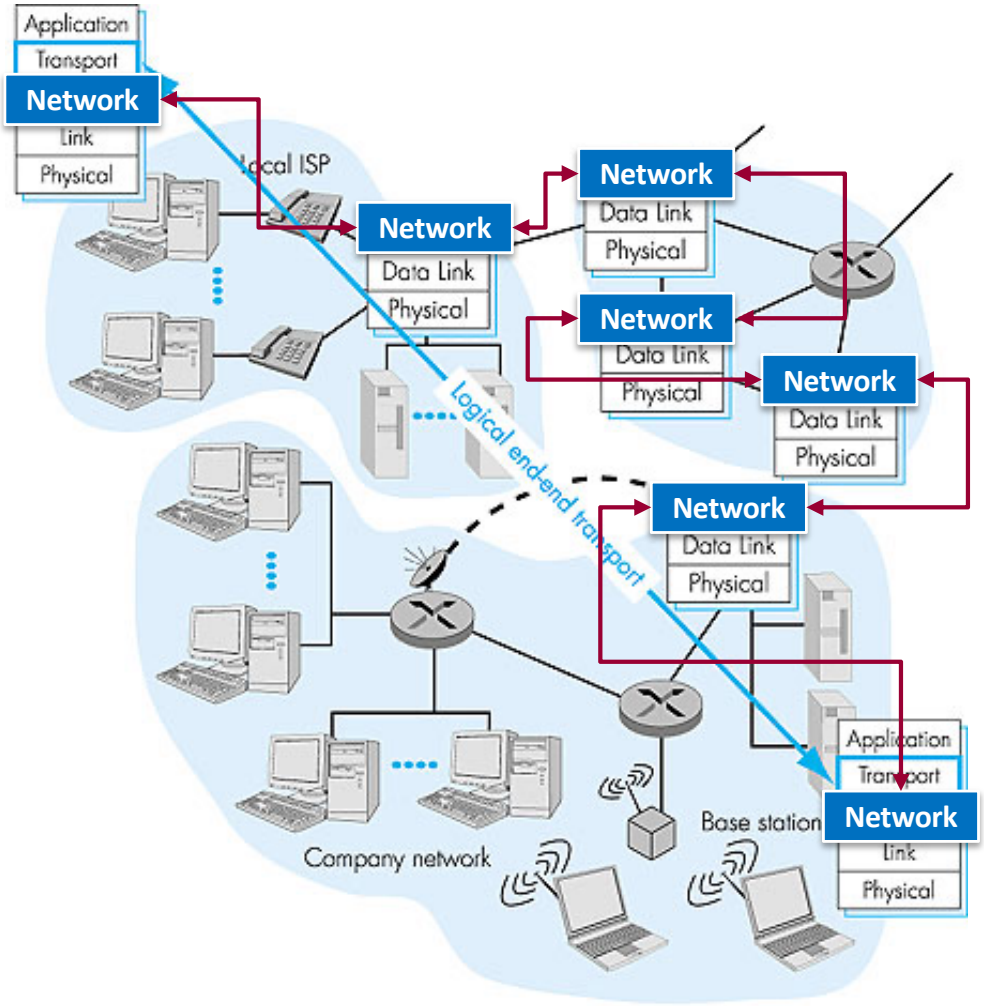
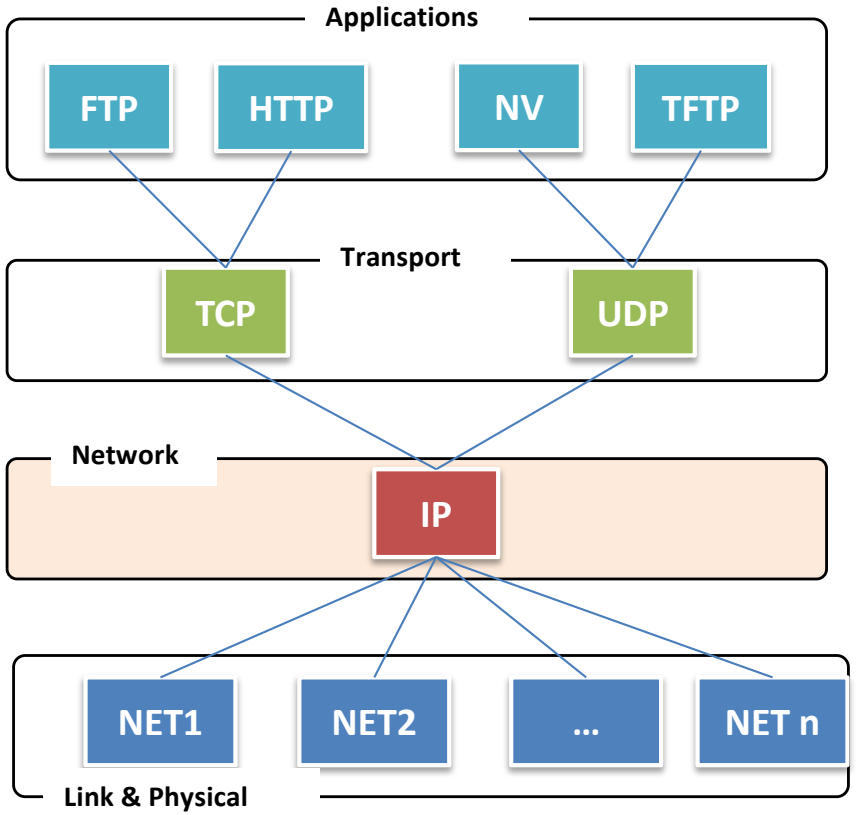
University of BRISTOL

Lecture 7

Basics of network protocols in DCN

Reflections of IP protocol stack

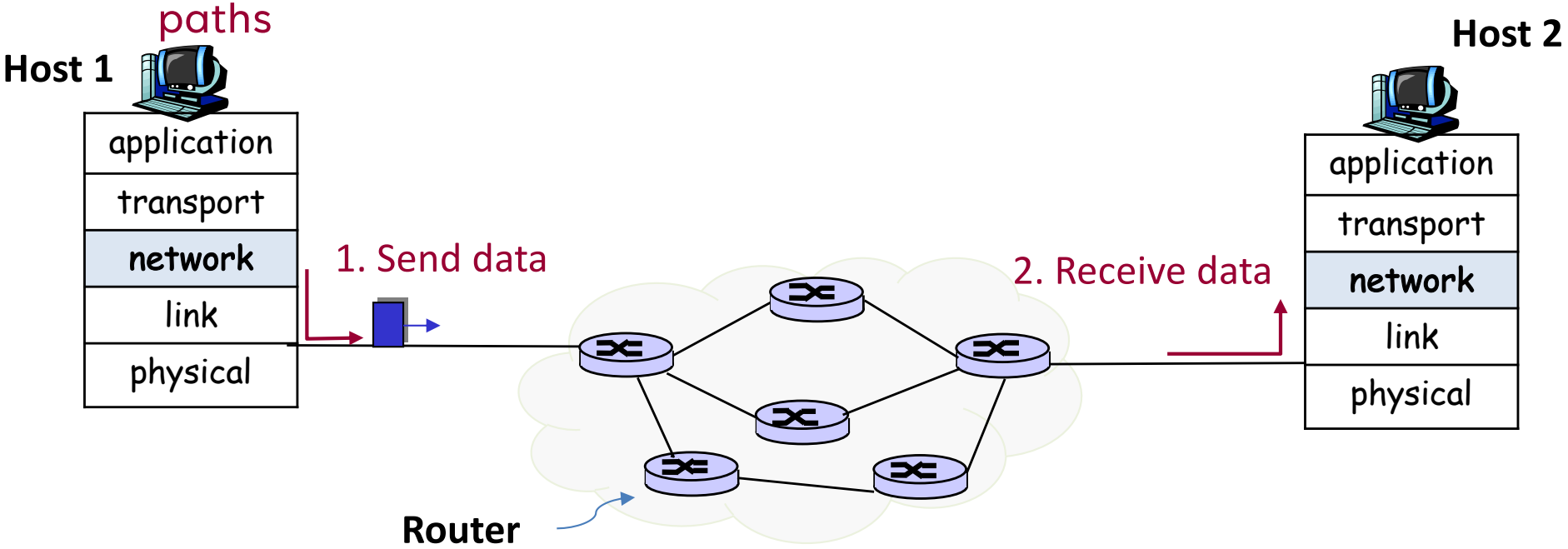




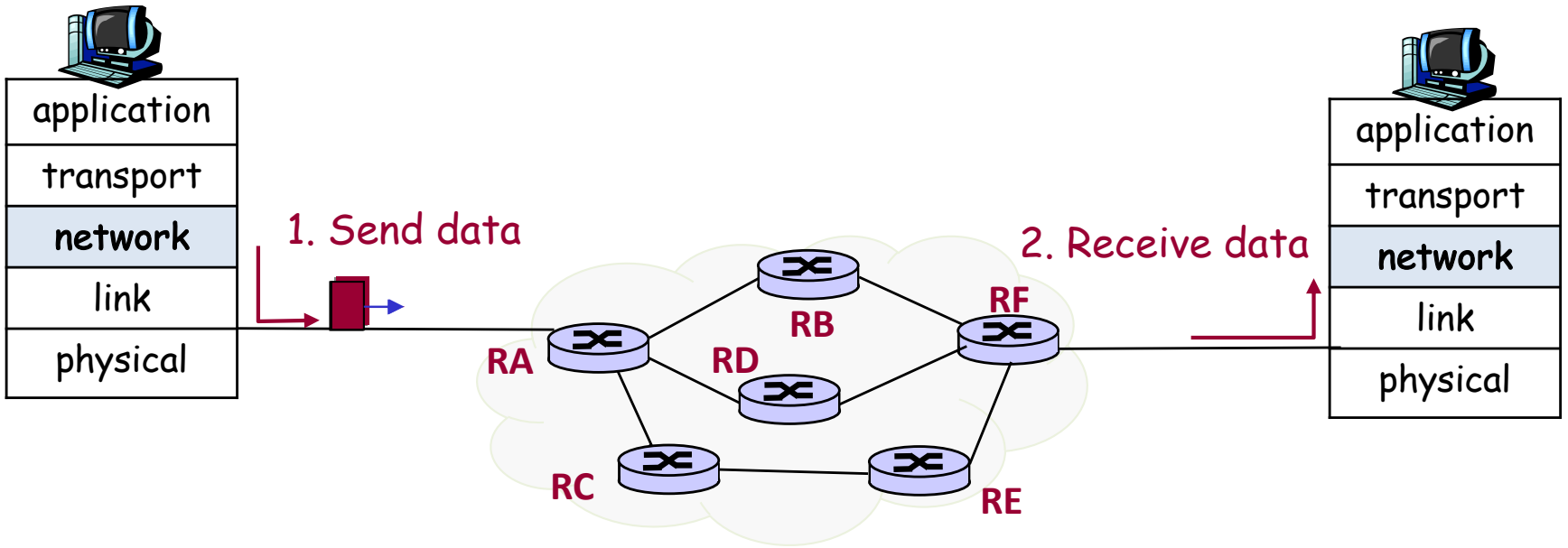
The waist facilitates interoperability.

