(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107204931 A
(43)申请公布日 2017.09.26

(21)申请号 201710137920.1

(22)申请日 2017.03.09

(30)优先权数据
15/075,158 2016.03.20 US

(71)申请人 迈络思科技TLV有限公司
地址 以色列赖阿南纳

(72)发明人 巴拉克·加夫尼

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 陆建萍 郑霞

(51)Int.Cl.
H04L 12/801(2013.01)
H04L 12/863(2013.01)
H04L 12/823(2013.01)
H04L 12/833(2013.01)
H04L 12/935(2013.01)
H04L 12/861(2013.01)

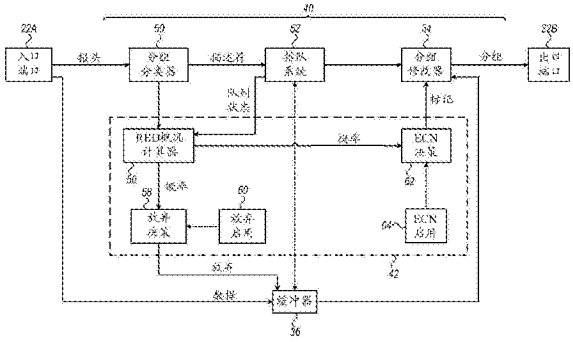
权利要求书2页 说明书5页 附图2页

(54)发明名称

拥塞控制措施的灵活应用

(57)摘要

本发明公开了拥塞控制措施的灵活应用。通信装置包括配置成连接到分组数据网络的多个接口和存储器，存储器耦合到接口并被配置为缓冲器以将通过入口接口接收的数据分组容纳在多个队列中同时等待经由出口接口到网络的传输。拥塞控制逻辑包括配置成响应于队列的状态而放弃缓冲器中的至少第一队列的数据分组的第一部分的分组丢弃机器和配置成响应于队列的状态而将拥塞通知应用于缓冲器中的至少第二队列的数据分组的第二部分的分组标记机器。机器控制电路被耦合以选择性地启用和禁用至少分组丢弃机器。



CN 107204931 A

1.一种通信装置,包括:

多个接口,其配置成连接到分组数据网络以便在通过所述装置从所述网络接收数据分组以及向所述网络转发数据分组中用作入口接口和出口接口;

存储器,其耦合到所述接口并被配置为缓冲器以将通过所述入口接口接收的数据分组容纳在多个队列中同时等待经由所述出口接口传输到所述网络;以及

拥塞控制逻辑,其包括:

分组丢弃机器,其配置成响应于所述队列的状态而放弃所述缓冲器中的至少第一队列的数据分组的第一部分;

分组标记机器,其配置成响应于所述队列的状态而将拥塞通知应用于所述缓冲器中的至少第二队列的数据分组的第二部分;以及

机器控制电路,其被耦合以选择性地启用和禁用至少所述分组丢弃机器。

2.如权利要求1所述的装置,其中所述机器控制电路还被耦合以选择性地启用和禁用所述分组标记机器。

3.如权利要求1所述的装置,其中所述分组丢弃机器和所述分组标记机器配置成放弃在所述队列的相同的一个或多个队列中的数据分组的相应部分以及将所述拥塞通知应用于在所述队列的相同的一个或多个队列中的数据分组的相应部分。

4.如权利要求1所述的装置,其中所述拥塞通知包括在所述数据分组的报头中设置显式拥塞通知(ECN)字段。

5.如权利要求1所述的装置,其中所述拥塞通知包括在所述数据分组的报头中设置业务类(TC)字段。

6.如权利要求1所述的装置,其中所述拥塞控制逻辑包括概况计算器,所述概况计算器配置成响应于所述第一队列和所述第二队列的相应状态而计算所述第一部分和所述第二部分。

7.如权利要求6所述的装置,其中所述概况计算器配置成通过比较所述队列的长度与在所述存储器中的所述队列的相应缓冲器分配来计算所述第一部分和所述第二部分。

8.如权利要求6所述的装置,其中所述概况计算器配置成基于所述队列的相应传输速率来计算所述第一部分和所述第二部分。

9.如权利要求6所述的装置,包括分组分类逻辑,所述分组分类逻辑配置成将通过所述入口接收的数据分组分配到所述多个队列,并将关于所接收的数据分组的信息传送到所述概况计算器。

10.一种用于通信的方法,包括:

