Universal RDMA: A Unique Approach for Heterogeneous Data-Center Acceleration

By Bob Wheeler
Principal Analyst

June 2017

The Linley Group

www.linleygroup.com

Universal RDMA: A Unique Approach for Heterogeneous Data-Center Acceleration

By Bob Wheeler, Principal Analyst, The Linley Group

As storage performance increases, new techniques and data rates will ensure that Ethernet does not become a bottleneck. Now available in 10G, 25G, 50G, and 100G Ethernet server adapters, RDMA reduces network processing overhead and minimizes latency. Cavium's FastLinQ stands out by handling both RoCE and iWARP, which are alternative protocols for RDMA over Ethernet.

The Growing Importance of RDMA

Remote direct-memory access (RDMA) is a technology that has been in use for more than a decade. In server connectivity, data copying is a major source of processing overhead. In a conventional networking stack, received packets are stored in the operating system's memory and later copied to application memory. This copying consumes CPU cycles and also introduces latency. Network adapters that implement RDMA enable writing data directly into application memory, as Figure 1 shows. Applications that transfer large blocks of data, such as networked storage and virtual-machine migration, reap the greatest efficiency gains from RDMA.

Because the network adapter needs to know where to place data in memory, the network protocol must explicitly support RDMA. The first such protocol to achieve widespread adoption was InfiniBand, which was designed around RDMA from the beginning. The techniques developed for InfiniBand were later adapted to Ethernet networks using new RDMA protocols—first iWARP and later RoCE—both of which we describe in more detail below.

High-performance computing (HPC) was first to adopt RDMA due to the performance benefits of reducing latency in applications that implement parallel processing. Next, enterprise-storage vendors began to use RDMA as they developed new clustered systems. Rather than implementing RDMA on server-facing ports, these systems used RDMA in the "back end" network that aggregated multiple storage nodes into one logical rack-level system. Most early RDMA implementations, however, used InfiniBand rather than Ethernet.

A lack of server-software compatibility was a major inhibitor to broader adoption of RDMA protocols, which initially required the use of specialized APIs (such as MPI in HPC). This situation has changed over the last few years as operating systems have begun to support RDMA-based storage protocols. For the first time, end users could implement RDMA-based protocols without any changes to applications. One major milestone was Microsoft's addition of the SMB Direct protocol to Windows Server 2012, which enables the SMB network file sharing protocol, and VM live migration, to use RDMA. Building on SMB Direct, Microsoft added to Windows Server 2016 a new feature called Storage Spaces Direct (S2D), which creates scalable software-defined storage systems using commodity servers and direct attached storage.

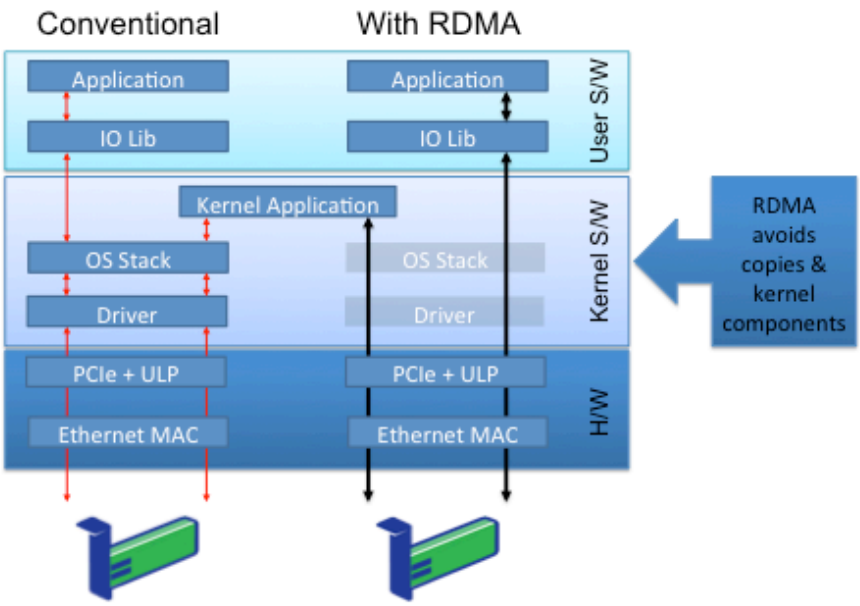

Figure 1. Conventional versus RDMA network stack.

