NFS/RDMA over 40Gbps Ethernet

Boosting NFS with iWARP RDMA Performance and Efficiency

Executive Summary

NFS over RDMA is an exciting development for the trusted, time proven NFS protocol, with the promise of high performance and efficiency brought in by the RDMA transport. The unprecedented price and adoption curve of 40Gbps Ethernet, along with the focus on high efficiency and high performance in an era of Big Data and massive datacenters, are driving the interest in performance optimized transports, such as RDMA. RDMA is also particularly interesting when mated to high throughput, low latency SSDs, where it allows extracting the most performance out of these new devices.

This paper presents early performance results for NFS over 40Gbps iWARP RDMA, using Chelsio's Terminator 5 (T5) ASIC. The results show line rate performance with standard Ethernet frames, and a significantly better curve and better CPU utilization than NIC. Thanks to its TCP/IP foundation, iWARP allows using standard Ethernet equipment, with no special configuration and without requiring a fabric overhaul or additional acquisition and management costs.

Overview

The Remote DMA (RDMA) protocol allows efficient memory-to-memory communication between two systems. With RDMA, all network protocol processing, protection and security checking is handled, in hardware, by the RDMA adapter. Chelsio's T5 RDMA implementation is a high performance, third generation design, which benefits from experience with hundreds of thousands of chips deployed in the field, across multiple generations.

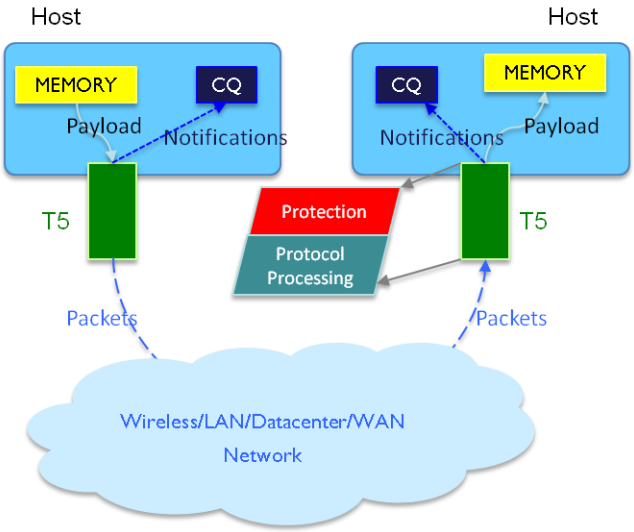


Figure 1 – RDMA Overview

The RDMA API – called the Verbs layer – provides an asynchronous communication model that allows user space access to hardware, and application polling for lowest communication latency. Traditional socket networking applications must be rewritten in order to make use of RDMA and benefit from it.

In addition to low latency, RDMA benefits include kernel and CPU bypass, as work requests are handed out directly from the application to the hardware, and CPU involvement is typically only required when an end of data transfer notification is received. Last but not least, RDMA enables zero copy transfer, as data get sourced and placed directly from/into application buffers.

The Internet Wide Area RDMA Protocol (iWARP) is the IETF standard for RDMA over Ethernet. It builds upon the proven TCP/IP foundation and benefits from its routability, scalability, reliability, flexibility and resilience to adverse network conditions. Unlike InfiniBand, users of iWARP can preserve their investments in network functions, such as security, load balancing and monitoring appliances, and infrastructure in general. Thanks to TCP/IP, iWARP can natively run over regular Ethernet switches and routers, as well as operate over long distance links.

NFS over RDMA

Storage protocols like NFS and SMB are particularly well suited to using and benefiting from RDMA, because of their characteristics and performance requirements. Thanks to transport abstraction layers, the changes are limited from the point of view of the application itself. The development expense of utilizing the RDMA verbs interface is thus contained to the RDMA specific transport library. The following diagram shows the software layering for NFS/RDMA at the client side over a Chelsio T5 NIC.

