September 2008

InfiniBand for Storage Applications

Storage solutions can benefit today from the price, performance and high availability advantage of Mellanox's industry-standard InfiniBand products

1.0 An Overview of InfiniBand

As the I/O technology with the largest installed base of 10, 20 and 40Gb/s ports in the market (over 3.7 million ports as of June 2008), InfiniBand has clearly delivered real world benefits as defined and envisioned by the InfiniBand Trade Association (www.infinibandta.org), an industry consortium formed in 1999.

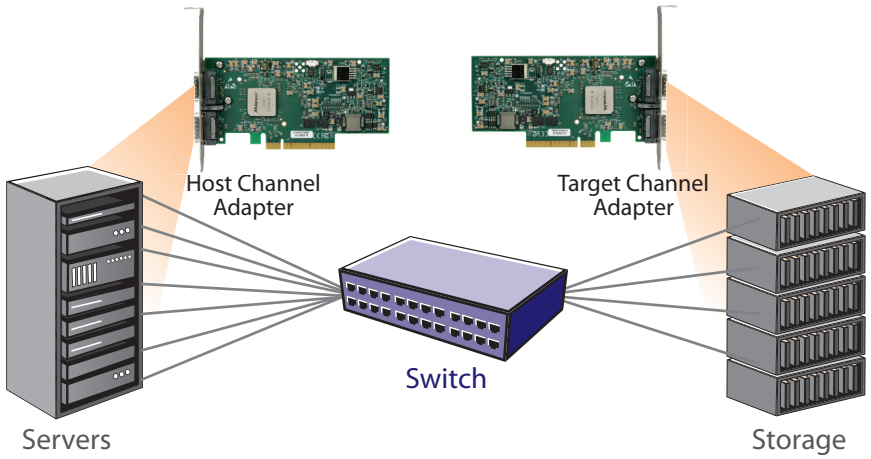
There are several factors that have enabled InfiniBand's adoption in data centers and technical compute clusters to quickly ramp and explain why it will continue to be the performance computing and storage fabric of choice.

1.1 The Basics of the InfiniBand Fabric

InfiniBand fabrics are created with host channel adapters (HCA) and target channel adapters (TCA) that fit into servers and storage nodes and are interconnected by switches that tie all nodes together over a high-performance network fabric.

The InfiniBand Architecture (IBA) is a fabric designed to meet the following needs:

- High bandwidth, low-latency computing, storage and management over a single fabric
- Cost-effective silicon and system implementations with an architecture that easily scales from generation to generation
- Highly reliable, available and scalable to tens-of-thousands of nodes
- Exceptionally efficient utilization of compute processing resources
- Industry-standard ecosystem of cost-effective hardware and software solutions



InfiniBand is the only fabric that meets all of these criteria. Standards based InfiniBand server-to-server and server-to-storage connections today deliver up to 40Gb/s of bandwidth. InfiniBand switch-to-switch connections deliver up to 120Gb/s. This high-performance bandwidth is matched with world-class application latency performance of 1µs and switch latencies of 100ns per hop that enable efficient scale-out of compute and storage systems.

With a true cut-through forwarding architecture and well defined end-to-end congestion management protocol, InfiniBand defines the most cost-effective and scalable I/O solutions in the market. A single switch silicon device supports up to thirty-six 10Gb/s, 20 or 40Gb/s InfiniBand ports, which equates to nearly three terabit per second of aggregate switching bandwidth.

Switches and adapters support up to 16 virtual lanes per link to enable granular segregation and prioritization of traffic classes for delivering Quality of Service (QoS). With integrated SerDes on all ports, InfiniBand is generations ahead of other switching solutions and has enabled the industry's densest switching systems (up to 3456 ports in a single chassis), significantly reducing the cost per port for large fabrics.

InfiniBand also defines an industry-standard implementation of remote direct memory access (RDMA), protocols and kernel bypass to minimize CPU overhead allowing computing resources to be fully used on application processing rather than network communication.

InfiniBand is clearly driving the most aggressive performance roadmap of any I/O fabric, while remaining affordable and robust for mass industry adoption.

1.2 Industry Standard

The importance of an open industry-standard specification for IBA cannot be understated. By gaining acceptance from industry-leading solution providers from its initial inception, InfiniBand has garnered wide support for both hardware and software-based solutions. All major server vendors in the industry are shipping InfiniBand PCI-X and PCI Express adapters and embedded solutions for their Blade Server architectures. PCI Express Gen2 InfiniBand adapters are now available and are compatible with industry leading PCIe Gen2 capable servers and storage platforms.

