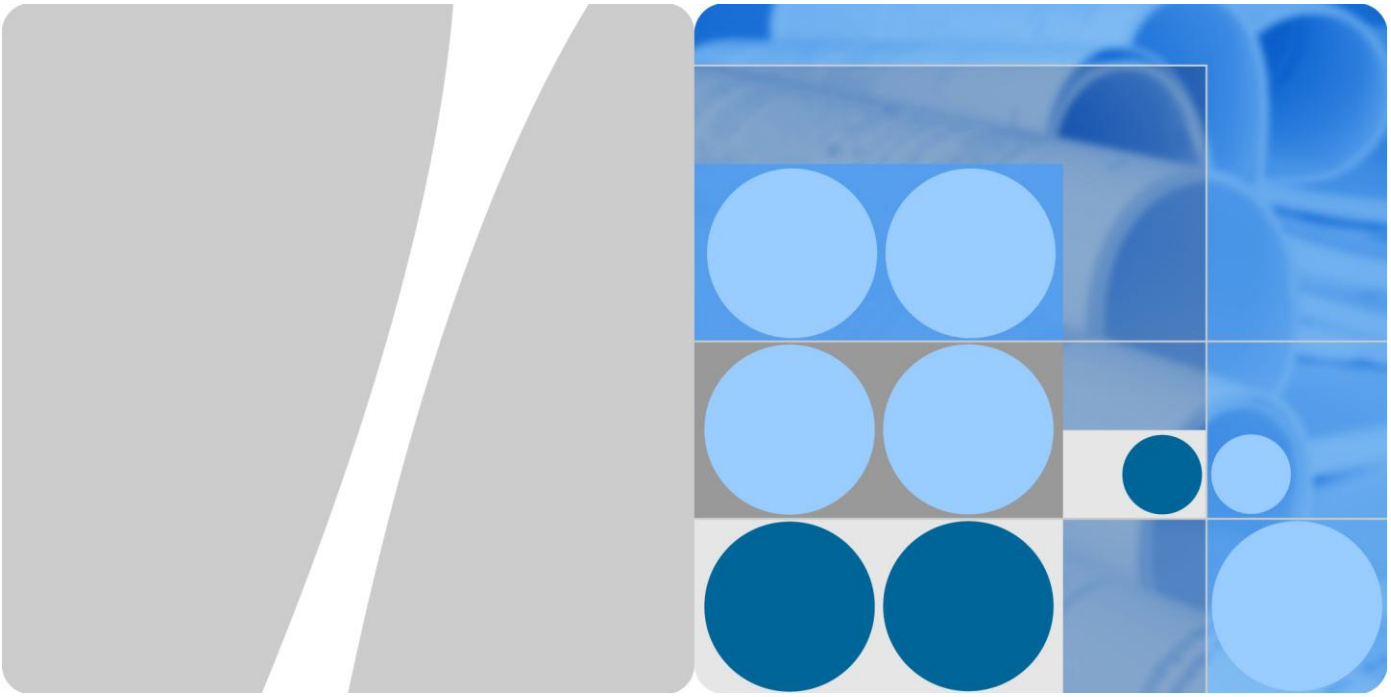


技术白皮书-FCoE/DCB

文档版本 01

发布日期 **2013-04-09**

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://enterprise.huawei.com

客户服务邮箱： ChinaEnterprise_TAC@huawei.com

客户服务电话： 4008229999

目 录

插图目录

表格目录

1 FCoE

1.1 介绍

定义

以太网光纤通道 FCoE（Fibre Channel over Ethernet）是由美国国家标准委员会 ANSI（American National Standards Institute）定义的一种融合网络技术，是以光纤通道 FC（Fibre Channel）存储协议为核心的 I/O 整合方案。

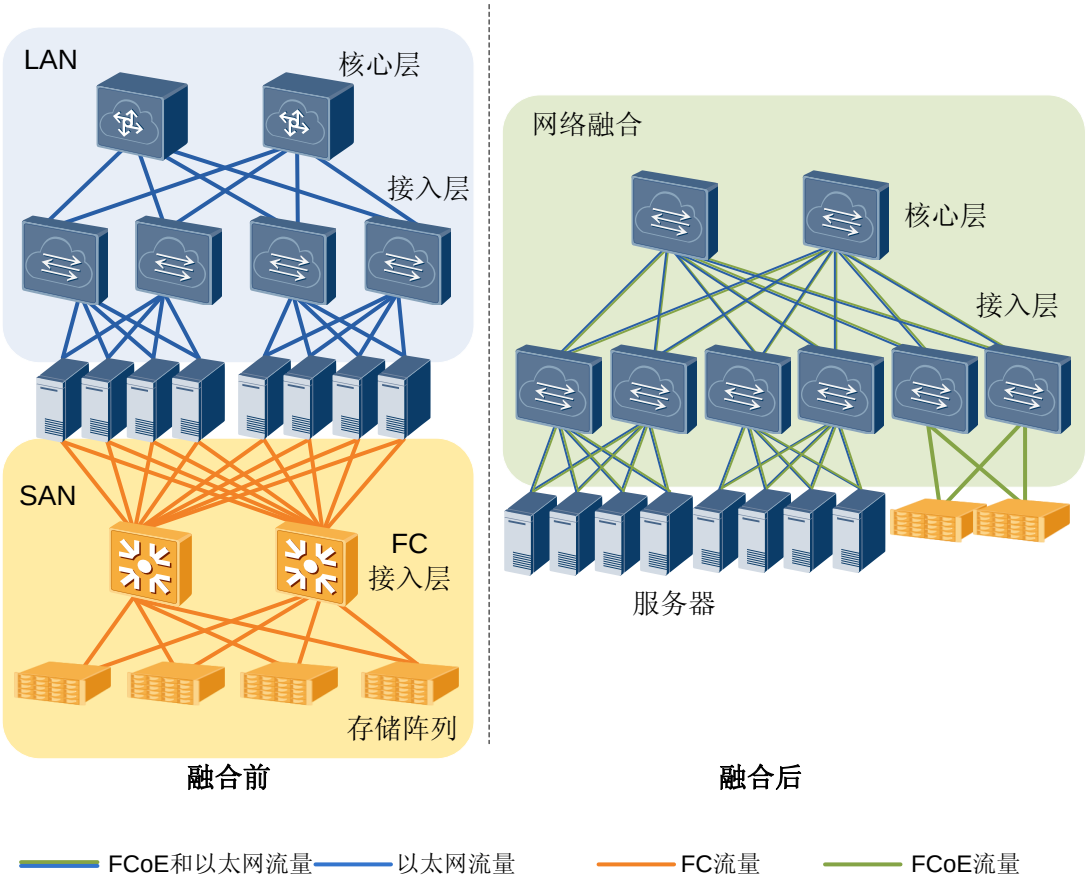
目的

如图 1-1 所示，传统数据中心组网中，以太网 LAN（Local Area Network）用于服务器与服务器和客户端与服务器之间的通信，存储区域网络 SAN（Storage Area Network）用于服务器与存储设备之间的通信。LAN 和 SAN 的部署和维护相互独立。

随着数据中心的飞速发展和服务器数量的激增，LAN 和 SAN 的独立部署存在如下问题：

- 网络复杂：LAN 和 SAN 的相互独立导致业务部署的灵活性差，网络扩展困难，网络的维护和管理成本高。
- 能效比低：服务器上至少配置 4～6 块网卡，包括用于接入 LAN 的网络接口卡 NIC（Network Interface Card）和用于接入 SAN 的主机总线适配器 HBA（Host Bus Adapter）。服务器的多类型的网卡使得整个数据中心的电力消耗和冷却成本增加。

