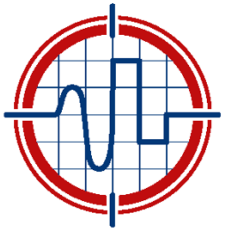
Networking for Data Acquisition Systems
Vesa Simola - 2022 -

vesa.simola@cern.ch



ISOTDAQ

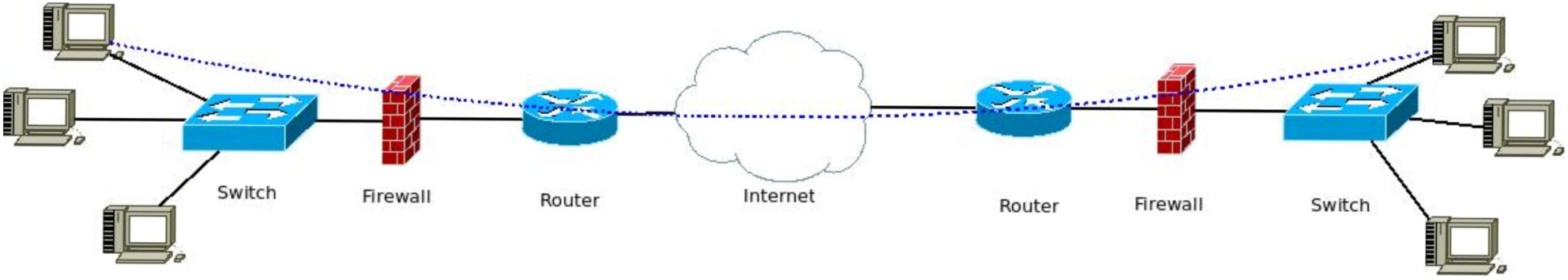
International School of Trigger and Data Acquisition

Agenda

- General networking concepts and terminology
- OSI model
- Ethernet
- IP and routing
- Protocols: TCP & UDP
- Data acquisition and networking and a 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 data and send it to other end devices for consumption.
- Transit devices forward data units between sources and destinations over various transmission mediums.
- Protocols are software constructs that enable the exchange of information to transmit the payload packaged into predefined formats such as frames and packets.



Different types of networks

- Networks can be classified by various characteristics, for example:
  - Physical and logical topology
    - Ring, star, mesh, clos etc.
  - Requirements for features:
    - Reliability, security, bandwidth latency, cost etc.
    - In some cases these requirements do not fit well together.
  - Communication patterns:
    - Unicast, Multicast, Broadcast and Anycast.
    - Many networks are mixtures of the above.
  - Purpose of the network: enterprise networks support some internal activity while carrier networks exists to make money.
  - Size:
    - LAN (Local Area Network)
    - WAN (Wide Area Network)
    - Internet, the network of networks

OSI model

- OSI stands for Open Systems Interconnection, a conceptual model 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 (synchronization points)
Transport	TCP, UDP
Network	IP addresses, networks
Data-link	MAC (media access control) and LLC sub-layers (logical link control)
Physical	Cables connectors and signals

Utilizing OSI model for communication

- Application is on the top of the OSI stack as it produces the data.
- Data is passed down across 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 place, 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 by Internet Engineering Task Force (IETF)
- 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 the transport layer
  - Application layer – houses the OSI session, presentation and application layer. Contains the application payload such as DNS. Also, any 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 unique 6 byte MAC (media access control) addresses to identify, locate and communicate over the network.
- Builds broadcast domains where every end device can talk to every other end device.
- Data is encapsulated inside frames.

