(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 113572655 A
(43)申请公布日 2021.10.29

(21)申请号 202110665574.0

(22)申请日 2021.06.16

(71)申请人 清华大学
地址 100084 北京市海淀区双清路30号清华大学

(72)发明人 任丰原 张乙然 刘一凡 孟晴开

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
代理人 肖艳

(51)Int.Cl.
H04L 12/26(2006.01)
H04L 12/801(2013.01)
H04L 12/861(2013.01)
H04L 12/875(2013.01)

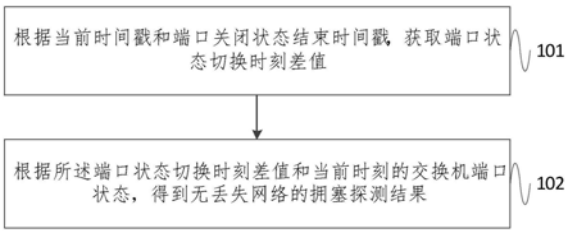
权利要求书2页 说明书9页 附图4页

(54)发明名称

无丢失网络的拥塞探测方法及系统

(57)摘要

本发明提供一种无丢失网络的拥塞探测方法及系统，该方法包括：根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。本发明通过检测交换机端口开启时期的时长和数据包队列长度演化特征，可准确探测识别无丢失网络中交换机端口的三种状态间转换，并为数据包进行正确标记，使得拥塞控制算法更加效率的完成调速决策，提高流完成时间和吞吐量。



CN 113572655 A

1.一种无丢失网络的拥塞探测方法,其特征在于,包括:

根据当前时间戳和端口关闭状态结束时间戳,获取端口状态切换时刻差值;

根据所述端口状态切换时刻差值和当前时刻的交换机端口状态,得到无丢失网络的拥塞探测结果。

2.根据权利要求1所述的无丢失网络的拥塞探测方法,其特征在于,所述方法还包括:

若无丢失网络采用基于缓存的逐跳流量控制技术,则所述端口关闭状态结束时间戳为上一次接收到RESUME帧的时间戳;

若无丢失网络采用基于定时器的逐跳流量控制技术,则所述端口关闭状态结束时间戳为上一次信用值为零的状态下,接收到新的信用值时的时间戳,所述信用值是通过信用限制值和交换机已发送数据块数量之间的差值获取得到的,所述信用限制值为下游交换机当前已分配缓存量和已接收到数据块之和。

3.根据权利要求1所述的无丢失网络的拥塞探测方法,其特征在于,所述根据所述端口状态切换时刻差值和当前时刻的交换机端口状态,得到无丢失网络的拥塞探测结果,包括:

将所述端口状态切换时刻差值和预设时长阈值进行比较判断;

根据判断结果和当前时刻的交换机端口状态,对所述交换机端口状态进行更新,并对数据包进行标记;

所述交换机端口状态包括非拥塞态、拥塞态和待定态,其中,所述非拥塞态为端口开启且无数据包队列累积的状态,所述拥塞态为端口开启且存在数据包队列累积的状态,所述待定态为端口处于开启与关闭多次切换的状态;

根据更新结果和标记结果,得到无丢失网络的拥塞探测结果。

4.根据权利要求3所述的无丢失网络的拥塞探测方法,其特征在于,在所述将所述端口状态切换时刻差值和预设时长阈值进行比较判断之前,所述方法还包括:

若无丢失网络采用基于缓存的逐跳流量控制技术,则所述预设时长阈值的公式为:

$$T_{max}=\frac{2(X_{off}-X_{on})+\tau C}{2\varepsilon C}+\tau;$$

$$\tau=2*\text{link delay}+\frac{2*MTU}{C};$$

其中,T_{max}表示预设时长阈值,X_{off}表示触发PAUSE帧发送的预设端口入队列长度阈值,X_{on}表示触发RESUME帧发送的预设端口入队列长度阈值,C表示链路带宽,τ表示响应时间,ε表示固定参数,link delay表示链路延迟,MTU表示最大传输单元;

若无丢失网络采用基于定时器的逐跳流量控制技术,则所述预设时长阈值为定时器的预设超时时间。

5.根据权利要求3所述的无丢失网络的拥塞探测方法,其特征在于,所述根据判断结果和当前时刻的交换机端口状态,对交换机端口状态进行更新,并对数据包进行标记,包括:

当所述端口状态切换时刻差值大于等于预设时长阈值,且当前时刻的交换机端口状态为非待定态时,

若数据包队列长度大于拥塞队列长度阈值时,将所述交换机端口状态更新为拥塞态,并为数据包标记经历拥塞态,将数据包出队;

若数据包队列长度小于等于所述拥塞队列长度阈值时，将所述交换机端口状态更新为非拥塞态，并将数据包出队。

6.根据权利要求3所述的无丢失网络的拥塞探测方法，其特征在于，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

当所述端口状态切换时刻差值大于等于所述预设时长阈值，且当前时刻的交换机端口状态为待定态时，

若数据包队列长度降低，且数据包队列长度大于拥塞队列长度阈值时，则将数据包出队；

若数据包队列长度降低，且数据包队列长度小于等于所述拥塞队列长度阈值，则将所述交换机端口状态更新为非拥塞态，并将数据包出队；

若数据包队伍长度未降低，且数据包队列长度大于所述拥塞队列长度阈值，则将所述交换机端口状态更新为拥塞态，并为数据包标记经历拥塞态，将数据包出队。

7.根据权利要求3所述的无丢失网络的拥塞探测方法，其特征在于，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

当所述端口状态切换时刻差值小于所述预设时长阈值，将所述交换机端口状态更新为待定态，若数据包未被标记为经历拥塞态，则为数据包标记经历待定态；

若数据包已被标记为经历拥塞态，则将数据包出队。

8.一种无丢失网络的拥塞探测系统，其特征在于，包括：

端口切换处理模块，用于根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；

拥塞探测模块，用于根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

9.一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至7任一项所述无丢失网络的拥塞探测方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述无丢失网络的拥塞方法的步骤。

