Azure Stack HCI Configuration using iWARP RDMA

Quick Start Guide for Windows

Overview

Chelsio's fifth/sixth generation (T5/T6), high performance iWARP RDMA 10/25/40/50/100GbE adapters enable incremental, non-disruptive server installs, and support the ability to work with any standard Ethernet switch, delivering a brownfield strategy to enable high performance, low cost, scalable Azure Stack Hyper Converged Infrastructure (HCI) deployments. Major benefits include cost savings on switches at higher speeds with each deployment. Windows SMB Direct over iWARP RDMA provides higher performance by giving direct access to the data residing on hyper-converged or disaggregated storage, while the CPU reduction enables a larger number of VMs per Hyper-V server, enabling savings in power dissipation, system configuration and deployment scale throughout the life of the installation. They prove to be a best fit for both networking and virtualization requirements, as well as hyper-converged scalable storage solutions like Storage Spaces Direct (S2D), a core storage feature for Azure Stack solution. This document provides quick steps to configure S2D using Chelsio iWARP RDMA adapters.

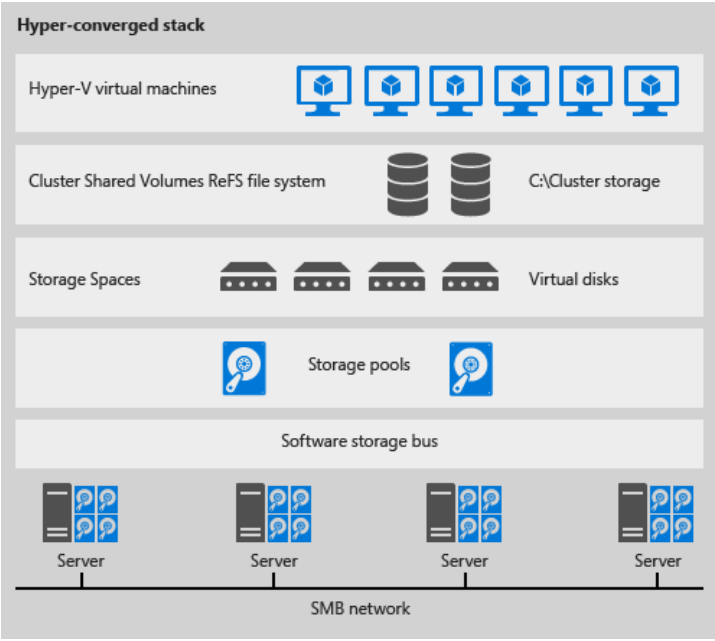


Figure 1 – Azure Stack HCI S2D

Chelsio iWARP RDMA Solution for Azure Stack HCI

iWARP has been an IETF standard (RFC 5040) since 2008, TCP/IP has been an IETF standard (RFC 793, 791) since 1982. iWARP inherits the loss resilience and congestion management from underlying TCP/IP stack and enables a very high performance, extremely low latency, high

bandwidth and high message rate solution. iWARP presents no surprises, no fine print, and is a plug and play solution. It is scalable to wherever the datacenter can scale to.

Network QoS is used in HCI configurations to ensure that the Software-Defined-Storage system has enough bandwidth to communicate between the nodes to ensure resiliency and performance. Chelsio's iWARP RDMA enabled Unified Wire Ethernet adapters with enhanced rate-limiting (network QoS) features offload bandwidth allocation to the adapter bypassing the operating system. This eliminates the need for a DCB enabled Ethernet switch and configuring complex DCB, ETS, PFC, ECN etc, resulting in reduced total ROI and simplified management.

Microsoft also recommends and prefers to use iWARP RDMA as it is easier to configure/setup, scalable, routable and works with any standard ethernet switches.

- Microsoft recommends iWARP for S2D
- Microsoft Recommendation on the RDMA alternatives in Windows
- Hyper-converged solution using Storage Spaces Direct in Windows Server 2016

Configuration

Follow the steps mentioned below to configure S2D in an Azure Stack HCI environment:

1. Install Chelsio adapters on all the nodes in PCI Gen 3 x8 or x16 slots.
2. Connect all the ports of the Chelsio adapters to a Switch.

