TECHNOLOGY BRIEF
NetEffect™ Ethernet Server Cluster Adapters
Internet Wide Area RDMA Protocol (iWARP)

Understanding iWARP:

Delivering Low Latency to Ethernet

For years, Ethernet has been the de facto standard local area network (LAN) technology for connecting users to each other and to network resources. Low cost and increasing Ethernet data rates have simplified growth for existing data networking applications and removed the wire-speed barriers to deployment in storage and clustering environments. However, for certain types of applications such as large-scale financial services, cloud computing, and high-performance computing (HPC), Ethernet's inherent latency and limited message-processing rates have presented unacceptable performance barriers.

As the data center evolves toward a more virtualized model, with a higher degree of abstraction applied to the network and servers, Internet Wide Area RDMA Protocol (iWARP) has emerged as an important enabler to help businesses get the full network throughput benefits of the latest 10 Gigabit Ethernet (10GbE) technologies.

What is iWARP?

iWARP delivers converged, low-latency fabric services to data centers through Remote Direct Memory Access (RDMA) over Ethernet.

The key iWARP components that deliver low-latency are as follows:

- **Kernel Bypass.** Removes the need for context switching from kernel-space to user-space
- **Direct Data Placement.** Eliminates intermediate buffer copies by reading and writing directly to application memory
- **Transport Acceleration.** Performs transport processing on the network controller instead of the processor

The iWARP specification, maintained by the Internet Engineering Task Force (IETF), supports transmissions over TCP and is implemented on top of IP networks using an existing Ethernet infrastructure.

Capitalizing on the Benefits of iWARP

NetEffect™ Ethernet Server Cluster Adapters from Intel use iWARP technology to decrease Ethernet's latency for storage and clustering networks. By addressing the key sources of Ethernet overhead, iWARP provides these benefits:

- **Fabric consolidation.** With iWARP technology, a true unified network of LAN, SAN, and RDMA traffic can pass over a single wire. Moreover, application and management traffic can be converged, reducing cables, ports, and switches.
- **IP-based management.** Network administrators can use standard IP tools to manage traffic in an iWARP network, taking advantage of existing skill sets and processes to reduce the overall cost and complexity of operations.
- **Native routing capabilities.** Because iWARP uses Ethernet and the standard IP stack, it can use standard equipment and be routed across IP subnets using existing network infrastructure.
- **Existing switches, appliances, and cabling.** The flexibility of using standard TCP/IP Ethernet to carry iWARP traffic means that no changes are required to Ethernet-based network equipment.

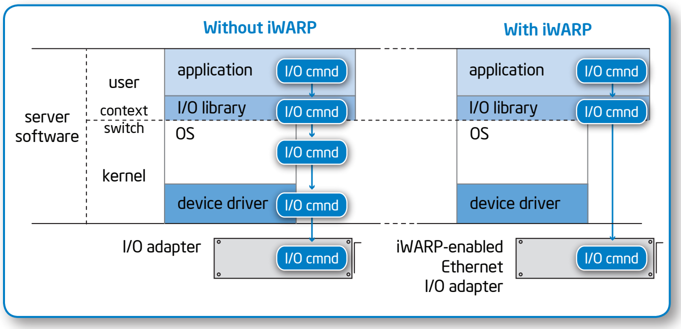
iWARP Brings Benefits to the Data Center

NetEffect Ethernet Server Cluster Adapters from Intel use iWARP to virtually eliminate processor overhead associated with Ethernet networking. TCP/IP continues to be the core protocol stack as data centers transition to cloud computing, and iWARP allows LAN, SAN, RDMA, and standard IP management traffic to pass over a single wire in TCP/IP networks. All of this traffic is natively routable and uses standard Ethernet switches, for a simplified, cost-effective holistic solution.

Because of these benefits, server vendors are increasingly standardizing on iWARP for clustered systems. NetEffect Ethernet Server Cluster Adapters from Intel fully implement iWARP extensions to TCP/IP, helping to bring dramatic improvements in networking performance at low cost using existing equipment and processes.