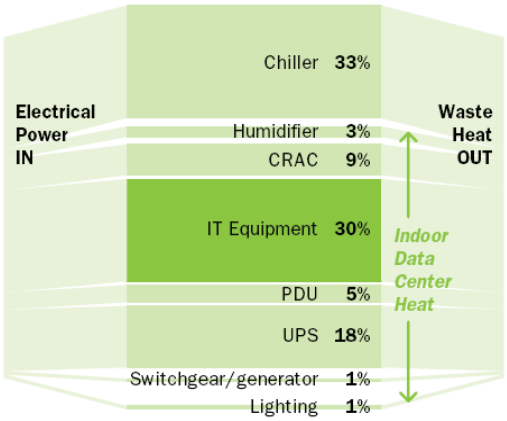
InfiniBand is also widely used in embedded and communication applications and is becoming the ubiquitous performance fabric. Several InfiniBand solutions are available in industry-standard form factor chassis that houses these applications such as VME and Advanced TCA. Other specialized chassis take advantage of the performance of InfiniBand fabrics for networking, industrial, medical and military applications.

1.3 The InfiniBand Power Advantage

Mellanox's InfiniBand adapters not only provide a low cost, high-performance interconnect solution, they also require very low power, less than 6W per 40Gb/s InfiniBand port. Coupled with high-performance and the ability to consolidate clustering, networking and storage, a single InfiniBand adapter can replace multiple legacy Clustering, Ethernet and Fibre Channel adapters to provide significant power saving to the data center. These advantages are making InfiniBand a vital interconnect for server blades.

InfiniBand has a compelling price / performance advantage over other I/O fabrics. A low cost single 20Gb/s port Mellanox InfiniBand HCA can meet the performance needs of up to ten 4Gb/s Fibre Channel HBAs or Sixteen Gigabit Ethernet NICs.

1.4 The InfiniBand Cost Advantage



InfiniBand and Fibre Channel energy costs

Note: Total power calculated based on formulas in the "Guidelines for Energy Efficient Data Centers" www.thegreengrid.org. Energy costs based on $0.10 per kWh

	Typical Watts per Adapter	Adapters required	Watts per system	Total power*	kWh / Year	Energy cost over 5 years
20Gb/s IB HCA	4	1	4	13	117	$58.34
4Gb/s FC HBA	5	5	25	83	729	$364.64

Coupled with low power consumption and the ability to consolidate network, storage and clustering onto a single fabric, InfiniBand provides not only significant savings in initial purchase, but also considerably reduces system management overheads and energy costs.

1.5 The InfiniBand Performance Advantage

One of the key reasons that data centers are deploying industry-standard InfiniBand is the total application level performance the fabric enables. First, InfiniBand is the only shipping solution that supports 40Gb/s host connectivity and 120Gb/s switch to switch links. Second, InfiniBand has world-class application latency with measured delays of 1µs end to end. Third, InfiniBand enables the most efficient use of all of the processors and memory in the network by offloading all of the data transport mechanisms in the adapter card and reducing memory copies. These three metrics combine to make InfiniBand the industry's most powerful interconnect.

The performance benefits are echoed in the trends of the Top500.org list that tracks the world's most powerful supercomputers. Published twice a year, this list is increasingly used as an indication of what technologies are emerging in the clustered and supercomputing arena.

1.6 InfiniBand Software Solutions

OPENFABRICS ALLIANCE

InfiniBand has garnered support from every mainstream operating system including Linux, Windows, Solaris, HPUX, AIX, BSD, VMware and VxWorks.

Open source and community-wide development of interoperable and standards-based Linux and Windows stacks are managed through the OpenFabrics Alliance. This alliance, consisting of solution providers, end-users and programmers interested in furthering development of the Linux or Windows stacks, has successfully driven InfiniBand support into the Linux kernel and gained WHQL qualification for Microsoft's Windows Server. The successful inclusion of InfiniBand drivers and upper layer protocols in the Linux kernel insures interoperability between

different vendor solutions and will ease the deployment of InfiniBand fabrics in heterogeneous environments.

