WHITE PAPER

Building a Scalable Storage with InfiniBand

The Problem

It will come as no surprise to those working in data centers today that an increasing amount of capital and operational expense is associated with building and maintaining storage systems. Many factors drive the need for increased storage capacity and performance. Increased compute power and new software paradigms are making it possible to perform useful analytics on vast repositories of data. The lowering cost per Gigabyte is making it possible for organizations to store more granular data and to keep data for longer periods of time.

The mosaic above shows some examples of both structured data (Wal-Mart's 500TB database) and unstructured data (Traffic and Security Camera's 8Tb/day). Advents in the way we process unstructured data, such as Hadoop, have made saving dynamically generated data such as web logs, environmental sensor data and email increasingly useful, thereby creating a spike in demand for storage.

It is no longer possible for manufacturers to create a single node with the capacity and performance needed to keep up with the compute performance in modern systems. As a result storage is increasingly scaling out not up. This methodology requires a high performance Storage Area Network in order to stay at parity with databases, analytics or applications.

Moreover, both unstructured data applications and traditional structured databases are moving to distributed parallel processing systems as well; most notably Hadoop and Oracle RAC. To support applications with a high degree of scalability, the storage tier itself must be able to scale.

Not only must storage systems have the capacity to move large volumes of data, but they also need the ability to quickly find and access individual points of data. This property is typically measured as access time. In the distributed database and analytics models, latency is as important as bandwidth in overall query performance.

The distribution of our databases and applications is not a trend but inevitably a way of the future. The ability to tackle larger problem spaces will continue to come from scale out configurations. Having an interconnect capable of supporting scale out designs will be the key differentiator in the era of "scalable big data".

Traditional Solutions and their Inherent Problems

There are a variety of storage technologies and protocols available on the market today, including SATA, SAS, SSD, iSCSI, InfiniBand, FCoE and Traditional Fibre Channel. A modern data center will likely employ many of these in order to balance storage performance, cost and availability.

It is widely believed that physical Fibre Channel is in its last days. Fibre Channel failed to keep up with the performance required by today's applications. At 8Gb/s, Fibre Channel is already out performed by Ethernet. Maintaining a dedicated SAN for storage separate from the network is an increased capital and operational cost, something that will not scale. The industry today is moving toward converged fabric solutions, using the same network interconnect for both the network and for storage.

Despite this trend, Fibre channel has advantages. Fibre Channel, by design, is a dedicated path over a lossless fabric. In contrast to Ethernet, which is unreliable, data delivered over Fibre Channel will have consistent access times because it is not subject to loss and retransmission. Secondly, because of the dedicated infrastructure, Fibre Channel doesn't suffer from congestion caused by other traffic classes. Finally, it is not subject to network events from routing protocols which share the same physical network. These are important characteristics that are very difficult to replicate in Ethernet, particularly as we bring systems to scale.

Scaling a storage system requires both a high performing interconnect and a sophisticated interconnect with support for replication, multipathing and high-availability. Of course, the complexity of these solutions needs to be managed through the use of good provision software to reduce the risk of operation mistakes resulting in data loss or down time.

To engineer storage systems that meet these requirements while maintaining ease of use, most vendors are moving toward an appliance strategy. The storage appliance implements a fast and reliable channel interconnect in the backplane, such as SAS, Fibre Channel, or InfiniBand, but will provide network based connectivity from the front end, most commonly Ethernet or InfiniBand. The most well known examples are Oracle Exadata, EMC Symmetrix, Isilon, Panasas and Data Direct Network.

Company	Backplane	Network
Oracle Exadata	InfiniBand	InfiniBand
EMC Symmetrix	Fibre Channel	Ethernet
Isilon	InfiniBand	Ethernet
Panasas	SAS	Ethernet, InfiniBand
Data Direct Networks	InfiniBand, SAS	InfiniBand

Each of these appliances utilize a channel interconnect to provide the storage clustering in the backplane and a network interconnect to provide the connectivity for application nodes. Note that InfiniBand is the only interconnect that is used as both a channel and network interconnect.

