An Oracle Technical White Paper
May 2012

Delivering Application Performance with Oracle’s InfiniBand Technology

A Standards-Based Interconnect for Application Scalability and Network Consolidation

Introduction

This document serves as an introduction to InfiniBand technology, with a focus on its ability to scale application performance and consolidate network infrastructure.

For applications that require a high-performance and scalable interconnect, InfiniBand technology can provide significant advantages over existing technologies such as 10 gigabit Ethernet (GbE) and Fibre Channel:

- InfiniBand is an open and standard technology that provides a compelling OpenFabrics software stack.
- Standard upper-level protocol support means that existing applications can use InfiniBand fabrics transparently.
- Because InfiniBand is a lossless fabric with embedded mechanisms for service differentiation, it is ideal for helping organizations consolidate their network infrastructure.
- InfiniBand provides very low compute overhead as well as the industry's lowest latency and highest throughput — all key for application scalability and server virtualization.

Oracle recognizes the importance of InfiniBand technology in the future of the data center, and has invested heavily in InfiniBand competency. In fact, InfiniBand is at the core of Oracle’s engineered systems offerings.

Challenges with Clustered Approaches

Clustered applications — distributed databases, Web 2.0 applications, compute clusters, storage clusters, and cloud computing — are here to stay. While these clustered approaches have solved fundamental problems, issues remain:

- With the growth in data, and the necessity of moving data between multiple systems, I/O infrastructure now frequently represents the performance bottleneck.
- The physical complexity of multiple interconnects per system has hobbled many data centers with multiple cabling systems, multiple switches, excess power consumption, and rising management costs.
- The computational overhead for software-based protocols is becoming burdensome, particularly as it cuts into available computational capacity.
- While traditional network and I/O interconnect technologies have served well, they usually provide insufficient bandwidth and excessive latency, and they can represent inflexible topologies that don't serve the needs of the modern data center.

The inherent advantages of InfiniBand technology make it ideal to address these issues, both for the largest supercomputers and for enterprise computing. While InfiniBand, 10 GbE, and Fibre Channel will continue to coexist, InfiniBand is ideal for a range of modern applications, including the following:

- Parallel database scaling
- Virtualized cloud infrastructure
- Computational HPC clustering
- Storage connectivity
- Transactional systems with low-latency requirements, such as financial trading systems and reservation systems

Application Scalability

As applications have scaled beyond individual systems, interconnect technology has become key to providing scalable application performance. Not only do applications need more bandwidth and lower latency, but the level of software protocol processing has grown with the level of interconnect traffic. Though some have sought third-party TCP offload processors and similar approaches, the overhead of software-based protocol processing cuts into available compute cycles available for applications. This burden can be especially pronounced for systems employing virtualization technology.

Not only does InfiniBand offer high bandwidth and low latency, but InfiniBand technology can also be key for providing additional application scalability. Unlike traditional software-based transport protocol processing, InfiniBand provides hardware support for all of the services required to move data between hosts. By providing reliable end-to-end transport protocols in hardware and user-level Remote Direct Memory Access (RDMA) capability, interconnected systems have more cycles available for processing, and scalable applications can require fewer numbers of systems.

In particular, RDMA provides efficient and direct access to host or client memory without involving processor overhead. Because InfiniBand supports a broad range of upper-level protocols, it lends RDMA capabilities to protocols such as IP, iSCSI, NFS, and others. InfiniBand also provides quality of service (QoS) mechanisms that can help ensure application performance. Finally, InfiniBand is the only interconnect that provides congestion control — essential for cluster scalability from hundreds to thousands of nodes.

Not only does InfiniBand perform well today, but the InfiniBand Trade Association (IBTA) has also established a performance roadmap to accommodate future demands for bandwidth. InfiniBand links consist of groups of lanes (1x, 4x, and 12x) that support a number of different data rates. Supported data rates and link widths are given in Table 1.

TABLE 1. INFINIBAND DATA RATES

DATA RATES	PER-LANE AND PER-LINK BANDWIDTH		
	1X	4X	12X
Single Data Rate (SDR): 2.5 Gb/sec	2.5 Gb/sec	10 Gb/sec	30 Gb/sec
Dual Data Rate (DDR): 5 Gb/sec	5 Gb/sec	20 Gb/sec	60 Gb/sec
Quad Data Rate (QDR): 10 Gb/sec	10 Gb/sec	40 Gb/sec	120 Gb/sec
Fourteen Data Rate (FDR): 14 Gb/sec	14 Gb/sec	56 Gb/sec	168 Gb/sec
Enhanced Data Rate (EDR): 26 Gb/sec	26 Gb/sec	104 Gb/sec	312 Gb/sec
High Data Rate (HDR): TBD Gb/sec	TBD Gb/sec	TBD Gb/sec	TBD Gb/sec

Figure 1 illustrates the IBTA InfiniBand roadmap with an anticipated time line for these data rates. While InfiniBand has a credible scalability story into the future, actual technology availability is dependent on many factors, including market demand for additional bandwidth. More information on the InfiniBand physical layer is provided later in this document. (See the section titled "Physical Medium.")

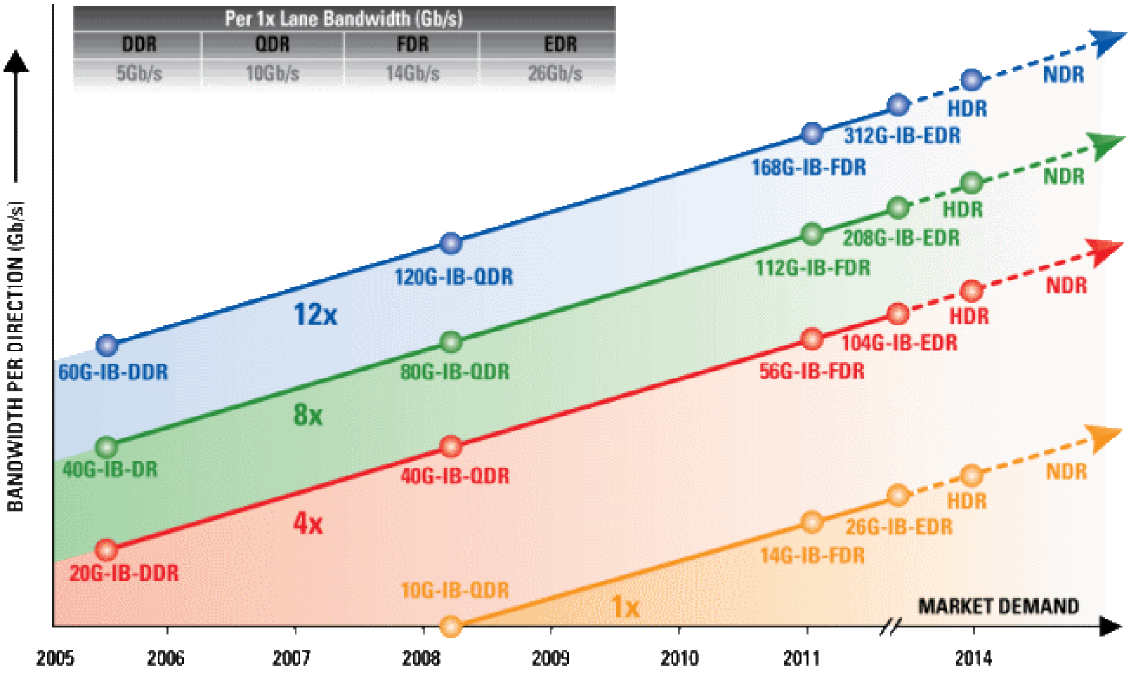


Figure 1. InfiniBand Trade Association speed roadmap.

Transparent Application Compatibility

In contemplating any new interconnect fabric, application compatibility is absolutely key. InfiniBand was designed from the outset to ease application deployment. For example, IP and sockets-based applications can benefit from InfiniBand performance without change. InfiniBand employs the OpenFabrics protocol stack (www.openfabrics.org), as shown in Figure 2. This software stack, or similar stack, is supported on all major operating systems including Oracle Solaris, Oracle Linux, and Microsoft Windows.

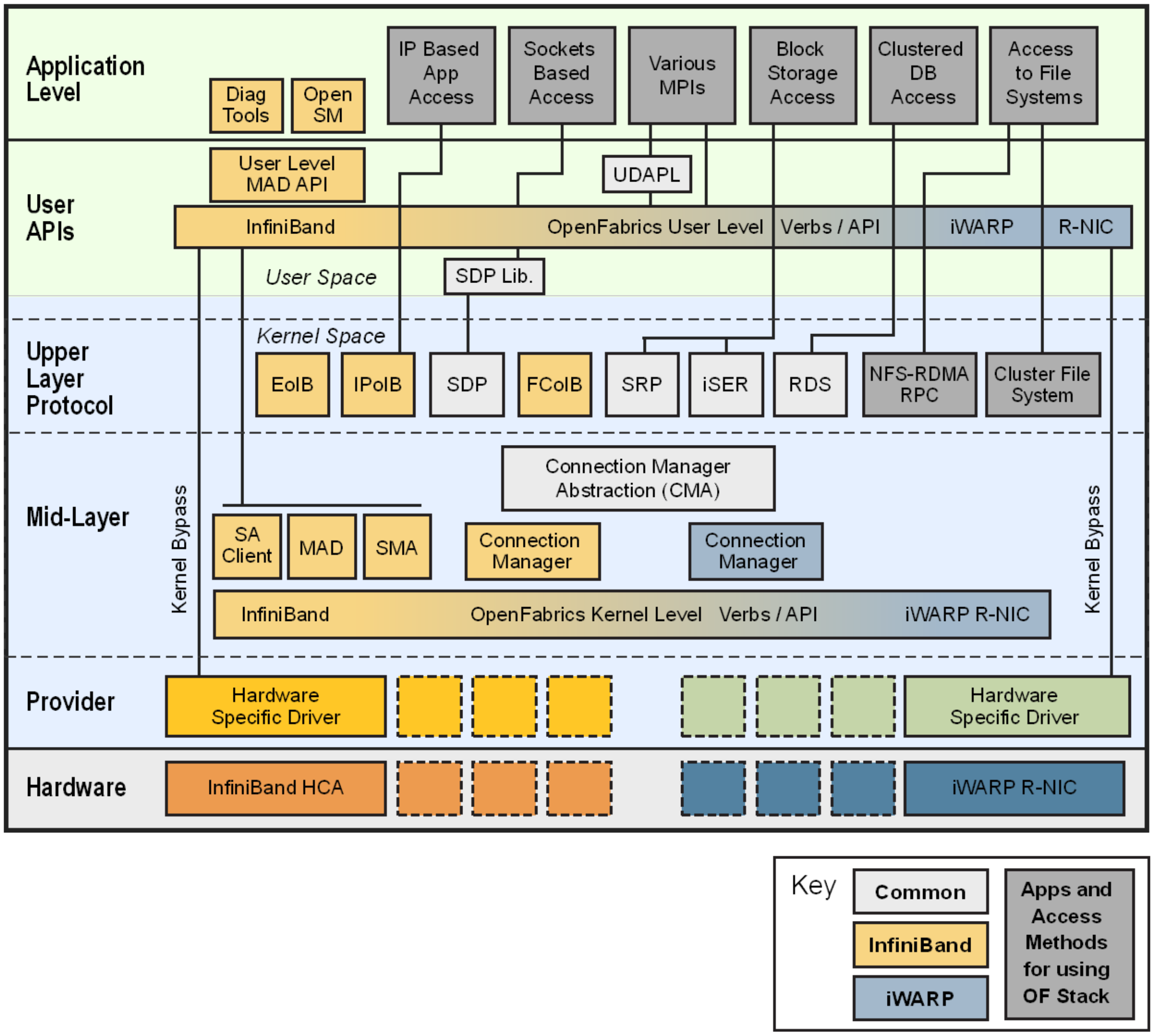


Figure 2. The OpenFabrics software stack.