   Note: Please refer Switch Configuration section for sample configurations.

3. Install Windows Server 2016/2019 on all cluster nodes.
4. Install Hyper-V and Failover cluster roles on all the nodes.
5. Set the following Registry entries for Server 2019 and reboot the nodes to enable S2D.

```
[HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\ClusSvc\Parameters]
"S2D"=dword:1

[HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\NetworkController]
"Enabled"=dword:1
```

6. Add the Nodes to a domain.
7. Install latest Chelsio Unified Wire Driver from Chelsio Download Center on all the nodes.

   Note: Please refer to the Support documents of Unified Wire packages for detailed steps on installation.

8. RDMA will be enabled by default for the Chelsio Ports. Verify using the following command.

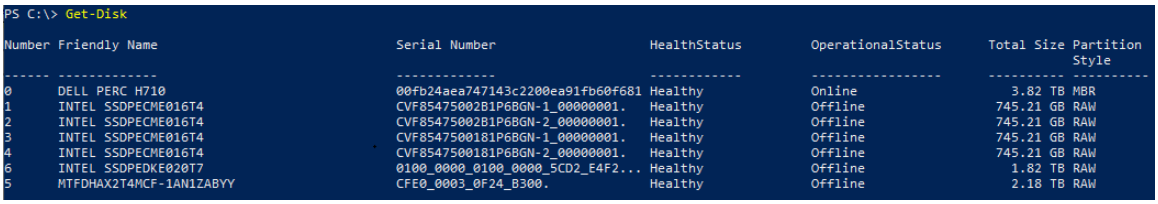
```
PS C:\> Get-NetAdapterRdma
```

```
PS C:\> Get-NetAdapterRdma

Name                    InterfaceDescription                Enabled    PFC        ETS
----                    --------------------                -------    ---        ---
ch--p1                  Chelsio Network Adapter #10         True       False      False
ch--p0                  Chelsio Network Adapter #9          True       False      False
```

9. The disks intended to be used for S2D need to be empty and without partitions or other data. If a disk has partitions or other data, it will not be included in the S2D system. Check the status of all the disks using the below command.

```
PS C:\> Get-Disk
```



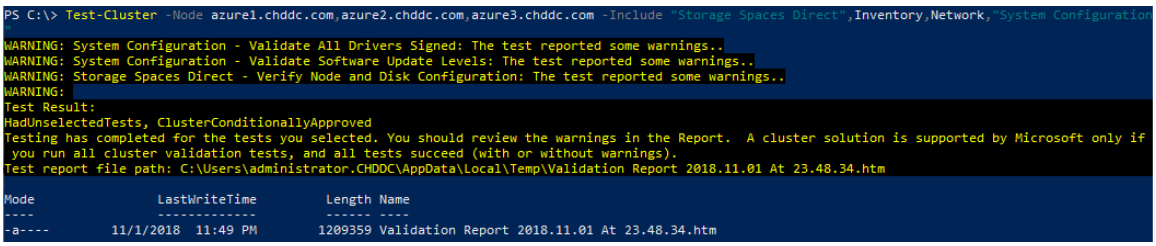
```
PS C:\> Get-Disk

Number Friendly Name                  Serial Number                    HealthStatus         OperationalStatus      Total Size Partition
                                                                                                                   Style
------ -------------                  -------------                    ------------         -----------------      ---------- ----------
0      DELL PERC H710                 00fb24aea747143c2200ea91fb60f681 Healthy              Online                   3.82 TB MBR
1      INTEL SSDPECME016T4            CVF85475002B1P6BGN-1_00000001.   Healthy              Offline                745.21 GB RAW
2      INTEL SSDPECME016T4            CVF85475002B1P6BGN-2_00000001.   Healthy              Offline                745.21 GB RAW
3      INTEL SSDPECME016T4            CVF8547500181P6BGN-1_00000001.   Healthy              Offline                745.21 GB RAW
4      INTEL SSDPECME016T4            CVF8547500181P6BGN-2_00000001.   Healthy              Offline                745.21 GB RAW
6      INTEL SSDPEDKE020T7            0100_0000_0100_0000_5CD2_E4F2... Healthy              Offline                  1.82 TB RAW
5      MTFDHAX2T4MCF-1AN1ZABYY        CFE0_0003_0F24_B300.             Healthy              Offline                  2.18 TB RAW
```