Datagram networks in network layer

- No call setup at network layer
- Routers: no state about end-to-end connections
  - No network-level concept of “connection”
- Packets forwarded using destination host address
  - Packets between same source-destination pair may take different paths



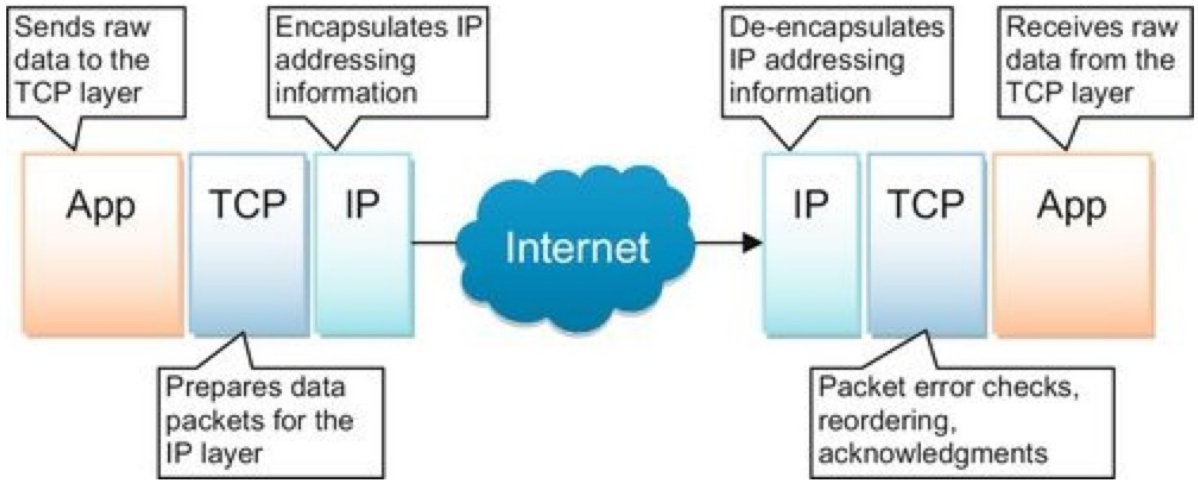
Routing and Forwarding



- Forwarding move packets from router's input to appropriate router's output
- Routing determine routes taken by packets from source to destination.
  - Routing algorithms

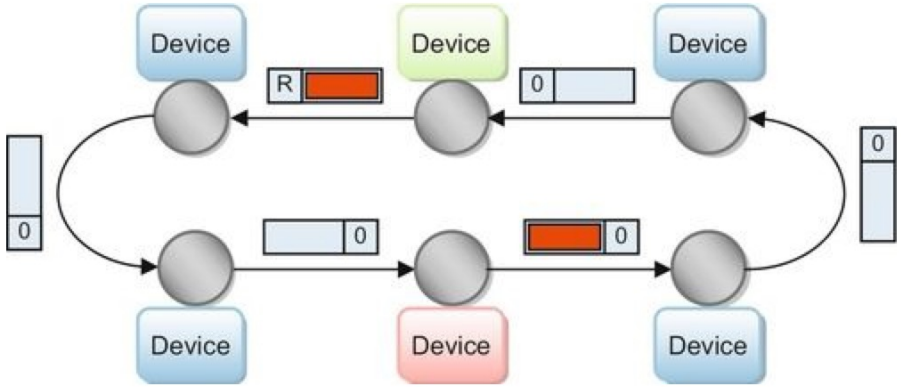
TCP/IP

- Application hands over data to be transmitted to TCP layer
  - This is generally a pointer to a linked list memory location within the CPU.
- TCP layer then segments the date into packets (if larger than maximum packet size) and adds TCP header to each packet
  - Header includes info such as source/destination port that application is using, sequence number, acknowledgement number, checksum, and congestion management information
- IP layer deals with all of the addressing details and adds source/destination IP address to IP packet.



Token ring/token bus

- Token Ring (developed by IBM) in early 1980s.
  - Empty information frames are continuously circulated around the ring.
  - When a device wants to send data to another device, it grabs an empty frame and inserts both the packet data and destination address.
  - The frame is then examined by each successive device, and if the frame address matches a given address, it takes a copy of the data and sets the token to 0.
- The frame is then sent back around the ring to the sending device as an acknowledgment, which then clears the frame.
- This topology is fine for *lightly loaded* networks. If each node wants to continuously transmit data, it will only get 1/N of the link bandwidth where N is the number of nodes.
- Latency is higher than directly connected networks.
- Token Ring was replaced by Ethernet.



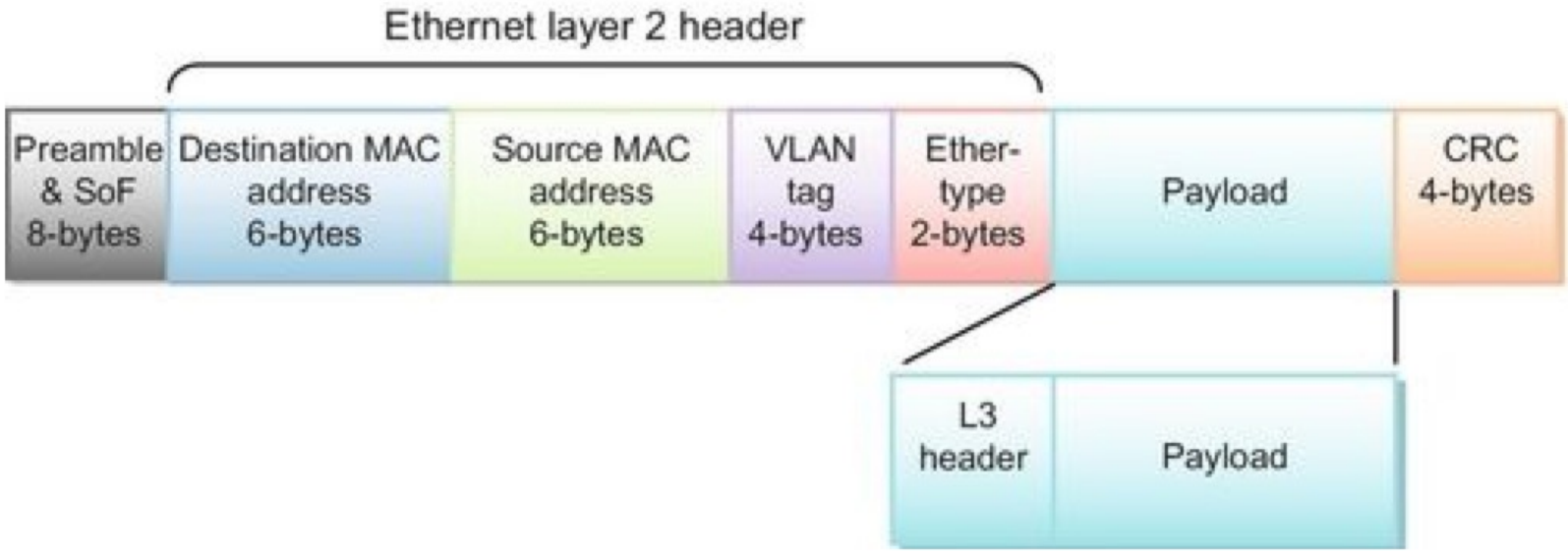
Ethernet

University of BRISTOL

- Introduced in 1980 and standardized in 1985.
- Ethernet started as shared media protocol where all hosts communicated over a single wire or channel.
  - If a host wanted to communicate on the channel, it would first listen to make sure no other communication was taking place.
  - It would then start transmitting and listen for any collisions with other host that may have started transmitting at the same time
  - If a collision was detected, each host would back off for a random time period before attempting again.
    - This is the CSMA/CD (Carrier Sense Multiple Access with Collision Detection) protocol.
- Each endpoint has a dedicated full-duplex connection to a switch that forwards the data to the correct destination address.

Ethernet

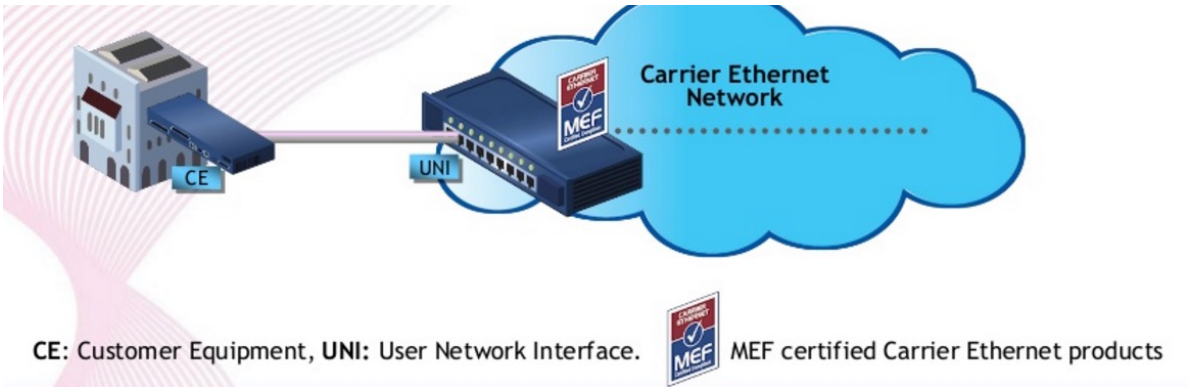
- Ethernet is a Layer 2 protocol.
- The original Ethernet IEEE 802.3 standard defined the minimum Ethernet frame size as **64 bytes** and the maximum as **1518 bytes**.



Carrier Ethernet networks

What do we mean by Metro Ethernet services?

- Use of Ethernet access tails
- Provision of Ethernet-based services across the MAN/WAN
  - Point-to-point
  - Point-to-multipoint
  - Multipoint-to-multipoint
- However, the underlying infrastructure used to deliver Ethernet services does NOT have to be Ethernet !!!



