ECN、PFC 技术白皮书

文档版本 V1.0

发布日期 2022-12-16

商标说明

Inspur 浪潮、Inspur、浪潮、Inspur NOS 是浪潮集团有限公司的注册商标。

本手册中提及的其他所有商标或注册商标，由各自的所有人拥有。

技术支持

技术服务电话：400-860-0011

地　　址：中国济南市浪潮路 1036 号

浪潮电子信息产业股份有限公司

邮　　箱：lckf@inspur.com

邮　　编：250101

前言

变更记录

版本	时间	变更内容
V1.0	2022-12-16	首版发布

目 录

1 概述

1.1 背景

大型网络中，每一组传输端和接收端之间的通信都要通过几台甚至几十台设备的协作来实现，所有的设备和应用程序同时运行时数据量是十分庞大的，一个设备要处理来自四面八方的报文。当设备的一个端口接收的报文量持续多于处理的数据量时，就会出现拥塞，设备将无法承载后续传来的报文，从而影响到整个网络环境的效率。

尽管不能完全避免网络应用程序中的拥塞现象，但也有一些方法可以减少设备在发生拥塞时对网络环境的影响。

- 尾部丢弃（Tail Drop）

交换机端口队列对于拥塞的最基本处理机制就是尾部丢弃，当交换机的端口队列容量用尽时，要进入队列的报文会被丢弃，直至有足够空间接收即将进入队列的报文。这种机制在队列容量用尽时才丢弃报文，在 TCP 环境中可能会导致全局同步（global synchronization）的现象发生，进而导致整个网络环境效率降低。

- 流量管制（Flow Control）

IEEE 802.3x 提出了一种针对双工模式的流量管制机制，用于处理拥塞情形，当侦测到设备端口入方向队列容量用尽出现拥塞时，会回传 Pause Frame 通知对接的设备拥塞发生，对接设备在收到 Pause Frame 后会依据报文内的信息停止转发报文一段时间。交换机端口通常包含多个队列，队列之间有优先权高低的差异，此种流量管制机制在收到 Pause Frame 时中断转发的单位是整个端口，可能出现低优先权的报文拥塞导致数量不多的高优先权的报文也一并停止转发的现象。

1.2 定义

1.2.1 ECN

ECN（Explicit Congestion Notification）是指流量接收端感知到网络上发生拥塞后，通过标记报文来通知流量发送端，使流量发送端降低报文的发送速率，允许拥塞控制的端对端通知来避免丢包。

1.2.2 PFC

PFC（Priority based Flow Control）是 PAUSE 机制的一种增强，一种无损传输和拥塞缓解

功能，其工作原理是为全双工以太网链路上的每个 IEEE 802.1p 代码点（优先级）提供精细的链路级流量控制。当交换机接口上的接收缓冲区填充到阈值时，交换机会向发送方（连接的对等方）发送一个 PAUSE 帧，以暂时阻止发送方发送更多帧。缓冲区阈值必须足够低，以便发送方有时间停止传输帧，并且接收方可以在缓冲区溢出之前接收已经在线路上的帧。交换机自动设置队列缓冲区阈值以防止帧丢失。

当拥塞迫使链路上的一个优先级暂停时，链路上的所有其他优先级继续发送帧，只有暂停优先级的帧不被传输。当接收缓冲区清空到另一个阈值以下时，交换机会发送一条消息，再次启动流。

1.2.3 ECN、PFC 基本环境

ECN（Explicit Congestion Notification，显示拥塞通知）和 PFC（Priority-based Flow Control，优先权基准流量控制）都是跨设备拥塞调整机制，相较于前述的随机早期检测机制或其衍生机制由单一设备丢弃报文来实现功能，ECN 通过传递特定报文内容来主动通知传送设备进行调整。如下图 1-1 所示，CE A 要与 CE B 传递数据，CE 和交换机通过在发送、转发及接收报文时调整 IP 和 TCP 字段的信息来实践 ECN 功能，主动调整传送端的传输速度以解决网络拥塞的情况，ECN 实现位置为出端口队列。如下图 1-2 所示，CE A 传递数据给 CE B，路径上经过的设备端口队列都有 PFC 相关配置，若发生拥塞，会从拥塞发生的队列反向回传 Pause Frame 来通知设备停止转发报文，PFC 实现位置为入端口队列。

图 1-1 ECN 基本环境运作架构

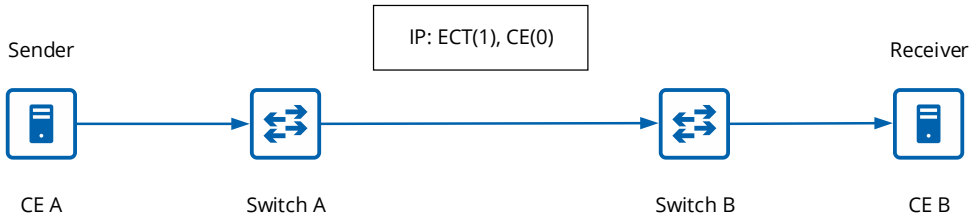
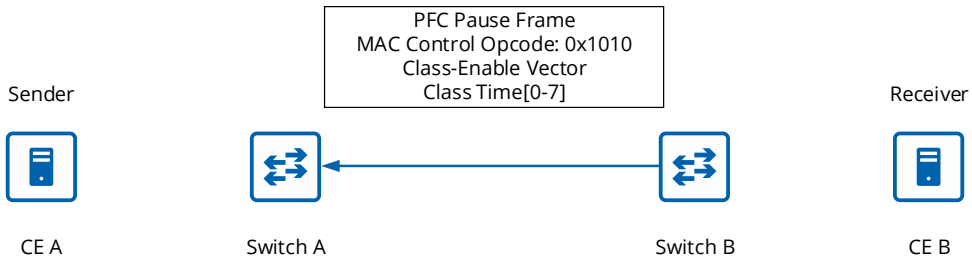


