Dell EMC Networking – RDMA over Converged Ethernet (RoCE v2) Cheat Sheet

RoCE v2 Configuration cheat sheet with Dell EMC OS10

Abstract

Detailed RoCEv2 configuration with Dell EMC networking and OS10

August 2018

Revisions

Date	Description
August 2018	Initial release

Acknowledgements

This paper was produced by the following members of the Dell EMC Networking Technical Marketing engineering team:

Author: Mario Chow

Other:

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

© Aug 2018 Dell Inc. or its subsidiaries. All Rights Reserved. Dell, EMC, Dell EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

Dell believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Table of contents

Revisions 2

Acknowledgements 2

1 Introduction 4

1.1 Objective 4

1.2 Audience 4

2 Setup 5

2.1 Spine Layer – (S4148Fs) 6

2.2 Leaf Layer – (S4048-3, S4048-4, S4148U) 6

3 Configuration details 7

3.1 Spine – S4148Fs 7

3.2 S4148F (OS10) 7

3.3 Leaf – S4048-3, S4048-4, S4148U 10

3.4 S4048-ON-3 (OS9) 10

3.5 S4048-ON-4 & S4148U (OS10) 11

1 Introduction

Your typical next generation data center is a dynamic and scalable asset that every enterprise is learning to harness to achieve a competitive edge. It is a convergence point of multiple types of workloads and technologies that together deliver business applications,

1.1 Objective

The objective of this document is to provide a set of valid Dell EMC networking configurations that apply to a converged environment consisting of RDMA, iSCSI, and LAN (Local Area Network) traffic.

1.2 Audience

The intended audience of this document is a network administrator, architect, or sales engineer. The information on this document is meant to be used as a reference. It is not meant to be an exhaustive deployment document covering all possible configurations.

2 Setup

Figure 1, describes the reference test environment used to deploy a converged data center. The test environment uses a distributed Spine-Leaf architecture, where the leaf layer acts as both Leaf and ToR (Top of Rack).

The connected servers to the Leaf/ToR switches simulate a rack within a data center. Each server is simulating FCoE, and RDMA traffic, however, for the sake of this document, only RDMA traffic is highlighted.

Besides iSCSI and LAN traffic, RDMA traffic is also being generated by the IXIA traffic generator to simulate a typical converged environment.

The following items were used to test RDMA, iSCSI, and LAN:

- Dell EMC Networking Layer 2/3 switches supporting DCB such as:
  - S4048-ON and S4148F-ON
- Dell EMC OS10.4.1, OS9.13.0
- QLogic – QL41254HLCU – Quad port 25GbE SFP+
- Windows 2016 Server
- IOMeter
- IMDisk
- IXIA – XG12, 10/40GE Modules
- Dell Poweredge R630s

