

- Infiniband Architecture
 - Overview - Mellanox white paper
 - (this is IB 1.1)

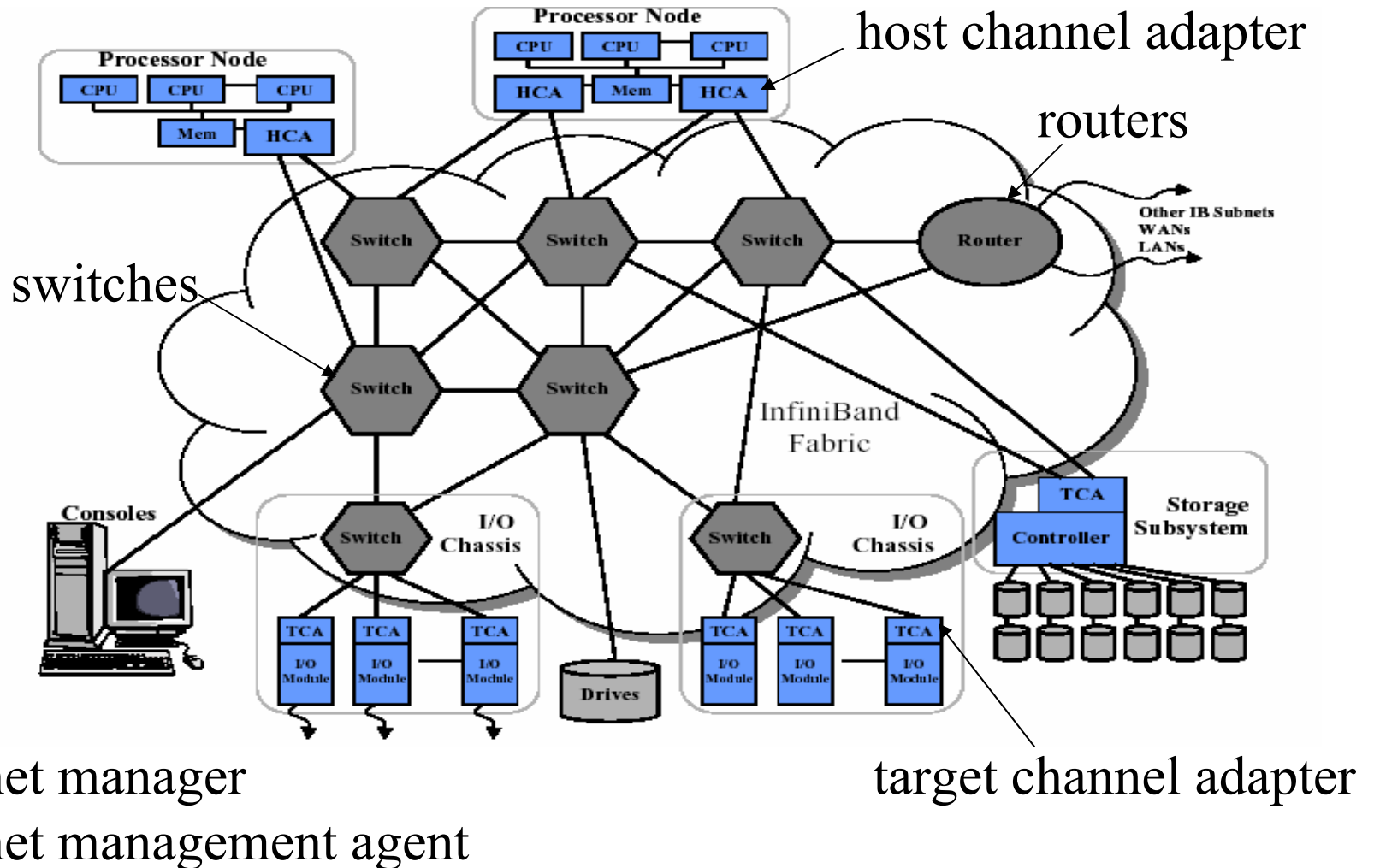
Origins

- Need for solution to integrate network, storage and interprocessor communication, and provide high performance, QoS and RAS...
- Virtual Interface Architecture (VIA)
- InfiniBand Architecture (IBA)
 - Consortium of big and small companies to establish specification
 - www.openib.org
 - www.infinibandta.org

