to a destination
  - Two nodes can be connected via different paths

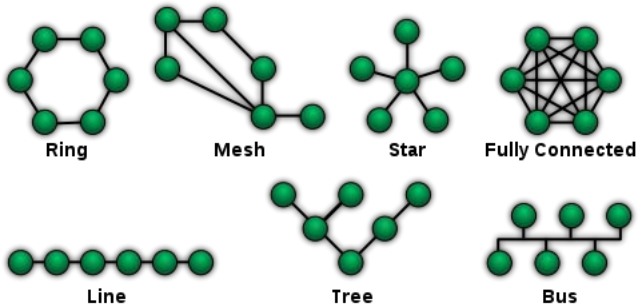


Network Types

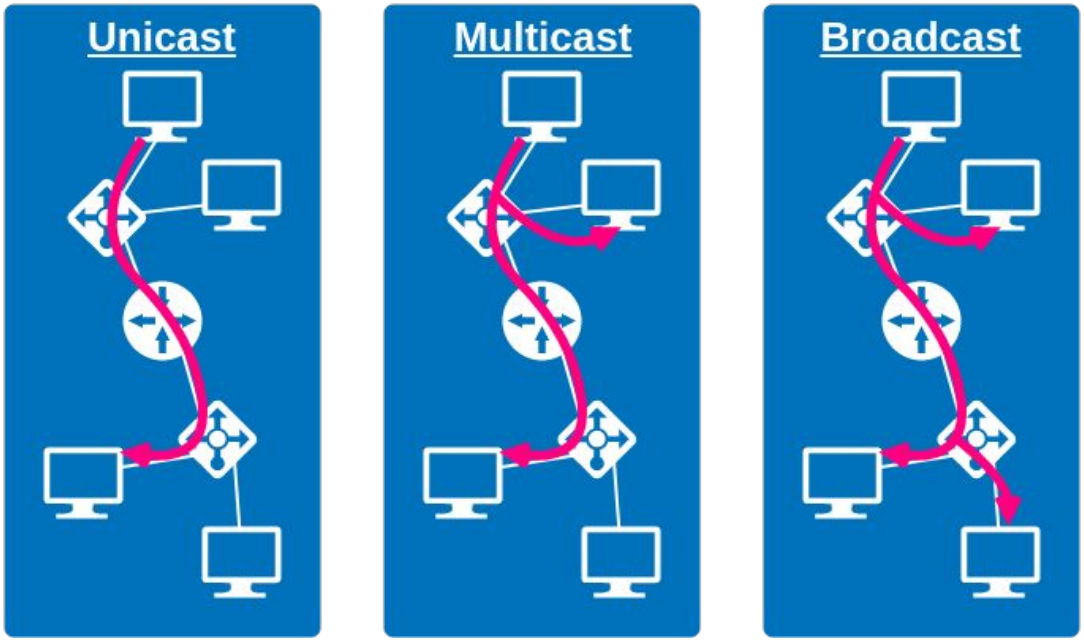
Networks are categorized according to different criteria.

- Physical size
  - **LAN**: Local Area Network
  - MAN: Metropolitan Area Network
  - **WAN**: Wide Area Network
- Topology:
  - Physical: position of nodes and links with respect to "cabling"
  - Logical: how signals propagate on the physical links
  - Example: in a building computers are usually connected to a central concentrator making a physical star topology, while at the logical level they use Ethernet which implements a bus topology (the concentrator propagates the signals to all links).





Communication Types

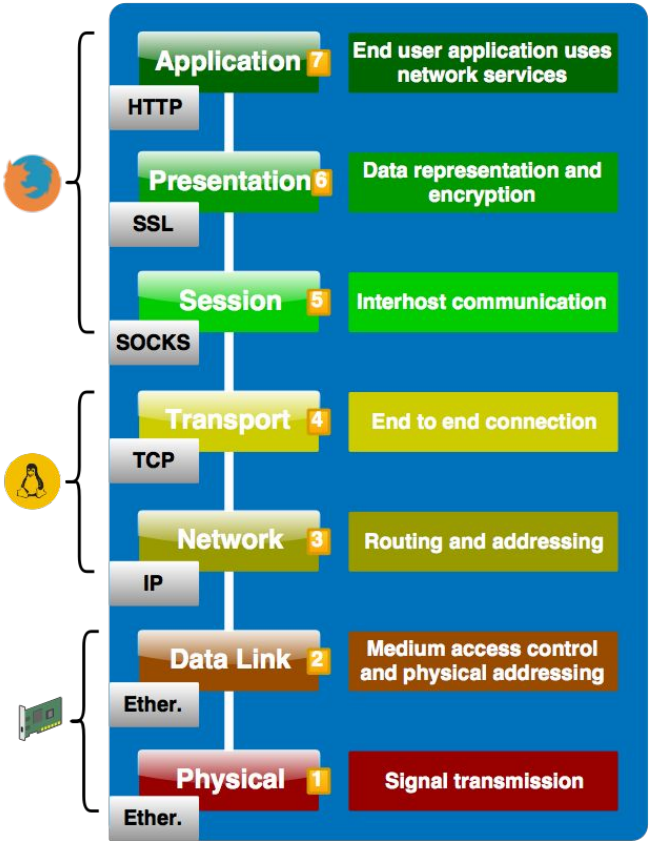


The OSI Model

The OSI Model

The ISO's (International Organization for Standardization) project OSI (Open Systems Interconnection) has defined a **conceptual model** (ISO/IEC 7498-1) that provides a common basis for coordination of standards development for the purpose of **systems interconnection**.

