



Dell Networking – RDMA over Converged Ethernet Technical Brief

Dell Networking RoCE Deployment Configuration Guide

Dell Networking – Data Center Technical Marketing
April 2015

Revisions (required)

| Date | Description |
|--------------|------------------------------|
| January 2015 | Initial release – Mario Chow |
| April 2015 | Added comments from team |

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2013 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

Dell, the DELL logo, and the DELL badge are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims any proprietary interest in the marks and names of others.



Table of contents

- Revisions (required) 2
- Introduction 4
- Objective 4
- Audience..... 4
- RDMA (Remote Direct Memory Access) 4
 - RoCE (RDMA over Converged Ethernet) – New Kid on the block 6
 - iWARP (Internet Wide Area RDMA Protocol)..... 6
- RDMA over Converged Ethernet (RoCE) 7
 - 802.1Qbb (PFC)..... 7
 - 802.1Qaz (ETS) 7
 - 802.1Qau (QCN) 7
 - DCBx – DCB exchange protocol 7
 - RoCE v1 and V2 8
- RoCE v1 and v2 Configuration 8
 - RoCE v1 Configuration – Pure Layer 2 solution 9
 - Dell S6K_1 configuration..... 10
 - MXL_1(3) Configuration..... 11



Introduction

More bandwidth, more processing power, faster access to critical applications, lossless data transfers; these and other increasing demands in today's enterprise network infrastructure have lead to the need for a better and more efficient way of transferring data across compute nodes in the data center.

Networking equipment vendors have realized the need to implement and deliver a solution to their customers; and an ecosystem between NIC (Network Interface Card) vendors and networking vendors have created a unique solution enhancing the necessary protocol within the data center to achieve this data transfer efficiency.

Dell Networking with its data center product family support DCB (Data Center Bridging). DCB comes with three additional features: PFC, ETS, and DCBx. These additional features enable the implementation of a granular and dedicated service that RDMA leverages and ensures its ability to deliver on the promises of speed and lossless services.

Objective

The following technical brief is intended for both internal and external consumption. It is written to provide a set of recommended step-by-step instructions used to deploy a converged data center environment based on proven practices using DCB end-to-end and its companion set of features. This document discusses RoCE v1 & 2 only.

Audience

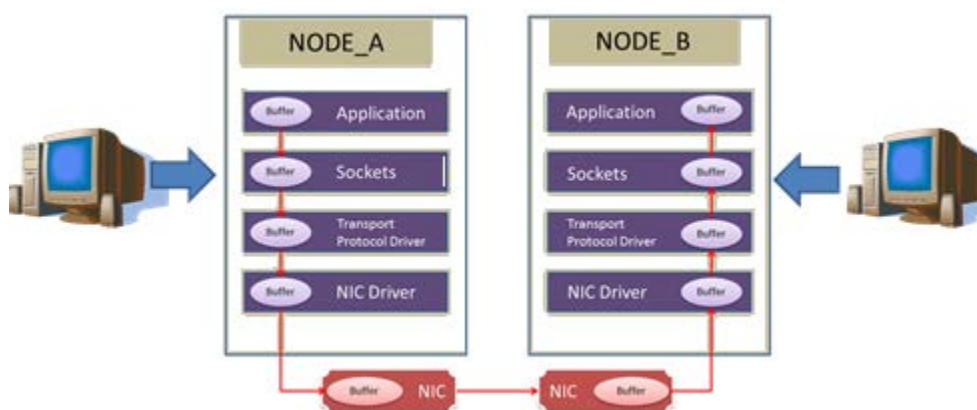
The intended audience for this document are system engineers, system administrators, network architects, or network operators that perform system design, or maintain and deploy a converged network infrastructure.

RDMA (Remote Direct Memory Access)

When two compute nodes decide to transfer data between them, the following basic steps take place (see Figure 1).