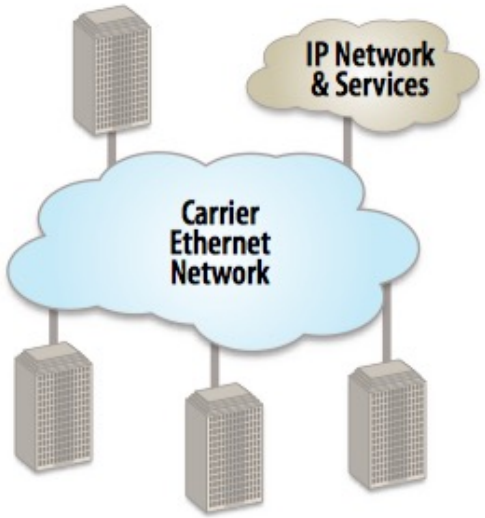
Carrier Ethernet vs Ethernet

Ethernet

- Each user connects to a dedicated Ethernet port on the LAN
- The LAN serves one organization
- The LAN is inside the building

Carrier Ethernet

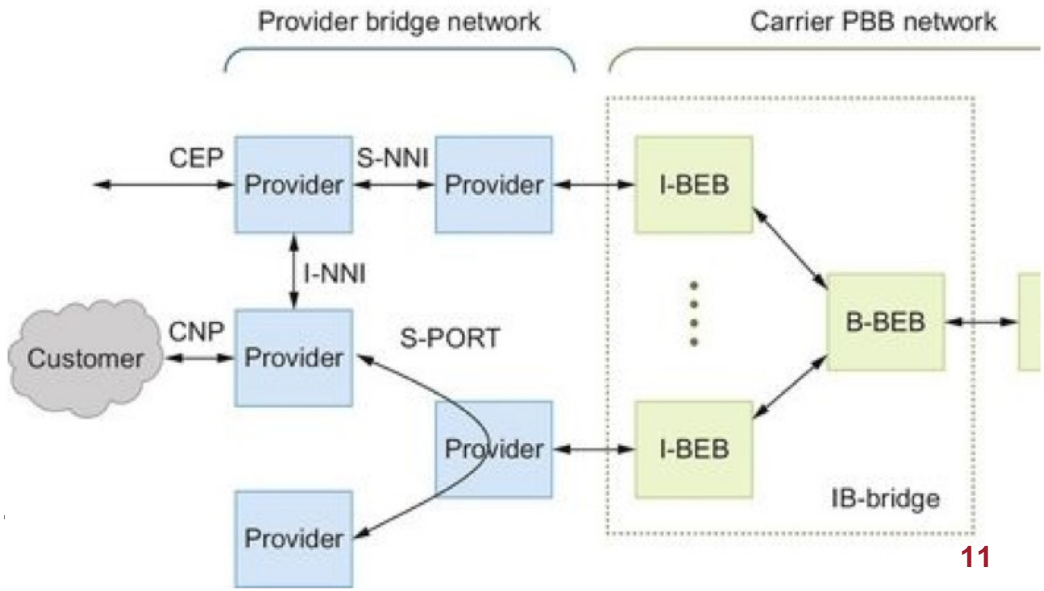
- An entire organization connects to a Carrier Ethernet "port" at a given subscriber location
- The Carrier Ethernet network serves many organizations
- The Carrier Ethernet network is outside the building across a wide area



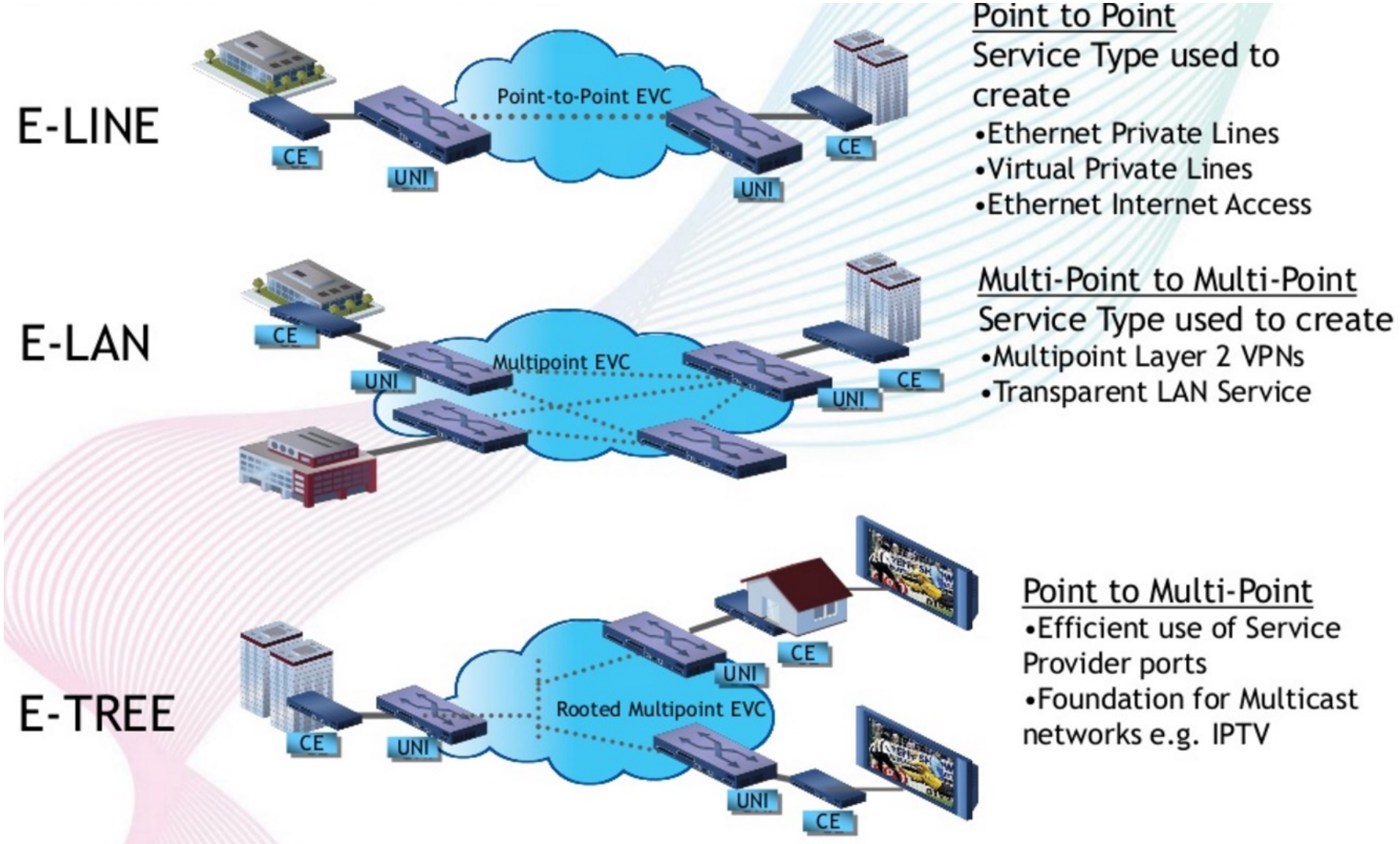
Carrier Ethernet Network

- Carrier Ethernet networks consist of Provided Bridge (PB) networks and Provided Backbone Bridge (PBB) network
  - Provided bridging utilizes an additional VLAN tag to tunnel packets between customers using several types of equipment. **Fundamental limitation of PB is the support of only 4096 special VLAN tags, limiting the scalability of the solution.**
- PBB is using an additional 48-bit MAC address (MAC-in-MAC) to tunnel packets between service providers.
  - I-BEEB (I-component Backbone Edge Bridge) add a service identifier tag and new MAC address based on info in the PB header.
  - B-BEB (B-component …) verifies

the service ID and forwards the packet using backbone VLAN tag.



Basic Carrier Ethernet Services



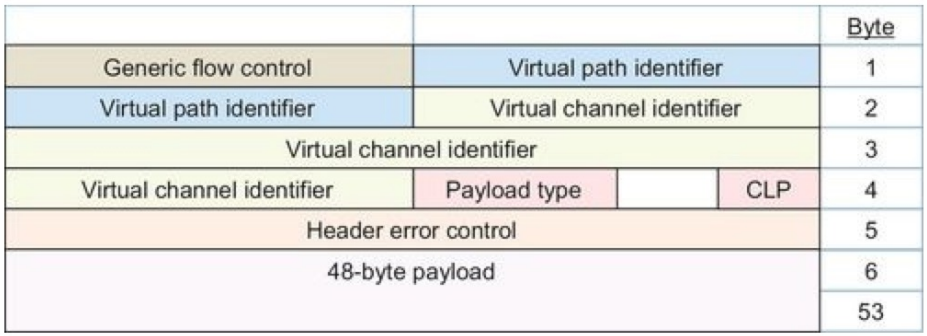
SONET/SDH (Synchronous Digital Hierarchy)

University of BRISTOL

- SONET/SDH was invented to carry multiple calls over a single line together with other types of data.
  - Synchronous Optical Network (SONET) was created as a circuit-switched network originally designed to transport both digitized DS1 and DS3 voice and data traffic over optical networks.
  - However, to make sure that all data falls within its dedicated time-slot, all endpoints and transmitting stations are time synchronized to a master clock, thus the name.
- SONET/SDH is using the concept of transport containers to move date throughout the network.
- SONET/SDH has been used extensively in telecommunication networks, whereas TCP/IP has been the choice for internet traffic. This led to development of IP over SONET/SDH.

Asynchronous Transfer Mode (ATM)

- In late 1980s, ATM emerged as promising new communication protocol.
  - Although ATM did gain some traction in the WAN, it never replaced Ethernet in the LAN.
- ATM frame shows the strong synergy with SONET/SDH.
  - Both use **fixed frames along with the concept of virtual paths and virtual channels**. ATM is also circuit-switched technology.
  - Data can be transferred using multiple virtual channels within a virtual path, and multiple ATM frames will fit within a SONET/SDH frame

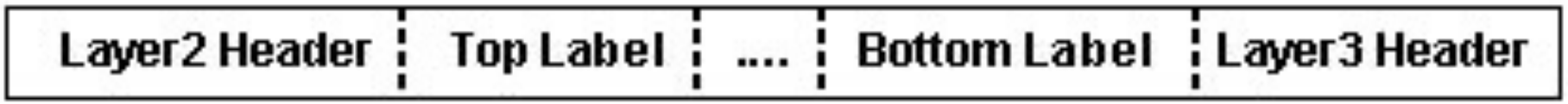


				Byte
Generic flow control	Virtual path identifier			1
Virtual path identifier	Virtual channel identifier			2
Virtual channel identifier				3
Virtual channel identifier	Payload type		CLP	4
Header error control				5
48-byte payload				6
				53

- Most IT networks use variable sized packets, and as link bandwidths increase it becomes more difficult to segment and reassemble data into 53-byte ATM frames, adding complexity and cost to the system. ATM header overhead percentage can be larger than packet-based protocols.
- These are the key reasons that ATM never found success in enterprise or data center networks.

