国 家 知 识 产 权 局

250014

山东省济南市历下区经十路 17703 号华特广场 B510 室 济南圣达知识产权代理有限公司

王雪(0531-68605722)

发文日：

2026 年 05 月 26 日

申请号：202610735428.3　　**发文序号：**2026052602167280

专 利 申 请 受 理 通 知 书

根据专利法第 28 条及其实施细则第 43 条、第 44 条的规定，申请人提出的专利申请已由国家知识产权局受理。现将确定的申请号、申请日等信息通知如下：

申请号：2026107354283

申请日：2026 年 05 月 26 日

申请人：山东省计算中心（国家超级计算济南中心）,齐鲁工业大学（山东省科学院）

发明人：谭立状,褚夫明,史慧玲,张玮,张志远

发明创造名称：一种基于可编程交换机的网络拥塞信号转换方法及系统

经核实，国家知识产权局确认收到文件如下：

权利要求书 1 份 1 页,权利要求项数 ： 10 项

说明书 1 份 7 页

说明书附图 1 份 1 页

说明书摘要 1 份 1 页

发明专利请求书 1 份 5 页

实质审查请求书 文件份数：1 份

申请方案卷号：2026703546

提示：

1.申请人收到专利申请受理通知书之后，认为其记载的内容与申请人所提交的相应内容不一致时，可以向国家知识产权局请求更正。

2.申请人收到专利申请受理通知书之后，再向国家知识产权局办理各种手续时，均应当准确、清晰地写明申请号。

国家知识产权局 专利申查业务章 1101081359734

审 查 员：自动受理

联系电话：010-62356655

审查部门：初审及流程管理部

纸件申请，回函请寄：100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局专利局受理处收

电子申请，应当通过专利业务办理系统以电子文件形式提交相关文件。除另有规定外，以纸件等其他形式提交的文件视为未提交。

权 利 要 求 书

1. 一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，包括：
实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文；
识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；
当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写；
将改写后的返程报文发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

2. 根据权利要求 1 所述的一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，根据 ECN 标识信息为对应流维护以下状态信息：FlowKey、窗口报文计数器P、带 ECN 标记报文计数器P_e、拥塞标志位 cong_flag、拥塞等级标志 Cong_Level 以及对应的 RTT 改写量索引。

3. 根据权利要求 1 所述的一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，拥塞事件的判定包括：
直接检测去程报文头中的 CE 标记，若 ECN 字段中的 CE 比特已置位，则判定该报文对应一次 ECN 拥塞事件。

4. 根据权利要求 1 所述的一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，拥塞事件的判定还包括：
可编程交换机根据本地队列长度、排队时延、缓存占用状态或预设的 ECN 标记策略自行判定该去程报文是否应视为拥塞事件。

5. 根据权利要求 1 所述的一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写，包括：
可编程交换机识别返程报文，判断该返程报文是否为 RTT 测量相关报文；若不属于则直接转发；若属于则解析 FlowKey；
读取该 FlowKey 的拥塞状态 Cong_Level，若 Cong_Level 为 0，则直接转发；若 Cong_Level 不为 0，若 Cong_Level 不为 0，则计算改写增量 ΔRTT，改写增量 ΔRTT 由 ECN−RTT 拥塞信号转换计算方法获得。

6. 根据权利要求 5 所述的一种基于可编程交换机的网络拥塞信号转换方法，其特征在于，ECN−RTT 拥塞信号转换计算方法，包括：
可编程交换机统计观测窗口内转发数据包数量P和带 ECN 标记数据包数量P_e，计算$e = P_e/P$；
可编程交换机通过查表函数计算$\Delta RTT = T(e)$。

7. 一种基于可编程交换机的网络拥塞信号转换系统，其特征在于，包括：
报文捕获模块：实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文；
拥塞状态判定模块：识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；
拥塞信号转换模块：当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 时间戳信息的测量值进行改写；
拥塞信号识别模块：将改写后的返程报文发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

8. 一种电子设备，其特征在于，包括：
存储器，用于非暂时性存储计算机可读指令；以及
处理器，用于运行所述计算机可读指令，
其中，所述计算机可读指令被所述处理器运行时，执行上述权利要求 1−6 任一项所述的一种基于可编程交换机的网络拥塞信号转换方法。

9. 一种存储介质，其特征在于，非暂时性存储计算机可读指令，其中，当非暂时性计算机可读指令由计算机执行时，执行权利要求 1−6 任一项所述的一种基于可编程交换机的网络拥塞信号转换方法。

10. 一种计算机程序产品，其特征在于，包括计算机程序，所述计算机程序当在一个或多个处理器上运行的时候用于实现上述权利要求 1−6 任一项所述的一种基于可编程交换机的网络拥塞信号转换方法。

