IP网络系列丛书

NoF+存储网络解决方案

主编：汪洋

HUAWEI

版权声明

主编： 汪洋

主要参与人员： 骆兰军、虞玲玲、祝春荣、张帆、高洋洋

发布日期： 2023-10-10

发布版本： 03

前言

主编简介

汪洋，华为企业解决方案专家，2011 年加入华为，拥有美国、欧洲等国家超过 10 年海外数通市场拓展经验，对数据中心、园区网络、广域网络等领域见解深刻并具备丰富的实践经验，现任瑞士企业解决方案部长，并积极参与相关解决方案书籍的开发工作。

本书内容

本书重点介绍了华为全无损以太存储网络解决方案（简称 NoF+存储网络解决方案）产生的背景、带来的价值、使用的关键技术等内容。

NoF+存储网络解决方案是华为超融合数据中心网络 CloudFabric 3.0 面向集中式存储场景的子方案。该方案基于 OceanStor Dorado 全闪存存储系统和 CloudEngine 数据中心存储网络交换机构建，可实现存储场景端到端数据加速，充分释放全闪存性能潜力，是全闪存时代的最佳选择。

读者对象

本书适合企业的中高层管理人员和网络工程师，对存储网络有所了解，以及对 NoF+ 存储网络解决方案感兴趣的读者阅读。

符号约定

对正文中重点信息的补充说明。“说明”不是安全警示信息，不涉及人身、设备及环境伤害信息。

表示如不避免则可能导致轻微或中度伤害的具有低等级风险的危害。

目录

第1章
NoF+存储网络简介

摘要

全闪存时代背景下，传统的FC（Fibre Channel，网状通道）存储网络已经无法满足全闪存数据中心的要求，NVMe（Non-Volatile Memory express，非易失性内存主机控制器接口规范）存储协议的出现极大提升了存储系统内部的存储吞吐性能、降低了传输时延，NoF（NVMe over Fabric）存储网络应运而生。在多种Fabric技术中，NVMe over RoCE（RDMA over Converged Ethernet）被广大存储厂商所接受，成为业界NoF的主流。华为推出的NoF+存储网络解决方案，相较于标准NoF方案，在性能、可靠性、易用性上均实现了颠覆性改进，是全闪存时代的最佳选择。

1.1 数据中心存储进入全闪存时代

数字经济时代，数据成为当之无愧的关键生产要素。数据的爆发式增长，推动了数据基础设施的革新。为了适应新兴业务的高并发浪涌和对数据响应速度的需求，数据中心存储已经全面进入更高性能的全闪存时代。

2018 年起 SSD（Solid-State Drive，固态硬盘）全球发货量超过传统 HDD（Hard Disk Drive，机械硬盘），读写性能提升百倍。Gartner 预测，2022 年 SSD 在存储中占比将达到 52%，超越半壁江山，成为真正的主流。存储介质的百倍性能提升，驱动了存储协议从传统串行 SCSI（Small Computer Systems Interface，小型计算机系统接口）协议发展到高速并行的 NVMe 协议。新一代存储网络基于 RDMA（Remote Direct Memory Access，远程直接存储读取）技术，保证了全闪存 NVMe 协议的高吞吐、低时延的特性。

智能时代各种新技术、新应用层出不穷，数据呈现爆发式增长，对存储网络的稳定性、可扩展性、易用性的要求逐步提高。首先就是更加强调高性能下的稳定性，存储作为数据中心底座，为了减少存储抖动放大对业务的影响，越是在高业务负荷承载的情况下越要保持稳定；其次，更多线下业务转型线上，数据流动性极大增强，这时就需要考虑如何能够满足网络大规模建设问题，未来必然需要扩展性更强大、产业可持续发展的网络；最后还要保障在大规模扩展建设情况下的易用性，以便很好地解决业务扩张带来的挑战。

1.2 传统 FC 存储网络成为存储产业瓶颈

FC 开发于 1988 年，初衷是用来提高硬盘协议的传输带宽，侧重于数据的快速、高效、可靠传输，早期应用于 SAN（Storage Area Network，存储局域网络）。到上世纪 90 年代末，FC 存储网络（FC SAN）开始得到大规模的应用。

FC 存储网络在高性能块存储网络独占鳌头已近 30 年，过去曾有 iSCSI（Internet Small Computer Systems Interface）、FCoE（Fibre Channel over Ethernet）等基于 IP 的新技术试图颠覆 FC 存储网络，但由于这些技术在性能和功能方面仍存在短板，所以只是替代了部分中低性能要求的场景，在高性能高可靠场景仍旧是 FC 存储网络的天下。

FC 存储网络具备高可靠、稳定低时延的特点：FC 内在的协议机制支持快速故障感知通告（百 ms 级），同时 FC B2B Credit 流控机制带来了稳定低时延性能。但 FC 本身机制也带来了很多问题，传统 FC 存储网络已成为存储产业瓶颈：

- 厂商垄断，网络技术封闭：业界唯二国外厂商，存在巨大的业务连续风险
- 带宽不足，存储性能瓶颈：FC 网络最大只有 32G 带宽，满足不了业务发展需求

- 运维复杂，依赖原厂支持：FC 运维人员稀缺，运维依赖原厂响应

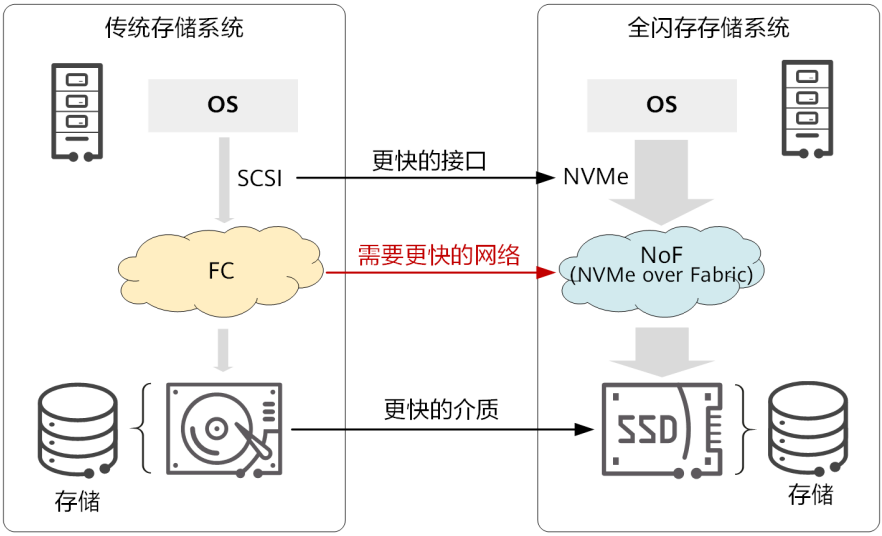
1.3 NVMe over RoCE 成为 NoF 主流

随着存储介质从 HDD 发展到 SSD，存储高性能吞吐与 SCSI 协议传输较低性能吞吐之间的矛盾日益严重，从而出现了 NVMe 存储协议。NVMe 规范了 SSD 访问接口，简化了协议复杂性，充分利用 PCIe（Peripheral Component Interconnect Express）通道的低延时以及并行性，利用多核处理器，通过降低协议交互时延，增加协议并发能力，并且精简操作系统协议堆栈，显著提高了 SSD 的读写性能。

全场景闪存化推动了数据中心的网络改革，NVMe 最大化释放了 SSD 介质的能力。更快的存储呼吁更快的网络。NoF 存储网络应运而生，通过使用 IP 网络对专用网络的创新性革新，实现了更高的带宽和更低的时延，同时也兼具 IP 易管理的优势，是更好地实现端到端 NVMe 存储网络的最佳方案。

NoF 将 NVMe 协议应用到服务器主机前端，作为存储阵列与前端主机连接的通道，可端到端取代 SAN 网络中的 SCSI 协议，构建全以太的存储 SAN 网络，如图 1-1 所示。

图1-1 NoF 的产生背景



NVMe over Fabric 中的“Fabric”，是 NVMe 的承载网络，这个网络可以是 RoCE、FC 或 TCP。具体说明如下：

- **NVMe over FC** 协议标准为 FC-NVMe，FC-NVMe 和 FC-SCSI 同样都基于 FCP，IO 交互基于 Exchange。FC-NVMe 基于传统的 FC 网络，通过升级主机驱动和交换机支持，FC-SCSI 和 FC-NVMe 能同时运行在同一个 FC 网络中。FC-NVMe 能最大化继承传统的 FC 网络，复用网络基础设施，基于 FC 物理网络发挥 NVMe 新协议的优势。
- **NVMe over TCP** 基于现有的 IP 网络，采用 TCP 协议传输 NVMe，在网络基础设施不变的情况下实现了端到端 NVMe。虽然 NVME over TCP 网络性能弱于 FC 和 RoCE，但整体性能通过 NVMe 得到提升，对比 iSCSI 仍有大幅度的提升。而且 NVME over TCP 对网络的要求比较低，具有更强大的兼容性，不需要单独建设无损网络，传统以太网即可支持，因此在不追求高性能的情况下，NVMe over TCP 将是未来市场的普遍选择。
- **NVMe over RoCE** 是 **NVMe over RDMA** 的一种，RDMA 是承载 NoF 的原生网络协议，RDMA 协议除了 RoCE 外还包括 IB（InfiniBand）和 iWARP（Internet

Wide Area RDMA Protocol）。其中，基于以太网的 RoCE 目前已成为 RDMA 的主流网络承载方式。NVMe over RDMA 协议比较简单，直接把 NVMe 的 IO 队列映射到 RDMA QP（Queue Pair）连接，通过 RDMA SEND，RDMA WRITE，RDMA READ 三个语义实现 IO 交互。NVMe over RoCE 基于融合以太网的 RDMA 技术承载 NVMe 协议。

NVMe over Fabric 作为集中存储网络的下一跳，会形成两个主要市场：一个是以 NVMe over RoCE/FC 为主，主打高性能和高可靠的市场；另一个是以 NVMe over TCP 为主，主打扩展性和兼容性的市场。

三种方案相比较，基于以太网的 RoCE 比 FC 性能更高（更高的带宽、更低的时延），同时兼具 TCP 的优势（全以太化、全 IP 化），因此 NVMe over RoCE 是 NoF 最优的承载网络方案，也已成为业界 NoF 的主流技术。

1.4 NoF+存储网络，全闪存时代最佳选择

如图 1-2 所示，基于以太网的 RoCE 在存储性能、带宽方面比 FC 有显著优势，但替换 FC，联接全闪存，标准的 NVMe over RoCE 还需在 3 个方面加强完善：

1. 网络性能：零丢包

   网络零丢包是存储网络的基本需求，传统以太网络拥塞易丢包。

2. 可靠性：秒级主备切换

   存储为了可靠性，会构建多个网络平面，切换时间需<1s。

3. 易用性：即插即用

   FC 存储网络场景单一、配置简单，当前以太网络还需针对存储场景适应性改进。

图1-2 RoCE 与 FC 的差异

存储网络关注焦点		FC	RoCE
网络性能	带宽	32/64G	400G
	丢包	稳定零丢包	拥塞易丢包
可靠性	升级中断时间	<1s	<1s
	主备切换时间	<1s	<8~15s
易用性	日常运维	智能定位	智能运维
	存储部署	即插即用	手工配置

基于当下业界主流的标准 NoF 方案，华为依靠在网络和存储领域的深厚积累，进一步从网络性能、可靠性和易用性这三点都进行提升，基于智能无损网络面向集中式存储场景提出了 **NoF+**解决方案，将数据中心存储网络进一步推向更广阔的发展空间。