在具有连接到分组数据网络以便用作入口接口和出口接口的多个接口并具有耦合到所述接口的存储器的网络元件中,将通过所述入口接口接收的数据分组放置在所述存储器中的多个队列中,同时所述数据分组等待传输到所述网络;

使用分组丢弃机器以及使用分组标记机器来将拥塞控制应用于被排队用于传输的数据分组,所述分组丢弃机器配置成响应于所述队列的状态而放弃所述缓冲器中的至少第一队列的数据分组的第一部分,所述分组标记机器配置成响应于所述队列的状态将拥塞通知应用于所述缓冲器中的至少第二队列的数据分组的第二部分;以及

选择性地启用和禁用至少所述分组丢弃机器,使得当所述分组丢弃机器被禁用时,所

述数据分组响应于由所述队列的状态指示的拥塞而不被所述网络元件丢弃。

11.如权利要求10所述的方法，包括选择性地启用和禁用所述分组标记机器。

12.如权利要求10所述的方法，其中应用所述拥塞控制包括放弃在所述队列的相同的一个或多个队列中的数据分组的相应部分以及将所述拥塞通知应用于在所述队列的所述相同的一个或多个队列中的数据分组的相应部分。

13.如权利要求10所述的方法，其中所述拥塞通知包括在所述数据分组的报头中设置显式拥塞通知(ECN)字段。

14.如权利要求10所述的方法，其中所述拥塞通知包括在所述数据分组的报头中设置业务类(TC)字段。

15.如权利要求10所述的方法，其中应用所述拥塞控制包括响应于所述第一队列和所述第二队列的相应状态而计算所述第一部分和所述第二部分。

16.如权利要求15所述的方法，其中计算所述第一部分和所述第二部分包括比较所述队列的长度与在所述存储器中的所述队列的相应缓冲器分配。

17.如权利要求15所述的方法，其中计算所述第一部分和所述第二部分包括评估所述队列的相应传输速率。

18.如权利要求15所述的方法，其中计算所述第一部分和所述第二部分包括在计算所述第一部分和所述第二部分时应用关于所接收的数据分组的信息。

拥塞控制措施的灵活应用

发明领域

[0001] 本发明通常涉及分组通信网络，且特别地涉及用于在这样的网络中的拥塞的控制的方法和系统。

[0002] 背景

[0003] 当需要在网络中的链路或节点运载比它能够传输或转发的更多的数据业务时，网络拥塞出现，结果是它的服务质量恶化。拥塞的一般结果包括排队延迟、分组丢失和新连接的堵塞。现代分组网络使用拥塞控制(包括拥塞避免)技术，以便在灾难性的结果开始之前减轻拥塞。

[0004] 很多拥塞避免技术在本领域中是已知的。例如在随机早期检测(RED，也被称为随机早期丢弃或随机早期放弃)中，网络节点例如交换机监控它们的平均队列大小并基于统计概率来丢弃分组：如果给定队列(或队列组)几乎是空的，则所有进入的分组被接受。当队列增长时，丢弃进入的分组的概率相应地增长，当缓冲器填充水平超过可适用的阈值时达到100％。加权RED(WRED)以类似的方式工作，除了不同的业务类被分配不同的拥塞避免阈值以外，使得对于给定队列长度，低优先级分组比高优先级分组具有更大的丢弃概率。对由统计概率确定的分组的部分操作的这种拥塞控制技术在本文被称为统计拥塞控制技术。

[0005] 另一拥塞避免技术是显式拥塞通知(ECN)，其为互联网协议(IP)和传输控制协议(TCP)的扩展。ECN最初由Ramakrishnan等人在“The Addition of Explicit Congestion Notification(ECN)to IP”中定义，该文作为Internet Engineering Task Force(2001)的请求注解(RFC)3168被出版并通过引用被并入本文。ECN通过在所传输的分组的IP报头中用信号通知即将发生的拥塞来提供网络拥塞的端对端通知。这种ECN标记的分组的接收方对发送方重复拥塞指示，这减小它的传输速率，好像它检测到放弃的分组一样。ECN功能最近扩展到其它传输和隧道协议。

