RoCE Exposed

The Trials of Deploying *Infiniband over Ethernet*

Executive Summary

Over the last several years, the problems with RoCE are becoming evident to everyone, including the proponents, who are scrambling to recover from their costly choices. In an attempt to stem the tide, Mellanox recently published a paper claiming to counter technical arguments against RoCE. This paper refutes the claims with facts to help shine a light on the truth, and expose additional attempts at deception within that paper itself.

Heeding the accumulating warnings from RoCE adopters, new RDMA deployments are avoiding limitations and dangerous network meltdowns by staying clear of RoCE and 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.

Introduction

This paper responds point-by-point to a paper published by Mellanox to address the mounting dissatisfaction with RoCE [1]. The responses reveal it to be a marketing pamphlet in the guise of a technical paper that confirms rather than refutes the arguments made by Chelsio and others against the RoCE specification, starting from its inception and continuing to date [for instance in 12,13,14,15]. In particular, Chelsio had been arguing the faults with RoCE since v1 and going through v1.5, v2 and the unofficial v3 [see 3,4,5,6,7,8,9].

On Outdated Information

Mellanox claims that Chelsio's published papers comparing Chelsio's 40Gb Ethernet solutions with Mellanox's 40GbE and FDR 56Gb/s InfiniBand products "*use outdated information and present data in unconventional ways*". This section addresses the claims in the paper that are listed in support of this statement, to show how they fail to stand to examination.

RoCE is not a Standard

Mellanox's assertion is that "*Chelsio's Website continues to promote documents with outdated information to bolster its claims. For example, in its RoCE FAQ […] Chelsio claims that RoCE is not the standard RDMA over Ethernet protocol. While iWARP may have been standardized first, RoCE is an open IBTA standard that runs on top of IETF standard UDP using an IANA assigned port number*". It is no surprise that the reasons for this assertion escape the authors of this paper. There are in fact three main reasons behind Chelsio's statement:

1. **Opaque Process.** As things stand today, the definition of standard is no longer applicable to specifications produced by the IBTA, which is effectively a Mellanox monopoly. The

fact that the specifications can be downloaded does not make them open. An open standard is one that is developed through **an open process**, providing **visibility** and opportunity for multiple entities to steer the definition to a **technically sound outcome**, not one that matches whatever the hardware features of the controlling company's product are. Ironically, there is no need to look further than the next paragraph of the same Mellanox paper to see an example of how RoCE fails the basic smell test: "*While the [routable RoCE] standard was only ratified in Q3 of 2014, products supporting routable RoCE have been shipping in high volume for much longer*". Of course, there is no level playing field when the company designing a product produces a corresponding "standard" a few years later. It is well known that this exact issue of uncertainty in the specification and last minute changes caused at least one of the RoCE aspirants to fall to the wayside.

2. **Technical Soundness –** a proper standardization process serves to improve the **quality** of the outcome. There is no longer need to rely on Chelsio's papers for an inventory of RoCE's failings. Suffice it to observe the successive, incompatible versions of the specification to realize that something is seriously amiss: RoCE is at version 3 today with version 4 in the works, and no end in sight. All the while the high profile adopters scrambling to hack together workarounds on their own [2].
3. **Completeness –** A standard also serves as a basis for building **interoperable solutions**. Not only has the RoCE specification been changed under the cover, and modified wildly in between revisions, it is an incomplete specification that leaves important details unspecified and unresolved. While the shortness of the RoCE specification has been marketed as a proof of its simplicity, it hides two large devils: the fact that the **InfiniBand stack is implicitly needed**, and that the details of **key mechanisms** needed for Ethernet are **missing**. The deception campaign started in high gear with the claim that RoCEv1 was routable, followed by others which will be discussed below.

Having established that RoCE fails to meet the basic attributes of an industry standard, Mellanox's response can be seen as a misleading marketing statement that mentions UDP and an IANA assigned number to legitimize RoCE, and create a false impression of an association with the IETF.