无丢失网络的拥塞探测方法及系统

技术领域

[0001] 本发明涉及拥塞控制算法技术领域，尤其涉及一种无丢失网络的拥塞探测方法及系统。

背景技术

[0002] 无丢失网络以其零丢包和低延迟的优点，成为数据中心中一个很有前景的发展方向。无丢失网络依赖于逐跳的流量控制技术(简称逐跳流控技术)，从而保证零丢包，现有以下两种逐跳流控技术：

[0003] 一、基于缓存的逐跳流控技术：当下游交换机端口的入队列长度超过一个阈值X_{OFF}时，该下游交换机向上游交换机发送PAUSE帧，使得上游交换机接收到PAUSE帧时，停止发送数据；当下游交换机端口的入队列长度降低到另一个阈值X_{ON}时，该下游交换机向上游交换机发送RESUME帧，使得上游交换机收到RESUME帧时，恢复正常发送。

[0004] 二、基于定时器的逐跳流控技术：下游交换机维护一个接收block数目的寄存器(Adjust Block Register.，简称ABR)来记录接收到的数据量。同时，下游交换机周期性向上游交换机发送一个信用限制值(Flow Control Credit Limit，简称FCCL)消息，该消息由下游交换机端口维护，包含下游交换机当前已分配的缓冲区大小和ABR(即已接收到的数据块数量)的总和。上游交换机维护一个流量控制总发送块寄存器(Flow Control Total Block Sent，简称FCTBS)，记录交换机已发送数据块数量。当上游交换机接收到FCCL消息后，FCCL和FCTBS之间的差值为可用信用值的数量，上游交换机只能在存在有可用信用值时发送数据包。

[0005] 上述两种逐跳流控技术，本质上采用相同的思想避免交换机上缓存溢出：交换机端口可以在开(ON)和关(OFF)之间交替；当端口被关闭时，后续的数据包排队等待下游交换机上可用的缓存。在无丢失网络中，逐跳流量控制技术在保证零丢包的同时，交换机端口状态在开-关(ON-OFF)间存在短时间内多次来回切换的状态(待定态)，现有技术中忽视这种新状态，导致端口的真实状态被错误地探测，未拥塞的数据包被错误标记，进而使得拥塞控制算法作出错误的调速决策，影响流完成时间和吞吐量。因此，现在亟需一种无丢失网络的拥塞探测方法及系统来解决上述问题。

发明内容

[0006] 针对现有技术存在的问题，本发明提供一种无丢失网络的拥塞探测方法及系统。

[0007] 本发明提供一种无丢失网络的拥塞探测方法，包括：

[0008] 根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；

[0009] 根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

[0010] 根据本发明提供的一种无丢失网络的拥塞探测方法，所述方法还包括：

[0011] 若无丢失网络采用基于缓存的逐跳流量控制技术，则所述端口关闭状态结束时间

戳为上一次接收到RESUME帧的时间戳；

[0012] 若无丢失网络采用基于定时器的逐跳流量控制技术，则所述端口关闭状态结束时间戳为上一次信用值为零的状态下，接收到新的信用值时的时间戳，所述信用值是通过信用限制值和交换机已发送数据块数量之间的差值获取得到的，所述信用限制值为下游交换机当前已分配缓存量和已接收到数据块之和。

[0013] 根据本发明提供的一种无丢失网络的拥塞探测方法，所述根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果，包括：

[0014] 将所述端口状态切换时刻差值和预设时长阈值进行比较判断；

[0015] 根据判断结果和当前时刻的交换机端口状态，对所述交换机端口状态进行更新，并对数据包进行标记；

[0016] 所述交换机端口状态包括非拥塞态、拥塞态和待定态，其中，所述非拥塞态为端口开启且无数据包队列累积的状态，所述拥塞态为端口开启且存在数据包队列累积的状态，所述待定态为端口处于开启与关闭多次切换的状态；

[0017] 根据更新结果和标记结果，得到无丢失网络的拥塞探测结果。

[0018] 根据本发明提供的一种无丢失网络的拥塞探测方法，在所述将所述端口状态切换时刻差值和预设时长阈值进行比较判断之前，所述方法还包括：

[0019] 若无丢失网络采用基于缓存的逐跳流量控制技术，则所述预设时长阈值的公式为：

[0020] $$T_{max} = \frac{2(X_{off} - X_{on}) + \tau C}{2\varepsilon C} + \tau;$$

[0021] $$\tau = 2 * \text{link delay} + \frac{2 * MTU}{C};$$

[0022] 其中，T_{max}表示预设时长阈值，X_{off}表示触发PAUSE帧发送的预设端口入队列长度阈值，X_{on}表示触发RESUME帧发送的预设端口入队列长度阈值，C表示链路带宽，τ表示响应时间，ε表示固定参数，link delay表示链路延迟，MTU表示最大传输单元；

[0023] 若无丢失网络采用基于定时器的逐跳流量控制技术，则所述预设时长阈值为定时器的预设超时时间。

[0024] 根据本发明提供的一种无丢失网络的拥塞探测方法，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，包括：

[0025] 当所述端口状态切换时刻差值大于等于预设时长阈值，且当前时刻的交换机端口状态为非待定态时，

[0026] 若数据包队列长度大于拥塞队列长度阈值时，将所述交换机端口状态更新为拥塞态，并为数据包标记经历拥塞态，将数据包出队；

[0027] 若数据包队列长度小于等于所述拥塞队列长度阈值时，将所述交换机端口状态更新为非拥塞态，并将数据包出队。

[0028] 根据本发明提供的一种无丢失网络的拥塞探测方法，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

