

VRD

RDMA OVER CONVERGED ETHERNET (ROCE) DESIGN GUIDE

REDUCING LATENCY, CPU OVERHEAD, AND ENABLING FAST TRANSFER
OF DATA

RDMA over Converged Ethernet (RoCE) design guide.....	1
Introduction.....	3
Modern data center challenges—what is the problem?.....	3
Remote Direct Memory Access (RDMA) Solution	3
RDMA transport technologies	4
What is RoCE?	5
RoCE aspects.....	5
RoCE use cases	6
Cloud computing.....	6
Data storage	7
Data warehousing.....	8
Financial services	8
Web 2.0 big data.....	9
RoCE design recommendations	9
Lossless network fabrics.....	9
Configuration guidance and details	10
Priority Flow Control (PFC).....	10
Enhanced Transmission Selection (ETS).....	12
Data Center Bridging Exchange (DCBx).....	13
Quantized Congestion Notification (QCN).....	13
IP Explicit Congestion Notification (ECN) and RoCE v2 Congestion Management (RCM)	14
Layer 3 interface considerations	18
Summary	19

Introduction

This guide provides information on Remote Direct Memory Access (RDMA) solutions and use cases as well as information about transporting and configuring RDMA over Converged Ethernet (RoCE) in a FlexFabric environment. The intended audience for this document is IT administrators and solution architects planning on deploying RDMA based RoCE solutions using HPE Networking.

Modern data center challenges—what is the problem?

Within today's enterprise, servers are required to handle massive amounts of data while providing 100% uptime. Over the years, adoption of server virtualization, big data analytics and the proliferation of mobile devices have continued to stress computing infrastructures. Users have noticed applications taking longer than they should to execute. When corporate and other users notice slowing of the systems, they become less productive. Many times, this type of delay happens because large amounts of data has to be processed by the CPU which then has to move from buffer spaces, down through the TCP stack, onto the wire between servers of the enterprise, and then back up the stack again on the other side. This transfer can cause the CPU to slow down processing of other tasks as the CPU is busy. Adding more servers may increase CPU processing power but it is not addressing the fact that the CPUs are getting over-utilized and it runs counter to the goal of doing more with less within today's organizations.

Remote Direct Memory Access (RDMA) Solution

RDMA enables the movement of data between servers with very little CPU involvement.

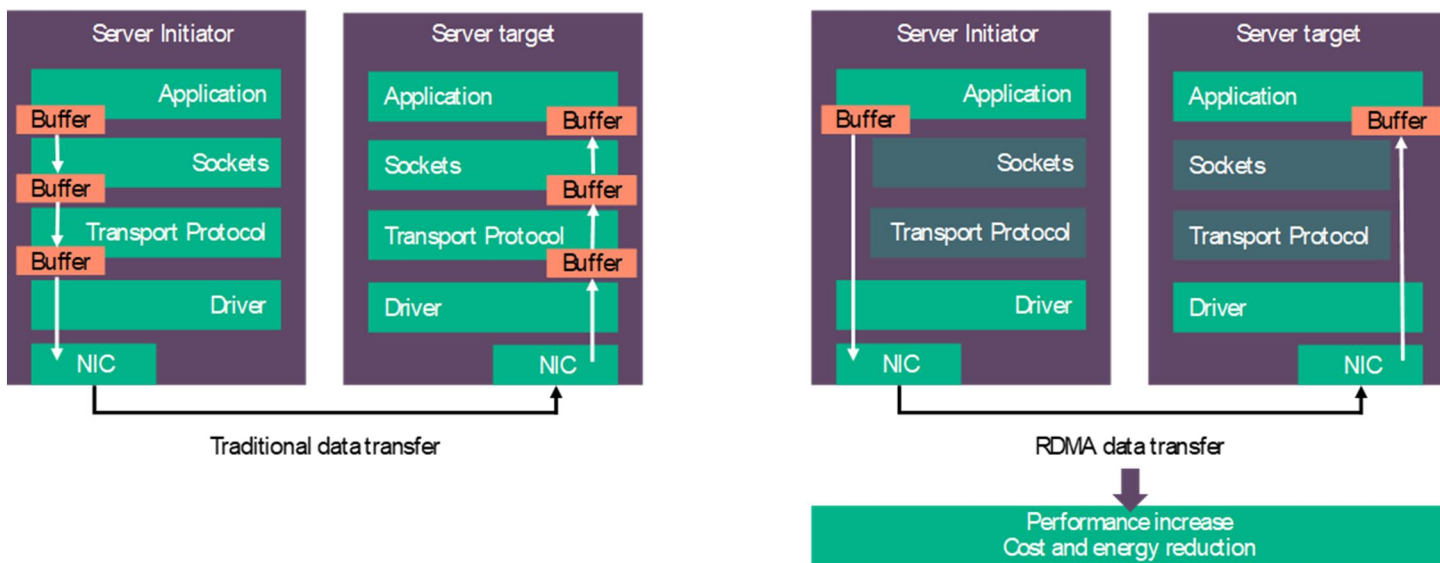


Figure 1. RDMA bypassing OS stack

Without RDMA, traditional movement of data utilizes TCP/IP buffer copies and significant overhead. Applications rely on the OS stack to move data from memory, virtual buffers through the stack, onto the wire, across the wire, and then again back up the wire. The receiving OS must retrieve the data and place it directly in the application(s)' virtual buffer space which leads to the CPU being occupied for the entire duration of read and write operations and is unavailable to perform other work.

RDMA solutions are able to bypass the OS stack. The OS is used to just establish a channel which applications use to directly exchange messages on. A network adapter transfers data directly to and from application memory eliminating the need to copy data between application memory and the data buffers within the operating system. Such communication requires no work to be done by CPUs, caches or context switches, and transfers continue in parallel with other system operations. When an application performs an RDMA Read or Write request, the application data is delivered directly to the network, reducing latency, CPU overhead, and enabling fast transfer of data.

Users running RDMA applications on an Ethernet network can see application performance improvements that derive from the offloading of data movement and higher availability of CPU resources to applications. Shifting the chore of data movement from the CPU makes both data movement and execution of applications more efficient. RDMA delivers performance and efficiency gains that are not available from any other communications protocol; including low latency, improved resource utilization, flexible resource allocation, fabric unification and scalability. Greater server productivity lowers the need for additional servers and lowers the total cost of ownership.

