



(12) 发明专利

(10) 授权公告号 CN 107005489 B

(45) 授权公告日 2021. 02. 02

(21) 申请号 201580062545.8
(22) 申请日 2015.12.02
(65) 同一申请的已公布的文献号
 申请公布号 CN 107005489 A
(43) 申请公布日 2017.08.01
(30) 优先权数据
 14/584,831 2014.12.29 US
 14/584,847 2014.12.29 US
(85) PCT国际申请进入国家阶段日
 2017.05.18
(86) PCT国际申请的申请数据
 PCT/US2015/063521 2015.12.02
(87) PCT国际申请的公布数据
 W02016/109105 EN 2016.07.07
(73) 专利权人 甲骨文国际公司
 地址 美国加利福尼亚
(72) 发明人 A·斯瑞尼瓦桑 C·卡斯蒂尔
(74) 专利代理机构 中国贸促会专利商标事务所
 有限公司 11038

(51) Int.Cl.
 H04L 12/861 (2013.01)
 H04L 12/863 (2013.01)
 H04L 12/931 (2013.01)
 H04L 12/933 (2013.01)
 H04L 12/935 (2013.01)
(56) 对比文件
 CN 101222422 A, 2008.07.16
 CN 1689283 A, 2005.10.26
 CN 102497310 A, 2012.06.13
 CN 101188556 A, 2008.05.28
 US 2008291825 A1, 2008.11.27
 US 2005036502 A1, 2005.02.17
 US 7391786 B1, 2008.06.24
 EP 2134037 A1, 2009.12.16
 US 2005076166 A1, 2005.04.07
 US 8391302 B1, 2013.03.05
 US 6611527 B1, 2003.08.26

审查员 朱文君

代理人 李晓芳

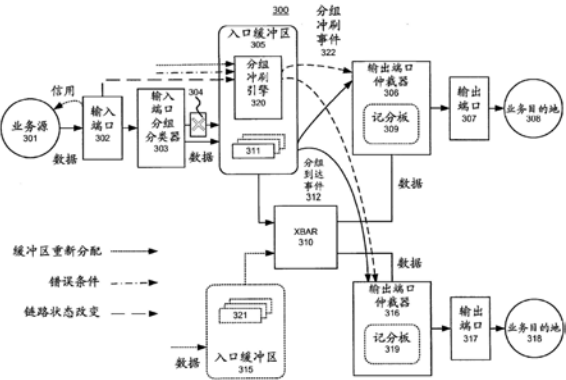
权利要求书4页 说明书10页 附图6页

(54) 发明名称

用于支持分组交换的系统、方法、介质和设备

(57) 摘要

系统和方法可以支持网络环境中的分组交换。系统可以包括联网设备上的入口缓冲区，其中包括一个或多个虚拟输出队列的入口缓冲区操作以存储在联网设备上的输入端口处接收到的一个或多个传入分组。此外，系统可以包括与入口缓冲区相关联的分组冲刷引擎，其中所述分组冲刷引擎操作以冲刷被存储在入口缓冲区中的所述虚拟输出队列中的分组，并且向一个或多个输出调度器通知分组被冲刷，其中每个输出调度器与输出端口相关联。



CN 107005489 B

1. 一种用于支持网络环境中的分组交换的系统,包括:

联网设备上的入口缓冲区;以及

入口缓冲区中的多个虚拟输出队列VOQ,其中所述多个VOQ包括共享存储器的资源并且操作以存储多个传入分组,所述多个传入分组在多个输入端口处被接收并且由所述入口缓冲区排队到所述多个VOQ中,其中所述入口缓冲区操作以通过针对被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组向所述多个输出端口的一个或多个调度器通告分组到达事件来向所述一个或多个调度器通知被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组;

多个输出调度器,其中每个输出调度器与所述多个输出端口中的一个输出端口相关联并且为所述多个VOQ中的一个或多个VOQ维护信用状态;以及

与所述入口缓冲区相关联的分组冲刷引擎,其中所述分组冲刷引擎操作以:

冲刷被存储在所述入口缓冲区中的所述虚拟输出队列中的第一选定分组,以及

通过通告分组冲刷事件来向所述多个输出调度器中的一个或多个输出调度器通知所述第一选定分组从所述多个VOQ中的所述VOQ被冲刷,所述分组冲刷事件将由所述多个输出调度器中的选定输出调度器用于实现所述信用状态以作为所述入口缓冲区和所述多个输出调度器之间的分组流的同步。

2. 根据权利要求1所述的系统,其中:

所述联网设备是网络交换机,所述网络交换机包括交叉开关结构。

3. 根据权利要求1或2所述的系统,其中:

所述多个输出调度器中的每个输出调度器操作以:

响应于所述入口缓冲区通告所述多个传入分组被排队到所述多个VOQ中,根据分组到达事件递增记分卡;

响应于所述多个传入分组中的分组从所述入口缓冲区被递送到输出端口,递减所述记分卡;以及

响应于由所述分组冲刷引擎通告所述第一选定分组的冲刷,根据所述分组冲刷事件递减所述记分卡。

4. 根据权利要求1或2所述的系统,其中:

每个所述输出调度器是输出端口仲裁器,其中所述输出端口仲裁器操作以:

调度所述输入端口以将分组转发到与所述输出端口仲裁器相关联的输出端口,以及更新与所述输出端口仲裁器相关联的记分板。

5. 根据权利要求1或2所述的系统,其中:

所述输出调度器操作以在从所述分组冲刷引擎接收到所述分组被冲刷的通知之后更新记分板。

6. 根据权利要求1或2所述的系统,其中:

所述分组冲刷引擎操作以使用分层进程从所述虚拟输出队列的头部开始遍历存储要被冲刷的分组的缓冲区链表。

7. 根据权利要求1或2所述的系统,其中:

所述入口缓冲区操作以释放与多播分组相关联的缓冲区,直到所述多播分组从所述多播分组的目的地列表中的每个虚拟输出队列出队。

8. 根据权利要求1或2所述的系统, 其中:

存储在所述入口缓冲区中的一个或多个分组被冲刷, 以处理所述输入端口处的链路状态改变。

9. 根据权利要求1或2所述的系统, 其中:

所述入口缓冲区中的每个虚拟输出队列与一个或多个定时器相关联, 其中当分组被插入到所述虚拟输出队列的头部中时, 所述一个或多个定时器开始。

10. 根据权利要求9所述的系统, 其中:

如果在所述定时器到期之前所述分组没有被调度以被发送到输出端口, 则所述分组被冲刷。

11. 一种用于支持联网设备中的分组交换的方法, 所述联网设备具有共享存储器、多个输入端口和多个输出端口, 所述方法包括:

在所述联网设备上提供入口缓冲区;

在所述入口缓冲区中维护多个虚拟输出队列VOQ, 其中每个VOQ包括所述共享存储器的资源;

利用与所述多个输出端口相关联的多个输出调度器为所述多个VOQ中的每个VOQ维护信用状态;

存储多个传入分组, 所述多个传入分组在输入端口处被接收并且由所述入口缓冲区排队到所述入口缓冲区的所述多个VOQ中;

由所述入口缓冲区通过针对被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组向所述多个输出端口的一个或多个调度器通告分组到达事件, 来向所述一个或多个调度器通知被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组;

经由与所述入口缓冲区相关联的分组冲刷引擎冲刷被存储在所述入口缓冲区中的所述多个VOQ中的VOQ中的第一选定分组; 以及

利用所述分组冲刷引擎通过通告分组冲刷事件来向所述多个输出调度器中的一个或多个输出调度器通知所述第一选定分组从所述多个VOQ中的所述VOQ被冲刷, 所述分组冲刷事件将由所述多个输出调度器中的选定输出调度器用于实现所述信用状态以作为所述入口缓冲区和所述多个输出调度器之间的分组流的同步。

12. 根据权利要求11所述的方法, 其中所述联网设备是网络交换机, 所述网络交换机包括交叉开关结构。

13. 根据权利要求11或12所述的方法, 还包括:

响应于所述入口缓冲区通告所述多个传入分组被排队到所述多个VOQ中, 根据分组到达事件递增输出调度器的记分卡;

响应于所述多个传入分组中的分组从所述入口缓冲区被递送到输出端口, 递减所述记分卡; 以及

响应于由所述分组冲刷引擎通告所述第一选定分组的冲刷, 根据所述分组冲刷事件递减所述记分卡。

14. 根据权利要求11或12所述的方法, 还包括:

允许每个所述输出调度器为输出端口仲裁器, 其中所述输出端口仲裁器操作以:

调度所述输入端口以将分组转发到与所述输出端口仲裁器相关联的输出端口, 以及

更新与所述输出端口仲裁器相关联的记分板。

15. 根据权利要求11或12所述的方法, 还包括:

经由所述输出调度器, 在从所述分组冲刷引擎接收到所述分组被冲刷的通知之后, 更新记分板。

16. 根据权利要求11或12所述的方法, 还包括:

经由所述分组冲刷引擎, 使用分层进程从所述虚拟输出队列的头部开始遍历存储要被冲刷的分组的缓冲区链表。

17. 根据权利要求11或12所述的方法, 还包括:

释放与多播分组相关联的缓冲区, 直到所述多播分组从所述多播分组的目的地列表中的每个虚拟输出队列出队。

18. 根据权利要求11或12所述的方法, 还包括:

冲刷被存储在所述入口缓冲区中的一个或多个分组, 以处理所述输入端口处的链路状态改变。

19. 根据权利要求11或12所述的方法, 还包括:

将一个或多个定时器与所述入口缓冲区中的每个虚拟输出队列相关联, 其中当分组被插入到所述虚拟输出队列的头部时, 所述一个或多个定时器开始, 以及如果在所述定时器到期之前所述分组没有被调度以被发送到输出端口, 则所述分组被冲刷。

20. 一种非暂态机器可读存储介质, 所述非暂态机器可读存储介质具有存储在其上的指令, 所述指令用于支持具有共享存储器、多个输入端口和多个输出端口的联网设备中的分组交换, 所述指令当被执行时, 使得系统执行步骤, 所述步骤包括:

在所述联网设备上提供入口缓冲区;

在所述入口缓冲区中维护多个虚拟输出队列VOQ, 其中每个VOQ包括所述共享存储器的资源;

利用与所述多个输出端口相关联的多个输出调度器为所述多个VOQ中的每个VOQ维护信用状态;

存储多个传入分组, 所述多个传入分组在输入端口处被接收并且由所述入口缓冲区排队到所述入口缓冲区的所述多个VOQ中;

由所述入口缓冲区通过针对被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组向所述多个输出端口的一个或多个调度器通告分组到达事件, 来向所述一个或多个调度器通知被排队到所述多个VOQ中的所述多个传入分组中的每个传入分组;

经由与所述入口缓冲区相关联的分组冲刷引擎冲刷被存储在所述入口缓冲区中的所述多个VOQ中的VOQ中的第一选定分组; 以及

利用所述分组冲刷引擎通过通告分组冲刷事件来向所述多个输出调度器中的一个或多个输出调度器通知所述第一选定分组从所述多个VOQ中的所述VOQ被冲刷, 所述分组冲刷事件将由所述多个输出调度器中的选定输出调度器用于实现所述信用状态以作为所述入口缓冲区和所述多个输出调度器之间的分组流的同步。

21. 一种非暂态机器可读存储介质, 所述非暂态机器可读存储介质具有存储在其上的指令, 所述指令用于支持具有共享存储器、多个输入端口和多个输出端口的联网设备中的

分组交换,所述指令当被执行时,使得系统执行如权利要求12至19中任一项所述的方法。

22.一种用于支持分组交换的设备,所述设备包括用于执行如权利要求11至19中任一项所述的方法的操作的装置。

用于支持分组交换的系统、方法、介质和设备

[0001] 版权声明：

[0002] 本专利文档的公开内容的一部分包含受版权保护的素材。版权拥有者不反对任何人对专利文档或专利公开内容按照在专利商标局的专利文件或记录中出现的那样进行传真复制，但是除此之外在任何情况下都保留所有版权权利。

技术领域

[0003] 本发明一般涉及计算机系统，并且尤其涉及云环境中的高性能系统。

背景技术

[0004] 随着更大的云计算架构被引入，与传统网络和存储相关联的性能和管理瓶颈已成为重要的问题。诸如Oracle工程系统之类的工程系统可以提供卓越的处理速度、明显更快的部署、用于深入分析的即时视图以及可管理的大数据能力。这是本发明的实施例旨在解决的总体领域。

发明内容

[0005] 本文描述了可以支持网络环境中的分组交换的系统和方法。系统可以包括联网设备上的入口缓冲区，其中包括一个或多个虚拟输出队列的入口缓冲区操作以存储在联网设备上的输入端口处接收到的一个或多个传入分组。此外，系统可以包括与入口缓冲区相关联的分组冲刷引擎，其中所述分组冲刷引擎操作以冲刷被存储在入口缓冲区中的所述虚拟输出队列中的分组，并且向一个或多个输出调度器通知分组被冲刷，其中每个输出调度器与输出端口相关联。此外，输入端口与多个源虚拟通道 (VL) 相关联。此外，入口缓冲区操作以将入口缓冲区中的缓冲区资源从与该多个源VL中的第一源VL相关联重新分配为与该多个源VL中的第二源VL相关联，并且将初始信用更新发送到用于第一源VL和第二源VL的输入端口。

附图说明

[0006] 图1示出了根据本发明的实施例的支持网络环境中的高性能系统的图示。

[0007] 图2示出了根据本发明的实施例的支持高性能系统中的网络交换机的图示。

[0008] 图3示出了根据本发明的实施例的管理高性能系统中的虚拟输出队列 (VOQ) 资源的图示。

[0009] 图4示出了根据本发明的实施例的用于管理高性能系统中的虚拟输出队列 (VOQ) 资源的示例性流程图。

[0010] 图5示出了根据本发明的实施例的支持高性能系统中的缓冲区重新分配的图示。

[0011] 图6示出了根据本发明的实施例的用于支持高性能系统中的缓冲区重新分配的示例性流程图。

具体实施方式

[0012] 在附图的图中通过示例的方式而非限制的方式示出本发明,在附图中相似的附图标记指示相似的元素。应当注意的是,在本公开中对“一个”或“一种”或“一些”实施例的引用不一定引用同一实施例,并且这样的引用意味着至少一个。

[0013] 以下本发明的描述使用InfiniBand (IB) 网络交换机作为高性能联网设备的示例。对于本领域技术人员将明显的是,可以使用其它类型的高性能联网设备而没有限制。

[0014] 本文描述了可以支持在诸如云环境之类的网络环境中的分组交换的系统和方法。

[0015] 高性能系统

[0016] 图1示出了根据本发明的实施例的支持网络环境中的高性能系统的图示。如图1所示,高性能系统100可以包括经由网络交换结构110互连的多个主机机器101-103 (或服务器)。

[0017] 高性能系统100中的网络交换结构110可以负责指引在各种主机机器101-103上运行的各种虚拟机 (VM) 111-113 (和/或虚拟化应用) 之间的业务移动。

[0018] 根据本发明的实施例,网络交换结构110可以基于InfiniBand (IB) 协议,IB协议可以管理对等 (peer-to-peer) 信用交换以及提供无损的端到端连接。因此,网络交换结构110中的各种联网设备可以在不同的条件下维持信用一致性以用于支持高性能系统100中的数据传送。

[0019] 此外,每个物理IB链路可以被划分成多个虚拟链路 (VL),以便为各种VM 111-113 (和/或应用) 之间的业务提供服务质量 (QoS)。例如,主机机器101-103之间的网络分组流120可以表示不同VM 111-113和应用可能期望的不同服务的聚合。此外,在不同的源和目的地对之间的聚合网络管道内发送的各个分组流120可以满足不同的服务要求 (或甚至冲突的服务要求)。