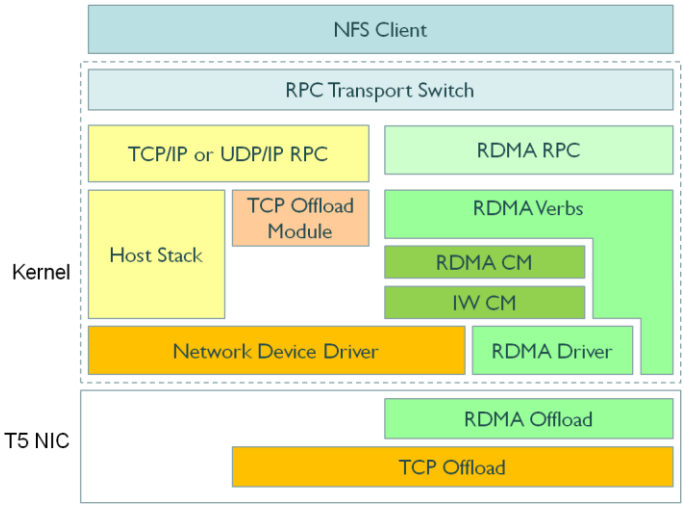


Figure 2 – NFS/RDMA Client Stack over T5

The RPC Transport Switch allows using the host software TCP/IP or UDP/IP stack, the offloaded TCP/IP stack on the T5 adapter, or the RDMA RPC module, which takes advantage of the T5 RDMA capability. The RDMA Verbs layer allows access to connection management (CM) functionality in the kernel, to access the RDMA driver for queue management, as well as direct

access to hardware for send and receive operation. T5 allows the flexibility of using all the different transports simultaneously over the same port.

Test Results

The following graphs compare NFS READ and NFS WRITE throughput and interrupts per second results for NFS over iWARP RDMA and regular NIC at different IO sizes using the **iozone** tool.

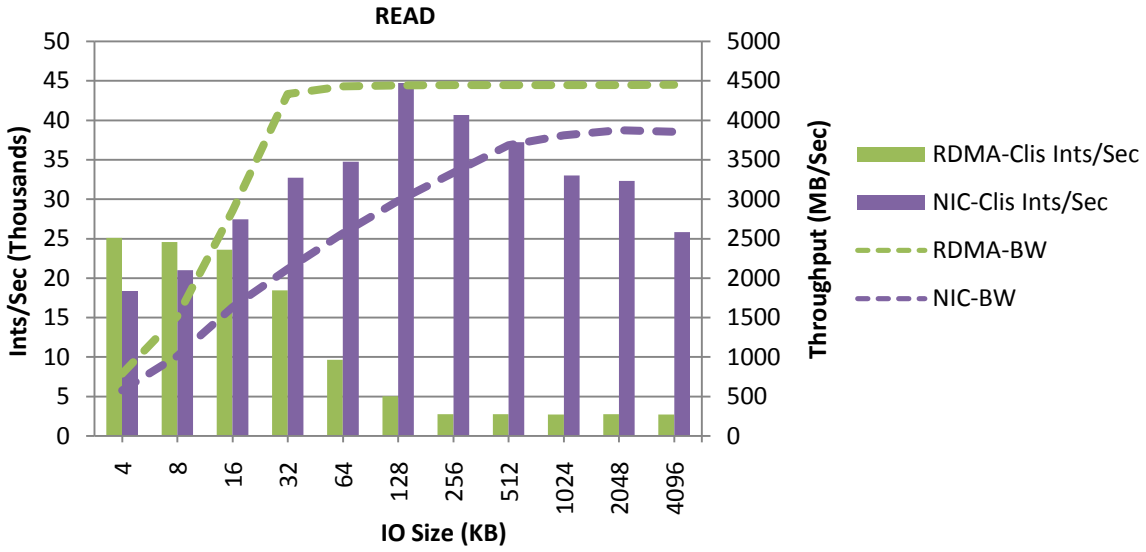


Figure 3 – NFS READ Throughput and Interrupts/sec vs. IO Size

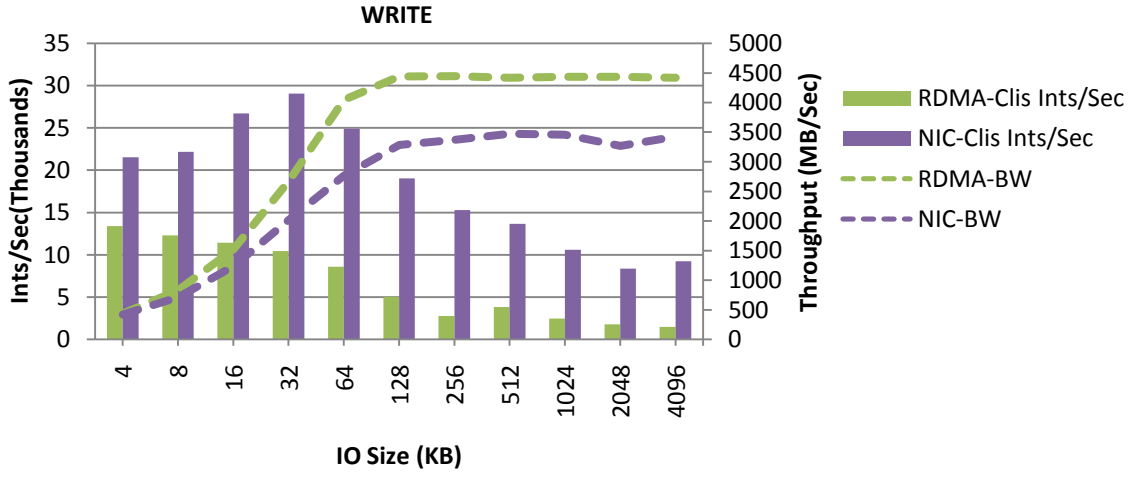


Figure 4 – NFS WRITE Throughput and Interrupts/sec vs. IO Size

The results show RDMA consistently beating NIC in throughput, as the latter fails to reach 40Gbps line rate, whereas RDMA does so at size 32KB for READ and 64KB for WRITE.