Basic features

- objective: low-latency IPC and high bandwidth I/O
- builds on top of lessons learned: source-based routing (though not exactly circuit established), OS-bypass, virtual channels, global VM, ...
- up to transport layer in hardware, address translation in hardware
- original target domain: SANs
- today – both, component-to-component, but also in data centers, clusters
 - (up to >10360 @Tokio IT, 3 of top 10)...

SAN Topology

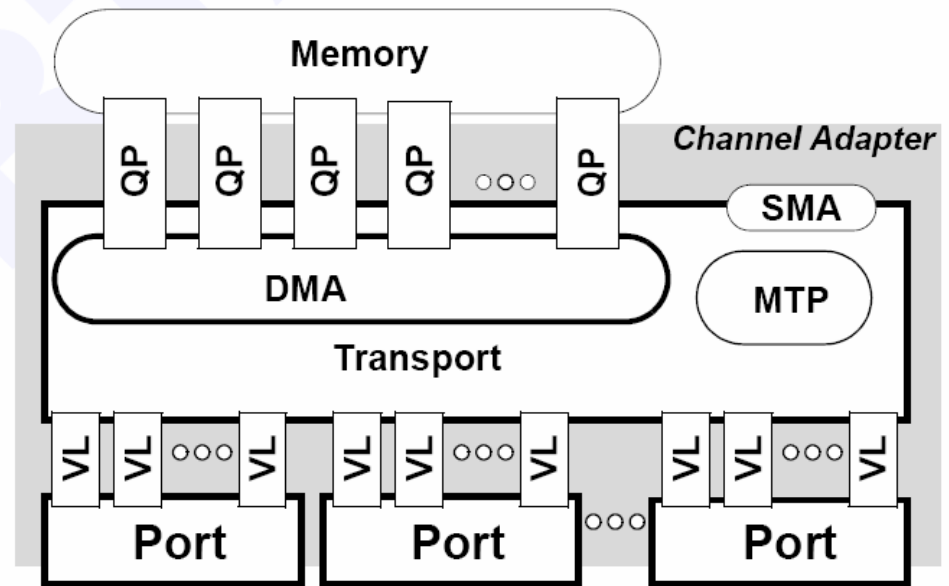


IB Link Layer

- Packets:
 - mgt or data packets
 - Local Route Header + up to 4kB transaction data
 - 2 CRCs – variant (including header) and invariant; checked on each link hop
- switching
 - within a subnet using 16b Local ID (no global ID)
 - globally using 64b Global Unique IDs (IPv6 addr)
 - routers translate to LID in new subnet
- QoS
 - up to 16 virtual links per physical link (VL15 reserved for mgt)
 - packets belong to up to 16 Service Levels
 - SL to VL mapping done by n/w mgr at each subnet
 - per VL credit-based flow control (credit = buffer size)

