ICS 35.110
M11

YD

中 华 人 民 共 和 国 通 信 行 业 标 准

YD/T XXXXX—XXXX

多路径远程直接内存访问（RDMA）传输协议技术要求

Multipath Remote Direct Memory Access (RDMA) Transport Protocol Technical Requirements

（草案稿）

XXXX - XX - XX 发布　　XXXX - XX - XX 实施

中华人民共和国工业和信息化部　　发 布

目　　次

前　言

本文件按照GB/T1.1-2020《标准化工作导则 第1部分：标准化公文的结构和起草规则》给出的规则起草。

注意本文件的某些内容可能涉及专利，本文件的发布机构不承担识别专利的责任。

本文件由中国通信标准化协会提出并归口。

本文件起草单位：北京邮电大学、中国移动通信有限公司研究院、中国电信集团有限公司、中国联合网络通信集团有限公司、中国信息通信研究院、山东省计算中心（国家超级计算济南中心）、北京交通大学、华为技术有限公司、山东浪潮科学研究院有限公司、鹏城实验室、之江实验室、中关村实验室、泉城实验室、西安交通大学、中国标准化研究院、浪潮电子信息产业股份有限公司、中国互联网络信息中心、迈普通信技术股份有限公司。

本文件主要起草人：张乙然、刘佳雪、任丰源、苏伟、郜帅、陈佳、范大卫、王莫为、林茂、王久霜、刘礼斌、谭立状。

多路径远程直接内存访问（RDMA）传输协议技术要求

1 范围

本文件规定了多路径RDMA传输协议的总体架构、报文格式、协议流程、功能要求、性能要求及接口规范。

本文件适用于支持RoCEv2协议的网卡（RNIC）、交换机及网络管理系统，用于规范多路径环境下的协议扩展、连接管理、流量调度、乱序处理与端网协同的技术实现。本文件所述协议扩展宜适用于iWARP协议。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

YD/T 4465-2023 无损网络总体技术要求

YD/T 4027-2022 基于RoCE协议的数据中心高速以太无损网络技术要求

YD/T 3902-2021 数据中心无损网络典型场景技术要求和测试方法

YD/T 4273-2023 无损网络应用场景与需求

YD/T 4627-2023 数据中心网络智能管控及运维系统技术要求

IETF RFC 5040 A Remote Direct Memory Access Protocol Specification

IETF RFC 7306 Remote Direct Memory Access (RDMA) Protocol Extensions

IETF RFC 3168 The Addition of Explicit Congestion Notification (ECN) to IP

IETF RFC 8200 Internet Protocol, Version 6 (IPv6) Specification

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件：

3.1.1 远程直接内存访问 Remote Direct Memory Access

一种在较少主机内核干预条件下，由网络接口直接完成本地与远端内存间数据访问的高速网络通信技术 。

3.1.2 多路径传输 Multi-Path Transmission

一种在源端节点与目的端节点之间，同时利用多条物理或逻辑网络路径进行数据转发的传输机制，用于提升端到端网络带宽利用率与传输可靠性。

3.1.3 路径熵 Path Entropy

数据报文头部中用于影响网络设备等价多路径转发决策的特征字段值，如UDP源端口号、IPv6流标签等。通过改变路径熵值，可使同一数据流的不同报文被转发至不同的网络路径。

3.1.4 逐包散射 Packet Spraying

一种以数据包为单位的流量调度策略，指发送端将属于同一数据流的连续数据包，依次或按特定算法分散到不同的网络路径上进行传输，以提升链路间的负载均衡程度。

3.1.5 流块 Flowlet

指同一个数据流中，时间间隔超过预设阈值（通常大于网络多条路径间的最大时延差）的一组连续数据包序列。以流块为粒度进行路径切换，可降低接收端的报文乱序概率。

3.1.6 自适应路由 Adaptive Routing

一种由网络设备（如交换机）主导的动态路由选择机制。网络设备实时感知各出端口的队列深度、拥塞状态或链路利用率，动态为数据包或流块选择当前负载较轻的转发路径，而非依赖静态哈希结果。