For Linux servers, NFS over RDMA has been available in the Linux kernel since 2008. Storage targets that handle block-level protocols can implement iSER (iSCSI over RDMA) using stacks from several Linux target projects.

Several factors contribute to the growing importance of RDMA and its increased adoption outside of HPC. Server virtualization has increased average CPU utilization to the point where the CPU cycles to handle networking are no longer free. Server connectivity has moved from Gigabit Ethernet to 10G Ethernet and now 25G Ethernet and faster rates, multiplying the network processing overhead. The rapid adoption of solid-state storage (SSDs) within data centers means that storage media is often no longer the performance bottleneck. The performance of SSDs increases further with the move from traditional SAS interfaces to NVM Express (NVMe), which is built on PCI Express and removes the overhead of the SCSI layer from the storage stack.

The rapid adoption of solid-state storage within data centers means that storage media is often no longer the performance bottleneck.

In 2016, the NVM Express organization extended the NVMe protocol across the network with the release of its NVMe Over Fabrics (NVMe-oF) specification. There exist multiple NVMe-oF protocols that work with Ethernet, InfiniBand, and Fibre Channel networks. The NVMe-oF protocol for Ethernet requires an RDMA transport in the form of iWARP or RoCE. Demonstrating the architectural elegance of the new standard, the majority of driver code in the Linux kernel is common between native NVMe and NVMe-oF. The common NVMe driver should ensure that NVMe-oF becomes a part of commercial Linux distributions as those consume new kernels. Although NVMe-oF products are only now reaching the market, the combination of NVMe and RDMA promises very high performance networked storage.

In parallel with broadening software support and development of new standards, a greater number of Ethernet adapter (NIC) vendors now support RDMA. In 2016, Broadcom, Cavium, and Mellanox offered RoCE, whereas Cavium and Chelsio handled iWARP. We expect Intel will introduce iWARP support on its code-name Purley server platform, which is ramping but not announced as of early 2017. Note that only Cavium's FastLinQ supports both the RoCE and iWARP protocols, and it supports both concurrently.

Only Cavium's FastLinQ supports both the RoCE and iWARP protocols, and it supports both concurrently.

Whereas RDMA was once available in only costly specialized NICs, vendors now support it in high-volume Ethernet NICs without requiring additional licenses or software packages. Overall, RDMA adoption is poised to expand from specialized applications running on InfiniBand to broad use cases operating over general-purpose Ethernet.

RoCE Grows from InfiniBand Roots

The InfiniBand Trade Association (IBTA) developed RoCE (pronounced "rocky") by adapting InfiniBand technology to Ethernet. Because the InfiniBand transport protocol expects guaranteed delivery, RoCE requires the underlying Ethernet network to be essentially lossless. Thus the name RDMA over Converged Ethernet, or RoCE for short. Creating a lossless Ethernet network requires the use of data center bridging (DCB) protocols, especially priority-based flow control (PFC). All switches in the RoCE data path must support PFC and have the protocol enabled and configured.

The IBTA released the first version of RoCE, sometimes referred to as RoCEv1, in 2010. As Figure 2 shows, RoCEv1 uses the InfiniBand network and transport layers over an Ethernet (Layer 2) network. The problem with this approach is that it restricts RoCEv1 traffic to a single subnet, which limits scalability. In 2014, the IBTA addressed this problem by delivering RoCEv2, which replaces the InfiniBand network layer with IP and UDP. The use of an IP header enables Layer 3 routing, whereas the addition of UDP allows for ECMP load balancing. These changes make RoCEv2 suitable for large-scale data centers, which implement Layer 3 fabrics and typically use ECMP to load balance traffic across multiple paths.

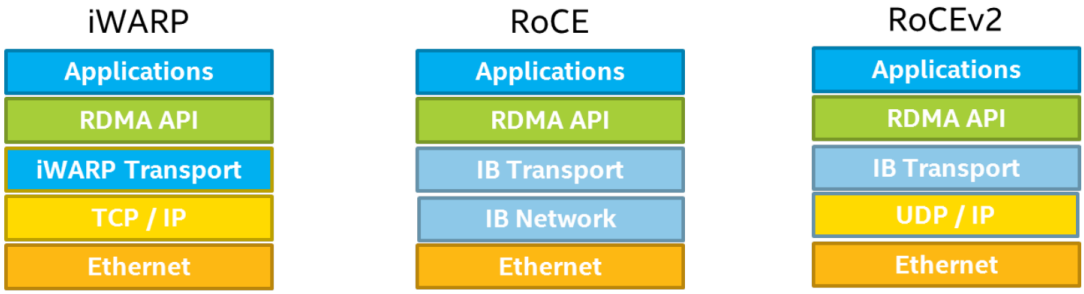

Figure 2. Comparison of iWARP and RoCE protocol stacks.