[0020] InfiniBand (IB) 网络交换机

[0021] 图2示出了根据本发明的实施例的支持高性能系统中的网络交换机的图示。如图2所示,诸如高性能系统200中的IB网络交换机220之类的网络设备可以负责将来自各种业务源201和211的数据业务引导到各种业务目的地208和218。

[0022] 例如,支持大量端口 (诸如输入端口202和212以及输出端口207和217) 的IB网络交换机220可以基于交叉开关 (XBAR, crossbar) 结构210。

[0023] 如图2所示,输入端口202可以使用源VL 221接收来自业务源201的各种传入数据分组,并且输入端口212可以使用源VL 231接收来自业务源211的各种数据分组。另外,输出端口207可以使用目的地VL 227将传出数据分组发送到业务目的地208,并且输出端口217可以使用目的地VL 237将传出数据分组发送到业务目的地218。

[0024] 此外,IB交换机220可以满足支持可用网络结构资源的最优使用的不同QoS需求。例如,IB交换机220可以基于与应用相关联的业务的服务级别 (SL) 将用于分组的传入VL (即,源VL) 重新映射到用于分组的不同的传出VL (即,目的地VL)。

[0025] 根据本发明的实施例,输入端口202或212中的每一个可以利用输入端口分组分类器203或213,输入端口分组分类器203或213可以确定用于每个传入分组的输出端口。例如,输入端口分组分类器203可以确定用于在输入端口202处接收到的每个分组的输出端口 (并且可以使用端口过滤器204来移除一个或多个分组),而输入端口分组分类器213可以确定

用于在输入端口212处接收到的每个分组的输出端口(并且可以使用端口过滤器214来移除一个或多个分组)。

[0026] 此外,输入端口分组分类器203或213可以确定用于到达输入端口202或212的(诸如用于多播和广播的)每个多目的地分组的多个输出目的地端口。端口过滤器204或214可以从给定分组的端口列表中移除一个或多个目的地端口。此外,如果所有的目的地端口都从列表中被移除,则多目的地分组可以被丢弃。否则,分组可以为了可用目的地端口而排队,该可用目的地端口可以是(由输入端口分组分类器203或213)最初分类的端口列表的子集。

[0027] 在将接收到的分组(例如经由输出端口207或217)发送到业务目的地208或218之前,输入端口202或212可以在每输入端口的基础上将接收到的分组存储在入口缓冲区(例如,虚拟输出队列(VOQ) 205或215)中。如图2所示,在输入端口202处接收到的分组可以被存储在VOQ 205中,而在输入端口212处接收到的分组可以被存储在VOQ 215中。

[0028] 此外,入口缓冲区(例如,VOQ 205或215)中的每一个可以包括若干队列,这些队列中的每个队列可以负责处理针对与输出端口相关联的目的地VL(例如,输出端口207上的VL 227和输出端口217上的VL 237)的分组。因此,在每输入端口的基础上,队列的总数可以是输出端口的数量与每个输出端口上支持的目的地VL的数量的乘积。因此,如果端口的数量和每个端口上支持的VL的数量大,则系统可能需要大量的队列以用于每个输入端口202或212。

[0029] 根据本发明的实施例,可以使用共享存储器结构来实现VOQ 205和215,并且VOQ 205和215中的每个队列的利用率可以取决于业务。例如,VOQ资源可以表示当传入分组排队(即,接收网络分组)时消耗的以及当分组出队列(即,将分组递送到输出端口)时最终释放的存储器块的数量。因此,VOQ资源的利用率可以与业务模式有关。

[0030] 根据本发明的实施例,系统可以调度输入端口202和212,并且引导存储在VOQ 205和215中的分组朝着输出端口207和217移动。入口缓冲区中的每个队列的排出速率可以取决于分组针对的目的地VL和输出端口。

[0031] 如图2所示,每个输出端口207或217可以利用输出调度器(诸如输出端口XBAR仲裁器206或216)。输出端口XBAR仲裁器206或216可以基于诸如各个VOQ的充满度和目的地VL上的可用信用之类的各种标准来做出与分组移动相关的决定。

[0032] 根据本发明的实施例,IB网络交换机220可以在不同条件下维持信用一致性。如图2所示,在IB网络交换机220的接收侧,可以基于传入分组的传入源VL 221和231来一致地维持信用;在IB网络交换机220的发送侧,可以基于传出分组的目的地VL 227和237来一致地维持信用。

[0033] 此外,在每输入端口的基础上,每个传入分组的排队可以基于分组的源VL来执行。因此,系统可以基于源VL 221或231来执行各种信用计费操作。例如,为了信用计费的目的,可以向IB网络交换机220中的每个源VL指派VOQ集合。

[0034] 管理VOQ资源

[0035] 根据本发明的实施例,系统可以提供用于在不同条件(诸如错误处理、链路状态改变和缓冲区重新分配)下管理VOQ资源的框架。

[0036] 图3示出了根据本发明的实施例的管理高性能系统中的VOQ资源的图示。如图3所

示,IB网络交换机300中的输入端口302可以向远程发送者(诸如业务源301)通告一个或多个信用以及从远程发送者接收一个或多个数据分组。

[0037] 输入端口302可以利用输入分组分类器303,分组分类器303可以为每个传入分组确定一个或多个目的地(例如,输出端口307或317)(并且可以使用端口过滤器304来移除一个或多个分组)。此外,输入端口302可以在将接收到的分组经由交叉开关(XBAR)结构310转发到不同的远程接收者(诸如业务目的地308和318)之前将分组存储在包括一组VOQ的入口缓冲区305中。

[0038] 如图3所示,入口缓冲区305中的VOQ 311与分组冲刷引擎320相关联,分组冲刷引擎320提供了在冲刷分组的低级别过程(包括管理用于冲刷分组的各种类型的事件)之上的抽象层。另外,入口缓冲区315中的VOQ 321可以与不同的分组冲刷引擎(未示出)相关联。

[0039] 例如,被存储在VOQ 311中的分组可以引发(spawn)多个缓冲区,该多个缓冲区可以被链接为链表。基于分层进程的分组冲刷引擎320可以从VOQ 311的头部开始遍历缓冲区的链表,并且逐个使缓冲区的链表出队列。一旦整个分组从VOQ 311中释放出来,则分组冲刷引擎320可以(例如通过向输出端口仲裁器306和316宣告分组冲刷事件322)通知各种输出端口307和317。

[0040] 此外,当多播分组要从入口缓冲区305出队列时,多播分组需要顺序地在目的地列表中的各种VOQ中排队以及从目的地列表中的各种VOQ出队列。因此,直到多播分组从作为目的地列表的一部分的最后一个VOQ出队列,分组冲刷引擎320才可以释放VOQ 311中的缓冲区。

[0041] 根据本发明的实施例,系统可以提供各种输出调度器(诸如输出端口仲裁器306和316)的抽象,以使得输出端口仲裁器306和316可以对在VOQ资源变化后面的不同条件是不可知的。例如,VOQ资源可以被回收以用于处理各种错误条件,诸如生命周期限制的头部(HLL)超时的到期以及各种链路状态改变。

[0042] 为了(在端口和相关联的VL之间)提供非阻塞行为,在单个输入端口302上的VOQ资源的回收可以结合不同的输出调度器(诸如具有从输入端口302调度的业务的输出仲裁器306和316)来执行。因此,抽象可以使输出端口仲裁器306和316与入口缓冲区305中的VOQ保持同步。

[0043] 如图3所示,当分组到达时,入口缓冲区305可以将分组排队到入口缓冲区305中的缓冲区中(例如,VOQ 311中),并且(例如,通过向输出端口仲裁器306和316宣告分组到达事件312)通知各种输出端口307和317。

[0044] 此外,输出端口仲裁器306或316可以相应地更新它自己的记录。例如,输出端口仲裁器306可以使记分板309中的分组计数(或长度)递增,并且输出端口仲裁器316可以使记分板319中的分组计数(或长度)递增。随后,在分组被调度并且已移动通过交叉开关310结构之后,输出调度器306和316可以例如通过使记分板309或319中的分组计数(或长度)递减来再次修改记分板。

