RoCE Fails to Scale

Repetitive Protocol Surgery or the Dangers of Sliding Down the RoCE Path

Executive Summary

A recent paper published by a public cloud vendor and others reveals the trials and tribulations of their RoCE experience, and the scale of their deployment woes [1].

The authors, including the RoCE vendor and some of the staunchest of RoCE proponents, reveal the severity of the problems faced, with "*poor application performance*", "*head-of-line blocking*", "*unfairness*", "*congestion that spreads*" in the network, and "*performance that degrades*" under load.

The authors then attempt to patch together pieces from Ethernet and DCTCP into a sideband congestion control scheme (DCQCN). However, DCQCN remains an unsuccessful attempt to address the issues, as it continues to require the same blunt pause mechanisms that expose the network to congestion collapse. Thus, the paper serves as a proof that the RoCE idea of airlifting the "*incompatible InfiniBand*" into the Ethernet space is doomed to fail, due to missing critical stability mechanisms.

Heeding the warnings from this paper and other known experiences, users are avoiding scalability limitations and dangerous network meltdowns by staying clear of RoCE. Instead, many are selecting the iWARP RDMA over Ethernet standard. *iWARP is a scalable, easy to use, plug-and-play protocol, which leverages a proven and mature TCP/IP foundation, and originates from the fully open IETF standards process*. There is no reason to slide down the RoCE path, when a stable, robust, cloud ready alternative is available and provides competitive performance and benefits.

Introduction

RoCE started life as the InfiniBand version of Fibre Channel over Ethernet (FCoE). While the latter has not been successful, the same conclusion hasn't yet been accepted for its twin. By examining the results of a recent publication by proponents of RoCE [1], this paper follows up on a number of earlier studies of the InfiniBand over Ethernet specification, and its many incompatible iterations, known as RoCE, Routable RoCE, RoCEv2 and an effective RoCEv3. Unlike the open IETF standard for RDMA over Ethernet (iWARP), RoCE is an ambiguous Annex to the InfiniBand specification that leaves much to interpretation and exposes adopters to wide interoperability gaps. More importantly, it is an incomplete specification with severe shortcomings that is produced in an opaque fashion by the InfiniBand Trade Association, (IBTA, an effective monopoly), invariably in conjunction with new silicon that implements it, from a single source.

It is important to note that the two RDMA over Ethernet alternatives are known to differ in key aspects:

1. **Openness** – although RoCE specifications are emitted by the IBTA, a technical review of these specifications shows they are severely lacking in details, and continue to miss critical functionality. On the non-technical side, the specifications are systematically released after corresponding products start shipping from a particular vendor. Until this approach of the InfiniBand Association ends, RoCE standards will effectively remain closed to real examination and will continue to stifle competition. On the other hand, **iWARP is an IETF standard, and therefore goes through practically the most open standard development process in the industry**. Competitive iWARP adapters are available or announced from multiple vendors.
2. **Scalability** – the recent paper [1] confirms the concerns raised about RoCE since its inception [2,3,4,5,6,7] and summarized in [8]. While there are technically no surprises here, the key take-away is that the RoCE supporters have finally acknowledged the issues. RoCE is overly sensitive to packet drops, and by relying on the blunt Priority Flow Control to avoid them, is shown to fall apart under load. RoCE thus fails to scale beyond a limited, direct attach or at most single hop scenario. In contrast, **iWARP scales transparently in all dimensions: distance, network architecture, link speeds and number of users**.

This paper summarizes the findings of [1] in the next section and discusses its argumentation in the conclusion.

Congestion Control for RoCE?

The abstract of the paper [1] summarizes the dilemma faced by the adopters of RoCE. Having chosen it because it was **deceptively lightweight** and did not require an end-to-end congestion control scheme, they are now forced to **reinvent and retrofit the same mechanisms** they claimed were not necessary in a "lossless Ethernet". This choice is proving **very costly**, as evidenced by vast amount of effort that is going into inventing new protocols and replacing older generations of RoCE adapters to overcome issues. In essence, the entire 30-year maturity cycle of TCP/IP is being repeated, addressing issues that have already been addressed by TCP/IP.

Not only has the RoCE specification had to undergo several painful revisions that **completely broke backward compatibility just to route beyond one hop**, a new version is inevitable as the mechanisms being developed will prove **inadequate and limiting**.

The following quotes are literal excerpts from the paper at hand [1]:

- "*The IB networking stack cannot be easily deployed in modern datacenters*", as "*the IB stack is incompatible with IP and Ethernet technologies*".
- "*RoCEv2 relies on Priority-based Flow Control (PFC) to enable a drop-free network*" but "*PFC is a coarse grain mechanism…*", "*…that can lead to poor application performance due to problems like head-of-line blocking and unfairness*"
- "*Because PAUSE frames can have a cascading effect, a flow can be hurt by congestion that is not even on its path*", otherwise known more ominously as **congestion propagation**. "*Flows in RoCEv2 deployments may see lower throughput and/or variability due to PFC's congestion-spreading characteristics*". "*The reason is the damage*

caused by PFC. As the degree of in cast goes up, more PAUSE messages are generated. As these cascade through the network, they wreak havoc on user traffic".