Figure 1 TCP/IP Interconnect



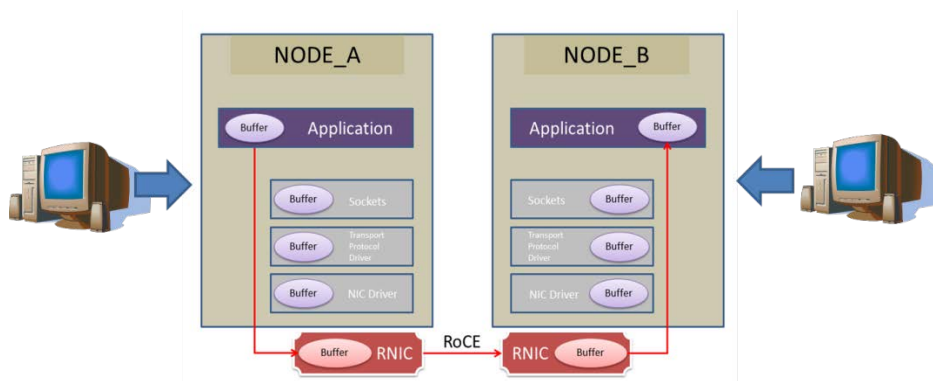
Steps 1 – 5:

- Data starts at the application layer, it is buffered
- Data passes to the socket layer, it is buffered
- Goes to the transport layer, it is buffered
- Data moves onto the NIC driver and finally hits the NIC itself.

The steps mentioned above have the disadvantages of throughput, and efficiency (latency) especially when dealing with high performance computing environments or applications. Without RDMA, multiple memory copies take place and this becomes a bottleneck in an HPC (High Performance Computing) environments primarily because the memory speeds have not kept up with the current increased rate of CPUs and interconnects. In addition to this bottleneck, there is also the extra amount of processing required for TCP/IP transfers. The typical TCP/IP network stack is implemented in the host operating system which means that the host CPU is required to perform all the necessary operations such as packet processing, checksum, and handling of system interrupts.

To address these issues, the creation of the RDMA Consortium was created; where RDMA technology enables a more efficient way of data transfer by removing the basic requirements of copies, operating system and, CPU involvement. As a result, reducing latency and increasing throughput by allowing one compute node to directly place data in another node's memory with minimal demands on memory bus bandwidth and CPU processing overhead. (see Figure 2).

Figure 2 Typical RDMA Layers



Currently, the following major NIC vendors support RDMA and similar technologies. (see Table 1).

Table 1 Network Interface Card – RoCE and iWARP support

| Vendor | Network Interface Card (model) | RoCE Support | iWARP |
|----------|--------------------------------|---------------|-----------|
| Mellanox | Mellanox ConnectX-3 Pro | Supported | |
| Emulex | OCe14000 | Supported | |
| Intel | | Not supported | Supported |
| Chelsio | | Supported | Supported |

There are two commonly known RDMA technologies that run over Ethernet:

- RoCE – RDMA over Converged Ethernet
- iWARP – Internet Wide Area RDMA Protocol (Ethernet)

RoCE (RDMA over Converged Ethernet) – New Kid on the block

RoCE as the title states, it is simply RDMA over a converged Ethernet network. It is a network protocol that provides all the benefits of RDMA technology and the familiarity of Ethernet. It highly leverages a data center feature – DCB or Data Center Bridging. With DCB, a set of complementary features such as PFC (Priority-based Flow Control), ETS (Enhanced Transmission Selection), and DCBx (Data Center Bridging eXchange) help provide a lossless service for RDMA type of applications such as high performance computing or any latency and throughput sensitive applications.

With RoCE, DCB is needed to be deployed from the host end, across the network infrastructure, to the other host end. In other words, RoCE has a network element to it and it works together.

There are two versions, 1 and 2. Version 1 is a pure Layer 2 solution, and version 2 which is still under review is a Layer 3 solution.

iWARP (Internet Wide Area RDMA Protocol)

iWARP delivers RDMA over standard based Ethernet using unchanged IP networks, and standard TCP/IP Ethernet data link services, i.e. is it connection-oriented. Unlike RoCE v1.0, iWARP packets are routable. It is not limited to just Ethernet. It is backed by the IETF standards body, while RoCE has the Infiniband Trade Association's support. Similar to RoCE, iWARP is also a creation of the RDMA Consortium.

