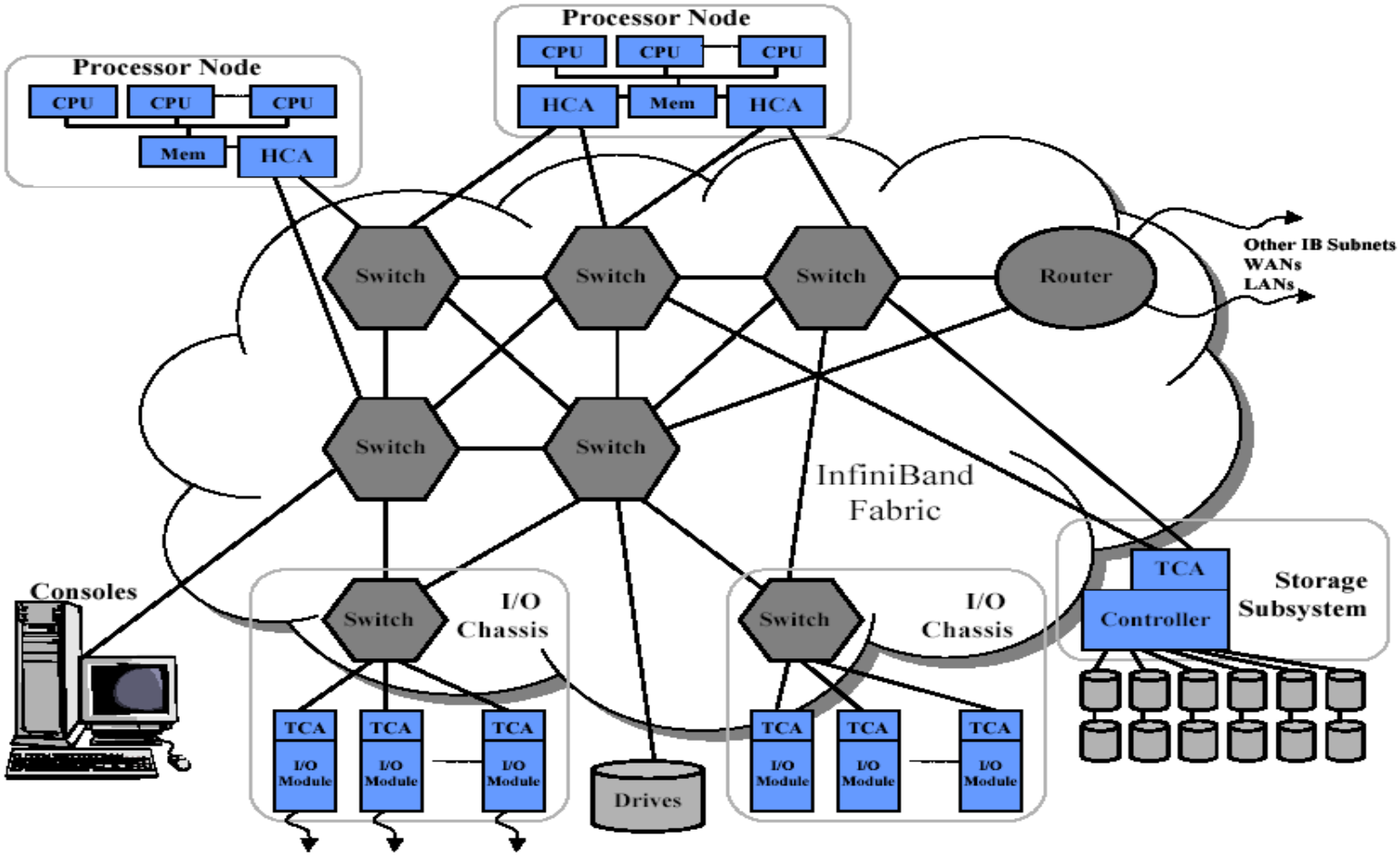
INFINIBAND OVERVIEW

Why InfiniBand?

- Open and comprehensive standard with broad vendor support
- Standard defined by the InfiniBand Trade Association (Sun was a founder member, along with Intel, HP, Compaq and Microsoft)
  - Intel made initial 2.5Gbps prototypes, but then concentrated on PCI-E which borrows substantially from the IBA.
- Optimised for low latency/high bandwidth
- Dual port 40Gbps PCI-e
- Low CPU overhead – only non-proprietary solution for RDMA
- Current standard is IBA 1.2, 900 page specification

  http://www.infinibandta.org

  Annual interop event to test and prove interworking



What is InfiniBand

InfiniBand is a I/O protocol designed to provide high bandwidth, low-latency interconnect for clustering.

It has been designed to offload CPU overhead by incorporating powerful RDMA (Remote Direct Memory Access) engines.

API's have been defined which allow transfer directly from User space, bypassing kernel memory copies associated with traditional protocols.

IB Switch functionality allows cut-through packets at wire speed with link and end-to-end data integrity

Implicit trunking across multiple serial lanes provides protocol independent speed improvements

