The Case Against iWARP

iWARP, the industry standard for RDMA over Ethernet, is now available at 40Gbps speeds from multiple vendors with performance that rivals the fastest high performance RDMA interconnect technologies.

This paper addresses myths and allegations disseminated in the process of marketing the competing InfiniBand over Ethernet (known as RDMA over Converged Ethernet – RoCE) specification. These claims are shown to be technically unsubstantiated talking points meant to spread uncertainty and doubt, while diverting attention away from the true limitations of RoCE.

Introduction

Remote DMA (RDMA) provides the capability for computer systems to *efficiently transfer data* between local and remote host memory, with low latency and high speed and without involving the host CPUs at either end. RDMA has traditionally been the privilege of esoteric fabrics, such as InfiniBand. As power and compute efficiency considerations started to dominate the data center and cloud networking space, interest in RDMA has grown tremendously in the recent years. With Ethernet's ubiquity and with speeds that have risen to 40Gbps and beyond, RDMA over Ethernet is fast becoming a requirement for the new data center.

Today, two competing RDMA over Ethernet technologies are available in the marketplace. The established standard, iWARP, has been in use for more than 8 years, with mature implementations and multiple vendor offerings. The InfiniBand over Ethernet (RoCE) protocol, on the other hand, is a still evolving specification that benefits from InfiniBand's software drivers, but lacks maturity and suffers from basic issues that remain unresolved. Other papers have exposed the limitations and pitfalls associated with the RoCE specification (see [1,2,3,4,5,9]). In summary, the arguments against RoCE[1] are:

RoCE is an incomplete specification
RoCE does not scale
RoCE does not route
RoCE is hard to deploy
RoCE is hard to use and manage
RoCE impacts network-wide QoS
RoCE lacks congestion control
RoCE performance is sensitive to network variability
RoCE is not robust in a real network environment

The main advantage touted by the RoCE proponents is simplicity, due to the absence of TCP in its definition, an "expensive" protocol to implement and operate. This paper debunks this and other myths disseminated to undermine iWARP and divert attention away from RoCE's fundamental limitations.

[1] RoCEv2 replaces the IB Global routing layer with UDP/IP to provide routability, at the cost of requiring lossless IP networks.

The first step in this process is to put to rest the claim of lossless operation over Ethernet that is the underlay of the RoCE edifice. Unlike InfiniBand, where virtual circuit flow control can provide loss-free operation, Ethernet is a traditional packet switched network with no access control, where congestion is a normal phenomenon. While switched Ethernet does provide the capability of suspending transmission on a per-link basis when the receiving end falls behind (PAUSE scheme), this scheme has known limitations (for more details see [2,3]).

Furthermore, "RDMA over Converged Ethernet" implies that RoCE is made possible by Converged Ethernet, which is a "lossless" network. Simply put, CE switches have larger packet buffers to reduce packet loss during transmission bursts, while simultaneously pausing the neighboring nodes. However, they do not eliminate congestion and instead result in increased latency. In addition, using Ethernet PAUSE indiscriminately can lead to congestion propagation from any hotspot to the whole network. For this reason, PAUSE is typically turned off in large networks, where congestion isolation is critical.

Therefore, there is need for a transport layer that handles congestion avoidance and control. In the IP/Ethernet world, this function is performed by the TCP protocol.

The following sections go into the details of the various claims made against iWARP, and point out the fact that each is trying to divert attention from.

First Myth: iWARP is Old

"iWARP was designed in 2002, therefore it is low performance"

This talking point is arguably the weakest of the list, since it ties performance to the age of the specification rather than that of the equipment used in benchmarking!

Today's leading iWARP implementations offer line rate 40Gbps performance with comparable end-to-end latency to the best InfiniBand offerings. InfiniBand itself was designed several years prior to iWARP.

In general, application level performance benchmarks show parity or a slight edge to iWARP in most common applications [8].