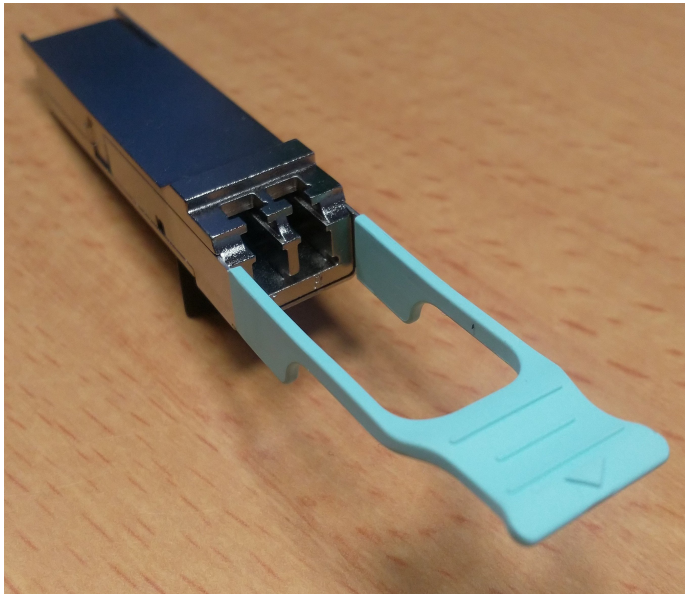
Ethernet frame structure

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.
- 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 IPv4, 0x86DD for IPv6 and 0x8100 for VLAN tagged frame and so forth.
- 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.

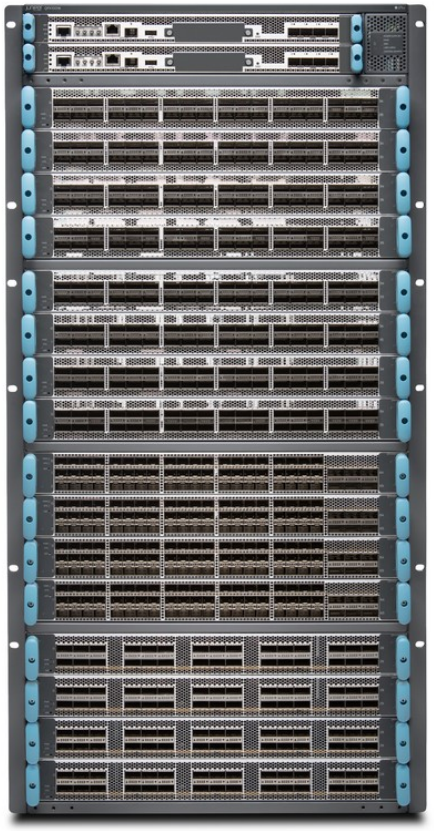
Ethernet implementations

- Ethernet comes in different shapes and speeds. Original experimental version had bandwidth of 2.94Mbps, whereas today we have 400Gbps available.
- 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 a MAC address table. This table is stored to a CAM (content addressable memory) or other fast medium. Switches usually have a general purpose CPU for management purposes but the actual traffic switching is done by an ASIC.
- MAC address table is used to transport traffic out via the correct port.
- Traffic towards an 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). When the possible reply arrives, switch has an opportunity to learn the previously unknown address.



Virtual LAN

- Switches can use Virtual LANs (VLAN) to create logically separated broadcast domain.
- 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 introduce couple of benefits:
  - Reduced size of broadcast domains leads to less “line noise”.
  - Improved isolation as it is not trivial to directly communicate with a device in another VLAN or broadcast domain without crossing a router.
  - It is possible to logically group end devices based on any characteristic.
  - VLANs can also have a role in traffic classification on quality of service (QoS).

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 in hexadecimal format. This 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.
  - It is difficult to get new IPv4 address blocks in the publicly routable address space.
- 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 does not implement broadcast but instead it relies on multicast to improve efficiency and to cut down line noise. Both IP versions have the concept of network address, the first “host address”, network and broadcast addresses should not be configured on end devices except in special cases like point-to-point networks.

