Education

InfiniBand Technology Overview

Dror Goldenberg, Mellanox Technologies

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced without modification
  - The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

Abstract

InfiniBand Technology Overview

- The InfiniBand architecture brings fabric consolidation to the data center. Storage networking can concurrently run with clustering, communication and management fabrics over the same infrastructure, preserving the behavior of multiple fabrics. The tutorial provides an overview of the InfiniBand architecture including discussion of High Speed – Low Latency, Channel I/O, QoS scheduling, partitioning, high availability and protocol offload. InfiniBand based storage protocols, iSER (iSCSI RDMA Protocol), NFS over RDMA and SCSI RDMA Protocol (SRP), are introduced and compared with alternative storage protocols, such as iSCSI and FCP. The tutorial further enumerates value-add features that the InfiniBand brings to clustered storage, such as atomic operations and end to end data integrity.

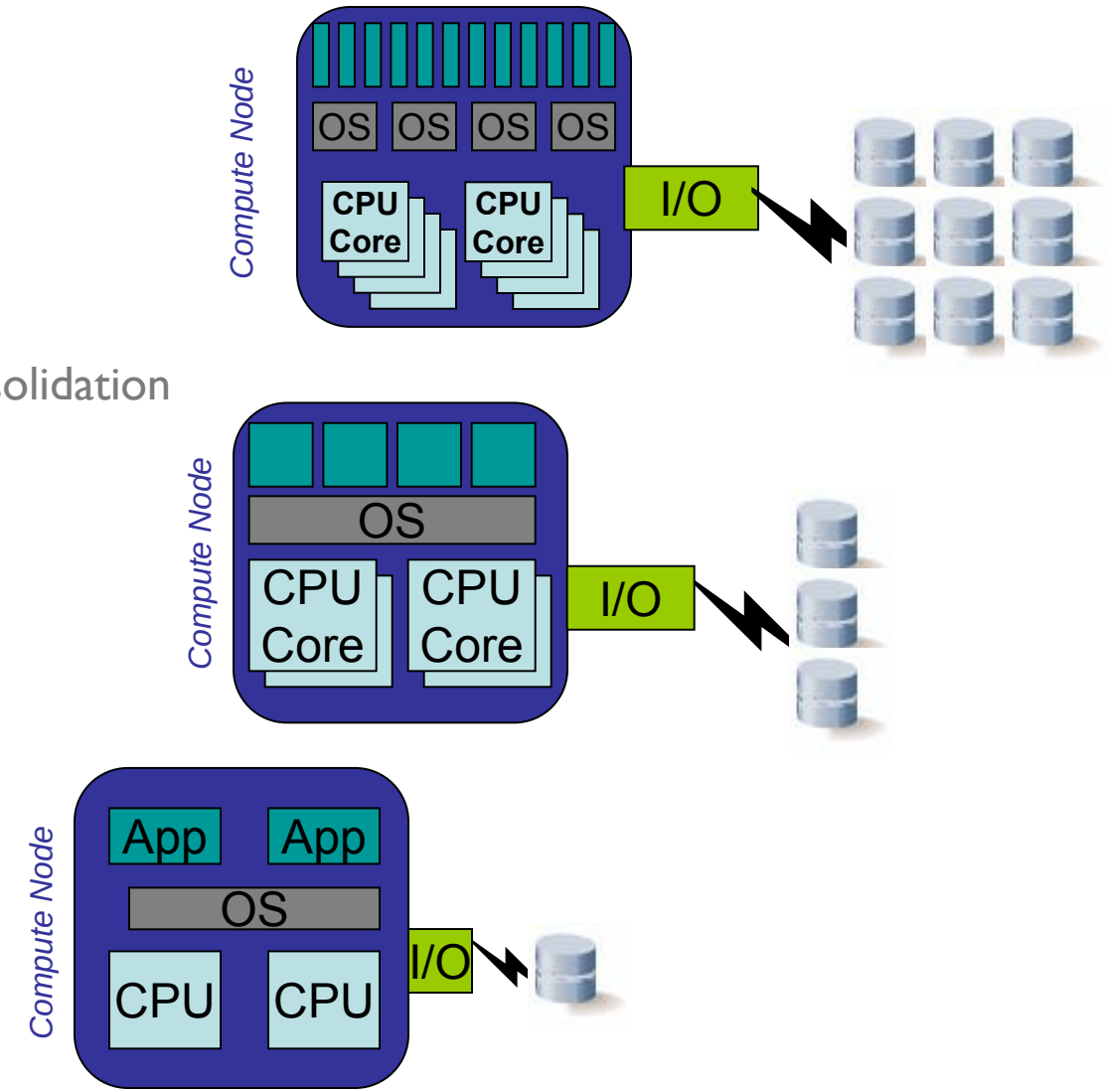
- Learning Objectives:
  - Understand the InfiniBand architecture and feature set.
  - Understand the benefits of InfiniBand for networked storage.
  - Understand the standard InfiniBand storage protocols.

Agenda

- Motivation and General Overview
- Protocol Stack Layers
- Storage Protocols over InfiniBand
- Benefits

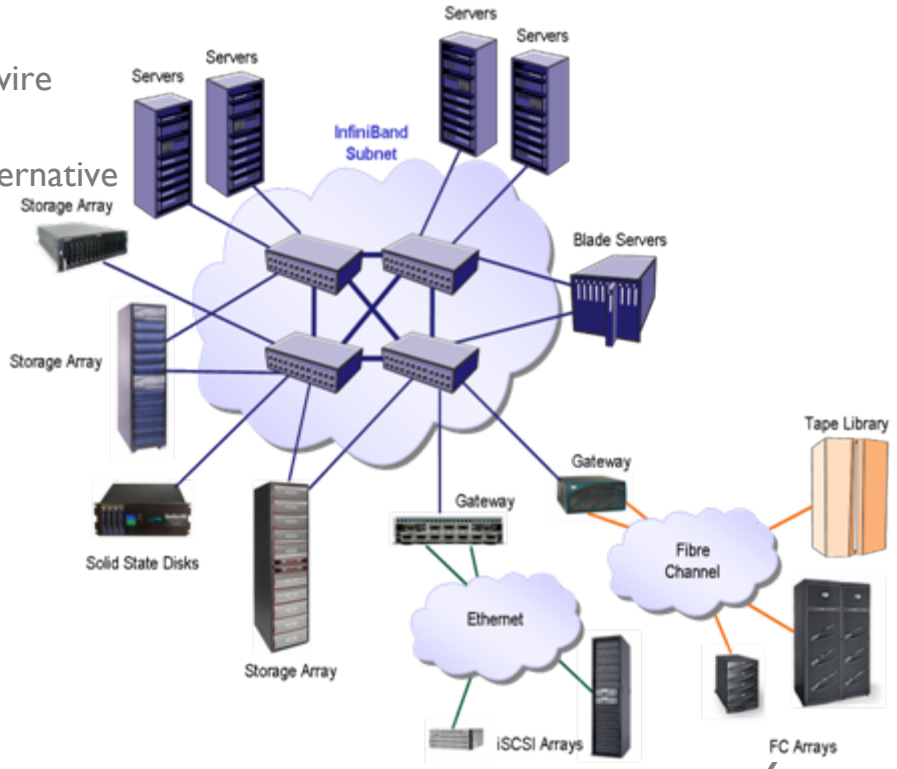
The Need for Better I/O

- Datacenter trends
  - Multi-core CPUs
  - Bladed architecture
  - Fabric consolidation
  - Server virtualization & consolidation
  - Increasing storage demand
- Better I/O is required
  - High capacity
  - Efficient
    - Low latency
    - CPU Offload
  - Scalable
  - Virtualization friendly
  - High availability
  - Performance
  - Low power
  - TCO reduction



InfiniBand Storage Solutions

- The InfiniBand standard - strong alternative to
  - Fibre Channel (SAN)
  - Ethernet (NAS)
- Superior performance
  - 20Gb/s host/target ports (moving to 40Gb/s 2H08)
  - 60Gb/s switch to switch (moving to 120Gb/s 2H08)
  - Sub 1µs end to end latencies
- Unified fabric for the Data Center
  - Storage, networking and clustering over a single wire
- Cost Effective
  - Compelling price/performance advantage over alternative technologies
- Low power Consumption – Green IT
  - Less than 5W per 20Gb/s port InfiniBand
- Mission Critical
  - Highly reliable fabric
  - Multi-pathing
  - Automatic failover
  - Highest level of data integrity

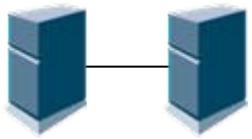


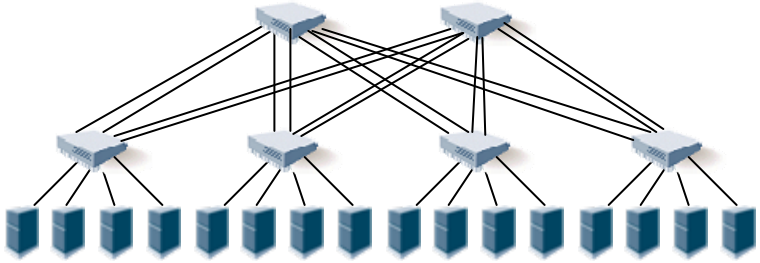
Fabric Technologies Comparison

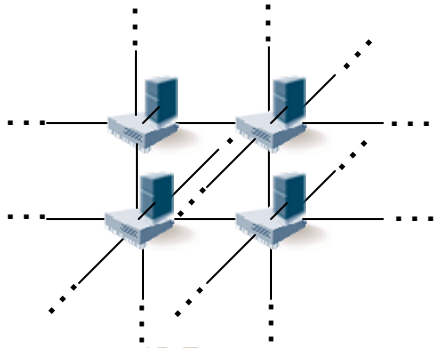
Feature	Fibre Channel	Standard 10 GbE	InfiniBand
Line Rate (GBaud)	4.25 (4GFC) 8.5 (8GFC) 10.51875* (10GFC)	10.3125*	20 (4x DDR) 40 (4x QDR)
Unidirectioanl Throughput (MB/s)**	400 (4GFC) 800 (8GFC) 1,200 (10GFC)	1,250	2,000*** (4x DDR) 4,000 (4x QDR)
Reliable Service	Practically no	No	Yes
Fabric Consolidation	Practically no	In the future (FCoE)	Yes
Copper Distance	15m	10GBase-CX4 15m 10GBase-T 100m	Passive SDR 20m/ DDR 10m Active DDR 25m
Optical Distance****	100m	10GBase-SR 300m 10GBase-LRM 220m	300m (SDR) 150m (DDR)

