



(12)发明专利

(10)授权公告号 CN 104246700 B

(45)授权公告日 2017.05.10

(21)申请号 201380022068.3

(22)申请日 2013.05.08

(65)同一申请的已公布的文献号

申请公布号 CN 104246700 A

(43)申请公布日 2014.12.24

(30)优先权数据

61/646,107 2012.05.11 US

13/889,088 2013.05.07 US

13/889,123 2013.05.07 US

(85)PCT国际申请进入国家阶段日

2014.10.27

(86)PCT国际申请的申请数据

PCT/US2013/040212 2013.05.08

(87)PCT国际申请的公布数据

W02013/169949 EN 2013.11.14

(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 B·博格丹斯基 B·D·约翰森

(74)专利代理机构 中国国际贸易促进委员会专

利商标事务所 11038

代理人 袁玥

(51)Int.Cl.

G06F 9/44(2006.01)

H04L 29/12(2006.01)

(56)对比文件

US 2004/0024903 A1,2004.02.05,

US 2008/0285562 A1,2008.11.20,

US 2010/0008222 A1,2010.01.14,

US 8175107 B1,2012.05.08,

US 2009/0216853 A1,2009.08.27,

审查员 覃婧婵

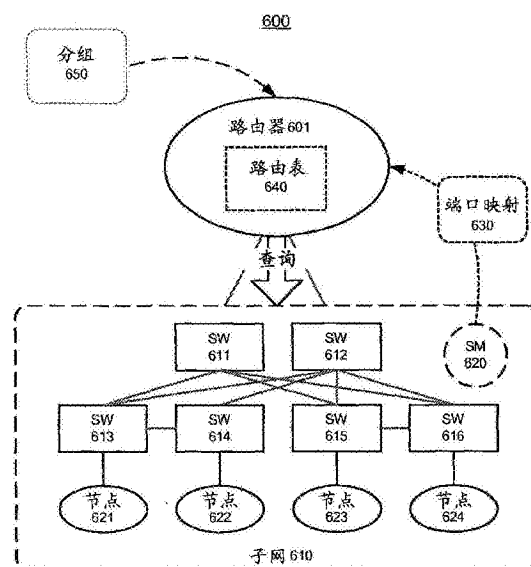
权利要求书3页 说明书10页 附图11页

(54)发明名称

用于基于胖树路由在不同无限带宽子网间路由流量的系统和方法

(57)摘要

一种可以在网络环境中的不同子网间路由流量的系统和方法。连接到比如无限带宽(IB)子网的不同子网的路由器可以接收该路由器负责路由一或多个分组所到的目的地的列表。此外,该路由器能够从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息,以及基于所获取的信息构建路由表。



1. 一种用于在具有第一胖树拓扑的第一子网中的多个源节点和具有第二胖树拓扑的第二子网中的多个目的地节点之间路由一个或多个分组的方法,其中,所述第一子网和第二子网通过路由器直接连接,所述路由器具有连接到第一子网的多个入端口和连接到第二子网的多个出端口,所述方法包括:

在所述路由器处,从第一子网接收所述目的地节点的列表;

在所述路由器处,从第二子网的子网管理器接收所述第二子网的所述第二胖树拓扑的端口映射列表,其中所述端口映射列表包括所述多个目的地节点;

从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够作为去往所述第二子网中的所述多个目的地节点中的一个或多个目的地节点的主要路径而用于路由该一或多个分组的信息;以及

基于所获取的信息构建路由表。

2. 根据权利要求1所述的方法,进一步包括:

在连接到所述第一子网和第二子网的所有可用路由器间执行轮询路径分配。

3. 根据权利要求1或2所述的方法,进一步包括:

查询连接到每个路由器的交换机以了解哪个交换机能够用作去往特定目的地的主要路径。

4. 根据权利要求1或2所述的方法,进一步包括:

基于所获取的信息确定为了最大性能能够使用路由器的哪些向下输出端口。

5. 根据权利要求1或2所述的方法,进一步包括:

从子网管理器接收在一或多个目的地全局标识符GID和一或多个路由器端口间的端口映射。

6. 根据权利要求1或2所述的方法,进一步包括:

使用一或多个附加的路由器将该第一子网和第二子网连接到网络环境中的一或多个相邻子网。

7. 根据权利要求1或2所述的方法,进一步包括:

允许每个所连接的子网是多级胖树的分支。

8. 根据权利要求1或2所述的方法,进一步包括:

允许所述第一子网和第二子网以胖树拓扑被配置。

9. 根据权利要求1或2所述的方法,进一步包括:

使用胖树路由算法以用于路由该至少一个子网。

10. 一种用于在具有第一胖树拓扑的第一子网中的多个源节点和具有第二胖树拓扑的第二子网中的多个目的地节点之间路由一个或多个分组的系统,其中,所述第一子网和第二子网通过路由器直接连接,所述路由器具有连接到第一子网的多个入端口和连接到第二子网的多个出端口,所述系统包括:

一或多个微处理器;

在一或多个微处理器上运行的路由器,其中所述路由器操作为

在所述路由器处,从第一子网接收所述目的地节点的列表;

在所述路由器处,从第二子网的子网管理器接收所述第二子网的所述第二胖树拓扑的端口映射列表,其中所述端口映射列表包括所述多个目的地节点;

从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够作为去往所述第二子网中的所述多个目的地节点中的一个或多个目的地节点的主要路径而用于路由该一或多个分组的信息;以及

基于所获取的信息构建路由表。

11. 根据权利要求10所述的系统,其中:

轮询路径分配在连接到所述第一子网和第二子网的所有可用路由器间被执行。

12. 根据权利要求10或11所述的系统,其中:

所述路由器操作为查询连接到所述路由器的交换机以了解哪个交换机能够用作去往特定目的地的主要路径。

13. 根据权利要求10或11所述的系统,其中:

所述路由器操作为基于所获取的信息确定为了最大性能能够使用路由器的哪些向下输出端口。

14. 根据权利要求10或11所述的系统,其中:

所述路由器操作为从子网管理器接收在一或多个目的地全局标识符GID和一或多个路由器端口间的端口映射。

15. 根据权利要求10或11所述的系统,其中:

该第一子网和第二子网被使用一或多个附加的路由器连接到网络环境中的一或多个相邻子网。

16. 根据权利要求10或11所述的系统,其中:

每个所连接的子网是多级胖树的分支。

17. 根据权利要求10所述的系统,其中:

所述第一子网和第二子网能够以胖树拓扑被配置,并且胖树路由算法能够被用于路由该至少一个子网。

18. 一种用于在具有第一胖树拓扑的第一子网中的多个源节点和具有第二胖树拓扑的第二子网中的多个目的地节点之间路由一个或多个分组的系统,其中,所述第一子网和第二子网通过路由器直接连接,所述路由器具有连接到第一子网的多个入端口和连接到第二子网的多个出端口,所述系统包括:

用于在所述路由器处,从第一子网接收所述目的地节点的列表的装置;

用于在所述路由器处,从第二子网的子网管理器接收所述第二子网的所述第二胖树拓扑的端口映射列表的装置,其中所述端口映射列表包括所述多个目的地节点;

用于从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够作为去往所述第二子网中的所述多个目的地节点中的一个或多个目的地节点的主要路径而用于路由该一或多个分组的信息的装置;以及

用于基于所获取的信息构建路由表的装置。

19. 根据权利要求18所述的系统,进一步包括:

用于在连接到所述第一和第二子网的所有可用路由器间执行轮询路径分配的装置。

20. 根据权利要求18或19所述的系统,进一步包括:

用于查询连接到每个路由器的交换机以了解哪个交换机能够用作去往特定目的地的主要路径的装置。

21. 根据权利要求18或19所述的系统,进一步包括:
用于基于所获取的信息确定为了最大性能能够使用路由器的哪些向下输出端口的装置。
22. 根据权利要求18或19所述的系统,进一步包括:
用于从子网管理器接收在一或多个目的地全局标识符和一或多个路由器端口间的端口映射的装置。
23. 根据权利要求18或19所述的系统,进一步包括:
用于使用一或多个附加的路由器将该第一子网和第二子网连接到网络环境中的一或多个相邻子网的装置。
24. 根据权利要求18或19所述的系统,进一步包括:
用于允许每个所连接的子网是多级胖树的分支的装置。
25. 根据权利要求18或19所述的系统,进一步包括:
用于允许所述第一子网和第二子网以胖树拓扑被配置的装置。
26. 根据权利要求18或19所述的系统,进一步包括:
用于使用胖树路由算法以用于路由该第一子网和第二子网的装置。