IPv4 packet structure

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

- Version, IPv4
- 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

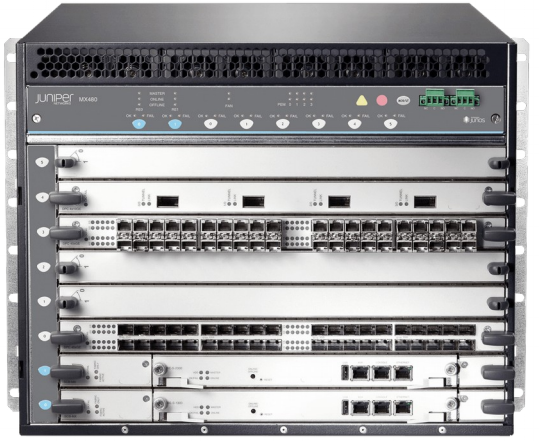
Further headers (TCP, UDP etc.) are encapsulated via the payload.

- Think of this recursive encapsulation like sending letters:
- (Application, brain) produces the actual payload.
- Person writes it down (presentation layer, language) in a suitable medium (transport, letter).
- Stamp and address (network layer) is added on an envelope (data-link).
- Letters moves via mail (physical layer).

Subnets

- IP address configuration consists of two 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
- To get the usable address space: 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 us the more human friendly subnet mask configuration parameter.

IP Routers



- 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.

IP routing steps

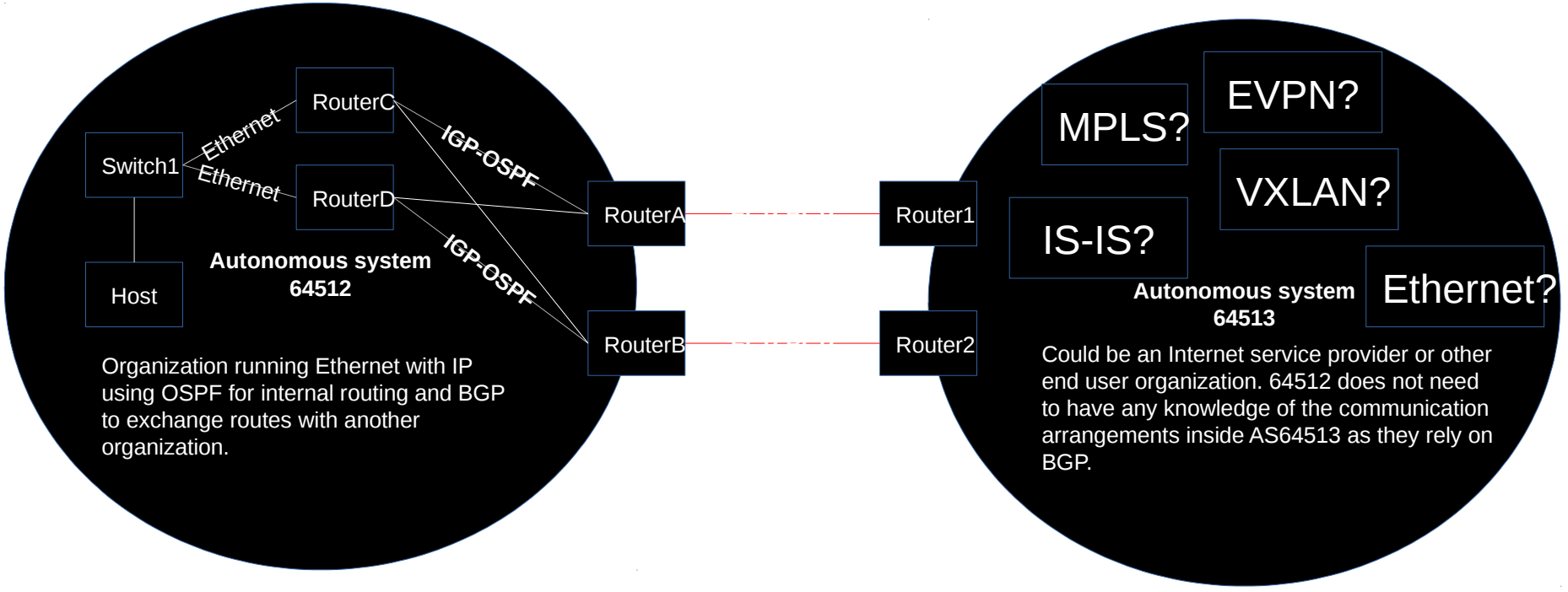
- Simplified steps taken by the network processor on a transit router:
  - Remove the link-layer header containing the routers MAC address
  - Look-up for the destination address from the IP header
  - Possible policing/filtering based on IP and transport layer headers:
    - Protocol, port, flags, TTL, addresses etc.
    - Near line-rate on modern hardware, capabilities depend on hardware
  - Look-up for the destination address from the routing table
  - Look-up for 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.