图1-1 数据中心网络融合前后对比



数据中心网络融合后，存储网络 SAN 和以太网 LAN 可共享同一个单一的、集成的网络基础设施，解决了不同类型网络共存所带来的问题，实现了网络基础设施整合、精简的目标。

对比 SAN 和 LAN 网络模型可知，当网络融合后会出现以下问题：

- 网络融合后 FC 流量无法正常转发。
- 现有以太网无法达到 FC 网络中无丢包转发的需求。

为解决上述问题，以太网光纤通道 FCoE 和数据中心桥接 DCB（Data Center Bridging）应运而生。其中：

- FCoE 实现了以太网帧承载 FC 帧，并控制 FCoE 转发，从而使得 LAN 和 SAN 能够共享网络资源，实现融合网络。
- DCB 实现了在数据中心网络中构建一个无丢包以太网，使得传统以太网实现 FC SAN 网络中的拥塞控制，为 FCoE 融合业务提供了传输质量保证。

受益

FCoE 可以为数据中心带来如下受益：

- 更低的总体拥有成本 TCO（Total Cost of Ownership）：LAN 和 SAN 网络通过 FCoE 技术共享网络资源，整合并更有效的利用以前分散的资源，减少对于 SAN 网络基础设施的投资，简化了网络复杂度，降低网络的管理和维护成本；服务器采用融合网络适配器 CNA（Converged Network Adapter），减少数据中心的电力和冷却成本。
- 强大的投资保护：FCoE 可以和数据中心现有的以太网及 FC 基础设施无缝互通，保护客户在现有 FC SAN 上的投资（如 FC 设备、工具成本和管理成本）。
- 增强的业务灵活性：FCoE 使得所有的服务器共享存储资源，满足虚拟机迁移的需求，这样也提高了系统的灵活性和可用性。

1.2 参考标准和协议

本特性的参考资料清单如下：

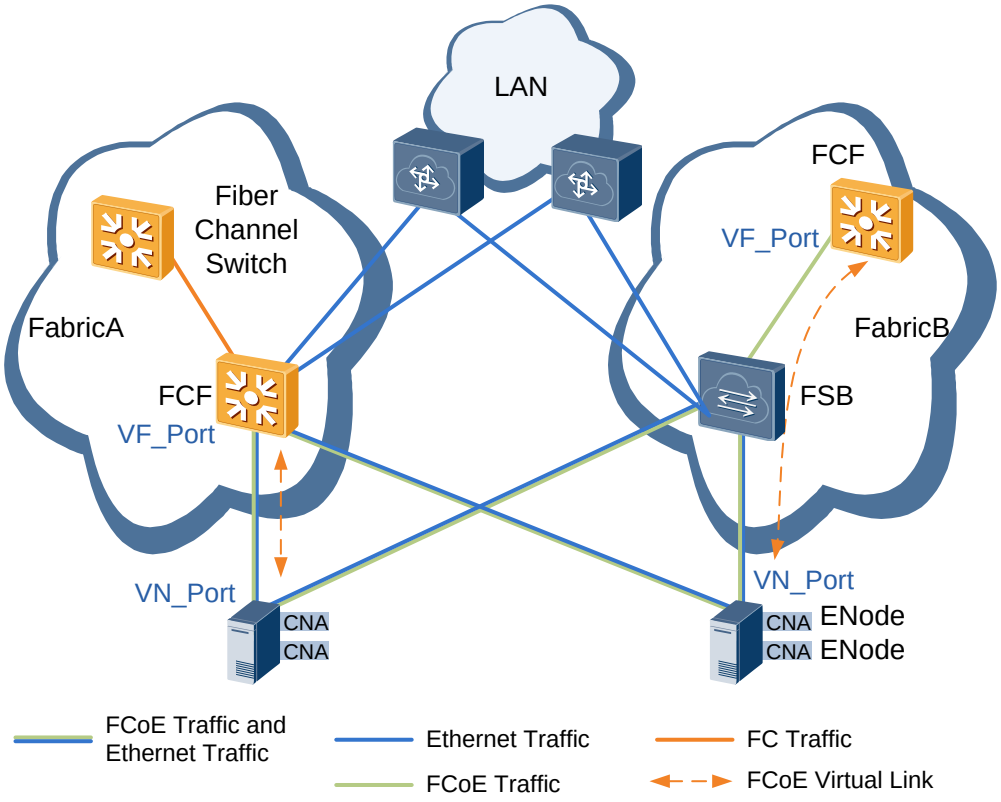
文档	描述	备注
FC-BB-5	Fibre Channel Backbone - 5 Rev 2.00	-

1.3 原理描述

1.3.1 FCoE 基本概念

如图 1-2 所示的 FCoE 组网中，FCoE 中存在如下基本概念：ENode、FCF、FSB、Fabric、FCoE 虚链路（FCoE Virtual Link）、FIP 协议、端口角色和 FCoE VLAN。

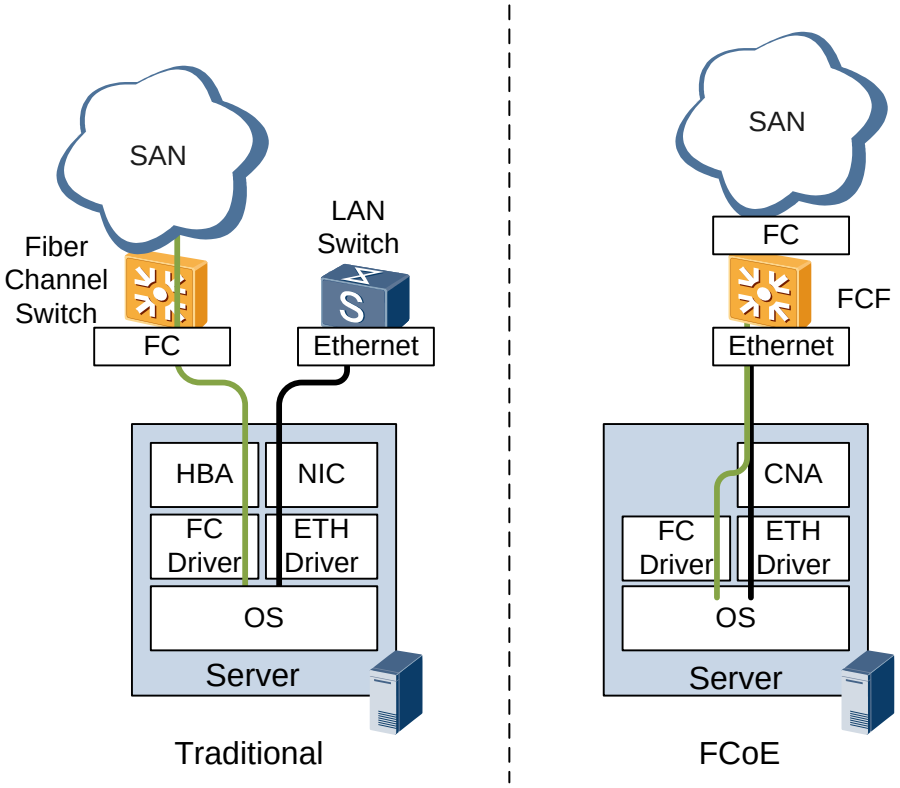
图1-2 FCoE 典型组网图



- ENode

  ENode 是指服务器上同时支持 FCoE 协议栈和 FC 协议栈的 CNA。如图 1-3 所示，传统的服务器中，存在两块网卡：用于接入 LAN 的 NIC 和用于接入 SAN 的 HBA。而 ENode 中的 CNA 替代了 NIC 和 HBA 的功能，在完成以太数据转发的同时也完成 FCoE 报文的上层处理和 FCoE 封装/解封装。