3.1.7 乱序重排 Out-of-Order Reordering

接收端处理多路径传输引发的乱序到达报文的过程。通过在专用缓冲区内暂存乱序报文，并依据包序列号（PSN）将其恢复为原始发送顺序后，再交付给上层应用。

3.2 缩略语及符号

下列缩略语及符号适用于本文件：

RDMA	远程直接内存访问	Remote Direct Memory Access
RNIC	RDMA 网络接口卡	RDMA Network Interface Card
AR	自适应路由	Adaptive Routing
PSN	包序列号	Packet Sequence Number
RTT	往返时延	Round Trip Time
BDP	延迟带宽积	Bandwidth-Delay Product
ECMP	等价多路径	Equal-Cost Multi-Path
QP	队列对	Queue Pair
PFC	优先级流量控制	Priority Flow Control

4 概述

4.1 技术背景

RDMA技术通过内核旁路与网卡协议卸载，实现高吞吐、低时延的网络数据传输，已广泛应用于数据中心内部及广域网络。

传统基于RoCEv2的RDMA网络通常依赖等价多路径(ECMP)等静态哈希机制进行流量调度。静态哈希根据报文头五元组计算转发路径，同一条数据流的所有报文始终沿同一路径传输。当网络中存在大容量、长生命周期的数据流（如AI训练中的数据并行或流水线并行通信流）时，多条大流可能被哈希映射至同一条物理链路，导致该链路拥塞，而其余等价链路处于空闲状态，即哈希极化现象。此外，多条流的发送窗口在时间上重叠还会引发微突发拥塞，进一步加剧排队时延。

面对上述由流量分布不均导致的结构性拥塞，传统拥塞控制算法通常采取源端降速或触发基于优先级的流量控制（PFC）暂停机制来缓解。然而，源端降速会直接降低吞吐率；PFC暂停帧则可能沿反压路径向上游扩散，造成无关流量被连带阻塞。

RDMA多路径传输协议针对上述问题，在现有RoCEv2协议基础上进行扩展，通过定义多路径扩展头、子路径管理机制和协议交互流程，使单个QP连接能够利用多条物理路径并发传输。其核心目标包括：降低单路径带宽瓶颈对吞吐的制约，减少因流量集中引起的局部拥塞，以及提升网络整体的链路带宽利用率。

4.2 场景描述

RDMA多路径传输协议主要面向两类典型应用场景：数据中心内场景以及广域网下跨数据中心场景。两类场景的底层网络在拓扑结构、链路时延、路径对称性等方面存在差异，对应的多路径调度机制与调度粒度也有所不同。

4.2.1 数据中心内场景

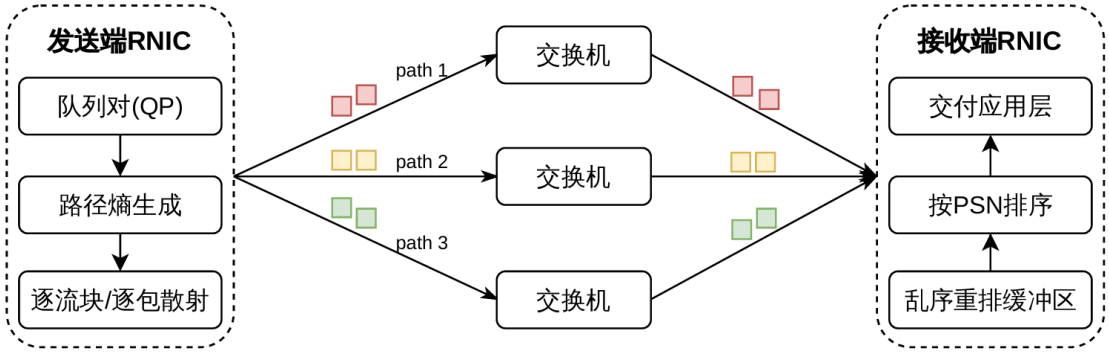
在数据中心内场景中，网络通常采用多级Clos架构（如Leaf-Spine组网），任意两台服务器之间存在多条等价的高带宽物理路径，往返时延普遍在微秒级别，各等价路径之间的时延差较小。

