WHITE PAPER

December 2015

RoCE is the RDMA Winner

Introduction 1

Inaccurate Attack on RoCE 1

Conclusion 3

Introduction

The successful large scale deployment of RoCE in multiple hyperscale data centers has resulted in Mellanox capturing dominant market share of the 40GbE market:

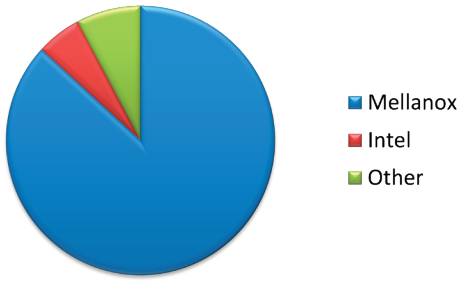

Figure 1. *40GbE Adapter Vendor Market Share (Crehan Report Q2/15)*

New advanced RoCE devices are now shipping that support the latest 25, 50, and 100Gb/s Ethernet speeds and promise further gains in market share.

Inaccurate Attack on RoCE

There is nothing that makes a failed competitor angrier than success, driving some companies into making irrational and unsubstantiated claims. Recently the sole remaining vendor of the failed iWARP technology issued an attack of a Sigcomm paper discussing advanced congestion mechanisms for RoCE.

The attack included many blatant inaccuracies including:

1. *IBTA is an effective monopoly*

This is false. The IBTA is a multi-vendor organization with over 40 members with governance rules similar to other standards bodies. The organization has well defined processes for developing and publishing standards, with IP required to be made available under reasonable and non-discriminatory licensing terms. Multiple vendors cooperated to develop the RoCE standard within the IBTA and then deliver interoperable products to the market – and these products are typically competitive. A multi-

vendor organization where member companies develop interoperable but competitive products can hardly be considered a monopoly.

2. *There is only a single source of RoCE hardware*

This is false. In fact multiple vendors have announced RoCE hardware with several vendors having passed testing as part of the most recent compliance and interoperability workgroup plugfest. In addition to being false, this attack is ironic given that there is in fact only a single vendor offering iWARP technology. Furthermore there are no products available at the more advanced Ethernet signaling rates of 25, 50, and 100Gb/s such as there are for RoCE. It is always easy to claim interoperability when there is a single vendor not able to deliver products at the latest performance levels.

3. *RoCE is ambiguous and just an annex*

This is false. In fact there is no ambiguity in the RoCE specification, allowing for interoperable solutions. The RoCE specification incorporates the entire body of the RDMA mechanisms as defined in the InfiniBand specification. The claim of ambiguity and having the RoCE additions defined separately from the core specification is silly and irrelevant, particularly given that the iWARP specification is defined across dozens of different documents.

4. *There are many incompatible versions of RoCE … Completely broke backward compatibility … RoCEv3*

This is false. By defining the protocols precisely and conforming to strict layering models the RoCE protocol is in fact backwards compatible. In fact there are existing deployments which mix multiple versions of RoCE with backwards compatibility.

The claim of a RoCEv3 standard is also false. There is no RoCEv3 being standardized in the trade association, nor is there any need to define a new version of RoCE. The existing RoCE specification and mechanisms are well defined to support large deployments.

5. *DCQCN is limited to a few flows per NIC*

This is false. RoCEv2 congestion management does not define any limits to the number of flows per NIC in fact hundreds of thousands of flows can be and are being supported. In practice the number flows is not limited other than by system memory. The key to scalability is to minimize the required per flow state and RoCE congestion management implements a lightweight protocol, precisely in order to offer better scalability compared to heavyweight stateful protocols such as iWARP.

6. *DCQCN adds ECN and RED*

This is false. RoCEv2 adds neither ECN (Explicit Congestion Notification) nor RED (Random Early Detection). Both ECN and RED are features defined elsewhere and available on most advanced switches today. RoCEv2 does not require and in fact is defined to avoid RED induced packet drop. RoCEv2 does in fact leverage switches with ECN capabilities, in order to accelerate congestion notification and greatly improve the congestion feedback loop timing – thereby eliminating packet loss. This is a fundamental advantage vs. protocols using implicit congestion notification that rely on intentional packet loss and sender side timeouts as a means of signaling congestion. Such mechanisms are useful in uncontrolled, unreliable, and heterogeneous networks (i.e. the internet) but are simply not useful to deliver high performance, low latency, hardware accelerated RDMA services in a well-controlled modern data center. Intentionally dropping packets in a 100Gb/s Ethernet network is like rear-ending someone on the freeway

to determine there is congestion. ECN is similar to using brake lights to allow traffic to slow down before congestion occurs.

7. *RoCE has problems that prevent it from being deployed at large scale.*

This is false. RoCE is the only Ethernet based RDMA technology being deployed at large scale. These attacks have highlighted issues that RoCEv2 actually *overcomes*. These attacks take these issues out of context and present as if they were problems with RoCEv2 preventing it from being deployed at large scale.

Conclusion

Both Microsoft and Google have implemented innovative and effective algorithms to manage congestion in RDMA based networks that overcome scalability issues. Specifically a Google paper at the same conference assessing RoCEv2 and its congestion management protocol says:

"Evaluations demonstrate that it addresses the HoL blocking and unfairness problems with PFC, thus making RoCE viable for large scale deployment."

There is room for continued innovation in this area with no need for additional mechanisms beyond those available today within state-of-the are Ethernet and RoCE protocol frameworks. The condescending arguments that RoCE does not scale is at odds with reality and belittles the significant contributions of vendors that have addressed the challenges and deployed large scale networks today.

Aggressive and sloppy marketing often occurs when a competitor is frustrated at their lack of success, however gross misrepresentation of facts goes beyond the pale of accepted industry practice.

Any successful technology will evolve, and indeed it is encouraging to see the rapid advancement of RoCE to enable large scale IP based routable deployments. As always new technology builds on previous developments, and success is determined not by hyperbolic rhetoric, but rather by effective deployments that deliver value to customers. RoCE has proven its value in large scale deployments and will continue to gain market share at the expense of traditional NICs as well as the failed iWARP technology.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

15-5658WP
Rev 1.0