图 1-2 PFC 基本环境运作架构



1.3 优点

与旧有的或基础的拥塞处理机制相比，ECN 可以在不主动丢弃报文的情况下处理拥塞现象，不丢弃报文的机制主要有以下几个优点：

- 更好的传输效能

旧有的或基本的处理机制在侦测到拥塞状况出现时会主动的丢弃报文，丢弃报文的操作势必会造成不必要的效能浪费，而 ECN 和 PFC 可以在侦测到拥塞时通知传送端调整传输速度来解决拥塞问题，从而有效提升传输的效能。

- 降低队头阻塞（Head-of-lice blocking，HOL blocking）情形发生

许多网络传输协议都需要依照报文顺序处理服务，若单个报文的片段因为拥塞机制被丢弃，会导致接收端为了等待传送端重送被丢弃的片段而发生队头阻塞的现象。很明显采用 ECN 和 PFC 技术可以有效的降低队头阻塞现象的发生。

- 降低逾时重送（Retransmission Timeout，RTO）的发生几率

正如前面提到的那样，当封包遗漏或被丢弃时，传输协议会启动重送机制，而随着丢弃的几率的增加，传送端的重送机制发生的次数也会增加，从而影响传输效率。相较于 IEEE 802.3x 提出的流量管制（Flow Control）针对整个端口做速度的调整，PFC 可以将调整影响的范围限制在特定的端口队列，使得应用服务更有弹性。此外，对于 PFC 的侦测还提供了看门狗机制，可在拥塞连续发生时选择丢弃或转发报文。

2 缩写和术语

如图 1-1 和图 1-2 所示：

缩写和术语	解释
ECN	Explicit Congestion Notification，显示拥塞通知
PFC	Priority-based Flow Control，优先权基准流量控制
DSCP	Differentiated Services Code Point，差分服务代码点，8个标识字节进行编码，来划分服务类别，区分服务的优先级
Tail Drop	尾部丢弃，队列拥塞处理机制，队列空间用尽时丢弃收到的报文
RED	Random Early Detection，队列拥塞处理机制，依据队列使用率采取几率性地丢弃报文
WRED	Weight Random Early Detection，队列拥塞处理机制，依据报文分类结果以及队列使用率采取几率性地丢弃报文机制
Flow Control	流量控制，IEEE 802.3x针对全双工模式提出的拥塞处理机制
CNP	Congestion Notification Packets，拥塞通知报文，服务器收到ECN拥塞报文，将发送CNP拥塞通知报文
ECT	ECN-Capable Transport，IP标头中用来代表报文是否预期使用ECN机制的字段
CE	Congestion Experienced，IP标头中用来代表通讯路径发生拥塞的字段

3 技术介绍

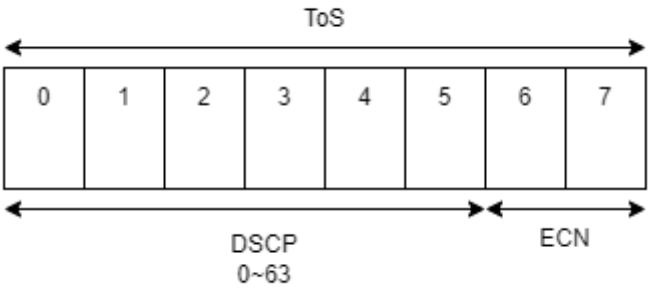
3.1 ECN 显示拥塞通知

3.1.1 背景知识与需求限制

相较于旧有的或基础的拥塞处理机制，ECN 采用由携带特定信息的报文取代直接丢弃报文的做法，为了实现这样的做法 ECN 在实践上有一些限制。ECN 需要结合其他基础的拥塞检测结果将丢弃报文的行为改为修改报文字段，常见的实现方式是搭配 RED 或 WRED 机制来运行 ECN。为了让设备间能同步整个路径上的拥塞情形，ECN 需要 IP 标头字段在设备间传递信息。因为 ECN 不在发生拥塞的端口直接丢弃报文而是直接让传送端进行调整，ECN 传送端设备在收到拥塞回报后要有对应的流量调整机制。

3.1.2 IP 中的 ECN 操作

图 3-1 DSCP/ECN 字段



ECN 使用 IPv4 首部或 IPv6 首部中 ToS（Type of Service）字段的两个最低有效位（最右侧的位编码）来表示四个状态代码：

00 – 不支持 ECN 的传输，非 ECT(Non ECN-Capable Transport)

10 – 支持 ECN 的传输，ECT(0)

01 – 支持 ECN 的传输，ECT(1)

11 – 发生拥塞，CE(Congestion Encountered)

交换机通过将 ECN 字段置为 11，就可以通知流量接收端交换机是否发生了拥塞。当流量接收端收到 ECN 字段为 11 的报文时，就知道网络上出现了拥塞，且该 IP 报文不会被 WRED 机制丢弃。如果接收服务器发现 IP 报文的 ECN 字段被标记成 11，就立刻产生 CNP

（Congestion Notification Packets）拥塞通知报文，并将该报文发送源服务器。CNP 消息里包含了拥塞的数据流信息，远端服务器接收到后，通过降低相应的数据流发送速率，缓解网络设备拥塞，从而避免发生丢包。

当网络中拥塞解除时，流量接收端不会收到 ECN 字段为 11 的报文，也就不会往流量发送端发送用于告知其网络中存在拥塞的协议通告报文。此时，流量发送端收不到协议通告报文，则认为网络中没有拥塞，从而会恢复报文的发送速率。

3.1.3 ECN 运作流程范例

图 3-2 发送端发送 IP 报文标记 ECN(ECN=10)