1x = 2.5 Gbps 10B/8B = 2.0 Gbps data

4x = 10 Gbps == 1G Byte/second data transfer

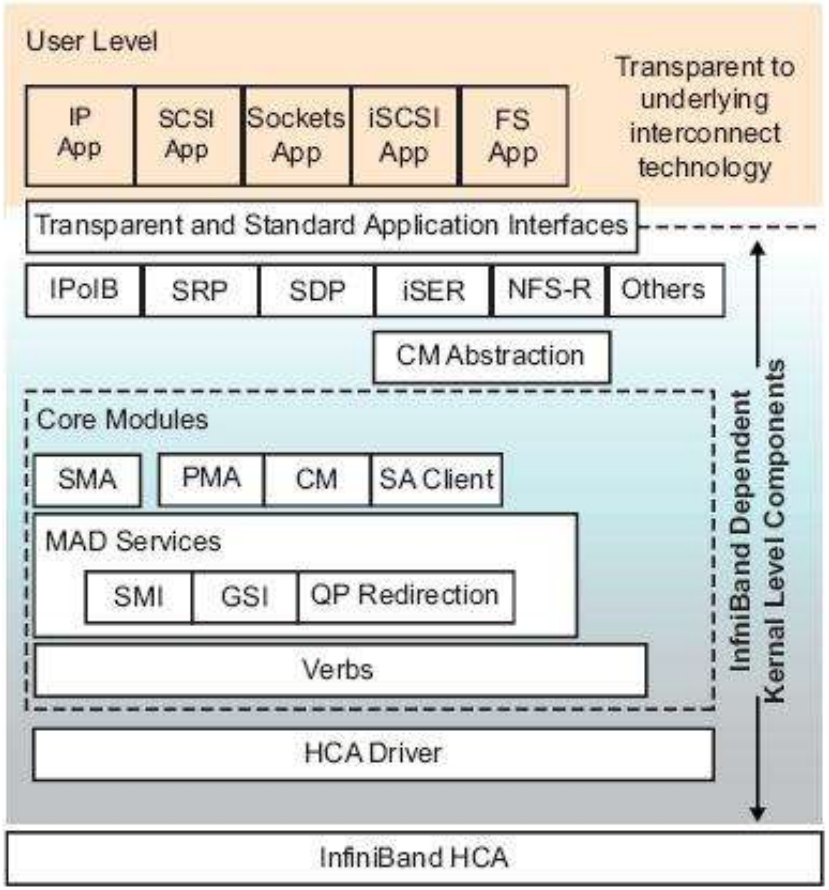
12x = 30 Gbps == 3G Byte/second data transfer

Double Data Rate = 5 Gbps

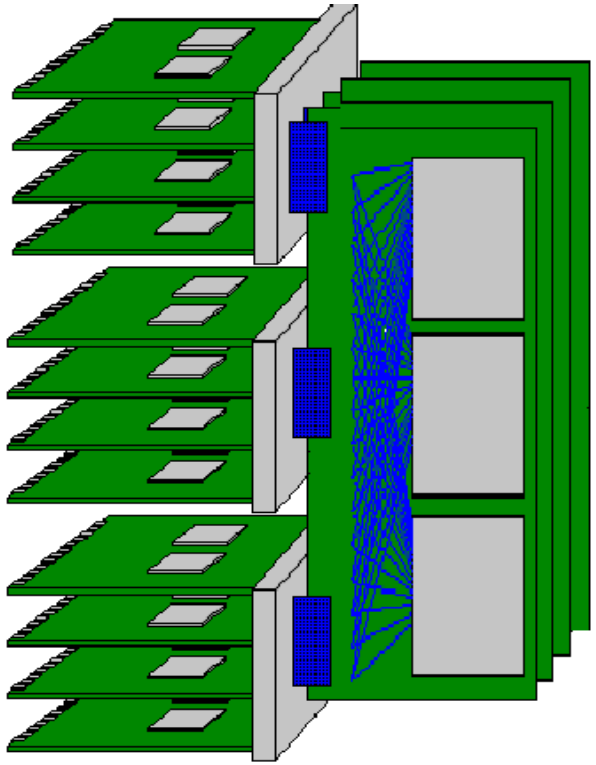
4xDDR = 20 Gbps == 2G Byte/second data transfer

Quad Data Rate = 10 Gbps

4xQDR == 4G Byte/second transfer

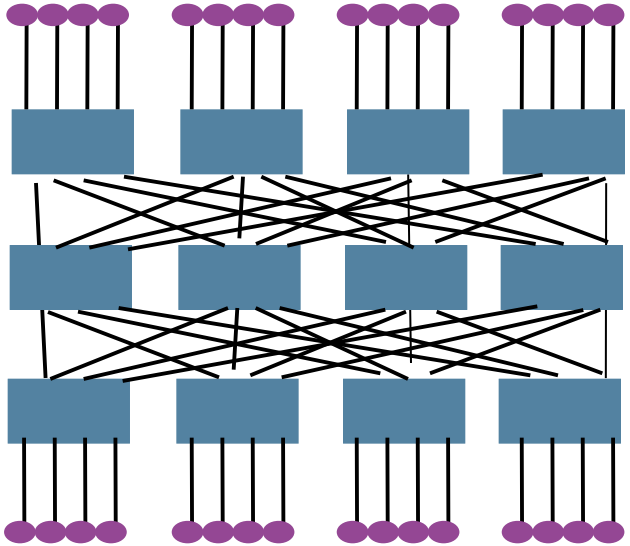


- Port densities of 24 - 3052
- Mellanox silicon
  - InfiniScale-IV chips supports 40 Gbps per port (QDR), 36-ports.
    - IS5600 director switch supports 648 fully non-blocking ports, 51.8T bps bisectional switch B/W
- Qlogic silicon
  - TrueScale supports 40 Gbps per port (QDR), 36-port QDR chip
    - 12800 Director switch supports 648 fully non-blocking, 51.8T bp/s bisectional switch B/W



IB Switch vendors – Mellanox, Qlogic, Voltaire

- Clos configuration with constant bi-sectional bandwidth
- Switch port count for a 3 tier ASIC topology = $n^2/2$
- Mellanox InfiniScale III is 24 port.
  - 24, 144, 288, 3052 port switches
- InfiniScale IV is 36 port
  - 36, 324, 648 port switches