[0045] 另一方面,当分组被冲刷(或移除)时,入口缓冲区305(例如,VOQ 311)可以使该分组从它的缓冲区出队列并且(例如,通过向输出端口仲裁器306和316宣告分组冲刷事件322)通知各种输出端口307和317。然后,输出端口307或317可以使来自记分板309或319的分组计数(或长度)递减(或相应地修改记分板309或319中的内容)。

[0046] 根据本发明的实施例,系统可以使用上面的抽象来处理诸如HLL到期条件之类的各种错误条件以及从各种错误条件恢复。例如,HLL到期可以在输出端口饥饿(starve)延长的时间段时发生。另外,HLL到期可以在上游端口上的拥塞和/或VL的停滞已发生延长的时间段之后发生。

[0047] 为了处理HLL到期,系统可以以优雅的方式在不阻塞其它VOQ和端口的情况下从VOQ的头部删除停滞的分组。为了支持HLL报告,为了处理HLL到期而被删除的分组可以具有包括各种分组统计信息的印迹(footprint)并且可以仅消耗与各种所需逻辑相关联的必需的缓冲区信用。

[0048] 此外,重复的HLL到期可以完全禁用队列,并且可能导致整个VOQ被冲刷。例如,系统可以通过重复使用分组冲刷引擎320来递归执行分组冲刷操作。

[0049] 此外,系统可以结合多播复制来处理HLL到期。在基于硬件的多播复制中,可以在每VOQ的基础上支持基于HLL的分组冲刷。此外,在目的地列表中的输出端口被服务之后,分组可以从一个VOQ中出队列并且排队到另一个VOQ中。系统可以利用忽略队列中出现的分组的类型的上面的抽象。因此,可以以与处理单播分组相同的方式来处理多播分组。

[0050] 根据本发明的实施例,系统可以为每个VOQ提供一组定时器(诸如HLL定时器)。例如,可以检查分组在队列的头部处已暂存(stage)了多长时间的定时器可以每当新的分组被插入到队列的头部时开始。

[0051] 如果在分组被调度以用于朝着输出端口传输之前定时器到期,则系统可以将分组标记为要被删除并且可以使用分组冲刷引擎320来冲刷分组。另一方面,如果在定时器到期之前分组被调度以用于传输,则定时器可以被重置并且等待下一个分组的到达。

[0052] 根据本发明的实施例,系统可以使用上面的抽象来处理各种链路状态改变。

[0053] 如图3所示,输入端口302可以将接收到的分组存储在入口缓冲区305中,并且多个输出端口307和317可以由同一输入端口302服务。当与输入端口302相关联的入口链路变得可用和不可用时,(例如,当输出端口在入口链路变回可用之前停滞延长的时间段时)在入口缓冲区305中可能出现陈旧数据。

[0054] 此外,当输入端口302上的链路状态改变时,诸如可操作的VL(例如所支持的VL的数量)和它们的初始信用分配(例如端口MTU等)之类的各种链路级别参数可能改变,因为这些链路级别参数可能与VOQ如何被设置以用于操作紧密耦合。

[0055] 为了处理各种链路状态改变,系统可以使用分组冲刷引擎320来冲刷出VOQ中出现的陈旧分组。例如,可以通过暂时将定时器减小到小的值来执行冲刷操作。此外,分组冲刷操作可以跨所有VOQ递归地执行,直到入口缓冲区305变得完全为空为止。因此,系统可以确保入口缓冲区305(或VOQ)不具有出现的陈旧数据。

[0056] 此外,当在输入端口302处发生链路状态改变时,多播分组可以在每次它被移动到目的地列表中的新VOQ中之后被冲刷。另外,入队操作和冲刷操作可以被顺序执行,以便使输出端口仲裁器306和326以及入口缓冲区305中的VOQ保持同步。

[0057] 图4示出了根据本发明的实施例的用于管理高性能系统中的VOQ资源的示例性流程图。如图4所示,在步骤401处,联网设备上的入口缓冲区可以存储在联网设备上的输入端口处接收到的一个或多个传入分组,其中入口缓冲区包括一个或多个虚拟输出队列。此外,在步骤402处,系统可以向一个或多个输出调度器通知分组已到达,其中每个输出调度器与

输出端口相关联。然后,在步骤403处,与入口缓冲区相关联的分组冲刷引擎可以冲刷被存储在入口缓冲区中的所述虚拟输出队列中的分组。此外,在步骤404处,系统可以向一个或多个输出调度器通知分组被冲刷,其中每个输出调度器与输出端口相关联。

[0058] 缓冲区的实时重新分配

[0059] 图5示出了根据本发明的实施例的支持高性能系统中的缓冲区重新分配的图示。如图5所示,输入端口502可以向远程发送者(诸如业务源501)通告一个或多个信用。然后,输入端口502可以经由不同的源VL(例如VL 521-523)接收来自远程发送者的一个或多个分组。

[0060] 输入端口502可以利用输入分组分类器503,分组分类器503可以为每个传入分组确定一个或多个目的地(例如,输出端口507或517)(并且可以使用端口过滤器504来移除一个或多个分组)。另外,输入端口502可以使用信用状态管理进程512(即,信用状态管理单元或块)来维护在IB链路上被通告给外部源的信用。

[0061] 此外,在将接收到的分组经由交叉开关(XBAR)结构510转发到不同的远程接收者(例如,业务目的地508和518)之前,输入端口502可以将这些分组存储在包括一组VOQ(诸如VOQ 511)的入口缓冲区505中。另外,另一个输入端口(未示出)可以将接收到的分组存储在同样包括一组VOQ的不同的入口缓冲区515中。

[0062] 此外,每个入口缓冲区505或515可以维护不同的信用状态,诸如每个源VL的当前缓冲区占用率(该当前缓冲区占用率可以基于每个缓冲区的充满度或空置度来定义)、每个源VL所允许的最大缓冲区、每个源VL的当前参与的VOQ列表以及它们的相对队列深度。

[0063] 根据本发明的实施例,系统可以基于如先前的部分中描述的抽象来支持VOQ资源的重新分配和/或重新分区。

[0064] 为了提供(各种端口和VL之间的)非阻塞行为,可以结合所有输出调度器(诸如具有从输入端口502调度的业务的输出端口仲裁器506和516)来管理输入端口502上的VOQ资源的重新分配。

[0065] VOQ资源的重新分配和/或重新分区可以由于不同的原因而发生,这些原因诸如是在每个源和/或每个目的地VL的基础上的缓冲要求的变化以及对应用关于它的QoS要求的需求的变化,这可能导致更多的缓冲区/信用被分配给VL而以其它VL为代价。

[0066] 此外,VOQ资源的重新分配和/或重新分区可以在链路状态改变之后发生。在链路状态改变的情况下,VOQ资源的重新分配和/或重新分区可以涉及将陈旧分组从VOQ冲刷出去。例如,在链路恢复之后新的可操作VL集合被协商的情况下,由于缓慢的排出速率,旧的分组可能保留在队列中。另一方面,应用可能希望保持在链路变得不可用之前的最后一组分组,而不是将陈旧分组完全冲刷出去。

[0067] 如图5所示,入口缓冲区505可以向物理输入端口502中的链路信用管理进程512通告初始信用更新。例如,比如当存在对每个源VL的最大可用缓冲区空间的更新时,VOQ 511可以向物理输入端口502通告每个源VL的可用缓冲区空间。然后,链路信用管理进程512可以保持所通告的最大允许值并且相应地管理链路信用。

[0068] 根据本发明的实施例,初始信用更新可以在不同场景中被触发,诸如在初始化期间以及在缓冲区的重新分区完成之后被触发。对于每种情况,物理输入端口502中的链路信用管理进程512可以对发起更新的进程是不可知的。

[0069] 当入口缓冲区505接收可以由软件和/或固件发起的重新分区请求时,系统可以选取为其减小缓冲区大小的VL。随后,系统可以通过从VOQ的头部冲刷一个或多个分组来减小与VL相关联的缓冲区大小。例如,系统可以(例如,基于队列的充满度或空置度)选取与输出端口上具有最高占用率的VL相关联的VOQ。

