Technology Brief

iWARP* RDMA
Here and Now

Driving factors for low-latency Ethernet

Ethernet is the de facto standard for data center server and storage connectivity and is by far the most agile and general-purpose network.

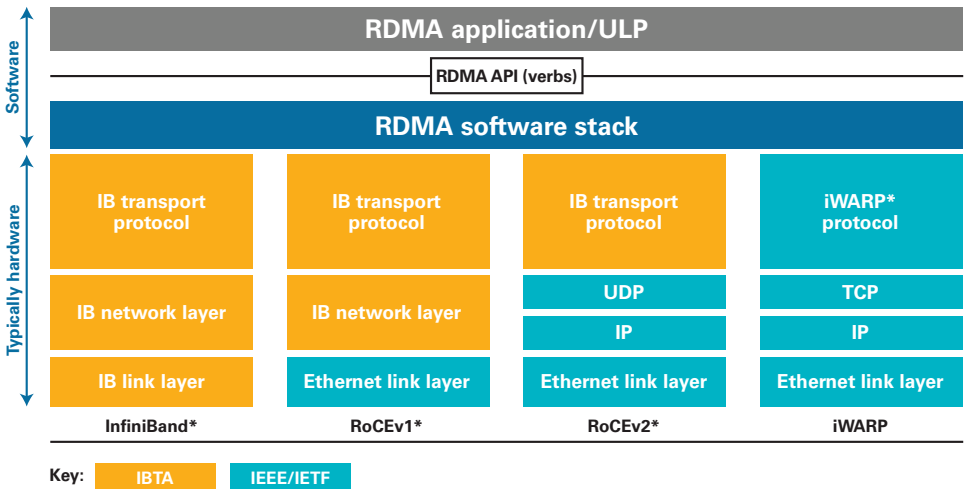
There are more workloads, like storage, that used to run over dedicated fabric that are now moving to Ethernet. These workloads require high data throughput and a low-latency network. The advent of NVMe* and NVMe over Fabrics* further drives the demand for higher-speed Ethernet.

Higher-speed Ethernet, 25/40/50/100GbE, significantly increases the network overhead, mainly the TCP/IP stack process, memory copies, and application context switching.

Remote direct memory access (RDMA) is one of the technologies that relieves Ethernet overhead for high-speed applications.

RDMA and RDMA options

RDMA is a host-offload, host-bypass technology that enables a low-latency, high-throughput direct memory-to-memory data communication between applications over a network.



Today there are three options for RDMA:

InfiniBand*: Requires deploying a separate infrastructure in addition to the requisite Ethernet network.

RoCE* (RDMA over Converged Ethernet): Developed in 2009 by the InfiniBand Trade Association (IBTA), RoCEv1 substituted the InfiniBand physical layer and data link layer with Ethernet, and RoCEv2 further changed to operate on top of UDP/IP. RoCEv2, though similar to RoCEv1, requires lossless Ethernet, and is routable over IP networks within data center boundaries.

iWARP*, IETF standard protocols based: Delivers RDMA on top of the pervasive TCP/IP protocol. iWARP RDMA runs over standard network and transport layers and works with all Ethernet network infrastructure. TCP provides flow control and congestion management and does not require a lossless Ethernet network. iWARP is a highly routable and scalable RDMA implementation.