- RDMA is the direct read from or write to an application's memory
- Hardware offload moves data faster with significantly less overhead allowing the CPU to work on other applications
- CPU initiates the transfer and processes other operations while the transfer is in progress
- Ultra-low latency through stack bypass and copy avoidance
- Reduces CPU utilization
- Reduces memory bandwidth bottlenecks
- Enables high bandwidth and I/O utilization
- RDMA is useful when CPU cannot keep up and needs to perform other useful work

RDMA transport technologies

There are three main transport types of solutions that can be used to transport RDMA over an Ethernet network.

- InfiniBand (IB)
 - Protocol which supports RDMA natively from the beginning
 - Requires dedicated NICs and switches that supports this technology
 - Pure InfiniBand solutions can provide high performance at cost of dual networking fabrics
- Internet Wide Area RDMA Protocol (iWARP)
 - RDMA over TCP
 - iWARP defined by IETF and uses the TCP/IP stack in order to be compatible with any Ethernet/IP infrastructure
 - Data Center Bridging (DCB) Ethernet helps avoid congestion, but it is not required by the standard
 - Supports offload to the NIC
 - Goes up the TCP/IP stack to achieve protection for loss
- RDMA Over Converged Ethernet (RoCE)
 - Data Center Bridging (DCB) Ethernet should be configured, but it is not required by the standard
 - Requires a DCB switch to provide for a lossless fabric
 - NICs should support RoCE and offloading
 - Lower level Ethernet mechanisms used to protect for loss:
 - Priority Flow Control (PFC) to stave off loss
 - Enhanced transmission selection (ETS) to protect traffic classes (TC)
 - Uses upper InfiniBand layers in case of need for retransmission to recover from loss.

InfiniBand solutions use a separate dedicated network and still may be applicable to certain markets. However, when converging the solution onto Ethernet, RoCE based solutions have seen more market adoption than iWARP. With that in mind, this document focuses on RoCE based networking.

What is RoCE?

RoCE is a network protocol that allows RDMA over Converged Ethernet, or RoCE (pronounced “rocky”). This critical technology is now expanding into enterprise markets where Ethernet networks are ubiquitous. RoCE is geared for high performance within an advanced data center architecture eliminating dedicated storage area networks (SANs) by converging compute, network, and storage onto a single fabric. Utilizing advanced reliable Ethernet and DCB with RDMA techniques, RoCE provides lower CPU overhead and increases enterprise data center application performance.

Today’s dynamic evolving enterprise, let it be local, remote cloud, or hybrid data centers, require high performance technologies like RoCE to support increasingly data intensive applications and the move to hyper converged scale-out solutions which leverage distributed computing/storage models.

Hewlett Packard Enterprise currently supports RoCE solutions using FlexFabric Products.

The benefits of implementing RoCE are:

- Lower cost of ownership
- Greater return on investment that span traditional and today’s hyper converged infrastructures
- Reduces CPU over utilization
- Reduces Host Memory Bottlenecks
- Helps to better leverage the storage media evolution which has brought 10,000x performance improvement factor
- Offloads memory access process
- Increases throughput and lowers latency between compute and storage systems

RoCE aspects

Within the technology industry, Ethernet naturally envisions the complete IP architecture consisting of TCP, IP and Ethernet. However, the truth is that RoCE and InfiniBand are really the sibling technologies....RoCE is only a distant cousin to TCP/IP/Ethernet.

The initial RoCE v1 solution simply replaced the IB Link Layer with an Ethernet link layer. In this solution RoCE was a Layer 2 based Ethernet solution. The latest version of RoCE, which is called RoCE v2, replaced the IB Network layer with a standard IP and UDP Header so traffic is routable now.

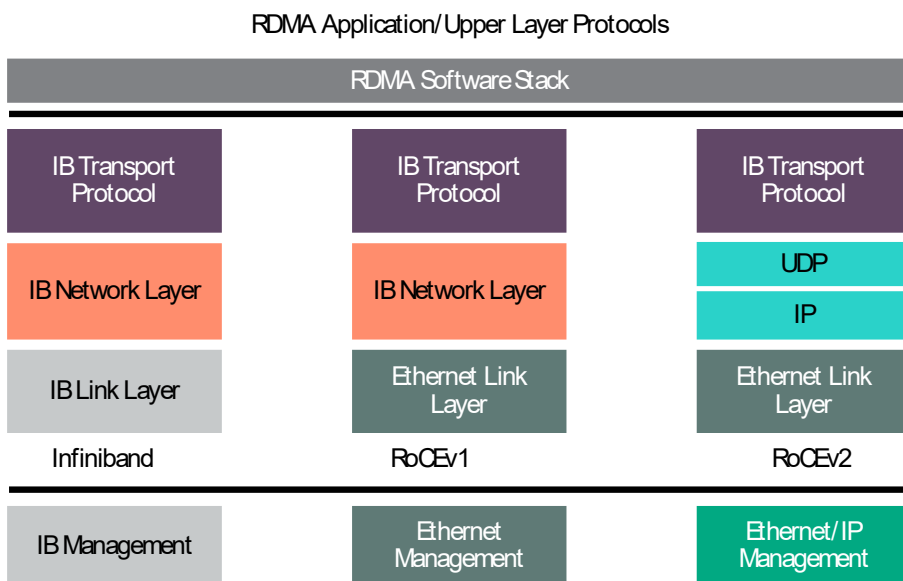


Figure 2. IB/RoCE evolution

The InfiniBand Annexes for RoCEv1 and RoCEv2 do not actually mandate that no loss occur on the Ethernet network. The actual InfiniBand annex state:

- InfiniBand annex RoCE V1: “Although this annex does not specifically require a ‘lossless’ Ethernet fabric, it is likely that a port implementing this annex which is connected to an Ethernet fabric will operate more efficiently (e.g. with a higher ratio of delivered payload to offered payload) if the underlying Layer 2 fabric does not routinely drop packets”
- InfiniBand annex RoCE V2: “As with RoCE, the underlying networks for RoCE v2 should be configured as lossless. In this context, lossless doesn’t mean that packets are absolutely never lost. Moreover, the Transport Protocol in RoCE v2 includes an end-to-end reliable delivery mechanism with built-in packet retransmission logic. This logic is typically implemented in HW and is triggered to recover from lost packets without the need for intervention by the software stack. The requirement for an underlying lossless network is aimed at preventing RoCE v2 packet drops as a result of contention in the fabric.