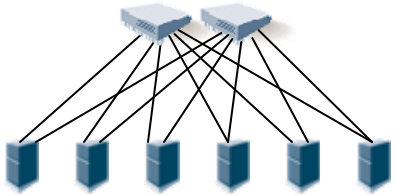
* Serial interface signaling rate
** MB/s = 10^6 Bytes/sec
*** 1,940 MB/s measured
**** Datacenter oriented media

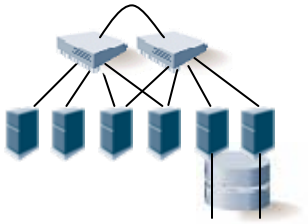
InfiniBand Topologies


Back to Back


2 Level Fat Tree

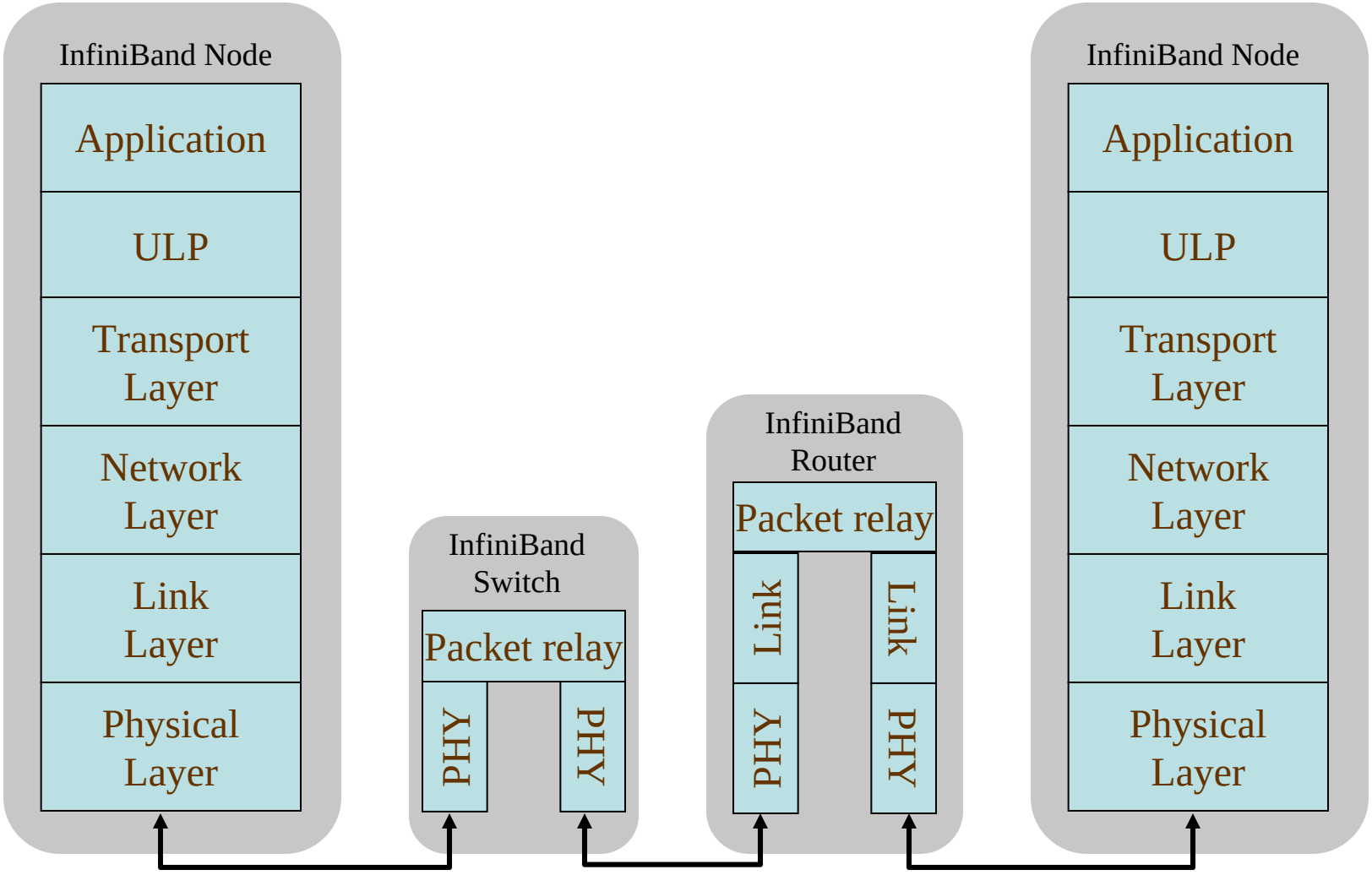

3D Torus


Dual Star


Hybrid

- Example topologies commonly used
- Architecture does not limit topology
- Modular switches are based on fat tree architecture

InfiniBand Protocol Layers



Physical Layer

- Width (1X, 4X, 8X, 12X) including auto-negotiation
- Speed (SDR/DDR/QDR) including auto-negotiation
  - 4X DDR HCAs and switches are currently shipping
- Power management
  - Polling / Sleeping
- Connector
  - Fixed: MicroGiGaCN
  - Pluggable: QSFP
- 8/10 encoding
  - Maintain DC Balance
  - Limited run length of 0's or 1's
- Control symbols (Kxx.x)
  - Lane de-skew, auto negotiation, training, clock tolerance, framing

Link Speed (10^9 bit/sec)

Lane Speed → / Link Width ↓	SDR (2.5GHz)	DDR (5GHz)	QDR (10GHz)
1X	2.5	5	10
4X	10	20	40
8X	20	40	80
12X	30	60	120

* MicroGiGaCN is a trademark of Fujitsu Components Limited

Physical Layer – Cont'd

Copper Cables*:

Width	Speed	Connector	Min Reach	Type / Power**
4X	SDR/ DDR	Micro-GiGaCN	20m/ 10m	Passive
4X	DDR	Micro-GiGaCN	15-25m	Active 0.5-0.75W
12X	SDR/ DDR	24pin Micro-GiGaCN	20m/ 10m	Passive

4X – MicroGiGaCN→

12X – 24 pair MicroGiGaCN→

Fiber Optics*:

Width	Speed	Connector	Type	Min Reach	Power **	Fiber Media
4X	SDR/ DDR	Micro-GiGaCN	Media Converter	300m/ 150m	0.8-1W	12 strand MPO
4X	DDR	Micro-GiGaCN	Optical Cable	100m	1W	12 strand attached

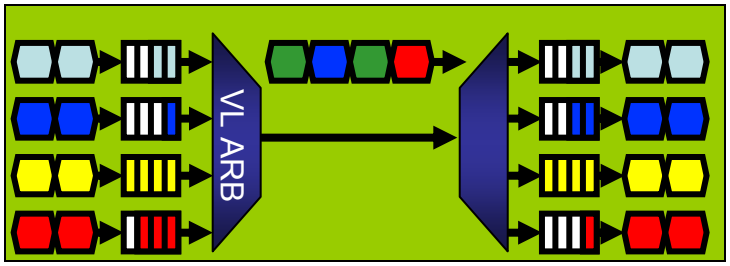
4X - MicroGiGaCN MPO Media Converter →

4X - MicroGiGaCN Optical Cable →

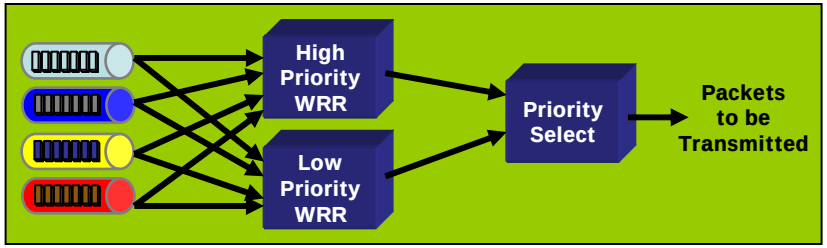
* currently deployed
** per end

Link Layer

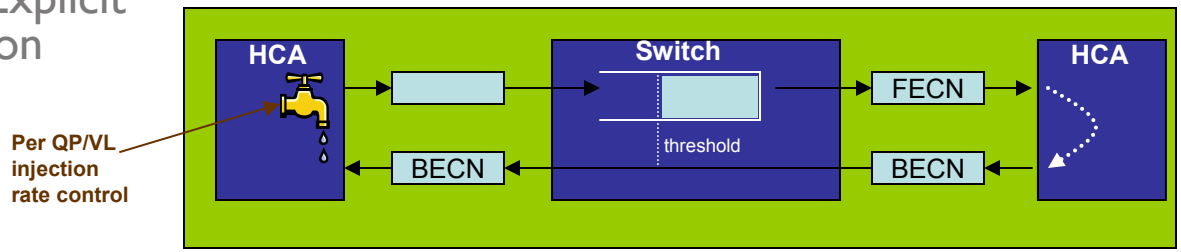
- Addressing and Switching
  - Local Identifier (LID) addressing
  - Unicast LID - 48K addresses
  - Multicast LID – up to 16K addresses
  - Efficient linear lookup
  - Cut through switching supported
  - Multi-pathing support through LMC
- Independent Virtual Lanes
  - Flow control (lossless fabric)
  - Service level
  - VL arbitration for QoS
- Congestion control
  - Forward / Backward Explicit Congestion Notification (FECN/BECN)
- Data Integrity
  - Invariant CRC
  - Variant CRC