However, this point does play a role in diverting attention from the rushed RoCE specification, apparent when looking beyond the marketing hype [9]. Recently, a specification for a new RoCE version (RoCEv2), incompatible with the first one, has been published and adds UDP and IP headers to finally provide routability. This further highlights this fact:

FACT 1: RoCE is a new specification that lacks maturity and is still undergoing fundamental changes

Second Myth: iWARP is Fat

"iWARP is RDMAP over DDP over MPA over TCP/IP"

While widely used and highlighted as a technical statement, this talking point is particularly easy to set straight. The figure below shows the iWARP headers that are present in one packet. It turns out that MPA (yellow) is effectively a packet length field and a CRC with perhaps a few bytes of padding[2], while DDP (red) and RDMA (green) each contribute 8 bits of flags or control information.

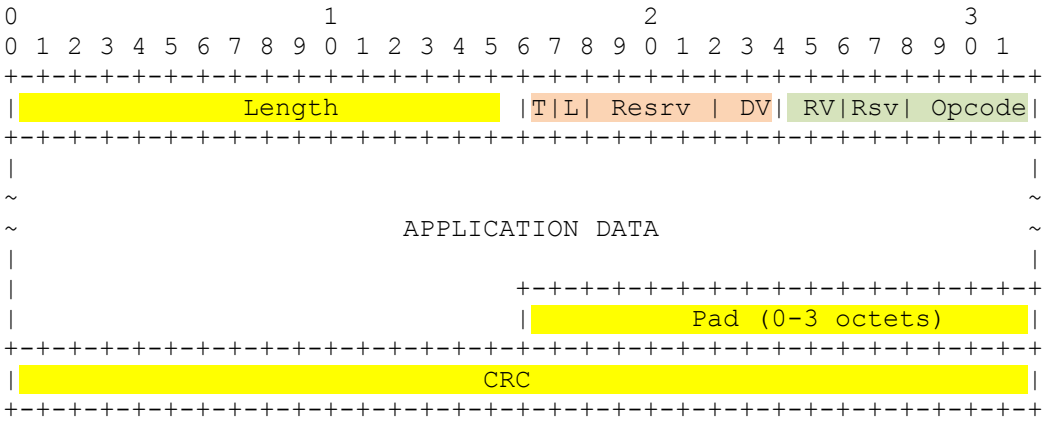

Figure 1 – iWARP Headers in an RDMA Packet

It is evident that these layers serve the purpose of modularizing and clarifying protocol operation, but have little impact on the efficiency of the implementation. Layering is a well known facet of protocol design and iWARP is no different than other protocols in that regard.

In contrast, the RoCE specification is a simple annex to the InfiniBand architecture and leaves out basic aspects of operation over Ethernet, such as how MAC addresses are resolved, or how VLANs are used (see [7] for one list).

Ironically, RoCE is heavier than iWARP in total header overhead[3], as shown in the figure below.

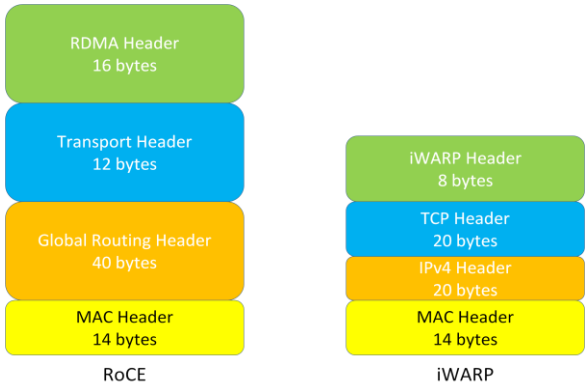

Figure 2 – RoCE vs. iWARP Header Sizes

[2] MPA markers are practically never used.
[3] The RoCEv2 Transport and Routing headers are equal to iWARP's, leaving the RDMA header size difference.

In attempting to portray normal layering in the iWARP protocol stack as a disadvantage, an important fact is getting swept under the rug: if RoCE seems deceivingly simple, it is because:

FACT 2: The RoCE specification is incomplete

Third Myth: TCP is Heavy

"TCP protocol processing is complex"

The main myth put forth to justify RoCE is that TCP is an expensive protocol to execute, and therefore cannot achieve low latency. A related false implication is that since InfiniBand was designed to be hardware implementation friendly, it is the only protocol that can be efficiently implemented and must be compared to software implementations of other protocols.