Importantly, all InfiniBand functionality is provided in Host Channel Adapter (HCA) hardware, and the host CPU is not interrupted or used for InfiniBand transport. The HCA typically connects through a PCI Express (PCIe) interface. To access this hardware protocol support in the HCA, applications have a choice of protocols. For instance, distributed applications that leverage MPI can write directly to the InfiniBand Verbs API (VAPI) or they can use a user Direct Access Programming Library (uDAPL).

The OpenFabrics stack provides significant upper-level protocol support that lets existing applications take advantage of InfiniBand's hardware acceleration. Key upper-level protocol support includes the following:

- *IPoIB (IP over InfiniBand)* — IPoIB provides a simple encapsulation of IP over InfiniBand, offering a transparent interface to any IP-based applications.
- *SDP (Sockets Direct Protocol)* — For applications that use TCP sockets, the Sockets Direct Protocol provides a high-performance interface that bypasses the software TCP stack. SDP implements zero copy and asynchronous I/O, and it transfers data using RDMA and hardware-based transfer mechanisms.
- *EoIB (Ethernet over InfiniBand)* — The Ethernet over InfiniBand protocol is a network interface implementation over InfiniBand. EoIB encapsulates layer-2 (L2) datagrams over an InfiniBand Unreliable Datagram (UD) transport service. The InfiniBand UD datagrams encapsulates the entire Ethernet L2 datagram and its payload. The EoIB protocol also enables routing of packets from the InfiniBand fabric to a single GbE or 10 GbE subnet.
- *SRP (SCSI RDMA Protocol)* — The SCSI RDMA Protocol was designed to provide block store capabilities by tunneling SCSI request packets over InfiniBand.
- *iSER (iSCSI over RDMA)* — iSER eliminates the traditional iSCSI and TCP bottlenecks by enabling zero-copy RDMA to offload CRC calculations in the transport layer and by working with message boundaries instead of streams.
- *Network File System (NFS) over RDMA* — NFS over RDMA extends NFS to take advantage of the RDMA features of InfiniBand and other RDMA-enabled fabrics.
- *Reliable Datagram Sockets (RDS)* — Reliable Datagram Sockets provides reliable transport services and can be used by applications such as Oracle Real Application Clusters (Oracle RAC) for both interprocess communication and storage access.

Oracle supports the full suite of InfiniBand upper layer protocols (ULPs) in both Oracle Linux and Oracle Solaris. In addition, to enable high-performance storage, Oracle has implemented NFS/RDMA, iSER, and SRP targets in Oracle's Sun ZFS Storage Appliance systems. Sun ZFS Storage Appliance systems are leading enterprise-class storage systems with extensive InfiniBand software and hardware support for both file and block storage.

Oracle leverages the performance advantages of InfiniBand ULPs in its engineered systems including, Oracle Exadata Database Machine, Oracle SPARC SuperCluster, Oracle Big Data Appliance, and Oracle Exalogic Elastic Cloud. Oracle invented the RDS protocol to support high performance and scalability with Oracle RAC. RDS is used extensively to deliver extreme IPC and storage performance in Oracle Exadata Database Machine. Oracle Exalogic Elastic Cloud leverages kernel bypass for high-performance messaging in Oracle WebLogic Server clusters with SDP. Furthermore, optimized application–to-database tier communication is enabled by GridLink, which is also supported by SDP.

EoIB enables the InfiniBand backplane within Oracle Exalogic Elastic Cloud and Oracle Big Data Appliance to function as a converged I/O fabric, carrying client traffic, storage traffic, and interprocess communication over a single, unified fabric.

Virtualization

Many organizations are now embracing server virtualization for the improved utilization and enhanced flexibility it brings. To make the most of available resources, virtualization technology is allowing substantial consolidation of compute systems, and yielding considerable deployment agility — letting organizations deploy and redeploy multiple operating systems and applications on the same physical hardware. At the same time, many are also realizing that I/O can represent a very real bottleneck that constrains storage bandwidth in virtualized systems. In addition, significant CPU overhead and latency can result as the virtualized host attempts to deliver I/O to the various running guest operating systems and their resident applications. Even as available bandwidth grows, this overhead can quickly overwhelm even fast buses and processing resources.

Server virtualization environments, such as Oracle VM, allow multiple virtual machines (VMs) to be deployed onto a single physical system. Each virtual machine runs an independent operating system. In each operating system instance, virtual Network Interface Controllers (NICs) and/or virtual storage interfaces are configured through the hypervisor, and each is associated with a physical NIC or physical storage interface in the system.

The PCI-SIG (www.pcisig.org) has recently added I/O Virtualization (IOV) extensions to the PCIe specification suite to let multiple virtual hosts share resources in a PCIe device (for example, a NIC, HCA, or HBA). InfiniBand complements this functionality by providing more fine-grained resource sharing through channel-based I/O with inherent isolation and protection. In addition, hypervisor and virtualization stack offloads save data copies and reduce context switching. This approach allows direct I/O access from the virtual machine and matches native InfiniBand performance.

Guest I/O in Virtualized Servers

Two I/O models are available to virtual servers resident on InfiniBand attached physical servers:

- Hardware sharing — Direct hardware access leveraging hardware and software supporting SR-IOV
- Software sharing — Software-mediated sharing of hardware through Virtual Bridging

Hardware Sharing: SR-IOV

Virtualized environments employing SR-IOV allow virtual servers to take full advantage of the underlying InfiniBand fabric. Each virtual server has the InfiniBand stack instantiated and applications can capitalize on features of the InfiniBand architecture and RDMA-enabled protocols to enhance performance. It is not required that all applications be "InfiniBand-aware" because InfiniBand provides support for IP and Ethernet via the IPoIB and EoIB protocols. Thus, high-performance I/O services are available to both InfiniBand-aware and legacy applications.

Hardware supported by SR-IOV exposes "Virtual Functions" (VF) enabling native InfiniBand access for guest virtual servers. Each of these InfiniBand virtual functions represents a lightweight HCA with its own GID. Virtual functions share a physical HCA, physical port, LID, and LMC. The Physical Function (PF) is hosted in Dom0 and is responsible for dynamically allocating resources (P_Keys, QPs, CQs, SRQs, memory regions, and so on) to the VFs. The PF owns QP0 and virtualizes QP1, making it available to guest virtual servers through the HCA mbox channel. Figure 3 illustrates the interfaces to the virtual servers provided by SR-IOV and the association of the PF driver in Dom0 with the HCA resources.

Live migration for InfiniBand-aware applications is not supported at this time, because InfiniBand emulators and paravirtualized interfaces for InfiniBand are not presently available.

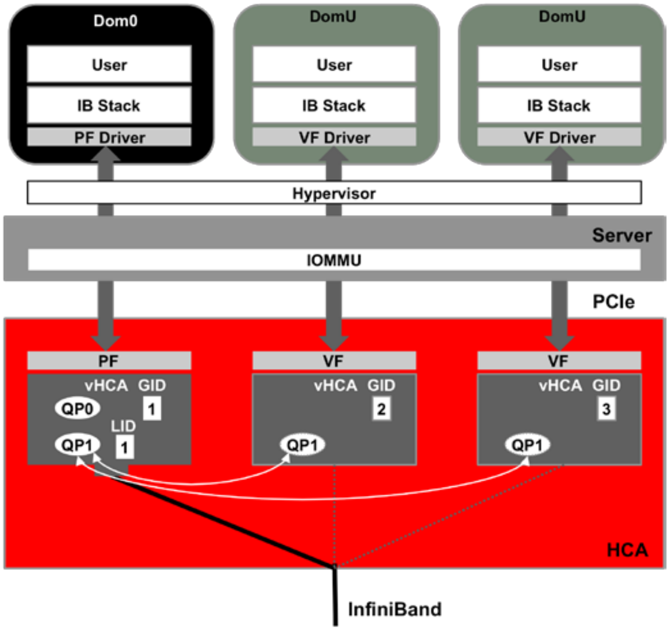


Figure 3. Virtual servers can access the InfiniBand hardware through SR-IOV.

Oracle VM Server supports InfiniBand when deployed within Oracle Exalogic Elastic Cloud. Oracle VM utilizes hardware sharing, leveraging SR-IOV to make all the performance and offload capabilities afforded by InfiniBand available to applications resident on Oracle Exalogic servers. Hardware-enabled I/O virtualization in Oracle Exalogic eliminates potential bottlenecks inherent in other I/O virtualization schemes where all virtual machine I/O must be proxied by a process residing in Dom0 that manages the physical function.

Software Sharing: Virtual Bridging

Virtualized environments employing Virtual Bridging allow virtual servers to be fabric agnostic. Virtual servers (DomU) use an emulated, generic Ethernet NIC (e1000) for server I/O. The only burden on the guest OS is that it must provide driver support for the e1000. The emulated e1000 communicates with the InfiniBand HCA through a Virtual Bridging function, which resides in Dom0.

The InfiniBand driver stack is instantiated only in Dom0. The InfiniBand hardware and driver stack provides virtualized "Ethernet services" to the guest virtual servers by attaching the Ethernet over InfiniBand (EoIB) driver to the Virtual Bridge (virbr). All host-to-host and host-to-gateway (exposed hardware Ethernet port on the Sun Network QDR InfiniBand Gateway Switch for external LAN connectivity) communication within an InfiniBand cluster that supports server virtualization via Virtual Bridging uses the EoIB protocol. Guest-to-guest communication for guests residing on the same physical server is facilitated by the Virtual Bridge. Figure 4 illustrates the interfaces to the virtual servers provided by Virtual Bridging and the association of the InfiniBand stack in Dom0 with the physical HCA.

Applications running on virtual servers in an environment supported by Virtual Bridging cannot leverage native InfiniBand features or RDMA-enabled protocols. Though Virtual Bridging results in lower I/O performance due to the fact that all virtual server I/O must be proxied by a single software stack in Dom0, emulation offers the benefits of wide OS support and enables transparent live migration.

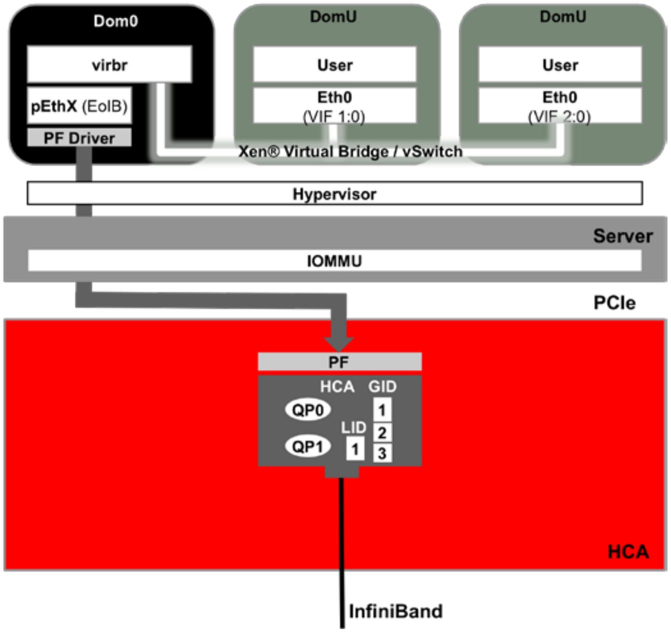


Figure 4. Virtual Bridging allows guests that are not InfiniBand-aware to transparently share InfiniBand resources.

Fabric Convergence and Consolidation: Virtualizing the Interconnect

The modern enterprise data center is at a major crossroads. The drive to solve more complex and demanding business problems has led to the deployment of ever more systems to meet the needs of business. As more and more systems, cables, and switches are added, the data center has become crowded, complex, hot, and power-consumptive. Ironically, each well-intended technology addition can actually deprive the organization of future flexibility.

Much of this problem is driven by having multiple, redundant interconnects connected to each server. For example, an individual server might be simultaneously connected to a cluster interconnect for computational activities, an Ethernet network for management and LAN access, and a Fibre Channel SAN for storage access (Figure 5). In addition to having multiple Fibre Channel HBAs, InfiniBand HCAs, and Ethernet NICs, each of these networks and interconnects brings along its own dedicated cabling, switch boxes, and management infrastructure. The result is often a bulky, complex, and tangled infrastructure that is very difficult to scale and change.

