High Performance S2D with Chelsio 40GbE

Chelsio T5 iWARP RDMA solution for Windows Storage Spaces Direct

Overview

Microsoft **Storage Spaces Direct** (S2D) is a feature introduced in Windows Server 2016 Technical Preview, which enables building highly available and scalable storage systems by pooling local server storage. You can now build HA Storage Systems using storage nodes with only local storage, which is either disk devices that are internal to each storage node. This eliminates the need for a shared SAS fabric and its complexities, but also enables using devices such as SATA solid state drives, which can help further reduce cost or NVMe solid state devices to improve performance. Storage Spaces Direct leverages SMB3 for all intra-node communication, including SMB Direct and SMB Multichannel, for low latency and high throughput storage. This guide presents S2D performance results using Chelsio iWARP RDMA technology in a **hyper-converged** deployment scenario.

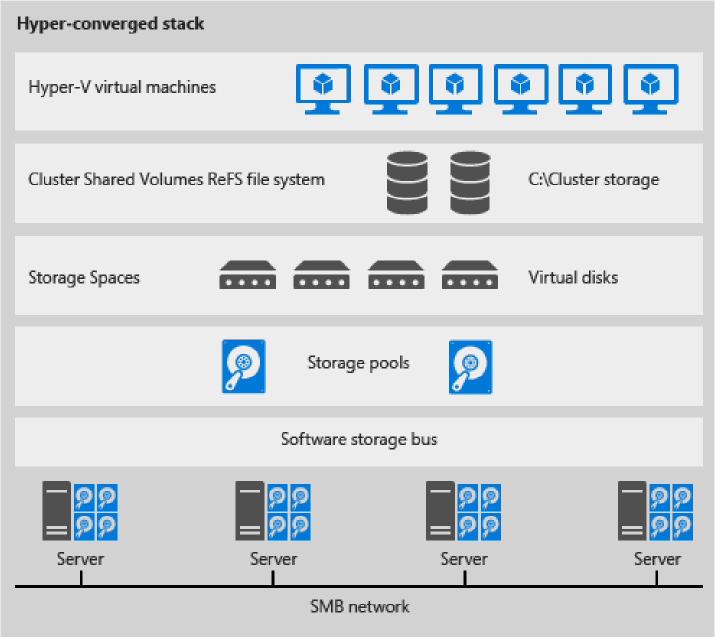


Figure 1 – S2D Hyper-Converged Stack

Why Chelsio iWARP RDMA Solution for S2D

Chelsio fifth generation, high performance RDMA 10/40Gbps 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” – see Microsoft blog Hardware options for evaluating S2D.
- 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.
    - Saves significant CPU cycles.
        - Enables cheaper CPU’s for equivalent performance.
        - Enables significantly lower datacenter and utilities.
- 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.

Following table shows a high level difference between iWARP and RoCE v2 (another RDMA over Ethernet alternatives).

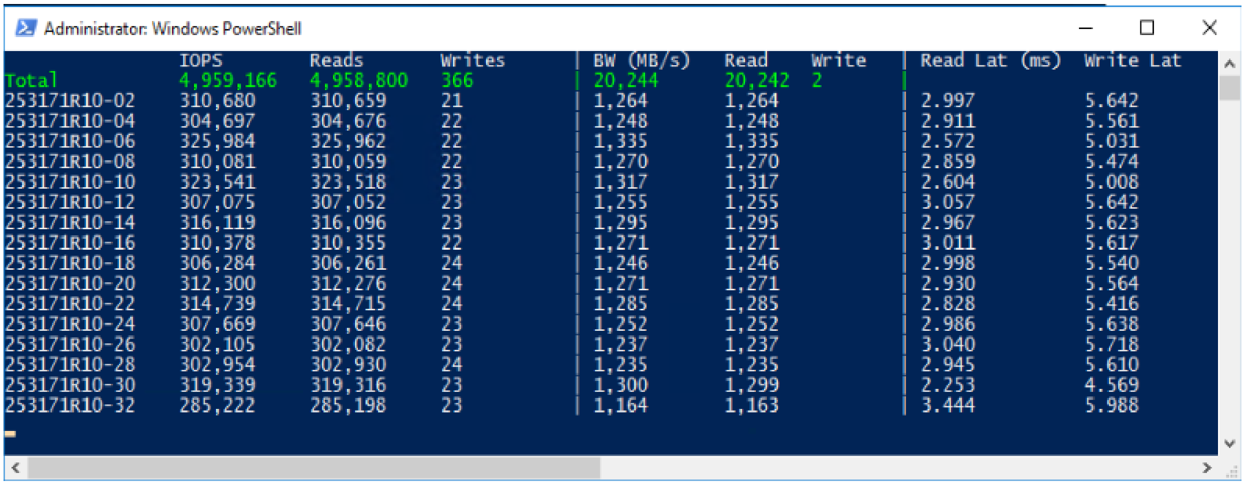
iWARP	RoCE
Native TCP/IP over Ethernet, no different from NFS or HTTP	Difficult to install and configure - “needs a team of experts” - **Plug-and-Debug**
Works with **ANY Ethernet Switch**	Requires DCB - **expensive** infrastructure equipment upgrade
Works with **ALL Ethernet equipment**	Poor **interoperability** - may not work with switches from different vendors
No need for special configuration - **TRUE Plug-And-Play**	**Fixed** QoS configuration - Data Center Bridging (DCB) must be setup identically across all switches
TCP/IP allows reach to cross **server racks and cloud scale**	Does not **scale** - requires Priority Flow Control (PFC), limited to **single subnet, intra-server rack communications**
No distance limitations. Ideal for **remote** communications and Windows DR support out-of-the-box	Short **distance** - PFC range is limited to a few hundred meters’ maximum - Requires Metro Router for Windows
WAN **routable**, users and IP Infrastructure	RoCEv1 **not routable**. RoCEv2 requires lossless IP infrastructure and restricts router configuration