Note: Storage Spaces Direct does not support disks connected via multiple paths, and the Microsoft Multipath MPIO software stack.

10. Before creating the cluster, validate the nodes using the cluster validation tool.

```
PS C:\> Test-Cluster –Node <Node1,Node2,...> –Include “Storage Spaces
Direct”,Inventory,Network,”System Configuration”
```



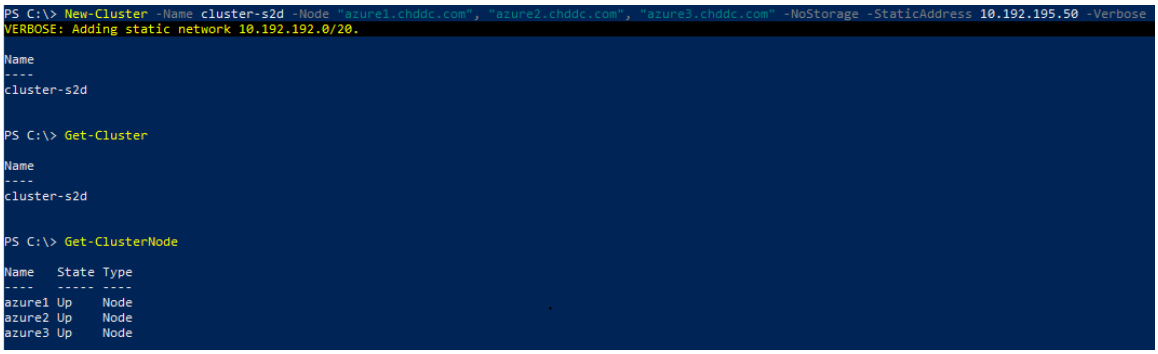
```
PS C:\> Test-Cluster -Node azure1.chddc.com,azure2.chddc.com,azure3.chddc.com -Include "Storage Spaces Direct",Inventory,Network,"System Configuration"

WARNING: System Configuration - Validate All Drivers Signed: The test reported some warnings..
WARNING: System Configuration - Validate Software Update Levels: The test reported some warnings..
WARNING: Storage Spaces Direct - Verify Node and Disk Configuration: The test reported some warnings..
WARNING:
Test Result:
HadUnselectedTests, ClusterConditionallyApproved
Testing has completed for the tests you selected. You should review the warnings in the Report.  A cluster solution is supported by Microsoft only if
 you run all cluster validation tests, and all tests succeed (with or without warnings).
Test report file path: C:\Users\administrator.CHDDC\AppData\Local\Temp\Validation Report 2018.11.01 At 23.48.34.htm

Mode                LastWriteTime         Length Name
----                -------------         ------ ----
-a----        11/1/2018  11:49 PM        1209359 Validation Report 2018.11.01 At 23.48.34.htm
```

11. Create a cluster using the nodes validated in the previous step.

```
PS C:\> New-Cluster –Name <ClusterName> –Node <Node1,Node2,...> –NoStorage -
StaticAddress <static_ip> -Verbose
```



```
PS C:\> New-Cluster -Name cluster-s2d -Node "azure1.chddc.com", "azure2.chddc.com", "azure3.chddc.com" -NoStorage -StaticAddress 10.192.195.50 -Verbose
VERBOSE: Adding static network 10.192.192.0/20.

Name
----
cluster-s2d


PS C:\> Get-Cluster

Name
----
cluster-s2d


PS C:\> Get-ClusterNode

Name   State Type
----   ----- ----
azure1 Up    Node
azure2 Up    Node
azure3 Up    Node
```

Note: Without the *–NoStorage* parameter, the disks may be automatically added to the cluster and you will need to remove them before enabling S2D. Otherwise they will not be included in the S2D pool.

