Introduction to InfiniBand (IB)

Presented by: Mohammad Parsa Bashari

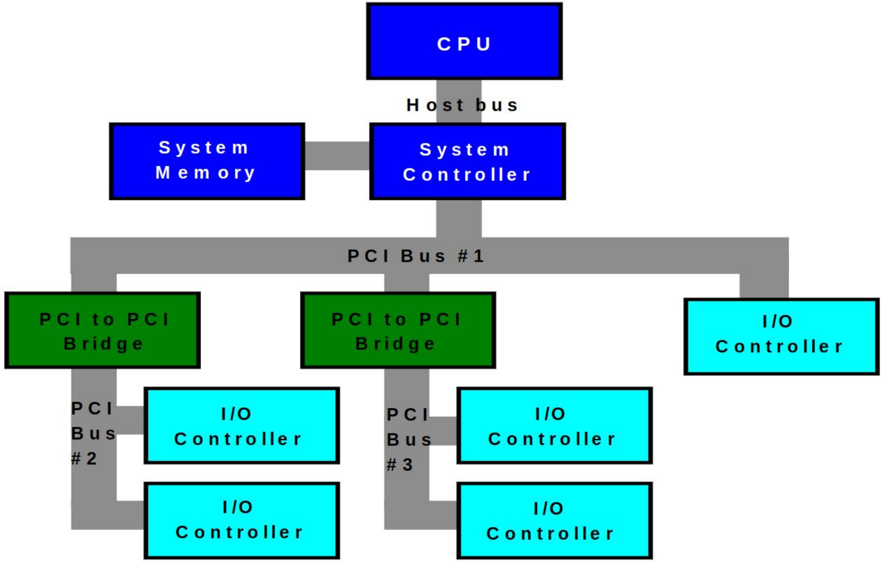
Interface Circuits (Fall 2024)

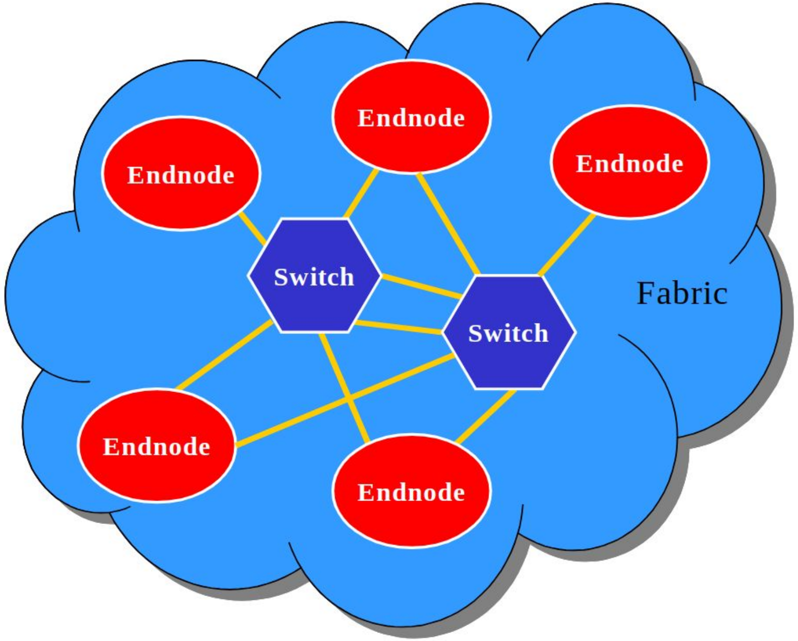
What is InfiniBand?

- InfiniBand is an industry standard, channel-based, switched fabric interconnect architecture for **server and storage connectivity** (Used in **HPC**, **AI**, and **cloud data centers**).
- It is used both for inter- and intra-computer communication.
- InfiniBand offers high data transfer rates, ranging from **10 to 400Gb/s**.
- Most of the world’s fastest supercomputers leverage InfiniBand, connecting **63 of the top 100** supercomputers on the TOP500 list.
- It implements **RDMA (Remote DMA)**.
- InfiniBand is a layered protocol.

