Meeting Today's Datacenter Challenges

Produced by Tabor Custom Publishing
in conjunction with:

HPC wire

Chelsio Communications
Accelerate

Introduction

In this era of Big Data, today's HPC systems are faced with unprecedented growth in the complexity and size of the data sets they are handling along with runaway memory requirements. This deluge of data is being fueled by the exponential growth in the parallel processing capabilities of new GPUs being introduced by companies like NVIDIA as well as CPUs being introduced by companies like Intel, in particular the latest addition to the Intel Xeon Phi family, code named Knights Landing.

These developments lead to the rapid growth of large HPC clusters with computation capabilities of many core, many socket systems as an integral part of today's Big Data escalation. This includes the development of large HPC clusters that leverage the computational capabilities of modern CPUs and GPUs.

With these processor enhancements there is a need to speed-up movement of large data sets between the memory and processors and between other machines. As a result, companies have turned to advanced network fabrics that incorporate RDMA (Remote Data Memory Access) to create a large data pipe that can be used to move data quickly with low latency, high bandwidth and minimum CPU resources.

Not only can RDMA move data efficiently between machines with the recently released specification for NVMe over Fabrics (NVMeF) by NVM Express, Inc, a new storage interface protocol has been created that optimizes access for non-volatile storage systems as well as provides network access beyond the host machine to access remote storage. To deliver the expected performance, NVMeF requires a low latency and highly efficient network like a network that includes RDMA.

Along with Big Data, the growth of Hybrid and Private cloud implementations also presses the need for a truly converged and an all in-boxed I/O solution to meet today's demand of growing enterprise and cloud implementations. Microsoft Storage Spaces Direct (S2D) is a leading example, which enables building highly available and scalable storage systems by pooling local server storage. This is an effort to build Highly Available Storage Systems using storage nodes with only local storage. S2D leverages the SMB3 protocol, which includes SMD Direct and SMB Multichannel, for all intra-node communications thereby creating a low latency and high throughput environment for storage.

Meeting the Challenges

There are a number of excellent tools available to help datacenter managers meet these specific challenges. They include solutions such as the iWARP and and other Remote DMA (RDMA) methods , both of which speed up the movement of data between a system's CPUs and memory.

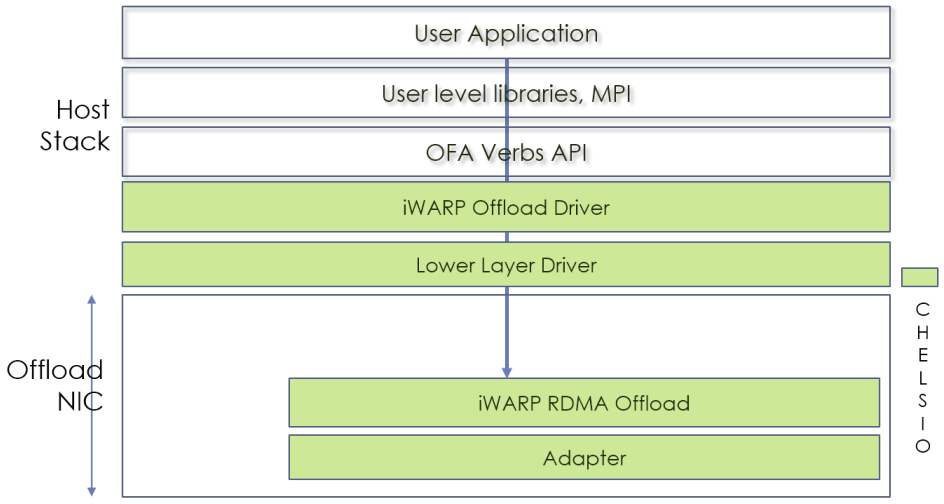
iWARP – an Essential Standard

This standard was developed by the iWARP Consortium and standardized by IETF in 2007. It layers RDMA on top of TCP/IP – the main transport protocol used in the Internet, datacenters, cloud installations and in Ethernet networks in general. According to network statistics, TCP carries over 90% of Internet traffic. iWARP allows RDMA and MPI applications to be ported from machines that use the InfiniBand/RoCE interconnect to one that uses iWARP on ethernet in a seamless fashion.

iWARP enables a brownfield strategy and provides a high performance and low latency RDMA transport. It preserves investments in Ethernet network functions, such as security, load balancing and monitoring appliances, and network infrastructure in general, all without the need for an expensive gateway, special configurations or additional management costs.

Thanks to the iWARP hardware TCP/IP foundation, it provides low latency and all the benefits of RDMA, with routability to scale to large clusters and long distances.