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 build link-state databases by applying the Dijkstra's algorithm to the received announcements that are sent whenever there is a change in the network.
    - RIP simply exchanges routes with its neighboring routers periodically.
    - Link-state protocols converge fast than RIP.
  - 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.

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

- These are 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

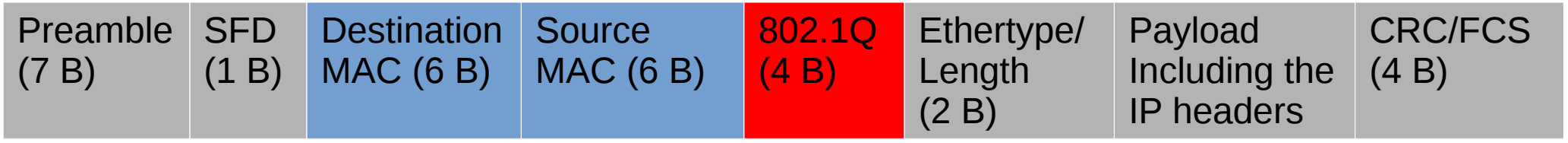
- UDP has the following characteristics:
  - Unreliable but guarantees data integrity.
    - Means that it has no mechanism to detect packet loss, application has to take care of this.
    - UDP packets can arrive in different order than intended and application has to be prepared for this.
  - UDP is 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
    - TCP is able to detect and react to packet loss.
    - 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 a three way hand-shake before data is being transferred.
    - Implements sessions supporting both flow control and congestion management.
  - Example applications: HTTP, FTP and SSH.

Using IP in LAN

- Application decides that it wants to send data to some IP destination. For this, the end node needs to figure out where to send the data:
  - Application does a name resolution to figure out the IP address where to send the data and uses a predefined destination port number (socket).
  - 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.
- Host where the application is running uses the address resolution protocol or neighbor discovery to determine the MAC address of the destination or if routing is required, the MAC address of the router.
- Source and destination MACs are changed at every transport router, while IP information stays the same. Also, VLAN tag might change.



Applying networking concepts to data acquisition

Characteristics of DAQ network

- Several small subnets with routing in between over high speed transport.
- Experiment data is unique and valuable so special care is taken to avoid dropping data while on 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. Custom software solutions may be required on the application level to minimize this.