图1-3 传统服务器和 FCoE 服务器的区别



- FCF

  FCoE 数据转发器 FCF（FCoE Forwarder）是指同时支持 FCoE 协议栈和 FC 协议栈的双协议栈交换机，主要用于连接传统 SAN 网络和 LAN 网络。FCF 能够转发 FCoE 报文，同时具有 FCoE 封装/解封装功能。

- FSB

  FSB（FCoE Initialization Protocol Snooping Bridge）是指运行 FIP Snooping 功能的交换机，交换机本身不支持 FC 协议栈。交换机利用 FIP Snooping 功能侦听 FIP 协议，控制 FCoE 虚链路的建立，预防恶意攻击。

- Fabric

  Fabric 指网络节点通过一台或多台交换机互联的网络拓扑结构。

- FCoE 虚链路（FCoE Virtual Link）

  FCoE 虚链路是指连接 FCoE 链路终端设备（如 ENode 和 FCF）之间点到点的逻辑链路。因为 FCF 和 ENode 之间可能通过无丢包以太网连接起来，此时 FCF 和 ENode 之间的点到点关系将被破坏，由此引入 FCoE 虚链路。

- FIP 协议

  FCoE 初始化协议 FIP(FCoE Initialization Protocol)主要用于 FCoE 网络中 FC 终端发现、Fabric 登录和 FCoE 虚链路建立的二层协议。通过 FIP 协议，ENode 可登录到 Fabric，实现和目标 FC 设备通信。FIP 协议同时也维护 FCoE 虚链路。

- 端口角色

  在传统 FC 网络中，FC 设备之间通过 FC 端口进行交互。其中 FC 端口分为 N_Port 和 F_Port。

- N_Port（Node Port）：是指 FC 主机（如服务器或存储设备）上连接 FC 交换机的端口。
- F_Port（Fabric Port）：是指 FC 交换机侧连接 FC 主机的接口，主要是为 FC 主机提供 Fabric 接入服务。

FCoE 保留 FC 协议中的端口角色的概念，即在 ENode 和 FCF 之间的 FCoE 虚链路上，ENode 侧的接口为 VN_Port（Virtual Node Port），FCF 侧的接口为 VF_Port（Virtual Fabric Port）。

- FCoE VLAN

  FC-BB-5 协议规定，FCoE 的报文需要在特定的 VLAN 中进行转发。在 FC 协议栈中，FC 设备支持多个 VSAN（类似以太网中的 VLAN），运行在不同的 VSAN 中的 FC 流量经过 FCoE 封装时需要用不同的 FCoE VLAN 来区分。

  虚链路仅存在在同一个 FCoE VLAN 中。FCoE VLAN 仅承载 FCoE 流量，而不承载任何以太网流量（如 IP 流量）。

1.3.2 FCoE 封装

FCoE 通过将 FC 帧封在普通以太帧中，从而实现 FC 流量在以太网中传输的功能。从 FC 协议的角度看，FCoE 仅仅是将 FC 流量承载在另一种链路上传输；从以太网协议的角度看，FCoE 仅仅是以太帧承载不同的上层协议。

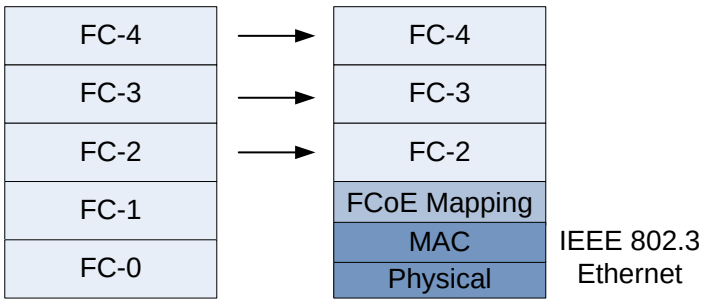
FCoE 的协议栈

如图 1-4 所示，在传统的 FC 协议中，FC 协议栈共分为 5 层：

- FC-0 定义承载介质类型
- FC-1 定义帧编解码方式
- FC-2 定义分帧协议和流控机制
- FC-3 定义通用服务
- FC-4 定义上层协议到 FC 的映射

在 FCoE 协议栈中，FC-0 和 FC-1 被映射成为 IEEE 802.3 Ethernet 协议的 Physical 和 MAC，并添加了 FCoE Mapping 作为上层 FC 协议栈与底层 Ethernet 协议栈之间的适配层。

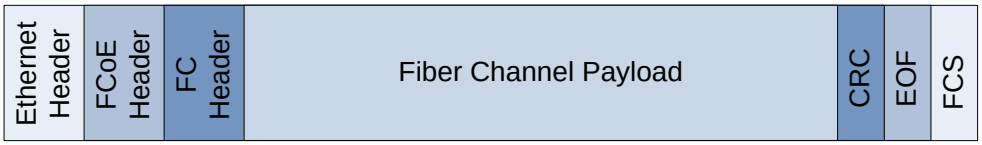
图1-4 从 FC 到 FCoE 的映射关系



报文封装

FCoE 协议将一个完整的 FC 帧封装在以太帧中，封装形式如图 1-5 所示。

图1-5 FCoE 的报文封装



其中：

- Ethernet Header 中指定了报文的源、目的 MAC 地址、以太帧类型和 FCoE VLAN。
- FCoE Header 指定了 FCoE 帧版本号和控制信息。
- FC Header 和传统 FC 帧相同，即指定了 FC 帧的源、目的地址等信息。

1.3.3 FIP 协议

基本原理

FCoE 初始化协议 FIP（FCoE Initiation Protocol）是 FCoE 的控制协议，用于成对 FCoE 设备（如 ENode 和 FCF）之间 FCoE 虚链路的建立和维护。

虚链路建立阶段：

- FIP 发现 FCoE VLAN 和对端的 FCoE 虚接口。
- FIP 执行 FCoE 虚链路的初始化功能，如：FLOGI（Fabric Login）和 FDISC（Fabric Discovery）。