[0006] 概述

[0007] 在下文所述的本发明的实施方式提供用于在网络中的拥塞控制的改进的方法和实现这样的方法的装置。

[0008] 因此根据本发明的实施方式提供通信装置，其包括多个接口，该多个接口配置成连接到分组数据网络以便用作通过该装置在来自网络和到网络的数据分组的接收和转发中的入口接口和出口接口。存储器耦合到接口并被配置为缓冲器以将通过入口接口接收的数据分组容纳在多个队列中同时等待经由出口接口到网络的传输。拥塞控制逻辑包括配置成响应于队列的状态而放弃来自缓冲器中的至少第一队列的数据分组的第一部分的分组丢弃机器和配置成响应于队列的状态而将拥塞通知应用于来自缓冲器中的至少第二队列的数据分组的第二部分的分组标记机器。机器控制电路被耦合以选择性地启用和禁用至少分组丢弃机器。

[0009] 在一些实施方式中，机器控制电路还被耦合以选择性地启用和禁用分组标记机器。

[0010] 在所公开的实施方式中，分组丢弃机器和分组标记机器配置成放弃在相同的一个

或多个队列中的数据分组的相应部分并将拥塞通知应用于在相同的一个或多个队列中的数据分组的相应部分。

[0011] 在一些实施方式中，拥塞通知包括在数据分组的报头中设置显式拥塞通知(ECN)或业务类(TC)字段。

[0012] 在所公开的实施方式中，拥塞控制逻辑包括概况计算器(profile calculator)，该概况计算器配置成响应于第一和第二队列的相应状态而计算第一和第二部分。一般，概况计算器配置成通过比较队列的长度与在存储器中的队列的相应缓冲器分配和/或基于队列的相应传输速率来计算第一和第二部分。此外或可选地，装置包括分组分类逻辑，其配置成将通过入口接收的数据分组分配到多个队列，并将关于所接收的数据分组的信息传送到概况计算器。

[0013] 还根据本发明的实施方式提供了用于通信的方法，其包括在具有连接到分组数据网络以便用作入口和出口接口的多个接口和耦合到接口的存储器的网络元件中，将通过入口接口接收的数据分组放置在存储器中的多个队列中，同时数据分组等待传输到网络。使用配置成响应于队列的状态而放弃缓冲器中的至少第一队列的数据分组的第一部分的分组丢弃机器并使用配置成响应于队列的状态将拥塞通知应用于缓冲器中的至少第二队列的数据分组的第二部分的分组标记机器，拥塞控制被应用于被排队用于传输的数据分组。至少分组丢弃机器被选择性地启用和禁用，使得当分组丢弃机器被禁用时，数据分组响应于由队列的状态指示的拥塞而不被网络元件丢弃。

[0014] 本发明将从连同附图一起理解的其实施方式的下面的详细描述中被更充分理解，其中：

[0015] 附图的简要描述

[0016] 图1是示意性示出根据本发明的实施方式的具有拥塞控制能力的交换机的方框图；以及

[0017] 图2是示意性示出根据本发明的实施方式的在交换机中的分组处理逻辑的细节的方框图。

[0018] 实施方式的详细描述

[0019] 在本领域中已知的网络元件例如交换机中，由ECN标记的分组结合由RED(包括WRED)放弃的分组在单个逻辑拥塞避免机器的控制下根据在上面提到的RFC 3168中定义的模型来操作。因此，ECN分组标记不能对可应用的分组启用，如果也不允许拥塞避免机器在拥塞严重的情况下放弃不受到ECN标记的分组的话。相反，当必须避免放弃某种类型的分组例如TCP控制分组(例如SYN和SYN/ACK分组)或其它无损业务类时，为了拥塞避免的目的的分组的标记也被禁用。

[0020] 本文所述的本发明的实施方式提供用于拥塞避免的更灵活的模型，其中分组丢弃和分组标记机制单独地和独立地被应用。在所公开的实施方式中，在通信装置例如网络交换机中的拥塞控制逻辑包括分组丢弃机器和分组标记机器。(如在本描述中和在权利要求中使用的术语“机器”指执行某个定义明确的任务的不同逻辑电路。)在装置中的机器控制电路被耦合以选择性地启用和禁用至少分组丢弃机器和可能也有分组标记机器。

[0021] 分组丢弃机器和分组标记机器的这个分离使系统操作员能够配置该装置以用于不同种类的拥塞响应：在拥塞的情况下仅标记、仅丢弃或标记和丢弃分组的适当部分。此

外，机器控制电路可设置分组丢弃和标记机器以将不同的拥塞响应应用于不同的队列以及不同类型的业务，使得TCP控制分组在拥塞的情况下例如被标记(但不被放弃)，而其它种类的分组可被放弃。分组丢弃机器和分组标记机器的分离也可增强拥塞控制的效率，因为分组丢弃可例如在网络交换机的处理管线中的早期被应用，以便立即释放缓冲器空间，而分组标记可在处理管线中的后期被应用以实现对在拥塞水平中的变化的快速响应。