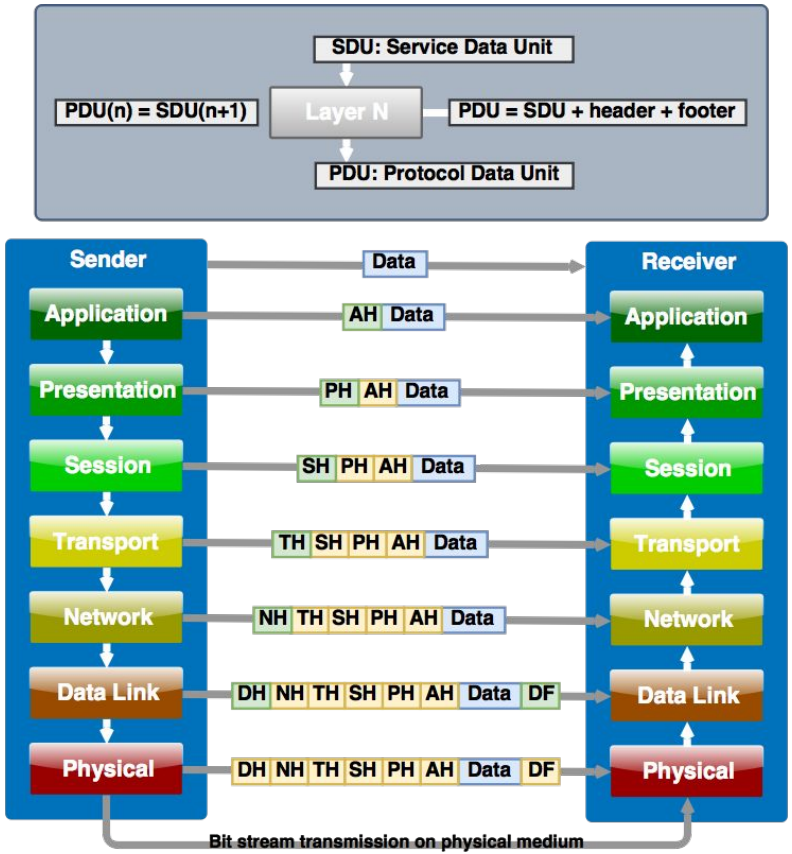
- Defines **7 layers** that splits responsibilities and functionalities of networking communication
- Layer interfaces allow actors of the "industry" to develop functionalities independently
- It's a framework not an actual implementation nor a strict guide
- Most network technologies reflect this layered structure



OSI Communications

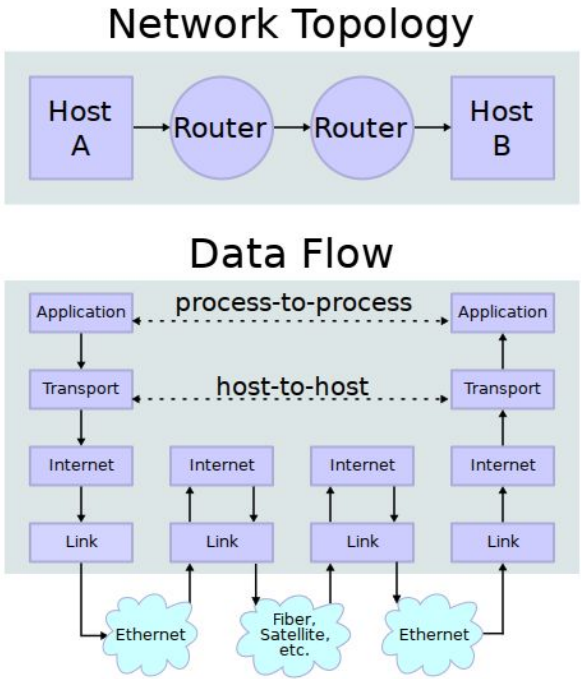
1. Data is produced at the topmost layer.
2. Data is passed to the layer N-1.
3. Layer N can add a header/footer to the data received from layer N+1.
4. At the physical layer, the data are transmitted to the receiving node.
5. On the receiving node, each layer removes its header/footer and passes it to the layer N+1

Sender's and receiver's entities of layer N exchange together **PDU** (Protocol Data Units) via a specific **protocol**.



The TCP/IP Model

- Simpler than OSI
- Most used protocol stack
- 4-layer model:
  - Link: communications within a LAN
  - Internet: inter-LAN communications
  - Transport: host-to-host communications
  - Application: process-to-process communications
- Internet layer is implemented with IP
- Transport layer is implemented with TCP (or UDP)



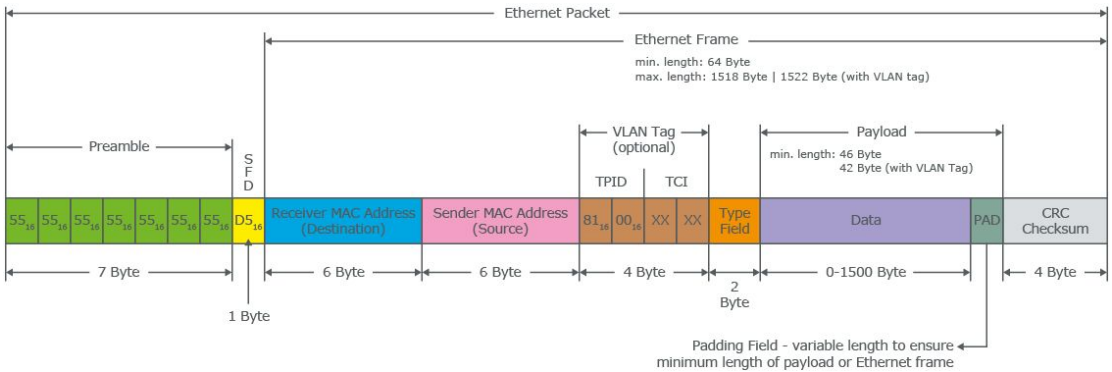
Ethernet and Local Area Networks

Ethernet

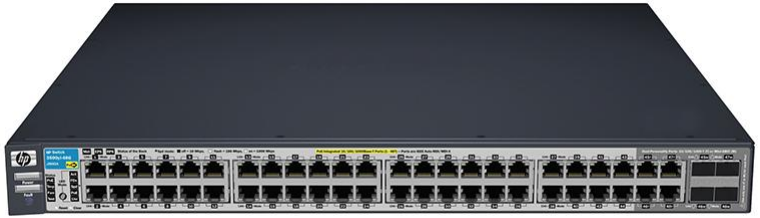
- Set of technologies that defines physical link protocols (layer 1) and data-link protocol (layer 2)
- Most widespread technology used for LAN communications
- **LAN <=> broadcast domain**
  - Every node can talk directly to every other node
  - Every broadcast frame will reach all other nodes
- Ethernet is a **logical bus**
  - Nodes compete for the physical resource
  - Need to synchronization of nodes transmitting at the same time (collisions)
  - This is the reason why there is a maximum (and minimum) packet length