该场景下，ECMP等静态哈希机制在承载大量长生命周期数据流时容易引发链路负载不均与微突发拥塞。为此，多路径传输协议通常采用细粒度调度策略，主要包括两种方式：

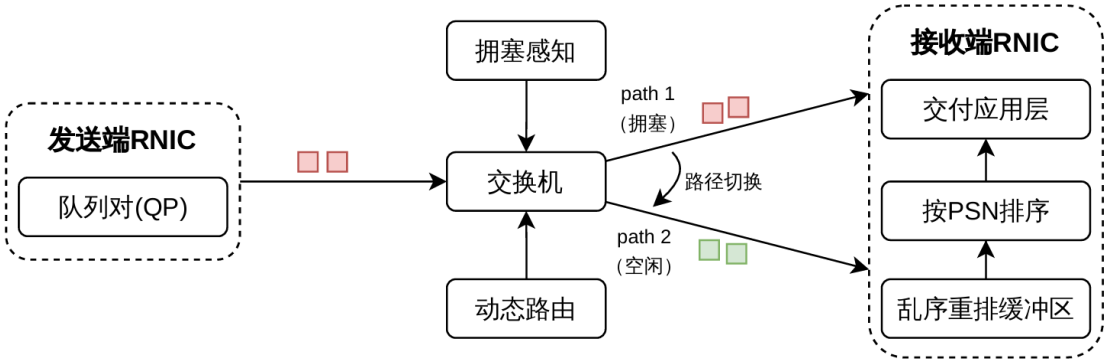
a） 端侧驱动的流量散射：发送端RNIC通过修改报文头部的路径熵字段（如UDP源端口号、IPv6流标签），将同一QP内的报文以更细的粒度（如流块、数据包等）分散至不同的子路径。

b） 网侧驱动的自适应路由：交换机实时感知各出端口的队列深度与拥塞状态，动态为报文或流块选择负载较轻的转发路径。

通过上述机制，单个QP连接可利用多条等价链路的聚合带宽进行并发传输。由于不同子路径的传播时延存在微小差异，接收端可能收到乱序报文，需由接收端RNIC的乱序重排缓冲区根据PSN将报文恢复为有序状态后交付上层应用。



（a） 端侧驱动的流量散射



（b） 网侧驱动的自适应路由

图1 RDMA多路径传输的主要两种方式

4.2.2 广域网下跨数据中心场景

在广域网与跨数据中心场景下，网络链路跨越较长物理距离，具有高带宽时延积特征。不同路径可能经过不同的中间网络或运营商，各路径之间的时延差异通常达到毫秒级别，且路径质量波动较大。

在此条件下，若采用数据包级或流块级的细粒度调度，不同子路径之间的时延差会导致接收端出现大幅度报文乱序，增加重排缓冲区溢出和触发大范围重传的风险。因此，广域网场景下的多路径传输协议通常采用逐流的粗粒度调度策略。系统根据各广域链路的实时质量（如可用带宽、丢包率、时延）与业务需求，将完整的数据流映射至不同的子路径进行转发。单条数据流的所有报文在同一子路径内按序传输，避免跨路径乱序。当某条子路径出现质量劣化或故障时，系统将该子路径上的数据流迁移至其他可用子路径，实现链路级的负载分担与冗余保护。

5 协议总体架构

5.1 与传统单路径 RDMA 协议的关系

本文件所述多路径传输协议是在现有RoCEv2协议基础上的扩展，不替代或修改RoCEv2的基础传输机制。

多路径扩展通过在RoCEv2报文中引入可选的多路径扩展头（MPEH），在单个QP连接内建立多条子路径，实现报文的多路径并发传输。当通信双方未协商启用多路径能力时，协议行为应与标准RoCEv2完全一致，确保向后兼容。