100001
2023.03

说 明 书

一种基于可编程交换机的网络拥塞信号转换方法及系统

技术领域

[0001] 本发明涉及电子信息与数据中心网络技术领域，特别是涉及一种基于可编程交换机的网络拥塞信号转换方法及系统。

背景技术

[0002] 拥塞控制是保障远程直接内存访问 RDMA 网络传输稳定性的关键机制，行业主流拥塞控制信号包括基于往返时间 RTT 的拥塞控制信号、基于显式拥塞通知 ECN 的拥塞控制信号及基于带内网络遥测 INT 的拥塞控制信号。为提升拥塞控制精度，业界普遍采用 ECN 与 RTT 联合部署方案。

[0003] 但由于二者触发机制、反馈通路及反馈粒度存在固有差异，因此采用 ECN 与 RTT 联合部署方案导致发送端需维护多套处理逻辑，CPU 与网卡负载显著升高; 并且不同信号的拥塞判断标准不一致易引发调速冲突，会破坏网络稳定性，广域等非全量 ECN 支持场景兼容性差，部署维护成本高。

[0004] 现有针对 ECN 与 RTT 联合部署方案的改进或采用信号优先级调度,或通过软件优化合并部分逻辑，均未从根本上减少发送端逻辑维护数量，也未能改善拥塞响应延迟，无法充分利用返程报文的传输特性。

发明内容

[0005] 为了解决现有技术中针对 ECN 与 RTT 联合部署方案的不足，本发明提供了一种基于可编程交换机的网络拥塞信号转换方法及系统，通过网络侧实现的拥塞信号转换机制，在不增加端侧处理逻辑的前提下，实现发送端对网络拥塞的及时、一致感知。

[0006] 一方面，提供了一种基于可编程交换机的网络拥塞信号转换方法，包括：

实时捕获 RDMA 网络中传输数据流的去程数据报文；

识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；

当判定某条流处于拥塞状态时，获取该流最近到达可编程交换机的返程报文，提取返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写；

将改写后的 RTT 值发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

[0007] 进一步的，根据 ECN 标识信息为对应流维护以下状态信息：FlowKey、窗口报文计数器P、带 ECN 标记报文计数器P_e、拥塞标志位 cong_flag、拥塞等级标志 Cong_Level 以及对应的 RTT 改写量索引。

[0008] 进一步的，拥塞事件的判定包括：直接检测去程报文头中的 CE 标记，若 ECN 字段中的 CE 比特已置位，则判定该报文对应一次 ECN 拥塞事件。

[0009] 进一步的，拥塞事件的判定还包括：可编程交换机根据本地队列长度、排队时延、缓存占用状态或预设的 ECN 标记策略自行判定该去程报文是否应视为拥塞报文。

[0010] 进一步的，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写，包括：

可编程交换机识别返程报文，判断该返程报文是否为 RTT 测量相关报文；若不属于则直接转发；若属于则解析 FlowKey；

读取该 FlowKey 的拥塞状态 Cong_Level，若 Cong_Level 为 0，则直接转发；若 Cong_Level 不为 0，若 Cong_Level 不为 0，则计算改写增量ΔRTT，改写增量ΔRTT 由 ECN−RTT 拥塞信号转换计算方法获得。

[0011] 进一步的，ECN−RTT 拥塞信号转换计算方法，包括：

可编程交换机统计观测窗口内转发数据包数量P和带 ECN 标记数据包数量P_e，计算$e = P_e/P$；

可编程交换机通过查表函数计算$\Delta RTT = T(e)$。

[0012] 另一方面，提供了一种基于可编程交换机的网络拥塞信号转换系统，包括：

报文捕获模块：实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文；

拥塞状态判定模块：识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；

拥塞信号转换模块：当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 时间戳信息的测量值进行改写；

拥塞信号识别模块：将改写后的返程报文发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

[0013] 再一方面，还提供了一种电子设备，包括：

存储器，用于非暂时性存储计算机可读指令；以及

处理器，用于运行所述计算机可读指令，

其中，所述计算机可读指令被所述处理器运行时，执行上述第一方面所述的方法。

100002
2023.03

[0014] 再一方面，还提供了一种存储介质，非暂时性存储计算机可读指令，其中，当非暂时性计算机可读指令由计算机执行时，执行第一方面所述方法。

[0015] 再一方面，还提供了一种计算机程序产品，包括计算机程序，所述计算机程序当在一个或多个处理器上运行的时候用于实现上述第一方面所述的方法。

[0016] 上述技术方案具有如下优点或有益效果：

本发明提出一种基于可编程交换机的拥塞信号转换方法及系统，可编程交换机对每条流维护拥塞状态，当在去程路径检测到该流的 ECN 拥塞事件时，在返程路径对该流的 RTT 测量相关报文中的时间戳字段进行受控改写，使发送端仅依据 RTT 拥塞控制逻辑即可提前降速或抑制增速，减少发送端 ECN 处理负担并降低多信号冲突风险。本发明将网络侧的 ECN 类拥塞信息转换为端侧可直接利用的 RTT 类拥塞信息，在交换机不拥塞时保持透明转发，在交换机发生拥塞或即将拥塞时提前发出统一反馈，降低发送端对 ECN/CNP 处理逻辑的依赖，减少多种拥塞信号并存所导致的冲突风险，并能够适配中间网络对 ECN 支持不完整的部署场景。