Unlike RoCE, iWARP's based TCP/IP implementation causes it to not have any special requirements in order to be supported by any Layer 2 or Layer 3 networking devices.

A quick comparison between iWARP and RoCE is shown in table 1.

Table 2 iWARP and RoCE key comparisons

| | iWARP | RoCE |
|-----------------------|-----------------------------|----------------------------|
| Networking dependency | None | Data Center Bridging (DCB) |
| Scalability | Layer 2 and Layer 3 support | Layer 2 (v1), Layer 3 (v2) |
| Latency | Same as RoCE | Single digit micro-seconds |
| Lossless | | Yes, with DCB enabled |



RDMA over Converged Ethernet (RoCE)

Prior to RoCE, the de-facto-standard or solution for any HPC or HPC-like environment was IB. As Ethernet technology began to evolve, the gap that existed between an IB and Ethernet solution to address the requirements of an HPC/HPC-like environment has narrowed to a point where Ethernet can now be considered a viable alternative to what IB offers. With the inception of the RDMA Consortium and the advances in Ethernet, the resulting child is RoCE.

So what changed? Ethernet changed. It evolved from the basic 802.1 to what we know as 802.1Qx. RoCE and 802.1Qx work together and we cannot mention one without the other.

This evolution provides the necessary tools to address RDMA transport requirements (see Table 2).

Table 3 Ethernet Data Center Evolution

| Ethernet | 802.1 | 802.1Qx |
|-----------------------|-------|---------------------|
| Lossless | No | Yes, 802.1Qbb (PFC) |
| Traffic Class Types | No | Yes, 802.1Qaz (ETS) |
| Congestion Management | No | Yes, 802.1Qau (QCN) |
| DCB exchange protocol | No | No |

Together PFC, ETS, and QCN make up what is known as Data Center Bridging or DCB. In addition to PFC, ETS, and QCN, DCBx is another important component of what DCB offers even though it is not defined as one of the 802.1 extensions.

802.1Qbb (PFC)

802.1Qbb provides a link level flow control mechanism that can be controlled independently using the 802.1p Class of Service values ranging from 0-7. The goal of this enhancement is to ensure no packet loss due to congestion in the wire in a DCB enabled network. With this enhancement QoS (Quality of Service) over Ethernet is delivered.

802.1Qaz (ETS)

802.1Qaz continues in the spirit of delivering solid quality of service by guaranteeing a percentage of the link bandwidth to a specific traffic class. For example, voice, video, and data can be assigned a different amount of guaranteed bandwidth.

802.1Qau (QCN)

802.1Qau provides end-to-end congestion mechanism control for upper layer protocols that otherwise do not have any native built-in congestion mechanism. 802.1Qau together with Qbb and Qaz make Ethernet a "lossless" technology.

DCBx – DCB exchange protocol

DCBx is an extension of the traditional LLDP protocol where LLDP is used to discover neighboring devices and its capabilities. DCBx is an exchange configuration protocol between DCB compliant networking devices where a source and client relationship is used, to synchronize the DCB parameters such as PFC, ETS, and QCN. The traditional configuration has the source device transmitting the DCB parameters to the client device, and the client device in a *willing* mode to listen and accept the parameters.



There are two RoCE versions (v1 and v2).