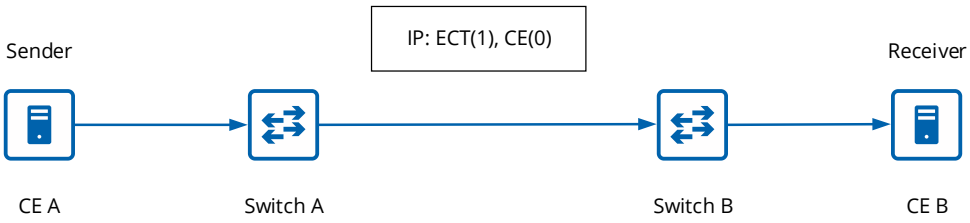


图 3-3 交换机在队列拥塞的情况下收到该报文，将 ECN 字段修改为 11 并转发出去

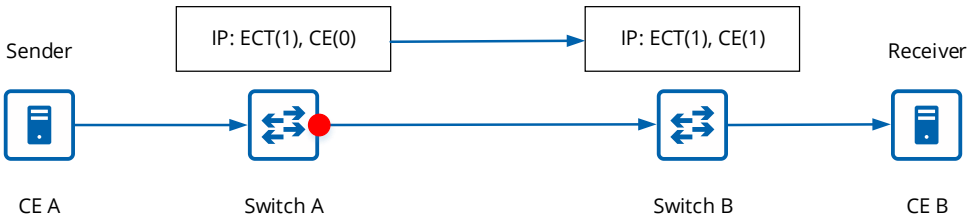


图 3-4 接收服务器收到 ECN 为 11 的报文发送拥塞，正常处理该报文

图 3-5 接收端产生拥塞通告，周期发送 CNP（Congestion Notification Packets）报文，ECN 字段为 01，要求报文不能被网络丢弃

图 3-6 交换机收到 CNP 报文后正常转发该报文

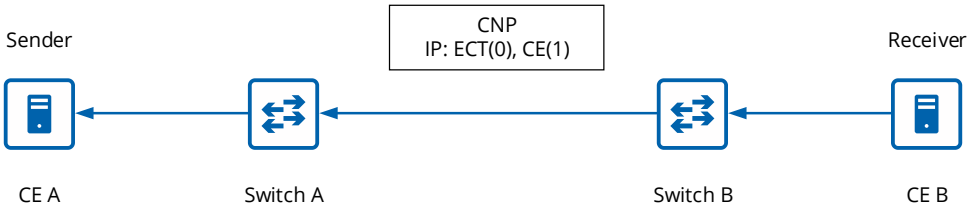


图 3-7 发送服务器收到 ECN 标记为 01 的 CNP 报文解析后对相应的数据流限速算法

3.1.4 ECN 支持配置内容

我们将使用有更多弹性的 WRED 做为 ECN 拥塞侦测的参考依据，可针对三种不同的报文（绿色、黄色和红色）分别独立配置 WRED 的数据。每个颜色拥有独立的最小阈值（min-threshold）、最大阈值（max-threshold）以及几率值（probability）三个数值供用户配置，此部份配置概念与 WRED 一致，当配置开启 ECN 功能时会在判断拥塞发生而处理的报文的 IP 标头中将 CE 字段设定为 1，当配置关闭 ECN 功能时则不执行 ECN 处理。详细配置指南可参考典型应用指南章节。

3.2 PFC 优先权基准流量控制

3.2.1 背景知识与需求限制

标准协议 IEEE 802.3x 的规范中针对全双工模式（Full Duplex）提出流量控制（Flow Control）的机制来处理拥塞情形，而流量控制主要由 Pause Frame 报文来完成两相连设备间的沟通，PFC 藉由在流量控制机制的 Pause Frame 中夹带进阶信息来实现。除了侦测到拥塞的设备要能够组成 PFC 所需要的进阶信息外，传送端在收到此 Pause Frame 时也需要有识别能力做对应的处理。

3.2.2 PFC Pause Frame 报文介绍

支持 PFC 的 Pause Frame 报文和 Flow Control 的 Pause Frame 报文一样带有 MAC Control Type 0x8808（如下图 3-8），不同的地方在于 PFC 的 Pause Frame （如下图 3-

9) 接续在 MAC Control Type 字段之后的内容依序为，2 Byte 长度的 MAC Control Opcode 0x0101，2 Byte 长度的 Priority-Enable Vector（首 8 个 bit 固定为 0 ，接着的 8 个 bit 分别代表 8 个优先权队列 PFC 功能的开启与关闭状态，0 为不采用 PFC 机制，反之 1 代表此队列采用 PFC 机制），接续 16 Byte 长度分别代表 8 个队列的 pause_time 数值，每个队列有 2 Byte 长度的 pause_time 可以配置。在我们的设计中除了设置 pause_time 字段告知相连设备的指定队列停止转发报文外，这类 Priority-Enable Vector 设定为 1 且对应 pause_time 不为 0 的 PFC Pause Frame 报文我们定义为 xoff 报文，主要目的为告知相连接端口队列有拥塞发生需要停止转发，若 PFC Pause Frame 报文的 Priority-Enable Vector 设定为 1 且对应的 pause_time 为 0 时我们定义为 xon 报文，其目的为主动告知相连设备端口队列可恢复报文的转发，此设计可以更快的对环境做出反应，不必每次都确实等待计时终了。

图 3-8 IEEE 802.3x PAUSE FRAME 报文格式

Preamble(7)	SFD(1)	DMAC(6) 01-80-C2-00-00-01	SMAC(6)	MAC Control Type(2) 0x8808	MAC Opcode(2) 0x0001	pause_time(2)	Padding(42)