- **网络性能增强**：华为 NoF+方案改变了传统以太静态水线方式，对网络预测性能力进行专项优化，通过样本计算，针对特定场景，通过算法进行精准的控制，从而预判业务对网络的诉求，提前做出优化，实现高吞吐带宽，进一步提升性能。

  同时，华为 NoF+突破 TCP 场景，基于存储 TCP Flexbuffer 调优算法，识别存储业务大流小流，动态调整缓存占用率，减少 TCP 流完成时间，对比传统以太网络提升 IOPS 性能 10%+，比 iSCSI 提升 25%。

- **可靠性增强**：保障业务系统可靠是存储的根基，比如存储的秒级切换功能就是可靠性的关键保障之一，标准以太缺乏故障主动发现和通知能力，华为 NoF+实现了从事后被动响应到主动通知、提前识别拥塞和故障。当一个节点出现故障，业务会以亚秒级响应速度切换，在高性能运行的前提下，也能维持系统的稳定可靠。
- **易用性增强**：华为的存储与网络产品强强联合，打造了“即插即用”的方案，实现了一键式扩容，自动化管理，增强了在未来建设时的易用性。

华为 NoF+方案基于全闪存数据中心和超融合以太网络，通过最新一代 OceanStor Dorado 全闪存存储系统和 CloudEngine 数据中心存储网络交换机的联合创新，针对传统专用存储网络，在性能、可靠性、易用性上均实现了颠覆性改进，是全闪存时代的最佳选择。

第2章
NoF+存储网络价值

摘要

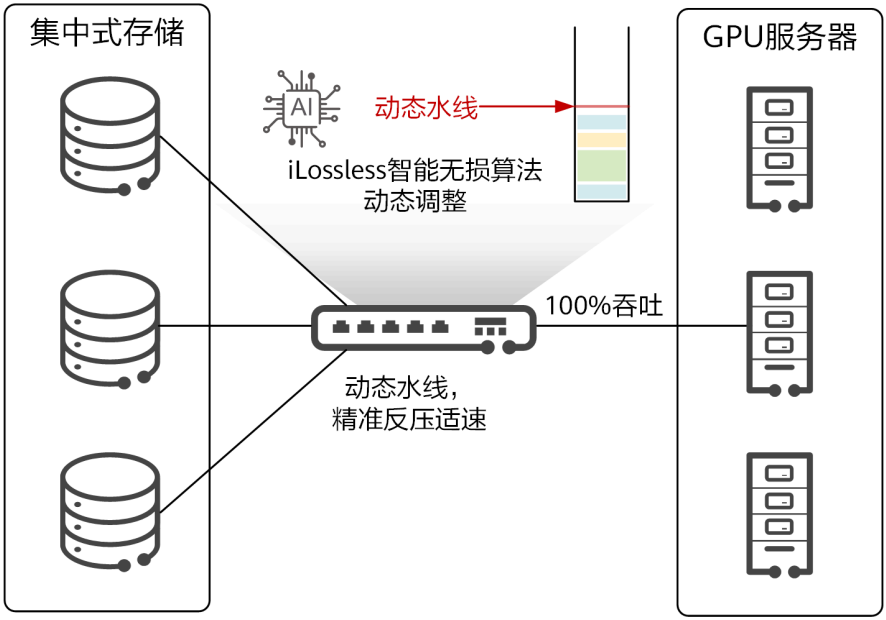
NoF+存储网络解决方案在可靠性上做到从事后被动响应到主动通知、提前识别拥塞和故障，配合华为OceanStor Dorado全闪存存储实现故障秒级切换；在性能上改变传统静态水线方式，针对特定场景，通过算法实现预判，提前做出优化，可进一步释放华为OceanStor Dorado全闪存强大的性能。同时还打造“即插即用”方案，实现一键式扩容，自动化管理，提升了在未来建设时的易用性。

2.1 本地智能无损零丢包，全网流量智能调优

传统以太网天然有丢包，0.1%的丢包可造成网络吞吐下降 50%。NoF+存储网络解决方案推出 iLossless 智能无损算法，通过实时感知网络流量模型，动态调整水线，可实现存储网络 100%吞吐下的零丢包，降低丢包引起的等待损失。IOPS

（Input/output Operations Per Second，每秒进行读写操作的次数）较 FC 网络提升 93%，时延降低 49%。

图2-1 全网流量智能调优

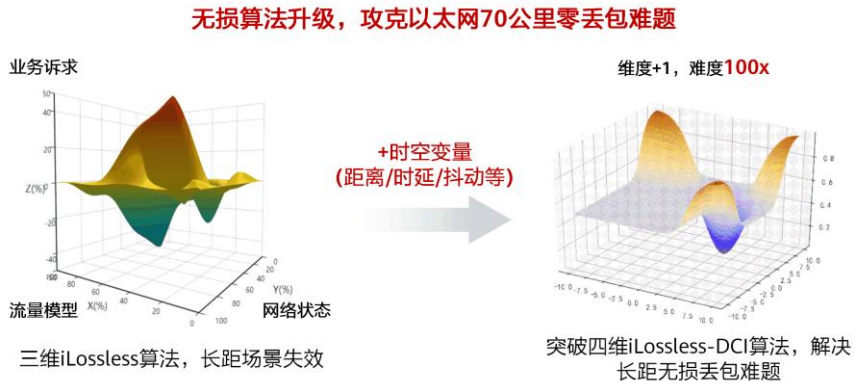


2.2 跨 DC 存储双活零丢包，攻克长距无损难题

NoF+存储网络在短距 iLossless 智能无损交换算法基础上，增加了时空变量，突破四维 iLossless-DCI 算法，解决长距无损丢包难题。基于大数据的卷积预测将随机流量确定化，提前应对流量变化，从而实现了以太网在长距范围的无损传输。

NoF+存储网络，面向存储场景提供双活全以太存储网络，针对同城双活等极端性能场景，支持 100GE 链路 70 公里长距无损传输，跨 DC 链路数量较 FC 可减少 90%。

图2-2 长距无损

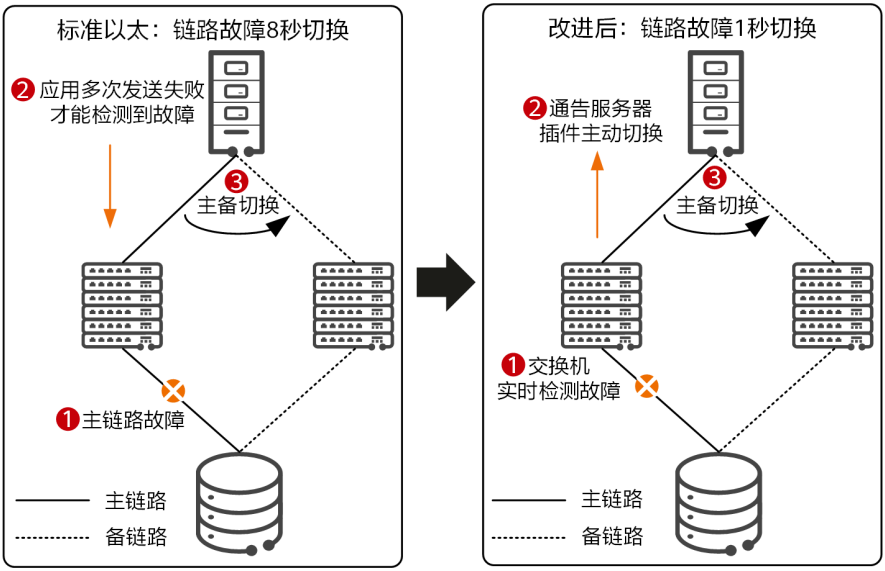


2.3 故障主动发现，主备服务器秒级切换

NoF+存储网络可实时监测链路健康状态，通过全网通告，实现主备服务器秒级切换，保证整网 100%零丢包。

传统以太网缺乏故障主动发现和通知机制，发生故障时链路切换时间长，造成存储业务中断。NoF+存储网络解决方案推出智能感知特性，交换机可毫秒级主动通告故障，与存储设备联动完成亚秒级故障倒换，实现网络单点故障下存储业务零影响。

图2-3 故障主动感知

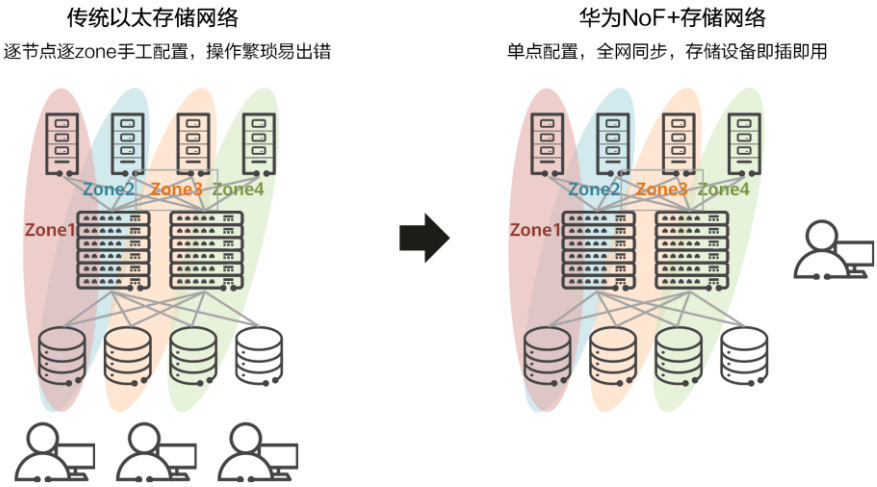


2.4 存储即插即用，一键式扩容

存储场景下传统以太网络需要逐点手工配置，操作复杂且易出错。NoF+存储网络解决方案实现存储和交换机的智能联动，支持存储设备的即插即用和一键式扩容。业务变更只需在单点配置，即可自动同步到全网，业务发放效率显著提升。

华为提出的 iNOF（Intelligent Lossless NVMe Over Fabric，智能无损存储网络）技术与 OceanStor Dorado 存储的 SNSD（Storage Network Smart Discovery，存储网络智能发现）特性联动，支持即插即用，一键安装建链，简单高效。SNSD 开关开启后，主机会感知到该 RoCE 端口下所有逻辑端口的状态变化，帮助主机根据逻辑端口状态去判断是否自动建立或断开连接。