[0029] 当所述端口状态切换时刻差值大于等于所述预设时长阈值，且当前时刻的交换机

端口状态为待定态时，

[0030] 若数据包队列长度降低，且数据包队列长度大于拥塞队列长度阈值时，则将数据包出队；

[0031] 若数据包队列长度降低，且数据包队列长度小于等于所述拥塞队列长度阈值，则将所述交换机端口状态更新为非拥塞态，并将数据包出队；

[0032] 若数据包队伍长度未降低，且数据包队列长度大于所述拥塞队列长度阈值，则将所述交换机端口状态更新为拥塞态，并为数据包标记经历拥塞态，将数据包出队。

[0033] 根据本发明提供的一种无丢失网络的拥塞探测方法，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

[0034] 当所述端口状态切换时刻差值小于所述预设时长阈值，将所述交换机端口状态更新为待定态，若数据包未被标记为经历拥塞态，则为数据包标记经历待定态；

[0035] 若数据包已被标记为经历拥塞态，则将数据包出队。

[0036] 本发明还提供一种无丢失网络的拥塞探测系统，包括：

[0037] 端口切换处理模块，用于根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；

[0038] 拥塞探测模块，用于根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

[0039] 本发明还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现如上述任一种所述无丢失网络的拥塞探测方法的步骤。

[0040] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如上述任一种所述无丢失网络的拥塞探测方法的步骤。

[0041] 本发明提供的无丢失网络的拥塞探测方法及系统，通过检测交换机端口开启(ON)时期的时长和数据包队列长度演化特征，可准确探测识别无丢失网络中交换机端口的三种状态间转换，并为数据包进行正确标记，使得拥塞控制算法更加效率的完成调速决策，提高流完成时间和吞吐量。

附图说明

[0042] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0043] 图1为本发明提供的无丢失网络的拥塞探测方法的流程示意图；

[0044] 图2为本发明提供的三种状态的转换示意图；

[0045] 图3为本发明提供的仿真配置的网络拓扑示意图；

[0046] 图4为本发明提供的基于缓存的逐跳流控技术的两种场景下交换机出端口队列长度演变情况和标记情况的示意图；

[0047] 图5为本发明提供的基于定时器的逐跳流控技术的两种场景下交换机出端口队列长度演变情况和标记情况的示意图；

[0048] 图6为本发明提供的拥塞探测机制配合拥塞控制算法与传统拥塞控制算法在大规模仿真实验中的流完成速度的比较示意图；

[0049] 图7为本发明提供的无丢失网络的拥塞探测系统的结构示意图；

[0050] 图8为本发明提供的电子设备的结构示意图。

具体实施方式

[0051] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0052] 无丢失网络中，交换机需要准确探测与拥塞相关的状态，具体地，无丢失网络中交换机端口有三种状态：非拥塞态(0)：端口始终为开(ON)，没有队长累积(即数据包队列长度累积)；拥塞态(1)：端口始终为开(ON)，存在有队长累积，且队长累积只是由于拥塞引起的而不是由于端口被关闭产生的；待定态(/)：端口在开-关间(ON-OFF)切换。

[0053] 在无丢失网络中，逐跳流量控制技术在保证零丢包的同时，导致交换机端口状态在开-关(ON-OFF)间存在短时间内多次来回切换的状态，现有技术中忽视这种新状态，以致于端口的真实状态被错误地探测，未拥塞的数据包被错误标记，进而使得拥塞控制算法作出错误的调速决策，影响流完成时间和吞吐量。

[0054] 图1为本发明提供的无丢失网络的拥塞探测方法的流程示意图，如图1所示，本发明提供了一种无丢失网络的拥塞探测方法，包括：

[0055] 步骤101，根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；

[0056] 步骤102，根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

[0057] 在本发明中，将无丢失网络中的交换机作为执行主体进行说明。对无丢失网络中交换机端口的三种状态(非拥塞态、拥塞态和待定态)进行识别，其中，针对待定态的识别，是基于端口处于开-关(ON-OFF)状态下的特征，即，当交换机端口在结束OFF时期后的ON时期，其ON时期的时长通常较小。

[0058] 进一步地，首先，交换机记录下端口关闭状态结束时的时间戳t_{off}；当端口再次打开后，根据当前时间戳t与t_{off}之间的差值，即端口状态切换时刻差值，若该差值小于预设时长阈值T_{max}，则表示端口进入待定状态，此时将端口的状态更新为待定态；当前端口处于待定状态时，如果检测到当前时间戳t与t_{off}之间差值大于等于阈值T_{max}，则根据数据包队列长度是否存在增长，判断交换机是进入拥塞状态还是非拥塞状态。图2为本发明提供的三种状态的转换示意图，交换机中三种拥塞状态的切换可参考图2所示。

[0059] 具体地，在上述实施例的基础上，所述方法还包括：

[0060] 若无丢失网络采用基于缓存的逐跳流量控制技术，则所述端口关闭状态结束时间戳为上一次接收到RESUME帧的时间戳；

[0061] 若无丢失网络采用基于定时器的逐跳流量控制技术，则所述端口关闭状态结束时间戳为上一次信用值为零的状态下，接收到新的信用值时的时间戳，所述信用值是通过信

用限制值和交换机已发送数据块数量之间的差值获取得到的，所述信用限制值为下游交换机当前已分配缓存量和已接收到数据块之和。

[0062] 需要说明的是，对于基于缓存的逐跳流控技术，t_{off}为交换机上一次接收到RESUME帧的时间，具体地，在收到PAUSE帧后，端口被关闭，直到收到下一个RESUME帧时，端口关闭状态结束，端口重新开启，才可以传送数据包，此时的时间戳为端口关闭状态结束时间戳。端口开启后才能对数据包进行读取、标记和出队等操作，这样每个数据包的当前时间戳与端口关闭状态结束时间戳作差，得到的是端口目前为止处于开启状态的持续时间长度，进一步根据这个持续时间长度的大小，区分端口是一直打开的状态还是一段时间内开关多次切换的状态。