In addition to providing all of the total cost of ownership benefits of Ethernet, iWARP delivers several distinct advantages:



- It is an established IETF standard
- Enables incremental, non-disruptive server installs and supports the ability to work with any legacy (non-DCB) switch infrastructure

- It's plug-play by nature - It has equivalent network switch configuration requirements such as "non-RDMA NICs"
- Enables a decoupled server and switch upgrade cycle and a brownfield strategy to enable high performance, low cost RDMA enablement
- It is built on top of TCP/IP, making it routable, reliable, and scalable from just a few nodes to thousands of collocated or geographically dispersed endpoints
- It uses the familiar TCP/IP/Ethernet stack and therefore leverages all the existing traffic monitoring and debugging tools
- It allows RDMA and MPI applications to be ported from InfiniBand interconnect to IP/Ethernet interconnect in a seamless fashion

Chelsio's Terminator 6 ASIC offers a high performance, robust fourth generation implementation of iWARP RDMA (Remote Direct Memory Access) over 10/25/40/50/100Gb Ethernet Unified Wire adapters, delivering end-to-end RDMA latency that is comparable to InfiniBand, using a standard Ethernet infrastructure. Chelsio's iWARP is in production today in GPU applications, in storage applications as a fabric for clustered storage, in Lustre and other storage applications, in HPC applications, and in remote replication and disaster recovery. It is a high performance, robust, reliable, and mature protocol that enables direct data placement, CPU savings, and RDMA functionality over TCP/IP and legacy Ethernet switches and internet with no performance penalties.

Chelsio Unified Wire Ethernet adapters also concurrently enable a full suite of networking and storage protocols, including user space I/O with Chelsio's WireDirect, full offload of

TCP/IP and UDP/IP, iSCSI and FCoE, all traffic managed and firewalled.

RDMA for Communication Efficiency

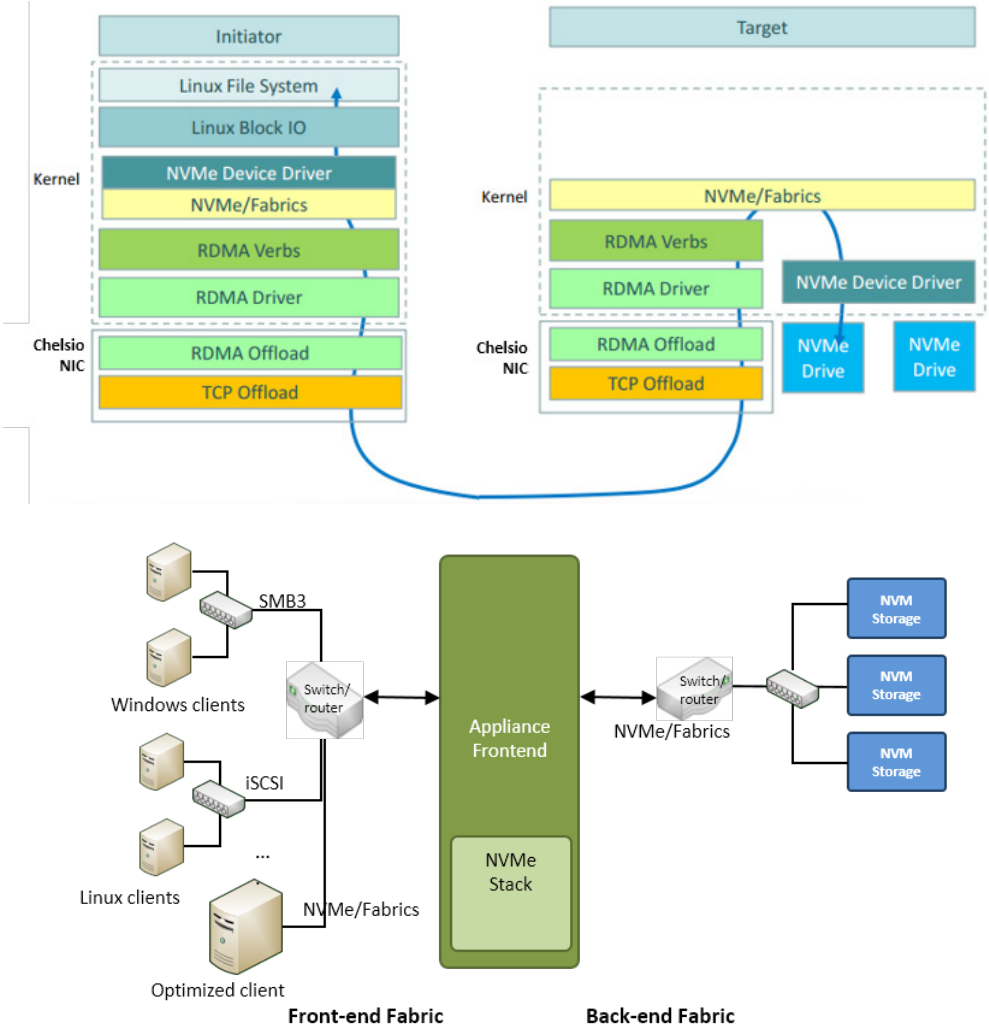
RDMA achieves unprecedented levels of communication efficiency by using a direct system or application memory-to-memory transfer without CPU involvement. All packet and protocol processing required for communication is handled by the network adapter.