图2-4 存储设备即插即用



第3章
NoF+存储网络架构

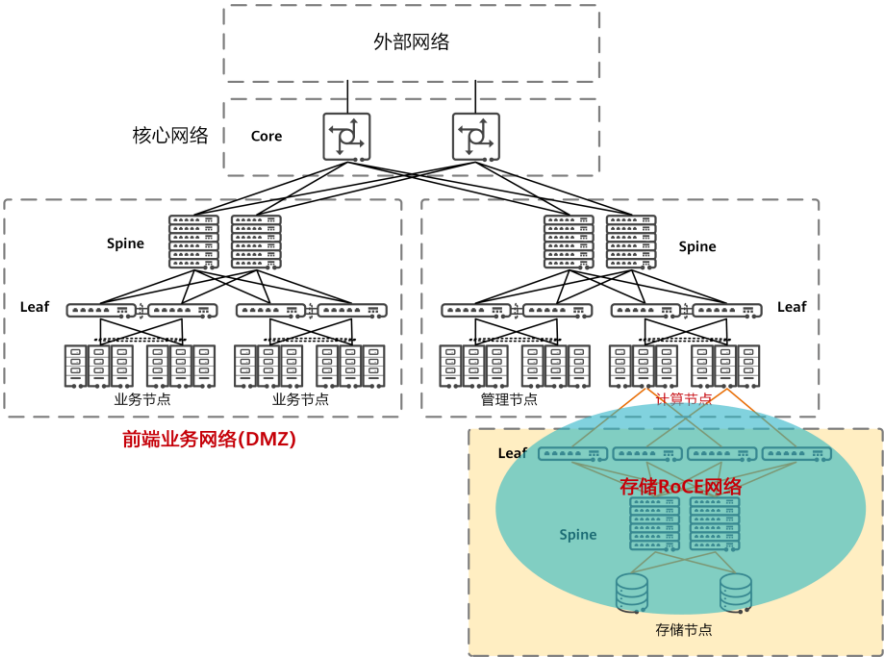
摘要

本章介绍了华为NoF+存储网络的网络架构以及核心组件。

3.1 网络架构

在数据中心常规组网里面，存储网络只是其中的一部分，集中式存储是一个独立的网络，与业务网络在物理上隔离，如图 3-1 所示。

图3-1 数据中心集中式存储网络架构图

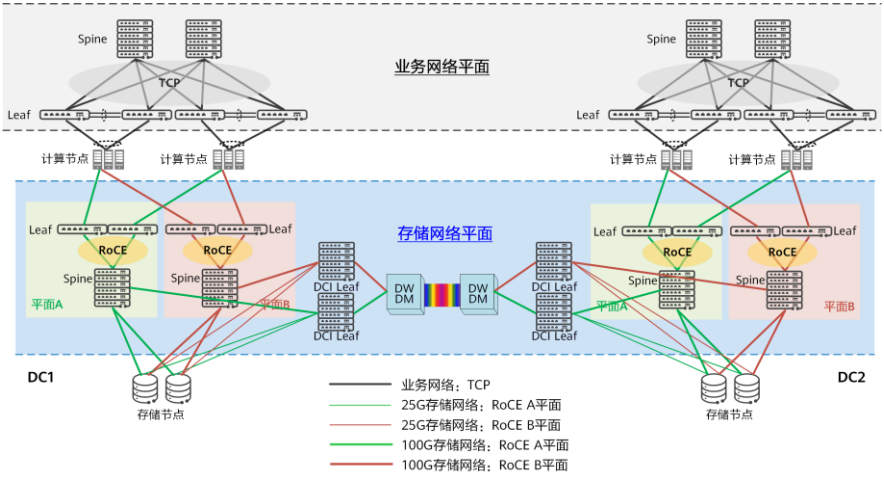


- 业务网络：是指业务服务器对外提供服务通信网络，该网络与外部网络互连互通。
- 计算网络：是指运行 OLTP/OLAP（Online Transaction Processing/Online Analytical Processing）数据库的后台服务计算节点所组成的物理网络，使用不同的网卡连接业务网络和存储网络，实现业务网络和存储网络之间物理隔离，避免相互影响。
- 存储网络：是指计算服务器访问存储数据时使用通信网络，该网络一般是独立的物理网络。为了保证数据高可靠，存储网络支持 DC 级容灾，支持同城双活存储网络，确保业务系统发生设备故障、甚至单数据中心故障时，业务无感知自动切换，实现 RPO（Recovery Point Objective）=0，RTO（Recovery Time Objective）≈0。

数据中心为了容灾考虑，需要实现多数据中心互通。同城两个数据中心互为备份，且都处于运行状态。当一个数据中心发生设备故障，甚至数据中心整体故障时，业务自动切换到另一个数据中心，解决了传统灾备中心不能承载业务和业务无法自动切换的问题。提供给用户高级别的数据可靠性以及业务连续性的同时，提高存储系统的资源利用率。

在集中式存储下，DC 间同城互联的一般组网如图 3-2 所示。为了实现同城读写支持 NVME over ROCE，需要实现同城无损网络，即需要一套跨 DC 的无损网络，每个 DC 部署两台支持智能长距无损的 DCI Leaf，中间通过波分设备或者裸光纤直连实现双平面，实现端到端的 ROCE 无损网络。

图3-2 DC 间同城互联一般组网示意图

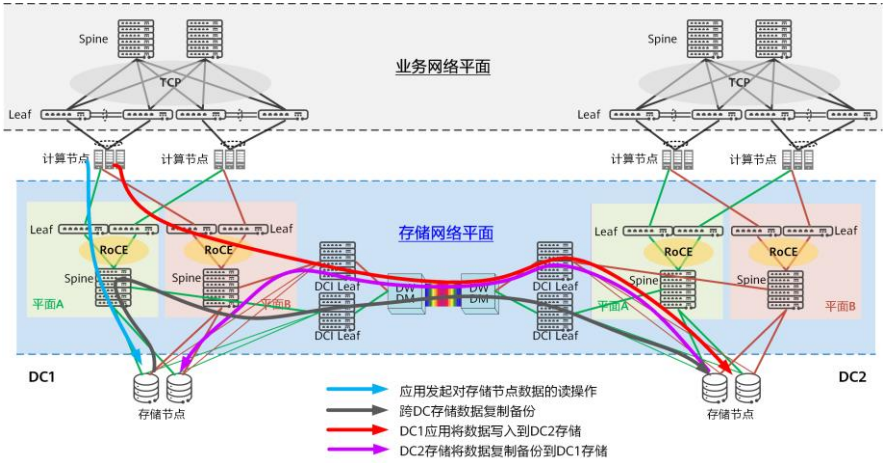


在本场景中，常见的流量有以下几种类型（图 3-3 中箭头表示主动发起的方向）:

- 由应用发起对存储节点数据的读操作，此时计算节点访问同 DC 中的存储节点，存储节点返回相应数据，如图中蓝色线条所示。

- 当应用同时还需要对存储写数据时，除了对本 DC 内的存储节点执行写操作，在存储系统之间，同时也会发起对另外 DC 中存储的写操作，作为数据的复制备份，如图中黑色线条所示。
- 当应用在写本 DC 存储时，会先探测本 DC 内存储节点是否可用，如果不可用，则应用会将数据写入到 DC2 中的存储节点中，如图中红色线条所示。然后 DC2 中的存储节点，再尝试将数据复制写如到 DC1 的存储节点中，如图中的紫色线条所示。

图3-3 DC 间同城互联存储业务流量模型



RoCEv2 协议将 RDMA 迁移到了 ETH/IP 网络，使得 ETH/IP 网络支持 HPC、AI、分布式存储和集中式存储。NoF+存储网络解决方案借助 RoCEv2 技术改变了传统数据中心前端业务网采用以太网、计算网采用 IB 网、存储网采用 FC 网的异构模式，让智能无损网络实现三网合一成为可能，全部采用以太的方式部署。

3.2 核心组件

CloudEngine 数据中心存储网络交换机

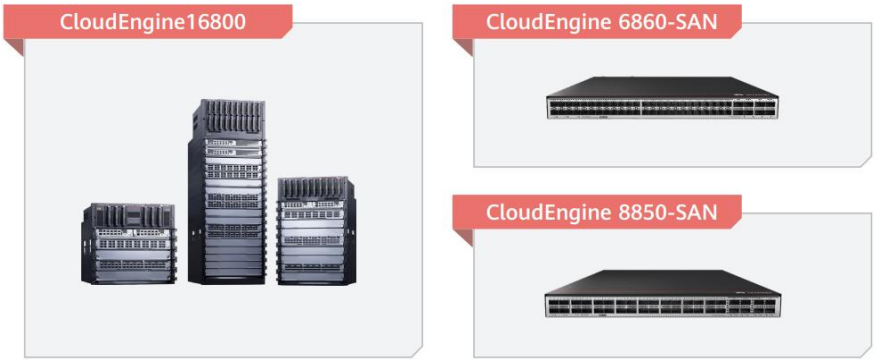
CloudEngine 数据中心存储网络交换机是华为公司面向数据中心全闪存存储网络推出的新一代高性能、高可靠、低时延、易运维的交换机，不仅支持独创 iLossless 智能算法，让 NVMe 运行更高效，全面释放全闪存潜力；同时支持 NoF+技术，实现存储网络即插即用，故障快速感知。

CloudEngine 数据中心存储网络交换机包含以下型号：

- CloudEngine 16800 配套 SAN 系列接口板（CEL72XS-SAN、CEL48CQ-SAN）
- CloudEngine 8850-SAN
- CloudEngine 6860-SAN

CloudEngine 6860-SAN、CloudEngine 8850-SAN 与 CloudEngine 16800 配合构建智能无损 DCN 方案，满足全闪存时代存储网络的需求。

图3-4 CloudEngine 数据中心存储网络交换机



OceanStor Dorado 全闪存存储系统

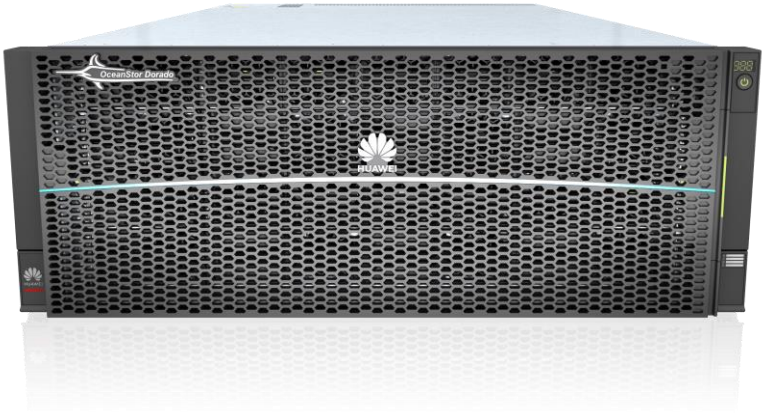
OceanStor Dorado 6800/18500/18800 V6 高端智能全闪存

OceanStor Dorado 6800/18500/18800 V6 存储系统是华为根据存储产品应用现状和存储技术未来发展趋势，针对企业大中型数据中心，推出的新一代全闪存高端存储系统，聚焦于大中型企业核心业务（企业级数据中心、虚拟数据中心以及云数据中心等），能够满足大中型数据中心高性能、高可靠、高效率的业务需求。

OceanStor Dorado 采用全新一代的 SmartMatrix 智能矩阵架构，该架构能实现业界唯一的控制框 2 坏 1 业务不中断，控制器 8 坏 7 业务不中断，能够满足大中型企业核心业务可靠性的要求。同时 OceanStor Dorado 由 AI 智能芯片加持，能够满足数据中心大型数据库 OLTP/OLAP、高性能计算（HPC，High-performance Computing）、数字媒体、Internet 运营、集中存储、备份、容灾和数据迁移等不同业务应用的需求。

OceanStor Dorado 不但能够为数据中心提供性能出色的存储服务。同时，提供各种完善的数据备份和容灾方案，保证数据业务顺利、安全的运行。除此之外，OceanStor Dorado 还提供易于使用的管理方式和方便快捷的本地/远程维护方式，大大降低了设备管理和维护的成本。