Independent Virtual Lanes (VLs)



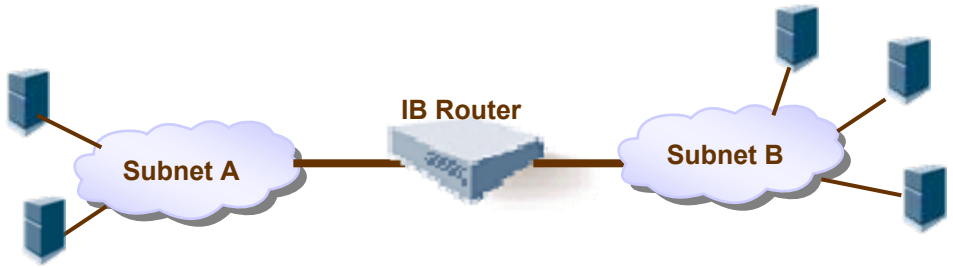
H/L Weighted Round Robin (WRR) VL Arbitration



Efficient FECN/BECN Based Congestion Control

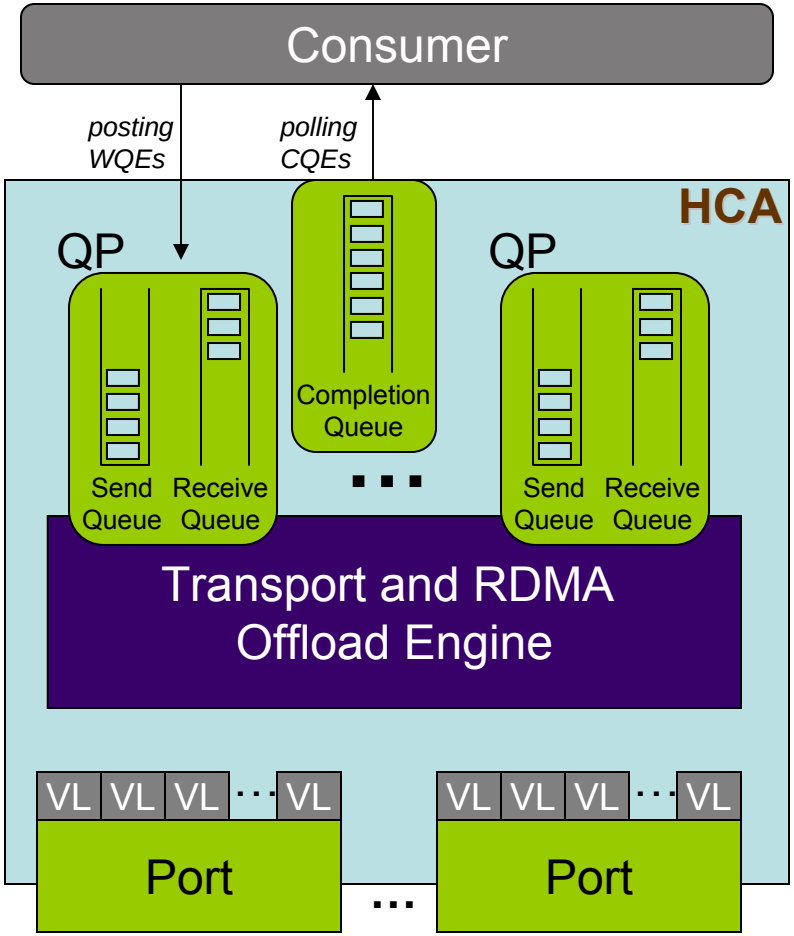
Network Layer

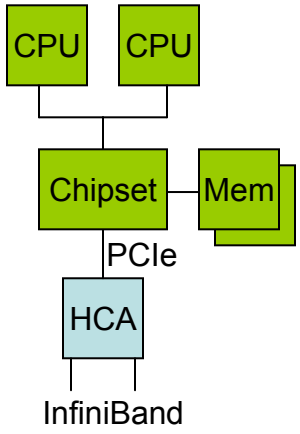
- Global Identifier (GID) addressing
  - Based on IPv6 addressing scheme
  - GID = {64 bit GID prefix, 64 bit GUID}
    - GUID = Global Unique Identifier (64 bit EUI-64)
    - GUID 0 – assigned by the manufacturer
    - GUID 1..(N-1) – assigned by the Subnet Manager
- Optional for local subnet access
- Used for multicast distribution within end nodes
- Enables routing between IB subnets
  - Still under definition in IBTA
  - Will leverage IPv6 routing algorithms

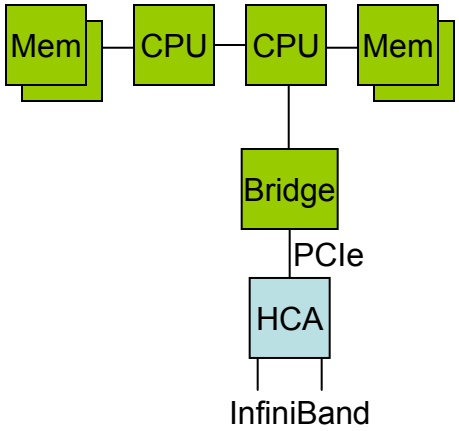


Transport – Host Channel Adapter (HCA) Model

- Asynchronous interface
  - Consumer posts work requests
  - HCA processes
  - Consumer polls completions
- Transport executed by HCA
- I/O channel exposed to the application
- Transport services
  - Reliable / Unreliable
  - Connected / Datagram



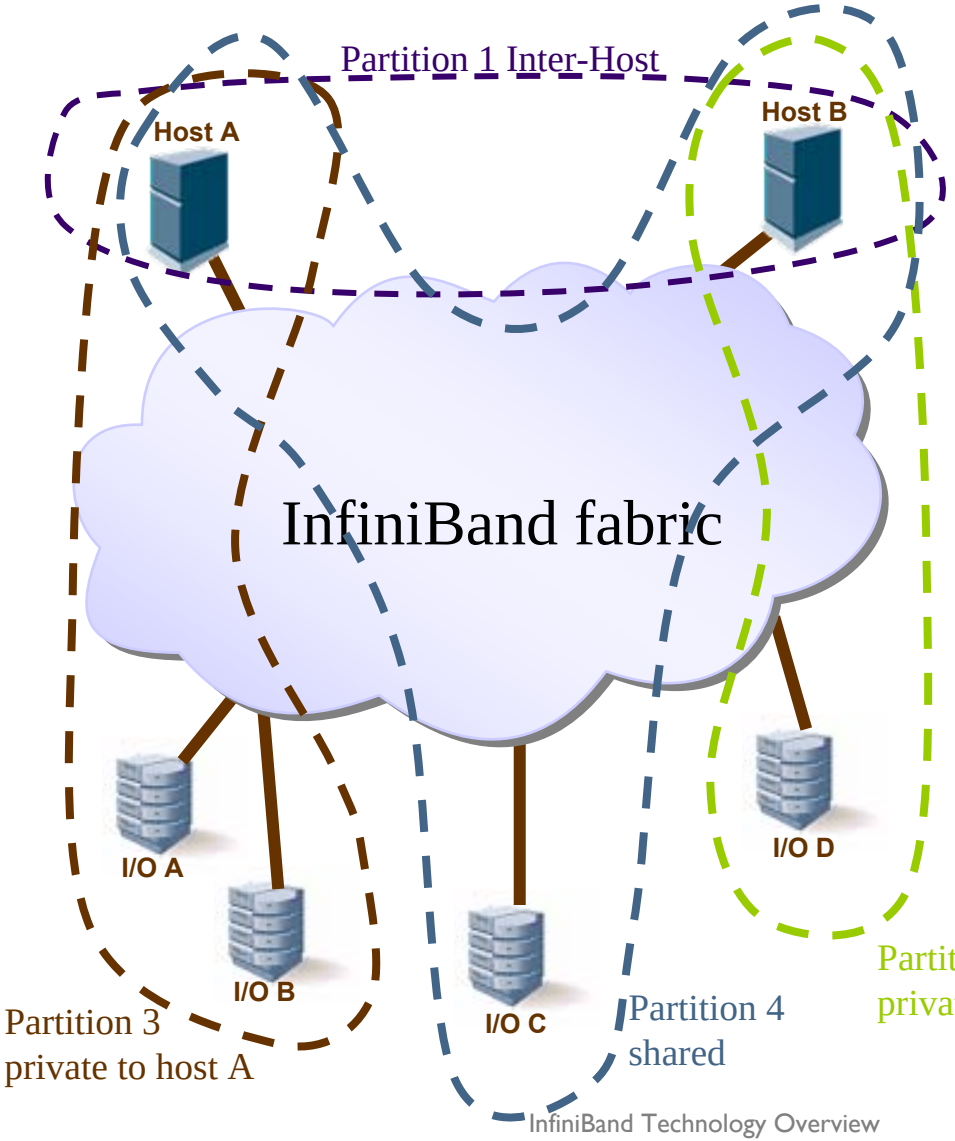




Transport Layer

- Queue Pair (QP) – transport endpoint
  - Asynchronous interface
    - Send Queue, Receive Queue, Completion Queue
  - Full transport offload
    - Segmentation, reassembly, timers, retransmission, etc
  - Operations supported
    - Send/Receive – messaging semantics
    - RDMA Read/Write – enable zero copy operations
    - Atomics – remote Compare & Swap, Fetch & Add
    - Memory management - Bind/Fast Register/Invalidate
- Kernel bypass
  - Enables low latency and CPU offload
  - Exposure of application buffers to the network
    - Direct data placement / zero copy
  - Enabled through QPs, Completion Queues (CQs), Protection Domains (PD), Memory Regions (MRs)
- Polling and Interrupt models supported