e.g a 8-port ASIC

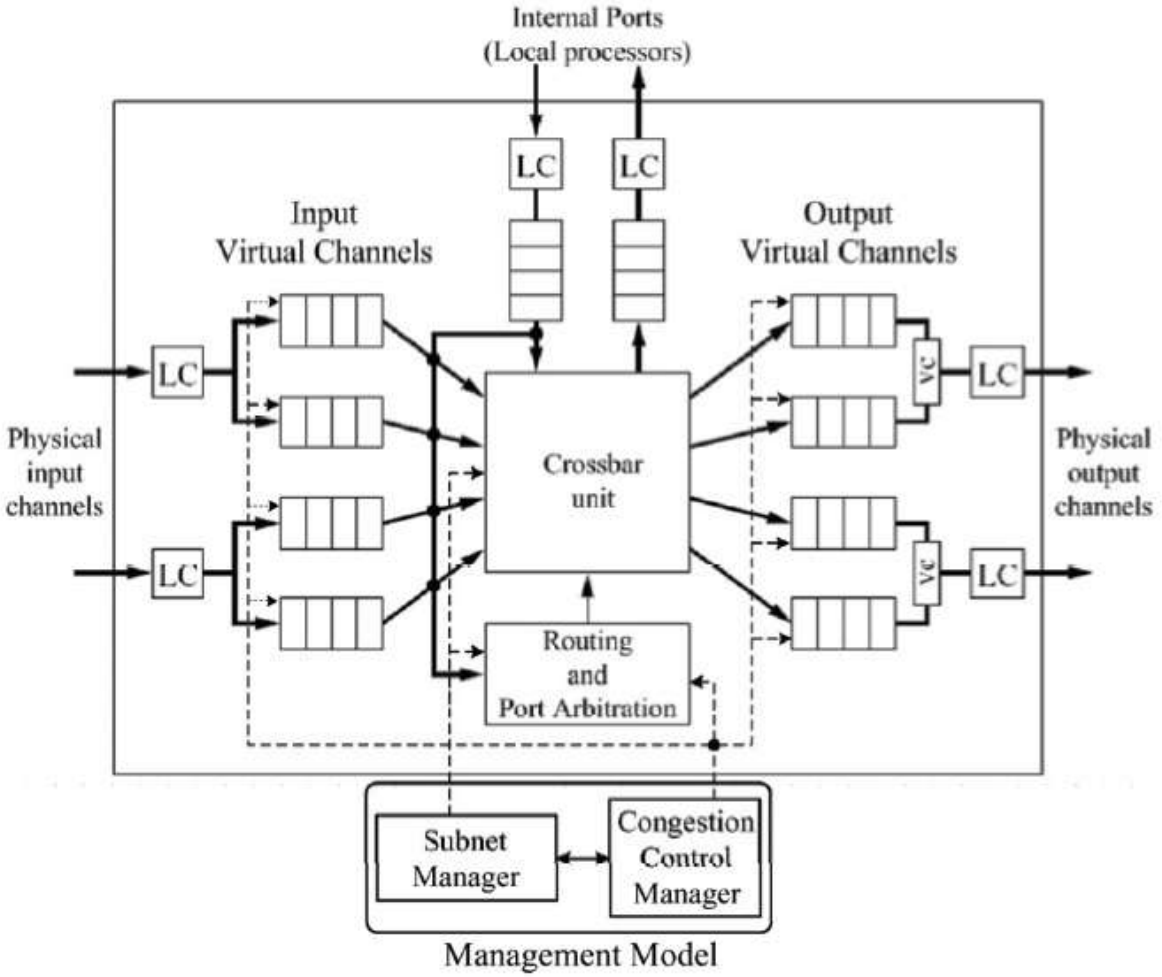
creates a 32 port Clos switch

Full non-blocking

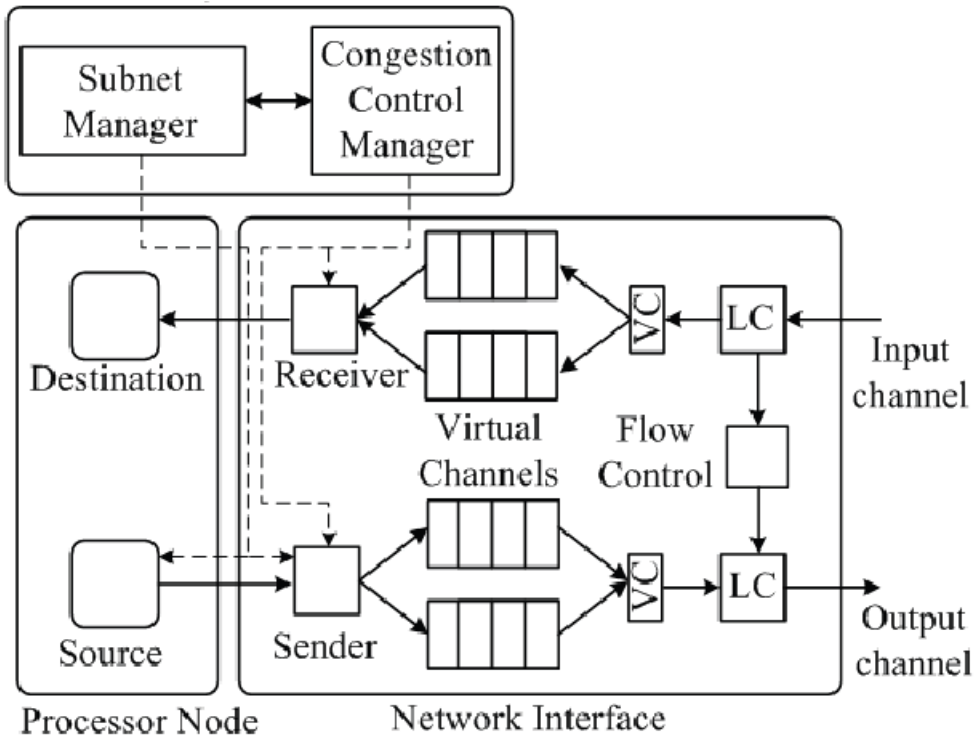
Any port to any port

max. of 2 hops between any 2 ports

- Crossbar unit provides full connectivity between ports
- Routing based on forwarding table. Forwarding to adjacent ports managed by the Link Controller (LC)
- LC buffers when then the destination port is busy and then sends based on VL priority
- Switch GUID so that the SM is aware that a collection of ports are within one unit
- Congestion Manager provides detection and notification of congestion