图3-5 OceanStor Dorado 6800/18500/18800 V6



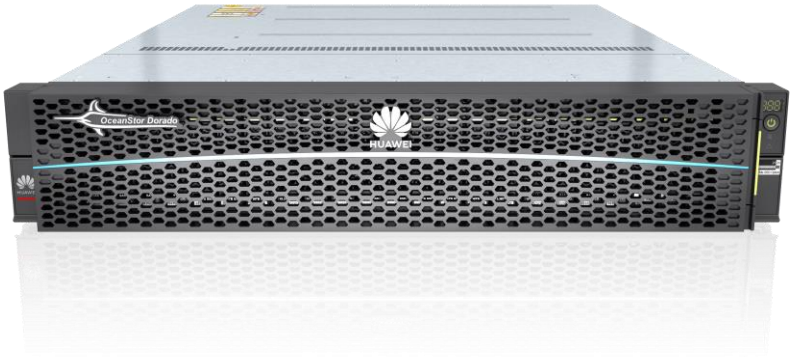
OceanStor Dorado 5300/5500/5600 V6 中端智能全闪存

OceanStor Dorado 5300/5500/5600 V6 存储系统是华为采用专为闪存设计的FlashLink 技术，面向企业关键业务打造的新一代全闪存存储产品，能够满足大中型企业用户对大容量数据存储、高速数据存取、高可用性、高利用率、绿色环保和易于使用的要求。

OceanStor Dorado 以业界领先的性能、多种效率提升机制为支撑，为用户提供了高性能、全方位的解决方案，使用户投资收益比最大化，能够满足大型数据库OLTP/OLAP、高性能计算、服务器虚拟化和虚拟桌面（VDI，Virtual Desktop Infrastructure）等不同业务应用的需求。

OceanStor Dorado 不但能够为企业用户提供高性能、高存储效率的存储服务，而且支持各种先进的数据备份和容灾技术，保证数据业务顺利、安全的运行。除此之外，OceanStor Dorado 还提供易于使用的管理方式和方便快捷的本地/远程维护方式，大大降低了设备管理和维护的成本。

图3-6 OceanStor Dorado 5300/5500/5600 V6



第4章
NoF+存储网络关键技术

摘要

华为NoF+方案通过流量控制技术、拥塞控制技术、智能无损存储网络技术，攻克了传统以太网丢包可靠性等难题，具备高性能、高可靠、易维护三大亮点。

4.1 流量控制技术

流量控制是端到端的，需要做的是抑制发送端的发送速率，以便接收端来得及接收，防止设备端口在拥塞的情况下出现丢包。华为提供了 PFC 死锁检测和死锁预防，提前预防 PFC 死锁的发生。

PFC 优先级流量控制

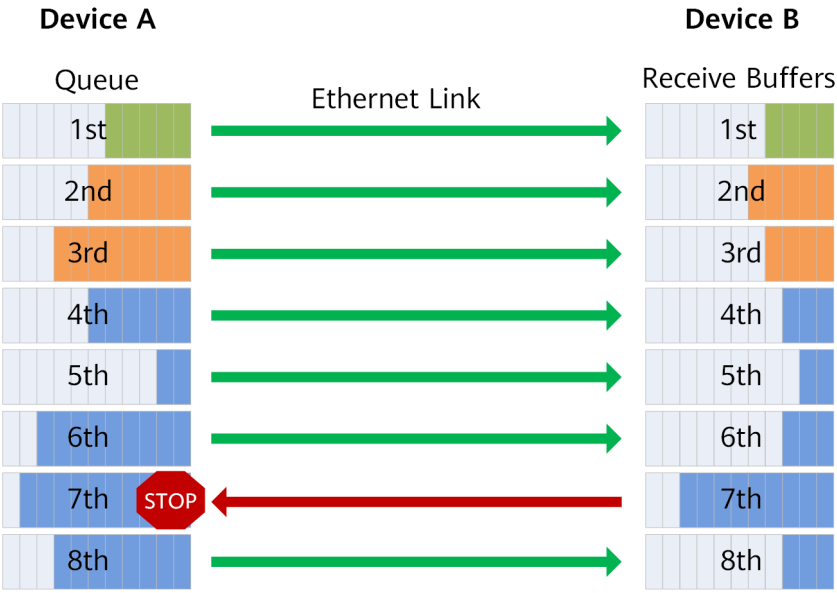
PFC（Priority-based Flow Control，基于优先级的流控制）也称为 Per Priority Pause 或 CBFC（Class Based Flow Control），是对 Pause 机制的一种增强。当前以太 Pause 机制（IEEE 802.3 Annex 31B）也能达到无丢包的要求。当下游设备发现接

收能力小于上游设备的发送能力时，会主动发 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 机制的流量暂停是针对整个接口，即在出现拥塞时会将链路上所有的流量都暂停。

而 PFC 允许在一条以太网链路上创建 8 个虚拟通道，并为每条虚拟通道指定一个优先等级，允许单独暂停和重启其中任意一条虚拟通道，同时允许其它虚拟通道的流量无中断通过。这一方法使网络能够为单个虚拟链路创建无丢包类别的服务，使其能够与同一接口上的其它流量类型共存。

如图 4-1 所示，DeviceA 发送接口分成了 8 个优先级队列，DeviceB 接收接口有 8 个接收缓存（buffer），两者一一对应，形成了网络中 8 个虚拟化通道，缓存大小不同使得各队列有不同的数据缓存能力。

图4-1 PFC 的工作机制



当 DeviceB 的接口上某个接收缓存产生拥塞时，即某个设备的队列缓存消耗较快，超过一定阈值（可设定为端口队列缓存的 1/2、3/4 等比例），DeviceB 即向数据进入的方向（上游设备 DeviceA）发送反压信号“STOP”。

DeviceA 接收到反压信号，会根据反压信号指示停止发送对应优先级队列的报文，并将数据存储在本地接口缓存。如果 DeviceA 本地接口缓存消耗超过阈值，则继续向上游反压，如此一级级反压，直到网络终端设备，从而消除网络节点因拥塞造成的丢包。

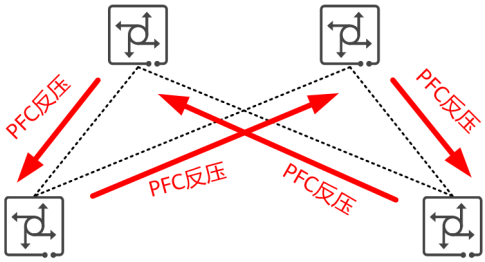
报文在以太网络中的无丢包传输是通过 PFC 流控机制实现的。设备会为端口上的 8 个队列设置各自的 PFC 门限值，当队列已使用的缓存超过 PFC 门限值时，则向上游发送 PFC 反压通知报文，通知上游设备停止发包；当队列已使用的缓存降低到 PFC 门限值以下时，则向上游发送 PFC 反压停止报文，通知上游设备重新发包，从而最终实现报文的无丢包传输。

PFC 死锁检测

PFC 死锁（PFC DeadLock），是指当多个交换机之间因为环路等原因同时出现拥塞，各自端口缓存消耗超过阈值，而又相互等待对方释放资源，从而导致所有交换机上的数据流都永久阻塞的一种网络状态。

正常情况下，PFC 中流量暂停只针对某一个或几个优先级队列，不针对整个接口进行中断，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。然而当发生链路故障或设备故障时，在路由重新收敛期间，网络中可能会出现短暂环路，会导致出现一个循环依赖缓冲区（Cyclic Buffer Dependency）。如图 4-2 所示，当 4 台交换机都达到 PFC 门限，都将同时向对端发送 PFC 反压帧，这个时候该拓扑中所有交换机都处于停流状态。

图4-2 PFC 死锁示意图



PFC 死锁检测通过以下几个过程对 PFC 死锁进行全程监控，当设备在死锁检测周期内持续收到 PFC 反压帧时，将不会响应。

1. **死锁检测**

   Device2 的端口收到 Device1 发送的 PFC 反压帧后，内部调度器将停止发送对应优先级的队列流量，并开启定时器，根据设定的死锁检测和精度开始检测队列收到的 PFC 反压帧。

2. **死锁判定**

   若在设定的 PFC 死锁检测时间内该队列一直处于 PFC-XOFF（即被流控）状态，则认为出现了 PFC 死锁，需要进行 PFC 死锁恢复处理流程。

3. **死锁恢复**

   在 PFC 死锁恢复过程中，会忽略端口接收到的 PFC 反压帧，内部调度器会恢复发送对应优先级的队列流量，也可以选择丢弃对应优先级的队列流量，在恢复周期后恢复 PFC 的正常流控机制。若下一次死锁检测周期内仍然判断出现了死锁，那么将进行新一轮周期的死锁恢复流程。

4. **死锁控制**

   若上述死锁恢复流程没有起到作用，仍然不断出现 PFC 死锁现象，那么可以设定在一段时间内出现多少次死锁后，强制进入死锁控制流程。比如设定一段时间内，PFC 死锁触发了一定的次数之后，认为网络中频繁出现死锁现象，存在极大风险，此时进入死锁控制流程，设备将自动关闭 PFC 功能，需要手动恢复。

PFC 死锁预防

PFC 死锁检测功能在死锁检测周期内持续收到 PFC 反压帧时，交换机设备可以通过不响应反压帧的方式去解除 PFC 死锁现象。然而这种事后解锁的方式只能解决极低概率出现 PFC 死锁的场景，对于一些由于多次链路故障等原因出现环路的网络，在 PFC 死锁恢复流程后瞬间又会进入 PFC 死锁状态，网络吞吐将受到很大影响。

PFC 死锁预防正是针对典型 CLOS 组网的一种事前预防的方案，通过识别易造成 PFC 死锁的业务流，修改队列优先级，改变 PFC 反压的路径，避免 PFC 反压帧形成环路，从而预防 PFC 死锁的发生。

4.2 点刹式流控

点刹式流控（也称为 ABS PFC 流控）是一种基于优先级的流量控制技术。周期性扫描接口优先级队列的缓存占用情况，通过向上游设备发送 PFC 反压帧控制周期内需要上游设备停止发送流量的时长，以持续调整流量的发送与暂停。

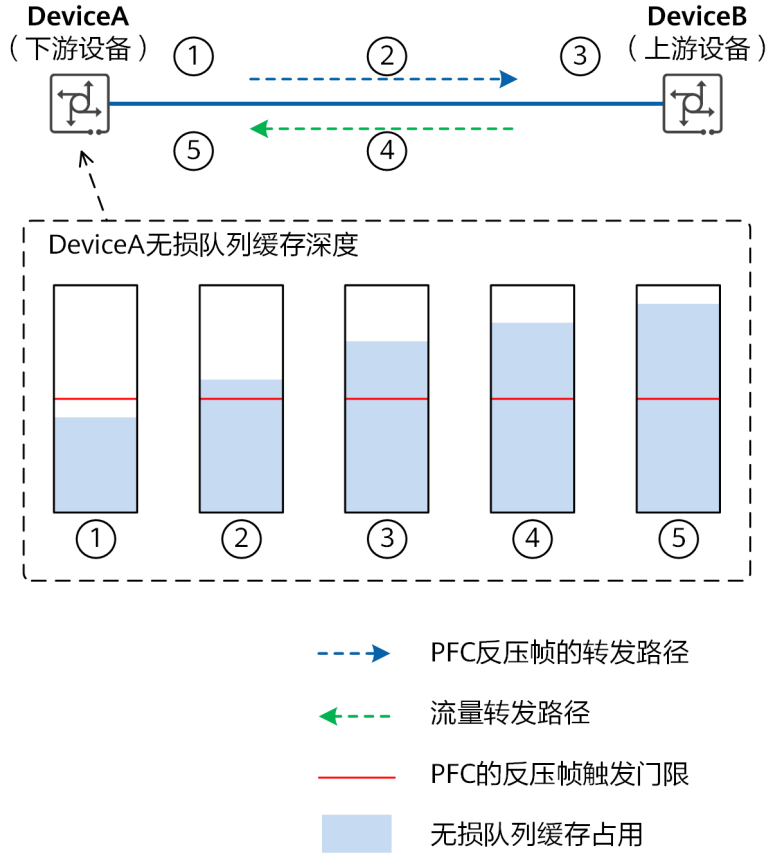
在数据中心互联的长距场景中，两个数据中心的入口设备间距离很远，若一端设备出现缓存拥塞，从该设备发送 PFC 反压帧给对端设备到停止接收对端设备发来流量的时间差内，设备需要有足够的缓存空间吸收这段时间内对端设备发来的流量，以保证长距无损。在缓存空间大小和带宽一定的情况下，点刹式流控依靠短周期、高频率、持续少量调节流量发送与暂停的机制，能够比传统 PFC 支持更长距离的长距无损场景。