[0022] 图1是示意性示出根据本发明的实施方式的具有拥塞控制能力的网络交换机20的方框图。交换机20包括连接到分组数据网络24并配置成在来自网络或到网络的数据分组26、28…的接收和转发中用作入口和出口接口的多个接口22，例如交换机端口。耦合到接口22的存储器36用作缓冲器以从入口接口接收分组并将分组保持在多个队列中，同时等待经由出口接口传输到网络24。在所示例子中，存储器36被配置为共享缓冲器，其中每个队列接收相应的分配38。可选地，本发明的原理可同样在网络元件中被应用，网络元件中不同的接口具有它们自己的单独缓冲器，或其中使用其它缓冲方案，例如在入口或出口端口的一部分之间的共享，作为图1所示的共享缓冲器的补充或替代。

[0023] 分配38(即队列被允许使用的缓冲器的量，或等效地，为了拥塞控制的目的的控制阈值)可以是静止的，或它们可随着时间的过去而改变。此外，不同的队列可接收不同大小的相应分配38，例如取决于业务优先级水平或其它系统考虑因素。指向同一出口接口的多个不同队列可接收它们自己的单独分配38。可选地或此外，存储器分配可在指向同一出口接口或甚至多个不同的出口接口的多个队列当中被共享。各种类型的动态缓冲器分配可由交换机20中的决策和排队逻辑40操纵，并将对由在交换机中的拥塞控制逻辑42应用的阈值有影响，但这些缓冲器分配机制本身在本描述的范围之外。例如在2015年3月30日提交的美国专利申请14/672,357中描述了可在这个上下文中使用的缓冲器分配机制，其公开通过引用被并入本文。

[0024] 拥塞控制逻辑42在这个例子中基于统计或其它拥塞控制标准来将拥塞控制例如ECN和/或WRED应用于被排队用于从存储器36中的每个队列传输到网络24的分组的相应部分。逻辑42一般基于在队列的长度和相应分配38的大小之间的关系在任何给定时间对每个队列将分组的该部分设置为在这个上下文中被标记或放弃。因此，响应于队列的状态并根据拥塞条件，拥塞控制逻辑42可放弃缓冲器中的某个队列或队列组的数据分组的某个部分，同时将拥塞通知标记应用于另一队列或队列组的数据分组的另一部分。这两组队列可交叉，意味着在一些或所有队列中，一些分组可被放弃，而其它分组用拥塞通知来标记。

[0025] 在图1所示的例子中，在交换机20中从网络24接收的分组26、28…包括报头30和有效载荷数据32，如在本领域中已知的。报头30在这个例子中被假设为IP报头，且因此包含ECN字段34，如在RFC 3168中规定的。决策和排队逻辑40将分组26和28放置在存储器36中的相应队列中，这两者在这个例子中都被假设为拥塞的。基于缓冲器填充水平和机器控制设置，拥塞控制逻辑42放弃分组28，并通过将ECN字段34设置为值“11”来标记分组26以当交换机20将这个分组转发到网络24时指示拥塞。

[0026] 虽然本描述为了具体性和清楚起见涉及图1所示的特定交换机20，本发明的原理可加以必要的变更来类似地应用于实现本文讨论的拥塞控制技术的种类的任何网络元件。因此在可选的实施方式中，这些原理可不仅应用在不同类型的切换装置例如路由器和桥中，而且例如应用在将主机计算机连接到网络的高级网络接口控制器中。此外，虽然本实施

方式特别涉及在IP网络中的拥塞控制并利用特别对这样的网络定义的技术例如ECN，本发明的原理可以可选地应用在其它种类的网络中和在统计(或可能非统计)拥塞控制所相关的不同协议例如MPLS、InfiniBand和以太网下。

[0027] 图2是根据本发明的实施方式示意性示出在交换机20中的分组处理逻辑的细节方框图。为了清楚和具体性起见，该附图示出决策和排队逻辑40的一个可能的实现，包括拥塞控制逻辑42，但其它实现将对阅读了本描述之后的本领域中的技术人员明显且被考虑为在本发明的范围内。虽然逻辑40和42的元件在图2中被示为单独的功能部件，实际上这些部件可一起在单个芯片或芯片组中的定制或可编程硬件逻辑中实现。