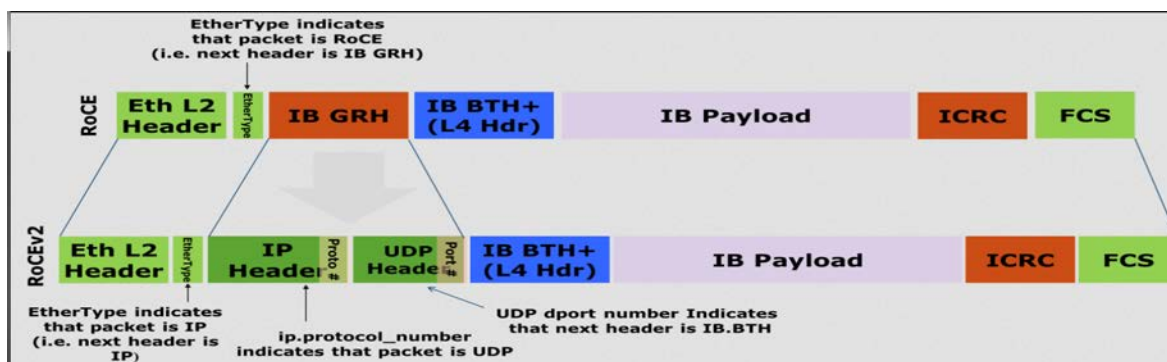
RoCE v1 and V2

Version 1 is the first version that was made available and it was defined as a link layer protocol allowing two hosts in the same broadcast domain (VLAN) to communicate. This version of the protocol uses ethertype 0x8915, which means that the frame length is limited by the standard Ethernet protocol definition, i.e. 1500 bytes for a standard Ethernet frame and 9000 bytes for an Ethernet jumbo frame.

Starting with Dell Networking OS release 9.0 RoCE v1 support was introduced on the S6000 data center switch allowing for HPC (High Performance Computing) and similar applications. With release OS 9.5 and above version 2 was introduced.

Version 2 is referred as Routable RoCE (RRoCE). Version 2 overcomes the limitation of version 1 being bounded to a single broadcast domain (VLAN). With version 2, IPv4 and IPv6 are supported allowing for the delivery of RoCE across different subnets bringing scalability into the picture.

Figure 3 RoCE v1 vs v2 Packet Format



RoCE; continues to enhance itself and with the ubiquity of Ethernet and the constant speed improvements and cost reductions per port it is poised to be a strong and certainly a viable alternative to Infiniband. Certainly there will always be cases where IB is more relevant than RoCE or iWARP, but for the most part the inherent HPC or HPC-like environments requirements can be met by the current Ethernet standards enhancements.

RoCE v1 and v2 Configuration

Figure 4, shows the test setup used to demonstrate a typical converged environment in a data center with LAN, storage (iSCSI), and IPC (Inter-Processor Communication) traffic.

Using this setup, RoCE v1 and v2 were configured and deployed using the following items:

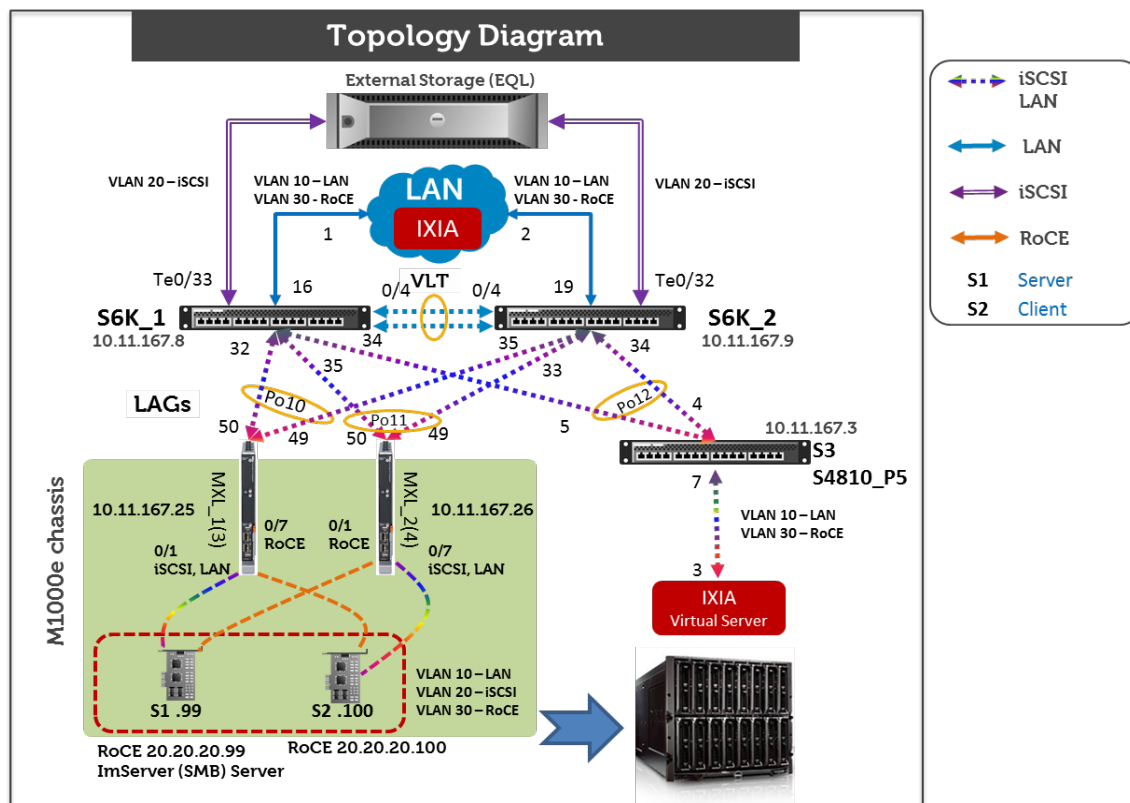
1. Dell Networking L2/L3 switches supporting DCB such as:
 - a. S4800 series, S6000, Z9500, S5000, and blade switches (MXL, and IOA)
2. Dell Networking Operating System 9.5.0.1
3. Dell M1000e with Dell Networking blade switch MXL
4. Dell Networking blade servers M620s
5. Emulex OCm14102-U3-D, 2-port 10GbE