Ethernet

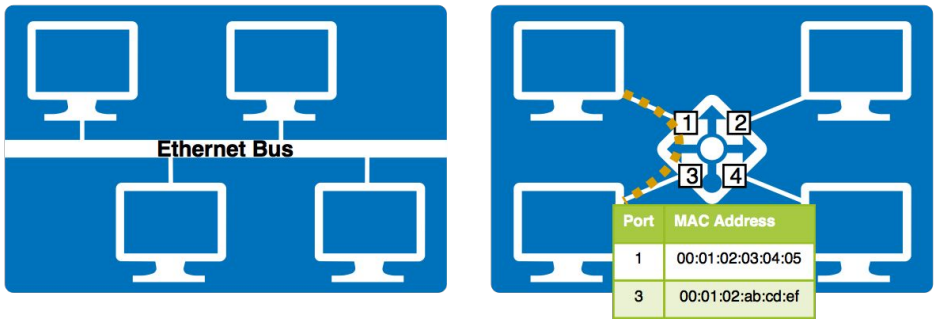
- Physical addressing: 48-bit MAC (Media Access Control) address
  - Unique by Ethernet device
- Currently up to 100 Gb/s (copper up to 10 Gb/s, fiber up to 100 Gb/s)
- Frame format:



Ethernet Switch

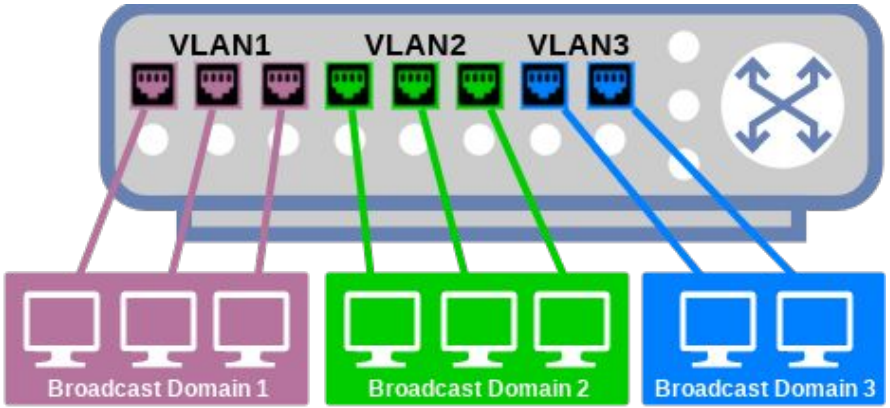


- Layer 2 device
- Transmit each frame to its specific destination only
  - Better unicast **traffic isolation** reduces collision rate, improves network efficiency
- **Learns** the association between physical ports and MAC addresses by observing traffic
- Some traffic is still broadcasted:
  - Broadcast frames
  - Unknown destination address



Ethernet Virtual LAN

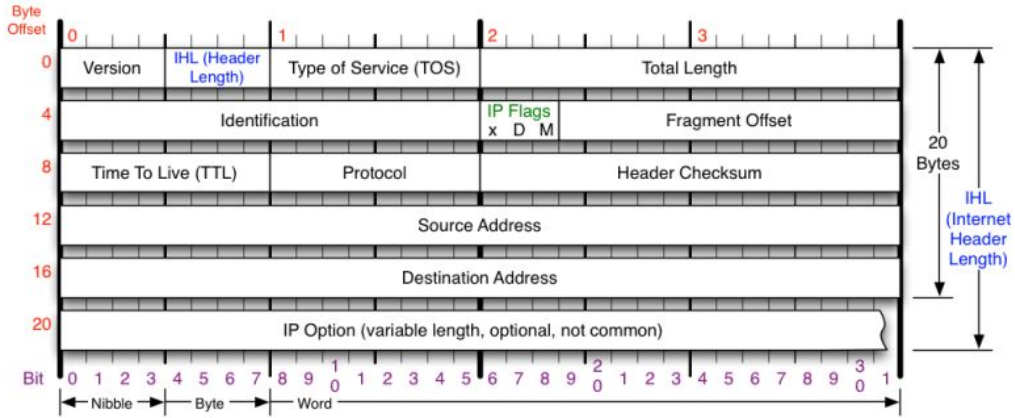
- Logical subdivision of Ethernet switches into several LANs
  - Reduce broadcast domain size
  - Improved efficiency
  - Improved security
  - Logical grouping of hosts



IP and Routing

Internet Protocol

- RFC 791
- Works with ICMP, Internet Message Control Protocol (RFC 792)
  - Error and control messages exchanged between routers
- Designed to connect LANs together
- Defines a hierarchical **addressing** system
- Provides **routing** functionality
  - Finds a path to reach destination from source
- Supports **fragmentation**
  - Packets can be split into smaller chunks
- **Connectionless**
  - No management of end-to-end connection
- Header
  - 20 bytes with no options



IP (v4) Addressing

- 32-bit address represented as four [0-255] numbers (10.20.30.40)
- An address is divided in two parts: network (upper bits) and host (lower bits)
  - The netmask determines where the boundary between network and host address lies
- And 3 ranges of private addresses (not routed by public routers, require network address translation):
  - 10.0.0.0/8
  - 172.16.0.0/12
  - 192.168.0.0/16

```
Address: 10.20.30.40 / Netmask: 255.224.0.0
Also: 10.20.30.40/11
Addr: 00001010.00010100.00011110.00101000
Mask: 11111111.11100000.00000000.00000000
Bitwise AND gives network address:
      00001010.00000000.00000000.00000000
    = 10.0.0.0
Broadcast address: set host bits to 1
      00001010.00011111.11111111.11111111
    = 10.31.255.255
Address range:
From  00001010.00000000.00000000.00000001
    = 10.0.0.1
To    00001010.00011111.11111111.11111110
    = 10.31.255.254
```

