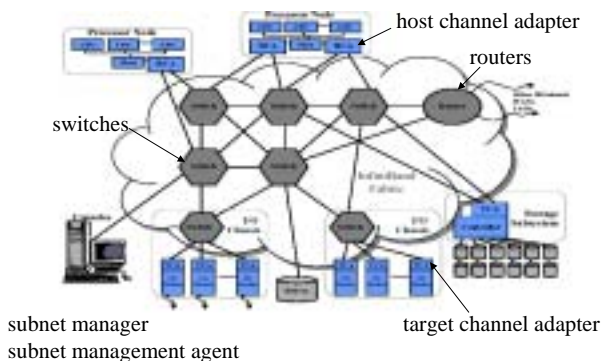


- Infiniband Architecture
 - Overview - Mellanox white paper

- Myrinet and Quadrics designed to address HPC domain, parallel/MPI applications...
 - OS bypass, NIC with protocol functionality in hardware (and software), source routing, virtual channels, global VM...
- Infiniband
 - objective: low-latency IPC and high bandwidth I/O
 - builds on top of lessons learned +
 - 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 >1100 nodes @VT)...

SAN Topology



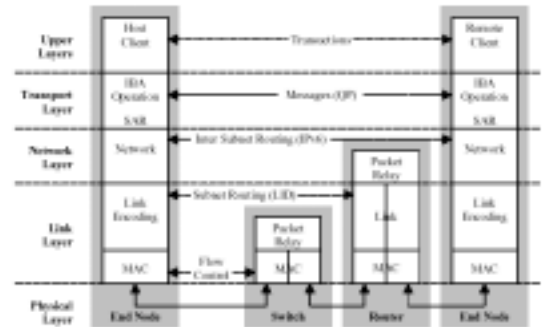
IB Link Layer

- 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)

IB transport

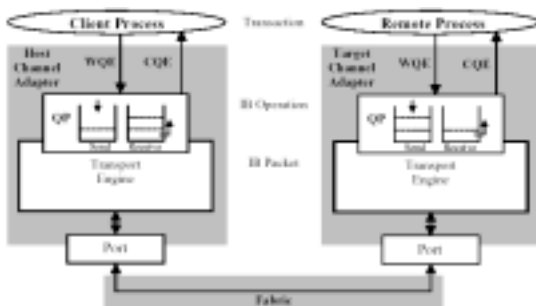
- in order packet delivery, channel multiplexing, data segmentation/reassembly, acknowledgements for reliable/unreliable services... (uses Base Transport Header)
- all in **HARDWARE!**
 - controlled environment, reliable link-layer... can do offload
- queue pairs for each “connection”
- optional support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

IP Protocol stack



• IB interface to above

- VIA compliant
- send/receive or remote DMA put/get



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 10Gbps or aggregate data rate of 16Gbps
- 12x -> 30Gbps (48Gbps)
- fiber and copper connectors