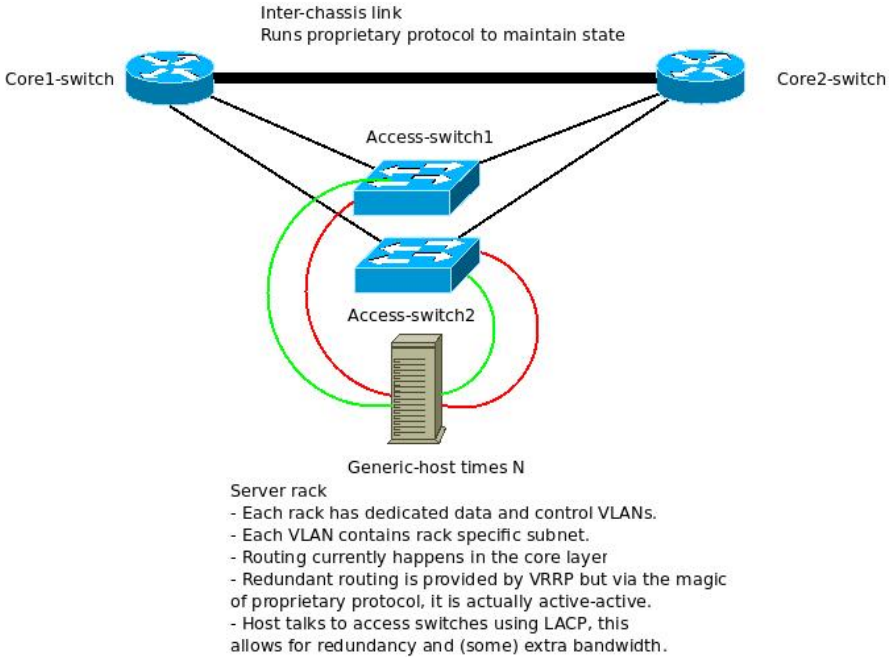
Data acquisition and networking

- 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 scalable means of providing connectivity as applications rely on IP.
- Applications that rely on Layer 2 are problematic and source of scaling issues. 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.



- Routing protocols used: Link-state routing and BGP.
- ECN, Spanning-tree, 802.1Qbb flow control.
- Proprietary extensions to LACP to allow active-active in L2.
- Linux bonding and teaming drivers are heavily utilized.
- In future DAQ applications might be containerized, this could mean overlay networking.

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 the CPU to the Host channel adapter (HCA).
  - RDMA is commonly found in high performance computing setups and comes in a few flavors, for example:
    - Infiniband comes in several speeds and provides possibly lowest latency.
    - RoCE relies on Ethernet and in the case of RoCEv2 also IP and UDP. Basically requires a “lossless” Ethernet setup implemented using Ethernet extensions (DCBX, PFC, ECN etc.)
    - IWARP runs on top of IP and TCP and does not necessarily require special networking arrangements due to TCP.
    - 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