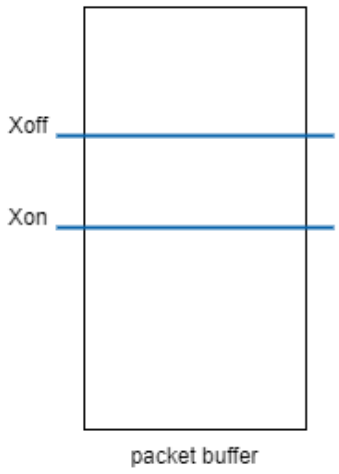
图 3-9 PFC 报文格式

Preamble(7)	SFD(1)	DMAC(6) 01-80-C2-00-00-01	SMAC(6)	MAC Control Type(2) 0x8808	MAC Opcode(2) 0x0101	Priority-Enable Vector(2)	pause_time[0-7](16)	Padding(20)

3.2.3 PFC 拥塞判断及 PFC 相关报文生成

报文的入端口的 Buffer 用于报文缓存，在使能 PFC 时，需要设置触发 Pause 帧的水线，也就是超过 Xoff Threshold 水线会触发停止对端发包的 Pause 帧，低于 Xon Threshold 水线会触发恢复对端发送的 Pause 帧。

图 3-10 PFC 水线示意图



XON：使能发送的 Pause 帧，代表对方可以发送数据了

XOFF: 使能发送的 Pause 帧，代表对方停止发送数据

PFC 水线是基于入端口 Buffer 进行触发的，入端口方向提供的 8 个队列可以将不同优先级的业务报文映射到不同队列上，从而实现不同优先级的报文分配不同的 Buffer。

3.2.4 PFC 报文处理

当设备端口收到 Pause Frame 时，检查 MAC Control Opcode 地址是否为 PFC 使用的 0x0101，若为 0x0101 则确认 Priority-Enable Vector 中设为 1 的 bit 地址，根据设为 1 的 bit 对应的队列编号截取后面对应的 pause_time 数值，若对应的 pause_time 数值不为 0 则代表对接设备端口的对应队列出现拥塞现象，则设备会暂停从此队列转发报文，若同一队列持续一段时间（Detection Time）都收到 xoff 报文，会进一步触发 PFC 看门狗机制运行对应的操作，PFC 看门狗机制在下一节详细解释。在队列停止转发报文之后如果经过 pause_time 的时间都未再收到 xoff 报文则此队列恢复正常报文转发行为。若 PFC Pause Frame 对应的 pause_time 数值为 0，则代表对接设备的拥塞情形已排除，此时不管先前 xoff 报文的 pause_time 时间终了与否都会恢复队列的报文转发。

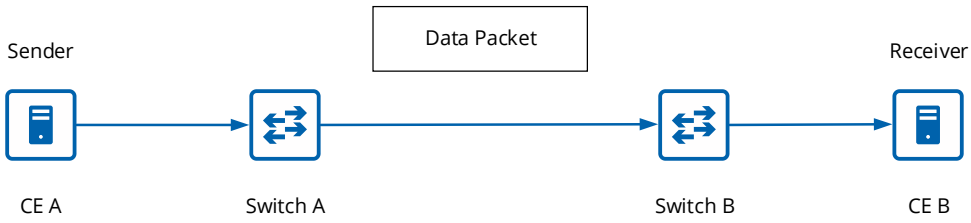
3.2.5 PFC 看门狗（PFCWD）机制

若同一队列持续一段时间（Detection Time）都收到 xoff 报文，就可能产生 PFC 暂停风暴，PFC 看门狗监控启用 PFC 的端口是否存在 PFC 暂停风暴。PFCWD 针对每个端口提供丢弃和转发报文两种行为选项，当触发 PFCWD 时队列会执行设备所配置的选项行为。PFCWD 有两个定时器可配置，分别是侦测时间（Detection Time）和恢复时间（Restoration Time），两个时间都针对 xoff 报文做侦测，当队列暂停转发报文的时间超过 Detection Time 时会触发 PFCWD 执行对应的操作。若超过 Restoration Time 都没收到 xoff 报文，就会停止 PFCWD 的操作选项。

3.2.6 PFC 运作流程范例

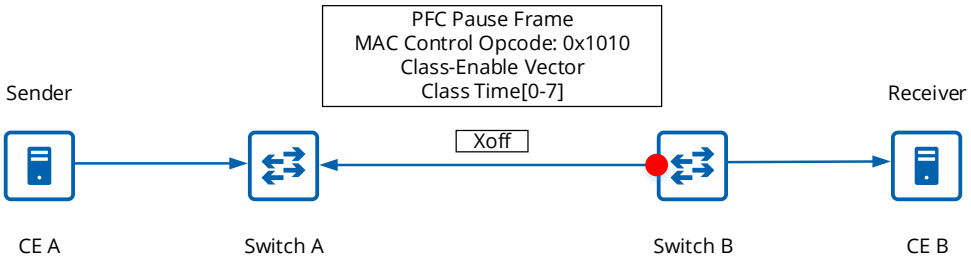
以下图 3-11 为例。CE A 传递数据给 CE B，假设路径中 Switch A 转发报文到 Switch B 端口入方向的队列 0 且队列并未出现拥塞的情形，此时 Switch B 不会回传 PFC Pause Frame。

图 3-11 CE A 传递信息给 CE B 正常运行状况



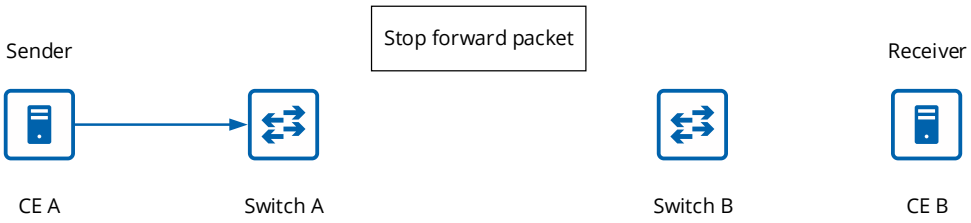
如下图 3-12，若此时在 Switch B 端口入方向的队列 0 出现了拥塞，这时 Switch B 会回传 PFC Pause Frame xoff 给 Switch A。

