RoCE is Dead, Long Live RoIP?

Official confirmation is pending, but reports are that RoCE is dead, and work is in progress on a successor that is actually routable. Until recently, RoCE had been a specification for running InfiniBand directly over Ethernet. The name RoCE (RDMA over Converged Ethernet) underscored both the raw Ethernet encapsulation and the need for a lossless fabric to run the InfiniBand protocol.

In effect, RoCE has finally been acknowledged to be not routable, despite repeated assurances of the contrary throughout its lifetime. This should not come as a surprise, since Internet routability requires the Internet Protocol (IP), inconspicuously (!) absent from the RoCE specification. As expected, an immutable law of the Internet has been upheld, and a RoCE successor that uses the IP protocol (RoIP?) is under consideration. Since the original name is no longer suitable, this paper will use the name RoIP to differentiate the new specification from the existing one.

The RoIP effort is the second step in a long process to port the InfiniBand stack to the Ethernet world. However, the solution will remain incomplete until it builds a proper networking stack that parallels iWARP's, the existing standard for RDMA over Ethernet. Fortunately, there is no need to wait for the process to complete: today, several leading vendors offer iWARP adapters that provide InfiniBand FDR performance levels that, unlike RoCE, are scalable, routable and robust.

Introduction

The paper continues on the line of argumentation followed over the past few years that exposed the limitations and pitfalls associated with the original RoCE specification (see [1,2,3,4,5]). In summary, the arguments against RoCE are:

1. RoCE does not scale
2. RoCE does not route
3. RoCE is hard to deploy
4. RoCE is hard to use and manage
5. RoCE impacts network wide QoS
6. RoCE lacks congestion control
7. RoCE performance is sensitive to network variability
8. RoCE is not robust in a real network environment

The new RoIP effort is expected to address the routability shortcoming. This paper argues that even if the protocol definition is changed to run over IP but forgoes TCP, RoIP will still lack critical congestion and reliability mechanisms, i.e. the reasons why the IETF standard for RDMA over Ethernet (iWARP) was designed to run over TCP/IP.

Indeed, the one advantage touted by the RoCE proponents is simplicity, due to the absence of TCP in its definition, a so called "expensive" protocol to operate (itself a myth that has been thoroughly debunked).

Notwithstanding the fact that RoCE assumes a lossless network is somehow available, the path RoIP is starting on inevitably leads to the inclusion of a TCP like transport layer. It may perhaps not yet be on the agenda and ultimately may not get called TCP, but it will need to perform the key scalability, reliability and congestion management functions that TCP is responsible for in the Internet. Given the evident opposition to using TCP, the process of designing and standardizing this new protocol is expected to take years, and will still falling short on integration with the existing TCP/IP infrastructure. The expected limitations thus defeat any reason to wait for an overhaul of the specification that spans two standard bodies (the InfiniBand Trade Association and Internet Engineering Task Force – IBTA and IETF). There is no need to suffer the expected delays and iterations required, in order to make use of an RDMA solution, whereas one (iWARP) exists today and runs at the same speeds as InfiniBand.

The following sections further discuss the reasons why simply adding the IP protocol layer is not sufficient to address the requirements for a robust RDMA over Ethernet solution.

Zero Loss Operation

The RoCE specification defers error handling to Ethernet, by assuming that the user can deploy a zero loss Converged Ethernet network. Other papers have expanded on the challenges in relying on Priority Flow Control in Ethernet for this purpose (e.g., see [2 and 3]). While the concept of a zero loss operation has at least been marketed in the Ethernet space, this is a fundamentally foreign concept in the context of an IP network. One of the basic tenants of IP is “Best Effort” delivery and attempting to shoehorn lossless operation onto it is unlikely to succeed.

The RoIP specification may presumably attempt to address this gap, but doing so outside of a transport layer constitutes a violation of the Internet protocol architecture.

Scalability Limits

Encapsulating RoCE within IP headers enables routability but will not address all of RoCE’s scalability limitations, unless the new specification breaks its fundamental reliance on a lossless network and addresses reliability through other means. Again, in the Internet protocol architecture, reliability requires a transport layer (such as TCP) that provides this service over the IP protocol.

In summary, within an internetwork environment, reliability is an end-to-end problem and is handled as part of the protocol stack. This layer is absent in RoCE today, and presumably in RoIP as well.

Management Challenges

Unless RoIP is weaned off its dependence on lossless Ethernet, and the use of priority flow control, it will remain critically exposed to network wide QoS configuration details, while greatly complicating that task.