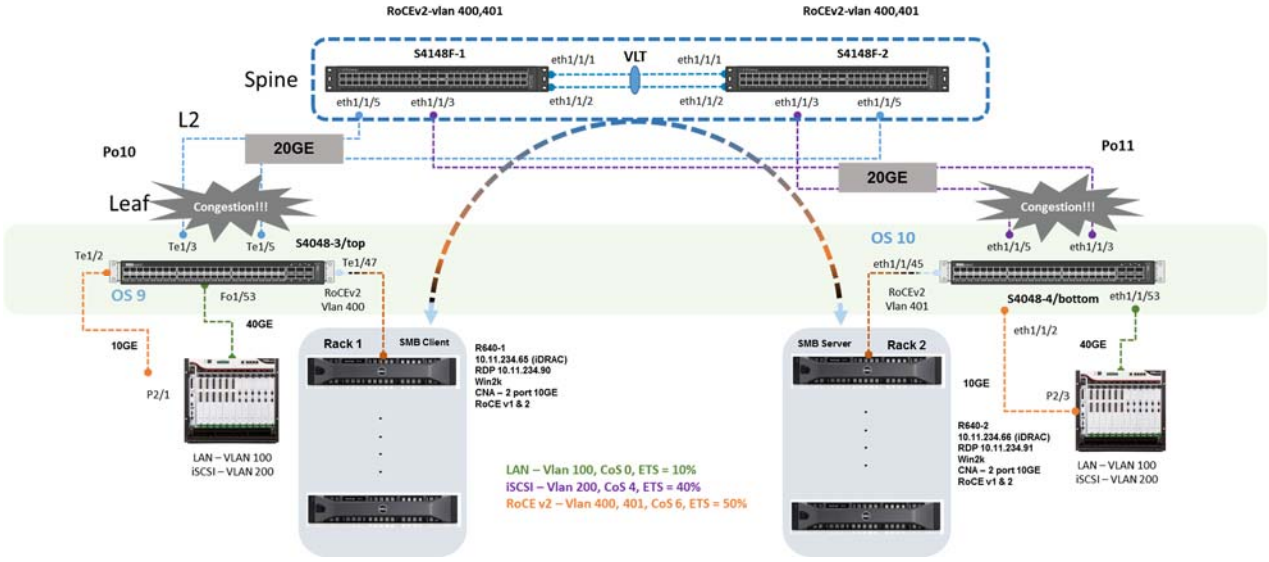
The following traffic parameters were used to ensure RDMA workload over iSCSI and LAN data traffic:

- LAN data traffic
  - 802.1p priority CoS = 0, vlan 100
  - ETS = 10%, PFC off
- iSCSI data traffic
  - 802.1p priority CoS = 4, vlan 200
  - ETS = 40%, PFC on
- RDMA
  - 802.1p priority CoS = 6 (RDMA v2), vlan 400
  - ETS = 50%, PFC on

Note: **RDMA v1 and v2 will be tested separately.**

RoCE traffic is crossing the spine layer via the congested 20GE port channels.

Figure 1 Dell EMC Networking RoCE v2 Reference Environment.



To demonstrate how Dell EMC networking ensures quality of service in a converged environment, congestion is created on the links between the leaf and spine (port-channels 10 & 20). A port-channel sized at 20GbE is configured between the leaf and spine interlinks.

Through the use of an IXIA traffic generator, LAN and iSCSI traffic are simulated. LAN traffic is running at 40Gbps, and iSCSCI is running at 10Gbps congesting the inter-links between the leaf and spine switches. RDMA traffic is then added to this congested set of links, through the use of simulated RDMA traffic between the servers using SMB 3.0 (fully integrated into Windows 2016 Server) in conjunction with ImDisk and IOmeter.

2.1 Spine Layer – (S4148Fs-ON)

The spine layer consists of two S4128Fs terminating all Layer 2 and Layer 3 traffic. Dell EMC's Virtual Link Trunk (VLT) is deployed creating a virtual switch at the spine layer. The connections between the spine and leaf are purposely congested to demonstrate end-to-end lossless connectivity for RoCE v1 data traffic.

2.2 Leaf Layer – (S4048-ON-3, S4048-ON-4)

At the leaf layer, two S4048s are running OS9 and OS10. The setup shows how an OS9 based system integrates well with the newer OS10 system.
The Leaf switches act as ToR (Top of Rack) switches providing end device connectivity to the fabric.

3 Configuration details

The following section depicts in full detail the entire switch configuration for both the spines and leaf. For brevity, only one spine configuration is shown. The configuration is repeated for the other spine.