传统 PFC 流控

传统 PFC 控制上游设备停止发送流量的过程

传统的优先级流量控制 PFC 控制上游设备停止发送流量的过程如图 4-3 所示。（B 表示接口带宽，T 表示 DeviceA 与 DeviceB 之间的单向转发时延。）

图4-3 PFC 控制上游设备停止发送流量的过程



1. DeviceA 的无损队列缓存占用小于反压帧触发门限时，DeviceB 持续向 DeviceA 发送流量，DeviceA 的无损队列缓存占用逐渐增大。
2. 当 DeviceA 的无损队列缓存占用超过反压帧触发门限，DeviceA 向 DeviceB 发送 PFC 反压帧通知 DeviceB 停止发送流量。在 PFC 反压帧未到达 DeviceB 的 T 时间内，DeviceB 仍在向 DeviceA 发送流量，DeviceA 的无损队列缓存占用继续增大。

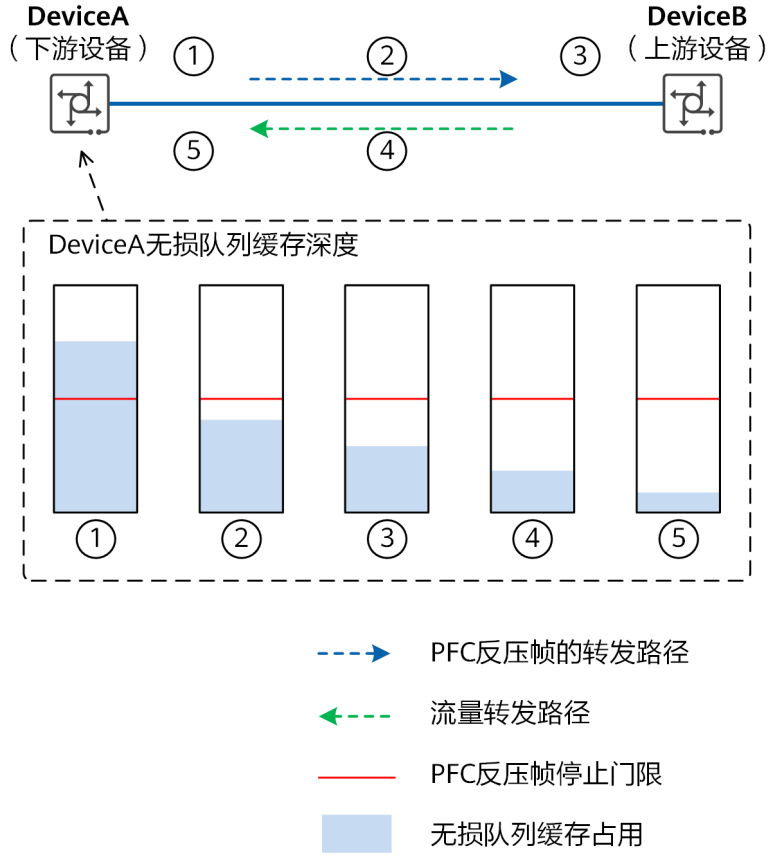
3. DeviceB 接收到 DeviceA 发来的 PFC 反压帧，停止向 DeviceA 发送流量。
4. 在 DeviceB 停止向 DeviceA 发送流量时，DeviceB 停流前发送的报文仍在发往 DeviceA，且需要经过 T 时间才会到达 DeviceA，DeviceA 的无损队列缓存占用在这 T 时间内继续增大。
5. DeviceA 停止接收到 DeviceB 发来的流量，DeviceA 的无损队列缓存占用达到最大，并开始逐渐减小。

在 PFC 控制上游设备停流的过程中，为保证无损队列不发生丢包，DeviceA 的无损队列缓存空间需要在超过反压帧触发门限后继续吸收 2T 时间内接收到的流量 B*2T。

传统 PFC 控制上游设备停流后重新发送流量的过程

传统的优先级流量控制 PFC 控制上游设备停流后重新发送流量的过程如图 4-4 所示。

图4-4 PFC 控制上游设备停流后重新发送流量的过程



1. DeviceA 的无损队列缓存占用大于反压帧停止门限时，DeviceB 一直不向 DeviceA 发送流量，DeviceA 的无损队列缓存占用逐渐减小。
2. 当 DeviceA 的无损队列缓存占用小于反压帧停止门限，DeviceA 向 DeviceB 发送 PFC 反压帧通知 DeviceB 开始发送流量。在 PFC 反压帧未到达 DeviceB 的 T 时间内，DeviceB 一直没有向 DeviceA 发送流量，DeviceA 的无损队列缓存占用继续减小。

3. DeviceB 接收到 DeviceA 发来的 PFC 反压帧，开始向 DeviceA 发送流量。
4. 在 DeviceB 开始向 DeviceA 发送流量时，DeviceB 发往 DeviceA 的流量需要经过 T 时间才会到达 DeviceA，DeviceA 的无损队列缓存占用在这 T 时间内继续减小。
5. DeviceA 开始接收到 DeviceB 发来的流量，DeviceA 的无损队列缓存占用达到最小，并开始逐渐增大。

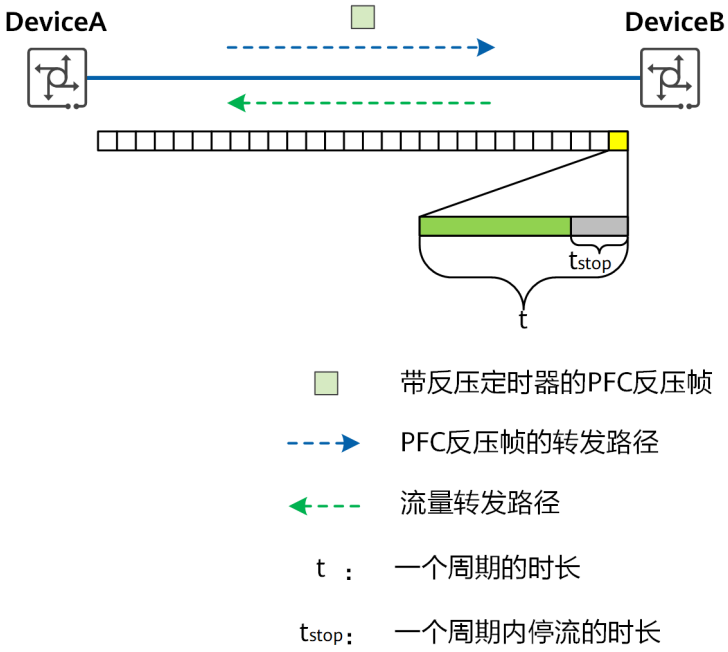
在 PFC 控制上游设备停流后重新发送流量的过程中，为保证无损队列不欠吞吐，DeviceA 的无损队列缓存空间占用需要在小于反压帧停止门限后的 2T 时间内一直大于 0，因此反压帧停止门限至少为 B*2T。

在流量控制过程中，为保证缓存占用减少，拥塞得到缓解后再停止反压，反压帧停止门限应小于反压帧触发门限。因此，为保证无损队列无丢包、高吞吐，队列缓存空间大小至少为 2*（B*2T）。

点刹式流控

点刹式流控周期性扫描无损队列的缓存占用情况，计算一个周期（时长为 t，us 级）内需要上游设备停止发送流量的时长 t_{stop}。若 $t_{stop}>0$，则通过向上游设备发送带反压定时器的 PFC 反压帧，控制上游设备在对应周期内停流时长 t_{stop} 后再发送流量。（一个周期时长为 t，t 远小于两设备间的单向转发时延 T，且 $0 \leqslant t_{stop} \leqslant t$。）

图4-5 点刹式流控工作原理图

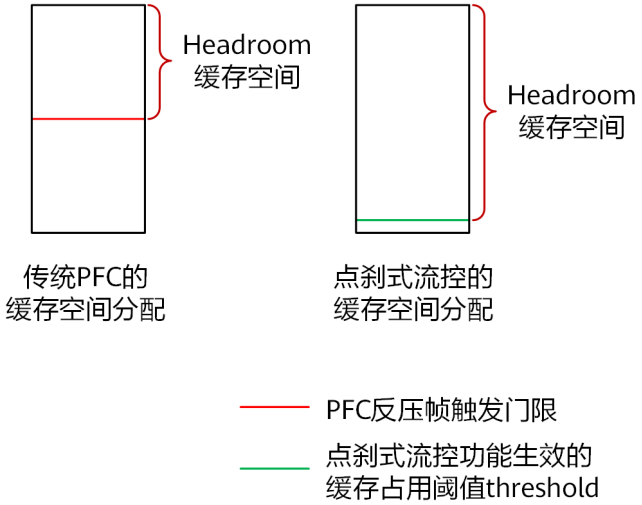


- 队列缓存深度上涨：当前周期需要上游设备停流一段时间，通过发送带停流时长的 PFC 反压帧，精准通知上游设备停流时长；上游设备收到该流控帧后，按照流控帧中的定时器设置停流时长，当定时器超时后就开始自动发送流量，不需要等待下游设备触发反压帧停止通告。
- 队列缓存深度下降或不变：不向上游设备发送 PFC 反压帧。

无损队列的 Headroom 缓存空间用于存储本队列发送 PFC 反压帧之后到停止接收上游设备发送的报文这段时间内收到的报文，以防这段时间内的报文被丢弃。根据上文对传统 PFC 流控机制的分析，传统 PFC 的反压帧触发门限值大小至少为 $B_{PFC}*2T_{PFC}$，Headroom 缓存空间大小至少为 $B_{PFC}*2T_{PFC}$，因此缓存空间占用至少需要 **$2*(B_{PFC}*2T_{PFC})$**。（B_{PFC} 表示接口带宽，T_{PFC} 表示两端设备之间的单向转发时延）

点刹式流控没有反压帧触发门限，当缓存占用超过阈值 threshold 时，点刹式流控开始生效。从一个周期结束时发送反压帧到该反压帧控制的流量发送到本设备需要经过的时长为 $2T_{ABS}$，为保证无损队列无丢包且不欠吞吐，Headroom 缓存空间大小至少需要 $B_{ABS}*2T_{ABS}$，因此缓存空间占用至少为 $B_{ABS}*2T_{ABS}$+threshold。为了达到最小缓存支持最大长距的效果，建议阈值 threshold 设置为 0，此时点刹式流控的缓存空间占用约为 **$B_{ABS}*2T_{ABS}$**。（B_{ABS} 表示接口带宽，T_{ABS} 表示两端设备之间的单向转发时延）

图4-6 传统 PFC 和点刹式流控的缓存空间分配对比



因此，在传统 PFC 和点刹式流控的带宽和设备距离一定的情况下，即 $B_{PFC}=B_{ABS}$、$T_{PFC}=T_{ABS}$，点刹式流控所需的缓存空间大小几乎是传统 PFC 的一半。

