(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 106031094 A
(43)申请公布日 2016.10.12

(21)申请号 201580009895.8

(22)申请日 2015.01.02

(30)优先权数据
61/922,990 2014.01.02 US

(85)PCT国际申请进入国家阶段日
2016.08.22

(86)PCT国际申请的申请数据
PCT/IB2015/050031 2015.01.02

(87)PCT国际申请的公布数据
WO2015/101952 EN 2015.07.09

(71)申请人 马维尔国际贸易有限公司
地址 巴巴多斯圣米加勒

(72)发明人 T·米兹拉希 Z·L·施米洛维希 G·纳沃恩

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅

(51)Int.Cl.
H04L 12/24(2006.01)
H04L 12/26(2006.01)

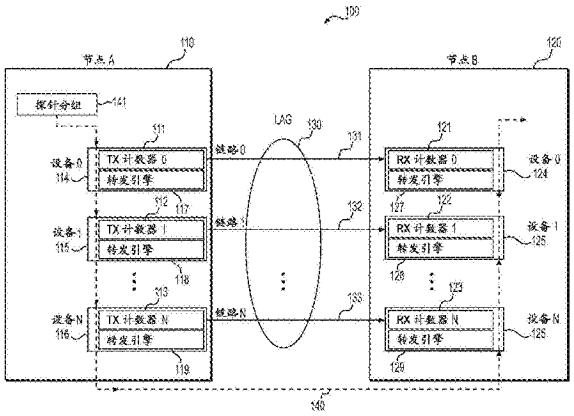
权利要求书2页 说明书10页 附图4页

(54)发明名称

分布式计数器的准确测量

(57)摘要

本公开的各方面提供了一种用于在具有多个分布式分组处理器的分组交换系统中收集分布式计数器值的方法。该方法包括在分组处理器接收探针分组，存储对应于由分组处理器处理的流的计数器值，用于随后向管理控制器递送，以及向下一分组处理器转发该探针分组。该下一分组处理器存储该下一分组处理器的计数器值，用于随后向管理控制器递送。



CN 106031094 A

1.一种方法,包括:

在具有多个分布式分组处理器的分组交换系统中的分组处理器处接收探针分组;

存储对应于由所述分组处理器处理的流的计数器值,以用于随后向管理控制器递送;以及

向所述分组交换系统中的下一分组处理器转发所述探针分组以存储所述下一分组处理器的计数器值,以用于随后向所述管理控制器递送。

2.根据权利要求1所述的方法,进一步包括:

在所述管理控制器处接收所述探针分组;以及

收集所述分组交换系统的所述多个分布式分组处理器的所述计数器值以确定针对所述分组交换系统的全局计数器值。

3.根据权利要求1所述的方法,进一步包括:

确定包括由所述管理控制器选择的分组处理器的转发路径;以及

在所述管理控制器处生成所述探针分组。

4.根据权利要求2所述的方法,进一步包括:

基于所述全局计数器值计算与所述分组交换系统有关的性能参数。

5.根据权利要求1所述的方法,其中存储所述计数器值包括:

将所述计数器值保存至所述探针分组的单独字段中。

6.根据权利要求1所述的方法,其中存储所述计数器值包括:

将所述计数器值聚合至所述探针分组的字段中。

7.根据权利要求1所述的方法,其中存储所述计数器值包括:

将所述计数器值写入每个相应分组处理器内的存储器;以及

响应于来自所述管理控制器的请求向所述管理控制器传输所述存储器的所述计数器值。

8.根据权利要求1所述的方法,进一步包括:将第一字段包括在所述探针分组中,所述第一字段将所述分组标识为探针分组,以及将第二字段包括在所述探针分组中,所述第二字段用于存储所述系统中的所述分组处理器的所述计数器值。

9.根据权利要求1所述的方法,进一步包括:对由所述分组处理器接收并传输的流的分组的数目进行计数,以获得所述计数器值。

10.根据权利要求1所述的方法,进一步包括:对由所述分组处理器接收并传输的流的字节的数目进行计数,以获得所述计数器值。

11.一种机架交换机,包括:

多个分布式分组处理器,每个分布式分组处理器包括维护对应于由所述分组处理器处理的流的计数器值的计数器;以及

管理控制器,所述管理控制器向所述交换机的所述多个分布式分组处理器传输探针分组,其中所述探针分组由第一分组处理器接收并且随后被向下一分组处理器转发,并且所述多个分布式分组处理器中接收所述探针分组的每个分布式分组处理器响应于接收所述探针分组而存储所述计数器值。

12.根据权利要求11所述的机架交换机,其中所述管理控制器进一步被配置为接收所述探针分组并且收集所述机架交换机的所述多个分布式分组处理器的所述计数器值以确

定针对所述机架交换机的全局计数器值。

13.根据权利要求11所述的机架交换机，其中所述管理控制器进一步被配置为确定包括所选择的分布式分组处理器的转发路径，并且生成所述探针分组。

14.根据权利要求12所述的机架交换机，其中所述管理控制器进一步被配置为基于所述全局计数器值计算与所述机架交换机有关的性能参数。

15.根据权利要求11所述的机架交换机，其中所述多个分布式分组处理器的每个分布式分组处理器通过将所述计数器值保存至所述探针分组的单独字段中来存储所述计数器值。

16.根据权利要求11所述的机架交换机，其中所述多个分布式分组处理器的每个分布式分组处理器通过将所述计数器值聚合至所述探针分组的字段中来存储所述计数器值。

17.根据权利要求11所述的机架交换机，其中所述多个分布式分组处理器的每个分布式分组处理器通过将所述计数器值写入每个相应分布式分组处理器内的存储器并且响应于来自所述管理控制器的请求向所述管理控制器传输所存储的值来存储所述计数器值。

18.根据权利要求11所述的机架交换机，其中所述管理控制器传输包括将所述分组标识为探针分组的第一字段和用于存储所述交换机中所述分布式分组处理器的所述计数器值的第二字段的所述探针分组。

19.根据权利要求11所述的机架交换机，其中所述分布式分组处理器被配置为对由所述分布式分组处理器接收并传输的所述流的分组的数目进行计数，以维护所述计数器值。

20.根据权利要求11所述的机架交换机，其中所述分布式分组处理器被配置为对由所述分布式分组处理器接收并传输的所述流的字节的数目进行计数，以维护所述计数器值。

分布式计数器的准确测量

[0001] 相关申请的交叉引用

[0002] 本公开要求2014年1月2日提交的、名称为“Accurate measurement of distributed counters”的第61/922,990号美国临时申请的权益，其全部内容通过引用并入于此。

背景技术

[0003] 本文所提供的背景技术描述是为了总体上给出本公开的背景。就该背景技术部分中所描述的工作的范围以及描述中在提交时并不以其他方式构成现有技术的方面而言，本发明人的工作既非明确也非隐含地被承认形成相对本公开的现有技术。