However, even though not mandatory or required by the standard, RoCE based solutions will perform poorly in lossy fabrics because in times of congestion devices will start to drop packets, causing retransmissions, reducing throughput, and increasing delay. In these situations, the solution will simply not get the RDMA benefits needed to reduce CPU load and latency.

HPE recommends, in production environments, RoCE based solutions should be deployed as a lossless fabric.

To achieve a lossless networking fabric, RoCE uses the lower level Ethernet mechanisms, namely DCB protocols (discussed later).

- RoCE is defined and specified by the ITBA
- End to end solution – require NIC and stack support as well as network support
- RoCE encapsulates IB transport and GRH headers in Ethernet packets bearing a dedicated ether type
- RoCE V2: Adds an IP/UDP header for routing capabilities
- Involves configuring explicit no drop classes on network devices to ensure as close to lossless operation as possible (typically PFC and sometimes ECN)
- Available at all speeds Ethernet; 10/25/40/50/100GbE

RoCE use cases

RDMA and RoCE have seen growing market adoption. Below is a summary of the markets which have started to leverage these solutions.

Cloud computing



Figure 3. Cloud computing

The cloud computing market has been actively leveraging the benefits of RDMA and RoCE. These environments obtain benefits, such as improved SLAs through deterministic performance, efficient clustering allowing for elastic/scale out computing. As indicated before, the benefits of implementing RoCE are a lower cost of ownership and greater return on investment that span traditional and modern hyper converged infrastructures.

These environments are able to use the following applications (and more) that leverage RDMA/RoCE:

- VMware®
 - VMware has published [Ultra-Low Latency on vSphere with RDMA](#) which compares RDMA vs regular TCP/IP stack in cloud environment showing following benefits:
 - Total vMotion traffic time is 36% faster.
 - 30% higher copy bandwidth ability.
 - CPU utilization is from 84% to 92% lower.
- Microsoft® Azure
 - [Video](#) describing how a 40GbE implementation RoCE solution provided 40Gbps throughput at 0% CPU utilization with Azure
- Red Hat® KVM
- Citrix® Xen
- Microsoft®
- Amazon EC2
- Google™ App Engine

Data storage



Figure 4. Data storage

Many data storage focused applications are gaining benefits from implementing RDMA/RoCE. A couple of examples are Microsoft SMB Direct, and Lustre. Data Storage protocols over RDMA deliver higher throughput and lowers latency.

Everyday application like Exchange, Lync® and SharePoint are also able to leverage and benefit from the use of RoCE.

- A Mellanox test paper described how 1.2 Tb/s storage throughput resulted on a 12 node DL380 cluster running Windows Server® 2016 and RoCE. This environment was utilizing the full 100 Gb/s Ethernet bandwidth between applications and the server-based storage.
- A Micron brief details how SMB network file shares using RDMA increased throughput by 2x and reduced processor utilization by up to 20%

Data warehousing



Figure 5. Data warehousing

Data warehousing applications such as Oracle RAC, IBM DB2 PureScale and Microsoft SQL will also benefit from RoCE environments. They can achieve significantly higher job operations, higher I/Os per second, linear scaling with cluster size, and maintain table scan time in the face of exponential growth in database table sizes.

- This [Accelerate Microsoft SQL Server on Hyper-V with HP and Emulex](#) paper provides data on an HPE FlexFabric solution with support for RoCE that accelerated application file storage I/O by up to 82 percent when compared to adapters lacking SMB Direct support

Financial services



Figure 6. Financial services

Financial Service market has been leveraging InfiniBand for some time now as those environments have a high demand for low latency. The following are applications which could see high performance and I/O benefits in RoCE solutions:

- Tibco
- Wombat/NYSE
- IBM WebSphere MQ
- Red Hat MRG
- 29West/informica

Web 2.0 big data



Figure 7. Web 2.0

Big data environment is another segment which can benefit from RoCE solutions. The goal in these big data environments is to minimize response time and increase I/O.

The following are applications which could see benefits in RoCE solutions:

- Hadoop
 - After [Hadoop](#) introduced a RoCE solution into their environment, they achieved 2x performance improvements using RoCE-based, RDMA-enabled file systems like HDFS, GPFS and Ceph
- Memcached
- Eucalyptus
- Cassandra

RoCE design recommendations

Lossless network fabrics

One of the objectives when designing network solutions that incorporate RoCE is to deploy a lossless fabric. Even though the RoCE standards do not necessarily demand lossless networks they will perform much better when deployed in a lossless manner, especially under heavy congestion. With that in mind, HPE recommends that a lossless fabric be considered a requirement for RoCE implementations.

Lossless fabrics can be built on Ethernet fabrics by leveraging the following DCB protocols.

- Priority- based flow control (PFC): IEEE standard 802.1Qbb, is a link-level flow control mechanism. The flow control mechanism is similar to that used by IEEE 802.3x Ethernet PAUSE, but it operates on individual priorities. Instead of pausing all traffic on a link, PFC allows you to selectively pause traffic according to its class.
- Enhanced Transmission Selection (ETS): ETS helps to ensure that when using PFC, lossless traffic does not get continually paused because other types of traffic are using the whole bandwidth of a link. With ETS you can tell an interface that a certain percentage of the bandwidth is guaranteed for each traffic type. This way each traffic type gets a minimum amount of bandwidth.
 - Data Center Bridging Exchange (DCBX): This protocol helps to ensure that the NIC and the Switch are configured properly. DCBx is a discovery and capability exchange protocol which discover peers and exchanges configuration information. The protocol allows auto exchange of Ethernet parameters and discovery functions between switches and end-points. If the server NIC has the DCBx willing bit turned on then after you configure the switch with the needed DCB and traffic marking rules, then DCBx will ensure the server NIC also knows how to mark and treat traffic on that link.

