A Rocky Road for RoCE

The industry standard clustering protocol over Ethernet and IP networks is the Internet Wide Area RDMA Protocol (iWARP), published by the IETF in 2007 as RFC 5040 and RFC 5041. iWARP specifies a Remote Direct Memory Access (RDMA) protocol over TCP/IP, which enables running unmodified InfiniBand (IB) applications using the same OpenFabrics Enterprise Distribution (OFED) software stack, with similar application performance levels.

Recently, a new Ethernet clustering protocol, named RDMA over Converged Ethernet (RoCE), has been proposed by an InfiniBand vendor. RoCE is essentially InfiniBand over Ethernet, where the layers of the IB stack that provide reliable and credit flow-controlled virtual circuit operation are replaced with raw Ethernet encapsulation. For an IB vendor, RoCE may be an easy transition from IB to Ethernet, but it comes with serious caveats to its potential users. To start, it is important to highlight a fundamental assumption in running IB over raw Ethernet without safeguards, namely that "the Ethernet network is lossless." Without lossless operation, performance hotspots quickly lead to performance collapse due to the lack of critical congestion avoidance and management layer, provided by TCP in the Ethernet world. This assumption rides on the tails of a recent push for mixing Fibre Channel traffic with other traditional networking traffic over "Converged Ethernet" (CE), where a similar guarantee is necessary for FCoE. The development of mechanisms for isolating FCoE traffic within an Ethernet network have led to popularizing the myth of "Lossless Ethernet." This paper argues that this basic assumption of RoCE does not hold up to scrutiny. Let us examine it in more detail.

RoCE Requires Converged Ethernet

1. The missing IB layers provide a reliable, flow controlled fabric. On the other hand, Ethernet has no such guarantees, and in fact, this simplicity has been one of its main strengths.
2. Converged Ethernet (also called Data Center Bridging or DCB) refers to a collection of proposed enhancements to traditional Ethernet:
    a. Per-Priority Pause (PPP), which allows specifying one out of eight classes of service (priority) in the PAUSE frames exchanged between two neighboring switches, in contrast to generic PAUSE, which applies to all traffic. PAUSE is the long available hop-by-hop flow control mechanism (standardized in 1997, as IEEE 802.3x) and currently supported by most, if not all switches and adapters.
    b. Enhanced Transmission Selection (ETS), used for provisioning bandwidth when multiple traffic types (e.g. storage and networking) are in use. ETS is not relevant in a network carrying a single traffic type, and in particular, high performance computing environments, where low latency is the critical metric.
    c. Congestion notification, which is a new end-to-end explicit notification scheme. Deploying this protocol requires replacing all network interface cards and switches. More importantly, it is not widely implemented or supported in the industry, mainly because of the doubts on its scalability, effectiveness and congestion messaging impact in large networks. In fact,

although explicit congestion notification has been proposed many times in the past, it never gained acceptance beyond hop-by-hop uses for these reasons.

3. The "Converged" part therefore boils down to PPP, which provides isolation, and therefore is only relevant when the same wire is carrying different traffic types (priorities). When a single traffic type is in use, there is no difference between "CE" and regular Ethernet with PAUSE support. There is no epiphany.

RoCE is Actually IB over Plain Ethernet

1. The same concerns that have prevented the widespread use of Ethernet PAUSE in large networks apply: congestion in one spot propagates quickly and results in gridlock throughout the whole network, which eventually runs at the speed of its slowest link, i.e. generalized poor performance.
2. Therefore, RoCE is really IB over good old Ethernet, and will have to fall back on inefficient error recovery as the network scales and packet loss increases.
3. Error recovery without congestion control exposes the network to congestion collapse, a well-known lesson, learnt as soon as networks grew beyond a few hosts.

In addition to the fundamental issues with running IB over raw Ethernet, RoCE suffers from a number of other serious limitations. First, since it is not routable, it requires flat L2 networks and as a result the issues above are amplified. Secondly, RoCE is not amenable to efficient load balancing and link aggregation within switches, which typically implement support for IP traffic only. Therefore, the flat L2 (Ethernet only) networks needed will have limitations at the core. Finally, the sensitivity to loss limits its use within virtualized data centers where mobility requires IP subnet boundary traversal.

Reality

In conclusion, by throwing overboard critical pieces of the IB and TCP stacks which provide stability and scalability, RoCE shines at simple micro-benchmarks in back-to-back or similarly limited deployment scenarios. However, it stands to fail in large clustered application performance, where all its limitations would be exposed.

Effectively, RoCE represents an attempt by InfiniBand vendors at enticing the customer with a good Ethernet clustering benchmark story, but switching to selling InfiniBand gear in the end.

With a solid and proven networking stack in place, hardened by years of deployment in increasingly larger clusters, iWARP is now a robust and scalable clustering protocol with no application level difference in performance compared to the fastest IB technology, and no compromises to make.

Related Links

IBM Research Report on IB and 10GbE Performance for HPC Applications
IBM/Blade Networks Presentation
Purdue University 10GbE Coates Cluster Whitepaper

References

[RFC 5040] Recio et al., "A Remote Direct Memory Access Protocol Specification", RFC 5040, October 2007.
[RFC 5041] Shah et al., "Direct Data Placement over Reliable Transports", RFC 5041, October 2007.