[0028] 当接收到进入的分组时，入口端口22A(例如在图1中的端口22之一)将分组有效载荷放置在存储器36中的缓冲器中并通知决策和排队逻辑40该分组为处理做好准备。分组分类器(packet classifier)50解析分组报头并产生确定出口端口22B(或多个端口)和队列的一个或多个描述符，该分组通过该出口端口22B(或多个端口)被传输，且该分组被放置在该队列中同时等待传输。描述符也可指示待应用于分组的服务质量(QoS)，即传输的优先级的水平，并可指示用于分组报头的修改的任何可应用的指令。分组分类器50将描述符放置在排队系统52中的适当的队列中，以等待经由指定出口端口进行传输。如早些时候提到的，排队系统52一般包含每个出口端口22B的专用队列或每出口端口的多个队列，每个QoS水平有一个队列。

[0029] 当描述符到达它的队列的报头时，排队系统52将描述符传递到分组修改器54用于执行。响应于描述符，分组修改器54从存储器36读取适当的分组数据，并做出在分组报头中要求的任何变化用于通过出口端口22B传输到网络24。这些变化可包括例如响应于来自拥塞控制逻辑42的指令通过将ECN字段34设置为拥塞通知来标记分组报头。

[0030] 拥塞控制逻辑42包括概况计算器56，其计算进入的分组可被分配到的每个队列的拥塞控制概率。这些概率被表示为分数，其从概况计算器56输入到分组丢弃机器58和分组标记机器62，为了在拥塞的情况下将被做出的放弃和ECN决策的目的。换句话说，对于任何给定队列在任何给定时间，由概况计算器56提供到分组丢弃机器58的概率值指示在队列中的分组的待放弃的部分；而被提供到分组标记机器62的概率值(其可与被提供到分组丢弃机器的概率值相同或不同)指示在队列中的分组的用拥塞通知标记的部分。

[0031] 概况计算器56基于由排队系统52提供的队列状态信息以及由分组分类器50分析的分组报头信息来计算并更新这些概率值。例如，分组分类器可为了这个目的参考指示业务类和拥塞状态的IP和传输报头字段。作为另一例子，当MPLS在使用中时，分组分类器可使用在MPLS报头中的相应字段(如由Davie等人的标题为“Explicit Congestion Marking in MPLS”的IETF RFC 5129提供的)和特别是在MPLS业务类(TC)字段中的QoS和拥塞通知信息(如由Andersson等人在IETF RFC 5462中定义的)。队列状态信息一般包括在讨论中的队列的长度和/或相应的传输速率，且概率值取决于这些长度与队列的可用缓冲器分配38的比较。相关的分组报头字段尤其包括在IP报头中的ECN和区分服务代码点(DSCP)字段。分组分类器50也可向分组丢弃机器58和分组标记机器62指示给定队列或分组类型是否适宜分组放弃、标记或这两者。

[0032] 拥塞控制逻辑42还包括机器控制电路，机器控制电路包括放弃启用电路60并可选地包括ECN启用电路64。放弃启用电路60被耦合以选择性地启用和禁用分组丢弃机器58，而

ECN启用电路64选择性地启用和禁用分组标记机器64。当放弃启用电路60禁用分组丢弃机器58时，例如拥塞控制逻辑42在拥塞的情况下仍将标记分组但将不丢弃分组。因此通过设置电路60和64，交换机20的系统操作员能够确定交换机将如何对拥塞做出响应：通过放弃分组、标记分组或这两者或没有这些功能中的任一个。这些设置可基于网络配置和状态以及其它系统要求随着时间的过去自动地或在直接操作员控制下改变。

[0033] 当被放弃启用电路60启用时，分组丢弃机器58基于来自概况计算器58的概率值来选择要从每个队列放弃的分组的适当部分。这些分组从存储器36和从在排队系统52中的相应队列删除。

[0034] 出于同样原因，当由ECN启用电路64启用时，分组标记机器62基于来自概况计算器58的概率值来选择在每个队列中的适当分组以用拥塞通知标记，并指示分组修改器54相应地修改分组报头。拥塞通知可被标记在例如IP报头的ECN字段中，如上面解释的，或在另一适当的报头字段例如MPLS TC字段中。分组然后经由出口端口22B被传输到网络24。