3.1 Spine – S4148Fs

3.2 S4148F (OS10)

1. Configure VLT
    a. **Switch#** conf
    b. **Switch(conf)#** vlt domain 1
    c. **Switch(conf-vlt-1)#** discovery-interface ethernet1/1/1-1/1/2
    d. **Switch(conf-vlt-1)#** end
    e. **Switch#**

2. Enable lldp
    a. **Switch#** conf
    b. **Switch(conf)#** lldp enable
    c. **Switch(conf-lldp)#** end
    d. **Switch#**

3. Enable DCBX
    a. **Switch#** conf
    b. **Switch(conf)#** dcbx en
    c. **Switch#**

4. Turn on ETS
    a. **Switch(config)#** system qos
    b. **Switch(config-sys-qos)#** ets mode on
    c. **Switch(config-sys-qos)#** end
    d. **Switch#**

5. Configure the qos-map to match the queues to the CoS values, 0 (LAN), 4 (iSCSI), and 5 (RDMA)
    a. **Switch#** conf t
    b. **Switch(config)#** qos-map traffic-class qmap
    c. **Switch(config-qos-map)#** queue 0 qos-group 0-3,6-7
    d. **Switch(config-qos-map)#** queue 5 qos-group 5
    e. **Switch(config-qos-map)#** queue 4 qos-group 4

f. **Switch(config-qos-map)#** end
g. **Switch#**

6. Start configuring the class maps to be used with all respective policy maps – first one "queueing".
   6a. **Queueing –** This queues refer to the queues configured in step 3.
   a. **Switch#** conf t
   b. **Switch(config)#** class-map type queueing q0
   c. **Switch(config-cmap-queuing)#** match queue 0
   d. **Switch(config-cmap-queuing)#** exit
   e. **Switch(config)#** class-map type queuing q4
   f. **Switch(config-cmap-queuing)#** match queue 4
   g. **Switch(config-cmap-queuing)#** exit
   h. **Switch(config)#** class-map type queuing q5
   i. **Switch(confg-cmap-queuing)#** match queue 5
   j. **Switch(config-cmap-queuing)#** exit
   k. **Switch(config)#**

   6b**. Network-qos** – This class map is used by the service policy input and applied as ingress traffic. (see "Interface configuration" in step 8)
   a. **Switch#** conf t
   b. **Switch(config)#** class-map type network-qos iSCSI
   c. **Switch(config-cmap-nqos)#** match qos-group 4
   d. **Switch(config-cmap-nqos)#** exit
   e. **Switch(config-cmap)#** class-map type network-qos RoCEv1
   f. **Switch(config-cmpa-nqos)#** match qos-group 5
   g. **Switch(config-cmap-nqos)#** end
   h. **Switch#**

7. Start configuring the relevant policy maps which will use the class-maps defined in steps 3-4 and self-generated class "class-trust".
   a. **Switch#** conf t
   b. **Switch(config)#** policy-map type qos trust_dot1p
   c. **Switch(config-pmap-qos)#** class class-trust
   d. **Switch(config-pmap-c-qos)#** trust dot1p
   e. **Switch(config-pmap-c-qos)#** end
   f. **Switch#**

   7a**. ETS policy map –** Assigning bandwidth percentage to each queue.
   a. **Switch#** conf t
   b. **Switch(config)#** policy-map type queuing q1

c. **Switch(config-pmap-queuing)#** class q0
d. **Switch(config-pmap-c-que)#** bandwidth percent 10
e. **Switch(config-pmap-c-que)#** exit
f. **Switch(config-pmap-queuing)#** class q4
g. **Switch(config-pmap-c-que)#** bandwidth percent 40
h. **Switch(config-pmap-c-que)#** exit
i. **Switch(config-pmap-queuing)#** class q5
j. **Switch(config-pmap-c-que)#** bandwidth percent 50
k. **Switch(config-pmap-c-que)#** end
l. **Switch#**

7b. **PFC policy map –** Policy map turning pfc "ON" on CoS 4 and 5 or iSCSI and RDMA flows respectively.

a. **Switch#** conf t
b. **Switch(config)#** policy-map type network-qos PFC_ON
c. **Switch(config-pmap-network-qos)#** class iSCSI
d. **Switch(config-pmap-c-nqos)#** pause
e. **Switch(config-pmap-c-nqos)#** pfc-cos 4
f. **Switch(config-pmap-c-nqos)#** exit
g. **Switch(config-pmap-network-qos)#** class RoCEv1
h. **Switch(config-pmap-c-nqos)#** pause
i. **Switch(config-pmap-c-nqos)#** pfc-cos 5
j. **Switch(config-pmap-c-nqos)#** end
k. **Switch#**