附图说明

[0017] 构成本发明的一部分的说明书附图用来提供对本发明的进一步理解，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。

[0018] 图 1 为实施例一所述的一种基于可编程交换机的网络拥塞信号转换方法流程图；

图 2 为实施例一所述的一种基于可编程交换机的网络拥塞信号转换方法的应用场景示意图。

具体实施方式

[0019] 应该指出，以下详细说明都是示例性的，旨在对本发明提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本发明所属技术领域的普通技术人员通常理解的相同含义。

[0020] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制本发明的示例性实施方式。术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0021] 在本发明实施例中，“和/或”仅仅是一种描述关联对象的关联关系，表示可以存在三种关系。例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，在本发明的描述中，“多个”是指两个或多于两个。

[0022] 另外，为了便于清楚描述本发明实施例的技术方案，在本发明实施例中，采用了“第一”、“第二”等字样对功能和作用基本相同的相同项或相似项进行区分。本领域技术人员可以理解“第一”、“第二”字样并不对数量和执行次序进行限定，并且“第一”、“第二”等字样也并不限定一定不同。

[0023] 在不冲突的情况下，本发明中的实施例及实施例中的特征可以相互组合。

[0024] 本实施例所有数据的获取都在符合法律法规和用户同意的基础上，对数据的合法应用。

[0025] 定义本发明所需定义的关键对象与术语，包括：

流：在交换机侧可被识别并维护独立状态的通信实体。可由以下方式之一进行标识：1）包括源 IP、目的 IP、源端口、目的端口、协议号的五元组。2）RDMA/RoCE 相关字段：如 QPN、PSN、UDP 端口加上 IP 头字段等组合。本公开采用 FlowKey 唯一标识一条流，FlowKey 可由上述标识字段通过哈希计算获得。

[0026] 去程方向与返程方向：去程方向指数据或请求从发送端到接收端的方向。返程方向指 ACK、RTT 响应报文或其他携带时间戳的反馈从接收端返回发送端的方向。

[0027] ECN 拥塞事件：交换机检测到去程报文 ECN 字段中的 CE 标记已置位。在可选实现中，ECN 拥塞事件还包括交换机依据本地端口队列长度、排队时延、缓存占用或标记概率策略对去程报文执行 ECN 标记的情形。

[0028] RTT 测量相关报文：可用于发送端计算 RTT 的报文类型，至少包含时间戳信息。典型形式包括，1）独立 RTT 探针报文和响应报文，响应报文携带发送时间戳与接收时间戳。2）ACK 携带时间戳字段，发送端基于发送时间戳与接收时间戳计算 RTT。本公开不限定具体报文格式，只要交换机可识别并可对用于 RTT 计算的时间戳字段执行受控改写。进一步地，在某些实现中，RTT 测量相关报文可携带由发送端写入并由接收端回显的时间戳字段，从而使 RTT 计算仅依赖发送端本地时钟域，不要求发送端与接收端时钟同步。

[0029] 时间戳字段：用于 RTT 计算的字段集合，抽象为T_s与T_r。T_s表示发送端写入或用于表征发送时刻的时间戳。T_r表示接收端写入或用于表征接收时刻的时间戳。RTT 计算通常可表述为$RTT = T_r - T_s$或其等价形式。

[0030] 此外，本公开不严格要求发送端和接收端执行时间同步，若发送端和接收端未执行时间同步，则发送端可采用相邻 RTT 测量报文计算连续两次 RTT 测量偏移量进行拥塞度量。

[0031] 实施例一

远程直接内存访问（Remote Direct Memory Access，RDMA）技术具有低延迟、高带宽、低 CPU 占用率的技术优势，是数据中心内部服务器间高速数据传输的核心网络技术，已广泛应用于人工智能训练与推理、分布式存储、高性能计算等领域。在 RDMA 网络中，拥塞控制是保障网络传输稳定性、避免链路超限、提升传输效率的关键机制。目前行业内主流的拥塞控制信号主要分为三类：基于往返时间（Round-Trip Time，RTT）的拥塞控制信号、基于显式拥塞通知（Explicit Congestion Notification，ECN）的拥塞控制信号，以及基于带内网络遥测（In-band Network Telemetry，INT）的拥塞控制信号。

[0032] 基于 RTT 的拥塞控制信号，通过检测数据包自发送端发送至接收端并返回响应的往返时间，判断网络拥塞程度，当 RTT 测量值增大时，判定网络发生拥塞，发送端相应降低传输速率。基于 ECN 的拥塞控制信号，由网络中的交换机检测链路拥塞状态，当链路发生拥塞时，交换机在数据包头部标记 ECN 标识，接收端识别该标识后，通过拥塞通告报文（Congestion Notification Packet，CNP）向发送端发送拥塞通知，发送端接收到 CNP 后执行速率下调操作。基于 INT 的拥塞控制信号，通过在业务数据包中嵌入网络链路的实时状态信息，如链路利用率、队列长度等，使发送端能够根据上述实时信息动态调整传输速率，适用于对拥塞感知精度要求较高的应用场景。