[0063] 对于基于定时器的逐跳流控技术，t_{off}为交换机在上一次零信用值的状态下，接收到新的信用值时的时间戳，在本发明中，信用值是通过下游交换机的FCCL和和上游交换机的FCTBS之间差值计算得到的，交换机端口获取可用的信用值：信用值＝FCCL-FCTBS，交换机每发送一个数据块，信用值减一，FCTBS加一，当信用值为零，不再发送数据。具体地，端口发送了部分数据包后导致信用值为零，端口此时被关闭，当交换机接收到新的信用更新值时，端口关闭状态结束，端口重新开启，此时的时间戳为端口关闭状态结束时间戳。

[0064] 本发明提供的无丢失网络的拥塞探测方法，通过检测交换机端口开启(ON)时期的时长和数据包队列长度演化特征，可准确探测识别无丢失网络中交换机端口的三种状态间转换，并为数据包进行正确标记，使得拥塞控制算法更加效率的完成调速决策，提高流完成时间和吞吐量。

[0065] 在上述实施例的基础上，所述根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果，包括：

[0066] 将所述端口状态切换时刻差值和预设时长阈值进行比较判断。

[0067] 在本发明中，首先计算当前时间戳和端口关闭状态结束时间戳之间的差值，然后将差值和预设时长阈值进行对比。

[0068] 根据判断结果和当前时刻的交换机端口状态，对所述交换机端口状态进行更新，并对数据包进行标记；

[0069] 所述交换机端口状态包括非拥塞态、拥塞态和待定态，其中，所述非拥塞态为端口开启且无数据包队列累积的状态，所述拥塞态为端口开启且存在数据包队列累积的状态，所述待定态为端口处于开启与关闭多次切换的状态；

[0070] 根据更新结果和标记结果，得到无丢失网络的拥塞探测结果。

[0071] 在本发明中，根据端口状态切换时刻差值和预设时长阈值的比较结果，以及交换机端口在当前时刻的状态，从而判断获知无丢失网络中交换机的拥塞状态，进一步地，对将交换机的拥塞态进行更新，并对数据包标记相应的状态标识。

[0072] 在上述实施例的基础上，在所述将所述端口状态切换时刻差值和预设时长阈值进行比较判断之前，所述方法还包括：

[0073] 若无丢失网络采用基于缓存的逐跳流量控制技术，则所述预设时长阈值的公式为：

[0074] $$T_{max}=\frac{2(X_{off}-X_{on})+\tau C}{2\varepsilon C}+\tau;$$

[0075] $$\tau = 2 * \text{link delay} + \frac{2 * MTU}{C};$$

[0076] 其中，T_{max}表示预设时长阈值；X_{off}表示触发PAUSE帧发送的预设端口入队列长度阈值，X_{on}表示触发RESUME帧发送的预设端口入队列长度阈值，在本实施例中，X_{off}和Xon之差设为两倍的最大传输单元MTU；C表示链路带宽，τ表示响应时间；ε表示固定参数，本实施例中，ε＝0.05；link delay表示链路延迟，MTU表示最大传输单元；

[0077] 若无丢失网络采用基于定时器的逐跳流量控制技术，则所述预设时长阈值为定时器的预设超时时间，在本实施例中，预设时长阈值为定时器的预设超时时间，即信用更新周期，具体地，定时器工作原理是预设一个超时时间后，当超时时长达到这个预设超时时间后触发某个事件。在本实施例中，下游交换机的定时器的超时时间就是信用更新周期，相当于每隔一个信用更新周期触发一次，向上游交换机发送FCCL消息，并重新设置下一次超时时间，其中，定时器的预设超时时间不超过65535符号时间。

[0078] 在上述实施例的基础上，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，包括：

[0079] 当所述端口状态切换时刻差值大于等于预设时长阈值，且当前时刻的交换机端口状态为非待定态时，

[0080] 若数据包队列长度大于拥塞队列长度阈值时，将所述交换机端口状态更新为拥塞态，并为数据包标记经历拥塞态，将数据包出队；

[0081] 若数据包队列长度小于等于所述拥塞队列长度阈值时，将所述交换机端口状态更新为非拥塞态，并将数据包出队。

[0082] 在上述实施例的基础上，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

[0083] 当所述端口状态切换时刻差值大于等于所述预设时长阈值，且当前时刻的交换机端口状态为待定态时，

[0084] 若数据包队列长度降低，且数据包队列长度大于拥塞队列长度阈值时，则将数据包出队；

[0085] 若数据包队列长度降低，且数据包队列长度小于等于所述拥塞队列长度阈值，则将所述交换机端口状态更新为非拥塞态，并将数据包出队；

[0086] 若数据包队伍长度未降低，且数据包队列长度大于所述拥塞队列长度阈值，则将所述交换机端口状态更新为拥塞态，并为数据包标记经历拥塞态，将数据包出队。

[0087] 在上述实施例的基础上，所述根据判断结果和当前时刻的交换机端口状态，对交换机端口状态进行更新，并对数据包进行标记，还包括：

[0088] 当所述端口状态切换时刻差值小于所述预设时长阈值，将所述交换机端口状态更新为待定态，若数据包未被标记为经历拥塞态，则为数据包标记经历待定态；

[0089] 若数据包已被标记为经历拥塞态，则将数据包出队。

[0090] 在上述实施例的基础上，通过一实施例，对本发明提供的无丢失网络的拥塞探测方式进行整体流程说明，具体步骤如下：

[0091] S1，计算交换机当前时间戳t与最近一次的端口关闭状态结束时间戳t_{off}之间的差值T_d，若T_d>＝T_{max}且端口当前状态为非待定态(即处于非拥塞态或拥塞态)，执行步骤S2；若