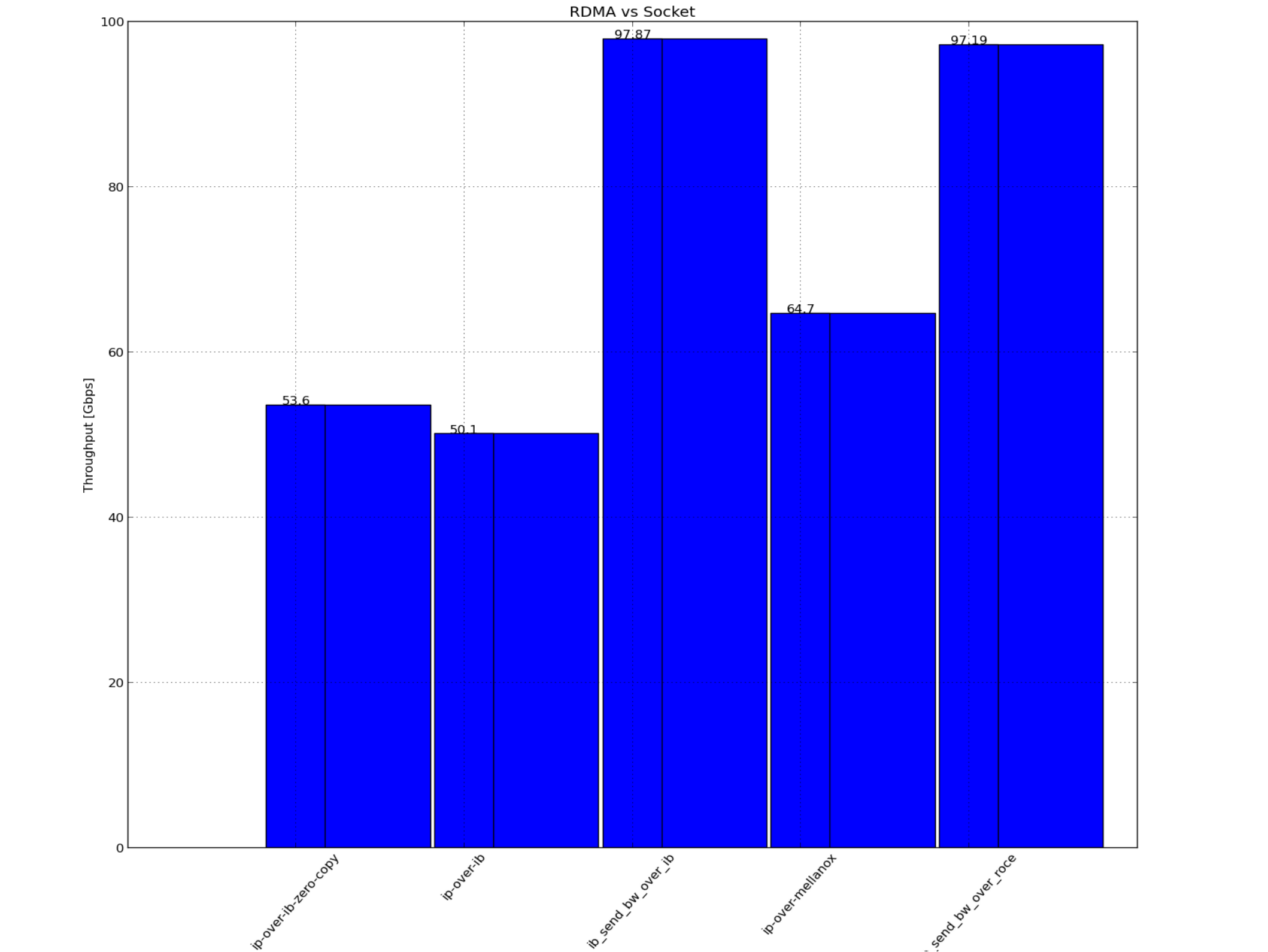
- RDMA connections come in flavors and selecting the correct application approach is essential. Choices have impacts on performance and availability of features and at layers those features need to be implemented on.
- Reliable connected (RC)
  - Guarantees that data is delivered, in-order and with integrity.
  - 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.

The cost of RDMA performance

Key thing to note is that there is no free lunch:

- Socket based applications need to be changed
- Potentially complex changes required on the hosts
- Potentially complex changes required on the network
- Risk of losing interoperability between hardware from different vendors