用于基于胖树路由在不同无限带宽子网间路由流量的系统和方法

[0001] 版权声明

[0002] 本专利文献中的一部分公开内容包含受版权保护的素材。版权所有人不反对任何人影印再现专利文献或专利公开,因为专利文献或专利公开出现在可公开获得的美国专利商标局文件或档案中,不过保留其它方面的所有版权。

技术领域

[0003] 本发明主要涉及计算机系统,更具体地,涉及中间件(middleware)机器环境。

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经成为了值得注意的问题。无限带宽(IB)技术已经作为云计算结构经历更多的部署。这是本发明的实施例旨在解决的主要领域。

发明内容

[0005] 此处描述的是可以在不同无限带宽(IB)子网间路由流量的系统和方法。连接比如无限带宽(IB)子网的不同子网的路由器可以接收该路由器负责路由一或多个分组所到的目的地的列表。此外,该路由器可以从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息,以及基于所获取的信息构建路由表。

[0006] 此处还描述了用于在网路环境中的不同子网间路由流量的系统。该系统包含用于在路由器处接收该路由器负责路由一或多个分组所到的目的地的列表的装置,其中该路由器连接到该网络环境中的至少一个子网。该系统进一步包括用于从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息的装置。该系统进一步包括用于基于所获取的信息构建路由表的装置。

[0007] 进一步,此处描述了用于在网络环境中不同子网间路由流量的系统。该系统包括路由器,该路由器被配置为接收该路由器负责路由一或多个分组所到的目的地的列表,其中该路由器连接到该网络环境中的至少一个子网。该系统还被配置为从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息,以及基于所获取的信息构建路由表。

附图说明

[0008] 图1示出了根据本发明的实施例在网络环境中不同无限带宽(IB)子网间路由流量的图解。

[0009] 图2示出了根据本发明的实施例在网络环境中路由器上支持分组转发的图解。

[0010] 图3示出了根据本发明的实施例在网络环境中为不同目的地选择路由器入端口的

图解。

[0011] 图4示出了根据本发明的实施例在网络环境中路由器处为接收到的分组选择路由器出端口的图解。

[0012] 图5示出了根据本发明的实施例在网络环境中路由器处用于支持子网际源路由 (ISSR) 算法的示例性流程图。

[0013] 图6示出了根据本发明的实施例在网络环境中路由器处支持子网际胖树路由 (ISFR) 算法的图解。

[0014] 图7-9图解了根据本发明的实施例利用不同拓扑在网络环境中路由。

[0015] 图10图解了根据本发明的实施例在网络环境中路由器处用于支持子网际胖树路由 (ISFR) 算法的示例性流程图。

[0016] 图11图解了示出根据本发明的实施例的特征的功能框图。

具体实施方式

[0017] 在附图的图表中以示例的方式而不是限制的方式图解本发明,其中在附图当中相同的附图标记表示相同的元素。应当注意,在本公开中参考的“一个”或“某个”或“某些”实施例并不一定是同一个实施例,且这样的参考意味着至少一个。

[0018] 本发明的如下描述使用无限带宽 (IB) 网络作为高性能网络的示例。对于本领域技术人员而言,很明显其他种类的高性能网络可以被无限制地使用。同时,本发明的如下描述使用胖树作为网络拓扑模型的示例。对于本领域技术人员而言,很明显其他种类的网络拓扑模型可以被无限制地使用。

[0019] 此处描述的是可以支持不同的网络间路由流量的系统和方法。

[0020] 无限带宽 (IB) 架构

[0021] IB架构是串联点到点全双工技术。随着IB集群大小和复杂度的增加,网络可以被分割为可管理的部分,这些部分可以被称为子网。IB子网可以包括使用交换机和点到点连接互联的一组主机。IB子网还可以包括至少一个子网管理器 (SM), 该管理器负责初始化和启动该网络,包括子网内所有交换机、路由器和主机通道适配器 (HCA) 的配置。

[0022] IB支持传输服务的富集以提供远程直接内存访问 (RDMA) 和传统发送/接收语义。独立于所使用的传输服务,该IB HCA使用队列对 (QP) 通信。QP产生于通信设置期间,且可以具有一组初始属性,比如QP数、HCA端口、目的地LID、队列大小和提供的传输服务。HCA可以处理多个QP,每个QP包括一对队列,比如发送队列 (SQ) 和接收队列 (RQ),且在参与通信的每个末端节点都出现这样的一对。该发送队列保持待被传输到远程节点的工作请求,而接收队列保持关于怎样处理从远程节点接收的数据的信息。在QP之外,每个HCA具有与一组发送和接收队列相关联的一或多个完成队列 (CQ)。该CQ为发布到发送和接收队列的工作请求保持完成通知。虽然通信的复杂度不为用户所知,QP状态信息被保存于HCA中。

[0023] 在不同无限带宽 (IB) 子网间路由流量

[0024] 图1示出了根据本发明的实施例在网络环境中不同无限带宽 (IB) 子网间路由流量的图解。如图1中所示,网络环境100可以基于无限带宽架构 (IBA), 并支持两层拓扑划分。

[0025] IB网络的较低层,或IB结构100可以被称为子网,例如子网101-104,每个子网包括使用交换机和点到点连接互联的一组主机。在较高层,在IB结构100中的一或多个子网101-

104可以使用例如路由器105-106的路由器互联。进一步,每个子网101-104可以运行它自己的子网管理器(SM),该子网管理器只配置局域子网上的端口且路由器105-106对于子网管理器(SM)不透明。

[0026] 每个子网101-104中的主机和交换机可以使用指定的局部标识符(LID)被寻址。大装置的尺寸可以由可用的局部标识符(LID)的数量限制。例如,单个子网可以被限制到49151个单播LID。扩展IB地址空间的一种方法是将LID寻址空间扩展到32比特,因为这一方法可能不是与较旧硬件反向兼容的,所以该方法的可用性可能受到限制。

[0027] 根据本发明的实施例,IB路由器105-106可以提供IB结构100中的地址空间可扩展性。如图1所示,当希望更多末端端口时,多个子网101-104可以使用一或多个IB路由器105-106被结合到一起。子网101-104中的每一个可以使用局部标识符(LID)地址空间111-114。由于在每个LID地址空间111-114中LID地址具有局部可见性,该LID地址可以在由路由器105-106连接的不同子网101-104中被重新使用。因此,可以为IB结构100提供地址空间可扩展性,并且这种方法可以理论上获得不受限制的寻址空间。

[0028] 另外,通过将大且复杂的网络100分割成多个子网101-104,该系统可以提供结构管理牵制。该结构管理牵制在1) 错误隔离、2) 增加安全性和3) 子网内路由灵活性中是有益的。

[0029] 首先,通过将大的网络100分解成多个较小子网101-104,错误或拓扑变化可以被包含到单独子网,且子网重新配置可以不经过路由器105-106传递给其他子网。这使重新配置时间变短且限制错误的影响。

[0030] 第二,从安全性的观点出发,使用路由器105-106将大结构100分割为子网101-104限制了对被攻击子网的大多数攻击的范围。

[0031] 第三,从路由的观点出发,结构管理牵制可以在支持更灵活的路由策略方面是有益的,这在包含两个或更多常规拓扑的混合结构的情况下是特别有利的。

[0032] 例如,混合结构100可以包括与网状或环状部分(或任意其他常规拓扑)互联的胖树部分。因为子网内路由算法可以只具有子网范围,所以可以不存在分别路由混合结构100的不同部分的直接方法。此外,对于IB没有可以为混合拓扑提供最优化性能的通用不可知路由算法。

[0033] 根据本发明的实施例,混合拓扑可以被分割成较小的常规子网101-104,每一个子网可以使用对特定子网最优的不同路由算法被路由。例如,胖树路由算法可以路由胖树部分而维序路由可以路由拓扑的网状部分。

[0034] 另外,在当最终目的地位于另一个子网之后的较多不规则网络(例如至少需要两个路由器跳)的情况下,超级子网管理器可以被用于在局部子网管理器间协调且可以创建穿过过渡子网的路径。

[0035] 因此,通过使用在IB子网间的路由,该系统可以在IB网络中提供地址空间可扩展性和结构管理牵制。

[0036] 本地无限带宽(IB)路由器

[0037] 图2示出了根据本发明的实施例在网络环境中的路由器上支持分组转发的图解。如图2所示,该网络环境200可以包括使用路由器210互联的多个子网,例如IB子网A-B 201-202。此外,IB子网A 201连接到末端节点X 203,该节点与主机通道适配器(HCA) 213相关联,

