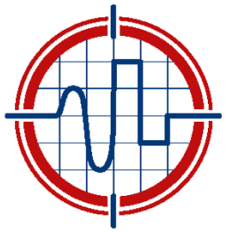
Networking for Data Acquisition Systems

Vesa Simola - 2020 -
vesa.simola@cern.ch



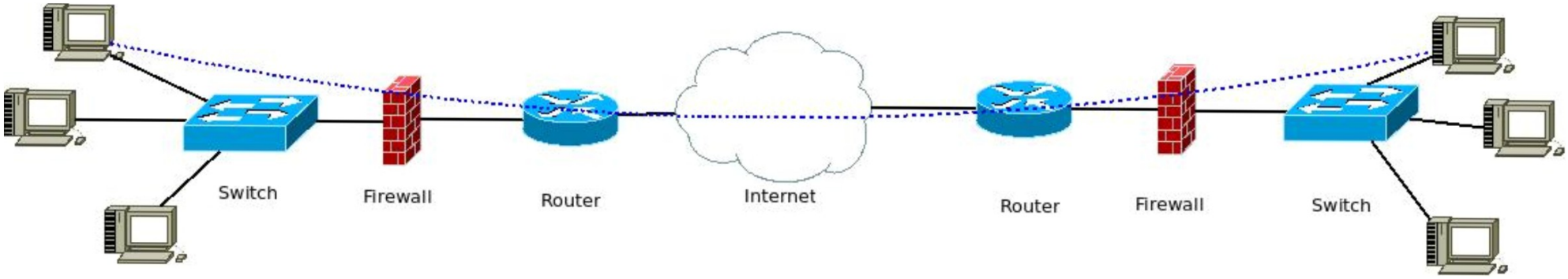
ISOTDAQ
International School of Trigger and Data Acquisition

Agenda

- General networking concepts and terminology
- OSI model
- Ethernet
- IP and routing
- Protocols: TCP & UDP
- Data acquisition and networking and bit of RDMA

Few words on terminology

- Computer network consists of end devices, transmission mediums, transit equipment, protocols and applications.
- End devices can run applications that act as a source and generate the data and send it to the destination end devices for consumption.
- Transit devices forward data units between sources and destinations over various transmission mediums.
- Protocols are software constructs that enable the transmission of data units packaged in predefined formats such as frames and packets.



Different types of networks

- Networks can be classified by various characteristics, for example:
  - Physical and logical topology
  - Purpose of the network: enterprise networks support some internal activity while carrier networks print money.
  - Size:
    - LAN (Local Area Network)
    - MAN (Metro Area Network)
    - WAN (Wide Area Network)
    - Internet, the network of networks

OSI

- OSI stands for Open Systems Interconnection, standardized by International organization for Standardization.
- OSI model is split into seven layers that provide the basic concepts for connecting between end devices.

OSI Layer	Example functions
Application	Application protocols such as NTP, SSH etc.
Presentation	Compression, encryption
Session	Authentication, checkpointing
Transport	TCP, UDP
Network	IP addresses, networks
Data-link	MAC and LLC sub-layers
Physical	Cables connectors and signals

Utilizing OSI model for communication

- Application is on the top of OSI stack as it produces the data.
- Data is passed down via all the layers in the OSI stack and each layer adds its headers until data finally hits the physical layer where signal is transported to the next device in the path.
- On the receiving end the same process takes places, except that it happens in reverse. Headers are stripped on each layer and data is passed to the receiving application.

TCP/IP model

- IP protocol maintained by Internet Engineering Task Force (IETF)
  - *Rough consensus and running code.*
- Only four layers:
  - Network access layer – similar to physical and data link layer of the OSI model, e.g cables and Media Access Control (MAC) addresses reside here.
  - Internet layer – similarly to OSI, this is where IP addresses appear.
  - Transport layer – Similar to the transport layer of the OSI model. Concepts such as protocols (TCP/UDP etc.) belong to transport layer
  - Application layer – houses OSI session, presentation and application layer. Contains the application payload such as DNS, possible encryption and other formatting happens here.