当虚链路建立完成后，FIP 协议还要履行 FCoE 虚链路的维护功能：

- 周期性检测 FCoE 虚链路两端的 FCoE 虚接口是否可达。
- FLOGO（Fabric Logout）拆除 FCoE 的虚链路。

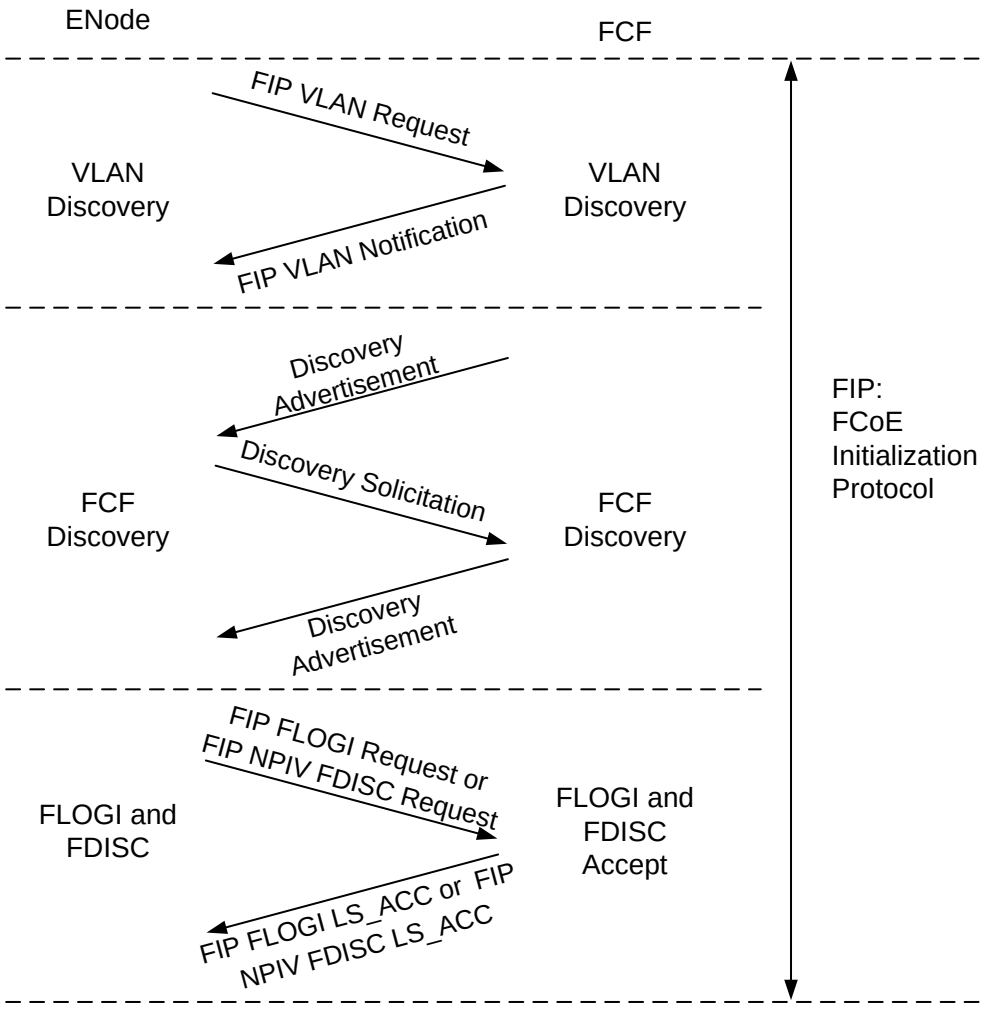
虚链路的建立

ENode 和 FCF 之间的 FCoE 虚链路的建立过程如图 1-6 所示，其中用于建立虚链路的是 FIP 帧，当虚链路建立后传输的是 FCoE 帧。

说明

FIP 协议中，所有的协议报文均由 ENode 端发起，而 FCF 也会主动发出未被请求的 FIP 通告报文，将在 **FIP FCF Discovery** 说明。

图1-6 FCoE 虚链路的建立过程



虚链路的建立主要分为三个阶段，分别是 **FIP VLAN Discovery**、**FIP FCF Discovery** 和 **FIP FLOGI and FDISC**。其中 FIP FLOGI 和 FDISC 的处理流程与传统 FC 协议中的 FLOGI 和 FDISC 类似。

1. FIP VLAN Discovery

   FIP VLAN Discovery 阶段主要功能是寻找用于承载 FCoE 帧的 FCoE VLAN，使 ENode 发现所有潜在的 FCoE VLAN，但是此时 ENode 并不能对 FCF 进行选择。

   FIP VLAN Discovery 的具体流程如下：

   - ENode 首先发送 FIP VLAN Discovery 的请求报文 FIP VLAN Request。其中报文的目的 MAC 地址是一个组播地址，这个地址又被称为 All-FCF-MAC（01-10-18-01-00-02），所有的 FCF 都会侦听这个组播地址的报文。
   - 所有的 FCF 都能通过普通 VLAN 向 ENode 反馈一个或者多个 FCoE VLAN，这些 FCoE VLAN 可被用于 ENode 的登录。

   FC-BB-5 协议规定，FIP VLAN Discovery 是可选的。FCoE VLAN 既可通过管理员手工配置，也可以通过 FIP VLAN Discovery 动态发现。

2. FIP FCF Discovery

FIP FCF Discovery 主要用于 ENode 发现可以登录的 FCF。

FIP FCF Discovery 的具体流程如下：

- FCF 周期性的在 FCoE VLAN 中发送 Discovery Advertisement 报文。报文的目的 MAC 地址为组播地址 All-ENode-MAC（01-10-18-01-00-01），以便所有 ENode 都能侦听到。报文携带的信息包括 FCF 的 MAC 地址、虚链路参数（包括 FIP 超时时间、FCF 优先级等）。
- ENode 从接收到的 Discovery Advertisement 报文中获取到可供登录的 FCF 信息，从中选择优先级最高 FCF 后向其发送单播的 Discovery Solicitation 报文。
- FCF 接收到 Discovery Solicitation 报文后，回应单播 Discovery Advertisement 报文，允许 ENode 登录。

除了接收周期性的 Discovery Advertisement 报文，新入网的 ENode 通常不需要等待所有的 FCF 发送 Discovery Advertisement 报文，协议规定 ENode 可向所有的 FCF 发送 Discovery Solicitation 报文，报文的目的地址是 All-FCF-MAC，FCF 收到请求报文之后响应一个单播 Discovery Advertisement 报文给 ENode。ENode 从接收到的 Discovery Advertisement 报文中选择优先级较高的 FCF 与其建立虚链路。

3. FIP FLOGI 和 FDISC

   当 ENode 选择一个 FCF 登录后，就会通过 FIP FLOGI 或 FIP FDISC 报文与该 FCF 的 VF_Port 建立虚链路，以便通过该虚链路传送 FCoE 帧。 FIP FLOGI 报文和 FIP FDISC 报文都是单播报文，分别替代 FC 协议中的 FLOGI 和 FDISC 报文，均用来为 ENode 分配 MAC 地址，以便其登录到 Fabric。

   FIP FIP FLOGI 和 FIP FDISC 的处理过程类似，两者唯一区别是：FIP FLOGI 指 ENode 首次登录 Fabric 时建立虚链路的过程；而 FIP FDISC 指，当 ENode 上存在多个 VM 时，为每个 VM 建立虚链路的过程。下面以 FIP FLOGI 为例来进行介绍。

   FIP FLOGI 的具体流程如下：

   - ENode 向 FCF 发送请求报文 FIP FLOGI Request。
   - FCF 响应 ENode 请求，并为 ENode 分配本地唯一 MAC 地址 FPMA（Fabric Provided MAC Address）或者 FCF 响应使用 ENode 自身指定本地唯一 MAC 地址 SPMA（Server Provided MAC Address）。

虚链路的维护

在传统 FC 网络中，如果物理链路发生故障，则 FC 协议可以马上感知到。但是在 FCoE 协议中，由于采用了以太封装，FC 协议无法感知链路层的故障，为此 FIP 提供了一种简单的 Keepalive 机制。

虚链路的监控机制如下：

- ENode 周期的向 FCF 发送 ENode FIP Keepalive 报文。如果 FCF 在 2.5 倍周期时间内没有收到已登录的 ENode 发送的 ENode FIP Keepalive 通告报文，则 FCF 认为该虚链路发生故障并断开该虚链路。
- FCF 周期的以 ALL-ENode-MAC 为目的地址向所有的 ENode 发送组播 Discovery Advertisement 报文。如果 ENode 在 2.5 倍周期时间内没有收到 FCF 发送的组播 Discovery Advertisement 报文，则 ENode 认为该虚链路发生故障并断开该虚链路。

如果 FCF 没有收到 ENode 发送的 ENode FIP Keepalive 报文，则向相应的 ENode 发送 FIP Clear Virtual Link 报文用于链路拆除。如果当 ENode 退出登录时，ENode 也可通过向 FCF 发送 Fabric Logout 请求来删除虚链路。

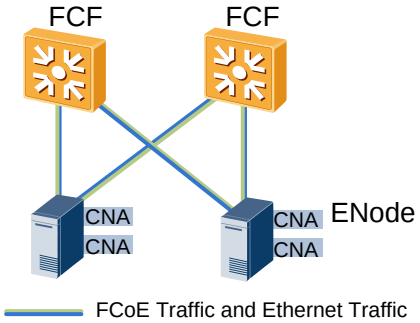
1.3.4 FIP Snooping

ENode 和 FCF 之间的连接模式分为两种：直连模式和远端连接模式。FIP Snooping 主要用来解决远端连接模式中引入的安全性问题。

直连模式

如图 1-7 所示，当 ENode 直连 FCF 时，虚链路和虚链路映射的物理链路均为点到点。在这种场景下，虽然在物理链路上转发的报文都会经过 FCoE 封装，但物理链路的两端都支持 FC 协议栈，FCoE 帧的转发和传统 FC 帧转发流程基本一致。