Some examples of iWARP RDMA applications (use cases) include:

- Microsoft Storage Spaces Direct: The Chelsio RDMA enabled Ethernet adapter delivers a high performance Storage Spaces Direct (S2D) solution using standard Ethernet infrastructure and enables datacenters to deploy S2D now by leveraging all-inboxed drivers. The ability to work with any non-DCBX switch, enables an immediate plug-and-play deployment. Support of iWARP protocol is enabled since Windows Server 2012-R2 release, has allowed for years of testing for a very robust, tested, and efficient deployment with Chelsio iWARP enabled Ethernet adapters.
- Chelsio supports Client RDMA capabilities in Windows 10 Enterprise and enables large file applications such as oil/gas and video post-production that require I/O-

intensive access to storage to see dramatic performance benefits using SMB Direct/ iWARP based client to storage communications.

- NVMe and NVMe over Fabrics – These are optimized standard interface for high performance SSD storage. The technology is supported by Chelsio's high performance iWARP RDMA over Ethernet. The technology provides a plug-and-play solution for connecting high performance SSDs over a scalable, congestion controlled and traffic managed traffic, with no special configuration needed.



- GPUDirect RDMA – This NVIDIA technology in combination with Chelsio's iWARP RDMA adapters, allows you to build powerful clusters and supercomputers. Access to the GPU is achieved at all supported speeds across a standard Ethernet network infrastructure.

- Hadoop – Chelsio unified wire adapters deliver a range of performance gains for Hadoop by bringing the Hadoop cluster networking into optimum balance with recent improvements in server and storage performance, minimizing the impact of high speed networking on the server CPU. The end result is improved Hadoop distributed

file system (HDFS) performance and reduced job execution times.

- NFS over RDMA and Lustre 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. Chelsio iWARP RDMA solution allows direct system or application memory-to-memory communication, without CPU involvement or data copies. iWARP RDMA uses a hardware TCP/IP stack that runs in the adapter, completely bypassing the host software stack, thus eliminating any inefficiencies due to software processing. The performance results show that iWARP at 40GbE is on par with IB-FDR, while utilizing standard Ethernet infrastructure, with no special configuration or management needed. Thanks to the resulting cost and management savings, iWARP is the most cost effective high performance RDMA transport available today.

Microsoft Storage Replica is a new feature in Windows Server 2016 that enables storage-agnostic, block-level, synchronous replication between clusters or servers for disaster preparedness and recovery, as well as stretching of a failover cluster across sites for high availability. Synchronous replication enables mirroring of data in physical sites with crash-consistent volumes, ensuring zero data loss at the file system level. Asynchronous replication allows site extension beyond metropolitan ranges.

iWARP and RDMA – Delivering the Benefits

The combination of iWARP and RDMA yield a number of specific benefits including:

- **Reliability** – Most packet networks (e.g. Ethernet and IP) are “best-effort” where packets can get dropped or re-ordered. TCP handles re-ordering and data retransmission, providing reliable data transfer in all environments, including long distance and next generation high speed wireless links, expected to reach 5 to 10Gb speeds in the near-term future.

- **Flow control** – TCP allows the receiver to flow control the sender to avoid over-subscribing its resources. This end-to-end control allows TCP to operate between vastly different endpoints, e.g. servers with 10x or 100x the network connectivity speed of clients.

- **Congestion control** – TCP implements algorithms to automatically adapt its transmission rate to the network capacity in order to avoid and react to congestion. This prevents collapse when load increases beyond trivial levels, and allows TCP to work at high performance in large scale or heterogeneous networks and across network boundaries.

- **Robust & Plug-Play** - The past decade of experience has enabled iWARP to mature and increase in robustness. The technology is now included in all major software distributions, all the while gaining in performance and capabilities. A true plug-and-play native of the Cloud and datacenter era, iWARP is the safe RDMA over Ethernet solution that is available today.

- iWARP's native support for reliability and congestion control mechanisms ensures maximum scalability, routability, reliability and robustness without requiring a lossless fabric or Ethernet PAUSE to be enabled. It also guarantees ease of deployment and use, and allows leveraging all existing infrastructure, including networking, monitoring, security and management with no changes required.
- iWARP is supported in the same OFED distribution as InfiniBand, the incumbent RDMA provider, and requires no changes to RDMA applications to run over Ethernet.

For more information on Chelsio's iWARP RDMA, please visit: http://www.chelsio.com/nic/rdma-iwarp/

Editor's Note: This compilation is based on information provided by Chelsio Communications, its partners and customers.