OSI compared to TCP/IP

OSI Model	TCP/IP Model	Contents
Application	Application	DNS,DHCP and other applications
Presentation		Compression, encryption and various conversions
Session		Session establishment/teardown
Transport	Transport	TCP,UDP,ICMP,SCTP etc.
Network	Internet	IP
Data-link	Network access	MAC, Ethernet
Physical		Cables, NICs, optics etc.

Ethernet

- Contains technologies from the first two layers of the OSI model: Physical and data-link layer.
- Uses supposedly unique 6 byte MAC (media access control) addresses to identify, locate and communicate over the network.
- Builds broadcast domains where every end devices can talk to every other end device.
- Data is encapsulated inside frames.

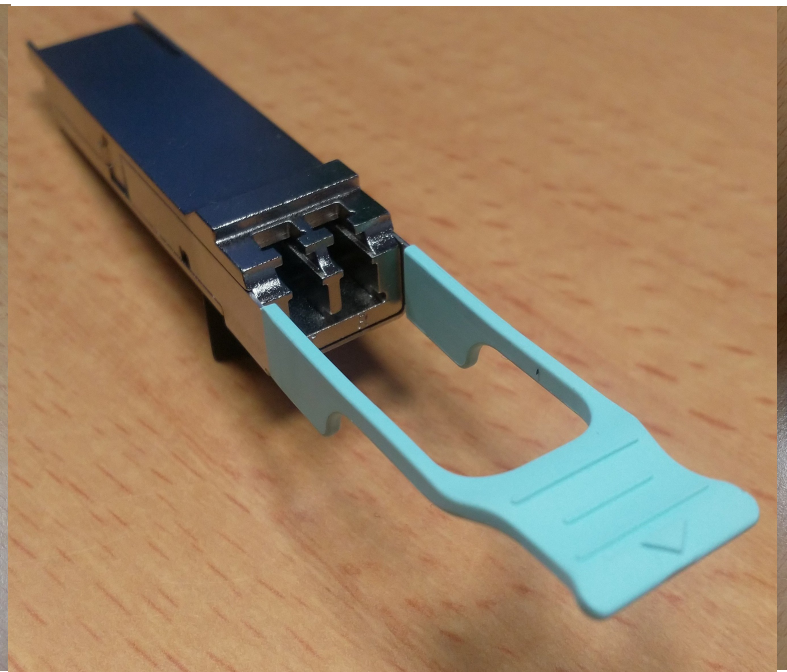
Ethernet frame structure

Preamble (7 B)	SFD (1 B)	Destination MAC (6 B)	Source MAC (6 B)	802.1Q (4 B)	Ethertype/ Length (2 B)	Payload (46-1500 B)	CRC/FCS (4 B)

- Preamble is a specific sequence that allows hardware to detect the new frame.
- Start frame delimiter tells that the destination address begins from the next byte.
- Destination and source MAC addresses follow the preamble.
- Non mandatory 802.1Q VLAN tag is used to identify or set the correct broadcast domain / VLAN for the frame: number between 0 and 4095.
- Ethertype signals the payload type that is encapsulated in the frame: 0x0800 for IP and 0x8100 for VLAN tagged frame.
- Payload is the actual application data. Larger than 1500 Byte payloads are also possible, but 1500 is the default.
- CRC checksums / frame check sequences allow the receiver to check that the frame arrived intact.

Ethernet implementations

- Ethernet comes in different shapes and speeds. Original experimental version had bandwidth of 2.94Mbps, whereas today we have 400Gbps available. More common variants being 1Gbps and 10Gbps.
- Commonly Ethernet can use either electric medium (copper cable) or optical (single mode or multimode). Optical mediums tend to support higher speeds over longer distances.


Ethernet switch

- Ethernet switch is a OSI layer two device that splits collision domains to allow full duplex operation.
- Switch monitors incoming traffic to learn MAC addresses to build MAC address table. This table is stored to a CAM (content addressable memory) and actual switching is (usually) done by an ASIC.
- MAC address table is used to transport traffic out via the correct port.
- Traffic towards unknown destinations is sent out of all but the original ingress port similarly to traffic with FF:FF:FF:FF:FF:FF destination address (Ethernet broadcast address). This behavior is essential concept in few of the more interesting malfunctions that can happen on an Ethernet network.

Virtual LAN

- Virtual LANs (VLAN) can be used to logically divide the switch into multiple broadcast domains.
- VLANs can share a medium and span multiple switches using VLAN tagging in a form of a 802.1Q header in the frame.