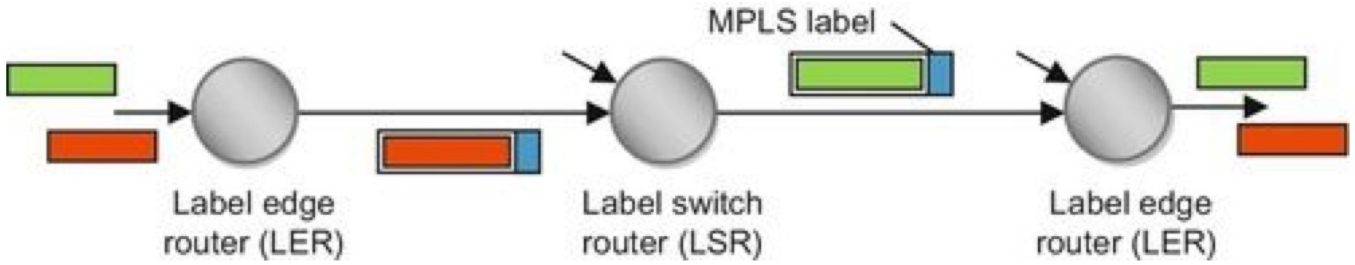
Multi-protocol label switching (MPLS)

- Mid-1990s a group of engineering from Ipsilon Networks thought of adding special labels to the packets (label switching), which the core routers can use to forward packets without the need to look into header details
  - Similar to postal sip code. When a letter is traveling through large postal centers, only the zip code is used to forward the letter. Not until the letter reaches the destination post office (identified by zip code) is the address information examined.
- This was the seed of Multi-Protocol Label Switching (MPLS) which is extensively used today.
  - This idea is also the basis for other tunneling protocols such as Q-in-Q, VXLAN, etc.



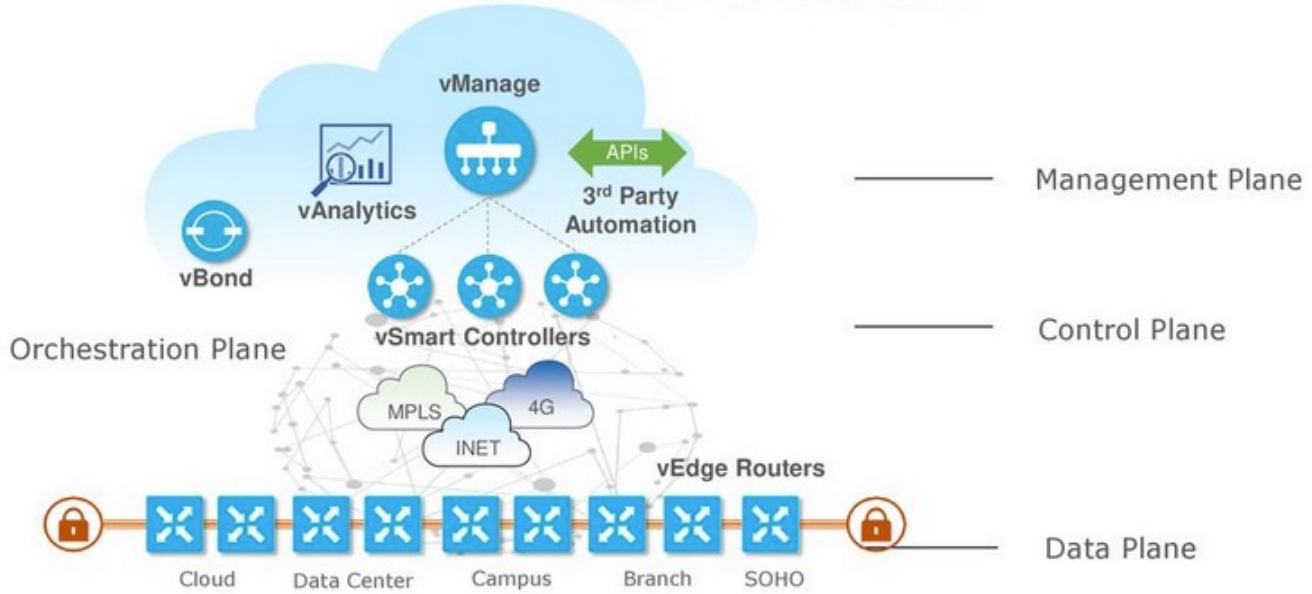
Multi-protocol label switching (MPLS)

- Packets enter an MPLS network through a Label Edge Router (LER) (usually at the edge of the network).



- They append an MPLS label in the packet header.
- Labels maybe assigned using a 5-tuple TCP/IP header lookup, where a unique label is assigned per flow.
- In the core of the network LSRs use the label to forward packets through the network and the egress LERs remove the labels and TCP/IP header info is used to forward packets to destination.

Software Defined WAN (SD-WAN)



- An SD-WAN simplifies the management and operation of a WAN by decoupling the networking hardware from its control mechanism.
- Centralized management or orchestration – the control plane
- Distributed data forwarding function – the data plane
- Application-driven traffic routing policies

Question

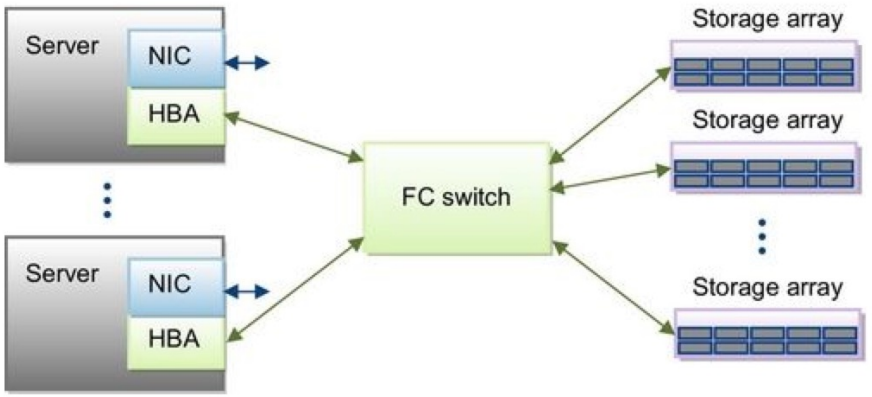
- Comparing SD-WAN vs. MPLS: Which is the Best Choice?

Lossless networks

- Data storage systems require higher resiliency and security due to potentially critical nature of business operations.
- Special applications:
  - iSCSI (Internet Small Computer System Interface) local traffic
  - LAN Backup
  - Virtual Machine Mobility
  - Cluster
  - HPC

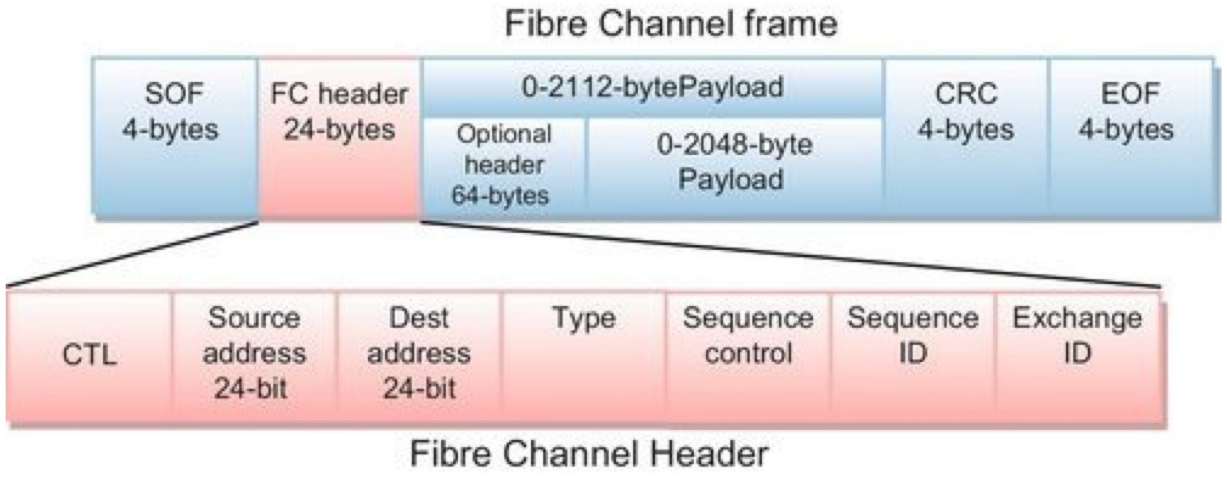
Fiber Channel (FC)

- Data storage systems in DCN require networks to offer higher resiliency and security due to potentially critical nature of business operations.
  - Storage traffic cannot tolerate such retransmission delays and for security reasons many IT managers want to keep storage on isolated networks.
  - Ethernet allows packet to be dropped under certain conditions, with the expectation that data will be retransmitted at a higher later such as TCP.
- FC requires the use of **Host Bus Adapters (HBAs)** that are similar to network interface cards (NICs) and are resident on the server, along with storage arrays that have FC interfaces.
- One or more FC switches are used to interconnect multiple servers to multiple storage arrays.



FC protocol

- FC is a serial protocol.
- Success in switch fabric topologies used in SANs providing rates of 1G, 2G, 4G, 8G, and 16G bit per second using optical cabling.
- Frame format:



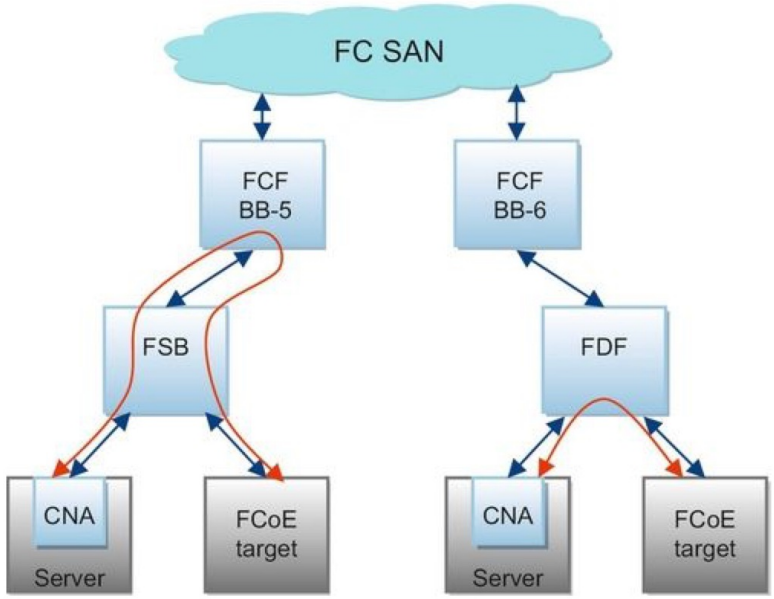
- FC header supports 24-bit address to cover over 16M ports.
- The payload can range from 0 bytes to over 2KB and the frame is protected by CRC(Cyclic Redundancy Check) field.
- Frame delivery order is not guaranteed so sequence control and sequence ID field is used along with an exchange ID.

