Storage at a Distance; Using RoCE as a WAN Transport

Paul Grun
Chief Scientist, System Fabric Works, Inc.
(503) 620-8757
pgrun@systemfabricworks.com

Why Storage at a Distance – the Storage Cloud

Following close on the heels of Cloud Computing, the notion of the Storage Cloud is a powerful concept now finding significant traction in the enterprise. In the last generation, SANs within the data center served to virtualize storage and in so doing disaggregated storage from the compute platforms hosting application processing. This yielded important benefits in the way data is managed and in the way in which applications access and share data. Among those benefits is the ability to flexibly relocate application processing as needed to satisfy demands for performance or load balancing. In the same way, federating information storage across multiple data centers into a storage cloud brings unprecedented flexibility and reliability to the enterprise. These benefits are driving progress in the realization of practical private storage clouds.

The Challenge of Storage over the WAN

Creating a storage cloud, by definition, requires an ability to keep data synchronized even though it is stored in two or more distinct geographies. This is a delicate balancing act that requires attacking three key issues:

1. The efficient transfer of large data blocks or files over long distances,
2. Caching technologies that can help to overcome some of the distance delays, and
3. Synchronization and coordination among the storage sites

The second and third items on the list are being addressed by a number of storage technology vendors such as EMC. But the issue of efficient high-speed transfer remains elusive.

Users have become accustomed to accessing locally stored data at nearly electronic data rates. The problem with so-called cloud computing is that the delays in accessing storage at a distance can be orders of magnitude beyond what is experienced when accessing local storage. Some of this is due to speed of light delays and cannot be reduced, but a huge amount of it arises in the low level details of transmitting large blocks of data over long distance links using traditional networking technologies. If we can successfully address these issues, we can bring storage access delays down to a realm where they can be managed using caching and other advanced data management techniques. In fact, that should be our goal: When designing a long distance interconnect for storage, we should strive to raise the performance level of remote storage to the point where the performance delta between remote storage and local storage can be effectively concealed through caching techniques.

So what are the challenges that inhibit high-speed storage access over the WAN? We have already mentioned speed of light which can be expected to contribute delays. Even though this number cannot be reduced, there are other significant contributors that can and should be addressed.

Current methods for synchronizing or moving data between remote sites depend on a file copy application like FTP, which in turn depends on a reliable transport such as TCP. As part of its reliability

mechanism, TCP employs a sliding window protocol ("congestion window") to detect failures or collisions within the network. These failures or collisions result in dropped packets and a subsequent loss of reliability and they can impact the performance of the network. The congestion window determines the number of bytes that the sender can transmit before it must stop and wait for an acknowledgement from the receiver. Once an acknowledgement is received, the sender can again begin transmitting until the congestion window is again full.

A key determinant of the performance of any windowing protocol like TCP is the bandwidth-delay product of the interconnect. As its name implies, the bandwidth-delay product is calculated as the product of the bandwidth of the wire multiplied by the length of the wire measured in units of time; thus the bandwidth-delay product is a measure of the amount of data that can be stored 'in transit' on the wire. There is an obvious connection between the bandwidth-delay product of the wire and the TCP congestion window; if the maximum congestion window supported by the transport protocol (TCP) is much smaller than the bandwidth-delay product for the wire connection, the transport will not be able to keep the wire continuously full; it will have to stop transmitting periodically while it waits for bytes to propagate down the length of the wire and for acknowledgments to be returned. This results in large gaps in time where the wire is idle as the transmitter waits for acknowledgements to be returned from the far end. For LANs, the bandwidth-delay product of a reasonable wire length is on the order of magnitude of a typical TCP congestion window size. For WANs, however, the delay component of the product becomes increasingly significant. For example, the delay associated with an 8600 mile WAN represents about 163mS of roundtrip delay. For a 100Mb/s link, this represents a bandwidth-delay product of about 200KB; well within the size of a typical congestion window. As the wire speed increases to 1Gb/s, the bandwidth-delay product increases to 20MB and at 10Gb/s it increases again to 200MB. Given that typical TCP implementations use a default window size of 4MB, it is clear that the bandwidth-delay product swamps the available congestion window. Thus, as the bandwidth-delay product of the wire increases, either due to increasing length or increasing bandwidth, so too must the congestion window size.

