RDMA 技术白皮书

目　录

1 概述 1
1.1 RDMA 产生背景 1
1.2 RDMA 技术优势 1
1.3 RDMA 技术分类 2
2 RDMA 技术概述 2
2.1 IB 技术 2
2.1.1 IB 技术简介 2
2.1.2 IB 技术特点 2
2.2 iWARP 技术 3
2.2.1 iWARP 简介 3
2.2.2 iWARP 技术特点 3
2.3 RoCE 技术 3
2.3.1 RoCE 简介 3
2.3.2 RoCE 技术特点 4
3 构建无损以太网 4
3.1 无损以太网关键特性 4
3.2 PFC 5
3.2.1 PFC 简介 5
3.2.2 PFC 工作机制 6
3.2.3 PFC 扩展功能 8
3.3 ECN 功能 11
3.3.1 ECN 简介 11
3.3.2 ECN 工作机制 11
3.4 DCBX 12
3.4.1 DCBX 简介 12
3.4.2 DCBX 工作机制 12
3.5 ETS 13
3.5.1 ETS 简介 13
3.5.2 ETS 工作机制 13
4 构建无损以太网配置举例 14

1 概述

1.1 RDMA产生背景

随着高性能计算、大数据分析、人工智能以及物联网等技术的飞速发展，集中式存储、分布式存储以及云数据库的普及等原因，业务应用有越来越多的数据需要从网络中获取，这对数据中心网络的交换速度和性能要求越来越高。

传统的 TCP/IP 软硬件架构及应用存在着网络传输和数据处理的延迟过大、存在多次数据拷贝和中断处理、复杂的 TCP/IP 协议处理等问题。RDMA（Remote Direct Memory Access，远程直接内存访问）是一种为了解决网络传输中服务器端数据处理延迟而产生的技术。RDMA 将用户应用中的数据直接传入服务器的存储区，通过网络将数据从一个系统快速传输到远程系统的存储器中，消除了传输过程中多次数据复制和文本交换的操作，降低了 CPU 的负载。

图1 传统 TCP/IP 数据传输过程

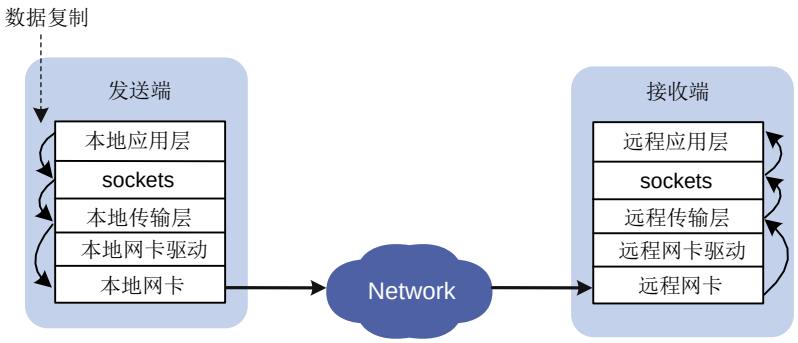
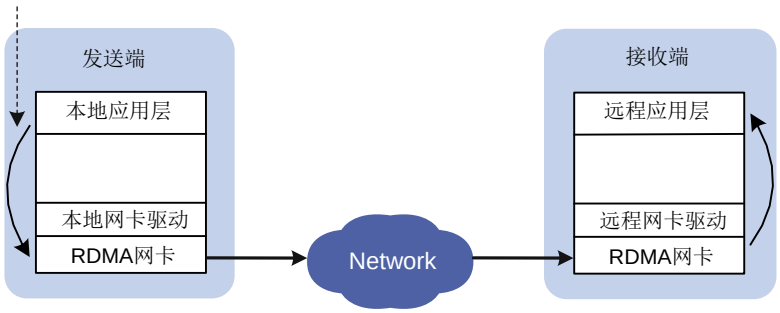


图2 RDMA 数据传输过程



1.2 RDMA技术优势

RDMA 技术实现了在网络传输过程中两个节点之间数据缓冲区数据的直接传递，在本节点可以直接将数据通过网络传送到远程节点的内存中，绕过操作系统内的多次内存拷贝，相比于传统的网络传输，RDMA 无需操作系统和 TCP/IP 协议的介入，可以轻易的实现超低延时的数据处理、超高吞吐量传输，不需要远程节点 CPU 等资源的介入，不必因为数据的处理和迁移耗费过多的资源。

1.3 RDMA技术分类

RDMA 技术主要包括：

- IB（InfiniBand）：基于 InfiniBand 架构的 RDMA 技术，由 IBTA（InfiniBand Trade Association）提出。搭建基于 IB 技术的 RDMA 网络需要专用的 IB 网卡和 IB 交换机。
- iWARP（Internet Wide Area RDMA Protocal）：基于 TCP/IP 协议的 RDMA 技术，由 IETF 标准定义。iWARP 支持在标准以太网基础设施上使用 RDMA 技术，但服务器需要使用支持 iWARP 的网卡。
- RoCE（RDMA over Converged Ethernet）：基于以太网的 RDMA 技术，也是由 IBTA 提出。RoCE 支持在标准以太网基础设施上使用 RDMA 技术，但是需要交换机支持无损以太网传输，需要服务器使用 RoCE 网卡。