- Quantized Congestion Notification (QCN): This protocol provides a means for a switch to notify a source that there is congestion on the network. The source will then reduce the flow of traffic. This help to keep the critical traffic flowing while also reducing the need for pauses. This is only supported in pure Layer 2 environments and seen very rarely now that RoCEv2 is the predominant RoCE solution.
- IP Explicit Congestion Notification (IP ECN): IP explicit congestion notification is not officially part of the DCB protocol suite, however, it can be leveraged to enable end-to-end congestion notification between two endpoints on TCP/IP based networks. The two endpoints are an ECN-enabler sender and an ECN-enabler receiver. ECN must be enabled on both endpoints and on all the intermediate devices between endpoints for ECN to work properly. ECN notifies networks about congestion with the goal of reducing packet loss and delay by marking the sending device to decrease the transmission rate until congestion clears, without pausing packets.

Note:

IP ECN is not specifically a DCB standard, but it can be useful in a RoCEv2 solution (see details later).

Configuration guidance and details

Priority Flow Control (PFC)

PFC enables flow control over a unified 802.3 Ethernet media interface, for local area network (LAN) and storage area network (SAN) technologies. PFC is intended to eliminate packet loss due to congestion on the network link. This allows loss sensitive protocols, such as RoCE to coexist with traditional loss-insensitive protocols over the same unified fabric.

Before the development of PFC a global pause was used which is port based. This means that in times of congestion all traffic on that port would get paused. PFC is port and queue based, which means it allows us to pause only traffic in a certain queue on a port. This way PFC simply pauses traffic that is in a lossless queue, so that no packets get dropped due to buffer congestion. All other traffic that is in the other queues on the port are not included for this type of checking and may be dropped as congestion occurs.

If you need to ensure that zero packets get dropped, then PFC must be deployed.

PFC details:

- Must be enabled on all endpoints and switches in the flow path
- Enables pause per hardware queue on an Ethernet device
- PFC is port and queue based (not flow based)
- Uses 802.1p CoS “Class of Service” values in 802.1Q VLAN tag to differentiate up to eight levels of CoS
- On L3 interfaces PFC requires preservation of 802.1Q tags
- On L3 interfaces if 802.1Q tags are not possible then traffic needs to be remarked to DSCP values
- Pause frames propagate hop-by-hop, without knowledge of the flows that are causing the congestion

Configuration examples

Below are simple configuration examples on HPE Comware devices. The idea is to configure PFC on an interface so that it knows not to drop traffic marked with a specific CoS value.

```
interface Ten-GigabitEthernet1/0/10
priority-flow-control auto
priority-flow-control no-drop dot1p 5
qos trust dot1p
```

Possible adverse effects due to PFC

Because PFC is port and queue based it is possible that PFC pauses affect unrelated traffic flows. Understanding possible adverse effects can help designers ensure that proper bandwidth and link distribution has been built into the design. Below are a couple of examples which can be considered 'victim flows' and 'unfairness'.

In the image below, hosts A to E are all sending traffic to port R. This is causing congestion at port R. In this example, because of the congestion at port R, PFC will send a pause frame to the select queue on all physical interfaces that are contributing to the congestion. The example also has a flow conversation occurring between Host X and Host Y. Even though there is no congestion on the switch to switch link, if the X-Y flow is marked with the same CoS values that PFC has been set to no-drop, then that flow would also receive a pause frame. This means that the X-Y flow becomes a 'victim flow'.

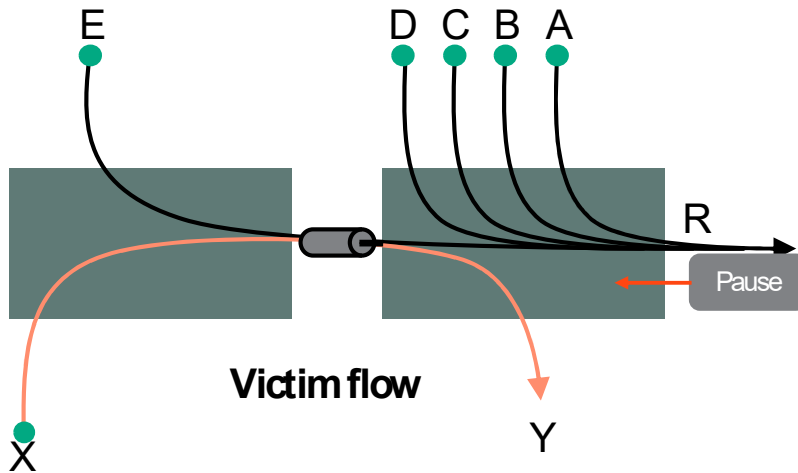


Figure 8. Victim flow

Lossless environments can also lead to "Unfairness" scenarios. In times of congestion, pause frames will be evenly distributed among ports contributing to congestion, with no consideration to the number of traffic flows over a port.

In this example below, traffic is coming from Host ABC and D in the direction of Host R, causing congestion at port R. Since PFC is port and queue based, it recognizes that there are only two ports contributing to congestion in this scenario. PFC does not recognize that there are three flows coming in on one of those ports. In this example, PFC is going to distribute pause frames evenly to the two ports.

Assuming these links are all 40GbE interfaces, Host D is going to obtain about 20GbE bandwidth on its flow to Host R, while Hosts ABC will all share the remaining bandwidth. This means they will only be getting roughly 6G of bandwidth to port R, even though the switch to switch link is not fully utilized.

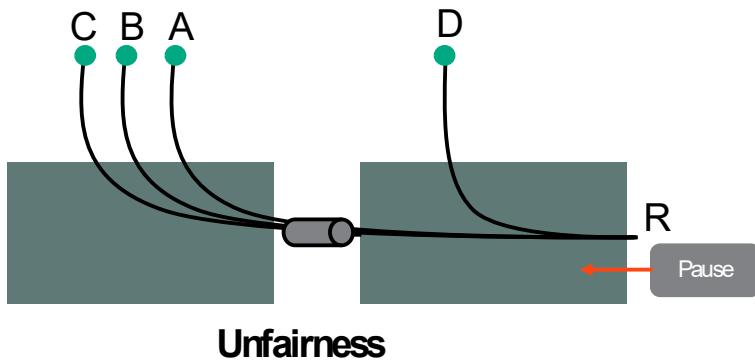


Figure 9. Unfairness

Enhanced Transmission Selection (ETS)

In converged environments, we need to ensure that PFC based lossless traffic does not get starved out due to congestion. The switches and interfaces can be configured with ETS to ensure that certain traffic classes on a switch can get a minimum amount of bandwidth. This ensures that even if the link had a very bandwidth hungry traffic flow then that flow will not cause pauses to starve out the lossless traffic class.