IP Routing

- Networks interconnected via routers (or gateways)
- Each router know its near neighbours
- Uses routing tables: association between networks and next IP router
- Routing algorithm
  - If (destination is part of local network) send directly to destination
  - else if (destination network matches an entry in the routing table) forward packet to next gateway
  - else send packet to default gateway (route to 0.0.0.0)
  - (or send ICMP error message if no default gateway)
- IP does not defines how the routing tables are established and maintained
  - Static tables
  - OSPF, BGP

Destination	Network mask	Gateway
127.0.0.0	255.0.0.0	127.0.0.1
127.0.0.1	255.255.255.255	127.0.0.1
192.168.0.0	255.255.255.0	192.168.0.1
192.168.0.1	255.255.255.255	127.0.0.1
192.168.0.255	255.255.255.255	192.168.0.1
192.168.10.0	255.255.255.0	192.168.10.1
192.168.10.1	255.255.255.255	127.0.0.1
192.168.10.255	255.255.255.255	192.168.10.1
224.0.0.0	240.0.0.0	192.168.10.1
224.0.0.0	240.0.0.0	192.168.0.1

IP on LAN

- Once the destination has been determined (via routing table), the physical address associated to the logical IP address must be determined
- ARP: Address Resolution Protocol
  - Broadcast message: who has IP w.x.y.z?
  - Node with w.x.y.z sends a response that contains its physical address
  - Associations are cached

- Assigning IP to node: DHCP, Dynamic Host Configuration Protocol
- A DHCP server provides IP configuration to requesting node:
  - IP address and netmask
  - Default gateway
  - DNS servers

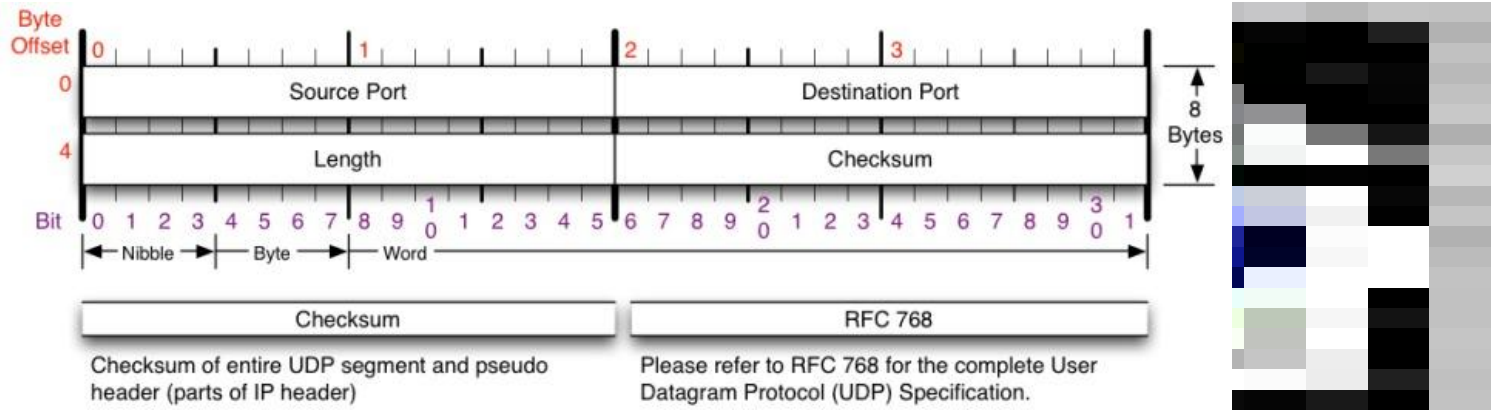
TCP, UDP and Transport

Transport (Layer 4)

- Provides end-to-end transport service to applications
- An application is associated to a port
  - Communication endpoint: IP address + port
  - From the application the connection is used via a "socket"
- Two major protocols: UDP, TCP

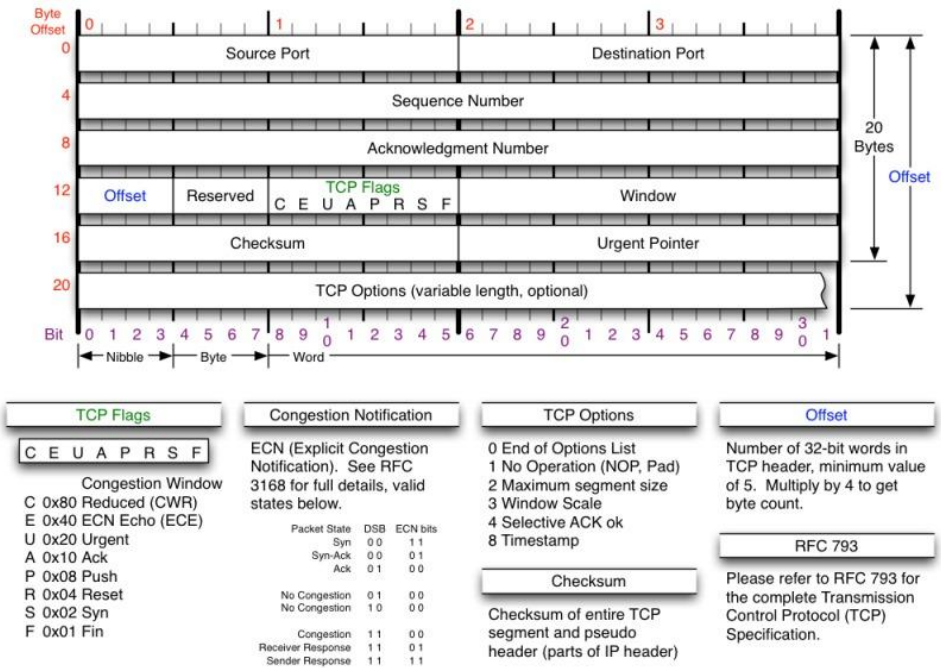
UDP (User Datagram Protocol)