6. Microsoft Windows 2012 with SMB (Server Message Block) v3.0

Following are the tested configuration for each of the different type of traffic found in this typical data center deployment. The following converged data traffic characteristics are:

- LAN data traffic
 - 802.1p priority 0, vlan 10
 - ETS = 50%
- iSCSI data traffic
 - 802.1p priority 4, vlan 20
 - ETS = 25%
- RoCE (RDMA over Converged Ethernet)
 - 802.1p priority 5, vlan 30
 - ETS = 25%

Figure 4 RoCE, iSCSI, LAN Deployment Diagram



RoCE v1 Configuration – Pure Layer 2 solution

The following configurations need to be repeated for both S6Ks, and MXLs. The configuration section includes only one of the devices since the configurations are identical.

Dell S6K_1 configuration

1. Enable DCB

S6K_1#conf

S6K_1(conf)#dcb enable pfc-queues 4 → *enable DCB and pfc with 4 queues*

2. Configure the DCB map and priority groups. Turn pfc on or off per traffic class type

S6K_1(conf)#dcb-map ALL → *configure the dcb map and turn on pfc and ets on specific traffic class*

S6K_1(conf-dcbmap-ALL)#priority-group 0 bandwidth 50 pfc off → *LAN traffic, pfc is off, assign 50% BW*

S6K_1(conf-dcbmap-ALL)#priority-group 1 bandwidth 25 pfc on → *iSCSI traffic pfc is on, assign 25% BW*

S6K_1(conf-dcbmap-ALL)#priority-group 2 bandwidth 25 pfc on → *RoCE traffic pfc is on, assign 25% BW*