Server virtualization increases the I/O demands on host servers to meet the requirements of multiple guest operating systems. The Open Fabrics OFED source code is used as the base for both Xen and VMware community lead virtualization solutions and InfiniBand is supported in the latest VMware ESX 3.5 release. The high-performance, low latency and I/O channel-based communication available in the InfiniBand architecture is perfectly suited for I/O virtualization enabling unprecedented levels of resource utilization and flexibility in the allocation of compute, storage and network resources based on dynamic data center demands.

From an application point of view, InfiniBand has support for a plethora of applications in both enterprise and high-performance computing environments. In the enterprise environment, InfiniBand is being used for grid computing and clustered database applications driven by market leaders Oracle and IBM DB2 for retail, enterprise resource planning and customer relationship management. In the commercial high-performance computing field, InfiniBand provides the fabric connecting servers and storage to address a wide range of applications including oil and gas exploration, automotive crash simulations, digital media creation, fluid dynamics, drug research, weather forecasting and molecular modeling just to name a few.

1.7 InfiniBand's Growing Role in the Data Center

Data centers simultaneously run multiple applications and need to dynamically reallocate compute resources between applications depending on end user workload. To meet these needs the network fabric must seamlessly support compute, storage, inter-process communication, and management traffic.

The emergence of virtual and grid computing solutions in addition to robust software solutions have set the stage for mass deployment of InfiniBand in business and utility computing environments.

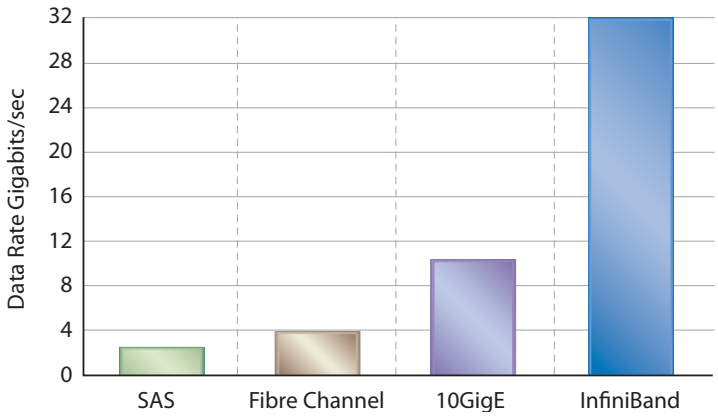
Industry-standard InfiniBand has the performance, proven reliability, manageability and widely available software solutions making it ready for prime time.

2.0 InfiniBand Storage

There are many performance sensitive applications that are driving the need for InfiniBand storage:

- Backup / Diskless Backup
- Server Clustering
- Replication / Snapshot / Data Check-pointing
- Streaming Video / Graphics
- Clustered Storage for Disaster Recovery
- Online Transaction Processing
- Data Warehousing

Compared to alternative storage interconnect technologies like Fibre Channel, InfiniBand offers significant performance and price/performance improvements. This translates into real world customer advantages such as increased application performance, reduced backup times, greater system consolidation, lower power consumption and lower total cost of ownership (TCO).

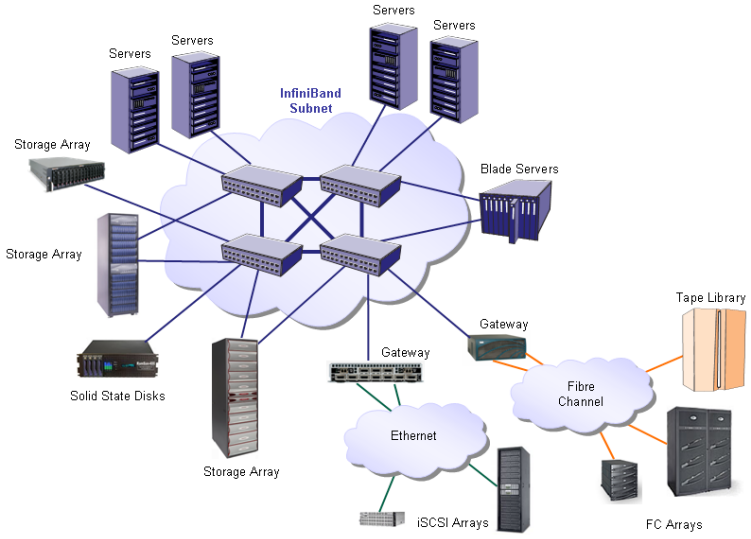


2.1 Storage Area Networking

The benefits of Storage Area Networks are well known and include:

