RoCE: Autopsy of an Experiment

The name RoCE (RDMA over Converged Ethernet) highlights the need for a lossless fabric to run the InfiniBand protocol, and Converged Ethernet is supposed to provide the required lossless operation. However, upon closer examination, the CE component of the name is revealed to be a misnomer at best, since in a dedicated fabric, the CE suite of protocols (also called DCB) effectively boils down to Ethernet's PAUSE. In other words, the CE mechanisms relevant to packet loss in a clustered fabric, where mixing storage and networking using different traffic priorities isn't required, have been available in "normal" Ethernet networks since 1997, when the IEEE 802.3x PAUSE standard was introduced.

Zero Loss Ethernet

PAUSE works on a link level (back-to-back) basis to temporarily stop the sender from transmitting and overrunning the receiver. However, it is well recognized that enabling PAUSE in a large network is fraught with risk of congestion propagation, and there are numerous online postings of IT staff reporting such occurrences, i.e. a hotspot somewhere in the network results in a generalized state of paralysis in the whole network (see [1], [2], [3] and [4]). These lead most switch vendors to recommend restricting PAUSE to first level switches [2], if enabled at all, effectively limiting its utility in multi-tier topologies. There is no reason to believe that such recommendations no longer apply today.

Given the concerns about deploying PAUSE, the Quantized Congestion Notification protocol is the only remaining potentially relevant component of the CE suite. However, assuming that QCN would provide lossless operation is misconstruing the purpose of the protocol [5]. In fact, QCN is an un-deployed, unverified, and unusual way of effecting congestion control, which involves all switches in the network sending explicit congestion notification frames back to the sources, requesting rate control changes. In a network with 10s of thousands of nodes, this traffic is going to be substantial, especially at times where the network is congested, i.e. in trouble. Furthermore, every single one of the sources must be changed to implement TCP like congestion avoidance behavior at the MAC layer. Otherwise, the scheme falls apart. Finally, QCN is limited to a single subnet, i.e. it does not work across IP network boundaries. It is unwise to assume that such explicit congestion signaling has never been thought of before, or that the reason it hasn't been adopted is unrelated to the problems it introduces. In fact, these considerations have resulted in a split between proponents of RoCE with QCN and those against it. Like all new protocol development that spans the infrastructure, QCN is expected to take some number of years before the feasibility, interoperability, scalability issues are worked out.

Scalability Concerns

Concerns about the feasibility of lossless Ethernet operation are further heightened once traffic goes over an IP hop. It may be useful to recall a basic tenet of the Internet Protocol design and implementation: best effort delivery. RFC 791 clearly states in the introduction that "there are no mechanisms to augment end-to-end data reliability, flow control, sequencing, or other services commonly found in host-to-host protocols". Assuming lossless operation across IP boundaries in a multi-tier network is stretching reality to a whole new level. In multi-tenant data centers, encapsulation protocols such as NVGRE and VXLAN all require going over an IP layer as a means to virtualize the network infrastructure. This effectively precludes RoCE and similar protocols requiring absolute zero loss from operating in such an environment.

Inevitably, to get out of the impasse, a topology with a first tier RoCE connectivity and an IB backbone (which has to also carry IP traffic) is forced (or perhaps premeditated by IB vendors), and one is left with a dysfunctional virtualization infrastructure, and a management and troubleshooting nightmare. Add to that the fact that RoCE does not scale easily past one layer 2 subnet (or worse – the only known RoCE installations are limited to a single switch), and that it requires a brand-new, expensive switch infrastructure to further exacerbate the scaling issues across the data center. Finally, it is interesting to note that the absence of IGMP means multicast is implemented via broadcast in RoCE, where every multicast packet results in a packet flood, and it does not take very many nodes attempting multicast in a data center to completely choke the fabric.

iWARP Solution

In contrast, all data centers today use TCP/IP, and iWARP is indistinguishable on the wire from other TCP/IP applications in both format and behavior. Therefore, it does not introduce unknowns or new challenges like RoCE does, and it does not require expensive switches, special treatment or separate debugging tools. Finally, TCP congestion control has been fine tuned over 3 decades to make efficient use of all available capacity in a network, whereas experimental studies at a major OEM showed that a network could not be run at more than 40% of capacity to accommodate RoCE. Moreover, a TCP Offload Engine that is capable of handling packet loss at silicon speeds is even more adept at enabling maximum utilization of a fabric, in addition to increasing host CPU efficiencies. Thus, the fact that iWARP is built on offloaded TCP/IP means that it will bring about both of these efficiencies, as it inherits all TCP/IP's attributes and levels of maturity acquired through decades of use in a wide range of the most demanding environments.

With a proper packet processing engine architecture, it is possible to build a very efficient and high performance TCP offload engine. The proverbial "TCP processing overhead" is composed of 2 clocks at 500MHz speed in today's state of the art pipelined implementation, a cost far lower than other basic components of a modern NIC. An iWARP implementation that is built on an efficient TOE, scales to 40G and 100G with latency comparable to IB, avoiding the need to design and iterate for years on a brand new set of mechanisms to prop RoCE up, or to replace the entire network infrastructure. It is a safe and proven choice for building a scalable high performance clustering fabric, and is available today.

Summary

In summary, iWARP is the no-risk path for Ethernet clustering, using TCP/IP's proven congestion control, scalability and routability, leveraging existing hardware and requiring no new protocols, interoperability, or long maturity period. RoCE by contrast is a journey of experimentation that is projected to take years to complete, if it ever does.

References

[1] *http://www.networkworld.com/netresources/0913flow2.html*
[2] *http://www.dell.com/downloads/global/products/pwcnt/en/Flow-Control-and-Network-performance-with-PowerConnect.pdf*
[3] *http://virtualthreads.blogspot.com/2006/02/beware-ethernet-flow-control.html*
[4] *http://www.gossamer-threads.com/lists/cisco/nsp/121117*
[5] *http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-630674.html*