High Performance Storage Replica at 40GbE

iWARP RDMA benefits for Windows Server 2016 Storage Replica

Executive Summary

Storage Replica (SR) is a Windows Server 2016 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.

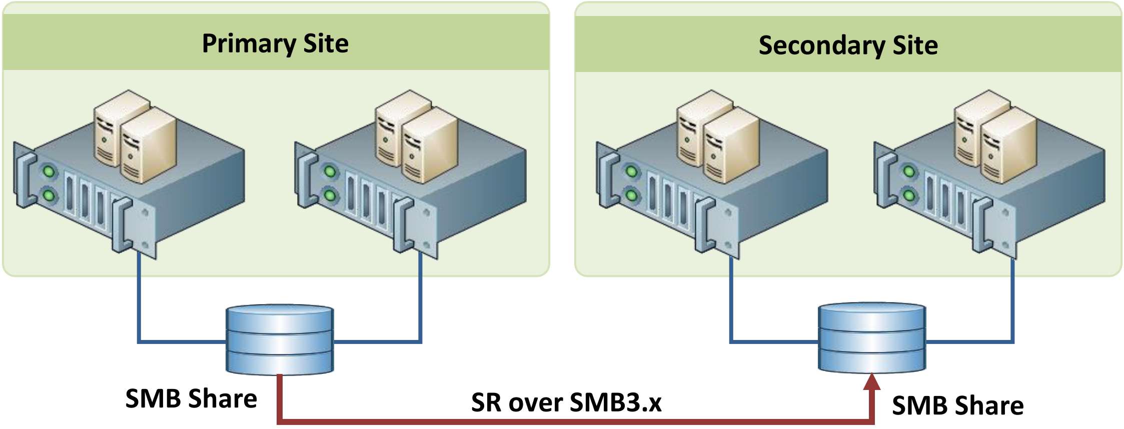

Figure 1 – Microsoft Storage Replica Setup

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^{TM} 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.

Why Chelsio iWARP RDMA Solution for Storage Replica

Chelsio's fifth generation, high performance RDMA Ethernet adapters, utilizing iWARP:

- Enable incremental, non-disruptive server installs.
  - Support the ability to work with any legacy (non-DCB) switch infrastructure.
  - Enable a decoupled server and switch upgrade cycle and a brownfield strategy to enable high performance, low cost S2D enablement.
- Are easy to use and install.
  - Have equivalent network switch configuration requirements "as non-RDMA NICs"
- Are cheaper to deploy → end user can purchase more compute servers for the same investment amount.
  - Do not require gateways or routers to connect to the TCP/IP world.
  - Save significant CPU cycles.
    - Enable cheaper CPUs for equivalent performance.
    - Enable significantly lower datacenter and utilities expenses.
- Utilize very robust and stable protocols.
  - iWARP has been an IETF standard (RFC 5040) for 9 years, TCP/IP has been an IETF standard (RFC 793, 791) for 35 years.
    - No surprises, no fine print, plug and play.
  - Have multi-vendor support.
- Are supported in other Windows products.
  - Client RDMA in Windows 10 enables more deployment options.
  - Storage Replica is natively supported to enable disaster recovery.
  - Network Direct and Nano Server are natively supported.
  - iSCSI HW offloaded initiator is natively supported.
- Are scalable to wherever the datacenter can scale to.
  - Inherit the loss resilience and congestion management from underlying TCP/IP.
- Are very high performance.
  - Extremely low latency, high bandwidth, high message rate.

Below table summarizes the differences of iWARP & two recent variants of RoCE (another RDMA over Ethernet alternative).

Metric	iWARP	RoCEv2 with DCB	RoCEv2 with ECN
Allows Packet Loss	Yes, performance is maintained by hardware retransmission, microsecond timers and TCP tunables to disable slow start and enable fast retransmit	No for Performance Yes, for Poor Performance	No for Performance Yes, for Poor Performance
Ease of Use	Easy 0 extra steps per node	Complicated DCB ~50 extra steps per node	Complicated ECN ~50 extra steps per node

Backward Compatibility	Use on Brownfield and Greenfield with any switch or wireless link	Only supported on DCB switches. ConnectX-3 cards don't support RoCEv2. Only ConnectX-3 Pro does.	Only supported on Mellanox adapters and switches to separate TCP, RDMA, CNP and other traffic. ConnectX-3 Pro cards do not support RoCEv2 with ECN (i.e. Resilient RoCE). Only ConnectX-4 does.
Performance	100Gb/s sub 1µs latency	100Gb/s sub 1µs latency	100Gb/s sub 1µs latency
Cost	Lower due to economies of scale. Works with legacy installs (decouples server/switch upgrade cycles). Does not need gateways or routers.	Higher due to needing specific switches for DCB and gateways to talk to the outside TCP/IP world. Switches must be the same brand (if the same set of ~50 configuration steps required)	Higher due to needing specific switches and gateways to talk to the outside TCP/IP world. Switches must be the same brand (if the same set of ~50 configuration steps required)
Support	No increase in support team size as everything works on plug-n-play TCP/IP/Ethernet	Large support team. More hours to get things up, running and to stay running. Modifications needed as cluster grows or traffic pattern changes	Large support team. More hours to get things up, running and to stay running. Modifications needed as cluster grows or traffic pattern changes

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 Atlanta, GA. 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, 256GB of RAM, Windows Server 2016 Datacenter edition, 1 NVMe Intel P3700 1.45TB SSD and Chelsio inbox driver v6.1.14.200.

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
Hyper-Converged Scale-Unit with Chelsio 40GbE
iWARP/RDMA Benefits in Windows 10
High Performance S2D with Chelsio 40GbE
Storage Replica with RDMA and 25km Fiber in Windows Server 2016 Technical Preview 2