InfiniBand Architecture

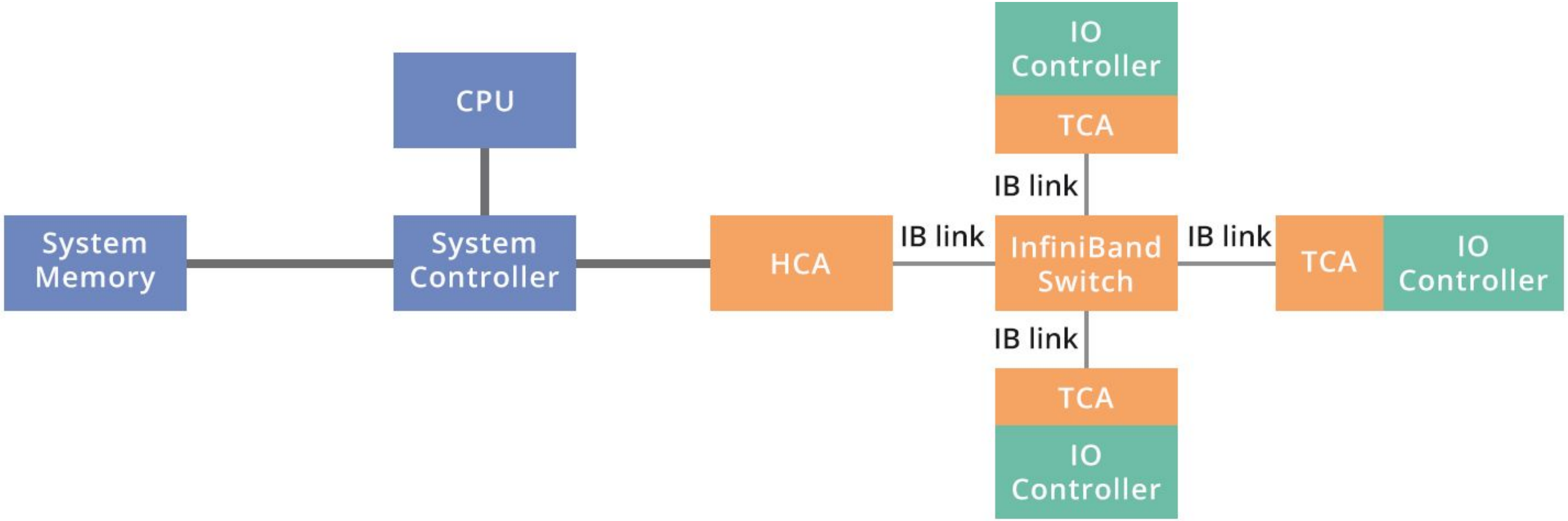
- Shared Bus Architecture vs. Switched Fabric





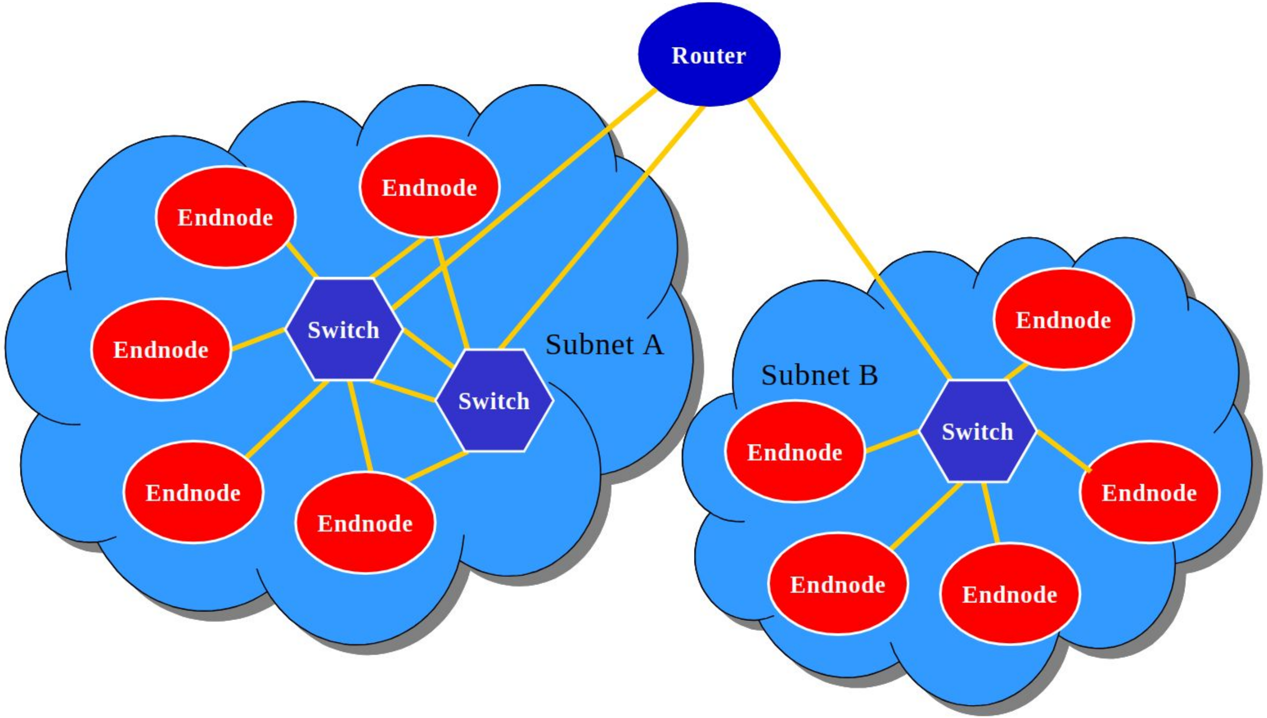
InfiniBand Architecture (cont.)