5.2 协议栈位置

多路径传输协议的扩展位于RoCEv2协议栈的传输层，介于基础传输头（BTH）处理和上层应用语义之间。其协议栈位置如下：

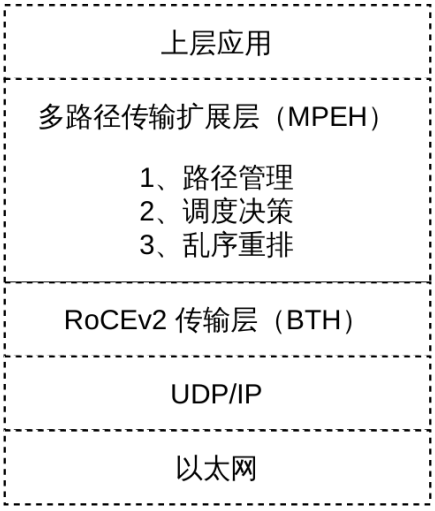


图2 多路径传输协议位置

多路径扩展层接收上层应用的发送请求，根据调度策略选择子路径、填充多路径扩展头字段，再交由RoCEv2传输层进行标准的BTH封装和发送。在接收端，多路径扩展层从BTH后解析MPEH，执行乱序重排后，再将有序数据交付上层应用。

5.3 多路径能力模型

一个启用多路径的QP连接包含以下逻辑结构：

- 一个主连接：对应传统RoCEv2的QP连接，维护全局PSN空间和连接状态。

- N条子路径：每条子路径拥有独立的路径标识、路径熵值、路径质量状态和拥塞控制状态。子路径数量通过建链阶段协商确定。
- 一个调度器：根据配置的调度粒度（逐流/逐流块/逐包）和各子路径的实时状态，决定每个报文或流块的转发子路径。
- 一个重排组件：在接收端维护基于PSN的乱序重排缓冲区，将跨子路径到达的乱序报文恢复为有序状态。

6 协议报文格式

6.1 报文格式总览

启用多路径传输后，RoCEv2数据报文的封装格式在BTH之后增加一个可选的多路径扩展头（MPEH）。完整的报文结构如下：

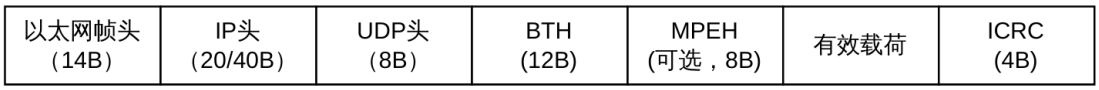


图3 报文格式总览

当通信双方在建链阶段未协商启用多路径能力时，报文中不包含MPEH字段，格式与标准RoCEv2报文完全一致。

MPEH的存在与否通过BTH中的OpCode扩展或保留位进行标识。接收端根据该标识判断是否需要解析MPEH。

6.2 多路径扩展头（MPEH）格式

MPEH为固定长度8字节，格式定义如下：

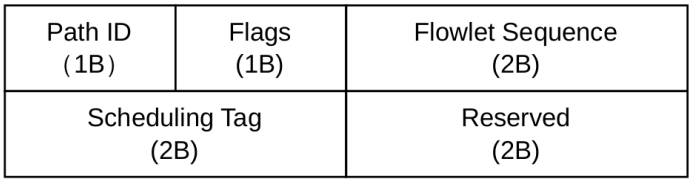


图3 多路径扩展头（MPEH）格式

各字段说明：

- Path ID： 子路径标识。标识该报文被调度至的子路径编号，取值范围为0至N-1（N为协商确定的子路径数量）。接收端可根据此字段区分不同子路径到达的报文，用于路径级统计和乱序分析。
- Flags： 多路径控制标志位，各位定义如下：

