ICS 35.110
M11

YD

中 华 人 民 共 和 国 通 信 行 业 标 准

YD/T XXXXX—XXXX

基于位图（Bitmap）的 RDMA 网络丢包检测技术要求

Requirements for Bitmap-based RDMA packet loss detection technology

（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、中国信息通信研究院、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、北京交通大学、华为技术有限公司、鹏城实验室、之江实验室、中关村实验室、泉城实验室、中国标准化研究院、浪潮电子信息产业股份有限公司、山东省计算中心（国家超级计算济南中心）。

本文件主要起草人：

基于位图（Bitmap）的RDMA网络丢包检测技术要求

1 范围

本文件规定了基于位图(Bitmap)机制的RDMA网络丢包检测技术的术语定义、总体架构、功能要求、性能要求、协议与接口要求、设备实现要求、测试方法等内容。

本文件适用于可靠连接类RDMA业务流的丢包检测与状态维护，也可为面向高带宽时延积、乱序容忍和多路径转发的RDMA增强型可靠传输机制提供实现参考。

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 4273-2023 无损网络应用场景与需求

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

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

IETF RFC 5041 Direct Data Placement over Reliable Transports

IETF RFC 5040 A Remote Direct Memory Access Protocol Specification

IETF RFC 4297 Remote Direct Memory Access (RDMA) over IP Problem Statement

IETF RFC 4296 The Architecture of Direct Data Placement (DDP) and Remote Direct Memory Access (RDMA) on Internet Protocols

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 位图 Bitmap

以二进制比特位序列表示窗口内分组状态的数据结构，每一比特位与一个或一组包序列号建立对应关系，用于记录数据包的接收、缺失、待确认或待重传状态。

3.1.2 丢包检测 Packet Loss Detection

通过位图状态变化、序列号推进关系、定时器机制、乱序容忍策略、反馈事件等手段，识别传输过程中数据包缺失的过程。

3.1.3 接收位图 Receive-side Bitmap

由接收端维护的位图，用于描述接收窗口内各包序列号对应分组的到达状态，支撑乱序接收判断、缺口识别和完整性交付判定。

3.1.4 发送位图 Send-side Bitmap

由发送端维护的位图，用于描述发送窗口内分组的确认状态、疑似丢失状态、重传状态或待处理状态，支撑精细化重传决策。

3.1.5 全局位图 Global Bitmap

在单一逻辑接收窗口范围内统一维护的位图结构，用于描述全局包序列号空间中的状态信息。

3.1.6 路径级位图 Path-level Bitmap

针对多路径场景中各子路径分别维护的位图结构，用于区分跨路径乱序与真实丢失，并支撑路径相关统计与反馈。

3.1.7 变长位图 Variable-length Bitmap

可根据往返时延、带宽时延积、拥塞状态、窗口大小等因素动态调整长度的位图结构。

3.1.8 分段位图 Segmented Bitmap

将较大的逻辑窗口划分为多个连续片段进行管理的位图组织方式，用于降低单次访问复杂度并提升可扩展性。

3.1.9 丢包缺口 Loss Gap

在连续序列号推进过程中，由位图中未置位区间或状态不连续区间形成的疑似丢包区段。

3.1.10 快速反馈 Fast Feedback

在无需等待粗粒度超时的条件下，基于位图缺口、乱序阈值或触发条件向发送端反馈疑似丢失状态的机制。

3.1.11 远程直接内存访问 RDMA

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。

3.2 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
RoCEv2	基于融合以太网的 RDMA 网络第二版	RDMA over Converged Ethernet v2
PSN	包序列号	Packet Sequence Number
QP	队列对	Queue Pair
BDP	延迟带宽积	Bandwidth-Delay Product
SACK	选择性确认	Selective Acknowledgement

4 概述

RDMA网络已广泛应用于数据中心存储、分布式计算、人工智能训练、内存池化及高性能计算等场景。在当前工程实现中，可靠连接类业务通常基于包序列号、确认反馈和重传机制保障数据正确交付。随着组网规模增大和业务对吞吐、时延、抖动的要求持续提高，RDMA运行环境逐步呈现高带宽、高并发、大窗口、多路径和一定乱序比例等特征。