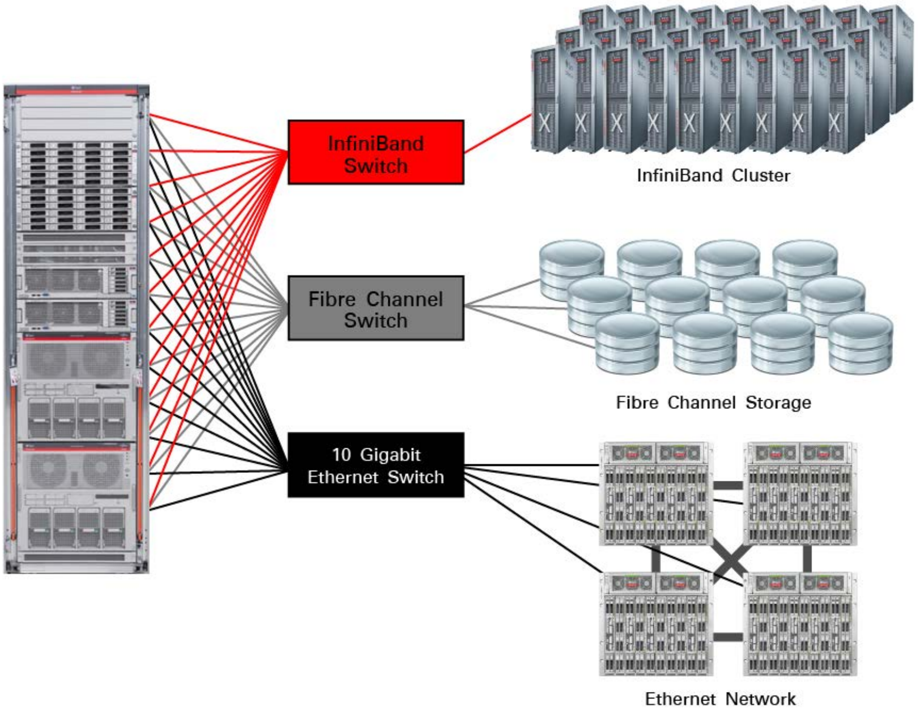


Figure 5. Many data centers now have redundant networks and interconnects.

Due to the complexities described, organizations have started to look into fabric consolidation. Benefits of a consolidated approach include the following:

- Reduced TCO
- A single consolidated network and I/O fabric to manage
- Lower power consumption
- A simpler cable plant

InfiniBand represents a significant opportunity for consolidation because it provides the key required mechanisms, such as reliable lossless guaranteed delivery, congestion control, and service differentiation. A converged interconnect using InfiniBand can consolidate complex IP networks within the data center and eliminate storage area networks (SAN) — all without requiring changes to applications and storage.

Oracle's engineered systems, including Oracle Exadata Database Machine, Oracle Exalogic Elastic Cloud, Oracle SPARC SuperCluster, and Oracle Big Data Appliance, utilize InfiniBand as a converged fabric, aggregating storage, network, and interprocess communication over a low-latency, 40 Gb/sec InfiniBand fabric. In comparison to traditional SAN/NAS approaches, InfiniBand provides up to 4x the bandwidth for storage traffic within Oracle's engineered systems. Furthermore, RDMA allows storage I/O to be managed directly by the InfiniBand hardware, eliminating OS calls, interrupts, and buffer copies.

To support fabric convergence in high-performance cluster and cloud computing environments, the Sun Network QDR InfiniBand Gateway Switch combines the functions of an InfiniBand leaf switch with an Ethernet gateway. The unique implementation of the Sun Network QDR InfiniBand Gateway Switch will not disrupt the operations and policies of existing LAN administration, because the Ethernet interface appears to both applications and the LAN as an Ethernet NIC. Thus, InfiniBand-attached servers can connect to an Ethernet LAN using standard Ethernet semantics. No application modifications are required for applications written to use standard Ethernet. Furthermore, LAN administrators simply treat servers and virtual machines residing with an InfiniBand fabric as if they were directly connected to the LAN using Ethernet NICs.

Network Virtualization and Isolation in InfiniBand Fabrics

The InfiniBand architecture facilitates the construction of rich network environments in support of isolation and security for fabric-resident applications. The fabric enables the creation of logically isolated subclusters through InfiniBand features such as partitions and subnets while preserving those isolation features provided by Ethernet and IP for Ethernet and IP traffic carried over the InfiniBand fabric. The InfiniBand architecture provides for traffic isolation through the following:

- InfiniBand partitions and Ethernet VLANs. Partitions, like VLANs, define a set of nodes that are allowed to communicate with one another. Partitions have two types of membership, full and limited. Full members can communicate with other full members and all limited members. Limited members cannot communicate with other limited members.
- InfiniBand, Ethernet, and IP subnets:
  - Like Ethernet and IP, InfiniBand supports subnets. In order for nodes on different subnets to communicate with each other, subnets must be joined by a router function.
  - A single InfiniBand subnet can contain up to 48,000 nodes.

Figure 6 illustrates a virtualized network infrastructure, providing isolation of traffic through the use of InfiniBand partitions, Ethernet VLANs, and "virtual switches," superimposed on the InfiniBand fabric. Servers and storage devices are connected to a common InfiniBand fabric. The fabric is configured to isolate client traffic, database traffic, and storage traffic.

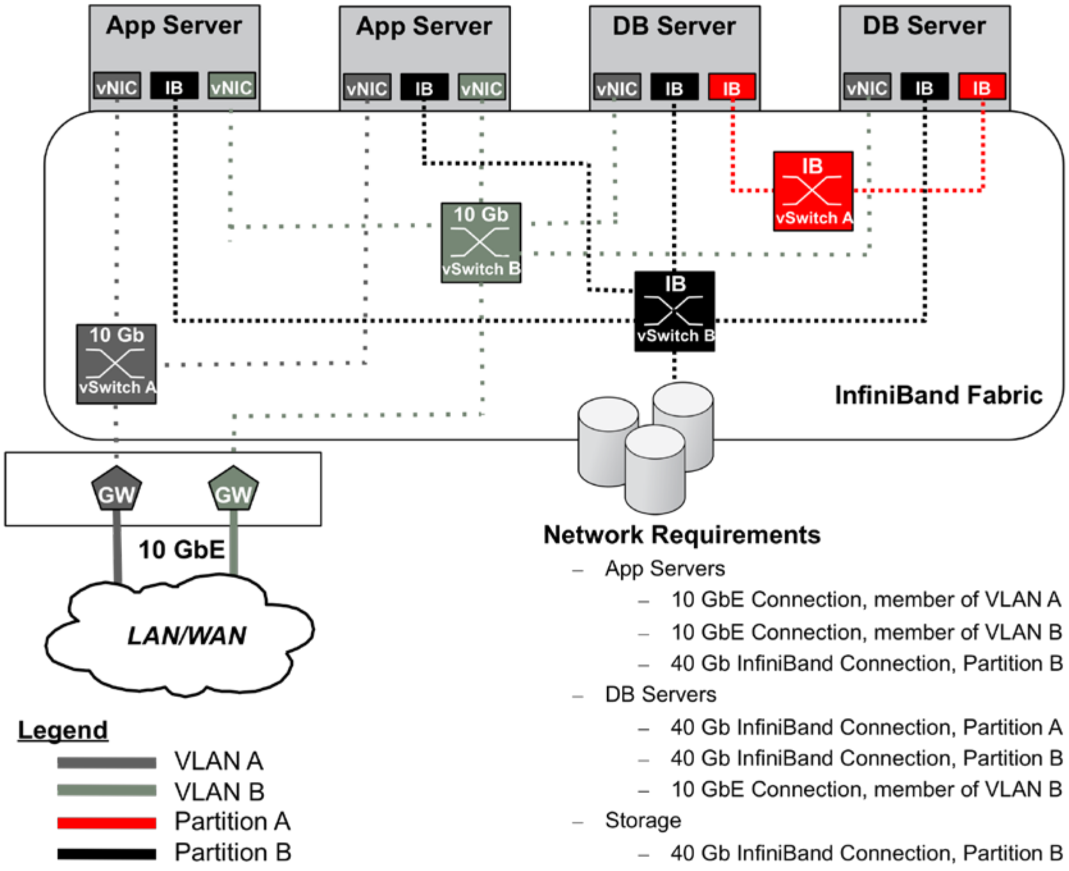

App Server
vNIC
IB
vNIC
App Server
vNIC
IB
vNIC
DB Server
vNIC
IB
IB
DB Server
vNIC
IB
IB
IB vSwitch A
10 Gb vSwitch B
IB vSwitch B
10 Gb vSwitch A
InfiniBand Fabric
GW
GW
10 GbE
LAN/WAN
Network Requirements
– App Servers
– 10 GbE Connection, member of VLAN A
– 10 GbE Connection, member of VLAN B
– 40 Gb InfiniBand Connection, Partition B
– DB Servers
– 40 Gb InfiniBand Connection, Partition A
– 40 Gb InfiniBand Connection, Partition B
– 10 GbE Connection, member of VLAN B
– Storage
– 40 Gb InfiniBand Connection, Partition B
Legend
VLAN A
VLAN B
Partition A
Partition B

Figure 6. InfiniBand partitions, Ethernet VLANs, and "virtual switches" enable a virtualized network.

Oracle's complete line of InfiniBand-enabled hardware and software products, including Oracle's family of InfiniBand switches, Host Channel Adapters (HCAs), operating systems, hypervisors, and Sun ZFS Storage Appliance systems, supports the isolation capabilities afforded by InfiniBand partitions and IP subnets. InfiniBand fabrics constructed with Oracle's InfiniBand-enabled products support secure isolation for multitenant deployments.

Network Services

All network services provided in an enterprise data center can be made available to applications running within the InfiniBand fabric. For source-destination connections that lie within the InfiniBand fabric, the fabric provides hardware-based, layer-2 services for InfiniBand as well as software-based services for Ethernet and IP. Table 2 outlines the network services provided through Oracle's InfiniBand fabric.

TABLE 2. NETWORK SERVICES PROVIDED THROUGH ORACLE'S INFINIBAND FABRIC

OSI LAYER		SERVICES
2	InfiniBand	• Switching/forwarding of packets destined for local subnet. • Forwarding of packets to router function (InfiniBand routing not currently supported). • InfiniBand multicast within an InfiniBand partition.
	Ethernet	• "Software-based" switching/forwarding of frames destined for InfiniBand-connected hosts in local subnet (EoIB). • Ethernet routing (inter-VLAN) needs to be provided via external network.
3	IP	• "Software-based" switching/forwarding of frames destined for InfiniBand-connected hosts in local subnet (IPoIB). • IP multicast for InfiniBand-connected hosts in local subnet (IPoIB). • Any other layer-3 services need to be provided via external network.
4–7		• Services need to be provided via external network.

Ethernet over InfiniBand

In addition to upper-level protocol support, InfiniBand also supports Ethernet functionality by tunneling encapsulated packets within the InfiniBand payload (Figure 7) — referred to as Ethernet over InfiniBand (EoIB) in the software stack illustration.

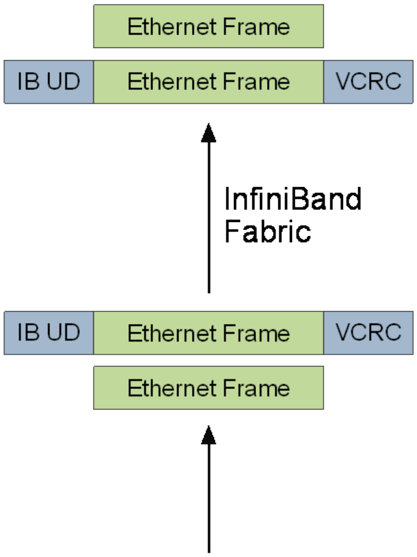


Figure 7. InfiniBand supports Ethernet functionality by tunneling encapsulated packets within the InfiniBand payload.

Ethernet frames can be easily routed across an InfiniBand interconnect through the use of InfiniBand-to-Ethernet gateways. In this case, the Ethernet frame is encapsulated into an InfiniBand packet and sent across the fabric. InfiniBand technology is "hidden" from the perspective of the Ethernet frame (Figure 8).