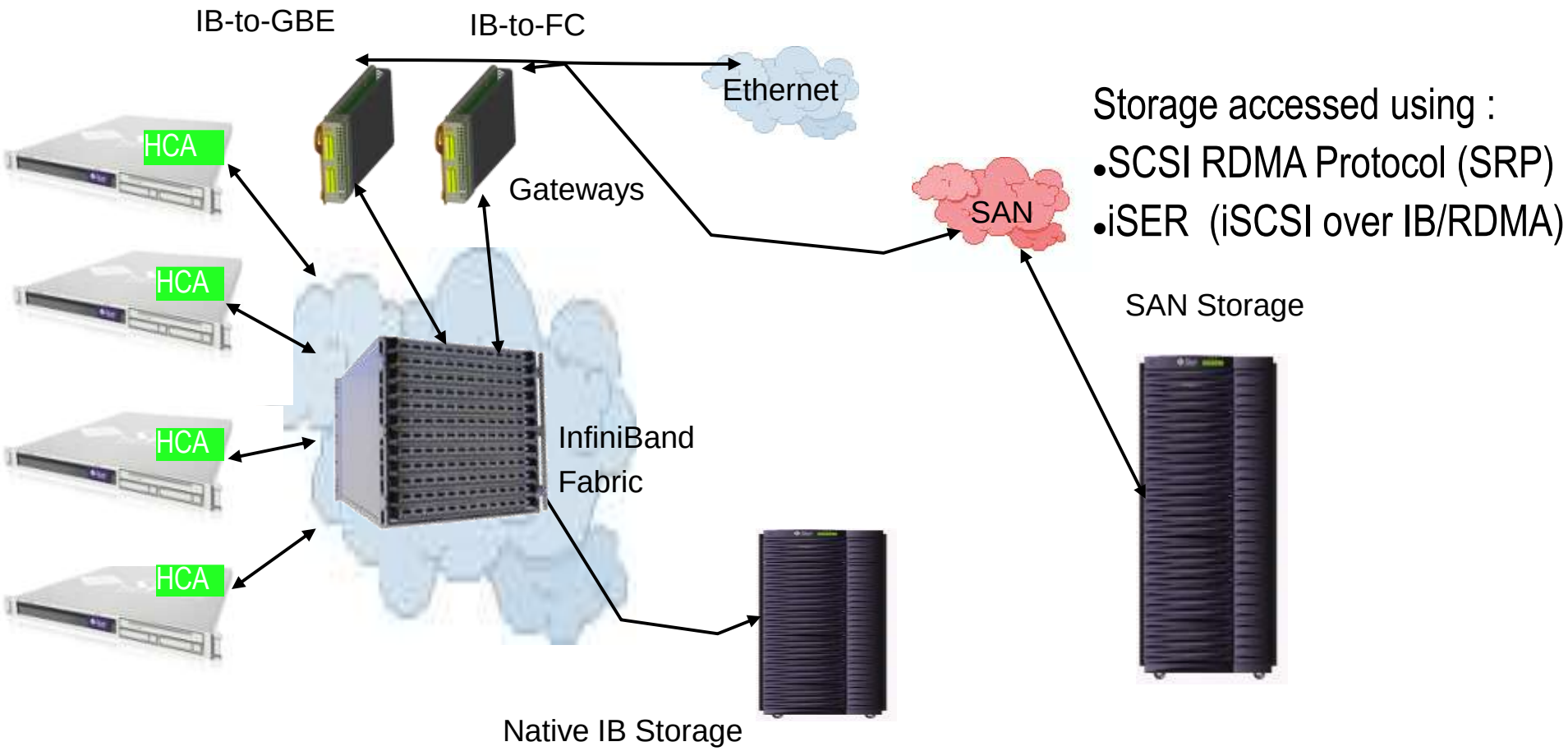


- Applications and drivers provide source and destination of packets
- Sender splits packets based on MTU size
- Packets are numbers to detect out of order delivery caused by adaptive routing schemes
- Virtual channels to manage priority in both directions
- Congestion Agent steps back transmit rate if notified and initiates notification on receiver if overloaded



Provides a Universal I/O channel capable of supporting multiple protocols

IP over InfiniBand (IPoIB) for compatibility and Ethernet gateways for connectivity



Storage accessed using :

- SCSI RDMA Protocol (SRP)
- iSER (iSCSI over IB/RDMA)

- 3 protocol layers inside each packet provide full management
- Link level and end-to-end CRC's improve reliability