Furthermore, the lower interrupt rate at the client clearly shows the benefit of RDMA with a linear decrease as the IO size increases, as expected.

The following graphs compare NFS READ and NFS WRITE throughput and CPU idle percentage for iWARP RDMA and FDR InfiniBand at different IO sizes using the **iozone** tool.

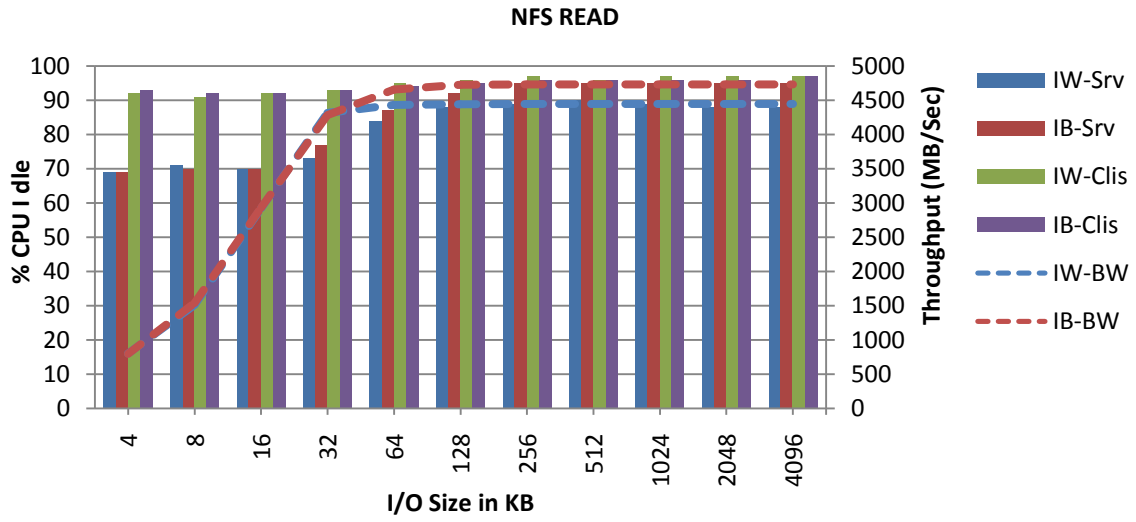


Figure 5 – READ Throughput & CPU Idle % vs. IO Size

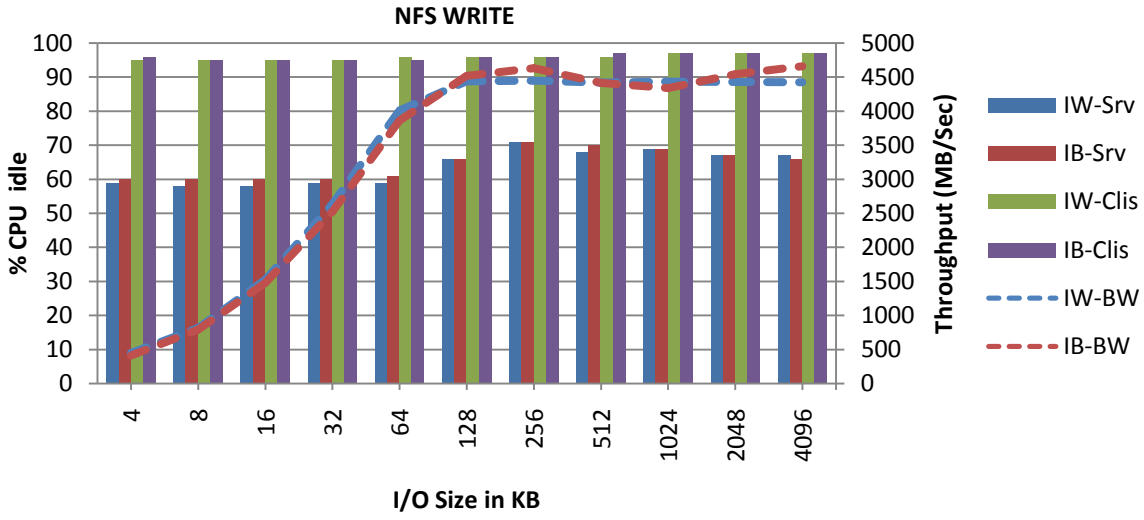


Figure 6 – WRITE Throughput & CPU Idle % vs. IO Size

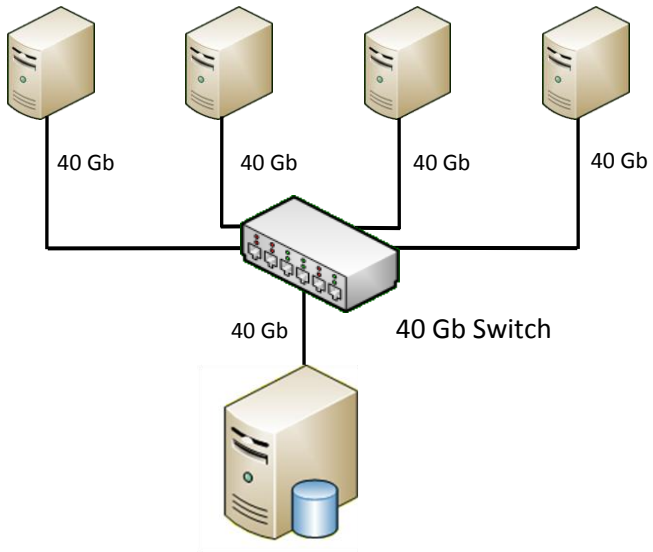
The results show that iWARP at 40Gbps and IB FDR (56Gbps) provide virtually the same NFS performance, in throughput and CPU utilization, although the latter is theoretically capable of higher wire rate. This result is another that affirms this conclusion; shown in all previous studies of the two fabrics, making iWARP at no compromise, drop-in Ethernet replacement for IB.

Test Configuration

The following sections provide the test setup and configuration details.

Topology

NFS Clients with T580-CR adapters running on RHEL 6.5 (3.16.0 kernel)



NFS Server with T580-CR running on RHEL 6.1 (3.16.0 kernel)

Figure 7 – Test Setup

Network Configuration

The network configuration consists of an NFS server connected to 4 client machines through a 40Gbps switch using a single port. Standard MTU of 1500B is configured.

The server machine is setup with two Intel Xeon CPU E5-2687W v2 8-core processors clocked at 3.10GHz and 64 GB of RAM. Inbox iWARP driver is installed in the system with RHEL 6.1 (3.16.0 Kernel) operating system. The setup included the latest NFS/RDMA patches available from the developers.

Each of the 4 client machines is setup with two Intel Xeon CPU E5-2687W v2 8-core processors clocked at 3.40GHz and 64 GB of RAM. Inbox iWARP driver is installed in each system with RHEL 6.5 (3.16.0 Kernel) operating system.