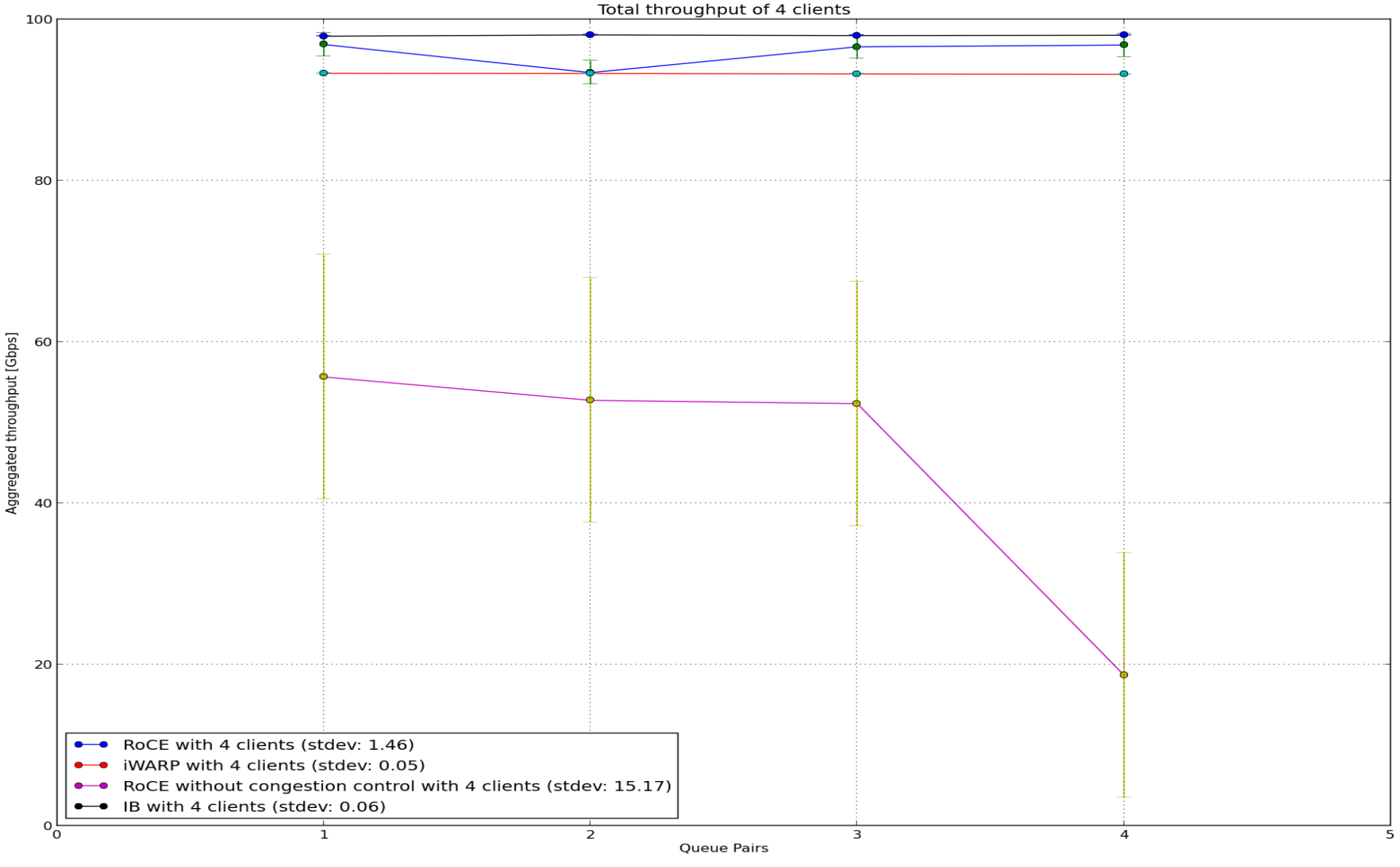
However, the performance improvements in terms of bandwidth and saved CPU cycles can be fairly significant as seen in the next couple of slides..


RDMA vs Socket
Throughput [Gbps]
0
20
40
60
80
100
53.6
50.1
97.87
64.7
97.19
ip-over-ib-zero-copy
ip-over-ib
ib_send_bw_over_ib
ip-over-mellanox
send_bw_over_roce

RoCEv2, what could go wrong?

- Firmware versions of the NICs have to be cross checked against kernel and driver versions.
- It is not enough to configure the network correctly, also the host network stack has to be tuned.
- Congestion control, QoS and the like have proven to be non-trivial for vendors to implement and cross-vendor compatibility has to be tested.
- Keeping mind the above the performance and reliability of the RDMA connections is abysmal when things go wrong as shown in the next slide..

