How-to: Configure Windows Server 2016 with Switch Embedded Teaming for RoCEv2 lossless fabric.

Created on Aug 8, 2019

Related Documents

- How-to: Configure Windows Server 2016 with Switch Embedded Teaming for RoCEv2 lossless fabric.

Introduction

This document provides the detailed steps to deploy Windows Server 2016 with SET technology over RoCEv2 lossless fabric.

Switch Embedded Teaming (SET) is an alternative NIC Teaming solution that you can use in environments that include Hyper-V and the Software Defined Networking (SDN) stack in Windows Server 2016. SET integrates some NIC Teaming functionality into the Hyper-V Virtual Switch. SET allows you to group between one and eight physical Ethernet network adapters into one or more software-based virtual network adapters. These virtual network adapters provide fast performance and fault tolerance in the event of a network adapter failure.

Table 1: Abbreviation

Definitions /Abbreviation	Description
RoCE	RDMA over Converged Ethernet
RoCEv2	Internet layer protocol which means that RoCE v2 packets can be routed
TOR	Top of Rack Switch
vSwitch	Hyper-V Virtual Switch
hNIC	Host vNIC – Virtual NIC from vSwitch
pNIC	Physical NIC
SET	Switch Embedded Teaming

References

HowTo Configure Mellanox Spectrum Switch for Lossless RoCE

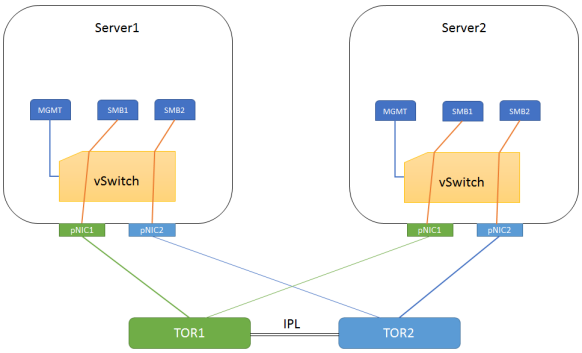
How To Configure MLAG on Mellanox Switches

HowTo Configure MAGP on Mellanox Switches

Solution Overview

Solution Logical Design

The following illustration shows an example configuration.



Deployment Guide

Network / Fabric Deployment and Configuration

Please use this document in order to configure LossLess RoCE for NVIDIA Spectrum switches.

Please use this document in order to configure IPL for NVIDIA switches.

You must configure MAGP for each VLAN ID which used in the Host vNIC adapters. Please use this document in order to configure MAGP on NVIDIA switches.

Host Deployment and Configuration

Physical Host QoS configuration

1. Turn on Data Center Bridging.

Install-WindowsFeature Data-Center-Bridging

Success Restart Needed Exit Code Feature Result

------- -------------- --------- --------------

True No Success {Data Center Bridging}

2. Set the policies for SMB-Direct

New-NetQosPolicy "SMB" -NetDirectPortMatchCondition 445 -PriorityValue8021Action 3

Name : SMB

Owner : Group Policy (Machine)

NetworkProfile : All

Precedence : 127

JobObject :

NetDirectPort : 445

PriorityValue : 3

3. Set policies for other traffic on the interface

New-NetQosPolicy "DEFAULT" -Default -PriorityValue8021Action 0

Name : DEFAULT

Owner : Group Policy (Machine)

NetworkProfile : All

Precedence : 127

Template : Default

JobObject :

PriorityValue : 0

4. Turn on Flow Control for SMB

Enable-NetQosFlowControl -priority 3

Get-NetQosFlowControl

Priority	Enabled	PolicySet	IfIndex	IfAlias
0	False	Global		
1	False	Global		
2	False	Global		
3	True	Global		
4	False	Global		
5	False	Global		
6	False	Global		
7	False	Global		

5. Disable traffic FlowControl for classes other than 3

Disable-NetQosFlowControl -priority 0,1,2,4,5,6,7

Get-NetQosFlowControl

Priority	Enabled	PolicySet	IfIndex	IfAlias
0	False	Global		
1	False	Global		
2	False	Global		
3	True	Global		
4	False	Global		
5	False	Global		
6	False	Global		
7	False	Global		

Create a vSwitch in SET mode and hNICs

1. Create a vSwitch in Switch Embedded Teaming mode

New-VMSwitch –Name "VSWSET" –NetAdapterName "pNIC1","pNIC2" -EnableEmbeddedTeaming $true -AllowManagementOS $true