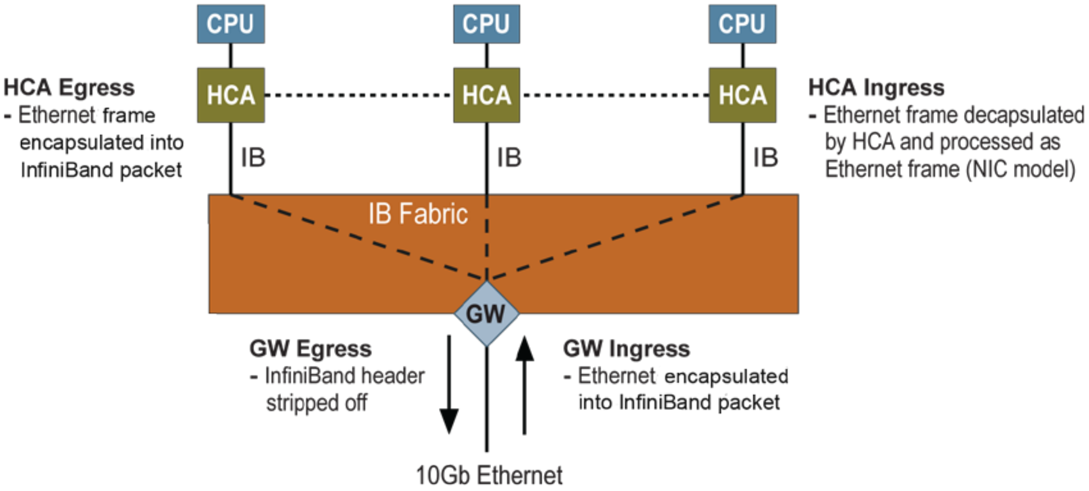


Figure 8. Encapsulated Ethernet over InfiniBand.

If the packet is destined for an InfiniBand host, the datagram is de-encapsulated by the respective HCA, and the packet is then processed as an Ethernet frame. Packets generated from InfiniBand-based hosts are likewise encapsulated in an InfiniBand packet, with the header stripped off if they leave via a gateway.

As discussed in the section on server virtualization, a virtualized system allows a physical NIC or storage interface that resides in the server to be used by multiple virtual machines (VMs). In a clustered system, the physical shared Ethernet port(s) reside at a gateway in the fabric that connects to the servers, while the actual NIC resides at the host adapter at the other end of the fabric. The gateway represents hundreds to thousands of virtual NICs — and these can be utilized by multiple VMs residing in different servers.

The ability of InfiniBand to offload protocol processing offers real rewards to virtualized environments, where processing for each of the virtualized guest operating systems and applications is also offloaded to the fabric. The result is that applications experience fewer real I/O bottlenecks, and host systems have more available processing resources to allocate to their VMs. Finally, InfiniBand's ability to tunnel Ethernet frames — together with appropriate gateways — provides the basis for utilizing InfiniBand as a general-purpose Ethernet switch. This ability provides the essential mechanisms for a real converged and consolidated data center fabric infrastructure.

Oracle's Sun Network QDR InfiniBand Gateway Switch partitions the Ethernet encapsulation and frame-forwarding functions between the hardware embedded within the switch and the host-resident InfiniBand software stack. This division of functionality ensures that traffic is efficiently forwarded to the intended destination. Ethernet and IP traffic that is destined for a server or storage device that is attached to the InfiniBand fabric is routed directly to that device without the intervention of the InfiniBand gateway. Only traffic destined for an Ethernet or IP network external to the InfiniBand fabric is processed by the InfiniBand gateway.

Performance

With distributed network applications, interconnect performance is absolutely key to application performance. A number of metrics are used as performance indicators, including the following:

- *Bandwidth* — The amount of data per second that is transmitted from a node
- *Latency* — The time it takes to transmit data from one node through the network to another node
- *Overhead* — The time (or number of cycles) that a node (CPU) uses for communication; high overhead implies less time (cycles) for application computation
- *N/2* — The message size required to reach half of the peak bandwidth

N/2 is a number that is used to quickly get an idea of how well the network handles small messages. Many parallel applications send fairly small messages. A small N/2 value means the messages that are sent take advantage of the bandwidth, decreasing the time for the message to be transmitted. A smaller N/2 generally makes the application run faster. Table 3 lists performance metrics for single GbE, 10 GbE, and various InfiniBand data rates.

TABLE 3. PERFORMANCE METRIC

METRIC / TECHNOLOGY	1 GBE	10 GBE	INFINIBAND QDR
Bandwidth	1 Gb/sec	10 Gb/sec	32 Gb/sec (4x) 96 Gb/sec (12x)
Latency (MPI ping)	Tens to hundreds of µs	Tens of µs	1.2 µs
Overhead	80 percent	80 percent	3 percent
N/2	~8,000 bytes	~8,000 bytes	256 bytes

InfiniBand Technology Overview

Since it was first standardized in October 2000, InfiniBand has emerged as an attractive fabric for building large virtualized compute clouds, supercomputing grids, clusters, and storage systems — where high bandwidth and low latency are key requirements. As an open standard, InfiniBand presents a compelling choice over proprietary interconnect technologies that depend on the success and innovation of a single vendor. Similar to single GbE and 10 GbE, InfiniBand represents a serial point-to-point full-duplex interconnect. InfiniBand also presents a number of significant technical advantages, including the following:

- InfiniBand is a lossless network with flow control to avoid packet loss due to buffer overflow, negating the need for retransmission and improving general performance.
- The InfiniBand standard includes congestion management to improve performance and avoid blocking fabrics — essential for scaling networks to hundreds or thousands of nodes.
- Service differentiation provided through Virtual Lanes (VLs) and Service Levels (SLs) helps enable quality of service (QoS).
- Multipath routing provides effective load balancing and redundancy.
- Host channel adapters with reliable transport protocols and RDMA support in hardware offload communications processing from the operating system, leaving more process resources available for computation.
- High data integrity is provided through link-level 16-bit VCRC, end-to-end 32-bit ICRC, and multiple protection keys.

The sections that follow provide an overview of InfiniBand technology.

Quality of Service (QoS)

Predictable QoS is increasingly vital to applications, and notions of QoS must logically extend to interconnect technologies. When discussing QoS in interconnection networks, there are three properties of significant importance:

- Bandwidth
- Latency
- Packet loss

In most interconnect technologies, there is a strict guarantee of no packet loss that is in effect for all data traffic.[1] With regard to latency and bandwidth, a combination of mechanisms is often defined, ranging from strict guarantees for single streams, to relative guarantees for classes of traffic, to no guarantees and/or overprovisioning. The capabilities that influence a given technology's ability to leverage QoS guarantees and differentiated treatment of traffic fall into the categories of flow control, congestion management, service differentiation, and admission control.

Point-to-Point Credit-Based Flow Control and Virtual Lanes

The main purpose of link-level flow-control is to eliminate packet loss as a result of contention and receive-buffer overflow in switches. InfiniBand uses a point-to-point credit-based flow control scheme. With this approach, the downstream side of a link keeps track of the available buffer resources (credits) by decreasing a credit counter whenever buffer space is allocated and increasing the credit counter whenever buffer space is deallocated. Similarly, the upstream node keeps track of the available credits (that is, the number of bytes it is allowed to send) and decreases this amount whenever it sends a packet. Whenever credits arrive from the downstream node, the upstream node increases the amount of available credits.

InfiniBand Virtual Lanes (VL) allow a physical link to be split into several virtual links — each virtual link with its own buffering, flow-control, and congestion management resources.

The result of this approach is that a packet is never sent downstream unless there is buffer space available for it. At regular intervals, the downstream node details credit availability to the upstream node. The update interval depends on the load — high loads increase the frequency of updates, while low loads reduce the frequency. The use of flow control helps ensure that packet loss is the result of only link transmission errors. The available link bandwidth is used effectively since retransmissions as a result of buffer overflow are not necessary. Because InfiniBand is a layered networking technology, flow control is performed on a virtual lane basis.

[1] Ethernet can be regarded as an exception to this rule, because it has no lossless transport and no flow control limitations. As a result, higher-layer protocols must accommodate packet loss and initiate retransmission.

Service Differentiation

InfiniBand Virtual Lanes are a key feature for supporting service differentiation. Virtual Lanes can be grouped as layers, making it possible to build virtual networks on top of a physical topology. These virtual networks, or layers, can be used for various purposes, such as efficient routing, deadlock avoidance, traffic separation, and service differentiation, provided by Service Levels and Virtual Lane Arbitration.

- *Service Levels (SLs)* — Echoing the notion of real-world service levels, the InfiniBand Service Level (SL) is a field in the packet header that denotes what type of service a packet receives as it travels towards its destination. A service level is similar to the packet marking approach used in Differential Services (DiffServ) as specified by IETF. At each InfiniBand port, there is a mapping table between a Service Level and a Virtual Lane (SL2VL mapping).
- *Virtual Lane Arbitration* — VL arbitration is a per-port transmission arbitration mechanism. Each VL can be configured with a weight and a priority (low or high), where the weight is the proportion of link bandwidth that the VL is allowed to use, and the priority is either high or low.

Figure 9, depicts the overlay of virtual lanes onto an InfiniBand physical link. Bandwidth is allocated based upon the VL arbitration mechanism previously described. Credit-based flow control is also implemented for each virtual lane. This scheme enables Quality of Service differentiation per Service Level/Virtual Lane, and ensures secure forward progress.

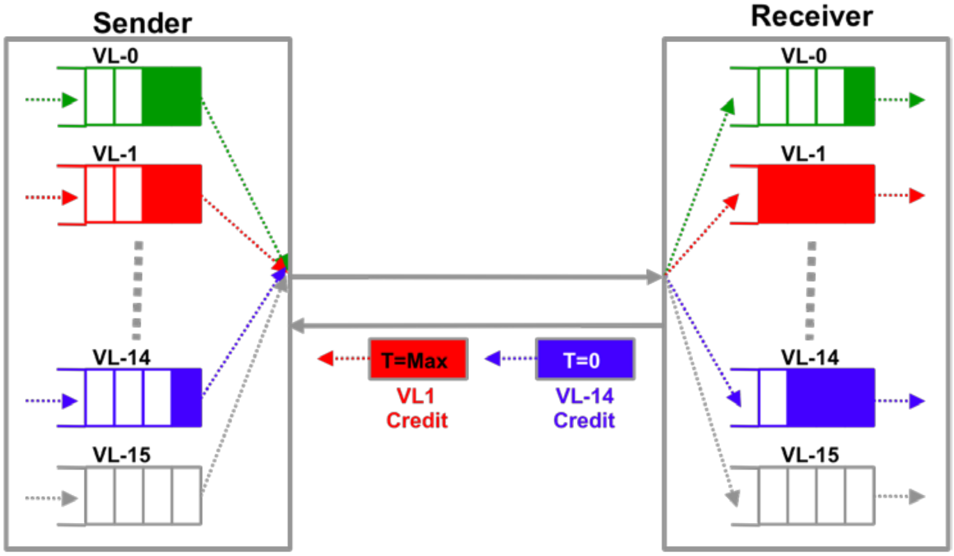


Figure 9. Overlay of Virtual Lanes on InfiniBand physical link.

Congestion Control

Just as flow control represents a mechanism for avoiding packet loss due to buffer overflows, congestion management is designed to help switches and links in the network avoid becoming overloaded with depleted credit supplies. As fabric utilization increases, depletion of credit supplies starts and the switch queues begin to fill. This process can spread upstream through the network and can result in the creation and growth of congestion trees that eventually terminate at the end nodes.

InfiniBand V1.2 added early congestion notification to help alleviate congestion spreading. Early congestion notification reduces the source injection rate to the available throughput. Forward explicit congestion notification (FECN) is used to inform the destination that the packet was subjected to congestion while traversing the network. Each switch monitors its own buffers, and when the buffer fill ratio exceeds a threshold, the switch will set a FECN flag in the packet's header. This flag is observed by the destination, which can signal the source about congestion by setting the backward explicit congestion notification (BECN) flag in the next acknowledgment packet or in a special packet for unacknowledged transports.