The T580-CR adapter is used in server and client machines in the Chelsio setup, whereas the MT27500 Connect-X FDR adapter is used in the InfiniBand setup.

Storage Topology and Configuration

The Server exposes a 32 GB ramdisk share with ext2 filesystem.

I/O Benchmarking Configuration

Iozone is used to assess the I/O capacity of the configuration. This test uses sample IO sizes varying from 4KB to 4096KB. Buffering is set to none, and the I/O access pattern used is sequential READs and WRITEs.

Commands Used

```
[root@host]# iozone -+n -i [0|1] -M -c -e -w -I -+u -t 16 -s 1g -r <IO Size>
```

Conclusions

The RDMA fabric offers potential for improved efficiency, as recently shown with the SMB v3.0 SMB Direct RDMA transport, sparking interest in enabling RDMA for other storage protocols. There is thus renewed interest in NFS over RDMA, where initial benchmarks already show performance benefits compared to host stack running over a standard server NIC.

Thanks to its TCP/IP foundation, iWARP RDMA is ready for deployment over wireless, LAN, datacenter and Cloud infrastructure, without modifications or special configuration. As a result, it offers exceptional savings in acquisition and operation costs, as opposed to other RDMA transports, which require specialized infrastructure, management tools, and administrative expertise.

Chelsio's T5 iWARP RDMA over Ethernet is shipping at 40Gbps, and is part of a high performance Unified Wire over Ethernet alternative to InfiniBand, enabling simultaneous operation of RDMA, NIC, TOE, iSCSI and FCoE. With superior performance across the board, as recognized in independent studies, Chelsio's adapters remain "great all-in-one adapters".

Related Links

The Chelsio Terminator 5 ASIC
40Gb TOE vs. NIC Performance
iSCSI over 40Gb Ethernet
FCoE at 40Gb with FC-BB-6
Helen Chen paper from 2007