由于点刹式流控这种短周期、高频率、持续少量控制流量发送与暂停的特点，在带宽和设备距离一定的情况下，点刹式流控比 PFC 占用的缓存空间少很多；在缓存空间大小和带宽一定的情况下，点刹式流控能够比 PFC 支持更长距离的长距无损场景。

4.3 AI ECN

AI ECN（Artificial Intelligence Explicit Congestion Notification）是一种根据现网流量模型智能地调整无损队列的 ECN 门限的功能，可以保障零丢包下的低时延和高吞吐，让无损业务达到最优性能。

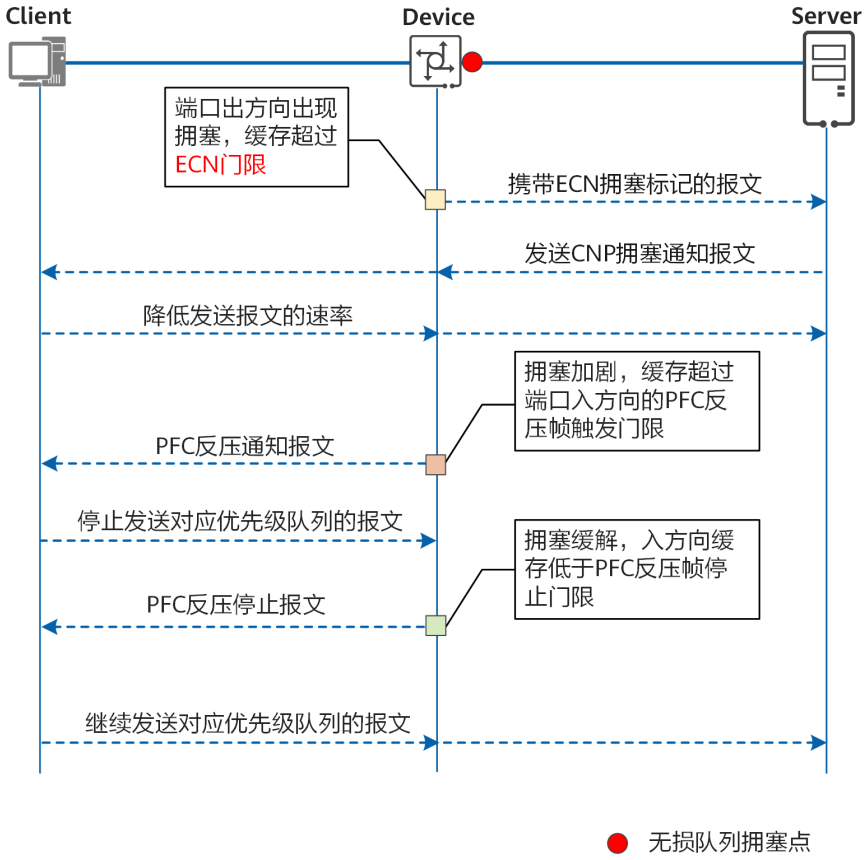
无损队列的动态 ECN 门限功能可以根据网络流量 N 对 1 的 Incast 值、大小流占比来动态调整无损队列的 ECN 门限，在尽量避免触发网络 PFC 流控的同时，尽可能的兼顾时延敏感小流和吞吐敏感大流。然而现网中的流量场景复杂多变，动态 ECN 门限功能并不能一一覆盖所有流量场景，无法帮助无损业务达到最优性能。而结合了 AI 算法的无损队列的 AI ECN 门限功能可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行队列的精确调度，保障整网的最优性能。

PFC 门限和 ECN 门限

设备通过缓存门限来度量队列的缓存使用情况。为了实现对无损队列的流量控制，减缓无损队列的缓存拥塞，可以为无损队列设置两种缓存门限——PFC 门限和 ECN 门限。PFC 门限是入方向的队列缓存阈值，ECN 门限是出方向的队列缓存阈值。实际上，如果出方向一直不拥塞，入方向是很难拥塞的，报文到达后会被马上转发，所以发生拥塞时，可以通过先触发 ECN 门限通知源端降速，让拥塞缓解，避免过多的触发 PFC。

以图 4-7 为例，介绍 PFC 门限和 ECN 门限的作用。

图4-7 PFC 门限和 ECN 门限减缓拥塞原理图



1. 当 Device 的无损队列出现拥塞，队列已使用的缓存超过 ECN 门限时，Device 在转发报文中打上 ECN 拥塞标记。
2. Server 收到携带 ECN 拥塞标记的报文后，向 Client 发送 CNP 拥塞通知报文。Client 收到 CNP 拥塞通知报文后，降低发包速率。

3. 当 Device 的无损队列拥塞加剧，队列已使用的缓存超过 PFC 反压帧触发门限时，Device 向 Client 发送 PFC 反压通知报文。Client 收到 PFC 反压通知报文后，停止发送对应优先级队列的报文。
4. 当 Device 的无损队列拥塞缓解，队列已使用的缓存低于 PFC 反压帧停止门限时，Device 向 Client 发送 PFC 反压停止报文。Client 收到 PFC 反压停止报文后，继续发送对应优先级队列的报文。

说明

本章节描述的所有“PFC 门限”均指 PFC 反压帧触发门限，即 PFC-XOFF，PFC 反压帧停止门限 PFC-XON 不在本节讨论范围内。取值上，PFC-XON 应该小于 PFC-XOFF，确保已占用的缓存减少（拥塞已缓解）后再停止反压。

由上面的过程可以看出，从 Device 发现队列缓存出现拥塞触发 ECN 标记，到 Client 感知到网络中存在拥塞降低发包速率，是需要一段时间的。在这段时间内，Client 仍然会按照原来的发包速率向 Device 发送流量，从而导致 Device 队列缓存拥塞持续恶化，最终触发 PFC 流控而暂停流量的发送。因此，需要合理设置 ECN 门限，使得 ECN 门限和 PFC 门限之间的缓存空间能够容纳 ECN 拥塞标记之后到 Client 降速之前这段时间发送过来的流量，尽可能的避免触发网络 PFC 流控。

并且网络中同时存在着时延敏感小流和吞吐敏感大流，ECN 门限也需要同时兼顾：

- ECN 门限设置偏高时，可以延缓触发 ECN 拥塞标记，保障流量发送的速率和队列内用来吸收突发流量的缓存空间，满足吞吐敏感的大流的流量带宽。但是，在队列拥塞时，报文在缓存空间内排队，会带来较大的队列时延，对时延敏感的小流无益。
- ECN 门限设置偏低时，可以尽快触发 ECN 拥塞标记，通知 Client 降速，从而使队列内的缓存空间维持在较低的缓存深度，减少报文排队，降低队列时延，对时延敏感的小流有益。但是，过低的 ECN 门限会影响吞吐敏感的大流，限制了大流的流量带宽，无法满足大流的高吞吐。

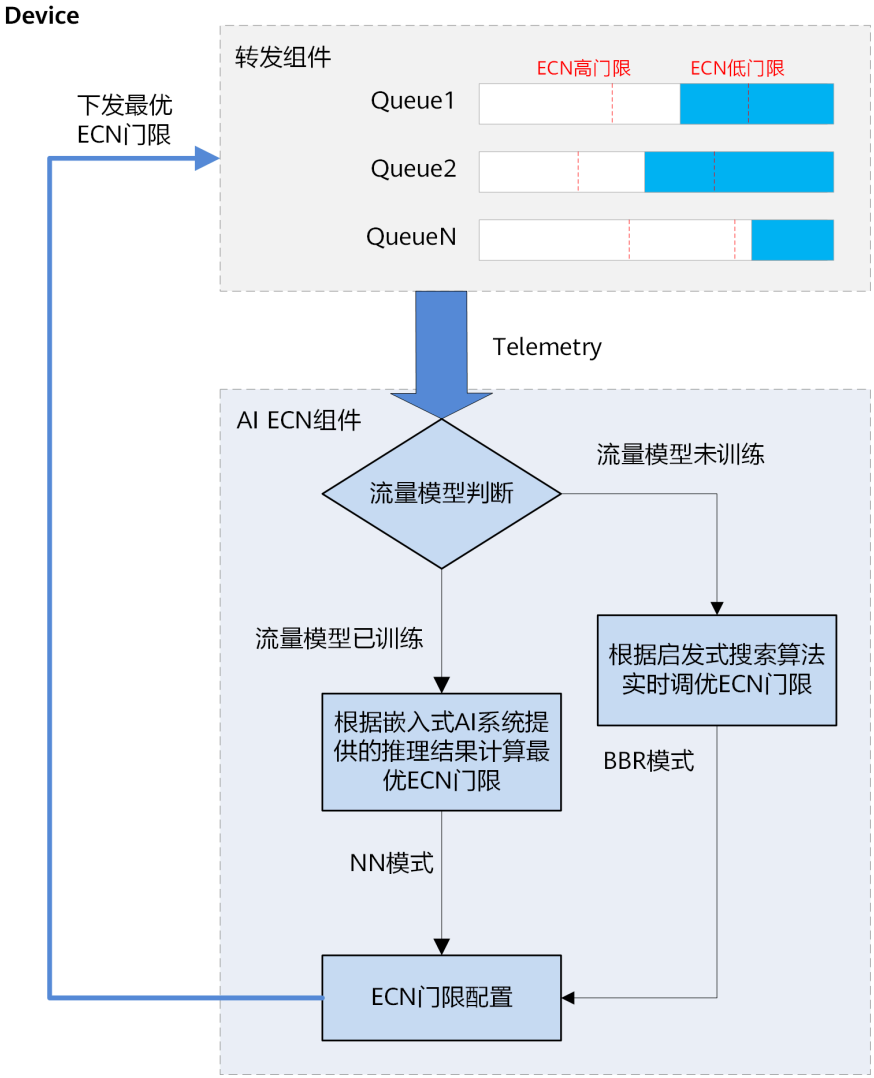
AI ECN 功能

静态 ECN 功能中，其 ECN 门限值是通过手工配置的，对于需要无丢包传输的无损业务，无法使 ECN 门限适应队列中不断变化的缓存空间，在兼顾时延敏感小流和吞吐敏感大流的情况下尽量避免触发 PFC 流控。

AI ECN 可以很好地解决上述问题。无损队列的 AI ECN 功能结合了智能算法，可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行无损队列缓存的精确管控，保障整网的最优性能。

如图 4-8 所示，设备会对现网的流量特征进行采集并上送至 AI ECN 组件，AI ECN 组件将根据嵌入式 AI 系统的推理结果，智能的为无损队列设置最佳的 ECN 门限，保障无损队列的低时延和高吞吐，从而让不同流量场景下的无损业务性能都能达到最佳。

图4-8 无损队列的 AI ECN 功能实现原理



1. Device 设备内的转发组件会对当前流量的特征进行采集，比如队列缓存占用率、带宽吞吐、当前的 ECN 门限配置等，然后通过 Telemetry 技术将网络流量实时状态信息推送给 AI ECN 组件。
2. AI ECN 功能启用后，将自动订阅嵌入式 AI 系统的服务。依据嵌入式 AI 系统，AI ECN 组件收到推送的流量状态信息后，将智能的对当前的流量模型进行判断，识别当前的网络流量场景是否是已知场景。