且该IB子网B 202连接到末端节点Y 204,该节点与主机通道适配器 (HCA) 214相关联。

[0038] 根据本发明的实施例,IB路由器210可以在IB寻址层级的层3操作。此外,在LID之外,每个IB装置也可以具有64比特全局唯一标识符 (GUID),该标识符可以被用于形成GID、IB层3地址。例如,可以通过将64比特子网ID与64比特GUID级联以创建GID以形成类似IPv6的128比特地址。额外地,术语GUID也可以被用于指代端口GUID,即,被分配至IB结构中每个端口的GUID,且该GUID可以被烧至非易失存储器中。

[0039] 如图2所示,例如末端节点X 203的末端节点可以通过路由器210将分组发送到例如末端节点Y 204的另一个末端节点。该地址解析机制可以使局部路由器210对末端节点X-Y 203-204可见。

[0040] 例如,在路由器210处接收的分组可以包括进入的路由报头,该报头包括局部路由报头 (LRH) 223和全局路由报头 (GRH) 225。在转发分组之前,路由器210可以将进入的路由报头221修改为外出的路由报头222,该报头包括LRH 224和GRH 226。

[0041] 如图2所示,末端节点X 203可以在LRH 223中将例如X的局部HCA的LID地址作为源LID,即srcLID放入,且将例如A的局部路由器的LID地址作为目的地LID,即dstLID放入。此外,末端节点X 203可以在GRH中将例如1234的它的层-3地址 (GID) 作为源LID,即srcGID,且可以将例如5678的最终目的地地址 (GID) 作为目的地地址GID,即dstGID。

[0042] 当分组到达路由器210时,分组转发机制220可以更新并替换分组字段。例如,该系统可以将局部路由报头 (LRH) 224中的源LID,即srcLID替换为例如B的路由器出端口的LID。此外,该系统可以将目的地LID,即dstLID替换为例如Y的目的地的LID地址。

[0043] 可替代地,如果分组被从另一个路由器转发进来,该系统可以将前跳路由器的出端口的LID作为源LID,即srcLID。此外,如果需要进一步转发分组,该系统可以将目的地LID,即dstLID替换为下一跳端口的LID。在任一种情况下,该系统可以在将分组转发到下一跳之前计算出循环冗余校验 (CRC)。

[0044] 根据本发明的实施例,不同的方法可以被用于在不同的无限带宽 (IB) 子网间路由流量。这些方法可以被用于回答子网际路由相关的许多问题,比如应该为特定的目的地选择哪个路由器 (第一路由阶段) 和路由器应该选择哪条路径以到达该目的地 (第二路由阶段)。

[0045] 例如,这些方法可以包括简单经典路由方法、为不同拓扑提供良好性能的源路由方法,以及被指定的且为基于胖树的拓扑提供最优性能的最优路由方法。源路由方法和最优路由方法都可以较经典路由方法提供潜在更好的性能。此外,最优路由方法允许为下层胖树拓扑获得最优性能。额外地,两种方法支持使用任意多端口路由器,而经典路由即使在不限限制可用端口的数量时,可能不能有效利用所有端口。

[0046] 因此,使用这些方法,本地IB路由器允许通过将子网连接在一起构建更复杂的IB结构,同时性能上没有明显下降。

[0047] 子网际源路由 (ISSR)

[0048] 根据本发明的实施例,比如子网际源路由 (ISSR) 路由算法的通用路由算法可以被用于路由具有混合子网的复杂网络。该ISSR路由算法可以包括两个阶段,第一阶段用于为特定的目的地选择路由器入端口,且第二阶段用于选择路由器出端口。

[0049] 图3示出了根据本发明的实施例在网络环境中为不同目的地选择路由器入端口的

图解。如图3所示,网络环境300可以包括子网310,该子网与例如路由器A 301和路由器B 302的一或多个路由器相连接。

[0050] 例如子网310中的SM 303的子网管理器(SM)可以包括映射文件304,该文件可以被用于为不同目的地选择路由器入端口。此外,路由器入端口的选择可以基于在连接到局部子网的所有可用路由器间的轮询路径分布。例如,为ISSR算法选择局部路由器端口的find_router()函数可以以类似OpenSM路由算法的方式实现。

[0051] 下面是映射文件304的高水平示例

[0052]

表 1: 用于 ISSR 和 ISFR 算法的映射文件的高水平示例

1: dst_gid1	router A port 1 guid
2: dst_gid2	router B port 1 guid
3: dst_gid3	router A port 2 guid
4: dst_gid4	router B port 2 guid
5: #default route	
6: *	router A port 1 guid
7: *	router B port 1 guid

[0053] OpenSM路由可以只将一整个子网与单独路由器端口匹配,与OpenSM路由不同,该系统可以将目的地末端端口映射到不同的路由器端口。此外,该映射文件的设置可以与OpenSM中提供的映射文件不同。如上表1所示,映射文件304包含完全合格的端口GID,而不仅仅是如同用于OpenSM子网际路由的子网前缀,该端口GID允许系统在选择路由器入端口时提供全粒度。

[0054] 此外,如图3所示,相等(或近似)数量的目的地可以以例如轮询方式被映射到多个端口。例如,目的地A 311(dst_gid1)和目的地C 313(dst_gid3)可以通过路由器A 301上的端口1和端口2被路由,且目的地B 312(dst_gid2)和目的地D 314(dst_gid4)可以通过路由器B302上的端口1和端口2被路由。另外地,映射文件304可以指定备用和默认路线。

[0055] 根据本发明的实施例,路由器出端口的选择可以在路由器固件上使用基于模的散列被实现。散列可以采用使用源和目的地LID产生的随机数(或其他任意有用的参数,比如路由器上的入端口数量)。路由器输出端口中的一个可以使用散列被选择。此外,ISSR路由算法是一直为同一节点对使用同一路径的确定性遗忘路由算法。

[0056] 此外,由于路由器可以被附着到多于两个的子网,所以可以使用两步端口验证方法以选择路由器出端口。首先,该系统可以选择附着到目的地所在的子网(或子网方向)的一组可能端口。之后,该系统可以使用基于模的简单散列函数来选择出端口。

[0057] 图4示出了根据本发明的实施例在网络环境中路由器处为接收到的分组选择路由器端口的图解。如图4所示,网络环境400可以包括例如路由器A 401的路由器,该路由器可以将一或多个分组路由到不同子网中的多个目的地。

[0058] 例如,路由器A 401可以在端口1接收来自源A 411的分组,该分组去往子网403中

的目的地X 421。另外,路由器A 401可以在端口2接收来自源C 413的另一个分组,该分组去往子网402中的目的地Y 422。

[0059] 下面用于选择路由器出端口的算法1可以在SM和路由器固件中实现。

[0060]

算法 1: ISSR 中 choose_egress_port()函数

```
1: if received_intersubnet_packet()then
2:     dstLID = get_next_LID(dGID)
3:     srand(srcLID+dstLID)
4:     port_set=choose_possible_out_ports()
5:     e_port = port_set[(rand()%por_set:size)]
6: end if
```

[0061] 该路由决策可以基于源LID和目的地LID。源LID可以是初始源或者在传输子网情况下前一跳路由器的出端口中的任一个,而目的地LID可以是最终目的地LID或下一跳的路由器入端口的LID的任一个。

[0062] 因为路由器A 401可以看到所附着的子网,所以它知道源LID和目的地LID的值。为了获得目的地LID,需要比如算法1(行2)中的get_next_LID()的映射函数,该函数可以被用于将目的地GID映射到目的地LID或基于位于GID中的子网前缀返回下一跳LID。该映射函数可以执行以下操作的一或多个:内容可寻址的存储器查找,从全局路由报头(GRH)字段解码最终目的地局部标识符(LID),从全局路由报头(GRH)字段解码下一跳局部标识符(LID)。

[0063] 根据本发明的实施例,选择路由器出端口的算法可以基于源和目的地LID计算随机数。该随机数可以被用于从一组可能端口中选择单个出端口。随机数的生成可以以确定性的方式完成以使得给定的源-目的地对可以一直产生同样的数值。因此,该系统可以当在子网间路由时阻止乱序投递,并且,不像用于选择出端口的轮询方法,该系统确保每一源-目的地对一直使用同一路径通过网络。

[0064] 图5示出了根据本发明的实施例在网络环境中路由器处用于支持子网际源路由(ISSR)算法的示例性流程图。如图5所示,在步骤501路由器可以接收该路由器负责路由一或多个分组所到的目的地的列表,其中该路由器连接到网络环境中的至少一个子网。然后,在步骤502,该路由器可以基于与该一或多个分组相关联的源本地标识符(LID)和目的地LID产生随机数。此外,在步骤503,该路由器可以使用基于模的散列以从路由器的输出路由器端口选择一个端口。

[0065] 子网际胖树路由(ISFR)

