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.

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 performance network. This performance is especially important for applications that require low latency and high throughput, such as gaming, streaming, and high performance computing.

Figure 1 – Azure Stack HCI S2D

![Figure 1 – Azure Stack HCI S2D](image)
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.

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.

   ```powershell
   [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](https://www.chelsio.com/products/downloads) 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.

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

<table>
<thead>
<tr>
<th>Name</th>
<th>InterfaceDescription</th>
<th>Enabled</th>
<th>PFC</th>
<th>ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch-p1</td>
<td>Chelsio Network Adapter #10</td>
<td>True</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>ch-p0</td>
<td>Chelsio Network Adapter #0</td>
<td>True</td>
<td>False</td>
<td>False</td>
</tr>
</tbody>
</table>
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
```

![Disk Status](image)

**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”
```

![Cluster Validation](image)

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
```

![Cluster Creation](image)

**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
```

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

Copyright 2019. Chelsio Communications Inc. All rights reserved.
Dell EMC S5148F-ON

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

Cisco Nexus 5010

```bash
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

```bash
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](https://docs.microsoft.com/en-us/WindowsServer/deploy-storage/deploy-storage-spaces-direct) (Microsoft Deployment document)
- [Storage Spaces Direct throughput with 100GbE iWARP](https://azure.microsoft.com/en-us/blog/storage-spaces-direct-throughput-with-100gb-e-iwarp/) (Microsoft Blog)
- [High Performance 25G S2D for AMD EPYC](https://azure.microsoft.com/en-us/blog/high-performance-25g-s2d-for-amd-epyc/)
- [S2D performance with Chelsio 25GbE](https://www.chelsio.com/blog/s2d-performance-with-chelsio-25gbe)
- [Axellio demos WSSD cluster with Chelsio 100GbE](https://www.axellio.com/products/wssd-cluster-with-chelsio-100gb-e/)
- [S2D Performance with Network QoS](https://www.chelsio.com/blog/s2d-performance-with-network-qos)
- [S2D Performance with iWARP RDMA](https://www.chelsio.com/blog/s2d-performance-with-iwarp-rdma)
- [iWARP RDMA – Best Fit for Storage Spaces Direct](https://www.chelsio.com/blog/iwarp-rdma-best-fit-for-storage-spaces-direct)