- High Availability
- Increased Scalability
- Storage Consolidation
- Improved Manageability
- Centralized Backup

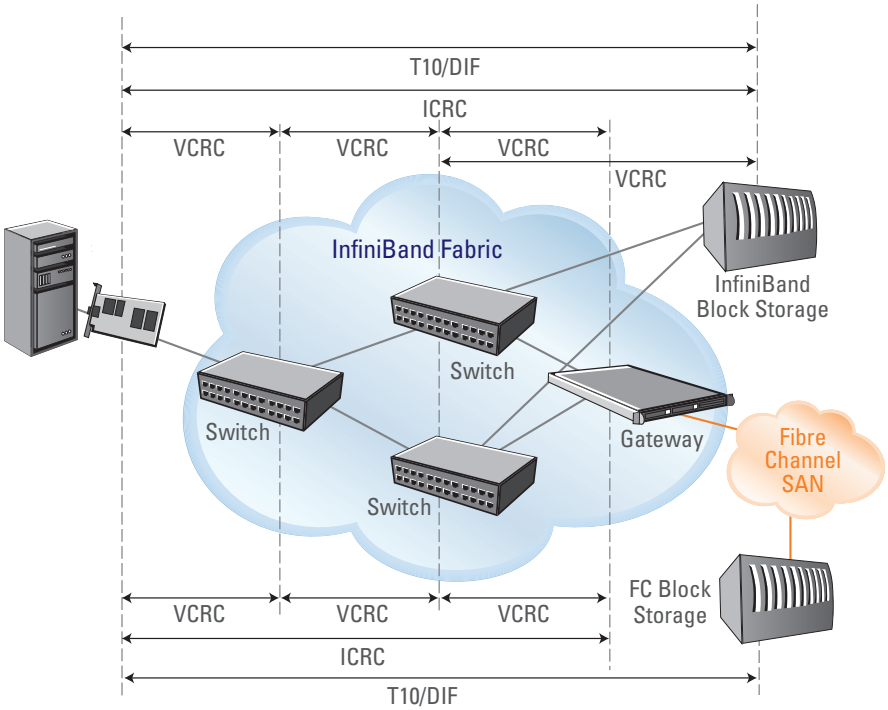
With proven reliability, scalability and 40Gb/s performance along with ultra low latencies, RDMA, QoS and OS kernel bypass, InfiniBand excels at meeting these needs and is set to become the interconnect standard for Storage Area Networks (SANs).



InfiniBand Storage Area Networks can be seamlessly implemented, while protecting previous investments in legacy Fibre Channel, iSCSI and NAS storage devices by using IB to FC and IB to IP gateway products from leading vendors like Cisco, Qlogic and Voltaire.

2.2 Data Integrity

The mission critical enterprise requires that data is correctly written to the storage device and has not been corrupted as it travels from server to storage media. InfiniBand enables the highest levels of data integrity by performing cyclic redundancy checks (CRCs) at each fabric hop and end to end across the fabric to ensure the data is correctly read and written between server and storage device.



Data on storage devices can be fully protected by application level standards such as T10/DIF (Data Integrity Feature), which adds 8 bytes of protection data to each storage block including a CRC field and a reference tag to ensure the correct data is written to the right location. Mellanox's ConnectX HCA fully offloads the DIF calculations to maintain peak performance.

2.3 High Availability and Redundancy

With the cost of losing access to data estimated at $100,000 an hour by International Data Corporation (IDC), the enterprise data center demands continuous availability of both applications and data.

InfiniBand is perfectly suited to meet the mission critical needs to today's enterprise data centre by enabling fully redundant I/O fabrics, with automatic path failover and link layer multi-pathing abilities to meet the highest levels of availability.

Each InfiniBand subnet is managed to control configuration and maintenance including error reporting, link failover, chassis management and other services to ensure a solid fabric.

Mellanox InfiniBand adapters are also available with multiple ports which provide greater protection against link level failures.