Upon reception of a BECN, the producer HCA will reduce the injection rate. The arrival of successive BECNs leads to further reductions, while a subsequent rate increase is based on a time-out interval that is relative to the latest congestion notification. The available static rates are selected from a congestion control table, where each entry defines an acceptable injection rate.

Routing and Topologies

The sections that follow describe several approaches to InfiniBand topologies and provide background on packets, switching, and multipath routing.

InfiniBand provides for two basic types of packets, management packets and data packets. Management packets are used for fabric discovery, configuration, and maintenance. Data packets carry up to 4 KB of data payload. A packet contains a header consisting of routing information, transport protocol information, QoS information, and payload. Both the header and payload are protected by link-level CRC and end-to-end CRC.

Switching

Within a given InfiniBand subnet, each node is assigned a 16-bit Local ID (LID) by the Subnet Manager. A packet that is injected into the fabric contains the source LID (SLID) and the destination LID (DLID). The InfiniBand switch forwards packets to the end node specified by the DLID within the packet. InfiniBand includes deterministic routing that preserves packet ordering through the fabric. This functionality eliminates the need for reordering at the receiving node. Both unicast and multicast routing are supported.

The network layer handles routing of packets from one subnet to another. Within a given subnet, the network layer is not required. Packets sent between subnets contain a Global Route Header (GRH) that contains the 128-bit IPv6 address for the source and destination of each packet.

Multipath Routing

Routing packets across an expansive fabric presents several challenges:

- Achieving high utilization of the topology in terms of throughput and latency
- Keeping the fabric connected throughout the lifetime of the application it serves
- Separating multiple applications running on the same system and fabric

Multipath routing is considered an efficient mechanism for increasing network performance and providing fault tolerance. Multipathing in InfiniBand is achieved by assigning multiple LIDs to an end point. Upper-level protocols, such as MPI, utilize this by striping across multiple paths (referred to as MPI multirailing).

Statistically nonblocking Clos fabrics are ideal topologies for the application of multipath routing. As long as there is no end-point contention, distributing all traffic from all sources over all possible paths will guarantee minimal contention and, thus, maximal throughput.

Clos Topologies and the Impact of Routing Algorithms

Clos networks have been the basis for the fabrics of many scalable, high-performance systems that require constant bandwidth to every node. First described by Charles Clos in 1953, Clos networks have long formed the basis for practical multistage telephone switching systems. Clos networks allow complex switching networks to be built using many fewer cross-points than if the entire system were implemented as a single large crossbar switch.

Clos switches typically comprise multiple tiers and stages (hops), with each tier comprising a number of crossbar switches. Connectivity exists only between switch chips on adjacent tiers. Clos fabrics have the advantage of being nonblocking fabrics in which each attached node has a constant bandwidth. In addition, an equal number of stages between nodes provides for uniform latency.

The fabric's logical topology is critical to both latency and bandwidth in multistage networks. The network performance of fabrics that are physically identical can be drastically impacted by the logical topology realized. Ethernet layer-2 networks typically use the Spanning Tree Protocol (STP), which is defined in IEEE 802.1D, to eliminate loops in the network. In order to achieve this, STP disables all redundant interswitch links. Alternatively, InfiniBand networks are frequently implemented as a fat tree. Fat trees provide increased bandwidth toward the root of the tree (thus the term "fat") and utilize all available interswitch links.

Figure 10 diagram A, shows a 2-tier/3-stage fabric composed of non-blocking, 8-port switches. The bandwidth capacity of the network is equivalent to 16x the bandwidth of the individual links. The effect of the Spanning Tree Protocol can be seen in diagram B, where all redundant links between the level-1 switches and the level-2 switches have been disabled. The result is a drastic reduction in total available bandwidth in the fabric, yielding just one quarter of the total bandwidth capacity. Diagram C shows a fat tree InfiniBand implementation. All network links are utilized and the fabric's full bandwidth capacity is available.

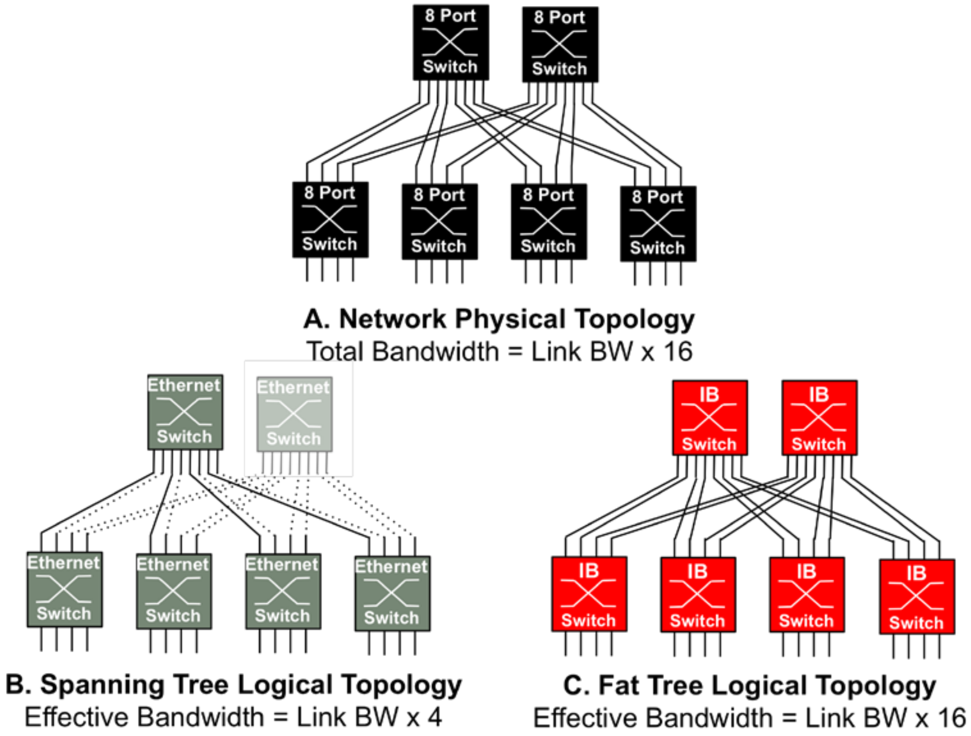


Figure 10. Effects of logical topology on effective bandwidth.

InfiniBand fabrics deployed in Oracle's engineered systems, such as Oracle Exadata Database Machine, Oracle Exalogic Elastic Cloud, and Oracle SPARC SuperCluster, utilize fat tree topologies, ensuring that all the fabric's bandwidth is available to support fabric-resident applications.

Transport Services and Protocols

The transport layer is responsible for in-order packet delivery, partitioning, channel multiplexing, and transport services. Legacy NICs rely upon the CPU to process TCP for traffic multiplexing and demultiplexing, data movement, and identification of the application for which the data is intended — often incurring significant overhead. In contrast, the InfiniBand HCA provides all transport services in hardware (reliable data delivery, multiplexing of streams, and so on). This approach improves CPU efficiency by offloading data movement, protocol stack processing, and multiplexing operations from the CPU.

Remote Direct Memory Access (RDMA)

InfiniBand addresses the problem of data movement between memory and I/O devices by using Remote Direct Memory Access (RDMA) to offload data movement from the server CPU to the InfiniBand HCA. RDMA enables data to be moved from one memory location to another, even if that memory resides on another device or server. With RDMA, the user application informs the DMA hardware of the memory location where data resides and the memory location where the data is to be moved to. Once RDMA instructions are sent, the user application can move on to process other threads while the DMA hardware moves the data.

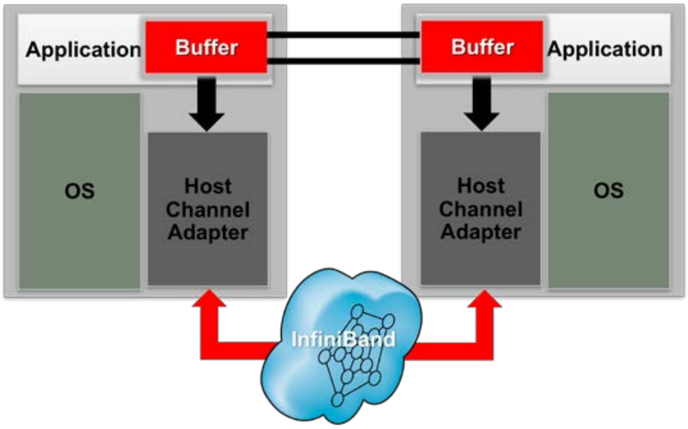


Figure 11. RDMA offloads data movement from the server CPU to the InfiniBand HCA.

Connection/Channel Types and Queue Pairs

RDMA solves the problem of data movement, but it does not address multiplexing and reliable sequenced delivery. For these purposes, InfiniBand uses the concept of queue pairs (QPs). Each QP is a virtual interface associated with the HCA hardware that is used to send and receive data. For interprocess communication, applications create one or more QPs and then assign registered memory regions as the source and destination for direct memory data transfers. Applications may have single or multiple QPs, depending on their needs.

InfiniBand specifies multiple transport types for data reliability. Each queue pair uses a specific transport type (Table 4). Based on the maximum transmission unit (MTU) of the path, the transport layer divides the data into packets of the proper size. The receiver delivers the data to memory in order.

TABLE 4. INFINIBAND TRANSPORT TYPES

CLASS OF SERVICE	DESCRIPTION
Reliable Connection	Acknowledged, connection-oriented
Reliable Datagram	Acknowledged, multiplexed
Unreliable Connection	Unacknowledged, connection-oriented
Unreliable Datagram	Unacknowledged, connectionless
Raw Datagram	Unacknowledged, connectionless

Data Integrity

Data integrity and silent data corruption in particular are of increasing concern in computer systems and storage. Silent data corruption also exists in data communication, especially as complex clustered solutions grow and transfer ever-larger amounts of critical data. While traditional networking has been lacking in end-to-end data integrity, InfiniBand provides considerable improvement.

Silent Data Corruption

Recognition of silent data corruption is on the rise. In fact, one study[2] has shown that between one in 1,100 to one in 32,000 IP packets fails the TCP checksum, even on links where link-level CRCs should catch all but one in four billion errors. These results point to considerable corruption taking place while data is in transit. Further, the damage is not taking place only on transmission links, where it would have been caught by the CRC, but rather in intermediate systems including routers, bridges, or the sending and receiving hosts (Figure 12).

Sources of errors include the following:

- Errors in end-host hardware and/or software
- Errors at the link level or in-network interface hardware
- Errors in switch/router memory (alpha particles or cosmic rays)

[2] Jonathan Stone and Craig Partridge, "When the CRC and TCP Checksum Disagree," 2000

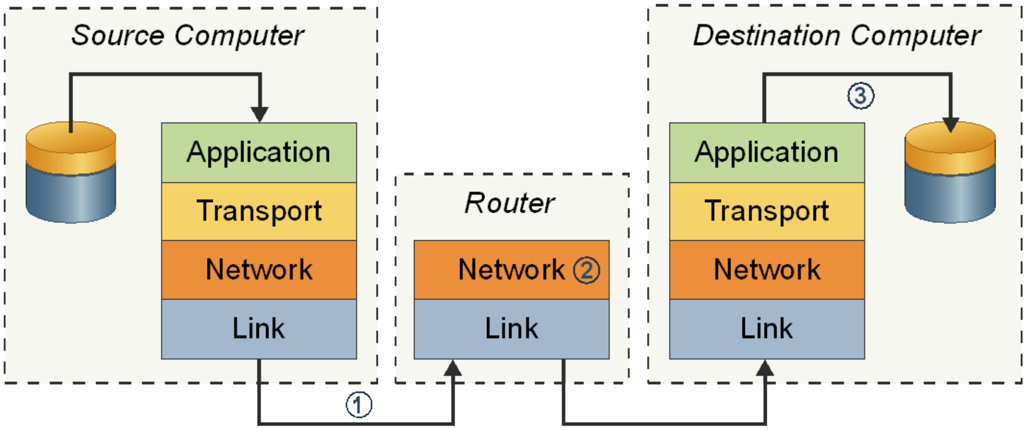


Figure 12. Errors can occur at any point along the path from system to system, including links (1), network and interconnect devices (2), and systems and storage devices (3).