图 3-12 Switch B 侦测到拥塞情形回传 xoff 报文



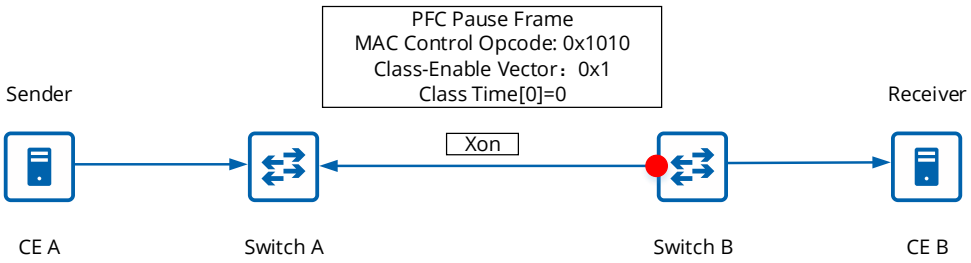
如下图 3-13，Switch A 在收到 xoff 后停止转发队列 0 报文到 Switch B 端口。

图 3-13 Switch A 收到 xoff 报文后停止端口队列 0 的报文转发



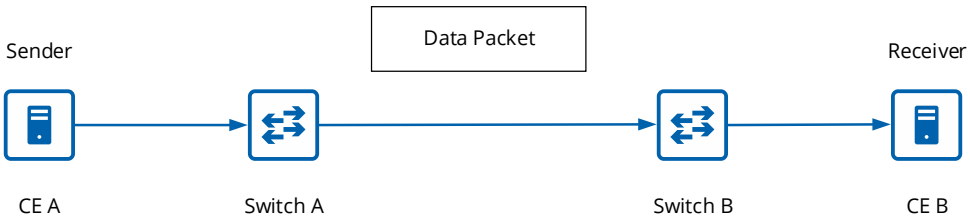
如下图 3-14，此时若 Switch B 端口入方向队列 0 已排除拥塞情行时，发送 xon 报文给 Switch A。

图 3-14 Switch B 端口队列 0 确认拥塞情况已排除，发送 xon 报文给 Switch A



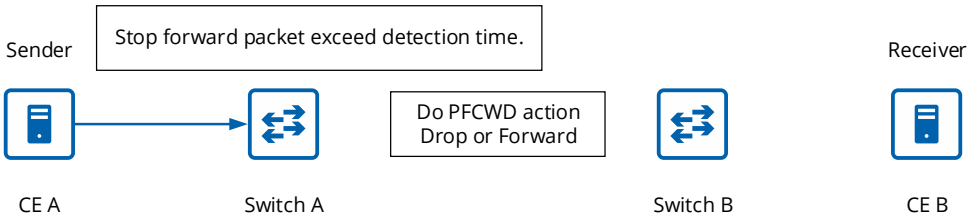
如下图 3-15，Switch A 收到 xon 报文时恢复队列的报文转发，重新开始转发队列 0 报文到 Switch B 端口。

图 3-15 Switch A 收到 xon 报文后端口队列恢复正常转发行为



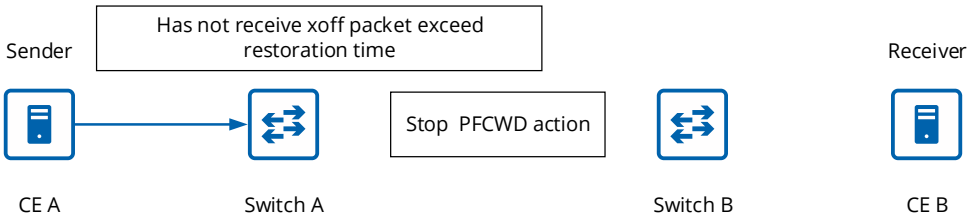
PFCWD 的状况如下图 3-16，若 Switch A 持续超过 Detection Time 未恢复转发状态则执行 PFCWD 配置行为。

图 3-16 Switch A 端口队列持续未恢复转发状态，触发 PFCWD 行为



如下图 3-17，当 Switch A 执行 PFCWD 配置行为时，若超过 Restoration Time 未收到 xoff 报文则停止 PFCWD 行为，此时队列转发与否由 PFC 判断机制判断。

图 3-17 Switch A 持续未收到 xoff 报文后停止 PFCWD 行为



3.2.7 PFC 支持配置内容

如同前面章节描述，PFC 可配置公共缓存空间的大小、xoff 阈值、端口队列的大小、static_th 值、xoff 阈值、xon 阈值和 xon_offset 值，上述项目可决定队列在哪些情况需要发送 xoff 报文或是 xon 报文，除此之外还有 PFCWD 的行为、Detection Time 和 Restoration Time 针对 PFCWD 机制的配置。详细配置指南可参考典型应用指南章节。

4 主要特性

ECN 相关特性：

- ECN 为出端口队列的跨设备拥塞处理机制。
- 目前 ECN 的拥塞判断是基于 WRED 机制。
- 单一设备可建立 128 个 ECN-WRED 使用的 Queue Profile 配置。
- 每个 Queue Profile 配置有 WRED、ECN 或是不处理拥塞三种选择。
- 每个 Queue Profile 配置可针对三种报文上色结果（绿、黄和红）做配置。
- Queue Profile 配置中每个上色结果有最小阈值（min-threshold）、最大阈值（max-threshold）和几率（probability）可配置。
- Queue Profile 配置的绑定和端口队列的绑定为完全独立，同一端口的不同队列可绑定不同的 Queue Profile 配置。