H3C 以太网交换机能够支持 iWARP，其中部分系列（具体系列请咨询市场技术人员或查看产品配套资料）支持无损以太网传输的关键技术，能够支持 RoCE。

2 RDMA 技术概述

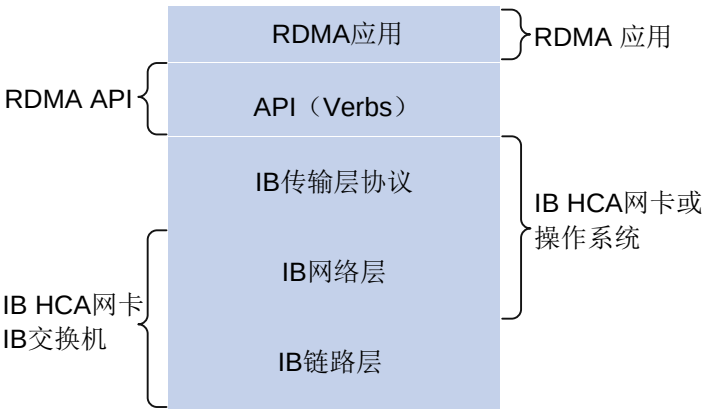
2.1 IB技术

2.1.1 IB 技术简介

InfiniBand 是一种基于 InfiniBand 架构的 RDMA 技术，它提供了一种基于通道的点对点消息队列转发模型，每个应用都可通过创建的虚拟通道直接获取本应用的数据消息，无需其他操作系统及协议栈的介入。InfiniBand 架构的应用层采用了 RDMA 技术，可以提供远程节点间 RDMA 读写访问，完全卸载 CPU 工作负载；网络传输采用了高带宽的传输；链路层设置特定的重传机制保证服务质量，不需要数据缓冲。

InfiniBand 必须运行在 InfiniBand 网络环境下，必须使用 IB 交换机及 IB 网卡才可实现。

图3 InfiniBand 架构



2.1.2 IB 技术特点

InfiniBand 技术具有以下特点：

- 应用层采用 RDMA 技术，降低了在主机侧数据处理的延迟。
- 消息转发控制由子网管理器完成，没有类似以太网复杂的协议交互计算。
- 链路层通过重传机制保证服务质量，不需要数据缓冲，无丢包。
- 具有低延迟、高带宽、低处理开销的特点。

2.2 iWARP技术

2.2.1 iWARP 简介

iWARP 是基于以太网和 TCP/IP 协议的 RDMA 技术，可以运行在标准的以太网基础设施上。

iWARP 由 MPA、DDP、RDMAP 三层子协议组成：

- RDMAP 层协议负责 RDMA 读、写操作和 RDMA 消息的转换，并将 RDMA 消息转发到 DDP 层。
- DDP 层协议负责将过长的 RDMA 消息分片分装成 DDP 数据包继续转发到 MPA 层。
- MPA 层在 DDP 数据段的固定标识位置增加转发后向标识、数据报文的长度以及 CRC 校验数据等字段构成 MPA 数据段交由 TCP 传输。

2.2.2 iWARP 技术特点

iWARP 从以下几个方面降低了主机侧网络负载：

- TCP/IP 处理流程从 CPU 卸载到 RDMA 网卡处理，降低了 CPU 负载。
- 消除内存拷贝：应用程序可以直接将数据传输到对端应用程序内存中，显著降低 CPU 负载。
- 减少应用程序上、下文切换：应用程序可以绕过操作系统，直接在用户空间对 RDMA 网卡下发命令，降低了开销，显著降低了应用程序上、下文切换造成的延迟。

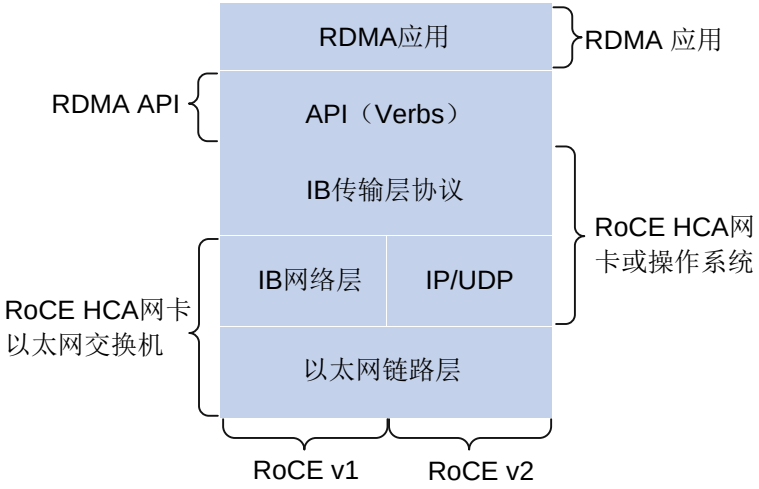
由于 TCP 协议能够提供流量控制和拥塞管理，因此 iWARP 不需要以太网支持无损传输，仅通过普通以太网交换机和 iWARP 网卡即可实现，因此能够在广域网上应用，具有较好的扩展性。

2.3 RoCE技术

2.3.1 RoCE 简介

RoCE 技术支持在以太网上承载 IB 协议，实现 RDMA over Ethernet。RoCE 与 InfiniBand 技术有相同的软件应用层及传输控制层，仅网络层及以太网链路层存在差异。

图4 RoCE 架构



RoCE 协议分为两个版本：

- RoCE v1 协议：基于以太网承载 RDMA，只能部署于二层网络，它的报文结构是在原有的 IB 架构的报文上增加二层以太网的报文头，通过 Ethertype 0x8915 标识 RoCE 报文。
- RoCE v2 协议：基于 UDP/IP 协议承载 RDMA，可部署于三层网络，它的报文结构是在原有的 IB 架构的报文上增加 UDP 头、IP 头和二层以太网报文头，通过 UDP 目的端口号 4791 标识 RoCE 报文。RoCE v2 支持基于源端口号 hash，采用 ECMP 实现负载分担，提高了网络的利用率。

2.3.2 RoCE 技术特点

RoCE 使得基于以太网的数据传输能够：

- 提高数据传输吞吐量。
- 减少网络延时。
- 降低 CPU 负载。

RoCE 技术可通过普通以太网交换机实现，但服务器需要支持 RoCE 网卡，网络侧需要支持无损以太网络，这是由于 IB 的丢包处理机制中，任意一个报文的丢失都会造成大量的重传，严重影响数据传输性能。