Fiber Channel over Ethernet (FCoE)

- Many Data Centers include dedicated FC SANs (storage area networks) in addition to their Ethernet data networks.
- There is a strong driving force to converge these two into a single network, for cost and operational purposes.
- FCoE was standardized (T11 FC-BB-5) in 2009 by the International Committee for Information Technology Standards.
  - It depends on IEEE DCB standards for efficient operation
- When using FCoE, the FC frame is encapsulated with both an FCoE header and an Ethernet header.
  - The FC frame itself is a transport for SCSI commands.
  - The Ethernet header contains an EtherType of 0x8906 to identify the frame as FCoE.
  - 4-bit version number is used along with Start of Frame (SOF), and End of Frame (EOF) indicators.
  - Reserve bytes are added to maintain the minimum size Ethernet frame when encapsulating smaller FC command frames

FCoE networking

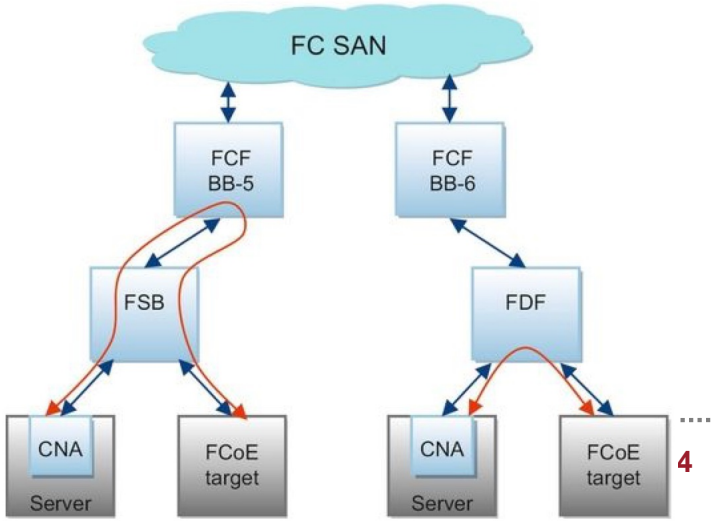
- In the servers, converged network adapters (CNAs) are used to connect to the Ethernet network and can provide the functionality of both traditional NIC as well as an FCoE HBA by generating and receiving FCoE frames.



- Special Ethernet switches called Fibre Channel Initiator Protocol (FIP) Snooping Bridges (FSBs) are used. FIP Snooping makes sure that only servers that have logged in to the FC network can have access to that network.

FCoE networking

- Fibre Channel Forwarders (FCFs) are switching devices used to forward all FCoE frames. These also act as bridges to traditional FC SANs by encapsulating and decapsulating FC frames.
  - There is a need of routing all data through FCFs and all FSBs must be connected to a FCF that increases congestion and limits the usefullness of FoE targets. i.e. an FCoE storage target connected to same FSB must have its data routes through the FCF (left side of figure)
  - FC-BB-6 standard introduced a new type of Ethernet switch called Fibre Channel Data Forwarder (FDF) capable of FCoE forwarding and zoning base on info provided by FCF (right side of figure). There is no need to directly connect an FDF to an FCF .



Fiber Channel over Ethernet (FCoE)

FIBRE CHANNEL OVER ETHERNET (FCOE)

Education SNIA

- From a Fibre Channel standpoint
  - FC connectivity over a new type of cable called… an Ethernet cloud
- From an Ethernet standpoint
  - Just another ULP (Upper Layer Protocol) to be transported,
    - but… a challenging one!
  - DCB designed to meet FCoE's requirements

FC-BB-5: VE-VE & VN-VF, FC-BB-6 adds VN2VN

Class 2, 3, and F carried over FCoE

Ethernet Support

Lossless – aka not allowed to discard because of congestion
Transit delay of no more than 500ms per forwarding element
Shall guarantee in order delivery

- Components
  - FCoE/FC Switches (or FCFs)
  - FCoE/FC Gateways (NPIV based)
  - FCoE Transit Switches (DCB plus FIP-Snooping)

Technical Overview of Data Center Networks

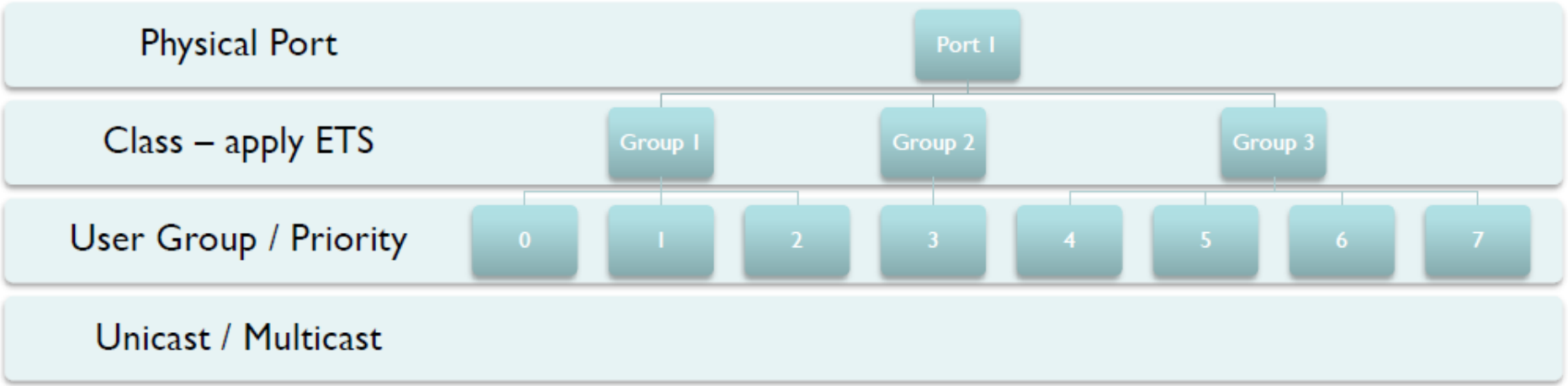
11

Data Centre Bridging (DCB)

University of BRISTOL

- Extending Ethernet's Capabilities in the Data Center
  - To improve and expand Ethernet networking and management capabilities in the data center. It helps ensure delivery over lossless fabrics and I/O convergence onto a unified fabric.
- Partitions the network into parallel planes
  - 8 largely independent lanes
  - Configurable separately
  - Leverages the 3-bit VLAN CoS (aka user priority)
- Designed for protocol/service separation
  - Individual protocols on each lane / class group