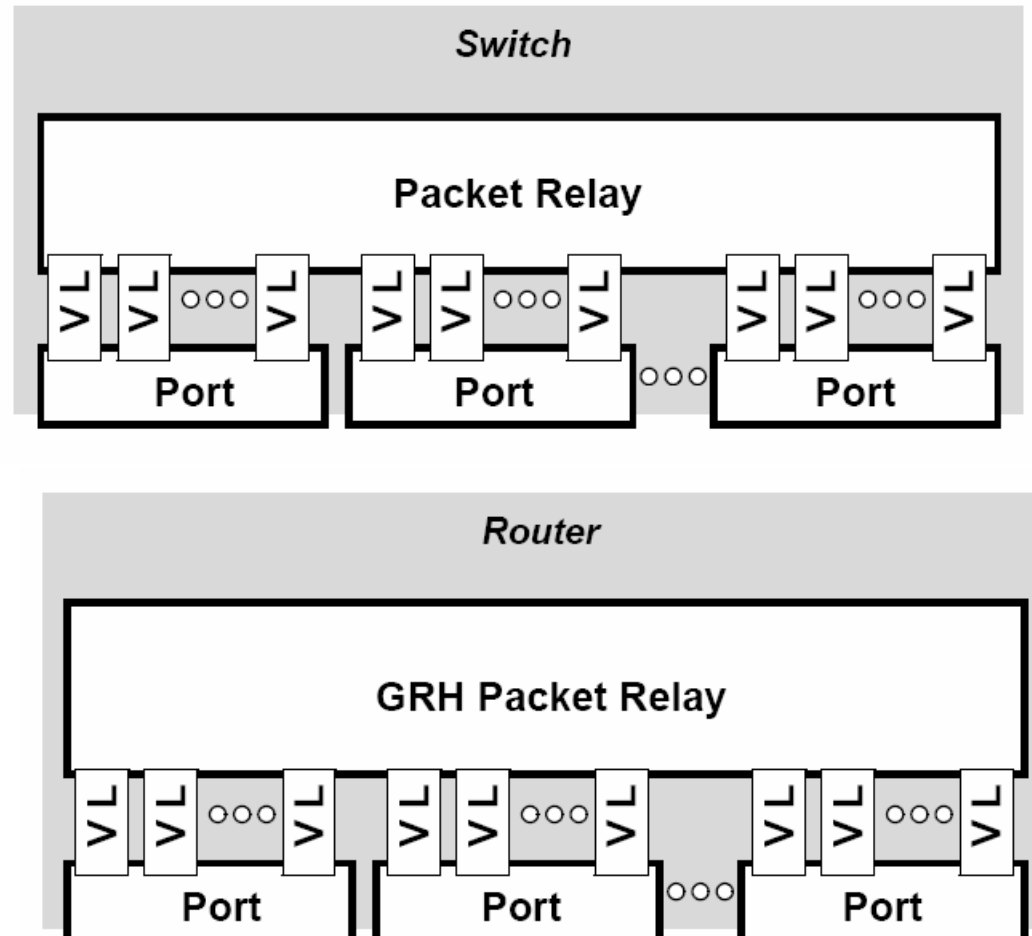
Channel Adapters

- Consume and generate IB packets
- Have programmable DMA engines, with memory protection features
- Multiple ports may exist, separate buffering per virtual lane
- Memory protection and translation tables, and transport in hardware
- Connection info in card or system memory -> MemFree cards