Furthermore, the absence of a transport header, be it TCP or UDP, significantly curtails the types of services and integration the RoIP traffic will receive in the network. For example, firewalls, NAT, traffic management and accounting infrastructure that is flow oriented have in large part relied on transport layer information to identify flows. RoIP will not operate properly without a standard transport header.

Congestion Control

One of the key functions of TCP is network congestion control, essential for handling network variability and load. TCP's congestion control mechanisms have been fine tuned over 3 decades to make efficient use of the available capacity in a network, regardless of the network type, speed or conditions.

No longer leveraging the confined environment of a small Ethernet subnet, RoIP will be required to implement congestion control mechanisms. Furthermore, any new mechanisms added to RoIP must be TCP compatible, and most likely simply replicate TCP's standard operation.

iWARP Solution

It should be clear that incorporating a fully featured transport layer into RoIP is necessary, although accepting this fact and its implications on the design may require several iterations of partial specifications and many years of experimentation.

All major data centers today rely on TCP/IP as a transport, which is the protocol that carries all Web traffic, and underlies Hadoop and popular Cloud stacks. The fact that iWARP is built on TCP/IP means that it inherits all the attributes and benefits of TCP/IP and levels of maturity acquired through decades of use in a wide range of the most demanding environments.

Today's state of the art iWARP adapters have highly efficient TCP/IP implementations that scale to 40Gb/s and 100Gb/s with latency comparable to IB.

There is no need to rediscover issues long resolved, or to design and iterate for years on a brand new set of mechanisms to enhance RoCE over and again, to make it routable or scalable. There is no need to replace the entire network infrastructure or to retrain IT staff. iWARP is a safe and proven choice for building a scalable high performance clustering fabric.

Summary

iWARP is the no-risk path for 40Gb/s Ethernet clustering, using TCP/IP's mature and proven design, with the required congestion control, scalability and routability, leveraging existing hardware and requiring no new protocols, interoperability, or long maturity period.

RoIP is just another step in RoCE's long journey of experimentation with new protocols, and frustration that would take years to complete, if it ever does. Until then, long live iWARP!

Postscript

The move from raw Ethernet to IP based encapsulation leaves existing RoCE users with a number of questions regarding their investment and the future of the technology. This section compiles a list of such questions and answers.

I invested in a RoCE installation, will it interoperate with RoIP?

NO. Given the fundamental nature of the changes, and the performance limitations of software or firmware implementations, RoIP will likely necessitate new hardware, and adapters in particular, in effect requiring another costly upgrade cycle.

I have an existing DCB infrastructure for RoCE, will it need to change for RoIP?

YES. RoIP is expected to require changes to the Layer-3 (IP) handling to reduce packet loss, most likely requiring expensive router upgrades.

I have an existing DCB infrastructure for RoCE, can I use it for iWARP?

YES. iWARP can operate over any Ethernet infrastructure. Unlike RoCE and RoIP, it does not require DCB capable switches or routers but can run perfectly well over such a network. Alternatively, one can avoid dealing with the complexity of configuring DCB by turning it OFF.

Is RoIP a standard?

NO. Work is in progress on RoIP, and it is expected to take years to get standardized. The fact that the work spans two organizations (IBTA and IETF) is likely to result in delays that are even longer than usual.

Is RoIP going to include a transport layer?

NO. Although it is a necessary requirement, it is unlikely that the RoIP specification will include a new transport layer. All indications are that the immediate goal is to simply address RoCE's routability problem while deferring the rest of the issues to the network, like RoCE did. Even if this approach changes, the resistance to using TCP means a completely new protocol would be pursued, which would take many years to develop, verify and tune. In fact, history has seen all such attempts fail to come to fruition or get any meaningful use.

Why is RoCE getting developed piecemeal?

RoCE is developed by the IBTA, whose main objective is to promote InfiniBand, rather than Ethernet technologies and Internet protocols. RoCE represented the minimum step needed to field a product that would work in an Ethernet environment, leaving out the necessary pieces for routability and scalability. The more pieces the specification adds the more it diverges from InfiniBand and the closer it gets to iWARP, which is the existing standard for RDMA over Ethernet and IP.

References

[1] Chelsio Communications, *A Rocky Road for RoCE*
[2] Chelsio Communications, *RoCE Autopsy of an Experiment*
[3] Chelsio Communications, *RoCE The Fine Print*
[4] Chelsio Communications, *RoCE FAQ*
[5] Chelsio Communications, *RoCE at a Crossroads*