Naturally, there is a tradeoff involved in increasing the congestion window; a larger window size means that more bytes are in-flight at any given point in time and thus at risk of loss due to congestion dropping or error. And, as the congestion window size increases, so too does the overhead associated with recovering from a lost byte. At some point it becomes impractical to increase the window size further. There is anecdotal and experimental evidence to suggest that TCP is satisfactory over the WAN at 1Gb/s speeds, and decreasingly less so at 10Gb/s speeds. Extrapolating from the empirical data suggests that significant problems will arise at the emerging generation of 10Gb/s and grow worse as 40Gb/s and 100 Gb/s links emerge during 2011. The DOE's ESnet, for example, which currently runs at 10Gb/s speeds, is expected to graduate to 40 and 100 Gb/s speeds during 2011.

RDMA over the WAN

Remote Direct Memory Access (RDMA) is a technology that is well established in the HPC space and is rapidly gaining traction in the enterprise space. An easy way to think of RDMA is as a message passing service; it enables applications to exchange messages with each other directly without requiring the involvement of the operating system. While the acronym 'RDMA' originally represented a particular memory access method and programming style, as the technology has emerged in the marketplace the simple acronym has come to represent much more than that. The nuance arises because of the 'R' in RDMA, which refers to remote, meaning outside the local server, and thus implies a networking technology. It turns out that efficiently implementing the RDMA access method over a network depends on a network transport mechanism which is suitable for transporting memory buffers (i.e. messages)

from one server to another. It is for this reason that the leading implementation of RDMA in the marketplace does not use the TCP transport which is considered to be inefficient for message based transfers.

The best-known implementation of RDMA is the InfiniBand Architecture, which is widely recognized in HPC and within data centers for its extremely low end-to-end latencies and its ability to reduce the CPU and memory bandwidth burdens that are common to standard network protocols. These performance characteristics are due largely to the remote memory access method, which is key to avoiding both buffer copies and CPU involvement in the endnodes. But what of the RDMA transport?

While well known within the data center, InfiniBand is less well known for its performance over the WAN. However, over the past several years the DoD has been engaged in a demonstration project dubbed *Large Data* which, among other major breakthroughs, has demonstrated great success in applying RDMA over the WAN for interconnecting widely distributed parallel storage systems. The Large Data system consists of multiple geographically distributed nodes; each node comprises a complete data center including servers for application processing and a large scale parallel file system and accompanying storage. At the heart of each of the nodes is a low latency InfiniBand 40Gb/s network.

The geographically distributed nodes are connected to one another via 10Gb/s WAN links to create a fully connected mesh. By executing the RDMA protocol over the WAN links instead of the better known TCP, the RDMA networks at the heart of each of the nodes are effectively merged into a large, high performance RDMA network. Because of this interconnection of the nodes, a user of the system, such as an intelligence analyst, is presented with a view of not only the data stored at his local node, but of all the data stored anywhere on the entire large data system. In effect, the Large Data system is a complete cloud architecture with an effective physical diameter on the order of 10,000 kilometers.

Experimental results clearly show the system's ability to transport storage data over the WAN at sustained rates very close to available wire speed – just under 10Gb/s. By comparison, when interconnected via conventional TCP/IP protocols running over the same 10Gb/s WAN links, the sustained data rates for more than one flow dropped precipitously to just a fraction of the available WAN wire speed. So what is it about InfiniBand that accounts for its remarkable performance over the WAN?

