SOFT-RoCE

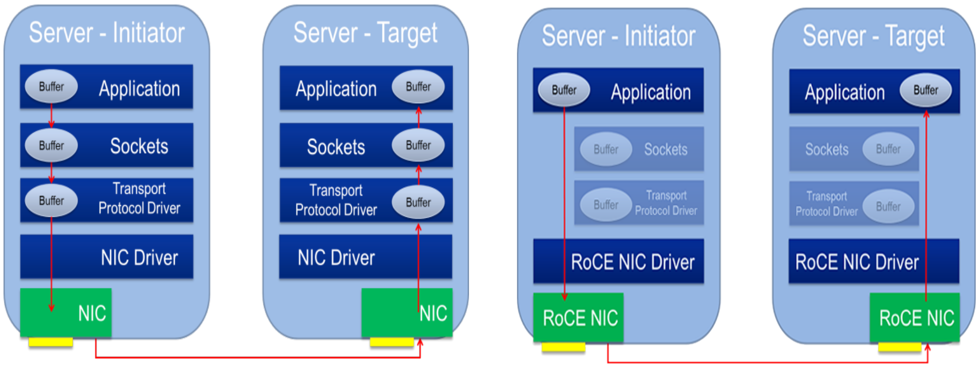
RDMA TRANSPORT IN A SOFTWARE IMPLEMENTATION

June 2015

In today's world, data growth is exploding at unprecedented rates, and fast data transfer within the data center is critical toward efficient information use. Interconnects supporting Remote Direct Memory Access (RDMA) technology are the ideal option for boosting data center efficiency, reducing overall complexity and increasing data delivery performance. RDMA enables data to be transferred from storage to server, server to server, and server to storage without the CPU and operating system directing all of the movement. This results in storage and servers attaining greater CPU and overall system efficiencies as the compute power is used for just that – computing, instead of processing network traffic.

RDMA enables bulk memory transfers with sub-microsecond latency and high bandwidth, translating to faster application performance, better storage and data center utilization, and simplified network management.

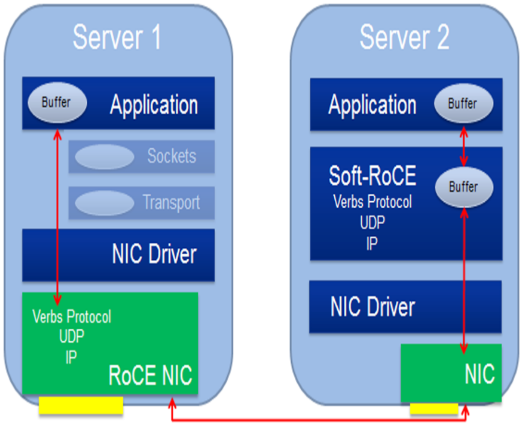
> *"The advent of the industry standard 'RDMA over Converged Ethernet' (RoCE) brought the benefits of RDMA to data centers that utilize an Ethernet or mixed-protocol fabric."*



Traditional server-to-server communication (*left*) versus RDMA over Converged Ethernet (RoCE) server-to-server communication (*right*).

The advent of the industry standard "RDMA over Converged Ethernet" (RoCE) brought the benefits of RDMA to data centers that utilize an Ethernet or mixed-protocol fabric. All of today's major operating systems already have RoCE drivers built in and Ethernet NICs with RoCE support interoperate with Layer 2 and Layer 3 Ethernet switches. RoCE is now a well-established protocol for enabling RDMA for Ethernet-based clusters.

However, until now RoCE has been confined to a hardware implementation. One new project changing this model is Soft-RoCE, a software implementation of the RDMA transport that makes RoCE technology available over any Ethernet-enabled servers. This project has been bolstered by open-source development in a Github community project, with primary contributions from IBM, Mellanox and System Fabric Works. Following rigorous testing and validation of the implementation, Soft-RoCE is now ready for Linux upstream submission.



Soft-RoCE implements the packet processing otherwise managed by the RoCE NIC.

Soft-RoCE leverages the same efficiency characteristics as RoCE, providing a complete RDMA stack implementation over any NIC. It is a transaction-oriented wire-protocol that decouples congestion control from reliability and enables data to be written directly to pinned application buffers where memory is registered. A Soft-RoCE device instance can be bound to any NIC in a lossless network.

"Soft-RoCE leverages the same efficiency characteristics as RoCE, providing a complete RDMA stack implementation over any NIC."

These characteristics make Soft-RoCE exceptionally efficient compared to TCP. Soft-RoCE avoids almost all system calls, providing zero-copy on send transactions and a highly efficient one-copy on receive, in which the destination buffer is guaranteed to be pinned and accessible to all CPUs. Therefore, Soft-RoCE will provide much lower latency and significantly higher bandwidth than TCP.

There are numerous real-world uses for Soft-RoCE that can take advantage of both its accessibility and the acceleration that it brings to the data center. For example, a data center can connect servers with simple Ethernet adapters to high performance storage appliances, which feature hardware-based RoCE, using iSCSI over RDMA (iSER). As the data center upgrades its servers over time, each can be equipped with a HW RoCE adapter for added performance improvement.

Soft-RoCE is also valuable in setups already implementing hardware-based RoCE for high performance servers that seek a less expensive, easy-to-implement software-based RoCE for their many client devices.

Another area in which Soft-RoCE can make a difference is virtualization. Virtual Machines have access to native RDMA performance through SR-IOV RoCE adapters. However, hardware virtual functions are exposed to the VMs in such solutions. With Soft-RoCE, VMs can benefit from RDMA performance without hardware exposure.

Soft-RoCE is the next important building block in the RDMA ecosystem. It provides an extremely high-performance software transport for both user-space and kernel applications, and it makes RoCE available to everyone, no matter the hardware vendor. Soft-RoCE adds a layer of accessibility to RDMA that was previously lacking, while continuing to offer the same efficiency that RDMA has always afforded.

About The RoCE Initiative

The RoCE Initiative promotes RDMA over Converged Ethernet (RoCE) awareness, technical education and reference solutions for high performance Ethernet topologies in traditional and cloud-based data centers. Leading RoCE technology providers are contributing to the Initiative through the delivery of case studies and white papers, as well as sponsorship of webinars and other events. For more information, visit www.RoCEInitiative.org.