[0070] 然后,VOQ可以发送具有用于被更新的VL(包括为其增加缓冲区大小的VL以及为其减小缓冲区大小的VL)的更新值的初始信用更新。此外,如果由于在任何时间点接收到一个或多个分组,与源VL 521-523相关联的缓冲区大小超过其最大允许的缓冲区大小,则这些分组可以被优雅地丢弃。

[0071] 根据本发明的实施例,每个VOQ可以与分组冲刷引擎520相关联,分组冲刷引擎520可以管理用于处理分组冲刷的各种类型的事件。当由软件/固件发起的重新分区请求被接收时,分组冲刷硬件引擎520可以执行如以下列表1中所示的序列。

```
[0072]   if (new_buffer_size is >= current_buffer_occupancy) {  
           update max_buffers_allowed with the specified new_buffer_size;  
           send init credit update with the new value and mark the process as completed  
       } else {  
           while (new_buffer_size is < current_buffer_occupancy) {  
               choose_a_voq_for_the given_src_vl();  
               flush_packet_from_head()  
[0073]           if the current occupancy drops below the new size exit, else continue  
               if the current occupancy increases over the new buffer size specified due to  
               incoming packets, packets are dropped gracefully before the queuing process.  
           }  
           update max_buffers_allowed with the specified new_buffer_size;  
           send initial credit update with the new value and mark the process as completed;  
       }
```

[0074] 列表1

[0075] 此外,在VOQ资源的改变之前和之后,系统可以提供了解不同要求的管理代理525。因此,系统可以确保大多数VL上的业务可以不间断地进行,并且仅损失最小的分组集合。

[0076] 图6示出了根据本发明的实施例的用于支持高性能系统中的缓冲区重新分配的示例性流程图。如图6所示,在步骤601处,系统可以在联网设备上的入口缓冲区中存储在输入端口处接收到的一个或多个分组,其中该输入端口与多个源虚拟通道(VL)相关联。此外,在步骤602处,系统可以将入口缓冲区中的缓冲区资源从与该多个源VL中的第一源VL相关联重新分配为与该多个源VL中的第二源VL相关联。然后,在步骤603处,入口缓冲区可以将初始信用更新发送到用于第一源VL和第二源VL的输入端口。

[0077] 在本发明的另一种实施例中,一种用于支持网络环境中的分组交换的方法包括:在联网设备上的入口缓冲区中存储在输入端口处接收到的一个或多个分组,其中该输入端

口与多个源虚拟通道 (VL) 相关联;将入口缓冲区中的缓冲区资源从与该多个源VL中的第一源VL相关联重新分配为与该多个源VL中的第二源VL相关联;以及将初始信用更新发送到用于第一源VL和第二源VL的输入端口。

[0078] 在本发明的另一种实施例中,该方法中联网设备是网络交换机,该网络交换机包括交叉开关结构。

[0079] 在本发明的另一种实施例中,该方法中入口缓冲区包括多个虚拟输出队列,其中每个所述虚拟输出队列操作以存储针对输出端口的一个或多个分组。

[0080] 在本发明的另一种实施例中,该方法还包括选择与第一源VL相关联并且具有最高占用率的虚拟输出队列。

[0081] 在本发明的另一种实施例中,该方法还包括从所述虚拟输出队列中冲刷一个或多个分组。

[0082] 在本发明的另一种实施例中,该方法还包括从所述虚拟输出队列的头部开始遍历存储所述一个或多个分组的缓冲区链表。

[0083] 在本发明的另一种实施例中,该方法还包括使用管理代理来发起缓冲区重新分配操作。

[0084] 在本发明的另一种实施例中,该方法还包括经由入口缓冲区向输入端口通告每个源VL的可用缓冲区空间。

[0085] 在本发明的另一种实施例中,该方法还包括经由输入端口基于初始信用更新来设置所允许的最大信用。

[0086] 在本发明的另一种实施例中,该方法还包括经由输入端口基于所允许的最大信用向外部源通告可用信用。

[0087] 在本发明的实施例中,一种用于支持网络环境中的分组交换的系统,该系统包括联网设备上的入口缓冲区,其中该入口缓冲区操作以:存储在联网设备上的输入端口处接收到的一个或多个传入分组,其中该输入端口与多个源虚拟通道 (VL) 相关联;将入口缓冲区中的缓冲区资源从与该多个源VL中的第一源VL相关联重新分配为与该多个源VL中的第二源VL相关联;以及将初始信用更新发送到用于第一源VL和第二源VL的输入端口。

[0088] 在本发明的另一种实施例中,该系统中联网设备是网络交换机,该网络交换机包括交叉开关结构。

[0089] 在本发明的另一种实施例中,该系统中入口缓冲区包括多个虚拟输出队列,其中每个所述虚拟输出队列操作以存储针对输出端口的一个或多个分组。

[0090] 在本发明的另一种实施例中,该系统中入口缓冲区操作以选择与第一源VL相关联并且具有最高占用率的虚拟输出队列。

[0091] 在本发明的另一种实施例中,该系统中入口缓冲区操作以使用分组冲刷引擎从所述虚拟输出队列中冲刷一个或多个分组。

[0092] 在本发明的另一种实施例中,该系统中分组冲刷引擎操作以从所述虚拟输出队列的头部开始遍历存储所述一个或多个分组的缓冲区链表。

[0093] 在本发明的另一种实施例中,该系统中管理代理操作以发起缓冲区重新分配操作。

[0094] 在本发明的另一种实施例中,该系统中入口缓冲区操作以向输入端口通告每个源

VL的可用缓冲区空间。

[0095] 在本发明的另一种实施例中,该系统中输入端口操作以:基于初始信用更新来设置所允许的最大信用,以及基于所允许的最大信用来向外部源通告可用信用。

[0096] 在本发明的实施例中,一种具有存储在其上的指令的非暂态机器可读存储介质,该指令当被执行时,使得系统执行步骤,这些步骤包括:在联网设备上的入口缓冲区中存储在输入端口处接收到的一个或多个分组,其中该输入端口与多个源虚拟通道(VL)相关联;将入口缓冲区中的缓冲区资源从与该多个源VL中的第一源VL相关联重新分配为与该多个源VL中的第二源VL相关联;以及将初始信用更新发送到用于第一源VL和第二源VL的输入端口。

[0097] 在本发明的实施例中,一种包括存储在计算机可读存储介质上的指令的计算机程序,该指令当被执行时,使得一个或多个计算机系统执行上面的方法。

[0098] 本发明的许多特征可以在硬件、软件、固件或其组合中执行,利用硬件、软件、固件或其组合执行,或者在硬件、软件、固件或其组合的辅助下执行。因此,本发明的特征可以利用(例如,包括一个或多个处理器的)处理系统来实现。

[0099] 本发明的特征可以在计算机程序产品中实现、利用计算机程序产品实现、或者在计算机程序产品的辅助下实现,该计算机程序产品是具有存储在其上/其中的、可以被用来对处理系统编程以执行本文所呈现的特征中的任何特征的指令的存储介质(媒介)或计算机可读介质(媒介)。存储介质可以包括但不限于任何类型的盘,包括软盘、光盘、DVD、CD-ROM、微驱动器、以及磁光盘、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、闪存存储器设备、磁卡或光卡、纳米系统(包括分子存储器IC)、或适合于存储指令和/或数据的任何类型的媒介或设备。

[0100] 被存储在机器可读介质(媒介)中的任何一种机器可读介质中,本发明的特征可以被结合到软件和/或固件中,以用于控制处理系统的硬件以及用于使得处理系统能够利用本发明的结果与其它机制交互。这样的软件或固件可以包括但不限于应用代码、设备驱动程序、操作系统和执行环境/容器。

[0101] 本发明的特征还可以利用例如硬件组件(诸如专用集成电路(ASIC))在硬件中实现。为了执行本文所描述的功能的硬件状态机的实现对相关领域的技术人员将是明显的。