图1-7 直连模式

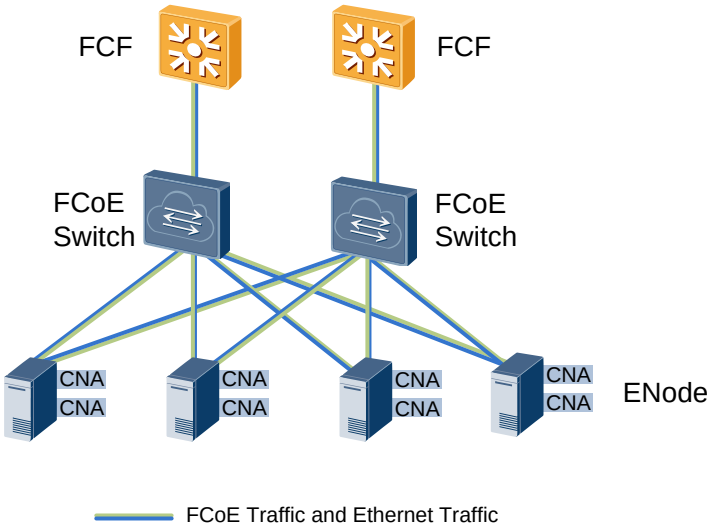


在直连模式下，FCoE 帧的处理过程除数据链路层的数据封装外均遵循 FC 协议。由此可知， FCoE 的安全性不会比传统 FC 协议增加任何风险。当采用 FCoE 技术后，SAN 管理员依旧可以采用原先的软件管理网络。

远端连接模式

虽然直连模式是最佳的组网模式，但是 FCF 价格相对昂贵，在数据中心中服务器的数量庞大的前提下，所有的服务器和 FCF 直连是不现实的。如图 1-8 所示，一种典型组网模型是在 FCF 和 ENode 之间增加接入交换机，即远端连接模式。此时，接入交换机作为 FCoE 交换机，不具备完全的 FCF 功能，如交换机作为 FSB（FIP Snooping Bridge）。

图1-8 远端连接模式



说明

在远端连接模式下，ENode 和 FCF 之间可能经过一台或多台 FCoE 交换机。

FIP Snooping

在 FC 网络中，FC 交换机是被当成可信任的设备，其他 FC 设备（如 ENode）必须先登录到 FC 交换机后才能接入 FC 网络，并由 FCF 交换机分配指定地址。同时，FC 链路是点到点的，FC 交换机能够完全控制 FC 设备接收/发送的流量，因此 FC 交换机能确保 FC 设备使用指定地址进行报文交互并保护设备不受恶意攻击。

而在 FCoE 下，特别是远端连接模式下，ENode 和 FCF 之间增加了 FCoE 交换机，因为 FCoE 交换机不支持 FC 协议，FCoE 帧在 FCoE 交换机上的转发基于以太网协议，转发的目的地址不一定是 FCF，ENode 和 FCF 之点到点的关系被破坏了。

为了达到和 FC 一样的强壮性，就需要在 FCoE 交换机上强制所有来自 ENode 的 FCoE 流量发往 FCF。FIP Snooping 功能就是让交换机通过侦听的 FIP 协议报文获取相应的虚链路信息，控制 FCoE 虚链路的建立，预防恶意攻击。

当 FCoE 交换机运行 FIP Snooping 功能时被称为 FSB（FIP Snooping Bridge）。CE6800 设备支持配置 FIP Snooping 功能。

步骤 1 配置 FC 实例

```
[~CE6800] fcoe FSB
[~CE6800-fcoe-FSB] vlan 2094
[~CE6800-fcoe-FSB] commit
[~CE6800-fcoe-FSB] quit
```

步骤 2 配置端口角色

```
[~CE6800-10GE1/0/1] fcoe role vnp
[~CE6800-10GE1/0/1] commit
[~CE6800-10GE1/0/1] quit
```

2 DCB

2.1 介绍

定义

数据中心桥 DCB（Data Center Bridging）协议是一组由 IEEE 802.1 工作组定义的以太网扩展协议。DCB 协议组主要用于构建无丢包以太网，以满足数据中心网络融合后的 QoS 需求。

目的

数据中心包括如下三种业务：SAN、LAN 和 IPC（Inter-Process Communication）。传统数据中心对每种业务部署一个网络，但随着现有数据中心规模的逐步增大，会带来如下问题：

- 每个服务器需要多个专用适配器（网卡），同时也需要不同的布线系统；
- 机房需要支持更多设备：空间、耗电、制冷；
- 多套网络无法统一管理，需要不同的维护人员；
- 部署/配置/管理/运维更加困难。

多网融合是解决上述问题的方向，但上述各种流量的 QoS 需求上存在较大差异。SAN 流量对丢包很敏感、且要求报文在传输过程中是保序的；LAN 流量允许丢包，只需要设备提供尽力而为的服务(BE)，丢包和乱序都可以由两端的主机来处理，不需要网络节点做过多的干预；IPC 用于服务器之间的通信，流量要求低时延。为了能在以太网满足上述各种流量（尤其是 SAN 流量）的 QoS 需求。由此 DCB 协议产生。

2.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
IEEE 802.1 Qbb	Priority-based Flow control	-
IEEE 802.1 Qaz	• Enhanced transmission selection • Data Center Bridging Exchange (DCBX) Protocol	

2.3 原理描述

在配置 DCB 特性时，配置 PFC 功能和配置 ETS 功能属于必选配置，无顺序关系。配置 DCBX 功能属于可选配置，当用户配置 PFC 的工作模式为 **auto** 模式时，请先 DCBX 功能。

DCB 协议的主要特性如表 2-1 所示。

表2-1 DCB 特性列表

特性	目的
PFC（Priority-based Flow control）	在共享链路上，提供针对优先级的流量控制能力。
ETS（Enhanced transmission selection）	在共享链路上，提高带宽利用率。
DCBX（Data Center Bridging Exchange）Protocol	自动协商链路两端的以太网参数，减少管理成本。

2.3.1 PFC

产生原因

网络融合后，SAN 流量在以太网中传输时要求不丢包。

现有以太 Pause 机制即可实现不丢包。以太 Pause 机制的原理如下：当下游设备发现接收能力小于上游设备的发送能力时，会主动发一个 Pause 帧给上游设备，要求暂停流量的发送，等待一定时间后再继续发送数据。但是以太 Pause 机制是将链路上所有的流量都暂停，即流量暂停是针对整个接口。而对 FCoE 而言链路共享至关重要。链路共享要求：