3 构建无损以太网

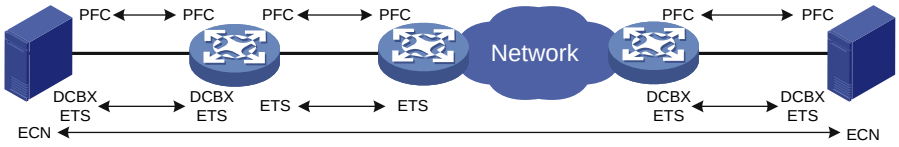
3.1 无损以太网关键特性

在 RoCE 网络中，我们需要构建无损以太网用于保证网络传输过程中不丢包。构建无损以太网需支持以下关键特性：

- （必选）PFC（Priority-based Flow Control，基于优先级的流量控制）：逐跳提供基于优先级的流量控制，能够实现在以太网链路上运行多种类型的流量而互不影响。

- （必选）ECN（Explicit Congestion Notification，显示拥塞通知）：设备发生拥塞时，通过对报文 IP 头中 ECN 域的标识，由接收端向发送端发出降低发送速率的 CNP（Congestion Notification Packet，拥塞通知报文），实现端到端的拥塞管理，减缓拥塞扩散恶化。
- （建议）DCBX（Data Center Bridging Exchange Protocol，数据中心桥能力交换协议）：使用 LLDP 自动协商 DCB 能力参数，包括 PFC 和 ETS 等。一般用在接入交换机连接服务器的端口，与服务器网卡进行能力协商。
- （可选）ETS（Enhanced Transmission Selection，增强传输选择）：将流量按服务类型分组，在提供不同流量的最小带宽保证的同时提高链路利用率，保证重要流量的带宽百分比。需要逐跳提供。

图5 构建无损以太网络关键特性组网示意图



在 RoCE 环境中，PFC 与 ECN 需要同时使用，以在无丢包情况下带宽得到保证。二者的功能对比如表 1 所示。

表1 PFC 与 ECN 对比

比较项目	PFC	ECN
网络位置	二层	网络层及传输层
作用范围	点到点	端到端
是否需要全网支持	是	否
被控制对象	网络中上一节点，如果服务器网卡支持PFC，PFC对网卡也能生效	发送主机
报文缓存位置	中间网络节点及发送端	发送端
受影响的流量	网络设备中8个转发队列中某个队列的所有流量	发生拥塞应用的连接
响应速度	快	慢

3.2 PFC

3.2.1 PFC 简介

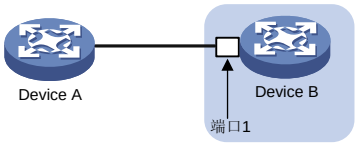
PFC 是构建无损以太网的必选手段之一，能够逐跳提供基于优先级的流量控制。设备在进行报文转发时，根据报文的优先级进入对应映射关系的队列中进行调度转发。当某一优先级报文发送速率超过接收速率，导致接收方可用数据缓冲空间不足时，设备通过 PFC PAUSE 帧反馈给上一跳设备，

上一跳设备收到 PAUSE 帧报文后停止发送本优先级报文，直到再收到 PFC XON 帧或经过一定的老化时间后才能恢复流量发送。通过使用 PFC 功能，使得某种类型的流量拥塞不会影响其他类型流量的正常转发，从而达到同一链路上不同类型的报文互不影响。

3.2.2 PFC 工作机制

1. PFC PAUSE 帧生成机制

图6 PFC 功能 PAUSE 帧产生示意图



如图 6 所示，PAUSE 帧产生过程：

(1) Device B 的端口 1 收到来自 Device A 的报文后，MMU（Memory Manage Unit，存储器管理单元）会为该报文分配 cell 资源，当设备的 PFC 功能处于开启状态时，会根据报文中的 dot1p 优先级统计占用的 cell 资源。

说明

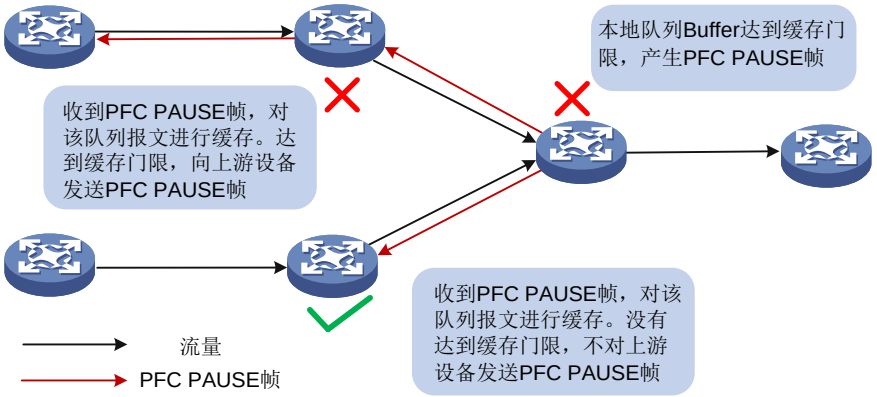
cell 资源：用来存储数据包的内容，端口会根据报文的实际大小占用相应大小的 cell 资源。比如一个 cell 资源是 208 字节，当发送的报文是 128 字节时，端口会给它分配一个 cell 资源，当发送的报文是 300 字节时，端口会给它分配两个 cell 资源。

---

(2) 当 Device B 端口 1 的某个优先级的报文占用的 cell 资源统计计数达到设置的门限后，再收到新的该优先级报文后，端口 1 会发送对应优先级的 PFC PAUSE 帧给 Device A。

(3) Device A 收到该优先级的 PFC PAUSE 帧后停止发送对应优先级的报文，对该优先级的报文进行缓存，如果触发了缓存门限，则也向其上游设备发送 PFC PAUSE 帧，如图 7 所示。

图7 多跳设备之间的 PFC PAUSE 帧处理



2. 报文优先级与队列映射关系

设备在进行报文转发时，将不同优先级的报文放入不同的队列中进行调度转发。报文优先级与队列映射关系与设备配置的优先级映射方式有关。设备支持的优先级映射配置方式包括：优先级信任模式方式、端口优先级方式。