12. Enable S2D and create a storage pool.

```
PS C:\>Enable-ClusterS2D -CacheState <State> -Verbose
```

```
PS C:\> Enable-ClusterS2D -CacheState Disabled -SkipEligibilityChecks -Verbose
VERBOSE: 2018/11/01-23:58:56.414 Ensuring that all nodes support S2D
VERBOSE: 2018/11/01-23:58:56.439 Querying storage information
VERBOSE: 2018/11/01-23:58:56.806 Sorted disk types present (fast to slow): NVMe. Number of types present: 1
VERBOSE: 2018/11/01-23:58:56.807 Checking that nodes support the desired cache state

Confirm
Are you sure you want to perform this action?
Performing operation 'Enable Cluster Storage Spaces Direct' on Target 'cluster-s2d'.
[Y] Yes  [A] Yes to All  [N] No  [L] No to All  [S] Suspend  [?] Help (default is "Y"): A
VERBOSE: 2018/11/01-23:59:10.841 Creating health resource
VERBOSE: 2018/11/01-23:59:11.179 Setting cluster property
VERBOSE: 2018/11/01-23:59:11.180 Setting default fault domain awareness on clustered storage subsystem
VERBOSE: 2018/11/01-23:59:11.244 Waiting until physical disks are claimed
VERBOSE: 2018/11/01-23:59:14.251 Number of claimed disks on node 'azure1': 0/6
VERBOSE: 2018/11/01-23:59:17.261 Number of claimed disks on node 'azure2': 0/6
VERBOSE: 2018/11/01-23:59:20.271 Number of claimed disks on node 'azure3': 6/6
VERBOSE: 2018/11/01-23:59:23.280 Number of claimed disks on node 'azure1': 6/6
VERBOSE: 2018/11/01-23:59:26.289 Number of claimed disks on node 'azure2': 6/6
VERBOSE: 2018/11/01-23:59:26.298 Node 'azure1': Waiting until cache reaches desired state (HDD:'Disabled' SSD:'Disabled')
VERBOSE: 2018/11/01-23:59:26.302 SBL disks initialized in cache on node 'azure1': 6 (6 on all nodes)
VERBOSE: 2018/11/01-23:59:27.307 Node 'azure2': Waiting until cache reaches desired state (HDD:'Disabled' SSD:'Disabled')
VERBOSE: 2018/11/01-23:59:27.311 SBL disks initialized in cache on node 'azure2': 6 (12 on all nodes)
VERBOSE: 2018/11/01-23:59:28.316 Node 'azure3': Waiting until cache reaches desired state (HDD:'Disabled' SSD:'Disabled')
VERBOSE: 2018/11/01-23:59:28.320 SBL disks initialized in cache on node 'azure3': 6 (18 on all nodes)
VERBOSE: 2018/11/01-23:59:29.324 Waiting until SBL disks are surfaced
VERBOSE: 2018/11/01-23:59:32.346 Disks surfaced on node 'azure1': 18/18
VERBOSE: 2018/11/01-23:59:32.365 Disks surfaced on node 'azure2': 18/18
VERBOSE: 2018/11/01-23:59:32.404 Disks surfaced on node 'azure3': 18/18
VERBOSE: 2018/11/01-23:59:35.655 Waiting until all physical disks are reported by clustered storage subsystem
VERBOSE: 2018/11/01-23:59:38.954 Physical disks in clustered storage subsystem: 18
VERBOSE: 2018/11/01-23:59:38.955 Querying pool information
VERBOSE: 2018/11/01-23:59:39.223 Starting health providers
VERBOSE: 2018/11/01-23:59:50.357 Checking that all disks support the desired cache state
VERBOSE: 2018/11/01-23:59:50.401 Required steps for this action completed successfully

Node   EnableReportName
----   ----------------
azure1 C:\Windows\Cluster\Reports\EnableClusterS2D on 2018.11.01-23.59.50.htm
```