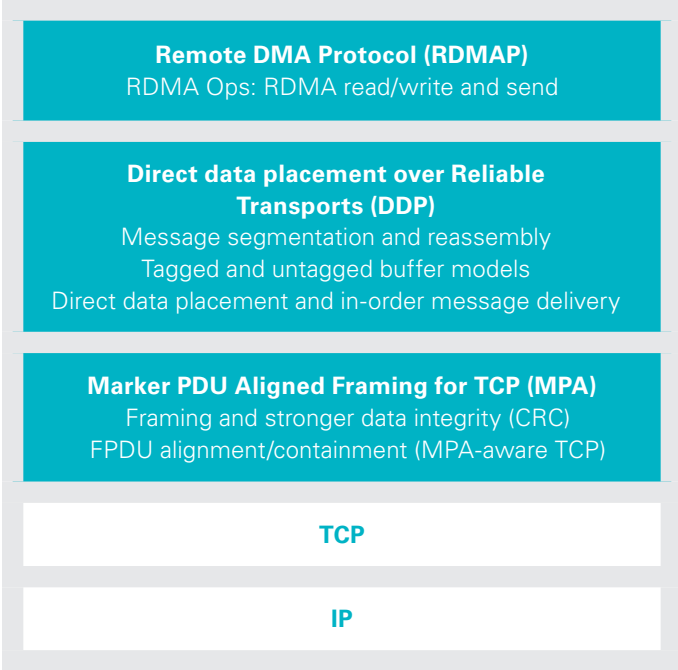
iWARP: Under the hood

iWARP extensions to TCP/IP were standardized by the Internet Engineering Task Force (IETF) in 2007. These extensions eliminated three major sources of networking overhead: TCP/IP stack process, memory copies, and application context switches.

Extension	Solution	Benefit
Offload TCP/IP	Offloads the TCP/IP process from the CPU to the RDMA-enabled NIC (RNIC)	Eliminates CPU overhead for network stack processing
Zero Copy	iWARP enables the application to place the data directly into the destination application's memory buffer, without unnecessary buffer copies	Significantly relieves CPU load and frees memory bandwidth
Less Application Context Switching	iWARP can bypass the OS and work in user space to post the command directly to the RNIC without the need for expensive system calls into the OS	Can dramatically reduce application context switching and latency

iWARP protocols

iWARP is comprised of a series of specifications: Remote DMA Protocol (RDMAP), Direct Data Placement (DDP) over Reliable Transports, and Marker PDU Aligned Framing (MPA) for TCP.



iWARP ecosystem

Operating systems vendors (OSVs) support

All types of RDMA protocols require integration and support in the operating system. iWARP has been integrated into key server operating systems and supports various storage-focused applications and HPC-style applications. Specifically:

- **Microsoft Windows*:** iWARP is fully supported in Microsoft Windows Server* 2012, Windows Server 2012 R2, Windows Server 2016, and Windows® 10 Enterprise and is a transport option for key applications like SMB Direct, Storage Spaces Direct, and virtual machine live migration.
- **Linux*:** Modern Linux kernels included in most major distributions provide end-to-end support for iWARP and enable applications like iSER, NFS over RDMA, or NVMe over Fabrics.

Intel, Cavium, and Chelsio are three key Ethernet solution providers that continue to work closely with other OSVs on building, extending, and enhancing support for iWARP technologies.

Key applications

Application	Benefits
SMB Direct	RDMA-enabled SMB, low latency, high data throughput, ready to be used by upper-level applications
Windows Server 2016 Storage Replica	Combines high-performance, block-level replication across metro areas with the high efficiency provided by the zero-copy and CPU bypass operation of the RDMA transport
NVMe over Fabrics	Defines a low-latency mechanism for NVMe drivers over RDMA networking in large-scale storage deployments
Accelerated VM live migration	Fast VM live migration without compression and not affected by the workload running in the VM
iSCSI over RDMA	Provides translation layer for operating the iSCSI protocol over RDMA/iWARP transports
NFS over RDMA	Allows the operation of the Network File System (NFS) protocol using RDMA-capable networking
High Performance Computing	Low-latency message passing over an Ethernet network

Hardware support

Intel has more than 35 years of experience providing Ethernet solutions. Intel has launched Intel® Ethernet Connection X722 featuring iWARP and will launch multiple Ethernet products featuring iWARP in the future.

Learn more at intel.com/ethernet

Chelsio 10/25/40/50/100GbE adapters for networking and storage within virtualized enterprise data centers, public and private hyperscale clouds, and cluster computing environments offer a 4th-generation implementation of iWARP.

Visit the company at chelsio.com

Cavium FastLinQ* 10/25/40/50/100GbE network adapters are purpose-built to accelerate and simplify data center networking by delivering enterprise-class reliability and acceleration for cloud, telco, and storage workloads leveraging RDMA.

Learn more at cavium.com/fastlinq

Summary

iWARP is a standards-based RDMA implementation running on top of TCP. iWARP addresses current network bottlenecks for high-speed Ethernet and provides a high-throughput, low-latency, and low-CPU utilization data communication. Because TCP provides flow control and congestion management, iWARP can run over current Ethernet infrastructure without lossless network support and is a highly scalable RDMA solution.

0717/ESD/CMD/PDF 336225-001US