HPCwire

The Leading Source for Global News and Information Covering the Ecosystem of High Productivity Computing

Home Page | Free Subscription | Advertising | About HPCwire

**Features:**

# Using iWARP to Accelerate Web Protocols

**by Dennis Dalessandro & Pete Wyckoff**

**Ohio Supercomputer Center**

The shortcomings of the way in which TCP/IP is currently handled by the CPU and operating system are well known, and the subject of many research studies. The reason for these shortcomings is clear when we look at the network adapter's reliance on the host operating system. Today's network interface controllers (NICs) simply move data on to, and off of the wire. The NIC relies on the host operating system to handle any protocol processing, error checking, and control. Thus, as network speeds rise, the CPU becomes much more involved, spending an increasingly disproportionate amount of time servicing network processing requests.

This, however, is only one aspect of the problem. The other major bottleneck is moving the data between user application buffers and the NIC. Generally, the CPU moves data from the buffer to the NIC by making a copy of the data in kernel space, then copying that data onto the NIC. This memory bottleneck is not something that can be solved by simply using a faster CPU. Furthermore, recent trends show network performance growing faster than the CPU clock rate.

There are a number of solutions already proposed, as this is not a new problem. Technologies such as InfiniBand, Myrinet and Quadrics, are all household names in the high performance computing (HPC) world, and aim to be a solution to the performance limitations of Ethernet communications. A key technique called Remote Direct Memory Access (RDMA), allows a host to move data to another host without involving the CPU on either end. It is, in effect, directly accessing a remote host's memory.

A drawback to these technologies is that they each use their own specialty hardware, none of which are based on TCP/IP. However, it is an undeniable fact that the world runs on TCP/IP. TCP is the network of choice for everything from your broadband connection, to the network connection in your office. This is where iWARP comes in, recognizing not only the need to maintain a TCP/IP-based network, but also the power of an RDMA solution. In combining these two ideas, iWARP has emerged as a particularly attractive network solution. In other words, iWARP takes the best features of networks like InfiniBand and makes it work over ordinary Ethernet.

Some of the most important applications today involve the World Wide Web. Built over TCP/IP networks, the Internet, impacts our lives more and more every day. It is a perfect vehicle to show the benefits of iWARP. By utilizing iWARP, web servers will show a dramatically decreased CPU load, be able to handle more client requests, and handle those requests faster than is possible with ordinary NICs. Interesting web pages are dynamic -- meaning that the content delivered is calculated on the server, possibly using languages like PHP or Ruby on Rails, and interacting with a database. All this work must proceed at the same time that data is transferred because an active web server handles requests for multiple clients at the same time.

At OSC, we have created a module for the popular web server, Apache. This module enables Apache to take advantage of iWARP for transferring data. The HTTP request and response are still sent over an ordinary TCP connection, but the bulk of the data in the web page is shipped using RDMA. Encoded in the HTTP header is the information necessary for the server to directly place data in the client's memory. Alternatively, a client can elect to use RDMA Read to retrieve data from the server's memory. There are instances in which one scheme is more attractive than the other, so both are provided. By using iWARP to transfer the data, the server is able to minimize its network processing load, and spend more time processing requests and rendering content.

Keep in mind, this only works if both the server and client speak iWARP. Since iWARP is in the early adoption phase, hardware can be quite costly, especially for ordinary users. The good news is that since iWARP is based on TCP/IP, an ordinary NIC is able to talk to an RDMA Enabled NIC (RNIC) by emulating the iWARP protocols in software. Through the use of "software iWARP," it is possible for client applications to be written to use the iWARP protocols, thus enabling the server to reap the benefits of using its RNIC. The client sees a bit of an overhead for doing this, but the benefit is realized on the server side as it is able to handle more clients, and process their requests faster. The client does see a benefit, in that the server can respond faster to its requests. Our software iWARP is freely available as open source, and meant to serve as a guide for others who wish to incorporate such functionality into their software.

This work is different from typical RDMA uses, in that it is not limited to the HPC domain. Anyone with a busy web server could potentially benefit from iWARP. The biggest hurdle is getting the clients to speak the iWARP protocol. Fortunately, it is possible to write a plug-in for your browser or web application of choice that lets it speak iWARP using an existing Ethernet card. As part of our work we have enabled the command line tool wget to use software iWARP, and we hope to do similar work for more popular browsers like Firefox.

The range of applications that could benefit from such a scenario is immense. Everything from search engines, which not only transfer data, but perform complex searches through databases, to streaming video, which not only sends data, but processes and often compresses it, can be improved. Even ordinary web sites that do not have a lot of dynamic content, but receive a lot of traffic, would benefit from offloading the burden of copying data to the network.

At SC06, we are showing a poster that explains this technology in more detail. We will also have a live demonstration of iWARP-enabled Apache running in OSC's booth at the show. The poster is based on work we have done with a gigabit iWARP adapter, while the live demo will use recently available 10 Gigabit adapters from NetEffect, connected through a Fujitsu 10 Gigabit Ethernet switch. In both instances, the pool of clients is made up of 1 gigabit hosts. For the poster, the clients are hardware based, and for the demo we will show clients which are software iWARP based.

## HPCwire Readers' and Editors' Choice Awards

2006 Readers' & Editors' Choice Awards

2005 Readers' & Editors' Choice Awards

2004 Readers' & Editors' Choice Awards

2003 Readers' & Editors' Choice Awards

Top of Page