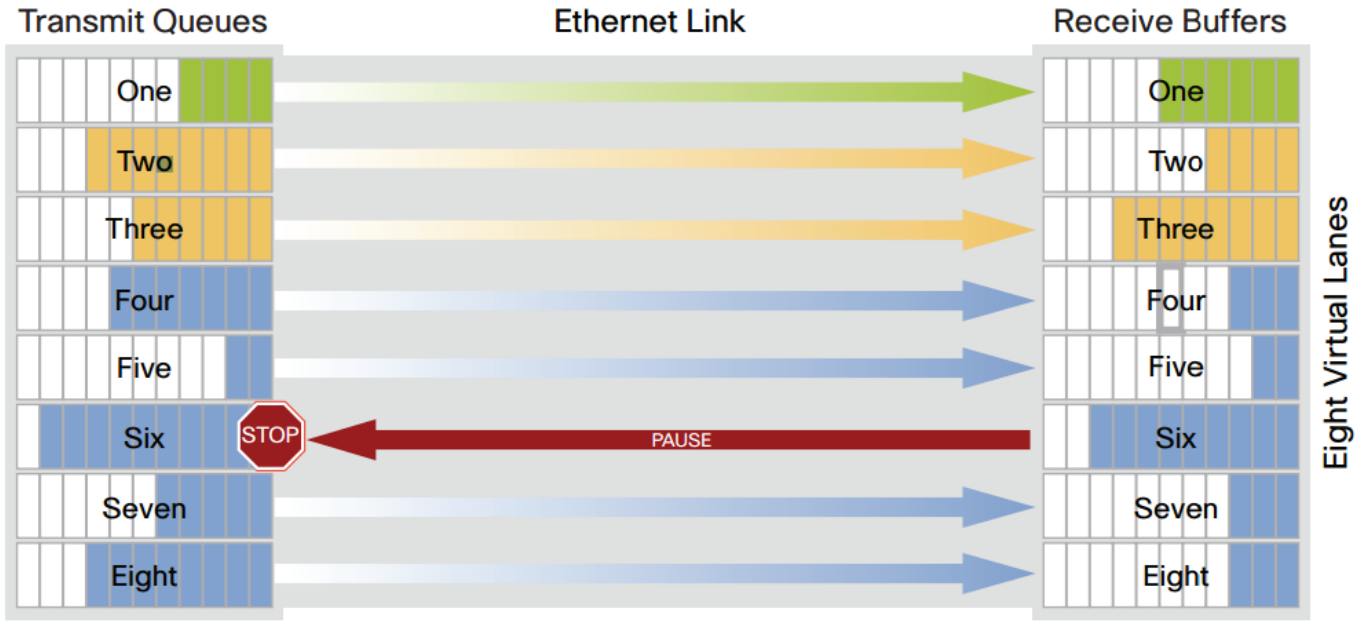
DCB Hierarchy – conceptual view



Implementations may have separation of multicast & unicast

DCB features

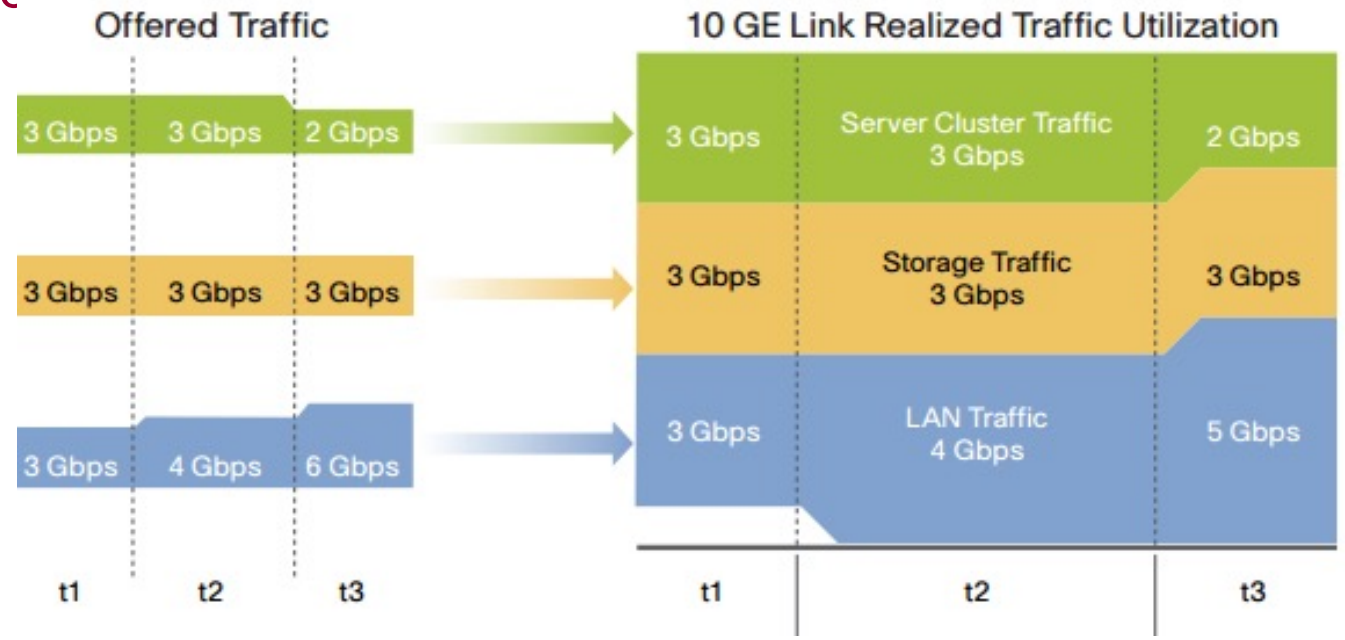
- Priority-based flow control (PFC)



- Calculations for Buffer Requirements When Using PFC PAUSE
  - Processing and queuing delay of the PFC PAUSE
  - Propagation delay across the media
  - Response time to the PFC PAUSE frame
  - Propagation delay across the media on the return path

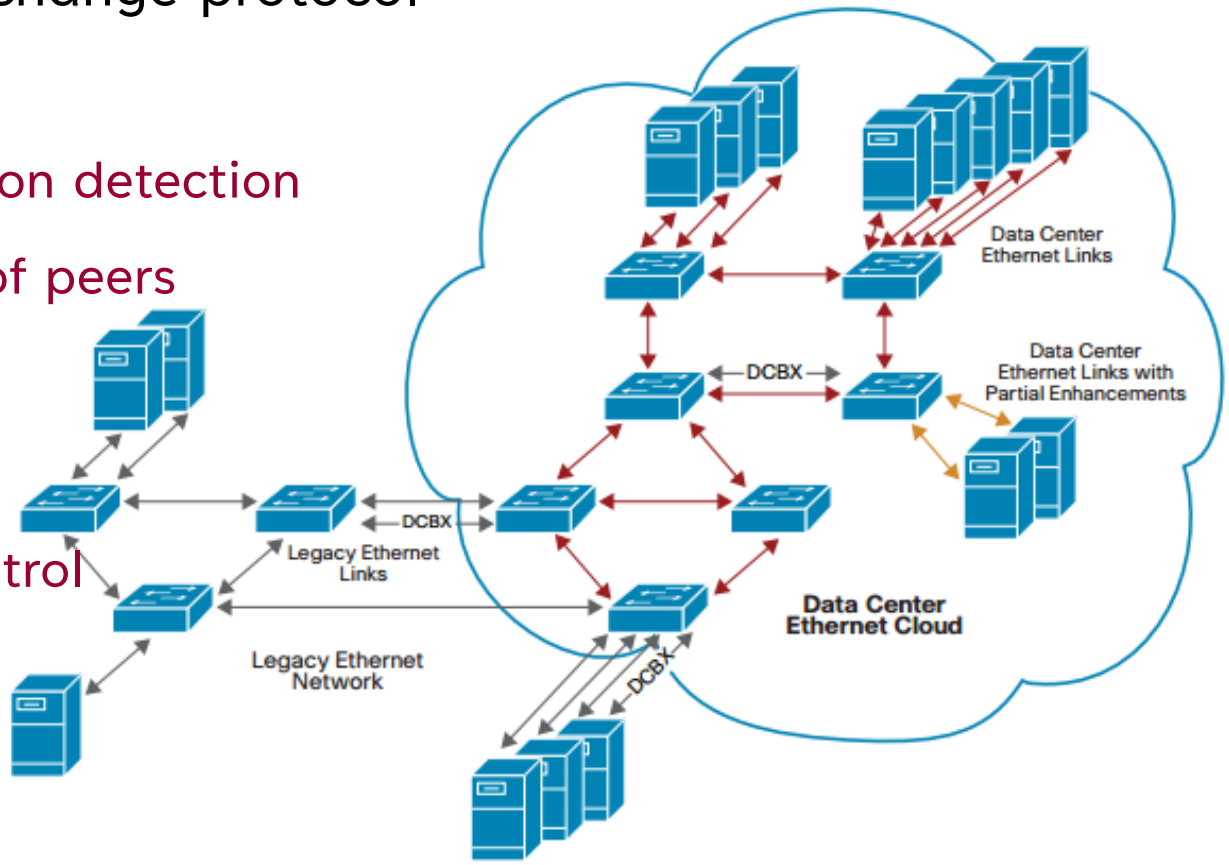
DCB features

- Enhanced transmission selection
  - ETS provides prioritized processing based on bandwidth allocation, low latency, or best effort, resulting in per-group traffic class alloc



DCB features

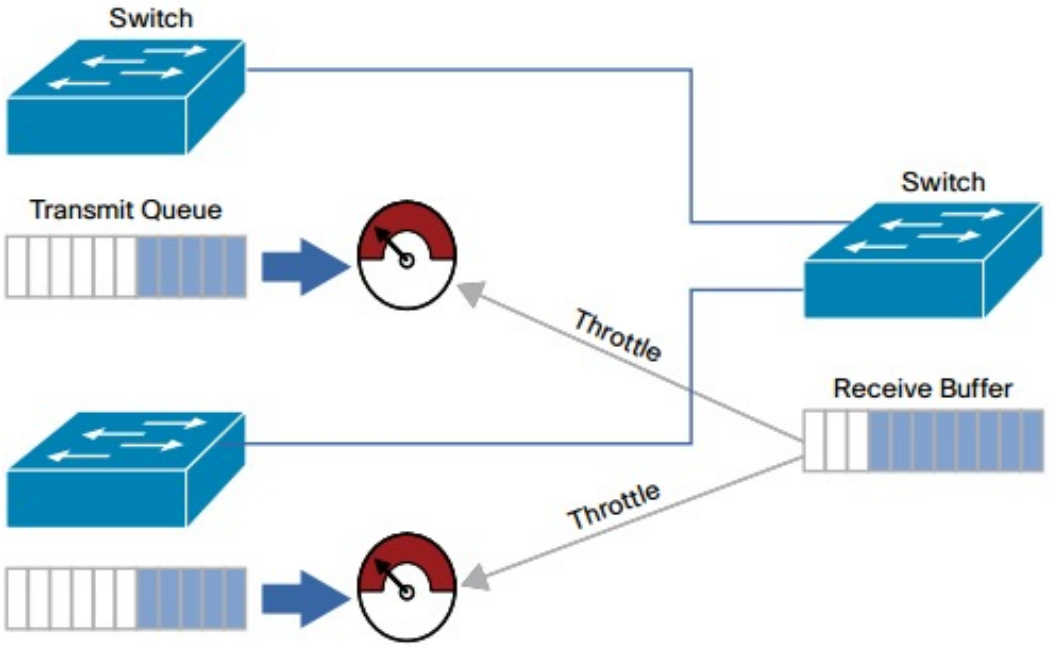
- Data center bridging exchange protocol
  - DCB peer discovery
  - Mismatched configuration detection
  - DCB link configuration of peers
- Information exchanged:
  - Priority groups in ETS
  - Priority based Flow Control
  - Congestion Notification
  - Applications
  - Logical link-down
  - Network interface virtualization



Cisco Unified Fabric Related Standards and Enhancements

DCB features

- Congestion notification
  - Congestion Notification is a Layer 2 traffic management system that pushes congestion to the edge of the network by instructing rate limiters to shape the traffic causing the congestion.



Protocols for Supercomputing

I/O architectures: Fabric vs bus

University of BRISTOL

- The shared bus architecture is the most common I/O interconnect today, although there are numerous drawbacks.
  - Clusters and networks require systems with high-speed fault-tolerant interconnects that cannot be supported properly with a bus architecture.
  - Thus, all bus architectures require network interface modules to enable scalable network topologies. To keep pace with systems, an I/O architecture must provide a high-speed connection with the ability to scale.
- Table below provides a simple feature comparison between a switched fabric architecture and a shared bus architecture.

Feature	Fabric	Bus
Topology	Switched	Shared bus
Pin count	Low	High
Number of endpoints	Many	Few
Max signal length	KMs	Inches
Reliability	Yes	No
Scalable	Yes	No
Fault tolerant	Yes	No

I/O architectures: Fabric vs bus

- The shared bus architecture
  - All communication shares the same bandwidth
  - The more ports added to the bus the less bandwidth
  - On parallel bus many pins are needed for each connection (64-bit PCI requires 90 pins)
    - Challenging PCB layout
  - At high frequencies, the distance of each signal is limited to short distances on the PCB board.
- Switched Fabric architecture
  - Is a point-to-point switch-based interconnect
  - Offers fault tolerance
  - Scalability by adding switches to the fabric and connecting more endnodes through the switches.
  - Each link has exactly one device connected at each end
    - Well controlled loading and termination characteristics.
  - Unlike a shared bus architecture, the aggregate bandwidth of a system increases as additional switches are added to the network.
    - Multiple paths between devices keep the aggregate bandwidth high and provide fail-safe redundant connections.

