iWARP Update: RDMA Over 40Gb Targets Data Center and Cloud Applications

iWARP, the IETF standard for remote direct memory access (RDMA) over Ethernet, made news in the industry several years ago when its performance and low latency had many industry pundits wondering if it would replace InfiniBand* (IB) as the leading network protocol for high-performance computing (HPC) and financial services industry (FSI) computing applications.

Since then, the demand for low latency network throughput for Big Data applications has spread to clouds and data centers. iWARP performance has improved, and ease of deployment—a significant concern for enterprise applications—has been dramatically simplified.

The result is that iWARP is emerging as a mainstream network technology that can have a dramatic impact on the performance and efficiency of many network applications.

iWARP is Ready for Broad Deployment Thanks to Performance Increases and Plug-and-Play Software Integration

David Fair
Intel Corporation

Asgeir Eiriksson
Chelsio

Introduction to iWARP

iWARP is an implementation of RDMA technology. RDMA is composed of three key technologies designed to lower latency and improve efficiency by offloading networking tasks from a server processor:

- Direct data placement, which reads and writes data directly to application memory
- Kernel bypass, which alleviates the need for context switching from kernel space to user space
- Transport acceleration, which leverages protocol engines on a network controller for fast transport processing

RDMA is deployed on specialized network controllers that have the capability to offload all of the processing of the network stack from a server's CPU, including connection context, segmenting and reassembling packets, and interrupt handling. With RDMA, the server processor no longer needs to copy data at each step as the payload makes its way from the receive buffer to the application buffer, eliminating a process that consumes server memory.

iWARP, ratified by the IETF in 2007, is RDMA implemented on top of the TCP/IP network transport. Thus, it is able to dramatically improve upon the most common and widespread Ethernet communications in use today and deliver on the promise of converged Ethernet to transport LAN, SAN, and RDMA traffic over a single wire with the unrestricted routability that is a key feature of IP.

RDMA can be deployed over other transport technologies, most prominently IB in HPC applications. Until recently, fourteen data rate (FDR) IB had a bandwidth advantage over

10Gbps Ethernet because it operates at 56 Gbps. However, with wide availability of 40Gbps iWARP Ethernet controllers, adapter-level performance is nearing parity (more on that in the test section below).

Another contender is RDMA over Converged Ethernet (RoCE), which deploys RDMA using IB messaging and Ethernet transport. This approach suffers from complexity in terms of required network protocols (lossless fabrics using data center bridging) and a lack of routability. Current RoCE implementations have been limited to a single one Layer-2 subnet, and the effort to make them routable by encapsulating the packets within IP headers further increases the network infrastructure complexity.

The Low-Latency Network Imperative

Lowering latency in networks is increasingly important for enterprises due to virtualization, Big Data applications, Web 2.0 services, and an increased number of real-time applications.

In many of these applications, completing a task is dependent on data calls to servers or storage in other parts of the data center. This east-west traffic is growing and requires a low-latency infrastructure in order for the application response time to be acceptable.

Running a virtualized process, for example, can involve data that is stored in several virtual storage locations, thus creating numerous latency sensitive east-west traffic flows. Similarly, changes in Web 2.0 services also create a significant amount of intra-data center traffic. One Web page request from a user can create many simultaneous requests for images, cookie information, video, or advertising from various servers in the data center. Page load speed depends on each of these data flows traversing the data center fabric and then heading "south" to the user.

In addition to latency, data transfer efficiency is increasingly important as more data centers see increased server utilization due to virtualization, which adds data copying and processing that impact the server processor. iWARP technologies such as zero copy, kernel bypass, and CPU bypass create a direct data pipeline from the virtual machine to the network controller, reducing the need for context switching in the server and thus freeing the system processor to handle more application work.

The New iWARP

All RDMA variants require a new application-programming model that has limited adoption of the protocol to networks serving HPC and FSI applications. RDMA replaces the network socket paradigm that is the standard for TCP/IP networks with a set of communication "verbs": asynchronous operations using RDMA's concept of a send-and-receive queue pair.

This complexity associated with using RDMA and iWARP diminished significantly when Microsoft added support for iWARP verbs to Microsoft Windows Server* 2012 R2, with built-in support for file server operations using Microsoft's Server Message Block (SMB) 3.0 technology and Hyper-V* virtual machine migration using Live Migration. Now network managers need to build only a network with iWARP controller cards, and Windows Server 2012 will automatically take advantage of available RDMA pathways. Because iWARP runs on TCP/IP, no changes are needed to an existing switch and router network infrastructure.

SMB Direct Delivers Plug-and-Play iWARP

Server Message Block (SMB) Direct is a data transport protocol that, for the first time, delivered plug-and-play support for iWARP within a leading server OS.