[0066] 根据本发明的实施例,比如子网际胖树路由(ISFR)路由算法的优化路由方法可以被用于在具有胖树拓扑的子网间路由。

[0067] 该优化路由方法可以包括三个阶段,比如在所有连接到局部子网的可用路由器间的轮询路径分配的第一阶段、询问连接到每个路由器的下行交换机以了解哪个交换机充当去往特定目的地的主要路径的第二阶段、使用从第一阶段和第二阶段获得的信息构建路由

表的第三阶段。

[0068] 图6示出了根据本发明的实施例在网络环境600中路由器处支持子网际胖树路由 (ISFR) 算法的图解。如图6所示, 路由器601负责将分组650路由到子网610, 该子网连接到一或多个末端节点621-624。该子网610可以包括一或多个交换机, 比如在胖树拓扑中的SW611-616。

[0069] 路由器601能够了解它的哪个端口可以被用于根据相关联的GID将分组路由到特定的目的地。例如, 路由器601可以从子网610中的子网管理器 (SM) 620接收端口映射630。端口映射630可以包括该路由器601负责路由源自子网610的分组的目的地列表。

[0070] 此外, 路由器601可以从交换机611-612了解使用哪个下行输出端口以在路由分组650时获得最大性能。例如, 路由器601可以为到子网610中的目的地节点的主要路径询问子网610中的交换机611-612。该询问允许路由器为分组650选择主要路径 (交换机611或交换机612), 因为这些交换机中只有一个可以具有去往所希望目的地的专用主要路径。

[0071] 之后, 路由器601可以使用以上获得的信息构建路由表640。例如, 本地OpenSM路由可以在交换机611或交换机612的路由表中标记到目的地节点的主要路径。

[0072] 下面用于选择路由器出端口的算法2可以在路由器固件中实现。

[0073]

算法 2: ISSR 中 query_down_for_egree_port()函数

```
1: if received_mapping_files then
2:   for all_port_in_down_ports do
3:     down_switch = get_node(port)
4:     lid = get_LID_by_GID(GID)
5:     if down_switch.routing_table[lid] ==
       primary_path then
6:       e_port = port
7:     end if
8:   end for
9: end
```

[0074] 以上ISFR算法可以使用之前定义的含有表1中GID到路由器端口映射的文件格式。像ISSR算法, ISFR算法可以在路由器装置中被实现。此外, ISFR算法可以只在胖树上工作并具有在每个子网中局部运行的胖树路由。而且, 如果需要, ISFR算法可以回退到ISSR。

[0075] 如图6所示, 路由器601可以被展示为位于适当的胖树拓扑顶部。因此, 在询问被执行之后, 路由器601可以在位于特定子网的每一个目的地在下行方向每个端口具有一条路径。在超额认购 (oversubscribed) 的胖树的情况下, 路径的数量可以等于超额认购的比率。

[0076] 另外, 可以存在多于一个的路由器连接到子网610。ISFR路由的性质使得如果所有脊 (spine) (顶) 路由器被替换为交换机, 用于ISFR路由 (使用路由器) 和用于局部路由 (使用

交换机)的路由表是相同的。

[0077] 根据本发明的实施例,ISFR路由算法可以基于在通过不透明路由器被连接的不同子网的子网管理器(SM)间建立的通信被实现。例如,可以在路由器中提供界面,SM可以穿透该界面通信,且通过该界面可以在位于相邻子网中的两个SM间实现握手。

[0078] 图7-9图解了根据本发明的实施例利用不同拓扑在网络环境中路由。每个拓扑可以被展示为例如在顶部放置路由器的三级胖树的多级胖树。例如,该三级胖树可以具有三个路由/交换级和一个节点级,每个子网(二级胖树)作为分支出现。此外,更大的拓扑也可以基于该ISFR路由算法被支持。

[0079] 根据本发明的实施例,网络环境可以被配置使得每个子网是胖树拓扑,该胖树拓扑可以中间没有传输子网地附着到其他子网。

[0080] 例如,在图7中,系统700可以使用六个路由器710连接两个胖树子网,例如子网A-B 701-702。在图8中,系统800可以使用六个路由器810连接三个胖树子网,例如子网A-C 801-803。在图9中,系统900可以使用十八个路由器910产生648个端口的三级胖树以连接六个胖树子网901-906。

[0081] 图10图解了根据本发明的实施例的在网络环境中路由器处用于支持子网际胖树路由(ISFR)算法的示例性流程图。如图10所示,在步骤1001,路由器可以接收该路由器负责路由一或多个分组所到的目的地的列表,其中该路由器连接到网络环境中的至少一个子网。之后,在步骤1002,该路由器可以从该至少一个子网中的一或多个交换机获得信息,该信息关于路由器的哪个下行输出端口可以被用于路由该一或多个分组。此外,在步骤1003,该路由器可以基于所获得的信息构建路由表。

[0082] 图11示出了功能性框图以示出本发明的实施例的特征。所呈现的特征可以被实现为用于在网络环境中不同子网间路由流量的系统1100。该系统1100包括一或多个微处理器和在该一或多个微处理器上运行的路由器1110。路由器1110操作为从接收单元1120接收该路由器负责路由一或多个分组所到的目的地的列表,其中路由器1110连接到网络环境中的至少一个子网;从获取单元1130从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息;和在构建单元1140上基于所获取的信息构建路由表。

[0083] 根据一个实施例,公开了一种用于在一或多个微处理器上的在网络环境中不同子网间路由流量的系统。该系统包括用于在路由器处接收该路由器负责路由一或多个分组所到的目的地的列表的装置,其中该路由器连接到网络环境中的至少一个子网。该系统包括用于从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息的装置。该系统包括用于基于所获取的信息构建路由表的装置。

[0084] 优选地,该系统包括用于在连接到至少一个子网的所有可用路由器间执行轮询路径分配的装置。

[0085] 优选地,该系统包括用于查询连接到每个路由器的交换机以了解哪个交换机能够用作去往特定目的地的主要路径的装置。

[0086] 优选地,该系统包括用于基于所获取的信息确定为了最大性能能够使用路由器的哪些向下输出端口的装置。

[0087] 优选地,该系统包括用于从子网管理器接收一或多个目的地全局标识符(GID)和一或多个路由器端口之间的端口映射的装置。

[0088] 优选地,该系统包括用于使用一或多个附加的路由器在网络环境中将该至少一个子网连接到一或多个相邻子网的装置。

[0089] 优选地,该系统包括用于允许所连接的子网成为不同拓扑的装置。

[0090] 优选地,该系统包括用于允许所每个所连接的子网成为多级胖树中的分支的装置。

[0091] 优选地,该系统包括用于允许该至少一个子网以胖树拓扑被配置的装置。

[0092] 优选地,该系统包括用于使用胖树路由算法路由该至少一个子网的装置。

[0093] 根据一个实施例,公开了用于在网络环境中不同子网间路由流量的系统。该系统包括路由器,所述路由器被配置为接收该路由器负责路由一或多个分组所到的目的地的列表,其中该路由器连接到网络环境中的至少一个子网。该路由器被配置为从该至少一个子网中的一或多个交换机获取关于路由器的哪些向下输出端口能够用于路由该一或多个分组的信息,并且还被配置为基于所获取的信息构建路由表。

[0094] 优选地,该系统能够使得循环路径分配在连接到该至少一个子网的所有可用路由器间执行。

[0095] 优选地,该系统能够使得所述路由器操作为查询连接到所述路由器的交换机以了解哪个交换机能够用作去往特定目的地的主要路径。

[0096] 优选地,该系统能够使得所述路由器操作位基于所获取的信息确定为了最大性能能够使用路由器的哪些向下输出端口。

[0097] 优选地,该系统能够使得所述路由器操作以从子网管理器接收在一或多个目的地全局标识符(GID)和一或多个路由器端口间的端口映射。

[0098] 优选地,该系统能够使得该至少一个子网使用一或多个附加的路由器连接到该网络环境中的一或多个相邻子网。

[0099] 优选地,该系统能够使得所连接的子网为不同拓扑。

[0100] 优选地,该系统能够使得每个连接的子网是多级胖树的分支。

[0101] 优选地,该系统能够使得该至少一个子网以胖树拓扑被配置,并且能够使用胖树路由算法用于路由该至少一个子网。

[0102] 本发明可以使用一或多个传统通用或专用数字计算机、计算设备、机器或微处理器,包括一或多个处理器、存储器和/或根据本公开的教导编程的计算机可读存储介质方便地实现。如对于软件领域技术人员将是明显地,合适的软件编码可以由熟练的编程人员基于本公开的教导容易地被准备。