Errors can be classified in two main categories, detected errors, and silent errors (undetected errors). Even detected errors can have an adverse effect on performance through required retransmissions. Undetected errors are more sinister yet, corrupting data without making themselves known. There are two main causes of errors, bit-flips within integrated circuits caused by cosmic rays or alpha particles (referred to as soft errors) and transmission errors. Soft error rates are increasing as integrated circuits geometries become smaller. A 10-fold increase in soft error rates has been observed in SRAM as it moved from a 130 nm process to 65 nm process, while an increase of 30-fold has been observed for flip-flops during the same transition.

By necessity, adequate protection, such as ECC and parity, has been added to memory and the storage structures of switches and HCAs. In addition, as transmission speed increases, so do transmission errors caused by power budget, dispersion, and nonlinearities. Though these noncorrected error rates can be below 10^{-12}, even that rate represents a one-bit error every 10 seconds on a 100 Gb/sec link. At that rate, a cluster with 1,000 links will see 100 errors per second.

Weak Ethernet and TCP Checksums

Complicating the rise in soft error rates, the 16-bit TCP checksum is relatively weak and will not detect many errors. The choice to provide a relatively weak checksum was made primarily to reduce software overhead, since these checksums are traditionally performed in software. The TCP checksum is also not end to end, providing only a packet-level checksum that is typically stripped off by the NIC.

Unfortunately, the Ethernet checksum is also not end to end. The Ethernet 32-bit CRC was added at the link level, but it is rewritten in every layer-3 switch. As a result, the Ethernet CRC can detect link errors but not end-to-end errors. Cosmic rays can easily flip bits in switch memory, resulting in bad packets with a good CRC.

InfiniBand Data Integrity

To address silent data corruption, InfiniBand provides strong data integrity functionality. A 16-bit link CRC (VCRC) is recomputed at every switch stage. In addition, a 32-bit end-to-end CRC (ICRC) is provided on a per-packet basis. Extensive protocol checking is performed in hardware before a packet is accepted. Multiple verification keys are embedded inside each packet, including a key to verify the 64-bit address and a key to verify that the device is in a partition that is allowed to receive the packet. As error rates increase, the net result is an increase in retransmissions. To avoid an even more adverse effect on performance, efficient retransmission is required. InfiniBand provides retransmission and automatic path migration in the host adapter hardware.

Physical Medium

InfiniBand supports a number of physical connections. Both copper and fiber-optic cables are provided. Typically, copper cables support distances up to 5 meters and multimode fiber-optic cables are used for distances up to 100 meters.

- **CX4**
  CX4 was initially defined by the IBTA as the InfiniBand 4x SDR/DDR copper cable interface, and it transmits over four lanes in each direction. CX4 supports distances up to 15 meters. IEEE has adopted CX4 in the 10G-Base-CX4 specification (four lanes of 3.125 Gb/sec).
- **SFP+**
  SFP was designed after the GBIC interface and allows greater port density (number of transceivers per inch along the edge of a mother board) than GBIC. SFP is expanding to SFP+ to support data rates up to 10 Gb/sec (XFI electrical specification). SFP+ is being used for 10 GbE and 8G Fibre Channel. Copper cables, fiber-optic modules, and optical active cables are all supported.
- **QSFP**
  QSFP (Quad Small Form-factor Pluggable) is a highly integrated pluggable module that is intended to replace four standard SFP transceiver modules, providing increased port density and total system cost savings. QSFP will support InfiniBand, Ethernet, Fibre Channel, and SONET/SDH standards with data rate options from 2.5 Gb/sec to 10 Gb/sec per lane. Copper cables, fiber-optic modules, and optical active cables are supported.

- **CXP**
  The InfiniBand 12x Small Form-factor Pluggable (CXP) connector consolidates three discrete InfiniBand 4x connectors, resulting in the ability to host a large number of ports on a very small physical footprint. In fact, the CXP connector supports three 4x links in the same physical footprint as a CX4 connector, and the CXP connector is considerably more robust. This connector was standardized by the InfiniBand Trade Association (IBTA). A splitter cable is available that converts one 12x connection to three 4x connections for connectivity to legacy InfiniBand equipment. Copper cables and optical active cables are supported.
- **MPO/MTP**
  "Multi-Fiber Push-On/Pull-Off" (EIA/TIA-604-5) terminated multimode fiber ribbon cables are supported in lengths up to 100 meters. QSFP and CXP optical transceivers are available with receptacles for MPO terminated cables. (Note that MTP is a branded version of the connector, which meets the MPO specification.)

Fabric Management

An InfiniBand fabric is defined by one or more InfiniBand subnets, each configured and controlled by a single, centralized Subnet Manager (SM). Multiple SMs may be present in an InfiniBand subnet, but only one may be active while the others are on standby. The Subnet Manager represents a strict administrative domain for physical cluster management, implemented as a single InfiniBand subnet. In-band management capabilities are provided by agents on each InfiniBand device.

InfiniBand Management Model

The Subnet Manager facilitates central control of policies as well as maintenance of the inventory of the nodes within the subnet. These capabilities help enable enterprise system features such as carefully controlled high availability, redundant operation, rolling upgrades, and required performance and scalability. The InfiniBand centralized management model combined with well-defined features for access control (partitioning), multiple traffic classes, and/or traffic isolation lends itself very well to an "owner/tenant" administrative model. In this model, the "cluster owner" (administrator) controls the Subnet Manager and is thereby able to provision resources to multiple "tenants" as multiple domains with different levels of isolation between the domains.

InfiniBand Subnet Manager Operation

At any point in time, the Subnet Manager (SM) and Subnet Administrator (SA) operate from a specific port within the subnet. The Subnet Manager is principally involved with configuring and routing the topology. The Subnet Administrator provides specific routes between nodes within the fabric and is responsible for communicating with client nodes (end nodes).

In-band discovery involves the discovery of arbitrary topologies using geographically addressed ("direct routed") Subnet Management Packets (SMPs). The Subnet Manager communicates with Subnet Manager Agents (SMAs) residing in individual HCAs and switches that provide both read-only as well as writable configuration data in addition to basic error status and counters.

To initialize the fabric, the Subnet Manager assigns LIDs (per-node Logical IDs) to end nodes and performs routing (end node-to-end node routes via switches) using Subnet Manager–specific policies and algorithms. Basic multicast routing is handled similarly. Routing is implemented by updating per-switch forwarding tables (LID-to-port mapping) using write requests to the switch Subnet Manager Agent (SMA).

In order to establish communication between end nodes, the relevant "path info" must be known. Normal data traffic is enabled when the Subnet Manager sets ports and links to an active state. Client nodes can learn about the presence of other nodes and what "path" exists by sending (in-band) requests to the Subnet Administrator (SA).

Establishing Connections and Datagram Traffic

The end nodes (client nodes) in an InfiniBand fabric use the information retrieved from the Subnet Administrator in order to determine what other nodes are present (and visible) and how to communicate with these nodes. The InfiniBand specification defines how Connection Managers (CMs) on different nodes can exchange information in order to determine what services are provided on a node, as well as to establish connections (reliable or unreliable) between the nodes.

Datagram-based communication has no connection context, but addressing is based on path information from the Subnet Administrator similar to that for connections. IP-based traffic is supported by IETF standards for how to implement Address Resolution Protocol (ARP) using InfiniBand multicast, as well as how addressing and IP packet encapsulation based on InfiniBand unreliable datagrams is to be implemented within a (logical) IPoIB link.

Fabric Security

The InfiniBand architecture uses various "keys" to provide isolation and protection. Keys are values assigned by an administrative entity that are used in messages in order to authenticate that the initiator of a request is an authorized requestor and that the initiator has the appropriate privileges for the request being made. Table 5 lists the keys provided in the InfiniBand architecture. These keys enable the architecture to provide the following:

- Management protection: M_Keys and B_Keys ensure that administrative control is retained by the respective management entities.
- Traffic isolation: P_Keys and Q_Keys ensure that virtual networks and communication channels are isolated from unauthorized participation.
- Data protection: L_Keys and R_Keys ensure that memory is protected from unauthorized access.

TABLE 5. KEYS PROVIDED IN INFINIBAND ARCHITECTURE

KEY	KEY NAME	FIELD SIZE (BITS)	DESCRIPTION
Baseboard Management Key	B_Key	64	Enforces the control of a subnet baseboard manager. Administered by the subnet baseboard manager and used in certain MADs. Each channel adapter port has a B_Key that the baseboard manager sets. The baseboard manager may assign a different key to each port. Once enabled, the port rejects certain management packets that do not contain the programmed B_Key. Thus only a baseboard manager with the programmed B_Key can alter a node's baseboard configuration. The baseboard manager can prevent the port's B_Key from being read as long as the baseboard manager is active. The port maintains a timeout interval so that the port reverts to an unmanaged state if the baseboard manager fails. There is one B_Key for a switch.
Memory Keys (Local Key and Remote Key)	L_Key and R_Key	32	Enable the use of virtual addresses and provide the consumer with a mechanism to control access to its memory. These keys are administered by the channel adapter through a registration process. The consumer registers a region of memory with the channel adapter and receives an L_Key and R_Key. The consumer uses the L_Key in work requests to describe local memory to the QP and passes the R_Key to a remote consumer for use in RDMA operations. When a consumer queues up an RDMA operation, it specifies the R_Key passed to it from the remote consumer and the R_Key is included in the RDMA request packet to the original channel adapter. The R_Key validates the sender's right to access the destination's memory and provides the destination channel adapter with the means to translate the virtual address to a physical address.
Management Key	M_Key	64	Enforces the control of a master subnet manager. Administered by the subnet manager and used in certain subnet management packets. Each channel adapter port has an M_Key that the SM sets and then enables. The SM may assign a different key to each port. Once enabled, the port rejects certain management packets that do not contain the programmed M_Key. Thus, only an SM with the programmed M_Key can alter a node's fabric configuration. The SM can prevent the port's M_Key from being read as long as the SM is active. The port maintains a timeout interval so that the port reverts to an unmanaged state if the SM fails. There is one M_Key for a switch.
Queue Key	Q_Key	32	Enforces access rights for reliable and unreliable datagram service (RAW datagram service type not included). Administered by the channel adapter. During communication establishment for datagram service, nodes exchange Q_Keys for particular queue pairs and a node uses the value it was passed for a remote QP in all packets it sends to that remote QP. Likewise, the remote node uses the Q_Key it was provided. Receipt of a packet with a different Q_Key than the one the node provided to the remote queue pair means that packet is not valid and, thus, it is rejected.

Partition Key	P_Key	16	Enforces membership. Administered through the subnet manager by the partition manager (PM). Each channel adapter port contains a table of partition keys, which is set up by the PM. QPs are required to be configured for the same partition to communicate (except QPO, QP1, and ports configured for raw datagrams) and, thus, the P_Key is carried in every IP transport packet. Part of the communication establishment process determines which P_Key that a particular QP or EEC uses. An EEC contains the P_Key for Reliable Datagram service and a QP context contains the P_Key for the other IBA transport types. The P_Key in the QP or EEC is placed n each packet sent and compared with the P_Key in each packet received. Received packets whose P_Key comparison fails are rejected. Each switch has one P_Key table for management messages and may optionally support par1ition enforcement tables that filter packets based on their P_Key.

Extending the InfiniBand Cluster with Gateways

As discussed previously, InfiniBand clusters can also connect to gateway (GW) nodes that interface to external LANs (Ethernet) or other InfiniBand subnets. In this fashion, the owner of the cluster can control the configuration and provisioning of external access in ways that are not radically different from the how the external access of a single virtualized (partitioned) server would be managed. In this case, the cluster administrator would configure the gateways based on agreements ("contracts") with the relevant external LAN administrator(s) and then provision the external access internally in the cluster based on the internal domaining policies and external resource contracts.

Oracle's InfiniBand Technology Leadership

As a founder[3] and Steering Committee member of the InfiniBand Trade Association — the standards body responsible for developing and maintaining the InfiniBand specification — Oracle recognizes the importance of InfiniBand technology in the future of the data center. Oracle is the only enterprise infrastructure provider making R&D investments in InfiniBand technology, enabling its complete hardware and software stack with the capability to take advantage of the performance and scalability afforded by InfiniBand.