SMB Direct is part of SMB 3.0, the file transfer functionality built into Windows Server 2012. With SMB Direct, once the network adapter driver is installed, all iWARP features are automatically enabled. This means that any file transfer that uses SMB 3.0 can benefit from the efficiency and low latency of iWARP.

In addition, Windows Server 2012 features multi-channel SMB technology, which allows the OS to consider latency and congestion of the network and to choose when to use the protocol to maximize performance and efficiency.

SMB Direct is one more reason why iWARP is ready for mainstream adoption in data center and cloud applications.

Another driving force behind RDMA's mainstream adoption is the rise of non-volatile RAM (flash)-based storage, which breaks performance barriers that have held back storage latency and I/O capacity to the speed of the slow, mechanical spinning disk spindles. RDMA allows the network fabric to match the performance of flash-based arrays, thus fully realizing their advantages. iWARP is again exceptionally well suited for this application, as a high-performance transport that can natively share the same Ethernet infrastructure with other storage protocols.

iWARP Performance Advantages

IB is still perceived by many as the performance leader in low-latency network design. So, how does today's iWARP match up against IB? Recent tests conducted using Chelsio Terminator T5* Unified Wire Adapters, demonstrate that iWARP and IB latency is near parity today.

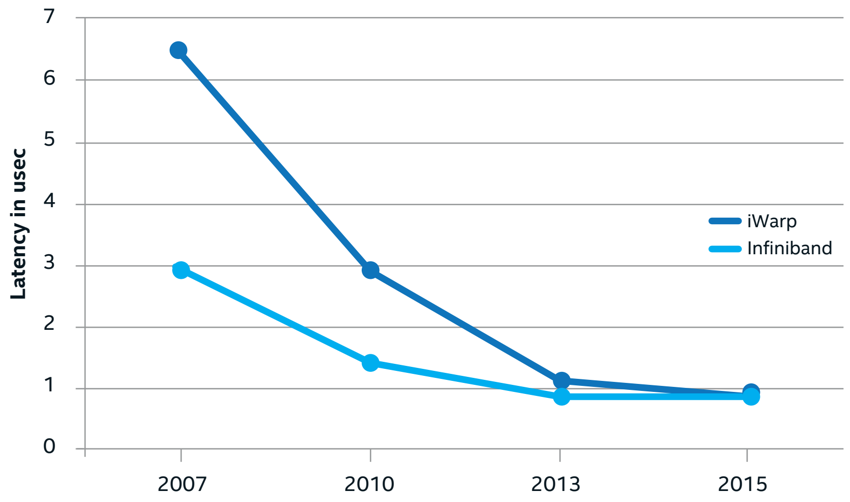

Figure 1. iWARP versus InfiniBand* latency.

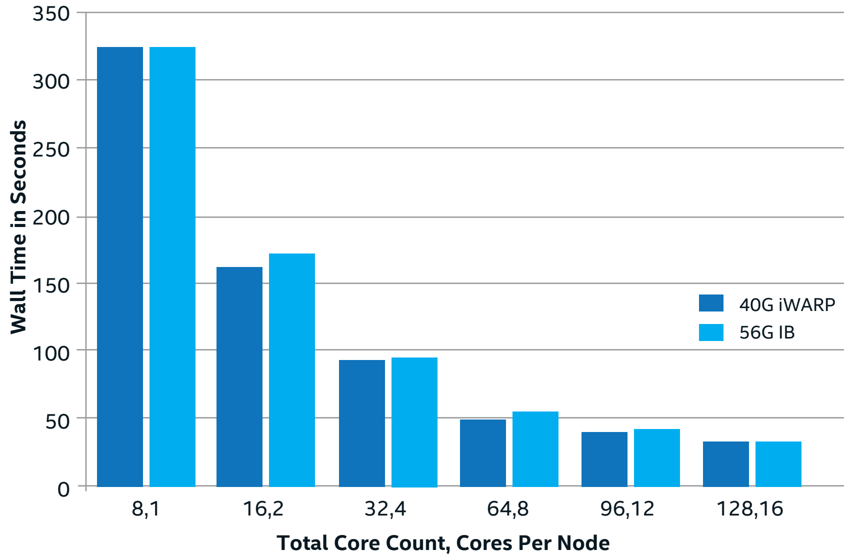

Figure 2. WRF execution time for 40Gb iWARP versus 56G InfiniBand* (lower is better).

IB proponents promulgate a myth that TCP technology itself keeps iWARP from reaching the latency levels of IB. They argue that IB has a more efficient link layer protocol that eliminates latency from higher layer protocols. However, the performance shown in Figure 1 dispels that myth.