T_d>=T_{max}且端口当前状态为待定态，执行步骤S3；若T_d<T_{max}，将端口的状态更新为待定态，并执行步骤S4；

[0092] S2，如果此时交换机中数据包队列长度超过拥塞队列长度阈值q_{len}，将端口当前状态更新为拥塞态，并执行步骤S6，需要说明的是，在本实施例中，拥塞队列长度阈值用于判断端口是否是拥塞，通常该阈值较小，例如一般在5KB～200KB之间选取。而用于触发PAUSE帧和RESUME帧的X_{off}和X_{on}值，通常大于拥塞队列长度阈值，例如X_{off}=320KB或者512KB，并且X_{on}值与X_{off}值相差很小，一般设为X_{off}=X_{on}+2KB；如果此时交换机中数据包队列长度小于等于拥塞队列长度阈值q_{len}，将端口当前状态更新为非拥塞态，执行步骤S7；

[0093] S3，如果在最近一次的预设时长阈值T_{max}时间段内，数据包队列长度降低，且当前数据包队列长度高于拥塞队列长度阈值q_{len}，执行步骤S7；如果当前数据包队列长度小于等于拥塞队列长度阈值q_{len}，将端口当前状态更新为非拥塞态，执行步骤S7；如果在最近一次的预设时长阈值T_{max}时间段内，数据包队列长度未降低且队列长度大于拥塞队列长度阈值q_{len}，将端口当前状态更新为拥塞态，执行步骤S6；

[0094] S4，如果数据包未被标记为经历拥塞态，执行步骤S5；否则执行步骤S7；

[0095] S5，对数据包标记为经历待定态(Undetermined Encountered，简称UE)，至步骤S8；

[0096] S6，对数据包标记为经历拥塞态(Congestion Encountered，简称CE)；

[0097] S7，数据包出队；

[0098] S8，重复执行步骤S1至S7，直至交换机结束工作。

[0099] 在另一实施例中，通过仿真结果表明，本发明提供的无丢失网络的拥塞探测机制能正确标记交换机端口的三种状态。具体地，图3为本发明提供的仿真配置的网络拓扑示意图，可参考图3所示，在该仿真实验中，链路为40Gbps，传播延迟为4us，MTU=1000B。对于使用基于缓存的逐跳技术的网络，通过上述实施例的公式，计算得到T_{max}值为100.4us，交换机X_{off}=320KB。对于使用基于定时器的逐跳流控技术的网络，Tc值为40us，交换机每端口缓存280KB。如图3所示，发送端S1将长流F1发送到接收端R1。作为突发流量的发送端A0至A14将并发的64KB短流发送到R1。假设F1在仿真开始时达到40Gbps。从时间戳0开始，并发短流开始并持续约3ms。短流开始后，长流F0和F2以恒定速率同时发送到接收端R0。图4为本发明提供的基于缓存的逐跳流控技术的两种场景下交换机出端口队列长度演变情况和标记情况的示意图，图5为本发明提供的基于定时器的逐跳流控技术的两种场景下交换机出端口队列长度演变情况和标记情况的示意图，可参考图3、图4和图5所示：

[0100] 在单拥塞点场景下，端口P3是唯一的拥塞点。流F1在开始时速率为40Gbps。短流开始后，流F0和F2以恒定的5Gbps速率发送。当流F0和F2启动时，F1的速率已在发送端拥塞控制算法作用下降至15Gbps以下。此时，端口P2和P1，从待定状态转换为非拥塞状态。

[0101] 在多拥塞点场景下，端口P3和P2均是拥塞点。流F1在开始时速率为40Gbps。短流开始后，流F0和F2以恒定的25Gbps速率发送。当流F0和F2启动时，F1的速率已在发送端拥塞控制算法作用下降至15Gbps以下。此时，端口P2，从待定状态转换为拥塞状态；端口P1为待定状态。

[0102] 通过上述实施例的大规模仿真结果表明，配合本发明提供的三状态拥塞探测和标记机制，拥塞控制算法可得到更好的性能，加快流完成速度。对于拥塞流，激进减速；对于待

定流，保持其速率不变。在本实施例中，网络拓扑是三层胖树拓扑，250台服务器。链路速度40Gbps，传播时延4us，MTU＝1000B。图6为本发明提供的拥塞探测机制配合拥塞控制算法与传统拥塞控制算法在大规模仿真实验中的流完成速度的比较示意图，如图6所示，以基于缓存的逐跳流控技术为例，T_{max}值为100.4us。统计每条流的降速(slowdown)，即：

[0103] $$\text{slowdown} = \frac{\text{流实际完成时间}}{\text{流基准完成时间}}。$$

[0104] 在本实施例中，流的基准完成时间可根据已知的流大小除以链路速度提前计算得到。slowdown越小，越接近于1，表明流完成时间性能越好。可参考图6所示，显示了不同流大小区间的slowdown指标，包括99百分位的slowdown和50百分位的slowdown。DCQCN(Data Center Quantized Congestion Notification)是一种常用的拥塞协议，它的拥塞探测机制只根据队列长度判断是否拥塞，忽视了端口的待定态，会造成误判。采用了本发明提供的三态拥塞探测方法后，能够正确区分拥塞端口和待定端口，经过拥塞端口的拥塞流及时减速，只经过待定端口的流不进行不必要的减速，因此提高流完成速度。结果显示DCQCN+三态拥塞探测方法(采用空心标记的曲线)，99百分位的slowdown曲线和50分位的slowdown曲线都位于DCQCN曲线下方，性能都得到了显著提升。

[0105] 图7为本发明提供的无丢失网络的拥塞探测系统的结构示意图，如图7所示，本发明提供了一种无丢失网络的拥塞探测系统，包括端口切换处理模块701和拥塞探测模块702，其中，端口切换处理模块701用于根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；拥塞探测模块702用于根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

[0106] 本发明提供的无丢失网络的拥塞探测系统，通过检测交换机端口开启(ON)时期的时长和数据包队列长度演化特征，可准确探测识别无丢失网络中交换机端口的三种状态间转换，并为数据包进行正确标记，使得拥塞控制算法更加效率的完成调速决策，提高流完成时间和吞吐量。