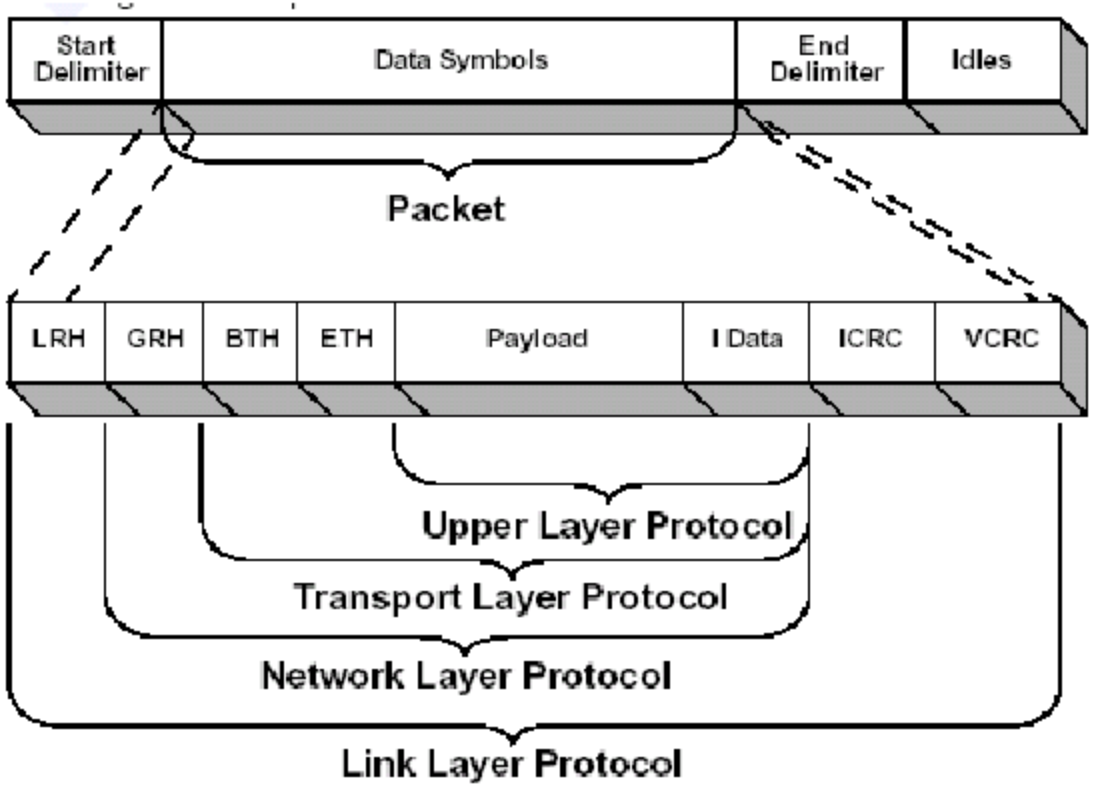


Diagram courtesy of Mellanox

Connection Type	Description	Message Size (Max)
Reliable Connection	Acknowledged-Connection Oriented	2 GB
Reliable Datagram	Acknowledged-Multiplexed	2 GB
Unreliable Connection	Unacknowledged-Connection Oriented	2 GB
Unreliable Datagram	Unacknowledged-Multiplexed	256 B-4 KB
Raw Datagram	Unacknowledged-Connectionless	256 B-4 KB

- Global Unique Identifiers (GUID) are 128 bit unique identifiers assigned by the manufacturer
- All packets in a subnet have a source and destination local identifier (LID).
- Arbitrarily assigned by SM so don't assume a value
- Ports can be assigned multiple LIDs
- Switches are typically only assigned a single LID
- LIDs are 16 bits with reserved addresses:
  - 0x0000 reserved
  - 0xFFFF permissive LID, used by SM before LIDs are assigned
  - 0x0001 – 0xBFFF , valid unicast LIDs (48K)
  - 0xC000 – 0xFFFE, valid multicast LIDs. Only valid for destination addresses (16K)
- Global Identifiers (GID) are 128-bit addresses used for routing addresses between subnets. Based on a IPv6 address format with 64-bit subnet and 64-bit port address

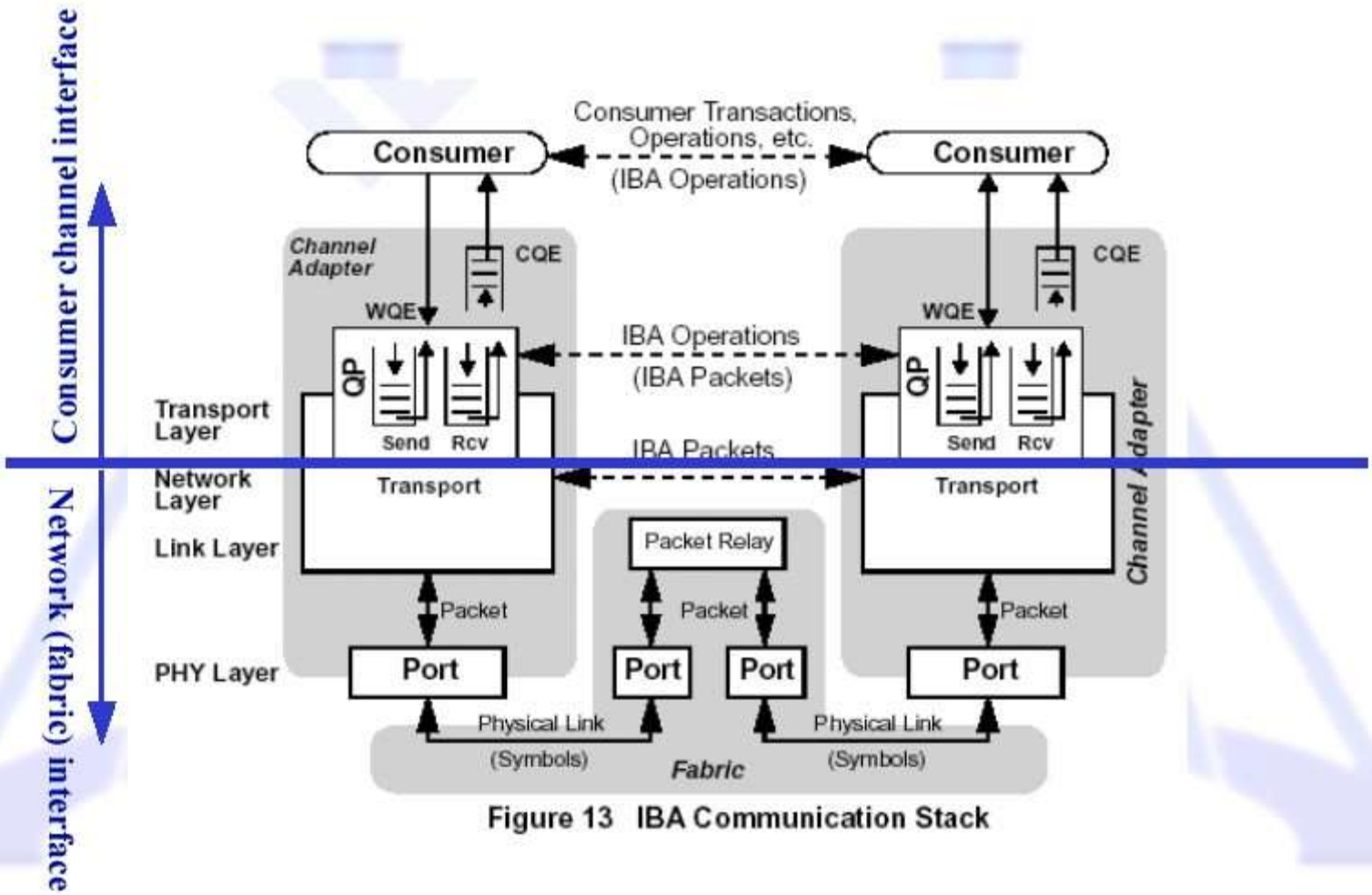


Figure 13 IBA Communication Stack

Credit based flow control at physical layer provides reliable delivery

Diagram courtesy of Mellanox

because no one approach is right all the time

Sends

- Low latency for small messages
- Saves overhead of RDMA setup
- Lowest latency
- Optimized for small messages