- HCA (connects host to InfiniBand)
- TCA (connects I/O controller to InfiniBand)

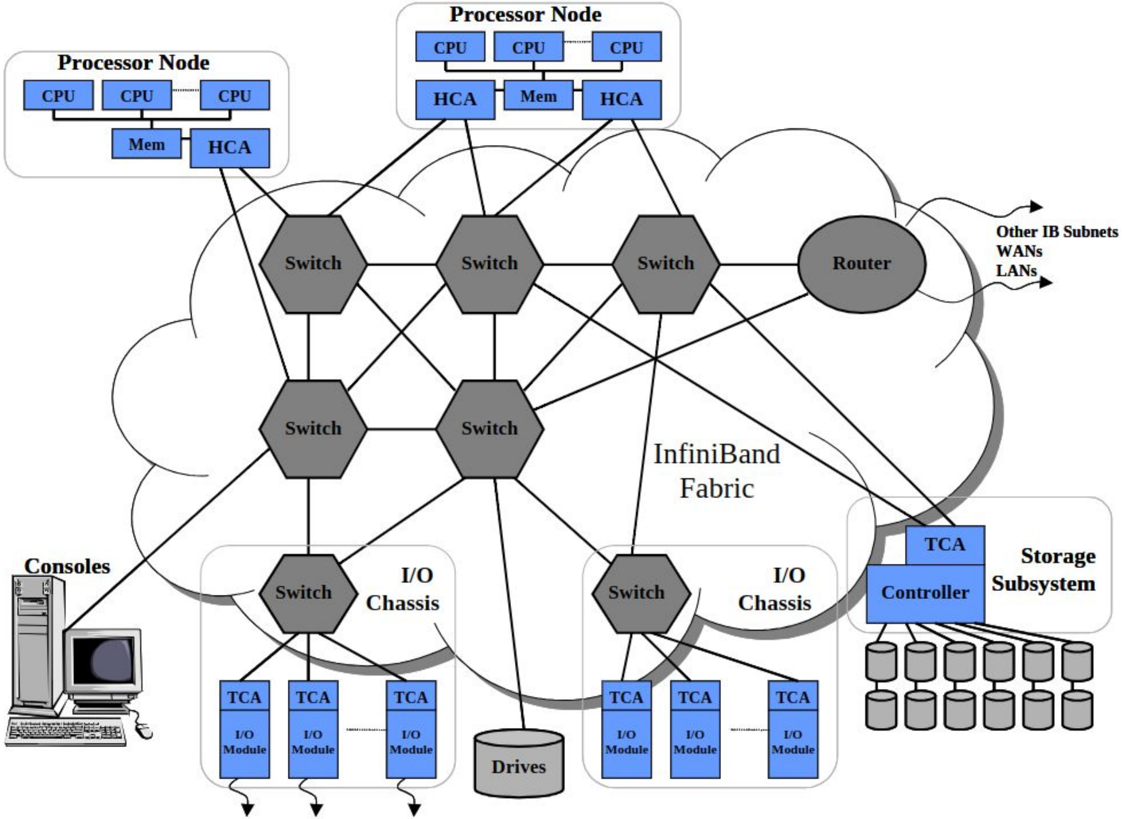


InfiniBand Architecture (cont.)