[0102] 此外,可以利用包括根据本公开的教导编程的一个或多个处理器、存储器和/或计算机可读存储介质的一个或多个常规的通用或专用数字计算机、计算设备、机器或微处理器来方便地实现本发明。如对软件领域的技术人员将明显的,适当的软件编码可以由熟练的程序员基于本公开的教导容易地准备。

[0103] 虽然上文已经描述了本发明的各种实施例,但是应当理解的是,它们通过示例的方式而不是通过限制的方式给出。对于相关领域的技术人员将明显的是,在不脱离本发明的精神和范围的情况下,其中可以进行形式和细节上的各种改变。

[0104] 上文已经借助示出指定的功能及其关系的执行的功能构建块描述了本发明。在本文中为了描述的方便通常任意定义这些功能构建块的界限。只要指定的功能及其关系被适当地执行,就可以定义可替代的界限。因此,任何这样的可替代界限在本发明的范围和精神之内。

[0105] 为了说明和描述的目的提供了本发明的上述描述。它不旨在是详尽的或者把本发

明限制到所公开的精确形式。本发明的宽度和范围不应当由上文描述的示例性实施例中的任何实施例来限制。许多修改和变化对本领域技术人员来说将是明显的。修改和变化包括所公开特征的任何相关组合。实施例被选择和描述以便于最佳地解释本发明的原理及其实地应用,从而使得本领域其他技术人员能够对于各种实施例并且利用适于预期的特定使用的各种修改来理解本发明。旨在由以下权利要求及其等价物来定义本发明的范围。