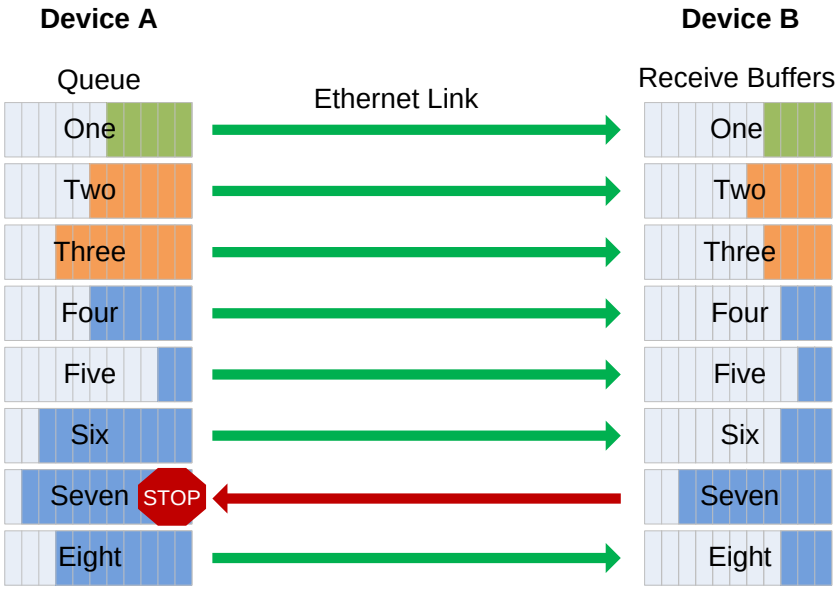
- 一种类型的突发流量不能影响其他类型流量的转发。
- 一种类型的流量大量积压在队列中不能抢占其他类型的流量的缓存资源。

为了解决现有以太 Pause 机制和链路共享之间的冲突，基于优先级流量控制 PFC 产生了。

基本原理

PFC 也称为 Per Priority Pause 或 CBFC（Class Based Flow Control），是对现有以太 Pause 机制的增强。PFC 是一种基于优先级的流控机制，如图 2-1 所示，DeviceA 发送接口分成了 8 个优先级队列，DeviceB 接收接口分成了 8 个接收缓存，两者一一对应。当 DeviceB 的端口上某个接收缓存即将产生拥塞时，发送一个反压信号“STOP”到 DeviceA，DeviceA 停止发送对应优先级队列的报文。

图2-1 PFC 的工作机制



“反压信号”实际上是一个以太帧，其具体报文格式如图 2-2 所示。

图2-2 PFC 帧格式

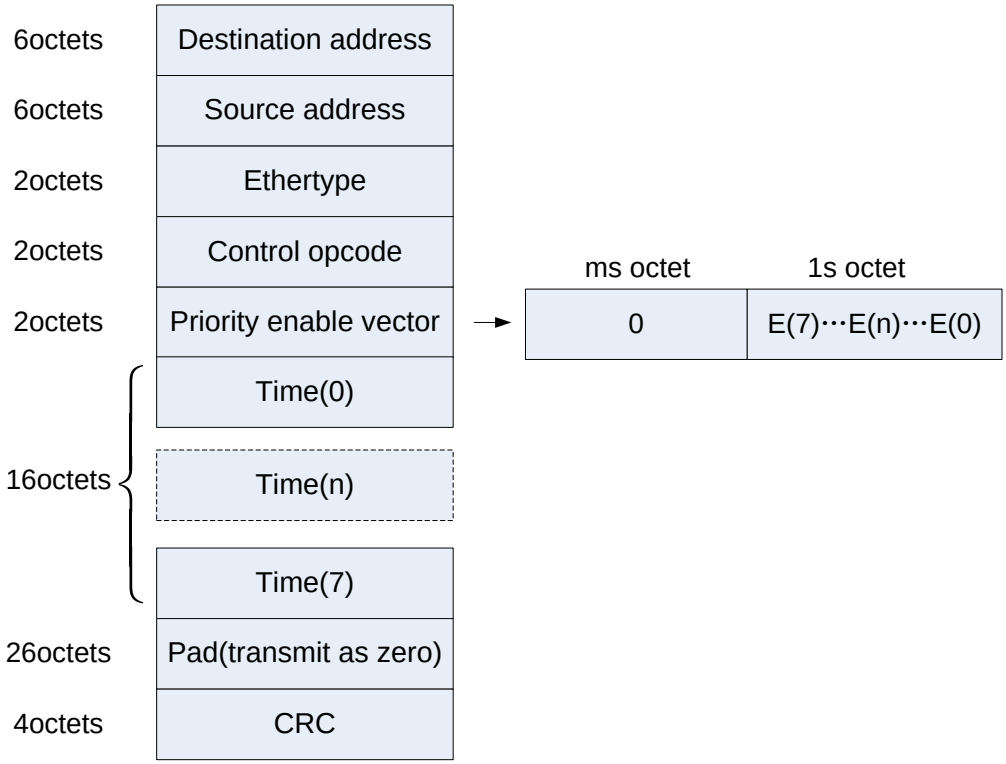


表2-2 PFC 帧定义

项目	描述
Destination address	目的 MAC 地址，取值固定为 01-80-c2-00-00-01。
Source address	源 MAC 地址。
Ethertype	以太网帧类型，取值为 88-08h。
Control opcode	控制码，取值为 01-01。
Priority enable vector	反压使能向量。 其中 E(n)和优先级队列 n 对应，表示优先级队列 n 是否需要反压。当 E(n)=1 时，表示优先级队列 n 需要反压，反压时间为 Time(n)；当 E(n)=0 时，则表示该优先级队列不需要反压。
Time(0)～Time(7)	反压定时器。当 Time(n)=0 时表示取消反压。
Pad(transmit as zero)	预留，传输时为 0。
CRC	循环冗余校验。

由此可见，流量暂停只针对某一个或几个优先级队列，不针对整个端口进行中断。每个队列都能单独进行暂停或重启，而不影响其他队列上的流量，真正实现多种流量共享链路。而对非 PFC 控制的优先级队列，系统则不进行反压处理，即在发生拥塞时将直接丢弃报文。

在 FCoE 环境下，管理员可指定 FCoE 流量对应的队列使能 PFC 保证不丢包。

PFC 配置

CE6800 设备支持配置 PFC 功能：

```
[~CE6800] interface 10ge 1/0/1
[~CE6800-10GE1/0/1] dcb pfc enable mode auto
[~CE6800-10GE1/0/1] quit
[~CE6800] commit
```

2.3.2 ETS

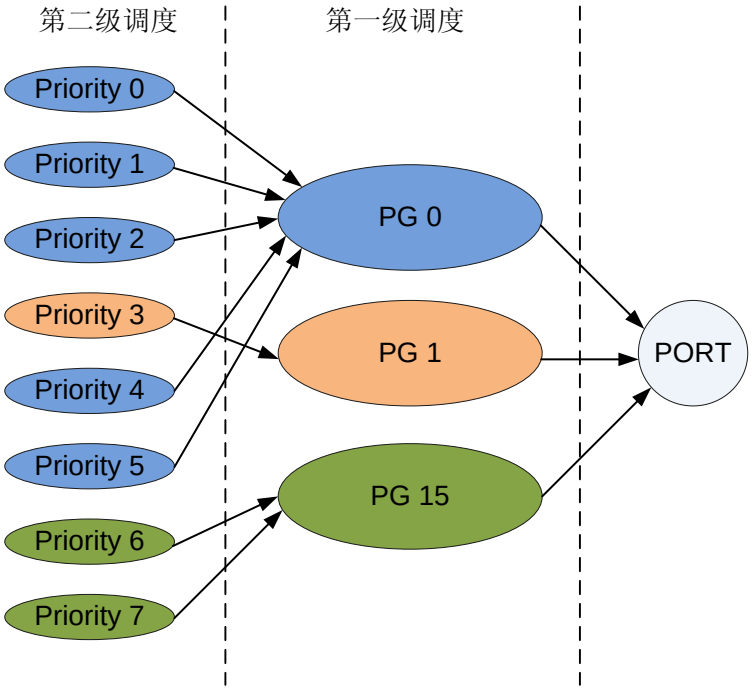
产生原因

数据中心网络融合后，在网络中存在三种流量：LAN 流量、SAN 流量和 IPC 流量。而融合网络中对 QoS 的要求很高。传统的 QoS 已经无法满足融合网络的需求，而增强传输选择 ETS 通过灵活的层次化的调度实现网络融合后的 QoS。