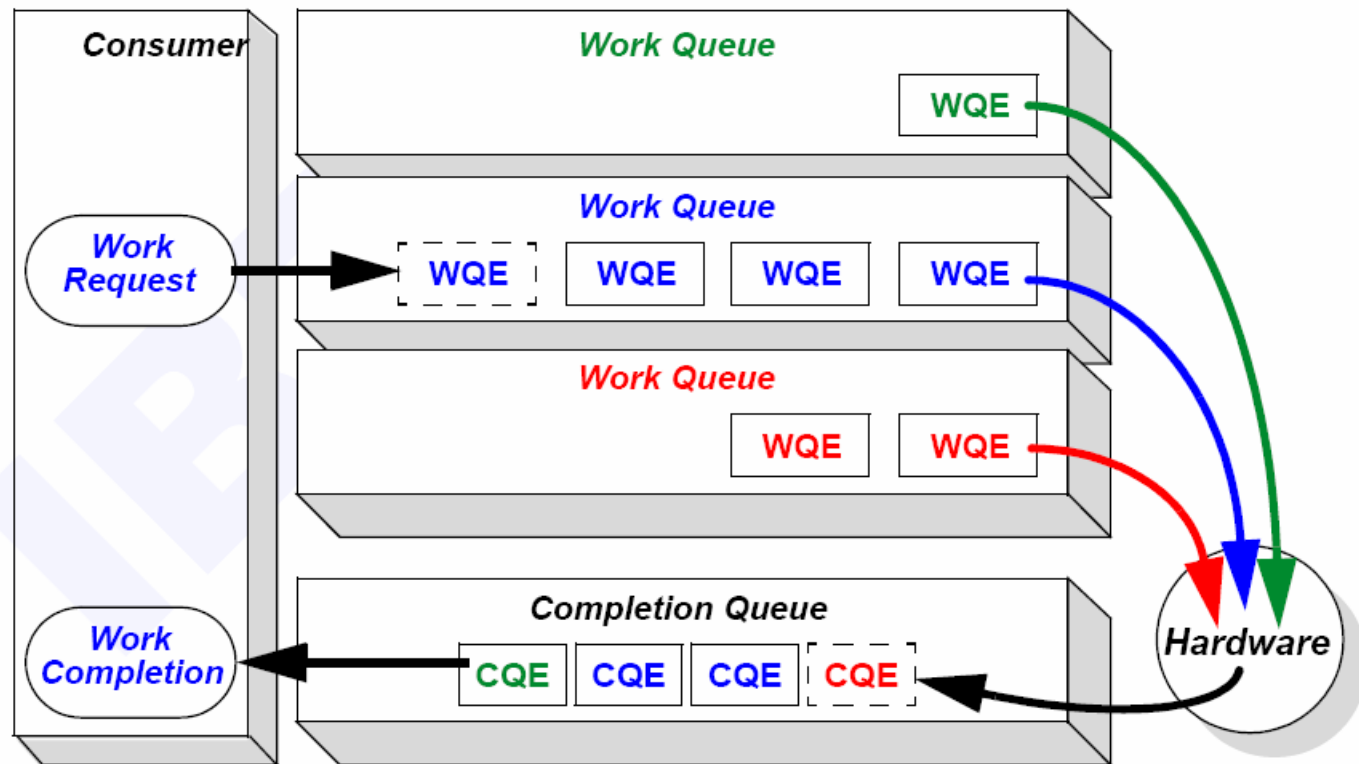


Switches and Routers

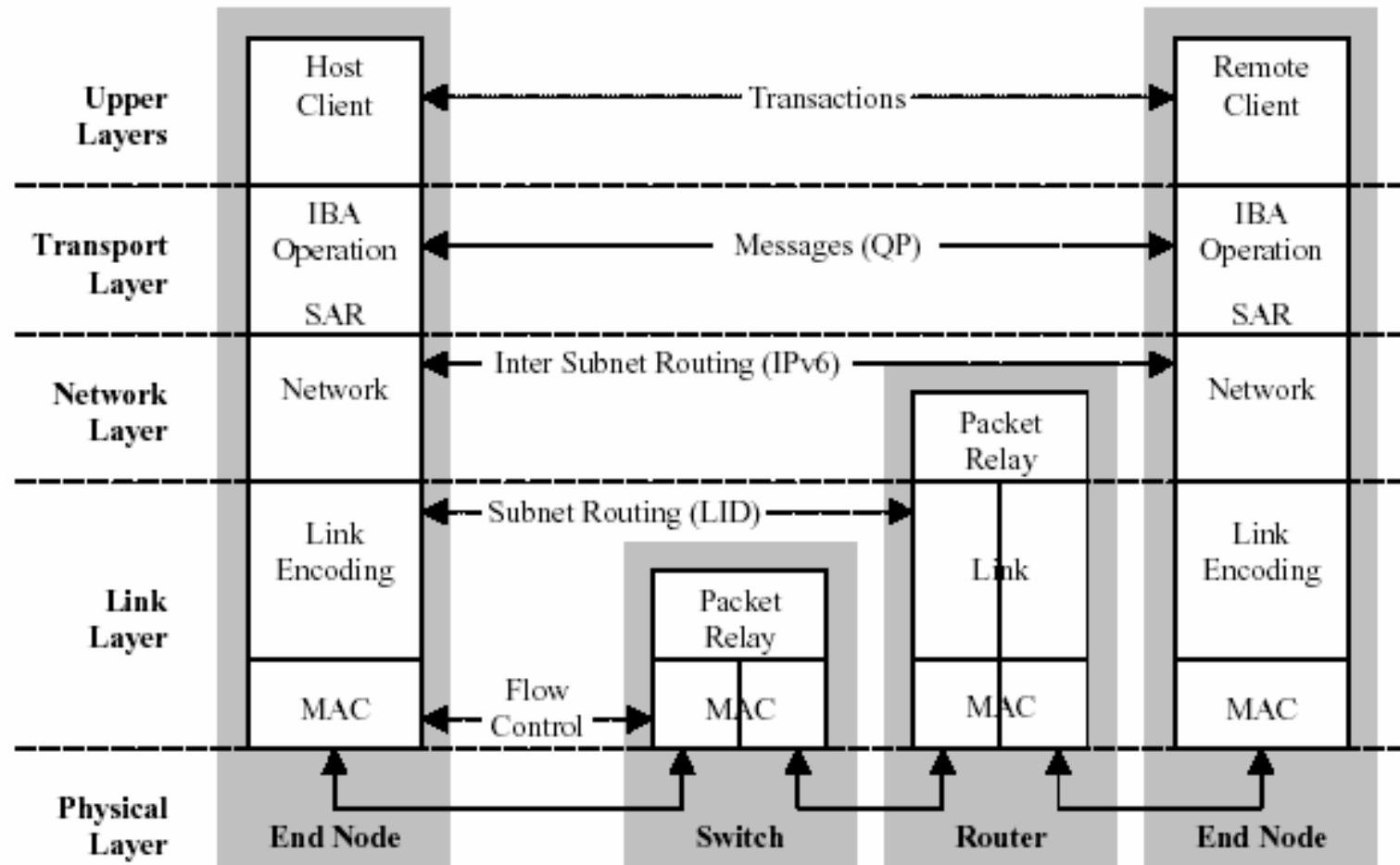
- Relay packets between ports
 - Switches: Inter-subnet
 - Routers: Intra-subnet; use global ID
 - Multicast supported
- Subnet manager responsible for configuration