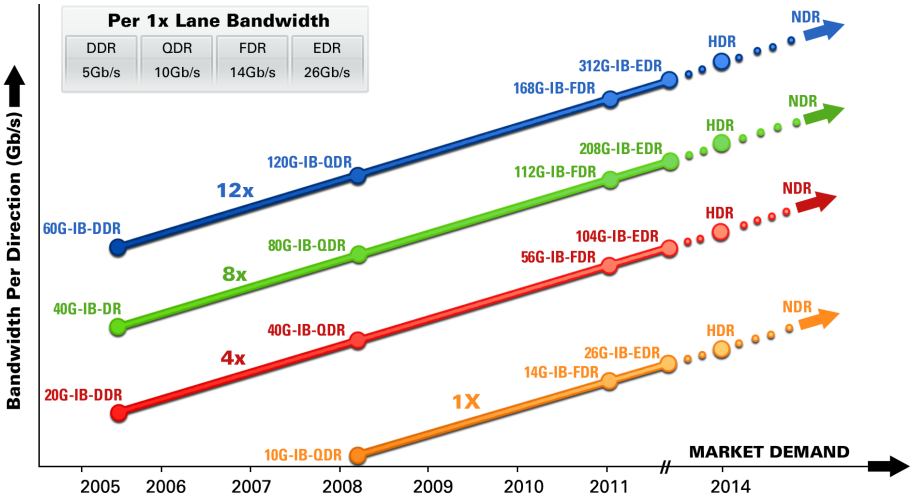
InfiniBand as a Key Advantage

So what is InfiniBand and why are more and more storage vendors moving to it for both the backplane and network connect? InfiniBand is a standards based protocol that came into existence circa 2000. InfiniBand was a merger of two technologies, NGIO and Future I/O, which were competing to be the PCI bus replacement technology. By design, InfiniBand has the characteristics of a bus technology. In fact, PCI Express, the eventual PCI replacement technology, is conceptually a subset of InfiniBand.

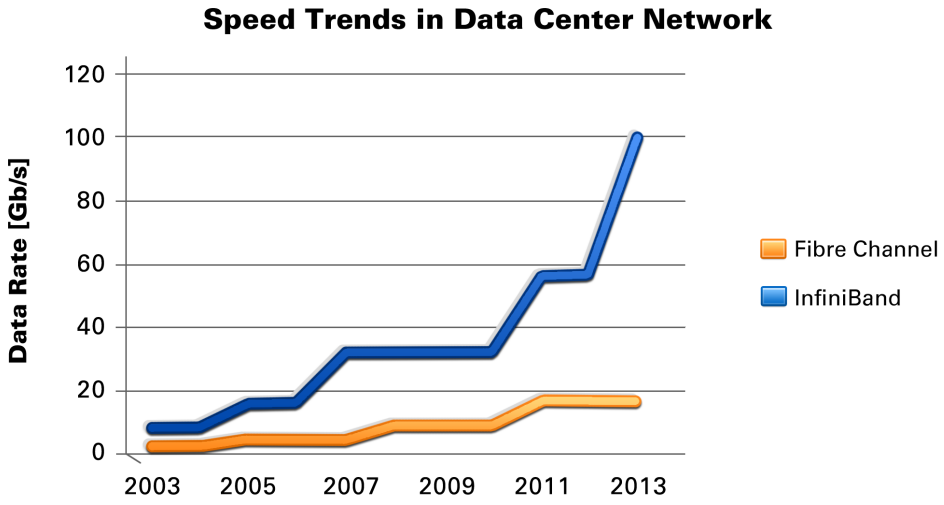
InfiniBand's core differentiator is twofold. First, it uses a credit based flow control system. This means that data is never sent unless the receiver can guarantee sufficient buffering. This makes InfiniBand a lossless fabric like Fibre Channel. Secondly, InfiniBand natively supports Remote Dynamic Memory Access (RDMA), the ability to move data between memory regions on two remote systems in a manner that fully offloads the CPU and operating system. A concept that is a legacy of its original bus design, RDMA is critical to allow distributed systems to scale. InfiniBand with RDMA enables a number of key advantages.

InfiniBand's physical signaling technology has always stayed well ahead of other network technologies, allowing the greatest bandwidth of any networking protocol. InfiniBand today runs at 56Gb/s with a road map to get to EDR (100Gb/s) in one-and-a half years.