Standardization of InfiniBand

- InfiniBand Trade Association (IBTA)
- Founded in 1999
- Actively markets and promotes InfiniBand from an industry perspective through public relations engagements, developer conferences and workshops
- InfiniBand software is developed under OpenFabrics Open Source Alliance
- http://www.openfabrics.org/index.html
- InfiniBand standard is developed by the InfiniBand Trade

InfiniBand (IB)

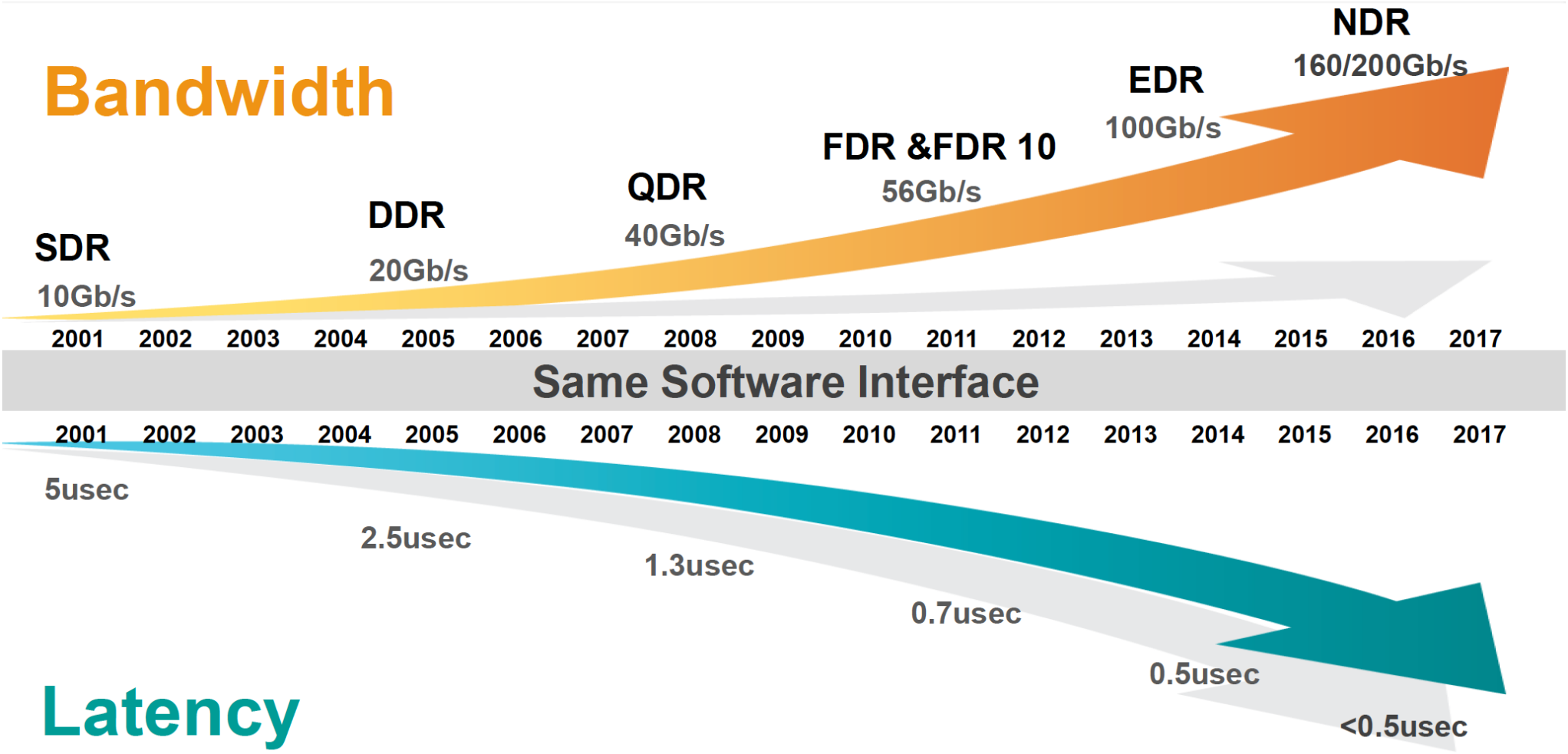
- InfiniBand is mostly used as fast interconnect for the inter-process communication within parallel cluster computers.
- InfiniBand replaces the PCI bus with the switch-based serial interconnect architecture.
  - Devices communicate by means of messages
  - An InfiniBand switch forwards the data packets to the receiver.
  - Communication is full-duplex and transmission rate of 2.5 Gbps or 10 Gbps in each direction is supported.
  - Unlike shared bus architectures, InfiniBand is a low pin count serial architecture that connects devices on the PCB and enables "Bandwidth out of the Box" spanning distances up to 17m over ordinary twisted pair copper wires.
  - Transmits data in packets of up to 4KB that are taken together to form a message.
  - Provides remote direct memory access (RDMA) for low CPU overhead

InfiniBand advantages

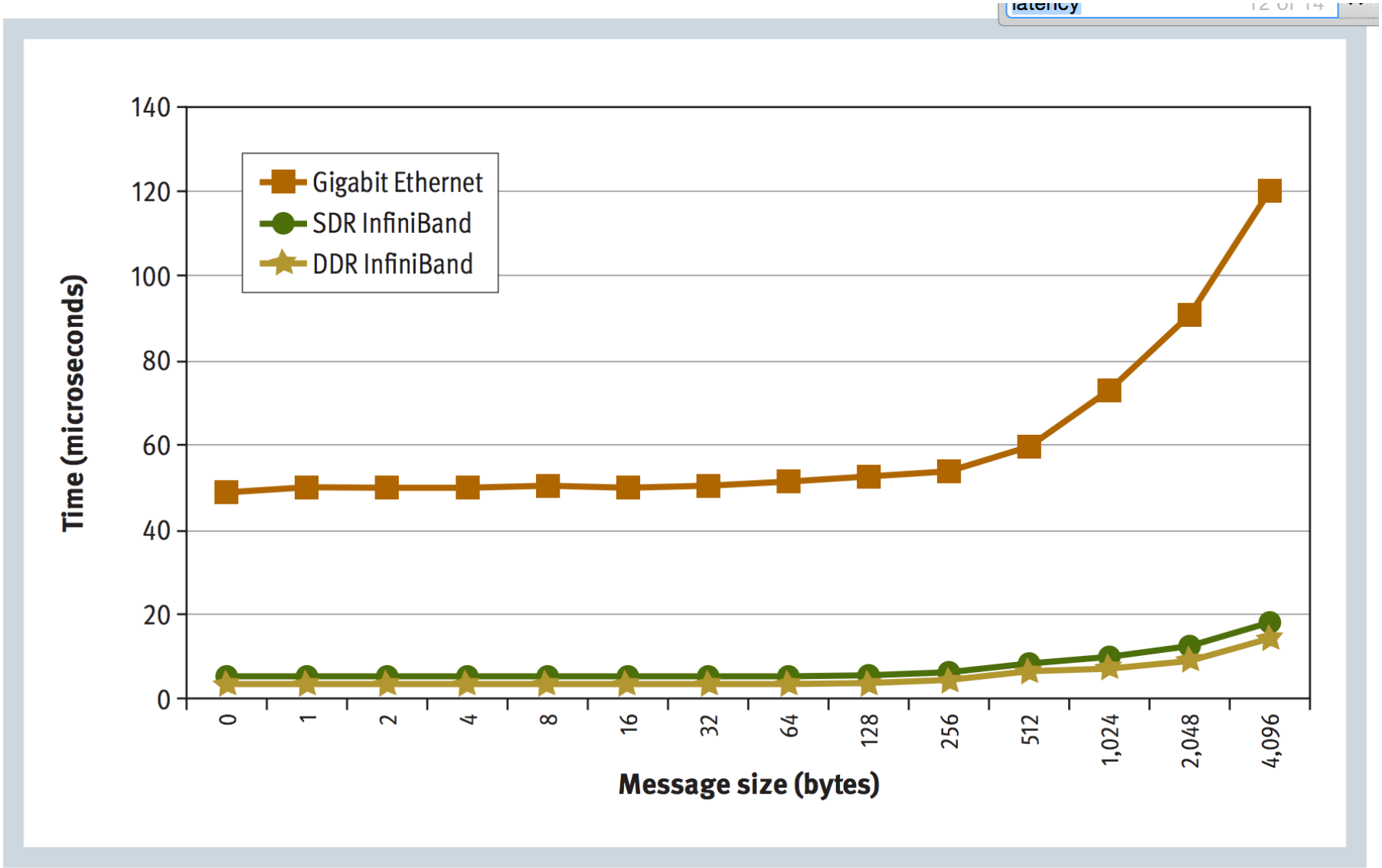
University of BRISTOL

- InfiniBand is a switch-based technology interconnecting CPUs and I/Os
- High Performance
  - High Bandwidth
  - Low latency < 1 microsecond (<200 nsec per switching hop)
  - Low CPU utilization with RDMA (Remote Direct Memory Access)
    - Unlike Ethernet, Traffic communication bypasses the OS and the CPU's.
- Originally designed for large-scale Grids and Clusters
- Increased application performance
- Single port solution for all LAN, SAN and application communication
- High reliability Cluster management (Redundant Subnet Manager)
- Automatic Cluster switches and ports configuration performed by the Subnet Manager Software (SW)