[0004] 分组交换系统中的分布式设备中存在若干类型的计数器，诸如基于机架的交换机或路由器。从那些分布式计数器获得准确的计数器值有助于监测操作。例如，帧丢失测量是由某些计算机网络操作和管理标准定义的机制，诸如标准ITU-T Y.1731、针对基于以太网网络的OAM功能和机制，其中OAM代表操作、管理和维护。方案通过使用节点中的传输和接收计数器对间隔上传输和接收的分组的数目进行计数来计算计算机网络中两个节点之间链路的帧丢失率。当若干链路被分组在一起作为链路聚合组(LAG)以形成不同分组处理器上的逻辑链路时，或者在多路径路由协议被部署的情况下，获得针对逻辑链路的传输和接收分组的准确计数器值是具有挑战的。

发明内容

[0005] 本公开的各方面提供了一种用于在具有多个分布式分组处理器的分组交换系统中收集分布式计数器值的方法。该方法包括在分组处理器接收探针分组，存储对应于由分组处理器处理的流的计数器值，用于随后向管理控制器递送，以及向下一分组处理器转发该探针分组。该下一分组处理器存储该下一分组处理器的计数器值，用于随后向管理控制器递送。该方法进一步包括在管理控制器接收探针分组，并且收集分组交换系统的多个分布式分组处理器的计数器值以确定针对分组交换系统的全局计数器值。该方法进一步包括确定包括由管理控制器选择的分组处理器的转发路径，并且在管理控制器生成探针分组。

[0006] 在一个实施例中，与分组交换系统有关的性能参数基于全局计数器值进行计算。

[0007] 在一个实施例中，分组处理器通过将计数器值保存至探针分组的单独字段来存储计数器值。在另一实施例中，分组处理器通过将计数器值聚合至探针分组的字段来存储计数器值。在又一实施例中，分组处理器将计数器值写入每个相应分布式分组处理器内的存储器，并且响应于来自管理控制器的请求向管理控制器传输存储器的计数器值。

[0008] 根据本公开的一方面，探针分组包括将分组标识为探针分组的第一字段，以及用于存储系统中分布式分组处理器的计数器值的第二字段。在一个实施例中，计数器对由所选择的分组处理器接收并传输的流的分组的数目进行计数，以获得计数器值。在另一实施例中，计数器对由所选择的分组处理器接收并传输的流的字节的数目进行计数，以获得计数器值。

[0009] 本公开的各方面提供了一种机架交换机。该交换机包括多个分布式分组处理器，其中每个分布式分组处理器包括维护对应于由分组处理器处理的流的计数器值的计数器。该交换机进一步包括管理控制器，该管理控制器向交换机的多个分布式分组处理器传输探针分组。该探针分组由第一分组处理器接收并且随后向下一分组处理器转发，并且多个分布式分组处理器中接收该探针分组的每个分布式分组处理器响应于接收该探针分组而存储计数器值。

[0010] 根据本公开的一方面，管理控制器进一步被配置为接收探针分组并且收集机架交换机的多个分布式分组处理器的计数器值以确定针对机架交换机的全局计数器值。

[0011] 根据本公开的一方面，管理控制器进一步被配置为确定包括所选择的分布式分组处理器的转发路径，并且生成探针分组。

[0012] 此外，管理控制器被配置为基于全局计数器值计算与机架交换机有关的性能参数。

附图说明

[0013] 被提议作为示例的本公开的各种实施例将参考以下附图进行详细描述，其中相似的标号表示相似的元件，其中：

[0014] 图1示出了根据本公开的实施例图示了计算机网络100的示图。

[0015] 图2示出了根据本公开的实施例图示了分组交换系统200的视图。

[0016] 图3A示出了根据本公开的实施例图示了由图2中探针分组处理器216执行的计数器值存储操作的第一选项技术300A的示图。

[0017] 图3B示出了根据本公开的实施例图示了计数器值存储操作的第二选项技术300B的示图。

[0018] 图3C示出了根据本公开的实施例图示了计数器值存储操作的第三选项技术300C的示图。

[0019] 图4示出了图示用于在图2中的分组交换系统200中的每个分布式分组处理器210-230处理探针分组的过程400的流程图。

[0020] 图5示出了图示用于在图2中的分组交换系统200中的管理控制器240生成并处理探针分组的过程500的流程图。

具体实施方式

[0021] 图1示出了根据本公开的实施例图示了计算机网络100的示图。在图1中，节点A 110和节点B 120是根据目的地地址和分组中承载的其他信息转发分组的分组交换系统。在一个示例中，分组交换系统是包括机架中安装的多个分布式线路卡或交换机模块的基于机架的交换机或路由器，并且每个线路卡或交换机模块包括多个分布式设备，诸如图1中所示的设备114-116和124-126。在一个示例中，多个分布式设备分别存在一个交换机模块中。在另一示例中，多个分布式设备被分发在多个交换机模块中。在一个实施例中，分布式设备是分组处理器。在各实施例中，分组处理器由通用处理器、专用集成电路(ASIC)、现场可编程门阵列(FPGA)或任意其他适当类型的硬件和软件实现。另外，分组处理器各种包括一个或多个转发引擎用于实现分组转发功能。例如，在图1中，分布式设备114-116和124-126各自

分别具有转发引擎117-119和127-129。

[0022] 在图1中，节点A和节点B的每个分布式设备包括不同类型的计数器用于不同用途。例如，取决于所使用计数器的类型，计数器将对通过分布式设备的特定分组流的分组或字节的数目进行计数，而另一类型的计数器用于对在分布式设备接收或传输的所有分组的分组或字节的数目进行计数。另外，可以存在对分布式设备处理的不同事件进行计数的不同计数器。每个计数器维护对应于分布式设备中处理的事件数目或流的分组或字节的数目的计数器值。在分组交换系统和网络中，分组流或流量流被定义为共享相同特性的分组序列。例如，分组流可以根据分组的源地址和目的地地址进行定义，并且具有相同源IP地址和目的地IP地址的分组属于相同分组流。在另一示例中，分组流根据与分组相关联的服务质量(QoS)的级别进行定义，并且与特定QoS级别相关联的分组在其报头中具有指示其需要的相应QoS级别的字段。通常，不同流的分组在分组交换系统中被不同地对待并处理。

[0023] 根据本公开的一方面，在一个实施例中，探针分组用于从分布式设备中收集计数器值。该探针分组使用分布式设备中包括的转发引擎的现有转发机制来生成并且转发通过分组交换系统中所有或选定子集的分布式设备。由探针分组触发，分布式设备将特定类型的计数器的计数器值并入探针分组或者将特定类型的计数器的计数器值保存到本地存储器用于稍后递送。用于使用探针分组来从分布式设备中收集计数器值的方法将在下文进一步详细描述。