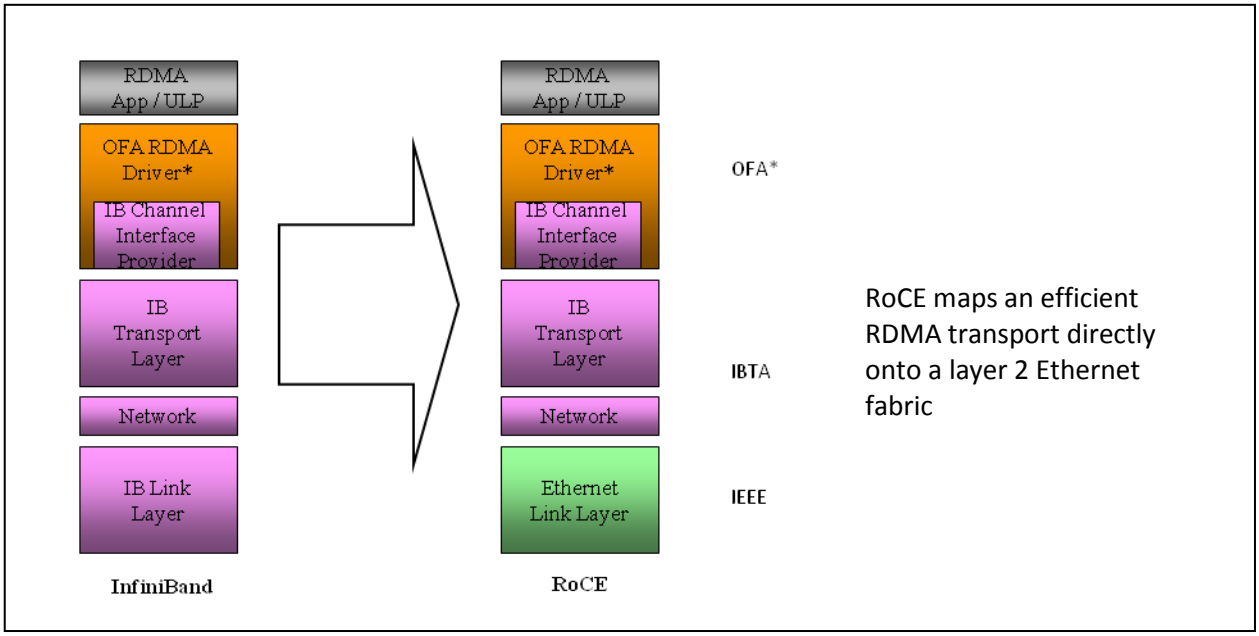
As described above, the bandwidth-delay product associated with the network, and its relationship to the transport's sliding congestion window, are the key determinants in achieving maximum sustained throughput over the WAN. The InfiniBand transport, like TCP, is based on a sliding congestion window. But unlike TCP, IB's congestion window is based on outstanding packets instead of outstanding bytes. With a packet sequence numbering range of more than 8 million packets, and each packet ranging in size from 256 bytes to 4K bytes, it is easy for the InfiniBand transport to completely fill even the longest wire long before its congestion window is filled. Is there a way to capture the features of the InfiniBand transport that make it suitable for the WAN without the requirement for a wholesale upgrade to an InfiniBand network? The answer is yes.

RDMA Implementations – RDMA over Ethernet – RoCE

As described so far, wringing the best possible bandwidth utilization from 10Gb/s and faster WAN links is facilitated through the unique features of the InfiniBand transport protocol; in particular, its message orientation coupled with its use of a packet-based sliding congestion window. Up until now, the only

route to accessing the InfiniBand transport has been to deploy a complete IB network, including the use of WAN gateway adapters to allow the extension of the RDMA protocols over the WAN.

Recently, the InfiniBand Trade Association announced the release of a new specification, called RoCE (RDMA over Converged Ethernet). This new specification defines a mechanism for layering InfiniBand's efficient message-based transport directly over an Ethernet layer 2 fabric, replacing the InfiniBand network, link and phy layers. The notion of RoCE was proposed at the OpenFabric Alliance's 2009 Sonoma workshop and publicly discussed by the InfiniBand Trade Association at the October 2009 T11 meeting using the working title, "IBXoE". The specification was ratified by the IBTA and released in April 2010 and can be found in the downloads area of the IBTA website. Non-members of the IBTA can register and freely download and use the specification at no cost. The diagram below is a notional representation of the RoCE architecture illustrating how IB's RDMA service is cleanly married with Ethernet's layer 2 link and phy layers.