When we configure ETS we tell the interface the minimum amount of bandwidth that each traffic class (TC) requires. This does not mean that each traffic class is only allowed to use up to that bandwidth. If there are no competing traffic flows, then each traffic class is allowed to use all of the available bandwidth on the link.

Configuration examples

The Comware example below configures ETS in two steps. The first is to modify the dot1p-lp mapping table. This is a locally significant configuration which tells the switch that all traffic marked with a CoS values of 5 will get placed into queue 1. In the example, all other CoS marked traffic will get placed into queue 0.

```
qos map-table dot1p-lp
import 0 export 0
import 2 export 0
import 3 export 0
import 4 export 0
import 5 export 1
import 6 export 0
import 7 export 0
```

The next step is to tell the physical interface that anything marked with a CoS value of 5 (now in the AF1 queue) will get a certain percentage of the bandwidth at minimum. The configuration uses a weight to represent a percentage. In the example, queue AF1 (queue 1) and queue BE (queue 0) will each get 50% of the bandwidth on the link. All other queues have Strict priority set although there will be no traffic in these queues thanks to the 82.1p-lp mapping table configuration.

```
interface Ten-GigabitEthernet1/0/10
qos wrr be group 1 byte-count 15
qos wrr af1 group 1 byte-count 15
qos wrr af2 group sp
qos wrr af3 group sp
qos wrr af4 group sp
qos wrr af group sp
qos wrr ca6 group sp
```

```
qos wrr ca7 group sp
```

Data Center Bridging Exchange (DCBx)

It is important that the RoCE related traffic sent into the switch is marked with the appropriate CoS priorities. This can be achieved by manually configuring the server NICs or by configuring rules on the switch which then get communicated via DCBx to the connected hosts.

In addition to CoS priorities, DCBx also exchanges VLAN tagging and ETS parameters to the NIC. DCBx helps to ensure that the NIC and the Switch are configured properly and it allows the auto exchange of Ethernet parameters and discovery functions between switches and end-points.

Configuration examples

The Comware example below shows a configuration for the same SMB direct solution. However, Comware requires a QoS policy to be created and applied to the interface. The QoS policy simply selects UDP port destination 4791 traffic and also remarks with a CoS value of 5.

```
acl number 3001
  Rule 0 permit udp destination-port eq 4791
#
traffic classifier ROCE_class operator or
  if-match acl 3001
#
traffic behavior ROCE_behavior
  remark dotlp 5
#
qos policy ROCE_policy
  Classifier ROCE_class behavior ROCE_behavior mde dcbx
#
interface Ten-GigabitEthernet1/0/10
  qos apply policy ROCE_policy outbound
```

Below is an example of turning on DCBx on a Comware switch so that the above marking rule as well as PFC, ETS and VLAN data get communicated to the server NIC.

```
lldp global enable
#
interface Ten-GigabitEthernet1/0/10
  lldp tlv-enable dot1-tlv dcbx
```

Quantized Congestion Notification (QCN)

QCN is an end-to-end congestion mechanism which helps to reduce packet loss and delay in Layer 2 networks only. When a switch starts to get congested, it simply sends a message towards the source of the congestion. The source will throttle back its traffic to help reduce congestion and avoid unnecessary pauses in the traffic pattern.

Because of this, QCN is one way to help mitigate or reduce the possible adverse effects that PFC may cause.

With QCN the Congestion Point (CP) periodically sample frames from queues that are enabled with QCN, and when congestion occurs it sends Congestion Notification Messages (CNMs) to the source end host that support QCN. The source end host reaction point (RP) that supports QCN will reduce their transmission rates when receiving CNMs. The source end host that supports QCN will also periodically probe the bandwidth and increase their transmission rates if they fail to receive CNMs for a specific period of time.

Note:

QCN is a pure end-to-end Layer 2 solution. Most modern data centers are building Layer 3 fabrics so QCN is less relevant in today's networks.

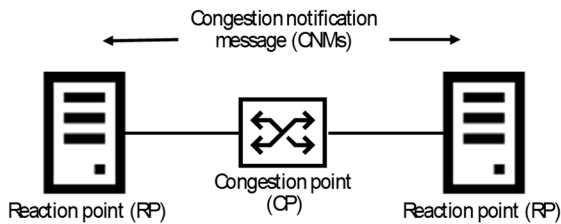


Figure 10. QCN example

Configuration examples

```
qcn enable
qcn priority 5 auto
#
interface Ten-GigabitEthernet1/0/10
  lldp tlv-enable dot1-tlv congestion-notification
```

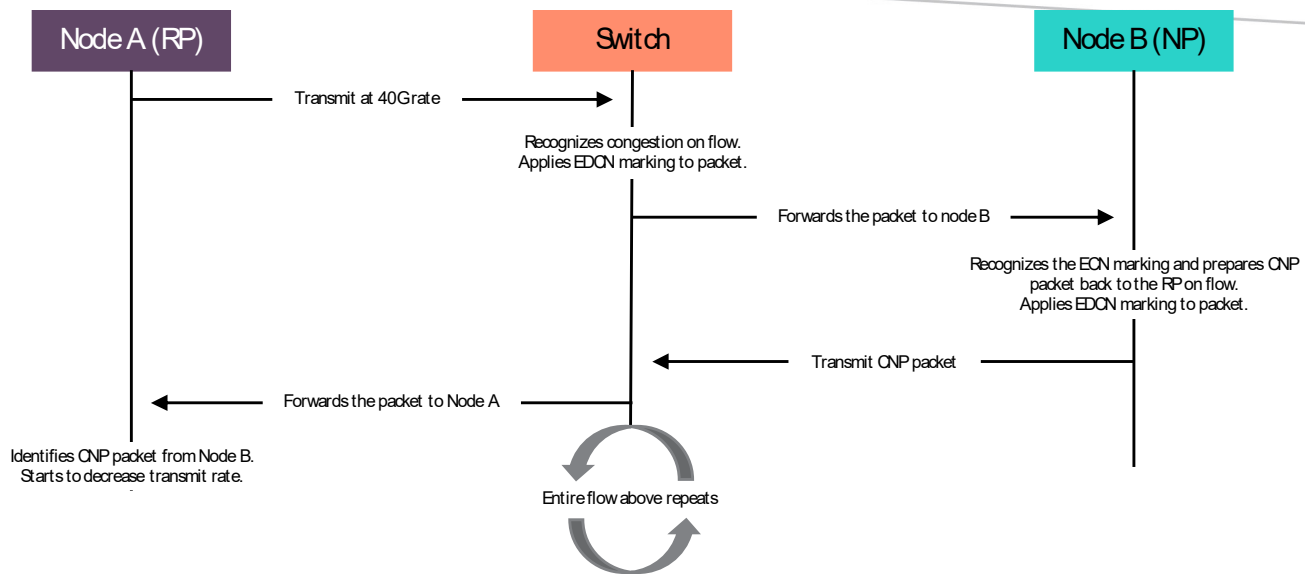
IP Explicit Congestion Notification (ECN) and RoCE v2 Congestion Management (RCM)