- Simple and light
  - Simply delivers data to a port that an application can bind
- Connectionless: does not maintain or manage a connection between endpoints, each packet is independent
- Unreliable, but guarantees data integrity (checksum)
- Can be used for broadcast and multicast



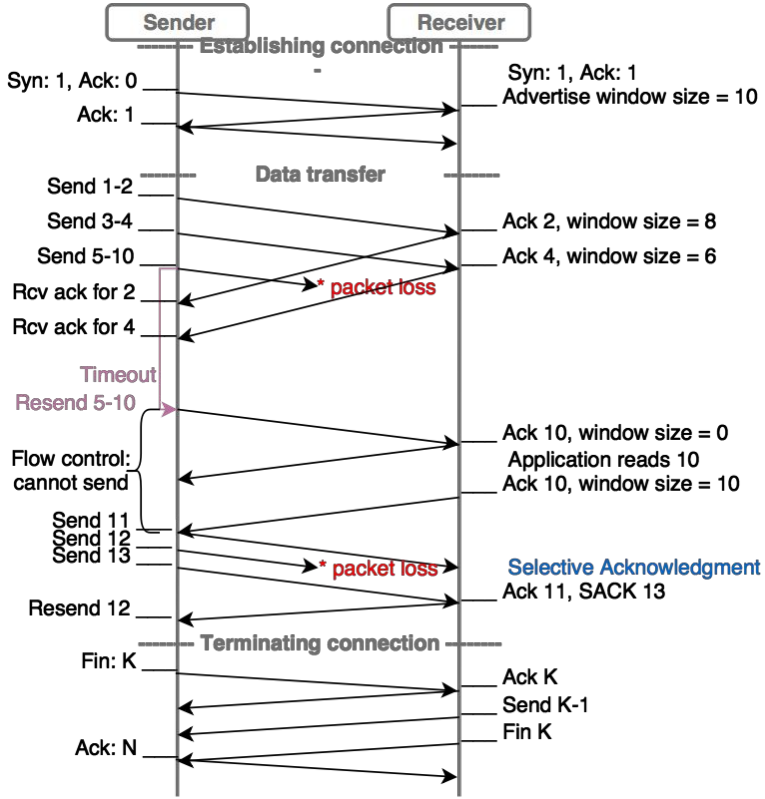
TCP (Transmission Control Protocol)

- Reliable, guarantees:
  - Data will reach the other endpoint
  - In order
  - Data integrity
- Connection-oriented
  - Flow control: adaptive to the receiver
  - Congestion control: adaptive to network occupancy
- TCP can fragment or assemble the application data as needed



Principles of TCP

- Sliding window, timers, and acknowledgements
- Specific handshakes for opening and closing connections
- Receiver advertise how much data it can receive (window size)
- Sender window size also depends on packet loss (congestion control)
- Sender start timers when transmitting a packet
- Receiver must acknowledge received packets
- Sender retransmit packets not acknowledged after timeout
    - Or upon request: selective acknowledgment



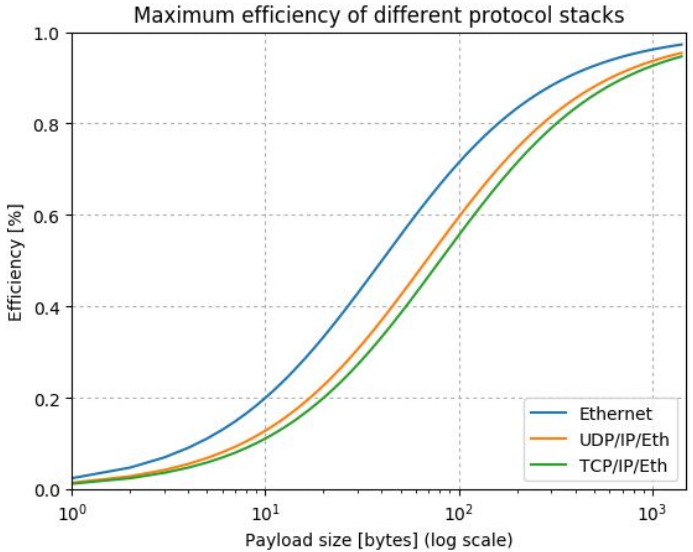
Networks Efficiency

Efficiency

- Static overheads: headers, footers, and maximum packet size set a maximum on the usable part of the available bandwidth:

$$\text{efficiency} = \frac{\text{useful data size}}{\text{useful data size} + \text{overhead}}$$

- Dynamic overheads: packet retransmission, protocol handshakes, packet fragmentation, etc.
  - Estimating the real efficiency of a TCP connection is not straightforward
  - Depends on operation conditions
- Ethernet MTU (Maximum Transmission Unit): 1500 bytes
- UDP, TCP, IP maximum packet size: 65536
- On local networks (and controlled environment) an extension called "jumbo frames" allow to reduce the fraction of the overhead: MTU up to 9000 bytes

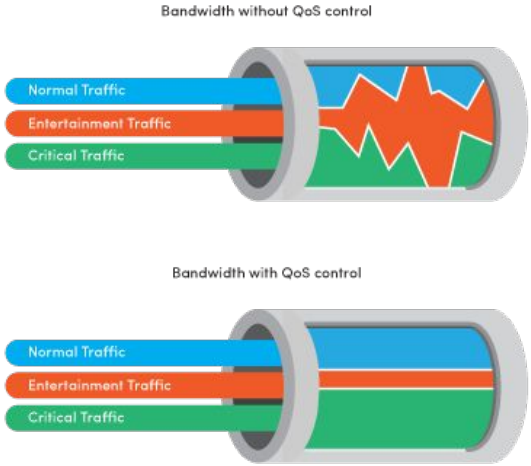


Quality of Service

- Resource allocation control
- Traffic prioritization

} To guarantee maximum latency, minimum bandwidth

- Different implementations:
  - Port based (layer 4)
  - IP address based (layer 3)
  - DiffServ (uses IP header field DSCP)
  - VLANs
  - IPv6
- Allow for sharing of the same physical network for different purposes
  - Already planned in next upgrade of LHC experiments