[0035] 将认识到，上面所述的实施方式作为例子被引证，以及本发明不限于在上文中特别示出和描述的内容。更确切地，本发明的范围包括在上文中所述的各种特征的组合和子组合以及本领域中的技术人员在阅读了前述描述时将想到的且在现有技术中未公开的其变形和修改。

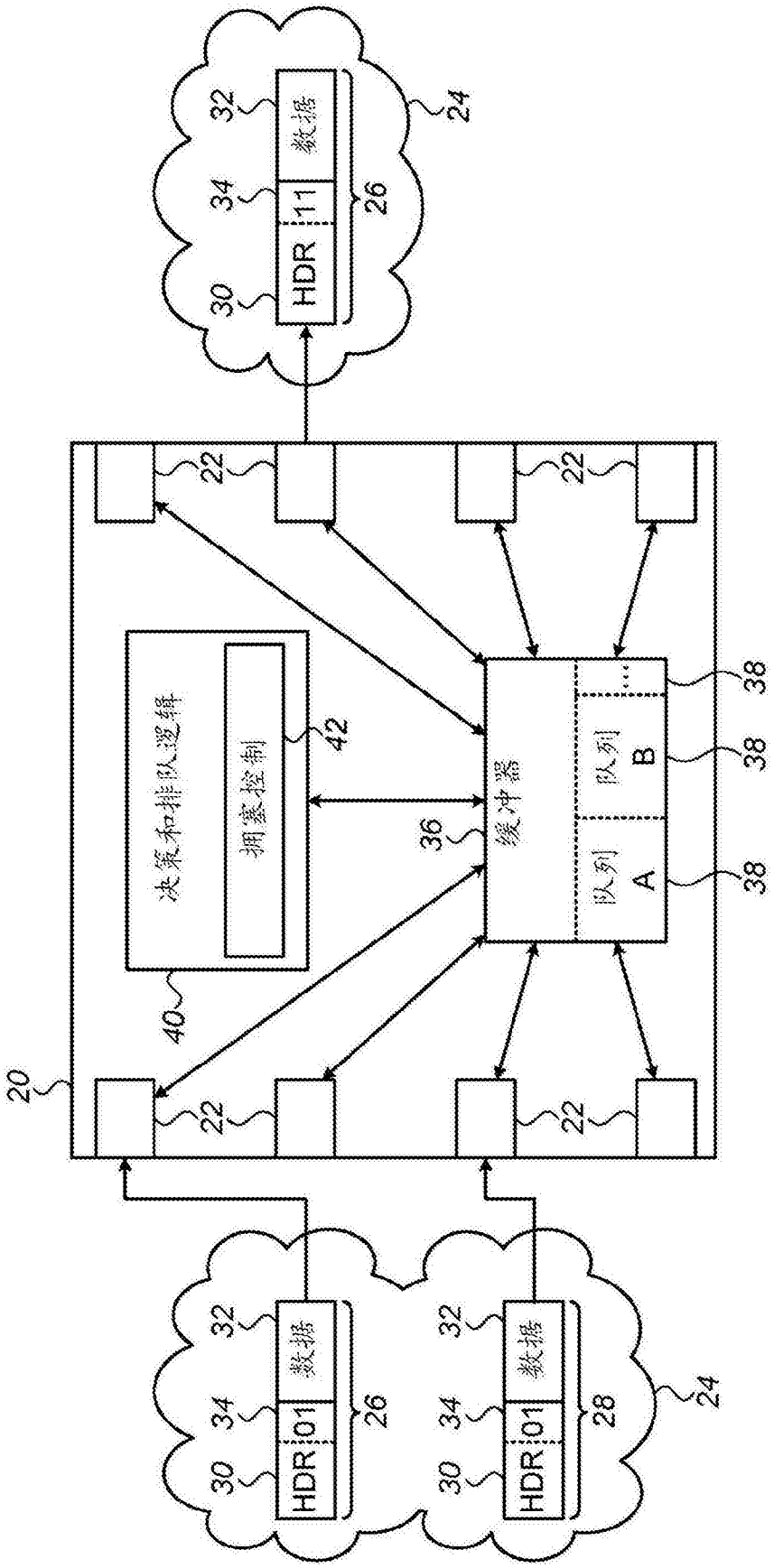

20
22
22
22
22
24
32
数据
34
11
26
30
HDR
决策和排队逻辑
拥塞控制
42
40
缓冲器
36
队列
B
队列
A
38
38
38
24
32
数据
34
01
26
30
HDR
32
数据
34
01
28
30
HDR