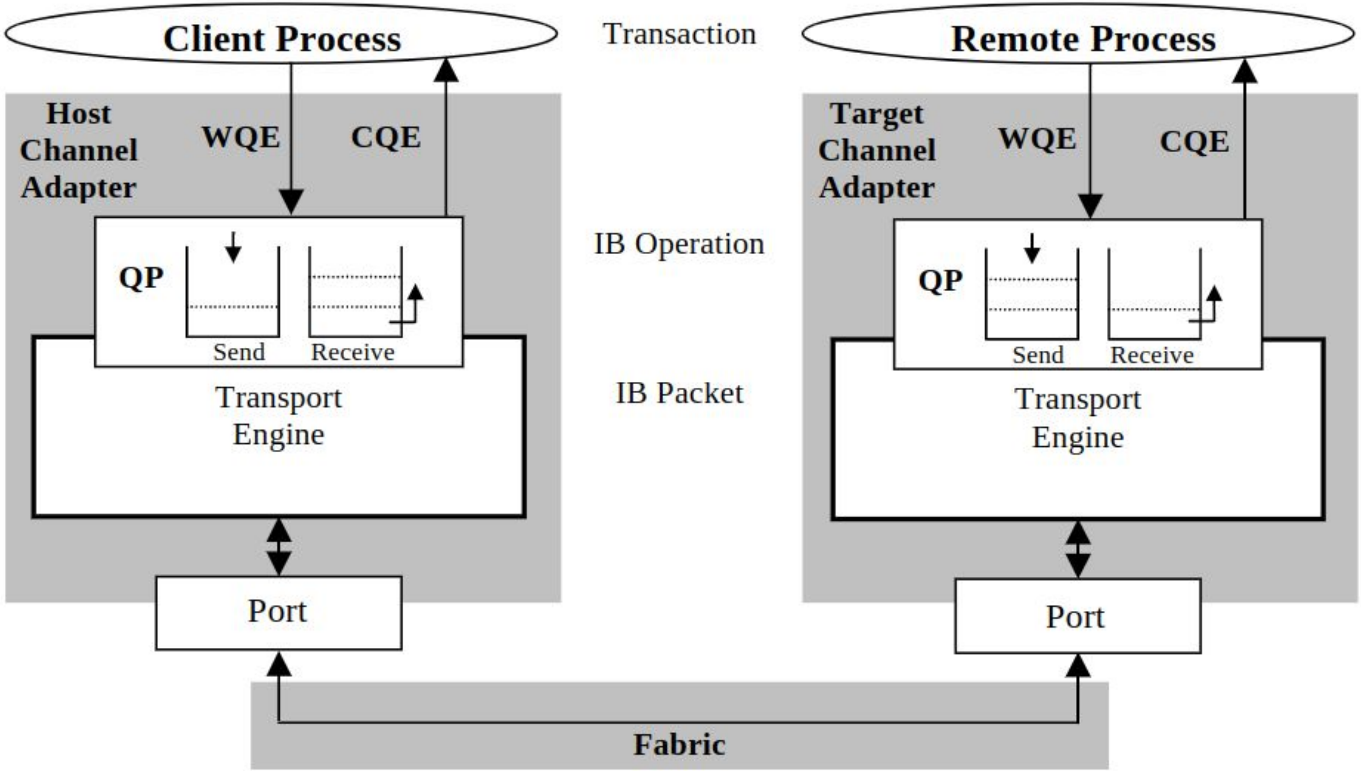
- HCA
- TCA
- Switch
- Router



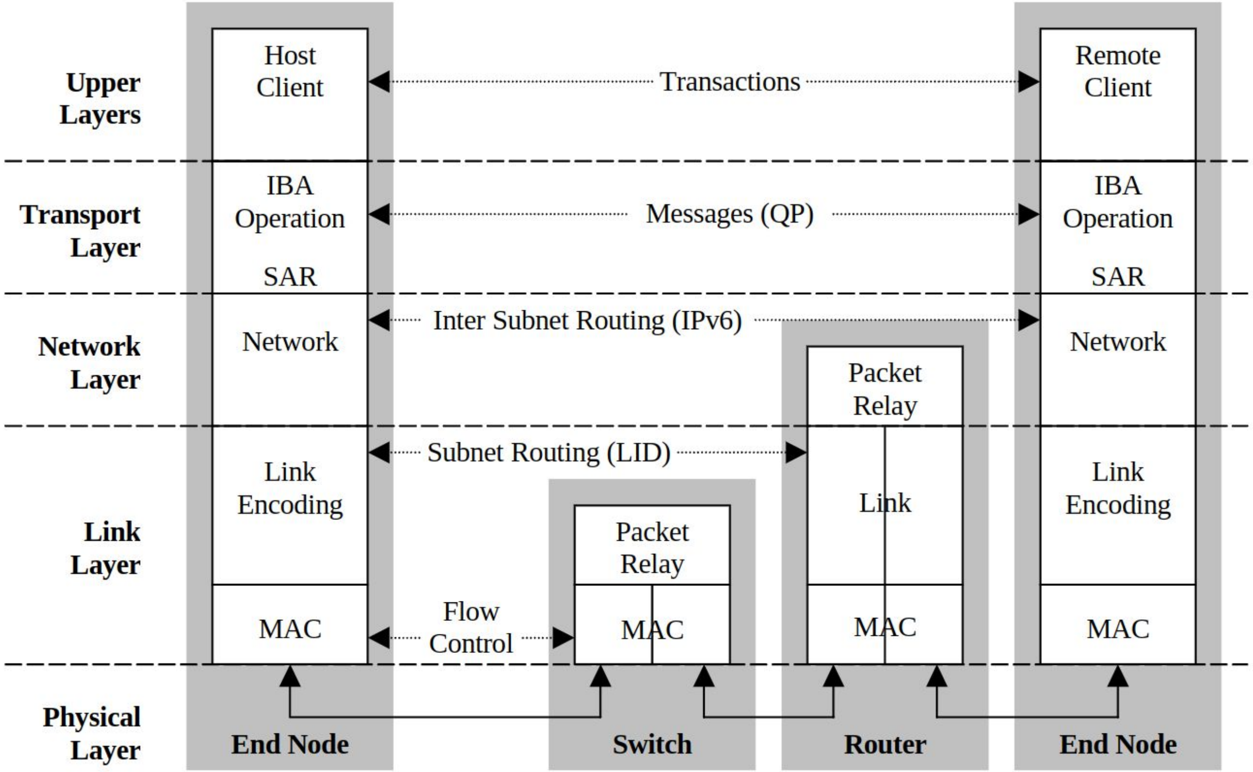
InfiniBand Architecture (cont.)



InfiniBand Architecture (cont.)