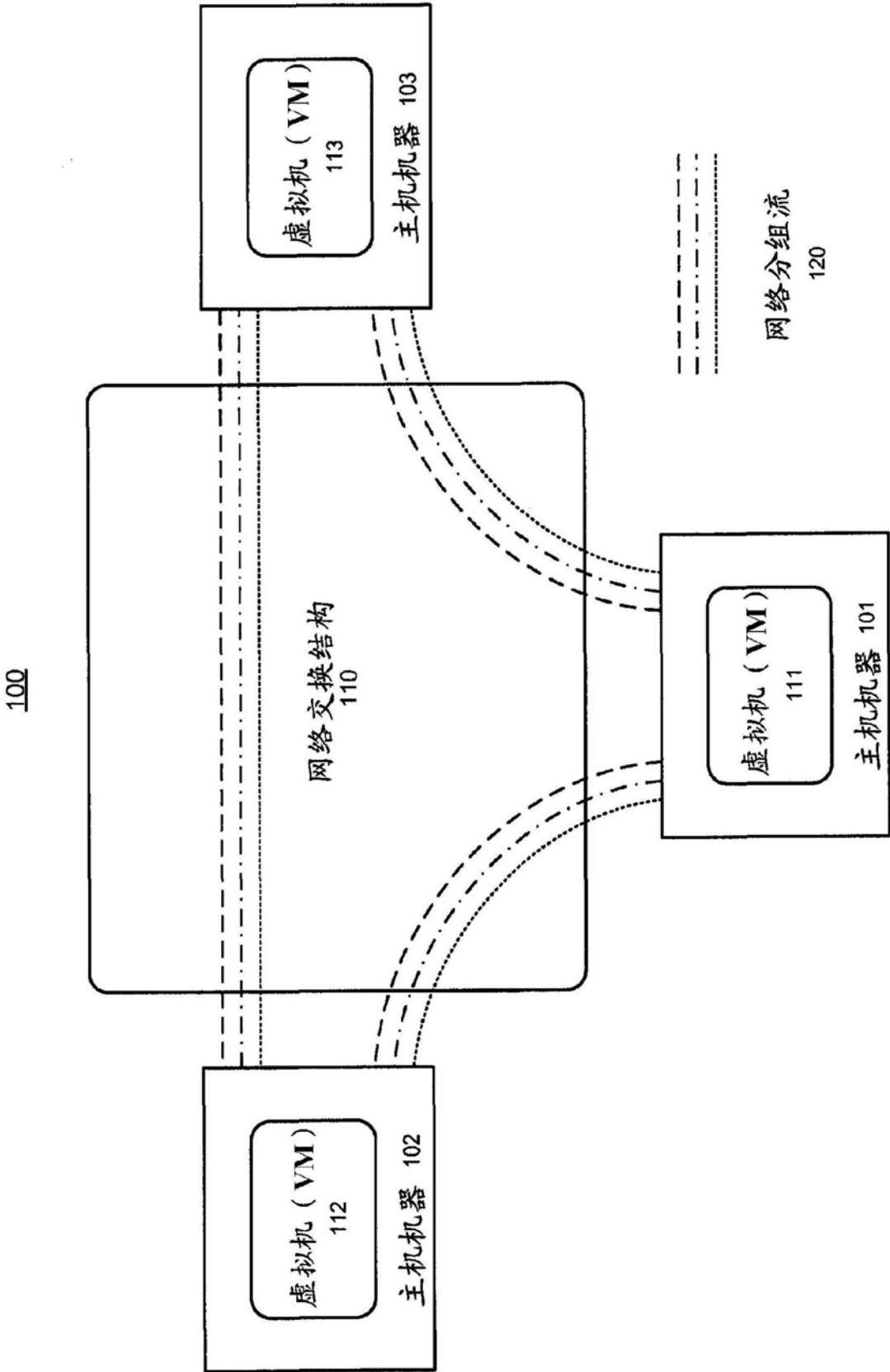


图1

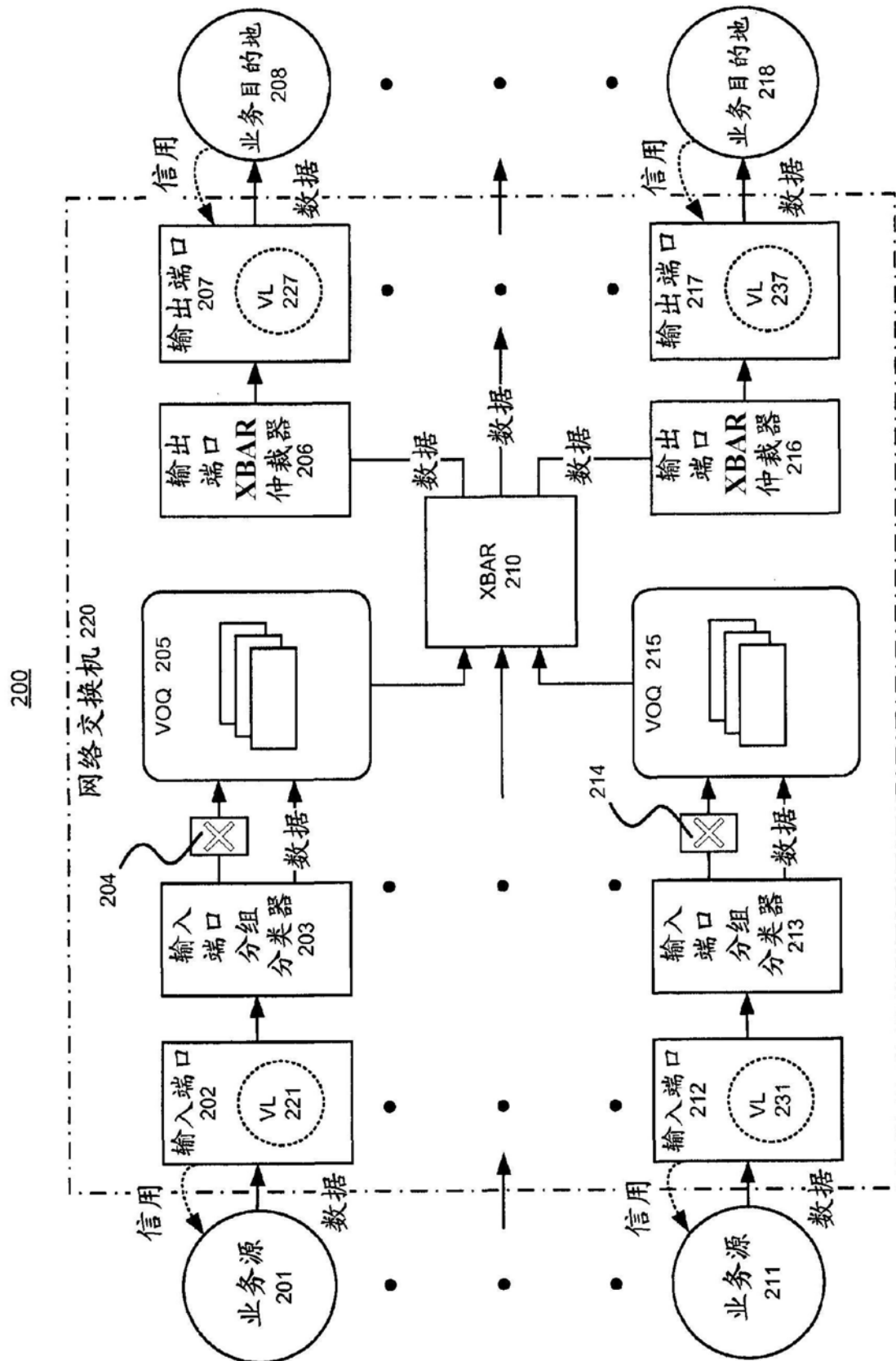


图2

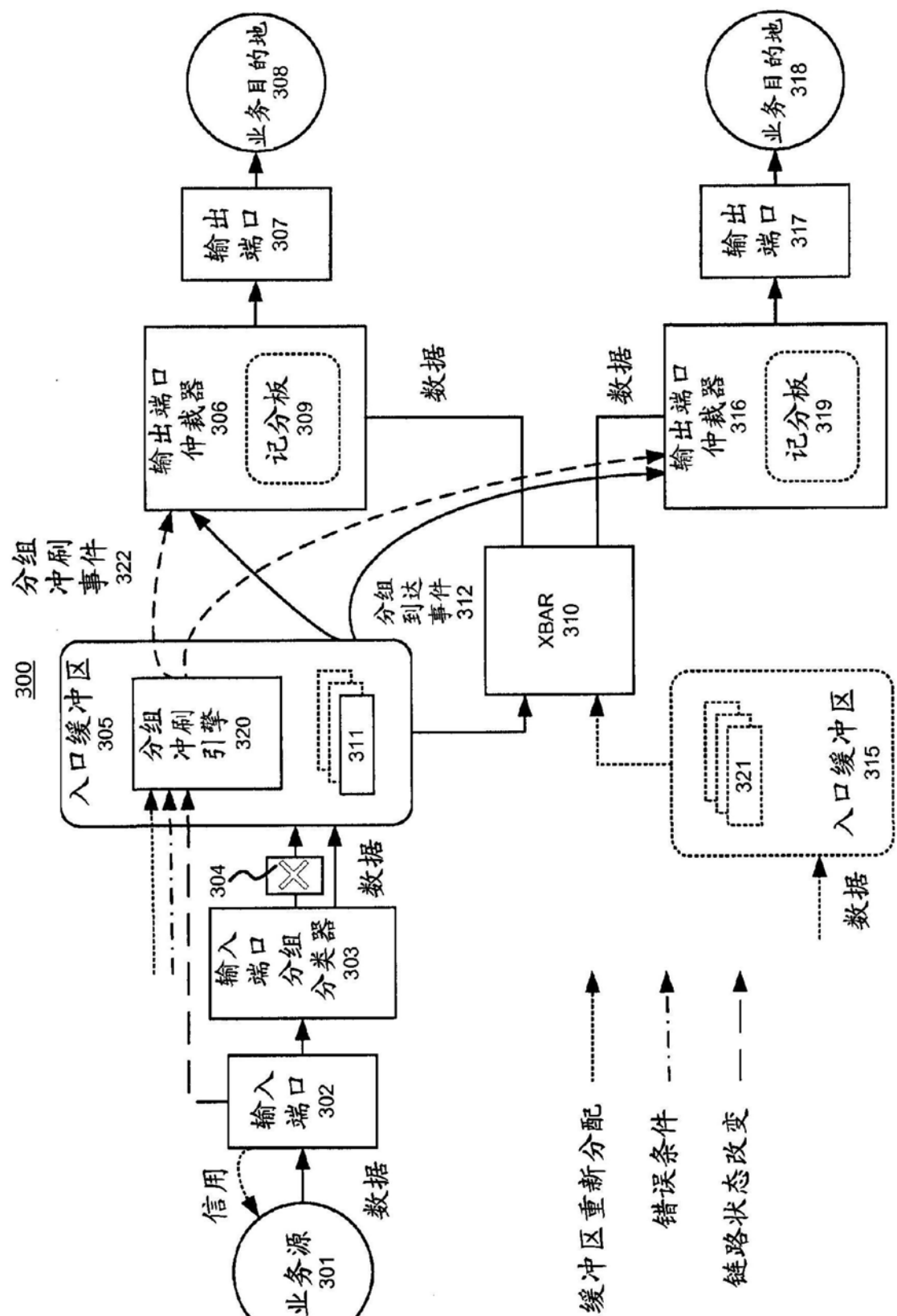


图3