2.4 Storage Protocols

There are 2 block level protocols available for storage applications on InfiniBand:

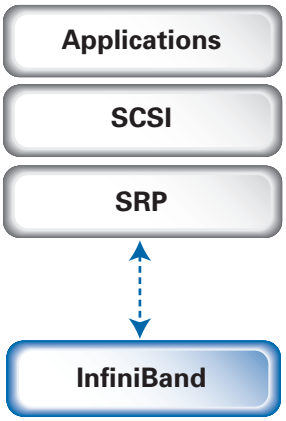
- SRP (SCSI RDMA Protocol)

- iSER (iSCSI Extensions for RDMA)

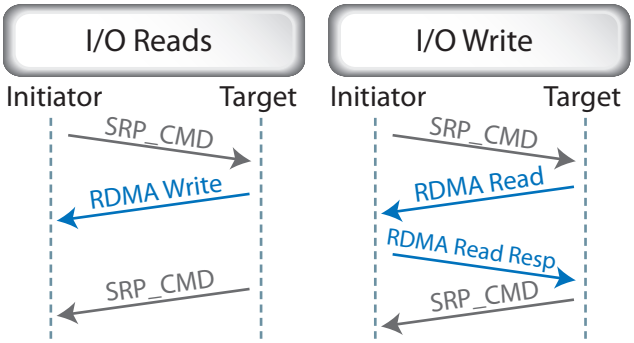
The NFSoRDMA protocol provides high-performance file level storage access on InfiniBand

2.4.1 SRP

SRP is an open ANSI T10.org standard protocol that encapsulates SCSI commands and controls data transfer over RDMA capable fabric such as InfiniBand.



SRP operates at the kernel level and uses RDMA to move data directly between the memory of computer system and storage device without involving the operating system. This enables high-performance, zero copy, low-latency communications between servers and storage systems.

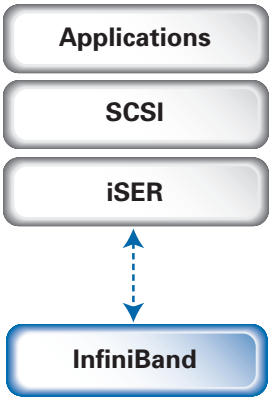


Performance measurements with the SRP protocol over InfiniBand 20Gb/s with PCIe Gen2 have demonstrated the following impressive results using 1MB I/O transfers.

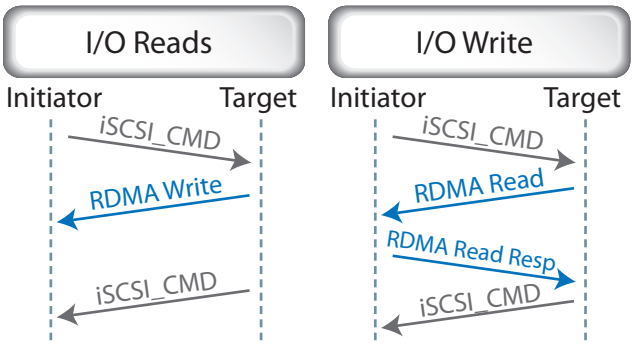
- I/O Read 1.8 Gigabytes / sec
- I/O Write 1.6 Gigabytes / sec

2.4.2 iSER

iSER is an IETF standard that maps the block level iSCSI protocol over RDMA capable fabrics such as InfiniBand. Like SRP this permits data to be transferred directly between server and storage devices without intermediate data copies or involvement of the operating system.



One key issue with traditional iSCSI is the handling of out of order TCP segments which have to be reassembled before the data can be used. This reassembly is very CPU intensive. The iSER protocol enables direct data placement of both in order and out of order data and eliminates the need for reassembly.

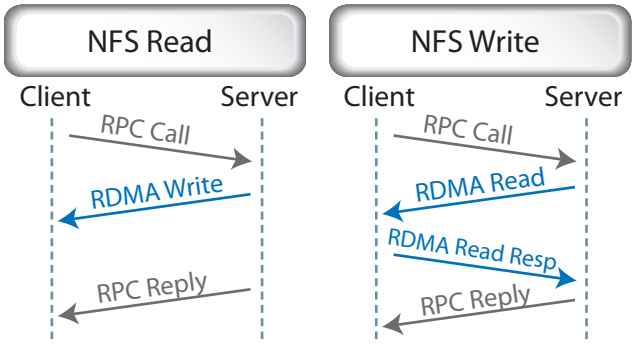


Initiators and target drivers for both SRP and iSER are also available from the OpenFabric Alliance OpenFabrics Enterprise Distribution (OFED) drivers set.