在传统顺序确认和粗粒度重传模式下，当网络存在突发拥塞、微突发排队、链路抖动或多路径乱序时，发送端和接收端对分组状态的维护复杂度上升，可能导致以下问题：

a) 丢包识别粒度较粗，无法快速定位具体缺失分组；

b) 难以区分乱序与真实丢包，易产生误判；

c） 丢包状态结构在大窗口场景下扩展性不足；

d） 重传触发滞后，影响端到端完成时延；

e） 硬件实现中的查找、插入和状态合并复杂度较高。

位图机制通过将窗口内包序列号与固定位置的比特位建立映射关系，可在有限状态空间内实现对分组接收状态、确认状态及缺失状态的精细描述。位图具有索引直接、更新简单、硬件流水线友好、适合并行访问等特点，适用于高性能网卡和交换设备实现。通过结合窗口滑动、缺口判定、乱序容忍、反馈与重传协同等机制，位图可提升RDMA丢包检测的准确性和时效性，为网络可靠性增强提供技术基础。

因此，本文件对基于位图的RDMA网络丢包检测技术提出统一要求，规范其术语、架构、功能边界、性能指标、接口行为和测试方法，为产业侧实现和互通验证提供依据。需要说明的是，接收位图为本文件描述的基础能力模型，发送位图为可选增强能力；实现方可采用等效状态机制替代。

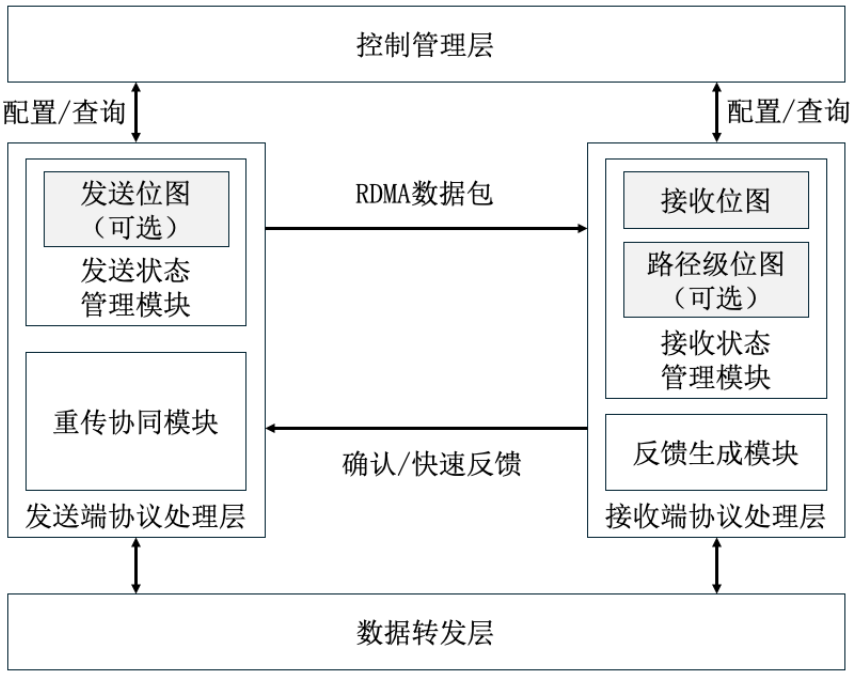

图1 基于位图的RDMA网络丢包检测总体架构图

5 核心参数

基于位图的RDMA网络丢包检测系统涉及的核心参数包括但不限于：

a） 位图长度；

b） 位图片段大小；

c） 窗口起始PSN；

d） 窗口滑动步长；

e） 快速反馈触发阈值；

f） 丢包判定等待时间；

g） 最大乱序容忍深度；

h） 路径级位图实例数；

i） 单QP状态上限；

j） 单设备最大并发QP数量。

系统实现应支持上述参数的配置、查询和能力声明，具体默认值和推荐值可由实现方根据设备定位、链路特征和场景规模确定，并在产品文档中公开说明。

6 技术架构要求

6.1 分层架构要求

系统宜采用控制管理层、协议处理层、数据转发层的分层设计，各层功能边界清晰。

层间接口应支持双向交互。控制管理层向协议处理层下发位图配置、阈值策略和查询请求；协议处理层向控制管理层反馈状态统计、事件和告警；协议处理层通过标准数据面报文与数据转发层协同承载业务流。