[0103] 在一些实施例中,本发明包括计算机程序产品,该产品是在其上/其中存储有可以被用于编程计算机以执行本发明任意过程的指令的计算机可读(多个)介质。该存储介质可以包括但并不限于任何种类的盘,包括软盘、光盘、DVD、CD-ROM、微驱动、磁-光盘、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、闪存装置、磁卡或光卡、纳米系统(包括分子存储器IC)、或适合于保存指令和/或数据的任意种类的介质或装置。

[0104] 出于举例和说明的目的,提供了本发明的上述说明。上述说明并不是详尽的,也不旨在将本发明局限于公开的具体形式。对本领域的技术人员来说,许多修改和变化是明显

的。选择和说明实施例是为了最好地解释本发明的原理及其实际应用,从而使本领域的其他技术人员能够理解本发明的多个实施例和适合于预期的特定应用的各种修改。本发明的范围由以下权利要求及其等同物限定。

100

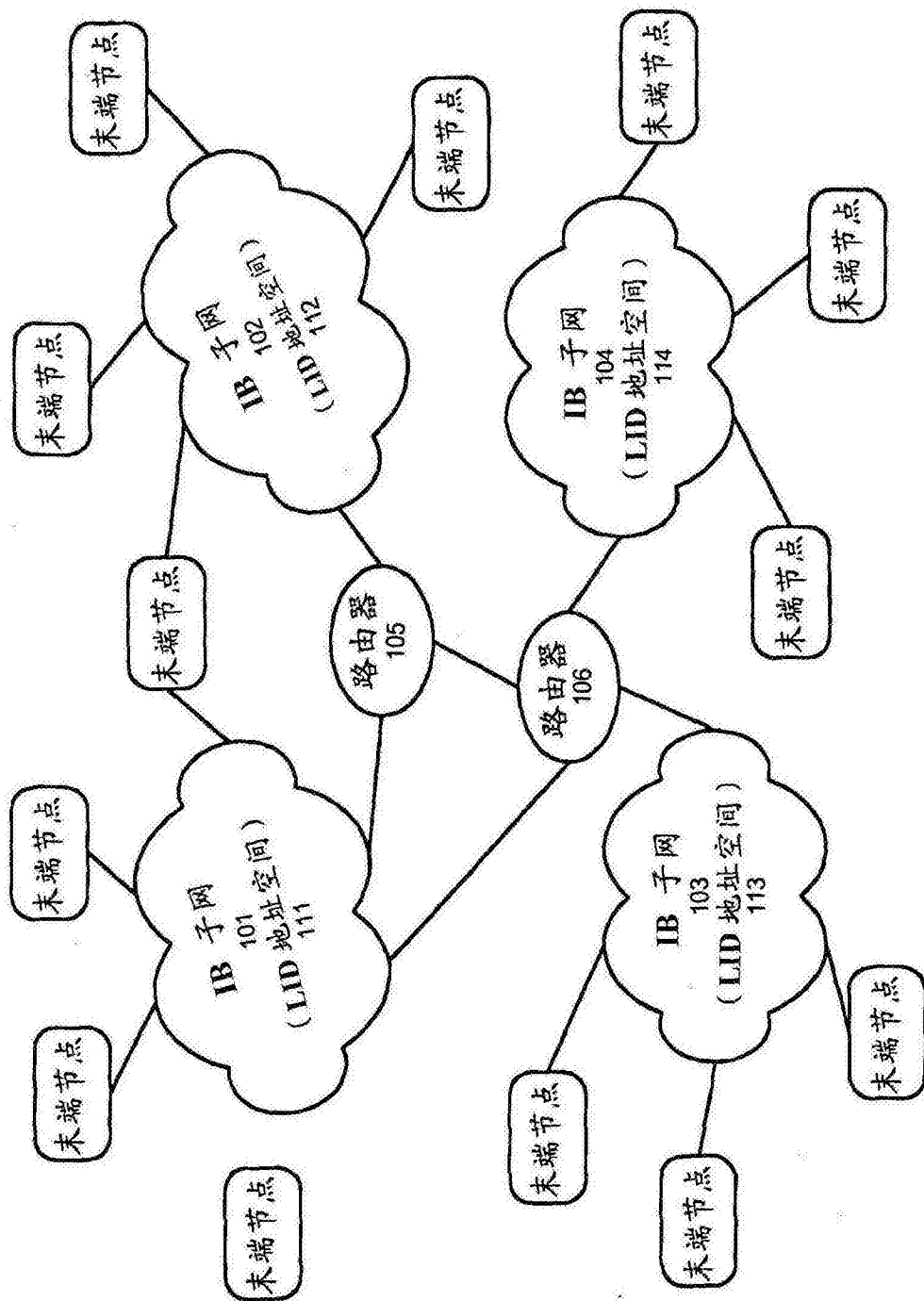


图1

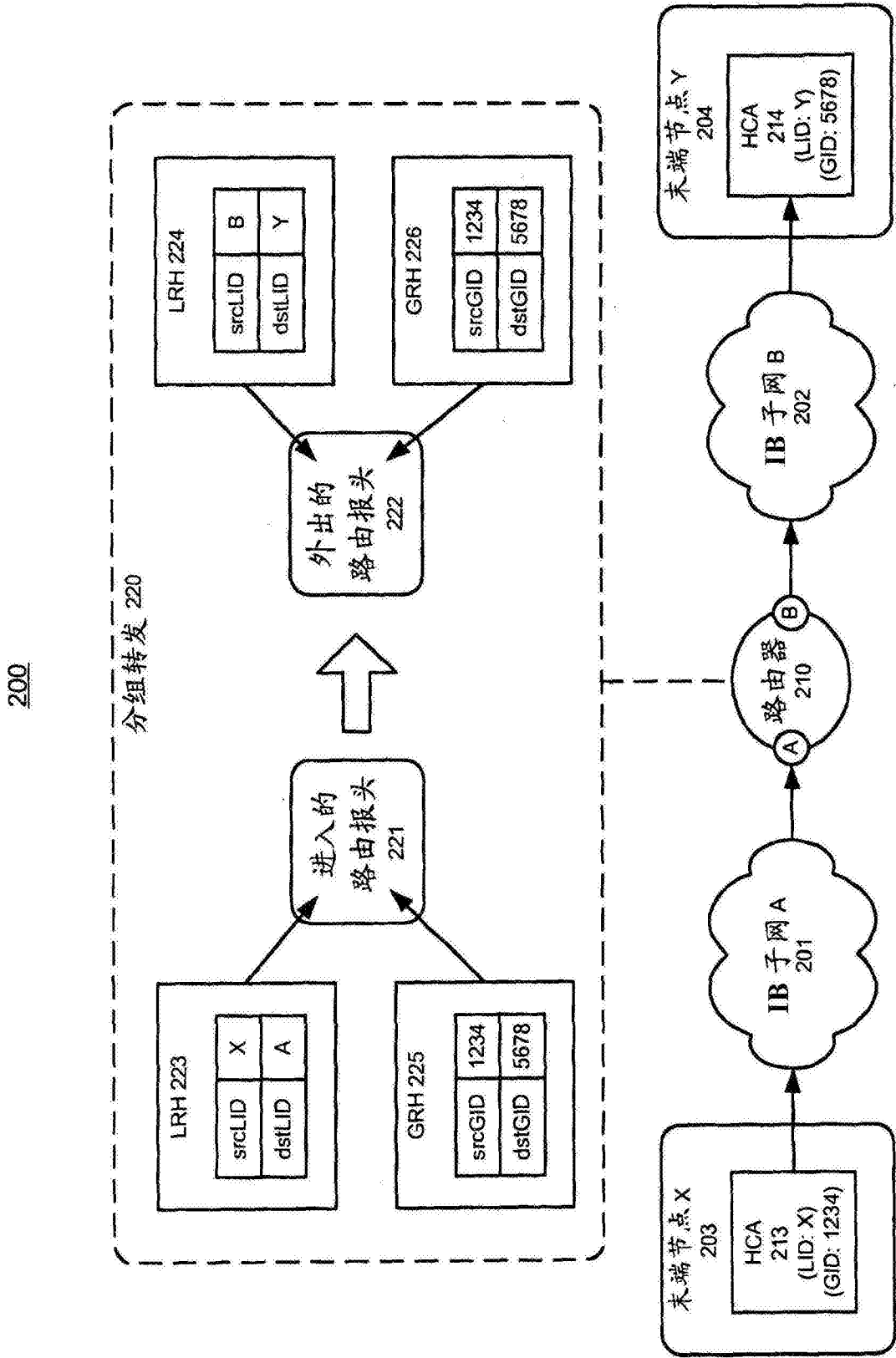


图2

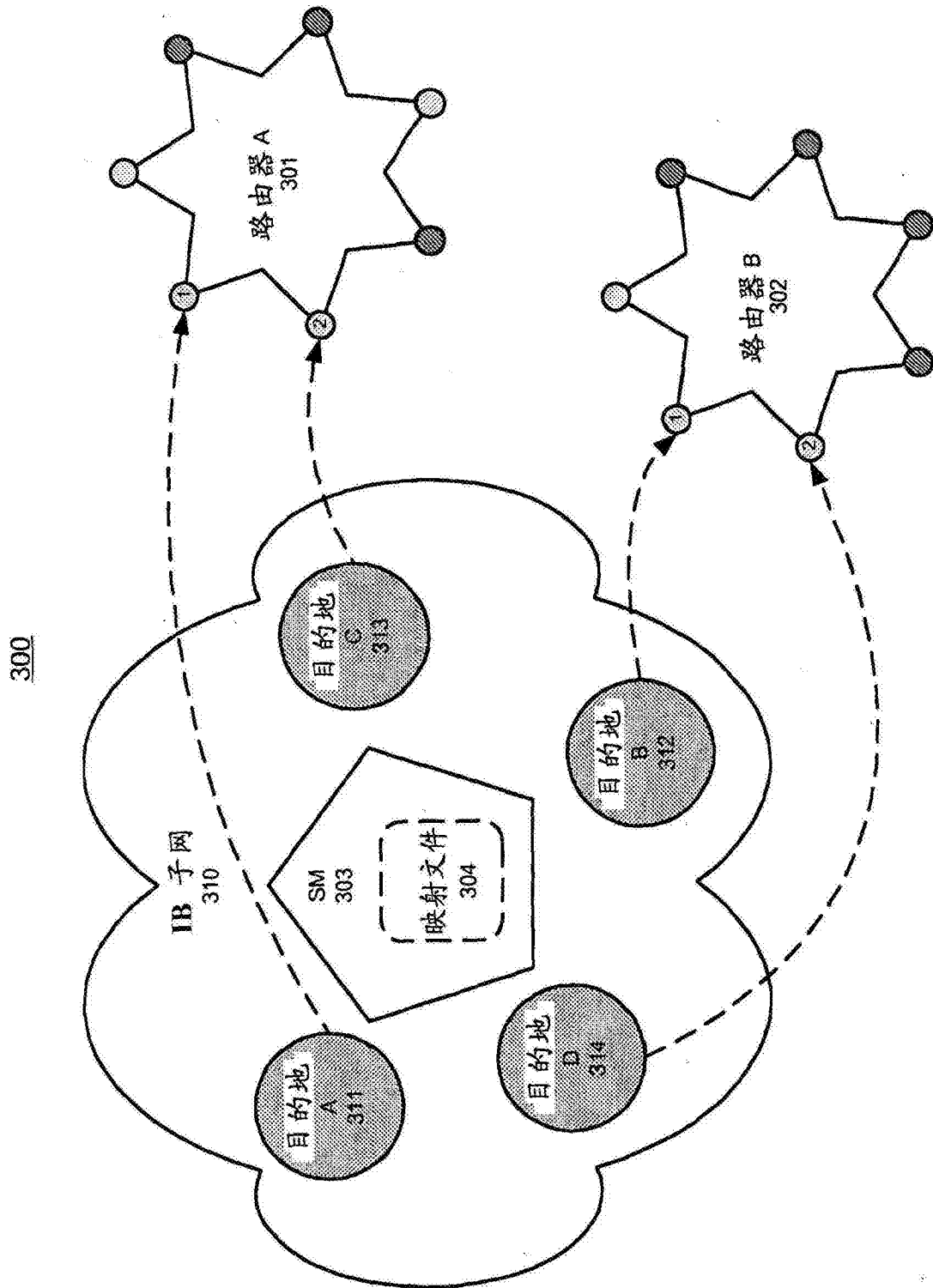


图3

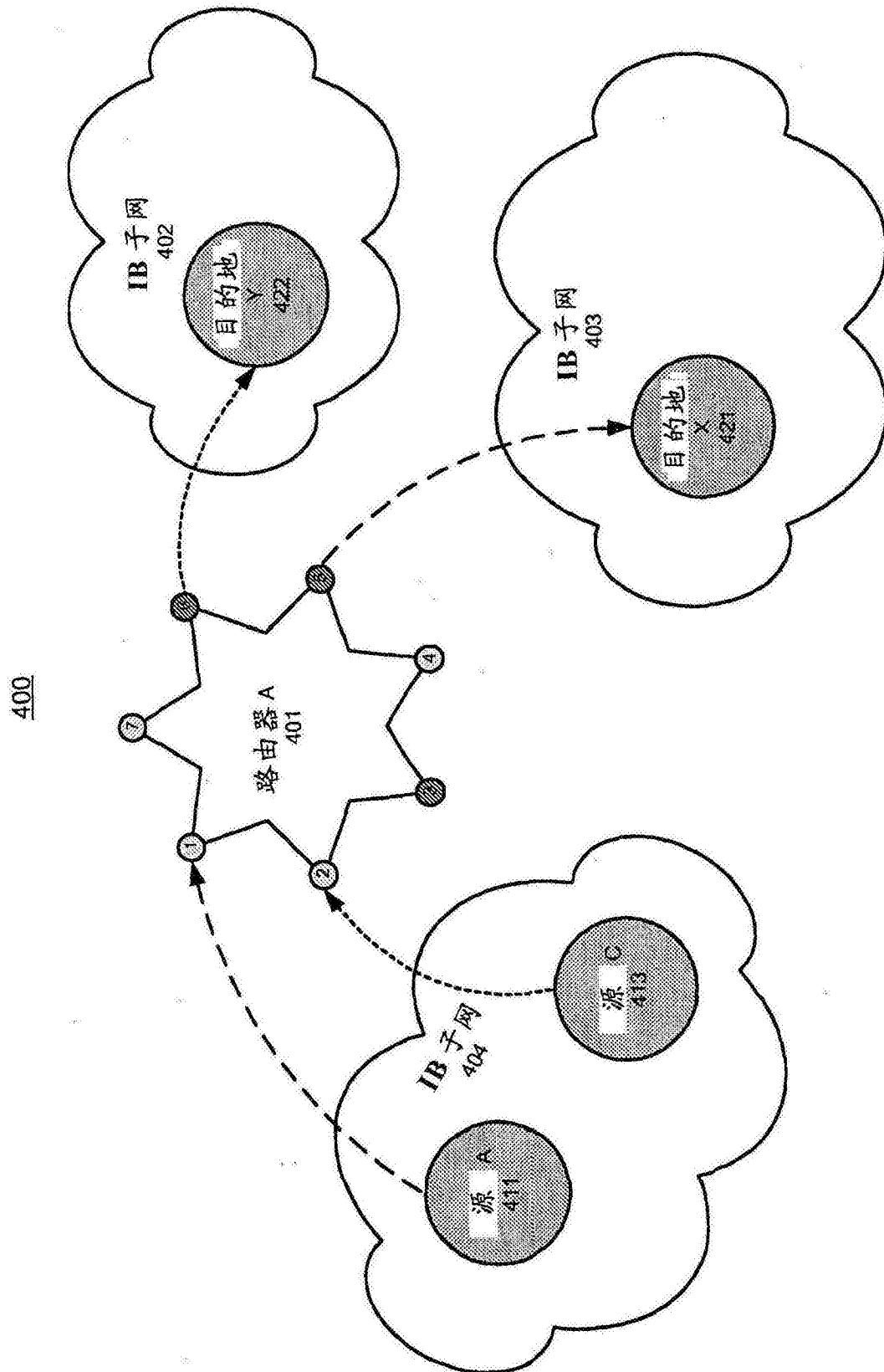


图4

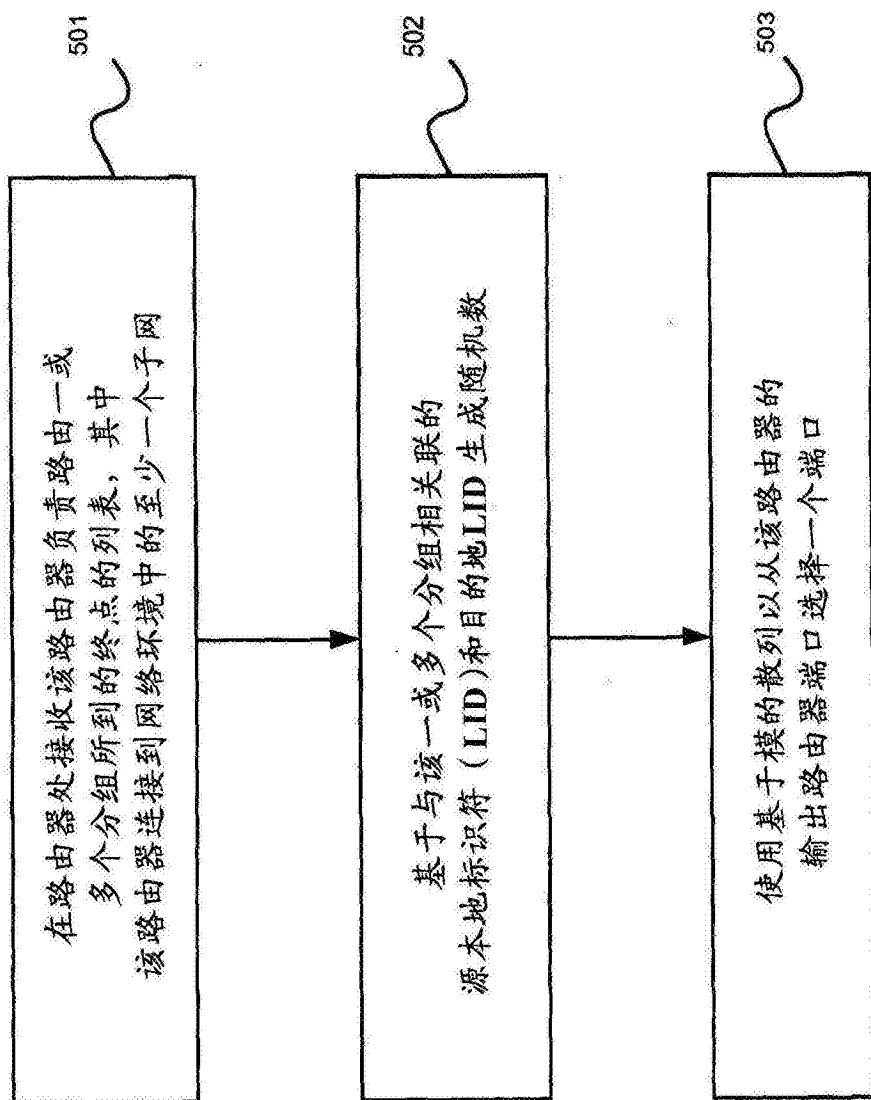