[0107] 本发明提供的系统是用于执行上述各方法实施例的，具体流程和详细内容请参照上述实施例，此处不再赘述。

[0108] 图8为本发明提供的电子设备的结构示意图，如图8所示，该电子设备可以包括：处理器(processor)801、通信接口(CommunicationsInterface)802、存储器(memory)803和通信总线804，其中，处理器801，通信接口802，存储器803通过通信总线804完成相互间的通信。处理器801可以调用存储器803中的逻辑指令，以执行无丢失网络的拥塞探测方法，该方法包括：根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值；根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果。

[0109] 此外，上述的存储器803中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，

Read-OnlyMemory)、随机存取存储器(RAM,RandomAccessMemory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0110] 另一方面,本发明还提供一种计算机程序产品,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,计算机能够执行上述各方法所提供的无丢失网络的拥塞探测方法,该方法包括:根据当前时间戳和端口关闭状态结束时间戳,获取端口状态切换时刻差值;根据所述端口状态切换时刻差值和当前时刻的交换机端口状态,得到无丢失网络的拥塞探测结果。

[0111] 又一方面,本发明还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各实施例提供的无丢失网络的拥塞探测方法,该方法包括:根据当前时间戳和端口关闭状态结束时间戳,获取端口状态切换时刻差值;根据所述端口状态切换时刻差值和当前时刻的交换机端口状态,得到无丢失网络的拥塞探测结果。

[0112] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0113] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0114] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

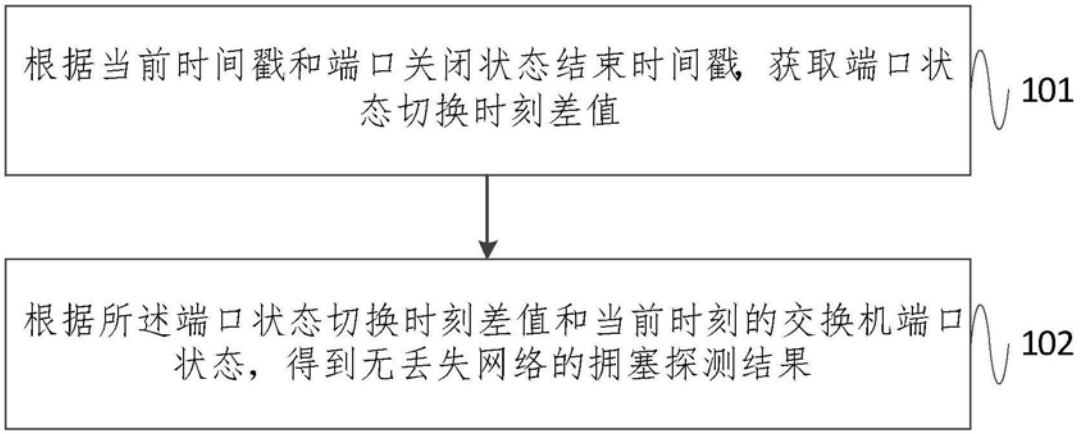

根据当前时间戳和端口关闭状态结束时间戳，获取端口状态切换时刻差值
101
根据所述端口状态切换时刻差值和当前时刻的交换机端口状态，得到无丢失网络的拥塞探测结果
102

图1

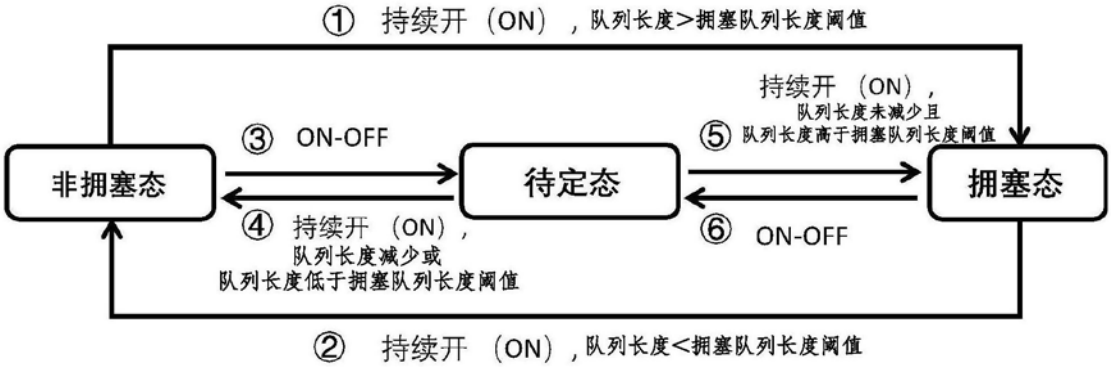

① 持续开（ON），队列长度>拥塞队列长度阈值
非拥塞态
③ ON-OFF
待定态
持续开（ON），
队列长度未减少且
⑤ 队列长度高于拥塞队列长度阈值
拥塞态
④ 持续开（ON），
队列长度减少或
队列长度低于拥塞队列长度阈值
⑥ ON-OFF
② 持续开（ON），队列长度<拥塞队列长度阈值

图2

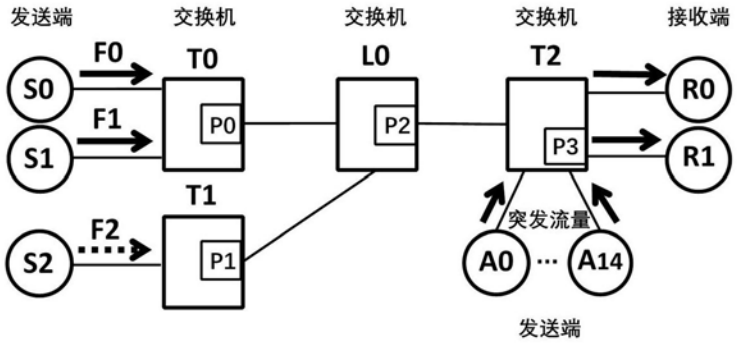

发送端
交换机
交换机
交换机
接收端
S0
S1
S2
F0
F1
F2
T0
T1
L0
T2
P0
P1
P2
P3
R0
R1
突发流量
A0
A14
发送端