Preamble (7 B)	SFD (1 B)	Destination MAC (6 B)	Source MAC (6 B)	802.1Q (4 B)	Ethertype/ Length (2 B)	Payload (46-1500 B)	CRC/FCS (4 B)

- VLANs have several benefits:
  - Reduced size of broadcast domains leads to less “line noise”.
  - Improved security as it is not trivial to directly communicate with a device in another VLAN.
  - It is possible to logically group end devices based on any characteristic.

Ethernet switch port config

- Below is an example of switch port configuration with few VLANs and MC-LAG:

```
native-vlan-id 307;
aggregated-ether-options {
    lacp {
        active;
        system-id a0:02:oh:no:00:xx;
        admin-key XYZ;
    }
    mc-ae {
        mc-ae-id XYZ;
    }
}
unit 0 {
    family ethernet-switching {
        interface-mode trunk;
        vlan {
            members [ DATA_VLAN_IP307 MON_VLAN_IP357 ];
        }
    }
}
```

Internet protocol - IP

- Internet protocol (IP) comes in two flavors:
  - IPv4 with 32-bit addresses consisting of four octets in a shape of numbers between 0 and 255: 193.166.3.17
  - IPv6 comes with 128-bit addresses in a form of eight groups of 16 bits each. Increased address space allows more unique addresses:

    1:5ee:bad:c0de:: (1:5ee:bad:c0de:0:0:0:0)
  - IPv4 and IPv6 are **not** compatible.
- In both cases, address is split into two parts:
  - Network part
  - Host part
- IPv4 network has two special addresses: network & broadcast. Broadcast address is the last address in the network. IPv6 relies on multicast instead to improve efficiency and cut down on line noise. Both IP versions have the concept of network address, it is the first “host address”, network and host addresses should not be configured on end devices except in special cases like /31 point-to-point networks.

IP packet structure

Version	Header size	ToS	Total length
Identifier	Flags	Fragment offset	
TTL	Protocol	Checksum	
Source IP			
Destination IP			
IP options			
Payload			

- Version, IPv4 or IPv6
- Header size
- ToS for ECN, DSCP etc. Qos
- Total length
- Identifier, fragmentation.
- Flags, DF
- Fragment offset, reassembly.
- TTL(Time to live)
- Protocol, TCP, UDP.
- Checksum

- Further headers (TCP, UDP etc.) are encapsulated via the payload.
- Think of this recursive encapsulation like sending letters:
- Brain (application) produces the actual payload.
- Person writes it down (presentation layer) in a suitable medium (transport).
- One adds a stamp and address (network layer) on an envelope (data-link).
- Letters moves via mail (physical layer).

Subnets

- IP address configuration consists of two essential elements: IP address and subnet mask.
- Subnet mask is used to determine the distribution of network and host address space. By moving bits between network and host portion we can manipulate the network size.
- Using 192.168.1.10/25 as an example:
- The first 25 bits represent the network, leaving the last 7 bits for host portion: 32-25=7
- 2^7=128 (0-127), 128-2 (network address 0 and broadcast 127) = 126 usable host addresses.
- Binary 11111111.11111111.11111111.10000000 turns into 255.255.255.128 in decimal giving the more human friendly subnet mask configuration parameter.

IP Routers

- DAQ networks are interconnected by routers.
- Routers utilize routing tables to build forwarding tables that in-turn are used by forwarding plane ASICs or network processors to forward transit traffic. So, forwarding table is (usually) a subset of the routing table.
- Buffering and storage can be on-chip CAM (fast, but expensive), external DRAM, HBM and the likes (bus speed might become an issue) or mixtures of the two.
- There are two main methods of populating routing tables:
  - Static routing maintained by hand
  - Routing protocols that exchange routing information between routers automatically.
  - Note that routing information is generally programmed using general purpose CPUs on the control plane.

IP routing steps

- Simplified steps taken by the network processor on a transit router:
  - Remove the link-layer header
  - Look-up for the destination address from the IP header
  - Possible policing/filtering based on IP and transport layer headers:
    - Protocol, port, flags, TTL, addresses etc.
    - Near line-rate on modern hardware, capabilities depend on hardware
  - Look-up for the destination address from the routing table
  - Look-up of destination link-layer address
  - Add correct link-layer header
  - Place the packet in the transit queue for sending
  - Decrement TTL by one and send – or drop, if congested – the packet to its way.