RCM provides the capability to avoid congestion hot spots and optimize the throughput of the fabric even over Layer 3 links.

The RoCE v2 Congestion Management feature is composed of three points:

- The congestion point (CP): Detects congestion and marks packets using DCQN bits.
- The notification point (NP) (receiving end node): Reacts to the marked packets by sending congestion notification packets (CNPs).
- The reaction Point (RP) (transmitting end node): Reduces the transmission rate according to the received CNPs.

With RoCE RCM, when congestion occurs the CP sees congestion but it continues to forward the traffic to the destination NP. Now that the NP destination knows there is congestion it replies to the source node RP and informs it that there is congestion. Now the source RP reacts by decreasing and later on increasing the Tx "transmit" rates according to the feedback provided. The source RP node keeps increasing the Tx rates until the system reaches a steady state of non-congested flow with traffic rates as high as possible.



The flow will continue in a loop until Node A has decreased the Tx rate to a point where:

- The switch will no longer feel congested and will stop ECN marking
- Node B will stop sending CNPs to Node A

Figure 11. RCM steps

On the networking side RoCE RCM is configured using IP ECN. Packets are marked with IP ECN bits by each switch at a configurable ECN threshold, allowing TCP to function normally and helping to prevent pauses. Both endpoints and each device in the traffic path needs to support this feature. ECN uses the DS field in the IP header to mark the congestion status along the packet transmission path.

ECN operates as follows:

- When the average queue size exceeds the lower limit, and is below the upper limit, before the device drops a packet which should be dropped according to the drop probability, the device examines the ECN field of the packet
 - If the ECN field shows that packet is sent out ECN-capable terminal, the device sets both the ECT bit and CE bit to 1 and forwards the packet
 - If the ECN field shows that a packet has experienced congestion (Both the ECT bit and CE bit are 1), the device forwards the packet without modifying the ECN field
 - If both ECT bit and CE bit are 0s, the device drops the packet
- When average queue size exceeds the upper limit, the device drops packet, regardless of whether the pack is sent from ECN-capable terminal

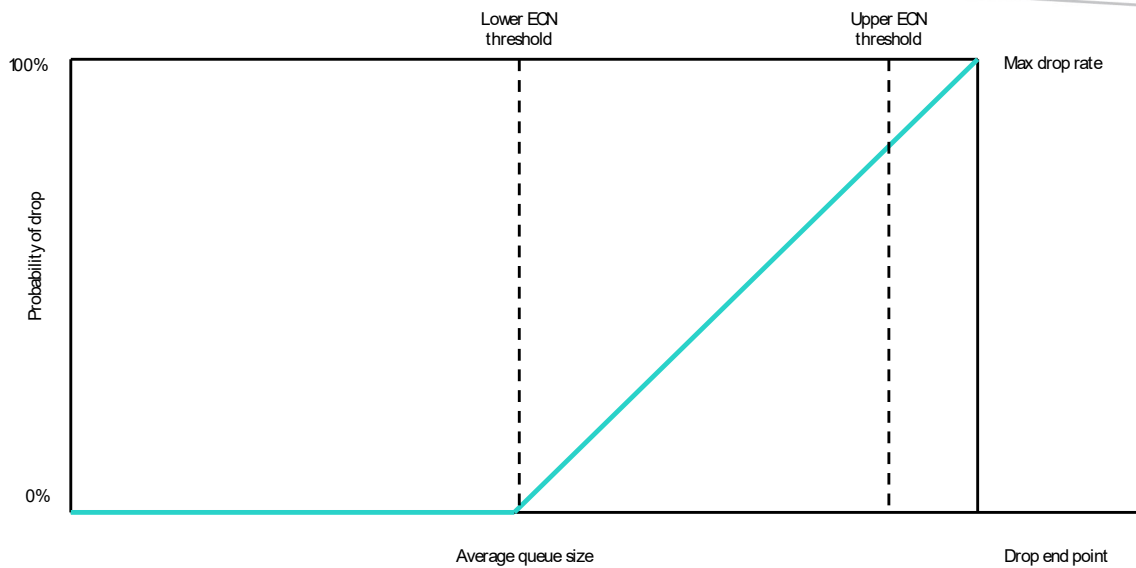


Figure 12. IP ECN thresholds

Configuration examples

The Comware example leverages a similar configuration however there is an additional discard-probability configuration. The Comware example also configures the queues based on whether traffic is considered Green (drop-level 0), Yellow (drop-level 1), or Red (drop-level 2). If the token bucket in the switch has enough tokens to forward the packet, the packet is colored green, otherwise the token is marked red. If token bucket C does not have enough tokens but bucket E has enough tokens, the packet is colored yellow.