位	名称	说明
0	F(Flowlet Boundary)	置1表示当前报文为流块的首包，接收端可据此辅助乱序判断
1	S (Sub-path Switch)	置1表示发送端已执行子路径切换，接收端应更新路径状态
2	P (Path Probe)	置1表示该报文为路径探测报文，接收端应在ACK中返回路径质量信息
3	C (Congestion Signal)	置1表示发送端已感知该子路径拥塞，通知接收端参考

4-7	Reserved	保留，应置0

- Flowlet Sequence： 流块序列号。在逐流块调度模式下，标识当前报文所属的流块编号。同一流块内的报文应具有相同的Flowlet Sequence值。在逐包调度或逐流调度模式下，该字段可置0。
- Scheduling Tag： 调度标签。由发送端调度器填充，用于承载调度相关的辅助信息。在端网协同场景下，网侧设备可读取该字段辅助转发决策。具体编码方式由实现方定义，应在产品文档中公开说明。
- Reserved： 保留字段，应置0。

6.3 多路径 ACK 报文格式

在多路径传输模式下，接收端返回的ACK报文需要携带子路径相关的反馈信息。ACK报文同样在BTH之后包含MPEH，其中各字段的使用方式与数据报文有所不同：

- Path ID： 填写接收端接收该ACK所确认数据的子路径标识。发送端可据此了解各子路径的确认进度。
- Flags：

位	名称	说明
0	R (Reorder Event)	置 1 表示接收端检测到跨子路径乱序事件
1	L (Loss Suspected)	置 1 表示接收端检测到疑似丢包（乱序深度超过容忍阈值）
2	P (Path Probe Response)	置 1 表示该 ACK 包含路径探测响应信息
3-7	Reserved	保留，应置 0

- Flowlet Sequence： 在ACK报文中可用于携带接收端当前已连续确认的最大流块序列号。
- Scheduling Tag： 在ACK报文中可用于携带接收端测量的路径质量信息（如单向时延估计值），具体编码方式由实现方定义。

6.4 路径熵字段的使用

除MPEH外，多路径传输协议还利用现有报文头部字段承载路径熵信息：

发送端应支持通过修改UDP源端口号生成路径熵，驱动网络中间设备的ECMP转发机制将不同报文分发至不同物理路径。

发送端宜支持通过修改IPv6流标签生成路径熵。

同一子路径内的所有报文应使用相同的路径熵值，确保其经过相同的物理路径。当子路径切换时，新子路径应使用不同的路径熵值。

路径熵值与Path ID之间应建立确定性的映射关系。

7 协议流程

7.1 建链阶段

建链阶段在传统RoCEv2的QP连接建立过程中增加多路径能力协商步骤。

7.1.1 多路径能力协商

在QP连接建立过程中，发送端与接收端应通过扩展的连接管理报文交换各自的多路径能力参数。协商参数应至少包括：

a）多路径支持标识：标识本端是否支持多路径传输扩展。

b）最大子路径数量：本端支持的最大并发子路径数。

c）调度粒度能力：本端支持的调度粒度类型（逐流、逐流块、逐包）。

d）乱序容忍深度：本端接收侧支持的最大乱序容忍PSN跨度。

e）重排缓冲区容量：本端接收侧可用于乱序重排的缓冲区大小。

当双方均支持多路径传输时，连接进入多路径模式，协商结果取双方能力的交集（如子路径数量取较小值）。当任一端不支持多路径传输时，连接应退化为标准RoCEv2单路径模式，后续报文中不应包含MPEH。