Router interface config

- Below is an example of router interface config
- IP-address, VLAN and VRRP (virtual router redundancy protocol).

```
apply-groups VRRP-MASTER;
vlan-id 634;
family inet {
    address 10.147.162.3/24 {
        vrrp-group 0 {
            virtual-address 10.147.162.1;
            accept-data;
        }
    }
}
```

Routing table part 1

- Example of routing table entry:

```
10.147.161.0/24    *[Direct/0] 1d 18:00:09
                    >  via ae33.633
                    [OSPF/150] 1d 17:59:40, metric 0, tag 0
                    >  to 192.168.49.70 via irb.11
10.147.161.1/32    *[Local/0] 1d 17:59:55
                       Local via ae33.633
10.147.161.3/32    *[Local/0] 1d 18:00:09
                       Local via ae33.633
10.147.162.0/24    *[Direct/0] 1d 18:00:09
                    >  via ae34.634
                    [OSPF/150] 1d 17:59:40, metric 0, tag 0
                    >  to 192.168.49.70 via irb.11
10.147.162.1/32    *[Local/0] 1d 18:00:05
                       Local via ae34.634
10.147.162.3/32    *[Local/0] 1d 18:00:09
                       Local via ae34.634
10.147.163.0/24    *[Direct/0] 1d 18:00:55
                    >  via irb.635
                    [OSPF/150] 1d 17:59:40, metric 0, tag 0
                    >  to 192.168.49.70 via irb.11
10.147.163.1/32    *[Local/0] 1d 18:00:46
```

Routing table part 2

- Example of more detailed routing table entry:

```
inet.0: 822 destinations, 902 routes (822 active, 0 holddown, 0 hidden)
10.147.68.0/24 (1 entry, 1 announced)
        *OSPF   Preference: 150
                Next hop type: Router, Next hop index: 0
                Address: 0xcc4e710
                Next-hop reference count: 9
                Next hop: 192.168.49.74 via irb.12
                Session Id: 0x0
                Next hop: 192.168.49.78 via irb.13, selected
                Session Id: 0x0
                State: <Active Int Ext>
                Age: 1d 18:03:39        Metric: 0
                Validation State: unverified
                        Tag: 0
                Task: OSPF
                Announcement bits (2): 0-KRT 4-Resolve tree 1
                AS path: I
```

IP routing protocols

Routing protocols fall into two main categories:

- Interior gateway protocols (IGP) are used to exchange routing information within single autonomous system. Roughly two variants: link-state (OSPF,IS-IS) and distance-vector (RIP).
  - Link-state protocols announce the status of routes to every other router in the autonomous system whereas distance-vector protocols announces only to neighbors. This leads to faster convergence when link-state routing protocols are used.
  - IGPs tend to auto-discover their neighbors.
- Exterior gateway protocols (EGP) are used to exchange routing information between autonomous systems. Autonomous system indicates a set of routers under the same administrative party. Internet runs on BGPv4 and BGP could be considered as a path-vector -protocol in comparison to distance-vector or link-state.
  - BGP requires explicit configuration of neighbors (or groups) and supports wide set of metrics and policing if needed.
  - BGP has support for several address families such as IPv4, IPv6, MPLS (L2VPN,VPNv(4|6)), EVPN etc.

