NETEFFECT

Understanding iWARP:

Eliminating Overhead and Latency in multi-Gb Ethernet Networks

For years, Ethernet has been the de facto standard LAN for connecting users to each other and to network resources. Ethernet sales volumes made it unquestionably the most cost effective data center fabric to deploy and maintain. The latest generation of Ethernet offers a 10-Gbps data rate (10GbE), simplifying growth for existing data networking applications and removing the wire-speed barriers to deployment in storage and clustering environments. Achieving 10-Gb wire speed, however, requires that Ethernet's long-standing overhead problems be resolved; problems that, in slower Ethernet generations, were adequately overcome by steadily increasing CPU clock speeds.

Founded to solve Ethernet's long-standing overhead problems, the RDMA Consortium released the iWARP extensions to TCP/IP in October 2002. Together these extensions eliminate the three major sources of networking overhead — transport (TCP/IP) processing, intermediate buffer copies, and application context switches — that collectively account for nearly 100% of CPU overhead related to networking. The iWARP extensions utilize advanced techniques to reduce CPU overhead, memory bandwidth utilization, and latency by a combination of offloading TCP/IP processing from the CPU, eliminating unnecessary buffering, and dramatically reducing expensive OS calls and context switches — moving data management and network protocol processing to an *accelerated Ethernet adapter*.

Source	% CPU overhead related to networking[1]	iWARP technique
Transport (TCP/IP) processing	40	Transport offload
Intermediate buffer copies	20	RDMA
Application context switching	40	OS bypass
	~100%	

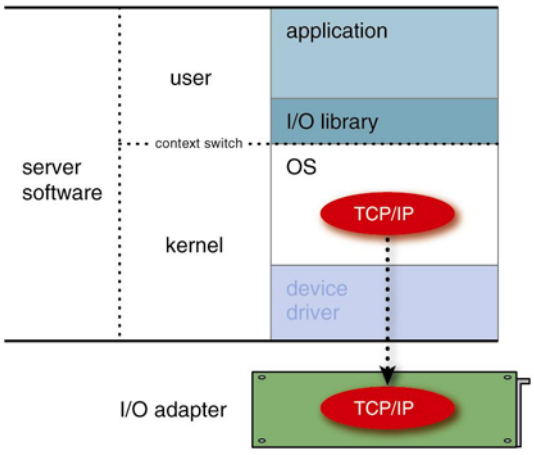
A rough estimate of the CPU overhead related to networking for a given Ethernet link without iWARP capabilities is: for every Mbps of network data processed, one MHz of CPU processing is required. For instance, a fully utilized 20-GHz CPU would be needed to drive a bi-directional 10GE link. A full implementation of the iWARP extensions eliminates virtually all of the CPU's networking overhead, returning these CPU cycles to the application.

iWARP BENEFITS

Offloading TCP/IP (Transport) Processing

In Ethernet, processing the TCP/IP stack is traditionally accomplished by software, putting a tremendous load on the host server's CPU. Transport processing includes tasks such as updating TCP context (sequence numbers, etc.), implementing required TCP timers, segmenting and reassembling the payload, buffer management, resource intensive buffer copies, interrupt processing, etc.

CPU load increases linearly as a function of packets processed. With the 10x increase in performance from 1GbE to 10GbE, packet processing also increases 10x, driving CPU overhead, related to transport processing, to increase by up to 10x as well. As a result, the CPU becomes burdened and eventually crippled by network processing well before reaching Ethernet's maximum throughput.