基本原理

ETS 提供两级调度，分别基于优先级组 PG（Priority Group）和优先级 Priority，如图 2-3 所示。接口首先对优先级组进行第一级调度，然后对优先级内的优先级队列进行第二级调度。

图2-3 ETS 的处理流程



相比普通 QoS，ETS 的优势在于提供了基于优先级组的调度，将同一类型的流量归入统一优先级组，使得同一类流量能够获得相同的服务等级。

优先级组的调度

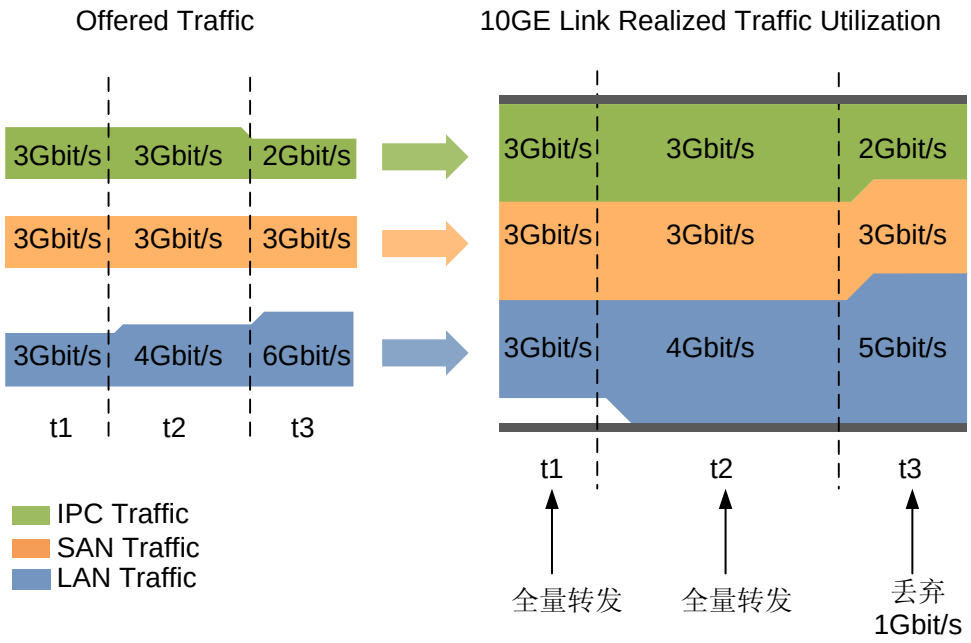
优先级组即一组拥有相同调度属性的端口优先级的队列，用户可通过设置将不同的优先级队列加入到优先级组中。基于优先级组的调度被称为第一级调度。

在 ETS 中定义了 3 个优先级组 PG0、PG1 和 PG15，分别代表是 LAN 流量、SAN 流量和 IPC 流量。

协议规定，PG0、PG1 和 PG15 的调度方式为是 PQ+DRR，其中由于 PG15 承载 IPC 流量，对延时要求很高，因此调度方式为是 PQ（Priority Queue）；PG0 和 PG1 的调度方式为赤字轮循队列调度 DRR（Deficit Round Robin）。另外，用户也可根据实际情况对优先级组划分带宽。

如图 2-4 所示，假设在出接口队列中，优先级为 3 的队列承载的是 FCoE 流量，则将优先级队列 3 划入 SAN 组（即 PG1）；优先级 0、1、2、4、5 的队列承载普通 LAN 流量，则划入 LAN 组（即 PG0）；优先级 7 的队列承载 IPC 流量，则划入 IPC 组（即 PG15）。接口总带宽是 10Gbit/s，PG1 和 PG0 各分配 50%的带宽限制，即 5Gbit/s。

图2-4 基于优先级组的拥塞管理



在 t1 和 t2 时刻，接口总流量不超过接口带宽时，所有流量都能转发；在 t3 总流量超过接口带宽，且 LAN 流量超过给定的带宽，按照 ETS 的参数进行调度，LAN 业务流量被丢弃 1Gbit/s。

另外，ETS 还提供基于优先级组的流量整形。优先级组的流量整形基于优先级组限制流量的突发，使该优先级组内的流量以比较均匀的速率向外发送。

优先级的调度

除了基于优先级组的调度外，对于同一优先级组内的队列，ETS 提供基于优先级的调度管理，称为第二级调度。

另外，ETS 还提供基于优先级的队列拥塞管理、队列整形、队列拥塞避免。

ETS 配置

CE6800 设备支持配置 ETS 功能。

步骤 1 配置 ETS 模板

```
[~CE6800] dcb ets-profile ets1
```

步骤 2 应用 ETS 模板

```
[~CE6800] interface 10ge 1/0/1
[~CE6800-10GE1/0/1] dcb ets enable ets1
[~CE6800-10GE1/0/1] quit
[~CE6800] commit
```

2.3.3 DCBX

产生原因

在数据中心网络融合场景下，为实现无丢包以太网，链路两端的 PFC 和 ETS 的参数配置需要保持一致。如果依靠管理员手工配置，不仅工作量庞大而且容易出错。数据中心桥接交换协议 DCBX（Data Center Bridging Exchange Protocol）作为一种链路发现协议，能够使链路两端的设备发现并交换 DCB 配置信息，大大减轻了管理员的工作量。

基本原理

DCBX 的具体功能包括：

- 发现对端设备的 DCB 配置信息。
- 发现对端设备的 DCB 配置错误。
- 远程配置对端设备的 DCB 参数。

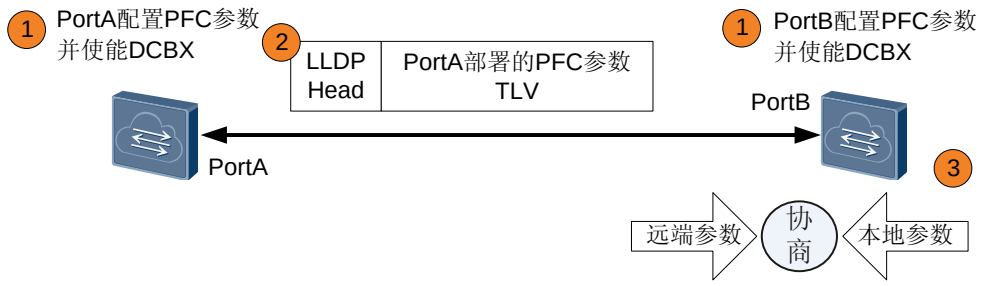
DCBX 能够交换的 DCB 配置信息如下：

- ETS 的优先级组信息
- PFC

DCBX 协议将需要交互的 DCB 配置信息封装入链路层发现协议 LLDP（Link Layer Discovery Protocol）中的 TLV 中，借由 LLDP 来进行链路两端设备的 DCB 配置交换。

下面以 DCB 中的 PFC 为例，介绍 LLDP 承载 DCBX 的实现过程。

图2-5 LLDP 承载 DCBX 的实现原理图



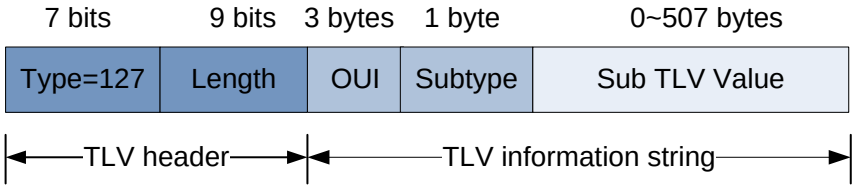
如图 2-5 所示，在 PortA 和 PortB 上分别使能 LLDP 功能，并且 PortA 上配置了允许发送 DCBX TLV 的前提下，实现过程如下：