[0033] 为进一步提升拥塞控制精度，当前在 RDMA 网络中逐渐采用 ECN 与 RTT 两类拥塞控制信号联合部署的方式，实现拥塞联合控制。例如，在高速数据中心网络中，RDMA 网卡支持硬件时间戳特性且开放可编程拥塞控制接口，允许用户定制开发高速拥塞控制算法，采用 RTT 与 ECN 联合拥塞控制获得优于单一拥塞控制信号的控制效果。在广域 RDMA 传输场景中，由于数据中心网络交换机支持 ECN 特性，而广域核心网不支持 ECN 特性，因此发送端通常采用以 RTT 测量为主、ECN 测量为辅的拥塞控制策略。然而，ECN 与 RTT 在触发机制、反馈通路及反馈粒度方面存在固有差异，导致端侧在联合使用两类信号时，需要额外设置协调机制以避免发生拥塞控制冲突。

[0034] 针对上述现有技术中多套拥塞控制信号共存的部署架构，结合实际应用场景进行分析可知，该部署方式存在以下核心技术缺陷：

首先，发送端负载显著增加，导致数据传输效率下降。由于发送端需同时维护基于 RTT、ECN 等多套拥塞控制信号的处理逻辑与速率调整逻辑，且每套逻辑均需独立的硬件资源及软件进程提供支撑，导致发送端 CPU、网络接口卡（Network Interface Card，NIC）等资源占用率显著升高。这不仅增加了发送端的硬件部署成本，还间接削弱了 RDMA 技术低延迟、高带宽的技术优势，在高并发、大数据量传输场景下，上述问题尤为突出。此外，在部分仅开放 RTT 类可编程拥塞控制接口、未开放 ECN 事件接口的硬件平台上，多信号共存的部署方式会进一步提高技术实现门槛。

[0035] 其次，不同拥塞控制信号之间易发生冲突，导致发送端执行错误的速率调整操作，破坏网络传输的稳定性。由于 ECN 与 RTT 等信号的拥塞判断标准、响应时机存在差异，可能出现同一时刻不同信号所传递的拥塞状态不一致的情形。例如，当交换机检测到链路发生轻微拥塞时，对数据包标记 ECN 标识并通知发送端降速，但此时发送端检测到的 RTT 值未发生明显变化，基于 RTT 的拥塞控制逻辑判定网络未发生拥塞，进而维持甚至提高传输速率；反之，当 RTT 值增大提示网络拥塞时，CNP 可能尚未生成或未到达发送端，导致发送端的速率调整逻辑产生冲突，最终执行错误的速率调整操作，加剧网络拥塞或造成链路带宽浪费，严重影响数据中心 RDMA 网络的传输稳定性与可靠性。此外，上述冲突还可能表现为端侧对同一拥塞阶段执行重复降速、降速幅度叠加过大，或在拥塞恢复阶段无法及时上调传输速率等问题。

[0036] 并且，基于多类拥塞控制信号的联合拥塞控制通常仅能获得联合性能增益，无法有效降低拥塞控制环路延迟。基于多类信号的联合拥塞控制，其理论最低响应延迟仍处于（1/2RTT，RTT）区间内，无法突破闭环控制延迟瓶颈；当网络发生突发拥塞时，发送端无法及时接收拥塞通知，易引发拥塞丢包。此外，现有技术方案对于中间网络非全量支持 ECN 的应用场景缺乏有效的解决方案；例如在广域传输网络中，由于广域核心网不支持 ECN 特性，发送端通常直接放弃 ECN 拥塞控制，无法获得联合拥塞控制的性能增益。

[0037] 最后，部署与维护成本高，设备兼容性差。多套拥塞控制信号共存的部署方式，需要对发送端、接收端及网络中的交换机进行复杂的配置与调试，以确保各套逻辑之间能够协同工作，存在配置复杂度高、兼容性要求高、部署成本高的问题。此外，不同厂商的设备对各类拥塞控制信号的支持程度存在差异，易发生信号识别错误、响应异常等问题，导致网络整体的维护与升级成本显著增加。尤其在异构数据中心环境中，不同网卡与交换机对 ECN、时间戳、反馈报文格式的支持差异，会进一步放大上述维护复杂度。

[0038] 为解决上述采用多套拥塞控制信号存在的问题，目前，业界已提出若干改进方案，但均未能从根本上解决上述问题。部分方案试图简化发送端的速率调整逻辑，采用优先级排序或改进型调度方式，设定某一类拥塞控制信号为最高优先级，如 ECN 信号优先，当多类信号发生冲突时，仅对最高优先级信号作出响应。但该类方案未能解决发送端需同时维护多套拥塞控制逻辑的资源负载问题，且优先级的静态设定易