[0024] 在图1中，节点A 110包括多个分布式设备114-116并且每个设备分别包括传输(Tx)计数器111-113和转发引擎117-119。类似地，节点B 120包括多个分布式设备124-126并且每个设备分别包括接收(Rx)计数器121-123和转发引擎127-129。当然应当理解，各实施例包括节点A或节点B中的任意数目的分布式设备。同样，分布式设备的备选实施例包括传输计数器和接收计数器两者。节点A和节点B由多个链路131-133连接。链路131从设备114开始并且在设备124终止。Tx计数器111对传输至链路131的分组或字节进行计数，并且Rx计数器121对从链路131接收的分组或字节进行计数。类似地，链路132将设备115与设备125连接，并且Tx计数器112和Rx计数器122对分别经由链路132传输和接收的分组或字节进行计数；链路133将设备116与设备126连接，并且Tx计数器113和Rx计数器123对分别经由链路133传输和接收的分组或字节进行计数。术语“分组”和“帧”通常分别与联网领域中的层3因特网协议(IP)网络和层2因特网网络一起使用，但是为了易于在此具体实施方式中描述的上下文中进行解释其可互换使用。如图所示，链路131-133被组合形成LAG130，从而分布式设备114-116和124-126与LAG 130相关联。类似地，计数器111-113和121-123还与LAG 130相关联。

[0025] 为了易于解释，在图1示例中，链路131-133被描述为单向链路，并且在此示例实施例中仅一个链路与每个设备耦合。然而，在其他实施例中，链路131-133是双工双向的，并且存在多个链路与每个设备114-116和124-126耦合。

[0026] 链路是各种物理类型。例如，链路的介质可以是非屏蔽双绞线，以及单模或多模光纤，并且协议可以是以太网、SONET(同步光纤网)和SDH(同步数字层级)。LAG 130可以通过各种方案进行配置，例如，其可以是由标准链路聚合控制协议(LACP)定义的LAG、将与模块化交换机或路由器中的不同交换机模块相关联的链路组合的基于交换机模块的LAG、或者将与不同交换机或路由器机架相关联的链路组合的基于机架的LAG。

[0027] 在一个实施例中，为了进行计算机网络100的性能监测，计算作为逻辑链路的LAG 130的帧丢失率参数。链路的帧丢失率通常被定义为帧丢失的数目与传输通过链路的帧的数目的比率。计数器在与链路耦合的两个节点处使用以对传输和接收的分组进行计数，并且计数器值通过周期性传输帧丢失测量分组在节点之间递送。丢失帧的数目通过比较经由链路传输的帧的数目与接收的帧的数目之间的差异来获得。

[0028] 在一个实施例中，作为逻辑链路的LAG 130具有与LAG 130相关联的多个Tx计数器111-113和多个Rx计数器121-123。例如，Tx计数器111和Rx计数器121与链路131以及分别经由该链路131传输和接收的计数分组相关联，而Tx计数器112和Rx计数器122与链路132以及分别经由该链路132传输和接收的计数分组相关联。类似地，Tx计数器113和Rx计数器123与链路133相关联。另外，计数器分别被分发在节点A和节点B的多个分布式设备114-116和124-126上。为了计算LAG 130的帧丢失率，需要获得多个分布式Tx计数器的第一总值和多个分布式Rx计数器的第二总值。第一总值是在帧丢失测量分组从节点A传输时捕获的分布式Tx计数器111-113的计数器值的总和。类似地，第二总值是在帧丢失测量分组到达节点B时捕获的分布式Rx计数器121-123的计数器值的总和。

[0029] 如图1中所示，在一个实施例中为了准确记录分布式计数器111-113和121-123的值，探针分组141被使用。在此实施例中，探针分组141执行将收集的计数器值从节点A递送到节点B的帧丢失测量分组的功能。在分组141被生成之前，节点A和节点B中分布式设备的子集被选择并且通过选定分布式设备的顺序被确定以创建针对探针分组的转发路径。另外，待被探测的计数器的类型也被确定。在探针分组被生成时，关于转发路径和待被探测的计数器的类型的信息被并入该探针分组。例如，为了计算LAG 130的帧丢失率，分布式设备114-116和124-126被选择包括在转发路径中，因为其与LAG 130相关联，并且Tx计数器111-113和Rx计数器121-123被确定进行探测，因为其正在对经由LAG 130传输的分组进行计数。接下来，探针分组141沿通过所有选定设备114-116和124-126的路由140进行发送。沿着路由140，探针分组141使得选定设备通过将计数器值并入分组或者将计数器值写入本地寄存器或相应设备中的其他类型的存储器来存储相应分布式设备的传输和接收计数器值用于稍后收集。

[0030] 具体地，在节点A侧，探针分组141顺序通过分布式设备114-116。当探针分组141到达分布式设备(例如，设备114)时，相应Tx计数器111的计数器值响应于接收探针分组141由设备114并入探针分组141，继而该分组由相应设备114中的转发引擎117转发至下一分布式设备115。通过此方式，在探针分组141通过节点A中的所有分布式设备114-116之后，其通过与最后一个分布式设备116相关联的链路133被传输到承载所存储的分布式Tx计数器114-116的计数器值的节点B。

[0031] 在节点B侧，类似于在节点A侧实现的过程，探针分组141在设备126接收，继而通过分布式设备126-124。在一个实施例中，当探针分组141通过每个设备时，相应Rx计数器123-121的计数器值响应于接收探针分组141由每个设备并入探针分组141。备选地，在不同于节点A中过程的另一实施例中，Rx计数器值被捕获并暂时写入相应分布式设备的本地存储器用于稍后收集。在路由140的末端，当探针分组141离开设备124时，所有分布式计数器值已经被并入探针分组141，使得其可以被递送到目的地，其中所述值可以被制成表，并且在一个实施例中，用于计算逻辑链路LAG 30的性能参数(诸如例如帧丢失率)。备选地，在节点B

侧，分布式设备124-126中存储的分布式Rx计数器值可以被稍后获取用于性能参数计算。

[0032] 注意，在上述过程期间，作为交换机或路由器的节点A的转发机制以线路速度操作，该线路速度对于探针分组足够快以立即捕获多个分布式计数器值，由此一致的计数器值被获得用于计算LAG 130的帧丢失率。

[0033] 还应当注意，上述方法被应用于LAG无关的设置。例如，在云计算领域中的数据中心网络中，存在若干服务器或交换机各自具有一个或多个计数器用于不同目的。在此情景下，为了收集分布式计数器值，使用基于分组转发决策通过每个设备的探针分组的方法比使用中央实体逐一获取计数器值的方法更有效。在其中多路径协议(诸如等价多路径(ECMP)路由协议)被部署在分组交换网络中的另一示例中，相同流的分组经由分组交换网络中的不同路由到达分组交换系统(诸如机架路由器)或者从分组交换系统(诸如机架路由器)离开。在此情景中，分组经由与不同交换机模块相关联的不同端口进入或离开分组交换系统，并且由分发在不同交换机模块的分组处理器进行处理。因此，对相同流的字节或分组进行计数的计数器被分发在分组交换系统中。使用探针分组来收集分布式计数器值的方法也可应用于此示例。

[0034] 图2示出了根据本公开的实施例图示了分组交换系统200的视图。捕获分布式设备的分布式计数器值的示例将参考图2更详细地描述。如图所示，分组交换系统200(诸如机架交换机或路由器)包括第一组多个分布式分组处理器210-230(各自包括多个计数器)，第二组多个分布式分组处理器281-283，管理控制器240以及交换结构250。