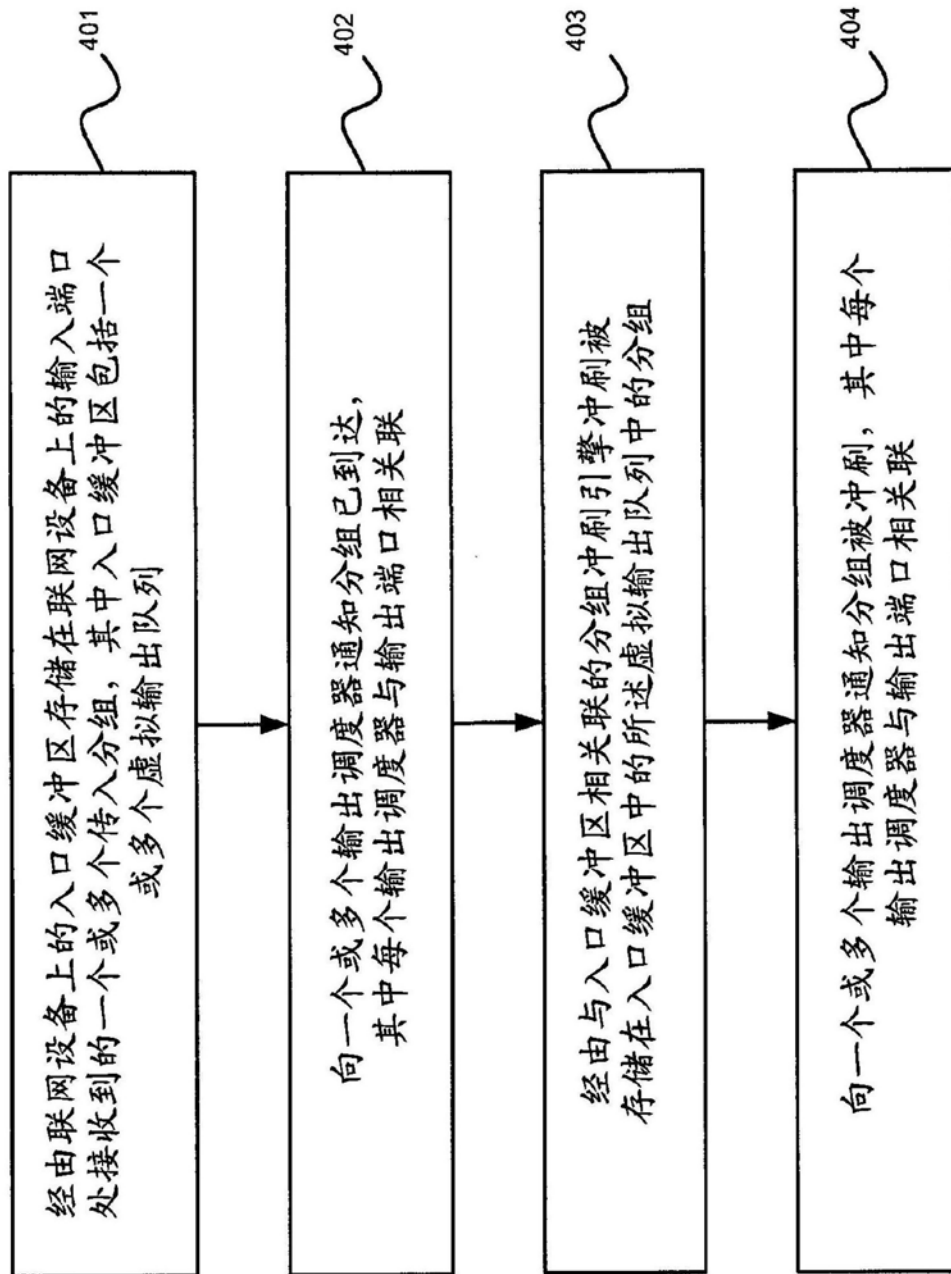


图4

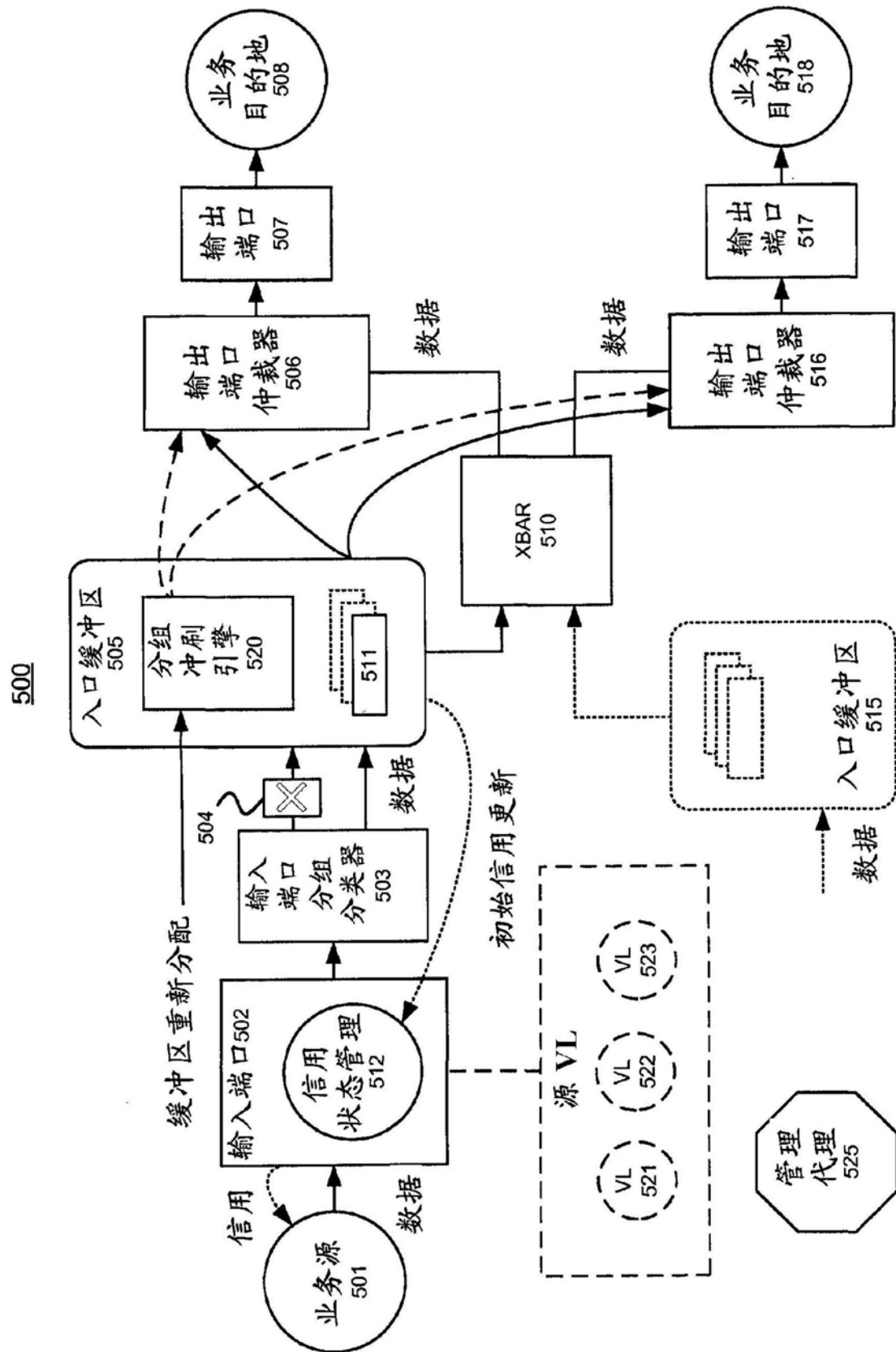


图5

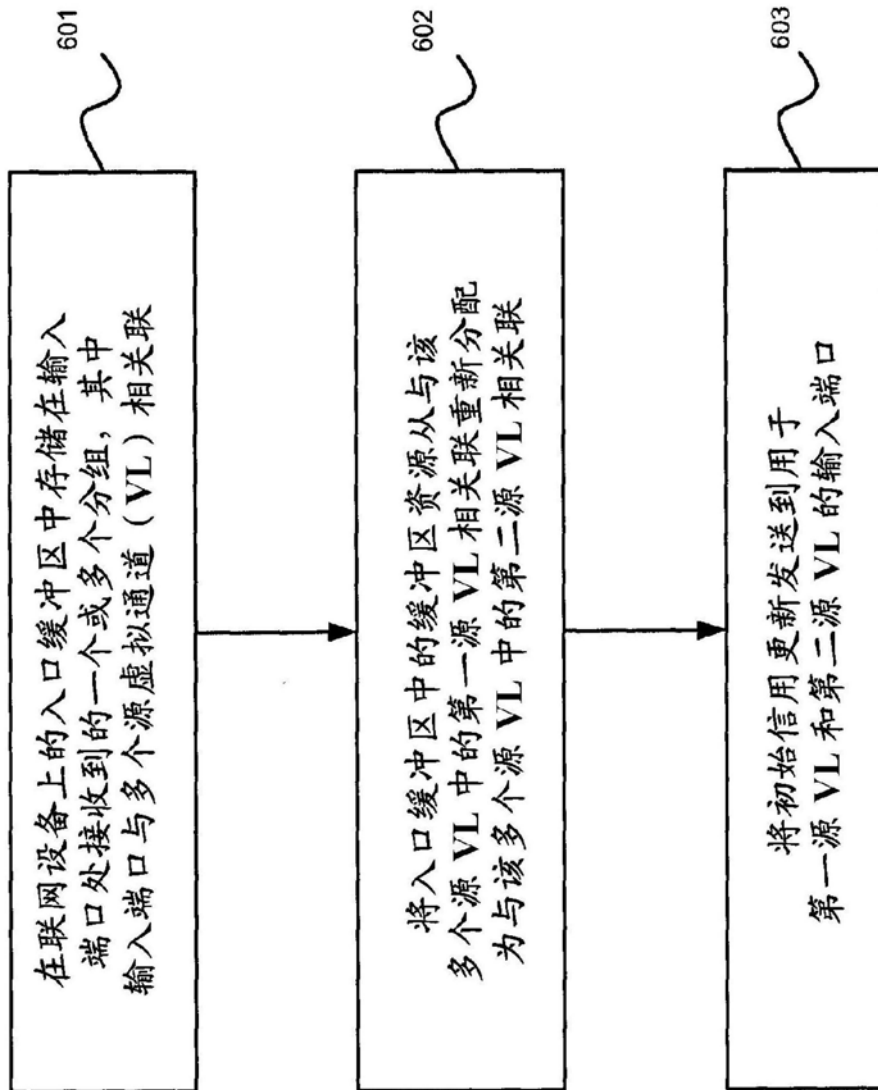


图6