RoCEv2 Congestion Management (RCM) is an optional feature that addresses concerns regarding the original RoCE protocol's reliance on PFC. In small deployments, some

network admins objected to the operational complexity of configuring VLANs across all switches in the network as required for PFC. Large data-center operators discovered that, even with PFC, their networks experienced some packet loss and congestion spreading. Because the InfiniBand transport protocol was designed for zero loss, it behaves poorly when packet loss occurs.

To address these concerns about PFC, RCM uses Explicit Congestion Notification (ECN). Whereas PFC operates on a point-to-point link, ECN enables an end-to-end congestion notification mechanism. When an ECN-capable switch experiences congestion on a given port, it notifies the receiving NIC on that port of congestion. The receiver then returns a congestion notification packet to the original sender, that is, the sending NIC. That NIC, in turn, reduces its transmission rate to mitigate the congestion.

Although RCM specifies the wire protocol for congestion management, it does not specify the algorithm used by the sender to adjust its transmission rate. One such algorithm is DCQCN, which was described in a paper presented at ACM Sigcomm 2015. Microsoft has deployed RoCEv2 with DCQCN at a very large scale in its Azure data centers. Although DCQCN reduces the number of PFC frames required in its network, the company still uses PFC as a "last defense" against packet loss. Microsoft's deployment has been successful, but it has required a great deal of engineering work, which is out of reach for the vast majority of data-center operators. On the plus side, much of the company's work is implemented by vendors and is ultimately made available to the rest of the industry.

We see RoCE as best suited to small-scale environments where all servers and storage nodes reside in the same rack or row.

The ongoing evolution of RoCE protocols and implementations illustrates the challenges of deploying the technology at scale. As a result, we see RoCE as best suited to small-scale environments where, for example, all servers and storage nodes reside in the same rack or row. Another fit is the back-end network for a cluster of storage nodes, where RoCE provides an Ethernet-based alternative to InfiniBand.

Thanks to its use of a nearly lossless Ethernet network, RoCE delivers deterministic latency. This bounded latency delivers maximum performance for storage applications, making RoCE well suited to networked flash arrays. In the near future, systems with the most extreme storage-performance requirements could be built using NVMe SSDs networked using NVMe-oF operating over RoCE, so long as they don't scale beyond a row.

iWARP Builds on TCP

The early development of iWARP dates back to 2002, when industry leaders including HP (now HPE), IBM, Intel, and Microsoft formed the RDMA Consortium to develop specifications for RDMA over TCP/IP. The specifications were standardized in 2007 through a series of IETF RFCs, with the protocol suite becoming known as iWARP. This suite includes the RDMA protocol, the Direct Data Placement (DDP) protocol, and Marker PDU Aligned (MPA) framing, which together form the iWARP transport layer

shown in Figure 2. The RDMA Consortium also developed iSCSI Extensions for RDMA, or iSER, which was released in 2003.

Although Microsoft provided early support for iWARP through its Winsock Direct interface and later the Network Direct API, most implementations focused on HPC applications. For this reason, most early iWARP application development focused on Linux environments. This situation changed in 2012, however, when Microsoft introduced SMB Direct. By that time, RoCEv1 was also available, setting up a battle between competing RDMA protocols.

Compared with RoCE, iWARP's major advantage is that it operates over standard Ethernet and IP fabrics without the need for DCB or special congestion-management techniques. The TCP layer provides a reliable connection as well as congestion control proven over decades in the field. Due to its ability to operate over any Ethernet infrastructure, iWARP reduces operating expenses by eliminating special configuration requirements for switches. The protocol enjoys the broadest software support, including client support in Windows 10. The iWARP protocol can even operate over metro-area or wide-area networks.

iWARP's major advantage is that it operates over standard Ethernet and IP fabrics without the need for DCB or special congestion-management techniques.

Although TCP recovers gracefully from packet loss, dropped packets cause a large increase in latency. Hardware (NIC) implementations use fast-retransmit algorithms to minimize this latency, but it is still measured in milliseconds. Thus, iWARP operating over standard Ethernet cannot match the bounded latency of RoCE operating over a network with PFC enabled.