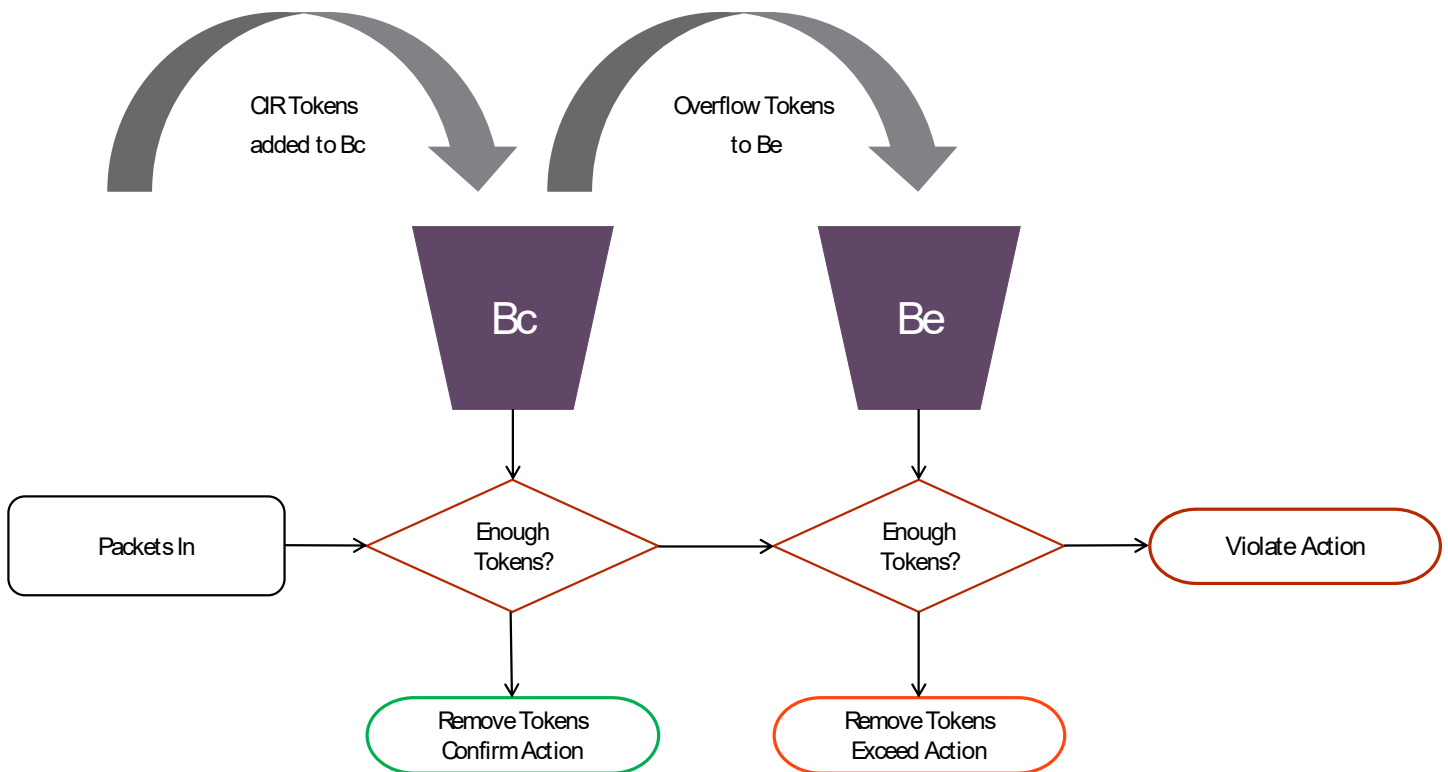


Figure 13. Token buckets (Comware)

The IP ECN configuration gets applied specifically to queue 1 to ensure this only gets applied to the selected RoCE traffic.

```
qos wred queue table ROCEv2
queue 0 drop-level 0 low-limit 1000 high-limit 18000 discard-probability 25
queue 0 drop-level 1 low-limit 1000 high-limit 18000 discard-probability 50
queue 0 drop-level 2 low-limit 1000 high-limit 18000 discard-probability 75
queue 1 drop-level 0 low-limit 18000 high-limit 37000 discard-probability 1
queue 1 drop-level 1 low-limit 18000 high-limit 37000 discard-probability 5
queue 1 drop-level 2 low-limit 18000 high-limit 37000 discard-probability 10
queue 1 ecn
#
interface Ten-GigabitEthernet1/0/10
qos wred apply ROCEv2
```

Congestion notification benefits

Going back to the PFC example of unfairness. In this example, when there is no congestion notification turned on, and the configured is using all 40GbE interfaces, the flow from Host D to Host R will get roughly 18 GB/s throughput while the other flows running across the switch to switch port will need to share the remaining bandwidth and only obtain about 6 GB/s throughput each.

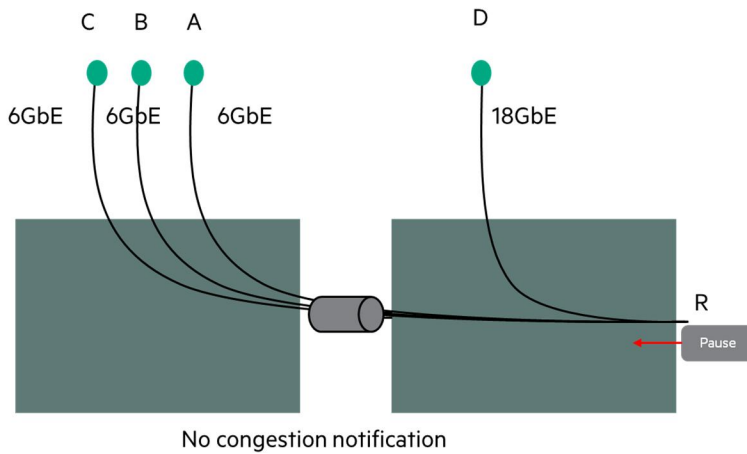


Figure 14. Unfairness with no CN

After configuring IP ECN, in this example, the results should show that all related flows are now evenly distributing the bandwidth. IP ECN is a good mechanism to help smooth out traffic flows and help alleviate these types of PFC based adverse effects.