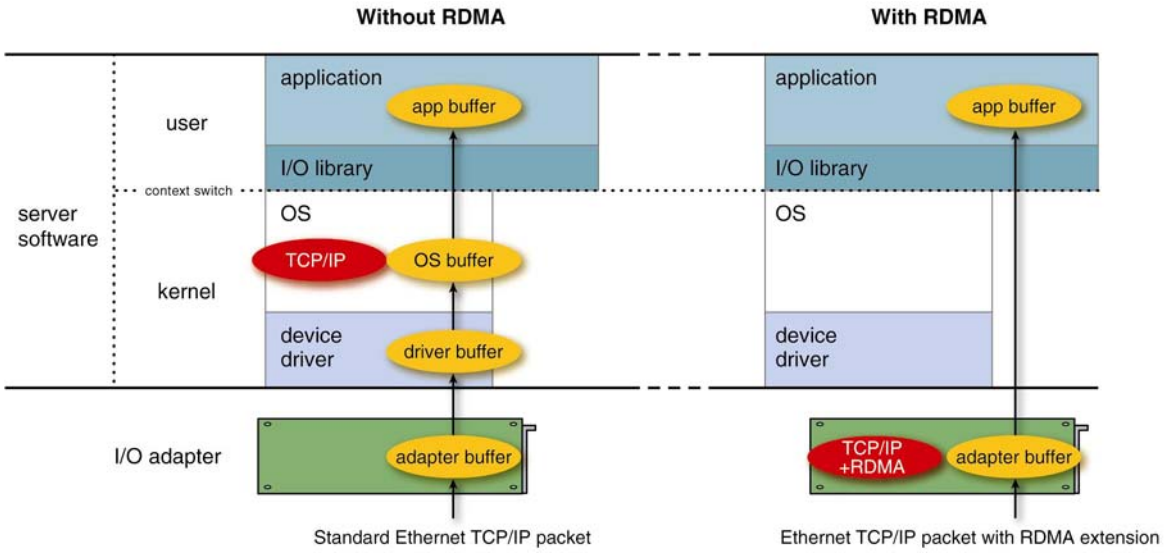


The iWARP extensions enable Ethernet to offload transport processing from the CPU to specialized hardware, eliminating 40% of CPU overhead attributed to networking. Transport offload can be implemented by a standalone transport offload engine (TOE) or embedded in an accelerated Ethernet adapter that supports other iWARP accelerations. Moving transport processing to an adapter also eliminates a second source of overhead –intermediate TCP/IP protocol stack buffer copies. Offloading these buffer copies from system memory to the adapter memory conserves system memory bandwidth and lowers latency.

Eliminating Intermediate Buffer Copies / Zero Copy

In a traditional software TCP/IP implementation, a received packet is moved to server memory via DMA (direct memory access) from the Ethernet NIC's receive buffer, then copied several times by the host CPU before arriving at its ultimate destination in an application buffer. Each buffer copy performed by the CPU requires a read followed by a write (e.g., read the OS buffer, write the application buffer), requiring 2x wire rate in memory bandwidth. In the simplified illustration below, without the ability to move the packet directly into application memory via RDMA (remote direct memory access), the bandwidth load on server memory for an incoming 10-Gbps stream would be 5x line rate: 50 Gbps (over 6 GBps!)

Some relief from buffer copy overhead is achieved through a fully optimized software networking stack. For example, in some OSs the need for a driver buffer-to-OS buffer copy has been dramatically reduced. However, two sources of intermediate buffer copies still exist in all mainstream OS TCP/IP implementations. When packets arrive off the wire, the OS temporarily stores them; performs TCP processing to check for errors and determine the destination application, then transfers the packet(s) into the specified application buffer. The only practical way to do this is with a copy operation, thus the copy from the OS buffer into the application buffer *cannot* be eliminated.



RDMA enables an accelerated Ethernet adapter to support direct memory reads from/writes to application memory eliminating buffer copies to intermediate layers.

The second source of intermediate buffer copies occurs within the application itself. This copy is often required when the application pre-posts buffers and when the application's transaction protocol integrates control information and data payload within the received byte stream, a characteristic of most popular application transaction protocols. The application examines the byte stream and separates control information from data payload, requiring a scatter copy operation from one application buffer into others. An application can eliminate this buffer copy by NOT pre-posting application buffers, but this can dramatically reduce performance.