There have indeed been poor implementations of TCP offload and iWARP that exhibited poor characteristics, but that applies to all poor implementations. There is similarly one very poor RoCE implementation, out of the two existing today.

In contrast, high performance hardware implementations clearly show that iWARP can get latencies comparable to InfiniBand itself. Furthermore, TCP processing is a minor contributor to end-to-end latency, and is dwarfed by MAC, PCI and other contributors *that are present in all adapters*. The following chart shows the relative contribution of software (SW), firmware (FW), physical cable (WIRE), and other hardware components in the 1.5usec application-to-application 1B RDMA latency in Chelsio's T5 based adapters.

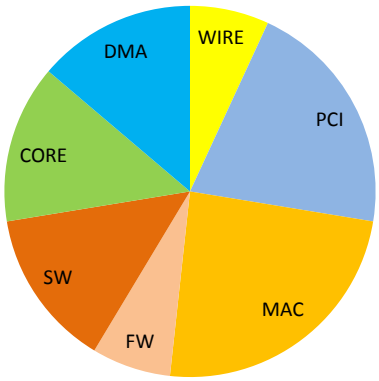

Figure 3 – End-to-End iWARP Delay Components

The chart shows that the full NIC processing (CORE) is about 10% of the latency, and the *TCP processing itself is actually a minimal part* of that slice. Focusing on the cost of implementing congestion control and reliability is an attempt to transform a bug into a feature:

FACT 3: RoCE dropped critical parts of the InfiniBand stack to operate over Ethernet

Fourth Myth: TCP is Slow

"TCP congestion control is slow, timers are large, etc…"

While this statement completely ignores the issues with assuming lossless operation is possible (or limiting its perceived benefits to RoCE), it is largely based on limitations with software TCP implementations.

iWARP adapters implement TCP in hardware, and in good implementations this includes all timers and retransmission mechanisms. A well-tuned hardware TCP stack can operate microsecond granularity timers, and can retransmit *orders of magnitude faster than any software stack* can. Hardware TCP can also forgo mechanisms, such as delayed ACKs, that have been added to mitigate the packet processing impact on host stacks, but may be detrimental to the original congestion avoidance principles behind TCP's operation.

Finally, in networks that support Explicit Congestion Notification (ECN), TCP can truly avoid packet loss in an end-to-end fashion without requiring huge buffers and large delays in all switches. This provides significant savings by allowing the use of existing infrastructure, or deploying simpler, cost optimized equipment.

In summary, *congestion control is simply not optional* and a good hardware TCP implementation can be a high performance and efficient embodiment of a set of trustworthy mechanisms that have been honed over several decades of use throughout the whole Internet. In contrast:

FACT 4: RoCE lacks essential network stability control, and requires a complex DCB infrastructure

Summary

This paper shows that the myths spread to promote RoCE do not stand up to close examination. Hinging on the assumption that iWARP over TCP is expensive and inefficient, these claims are quickly refuted by looking at protocol specifications and actual implementations that are shipping today.

When considering RDMA over Ethernet alternatives, iWARP stands out as the no-risk path for 40Gbps Ethernet clustering, using TCP/IP's mature and proven design, with the required congestion control, scalability and routability. iWARP leverages existing infrastructure and requires no new protocols, interoperability, or long maturity period to replace InfiniBand with the familiar Ethernet technology.

References

[1] Chelsio Communications, *A Rocky Road for RoCE*
[2] Chelsio Communications, *RoCE Autopsy of an Experiment*
[3] Chelsio Communications, *RoCE the Fine Print*
[4] Chelsio Communications, *RoCE FAQ*
[5] Chelsio Communications, *RoCE at a Crossroads*
[6] Chelsio Communications, *RoCE is Dead, Long Live RoIP?*
[7] Wikipedia, *RDMA over Converged Ethernet*
[8] IBM, *A Competitive Alternative to InfiniBand*
[9] Roland's Blog, *Two Notes on IBoE*
[10] Roland's Blog, *RDMA on Converged Ethernet*