Name SwitchType NetAdapterInterfaceDescription

---- ---------- ----------

VSWSET External Teamed-Interface

2. List the Physical adapters in logical switch with SET

Get-VMSwitchTeam -Name VSWSET | FL

Name : VSWSET

Id : bff9d6c4-4259-4db2-8432-4870117f0da3

NetAdapterInterfaceDescription : {Mellanox ConnectX-4 VPI Adapter, Mellanox ConnectX-4 VPI Adapter #2}

TeamingMode : SwitchIndependent

LoadBalancingAlgorithm : Dynamic

3. To prevent auto-tagging the egress traffic with incorrect VLAN ID from both physical NIC please remove the ACCESS VLAN Setting and filtering ingress traffic which doesn't match the ACCESS VLAN ID

Set-NetAdapterAdvancedProperty -Name "pNIC1" -RegistryKeyword VlanID -RegistryValue "0"

Set-NetAdapterAdvancedProperty -Name "pNIC2" -RegistryKeyword VlanID -RegistryValue "0"

4. Creating the Management NIC in order to use separate Host vNICs instances for RDMA and set VLANID 101

Rename-VMNetworkAdapter -ManagementOS -Name "VSWSET" -NewName "MGMT"

Set-VMNetworkAdapterVlan -VMNetworkAdapterName "MGMT" -VlanId "101" -Access -ManagementOS

5. Set Priority tagging on the Management Host vNIC

Set-VMNetworkAdapter -ManagementOS -Name "MGMT" -IeeePriorityTag on

6. Creating two Host vNIC for RDMA and set VLANID 102

Add-VMNetworkAdapter –SwitchName "VSWSET" –Name SMB1 –ManagementOS

Add-VMNetworkAdapter –SwitchName "VSWSET" –Name SMB2 –ManagementOS

Get-VMNetworkAdapter –ManagementOS

Name InterfaceDescription ifIndex Status MacAddress LinkSpeed

---- ---------- ------- ------ ---------- ---------

vEthernet (SMB1) Hyper-V Virtual Ethernet Adapter #1 6 Up 00-1D-D8-B7-1C-11 200 Gbps

vEthernet (SMB2) Hyper-V Virtual Ethernet Adapter #2 18 Up 00-1D-D8-B7-1C-0E 200 Gbps

vEthernet (MGMT) Hyper-V Virtual Ethernet Adapter 16 Up 00-1D-D8-B7-1C-01 200 Gbps

Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB1" -VlanId "102" -Access -ManagementOS

Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB2" -VlanId "102" -Access –ManagementOS

Get-VMNetworkAdapterVlan -ManagementOS

VMName VMNetworkAdapterName Mode VlanList

------ -------------------- ---- --------

SMB1 Access 102

MGMT Access 101

SMB2 Access 102

7. Because vNICs dedicated to storage are bound to a SET we need create an affinity between a vNIC and a pNIC ensures that the traffic from a given vNIC on the host (storage vNIC) uses a particular pNIC to send traffic so that it passes through the shorter path

Set-VMNetworkAdapterTeamMapping -ManagementOS -SwitchName VSWSET -VMNetworkAdapterName "SMB1" -PhysicalNetAdapterName "pNIC1"

Set-VMNetworkAdapterTeamMapping -ManagementOS -SwitchName VSWSET -VMNetworkAdapterName "SMB2" -PhysicalNetAdapterName "pNIC2"

8. Set Priority tagging (PCP mode) on the RDMA Host vNIC and enable RDMA on this vNICs

Set-VMNetworkAdapter -ManagementOS -Name "SMB1" -IeeePriorityTag on

Set-VMNetworkAdapter -ManagementOS -Name "SMB2" -IeeePriorityTag on

Enable-NetAdapterRdma -Name "vEthernet (SMB1)"

Enable-NetAdapterRdma -Name "vEthernet (SMB2)"

Performance Testing

Please assign an IP address to each RDMA enabled Host vNICs and provides testing with TEST-RDMA.PS1 PowerShell script. Please see below script execution examples:

1. With RDMA enabled (parameter **-IsRoCE $true**)

C:\TESTS\Test-RDMA.PS1 -IfIndex 6 -IsRoCE $true -RemoteIpAddress 192.168.2.2 -PathToDiskspd C:\TESTS\Diskspd-v2.0.17\amd64fre

2. With RDMA disabled (parameter **-IsRoCE $False**)

C:\TESTS\Test-RDMA.PS1 -IfIndex 6 -IsRoCE $False -RemoteIpAddress 192.168.2.2 -PathToDiskspd C:\TESTS\Diskspd-v2.0.17\amd64fre

Done.