引发拥塞判断偏差，仍可能导致发送端执行错误的速率调整操作；同时，该类方案未能有效改善拥塞响应延迟，亦未充分利用返程 RTT 测量通告报文或拥塞通告报文 CNP 的逐包响应及连续传输特性。

[0039] 部分方案试图通过软件优化手段合并部分处理逻辑，或采用联合拥塞控制策略提升控制性能增益，但上述方案均未能从根本上减少发送端所需维护的拥塞控制逻辑数量，发送端 CPU、NIC 等硬件资源负载过高的技术问题仍未得到有效解决。

[0040] 基于此，本发明提供了一种基于可编程交换机的网络拥塞信号转换方法，在发送端部署基于 RTT 的拥塞控制逻辑和速率调整逻辑，确保发送端能够正常解析返程 RTT 测量通告中的时间戳信息。可编程交换机负责 ECN 和 RTT 的拥塞信号转换，接收端负责数据接收和响应反馈。本发明通过由网络侧实现的拥塞信号转换机制，能够在不增加端侧处理逻辑复杂度的前提下，使发送端更早、更一致地感知网络拥塞状态并执行相应的拥塞控制操作。

[0041] 本实施例提供了一种基于可编程交换机的网络拥塞信号转换方法，如图 1 所示，包括：

实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文；

识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；

当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写；

将改写后的 RTT 值发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

[0042] 本实施例提供了一种基于可编程交换机的网络拥塞信号转换方法，具体步骤包括：

S1：实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文。

[0043] 本实施例通过可编程交换机捕获 RDMA 网络中传输数据流的所有数据报文，包括发送端向接收端发送的去程报文，以及接收端向发送端发送的返程报文，其中返程报文为连续的 RTT 测量响应报文。

[0044] S2：识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态，具体包括：

在本实施例中，拥塞事件的判定包括两种方式：第一种方式为直接通过可编程交换机读取去程报文的 ECN 标识信息，识别 ECN 标识信息中的 CE 字段，若 ECN 标识信息中的 CE 字段置位则判定拥塞事件。

[0045] 第二种方式为交换机根据本地队列长度、排队时延、缓存占用状态或预设的 ECN 标记策略自行判定该报文是否应视为拥塞事件。

[0046] 本实施例根据 ECN 标识信息为对应流维护以下状态信息：FlowKey、窗口报文计数器P、带 ECN 标记报文计数器P_e、拥塞标志位 cong_flag、拥塞等级标志 Cong_Level 以及对应的 RTT 改写量索引。其中，P 和 Pe 用于近似统计一个观测窗口内的 ECN 拥塞强度，Cong_Level 用于表示当前流处于无拥塞、轻微拥塞或严重拥塞状态。

[0047] 若判断为拥塞事件，则将该去程报文记为带 ECN 标记报文，并令 Pe 加 1。无论是否带 ECN 标记，每接收到一个去程报文，P 均加 1。根据识别的 ECN 标记信息，更新对应流的拥塞状态为未拥塞、轻微拥塞和严重拥塞。

[0048] 可编程交换机接收去程报文，解析 FlowKey；若 FlowKey 不存在，则注册 FlowKey，并置 Cong_Level 为 0。其中，FlowKey 采用五元组与 RDMA 队列标识相结合的方式进行构造。优选地，FlowKey 可由源 IP 地址、目的 IP 地址、源 UDP 端口、目的 UDP 端口以及 QPN 字段组合得到。

[0049] 在本实施例中，观测窗口采用固定包数窗口，优选为 8 个去程报文构成一个观测窗口。采用固定包数窗口的优点在于无需在交换机中执行复杂除法运算，便于在数据平面中直接通过整数计数与查表完成等效 RTT 增量的计算。具体地，对于每条 FlowKey，交换机在去程方向接收到一个报文时，先读取该流的 P 和 Pe。若该流首次出现，则创建 FlowKey 对应状态项，同时将 P、Pe 和 Cong_Level 初始化为 0。

[0050] 本实施例中，当交换机完成对当前去程报文的拥塞统计后，为避免发送端继续接收到原始 ECN 信号并触发额外的 ECN 处理逻辑，交换机将该去程报文中的 ECN 字段重置为 0 后再转发。这样，端侧主要依据后续转换得到的 RTT 拥塞信号进行调速，有利于减少 RTT 和 ECN 两类反馈同时存在时产生的冲突风险。

[0051] S3：当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，提取返程报文中的 RTT 时间戳信息，对 RTT 测量值进行改写；

对于返程报文，交换机将源地址与目的地址、源端口与目的端口进行互换后生成反向匹配键，从而与对应去程流关联，交换机能够对同一条 RDMA 业务流的去程报文和返程报文进行一致识别。

[0052] 对于处于“未拥塞”状态的流，不对其返程报文进行任何处理，直接转发。

[0053] 对于处于“轻微拥塞”或“严重拥塞”状态的流，根据该流最近到达的返程报文，识别其中的发送时间戳或接收时间戳，按照预设的改写规则，修改发送时间戳或接收时间戳，生成改写后的时间戳或，确保改写后的 RTT 值能够被发送端的 RTT 检测逻辑识别为拥塞信号，具体包括：

