High Performance Disaster Recovery at 40GbE

Chelsio T5 iWARP RDMA solution for Windows Storage Replica

Executive Summary

Storage Replica (SR) is a Windows Server 2016 Technical Preview 3 feature which enables block-level replication between clusters or individual servers for disaster recovery, and stretching of failover clusters to metropolitan (MAN) and wide area (WAN, US coast-to-coast) distances for high availability. SR provides two modes of operation: *synchronous* and *asynchronous* replication. Synchronous replication enables mirroring of data with zero data loss at the volume level, whereas asynchronous replication trades off full data replication guarantees for reduced latency by locally completing I/O operations.

This paper highlights how Storage Replica over Chelsio's iWARP RDMA solution combines high performance with the high efficiency provided by the zero copy and kernel/CPU bypass operation of the RDMA transport to provide a reliable, scalable and robust disaster recovery solution for mission critical workloads.

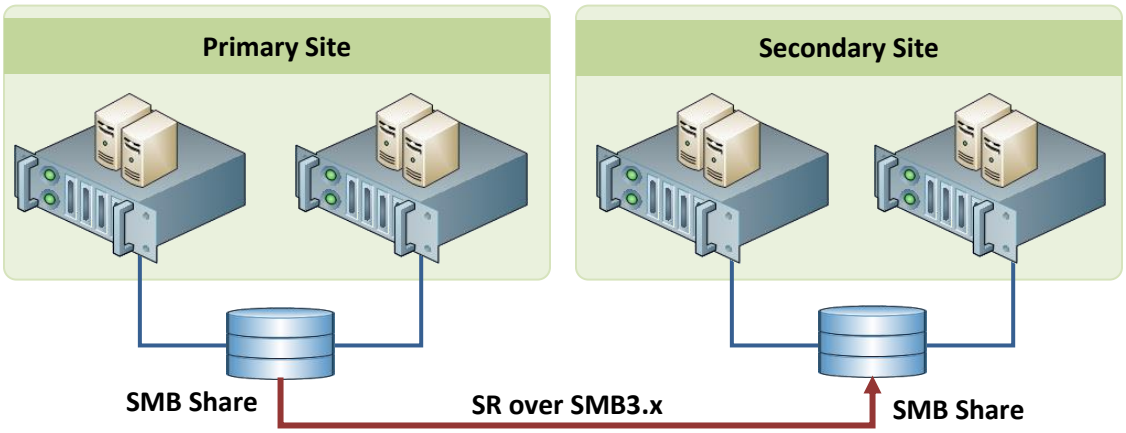

Figure 1 – Microsoft Storage Replica Setup

The Chelsio Terminator 5 ASIC

The Terminator 5 (T5) ASIC from Chelsio Communications, Inc. is a fifth generation, high-performance 2x40Gbps/4x10Gbps server adapter engine with Unified Wire™ capability, allowing **offloaded storage, compute and networking** traffic to run simultaneously.

Remote DMA (RDMA) is a technology that achieves unprecedented levels of efficiency, thanks to 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. iWARP RDMA provides all the benefits of RDMA, including CPU **bypass and zero copy**, while operating over standard, plug-and-play Ethernet.

Thanks to its TCP/IP foundation, iWARP allows using standard Ethernet equipment, with no special configuration (like DCB and PFC) and without requiring a fabric overhaul or additional acquisition and management costs. iWARP stands out as the standards-based, mature, routable, scalable and robust Plug-and-Play option that is shipping today at 40Gbps from multiple vendors. In contrast, alternative RDMA transports like RoCE and InfiniBand impose additional costs in equipment, such as DCB or Metro-X, and management complexity. In fact, RoCE is not supported by Storage Replica due to range limitations and lack of routability. The second, incompatible incarnation of the protocol – RoCEv2 – lacks critical congestion control capabilities to operate over long distance links. All versions of RoCE require special configuration and are hard to deploy and maintain. These attributes make T5 with iWARP the preferred solution for all storage networking needs: frontend, backend and high availability.

The Demonstration

Microsoft Corp. recently collaborated with Chelsio to showcase SR operating over a 50Km fiber loop, in synchronous mode at the Microsoft Ignite conference in Chicago, IL. The demonstration showed SR operating at 40Gbps using SMB3.1.1 over Chelsio's T580-LP-CR RDMA enabled NICs.

The setup consisted of a Server connected to a Client, using single 40Gbps link and standard MTU of 1500B. Both machines were configured with 2 Intel Xeon E5-2660 0 8-core processors running @ 2.20GHz, 256 GB of RAM, Windows Server 2016 Technical Preview 3 OS (Build 10158), 1 NVMe Intel P3700 1.45TB SSD and Chelsio inbox driver v5.3.22.140.

Results

The following results were obtained using SR over Chelsio 40GbE Network:

- Storage line rate of 1.75GB/s (limited by Intel P3700 NVMe) - Remote SR performance identical to local disk.
- Consistent replication time of 857 secs with 1.45 TiB of data.
- Only ~25 to 30% Network utilization- Considerable Bandwidth available for other applications.

Command used

```
# diskspd.exe -c1G -b<IO Size> -t8 -o8 -r -h -L -w10 -d60 -W5 -C5 testfile.dat
```

Summary

This paper presented an overview of how Microsoft's Storage Replica and Chelsio T5 iWARP RDMA solution offers an all-round disaster recovery solution for mission-critical applications during power outages. The results confirm iWARP's native TCP/IP ability to operate beyond a single datacenter environment, to extend the RDMA transport over long distance, and its robustness under SR's load pattern. Long distance replication was shown to provide near local access performance levels, with low impact on I/O performance and latency.

Related Links

iWARP: From Clusters to Cloud RDMA
Storage Replica with RDMA and 25km Fiber in Windows Server 2016 Technical Preview 2