IB Communications Model



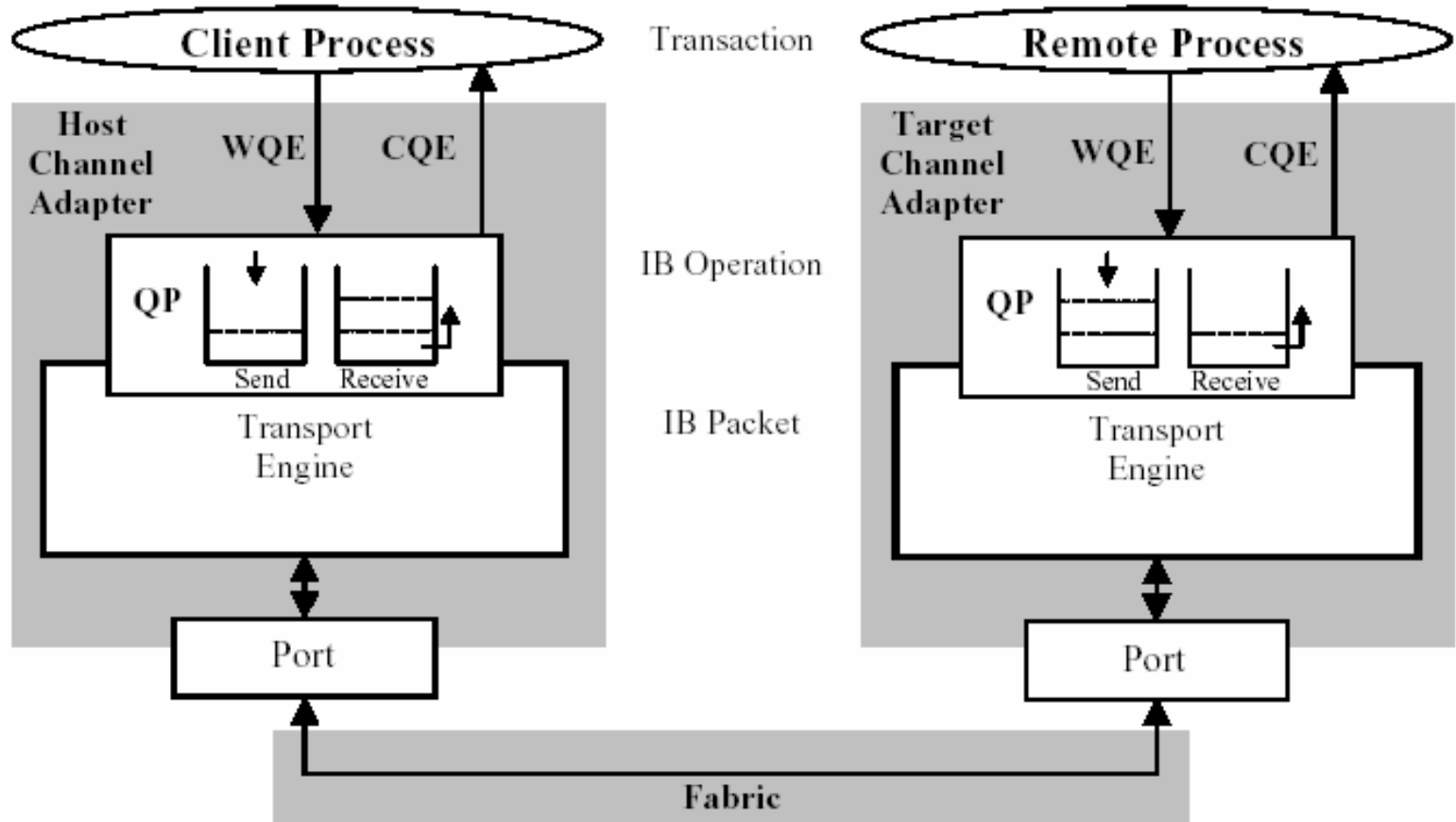
IP Protocol stack



- IB interface to above

- VIA compliant

- send/receive or remote DMA put/get



IB transport

- in order packet delivery, channel multiplexing, data segmentation/reassembly, acknowledgements for reliable/unreliable services... (uses Base Transport Header)
- all in HARDWARE!
 - controlled environment, reliable link-layer... can do offload
- queue pairs for each “connection”
- optional support for multicast
 - based on LID and GID; at-most-once guarantee and no loops

IB Service Type

Service Type	Connection Oriented	Acknowledged	Transport
Reliable Connection	yes	Yes	IBA
Unreliable Connection	yes	no	IBA
Reliable Datagram	no	Yes	IBA
Unreliable Datagram	no	no	IBA
RAW Datagram	no	no	Raw

- Raw is for legacy protocols and network stacks, dedicate QP

Physical Layer

- 1x – 4 wires, 2+2 in each direction, for total of 2.5 Gbaud per direction
 - 8b/10b encoding
 - => total 2Gbps of data in each direction
- 4x link -> 16 wires, total 10Gpbs or aggregate data rate of 16Gpbs data rate
- 12x -> 30Gbps (48Gpbs)
- fiber and copper connectors