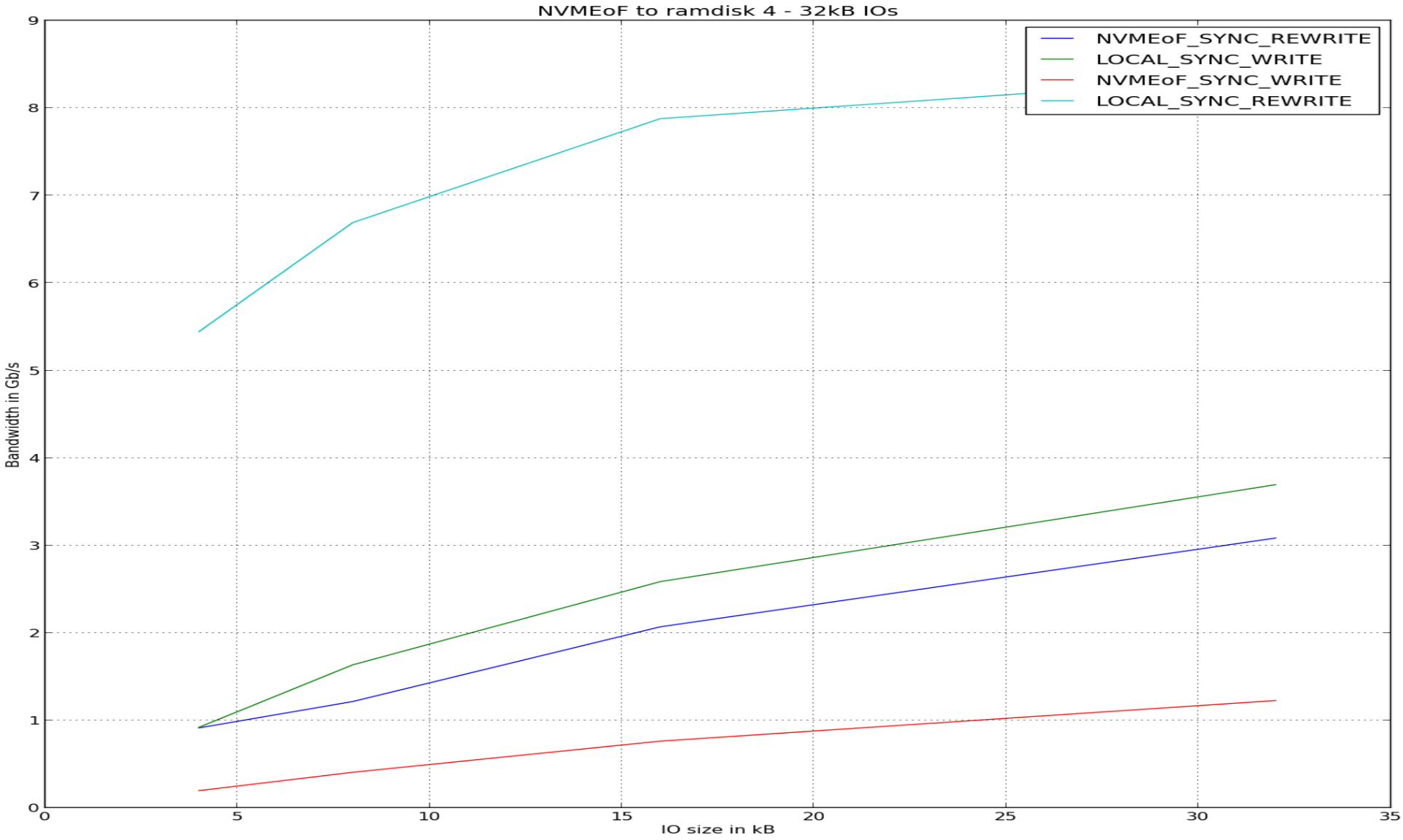
Importance of proper network config



Where might we see RDMA?

- RDMA already in use in ATLAS, likely even more so in the new ATLAS DAQ network. Currently we're using RoCEv2 but we're also evaluating iWARP for other potential future use cases.
- This means that the DAQ network needs to support 'lossless' connectivity via means such as ECN and pause frames.
- Applications relying on message passing interfaces (MPI) and storage (NVME-over-fabrics, iSER) are immediate winners also outside the DAQ domain such as HPC compute clusters.

Storage applications



Conclusions

- DAQ network relies largely on industry standard technologies found in HPC, ISP and enterprise sectors.
- But DAQ network is a challenging combination of 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?
- Containerization of applications might introduce needs for overlays, such as VXLAN. These aspects should maybe be taken into an account when designing new DAQ networks.
- 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?