S31：可编程交换机识别返程报文，判断该返程报文是否为 RTT 测量相关报文。若不属于则直接转发；若

属于则解析 FlowKey。

[0054] S32：读取该 FlowKey 的拥塞状态 Cong_Level。读取该 FlowKey 的拥塞状态 Cong_Level。若 Cong_Level 为 0，则直接转发；若 Cong_Level 不为 0，则计算改写增量 ΔRTT。ΔRTT的确定方式可为以下任一：

1）固定增量：ΔRTT=常量D。

[0055] 2）分档映射：$\Delta RTT = f(Cong\_Level) * D$，其中，$f(Cong_Level)$从 1 到 N 线性映射对应 1%到 100%。

[0056] 3）概率映射：$\Delta RTT = \max\{P_{max}, f(Cong\_Level)\} * D$，其中，$P_{max}$为交换机配置的 ECN 最大标记概率。

[0057] 4）根据本实施例所述的 ECN-RTT 拥塞信号转换计算方法获得。

[0058] 其中，常量 D 可由管理员手动配置或按交换机每端口期望转发延迟平均上界计算获得，即 $D = \frac{Buffer \times 8}{PortNum * PortRate}$。

[0059] 若 Cong_Level 不为 0，则更新该 FlowKey 状态，置 Cong_Level 为 Cong_Level+1。若 Cong_Level+1>N，则置 Cong_Level 为 N，N 为设定阈值；若未发生拥塞事件，更新该 FlowKey 状态，置 Cong_Level 为 Cong_Level-1。若 Cong_Level-1<0，则置 Cong_Level 为 0。

[0060] 为了使发送端收到被改写后的 RTT 测量相关报文后，产生与 ECN 拥塞强度相当的减速行为，本实施例提出 ECN-RTT 拥塞信号转换计算方法，使可编程交换机实现时间戳字段的合理改写。该方法可以表述为假设交换机在一个观测窗口内检测到某流的 ECN 拥塞强度为 e，则交换机在该流返程报文中注入一个等效 RTT 增量 ΔRTT，使得发送端基于 Timely 拥塞控制算法在该窗口触发的速率更新幅度，近似等于 ECN 反馈本应触发的速率更新幅度。本实施例所述的 ECN-RTT 拥塞信号转换计算方法，具体包括以下步骤：

S321：交换机统计观测窗口内转发数据包数量P和带 ECN 标记数据包数量P_e，计算$e = P_e/P$。

[0061] S322：交换机通过查表函数计算$\Delta RTT = T(e)$。

[0062] 其中，查表函数可设计如下表 1 所示。

[0063] 表 1 查表函数表

e	[0,0.2)	[0.2,0.4)	[0.4,0.6)	[0.6,0.8)	[0.8,1]
ΔRTT	0	D/8	D/4	D/2	D

[0064] 在本实施例中，当某条流的 P 达到预设窗口大小 8 时，交换机基于该窗口内的 ECN 拥塞强度计算对应的等效 RTT 改写量 ΔRTT。理论上，ECN 拥塞强度 e=Pe/P。由于本实施例中 P 固定为 8，因此无需执行除法，只需根据 Pe 的取值直接查表即可。优选地，采用如下查表规则：

当 Pe 为 0 或 1 时，对应 e 落入区间[0,0.2)，令ΔRTT=0；

当 Pe 为 2 或 3 时，对应 e 落入区间[0.2,0.4)，令ΔRTT=D/8；

当 Pe 为 4 时，对应 e 落入区间[0.4,0.6)，令ΔRTT=D/4；

当 Pe 为 5 或 6 时，对应 e 落入区间[0.6,0.8)，令ΔRTT=D/2；

当 Pe 为 7 或 8 时，对应 e 落入区间[0.8,1]，令ΔRTT=D。

[0065] 其中，D 为基础时间增量，用于表征交换机向发送端注入的一次标准拥塞反馈强度。在本实施例中，D 由交换机控制面预先配置，优选可依据交换机端口速率和期望缓冲占用上界计算得到；也可由管理员根据具体网络规模、链路速率和目标拥塞灵敏度直接设置。对于 100Gbps 端口的示例性部署，D 可取 2μs。

[0066] 在本实施例中，交换机完成查表后，将所得 ΔRTT 对应保存为该流当前的拥塞反馈量，并同步更新 Cong_Level。优选地，可将 Cong_Level 设置为 0、1、2、3、4 五个等级，分别对应 ΔRTT 取值为 0、D/8、D/4、D/2 和 D。随后，交换机将该流的 P 和 Pe 清零，开始统计下一个观测窗口。采用该方式，交换机仅需维护小位宽整数计数器和少量查表项，适合在 P4 交换机上部署实现。