InfiniBand Roadmap



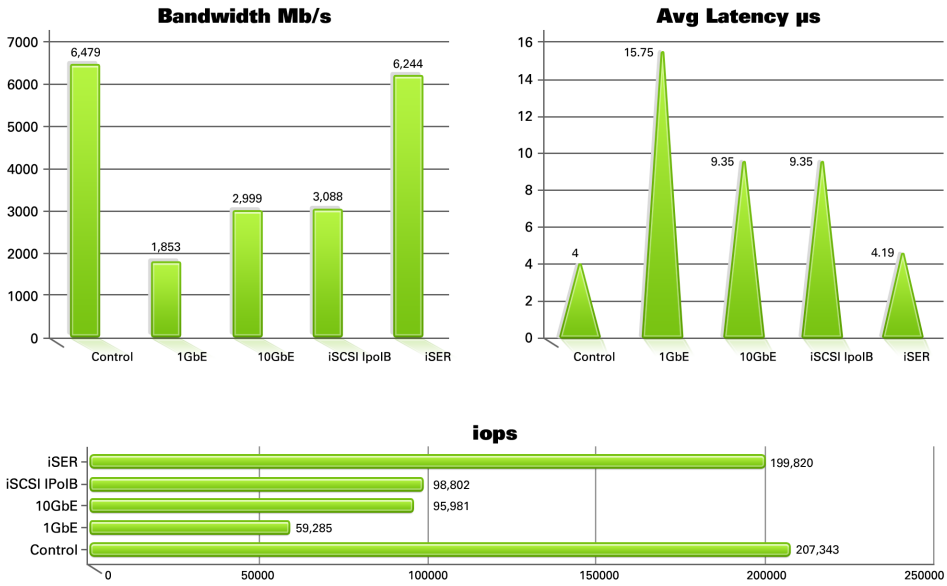
The name InfiniBand itself is a reference to the bandwidth promise. The InfiniBand roadmap is deliberately designed to guarantee that bandwidth of a single link will remain greater than the data rate of the PCI Express (PCIe) bus. This allows the system to move data over the network as fast as it can possibly generate it, without ever backing up to due to a network limitation. This effectively makes the bandwidth of InfiniBand infinite.



Although bandwidth is the probably the best known property of InfiniBand, the benefits of RDMA actually result in more performance gain for most storage applications. InfiniBand's ability to bypass the operating system and CPU using RDMA allow much more efficient data movement paths. The operating system is responsible for managing all resources of the system, including access to CPU and IO devices. The normal data path for protocols like TCP, UDPO, NFS and iSCSI, all have to wait in line with the other applications and system processes to get their turn on the CPU. This not only slows the network down it uses system resources that could be used for executing the jobs faster.

The RDMA bypass allows the data path for InfiniBand traffic to skip lines. Data is placed immediately when it is received without being subject to variable delays based on CPU load. This has three effects. First there is no waiting, so the latency of transactions is incredibly low. Raw RDMA ½ RTT latency is sub microsecond. In the diagrams below you can see access time for a storage application running over iSER, an RDMA enabled version of iSCSI. Note they are very close to the access times for local access. Secondly, because there is no contention for resources the latency will be consistent. Finally, by skipping the OS using RDMA results in a large savings of CPU cycles. With a more efficient system, those saved CPU cycles can be used to accelerate application performance.

