WHITE PAPER

October 2014

RoCE in the Data Center

Introduction 1
Background 1
The Case for RDMA 2
Why RoCE? 3
RoCEv2 3
Conclusion 3

Introduction

In a world of ever-increasing data, the speedy transfer of all that data is critical to the information being used efficiently. Interconnect based on Remote Direct Memory Access (RDMA) offers the ideal option for boosting data center efficiency, reducing overall complexity, and increasing data delivery performance. RDMA allows data to be transferred from storage to server without passing the data through the CPU and main memory path of TCP/IP Ethernet. Greater CPU and overall system efficiencies are attained because the storage and servers' compute power is used for just that –computing, instead of processing network traffic.

RDMA enables sub-microsecond latency and up to 56Gb/s bandwidth, translating to screamingly fast application performance, better storage and data center utilization, and simplified network management.

Until recently, though, RDMA was only available in InfiniBand fabrics. With the advent of RDMA over Converged Ethernet (RoCE), the benefits of RDMA are now available to data centers that are based on an Ethernet or mixed-protocol fabric as well.

Background

Traditional interconnect protocols based on TCP/IP or UDP are far less efficient than RDMA because of some very fundamental challenges.

First, these legacy protocols require data to be written to the memory buffer on both the send and receive sides of a data transfer. This takes valuable resources away from the CPU's primary compute responsibilities and dedicates them instead to repetitious copying and reading of input/output processes to and from the memory buffers.

Additionally, the Sockets API is used as the interface for an application to access the network, which requires two-sided communication. Before transferring can begin, a send request must be delivered and a response acknowledging and granting permission for the request must be received. This extra step in the interconnect process adds to the overall transfer time and burns compute resources on the remote device.

RDMA, on the other hand, was designed to address these challenges. By building OS bypass, zero copy, and CPU offloading into the architecture, RDMA was planned with a view to high performance.

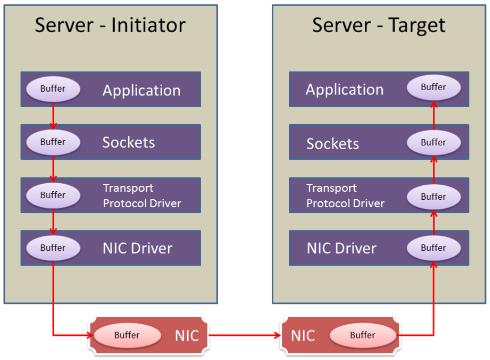

Figure 1. *TCP/IP communication*

OS bypass gives an application direct access to the network card, allowing the CPU to communicate directly with the I/O adapter, bypassing the need for the operating system to transition from the user space to the kernel. With RDMA, there is no need for involvement from the OS or driver, creating a huge savings in efficiency of the interconnect transaction.

RDMA also allows communication without the need to copy data to the memory buffer. This zero copy transfer enables the receive node to read data directly from the send node's memory, thereby reducing the overhead created from CPU involvement.

Furthermore, unlike in legacy interconnects, RDMA provides for the transport protocol stack to be handled by the hardware. By offloading the stack from software, there is less CPU involvement, and the transport is more reliable.

The overall effect of the significant reduction of CPU overhead that RDMA provides by way of OS bypass, zero copy, and CPU offloading is to maximize efficiency in order to provide lightning fast interconnect.

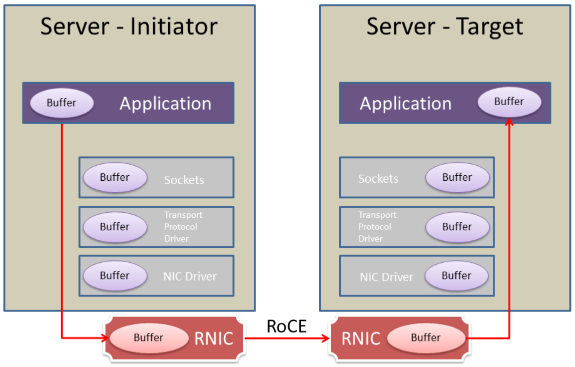

Figure 2. *RDMA communication*

The Case for RDMA

Today's data centers demand that the underlying interconnect provide the utmost bandwidth with extremely low latency. No matter the market, low latency has become an absolute necessity.

For example, mobile, gaming, and video-on-demand use low latency to ensure a real-time, consistent response. In financial markets where high performance computing is required, exceptionally low latency can mean the difference of millions of dollars. The scale-out of data centers requires higher performance, as do storage-to-server and server-to-storage transactions. Similarly, the move to Solid State Disk (SSD) storage has made latency relevant to the storage market as well.

While high bandwidth is important, without low latency bandwidth is not worth much. Moving large amounts of data through a network can be achieved with TCP/IP, but only RDMA can produce the low latency that avoids costly transmission delays. Moreover, RDMA offloading reduces jitter, which means that the low response time is that much more consistent.

Why RoCE?

CIOs and application writers have long recognized the advantages of RDMA and therefore advocated for InfiniBand infrastructure. Nonetheless, some IT managers prefer not to migrate from their existing Ethernet data centers or to learn a new protocol.

RDMA over Converged Ethernet (RoCE) allows all the advantages of RDMA, but on existing Ethernet networks. With RoCE there is no need to convert a data center from Ethernet to InfiniBand, saving companies massive amounts of capital expenditures. There is no difference to an application between using RDMA over InfiniBand or over Ethernet, so application writers who are more comfortable in an Ethernet environment are well-covered by RoCE.

Basically, RoCE finally brings RDMA technology into Ethernet-based data centers, enabling such data centers to benefit from the low latency of RDMA without having to adopt an InfiniBand-based network infrastructure.

RoCEv2

The latest version of RoCE adds even greater functionality. By changing the packet encapsulation to include IP and UDP headers, RDMA can now be used across both L2 and L3 networks. This enables Layer 3 routing, which brings RDMA to networks with multiple subnets. IP multicast is now also possible thanks to the updated version.

Conclusion

Until the advent of RoCE, there were two very limited options for solving poor data center performance. However, with RoCE there is an excellent option available, both in terms of performance and in terms of savings. RoCE enables efficient data transfer on existing Ethernet infrastructure, providing many of the benefits of InfiniBand without the expense of adding vast amounts of hardware or a large-scale conversion.

Thanks to RoCE, it is finally possible to experience the lowest available interconnect latency in a legacy Ethernet data center.

Mellanox TECHNOLOGIES

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