   如果该流量模型是嵌入式 AI 系统内已训练的模型，则判断当前网络流量场景为已知场景，AI ECN 组件将根据嵌入式 AI 系统推理的最优结果，计算出与当前网络状态匹配的 ECN 门限配置，这种模式称为模型推理模式，这种模式采用 NN（Neural Network）算法，因此也称为 NN 模式。

   如果该流量模型是嵌入式 AI 系统内未训练的模型，则判断当前网络流量为未知场景，AI ECN 组件将结合启发式搜索算法，基于现网状态，在保障高带宽、低时延的前提下，对当前的 ECN 门限不断进行实时修正，最终计算出最优的 ECN 门限配置，这种模式称为启发式推理模式，这种模式采用 BBR（Bottleneck Bandwidth and RTT）算法，因此也称为 BBR 模式。
3. 最后，AI ECN 组件将最优 ECN 门限下发到设备中，调整无损队列的 ECN 门限。
4. 对于获得的新的流量状态，设备将重复进行上述操作，从而保障无损业务的最佳性能。

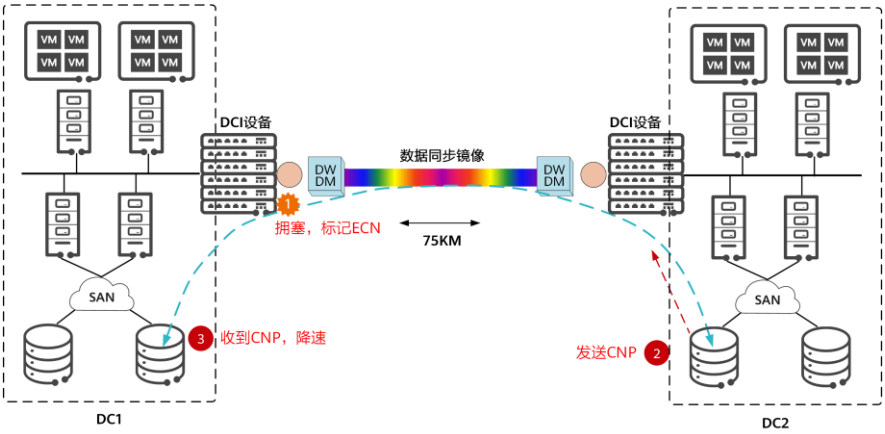
4.4 NPCC

NPCC（Network-based Proactive Congestion Control）是一种以网络设备为核心的主动拥塞控制技术，可以根据设备端口的拥塞状态，准确控制服务器发送 RoCEv2 报文的速率。

DCQCN（Data Center Quantized Congestion Notification）是目前 RoCEv2 网络应用最广泛的拥塞控制算法。传统 DCQCN 提供的拥塞控制机制，是在设备上发现拥塞后，设备会向接收端服务器发送携带拥塞标记的报文，接收端服务器随后向发送端服务器发送 CNP 拥塞通知报文（Congestion Notification Packets，简称 CNP 报文），以通知发送端服务器降低发送报文的速率，从而缓解拥塞。

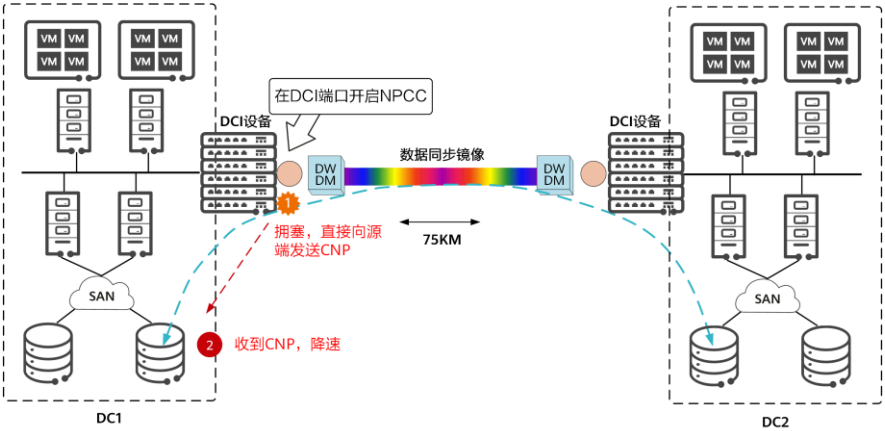
根据 DCQCN 原理，如图 4-9 所示，当 DC1 的 DCI 端口出现拥塞时，交换机对报文进行 ECN 标记，然后需要 DC2 的存储阵列收到标记了 ECN 的报文，然后才能反馈 CNP，并传送回 DC1，这样一去一回，由于同城距离长，消耗的时间就比较长，属于数毫秒级别的，因此无法达到及时降速的效果。

图4-9 传统 DCQCN 问题



因此，智能无损网络又提供了 NPCC 功能（即 NETWORK CC），支持在网络设备上智能识别拥塞状态，如图 4-10 所示，在本 DC 的 DCI 设备近端出现拥塞的情况下，由 DCI 设备直接发送 CNP 给本地的存储阵列，这样可以实现及时降速，避免拥塞加剧，缓解拥塞。

图4-10 开启 NPCC 场景



智能无损网络提供的 NPCC 功能，支持在网络设备上智能识别拥塞状态，主动发送 CNP 报文，准确控制服务器发送 RoCEv2 报文的速率，既可以确保拥塞时的及时降速，又可以避免拥塞已经缓解时的过度降速，最终确保数据中心互联这种长距场景中 RoCEv2 业务的低时延和高吞吐。

4.5 iNOF

iNOF（Intelligent Lossless NVMe Over Fabric，智能无损存储网络）是指通过对接入主机的快速管控，将智能无损网络应用到存储系统，实现计算和存储网络融合的技术。

NVMe over RoCE 在性能、成本、网络管理、技术发展等方面有显著优化，正逐步成为 NVMe over Fabric 的最佳应用。但 NVMe over RoCE 在开局布署&扩容、可靠性、易维护方面面临一些挑战：

1. 开局布署&扩容

   开局布署和扩容时，除了根据网络规划配置 IP、VLAN 等之外，还需要在每一台主机配置与存储的业务关系，配置管理复杂且容易出错。

2. 可靠性

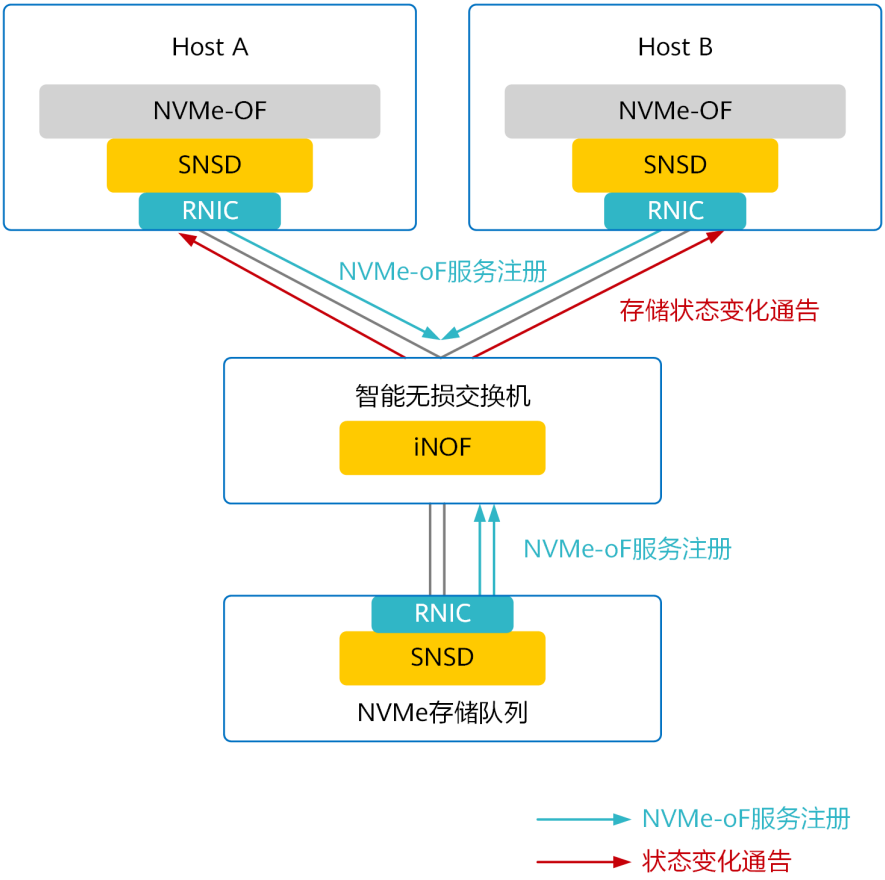
网络故障（如光纤故障、光模块故障、交换机故障、存储设备接口卡故障、存储设备控制器故障）时，NVMe over RoCE 依赖心跳超时（通常大于 5 秒）感知网络故障，然后再将业务切换至冗余路径，IO 归零时间通常在心跳超时时间-2*心跳超时时间（通常是 5-10 秒）之间。如果心跳超时时间更长，则 IO 归零时间更长，对于高可靠和低时延应用（如实时交易系统），这是不可接受的。

3. 易维护

在硬件故障（如光纤故障、光模块故障、接口卡故障、存储控制器故障、交换机故障）后，会对故障件进行更换（通常需要数小时至数天），此时受影响的 NVMe over RoCE 的业务连接已经中断，更换完备件之后，需要在每台主机操作恢复业务。

为了进一步提升在易用性、可靠性和易维护性方面的竞争力，为了让智能无损网络技术更好的服务于存储系统，华为提出了 iNOF 技术，通过对接入主机的快速管控，可以第一时间获知新接入的主机，智能的调整智能无损网络的相关配置，并且 iNOF 技术支持将主机信息和存储系统相互通告，可以协助存储系统管理主机。

图4-11 iNOF 与 SNSD 联动



通过网络设备的 iNOF 功能与主机/存储阵列的 SNSD 功能协同联动，可以实现以下功能：

1. 存储设备自动发现

开局和扩容部署：根据业务规划，在交换机上把 IP 业务域按照规划完成配置，主机和存储开启 SNSD 功能，主机与存储的业务连接即可自动建立业务连接。

故障维护场景：典型故障维护场景，例如更换光前、光模块、主机的 RoCE 卡、存储 RoCE 卡或者存储控制器单元等。在完成备件更换之后，主机与存储的业务连接自动恢复，不需要做额外的操作。