系统架构宜支持软硬件协同实现，允许位图状态管理逻辑部署在主机协议栈、RNIC固件、RNIC硬件流水线或其他可编程加速单元中。

系统应支持可扩展部署，在单机、多机柜、跨机房和多路径等组网场景中保持一致的功能模型。

6.2 核心组件功能要求

系统核心组件包括发送状态管理模块、接收状态管理模块、反馈生成模块、重传协同模块、管理与监测模块等。

6.2.1 发送状态管理模块

应支持按QP建立发送窗口状态。

应支持发送位图功能，至少支持以下能力：

a） 标识已发送分组状态；

b） 标识已确认分组状态；

c） 标识待重传分组状态或重传请求状态；

d） 支持窗口推进后的状态清理和复用。

应支持依据确认反馈、快速反馈或超时事件更新发送位图。

应支持在启用发送位图时按PSN精确定位待处理分组。

应支持与重传模块联动，避免对已确认分组重复重传。

6.2.2 接收状态管理模块

应支持按QP建立接收窗口状态。

应支持接收位图功能，至少支持以下能力：

a） 对窗口内接收分组进行置位；

b） 对重复到达分组进行识别；

c） 对乱序到达分组进行记录；

d） 对未到达位置形成缺口标识；

e） 对连续完整区间进行推进。

应支持基于接收位图区分乱序与持续缺失。

应支持在消息级交付前进行完整性检查。

应支持窗口滑动后对历史位图状态进行释放或复用。

接收位图宜支持全局位图模式，在多路径场景下宜支持路径级位图模式。

6.2.3 反馈生成模块

应支持生成基于位图状态的确认信息、快速反馈信息或统计信息。

应支持以下反馈触发条件中的一种或多种：

a） 连续完整区间推进；

b） 缺口持续存在达到阈值；
c） 乱序深度达到阈值；
d） 接收窗口停滞达到阈值；
e） 管理面主动查询。

反馈信息应至少能够反映以下内容中的一项或多项：
a） 确认边界；
b） 缺口位置；
c） 缺失区间长度；
d） 当前窗口基准；
e） 路径标识；
f） 时间戳或统计计数。

6.2.4 重传协同模块

应支持与发送状态管理模块协同触发精细化重传。
应支持按单个PSN、连续区间或策略合并方式生成重传请求。
应支持重传抑制机制，避免在反馈重复或状态未收敛情况下产生冗余重传。
应支持与拥塞控制机制协同，重传流量不应破坏基本拥塞控制约束。

6.2.5 管理与监测模块

应支持设备能力发现和参数查询。
应支持上报以下信息中的一项或多项：
a） 当前启用的位图模式；
b） 位图长度和片段数量；
c） 单QP状态占用；
d） 丢包事件次数；
e） 快速反馈次数；
f） 重传次数；
g） 误判修正次数；
h） 路径级统计。

应支持阈值告警与事件日志记录。

7 功能技术要求

7.1 位图状态管理要求

7.1.1 位图映射要求

位图应支持PSN到比特位的确定性映射关系。
映射关系应支持随窗口滑动进行更新，不得因窗口推进造成状态歧义。
当采用分段位图时，应明确段编号、段内偏移和段间切换规则。
当采用变长位图时，应支持长度调整后的状态迁移或重建，并确保窗口边界一致。

7.1.2 位图更新要求

接收端在收到有效分组后，应能够在对应位置完成置位操作。
接收端对重复分组应支持识别并记录，宜支持重复统计计数。
发送端在收到确认或快速反馈后，应支持更新相应状态位。