图3

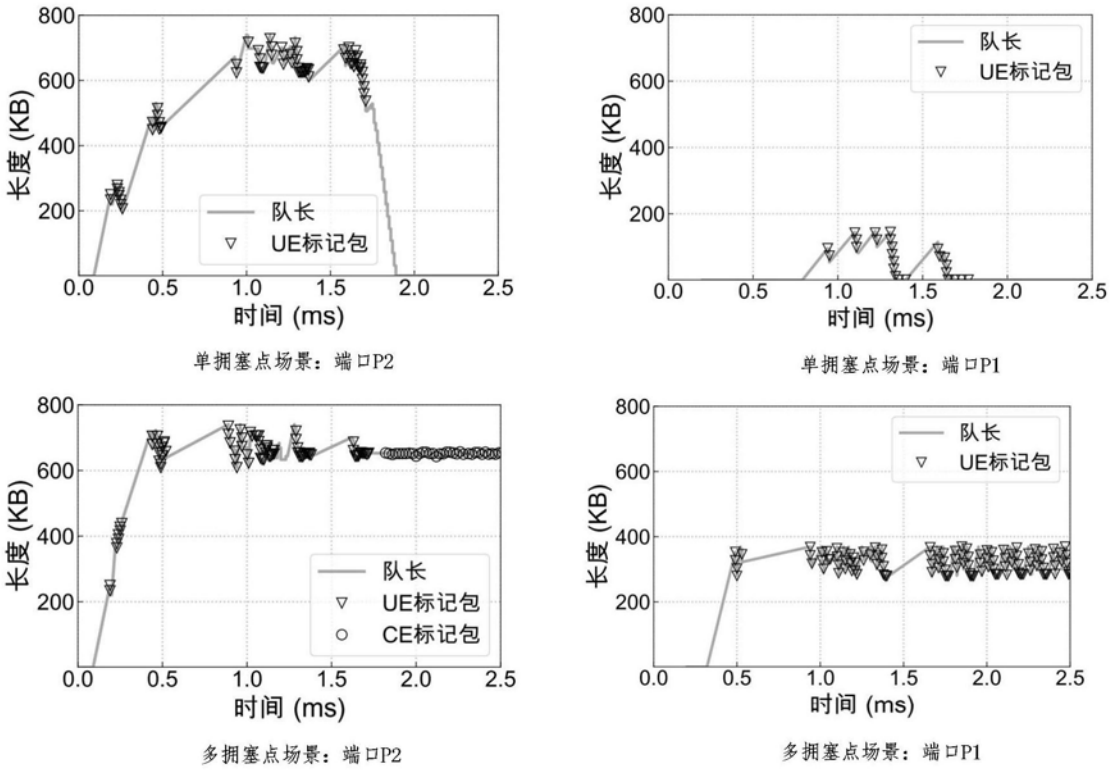

长度 (KB)
时间 (ms)
队长
UE标记包
CE标记包
单拥塞点场景：端口P2
单拥塞点场景：端口P1
多拥塞点场景：端口P2
多拥塞点场景：端口P1

图4

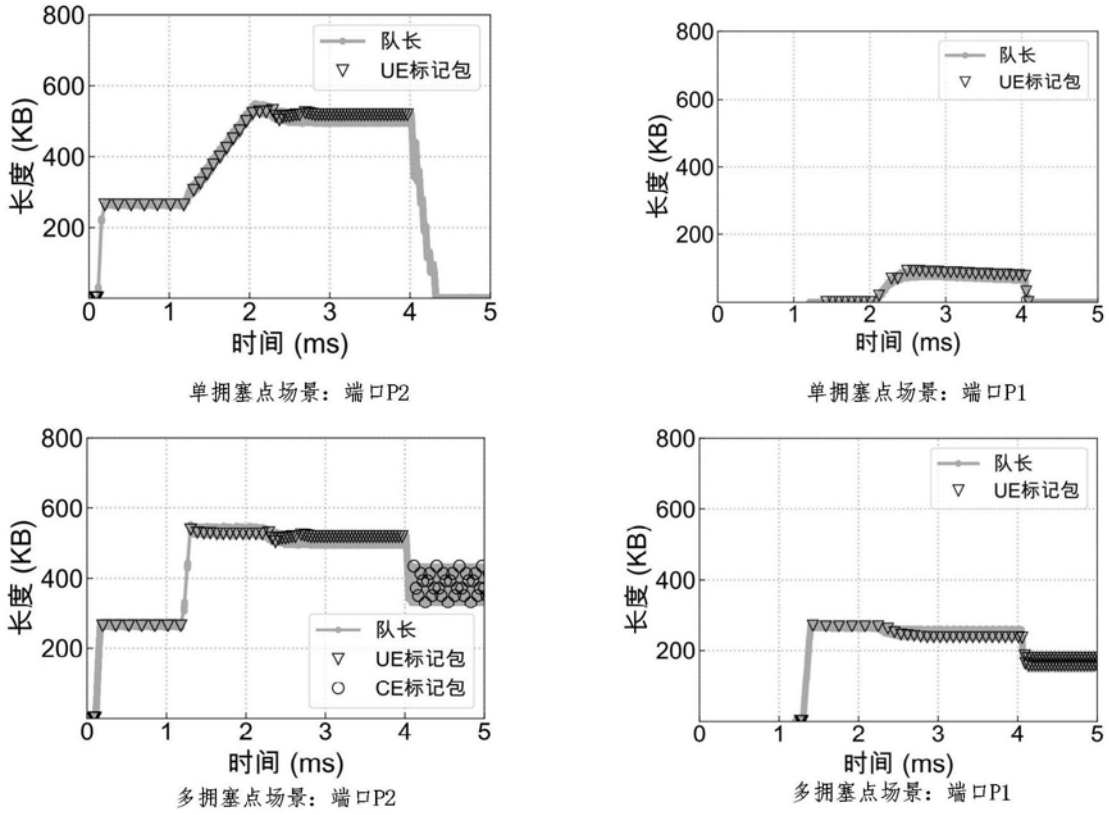

长度 (KB)
时间 (ms)
队长
UE标记包
CE标记包
单拥塞点场景：端口P2
单拥塞点场景：端口P1
多拥塞点场景：端口P2
多拥塞点场景：端口P1

图5

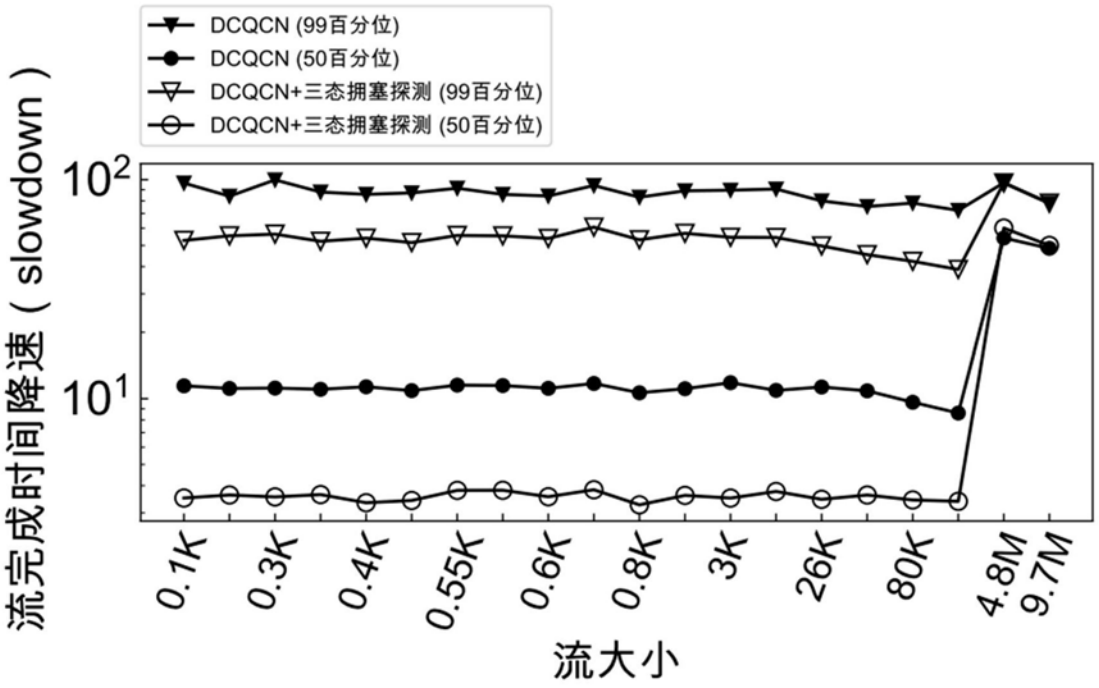