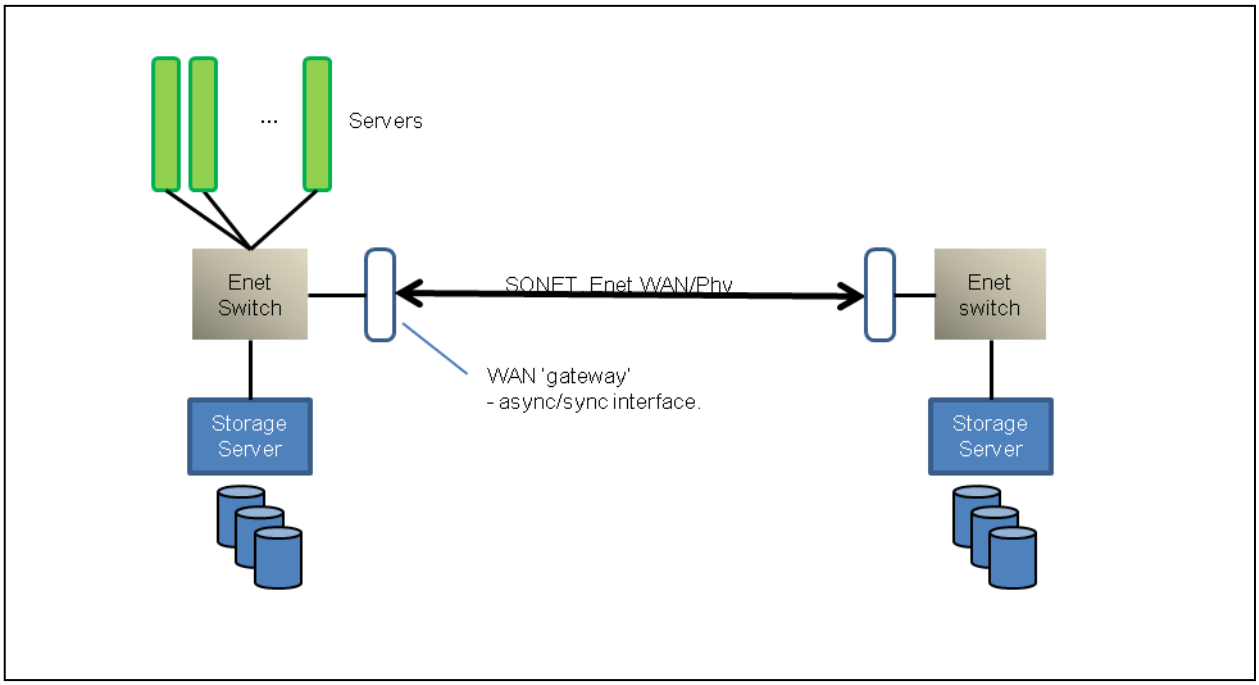
This specification for the first time permits the use of the proven RDMA transport that is at the heart of the InfiniBand Architecture directly on top of layer 2 Ethernet. This major advancement means that an IT user now has access to the performance and latency advantages associated with the best in RDMA technology, while preserving existing Ethernet cables and switches. Since RoCE conforms to InfiniBand's verbs definitions, it can be easily deployed using the Open Fabrics Enterprise Distribution (OFED) software which is freely available in open source from the Open Fabrics Alliance.

There are presently two different, but wholly interoperable implementations of the RoCE specification available. Mellanox' Connect-X HCA fully supports RoCE with a hardware-based transport engine. And System Fabric Works has made available through the OFA a software implementation of the transport engine, known as 'soft RoCE'. This implementation, which interoperates with the Mellanox hardware version, is used in conjunction with a standard 10Gb/s Ethernet NIC.