Remote Dynamic Memory Access

- Used for efficient, large data transfers
- Between registered memory regions connected as a Queue Pair
- Initial overhead in registering memory and establishing RDMA channel – once per session
- Highest throughput
- Optimized for large data transfers

Semi-persistent “channels” between nodes

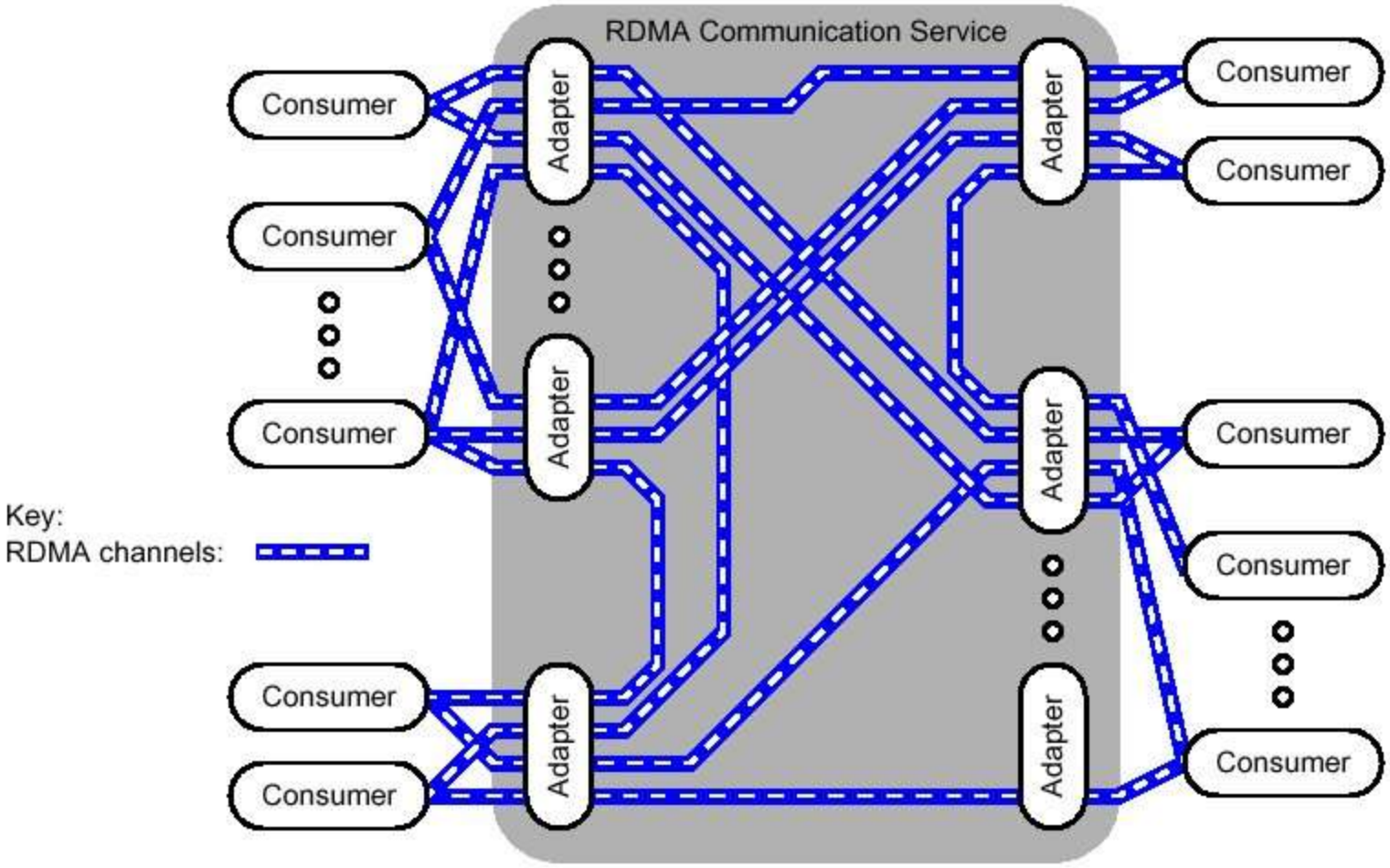
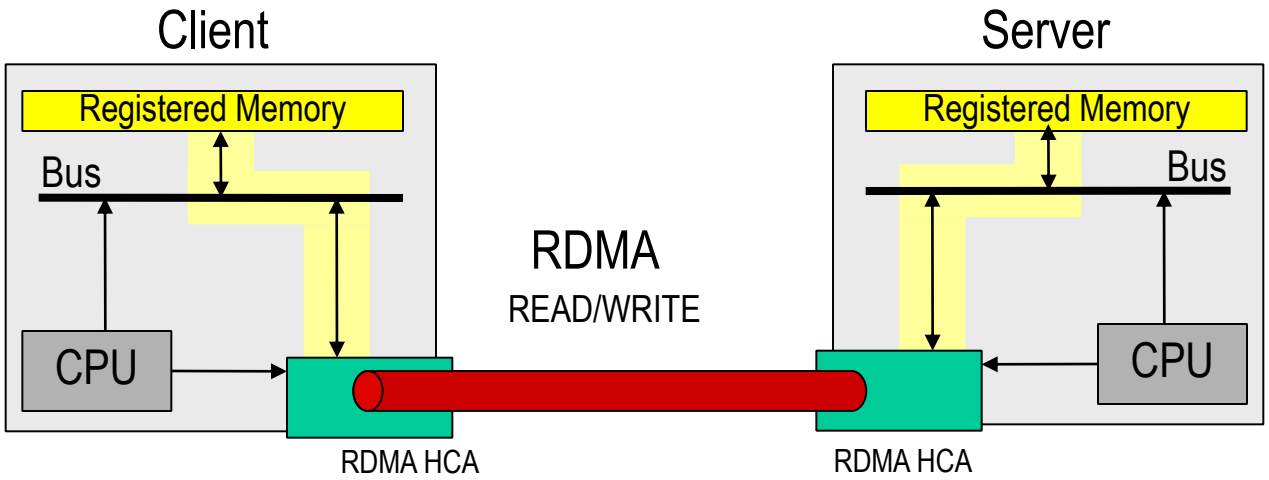