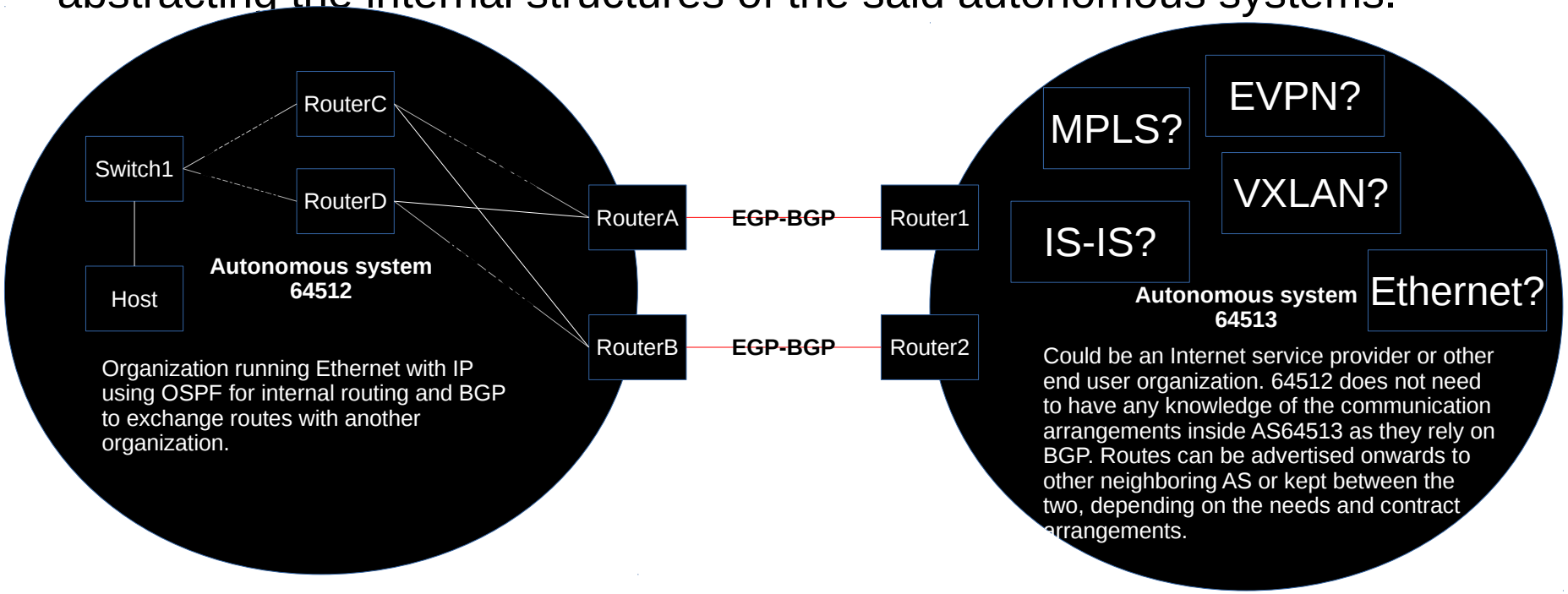
Example: OSPF adjacency

- IGP (OSPF) forms adjacency to neighbor router utilizing multicast and starts to flood LSA messages (Link State Advertisements). These LSA messages are used to construct independent view of the network on each participating router.
- Flooding of LSA is controlled by using concept of areas. Areas in OSPF come in few flavors:
  - Backbone, normal, stub, not-so-stubby-area area.
  - Each area has to directly connect with backbone area – with the exception of OSPF virtual link.

Address	Interface	State	ID	Pri	Dead
192.168.29.10	irb.11	Full	192.168.29.5	128	38
192.168.59.34	irb.12	Full	192.168.59.17	128	31
192.168.89.98	irb.13	Full	192.168.89.21	128	38
192.168.39.42	irb.14	Full	192.168.39.9	128	34
192.168.39.66	irb.15	Full	192.168.39.13	128	35

Interconnecting networks

- Ethernet provides LAN connectivity, IGPs provide internal routing and BGP takes care of advertising routes between autonomous systems while abstracting the internal structures of the said autonomous systems.



TCP, UDP and Sockets

- TCP and UDP are constructs of OSI transport layer.
- Protocols that provide end-to-end transmission of data, connections can rely on sockets for addressing, for example: 8.8.8.8:53 UDP-socket points to known Google public DNS resolver at IPv4 address 8.8.8.8, listening on port 53.
- Applications open sockets to send data to a known destination socket where another application is hopefully listening.
- Sockets come in flavors, for example: streaming socket (TCP), Datagram socket (UDP), Raw socket (ICMP).
- Think of sockets as the interface between application and the network, in terms of OSI layers socket could reside between transport and session layers.