Implementing RoCE in the Storage Cloud

A storage cloud can be implemented as two or more storage sites, usually combined with application processing capability and connected together with wide area links. Each node is built around the usual

hierarchy of Ethernet switches, with one or more of the Ethernet switch ports exiting to the WAN. The WAN port is used to keep the cloud nodes synchronized. The figure below describes the basic concept of a two node storage cloud.



By taking advantage of RoCE's transport characteristics in the WAN, it should be possible to dramatically increase throughput across the WAN links compared to FTP/TCP and in so doing reduce the performance delta between local storage and remote storage. Ideally, the delta can be reduced to the point that caching techniques can be effective in hiding performance differences. As a distinct side benefit, the use of RoCE as the WAN transport also improves the utilization of the relatively expensive WAN links, thus maximizing the return on investment in the WAN equipment and links.

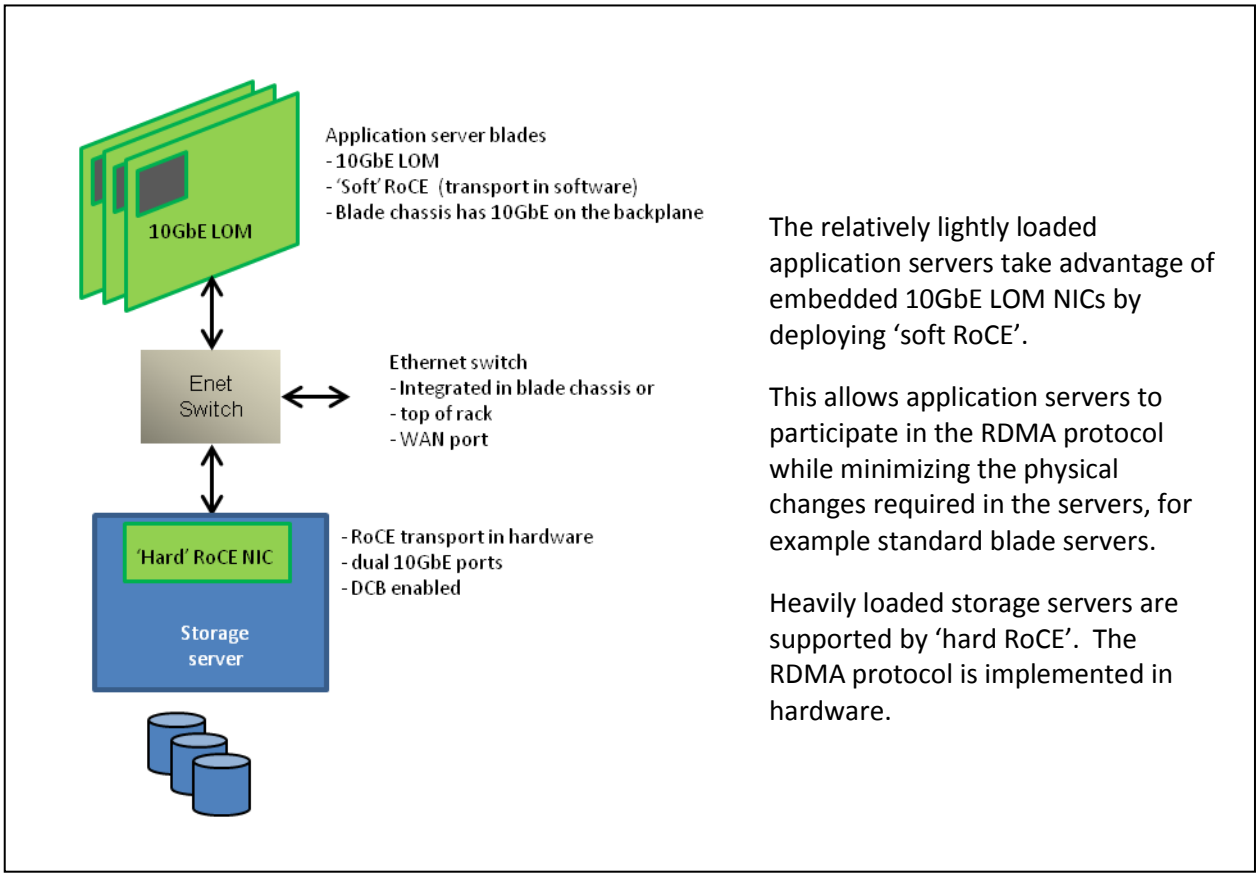
Running the RoCE protocol over the WAN can be effective for a range of different WAN choices, for example SONET or Ethernet WAN PHY are likely choices. In the SONET case, a WAN gateway device is required in order to provide the interface between the asynchronous Ethernet data center network and the synchronous SONET network. For Ethernet WAN PHY, a standard Ethernet switch could support an integrated WAN PHY port potentially eliminating the need for a relatively expensive WAN gateway device.