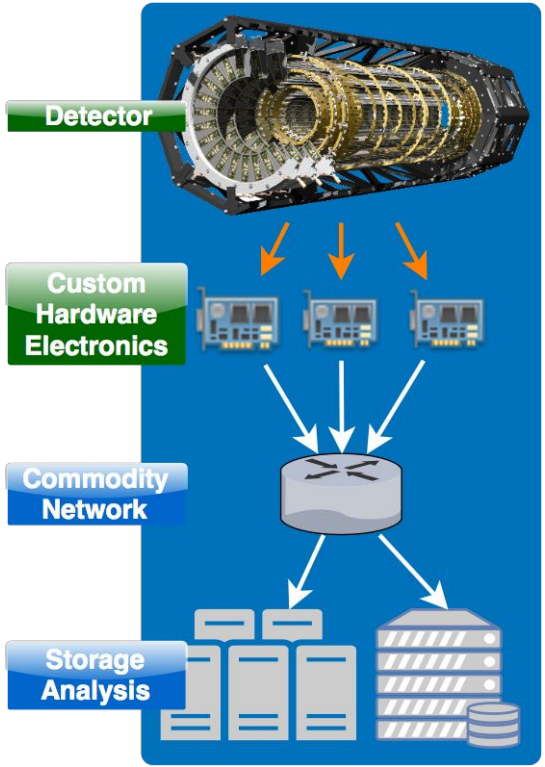
Monitoring

- SNMP (Simple Network Management Protocol) is the standard
- Supported by all professional network devices
- Variable based system:
  - Devices maintain variable up to date
  - Client request SNMP server to access the variables
- Covers a wide variety of metrics (traffic, errors, temperatures, etc.)
- Requires a logical and a presentation layer
- See lab for more details

Networking for DAQ Systems

Usage

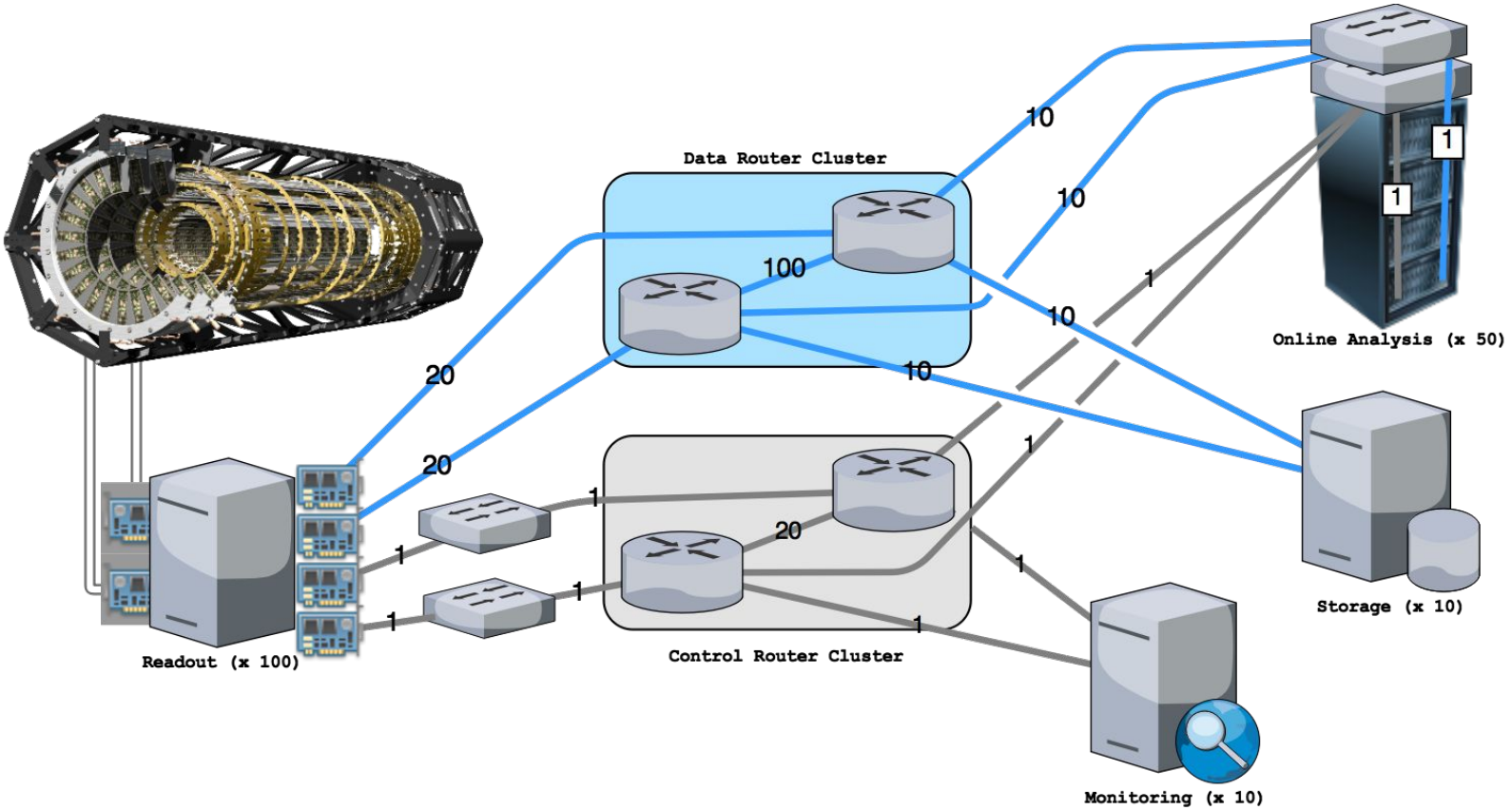
- Convey data from experiment to analysis and storage
- Other traffics: experiment control, monitoring
  - Different requirements
- Detectors usually use custom links/data format
  - A step of digitization, data formatting, data processing is required to use commodity network technologies
- Valuable data
- Use of commodity technologies wherever possible
  - Commodity networks make cost, maintenance and application development more efficient (compared to custom technologies)
  - Development of these technologies is driven by the global market
  - But DAQ systems of physics experiments is not a significant player in the market
  - So we need to adapt, follow technology
  - Growth of the high-performance computing drives the market in a favorable direction
    - Important DAQ systems share some characteristics with HPC