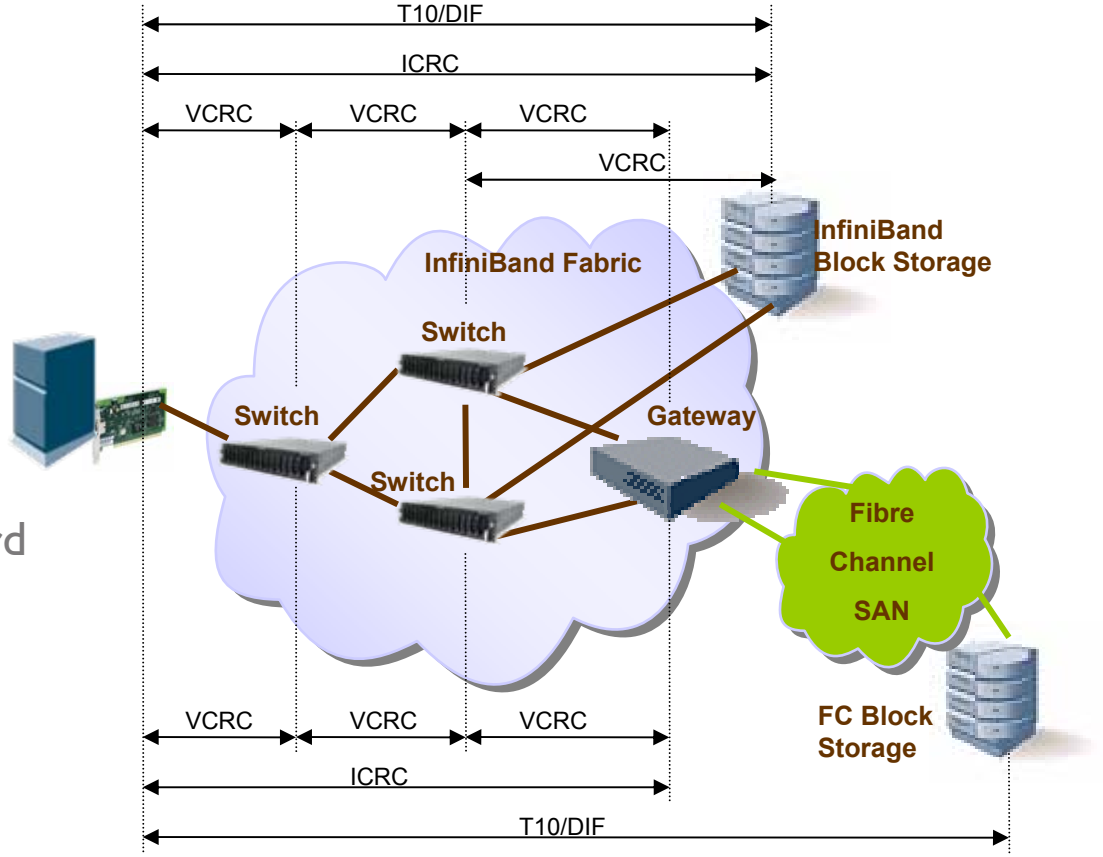
Partitions



- Logically divide the fabric into isolated domains
- Partial and full membership per partition
- Partition filtering at switches
- Similar to
  - FC Zoning
  - 802.1Q VLANs

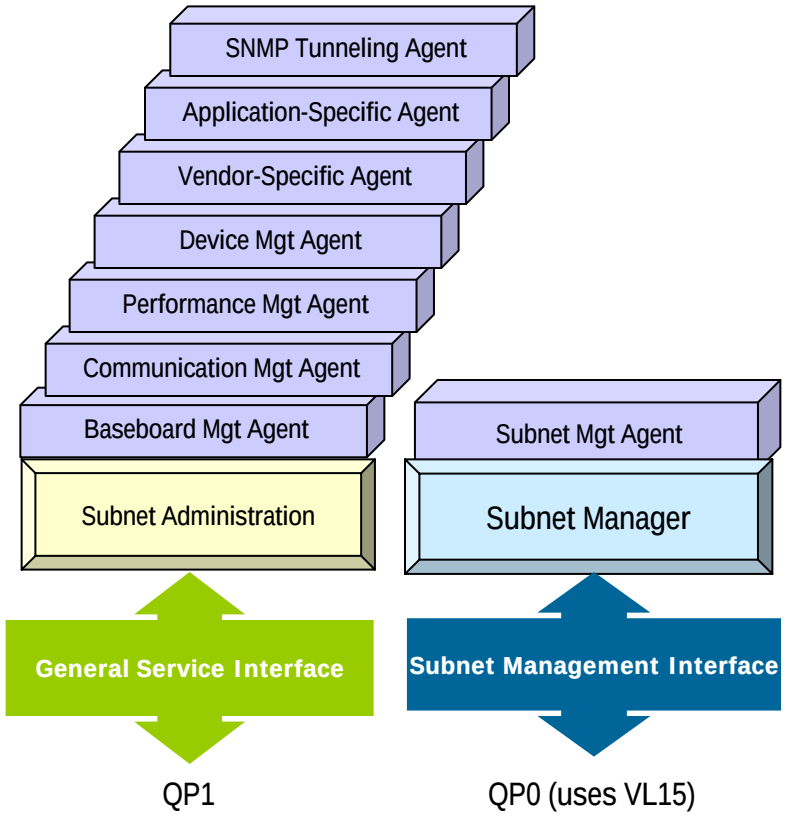
InfiniBand Data Integrity

- Hop by hop
  - VCRC – 16 bit CRC
  - CRC16 0x100B
- End to end
  - ICRC – 32 bit CRC
  - CRC32 0x04C11DB7
  - Same CRC as Ethernet
- Application level
  - T10/DIF Logical Block Guard
    - Per block CRC
  - 16 bit CRC 0x8BB7



Management Model

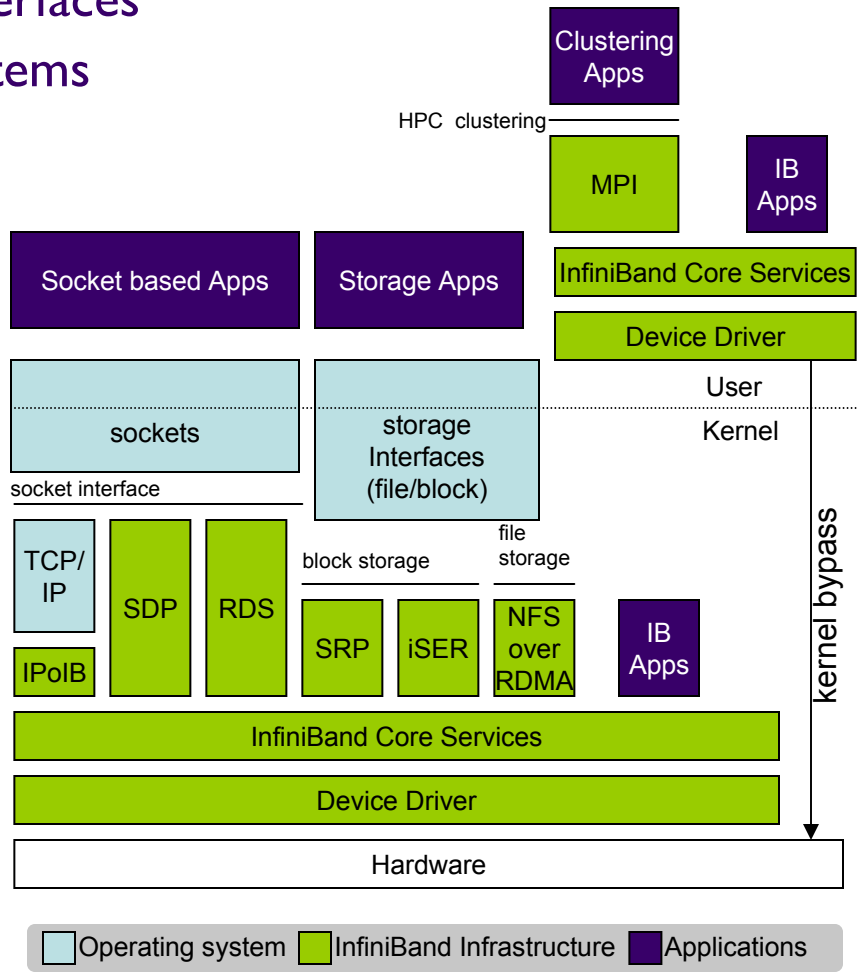
- Subnet Manager (SM)
  - Configures/Administers fabric topology
  - Implemented at an end-node or a switch
  - Active/Passive model when more than one SM is present
  - Talks with SM Agents in nodes/switches
- Subnet Administration
  - Provides path records
  - QoS management
- Communication Management
  - Connection establishment processing



