RoCE – Plug and Debug

In Search of a User Friendly RDMA over Ethernet Solution

Executive Summary

This paper discusses the advantages of RDMA and the alternatives available over Ethernet. Drawing on real life IT experience, it highlights the difficulties and costs associated with using RoCE (InfiniBand over Ethernet) in deployment and maintenance. RoCE comes out to be a Plug-and-Debug offering that suffers from debuggability limitations, which further exacerbates the problems. In contrast, iWARP is shown to be a true plug-and-pay solution that uses the familiar TCP/IP standards, and fully leverages existing tools and infrastructure.

Introduction

Remote DMA (RDMA) is a technology that achieves unprecedented levels of efficiency, thanks to direct system or application memory-to-memory communication, without CPU involvement. With RDMA enabled adapters, all packet and protocol processing required for communication is handled by the network adapter itself, typically in hardware for high performance. In return for the performance and efficiency benefits, RDMA does require application changes, from the popular socket paradigm to an asynchronous communication model based on a send and receive "queue pair" concept, using a set of communication "verbs" or operations.

In an era of Big Data, massive datacenters, pervasive virtualization and focus on "Green" operation and efficiency, RDMA use is rapidly gaining ground. Moreover, RDMA support is integrated into the very core of today's server operating systems. By providing high level, simplified communication abstractions, such integration further lowers the barrier to realizing the benefits of RDMA, and is further contributing to the acceleration in RDMA adoption. A clear example of this movement is seen in two key applications that have been identified and targeted in Windows Server 2012, namely high performance file storage (SMB) and Virtual Machine migration in virtualized systems. In fact, the latter builds upon the native RDMA support introduced into SMB to seamlessly achieve unprecedented levels of performance in Virtual Machine migration.

In the networking world, Ethernet is the most widely used and preferred technology, and has systematically replaced specialized fabrics. While InfiniBand has been the leading RDMA technology, for reasons both economic and technical, the vast majority of users would much rather deploy Ethernet based fabrics. Today, there are two competing technologies that provide RDMA capability over Ethernet networks: iWARP, the IETF standard for RDMA over Ethernet, and the IBTA's IB over Ethernet, or RoCE.

This paper discusses the background behind RoCE and the issues that it introduces. It relies on real world IT expert blogs, press releases, research papers and other data readily available online to show how deploying it beyond a trivial test-bench quickly turns into a "Plug and Debug" exercise, contrasting it to the robust and familiar TCP/IP foundation of iWARP, a tried and true Plug and Play solution that is "way easier" to use.

Data Center Bridging?

The "Converged Ethernet" in RoCE refers to the need to enable and configure DCB in the network. Most evaluations start with a simple back-to-back setup, where RoCE appears to work well. Then, the next step where a larger network must be configured immediately exposes users to the difficulty of configuring network, hosts and adapters. Referred to as "a pain" and "a fight", even when configuring switches from a single vendor, it turns into a real nightmare when multiple different switch vendors are in the mix and "many, many lonely hours" trying to make it work.

While DCB consists of multiple components, some claim only priority flow control PAUSE (PFC) is needed for RoCE operation, while others say only single priority PAUSE is sufficient. Both gloss over the fact that common wisdom is to "disable PAUSE outside of the first tier switches", i.e. in any reasonable scale network. Thus, unless you have "deep enough pockets" to "build non-blocking networks", choose iWARP and "get away with cheaper switches".

Touted as an advantage of RoCE, simplicity is in fact limited to the vendor's side and benefit. In practice, "RoCE is sensitive to NIC, switch and driver (host)" configuration as all the complexity is dumped into the lap of the IT staff and end users, who end up shouldering the weight and responsibility of the missing functionality.

Range Extenders?

When trying to use RoCE over long distance links, users quickly face additional limitations. While "iWARP routes, it's not bound by a single Ethernet broadcast domain", RoCE does not. RoCE's scalability limitations are not simply due to the lack of IP headers, but also the lack of effective packet drop avoidance and recovery beyond single hop PAUSE. Studies of RoCE performance over long distance show performance plummeting as packet drop or reordering rates increase. The lack of reliability mechanisms thus forces RoCE to use range extenders, or specialized equipment that allows dealing with network variability as it attempts to go over longer distances. These extenders impose additional network engineering, acquisition, maintenance and operation costs, and typically are of limited performance and represent a choke point and single point of failure in the system.

Thus, "configuring the network for RoCE currently takes a team of experts" and WAN operation remains a subject of academic study. However, what appears exciting in the realm of academic research may be a worst case time sink for real world IT staff.

Reliability Overlay?

The main and probably sole selling point of RoCE is a simple specification that "everyone can ship easily". However, what looks simple may simply be deceptive! Vendors offering RoCE based solutions are forced to add external TCP-like reliability layers to allow communication outside of the idealized world of lossless Ethernet. Similarly to range extenders, this additional functionality increases the costs for users, compounds the complexity and multiplies the failure points to debug.

Summary: Go iWARP

This paper discussed RoCE as an alternative to iWARP for deploying RDMA over an Ethernet network. By compiling available real world experience and evidence from the field, it shows RoCE as a difficult to deploy, difficult to operate and difficult to debug technology with gaping holes in reliability and scalability, that only addresses imaginary limitations in the iWARP solution. While it may have been a "simple specification to sketch" by the vendors, it surely isn't simple from the users' point of view.

> *When looking for a high performance and user friendly RDMA over Ethernet solution,* ***iWARP*** *stands out as the* ***standards-based, mature, scalable and robust Plug-and-Play*** *option that is shipping today* ***at 40Gbps*** *from multiple vendors.*
>
> *Choosing iWARP guarantees forward and backward compatibility thanks to its stable standard foundation, while performance results show iWARP over 40GbE* ***out-performing the fastest InfiniBand FDR gear****. With in-box support in all major operating systems, it is a true drop-in replacement for IB.*

There is no reason to travel down the rocky road of RoCE, with unpleasant discoveries at every corner, and a specification that is known to be incomplete and still undergoing structural changes. In effect, today's non-routable RoCE offerings are not compatible with future RoCE over IP products, and will need to be scrapped. There simply is no reason to incur all the costs and aggravation.

The final words of one of the IT trials sum it up perfectly: "if you want RDMA the easy way… go iWARP"!

References

All the quotes in this paper have been sourced from the following online material:

Working Hard in IT Blog, *RoCE Does Not Work Without DCB*
Working Hard in IT Blog, *The RoCE Path Over DCB...*
Working Hard in IT Blog, *RoCE Requires Tagged Non Default VLANs*
Aiden Finn, *Software Defined Storage*
Roland's Blog, *Two Notes on IBoE*
Roland's Blog, *RDMA on Converged Ethernet*
Press Release, *Indiana University and Orange Silicon Valley...*

Further Reading

Chelsio Communications, *RoCE the Fine Print*
Chelsio Communications, *RoCE FAQ*
Chelsio Communications, *RoCE at a Crossroads*
Chelsio Communications, *RoCE is Dead, Long Live RoIP?*
IBM, *A Competitive Alternative to InfiniBand*