- 优先级信任模式方式

  配置端口的优先级信任模式后，设备将信任报文自身携带的优先级。通过优先级映射表，使用所信任的报文携带优先级进行优先级映射，根据映射关系完成对报文优先级的修改，以及实现报文在设备内部的调度。端口的优先级信任模式分为：

  - **dot1p**：信任报文自带的 802.1p 优先级，以此优先级进行优先级映射。
  - **dscp**：信任 IP 报文自带的 DSCP 优先级，以此优先级进行优先级映射。

- 端口优先级方式

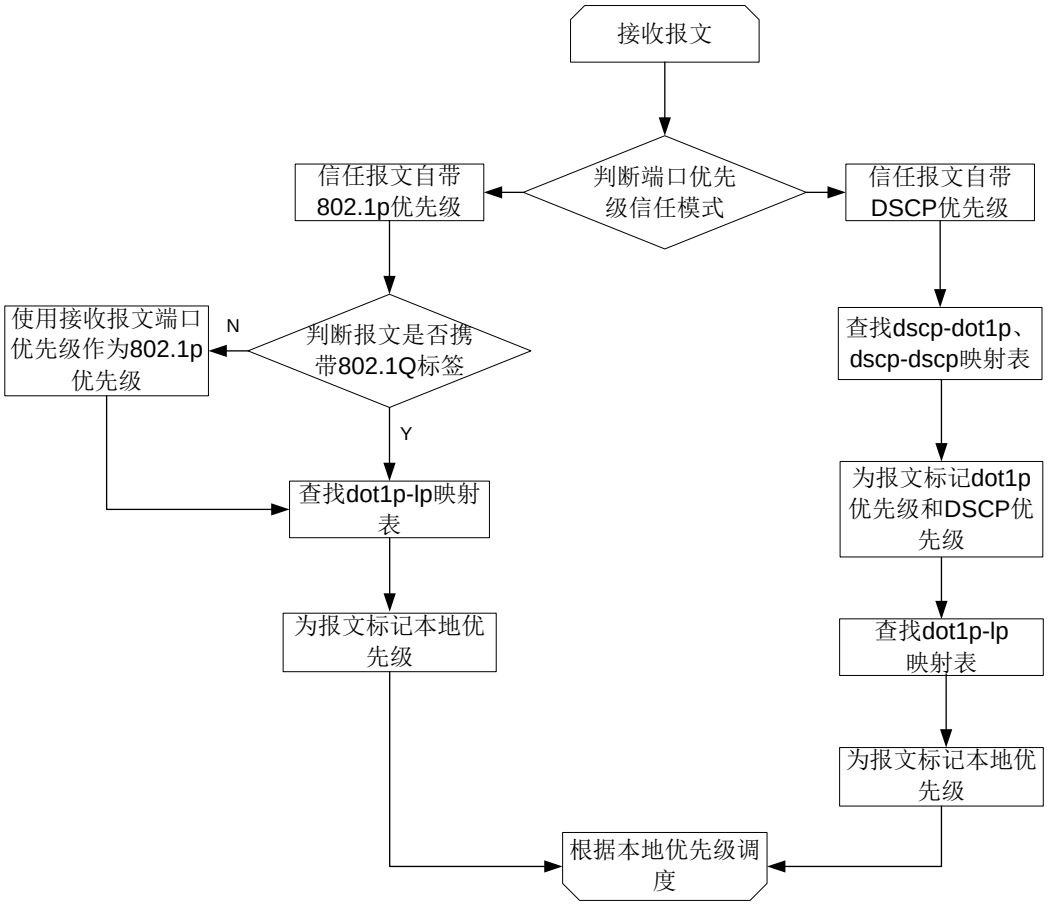
  未配置端口的优先级信任模式时，设备会将端口优先级作为报文自身的优先级。通过优先级映射表，对报文进行映射。用户可以配置端口优先级，通过优先级映射，使不同端口收到的报文进入对应的队列，以此实现对不同端口收到报文的差异化调度。

接口配置 PFC 功能时，必须配置接口信任报文自带的 802.1p 优先级或 DSCP 优先级。接口收到以太网报文，根据优先级信任模式和报文的 802.1Q 标签状态，设备为不同优先级的报文标记不同的本地优先级（LP），根据本地优先级进行队列调度，具体过程如图 8 所示。

说明

本文仅介绍接口信任报文自带的 802.1p 优先级或 DSCP 优先级的情况下，报文优先级到本地优先级的映射情况，关于端口采用端口优先级时的映射情况和报文丢弃时参考的丢弃优先级请参考产品配置指导。

图8 报文优先级与队列映射关系



需要注意的是：配置 PFC 功能时，必须配置接口信任报文自带的 802.1p 优先级或 DSCP 优先级，并且转发路径上所有端口的 802.1p 优先级与本地优先级映射关系以及 DSCP 优先级与 802.1p 优先级映射关系必须一致，否则 PFC 功能将无法正常工作。

3.2.3 PFC 扩展功能

1. PFC 门限配置

通过配置 PFC 缓存门限可以有效解决因缓冲空间不足和入流量队列数量过大，导致发送数据缓冲区尾丢弃等问题。

我们先来了解一下接口的缓冲空间设置。接口的缓冲空间分为以下几种：

- Guaranteed 存储空间：固定缓冲区，为每一个优先级队列和端口提供最小的缓存保证。系统会根据用户的配置给队列预留指定大小的空间，即便该队列没有报文存储需求，其他队列也不能抢占。给队列预留的空间均分给每个端口的，即使某端口的某队列没有报文存储需求，其他端口也不能抢占。
- Shared 存储空间：共享缓冲区，当端口或优先级的固定缓冲区不够用时，使用 Shared 存储空间，系统会根据用户配置以及实际需要收发报文的数量决定每个队列实际可占用的缓冲区的大小。如果某个队列没有报文存储需求，则其他队列会抢占该队列的配额。对于某个队列