Figure 2 - RDMA communication service example

Courtesy ANSI T.10 Project 1415-D

RDMA based communication



Removes CPU from being bottleneck

User space to User space remote copy - after memory registration

Any (registered) contiguous VM extent as a single RDMA operation

IBA includes RDMA atomic primitive for semaphores and locks

Shared keys created and exchanged for access rights

Allows a new class of distributed application based around low cost, high performance, low latencies

IBA Scope

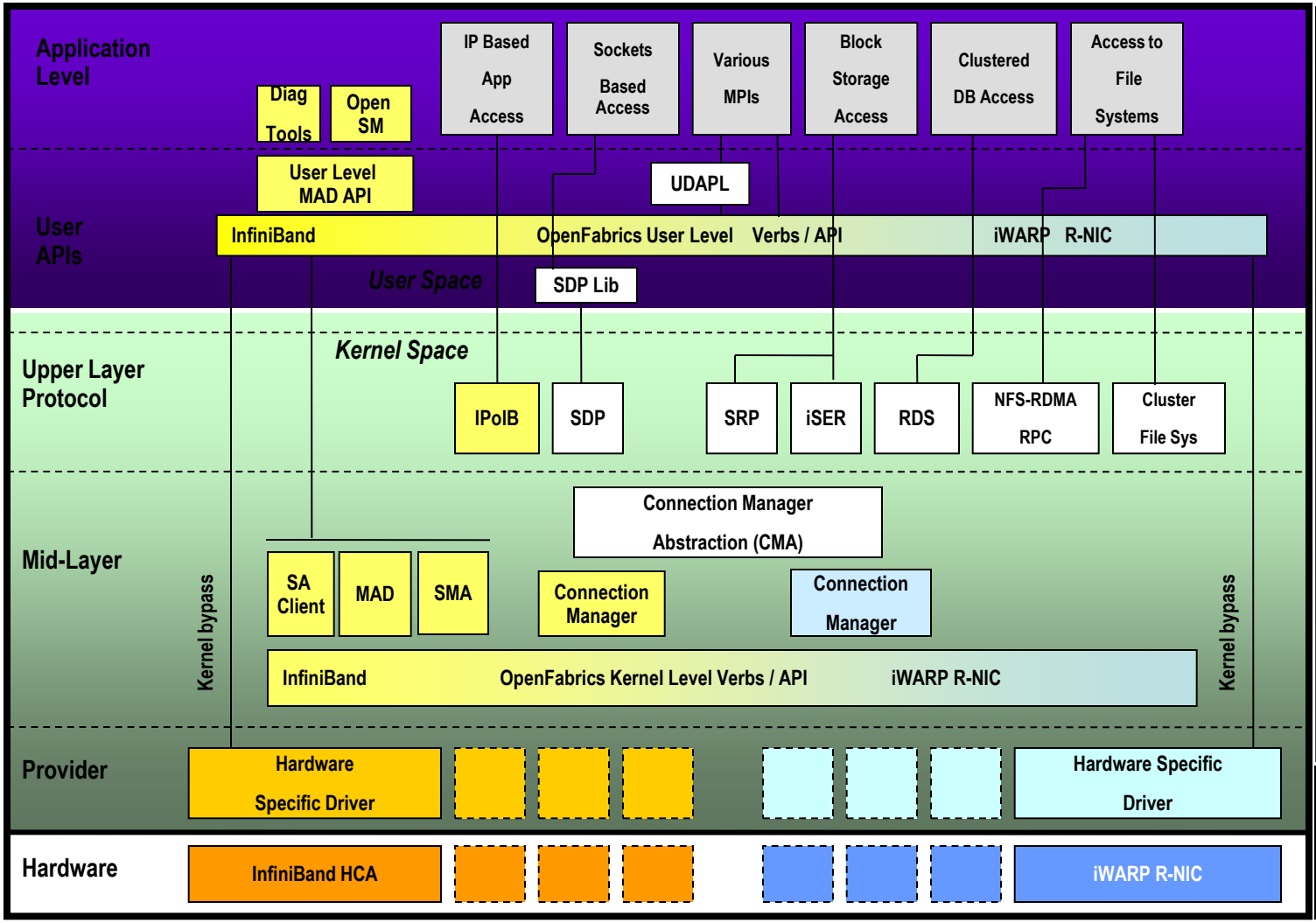
- Defines the InfiniBand protocol and Verbs
- Is not an implementation or assume a specific environment – no mention of Linux or other OS's
- Defines QoS based on Differentiated Services
- Defines path failover within fabric to alternate path
  - Does not define host failover across HCA's, since each HCA is a separate management entity
- Excludes Routing and path selection algorithms
- Defines management protocols and how devices are managed
- Defines the management entity - IBSM (Subnet Manager)
- Defines how subnets can be interconnected through routing
- Defines how a IBSM can be replicated, but not how it can interoperate with other IBSM implementations
  - Single vendor IBSM per fabric
- Must be one active IBSM to initialise fabric

Multicast Support

- One to many communication
- Used by RV, LBM, LLM, Grid and Jini applications
- Sent as unreliable datagrams, i.e. Packets can be dropped
- 16K multicast address space allocated
- Multicast support is fabricated inside the switches using forwarding tables which are setup by the Subnet manager
- Switching handled directly in silicon to minimize packet latency (InfiniScale IV forwards to all ports within 100nS)
- Nodes make explicit requests to create or join multicast groups
- Can be enabled/disabled on a per port basis by the Subnet Manager
- No port flooding or explicit enabling of groups on a port basis
- Supported through IPoIB or directly at the VERB layer
- Single L2 network so no need for routing support e.g. PIM