The diagram below illustrates one storage cloud node. Within the node, application and storage server(s) are interconnected via RoCE running over conventional Ethernet switches. As described, the set of nodes comprising the cloud benefits from the advantages of using RoCE as the WAN transport. But more than that, each individual node within the cloud also benefits from the performance advantages that RDMA brings to each server; for example, RDMA's ability to deliver data directly to a user's virtual address space dramatically reduces latency, eliminates buffer copies, reduces the demand on memory bandwidth and reduces CPU utilization considerably. These advantages are the traditional realm of RDMA technology and are independent of the use of RoCE as a WAN transport.

Note the illustration of two different types of 10GbE NICs in the diagram below. The heavily loaded storage servers are connected via 'hard RoCE' NICs. These NICs implement the complete RoCE RDMA

protocol in hardware and provide maximum performance, lowest latency and lowest CPU utilization. The application servers, on the other hand, are assumed to be more lightly loaded and can be connected via 'soft RoCE'. Soft RoCE consists of the RDMA protocol implemented in software coupled together with a conventional 10GbE NIC.

Although not required by the RoCE specification, one option is the use of the newly emerging Data Center Bridging variation of Ethernet. This variation provides certain features, such as link level flow control and congestion control mechanisms that are intended to reduce the instance of dropped packets in the Ethernet fabric.



Summary

The emergence of high-speed networks extended over long geographical distances demands a re-look at networking transport protocols. Although many approaches to fine tuning TCP's congestion window mechanism have been proposed (High Speed TCP, H-TCP, BIC and others) each of them represents a fairly sophisticated tuning of the TCP protocol. Although these improvements can yield some performance gains, the improvements are incremental. The opportunity for large scale advances in the well-known TCP transport appears to be greatly receding. Something new is required to transport data over long distances.

A by-product of the emergence of practical and effective RDMA technologies is the appearance of a transport protocol which is capable of tolerating high bandwidth-delay product environments. The InfiniBand Trade Association has recently released a specification for RoCE, which, for the first time,

combines the RDMA transport protocol with Ethernet. This development opens the possibility of applying RDMA technologies, including in the WAN, in environments where it is not practical or desirable to deploy a non-Ethernet fabric.

The ability to transfer large data blocks or files over long distances at 'reasonable' bandwidths and latencies will propel the trend toward greater data center flexibility. It can do this by making it possible to federate storage over long distances, effectively decoupling the data storage location from the application execution location. In addition, with the advent of fast long distance links, the barriers to migrating entire virtual machines among data centers are substantially lowered. Again, this is another important step in improving the flexibility with which IT resources can be allocated.

About System Fabric Works

System Fabric Works (www.systemfabricworks.com) is an Austin Texas based engineering company providing consulting, professional engineering services and products to assist our client in maximizing the value of his investment in applications, hardware and systems. By focusing on the application's needs for data movement and data storage, we are able to assist our client by recommending specific I/O and storage technologies, providing engineering services including software development and assisting with system integration and deployment. We also supply and support a highly customizable range of high performance, low cost storage and networking products. We are worldwide experts in the development and application of RDMA technologies and are widely recognized as leading experts in developing and deploying advanced software such as the Open Fabrics Enterprise Distribution (OFED) software stacks from the Open Fabrics Alliance.