的缓冲区，所有端口接收或发送的报文采用抢占的方式，先到先得，如果资源耗尽，则后到达的报文将被丢弃。

- Headroom 存储空间：Headroom 缓冲区，当端口 PFC 功能生效并触发 PFC 反压帧门限后，本端设备发送 PFC PAUSE 帧到对端设备让对端停止流量发送的过程中，已经在途的这部分流量的缓存空间，设备需要这些缓冲空间来保证 PFC 流程的不丢包。

PFC 目前提供以下门限设置：

- Headroom 缓存门限：Headroom 存储空间中某 802.1p 优先级报文的最大使用 cell 资源。当达到使用的 cell 资源后，该接口会丢弃收到的报文。
- 反压帧触发门限：Shared 存储空间中某 802.1p 优先级报文在该存储空间使用 cell 资源上限。达到上限后，会触发 PFC 功能发送 PAUSE 帧。反压帧触发门限又分为动态反压帧触发门限和静态反压帧触发门限：
  - 动态反压帧触发门限：设置某 802.1p 优先级报文触发 PFC PAUSE 帧的可用 cell 资源的百分比。
  - 静态反压帧触发门限：设置某 802.1p 优先级报文触发 PFC PAUSE 帧的可用 cell 资源门限为一个固定值。
- 反压帧停止门限与触发门限间的偏移量：当某 802.1p 优先级报文使用的 cell 资源减小了一个固定值时，停止发送 PFC PAUSE 帧，使对端设备恢复流量发送。
- PFC 预留门限：Guaranteed 存储空间中为某 802.1p 优先级报文预留的 cell 资源。
- Headroom 最大可用的 cell 资源：配置某缓存池（pool，产品具体支持的 poolID 与产品型号有关，请以设备的实际情况为准）中，分配给 Headroom 存储空间的 cell 资源的大小。

具体配置命令行请参见对应产品的配置指导和命令参考。

注意

Headroom 缓存门限的建议配置值与接口传输速率和距离有关，具体建议值请参考产品命令手册。对于其他门限值，开启指定 802.1p 优先级的 PFC 功能后，设备会为 PFC 的各种门限设置一个缺省值，此缺省值在一般的组网环境下是效果较好的参数组合，一般不建议调整。如组网环境或流量背景确实非常复杂，建议咨询专业人员进行调整。

---

2. PFC 死锁检测功能

当指定优先级的报文形成环路时，会导致数据缓冲区内报文无法转发，设备间反复发送和接收 PFC 帧，导致设备接口的缓冲区 cell 资源一直被占用无法释放，此时设备进入 PFC 死锁状态。设备处于 PFC 死锁状态后，采用关闭 PFC 功能或者忽略接收到的 PFC XOFF 帧（表示停止流量发送）的方式使设备继续转发数据报文即可解除死锁。

通过配置 PFC 死锁检测功能，可以定期检测设备是否处于 PFC 死锁状态。当设备检测到 PFC 死锁状态后，设备会在恢复周期内自动解除死锁状态。此时设备会自动暂时禁用 PFC 死锁检测功能，以便报文能够正常转发，解除死锁状态。

PFC 死锁状态解除后，用户可采用自动或手工方式来恢复 PFC 死锁检测功能。恢复 PFC 死锁检测功能会让 PFC 功能继续生效。所以，通常情况下，使用自动恢复方式即可。当报文环路无法消除，设备频繁处于 PFC 死锁状态时，用户可以进入以太网接口视图配置 PFC 死锁检测功能的恢复方式为手工恢复方式，并尽快排除故障，再手工恢复 PFC 死锁检测功能。

更多 PFC 死锁检测功能的配置命令请参见对应产品的配置指导和命令参考。

3. PFC 一键逃生功能

当设备的 PFC 功能出现紧急故障时，用户不用逐个接口去关闭 PFC 功能，可以通过命令行一键关闭所有接口的 PFC 功能。当故障恢复后，可以通过命令行一键开启所有接口的 PFC 功能。具体的命令行形式与设备的支持情况及版本有关，请以设备的实际情况为准。

4. PFC 报文的预警门限

用户可根据实际组网情况，配置接口入方向或者出方向 PFC 报文的预警门限。预警门限用于 PFC 报文传输速率处于正常范围内，但需要提醒用户提前关注的情况。

当接口接收或发送 PFC 报文的速率达到预警门限时，系统会生成 Trap 和日志信息来提醒用户，以提前发现网络中的一些异常问题。例如：

- 对端设备网卡故障，不停地持续高速发送 PFC 帧，可以配置入方向预警门限进行监控。
- 本设备故障后不停发送 PFC 帧，可以配置出方向预警门限进行监控。
- 如果有双向监控需求的，可以在入和出方向都配置预警门限进行监控。

PFC 报文的预警门限配置的具体命令介绍请参见产品命令手册。

5. 配合 gRPC 实现统计、告警信息上报

PFC 配合 gRPC 可以实现丢包主动上报告警，超限主动上报告警，同时提供各种丢包和瞬时使用值供统计查询。

支持上报的统计信息包括：

- ingress/egress 丢包总量
- RX/TX PFC 帧总量及速率
- ingress/egress buffer 使用
- headroom buffer 使用
- ingress/egress buffer 超限次数
- headroom 超限次数。
- 基于 XPE 统计的 buffer 利用率
- ECN 功能 marked 次数
- WERD 丢包总量

支持上报的告警信息包括：

- ingress/egress 丢包
- headroom buffer 超限
- ingress buffer 超限
- egress buffer 超限。

6. 丰富的诊断维护功能

使用 **`display priority-flow-control`** 命令可以显示查看 PFC 功能在端口上的配置情况及每端口和每队列的 PFC 帧的收发总量和收发速率。