[0035] 在一个实施例中，第一组分布式分组处理器被配置为与LAG 270相关联。为了计算LAG 270的帧丢失参数，多个分布式分组处理器(诸如分组处理器210-230)的计数器值需要在LAG 270的性能监测过程期间被收集。同样如图2所示，在功能和结构方面，第二组分组处理器与第一组分组处理器类似。然而，第二组分组处理器保持分离，因此不与LAG 270相关联。第二组分组处理器的内部结构和操作与第一组类似，当为了描述的清楚而没有在图2中示出。

[0036] 在图2中，交换结构250与第一组多个分布式分组处理器210-230和第二组多个分布式分组处理器281-283耦合，并且在一个实施例中，其提供分组交换系统200中支持结构的分组处理器之间的高速分组传输信道。交换结构250还与管理控制器240耦合，从而向管理控制器240提供信道以与分布式分组处理器210-230和281-283进行通信。交换结构250可以包括形成结构网络的多个交换结构。交换结构通常以交换结构芯片的形式实现。在某些实施例中，某些分组处理器是不支持结构的。在此情况下，交换机总线用于提供不支持结构的分组处理器之间的连接。另外，交换机总线用于经由交换结构与交换机总线之间的接口来提供不支持结构的分组处理器与支持结构的分组处理器之间的连接。

[0037] 在图2中，管理控制器240生成探针分组，诸如探针分组261。备选地，作为与LAG 270耦合的远程系统(未示出)的一部分的管理控制器也生成探针分组，诸如探针分组262。探针分组承载转发路径信息。为了确定转发路径，管理控制器240选择所有分布式设备或其子集，并且在一个实施例中决定通过选定分布式设备的顺序。例如，在图2所示LAG 270的示例中，分组交换系统200中的分布式分组处理器210-230被选择进行探测。类似地，远程系统中与LAG 270耦合的分布式分组处理器也被选择。同时，分组处理器210被确定为待通过的第一选定设备，分组处理器220被确定为待通过的第二选定设备以此类推。虽然未示出，探

针分组261也能够通过不与LAG相关联的设备281-283。

[0038] 另外，管理控制器240接收探针分组，诸如探针分组262。所接收到的探针分组承载分布式分组处理器的计数器值。此外，管理控制器240获取每个分布式分组处理器210-230的存储器中存储的计数器值作为对接收探针分组的响应。基于所接收到的探针分组中承载的或者从存储器中获取的计数器值，管理控制器240确定全局计数器值。在计算帧丢失率参数的示例中，管理控制器基于全局计数器值计算针对LAG 270的帧丢失率。全局计数器值可以以相应计数器值的总值的形式或者以包括每个收集的计数器值的列表的形式。管理控制器240通常存在于分组交换系统200的控制平面中并且以软件或硬件实现。例如，其是控制平面的中央处理单元(CPU)中运行的软件，或者设计用于执行管理控制器240的各功能的电路。

[0039] 探针分组通常具有与分组交换系统200中处理的常规分组相似的格式。然而，其可以基于某些特定的字段以便探针分组的可识别以及执行探针分组的各功能。例如，探针分组在其报头中具有指示其是探针分组的字段或标识符。

[0040] 另外，探针分组可以在其净荷中具有一个或多个字段用于在其通过分布式分组处理器时并入分布式计数器值。所述字段可以在探针分组在分布式分组处理器中处理时该探针分组被生成或附接时预先保留。例如，从远程系统传输的探针分组262承载有其与LAG 270耦合的远程系统中分布式分组处理器的净荷计数器值。

[0041] 此外，探针分组还可以在其报头中具有指示待被探测的目标计数器的类型的字段。例如，在图2示例中的LAG应用中，待处理的计数器是传输计数器或接收计数器，诸如分布式设备210-230中的Rx计数器或Tx计数器。在其他示例中，目标计数器是对特定分组流的字节或分组计数的计数器。因此，探针分组报头中的字段用于指示目标计数器的类型。

[0042] 如上文所述，在各实施例中，探针分组还承载转发路径的信息，诸如哪个选定的分组处理器待被探针以及以什么顺序。例如，当探针分组由管理控制器生成时，分组处理器地址或标识的信息以特定顺序加载到探针分组的报头的特定字段。备选地，上述转发路径信息被配置到转发引擎而不是由探针分组承载。例如，地址信息被管理控制器使用与将转发表信息下载到分布式分组处理器类似的方案分发到转发引擎。在操作中，分组处理器中现有的转发机制可以用于基于转发路径信息向下一分布式设备转发探针分组。例如，由接收探针分组触发，每个分组处理器中的转发引擎可以通过使用本地信息或探针分组中承载的信息来选择下一分组处理器。

[0043] 在图2中，在一个实施例中，分组交换系统200具有与LAG 270相关联的多个相似的分布式分组处理器210-230。分组处理器210下面作为示例用于图示分布式分组处理器中的探针分组处理过程。

[0044] 如图所示，分组处理器210具有多个功能块，包括：转发引擎214、Rx计数器211、Tx计数器212、探针分组处理器216和存储器218。转发引擎214与Rx计数器211、Tx计数器212和探针分组处理器216耦合。另外，探针分组处理器216与Rx计数器211、Tx计数器212和存储器218耦合。在某些实施例中，除了Rx计数器211和Tx计数器212，还存在多个不同类型的计数器。例如，存在对特定流的字节或分组进行计数的各类计数器。在某些实施例中，一个分布式分组处理器中存在多个转发引擎。例如，在一个实施例中，存在用于处理经由入口端口从相关联链路271接收的分组的转发引擎以及用于处理经由出口端口待向相关联链路271传

输的分组的转发引擎。然而，为了易于解释，图2示例中的每个分组处理器210-230中仅使用了一个转发引擎。如图2中所示，与每个分组处理器连接的所有链路都是LAG270的成员。然而，在其他实施例中，可以存在仅这些链路的一部分被配置为LAG 270的成员。

[0045] 转发引擎214通常执行以下功能，诸如转发查找表、分组修改、分组分类和流量管理。在操作中，转发引擎214经由能够与各种物理类型的链路一起操作的物理设备接口(未示出)从多个链路271接收分组。转发引擎214经由交换结构250将接收到的分组转发到第二组设备281-283中包括的对应目的地分组处理器。目的地分组处理器根据分组中承载的地址信息进行确定。在此转发过程期间，从链路271接收的分组的数目可以由Rx计数器211进行计数，或者对应于特定分组流的分组或字节的数目可以由其他类型的计数器进行计数。类似地，转发引擎214经由交换结构250从第二组分组处理器281-283接收分组，并且向链路271转发所接收到的分组。同时，向链路271传输的分组的数目可以由Tx计数器212进行计数，或者对应于特定分组流传输的分组或字节的数目可以由其他类型的计数器进行计数。