UDP – User Datagram Protocol

- UDP has the following characteristics:
  - Unreliable but guarantees data integrity.
    - Means that it has no mechanism against packet loss, application has to take care of this.
    - UDP packets can arrive in different order than intended and application has to be prepared for this.
  - UDP is connectionless
    - Each packet is independent.
    - No concept of connection setup or tear-down.
    - Supports unicast, multicast, anycast and broadcast.
  - Example applications: (most) DNS queries, SNMP and VXLAN.

TCP – Transmission Control Protocol

- TCP has the following characteristics:
  - Reliable
    - Data will reach the destination, or TCP will inform of this.
    - Data is delivered in order and integrity is guaranteed.
    - TCP utilizes sequence numbers to keep track of the data stream.
  - TCP is connection oriented
    - Connection is established using three way hand-shake before data is being transferred.
    - Implements sessions supporting both flow control and congestion management.
    - Still, some TCP-based services run in Internet over anycast.
  - Example applications: HTTP, FTP and SSH.

Using IP in LAN

- Application decides that it wants to send data to some IP destination, for this end node needs to figure out where to send the data:
  - Lookup on the local routing table to see if the destination address is local or if data has to be sent to a router for forwarding further.
- IP-addressing is a logical construct and relies on Ethernet layer for transport. In IPv4 broadcast based Address Resolution Protocol (ARP) is used to map IP-addresses to MAC-addresses. IPv6 relies on multicast based Neighbor Discovery Protocol (NDP) for similar purpose.
- Source and destination MAC is changed at every network layer transport hop, while IP information stays the same.

Applying networking concepts to data acquisition

Characteristics of DAQ network

- Several LANs with routing in between utilizing high speed transport.
- Experiment data is unique and valuable so special care is widely taken to avoid dropping data while in transit. Or at the very least, drops should be detected and data sent again.
- High throughput, low latency and reliability do not always go hand-in-hand. This leads to more complex configuration as quality of service (QoS) is required to manage different traffic profiles.
- Security aspects are heavily involved from data integrity stand point. QoS plays a role here as well since less critical data is marked for dropping first if the queues get full.
- Potentially challenging traffic patters introduced: many to few -traffic flows can create buffering problems and in-cast scenarios where receiving end device or transit devices start to have buffering issues.

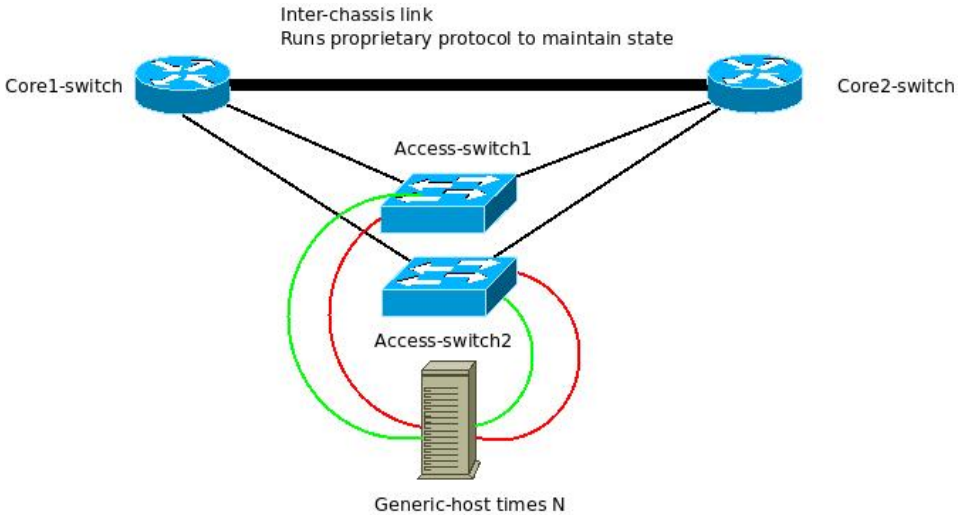
Data acquisition and networking

- Network is used to transport data from the detector read-out to online analysis and finally storage: multiple traffic patterns.
- Using commodity technologies as much as possible because of cost-efficiency and rapid development of Ethernet and Infiniband.