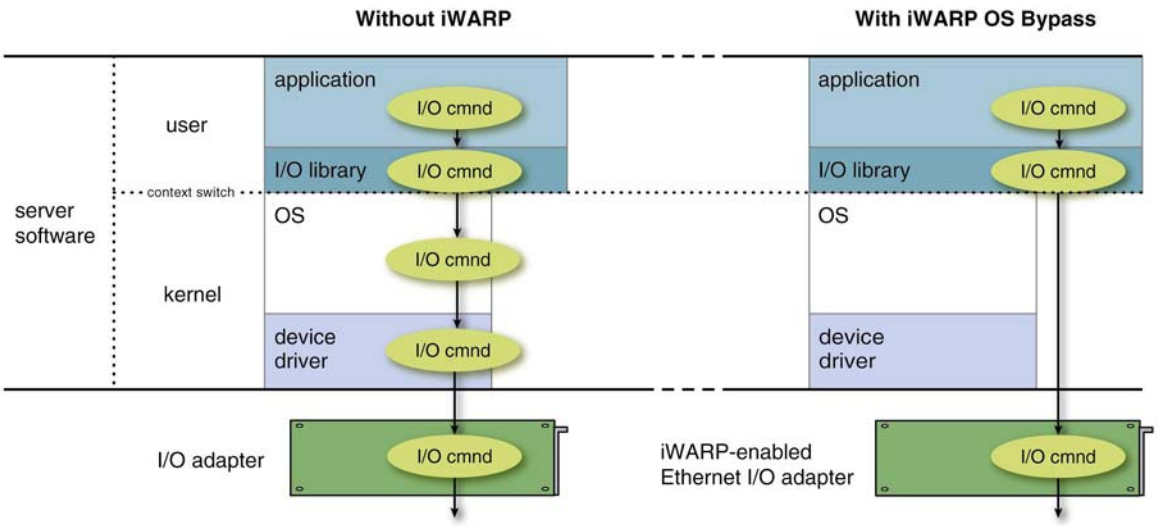
DMA techniques have been used for years and were most recently applied to network processing by InfiniBand. Repurposed for Internet protocols by the RDMA Consortium, Remote DMA (RDMA) and Direct Data Placement (DDP) techniques were formalized as part of the iWARP extensions. RDMA embeds information into each packet that describes the application memory buffer with which the packet is associated. This enables payload to be placed directly in the destination application's buffer, even when packets arrive out of order. Data can now move from one server to another without the unnecessary buffer copies traditionally required to 'gather' a complete buffer. This is sometimes referred to as the 'zero copy' model. Together

RDMA and DDP eliminate 20% of CPU overhead related to networking and free the memory bandwidth attributed to intermediate application buffer copies.

Avoiding Application Context Switching / OS Bypass

The third and somewhat less familiar source of overhead, context switching, contributes significantly to overhead and latency in applications. Traditionally, when an application issues commands to the I/O adapter (e.g. reads from or writes to an I/O device), the commands are transmitted through most layers of the application/OS stack. Passing a command from the application to the OS requires a compute-intensive context switch. When executing a context switch, the CPU must save the application context (the information needed to return to the application when finished with the command) in system memory including all of the CPU general-purpose registers, floating point registers, stack pointers, the instruction pointer, all of the memory management unit state associated with the application's memory access rights, etc. Then the OS context is restored by loading a similar set of items for the OS from system memory.

The iWARP extensions implement OS bypass (user-level direct access) enabling an application executing in user space to post commands directly to the network adapter. This eliminates expensive calls into the OS, dramatically reducing application context switches. Tasks typically performed by the OS – protection of the I/O device, memory protection checks, and virtual-to-physical memory translation, etc. – are handled by an accelerated Ethernet adapter. Such adapters are necessarily more complex than traditional unaccelerated NICs but can eliminate the final 40% of CPU overhead related to networking.



OS bypass eliminates expensive calls to the OS kernel typically required for context switching.

The iWARP Protocols

The iWARP protocol layers are formally described in a series of four specifications that standardize how transport offload, RDMA (zero copy), and OS bypass (user-level direct access) features are implemented. Three of these specifications defining packet formats and transport protocol for RDMA over Ethernet connections were publicly released by the RDMA Consortium in October 2002. They include Marker PDU Aligned Framing for TCP (MPA), Direct Data Placement over Reliable Transports (DDP), and Remote DMA Protocol (RDMAP).

Released in April 2003, the fourth specification (RDMA Protocol Verbs) defines the behavior of RDMA-enabled Ethernet network adapter hardware, firmware, and software as viewed by the host. An extension of InfiniBand Verbs, the RDMA Protocol Verbs specification provides OS bypass, enhancements for privileged consumers, better control over registered buffers, and better multi-processor event distribution.

NETEFFECT

RELATED PROTOCOLS

RDMA Consortium
In addition to the iWARP extensions, the RDMA Consortium has released two specifications that enable existing data center applications to take advantage of RDMA transparently:

iSER (iSCSI Extensions for RDMA) is an extension of the iSCSI storage networking standard. It maps the iSCSI application protocol over the iWARP protocol suite to take advantage of RDMA technology while preserving compatibility with the iSCSI infrastructure. Version 1.0 of the iSER specification was publicly released by the RDMA Consortium in July 2003.