PFC 相关特性：

- PFC 为入端口队列的跨设备拥塞处理机制。
- 各队列分别占用公用缓存空间，PFC 可配置公用缓存空间的尺寸以及触发 Pause Frame 的 xoff 数值。
- 单一设备可建立多个 Buffer Profile 并且配置队列相关的参数。
- Buffer Profile 可分为 Ingress 和 Egress 两种，PFC 应用方式以 Ingress 实现，详细应用方法可参考典型应用指南章节。
- Buffer Profile 配置中包含了缓存大小（buffer size）、公用缓存使用大小（static shared buffer size）、xoff-size、xon-size 和 xon-offset。
- Buffer Profile 配置的绑定和端口队列的绑定为完全独立，同一端口的不同队列可绑定不同的 Buffer Profile 配置。
- PFCWD 针对每个端口支持丢弃报文（Drop）和转发报文（Forward）两种行为配置，默认为丢弃报文。
- PFCWD 针对每个端口支持 Detection Time 和 Restoration Time 两个时间区间配置。

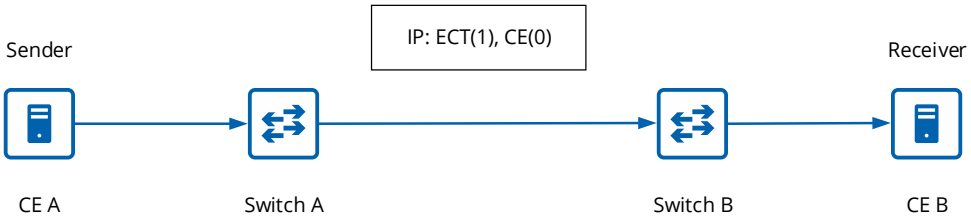
5 典型应用指南

5.1 ECN 典型组网方案

下图 5-1 是数据中心常用的典型拓扑，CE A 和 CE B 之间由交换机（Switch A 和 Switch B ）做连接，此范例以 CE A 传递数据到 CE B 为例。

- CE A、CE B、Switch A 和 Switch B 都需支持 ECN 功能
- Switch A 和 Switch B 分别建立各自的 Queue Profile。
- Switch A 和 Switch B 配置 Queue Profile 的 ECN 相关参数。
- Switch A 和 Switch B 分别将建立的 Queue Profile 与端口出方向队列绑定。

图 5-1 典型 ECN 配置拓扑范例



5.2 ECN 主要配置命令

1. qos queue-profile <profile-name>

no qos queue-profile <profile-name>

命令模式：

config 配置模式。

参数说明：

<profile-name> 长度限制 32 个字符且支持 a-z、A-Z、0-9、.、- 和 _ 字符。

命令说明：

建立 ECN-WRED 使用的 Queue Profile。

删除 ECN-WRED 使用的 Queue Profile。

2. random-detect

no random-detect

命令模式：

QoS queue profile domain 域配置模式。

参数说明：

无。

命令说明：

开启 WRED 功能，若同时开启 WRED 及 ECN 时会执行 ECN 操作行为。

关闭 WRED 功能。

3. ecn

no ecn

命令模式：

QoS queue profile domain 域配置模式。

参数说明：

无。

命令说明：

开启 ECN 功能，若同时开启 WRED 及 ECN 时会执行 ECN 操作行为。

关闭 ECN 功能。

4. green min-threshold <threshold>

yellow min-threshold <threshold>

red min-threshold <threshold>

命令模式：

QoS queue profile domain 域配置模式。

参数说明：

<threshold> 队列使用空间阈值，范围 0 ~ 32058368 单位为 byte。

命令说明：

设定 WRED 判断拥塞发生最小阈值，队列使用空间超过此数值时判定拥塞情形出现，从 0 开始线性递增 WRED 或 ECN 执行几率。

5. green max-threshold <threshold>

yellow max-threshold <threshold>

red max-threshold <threshold>

命令模式：

QoS queue profile domain 域配置模式。

参数说明：

<threshold> 队列使用空间阈值，范围 0 ~ 32058368 单位为 byte。

命令说明：

设定 WRED 判断拥塞发生最大阈值，队列使用空间到达此数值时报文依 probability 配置几率执行 WRED 或 ECN 操作，队列使用空间超过此数值时报文必定执行 WRED 或 ECN 操作。

6. green probability <probability>

yellow probability <probability>

red probability <probability>

命令模式：

QoS queue profile domain 域配置模式。

参数说明：

<probability> 队列使用率到达 max-threshold 数值时执行 WRED 或 ECN 操作的几率，范围为 0 到 100。

命令说明：

队列使用率到达 max-threshold 数值时执行 WRED 或 ECN 操作的几率。

7. service-policy queue-profile <profile-name> queue <queue-list>

no service-policy queue-profile <profile-name> queue <queue-list>

命令模式：

interface ethernet 配置模式。

参数说明：

<profile-name> Queue Profile 名称。

<queue-list> 队列列表。

命令说明：

绑定队列与 Queue Profile 关系。

解绑队列与 Queue Profile 关系。

5.3 ECN 具体配置

1. 建立并配置参数。

1.1. 需确认 QoS 功能已启动

```
Switch1# configure terminal
Switch1(config)# qos reload default
```

1.2. 建立 queue profile。

```
Switch1(config)# qos queue-profile profileA
```

1.3. 配置参数。

```
Switch1(config-qos-queue-profile-profileA)# random-detect
Switch1(config-qos-queue-profile-profileA)# ecn
Switch1(config-qos-queue-profile-profileA)# green min-threshold 1000
Switch1(config-qos-queue-profile-profileA)# green max-threshold 5000
Switch1(config-qos-queue-profile-profileA)# green probability 100
```

2. 绑定 Queue Profile 和队列。

```
Switch1(config)# interface ethernet 1
Switch1(config-if-ethernet1)# service-policy queue-profile profile queue 0-4
```

3. 解绑 Queue Profile 和队列。

```
Switch1(config-if-ethernet1)# no service-policy queue-profile profile queue 0-4
```