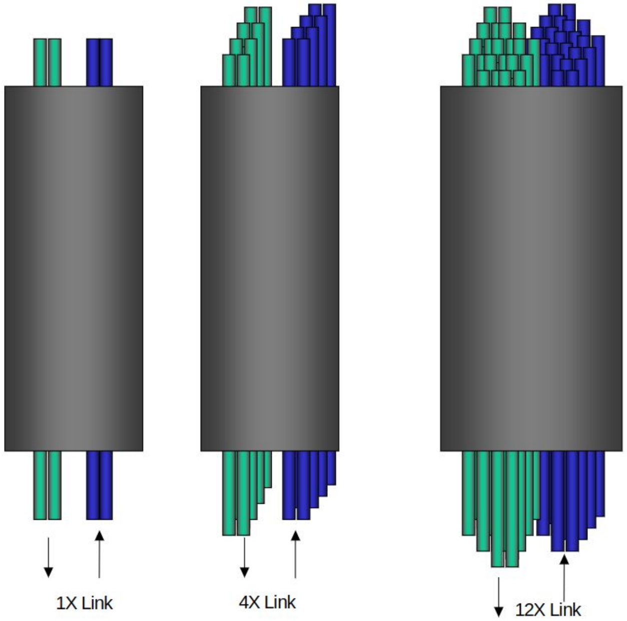


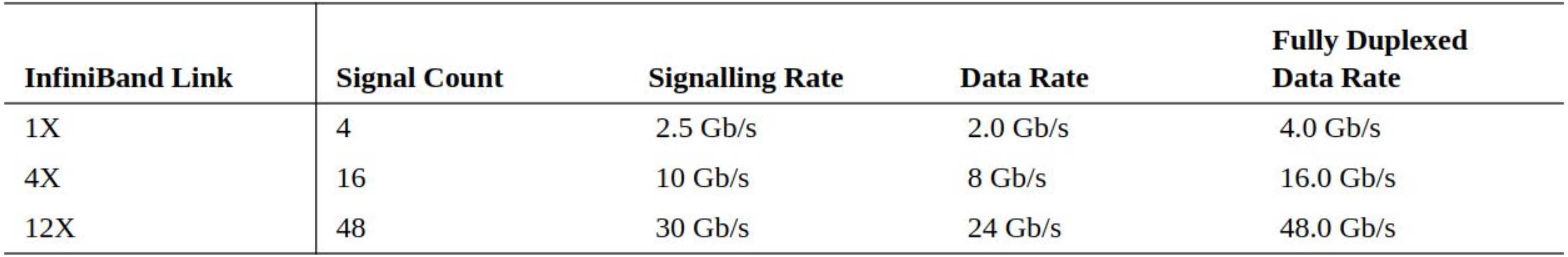
InfiniBand Protocol Stack



InfiniBand Protocol Stack (cont.)

- **Physical Layer:** InfiniBand defines three link speeds at the physical layer, 1X, 4X, 12X. Each individual link is a four wire serial differential connection (two wires in each direction) that provide a full duplex connection at 2.5 Gb/s.





				Fully Duplexed
InfiniBand Link	**Signal Count**	**Signalling Rate**	**Data Rate**	**Data Rate**
1X	4	2.5 Gb/s	2.0 Gb/s	4.0 Gb/s
4X	16	10 Gb/s	8 Gb/s	16.0 Gb/s
12X	48	30 Gb/s	24 Gb/s	48.0 Gb/s

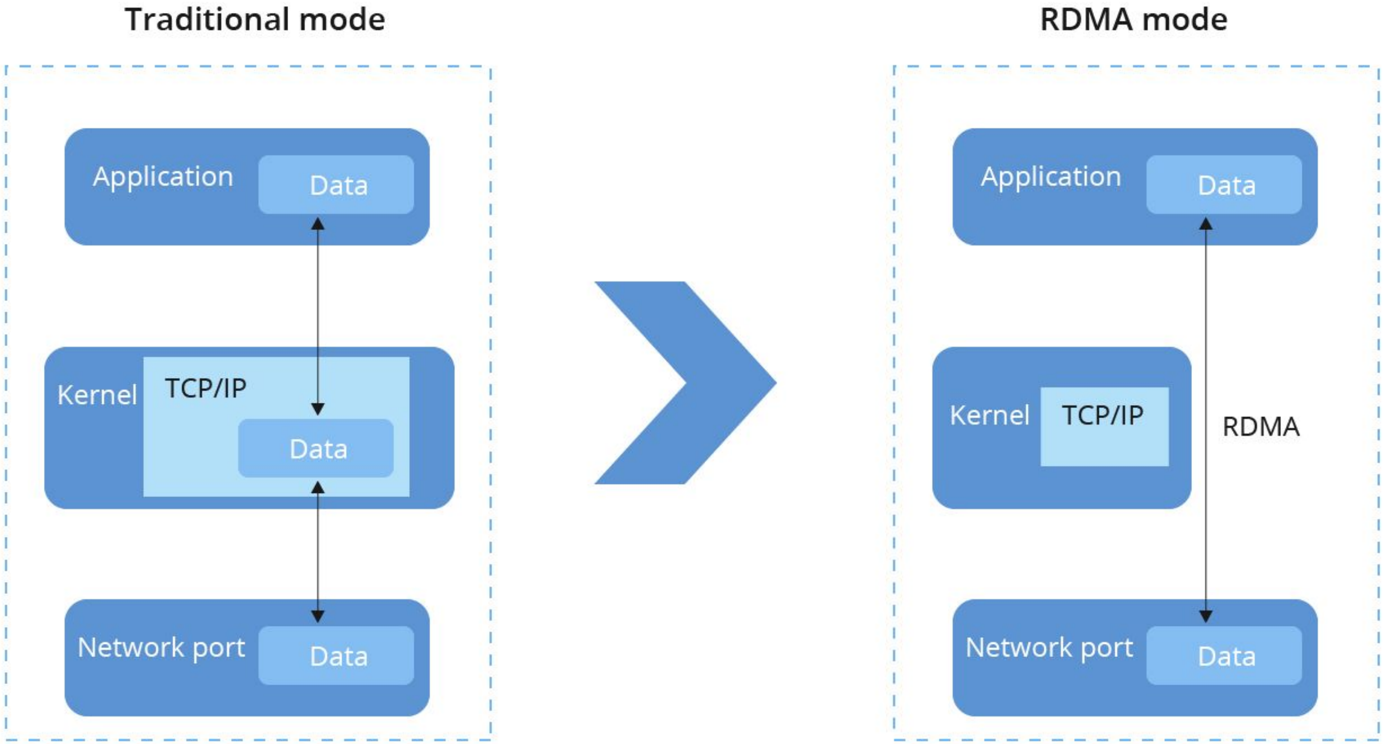
InfiniBand Protocol Stack (cont.)

