Which RDMA?

Executive Summary

Today, there are various options for enabling an RDMA solution over Fibre Channel, InfiniBand, Omnipath, or Ethernet. Of these, the non-Ethernet solutions are migrating to Ethernet since Ethernet speeds/economies at this point have overtaken all other physical medium. Today, two competing RDMA over Ethernet technologies are available in the marketplace. The established standard, iWARP, has been in use for more than 11 years, with mature implementations and multiple vendor offerings. InfiniBand over Ethernet (RoCEv2) 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. What is clear, is that since the RDMA solution is a wire protocol, both the source and destination must be capable of the exact same protocol. What is not as clear, is which protocol is preferable. While it is true that given a large amount of investment of time and money, that one can make any protocol work in any application, it is the case that some protocols are better suited than others, or perhaps better suited in different applications. We examine this question along below metrics.

Performance

Given the implementation of both iWARP and RoCEv2 is in silicon, the difference in performance is negligible. Both protocols have a verb interface/PCIe on one side of the hardware controller and send and receive ethernet packets on the other side. In a typical NVMe-oF comparison of iWARP vs. RoCEv2, only about a ~3% variation is observed in remote vs. local SSD access latency between the two protocols.

The following chart shows the relative contribution of software (SW), firmware (FW), physical cable (WIRE), and other hardware components in the 1.5µsec application-to-application 1B RDMA latency in Chelsio's based adapters.

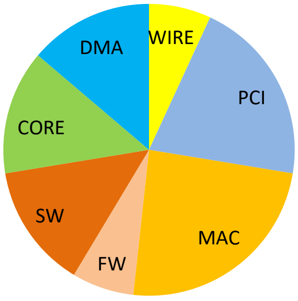

Figure 1 – End-to-End iWARP Delay Components

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.

Ease of Use

RoCEv2 implementations require DCB/ETS/PFC capable switches to operate. A RoCEv2 implementation typically cannot go past one switch hop without experiencing a severe performance degradation and requires specialized IT skills to enable. An iWARP implementation by comparison works with any legacy switch or link (including metro or wireless links), and as such is plug-and-play. Other papers have exposed the limitations and pitfalls associated with the RoCEv2 specification (see [1,2,3,4,5]). In summary, the arguments against RoCEv2 are:

1. RoCEv2 does not scale
2. RoCEv2 does not route
3. RoCEv2 is hard to deploy
4. RoCEv2 is hard to use and manage
5. RoCEv2 impacts network-wide QoS
6. RoCEv2 lacks congestion control
7. RoCEv2 performance is sensitive to network variability
8. RoCEv2 is not robust in a real network environment

Ease of Adoption

A RoCEv2 deployment requires a concurrent upgrade of the switch infrastructure cycle for the reasons mentioned above. In addition, given the 2 or 3 different variants of RoCE, one has to be sure that the peer also is RoCEv2 capable. As such, it impedes the sales cycle of typical server or storage OEMs. RoCEv2 essentially requires a greenfield installation. iWARP by comparison is based on TCP/IP and as such decouples the refresh cycle of the infrastructure from others and as such presents a brownfield installation to the server, storage, and OS vendors. Therefore, RoCEv2 has found traction in storage back-end applications where typically one can control the whole environment consisting of the JBOF, the NAS head, and the switch, and the required gateways, in a single-protocol environment. Given iWARP can operate in front-end, server-facing environments, or in long distance replication environments as well, it offers a solution that can be utilized in all cases easily.

In addition, iWARP has introduced a very high-performance software-only solution to enable any L2 NIC run iWARP protocol. This solution basically enables every single existing server to connect with iWARP – the adoption will be instant. Given the state of CPU technology today, a soft implementation provides adequate performance for most applications, while on the storage target side, a hardware offload solution will still be required. By the way of example, historically, iSCSI protocol adopted rapidly only after a software initiator solution was introduced in the market. The same is now expected with the introduction of the soft iWARP solution in 2Q19. This surprisingly high-performance software solution enables a simple migration path for RDMA applications from appliance to the cloud on any NIC. It also enables typical storage target solutions to use a cloud as a server initiator.

Ease of Support

Given the interoperability issues of RoCEv2, all sorts of field issues, will alias to a support call to the factory. This essentially forms a cost of sales, and it causes the link issues to reflect on the storage, server, or OS vendor's products. This is one of the main reasons why Microsoft for example has recommended use of iWARP over RoCEv2 deployments for their Storage Spaces Direct deployments (see [11, 12, 13]). Yes, RoCEv2 can work, but iWARP is easier, and it costs less, and one is less likely to get a support call.

Flexibility

An iWARP solution can work in front-end applications, back-end applications, disaster-recovery or replication applications. iWARP offers a robust software-only solution as well. It is also offered on the motherboard chipset of some high-end Intel servers. It operates with any switch, or router. It does not require gateways when leaving the data centers, and it easily operates across multiple hops. By comparison, RoCEv2 can only operate in constrained environment for short distances.

Cost

iWARP is on average a lower cost solution because the hardware CNA solutions are cost competitive, and because software iWARP implementation is free. iWARP does not need special gateways. iWARP has lower support costs and no need to upgrade to higher cost switches.

Summary

When considering RDMA over Ethernet alternatives, iWARP stands out as the no-risk path for a variety of applications, using TCP/IP's mature and proven design, with the required congestion control, scalability and routability inherent in TCP/IP. iWARP leverages existing infrastructure and requires no new protocols, interoperability, or long maturity period to replace InfiniBand or Fibre Channel with the familiar Ethernet or TCP/IP technology. At the same performance as RoCEv2, but with more ease of use, faster sales cycles, less support issues, and less cost, iWARP is the preferred RDMA protocol, despite the heavy marketing effort behind RoCEv2.

References

[1] Chelsio Communications, *A Rocky Road for RoCE*
[2] Chelsio Communications, *RoCE Autopsy of an Experiment*
[3] Chelsio Communications, *RoCE the Missing 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] Chelsio Communications, *iWARP Myths*
[10] Chelsio Communications, *100G SPDK NVMe over Fabrics*
[11] Microsoft, *Microsoft prefers iWARP*
[12] Microsoft, *Microsoft prefers iWARP*
[13] Microsoft, *Microsoft prefers iWARP*
[14] Chelsio Communications, *RoCE - Plug and Debug*
[15] Chelsio Communications, *RoCE fails to scale*
[16] Chelsio Communications, *RoCE Exposed*
[17] Chelsio Communications, *Resilient RoCEv4: The Experiment Continues*
[18] Chelsio Communications, *Resilient RoCE: Misconceptions vs. Reality*