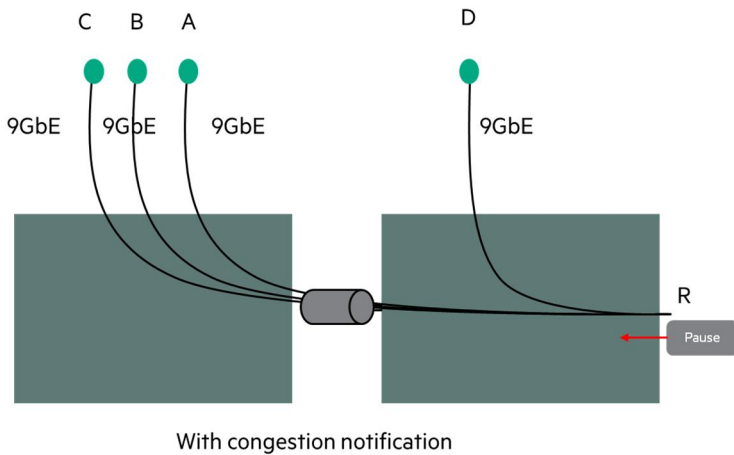


Figure 15. Unfairness with CN

Layer 3 interface considerations

As noted in the above configurations, PFC is applied to the switch by telling an interface not to drop packets marked with a specific 802.1p CoS value. This works well for any interface as long as the 802.1Q tag which carries the CoS value is retained. For these scenarios where a pure L3 routed link that does not carry 802.1Q tags is required then we need to add another layer to the config to ensure these packets that we don't want to drop are also marked properly. In this case, we need to remark traffic with a specific DSCP value.

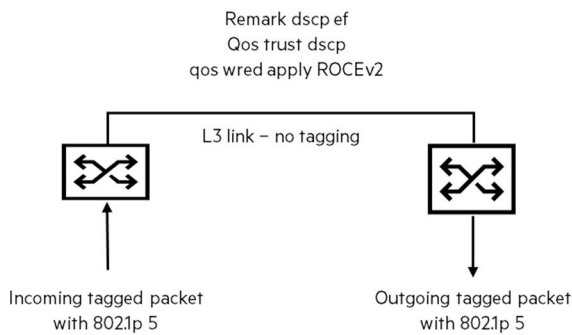


Figure 16. End to end marking

Configuration examples

The Comware example below creates a new QoS policy and it applies it to the L3 interface. There are many ways to identify and remark the required traffic. The example below identifies the RoCE traffic by selecting IP Source and IP Destinations which in this case match the SMB direct servers. The shown QoS policy remarks the traffic with a DSCP based value of EF which is carried in the IP header. Now that the selected traffic gets marked with DSCP value of EF we also need to modify the DSCP to dot1p mapping table.

The switch will see the DSCP value of EF and go to the DSCP-802.1p mapping table and understand that it equals traffic marked with 802.1p value of 5. It then looks at the 802.1p-lp mapping table (configured earlier) to understand that this traffic gets placed into queue 1.

PFC is applied as a no-drop to 802.1p value of 5 and IP ECN is applied to queue 1.

```
acl number 3003 name DSCP
```

```

rule 0 permit ip destination 9.1.1.10.0
rule 5 permit ip source 8.1.1.10.0
#
traffic classifier DSCP operator and
  if-match acl 3003
#
traffic behavior DSCP
  remark dscp ef
#
qos policy DSCP
  classifier DSCP behavior DSCP
#
qos map-table dscp-dot1p
#
...
import 8 export 0
import 9 export 0
import 46 export 5
...
#
interface Ten-GigabitEthernet1/0/10
  qos apply policy DSCP

```

Summary

In summary, utilizing RDMA solutions such as RoCE can lead to increased performance, reduced latency, as well as lower CPU utilization.

Converged solutions, such as RoCE, carry both lossless and lossy traffic and they are gaining in popularity in modern data centers. This is thanks in large part to the growth of Hyper-convergence, IP storage and of course RoCE based solutions which have also grown thanks to solutions like SMB direct, and more.

When designing any DC, but especially a converged DC, one main point should be to keep it simple. Convergence at the ToR is the first and obvious part of a network to converge. Modern servers are now sending up massive amount of various types of traffic, some requiring lossless behavior, while other types of traffic is ok with packet drops. Upwards of 80% of the cabling in data centers is at the ToR so converging the ToR helps to reduce capital and operational expenses. This is because of fewer switches, lower power consumption, reduced cooling, and reduced NIC/HBAs and cables.

From there the simplest solution is to break off the lossless traffic from the lossy traffic at the ToR. In storage environments, this is commonly achieved by using a SAN core for the lossless traffic and the normal LAN core for lossy traffic.

However, more and more solutions, like Hyper-converged solutions may require lossless traffic to traverse the main LAN network core. The HPE Networking environments can support this, however, keep in mind that even though we can configure the network to transport lossless traffic in an efficient manner we cannot forget to also provision the network for peak loads.

As discussed, protocols like PFC can create a cascading pausing affect which can affect unrelated traffic. Ensuring the network links are provisioned for peak load times will help to reduce excessive congestion.

Ensuring traffic is segregated into proper queues can also help. As an example, perhaps the environment has chatty servers, initiators, and targets. This might mean moving chatty servers, initiators and targets into a different queue so that pauses in that queue will not affect other traffic.

Table 1. Summary.

Feature	RoCEv1	RoCEv2	Notes
---------	--------	--------	-------

Priority-based Flow Control	YES	YES	Must use always. If not used, in times of congestion the RDMA advantages will not be achieved.
Enhanced Transmission Selection (ETS)	YES	YES	Must use in converged environments. If not used, lossless traffic classes can get starved for bandwidth.
Data Center Bridging Exchange (DCBX)	YES	YES	Not Mandatory but recommended.
Quantized Congestion Notification (QCN)	YES	NO	Highly recommended for L2 RoCEv1. CN helps to address pause unfairness and victim flow issues. When used with PFC, PFC acts as fast acting mechanism to address microbursts, while CN smooths out traffic flows helping to reduce pause storms under normal load.
IP Explicit Congestion Notification (ECN)	NO	YES	Highly recommended for L2 RoCEv2. ECN will help to address pause unfairness and victim flow issues. When used with PFC, PFC acts as fast acting mechanism to address microbursts, while CN smooths out traffic flows helping to reduce pause storms under normal load.

Resources, contacts, or additional links

[HPE Data Center Networking](#)

[HPE FlexFabric Networking](#)

[HPE Storage](#)

[HPE Servers](#)



a Hewlett Packard
Enterprise company

www.arubanetworks.com

3333 Scott Blvd. | Santa Clara, CA 95054

1.844.472.2782 | T: 1.408.227.4500 | FAX: 1.408.227.4550 | info@arubanetworks.com