使用 **`display packet-drop`** 命令可以对接收及发送端丢包的总信息及各端口丢包信息进行诊断查询。

使用 **display qos queue-statistics interface outbound** 命令可以显示端口队列出方向的统计信息。

3.3 ECN功能

3.3.1 ECN 简介

ECN 是构建无损以太网的必选手段之一。ECN 定义了一种基于 IP 层及传输层的流量控制及端到端拥塞通知机制。ECN 功能利用 IP 报文头中的 DS 域来标记报文传输路径上的拥塞状态。支持该功能的终端设备可以通过报文内容判断出传输路径上发生了拥塞，从而调整报文的发送方式，避免拥塞加剧。

3.3.2 ECN 工作机制

ECN 功能对 IP 报文头中 DS 域的最后两个比特位（称为 ECN 域）进行了如下定义：

- 比特位 6 用于标识发送端设备是否支持 ECN 功能，称为 ECT 位（ECN-Capable Transport）
- 比特位 7 用于标识报文在传输路径上是否经历过拥塞，称为 CE 位（Congestion Experienced）

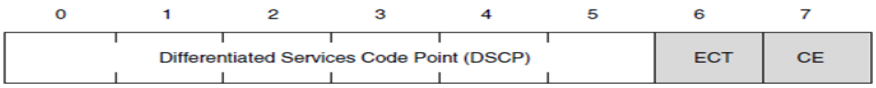
说明

在实际应用中，设备将 ECT 位为 1、CE 位为 0 的报文，以及 ECT 位为 0，CE 位为 1 的报文都识别为由支持 ECN 功能的终端发出的报文。

图9 DS 域位置信息

bit offset	0–3	4–7	8–15	16–18	19–31
0	Version	Header length	Differentiated Services	Total Length	
32	Identification			Flags	Fragment Offset
64	Time to Live		Protocol	Header Checksum	
96	Source Address				
128	Destination Address				
160	Options (if Header Length > 5)				
160 or 192+	Data				

图10 ECN 域位置信息



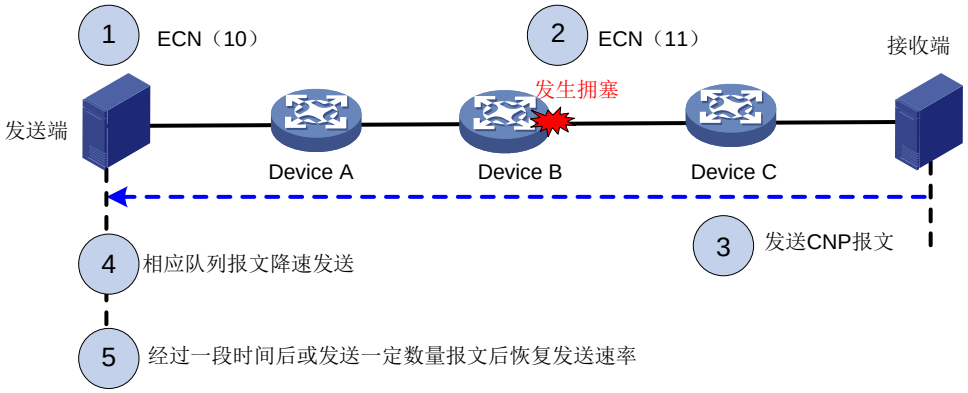
在设备上开启 ECN 功能后，拥塞管理功能将按如下方式对报文进行处理：

- 如果队列长度小于下限，不丢弃报文，也不对 ECN 域进行识别和标记。
- 如果队列长度在上限和下限之间，当设备根据丢弃概率计算出需要丢弃某个报文时，将检查该报文的 ECN 域。如果 ECN 域显示该报文由支持 ECN 的终端发出，设备会将报文的 ECT 位和 CE 位都标记为 1，然后转发该报文；如果 ECN 域显示报文传输路径中已经经历过拥塞

（即 ECT 和 CE 位都为 1），则设备直接转发该报文，不对 ECN 域进行重新标记；如果 ECT 位和 CE 位都为 0，设备会将该报文丢弃。

- 如果队列长度超过上限，将队列中所有报文的 ECN 域都标记为 11，当队列长度达到队列尾丢弃门限后,报文将被丢弃。

图11 ECN 工作机制示意图



ECN 功能工作机制：

(1) 发送端设置 ECN 域为 10，告知路径上的设备及接收端，发送端设备支持 ECN 功能。

(2) 中间设备发生拥塞并达到门限，拥塞设备将发生拥塞的报文 ECN 域设置为 11，报文正常转发。

(3) 接收端收到 ECN 置位为 11 的报文，由传输层发送 CNP（Congestion Notification Packet，拥塞通知报文）通知发送端。

(4) 发送端收到 CNP 报文，对对应的优先级的队列进行降速处理。

(5) 经过一段可配置的时间或者发送一定数量数据，发送端恢复原来的速率。

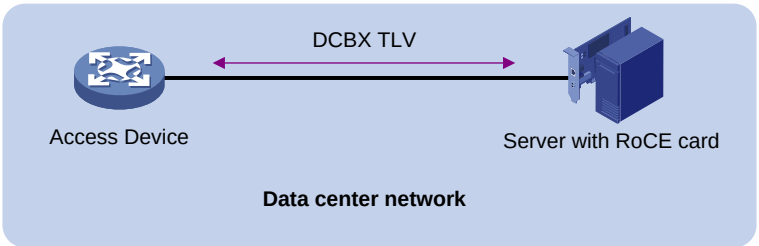
3.4 DCBX

3.4.1 DCBX 简介

DCBX 是实现无损以太网传输的关键手段之一，用于 DCE 中各网络单元进行桥能力协商以及远程配置。通过 DCBX，交换机之间以及交换机和网卡之间可以协商和自动配置 DCB 参数，以实现简化配置以及保证配置一致性的目的。

3.4.2 DCBX 工作机制

图12 图 2-2 DCBX 功能配置组网图



DCBX 通过 LLDP（Link Layer Discovery Protocol，链路层发现协议）来完成信息交互的过程，支持交互链路双方的 ETS、PFC 以及应用的优先级等配置信息。