Having developed three generations of InfiniBand host and switching products and having integrated InfiniBand into both mission-critical enterprise application environments as well as some of the world's largest supercomputers, today Oracle offers solutions for server and storage connectivity, as well as a family of enterprise-class switches.

[3] Sun Microsystems (acquired by Oracle in 2010) was a founding member of the IBTA.

Oracle was the first to introduce InfiniBand's RDMA capabilities into distributed databases with Oracle RAC 10*g* and has extended its leadership in RDMA-enabled enterprise software to the mid-tier with Oracle Fusion Middleware 11*g*. Oracle also leverages the efficiency in the network stack, the offload capabilities, and the extreme bandwidth of InfiniBand within its engineered systems. These capabilities enable Oracle engineered systems, such as Oracle Exadata, Oracle Exalogic, and SPARC SuperCluster, to deliver unparalleled performance in the enterprise.

InfiniBand in Oracle's Engineered Systems

As a complete enterprise solution provider, Oracle is uniquely positioned to innovate at all levels of the system — from switching fabric through to the application software stack. Oracle's holistic approach provides a highly integrated, high-performance, InfiniBand-based I/O solution together with applications and computational and storage solutions that function as a unified system. Best of all, organizations can obtain the performance benefits of InfiniBand without having to deploy InfiniBand themselves. Oracle's engineered systems utilize InfiniBand as an internal system "backplane," while exposing Ethernet interfaces for connection of the system into the data center infrastructure, making the migration to InfiniBand transparent.

InfiniBand as the Backbone for Oracle Engineered Systems

The capabilities described previously have made InfiniBand fabrics a great fit as the consolidated network fabric for Oracle engineered systems, such as the Oracle Exadata Database Machine and the Oracle Exalogic Elastic Cloud. Figure 13 illustrates the I/O and network design used in Oracle Exalogic. InfiniBand technology is used to provide a consolidated interconnect backbone that serves all of the hardware components of the Oracle Exalogic Elastic Cloud X2-2 system. The compute nodes and the storage device within Oracle Exalogic are provisioned with redundant InfiniBand connections. Seamless interoperability with the existing I/O infrastructure of the "external" data center network is provided through the Ethernet ports on the InfiniBand gateways. Note that there is an independent single GbE network for device management that connects to all components in the Oracle Exalogic system, including native-attached storage, InfiniBand switches, and InfiniBand-to-Ethernet gateways.

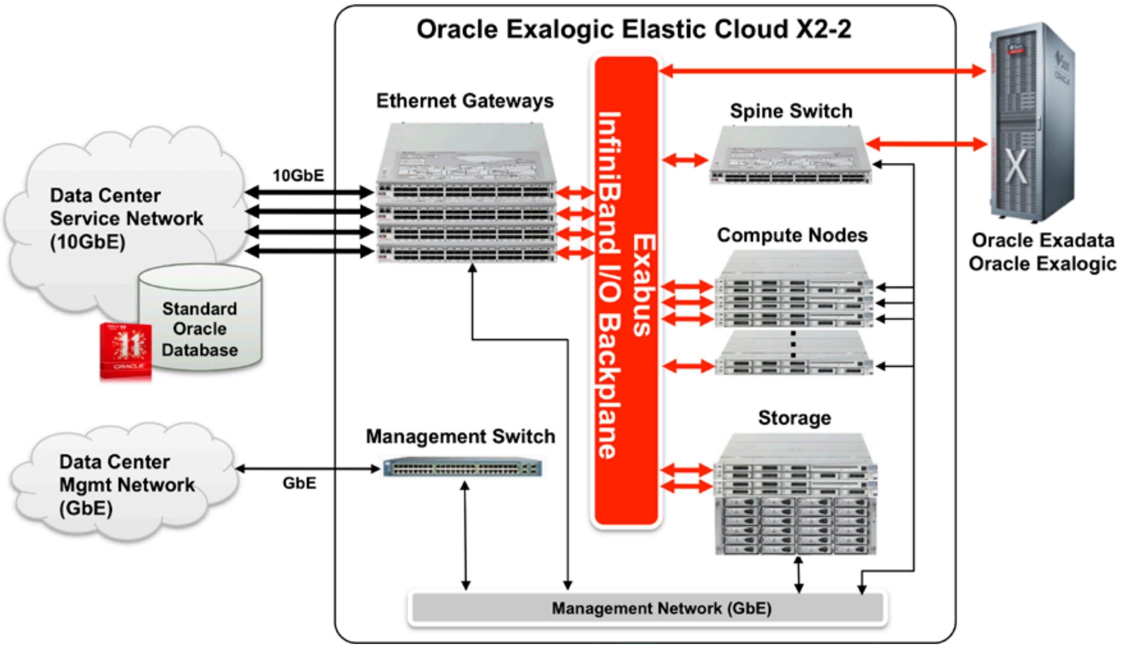


Figure 13. Oracle Exalogic consolidated I/O and network design.

System and Network Consolidation with Oracle Engineered Systems

Traditional approaches to network architecture for three-tiered data centers (Web/application/database tiers) employ a hierarchical topology, typically composed of three layers of network infrastructure; access switching, aggregation switching, and core switching. Figure 14 illustrates the network topology for a legacy three-tiered data center deployment.

This legacy model evolved due to limitations in the scalability of Ethernet L2 networks. The resulting hierarchical topology offered some benefits including broadcast domain isolation, the potential to scale beyond Ethernet L2 network limitations, and the ability to fit access-layer switching performance with workload requirements. However, the costs of deploying a complex, hierarchical, topology are switch sprawl, high network latency, oversubscription, and management complexity.

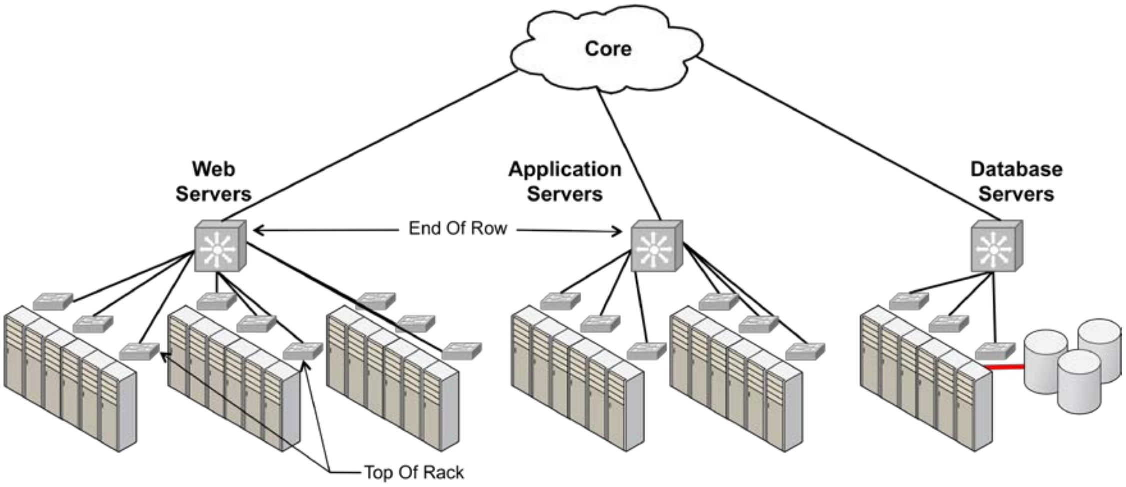


Figure 14. Network topology for a legacy three-tiered data center deployment.

Oracle's engineered systems utilize a flat, two-tier, leaf-and-spine network topology that eliminates the need for aggregation switching. The fabric delivers uniform bandwidth and latency to all nodes, which is critical to the performance of applications in cloud environments. Figure 15 illustrates the leaf-and-spine network topology employed in Oracle's engineered systems.

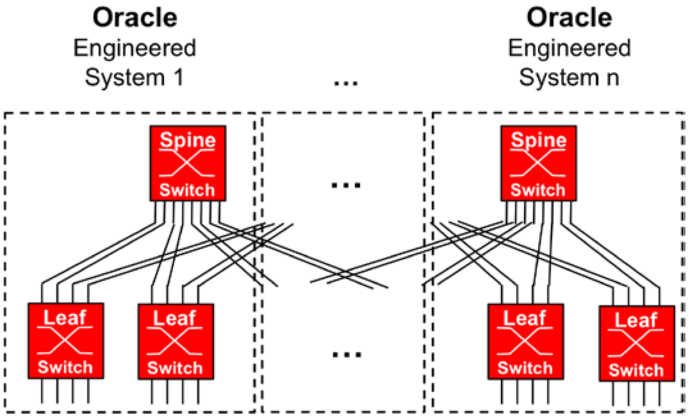


Figure 15. Leaf-and-spine topology.

Oracle's engineered systems include embedded leaf-and-spine switches as integral parts of the system. With Oracle's leaf-and-spine topology, there are at most three switch hops between any two nodes in the fabric (worst-case switch latency is 300 nanoseconds). Furthermore, providing uniform network building blocks within each engineered system allows Oracle's engineered systems to scale without the need for external switching components. Deploying a scalable application and database infrastructure is achieved by simply interconnecting Oracle Exalogic and Oracle Exadata cabinets. No external switching is required. Figure 16 illustrates deployment of the application and database tiers with Oracle Exalogic and Oracle Exadata.

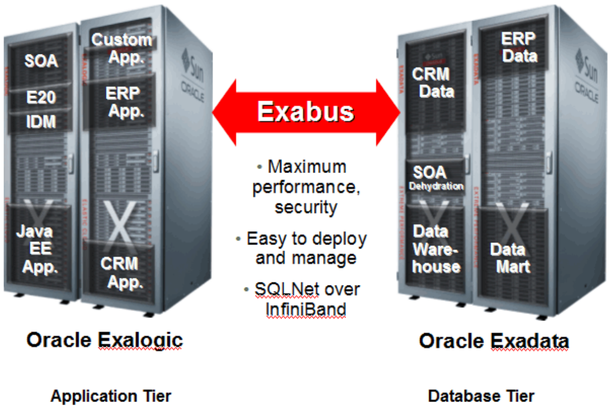


Figure 16. Application and database tiers with Oracle Exalogic and Oracle Exadata.

Oracle's Enterprise-Hardened Network Stack

Both Oracle Linux and Oracle Solaris provide enterprise-hardened InfiniBand stacks. Oracle Linux leverages the Open Fabrics Enterprise Distribution (OFED), open source, high-performance networking stack. Oracle has enhanced the stack components, including fabric management agents and Upper Layer Protocols (ULPs) that are leveraged in Oracle's engineered systems, such as SDP, RDS, IPoIB, and EoIB, to ensure high availability, enhance security, and deliver extreme performance for Oracle applications. Enterprise-ready, Oracle Solaris fully supports the InfiniBand ULPs leveraged within Oracle's engineered systems and, furthermore, it provides support for RDMA-enabled storage target ULPs, enabling both file and block storage.

Oracle's InfinBand-Enabled Storage

Oracle's Sun ZFS Storage Appliance systems deliver best-in-class NAS performance, efficiency, and integration with Oracle software. Supporting powerful analytics and extensive SAN capabilities, Sun ZFS Storage Appliance systems are high-performance enterprise storage platforms with system capacities ranging from 3.3 terabytes up to 1.73 petabytes.

Sun ZFS Storage Appliance systems are enabled with InfiniBand support for both file and block storage. Oracle Solaris, the operating system underpinning Sun ZFS Storage Appliance systems, includes an RDMA-enabled storage target for NFS, as well as targets supporting block storage protocols with SRP and iSER. Combining InfiniBand's bandwidth, low latency, and the efficiency of RDMA with Sun ZFS Storage Appliance systems produces storage solutions with unparalleled performance and ease of use.

Sun ZSF Storage Appliances are the fastest and most reliable external storage solutions for Oracle's engineered systems, such as Oracle Exadata Database Machine, Oracle Exalogic Elastic Cloud, and SPARC SuperCluster, achieving 4x faster backup and 2x faster recovery than competitive NAS or SAN solutions.