- **Link Layer:**
  - Packets:
    - Type 1) Management Packets
    - Type 2) Data Packets (up to 4K bytes)
  - Switching:
    - All devices within a subnet have a 16 bit Local ID (LID).
    - All packets sent within a subnet use the LID for addressing.
  - Flow Control:
    - Credit-based
  - Data Integrity:
    - Two CRCs per packet: (1) Variant CRC and (2) Invariant CRC.

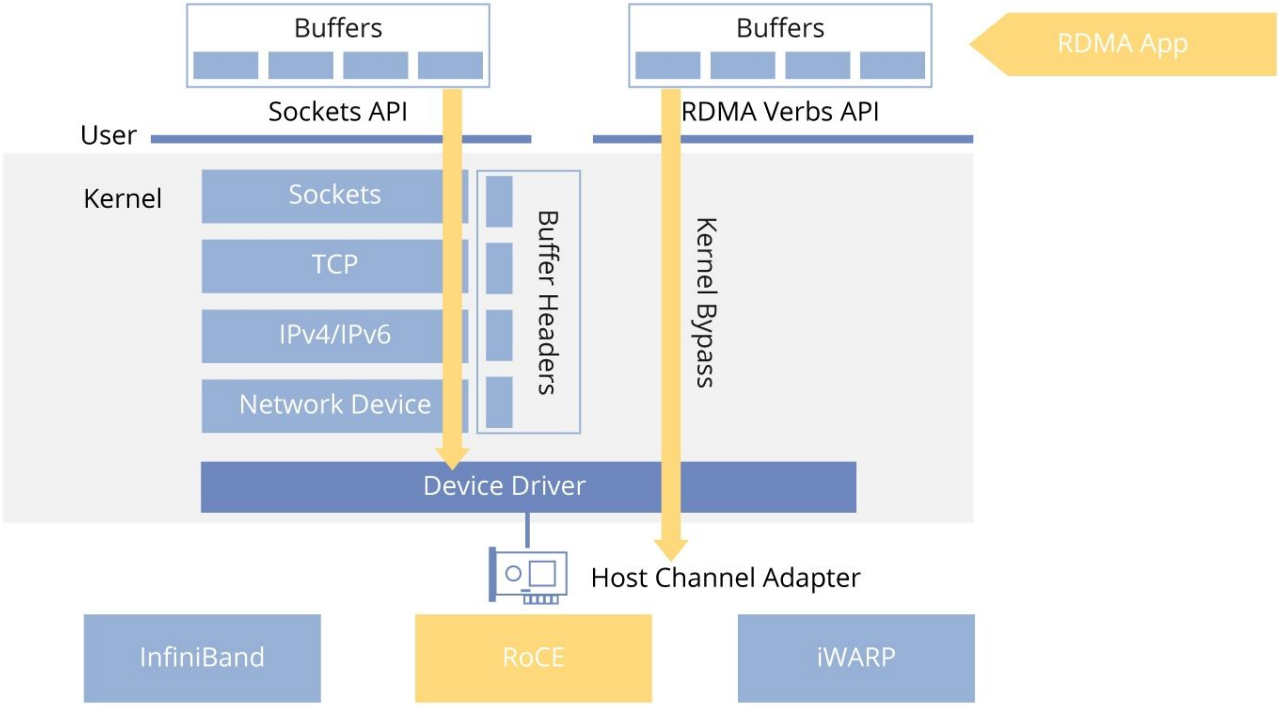
InfiniBand Protocol Stack (cont.)