在设备全局和接口上都开启 LLDP 功能并允许接口发布 DCBX TLV 即可开启设备的 DCBX 功能，再根据实际应用需求配置设备通过该接口发布 APP（Application Protocol，应用协议）、ETS 和 PFC 参数。在本文的应用需求中，主要用于发布 ETS 参数。

配置 DCBX 时，还需要注意配置 DCBX 的版本，可手工配置或自协商 DCBX 版本。DCBX 版本需要视对端设备支持的版本而定，要求两端端口的 DCBX 版本一致，否则版本无法兼容，将会导致 DCBX 无法正常工作。

3.5 ETS

3.5.1 ETS 简介

ETS 是基于优先级组的带宽分配处理，ETS 用于实现承诺带宽。设备通过 ETS 参数与对端进行协商，控制对端指定类型数据的发送带宽，保证其在接口的承诺带宽范围之内，从而不会因流量拥塞而导致数据丢失。

3.5.2 ETS 工作机制

ETS 机制将网络中的流量优先级分成不同的优先级组（Priority Group），为每个优先级组分配一定的带宽，如果一个优先级组未消耗为其分配的带宽其他优先级组可以使用这些未使用的带宽。保证重要流量在传输过程中具有承诺带宽。

配置 ETS 功能的具体实现过程为：

(1) 配置 802.1p 优先级到本地优先级的映射使报文进入特定的队列。有以下两种方式：

- QoS 策略方式。
- 优先级映射表方式。

如果同时配置了这两种方式，则前者的配置优先生效。有关 QoS 策略命令的详细介绍，请参见“ACL 和 QoS 命令参考”中的“QoS 策略”。有关优先级映射表命令的详细介绍，请参见“QoS 命令参考”中的“优先级映射”。

(2) 配置分组 WRR 队列，以实现不同队列带宽的分配。有关 WRR 队列命令的详细介绍，请参见“QoS 命令参考”中的“拥塞管理”。

为了更好的理解为什么 ETS 能够实现重要流量的带宽保证，这里我们详细介绍一下分组 WRR 队列。

我们先来了解一下 WRR 队列。WRR 队列在队列之间进行轮流调度，保证每个队列都得到一定的服务时间。以端口有 8 个输出队列为例，WRR 可为每个队列配置一个加权值（依次为 w7、w6、w5、w4、w3、w2、w1、w0），加权值表示获取资源的比重。如一个 100Mbps 的端口，配置它的 WRR 队列的加权值为 50、50、30、30、10、10、10、10（依次对应 w7、w6、w5、w4、w3、w2、w1、w0），这样可以保证最低优先级队列至少获得 5Mbps 的带宽。

WRR 队列还有一个优点是，虽然多个队列的调度是轮询进行的，但对每个队列不是固定地分配服务时间片——如果某个队列为空，那么马上换到下一个队列调度，这样带宽资源可以得到充分的利用。

WRR 队列分为：

- 基本 WRR 队列：基本 WRR 队列包含多个队列，用户可以定制各个队列的权重，WRR 按用户设定的参数进行加权轮询调度。
- 分组 WRR 队列：所有队列全部采用 WRR 调度，用户可以根据需要将输出队列划分为 WRR 优先级队列组 1 和 WRR 优先级队列组 2。进行队列调度时，设备首先在 WRR 优先级队列组 1 中进行轮询调度；优先级队列组 1 中没有报文发送时，设备才在优先级队列组 2 中进行轮询调度。当前设备仅支持 WRR 优先级队列组 1。

在分组 WRR 队列中，也可以配置队列加入 SP 分组，采用严格优先级调度算法（即只有较高优先级队列为空时，才会发送较低优先级队列中的分组，最大限度保证关键业务流量的发送）。调度时先调度 SP 组，然后调度其他 WRR 优先组。

在 ETS 的配置中，为了保证重要流量的发送带宽，我们可以采用如下两种方式之一配置分组 WRR 队列：

- 配置 WRR 优先组 1 的 WRR 队列调度权重，使重要流量所在队列拥有较高权重。

  qos wrr *queue-id* **group 1 byte-count** *schedule-value*
- 配置端口队列采用严格优先级调度算法，使重要流量所在队列优先调度。

  qos wrr *queue-id* **group sp**

4 构建无损以太网配置举例

1. 组网需求

如图 13 所示，服务器 Server 1、Server 2 和 Server 3 上均安装 RoCE 网卡，Server 1 和 Server 2 均通过以太网交换机 Device A 和 Device B 连接 Server 3。

为支持 RoCE 技术，现要求将整个网络搭建为无损以太网，具体要求为：

- 报文转发路径的所有端口都开启 PFC 功能，本例以实现 802.1p 优先级为 5 的报文的无损传输为例。
- 交换机连接服务器的端口开启 DCBX 功能，使设备和服务器网卡可以协商 ETS 和 PFC 参数。
- Device A 的 Twenty-FiveGigE1/0/3 和 Device B 的 Twenty-FiveGigE1/0/2 端口配置 ETS 功能，保证 802.1p 优先级为 5 的报文的发送带宽。

  说明

  本例所示的组网中，我们认为 Server 1、Server 2 发往 Server 3 的流量大于反向流量，所以仅在上述端口配置 ETS 功能，实际应用中如果无法预测流量发送情况，可以在组网中的所有端口上配置 ETS 功能。

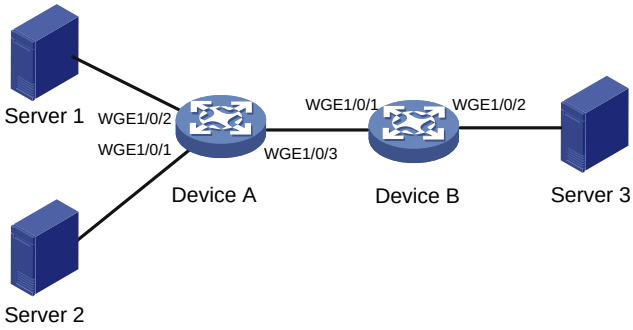
- Device A 的 Twenty-FiveGigE1/0/3 端口配置 ECN 功能，使设备在发生拥塞时能够通知发送端调整发送速率。