Characteristics

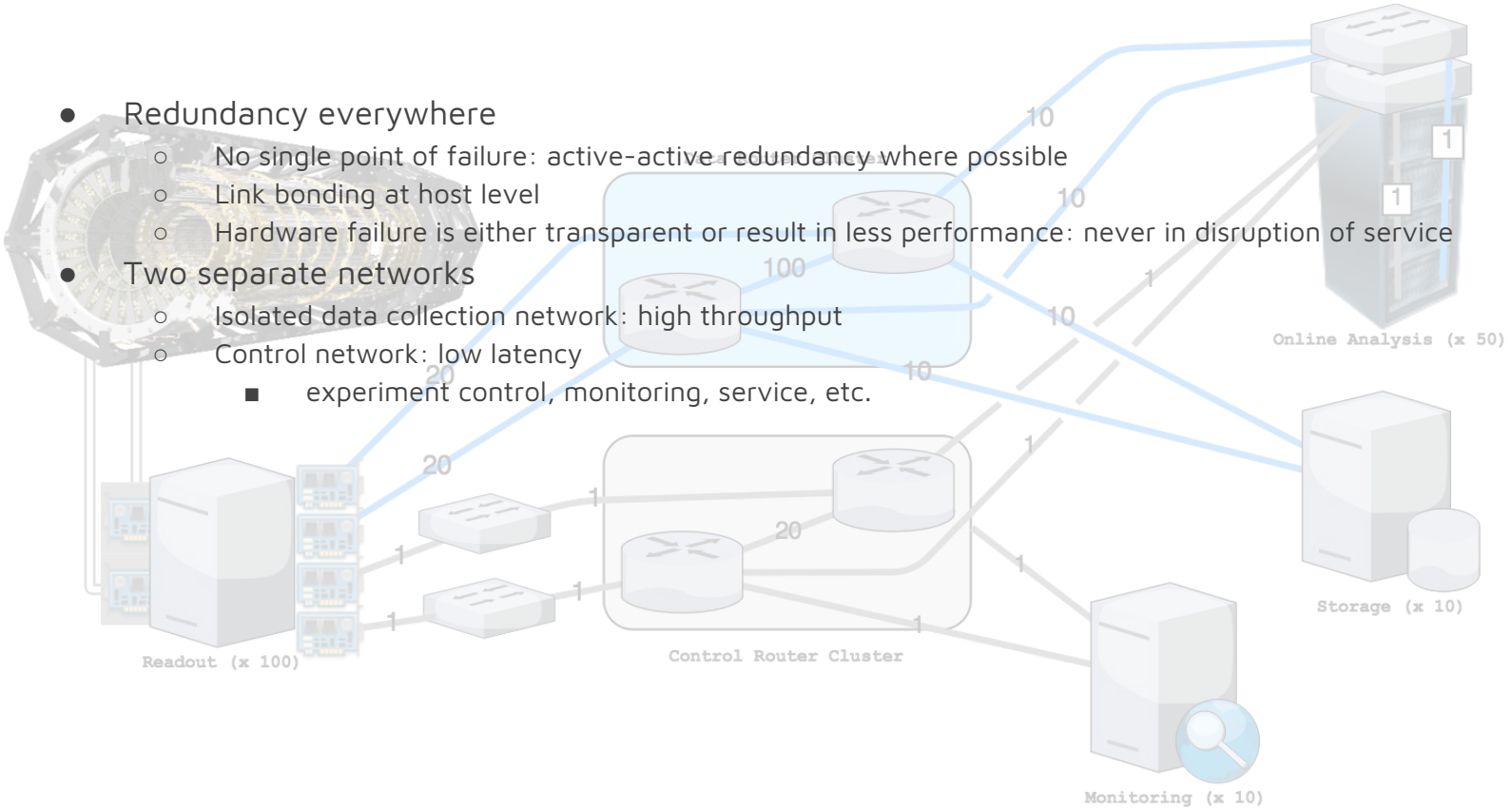
- LAN(s)
- Controlled environment
  - Private local networks
- Reliability, high availability
  - Advanced monitoring system
  - Redundancy (adapters, host, switches, routers)
  - Automatic failover
- Performance
  - High throughput
  - Low latency
  - Do not always go well together
- Security
  - Trusted environment but data integrity is critical

A Real Case (Almost)