[0046] 在操作中，探针分组(例如，探针分组261或探针分组262)到达分组处理器210。在此示例中，探针分组由转发引擎214首先处理。该探针分组到达经由不同路由从分组交换系统200的内部或外部到达分组处理器210。例如，探针分组261由管理控制器240生成并经由交换结构250传输到设备210。随后，分组处理器220经由交换结构250从分组处理器210接收探针分组。在另一示例中，探针分组262经由链路271从与LAG 270耦合的远程系统接收。在又一示例中，转发引擎214通过检查分组报头中的标识符来识别探针分组261或262，继而将探针分组261或262传向探针分组处理器216用于进一步处理。在其他实施例中，探针分组261或262被保存在存储模块(图2中未示出)中的存储器，并且仅分组报头中的相关信息被传向探针分组处理器216。

[0047] 在探针分组处理器216，当从转发引擎214接收探针分组261时，处理器216首先检查分组报头以确定多个类型的计数器中哪种类型的计数器将被处理。因此，Tx计数器212的值被确定将针对探针分组261进行存储。针对探针分组262被接收的情况，Rx计数器211的值被确定进行存储。继而，探针分组处理器216执行计数器值存储操作。在一个实施例中，探针分组处理器216例如通过将来自每个设备的计数器值保存在探针分组的单独字段或者通过将计数器值添加到被配置为包含聚合计数器值的专用字段来将目标计数器值写入探针分组用于立即递送。在另一实施例中，目标计数器值被写入本地存储器218(例如，寄存器)用于稍后响应于接收探针分组向对应的管理控制器递送。在计时器值被存储之后，无论通过适当地更新探针分组还是通过向寄存器写入计数器值，探针分组261或262被返回转发引擎214用于向下一分组处理器220转发。

[0048] 在某些实施例中，每个分布式分组处理器中可以存在不止一个Tx计数器或Rx计数器。在此情况下，在计数器值存储操作期间，探针分组处理器216首先分别聚合多个Tx计数器或Rx计数器的值，继而将所述值的总和并入探针分组261或者将所述值的总和写入本地存储器。备选地，多个计数器值被并入探针分组中的多个字段或者写入存储器而不需要聚合。

[0049] 探针分组处理器216通常可以以硬件实现并且以线路速度操作。例如，其可以由ASIC、FPGA或其他类型的适当集成电路实现。探针分组处理器216可以驻留在同一芯片上的转发引擎旁，或者其可以在另一单独集成电路上实现。

[0050] 在转发引擎214，基于转发路径信息，转发引擎214将探针分组经由交换结构250转发至下一设备(例如，设备220)。针对不同实施例，上文所使用的转发路径信息被承载在探针分组或被本地存储。当探针分组通过与LAG 270相关联的最后一个分布式分组处理器230时，取决于探针分组的源，该探针分组被转发到不同目的地。例如，针对由管理控制器240本地生成的探针分组261，在分布式设备210-230的Tx计数器的计数器值被并入探针分组261之后，其经由与设备230耦合的链路沿图2中所示的路由263被传输到LAG 270的远程系统。而对于来自远程系统的探针分组262，在分布式设备210-230的Rx计数器的计数器值被并入探针分组262或被写入本地存储器之后，其经由交换结构250沿路由264被传输到管理控制器240。在某些其他实施例中，探针分组可以由系统的管理控制器生成并且在分布式计数器值被收集或存储之后发送回到同一管理控制器。

[0051] 注意，在探针分组262到达分组交换系统200之前，远程系统中分布式分组处理器的Tx计数器的计数器值通过与上文所述的过程类似的过程被并入探针分组262。

[0052] 在管理控制器240，当探针分组(例如，探针分组262)被接收时，通过检查探针分组262中的特定字段，控制器240确定计数器值已经被并入探针分组262，或者处理器210-230的某些Rx计数器值被存储在分组处理器中的本地存储器中等待由控制器240获取。

[0053] 在第一情景中，计数器值已经通过将计数器值保存在对应于相应设备的探针分组的单独字段中或者通过将计数器值聚合到专用字段中而并入探针分组262。因此，管理控制器240从分组的对应字段收集值，并且获得远程系统的总传输计数器值以及分组交换系统200中分布式设备210-230的总接收计数器值。

[0054] 在第二情景中，分组处理器210-230的Rx计数器值被存储在本地存储器中。因此，管理控制器240向分布式分组处理器210-230发送请求以获得所存储的计数器值。分组处理器210-230随后经由管理控制器240与分组处理器210-230之间的通信过程向控制器240传输所存储的计数器值作为对管理控制器的请求的响应。此通信过程经由交换结构250或者经由分组处理器210-230与控制器240之间的其他通信信道(例如，针对基于机架的路由器或交换机中控制平面通信的以太网网络)引导。当控制器240从分布式分组处理器接收所存储的计数器值时，其获得分组交换系统200中分组处理器210-230的Rx计数器的总值。另外，控制器240通过收集探针分组262中承载的Tx计数器值来获得远程系统中Tx计数器的总值。因此，基于上述过程中收集的计数器值，控制器240可以计算LAG 270的分组丢失率参数。

[0055] 针对探针分组261，其被传输到远程系统，其中探针分组261通过与LAG 270相关联的分布式设备并到达与控制器240类似的管理控制器。与在控制器240发生的过程相似的过程在远程系统的控制器处发生。

[0056] 参考图2所述由探针分组处理器进行的计数器值存储操作现将参考图3A、图3B和图3C进一步进行描述。如下所述，存在若干技术用于执行存储操作。

[0057] 图3A示出了根据本公开的实施例图示了由图2中探针分组处理器216执行的计数器值存储操作的第一技术300A的示图。如图所示，管理控制器生成探针分组310A，其顺序通过设备0和设备1并且到达管理控制器1。该探针分组310A包括报头以及包括字段0和字段1的多个单独字段。当分组310A通过设备0时，设备0的计数器0的计数器值由设备0中的探针分组处理器(图3A中未示出)保存至分组310A的单独字段0。类似地，当分组310A随后通过设备1时，设备1的计数器1的计数器值被保存至分组310A的单独字段1。接下来，探针分组310A

被转发到管理控制器1。因此，在第一技术中，分布式计数器值出于存储的目的被集成到单独字段。在某些实施例中，每个分布式设备中可以存在多个计数器。针对此情况，备选集成操作可以将多个计数器值中的每个计数器值并入单独字段。

[0058] 图3B示出了根据本公开的实施例图示了计数器值存储操作的第二技术300B的示图。探针分组310沿与图3A所示相同路由通过相同设备。然而，分组310具有不同的结构。具体地，分组310除了报头还包括值总和字段。当分组310B通过设备0时，计数器0的计数器值由设备0中的探针分组处理器(未示出)聚合至值总和字段中的值总和。接下来，类似地，当分组310B通过设备1时，计数器1的计数器值被聚合至由探针分组310B承载的值总和。最后，分组310B到达管理控制器1。因此，不同于第一技术，分布式计数器值被聚合至探针分组310B中的专用字段以便收集针对设备0和设备1的总计数值。