[0067] S33：对返程报文的 RTT 时间戳字段进行受控改写，改写发送时间戳$T_s' = T_s - \Delta RTT$或改写接收时间戳$T_r' = T_r + \Delta RTT$。

[0068] 在本实施例中，优选采用改写接收时间戳 Tr 的方式，即将返程报文中的接收时间戳改写为 Tr'=Tr+ΔRTT。采用该方式的优点在于实现逻辑较简单，且不易出现发送时间戳下溢的问题。在其他实施方式中，也可以改写发送时间戳。若报文协议需要校验和更新，则交换机在完成时间戳字段改写后同步

更新相应校验字段。

[0069] S4：将改写后的返程报文发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

[0070] 在本实施例中，发送端按照既有的 RTT 拥塞控制逻辑对改写后的返程报文进行处理。由于交换机在拥塞发生时人为增大了返程报文中的等效 RTT，发送端在计算 RTT 时将观测到 RTT 上升，从而触发 Timely 算法执行减速或抑制增速。这样，发送端无需额外处理 ECN 事件或 CNP 报文，即可借助现有 RTT 控制逻辑对拥塞作出响应。

[0071] 下面结合一个具体工作过程对本实施例的运行方式进行说明，如图 2 所示。假设发送端持续向接收端发送某一条 RDMA 业务流的去程报文，可编程交换机端口速率为 100Gbps 且支持本实施例所述的拥塞信号转换功能，基础时间增量 D 配置为 2us，观测窗口大小为 8 个去程报文。

[0072] 在第一个观测窗口内，可编程交换机共接收到该流的 8 个去程报文，其中 5 个报文已带 CE 标记或被交换机判定为应执行 ECN 标记，因此 P=8，Pe=5。根据上述查表规则，Pe=5 对应 e 处于[0.6,0.8)区间，故令 ΔRTT=D/2=1us，并将 Cong_Level 更新为对应等级。

[0073] 随后，接收端返回 ACK 报文，可编程交换机在返程方向识别到该 ACK 报文携带时间戳字段，查得该流当前 ΔRTT=1us，于是将所述 ACK 报文中的 Tr 改写为 Tr+1us 后转发给发送端。

[0074] 发送端接收到该 ACK 报文后，计算得到的 RTT 值相较于未改写时增大约 1us，从而判断网络出现拥塞并执行速率下降。需要特别说明的是，可编程交换机对时间戳修改是在正常 RTT 基础上额外增加 ΔRTT，从而实现拥塞通告。

[0075] 本发明适用于支持 RDMA（Remote Direct Memory Access，远程直接内存访问）传输的大型数据中心网络或广域传输网络，尤其适配采用 ECN 和 RTT 作为拥塞控制信号的 RDMA 网络场景，实现拥塞信号的转换。

[0076] 实施例二

本实施例提供了一种基于可编程交换机的网络拥塞信号转换系统，包括：

报文捕获模块：实时捕获 RDMA 网络中传输数据流的所有数据报文，包括去程报文和返程报文；

拥塞状态判定模块：识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；

拥塞信号转换模块：当判定某条流处于拥塞状态时，根据该流最近到达可编程交换机的返程报文，识别返程报文中的 RTT 时间戳信息，对 RTT 时间戳信息的测量值进行改写；

拥塞信号识别模块：将改写后的返程报文发送至发送端，由发送端的 RTT 检测逻辑识别为拥塞信号，并作出响应。

[0077] 上述实施例中对各个实施例的描述各有侧重，某个实施例中没有详述的部分可以参见其他实施例的相关描述。

[0078] 所提出的系统，可以通过其他的方式实现。例如以上所描述的系统实施例仅仅是示意性的，例如上述模块的划分，仅仅为一种逻辑功能划分，实际实现时，可以有另外的划分方式，例如多个模块可以结合或者可以集成到另外一个系统，或一些特征可以忽略，或不执行。

[0079] 实施例三

本实施例还提供了一种电子设备，包括：一个或多个处理器、一个或多个存储器、以及一个或多个计算机程序；其中，处理器与存储器连接，上述一个或多个计算机程序被存储在存储器中，当电子设备运行时，该处理器执行该存储器存储的一个或多个计算机程序，以使电子设备执行上述实施例一所述的方法。

[0080] 应理解，本实施例中，处理器可以是中央处理单元 CPU，处理器还可以是其他通用处理器、数字信号处理器 DSP、专用集成电路 ASIC，现成可编程门阵列 FPGA 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0081] 存储器可以包括只读存储器和随机存取存储器，并向处理器提供指令和数据、存储器的一部分还可以包括非易失性随机存储器。例如，存储器还可以存储设备类型的信息。

[0082] 在实现过程中，上述方法的各步骤可以通过处理器中的硬件的集成逻辑电路或者软件形式的指令完成。

[0083] 实施例一中的方法可以直接体现为硬件处理器执行完成，或者用处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器、闪存、只读存储器、可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器，处理器读取存储器中的信息，结合其硬件完成上述方法的步骤。为避免重复，这里不再详细描述。