- **Network Layer** (between subnets):
  - Packets that are sent between subnets contain a Global Route Header (GRH) which is a 128 bit **IPv6 address**.
  - The packets are forwarded between subnets through a router based on each device’s 64 bit **globally unique ID (GUID)**.
  - The router modifies the LRH with the proper local address within each subnet.
  - Therefore the last router in the path replaces the LID in the LRH with the LID of the destination port.
- **Transport Layer:** Based on the **Maximum Transfer Unit (MTU)** of the path, the transport layer divides the data into packets of the proper size and the receiver reassembles the packets.
  - Note: In transport layer of InfiniBand, all functions are implemented in hardware.

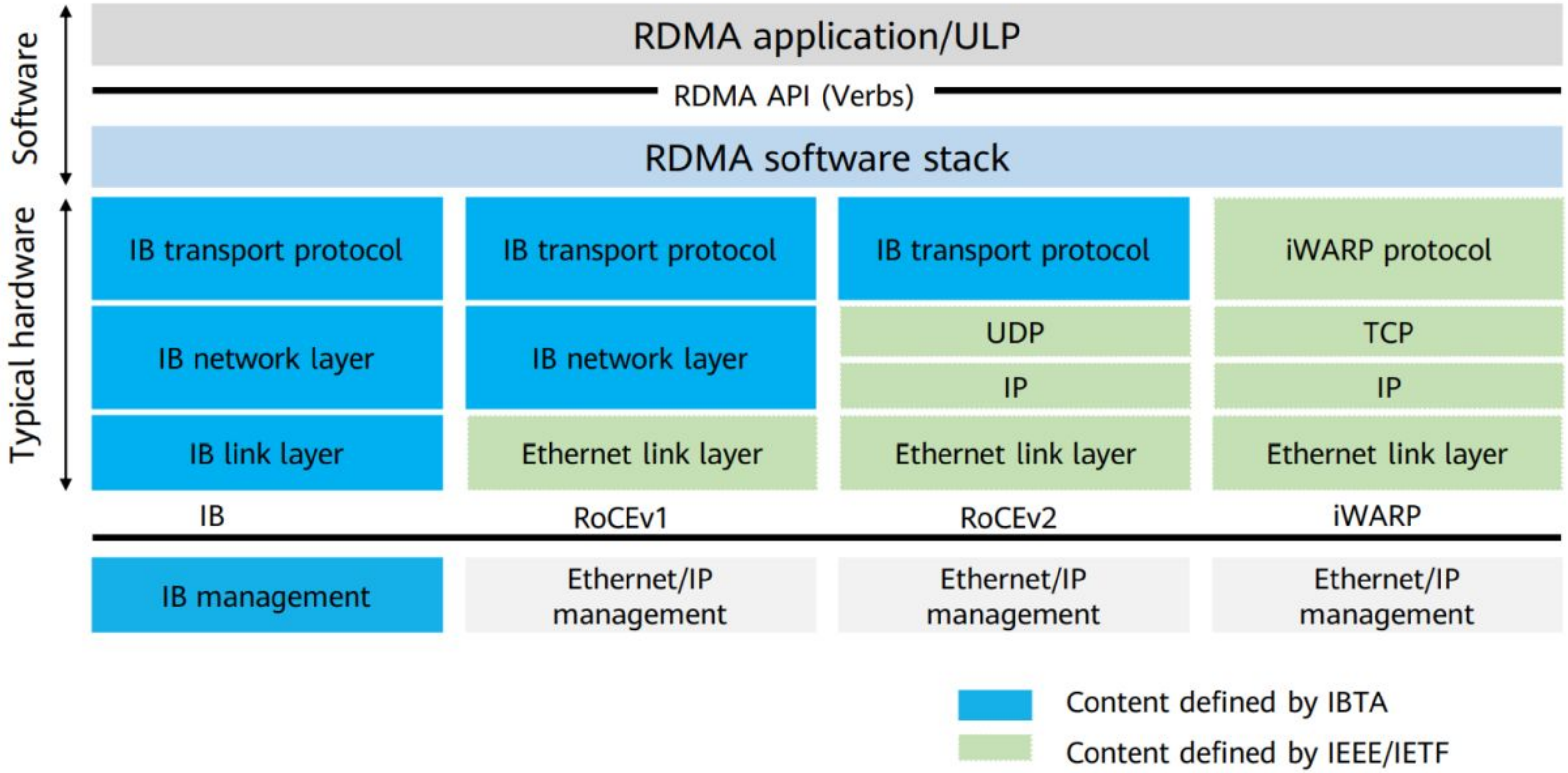
Remote Direct Memory Access



Remote Direct Memory Access (cont.)



Remote Direct Memory Access (cont.)