4. 删除 Queue Profile。

```
Switch1(config)# no qos queue-profile profile
```

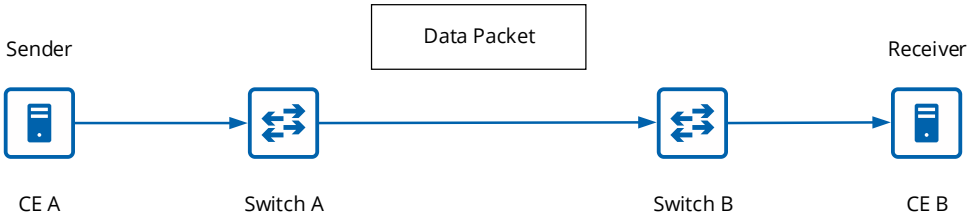
5.4 PFC 典型组网方案

下图 5-2 是数据中心常用的典型拓扑，CE A 和 CE B 之间由交换机（Switch A 和 Switch B ）做连接，此范例以 CE A 传递数据到 CE B 为例。

- Switch A 和 Switch B 支持 PFC 机制 。
- Switch B 配置公用缓存空间的 buffer-size 和 xoff-size。

- Switch B 开启端口队列 PFC 功能。
- Switch B 建立 Buffer Profile。
- Switch B 配置 Buffer Profile 内 PFC 相关参数。
- Switch B 将建立的 Buffer Profile 端口入方向队列绑定。
- Switch A 配置相连接端口的 PFCWD action、detection-time 和 restore-time。

图 5-2 典型 PFC 配置拓扑范例



5.5PFC 主要配置命令

1. qos buffer-pool ingress buffer-size <buffer-size> pause-threshold <xoff-size>

命令模式：

config 配置模式。

参数说明：

<buffer-size> 公用缓存可用大小，范围 0 ~ 32058368 单位为 byte。

<xoff-size> 公用缓存触发 Pause Frame 的阈值，范围 0 ~ 32058368 单位为 byte。

命令说明：

端口入方向队列公共缓存空间参数配置。

2. qos buffer-profile <profile-name>

no qos buffer-profile <profile-name>

命令模式：

config 配置模式。

参数说明：

<profile-name> 长度限制 32 个字符且支持 a-z、A-Z、0-9、.、- 和 _ 字符。

命令说明：

建立 PFC 使用的 Buffer Profile。

删除 PFC 使用的 Buffer Profile。

3. ingress static-th <static-shared-buffer-size> buffer-size <buffer-size> pause-threshold <xoff-size> resume-threshold <xon-size> resume-offset-threshold <xon-offset>

命令模式：

QoS buffer profile domain 域配置模式。

参数说明：

<static-shared-buffer-size> 队列优先使用公共缓存空间大小，范围 0 ~ 32058368 单位为 byte。

<buffer-size> 公共缓存保留队列使用大小，范围 0 ~ 32058368 单位为 byte。

<xoff-size> xoff 参数，范围 0 ~ 32058368 单位为 byte。

<xon-size> xon 参数，范围 0 ~ 32058368 单位为 byte。

<xon-offset> xon-off 参数，范围 0 ~ 32058368 单位为 byte。

命令说明：

队列可使用空间小于 xoff-size 时发送 PFC xoff 报文。

队列使用率低于 xon-size 和 static-shared-buffer-size-xon-offset 两者较大值时发送 PFC xon 报文。

4. service-policy buffer-profile ingress <profile-name> queue <queue-list>

no service-policy buffer-profile ingress <profile-name> queue <queue-list>

命令模式：

interface ethernet 配置模式。

参数说明：

<profile-name> Buffer Profile 名称。

<queue-list> 队列列表。

命令说明：

绑定队列与 Queue Profile 关系。

解绑队列与 Queue Profile 关系。

5. priority-flow-control priority <priority>

no priority-flow-control priority <priority>

命令模式：

interface ethernet 配置模式。

参数说明：

<priority> 开启/关闭 PFC 功能的队列列表。

命令说明：

开启队列 PFC 功能。

关闭队列 PFC 功能。

6. priority-flow-control watch-dog action {drop|forward} interval <detection-time> restore <restore-time>

no priority-flow-control watch-dog

命令模式：

interface ethernet 配置模式。

参数说明：

{drop} PFCWD 丢弃报文。

{forward} PFCWD 转发报文。

<detection-time> PFC detection time，范围 100 ~ 5000 单位为毫秒。

<restore-time> PFC restoration time，范围 100 ~ 5000 单位为毫秒。

命令说明：

配置 PFCWD 行为、Detection Time 和 Restoration Time。PFC 具体配置。

5.6PFC 具体配置

1. 建立并配置参数。

1.1. 需确认 QoS 功能已启动

```
Switch1# configure terminal
Switch1(config)# qos reload default
```

1.2. 配置公共缓存空间参数。

```
Switch1(config)# qos buffer-pool ingress buffer-size 28500000 pause-threshold 4500000
```

1.3. 建立 buffer profile。

```
Switch1(config)# qos buffer-profile profileA
```

1.4. 配置参数。

```
Switch1(config-qos-buffer-profile-profileA)# ingress static-th 19456 buffer-size 2048 pause-threshold 196608 resume-threshold 15872 resume-offset-threshold 3584
```

2. 绑定 Buffer Profile 和队列。

```
Switch1(config)# interface ethernet 1

Switch1(config-if-ethernet1)# service-policy buffer-profile ingress profileA queue 0-4
```

3. 队列开启 PFC 功能。

```
Switch1(config-if-ethernet1)# priority-flow-control priority 3-4
```

4. 端口配置 PFCWD。

```
Switch1(config-if-ethernet1)# priority-flow-control watch-dog action drop interval 400 restore 400
```

5. 队列关闭 PFC 功能。

```
Switch1(config-if-ethernet1)# no priority-flow-control
```

6. 解绑 Buffer Profile 和队列。

```
Switch1(config-if-ethernet1)# no service-policy buffer-profile ingress profileA queue 0-4
```