One challenge in iWARP NIC designs is the implementation complexity of TCP. Known as TCP offload engines, or TOE, these designs must implement the TCP stack in hardware/firmware. High cost and high power characterized early TOE NIC designs, but Moore's Law has reduced this penalty as process technology advanced. Stack maturity is also a concern, as the TOE design must be field-proven. Fortunately, the Cavium and Chelsio designs represent evolutions of architectures with a decade or more of customer deployments.

The maturity of iWARP as well as its ability to operate over standard Ethernet infrastructure makes it attractive for a broad range of use cases. It enables RDMA traffic to span large-scale networks, including router traversal, without any special configuration. It handles a wide variety of storage protocols including SMB Direct, iSER, and NFS over RDMA, as well as the emerging NVMe-oF. Whereas RoCE provides a high-performance back-end network for storage systems, iWARP is better suited to front-end storage networks. The main caveat in deploying iWARP is that customers should select a vendor that offers a proven underlying TOE design.

In summary, the majority of operating systems and applications support both iWARP and RoCE. The former protocol leads in ease of deployment, whereas the latter delivers superior performance when properly deployed. Customers are free to choose the protocol that best fits their current and projected needs.

FastLinQ Universal RDMA

In February 2016, Cavium (then QLogic) announced its FastLinQ 45000 family of 25GbE, 40GbE, and 100GbE adapters. These products built on its existing family of FastLinQ 10GbE NICs, which hold second place in worldwide market share. In 2017, the company added its FastLinQ 41000 family, which represents its second-generation 25GbE design and also supports 10GbE.

Although we focus here on RDMA, the FastLinQ 41000 and 45000 families handle a wide variety of protocols and offloads. They support server virtualization, network virtualization, and a range of storage-protocol offloads including iSCSI and Fibre Channel over Ethernet (FCoE). For network virtualization, the NICs offload tunneling protocols including VXLAN, NVGRE, and Geneve. Support for PCI Express SR-IOV allows up to 240 virtual machines to directly access the NIC, bypassing the hypervisor for maximum performance. As Figure 3 shows, the FastLinQ 41000 uses a single-chip controller design to minimize cost and power dissipation.

The FastLinQ 41000 and 45000 families offer Microsoft-certified compatibility with SMB Direct.

Cavium provides NIC drivers for a range of operating environments including Linux, Windows Server, VMware ESXi, XenServer, and FreeBSD. The same drivers and software work across the FastLinQ 41000 and 45000 families. Once a customer qualifies the software, they can make seamless speed upgrades or switch between RDMA protocols.

The FastLinQ 41000 and 45000 families offer Microsoft-certified compatibility with SMB Direct under Windows Server 2012 and 2016. Cavium's software for Windows supports RoCEv1, RoCEv2, and iWARP. In Linux environments, the company similarly supports all three RDMA-protocol variants. FastLinQ handles NFS over RDMA across various Linux distributions, and it also supports both user-mode and kernel-mode iSER target projects. The NICs also support NVMe-oF host and target drivers under Linux kernel 4.8, which was the first release to handle the new standard. Cavium supports RoCE and RoCEv2 under ESXi 6.5, but VMware's RDMA features are limited in this release.

In RDMA applications, the FastLinQ 45000 NICs stand out by handling RoCEv1, RoCEv2, and iWARP. Furthermore, the design can handle all three protocols simultaneously, which opens up new RDMA use cases. In a storage server, for example, concurrent protocol support allows initiators (clients) to use either RoCE or iWARP. Also, a dual-port NIC can simultaneously handle front- and back-end networks even when they use disparate RDMA protocols.

Windows Server 2016, which supports hyperconverged clusters with up to 16 nodes, provides a similar use case. In this release, the SMB Direct service can share the same NIC with VMs running under Hyper-V. In the cluster configuration, the back-end network can use RoCE for S2D and live-migration traffic between the cluster nodes, whereas the front-end network handles VM iWARP traffic. When an RDMA application is moved from a physical server to a VM, that VM can continue to use whatever RDMA protocol was previously employed.