OpenFabric Current Deliverables

- Derived from and superseded InfiniBand sourceforge project, formerly known as OpenIB. Implementation known as OFED
- Includes:
  - CM – Connection Manager
  - SDP – Socket Direct Protocol (IBA working group specification)
  - MVAPI – Mellanox VERB API
  - SRP Initiator – SCSI RDMA Protocol, an ANSI T10 specification
  - iSER initiator – iSCSI over RDMA, a Voltaire specification
  - IPoIB – Internet Protocol over IB
  - DAPL – Direct Access Provider Library, defined by the DAT Collaborative for RDMA interfaces
  - RDS – Reliable Datagram Sockets
  - OpenIB Subnet Manager
  - VNIC driver – remote access to Ethernet via a IB attached node
- Accepted by Torvolds and included incorporated into Linux 2.6.14 Kernel
- Windows implementation also provided but with fewer API's



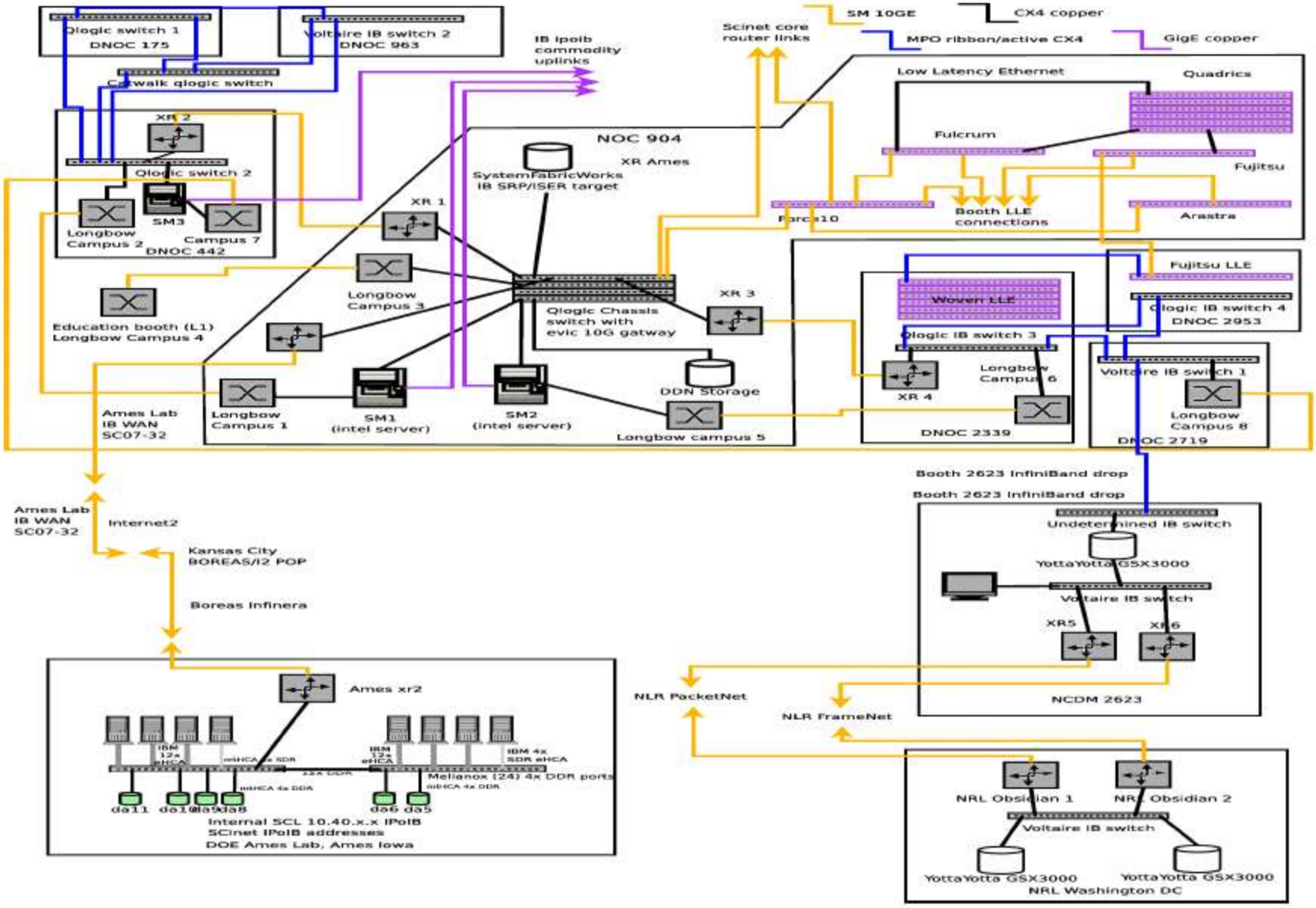
SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Key: Common; InfiniBand; iWARP; Apps & Access Methods for using OF Stack

InfiniBand vendors

- **Open Fabric Alliance** - Linux and Windows InfiniBand software
- **Mellanox** - HCA's , switches, gateways, GUI based management
- **Qlogic** – HCA's, switches, gateways, GUI based management
- **Voltaire** – HCA's, switches, gateway, GUI based management, software accelerators
- **Obsidian Research** – long haul switches
- **Xsigo** – Packaged virtual I/O switches with Ethernet and FibreChannel gateways
- **Oracle/Sun**
  - Solaris (Solaris 10 and later), SPARC and x64
    - Kernel independently developed including IPoIB
    - SDP, SRP, DAPL ported from open source
  - Magnum switch - 3052 port non-blocking DDR switch based on Mellanox silicon
  - InfiniBand integrated inside Oracle RAC Exadata solution
- **IBM**
  - AIX Power software stack
  - IBM z/OS - Used for channel I/O and sysplex clustering
  - Blade server switches (both Mellanox and Voltaire)
  - HCA
  - Native InfiniBand support included in both WebSphere LLM and GPFS
- **HP**
  - HP-UX iTanium Software stack
  - Blade server switches (both Mellanox and Voltaire)

Interoperability – Annual plugfest



This has only been a quick intro, see also:

- Ethernet v. InfiniBand
- InfiniBand vendors
- InfiniBand Management and Observability
- InfiniBand Monitoring and Diagnosis
- InfiniBand Gateways
- Long Distance InfiniBand
- Programming with OFED
- Low Latency Messaging Solutions