[0059] 图3C示出了根据本公开的实施例图示了计数器值存储操作的第三技术300C的示图。在技术300C中，探针分组300C由管理控制器0生成，通过设备0和设备1，并且到达管理控制器1。设备、管理控制器以及探针分组310通过的路由与图1和图2相同。然而，存储器0和存储器1分别包括在设备0和设备1中。当探针分组310C通过设备0和设备1时，探针分组310C上不执行任何操作。相反，计数器0和计数器1中的计数器值由设备0和设备1中的相应分组处理器(未示出)分别存储至存储器0和存储器1中。因此，在第三技术中，计数器值通过将至存储至存储器中而捕获。作为第三技术过程的最后阶段，管理控制器1从分布式设备0和设备1的存储器获取所存储的值。例如，在一个实施例中，管理控制器1通过单独请求获取存储器的内容。备选地，管理控制器1使用探针分组310C后的第二探针分组(未示出)获取存储器的内容。

[0060] 注意，虽然仅两个分布式设备被用于图示上文参考图3A、图3B和图3C所述的计数器值存储操作的技术，但是所述方法的能力可以通过重复单个分布式设备内的操作容易地扩展到不止两个分布式设备的情景。

[0061] 图4示出了图示用于在图2中的分组交换系统200中的每个分布式分组处理器210-230处理探针分组的过程400的流程图。在探针分组210处的过程被用作示例。过程从S401开始并且向S410前进。

[0062] 在S410，探针分组从链路271或交换结构250之一在转发引擎214处接收，并且其被识别为探针分组并由转发引擎214转发到探针分组处理器216。

[0063] 在S420，探针分组处理器216基于探针分组的报头中承载的信息来确定从哪个计数器取得计数器值，继而使用先前所述的三个存储技术之一来存储计数器值。

[0064] 在S430，探针分组通过转发引擎继而基于由探针分组承载的转发路径信息向下一分布式分组处理器220转发。备选地，在探针分组通过最后一个选定的分布式分组处理器230之后，其被经由交换结构250转发至本地管理控制器240或者被转发至与LAG 270耦合的远程系统中的远程管理控制器。接下来，过程前进到S499并终止。

[0065] 图5示出了图示用于在图2中的分组交换系统200中的管理控制器240生成并处理探针分组的过程500的流程图。过程从S501开始并且向S510前进。

[0066] 在S510，管理控制器240选择分组交换系统200中待探测的分布式分组处理器的子集并且确定通过选定分组处理器的顺序，随后确定探针分组261的转发路径。同时，选定分布式分组处理器中的计数器类型(诸如Tx计数器)也被确定。

[0067] 在S520，探针分组261被生成从而承载转发路径和计数器类型的信息。

[0068] 在S530，从与LAG 270耦合的远程系统确定的转发路径中包括的最后一个选定分布式分组处理器转发的探针分组262在管理控制器240被接收。

[0069] 在S540，管理控制器240通过从所接收探针分组的对应字段直接读取所并入的计数器值或者通过向分布式设备发送请求以获取分布式分组处理器中本地存储器中所存储的计数器值来收集分布式计数器值。

[0070] 在S550，全局计数器值可以基于在S540收集的计数器值获得。随后，基于全局计数器值，管理控制器计算LAG 270的分组丢失率。接下来，过程前进到S599并终止。

[0071] 根据本公开的一方面，用于通过利用探针分组通过选定分布式设备来收集分布式计数器值的方法可以在云计算的上下文中在数据中心的交换机网络中使用，其中若干服务器和交换机经由交换机网络连接。在此情景中，控制器(例如，作为交换机网络的管理角色的单独计算机)通过以上文所述相似的方式使用探针分组来从服务器或交换机收集流量计量计数器值。在此示例中，对每个服务器或交换机中的字节计数的流量计量计数器被用于监测网络流量并且收集统计以便控制流量或者改变作为不同服务器的用户的顾客。

[0072] 虽然本公开的各方面已经结合提供作为示例的特定实施例进行了描述，但是也可以进行对示例的备选、修改和变体。因此，本文所述的实施例旨在图示说明而不是进行限制。在不脱离以下所述权利要求书的范围的前提下，可以进行各种修改。