S6K_1(conf-dcbmap-ALL)#

3. Assign dcb map to physical interface

S6K_1(conf)#int range te0/32 – 34, te0/35, fo0/4

S6K_1(conf-if-range-te0/32-34, te0/35, fo0/4)#dcb-map ALL

S6K_1(conf-if-range-te0/32-34, te0/35, fo0/4)#no shut

S6K_1(conf-if-range-te0/32-34, te0/35, fo0/4)#end

4. Configure VLT on the Spine switches (S6Ks)

S6K_1#conf

S6K_1(conf)#int fo0/4

S6K_1(conf-if-fo0/4)#port-channel-protocol IACP

S6K_1(conf-if-fo0/4-lACP)#port-channel 1 mode active

S6K_1(conf-if-fo0/4-lACP)#no shut

S6K_1(conf-if-fo0/4-lACP)#end

S6K_1#conf

S6K_1(conf)# vlt domain 1

S6K_1(conf-vlt-domain)#peer-link port-channel 1

5. Configure VLT peer port-channels

S6K_1#conf

S6K_1(conf)#int te0/32

S6K_1(conf-if-te0/32)#port-channel-protocol lacp

S6K_1(conf-if-te0/32-lACP)#port-channel 10 mode active

S6K_1(conf-if-te0/32-lACP)#no shut

S6K_1(conf-if-te0/32-lACP)#end

S6K_1#conf

S6K_1(conf)#int po 10

S6K_1(conf-if-po-10)#portmode hybrid

S6K_1(conf-if-po-10)#switchport

S6K_1(conf-if-po-10)#vlt-peer-lag port-channel 10

S6K_1(conf-if-po-10)#no shut

The same set of commands would be configured on the S6K_2 port Te0/35 to create the VLT peer port-channel.

6. Configure the different vlans – traffic classes

S6K_1#conf

S6K_1(conf)#int vlan 10

S6K_1(conf-if-vl-10)#tag po10

Repeat the steps in # 6 for vlans 20 & 30.



MXL_1(3) Configuration

1. Enable DCB

```
MXL_1#conf
MXL_1(conf)#dcb enable
```

2. Re-configure the queue maps on the MXL, so that 802.1p 3 is assigned to the same queue (queue 3) and 802.1p 5 is assigned to queue 1. This re-configuration is necessary because the S6K has 4 pfc queues, while the MXL has 2 pfc queues usable. The dcb configuration is being pushed by the S6K as the source, and the dcb-map configuration calls for two pfc queues on which pfc is turned on.

```
MXL_1(conf)#do sh qos dot1p-queue-mapping → default 802.1p mapping
Dot1p Priority : 0 1 2 3 4 5 6 7
Queue : 0 0 0 1 2 3 3 3
```

The configuration line “**priority-pgid 0 0 0 0 1 2 0 0**” as part of the dcb-map, says turn on pfc on 802.1p 4 & 5, but the default queue mapping (see above) states that 802.1p 5 is part of queue 3 and so is 802.1p 6 & 7. This creates a dcb configuration conflict between the S6K (source) and MXL which requires the re-configuration of the queue map.

NOTE: The following data center product family (Z9100, 9500, S6000, S4048) have 4 “lossless” queues while the following product family (S4810, S4820T, S5000) have 2 “lossless” queues.

```
MXL_1#conf
MXL_1(conf)#service-class dot1p-mapping dot1p3 3 dot1p5 1
MXL_1(conf)#do sh qos dot1p-queue-mapping
Dot1p Priority : 0 1 2 3 4 5 6 7
Queue : 0 0 0 3 2 1 3 3
```

3. Configure the respective DCBx port-role on the interfaces facing the hosts.

```
MXL_1#conf
MXL_1(conf)#int range te0/1 , te0/7
MXL_1(conf-if-te-0/1,te-0/7)#portmode hybrid
MXL_1(conf-if-te-0/1,te-0/7)#switchport
MXL_1(conf-if-te-0/1,te-0/7)#protocol lldp
MXL_1(conf-if-te-0/1,te-0/7-lldp)#dcbx port-role auto-downstream
MXL_1(conf-if-te-0/1,te-0/7-lldp)#dcbx version auto
MXL_1(conf-if-te-0/1,te-0/7-lldp)#end
```

4. Configure the port-channel to the VLT domain and respective DCBx port-role on the upstream interfaces to the source.

```
MXL_1#conf
MXL_1(conf)#int range te0/49 – 50
MXL_1(conf-if-range-te-0/49-50)#port-channel-protocol lacp
MXL_1(conf-if-range-te-0/49-50-lacp)#port-channel 10 mode active
MXL_1(conf-if-range-te-0/49-50-lacp)#no shut
MXL_1(conf-if-range-te-0/49-50-lacp)#exit
MXL_1(conf-if-range-te-0/49-50-lldp)#protocol lldp
MXL_1(conf-if-range-te-0/49-50-lldp)#dcbx port-role auto-upstream
MXL_1(conf-if-range-te-0/49-50-lldp)#no shut
MXL_1(conf-if-range-te-0/49-50-lldp)#end
```

