Windows SMB 3.1.1 Performance at 100Gbps

Line Rate Performance with Chelsio T6 100GbE iWARP RDMA

Executive Summary

SMB Direct is an extension of the Microsoft Windows Server Message Block (SMB) file services, which seamlessly leverages Remote Data Memory Access (RDMA) enabled network adapters for improved storage bandwidth, latency and efficiency. The Direct part implies the use of various high speed RDMA methods to transfer large amounts of data with little CPU intervention. This paper provides an overview of Chelsio T6 100GbE iWARP RDMA solution and demonstrates line rate throughput and IOPs numbers in a two machine, back-to-back, server-client setup.

Overview

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.

One of the main advantages of the SMB 3.1.1 implementation is that once the network adapter driver is installed, all its features are automatically enabled and made available to the SMB application. Furthermore, with the multi-channel SMB technology, Windows can choose the best protocol to use at any time, as well as aggregate traffic over multiple different links using different protocols. Chelsio T6/T5 series of iWARP adapters seamlessly support SMB Direct over Ethernet for Microsoft's Windows Server 2016.

Why Chelsio iWARP RDMA Solution for SMB Direct

Chelsio's sixth 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	Higher due to needing specific switches for DCB and gateways to talk to the outside TCP/IP world. Switches must be the same brand	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

	cycles). Does not need gateways or routers.	(if the same set of ~50 configuration steps required)	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

SMB Direct Performance with 100GbE iWARP

This demonstration was conducted in conjunction with **Microsoft** (Press Release).

Test Results

Throughput Test:

Buffer Size/QD	Read IOPS	Read BW (Bytes)	Read BW (Bits)	Average Read (ms)	50th %ile Read (ms)	99th %ile Read (ms)	Read IOPS CoV
262144/2	45,406	11,902,792,499	95,222,339,994	0.176	0.174	0.228	0.1%

95.2 Gbps READ throughput was achieved at 256KiB I/O size using 1 thread/file (4 total), 2 outstanding random I/Os per thread (4 x 2 = 8 total) running to remote memory (file cache).

IOPs Test:

Buffer Size/QD	Read IOPS	Read BW (Bytes)	Read BW (Bits)	Average Read (ms)	50th %ile Read (ms)	99th %ile Read (ms)	Read IOPS CoV
4096/10	529,893	2,170,440,294	17,363,522,355	0.377	0.348	0.571	4.8%

530K READ IOPs was achieved at 4KiB I/O size using 5 threads/file (20 total), 10 outstanding random I/Os per thread (20 x 10 = 200 total) running to remote memory (file cache).

Topology

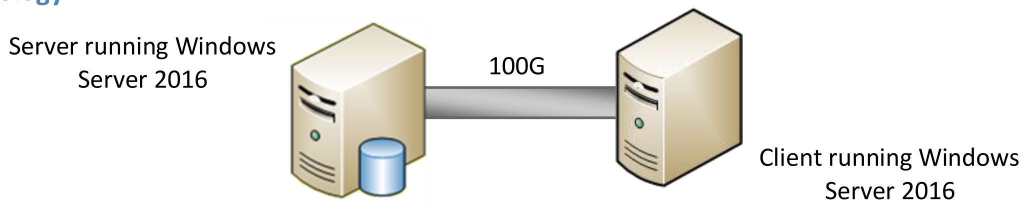
Server running Windows Server 2016
100G
Client running Windows Server 2016

Figure 1 – Simple Back-to-Back Test Topology

Network Configuration

The demo utilizes 2 nodes, connected back-to-back, using single 100GbE link. MTU of 9000B is used. Each node is equipped with the following hardware:

- 2x Intel Xeon E5-2660v3 2.6Ghz (20c, HT disabled, Turbo On)

- 256GB DDR4 RAM
- 1x Chelsio T62100-CR (Dual Port 100Gb PCIe 3.0 x16)
    - Chelsio network driver v6.3.3.0
    - Chelsio firmware v1.16.4.8
    - Single port connected/adapter

Storage Topology and I/O Benchmarking Configuration

The storage server is configured with 4 test files. DISKSPD is used to assess the I/O performance of the configurations.

Commands Used

Throughput Test:

```
C:\Users\Administrator> diskspd.exe -Rxml `-ag0,10,11,12,13,14,15,16,17,18,19 -t1 -o2 -b256K -r256K -Sr -fr -W10 -d60 -D -L <dir. of files>
```

IOPs Test:

```
C:\Users\Administrator> diskspd.exe -Rxml -t5 -o10 -b4k -r4k -Sr -fr -W10 -d60 -D -L <dir. of files>
```

Erin Chapple, Partner Director, Program Management for Windows Server, Microsoft said, ***"Windows Server 2016 and System Center 2016 offer our most cloud-ready server operating system ever, with exciting new innovation to help you transform your applications for the cloud, build a software-defined datacenter with cloud efficiencies, and keep your IT safer than ever. The complementary solutions and services from our partners are what truly brings the innovation to life for our customers as they transform their IT solutions for the cloud-first world."***

Summary

This paper provided performance results for SMB 3.1.1 running over Chelsio's T6 iWARP RDMA enabled Ethernet adapter, T62100-CR. Chelsio iWARP RDMA solution delivers ground breaking 95Gbps line rate performance and 530K IOPs using standard Ethernet infrastructure. With plug-and-play operation, enhanced reliability, higher efficiency and line rate performance, SMB over Chelsio's T6 series of RDMA enabled adapters is an extremely compelling solution for Windows Server 2016 storage networking. Support of iWARP protocol is enabled since Windows Server 2012-R2 release, thus allowing for years of testing for a very robust, tested, and efficient deployment with Chelsio iWARP enabled Ethernet adapters. In addition to SMB Direct, iWARP protocol also powers other aspects of Microsoft Windows installations such as Storage Replica for disaster recovery, Storage Spaces Direct, Client RDMA for bringing RDMA benefits to Windows 10 deployments, and Network Direct for Windows HPC deployments.

Related Links

iWARP: From Clusters to Cloud RDMA
iWARP/RDMA Benefits in Windows 10
Windows SMB 3.0 Performance at 40Gbps
SMBDirect Latency on Windows Server 2012 R2
SMBDirect 40 GbE iWARP vs 56G InfiniBand