Upper Layer Protocols

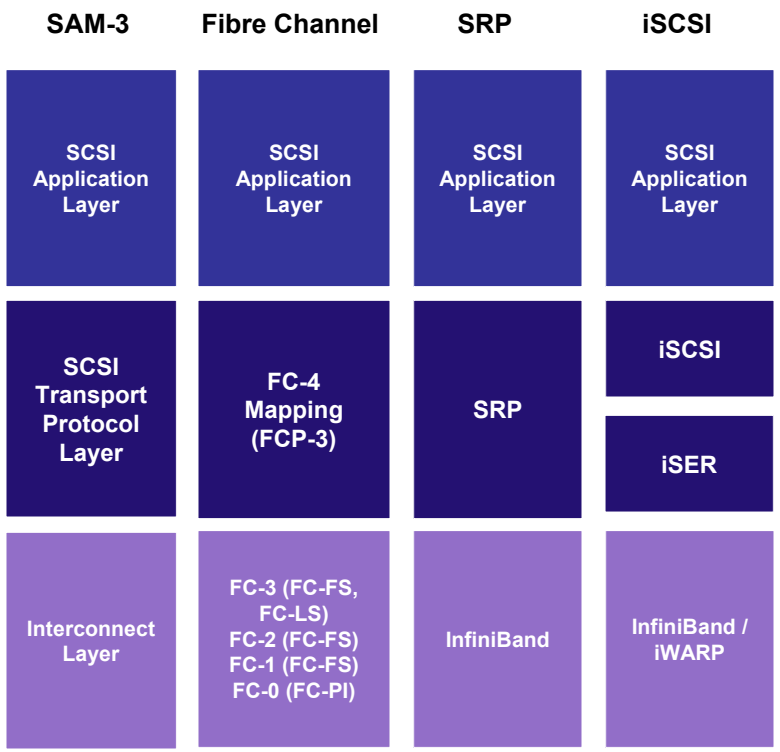
- ULPs connect InfiniBand to common interfaces
- Supported on mainstream operating systems

- Clustering
  - MPI (Message Passing Interface)
  - RDS (Reliable Datagram Socket)
- Network
  - IPoIB (IP over InfiniBand)
  - SDP (Socket Direct Protocol)
- Storage
  - SRP (SCSI RDMA Protocol)
  - iSER (iSCSI Extensions for RDMA)
  - NFSoRDMA (NFS over RDMA)



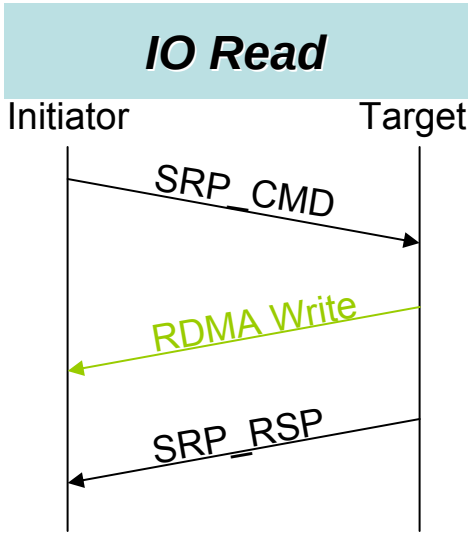
InfiniBand Block Storage Protocols

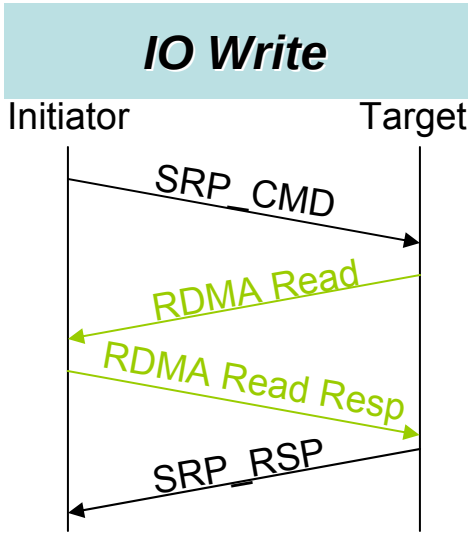
- SRP - SCSI RDMA Protocol
  - Defined by T10
- iSER – iSCSI Extensions for RDMA
  - Defined by IETF IP Storage WG
  - InfiniBand specifics (e.g. CM) defined by IBTA
  - Leverages iSCSI management infrastructure
- Protocol offload
  - Use IB Reliable Connected
  - RDMA for zero copy data transfer



SRP - Data Transfer Operations

- Send/Receive
  - Commands
  - Responses
  - Task management
- RDMA – Zero Copy Path
  - Data-In
  - Data-Out
- iSER uses the same principles
  - Immediate/Unsolicited data allowed through Send/Receive
- iSER and SRP are part of mainline Linux kernel



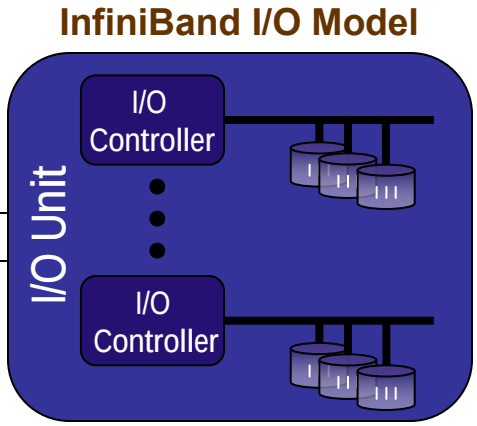


Data Transfer Summary

	SRP	iSER	iSCSI	FCP
Request	SRP_CMD (SEND)	SCSI-Command (SEND)	SCSI-Command	FCP_CMND
Response	SRP_RSP (SEND)	SCSI-Response (SEND)	SCSI-Response (or piggybacked on Data-In PDU)	FCP_RSP
Data-In Delivery	RDMA Write	RDMA Write	Data-In	FCP_DATA
Data-Out Delivery	RDMA Read RDMA Read Resp.	RDMA Read RDMA Read Resp.	R2T Data-Out	FCP_XFER_RDY FCP_DATA
Unsolicited Data-Out Delivery		Part of SCSI-Command (SEND) Data-Out (SEND)	Part of SCSI-Command Data-Out	FCP_DATA
Task Management	SRP_TSK_MGMT (SEND)	Task Management Function Request/ Response (SEND)	Task Management Function Request/ Response	FCP_CMND

SRP Discovery

- Discovery methods
  - Persistent Information {Node_GUID:IOC_GUID}
  - Subnet Administrator (Identify all ports with CapabilityMask.IsDM)
  - Configuration Manager (CFM)*
    - Locate the Device Administrator through Service Record
  - Boot Manager*
  - Boot Information Service*
- Identifiers
  - Per LUN WWN (through INQUIRY VPD)
  - SRP Target Port ID {IdentifierExt[63:0], IOC GUID[63:0]}
  - Service Name – SRP.T10.{PortID ASCII}
  - Service ID – Locally assigned by the IOC/IOU