说明

本例所示的组网中，拥塞的可能发生位置为 Device A 的 Twenty-FiveGigE1/0/3 端口，因此仅在该端口配置 ECN 功能。实际应用中如果无法预测拥塞发生的可能位置，可以在组网中的所有端口上配置 ECN 功能。

2. 组网图

图13 RoCE 功能配置组网图



3. 配置步骤

(1) 配置 Device A

在接口 Twenty-FiveGigE1/0/1、Twenty-FiveGigE1/0/2、Twenty-FiveGigE1/0/3 上配置接口信任报文自带的的 802.1p 优先级，开启接口的 PFC 功能，并对 802.1p 优先级 5 开启 PFC 功能。

```
<DeviceA> system-view
[DeviceA] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/3
[DeviceA-if-range] qos trust dot1p
[DeviceA-if-range] priority-flow-control enable
[DeviceA-if-range] priority-flow-control no-drop dot1p 5
[DeviceA-if-range] quit
```

全局开启 LLDP 功能。

```
[DeviceA] lldp global enable
```

在接口 Twenty-FiveGigE1/0/1、Twenty-FiveGigE1/0/2 上开启 LLDP 功能，并允许发布 DCBX TLV，配置接口 Twenty-FiveGigE1/0/1、Twenty-FiveGigE1/0/2 的 DCBX 版本为预标准版 1.01。

```
[DeviceA] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/2
[DeviceA-if-range] lldp enable
[DeviceA-if-range] lldp tlv-enable dot1-tlv dcbx
[DeviceA-if-range] dcbx version rev101
[DeviceA-if-range] quit
```

在接口 Twenty-FiveGigE1/0/3 上开启 WRR 队列，并按照每次轮询可发送的字节数进行计算，同时配置端口队列 5（802.1p 优先级 5 到本地优先级 5 为默认映射关系）采用严格优先级调度算法。

```
[DeviceA] interface twenty-fivegige 1/0/3
[DeviceA-Twenty-FiveGigE1/0/3] qos wrr byte-count
[DeviceA-Twenty-FiveGigE1/0/3] qos wrr 5 group sp
[DeviceA-Twenty-FiveGigE1/0/3] quit
```

创建 WRED 表 queue-table5，同时进入 WRED 表视图配置队列 5 的平均队列长度指数及 WRED 表参数并开启 ECN 功能。在接口 Twenty-FiveGigE1/0/3 上应用 WRED 表 queue-table5。

```
[DeviceA] qos wred queue table queue-table5
[DeviceA-wred-table-queue-table5] queue 5 weighting-constant 12
[DeviceA-wred-table-queue-table5] queue 5 drop-level 0 low-limit 10 high-limit 20
discard-probability 30
[DeviceA-wred-table-queue-table5] queue 5 ecn
[DeviceA-wred-table-queue-table5] quit
[DeviceA] interface twenty-fivegige 1/0/3
[DeviceA-Twenty-FiveGigE1/0/3] qos wred apply queue-table5
```

(2) 配置 Device B

在接口 Twenty-FiveGigE1/0/1、Twenty-FiveGigE1/0/2 上配置接口信任报文自带的 802.1p 优先级，开启接口的 PFC 功能，并对 802.1p 优先级 5 开启 PFC 功能。

```
<DeviceB> system-view
[DeviceB] interface range twenty-fivegige 1/0/1 to twenty-fivegige 1/0/2
[DeviceB-if-range] qos trust dot1p
[DeviceB-if-range] priority-flow-control enable
[DeviceB-if-range] priority-flow-control no-drop dot1p 5
[DeviceB-if-range] quit
```

全局开启 LLDP 功能。

```
[DeviceB] lldp global enable
```

在接口 Twenty-FiveGigE1/0/2 上开启 LLDP 功能，并允许发布 DCBX TLV，配置接口 Twenty-FiveGigE1/0/2 的 DCBX 版本为预标准版 1.01。

```
[DeviceB]interface twenty-fivegige 1/0/2
[DeviceB-Twenty-FiveGigE1/0/2] lldp enable
[DeviceB-Twenty-FiveGigE1/0/2] lldp tlv-enable dot1-tlv dcbx
[DeviceB-Twenty-FiveGigE1/0/2] dcbx version rev101
[DeviceB-Twenty-FiveGigE1/0/2] quit
```

在接口 Twenty-FiveGigE1/0/2 上开启 WRR 队列，并按照每次轮询可发送的字节数进行计算，同时配置端口队列 5（802.1p 优先级 5 到本地优先级 5 为默认映射关系）采用严格优先级调度算法。

```
[DeviceB] interface twenty-fivegige 1/0/2
[DeviceB-Twenty-FiveGigE1/0/2] qos wrr byte-count
[DeviceB-Twenty-FiveGigE1/0/2] qos wrr 5 group sp
```

4. 配置验证

在 Device B 上显示丢弃报文信息。

```
<DeviceB> display packet-drop summary
All interfaces:
  Packets dropped due to Fast Filter Processor (FFP): 0
  Packets dropped due to STP non-forwarding state: 0
  Packets dropped due to insufficient data buffer. Input dropped: 0 Output dropp
ed: 0
  Packets of ECN marked: 1622267130
  Packets of WRED droped: 0
```

以上信息表明，网络中丢包数为 0，报文零丢弃。

在 Device B 上显示端口 Twenty-FiveGigE1/0/2 的带宽利用率。

```
<DeviceB> display counters rate outbound interface Twenty-FiveGigE 1/0/2
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)   Broadcast (pps)   Multicast (pps)
WGE1/0/2                 100       2825427                --                --

 Overflow: More than 14 digits.
      --: Not supported.
```

以上信息表明，端口 Twenty-FiveGigE1/0/2 的带宽利用率为 100%。