13. Create virtual disks on the storage pool created.

```
PS C:\Users\Administrator>Get-ClusterNode |% { New-Volume -
StoragePoolFriendlyName s2d -FriendlyName $_ -FileSystem CSVFS_ReFS -Size
500GB -Verbose  }
PS C:\Users\Administrator>New-Volume -StoragePoolFriendlyName s2d -
FriendlyName Collect -FileSystem CSVFS_ReFS -Size 100GB -Verbose
```

14. Create or deploy VMs. The VM's files should be stored on the virtual disks.

Switch Configuration

Chelsio iWARP RDMA does not require any configuration of DCB, PFC, ETC, ECN etc. on the Switch. It is recommended to disable them and enable regular flow control on switch ports. The following section shows the sample configuration on few switches.

Dell/Force 10 S4810 Switch

```
Force10#configure
Force10(conf)#no dcb enable
Force10(conf)#interface fortyGigE 0/48
Force10(conf-if-fo-0/48)#flowcontrol rx on tx on
Force10(conf-if-fo-0/48)#shutdown
Force10(conf-if-fo-0/48)#no shutdown
```

Dell EMC S5148F-ON

```
OS10# configure terminal
OS10(config)# interface ethernet 1/1/6
OS10(conf-if-eth1/1/6)# no lldp transmit
OS10(conf-if-eth1/1/6)# no lldp receive
OS10(conf-if-eth1/1/6)# no priority-flow-control
OS10(conf-if-eth1/1/6)# flowcontrol receive on
OS10(conf-if-eth1/1/6)# flowcontrol transmit on
OS10(conf-if-eth1/1/6)# shutdown
OS10(conf-if-eth1/1/6)# no shutdown
```

Cisco Nexus 5010

```
ciscoswitchcert2# configure
ciscoswitchcert2(config)# interface ethernet 1/1
ciscoswitchcert2(config-if)# no lldp transmit
ciscoswitchcert2(config-if)# no lldp receive
ciscoswitchcert2(config-if)# priority-flow-control mode off
ciscoswitchcert2(config-if)# flowcontrol receive on
ciscoswitchcert2(config-if)# flowcontrol send on
ciscoswitchcert2(config-if)# shutdown
ciscoswitchcert2(config-if)# no shutdown
```

Mellanox 2410 Switch

```
isn2410 [standalone: master] > enable
isn2410 [standalone: master] # configure terminal
isn2410 [standalone: master] (config) # interface ethernet 1/1
isn2410 [standalone: master] (config interface ethernet 1/1) # shutdown
isn2410 [standalone: master] (config interface ethernet 1/1) # no lldp transmit
isn2410 [standalone: master] (config interface ethernet 1/1) # no lldp receive
isn2410 [standalone: master] (config interface ethernet 1/1) # no dcb priority-
flow-control mode
isn2410 [standalone: master] (config interface ethernet 1/1) # flowcontrol send
on
isn2410 [standalone: master] (config interface ethernet 1/1) # flowcontrol
receive on
isn2410 [standalone: master] (config interface ethernet 1/1) # no shutdown
```

Related Links

Deploy Storage Spaces Direct (Microsoft Deployment document)
Windows Server 2016 Converged NIC and Guest RDMA Deployment (A Microsoft Guide)
Storage Spaces Direct throughput with 100GbE iWARP (Microsoft Blog)
High Performance 25G S2D for AMD EPYC
S2D performance with Chelsio 25GbE
Axellio demos WSSD cluster with Chelsio 100GbE
Migrating to Microsoft Storage Spaces Direct
S2D Performance with Network QoS
S2D Performance with iWARP RDMA
iWARP RDMA – Best Fit for Storage Spaces Direct