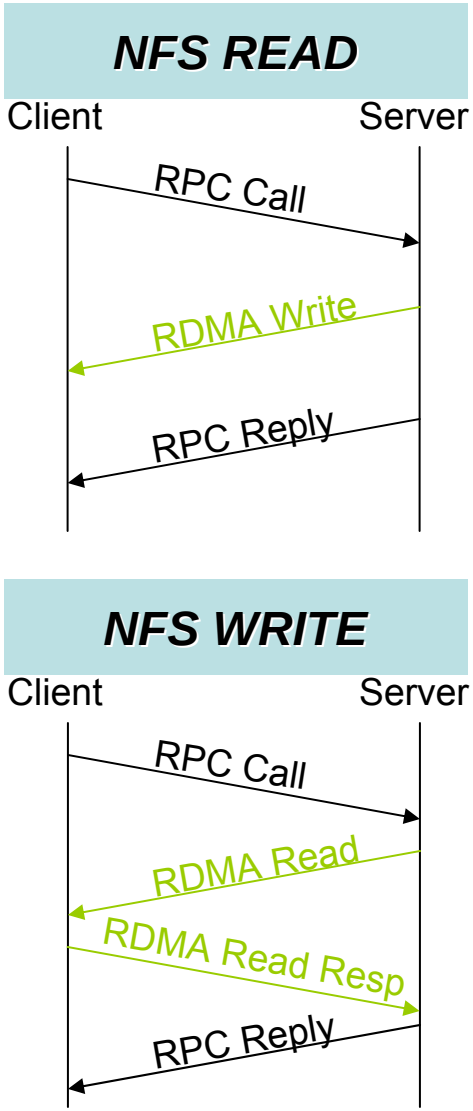
* Not deployed today

iSER Discovery

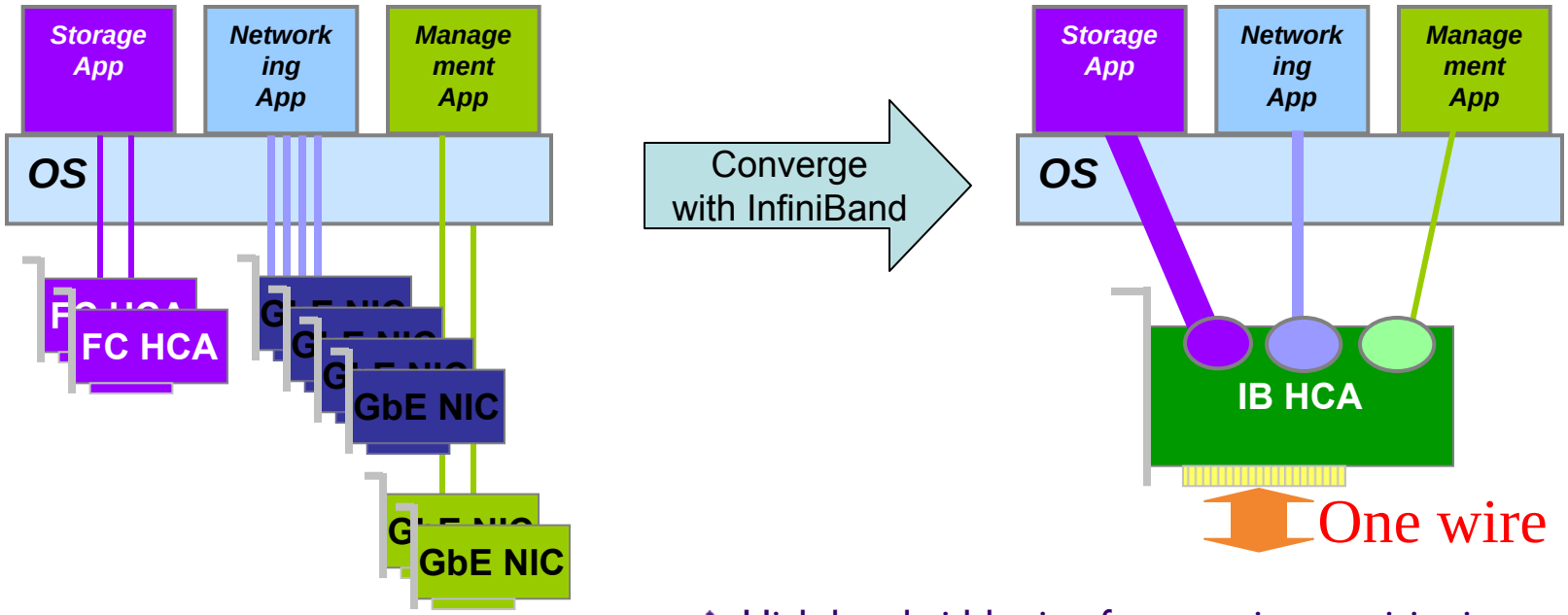
- Leverages all iSCSI infrastructure
  - Using IP over InfiniBand
- Same iSCSI mechanisms for discovery (RFC 3721)
  - Static Configuration {IP, port, target name}
  - Send Targets {IP, port}
  - SLP
  - iSNS
- Same target naming (RFC 3721/3980)
  - iSCSI Qualified Names (iqn.)
  - IEEE EUI64 (eui.)
  - T11 Network Address Authority (naa.)

NFS over RDMA

- Defined by IETF
  - ONC-RPC extensions for RDMA
  - NFS mapping
- RPC Call/Reply
  - Send/Receive – if small
  - Via RDMA Read chunk list - if big
- Data transfer
  - RDMA Read/Write – described by chunk list in XDR message
  - Send – inline in XDR message
- Uses InfiniBand Reliable Connected QP
  - Uses IP extensions to CM
  - Connection based on IP address and TCP port
  - Zero copy data transfers
- NFSoRDMA is part of mainline Linux kernel



I/O Consolidation

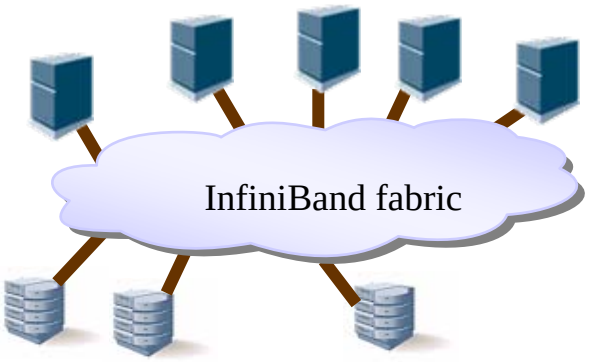


- Slower I/O
- Different service needs – different fabrics
- No flexibility
- More ports to manage
- More power
- More space
- Higher TCO

- High bandwidth pipe for capacity provisioning
- Dedicated I/O channels enable convergence
  - For Networking, Storage, Management
  - Application compatibility
  - QoS - differentiates different traffic types
  - Partitions – logical fabrics, isolation
- Gateways - Share remote Fibre Channel and Eth ports
  - Design based on average load across multiple servers
  - Scale incrementally – add Ethernet/FC/Server blades
  - Scale independently

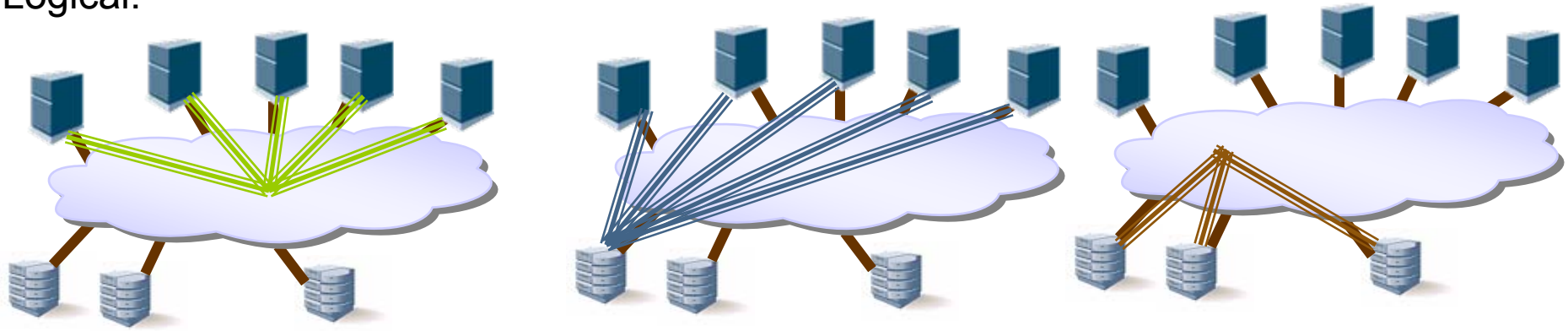
I/O Consolidation – VLs and Scheduling Example

Physical:



- VLs and scheduling can be dynamically configured and adjusted to match application performance requirements

Logical:



Low Latency VL
For Clustering

Mainstream Storage VL
Day - at least 40% BW
Night – at least 20% BW

Backup VL
Day – at least 20% BW
Night – at least 70% BW

High Availability and Redundancy

- Multi-port HCAs
  - Covers link failure
- Redundant fabric topologies
  - Covers link failure
- Link layer multi-pathing (LMC)
- Automatic Path Migration (APM)
- ULP High Availability
  - Application level multi-pathing (SRP/iSER)
  - Teaming/Bonding (IPoIB)
  - Covers HCA failure and link failure

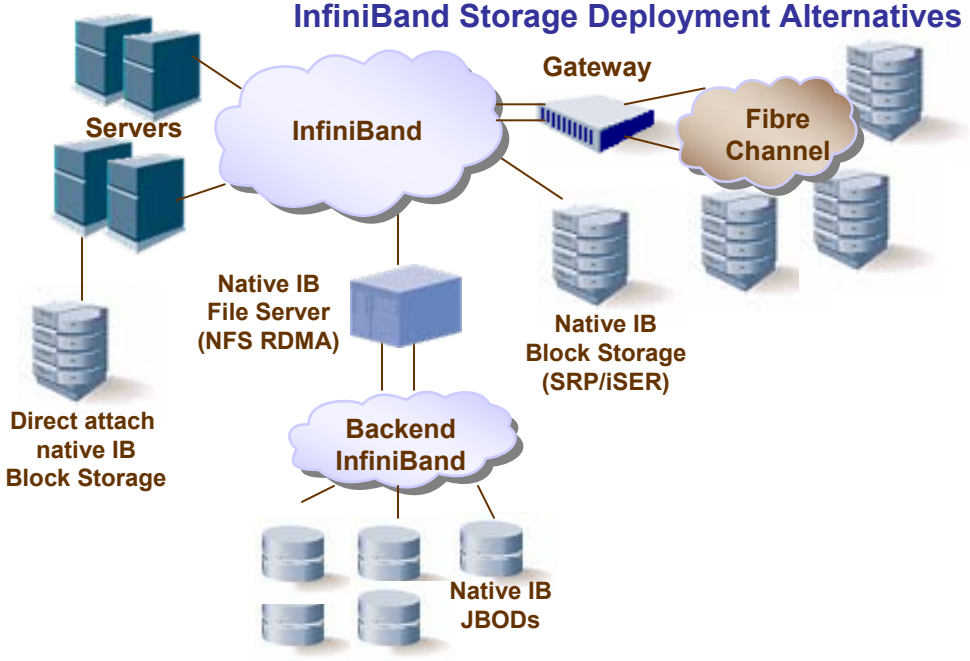
Performance Metrics

- IB Verbs
  - Latency
    - RDMA Write 0.99us
    - RDMA Read 1.87us (roundtrip)
  - Bandwidth
    - 1.5-1.9GB/s (unidirectional)
    - 2.9-3.7GB/s (bidirectional)
      - Depends on PCIe (2.5-5GT/s)
- Clustering (MPI)
  - Latency 1.2us
  - Message rate 30M msg/sec
- Block Storage (SRP)
  - Bandwidth (1MB I/O, no RAID)
    - I/O Read 1.7GB/s
    - I/O Write 1.6GB/s
    - Single DDR port
- File Storage (NFSoRDMA)
  - Read 1.6GB/s
  - Write 0.59GB/s
  - Single DDR port

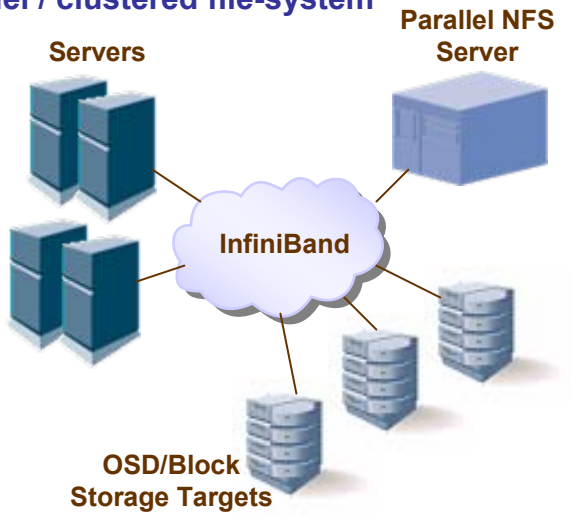
InfiniBand Storage Opportunities & Benefits