DCQCN (99百分位)
DCQCN (50百分位)
DCQCN+三态拥塞探测 (99百分位)
DCQCN+三态拥塞探测 (50百分位)
流完成时间降速（slowdown）
10^2
10^1
0.1K
0.3K
0.4K
0.55K
0.6K
0.8K
3K
26K
80K
4.8M
9.7M
流大小

图6

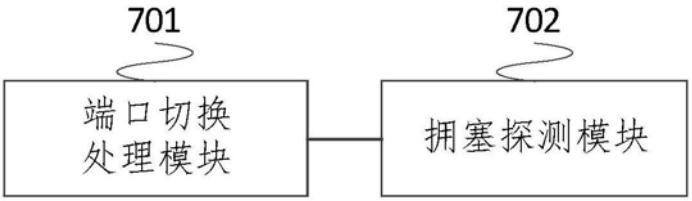

701
702
端口切换
处理模块
拥塞探测模块

图7

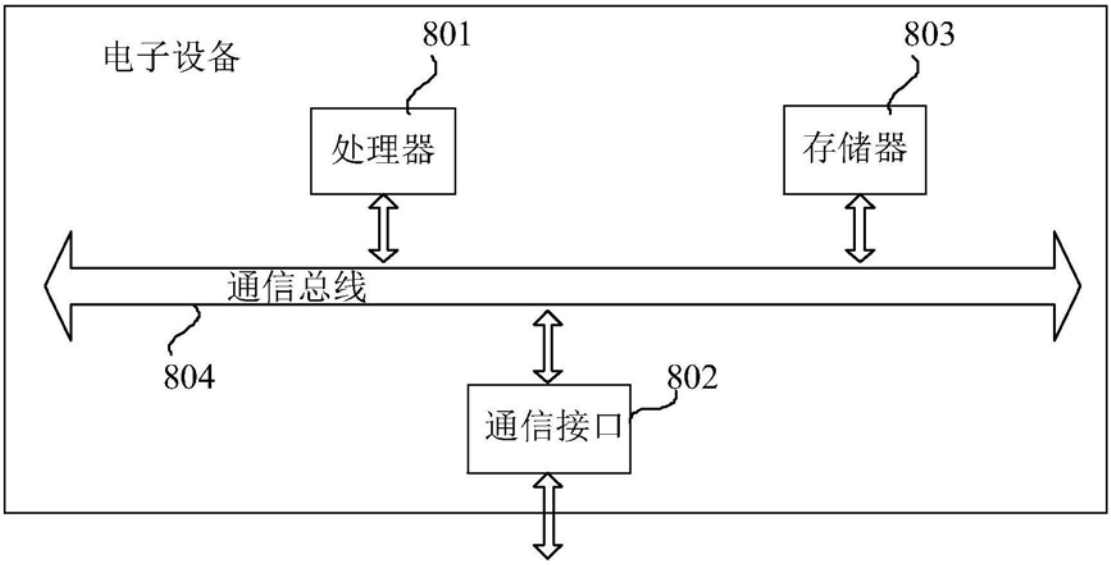


图8