Data rate (Ethernet)	Year	Data rate (Ethernet)	Year
10 Mb/s	1990	40 Gb/s	2015
1 Gb/s	1999	100 Gb/s	2016 (twolane)
10 Gb/s	2003	400 Gb/s	2018

Topology

- Routed interconnects between networks (racks) are a time-tested and Internet proof approach. Applications rely on IP.
- Applications that rely on Layer 2 are problematic and source of complexity. These should be minimized after initial read-outs.
- All racks represent themselves as a set of separate vlans and subnets. Ideally everything is connected in redundant fashion.



Server rack
- Each rack has dedicated data and control VLANs.
- Each VLAN contains rack specific subnet.
- Routing currently happens in the core layer
- Redundant routing is provided by VRRP but via the magic of proprietary protocol, it is actually active-active.
- Host talks to access switches using LACP, this allows for redundancy and (some) extra bandwidth.

- Routing protocols used: Link-state routing (OSPF, in some cases IS-IS) and BGP.
- ECN, Spanning-tree, 802.1Qbb flow control.
- Proprietary extensions to LACP to allow active-active in L2. These hopefully go away sooner rather than later.
- Linux bonding and teaming drivers are heavily utilized, some configurations (PXE boot) require special handling.

RoCE, Infiniband and RDMA

- Remote direct memory access (RDMA):
  - Improves performance by allowing direct access to the memory buffers of a remote system and offloads parts of the communication process from CPU to Host channel adapter (HCA).
  - RDMA is commonly found in high performance computing setups and comes in few flavors, for example:
    - Infiniband comes in several speeds and provides possibly lowest latency.
    - RoCE relies on Ethernet and in case of RoCEv2 IP and UDP. Basically requires a “lossless” Ethernet setup implemented using Ethernet extensions (DCBX, ECN etc.)
    - Several proprietary variants: Omnipath from Intel, Gemeni & Aries etc.
  - Even basic high availability in forms of link failover etc. is currently dependent on the implementation of the application.

Word on RDMA connection modes

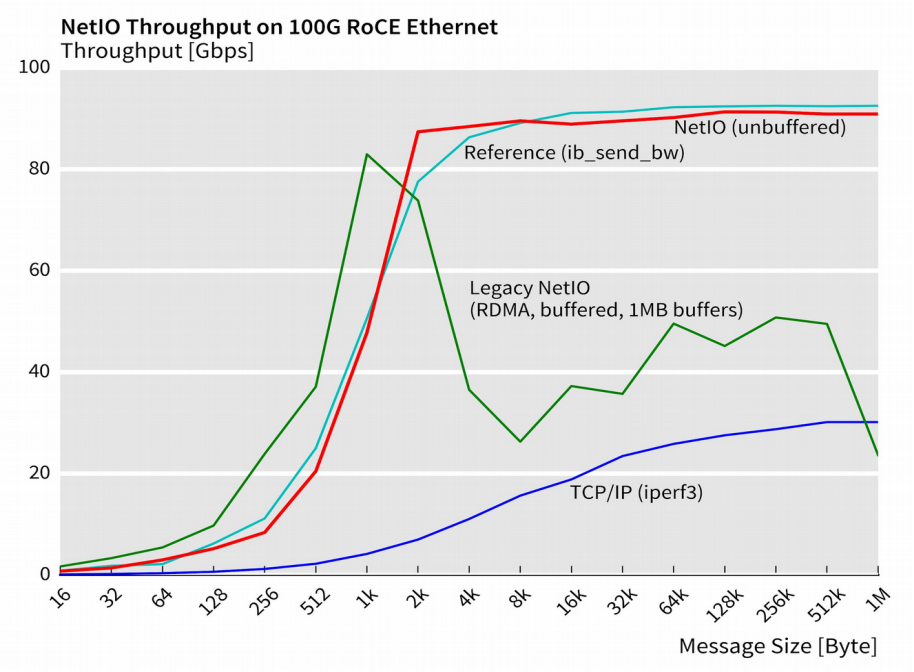
- RDMA connections come in flavors and selecting the correct application approach is essential. Choices have impacts on performance and availability of features and at layers those features need to be implemented.
- Reliable connected (RC)
  - Guarantees that data is delivered, in-order and with integrity.
  - Maybe think of this as somewhat similar to TCP.
  - One to one between queue pairs.
- Unreliable connected (UC)
  - One to one between queue pairs.
  - No guarantees.
- Unreliable datagram (UD)
  - Provides multicast support but no reliability.
  - Maybe think of this as somewhat similar to UDP.