7.1.2 子路径建立

多路径能力协商完成后，发送端应为该QP连接建立协商确定数量的子路径。子路径建立过程应完成以下操作：

为每条子路径分配唯一的Path ID（从0开始连续编号）。

为每条子路径生成对应的路径熵值（如不同的UDP源端口号），并建立Path ID与路径熵值的映射关系。

对各子路径执行初始路径质量探测，获取各子路径的RTT和可用带宽等基线数据。

根据探测结果为各子路径设置初始的流量分配权重。

接收端在收到首个包含MPEH的数据报文后，应根据Path ID建立对应的子路径接收状态。

7.1.3 参数同步

建链阶段结束前，通信双方应就以下运行参数达成一致：

a）生效的子路径数量及各子路径的Path ID。

b）调度粒度（逐流、逐流块或逐包）。

c）乱序容忍阈值（以PSN跨度或时间差表示）。

d）流块超时阈值（在逐流块调度模式下，用于判断流块边界的时间间隔）。

上述参数应支持在数据传输阶段通过管理面或控制面动态修改。

7.2 数据传输阶段

数据传输阶段描述多路径模式下报文的发送、转发和接收处理流程。

7.2.1 发送端处理流程

发送端RNIC在收到上层应用的发送请求后，按以下步骤处理：

步骤1——调度决策：调度器根据配置的调度粒度和各子路径的实时状态，为当前报文或流块选择目标子路径。调度依据包括各子路径的当前负载、RTT、拥塞状态和分配权重。

步骤2——PSN分配：在全局PSN空间内为当前报文分配PSN。PSN应保持全局单调递增，不因子路径分散而产生跳跃或重复。

步骤3——MPEH填充：填充多路径扩展头各字段，包括目标子路径的Path ID、当前流块的Flowlet Sequence、调度标签以及相应的标志位。

步骤4——路径熵设置：根据目标子路径的Path ID查找对应的路径熵值，填入UDP源端口号或IPv6流标签。

步骤5——BTH封装与发送：按照标准RoCEv2流程完成BTH封装，将报文发送至网络。

7.2.2 网络转发处理

网络中间设备（交换机）对包含MPEH的报文的处理分为两种模式：

透明转发模式：交换机不解析MPEH，仅根据报文外层头部（IP头、UDP头）的ECMP哈希进行转发。报文的路径由发送端设置的路径熵值决定。该模式不需要交换机支持多路径协议扩展。

自适应路由模式：交换机解析MPEH中的Scheduling Tag和Flags字段，结合本地拥塞感知信息进行转发决策。该模式下交换机宜支持以下能力：

a)读取Scheduling Tag辅助调度判断。

b)在发生子路径切换时（Flags.S=1），更新内部转发状态。

c)将拥塞信息写入报文的ECN字段或INT遥测字段，向端侧反馈路径状态。

7.2.3 接收端处理流程

接收端RNIC在收到包含MPEH的数据报文后，按以下步骤处理：

步骤1——MPEH解析： 解析多路径扩展头，提取Path ID、Flags和Flowlet Sequence等字段。

步骤2——子路径状态更新：根据Path ID更新对应子路径的接收统计信息，包括接收报文计数、最近接收时间戳等。

步骤3——乱序检测与缓冲：将当前报文的PSN与期望的下一个连续PSN进行比较。若PSN连续，直接进入步骤4；若PSN不连续，将报文暂存至乱序重排缓冲区。

步骤4——乱序判定：对于缓冲区中的缺口报文，根据乱序容忍阈值判断：若缺口持续时间或PSN跨度未超过阈值，保持等待，不触发丢包反馈；若超过阈值，将缺口升级为疑似丢包事件，生成携带Flags.L=1的ACK报文通知发送端。