位图更新操作应满足实现侧并发访问一致性要求，避免出现状态覆盖、错误清零或交叉污染。

7.1.3 位图滑动要求

当连续完整区间满足推进条件时，系统应支持窗口滑动。

窗口滑动应同步完成以下操作：

a） 更新窗口基准PSN；

b） 释放已无效的历史状态；

c） 保留未完成区间对应状态；

d） 维持新窗口映射关系正确性。

窗口滑动不应导致未完成缺口信息丢失。

7.1.4 位图模式要求

系统应至少支持全局位图模式。

在单路径场景下，系统应支持基于全局位图的统一状态维护与丢包检测。

在多路径场景下，系统宜支持路径级位图模式。启用路径级位图模式时，应能够按路径分别维护接收状态，并结合路径标识进行缺口识别与统计。

当路径级位图模式不可用、路径标识缺失或实现未启用该能力时，系统应退化为全局位图模式。

系统应支持查询当前启用的位图模式，并支持对不同模式下的检测结果进行区分统计。

7.2 丢包检测要求

7.2.1 基本检测要求

系统应支持基于位图缺口识别疑似丢失分组。

在接收窗口内，若某一PSN之前的后续分组已到达且对应位置置位，且该PSN在满足判定条件后仍未置位，则应可判定为疑似丢失。

系统应支持将疑似丢失与短时乱序进行区分，避免立即将全部缺口认定为真实丢失。

7.2.2 乱序容忍要求

系统应支持配置最大乱序容忍深度或等效阈值。

在未超过乱序容忍阈值前，系统宜保留缺口状态并延迟触发快速反馈。

在超过乱序容忍阈值、持续时间阈值或窗口停滞阈值后，系统应支持将缺口升级为丢包事件。

7.2.3 多路径检测要求

在多路径场景下，系统宜支持路径级位图或路径相关状态管理。

当启用路径级位图时，应支持：

a） 对不同子路径分别记录接收状态；

b） 结合路径标识统计乱序和缺口；

c） 降低跨路径乱序导致的误判概率。

在未启用路径级位图时，系统应至少支持基于全局位图的统一检测。

7.2.4 长距离大窗口检测要求

在高BDP场景下，系统应支持扩展位图长度、分段位图或等效状态扩展机制。

系统应支持在窗口较大时维持缺口识别能力，不应因窗口增大导致功能失效。

7.3 反馈与确认要求

7.3.1 确认要求

系统应支持基于连续完整接收区间生成确认信息。

确认信息应至少支持表示当前可确认的连续边界。

确认信息生成后，发送端应能够据此释放已完成状态或推进发送窗口。

7.3.2 快速反馈要求

当检测到满足快速反馈条件的疑似丢失事件时，系统应支持生成快速反馈。

快速反馈宜至少包含以下信息中的一项或多项：

a） 触发QP标识；

b） 缺口起始位置；

c） 缺口长度；

d） 当前接收边界；

e） 时间戳；

f） 路径标识。

系统应支持快速反馈抑制和合并，避免频繁反馈导致控制面开销过大。

7.4 重传协同要求

系统应支持以下重传触发来源中的一种或多种：

a） 快速反馈触发；

b） 确认停滞触发；

c） 定时器触发；

d） 管理面强制触发。

系统应支持按粒度控制重传范围，包括单包重传、区间重传和策略聚合重传。

系统应支持对同一缺口事件进行去重处理，避免重复下发重传动作。

7.5 设备管理功能要求

设备应支持位图模式启停、参数下发、运行状态查询和统计信息导出。

设备应支持按端口、按QP、按业务流或按租户维度进行策略配置。

设备应支持日志记录，包括但不限于参数变更、异常事件、反馈触发、重传协同和状态恢复事件。

8 性能技术要求

8.1 检测时效要求

在满足实现能力声明的前提下，系统应支持在不依赖粗粒度超时的条件下完成快速丢包识别。

位图状态更新路径应满足低时延处理要求，宜支持线速或准线速更新。

快速反馈生成时延应小于对应业务流粗粒度超时阈值，并宜与网络RTT保持同数量级。

8.2 检测准确性要求

系统应具备较高的丢包检测准确性。

在存在短时乱序场景下，系统应通过乱序容忍机制降低误报概率。

在启用路径级位图或等效机制后，多路径场景下的误判率应较未区分路径状态时明显降低。

8.3 资源占用要求

系统应声明单QP状态占用能力边界，包括但不限于位图长度、附加元数据和计数器占用。

位图实现应支持按设备能力进行资源上限控制，避免状态资源耗尽影响业务稳定性。

当设备资源接近上限时，应支持限流、降级、告警或拒绝新策略下发。

8.4 并发能力要求

设备应支持多QP并发维护位图状态。

设备应声明最大并发QP数量、单QP最大窗口能力及总状态容量。

在达到额定并发规模时，系统核心功能不应失效，性能退化应可预测并可监测。

9 协议与接口要求

9.1 管理接口要求

控制管理层与设备之间应支持标准化管理接口，可采用CLI、NETCONF、RESTful API、gRPC或等效机制。

管理接口应至少支持以下能力：

a) 位图模式配置；

b) 阈值和定时器配置；

c) 状态查询；

d) 统计导出；