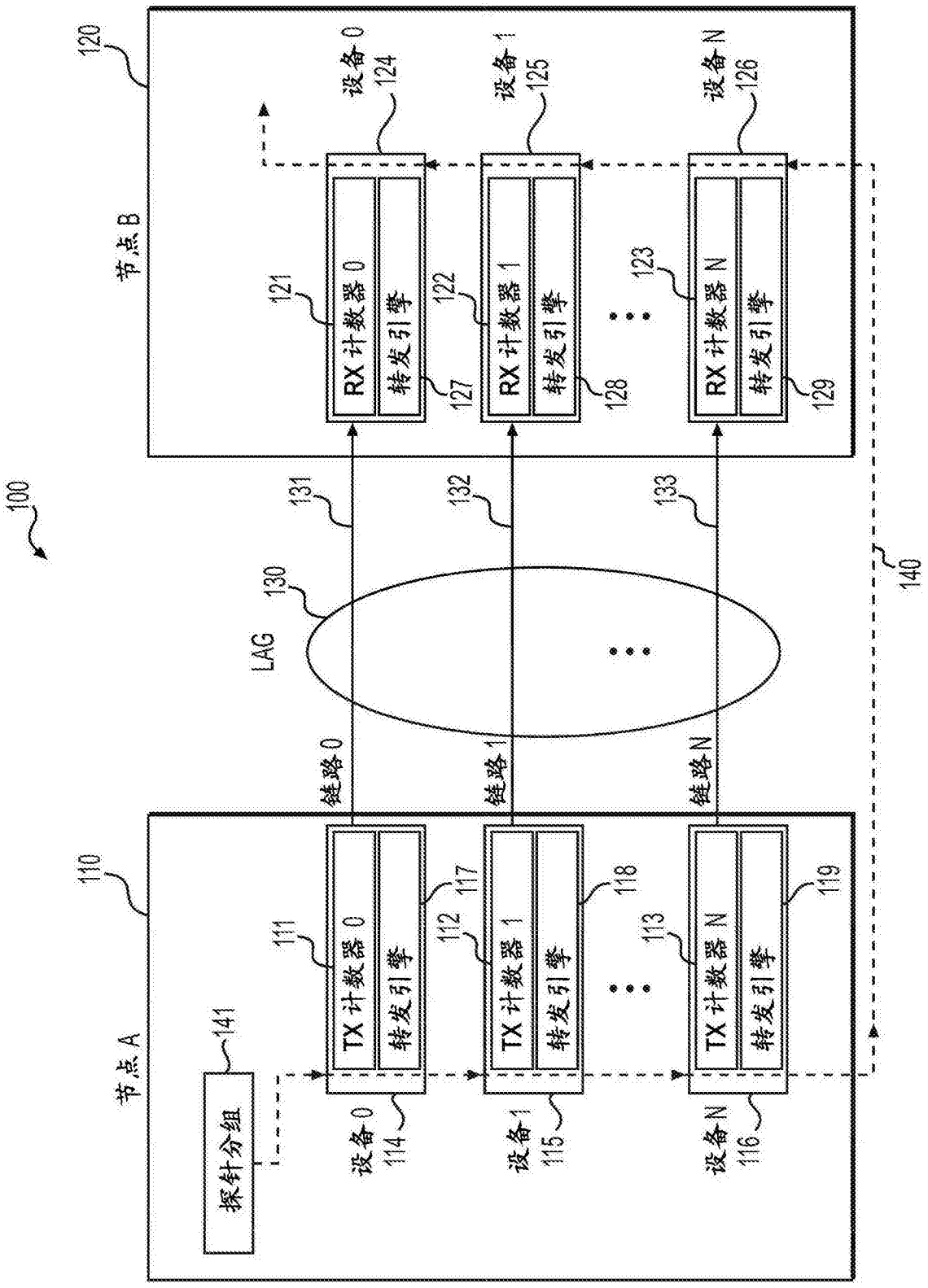

100
110
节点 A
120
节点 B
140
LAG
130
131
132
133
链路 0
链路 1
链路 N
TX 计数器 0
TX 计数器 1
TX 计数器 N
转发引擎
RX 计数器 0
RX 计数器 1
RX 计数器 N
111
112
113
117
118
119
121
122
123
127
128
129
设备 0
设备 1
设备 N
114
115
116
124
125
126
探针分组
141

图1

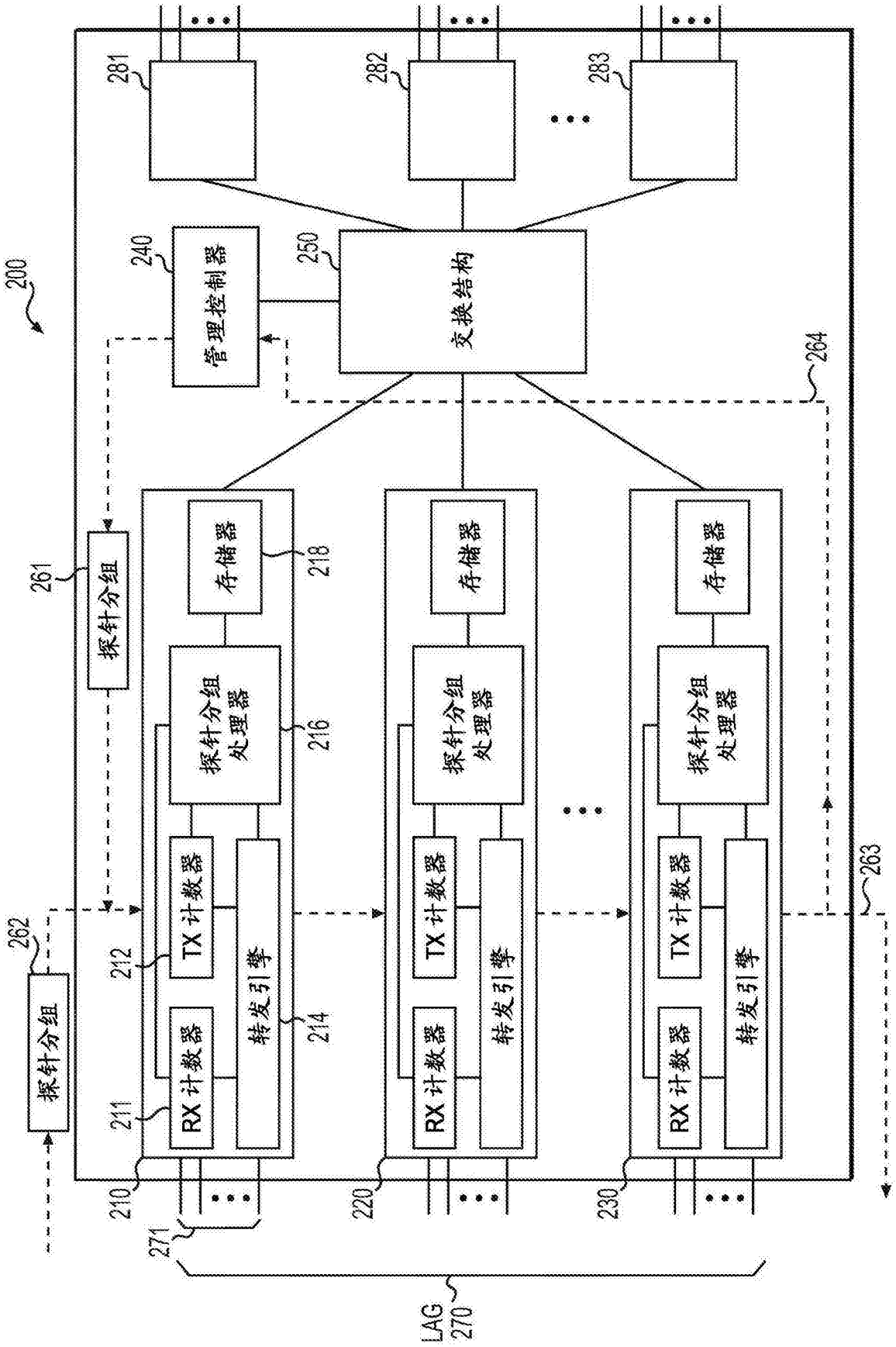

200
281
282
283
240
管理控制器
250
交换结构
264
261
探针分组
存储器
218
探针分组处理器
216
TX 计数器
212
RX 计数器
211
转发引擎
214
262
探针分组
263
210
220
230
271
LAG
270