RoCE Does Not Scale

Mellanox's paper continues to claim: "*similarly, in the RoCE FAQ, Chelsio posits that RoCE does not scale and has issues of interoperability with switches from other vendors. These claims have been overcome or proven incorrect long ago. Today virtually all advanced data center network equipment supports data center bridging technology that is required to fully take advantage of RDMA*" and that "*There are deployed RoCE-based networks with tens of thousands of nodes*". To refute these claims, one needs to turn to a recent paper [2], co-authored by no other than Mellanox, and which is discussed in [3]. The reader is encouraged to review both papers, to realize that the truth about RoCE's fundamental scalability problems cannot be covered by statements carefully crafted to deceive.

RoCE is not Routable

The final allegation in this section is that "*Chelsio also indicates that RoCE is not routable and is unrecognized by standard traffic management and monitoring tools. Again, both claims are based on outdated information and play on antiquated fears of potential customers. […] RoCE does* *support routable networks. The routable version of the standard was released September 2014 by the IBTA with multiple vendors supporting the announcement".* This statement portrays Mellanox as a victim of FUD, while the truth is the exact reverse. Chelsio released technical assessments of RoCE as early as version 1 which was marketed as a routable and scalable protocol, which it clearly was not. While adding IP and UDP headers allows traversing IP subnet boundaries, there is a lot more involved in producing an actual routable protocol, as clearly shown in [2,3]. Looking at the extensive reference section below, it is also clear that Chelsio continues to publish updated technical studies of RoCE as it goes through the progressive unmasking of the layers of deception.

On Inaccurate Information

The next section in the Mellanox paper [1] claims to identify inaccuracies in some Chelsio papers, in the process inserting more deceiving statements.

The first claim: "*Chelsio makes statements that seem to contradict demonstrated real-world results. For example, in the RoCE FAQ, Chelsio claims that the positive performance numbers seen in RoCE's micro-benchmarks do not match its real application performance. Yet public presentations have demonstrated a 10X performance improvement using RoCE in real-world applications such as virtual machine migration*". A characteristically obfuscated statement from a company known to compare 40Gb RoCE to 10Gb NICs in order to show RDMA providing a large performance boost! Chelsio's papers have consistently referred to comparing Chelsio's 40Gb iWARP and NIC performance to 40Gb RoCE, and this statement carries no relevant technical value in this context.

Another technically void argument is hidden in this next statement "*Chelsio posits that RoCE is limited to operation over short distances (of a few hundred meters). However, this is easily overcome with Layer 3 networking and various switch configurations*". The paper clearly glosses over the fact that inter-switch distances are limited when PAUSE or Priority Flow Control must be used, making RoCE a non-starter for long distance communications. And the suggestion to insert a router every few hundred meters is absurd.

The technical inaccuracies continue with "*Chelsio suggests that RoCE has no congestion management layer, depending entirely on the Priority Flow Control (PFC) Pause feature instead. In fact, the Pause feature is a Layer 2 mechanism that is unrelated to congestion management*". We will start by ignoring the second statement, leaving it as an exercise for the reader. Whether PFC is or is not a congestion control mechanism is not central to any argument made in a Chelsio paper. In fact, what Chelsio papers repeat is that RoCE has no working congestion management layer, simply because it does not have one, as clearly concluded in the paper co-authored by Mellanox [2]. From the start, RoCE depended and continues to depend on the Ethernet PAUSE mechanism to avoid network losses, and Chelsio's papers consistently warned about the dangers of relying on this scheme beyond a small, constrained environment, dangers which the aforementioned paper shows to have resulted in major network problems. Conveniently, the same paper [2] can be consulted to disprove the Mellanox claim that "*the latest update to the RoCE specification (RoCEv2) defines all the necessary mechanisms to address congestion*". Not to mention the enigmatic assertion that "*there are multiple schemes used in practice to manage congestion that are very effective in avoiding packet loss and retransmission*", which serves more to worry than to comfort the reader.

On Displaying Results

The next section of the Mellanox paper looks at Chelsio's published results and focuses on a few graphs where it claims "*unconventional ways*" were used to represent the data. This section is particularly illustrative of the tactics Mellanox uses to deceive unsuspecting audiences.