e) 告警订阅；

f) 版本与能力发现。

接口应支持权限控制、操作审计和版本管理。

9.2 数据面协同要求

本文件重点规范状态管理、检测逻辑和协同行为，不对现有RoCE基础报文格式提出统一修改要求，但设备实现应明确其用于确认、快速反馈、重传协同的报文承载方式。

当实现采用现有协议字段扩展、私有控制报文或管理队列承载反馈时，应在产品规范中公开说明其兼容边界、互通范围和异常处理行为。

当实现涉及发送端和接收端协同位图状态时，应保证双方对窗口基准、PSN映射和反馈语义的一致理解。

9.3 北向接口要求

面向网络管理平台、运维系统或控制器的北向接口应支持策略配置、状态查询和事件订阅。

接口返回结果宜至少包含以下信息中的一项或多项：

a) QP标识；

b) 当前窗口基准；

c) 位图长度；

d) 丢包统计；

e) 误报统计；

f） 快速反馈统计；
g） 当前资源占用。

9.4 南向接口要求

设备内部各模块之间的南向接口或内部调用路径应支持高效传递位图状态、缺口事件、反馈事件和重传请求。

在软硬件分离实现中，模块间接口应考虑同步机制、缓存一致性及异常恢复。

10 设备实现要求

10.1 网卡设备要求

RNIC应支持接收位图或等效状态机制，宜支持发送位图或等效重传状态机制。

RNIC应支持按QP维护状态，不同QP之间状态隔离。

RNIC应支持在硬件、固件或驱动层完成参数配置与状态上报。

RNIC应支持异常恢复，包括位图重建、窗口同步和状态清理。

10.2 交换设备要求

交换设备不强制实现位图逻辑，但宜支持以下增强能力中的一种或多种：

a） 提供拥塞与队列统计；
b） 提供路径标识辅助信息。

10.3 主机协议栈要求

当位图逻辑部署在主机协议栈或软件栈时，应支持与RNIC驱动、管理面和上层应用的协同。

软件实现应支持高并发访问保护，避免因锁竞争导致状态失真或性能严重下降。

10.4 管理系统要求

管理系统应支持策略模板、分级配置、批量下发、告警联动和历史分析。

管理系统应支持对位图相关统计进行展示，包括缺口分布、乱序分布、重传趋势和资源利用率。

11 测试方法

11.1 测试原则

测试应覆盖单路径、多路径、低时延、高时延、高BDP、低乱序、高乱序、突发丢包、持续丢包等典型场景。

测试应分别验证功能正确性、性能边界、资源占用和异常恢复能力。

11.2 测试环境

测试环境宜包括：

a） 至少两台支持RoCE的端系统；
b） 至少一台承载业务流的交换设备；
c） 可配置丢包、时延、乱序、路径切换的网络环境或流量仿真平台；
d） 支持抓包、日志采集和状态查询的管理系统。

11.3 基本功能测试

应验证以下内容：

a） 位图初始化是否正确；

b） 分组到达后对应位置是否正确置位；

c） 重复分组是否可识别；

d） 连续完整区间是否可正确推进；

e） 窗口滑动后映射是否正确；

f） 状态清理是否正确。

判定准则：上述测试项结果应与预期一致，不应出现错误置位、错误清零、窗口基准错位或未完成状态丢失。

11.4 丢包检测测试

在指定位置引入单包丢失、连续区间丢失和随机丢失，验证系统是否能正确识别缺口。

记录检测结果、触发时延、误报和漏报情况。

判定准则：缺口位置识别结果应与注入位置一致；未注入位置不应产生误报；检测结果应可查询或导出。

11.5 乱序容忍测试

在不引入真实丢包的条件下制造乱序，验证系统是否在设定阈值内保持缺口等待而不误触发丢包事件。

在超过阈值后，验证系统是否可按策略触发快速反馈或异常标识。

判定准则：当乱序深度或持续时间未超过配置阈值时，不应误报丢包；超过阈值后，应按配置触发相应事件。

11.6 多路径测试

在多路径场景下，验证路径级位图或等效机制是否可降低跨路径乱序误判。

应分别测试全局位图模式和路径级位图模式下的检测效果差异。

判定准则：启用路径级位图或等效机制后，误判率应较仅使用全局位图模式时降低，或不高于实现方声明值。

11.7 大窗口测试

通过增大RTT或提高发送速率构造高BDP场景，验证变长位图、分段位图或等效机制在大窗口下的功能完整性和资源可控性。

判定准则：在实现方声明的窗口规模内，丢包检测、状态维护和窗口滑动功能不应失效。

11.8 重传协同测试

验证快速反馈触发后，发送端是否能够按预期粒度执行重传。

验证重复反馈、延迟反馈、过期反馈下的抑制机制是否有效。

判定准则：重传范围应与触发条件一致；不应对已确认分组产生不必要的重复重传。

11.9 性能测试

测试项目宜包括：

a） 位图更新时延；

b） 丢包检测时延；

c） 快速反馈生成时延；

d） 单设备最大并发QP能力；

e） 单QP最大窗口能力；

f） 单QP状态占用；

g） 吞吐影响与CPU开销。

判定准则：测试结果应与实现方公开声明的能力边界一致，并应可重复验证。