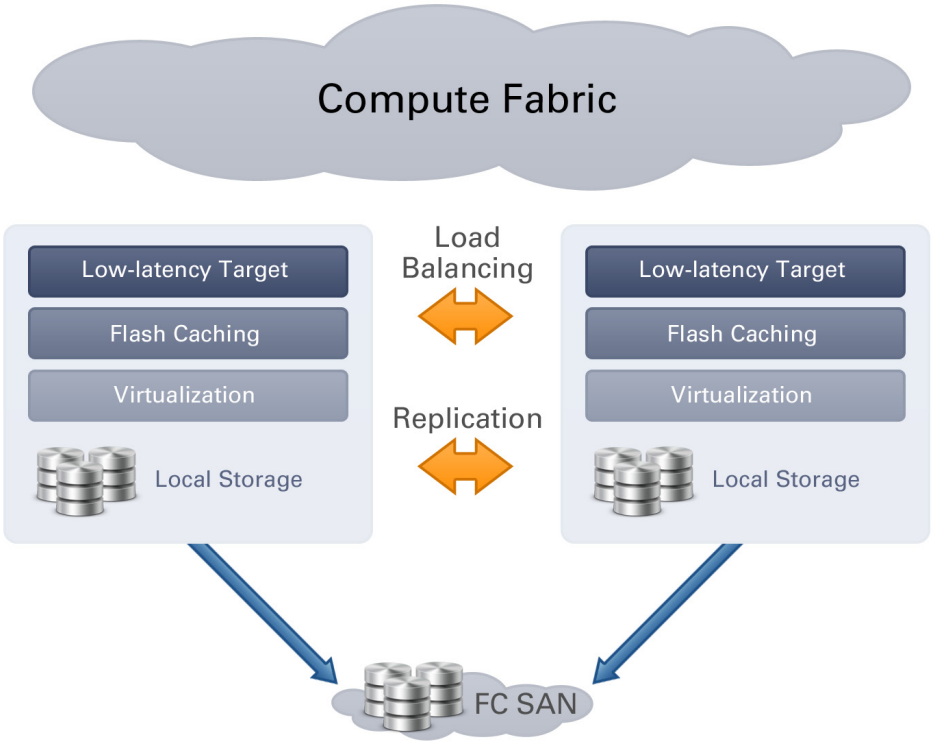
The effect of RDMA on storage performance can be seen in the test below. In this lab test an initiator and target server are connected through an InfiniBand switch (For the Ethernet portion of the testing the systems are connected back to back without a switch). First a control experiment is run locally on the storage target system. The goal of the control is to set a baseline for the maximum performance expected from the target system. The same test is then performed over a variety of transports, including iSCSI over 1 GigE, iSCSI over 10GigE, iSCSI over IPoIB (a non-RDMA InfiniBand transport) and iSER a RDMA enabled variety of iSCSI.



The results of this simple test are very telling. There is a large performance increase between iSCSI over 10GbE vs. 1GbE. However, we don't see as great an impact when we go from 10GgE to 40GgE InfiniBand. The reason is that the target system cannot saturate the full InfiniBand capacity. The network is not the bottleneck the CPU is. It takes many CPU cycles to process TCP and iSCSI operations. However, when the RDMA (iSER) protocol is used with the InfiniBand link, we are able take advantage of the additional capacity. With iSER a level of performance equal to 96% of our storage systems capability is possible.

VSA Enables Solutions from a Core Technology

Based on the performance of InfiniBand, it is no wonder it is becoming the interconnect of choice for the storage appliances. Demand for scalable applications, like distributed databases, Hadoop, clouds and HPC are creating a demand to bring InfiniBand connectivity directly to the application.



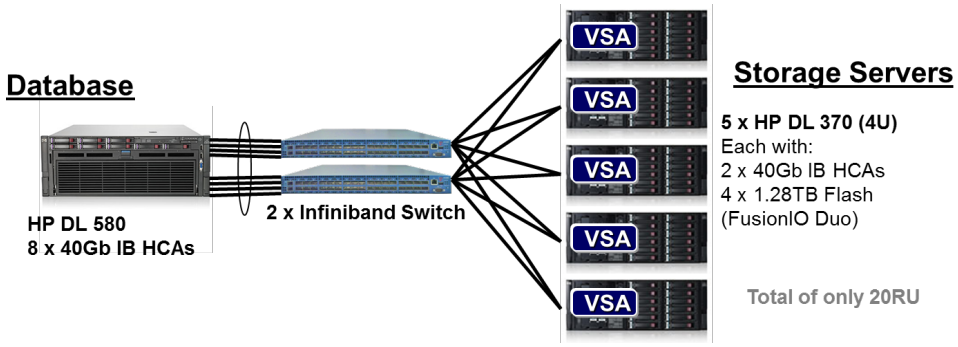
Mellanox Technologies, the world's leading InfiniBand vendor, believes strongly that InfiniBand is the storage interconnect for these applications. Mellanox provides to its partners and integrators a storage acceleration software product called VSA, which is a software platform built around the iSER technology. VSA simplifies the process for customers and appliance manufacturers to adopt InfiniBand for storage.