7. 删除 Buffer Profile。

```
Switch1(config)# no qos buffer-profile profileA
```

6 维护

下面以“典型应用指南”一章所举例子介绍如何监控 ECN、PFC 模块运行状态及进行相关的故障排查。

1. 查看所有 Queue Profile 内容

```
Switch1# show qos queue-profile
Profile: AZURE_LOSSLESS
Admin State: enable
Action: ecn
Color        Min-Threshold(Byte)     Max-Threshold(Byte)     Probability(%)
-------------  ------------------------------  ------------------------------  --------------------
Green                        104000                       312000                     100
Yellow                       104000                       312000                     100
Red                          104000                       312000                     100
Profile: profileA
Admin State: enable
Action: ecn
Color        Min-Threshold(Byte)     Max-Threshold(Byte)     Probability(%)
-------------  ------------------------------  ------------------------------  --------------------
Green                             0                     32058368                     100
Yellow                            0                     32058368                     100
Red                               0                     32058368                     100
```

2. 查看特定 Queue Profile 内容

```
Switch1# show qos queue-profile profileA
Profile: profileA
Admin State: enable
Action: ecn
Color        Min-Threshold(Byte)     Max-Threshold(Byte)     Probability(%)
-------------  ------------------------------  ------------------------------  --------------------
Green                             0                     32058368                     100
Yellow                            0                     32058368                     100
```

```
Red                                0                   32058368                  100
```

3. 查看所有 Queue Profile 与队列绑定关系

```
Switch1# show interface service-policy queue-profile
Profile: AZURE_LOSSLESS
Port       Queue
---------  ----------


Profile: profileA
Port            Queue
--------------  ---------
Ethernet2   0,1,2,3,4
```

4. 查看特定 Queue Profile 与队列绑定关系

```
Switch1# show interface service-policy queue-profile profileA
Profile: profileA
Port            Queue
--------------  ---------
Ethernet2   0,1,2,3,4
```

5. 查看公用缓存空间配置

```
Switch1# show qos buffer-pool
     pool_name           size         xoff
---------------------------  -----------  -------
  egress_pool         32058368     N/A
ingress_pool         28500000    4500000
```

6. 查看所有 Buffer Profile 内容

```
Switch1# show qos buffer-profile
                profile name       reserved size            mode      shared buffer/alpha value       xoff       xon      xon_offset
-------------------------------------  ----------------------------  ----------  ----------------------------------------  ----------  ---------  ------------
 egress_lossless_profile                     0   -         static_th                         32058368         N/A        N/A          N/A
      egress_lossy_profile                     0        dynamic_th                                1         N/A        N/A          N/A
ingress_lossless_profile                  2048          static_th                            19456      196608      15872         3584
    ingress_lossy_profile                     0        dynamic_th                                3         N/A        N/A          N/A
                 profileA                  2048          static_th                            19456      196608      15872         3584
```

7. 查看端口队列 PFC 配置状态

```
Switch1# show interface priority-flow-control priority config
   Interface       Lossless Priority
-------------------   ------------------------
   Ethernet1                    3,4
   Ethernet2                    3,4
   Ethernet3                    3,4
   Ethernet4                    3,4
   Ethernet5                    3,4
   Ethernet6                    3,4
   Ethernet7                    3,4
   Ethernet8                    3,4
   Ethernet9                    3,4
  Ethernet10                    3,4
…… 以下因篇幅省略 ……
```

8. 查看端口队列 PFC 报文转生成数据

```
Switch1# show interface priority-flow-control priority statistic
     Port Rx      PFC0      PFC1      PFC2      PFC3      PFC4      PFC5      PFC6      PFC7
-------------------   ----------   ----------   -----------   ----------   -----------   ----------   -----------   ------
  Ethernet1          0         0         0         0         0         0         0         0
  Ethernet2          0         0         0         0         0         0         0         0
  Ethernet3          0         0         0         0         0         0         0         0
  Ethernet4          0         0         0         0         0         0         0         0
  Ethernet5          0         0         0         0         0         0         0         0
  Ethernet6          0         0         0         0         0         0         0         0
  Ethernet7          0         0         0         0         0         0         0         0
…… 以下因篇幅省略 ……
```

9. 查看端口队列与 Buffer Profile 绑定关系

```
Switch1# show interface service-policy buffer-profile ingress
Interface          Queue     Profile
-------------------   ----------   ------------------------
Ethernet1                0     ingress_lossy_profile
```

```
Ethernet1               1    ingress_lossy_profile
Ethernet1               2    ingress_lossy_profile
Ethernet1               3    ingress_lossless_profile
Ethernet1               4    ingress_lossless_profile
Ethernet1               5    ingress_lossy_profile
Ethernet1               6    ingress_lossy_profile
Ethernet1               7    ingress_lossy_profile
Ethernet2               0    ingress_lossy_profile
Ethernet2               1    ingress_lossy_profile
Ethernet2               2    ingress_lossy_profile
Ethernet2               3    ingress_lossless_profile
Ethernet2               4    ingress_lossless_profile
…… 以下因篇幅省略 ……
```

10. 查看端口 PFCWD 配置状态

```
Switch1# show interface priority-flow-control watch-dog config
        PORT       ACTION       DETECTION TIME       RESTORATION TIME
------------------   ----------------   ----------------------------   ------------------------------
Ethernet1          drop                    400                         400
```

11. 查看端口 PFCWD

```
Switch1# show interface priority-flow-control watch-dog statistic
  QUEUE     STATUS     STORM DETECTED/RESTORED     TX OK/DROP     RX OK/DROP     TX LAST OK/DROP     RX LAST OK/DROP
----------------   ----------   ----------------------------------------   ----------------   ----------------   ----------------------------   ----------------------------
      3       stormed          3/4                          0/50           0/100          0/50                  0/100
```