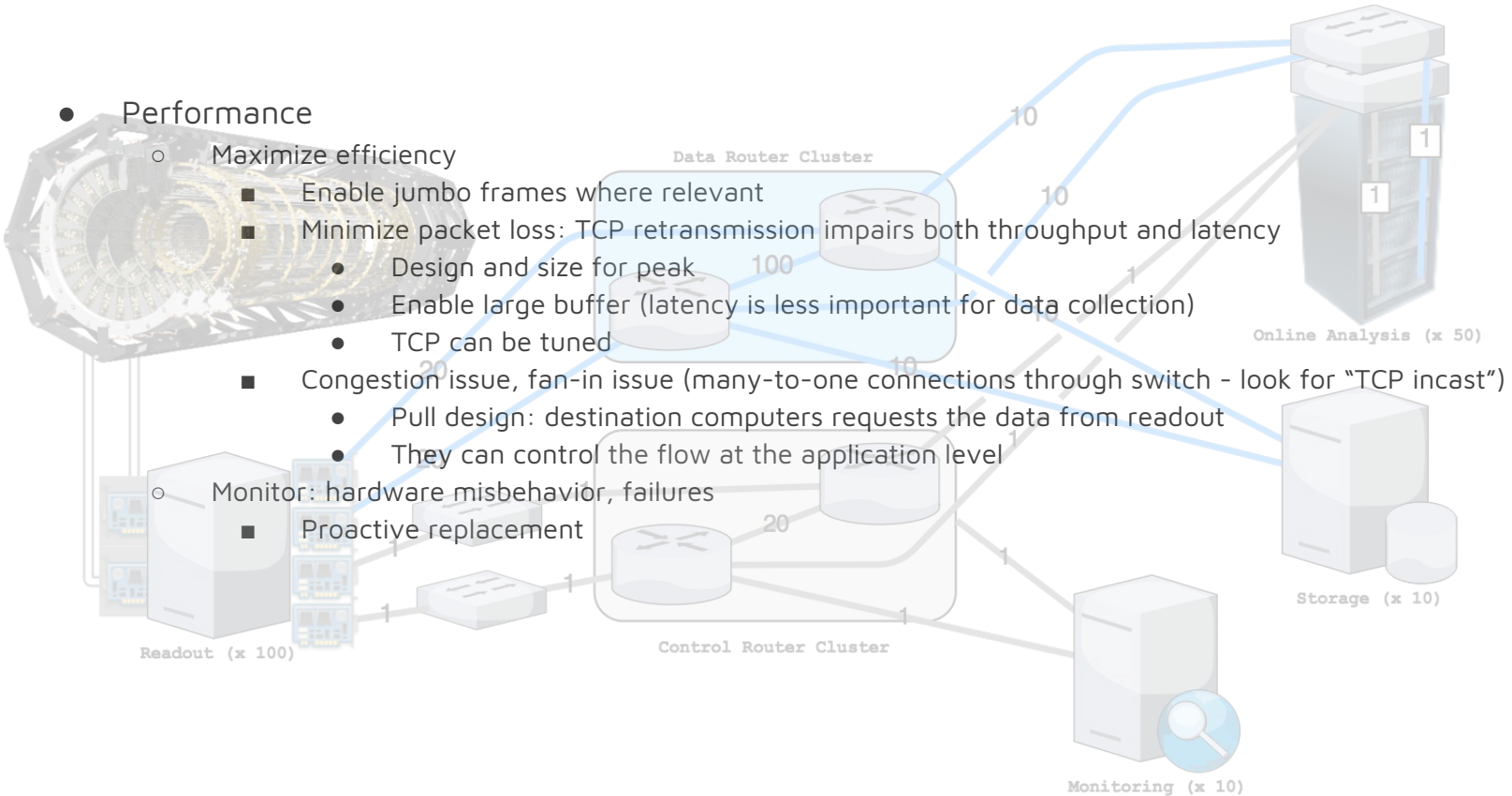
A Real Case (Almost)

- Redundancy everywhere
  - No single point of failure: active-active redundancy where possible
  - Link bonding at host level
  - Hardware failure is either transparent or result in less performance: never in disruption of service
- Two separate networks
  - Isolated data collection network: high throughput
  - Control network: low latency
    - experiment control, monitoring, service, etc.



A Real Case (Almost)

- Performance
  - Maximize efficiency
    - Enable jumbo frames where relevant
    - Minimize packet loss: TCP retransmission impairs both throughput and latency
      - Design and size for peak
      - Enable large buffer (latency is less important for data collection)
      - TCP can be tuned
    - Congestion issue, fan-in issue (many-to-one connections through switch - look for “TCP incast”)
      - Pull design: destination computers requests the data from readout
      - They can control the flow at the application level
  - Monitor: hardware misbehavior, failures
    - Proactive replacement



Going beyond TCP/IP

- HPC technologies: very high throughput and very low latency within data centers
- Standard: Infiniband, implemented mostly by Mellanox and Intel
- Replacement technologies for layers 1 to 4 at least
  - TCP is not suitable for intra-data center communications (timeout too long)
  - IP is often not needed
  - Ethernet has too much overhead
- RDMA, remote direct memory access:
  - Network packets are written directly into host memory
  - Minimal latency, no OS overhead

Conclusion

- The OSI Model to understand
  - Ethernet interconnect hosts in LAN
  - IP interconnect LANs
  - TCP reliably connects applications
- Efficiency
  - Know your protocols to use them at best
  - QoS allows for efficient sharing of common resources
  - Monitoring
- DAQ
  - Performance is the key but useless without reliability, high availability
  - Common stacks cover most of the DAQ systems
  - Optimization for the specific requirements
  - HPC technologies are growing and meet requirements of demanding experiments

Thank you

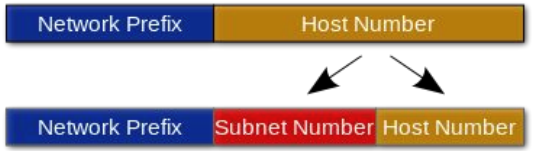
Questions ?

flegoff@cern.ch

See you in Lab 9

IP Subnetting

- Subnetting: divide a given network address range into multiple subnets by adding bits to the netmask
  - Better traffic control
  - Logical grouping of nodes and services




```
Address Range: 1.2.3.0 / 24
4 internal networks: add 2 bits to netmask
4 networks (with address range):
 - 1.2.3.0/26   : 1.2.3.1   -> 1.2.3.62
 - 1.2.3.64/26  : 1.2.3.65  -> 1.2.3.126
 - 1.2.3.128/26 : 1.2.3.129 -> 1.2.3.190
 - 1.2.3.192/26 : 1.2.3.193 -> 1.2.3.254
```

IPv6

- IP version 6
- New addressing system: 128-bit addresses
- Main advantages:
  - More addresses
  - Simpler header
  - Less configuration (exit DHCP)
  - Less fragmentation (additional protocol to determine path's MTU at the source)
  - Better routing (smaller routing table)
  - Better multicast support
  - Better QoS support
- Main drawback:
  - Transition from IPv4