Software

RDMA application/ULP

RDMA API (Verbs)

RDMA software stack

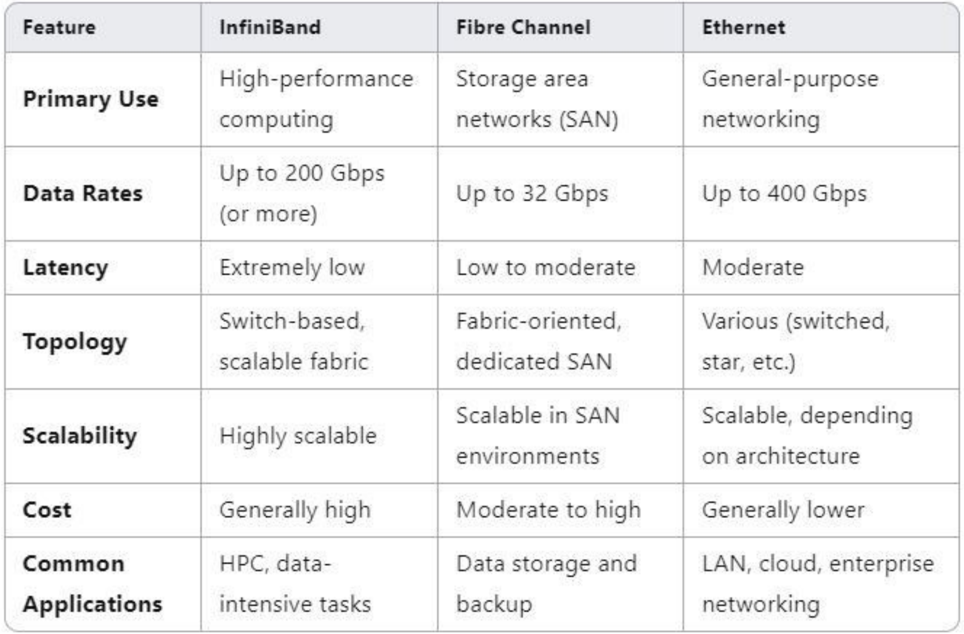
Typical hardware

IB	RoCEv1	RoCEv2	iWARP
IB transport protocol	IB transport protocol	IB transport protocol	iWARP protocol
IB network layer	IB network layer	UDP IP	TCP IP
IB link layer	Ethernet link layer	Ethernet link layer	Ethernet link layer
IB management	Ethernet/IP management	Ethernet/IP management	Ethernet/IP management

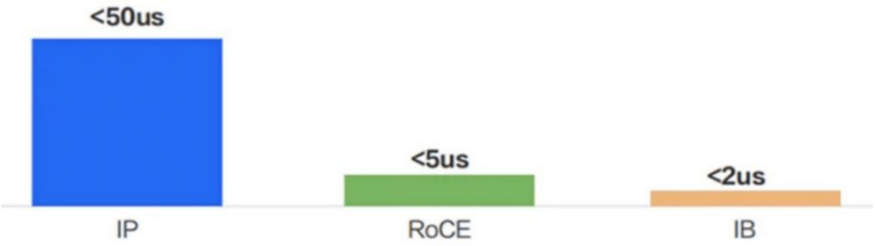
🟦 Content defined by IBTA

🟩 Content defined by IEEE/IETF

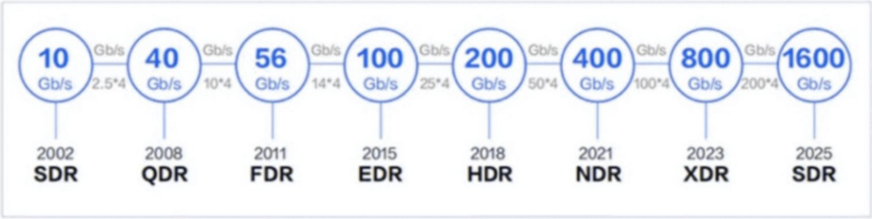
Comparison to Ethernet



Feature	InfiniBand	Fibre Channel	Ethernet
Primary Use	High-performance computing	Storage area networks (SAN)	General-purpose networking
Data Rates	Up to 200 Gbps (or more)	Up to 32 Gbps	Up to 400 Gbps
Latency	Extremely low	Low to moderate	Moderate
Topology	Switch-based, scalable fabric	Fabric-oriented, dedicated SAN	Various (switched, star, etc.)
Scalability	Highly scalable	Scalable in SAN environments	Scalable, depending on architecture
Cost	Generally high	Moderate to high	Generally lower
Common Applications	HPC, data-intensive tasks	Data storage and backup	LAN, cloud, enterprise networking



End-to-end communication latency of different technologies



Resources

- Nvidia white paper: https://network.nvidia.com/pdf/whitepapers/IB_Intro_WP_190.pdf
- https://www.fs.com/blog/infiniband-what-exactly-is-it-7714.html
- InfiniBand Trade Association website: https://www.infinibandta.org/
- Wikipedia: https://en.wikipedia.org/wiki/InfiniBand
- https://community.fs.com/encyclopedia/remote-direct-memory-access-rdma.html
- https://www.fibermall.com/blog/how-to-choose-between-infiniband-and-roce.htm
- LinkedIn Post by Pawan Sharma: https://www.linkedin.com/pulse/infiniband-vs-fiber-channel-ethernet-pawan-sharma-9qhtc

Thank you for your attention!