8. **Applying policy maps to interfaces**
   a. **Switch#** conf t
   b. **Switch(config#** inte ethe1/1/5 – interface part of port channel carrying all flows
   c. **Switch(config-if-eth1/1/5)#** switchport mode trunk
   d. **Switch(config-if-eth1/1/5)#** switchport trunk allowed vlan 100,200,300 – LAN, iSCSI, and RDMA traffic allowed in this interface.
   e. **Switch(config-if-eth1/1/5)#** service-policy input type network-qos PFC_ON
      Service policy that turns on PFC on CoS = 4 and 5. Policy map in step 5b.
   f. **Switch(config-if-eth1/1/5)#** service-policy input type qos trust_dot1p
      QoS policy that trusts the system's dot1p settings
   g. **Switch(config-if-eth1/1/5)#** service-policy output type queuing q1
      Egress service policy applying ETS on output queues defined in step 5a.
   h. **Switch(config-if-eth1/1/5)#** ets mode on
      Turns "ON" ETS
   i. **Switch(config-if-eth1/1/5)#** qos-map traffic-class qmap
      Assigns the system queues to the proper class of service value on ingress
   j. **Switch(config-if-eth1/1/5)#** priority-flow-control mode on

PFC is turned "ON" at the interface

k. **Switch(config-if-eth1/1/5)#** end
l. **Switch#**

3.3 Leaf – S4048-ON-3, S4048-ON-4

3.4 S4048-ON-3 (OS9)

1. Enable lldp
    a. **Switch#** conf
    b. **Switch(conf)#** protocol lldp
    c. **Switch(conf-lldp)#** no dis
    d. **Switch(conf-lldp)#** exit
    e. **Switch(conf)#**

2. Enable dcb and assign 4 priority queues
    a. **Switch#** conf
    b. **Switch(conf)#** dcb enable pfc-queues 4
    c. **Switch(conf)#** exit
    d. **Switch#**

3. Configure DCB, PFC, ETS profiles to be used for all three traffic patterns. Priority group
    a. **Switch#** conf
    b. **Switch(conf)#** service-class dynamic dot1p
    c. **Switch(conf)#** dcb-map RoCE_and_all
    d. **Switch(conf-dcbmap-RoCE_and_all)#** priority-group 0 bandwidth 10 pfc off
    e. **Switch(conf-dcbmap-RoCE_and_all)#** priority-group 1 bandwidth 40 pfc on
    f. **Switch(conf-dcbmap-RoCE_and_all)#** priority-group 2 bandwidth 50 pfc on
    g. **Switch(conf-dcbmap-RoCE_And_all)#** priority-pgid 0 0 0 0 1 2 0 0
    h. **Switch(conf-dcbmap-RoCE_and_all)#** end
    i. **Switch#**

4. Apply DCB map profile on all respective interfaces.
    a. **Switch#** conf
    b. **Switch(conf)#** int range te1/3, te/15, te1/2, fo1/53, te1/48
    c. **Switch(conf-int-range-te1/2…fo1/53)#** no shut
    d. **Switch(conf-int-range-te1/2...fo1/53)#** dcb-map RoCE_and_all
    e. **Switch(conf-int-range-te1/2…fo1/53)#** end
    f. **Switch#**

3.5 S4048-ON-4 (OS10)

To configure the S4048-ON-4 repeat steps 2 – 8 from the S4148F section.

Note: The configurations applied on interface eth1/1/5 should be repeated for all other interfaces. However, not all configurations apply. For example, interface eth1/1/53 is only receiving LAN traffic, therefore only vlan 100 would be configured. Interface 1/1/2 is only receiving iSCSI traffic, therefore only vlan 200 would be configured on this interface.