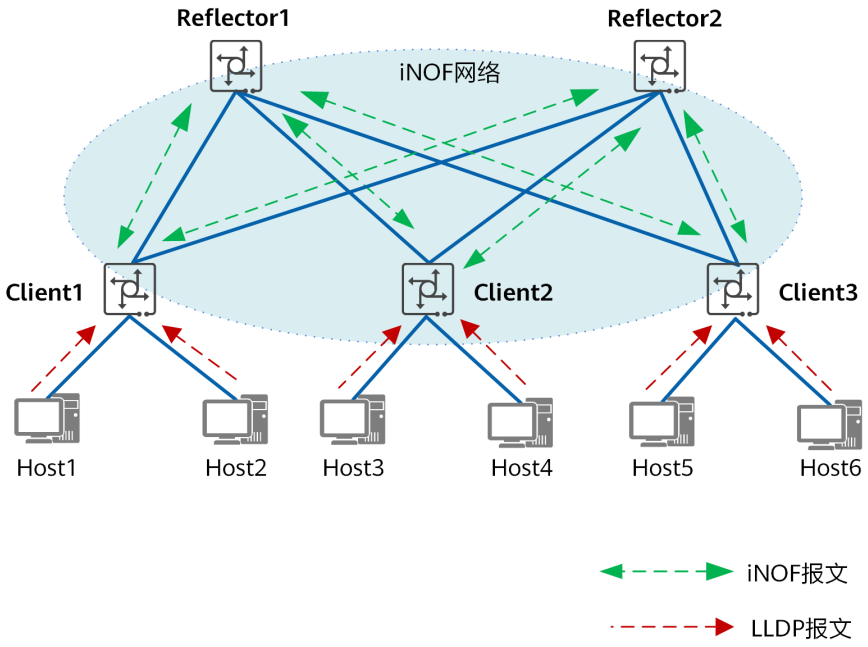
2. 链路故障自动快速切换 IO 路径

   故障快速感知的目的是在网络故障发生后，业务可以快速切换到冗余路径，降低对业务的影响。

iNOF 原理

如图 4-12 所示，在 iNOF 网络中，其中一台或两台设备可以作为反射器（Reflector），用于同步域信息，其余设备作为客户端（Client）。每个 iNOF 反射器和 iNOF 客户端之间都需要建立 iNOF 连接，从而可以传输 iNOF 报文。iNOF 客户端之间不需要建立 iNOF 连接，只需要与主机（Host）直连。iNOF 反射器之间不需要建立 iNOF 连接，两者互为备份。与主机直连的接入设备也可以作为 iNOF 反射器。建议首选 Spine 作为 iNOF 反射器，如若 Spine 节点不支持 iNOF 特性，需要选用 Leaf 作为 iNOF 反射器。

图4-12 iNOF 原理图



iNOF 报文是 TCP 封装的报文，包含 iNOF 关键信息的内容承载在 TCP 报文的 Data 字段内。客户端可以通过 iNOF 报文将 iNOF 关键信息发送给反射器，反射器汇总后再发往其他客户端。通过 iNOF 报文可以传输以下几类信息：

- 建连信息：iNOF 反射器和客户端之间需要通过互相交换 iNOF 报文来建立 iNOF 连接，具体的建立过程类似 TCP 建连。
- 域配置信息：iNOF 系统中，设备可以通过域（Zone）对接入的主机进行管理，iNOF 反射器上完成 iNOF 域的相关配置后，会通过 iNOF 报文把域配置信息发往各个客户端。
- 主机动态信息：主机的网卡和 iNOF 设备均需要启用 LLDP（Link Layer Discovery Protocol）功能，当有新的主机接入客户端或者离开客户端时，主机会主动向客户端发送 LLDP 报文，报文内记录了 LLDP 邻居信息的变化。客户端

可以通过 iNOF 报文将这些主机动态信息发送给反射器，反射器汇总后再发往其他客户端，使 iNOF 系统内的其他设备感知到主机动态信息。

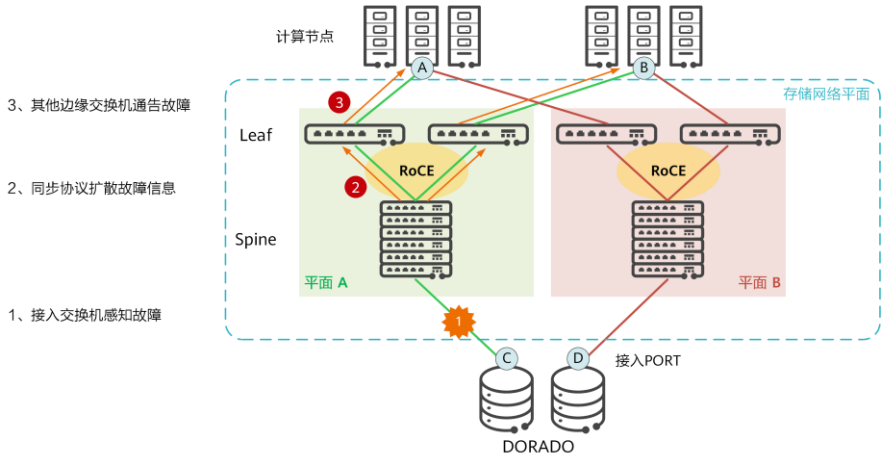
- 接口 Error-Down 信息：当 iNOF 设备因为 PFC 死锁、CRC 错误报文达到告警阈值等问题触发接口 Error-Down 后，iNOF 报文内会携带接口 Error-Down 信息，让 iNOF 系统内的其他设备迅速感知，及时调整路径信息。

iNOF 系统建立后，系统内的所有设备都可以第一时间感知到接入主机的变化，从而可以将信息反馈给各个智能无损网络功能去智能的调整相关配置，最终使网络达到低时延、无丢包和高吞吐的性能。

故障切换

故障场景下的流量切换如图 4-13 所示。故障主要包括用户侧故障、网络侧故障两类。

图4-13 故障切换示意图



用户侧故障

用户侧故障包括主机或者存储阵列节点故障、主机或者存储阵列与 TOR 接入交换机之间链路故障。

1. 故障点 1：与阵列直连端口故障

   当阵列直连端口故障时，主机侧并不一定能感知到端口故障（只有主机与阵列端口直连的场景才能感知），网络需要将端口故障信息通告给域内所有主机端口，主机端口根据策略进行多路径切换。

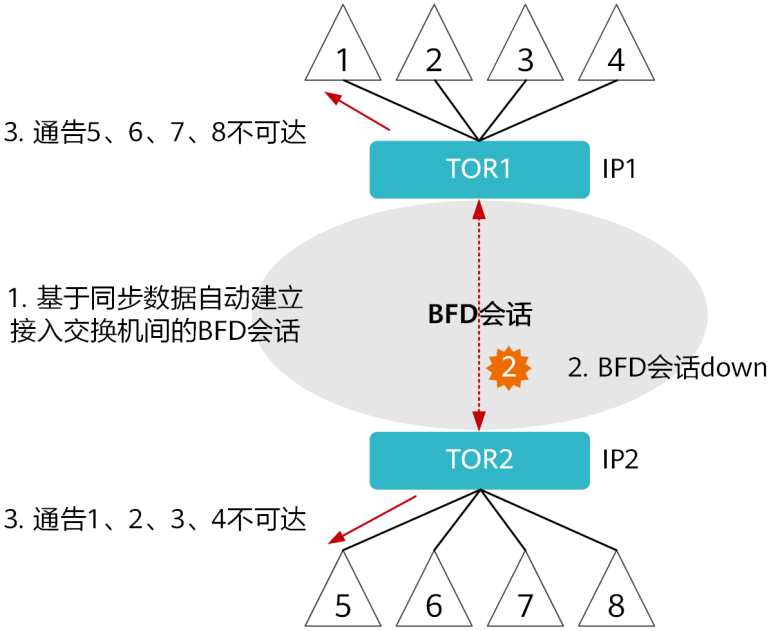
2. 故障点 2：与主机直连端口故障

   从存储业务角度上看，主机直连端口物理故障时，交换机与主机均能感知，主机感知到故障后即进行路径切换，实质上不需要网络再进行一次故障通告。

网络侧故障

网络侧故障包括网络节点故障和网络间链路故障。如图 4-14 所示，在 TOR2 出现故障，或者 TOR1 到 TOR2 的路径不可达时，由于 TOR1 无法快速感知 TOR2 的故障，导致前端的编号 1-4 的 HOST 继续往 5-8 的存储阵列节点写数据或者读数据，或导致 IO 不可用，HOST 的 IO 会出现归零的情况，这个时候如果无法快速完成路径切换，将大大影响 HOST 主机的读写性能，需要 HOST 主机自己感知故障然后完成路径切换，预计需要 5-10s 的时间。针对这类故障，我们可以通过双向转发检测 BFD（Bidirectional Forwarding Detection）来进行路径检测。

图4-14 网络侧故障



在 TOR 之间建立 BFD 会话，用于快速检测设备之间的通信故障，并在出现故障时通知上层应用。BFD for iNOF 就是将 BFD 和 iNOF 协议关联起来，将 BFD 对链路故障的快速感应通知 iNOF 协议，需要每两个接入设备之间两两都要建立 BFD 的会话。

BFD for iNOF 功能可以实现对 iNOF 系统中的链路故障进行快速检测，从而可以及时将链路的故障信息通告给同一个 iNOF 域中的主机，及时进行链路切换。

第5章
NoF+存储网络成功应用

华为助力某银行客户构建业界首个 RoCE SAN 网络，率先完成存储网络代际变革。

客户痛点

某银行客户在迈向智能时代，打造科技创新引领能力，全面推动数字化转型的过程中，现有的 FC 存储网络面临挑战：

- 性能瓶颈：FC 带宽低，最大带宽仅 32G；多 DC 间 100+链路，部署复杂
- 运维复杂：原有 FC SAN 网络封闭架构，运维依赖原厂，运维成本高、难度大
- 投资风险：客户每年 FC SAN 网络投资 X 千万，可能存在供应、商务风险

全闪存时代，客户希望有新的选择来新建存储网络。

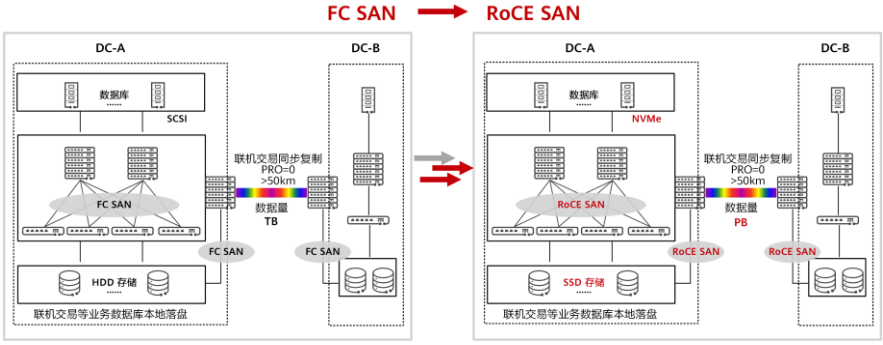
基于 NVMe over RoCE 的 NoF+存储网络

某银行客户选择了 RoCE 作为下一代存储网络，与华为联合创新的 NoF+存储网络解决方案，给银行联机交易等生产系统带来了三大升级：

- RoCE，25GE/100GE 接入，100GE/400GE 级联；日均复制量 PB 级

- 开放架构，基于标准 API 与第三方伙伴共建场景化服务
- LAN/SAN 合一，可实现统一、自动化运维

图5-1 构建业界首个 RoCE SAN 网络



第6章
NoF+存储网络未来展望

基于华为 NoF+存储网络和全闪存数据中心的解决方案，无论是在金融行业，还是在运营商、大企业，正在进行越来越多的商用部署。华为也希望联合业界的各种厂家以及合作伙伴，共同繁荣 NoF+的产业生态，能够向 NoF 的标准组织注入更多的活力，共同推动 NoF+的产业发展。

目前，华为 NoF+是业界唯一集合零丢包的以太网，大带宽、低时延、易维护，开放兼容，利于存储、网络、计算三种资源融会贯通，实现数据的实时共享。并且华为 NoF+方案已经同部分大型金融、运营商企业都进行了联合创新实践，为数据基础设施领域建设打开新格局。

面向未来，华为将继续秉承以客户为中心的发展理念，携手合作伙伴打造更多结合行业场景的存储网络优秀实践，为加速企业数字化转型做出积极贡献。

IP网络
系 列 丛 书

联系我们

networkinfo@huawei.com

获取更多 IP 网络系列丛书

https://e.huawei.com/cn/solutions/enterprise-networks/ip-ebook