1. PortA 和 PortB 上分别配置 PFC 参数，并使能 DCBX 功能。DCBX 模块通知 PortA 和 PortB 可以将各自配置的 PFC 参数封装到 LLDP 报文中发送给对端。
2. PortA 的 LLDP 模块根据自己的报文发送周期定期向 PortB 发送携带了 DCBX TLV 的 LLDP 报文。
3. PortB 接收到 LLDP 报文后解析出 DCBX TLV，将 PortA 的 PFC 参数通知给 DCBX 模块。DCBX 模块将 PortA 的 PFC 参数和本端配置的 PFC 参数进行比较，协商一致之后生成配置文件，保证两端配置一致。

DCBX TLV

如图 2-6 所示，DCB 的信息被封装在特定的 TLV 中。其中，Type 字段固定为 127；OUI 字段固定为 0x0080c2。

图2-6 DCBX 的 TLV 结构



DCBX TLV 的包括：ETS Configuration TLV、ETS Recommendation TLV 和 PFC Configuration TLV。具体内容如表 2-3 所示。

表2-3 DCBX TLV 的内容

TLV 名称	Subtype	Length	描述
ETS Configuration TLV	09	25	ETS 的本地配置。内容包括： • 优先级组的配置：PG ID 和优先级组的带宽占用率 • 优先级队列的配置：优先级队列 ID 和所属 PG ID
ETS Recommendation TLV	0A	25	ETS 的建议配置，通常用于协商 ETS 两端的配置，使其保持一致。内容包括： • 优先级组的配置：PG ID 和优先级组的带宽占用率 • 优先级队列的配置：优先级队列 ID 和所属 PG ID
PFC Configuration TLV	0B	6	PFC 的本地配置。内容包括： • 优先级队列 ID • 队列是否 PFC

DCBX 配置

CE6800 设备支持配置 DCBX 功能。

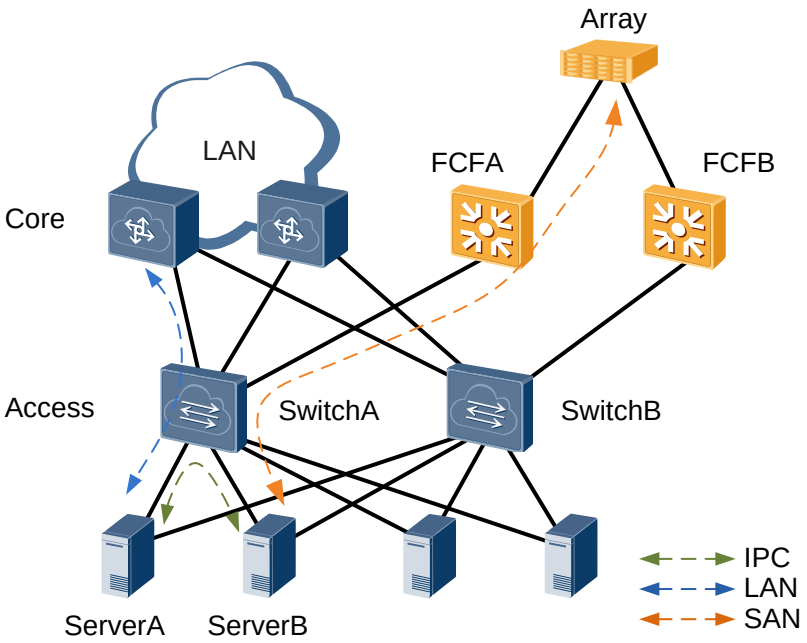
```
[~CE6800] lldp enable
[~CE6800] interface 10ge 1/0/1
[~CE6800-10GE1/0/1] lldp tlv-enable dcbx
[~CE6800-10GE1/0/1] quit
[~CE6800] commit
```

3 应用

FCoE/DCB 的典型组网应用

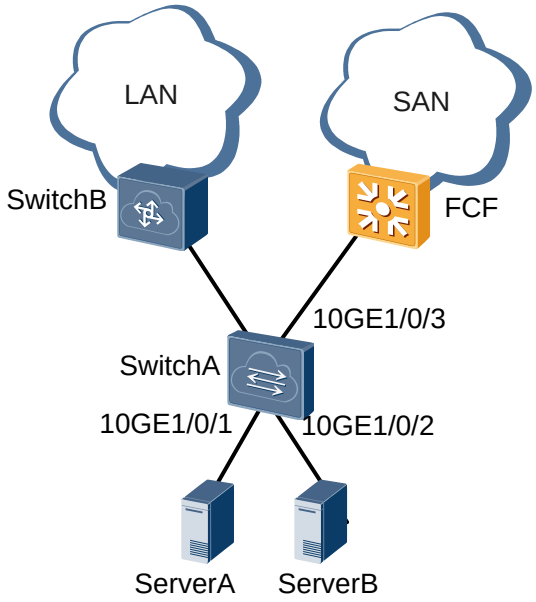
在数据中心融合网络场景中，采用 FCoE 技术实现网路融合，同时为降低用户投资成本，通常在服务器和 FCF 的之间部署接入交换机。如图 3-1 所示，Switch A 作为接入交换机，需要转发 LAN 流量、SAN 流量和 IPC 流量。组网中，为保证 ServerA 到 Array 之间的链路可靠性部署了两条独立的链路 ServerA-SwitchA-FCFA-Array 和 ServerA-SwitchB-FCFB-Array。

图3-1 FCoE/DCB 典型组网图



为了保证 SAN 流量的正确转发，在 SwitchA 上配置 FIP Snooping；为保证 LAN 流量、SAN 流量和 IPC 流量的 QoS，在 Switch A 上配置 DCB。具体配置如下所示：

图3-2 FCoE/DCB 配置组网图



SwitchA 的配置文件

```
#
sysname SwitchA
#
dcb pfc
#
dcb ets-profile ets1
 priority-group 0 queue 0 to 2 4 to 6
 priority-group 15 queue 7
 priority-group 0 drr weight 60
 priority-group 1 drr weight 40
#
fcoe FSB
 vlan 2094
#
lldp enable
#
diffserv domain ds1
 8021p-inbound 3 phb af1 green
 8021p-outbound af1 green map 3
#
interface 10GE1/0/1
 port link-type trunk
 port trunk allow-pass vlan 2094
```

```
 lldp tlv-enable dcbx
 trust upstream ds1
 dcb pfc enable mode auto
 dcb ets enable ets1
#
interface 10GE1/0/2
 port link-type trunk
 port trunk allow-pass vlan 2094
 lldp tlv-enable dcbx
 trust upstream ds1
 dcb pfc enable mode auto
 dcb ets enable ets1
#
interface 10GE1/0/3
 port link-type trunk
 port trunk allow-pass vlan 2094
 lldp tlv-enable dcbx
 fcoe role vnp
 dcb pfc enable mode auto
#
return
```

4 术语与缩略语

缩略语

缩略语	英文全称	中文全称
FCoE	Fibre Channel over Ethernet	以太网光纤通道
FC	Fibre Channel	光纤通道
SAN	Storage Area Network	存储区网络
NIC	Network Interface Card	网络接口卡

缩略语	英文全称	中文全称
HBA	Host Bus Adapter	主机总线适配器
DCB	Data Center Bridging	数据中心桥接
CNA	Converged Network Adapter	融合网络适配器
FCF	FCoE Forwarder	FCoE 数据转发器
FSB	FCoE Initialization Protocol Snooping Bridge	运行 FIP Snooping 功能的交换机
FIP	FCoE Initialization Protocol	FCoE 初始化协议
PQ	Priority Queue	优先级队列调度
DRR	Deficit Round Robin	赤字轮循队列调度
DCBX	Data Center Bridging Exchange Protocol	数据中心桥接交换协议
LLDP	Link Layer Discovery Protocol	链路层发现协议