[0084] 本领域普通技术人员可以意识到，结合本实施例描述的各示例的单元及算法步骤，能够以电子硬件或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的范围。

[0085] 实施例四

本实施例还提供了一种存储介质，用于存储计算机指令，所述计算机指令被处理器执行时，完成实施例一所述的方法。

[0086] 实施例五

本实施例还提供了一种计算机程序产品，包括计算机程序，所述计算机程序当在一个或多个处理器上运行的时候用于实现实施例一所述的方法。

[0087] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

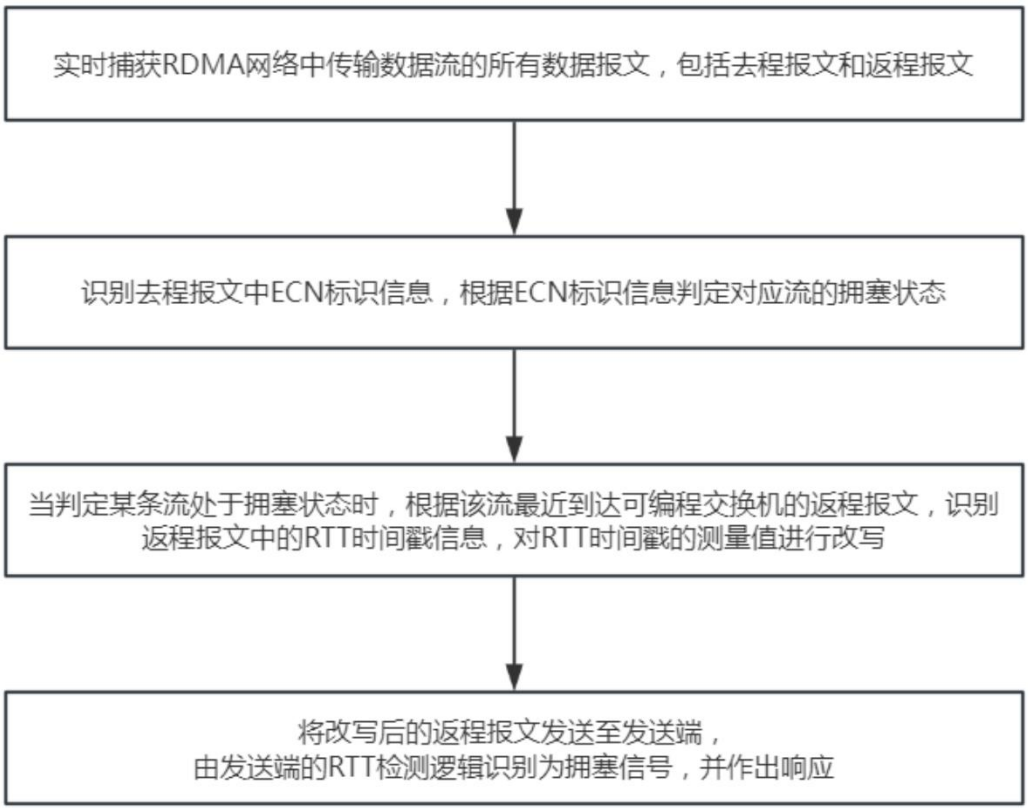

图 1

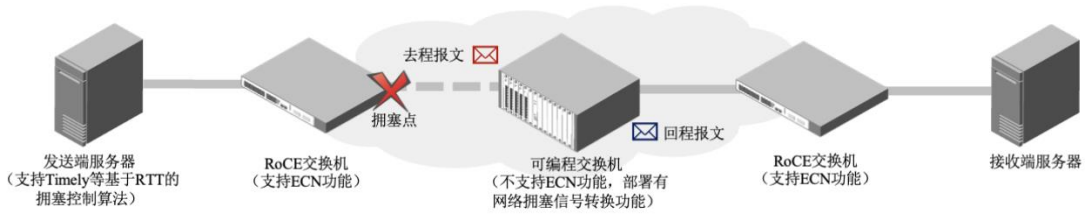

图 2

100003
2023.03

说 明 书 摘 要

本发明公开了一种基于可编程交换机的网络拥塞信号转换方法及系统，在去程方向对每条流进行识别，识别去程报文中 ECN 标识信息，根据 ECN 标识信息判定对应流的拥塞状态；在返程方向识别时间戳的 RTT 测量相关报文，并在拥塞状态有效时对用于 RTT 计算的时间戳字段进行受控改写，使发送端观测到增大的等效 RTT，从而仅依据 RTT 拥塞控制逻辑即可及时降速或抑制增速。本发明将网络侧的 ECN 类拥塞信息转换为端侧可直接利用的 RTT 类拥塞信息，在交换机不拥塞时保持透明转发，在交换机发生拥塞或即将拥塞时提前发出统一反馈，降低发送端对 ECN/CNP 处理逻辑的依赖，减少多种拥塞信号并存所导致的冲突风险。

100004
2023.03