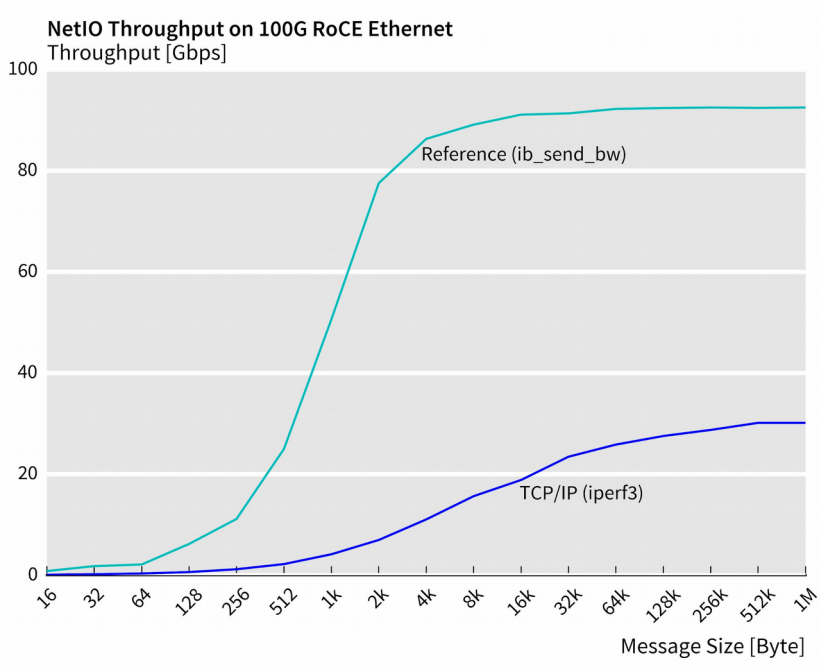
RDMA communications

- RDMA communication between hosts happens over queue pairs (QP), one for sending, one for receiving. Conceptually bit similar to sockets.
- Instead of packets or frames, hosts utilize *verbs* with the queues.
  - Two groups of verbs: Control and Data verbs.
  - Control verbs are used to: create, destroy, modify, query and they usually require context switching.
  - Data verbs don’t require context switching and are used for sending, receiving etc.
- Memory is required during the execution of a work request. Memory addresses for this are associated with the WR. The association comes in a form of a pointer to a region of memory where the host channel adapter (HCA, RDMA HW) has sufficient access to. Access to the memory region is controlled using L_Key and R_Key, L_Key for local memory access and R_Key for remote. R_Key may be shared among the communicating parties, when memory is “handed over” to the remote node.
- In order to send or receive data, work requests (WR) are placed into QP. Once the work request is completed a work completion (WC) is places into the completion queue (QC) that is relevant to the work queue in question.

RDMA performance

- Application performance does not come for free just by using RDMA capable network and IP over IB is not a solution to this.
- The below NetIO benchmarks illustrate the importance of application optimization when using RDMA connectivity.





Benchmarks by Jörn Schumacher, CERN

Where we might see RDMA?

- It is likely that RoCE version 2 will be widely supported in the new ATLAS DAQ network. Why RoCEv2? RoCEv2 works over routed links and it should work in combination of ECMP-routing. RoCEv2 is also widely supported by NIC vendors.
- This means that the new DAQ network needs to support ‘lossless’ connectivity via means such ECN and pause frames.
- Applications relying to messaging (MPI) and storage (NVME-over-fabrics,iSER) are immediate winners.

Conclusions

- DAQ network relies largely on industry standard technologies found in HPC, ISP and enterprise sectors.
- But DAQ network is a challenging combination high capacity, low latency and robustness.
- RDMA is likely playing increasingly important role in DAQ applications, binding connectivity and applications closer together than before.

Further topics

- DAQ network topology?
  - To encapsulate layer two to layer three or not, and why?
  - How to scale applications from networking perspective? Advertise service addresses (anycasted), how and why? FRR, BIRD, EXABGP.
  - DSCP / multified QoS starting from the application?
- How Internet works?
  - DNS system, robust large scale database.
    - Root, tld, first level and replication with caching.
    - Could this robust structure be used in DAQ for distributing some configuration information?