Storage Spaces Direct, Storage IOPS Performance with iWARP

This demonstration was published by the Microsoft team (**Storage IOPS Update with S2D - Microsoft Blog**) and utilizes a 16-node Storage Spaces Direct hyper-converged configuration, NVMe SSD technology, and DRAM connected across an iWARP RDMA-enabled cluster, attached to a 32 port Cisco 3132 switch. Each node was equipped with the following hardware:

- 2x Xeon E5-2699v4 2.3Ghz (22c44t)
- 128GB DRAM
- 4x 800GB Intel P3700 NVMe (PCIe 3.0 x4)
- 1x LSI 9300 8i
- 20x 1.2TB Intel S3610 SATA SSD
- 1x Chelsio T580-CR (Dual Port 40Gb PCIe 3.0 x8)
    - Chelsio Unified Wire driver v5.0.0.62
    - Chelsio FW v1.16.1.0
    - Dual port connected / adapter

A 44 virtual machines per node, for a total of 704 virtual machines configuration was used for this setup. Each virtual machine was configured with 1vCPU and VMFleet tool was used to run DISKSPD in each of the virtual machines with 1 thread, 4KiB random read with 32 outstanding IO.



```
              IOPS       Reads      Writes     | BW (MB/s)   Read     Write    | Read Lat (ms)   Write Lat
Total         4,959,166  4,958,800  366        | 20,244      20,242   2        |
253171R10-02  310,680    310,659    21         | 1,264       1,264             | 2.997           5.642
253171R10-04  304,697    304,676    22         | 1,248       1,248             | 2.911           5.561
253171R10-06  325,984    325,962    22         | 1,335       1,335             | 2.572           5.031
253171R10-08  310,081    310,059    22         | 1,270       1,270             | 2.859           5.474
253171R10-10  323,541    323,518    23         | 1,317       1,317             | 2.604           5.008
253171R10-12  307,075    307,052    23         | 1,255       1,255             | 3.057           5.642
253171R10-14  316,119    316,096    23         | 1,295       1,295             | 2.967           5.623
253171R10-16  310,378    310,355    22         | 1,271       1,271             | 3.011           5.617
253171R10-18  306,284    306,261    24         | 1,246       1,246             | 2.998           5.540
253171R10-20  312,300    312,276    24         | 1,271       1,271             | 2.930           5.564
253171R10-22  314,739    314,715    24         | 1,285       1,285             | 2.828           5.416
253171R10-24  307,669    307,646    23         | 1,252       1,252             | 2.986           5.638
253171R10-26  302,105    302,082    23         | 1,237       1,237             | 3.040           5.718
253171R10-28  302,954    302,930    24         | 1,235       1,235             | 2.945           5.610
253171R10-30  319,339    319,316    23         | 1,300       1,299             | 2.253           4.569
253171R10-32  285,222    285,198    23         | 1,164       1,163             | 3.444           5.988
```

Figure 3 – Storage Space Direct IOPS Numbers

As you can see from the above screenshot, this setup was able to demonstrate ~5M IOPS in aggregate into the virtual machines. This delivers ~7,000 IOPS per virtual machine!

“***This technology demonstration highlights how Storage Spaces Direct, combined with advanced flash-based storage connected by Chelsio’s 40GbE iWARP networking solution, helps users build faster, easy-to-scale and reliable storage for their private cloud deployments***,” said Erin Chapple, partner director of program management, enterprise cloud group, Microsoft Corp.

Summary

Chelsio RDMA enabled 40Gb Ethernet adapters deliver a high performance Storage Spaces Direct (S2D) solution using standard Ethernet infrastructure and enables datacenters to deploy S2D now by leveraging all-inboxed drivers with Chelsio Ethernet adapters. 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. In addition to Storage Spaces Direct, iWARP Protocol also powers other aspects of Microsoft Windows installations such as Storage Replica for disaster recovery, SMB Direct for high performance file access, Client RDMA for bringing RDMA benefits to Windows 10 deployments, and Network Direct for Windows HPC deployments.

Related Links

Storage IOPS Update with S2D - Microsoft Blog
5M IOPS with Storage Spaces Direct - Microsoft Tweet
ClearPointe case study using Chelsio iWARP RDMA Adapters
Windows Server 2016 Storage Spaces Direct
Configuring Storage Spaces Direct