图2

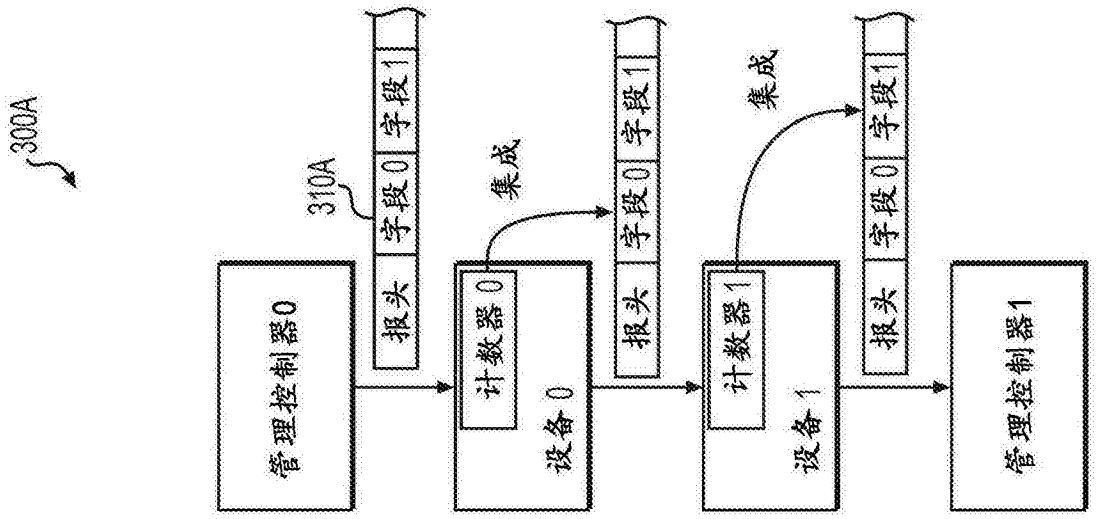

300A
310A
报头
字段0
字段1
集成
计数器0
设备0
计数器1
设备1
管理控制器0
管理控制器1

图3A

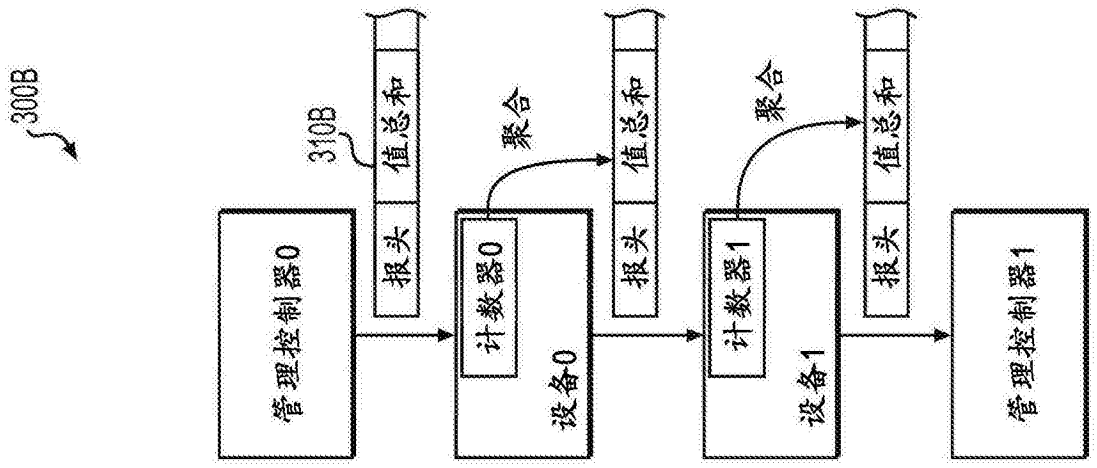

300B
310B
报头
值总和
聚合
计数器0
设备0
计数器1
设备1
管理控制器0
管理控制器1

图3B

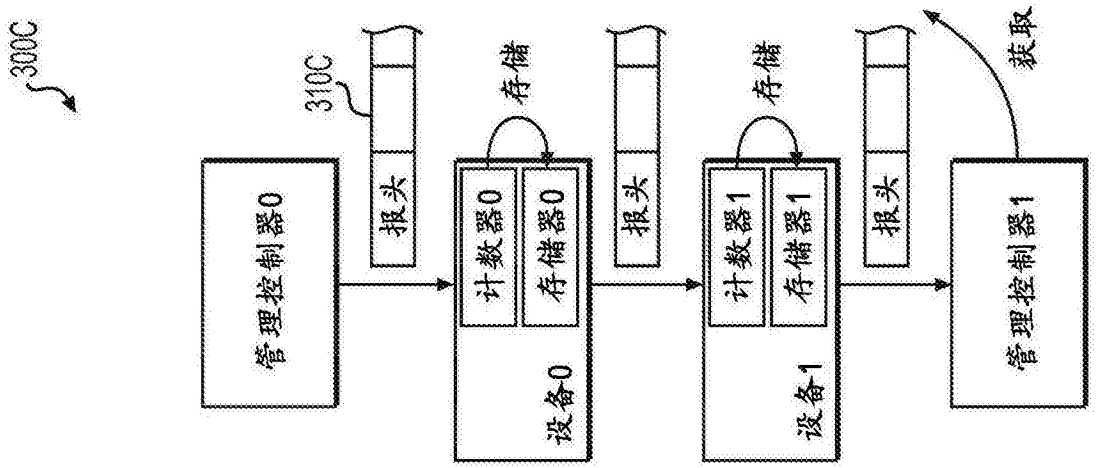

300C
310C
报头
存储
计数器0
存储器0
设备0
计数器1
存储器1
设备1
获取
管理控制器0
管理控制器1

图3C

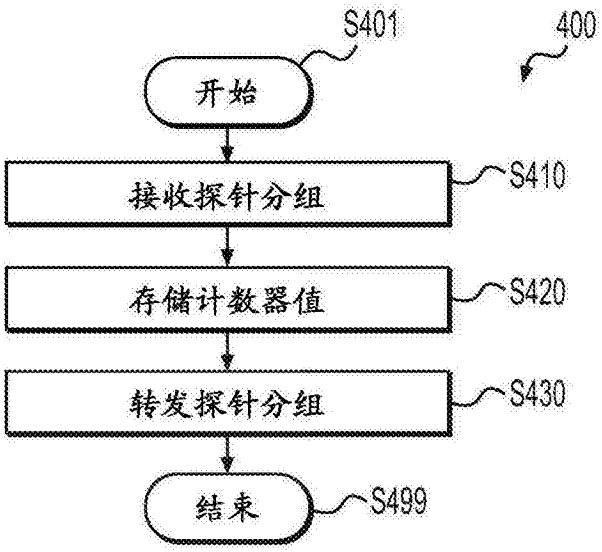

S401
400
开始
接收探针分组
S410
存储计数器值
S420
转发探针分组
S430
结束
S499

图4

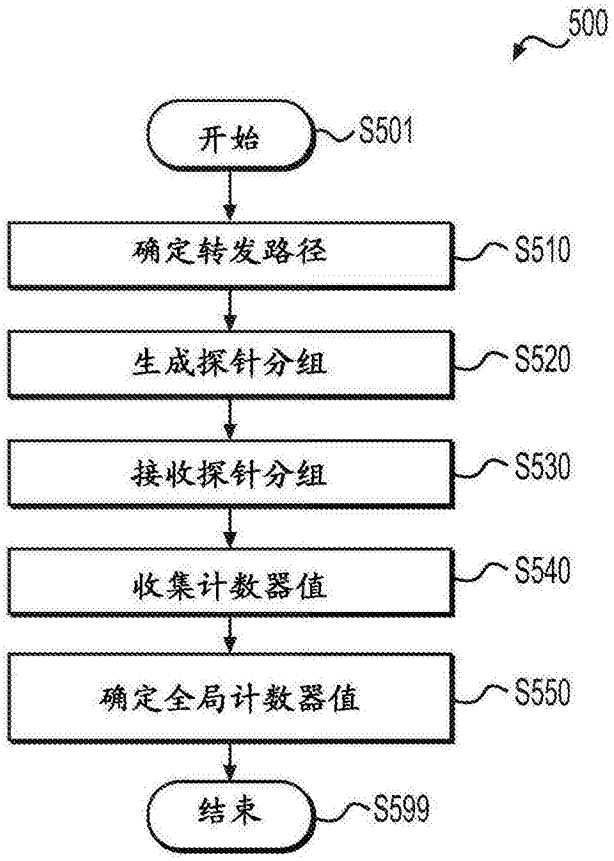

500
开始
S501
确定转发路径
S510
生成探针分组
S520
接收探针分组
S530
收集计数器值
S540
确定全局计数器值
S550
结束
S599

图5