SDP (Sockets Direct Protocol) is a transaction protocol enabling emulation of sockets semantics over RDMA. This allows applications to gain the performance benefits of RDMA without changing application code that relies on sockets today. Version 1.0 of the SDP specification was publicly released by the RDMA Consortium in October 2003.

Internet Engineering Task Force (IETF)
The Internet Engineering Task Force (IETF) is an open international community of network designers, operators, vendors, and researchers that develops and promotes Internet standards particularly those dealing, in particular, with standards of the TCP/IP and Internet protocol suite. It has authored several Internet standards related to iWARP. Versions of the MPA, DDP, RDMAP, and iSER specifications are all approved for publication as Internet standards, and there is work underway on a draft specification for an RDMA-enabled version of NFS which will reduce overhead associated with network file systems.

The publication of an IETF Internet standard indicates broad industry interest in the published technology. This enables development of interoperable products by multiple vendors, and ensures the technology can be easily deployed by end users. Examples of successful Internet standards are TCP/IP, iSCSI, and NFS.

Microsoft
Microsoft operating systems (OSs) enable users to eliminate all types of networking overhead:

> Windows Server 2003 and Windows Compute Cluster Server 2003 implement Winsock Direct, a precursor to SDP, for RDMA-enablement of legacy sockets applications. Winsock Direct uses all three techniques described above — transport offload, direct data placement, and OS bypass — to eliminate networking overhead. Winsock Direct is a mature protocol, implemented in Microsoft OSs since early 2001.

> Windows Server 2003 with Scalable Networking Pack supports a major enhancement to the Microsoft networking stack known as TCP Chimney. TCP Chimney uses transport offload to eliminate networking overhead. It can be used in situations where both servers in a peer-to-peer communication are not RDMA-enabled. TCP Chimney is a relatively new feature, implemented in Microsoft OSs since May 2006.

OpenFabrics
The OpenFabrics Alliance (www.openfabrics.org) was founded in June 2004 to develop open-source, cross-platform, transport-independent software for RDMA. Its membership of more than 30 companies includes silicon, server, networking, storage, and software vendors. A partial list of end-user benefits provided by the OpenFabrics software stack for Linux includes:

> RDMA-acceleration of high-performance computing (HPC) applications written for the Message Passing Interface (MPI) API. Various implementations of MPI middleware are written for OpenFabrics, including Open MPI, MVAPICH, MVAPICH2, HP MPI, and Intel MPI.

> RDMA-acceleration of popular network storage protocols, including block storage (iSER) and file storage (NFS-RDMA)

> RDMA-acceleration of Linux sockets applications, including legacy sockets applications via SDP, and clustered database applications via Reliable Datagram Sockets (RDS) protocol.

> RDMA-acceleration of user-level applications via the new OpenFabrics verbs API.

The OpenFabrics Alliance provides a tested version of the OpenFabrics software stack, known as the OpenFabrics Enterprise Distribution (OFED). OFED is integrated into Linux distributions from Novell and Red Hat, and portions of OFED have been included in the Linux kernel since mid-2005.

NETEFFECT

BENEFITS THROUGHOUT THE DATA CENTER

Deploying Ethernet adapters that fully implement the iWARP extensions to TCP/IP brings dramatic improvements in networking performance throughout the data center.

> In networking applications, 1Gb accelerated Ethernet adapters are compatible with existing Gigabit Ethernet infrastructures and can be deployed immediately to resolve bottlenecks and improve latency in applications where performance is needed most — without changes to existing cabling or switch infrastructure.

> NAS and iSER-based storage networks utilizing accelerated Ethernet adapters can deliver the highest-performance, lowest-overhead storage solutions at any given wire rate. At 10Gb, Ethernet-based storage networks offer a viable alternative to a Fibre Channel SAN. For storage vendors, iWARP will become a standard feature of NAS and iSCSI SAN networks.

> Server vendors are looking to iWARP to standardize the performance and features of clustered systems. Accelerated Ethernet adapters can be used in clustering applications without sacrificing latency or bandwidth performance.

Together, the iWARP extensions to TCP/IP eliminate virtually 100% of the CPU overhead due to networking in 1Gb and 10Gb Ethernet deployments, increasing bandwidth, lowering latency, and maximizing CPU availability for more productive work. This gives data centers the opportunity to reap the increased productivity and lowered total cost of ownership benefits of a ubiquitous, standards-based technology.