2.4.3 NFS over RDMA (NFSoRDMA)

NFS is network attached storage file access protocol layered on the RPC (remote procedure call) systems and is typically carried over UDP/TCP networks. NFS has been widely adopted for several years as a distributed file system. However NFS moves large chucks of data and incurs many copies with each transfer which inefficiently use system memory and CPU resources.

NFSoRDMA is an IETF RFC that leverages the benefits of RDMA to offload the protocol processing and avoid data copies to significantly lower CPU utilization and increase data throughput. NFSoRDMA is already part of the Linux Kernel and will be part of commercial Linux distributions in the first half of 2009



2.5 Storage Products

Storage vendors are quickly introducing solutions based on InfiniBand to address the strong market demand led by financial and clustered database applications as well as parallel technical computing applications. Both block level and file level storage solutions are available in the market today.

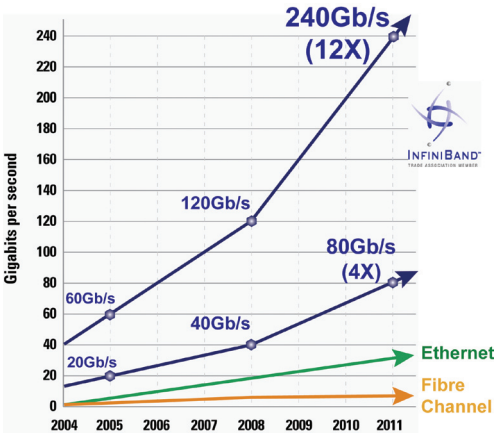
3.0 Fabric Convergence

As InfiniBand continues rapid adoption in both technical computing and data center environments, and computing resources are able to handle higher data processing throughputs, storage access has now become a bottleneck. Legacy interconnects (Fibre Channel and Ethernet) are insufficient. As a result, the industry is demanding storage solutions instead connect directly into the 40Gb/s InfiniBand fabric. The convergence of computing and storage access onto a single InfiniBand fabric has significant infrastructure investment savings and total cost of ownership (TCO) benefits, in addition to improved overall storage access performance.

4.0 InfiniBand Roadmap

One of the early goals of the IBA was to define a fabric protocol that is cost effective to implement in silicon and system solutions while being easily scalable from generation to generation. Vendors are successfully meeting the milestones on what is the most aggressive I/O technology roadmap in the industry and are scheduled to introduce 80Gb/s node links and 240Gb/s switch links in 2010. This exceptional roadmap will continue to depend on industry-standard, commodity based components to enabling world-class performance at mainstream interconnect prices. At the same time, these solutions will remain backward compatible to the eco-system being developed and deployed today.

Mellanox's InfiniBand adapters are perfectly aligned with the roadmap for PCI Express. Today servers ship with PCI Express Gen2 x8 slots capable of transmitting 40Gb/s and receiving 40Gb/s of data simultaneously (40+40Gb/s).



Number of IB Lanes	Per Lane Bandwidth		
	SDR 2.5Gb/s	DDR 5Gb/s	QDR 10Gb/s
4X	10Gb/s	20Gb/s	40Gb/s
8X	20Gb/s	40Gb/s	80Gb/s
12X	30Gb/s	60Gb/s	120Gb/s

5.0 InfiniBand – Proven for Real World Deployments Today

Since its introduction in the early 2000s, InfiniBand technology has matured and is emerging as the preferred fabric for enterprise data center and performance compute cluster deployments. Shipping in production today with 10, 20 and 40Gb/s adapter and switch solutions capable of up to 120Gb/s switch to switch links, the InfiniBand fabric is delivering on the value proposition defined by the industry back in 1999.

With virtually every server vendor shipping InfiniBand solutions, availability of native InfiniBand storage systems, far reaching operating system support, and a wide variety of enterprise and technical computing applications, all of the pieces are in place for continued mass market deployment of InfiniBand fabrics.

6.0 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

7.0 Reference

InfiniBand Architecture Specification Volume 1-2 Release 1.2
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 Specification
http://www.ietf.org/html.charters/ips-charter.html

SCSI RDMA Protocol, DIF
www.t10.org

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

Mellanox TECHNOLOGIES
2900 Stender Way, Santa Clara, CA 95054
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com