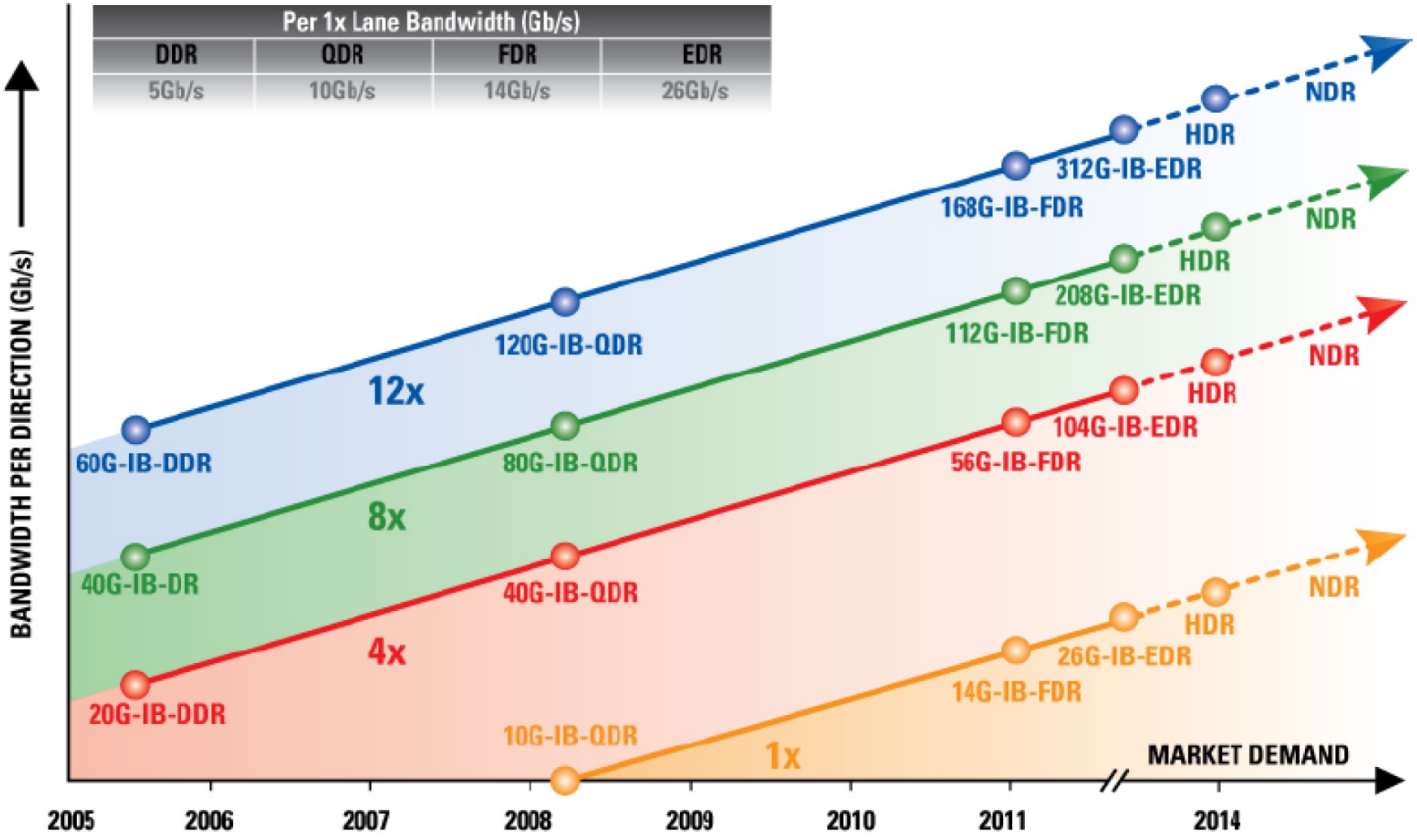
InfiniBand Performane



Latency comparison between GE and SDR/DDR InfiniBand

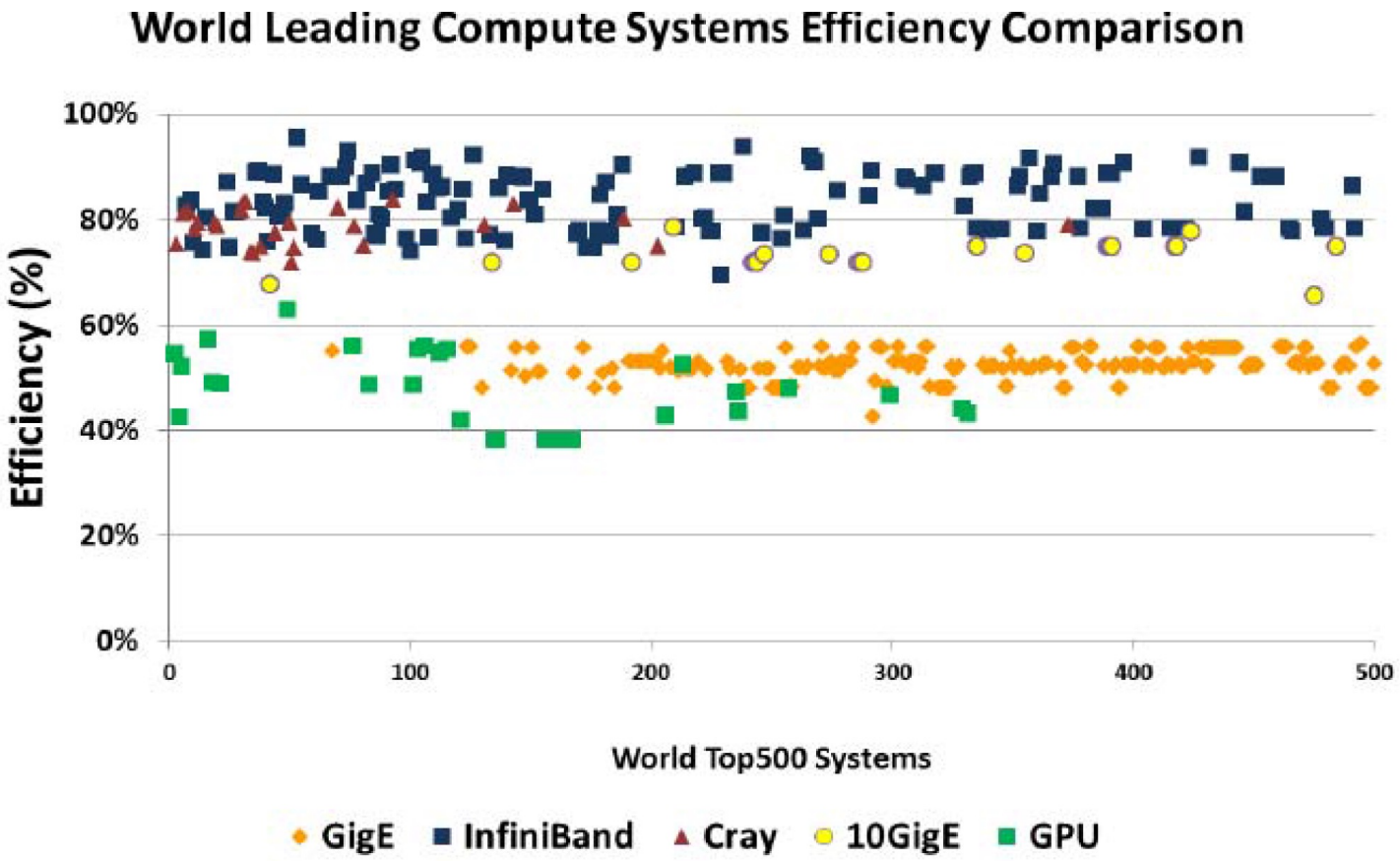
- 

InfiniBand Link Bandwidth Roadmap



Ref: Alan Benner, "Optical Interconnet Opportunities in Supercomputers and High End Computing", OFC 2014

Network protocols key to compute system efficiency

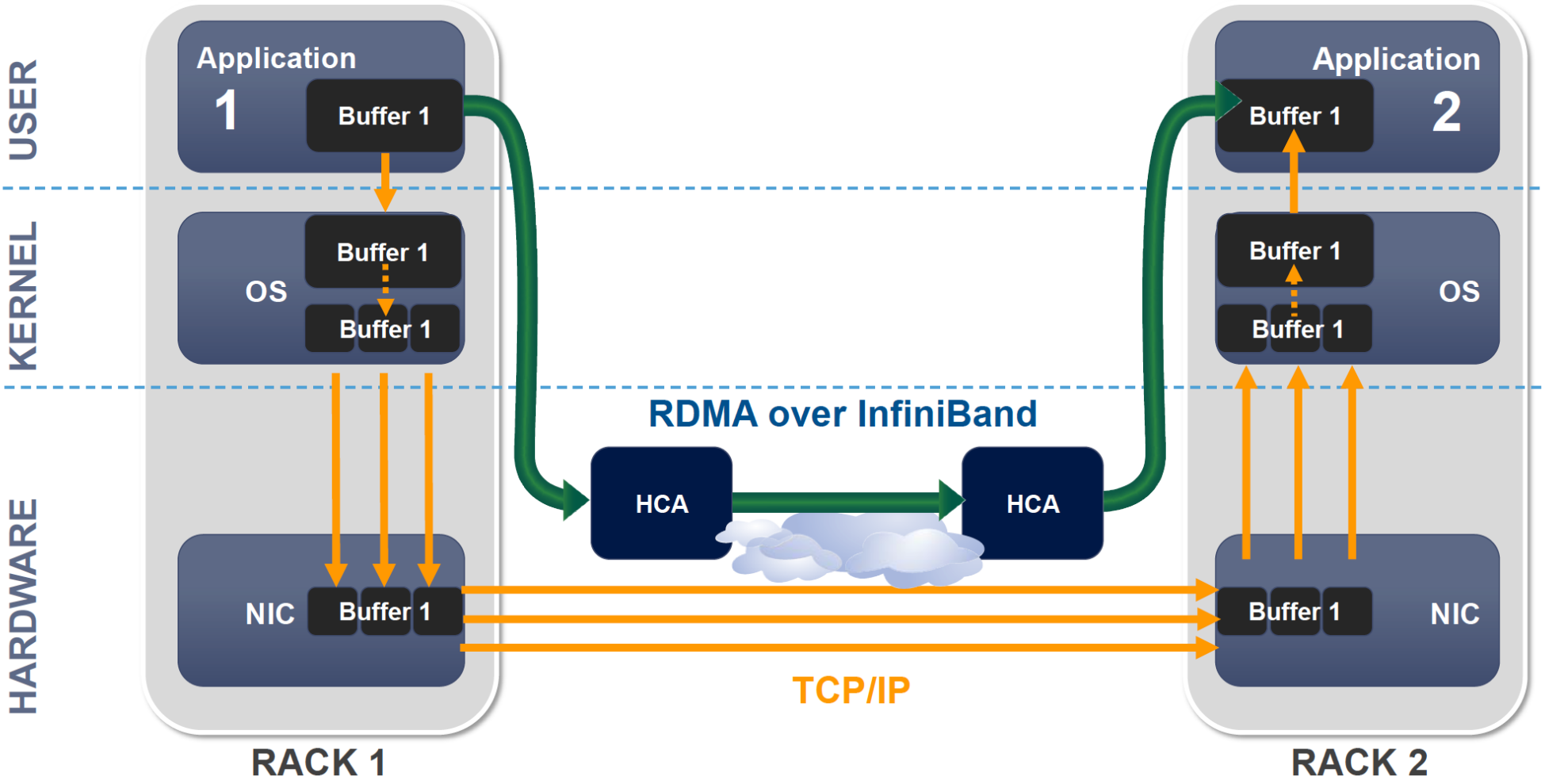


- TOP500 systems listed according to their efficiency
- InfiniBand is the key element responsible for the highest system efficiency
  - Up to 96% efficiency

Remote Direct Memory Access (RDMA)

University of BRISTOL

•



Final thoughts

- The large number of protocols in operation leads to many complex interactions
- The type of complexity outlined here is manageable with proper designs
- Several of the protocols are specifically targeted at mitigating negative effects

Reference list:

- MPLS-based Metro Ethernet Networks Tutorial by Khatri, link: https://www.slideshare.net/feb_989/mplsbased-metro-ethernet-networks-tutorial-by-khatri
- DCB:
  - http://www.cisco.com/c/dam/en/us/solutions/collateral/data-center-virtualization/ieee-802-1-data-center-bridging/at_a_glance_c45-460907.pdf
  - http://www.juniper.net/techpubs/en_US/learn-about/data-center-bridging.pdf
- A. Benner, “Optical interconnect opportunities in supercomputers and high end computing,” in OFC/NFOEC, 2012.