Part of the problem has always been an apples-to-oranges comparison: the IB tests were done using controllers with dedicated processors and highly optimized firmware, whereas the iWARP tests were conducted using controllers based on general-purpose processors with non-optimized firmware.

However, the Chelsio controllers used to generate the data shown here are built around a specialized cut-through processor which only requires 10 additional nanoseconds to fully process a NIC packet through the TCP/IP and RDMA stacks.

iWARP performance gets even better when it leaves the lab and is tested in real-world scenarios. The test result in Figure 2 shows weather research and forecasting execution time from an IBM study that compared 40Gb iWARP and 56G FDR IB.

The results show the two technologies delivering equal performance results with iWARP delivering slightly better performance in several scenarios.

Similar testing on four different applications conducted by researchers at Pennsylvania State University led to a report stating, "The testing conducted to date shows that an iWARP-enabled 10GbE Ethernet network is a credible competitor to a dedicated IB fabric for the purposes of computational applications."

So even though iWARP testing demonstrates some micro-benchmark latency differences with IB, real world testing demonstrates that iWARP can match or exceed competing IB gear in real-life application-level benchmarks.

iWARP For the Data Center and Cloud

The previous real-world tests focused on HPC applications. The future opportunity for iWARP technology is in data center and cloud applications, and the Chelsio labs test results in Figure 3 demonstrate how iWARP and IB compete in a file server application using Windows Server 2012 and SMB 3.0.

Even though FDR IB is capable of sustaining a higher bandwidth than iWARP-enabled Ethernet, the throughput advantage is not nearly as great. In fact, that advantage disappears as the size of the data packets grows.

Given the investment most data centers have in Ethernet and the in-boxed support for iWARP in Windows Server 2012, the diminishing advantage of IB is not worth the cost of installing and maintaining a completely separate infrastructure.

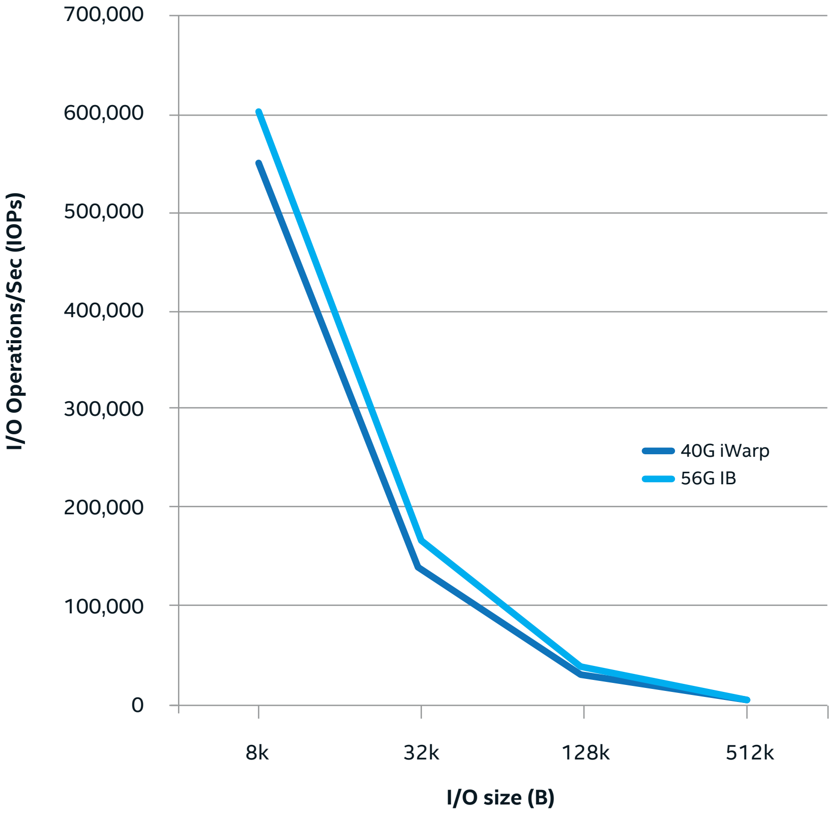

Figure 3. iWARP versus InfiniBand* SMB direct throughput and IOPS performance.

Conclusion

Data center and cloud and web technology trends will continue to put pressure on network managers to deliver low-latency network performance and get the most performance from their virtualized servers. At 40Gbps, iWARP is demonstrating that it is the best technology on the market to deliver network latency performance, routed scalability, and improved CPU efficiency for mainstream cloud and data center workloads. It can accomplish this without the need for new networking infrastructure or specialized network switch features. With performance and the simplicity of plug-and-play integration with Windows Server 2012, network managers need to take a close look at how iWARP can benefit their network.

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 at www.intel.com.

0814/BY/MESH/PDF ♻ Please Recycle 331053-001US