The first statement accuses Chelsio of unorthodoxy for using the apparently unusual logarithmic scale: "*another way that Chelsio plays with the data in its published papers is to display information on graphs that use a logarithmic scale instead of the more commonly used linear scale*". Then it goes about to show how this supposedly reduces the difference between Chelsio and Mellanox results, in one case simply moving the range where the two solutions look equivalent to the other end of the axis, and in another conveniently chopping the x-axis short to "prove" its point!

In fact, it is well accepted that a logarithmic scale is best used in comparing benchmarks across large scales to highlight the main trends of the data. Going through the performance benchmarks published by Chelsio, one sees that many papers do mention differences that exist at the micro-benchmark level, or at the single port level such as for 56Gb IB vs. 40Gb Ethernet. However, the key message in Chelsio's papers is actually consistent, that is to highlight the conclusion that **application level performance** is often identical for iWARP vs. IB. The papers also argue that practical results and observed trends do not justify Mellanox's repeated claims that iWARP is inherently higher latency and lower performance. In fact, any differences that exist today are easily negated by the fact that a RoCE NIC is single purpose, whereas a Chelsio adapter is a true converged NIC that supports a full suite of protocols at high performance. Finally, note that all of Chelsio's results provide the details of the configuration used to allow reproducibility.

Summary

Selecting an RDMA over Ethernet technology is a task that many organizations are facing today. In approaching it, it is important to make an informed selection that includes an assessment of the pros and cons of the technologies, as well as independently benchmarking the competing offerings **at the application level**, rather than accepting allegations made by an interested party, because, performance aside, the differences between RoCE and iWARP remain significant, and a wrong selection can prove costly, very quickly.

Before RoCE, Mellanox committed to InfiniBand and built products that found their place in High Performance Computing applications, effectively becoming the sole vendor of InfiniBand gear. With RoCE, Mellanox has built a name for itself as a marketer of unbaked, incomplete solutions that are hard to use, require massive investments in infrastructure, and complicated configuration, with repetitive, large scale rip-and-replace cycles to get basic functionality working. Mellanox relied on leverage and deception to push customers to use its InfiniBand products over Ethernet, a technology it clearly does not understand, and is neither incentivized nor desiring to serve well. In the process, Mellanox resorted to a concerted FUD campaign to undermine iWARP, the native RDMA over Ethernet technology.

Chelsio remains committed to a converged Ethernet solution that supports offloaded RDMA, storage and networking over a single wire, with no special switches or configuration needed. With iWARP, Ethernet has the right plug-and-play RDMA over Ethernet solution that is easy to deploy, scalable, robust and ready for the cloud.

References

[1] Mellanox, RoCE Facts You Should Know
[2] Yibo Zhu et al., SIGCOMM 2015, *Congestion Control for Large-Scale RDMA Deployments*
[3] Chelsio Communications, *RoCE Fails to Scale*
[4] Chelsio Communications, *RoCE The Grand Experiment*
[5] Chelsio Communications, *Rocky Road for RoCE*
[6] Chelsio Communications, *RoCE Plug and Debug*
[7] Chelsio Communications, *RoCE The Fine Print*
[8] Chelsio Communications, *RoCE at a Crossroads*
[9] Chelsio Communications, *RoCE is Dead, Long Live RoIP?*
[10] Chelsio Communications, *RoCE FAQ*
[11] Chelsio Communications, *The Case Against iWARP*
[12] Chelsio Communications, *iWARP Goes Mainstream*
[13] Intel, *iWARP Ready for Cloud*
[14] IBM, *iWARP for Microsoft SQL Server*
[15] IBM, *iWARP: A Competitive Alternative to Infiniband*
[16] Redmond Magazine, *Scale the datacenter with Windows Server SMB Direct*

Related Material

Chelsio Communications, *iWARP for Disaster Recovery*
Chelsio Communications, *iWARP for High Performance CUDA Cluster*
Chelsio Communications, *Lustre Performance with iWARP*
Jim Pinkerton, SDC 2015, *Moving to Ethernet connected JBOD (EBOD)*