- Clustering port can connect to storage
- High Bandwidth Fabric
- Fabric consolidation (QoS, partitioning)
- Efficiency – full offload and zero copy
- Gateways
  - One wire out of the server
  - Shared remote FC ports - scalability

InfiniBand Storage Deployment Alternatives



Parallel / clustered file-system



- Clustered/Parallel storage, Backend fabric benefits:
  - Combined with clustering infrastructure
  - Efficient object/block transfer
  - Atomic operations
  - Ultra low latency
  - High bandwidth

Summary

- Datacenter developments require better I/O
  - Increasing compute power per host
  - Server virtualization
  - Increasing storage demand
- InfiniBand I/O is a great fit for the datacenter
  - Layered implementation
  - Brings fabric consolidation
  - Enables efficient SAN, Network, IPC and Management traffic
  - Price/Performance
  - Gateways provide scalable connectivity to existing fabrics
- Existing storage opportunities with InfiniBand
  - Connectivity to HPC clusters, where IB is the dominant fabric

Other SNIA Tutorials

- **Check out SNIA Tutorials**
  - **Comparing Server I/O Consolidation Solutions: iSCSI, InfiniBand and FCoE**
  - **Fibre Channel over Ethernet**

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: tracknetworking@snia.org

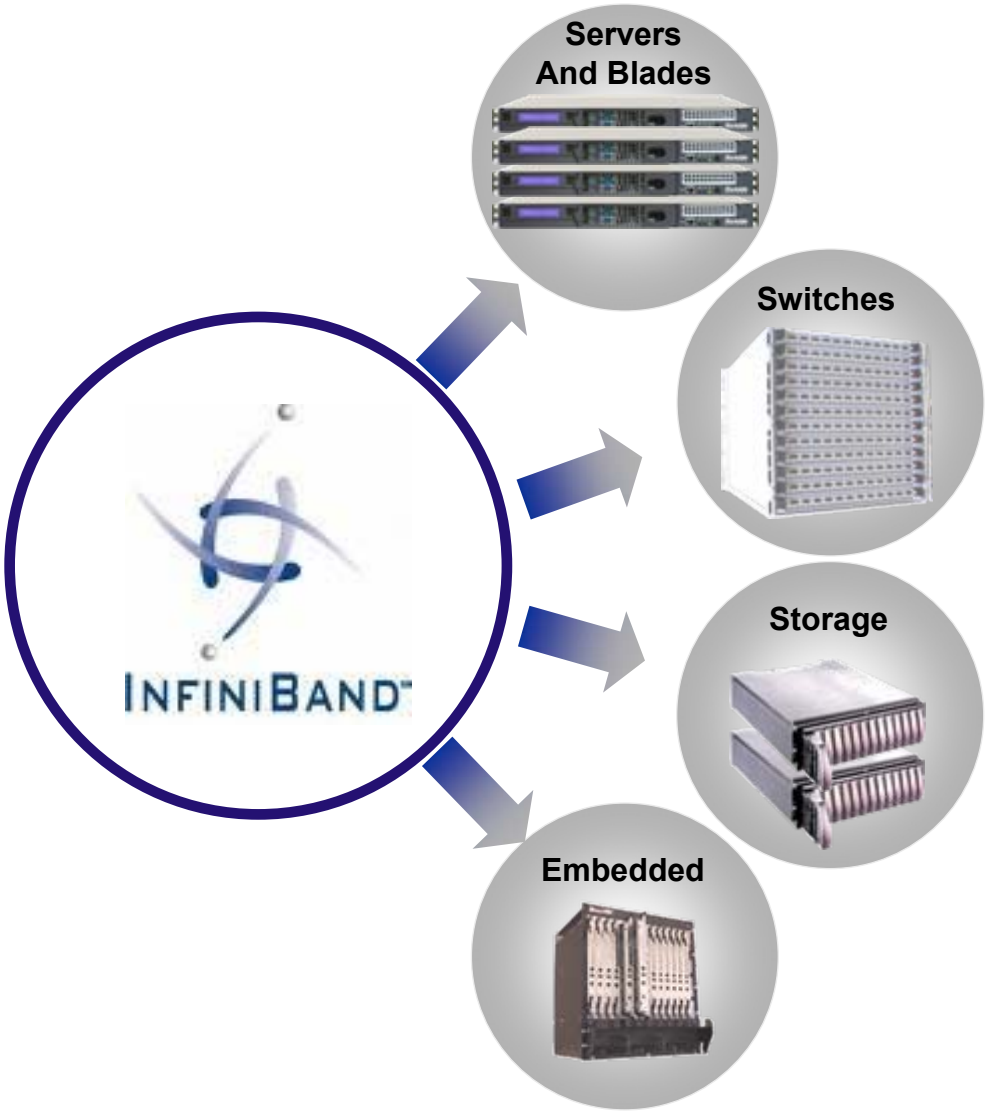
Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Bill Lee
Howard Goldstein
Sujal Das
Graham Smith
Ron Emerick
Walter Dey

Backup

Interconnect: A Competitive Advantage



End-Users

Enterprise Data Centers

- Clustered Database
- eCommerce and Retail
- Financial
- Supply Chain Management
- Web Services

High-Performance Computing

- Biosciences and Geosciences
- Computer Automated Engineering
- Digital Content Creation
- Electronic Design Automation
- Government and Defense

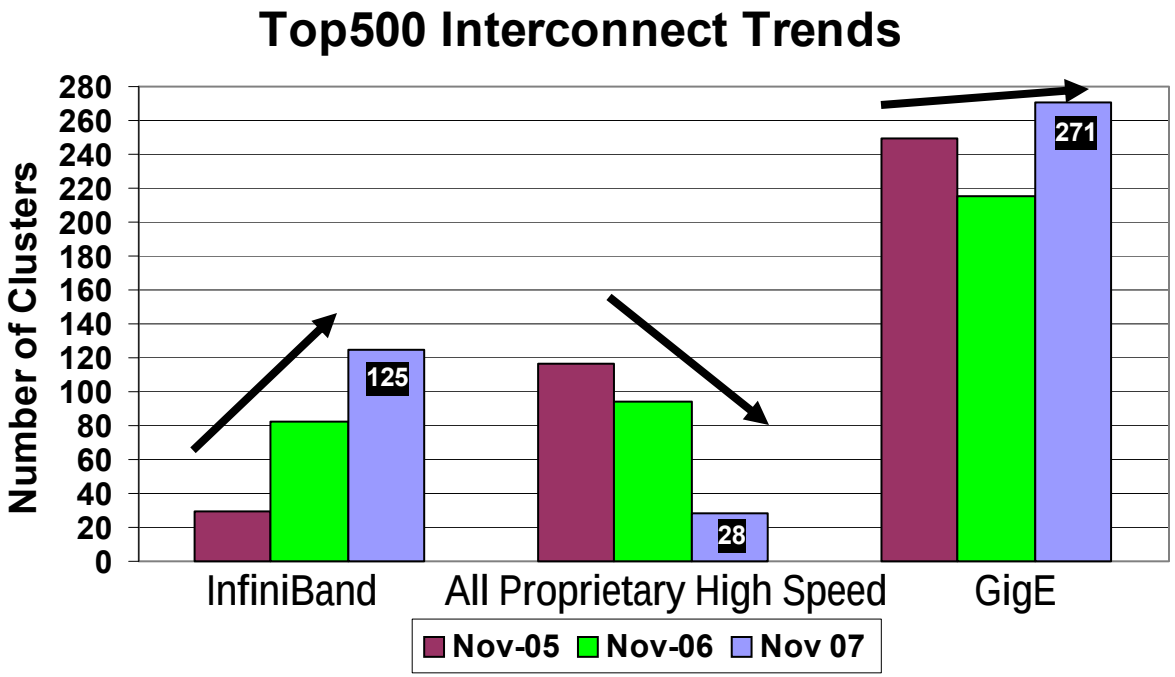
Embedded

- Communications
- Computing and Storage Aggregation
- Industrial
- Medical
- Military

Applicable Markets for InfiniBand

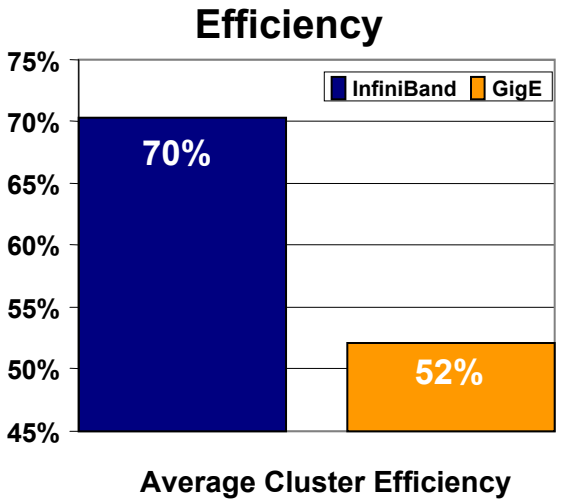
- Data Centers
  - Clustered database, data warehousing, shorter backups, I/O consolidation, power savings, virtualization, SOA, XTP
- Financial
  - Real-time risk assessment, grid computing and I/O consolidation
- Electronic Design Automation (EDA) and Computer Automated Design (CAD)
  - File system I/O is the bottleneck to shorter job run times
- High Performance Computing
  - High throughput I/O to handle expanding datasets
- Graphics and Video Editing
  - HD file sizes exploding, shorter backups, real-time production