- *"PAUSE messages that affect downlinks from spine switches can cause extensive damage, as many flows pass through these switches".*

Then the authors go on to design a Frankenprotocol called DCQCN that stitches together pieces from Converged Ethernet's Quantized Congestion Notification QCN [9] (which "does not work over L3 networks", and as most network experts would agree, does not work on any network) and Data Center TCP DCTCP [10]. The new protocol **not only continues to require PFC, but also adds ECN and RED** to implement some congestion control:

- "*The fundamental solution to PFC's limitations is a flow-level congestion control protocol*". This is a key component of TCP, and it is important to note that DCQCN does not actually address this requirement, as the computations it requires are "*expensive*" and support is limited to a few flows per NIC.

A **worrisome** aspect about the new protocol is that it is a contrived **point-in-time design dictated by the existing hardware**, since for the authors "*updating the chip design was not an option*", thus choosing alternatives simply because they are easier to implement.

Incidentally, the authors dismiss iWARP as requiring a Slow Start[1] phase, apparently oblivious to the fact that there are multiple IETF standard RFCs that alter this behavior, and furthermore, that iWARP implementations provide flexibility to tune TCP's behavior much beyond a simple change as this. The authors then proceed to note that their "*ultra-fast start*" (i.e. starting at full blast) results in burstiness that causes packet loss and poor performance, necessitating the continued use of PFC:

- "*DCQCN does not obviate the need for PFC. With DCQCN, flows start at line rate. Without PFC, this can lead to packet loss and poor performance"*.
- With DCQCN alone, "*some flows are simply unable to recover from persistent packet losses. This result underscores the need to use DCQCN with PFC".*

Another very worrisome aspect about the DCQCN scheme, is that it requires careful tuning of no less than **10 switch and NIC parameters**, some of which seem to be assigned arbitrary numbers. RoCE was never a plug-and-play protocol, and its **plug-and-debug** reputation is up for a further boost!

The paper closes with a reminder that Ethernet PAUSE issues are there to stay, as the authors are "*currently studying how to avoid outages that may be caused by a malfunctioning card".* If past experience is anything to go by, they will eventually reach the well-known conclusion: PFC is not a scheme that can be deployed across a datacenter.

[1] A widely known misnomer as it effectively is an exponential ramp-up, which is practically instantaneous in a data center.

Conclusion

In light of mounting evidence that users of RoCE are facing major deployment problems, the truth is finally emerging and exposing the misleading claims by the aggressive FUD campaign accompanying the push for InfiniBand over Ethernet.

This paper discussed the widespread congestion failures observed when RoCE is deployed at scale, as reported by Microsoft Azure, widely advertised as the main datacenter proof point for RoCE. This paper also discussed the DCQCN scheme put together to plug some of the gaping holes in InfiniBand's Ethernet incursion [1].

The perils of using PFC in a large-scale deployment are well known. The need for PFC alone should prevent one's slide down the RoCE path. Unfortunately, the repetition of baseless marketing fluff and technically meaningless statements such as "*due to end-to-end loss recovery, iWarp* [sic] *cannot offer ultra-low latency like RoCEv2*" and the shallowness of their understanding of iWARP, put into question the objectiveness of their analysis and conclusions.

Nevertheless, there is no hiding that the RoCE problems are real, and the successively attempted solutions haphazard and incomplete, with RoCE users repeatedly having to undergo emergency protocol surgery. These hard learnt lessons are helping others avoid the same mistake by selecting iWARP, and across the board, a sea change is in effect as the industry reels back from the edge of a RoCE abyss.

References

[1] Yibo Zhu et al., SIGCOMM 2015, *Congestion Control for Large-Scale RDMA Deployments*
[2] Chelsio Communications, *RoCE The Grand Experiment*
[3] Chelsio Communications, *Rocky Road for RoCE*
[4] Chelsio Communications, *RoCE Plug and Debug*
[5] Chelsio Communications, *RoCE The Fine Print*
[6] Chelsio Communications, *RoCE at a Crossroads*
[7] Chelsio Communications, *RoCE is Dead, Long Live RoIP?*
[8] Chelsio Communications, *RoCE FAQ*
[9] IEEE, *802.1Qau Congestion Notification*
[10] Alizadeh et al., SIGCOMM 2010, *Data Center TCP (DCTCP)*
[11] Chelsio Communications, *The Case Against iWARP*
[12] Intel, *iWARP Ready for Cloud*
[13] Chelsio Communications, *iWARP Goes Mainstream*
[14] IBM, *iWARP for Microsoft SQL Server*
[15] IBM, *iWARP: A Competitive Alternative to Infiniband*
[16] Chelsio Communications, *iWARP for Disaster Recovery*
[17] Chelsio Communications, *iWARP for High Performance CUDA Cluster*
[18] Chelsio Communications, *Lustre Performance with iWARP*
[19] Redmond Magazine, *Scale the datacenter with Windows Server SMB Direct*
[20] Jim Pinkerton, SDC 2015, *Moving to Ethernet connected JBOD (EBOD)*
[21] *iWARP Videos*