How iWARP Reduces Ethernet Overhead and Latency

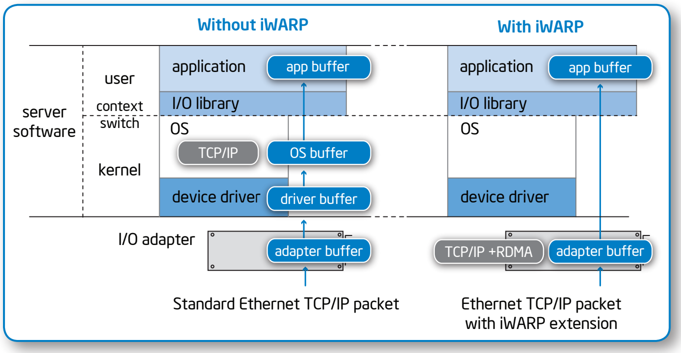
Kernel Bypass: Removes the need for context switching from kernel-space to user-space



Traditionally, when an application issues commands, such as reads and writes, to a server adapter, those commands are transmitted through user-space layers of the application to kernel-space layers in the OS stack. This requires a compute-intensive context switch between user space and the OS.

The iWARP extensions use RDMA to enable the application to post commands directly to the server adapter. This capability eliminates expensive calls into the OS, and that lower overhead reduces latency.

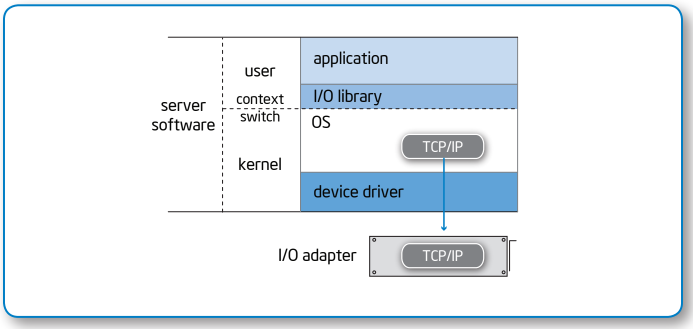
Direct Data Placement: Eliminates intermediate buffer copies by reading and writing directly to application memory



Under conventional Ethernet, data is copied (and re-cached each time) by the processor several times as it passes from the server adapter's receive buffer to the application buffer. Those operations consume time and memory bandwidth that the application could otherwise use.

Using RDMA, iWARP enables direct copies from the server adapter's receive buffer to the application buffer. This provides a direct data placement implementation that eliminates the intermediate operations, which significantly reduces latency.

Transport Acceleration: Performs transport processing on the network controller instead of the processor



With traditional Ethernet, the processor dedicates substantial resources to maintaining the network stack. It must maintain connection context, segment and reassemble payloads, and process interrupts. This overhead increases linearly with wire speed, limiting scalability.

The iWARP extensions enable the transport processing to be done in the network controller. This enables the processor to perform more application processing, providing a deterministic, low-latency solution that is optimized for applications that demand low latency.

Operating System Support

Support for iWARP is required in operating systems to bring the benefits of RDMA to HPC applications.

- **Windows* HPC Server 2008:** The Network Direct Interface for low-latency networking is validated for iWARP on NetEffect Ethernet Server Cluster Adapters from Intel, with support for the Microsoft Message Passing Interface (MS-MPI) for cluster computing.
- **Linux*:** The OpenFabrics Enterprise Distribution (OFED*) from Open Fabrics Alliance provides open source RDMA software for HPC applications. Linux distributors including Red Hat incorporate this software in their releases. OFED enables MPI APIs such as Open MPI, MVAPICH, MVAPICH2, Platform Computing MPI, and Intel® MPI, as well as popular network storage protocols, including block storage (iSER) and file storage (NFS-RDMA).

The robust deployment capabilities provided by Windows and Linux are complemented by third-party offerings, such as Clustercorp Rocks+*, which help to streamline installation and management of the software stack.

To learn more about iWARP, please visit:

www.intel.com/technology/comms/iWARP

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information. The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web Site http://www.intel.com/.

*Other names and brands may be claimed as the property of others.

0710/TS/MESH/PDF 324032-001US