Interconnect Trends – Top500



Growth rate from Nov 06 to Nov 07 (year)

- InfiniBand: +52%
- All Proprietary: -70%
- GigE: +26%



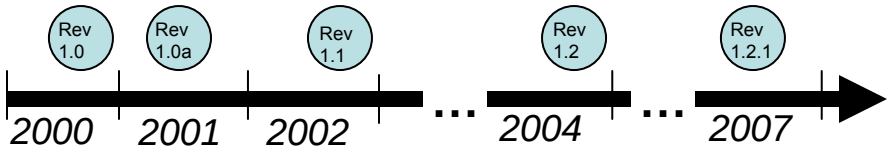
InfiniBand - the only growing high speed interconnect

- 52% growth from Nov 2006

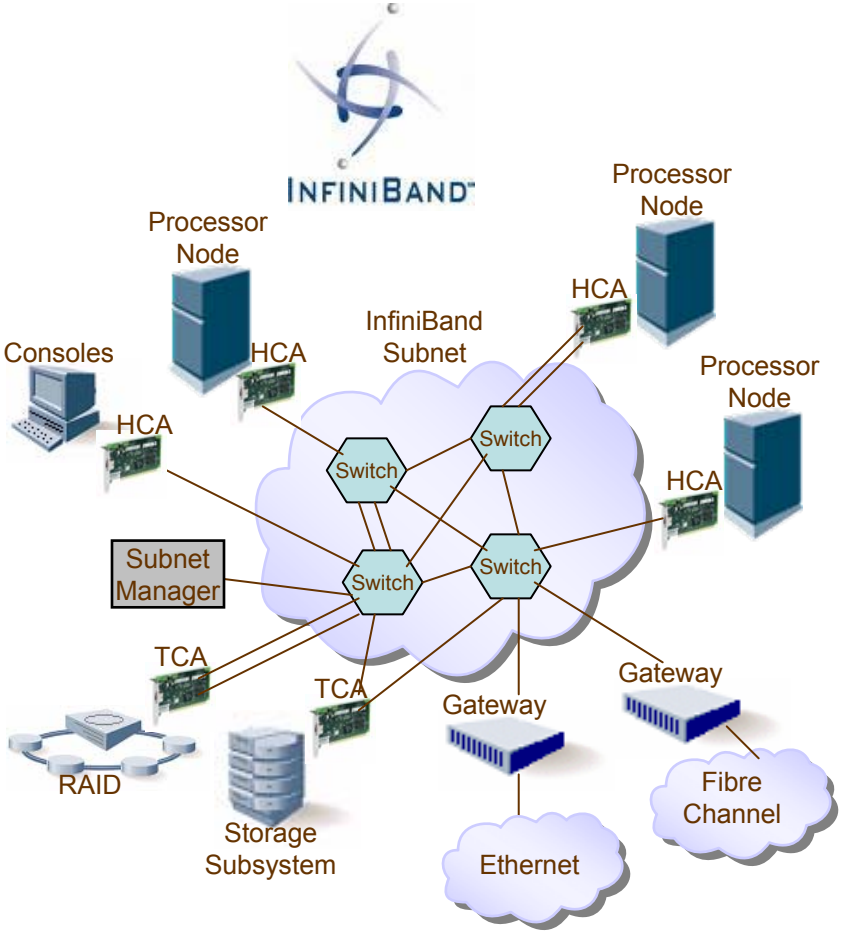
Source: http://www.top500.org/list/2007/06/
The TOP500 project was started in 1993 to provide a reliable basis for tracking and detecting trends in high-performance computing.

The InfiniBand Architecture

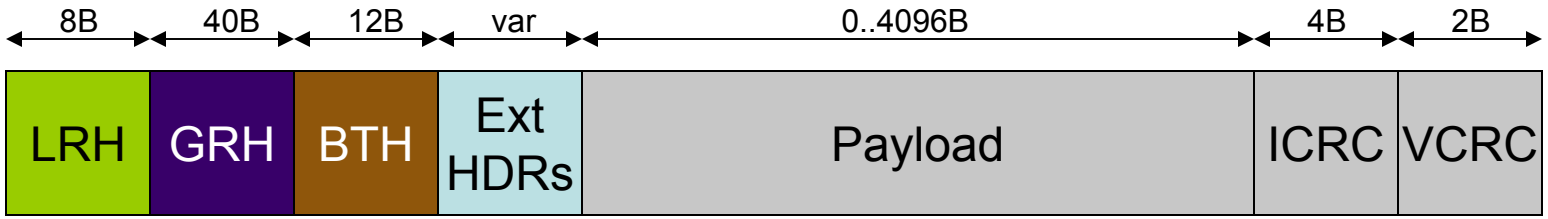
- Industry standard defined by the InfiniBand Trade Association
- Defines System Area Network architecture
  - Comprehensive specification:
    from physical to applications



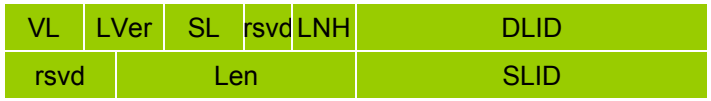
- Architecture supports
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload



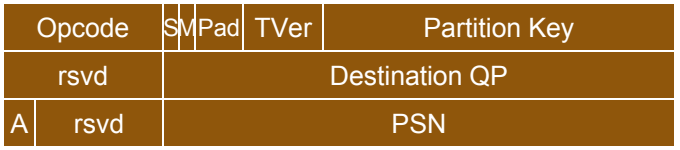
InfiniBand Packet Format



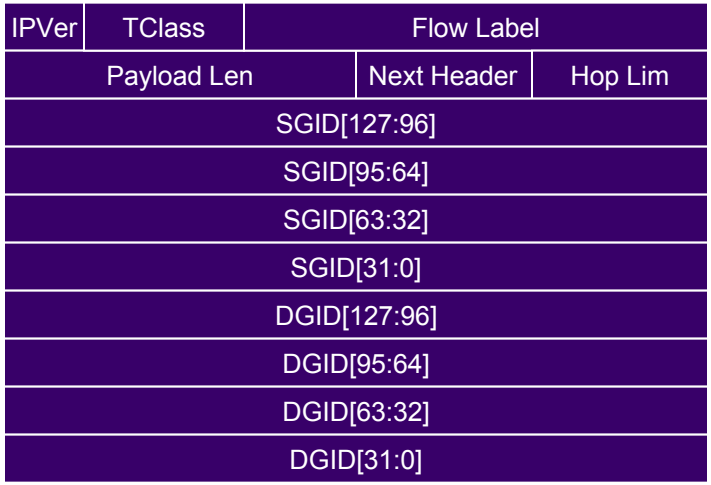
InfiniBand Data Packet



LRH



BTH



GRH (Optional)

Extended headers:

- Reliable Datagram ETH (4B)
- Datagram ETH (8B)
- RDMA ETH (16B)
- Atomic ETH (28B)
- ACK ETH (4B)
- Atomic ACK ETH (8B)
- Immediate Data ETH (4B)
- Invalidate ETH (4B)

InfiniBand Resources

InfiniBand software is developed under OpenFabrics Open source Alliance
http://www.openfabrics.org/index.html

InfiniBand standard is developed by the InfiniBand® Trade Association
http://www.infinibandta.org/home

Reference

- InfiniBand Architecture Specification Volumes 1-2 Release 1.2.1
  - www.infinibandta.org
- IP over InfiniBand
  - RFCs 4391, 4392, 4390, 4755 (www.ietf.org)
- NFS Direct Data Placement
  - http://www.ietf.org/html.charters/nfsv4-charter.html
- iSCSI Extensions for RDMA (iSER) Specification
  - http://www.ietf.org/html.charters/ips-charter.html
- SCSI RDMA Protocol (SRP), DIF
  - www.t10.org

Glossary

- APM - Automatic Path Migration
- BECN - Backward Explicit Congestion Notification
- BTH - Base Transport Header
- CFM - Configuration Manager
- CQ - Completion Queue
- CQE - Completion Queue Element
- CRC - Cyclic Redundancy Check
- DDR - Double Data Rate
- DIF - Data Integrity Field
- FC - Fibre Channel
- FECN - Forward Explicit Congestion Notification
- GbE - Gigabit Ethernet
- GID - Global IDentifier
- GRH - Global Routing Header
- GUID - Globally Unique IDentifier
- HCA - Host Channel Adapter
- IB - InfiniBand
- IBTA - InfiniBand Trade Association
- ICRC - Invariant CRC
- IPoIB - Internet Protocol Over InfiniBand
- IPv6 - Internet Protocol Version 6
- iSER - iSCSI Extensions for RDMA
- LID - Local IDentifier
- LMC - Link Mask Control
- LRH - Local Routing Header
- LUN - Logical Unit Number
- MPI - Message Passing Interface
- MR - Memory Region
- NFSoRDMA - NFS over RDMA
- OSD - Object based Storage Device
- OS - Operating System
- PCIe - PCI Express
- PD - Protection Domain
- QDR - Quadruple Data Rate
- QoS - Quality of Service
- QP - Queue Pair
- RDMA - Remote DMA
- RDS - Reliable Datagram Socket
- RPC - Remote Procedure Call
- SAN - Storage Area Network
- SDP - Sockets Direct Protocol
- SDR - Single Data Rate
- SL - Service Level
- SM - Subnet Manager
- SRP - SCSI RDMA Protocol
- TCA - Target Channel Adapter
- ULP - Upper Layer Protocol
- VCRC - Variant CRC
- VL - Virtual Lane
- WQE - Work Queue Element
- WRR - Weighted Round Robin