步骤5——按序交付：当缓冲区中的报文形成连续PSN序列时，按序提交给上层应用。

步骤6——ACK生成：生成ACK报文，填充MPEH反馈字段（确认进度、乱序事件、路径质量信息等），返回给发送端。

7.2.4 子路径动态管理

在数据传输过程中，系统应支持对子路径进行动态管理：

a)新增子路径：当网络拓扑变化或新的物理路径可用时，发送端宜支持在不中断现有传输的情况下新增子路径。新增子路径应完成路径质量探测后方可承载业务流量。

b)移除子路径：当某条子路径持续劣化或物理链路失效时，发送端应将该子路径上的后续报文迁移至其他可用子路径，待该子路径上所有在途报文完成确认后，释放相关资源。

c)权重调整：发送端应支持根据各子路径的实时质量数据动态调整流量分配权重。拥塞或时延升高的子路径应降低权重，质量良好的子路径应增加权重。

d)路径切换通知：发送端在执行子路径切换时，应在切换后的首个报文中将MPEH的Flags.S位置1，通知接收端更新路径状态。

7.2.5 拥塞控制协同

多路径模式下，拥塞控制机制应考虑子路径的独立性：

各子路径宜维护独立的拥塞控制状态（如独立的发送速率或拥塞窗口）。某条子路径的拥塞不应导致其他正常子路径的发送速率降低。

当接收端通过ECN或ACK反馈某条子路径的拥塞信号时，发送端应仅对该子路径执行拥塞响应（如降低该子路径的发送速率或权重），不应对其他子路径产生连带影响。

发送端宜支持将拥塞子路径的流量主动迁移至空闲子路径，而非仅依靠降速应对拥塞。

7.3 连接拆除阶段

连接拆除阶段负责安全关闭多路径QP连接，确保数据完整性和资源正确释放。

7.3.1 正常拆除流程

当上层应用请求关闭QP连接时，多路径连接的拆除应按以下顺序进行：

步骤1——停止新请求：发送端停止接受上层应用的新发送请求。

步骤2——排空在途报文：发送端等待所有子路径上的在途报文完成传输。接收端对所有已接收报文返回ACK确认。

步骤3——PSN一致性确认：发送端与接收端确认全局PSN空间内的所有报文已完成交付，重排缓冲区中无残留报文。

步骤4——子路径拆除：依次释放各子路径的资源，包括路径状态表项、路径熵映射和路径级拥塞控制状态。

步骤5——主连接关闭：按照标准RoCEv2的QP连接关闭流程完成主连接的拆除。

7.3.2 异常拆除处理

当某条子路径在传输过程中发生不可恢复故障时，系统应按以下方式处理：

发送端应立即停止向故障子路径发送新报文，将后续流量迁移至其他可用子路径。

对于故障子路径上的在途报文，发送端应在超时后在其他子路径上执行重传。

故障子路径的资源应在确认无残留状态后释放。

单条子路径的故障不应导致整个QP连接关闭。仅当所有子路径均不可用时，连接才应进入异常关闭流程。

8 功能技术要求

8.1 端侧功能要求

8.1.1 发送端要求

发送端RNIC应支持在建链阶段进行多路径能力协商，并根据协商结果启用或关闭多路径传输模式。

发送端应支持在全局PSN空间内为多条子路径上的报文分配连续递增的PSN。

发送端应支持至少逐流和逐流块两种调度粒度，宜支持逐包调度。

发送端宜支持根据各子路径的实时质量数据动态调整流量分配权重。

发送端应支持路径熵生成规则的配置与修改，包括路径熵生成算法、散射粒度和权重参数等。

8.1.2 接收端要求

接收端RNIC应支持解析MPEH，并根据Path ID维护子路径级的接收状态。

接收端应支持基于PSN的乱序重排，将跨子路径到达的乱序报文恢复为有序状态后交付上层应用。

接收端应支持区分乱序与丢包。在乱序深度未超过协商的容忍阈值前，应抑制丢包反馈。

接收端应支持在ACK报文的MPEH中携带子路径级的反馈信息。

接收端宜支持对乱序事件进行统计，包括乱序报文数量、最大乱序深度、乱序持续时间等。

8.2 网侧功能要求

网络设备应支持对包含MPEH的报文进行透明转发。

网络设备宜支持解析MPEH中的Scheduling Tag，结合本地拥塞感知信息进行自适应转发。

网络设备应支持对各出端口的拥塞状态进行评估，评估指标应包含出端口队列深度、端口带宽利用率及链路物理状态。

当网络启用PFC时，网络设备应支持将端口受PFC暂停的时间纳入拥塞评估。