图1

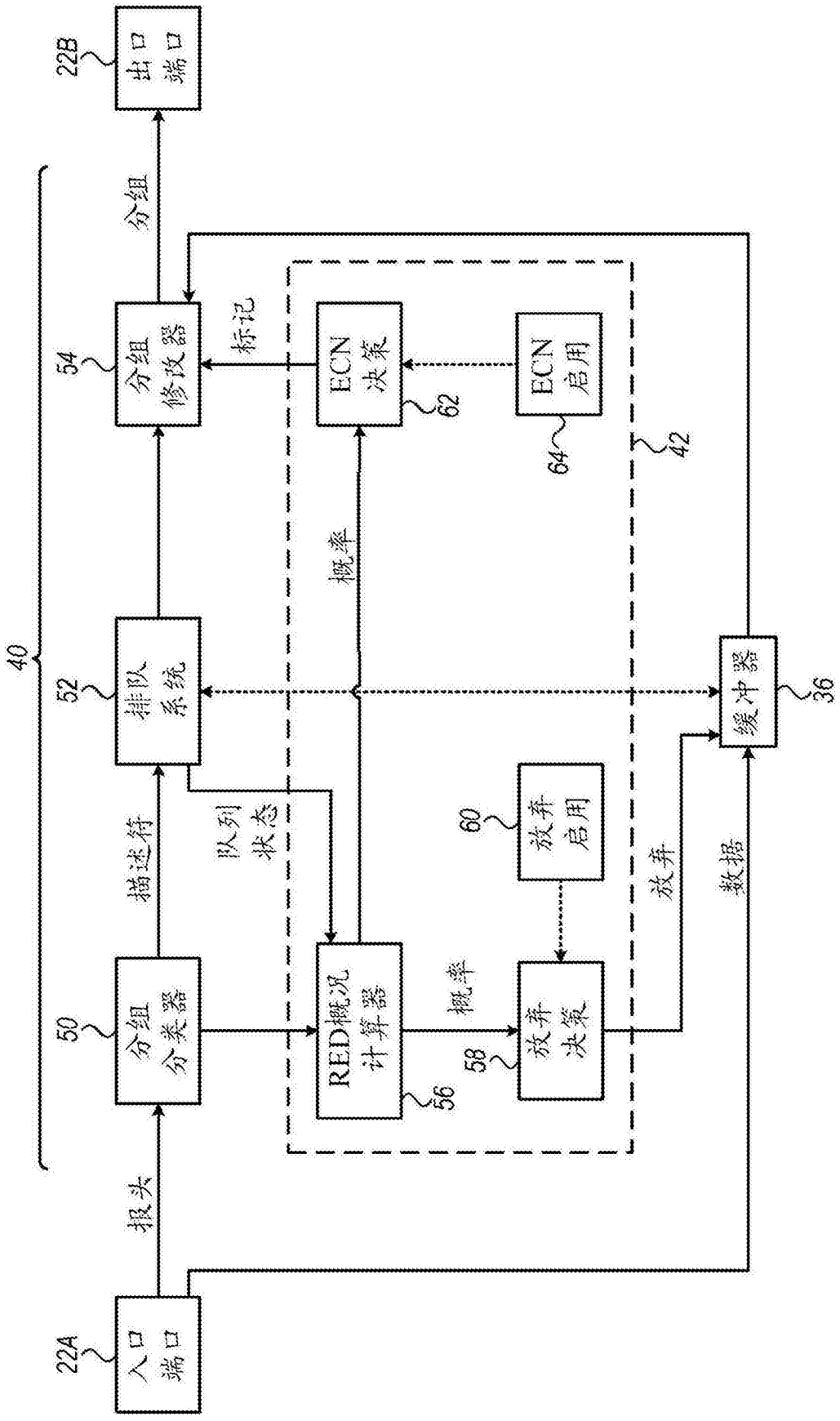

出口端口
22B
分组
分组修改器
54
标记
ECN决策
62
ECN启用
64
42
概率
40
排队系统
52
缓冲器
36
描述符
队列状态
60
放弃启用
放弃
数据
分组分类器
50
RED概况计算器
56
概率
放弃决策
58
报头
入口端口
22A

图2