Figure 17. Sun ZFS Storage Appliance systems support InfiniBand.

Oracle's Family of InfiniBand Switches and Host Channel Adapters

Oracle provides a complete line of QDR InfiniBand products, including switches, host adapters, cables, and transceivers. Oracle has engineered its InfiniBand hardware product line specifically for use in mission-critical enterprise application environments. With features supporting fabric resiliency, availability, and security, Oracle's InfiniBand switches provide an optimal platform for mission-critical applications demanding high performance.

Leveraging the properties of the InfiniBand architecture, Oracle's Sun Datacenter InfiniBand Switch 36 and Sun Network QDR InfiniBand Gateway Switch support the creation of logically isolated subclusters, as well as advanced InfiniBand features for traffic isolation and QoS management — preventing faults from causing costly service disruptions. Furthermore, both switches provide an embedded InfiniBand fabric management module that supports active/hot-standby dual-manager configurations, helping to ensure a seamless migration of the fabric management service in the event of a management module failure. Oracle's InfiniBand switches are provisioned with redundant power and cooling for high availability.

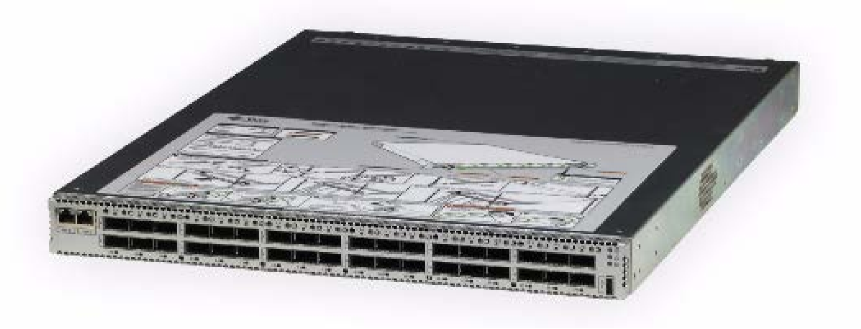


Figure 18. The Sun Datacenter InfiniBand Switch 36.

The Sun Datacenter InfiniBand Switch 36 is a fully nonblocking 36-port QDR InfiniBand switch that can act as a self-contained fabric solution for smaller InfiniBand clusters or as a switching "building block" for hierarchical fabric topologies supporting larger clusters of Sun Blade, Sun Fire, and SPARC servers from Oracle along with Sun ZFS Storage Appliance systems.

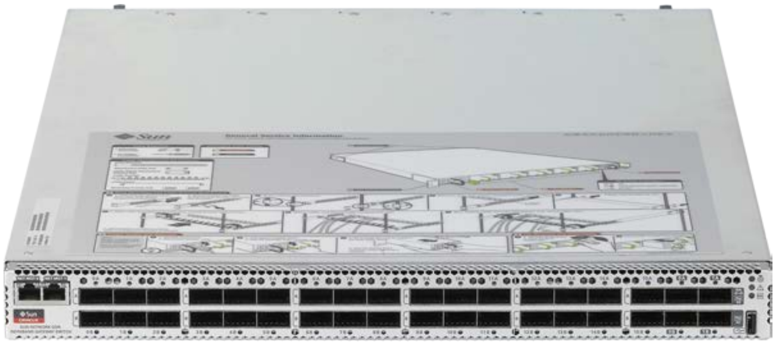


Figure 19. The Sun Network QDR InfiniBand Gateway Switch.

The Sun Network QDR InfiniBand Gateway is a high-performance QDR InfiniBand switch combined with an InfiniBand-to-Ethernet gateway. The Sun Network QDR InfiniBand Gateway Switch uniquely combines the functions of an InfiniBand leaf switch with an Ethernet gateway, which will not disrupt the operations and policies of existing LAN administration. In addition to carrying all InfiniBand traffic, the Sun Network QDR InfiniBand Gateway Switch enables all InfiniBand-attached servers to connect to an Ethernet LAN using standard Ethernet semantics. No application modifications are required for applications written to use standard Ethernet. The switch is ideal for top-of-rack deployment in server clusters or as a shared I/O resource in larger systems.

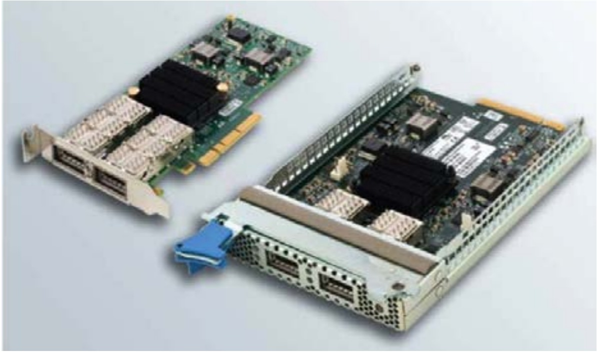


Figure 20. Sun Dual Port 4x IB QDR Host Channel Adapter PCIe M2 and Sun Dual Port 4x IB QDR Host Channel Adapter ExpressModule M2.

Oracle's family of PCIe QDR InfiniBand host channel adapters provides QDR InfiniBand connectivity for Oracle servers and ZFS storage devices. The adapters are available in both low-profile PCIe or modular hot-pluggable PCIe ExpressModule form factors.

Unified Network Management with Oracle Enterprise Manager Ops Center

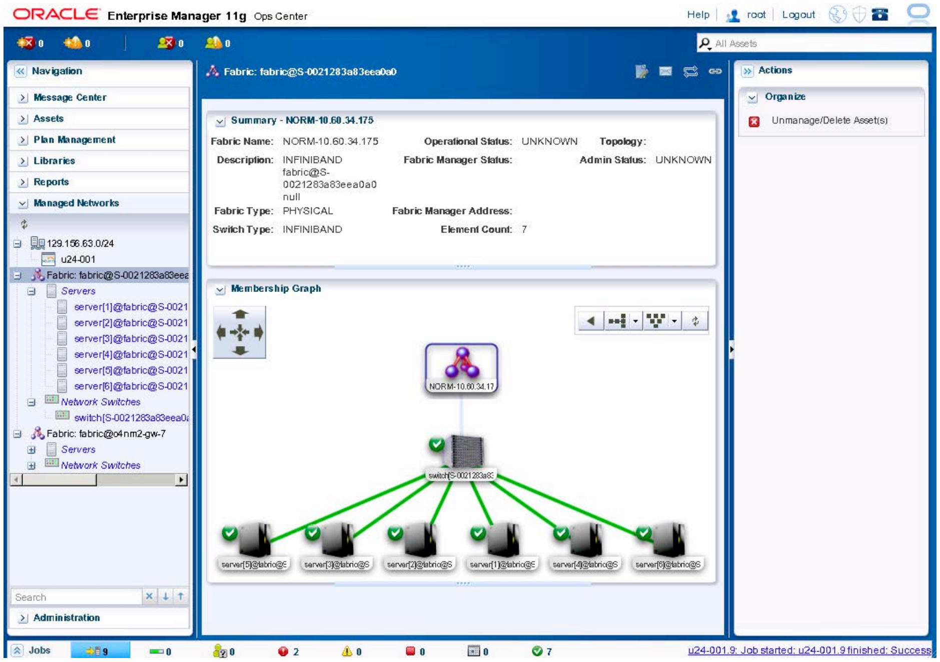
Oracle Enterprise Manager Ops Center's network management and provisioning features provide a unified platform for the administration of network resources supporting applications resident within Oracle's engineered systems. Oracle Enterprise Manager Ops Center's network management features include the following:

- Discovery: Discovers the network topology and all network elements including switches and end points
- Configuration and provisioning:
    - Configures the physical fabric including switches and server interfaces
    - Provisions logical fabric(s) overlaying the physical fabric
    - Provisions virtual server network interfaces
- Monitoring: Monitors the health and availability of all network elements
- Diagnostics: Provides a single console for diagnosing network hardware or configuration issues

Oracle Enterprise Manager Ops Center is a key component of Oracle Enterprise Manager, the management and provisioning framework for Oracle's engineered systems. Oracle Enterprise Manager Ops Center facilitates many common network provisioning tasks. For example, when an application is deployed within an InfiniBand fabric utilizing the Sun Network QDR InfiniBand Gateway Switch for 10 GbE connectivity, Oracle Enterprise Manager Ops Center automates the provisioning of the network infrastructure supporting the application, including the following:

- Virtualized Ethernet NICs (vNICs)
- Virtualized InfiniBand HCA(s)
- InfiniBand partition membership
- InfiniBand gateway resources for external access

In addition to network management and provisioning functions, Oracle Enterprise Manager Ops Center provides the same enclosure management services for Oracle's InfiniBand switches that are provided for the rest of Oracle's hardware stack.



Summary - NORM-10.60.34.175

Fabric Name: NORM-10.60.34.175
Operational Status: UNKNOWN
Topology:
Description: INFINIBAND fabric@S-0021283a83eea0a0 null
Fabric Manager Status:
Admin Status: UNKNOWN
Fabric Type: PHYSICAL
Fabric Manager Address:
Switch Type: INFINIBAND
Element Count: 7

u24-001.9: Job started: u24-001.9 finished: Success

Figure 21. Oracle Enterprise Manager Ops Center can be used to manage Oracle's InfiniBand switches.

Conclusion

With a long list of technology advantages, InfiniBand is ideally positioned to solve key problems for high-performance and scalable enterprise applications. Developed as a standards-based protocol, InfiniBand was designed to provide high bandwidth and low latency, key for horizontally scaled applications and cloud computing. In addition, hardware protocol processing offered by InfiniBand fabrics offloads system processors, providing more computational cycles for organizations' most important tasks. With today's virtualized server environments, InfiniBand can greatly reduce the I/O bottleneck, and promote considerable application scalability.

With its ability to support existing upper-level application protocols, InfiniBand also greatly simplifies interconnect transitions. Existing applications and operating systems can take advantage of accelerated performance without change. With gateways and the ability to encapsulate Ethernet frames, InfiniBand also provides an ideal fabric for consolidation of adapters, cables, and switches. The result can be cleaner deployments that are less costly to power, cool and support, and more flexible and agile to reconfigure.

As a system vendor with considerable networking and enterprise application experience, Oracle has taken a leading role in InfiniBand technology. Oracle's engineered systems, such as Oracle Exadata, Oracle Exalogic, and SPARC SuperCluster, exemplify this innovation, combining the innovation in Oracle's industry leading software assets with an InfiniBand fabric that is architected for deployment in mission-critical environments to deliver unsurpassed application performance and scalability.

For More Information

To learn more about Oracle’s InfiniBand products and the benefits of Oracle’s engineered systems, please contact an Oracle sales representative, or consult the related documents and Websites listed in Table 6.

TABLE 6. RELATED WEBSITES

WEBSITE URL	DESCRIPTION
http://www.oracle.com/us/products/servers-storage/networking/infiniband/index.html	Oracle’s InfiniBand network products
http://www.oracle.com/us/products/database/exadata	Oracle Exadata
http://www.oracle.com/us/products/middleware/exalogic	Oracle Exalogic
http://www.oracle.com/us/products/servers-storage/servers/sparc-enterprise/supercluster/supercluster-t4-4/overview/	SPARC SuperCluster
http://www.oracle.com/us/products/database/big-data-appliance/overview/index.html	Oracle Big Data Appliance
http://www.oracle.com/us/products/servers-storage/networking/infiniband/036206.htm	Sun Datacenter InfiniBand Switch 36
http://www.oracle.com/us/products/servers-storage/networking/infiniband/sun-network-qdr-ib-gateway-switch-171610.html	Sun Network QDR InfiniBand Gateway Switch
http://www.oracle.com/us/products/servers-storage/networking/infiniband-dual-port-4x-qdr-pcie-171241.html	Sun Dual Port 4x IB QDR Host Channel Adapter PCIe M2 and Sun Dual Port 4x IB QDR Host Channel Adapter ExpressModule M2

Delivering Application Performance with
Oracle's InfiniBand Technology
May 2012, Version 2.0

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com