网络设备宜支持通过ECN或INT向端侧传递拥塞信息。

8.3 端网协同要求

系统应支持端侧设备与网侧设备之间的调度状态互通，避免端网叠加调度引发过度乱序。

对于同一数据流，端侧与网侧不宜同时执行逐包级的细粒度调度。当一侧采用细粒度调度时，另一侧宜采用较粗粒度或关闭主动调度。

网侧设备在感知到特定子路径拥塞时，应支持通过ECN或INT标记报文向端侧传递拥塞信息。端侧在收到反馈后应调整对应子路径的流量分配。

宜支持应用层向网络层传递业务特征参数（如流块大小、通信模式、优先级等），网侧设备宜根据业务参数调整调度策略。

9 性能技术要求

9.1 传输性能要求

多路径传输系统的性能评估应基于以下指标。实现方应在产品文档中对各项指标的能力边界进行公开声明。

在多路径传输场景下，系统的有效吞吐量应高于同等条件下仅使用单路径传输时的有效吞吐量。

当多条子路径均处于空闲状态时，系统的聚合有效吞吐量宜随可用子路径数量近似线性增长。实现方应声明在典型组网条件下的聚合吞吐效率。

多路径调度不应导致短消息流的流完成时间相比单路径传输出现明显恶化。

多路径传输系统的尾部时延（P99/P99.9）应优于同等网络负载条件下仅使用ECMP静态哈希调度时的尾部时延。

9.2 乱序重排性能要求

乱序容忍深度应能覆盖当前网络中各子路径之间的最大时延差。实现方应声明设备支持的最大乱序容忍深度，并说明该值是否支持动态调整。

实现方应声明单QP的重排缓冲区容量上限以及单设备支持的总重排缓存容量。

重排附加时延宜与网络各子路径之间的时延差保持同数量级，不应因重排处理引入显著超出子路径时延差的额外延迟。

9.3 路径切换与响应性能要求

拥塞感知时延宜与网络RTT保持同数量级。

路径切换响应时间应小于业务层的超时重传阈值，避免因切换耗时过长导致上层应用触发不必要的重传。

实现方应声明在链路故障和拥塞切换两种场景下的路径切换响应时间。

10 接口要求

10.1 北向接口要求

应支持RESTful API或gRPC等行业标准接口协议。数据封装格式宜采用JSON或GPB。应提供完整的接口定义文档。

应支持多路径调度模式的按流或按租户配置，包括调度粒度选择、子路径数量、端侧散射与网侧自适应路由的启停控制等。

应支持多路径核心参数的动态修改，包括但不限于乱序容忍深度、子路径权重、流块超时阈值等。参数修改应支持在线生效。

应支持多路径传输运行状态的查询，包括各子路径的有效吞吐量、负载水平、乱序重排缓冲区占用率、路径切换事件记录等。

应支持异常事件的异步订阅与推送，包括拥塞告警、子路径失效、乱序深度超限等。

接口调用响应时间宜≤50ms，稳态下的调用成功率应≥99.9%。

10.2 南向接口要求

应支持gRPC、NETCONF或其他设备管理协议。配置模型应兼容行业主流的YANG模型规范。

应支持向端侧RNIC下发子路径配置、路径熵生成规则、调度粒度、路径权重等参数。

应支持向网侧交换机下发自适应路由的调度粒度、选路策略、拥塞检测阈值等配置。

应支持设备状态的周期性上报，上报内容应至少包括各子路径的质量指标、乱序程度、重排缓冲区水位等。

当数据面设备感知到子路径失效或持续拥塞时，应支持高优先级事件上报。

接口应具备兼容性，支持不同厂商、不同型号的设备接入。

参 考 文 献

[1] Lu Y, Chen G, Li B, et al. Multi-Path transport for RDMA in datacenters. NSDI’18. USENIX, 2018: 357-371.

[2] Li X, Zhang Y, Lyu S, et al. Load Balancing for LLM Traffic via Flow Block. LCN’25. IEEE, 2025: 1-9.

[3] Bonato T, Abdous S, Kabbani A, et al. Uno: A One-Stop Solution for Inter-and Intra-Data Center Congestion Control and Reliable Connectivity. SC’25. ACM, 2025: 1195-1210.