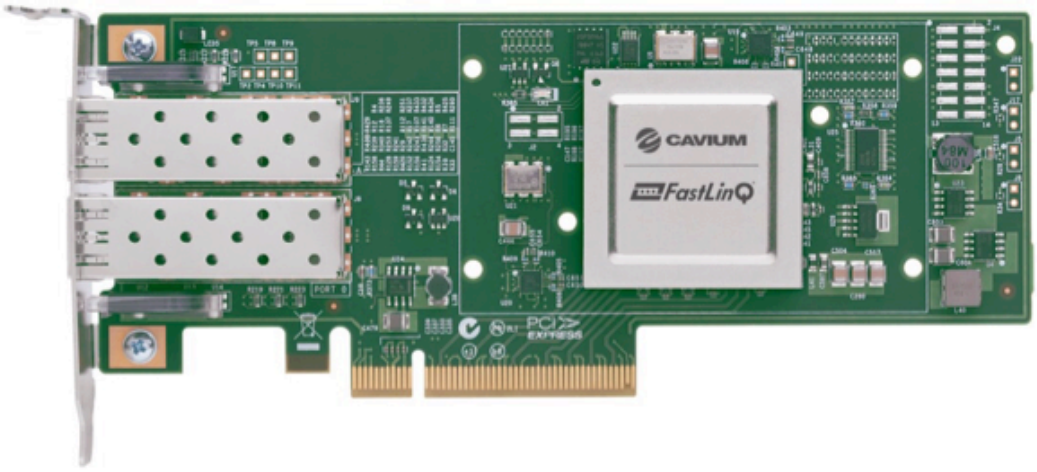

Figure 3. FastLinQ QL41000 dual-port 25GbE NIC with Universal RDMA. (Photo source: Cavium)

For enterprises that distribute workloads across multiple data centers, concurrent RoCE and iWARP support enables efficient VM migration. The user may employ RoCE for maximum performance within a data center, whereas iWARP can move VMs between data centers. No matter the use case, the flexibility of universal RDMA enables seamless interoperability with the RDMA hardware ecosystem.

Future Proofing RDMA Acceleration

RDMA is an example of an excellent technology that has struggled to achieve broad adoption. In Ethernet networks, RDMA requires a new wire protocol, and both ends of the connection must support that protocol. RDMA also requires support in the Ethernet controller chips used in NICs. Finally, it requires an ecosystem of operating systems that hide these differences from applications. Unfortunately, the competing RoCE and iWARP protocols have also fragmented the market.

In 2017, Ethernet-based RDMA may have finally reached critical mass. Multiple vendors now support each standard with cost-effective mainstream NICs. Both Windows Server and Linux environments offer RDMA-enabled network file systems. At the same time, the need for RDMA has grown due to the increasing performance of flash-based storage systems and the adoption of 25G, 50G, and 100G Ethernet. Most NIC vendors, however, remain entrenched in supporting only one RDMA protocol.

With its FastLinQ 41000/45000 designs, Cavium has taken the unique approach of handling both RoCE and iWARP. By delivering a protocol-agnostic design, the company is offering customers a future-proof path to implementing RDMA. If, over the longer term, the RoCE or iWARP camp gains greater market adoption, the customer is protected from NIC obsolescence. In addition, Cavium's support of 100G Ethernet gives customers a guaranteed roadmap to higher performance as it is required.

By selecting FastLinQ NICs, customers also avoid vendor lock-in, as these NICs interoperate with those from all other vendors offering RoCE or iWARP. If a customer ultimately decides to change from iWARP to RoCE or vice versa, those systems already deployed using FastLinQ need not be requalified. For storage systems, FastLinQ NICs can be mixed and matched between front-end and back-end networks, maximizing reuse and minimizing spares. For example, a customer can use a 25GbE NIC initially for a back-end connection and later reuse that NIC for a front-end link when the back end is upgraded to 50GbE or 100GbE.

In summary, products like Cavium's FastLinQ 41000/45000 families are critical enablers of broader RDMA adoption. The software ecosystem is in place for network file systems to adopt RoCE and iWARP now. Although immature, NVMe-oF promises the ultimate in block-level storage performance, extracting the full potential of SSDs. The time is right for data-center operators to evaluate RDMA for their Ethernet and IP networks.

Bob Wheeler is a principal analyst at The Linley Group and networking editor of Microprocessor Report. *The Linley Group is the leading vendor of technology analysis on networking, communications, mobile, and computing semiconductors, providing a unique combination of technical expertise and market knowledge. Our in-depth reports cover topics including Ethernet chips, base-station processors, server processors, embedded processors, and CPU IP. For more information, see our web site at www.linleygroup.com.*

The Linley Group prepared this paper, which Cavium sponsored, but the opinions and analysis are those of the author.