VSA is designed to enable a variety of storage architectures. It is a lightweight block based storage layer that provides high performance iSER and iSCSI targets. In addition, VSA centralizes the management of features like multimode provisioning, monitoring, replication and high availability. Among its features, VSA offers:

- Central Management of all Storage in a Cluster
- High Performance iSER/iSCSI Targets
- Support to Use Flash Memory or SSD as a Caching Tier
- Replication via RDMA
- Load Balancing
- High Availability
- Fibre Channel Bridging

Database

VSA based appliances are being used to enhance the performance of clustered databases by providing a "Fan-In" approach. High-end SMP machines, like the HP DL 580, can process more transactions than can be handled by local storage or a dedicated SAN. A solution is to provide a virtual RAID across disks from five physical storage servers. In this example, InfiniBand was able to saturate the IO bus of the DL580.



In this configuration, performance was 23Gb/s with 2.5M Random IOPs. An equivalent storage configuration with Fibre Channel would require 50 FC wires!

Data warehouse Petabyte Storage

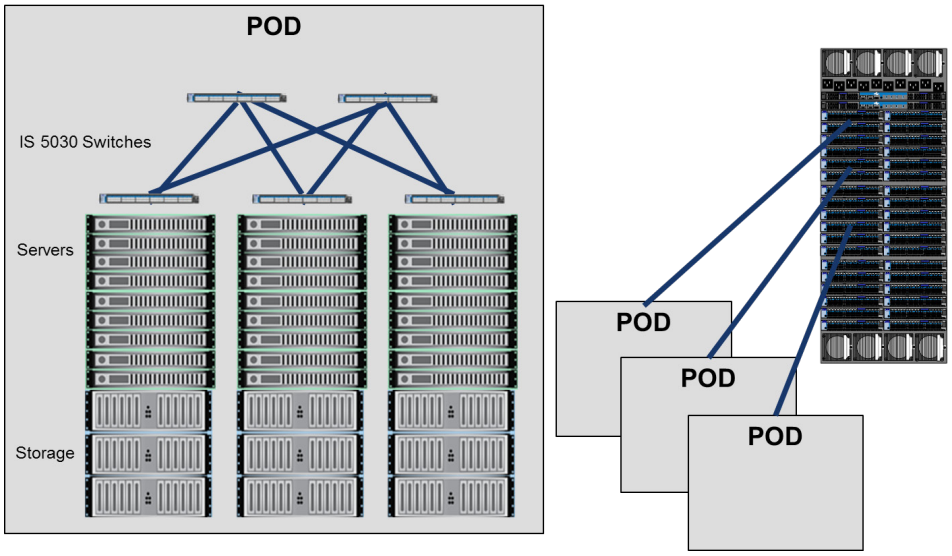
The volume of data required for web 2.0 applications, cloud storage, and big data applications as well as accounting and traceability standards are all factors that are driving the capacity requirements of data warehousing applications. The primary metric considered in data warehousing is dollars per gigabyte. VSA appliances are available today that provide not only the best capacity at the lowest prices, but also provide un-paralleled performance. The usage of VSA enables scalable storage design that provides linear performance scalability as well as central management.

Cloud Solutions

A strong cloud design is built on the principals of scalability and elasticity. Decoupling the storage from the hypervisor in a cloud architecture allows for better elasticity through faster provisioning, better use of resources and quicker live migrations. Traditional SAN/Network cloud architectures require separate dedicated adapters for storage, network, management and live migration. InfiniBand's capacity allows the same network used for storage to be used for live migration, management and standard network traffic. This allows the hypervisor nodes to contain single network connectivity, while still getting the benefit of having a dedicated SAN for storage. With full support for PXE boot over InfiniBand, the VSA allows

decoupled storage with diskless thin hypervisors. The RDMA offload of the CPU allows more VM's per hypervisor.

InfiniBand has the ability to scale to tens of thousands of nodes on a single layer 2 subnet, making provisioning for very large clouds simple.



Summary

Raising enterprise storage needs can be met with economical storage fabric of high bandwidth, low latency and high IOPs while keeping data center cost low by using Mellanox solutions. Mellanox provides InfiniBand based storage connectivity solutions for customers to converge their infrastructure connectivity on a single fabric of their choice while saving data center power and space.

Mellanox VSA, InfiniBand and Ethernet based storage solutions offer significant higher performance at a lower price. This translates into real world customer advantages such as maximize server utilization, increased application performance, reduced backup times, greater system consolidation, lower power consumption and lower total cost of ownership (TCO).

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com