图5

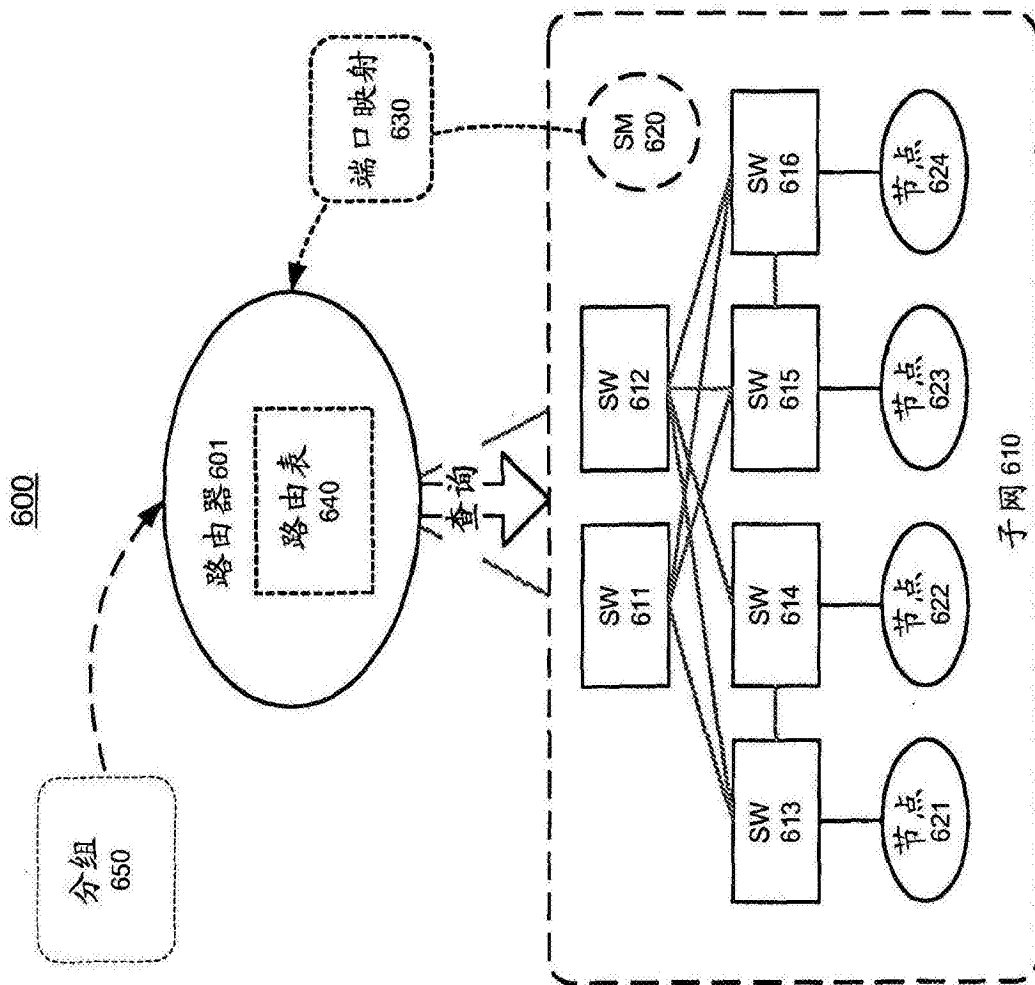


图6

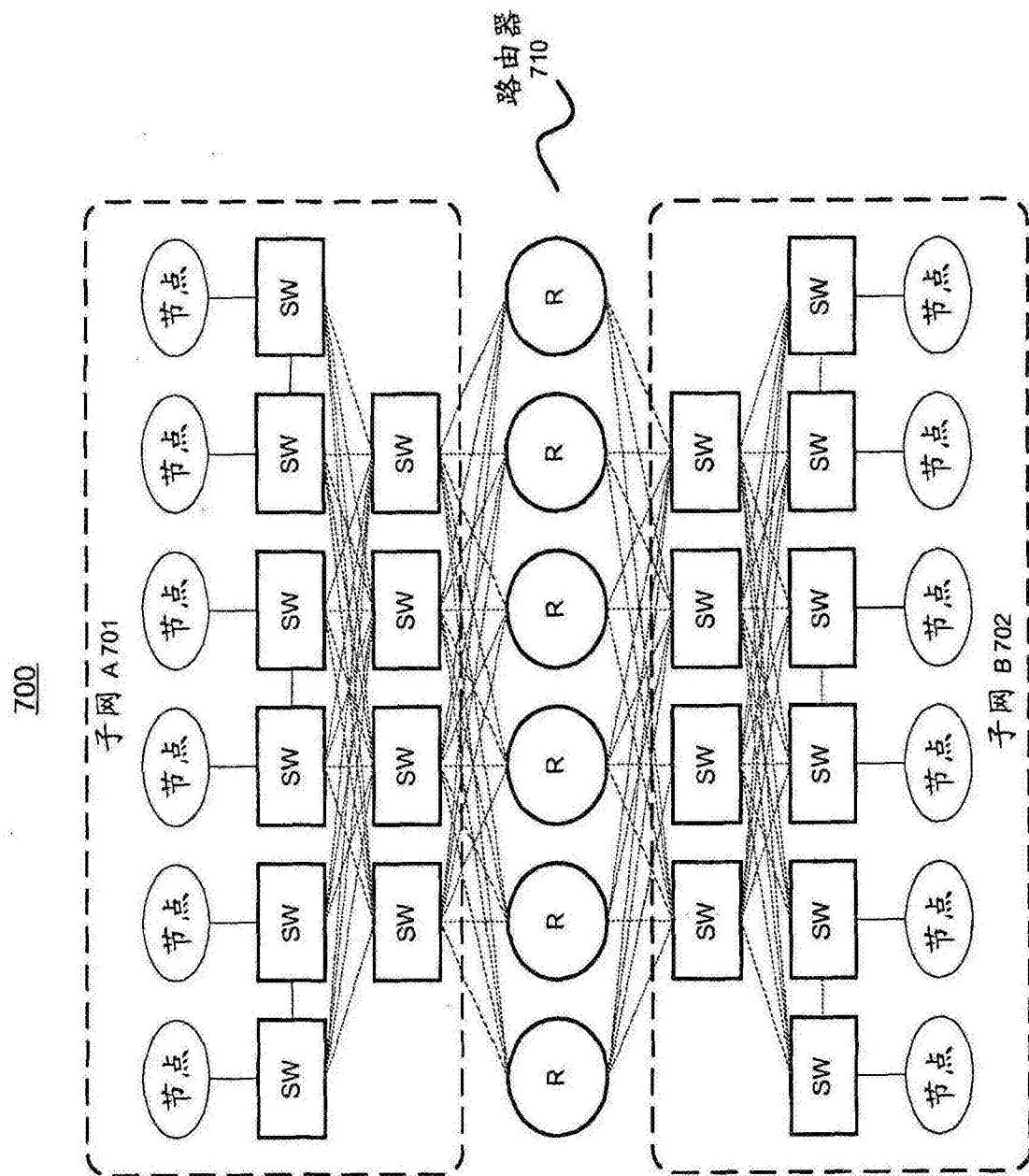


图7

800

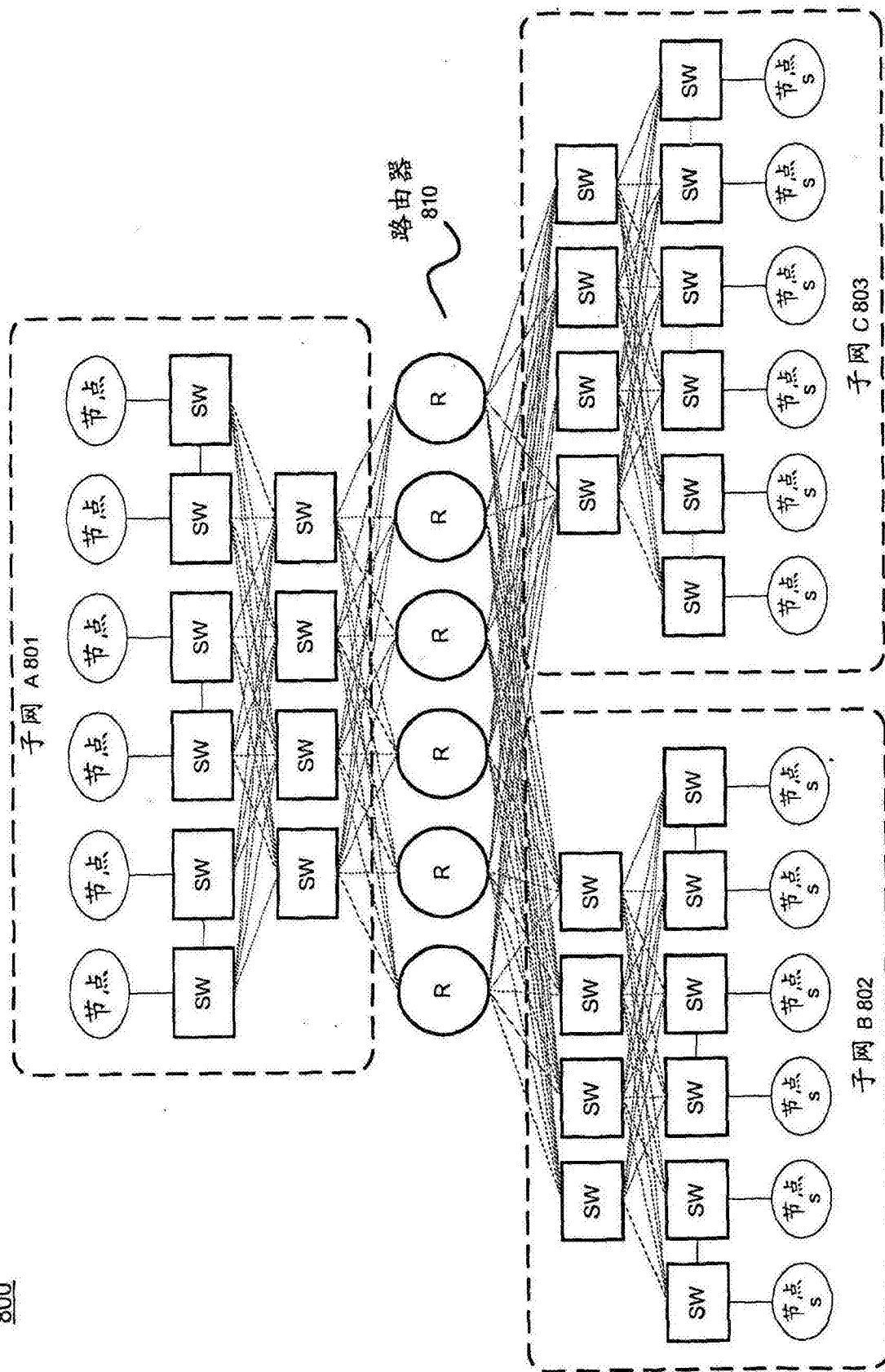


图8

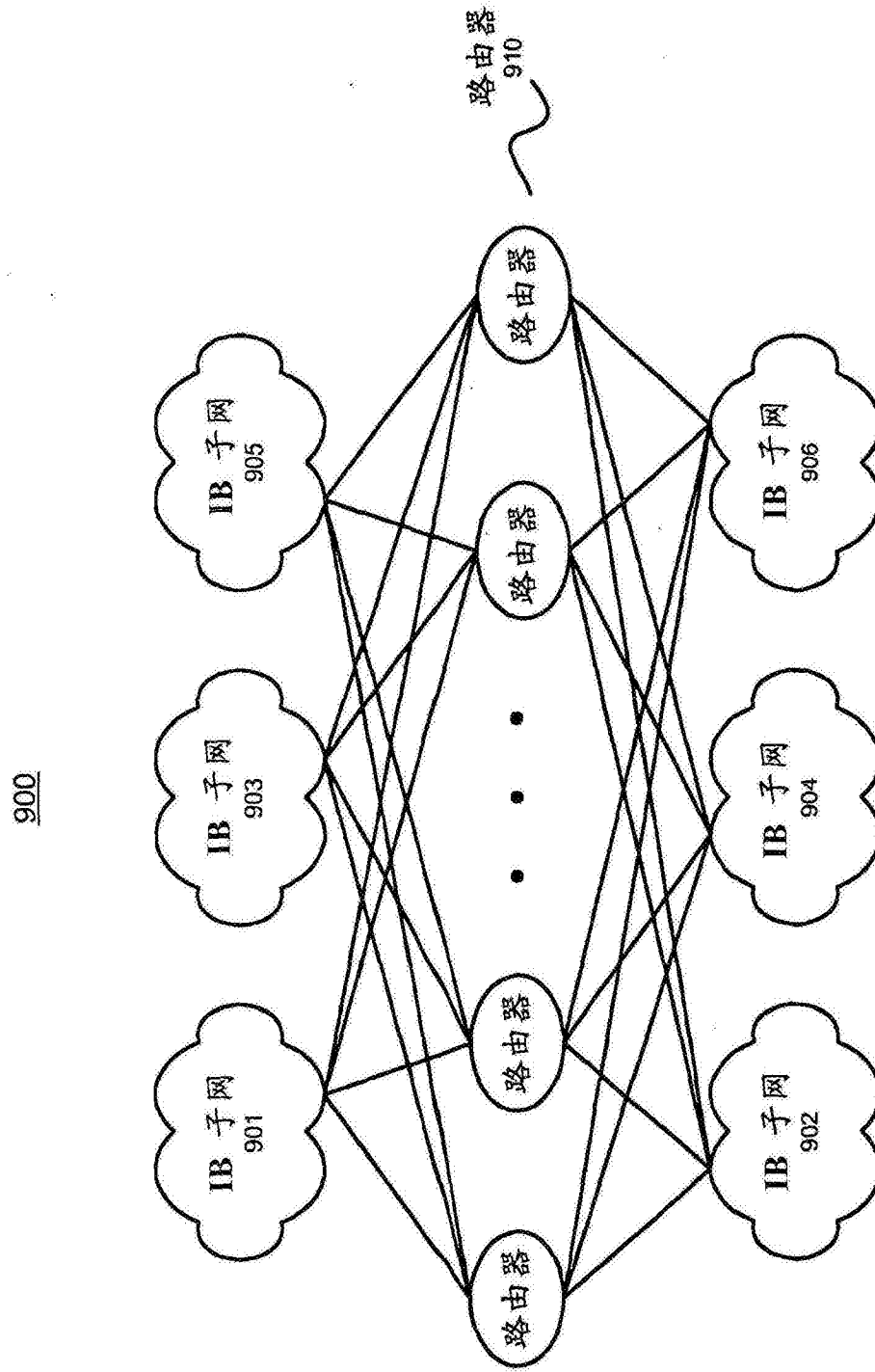


图9

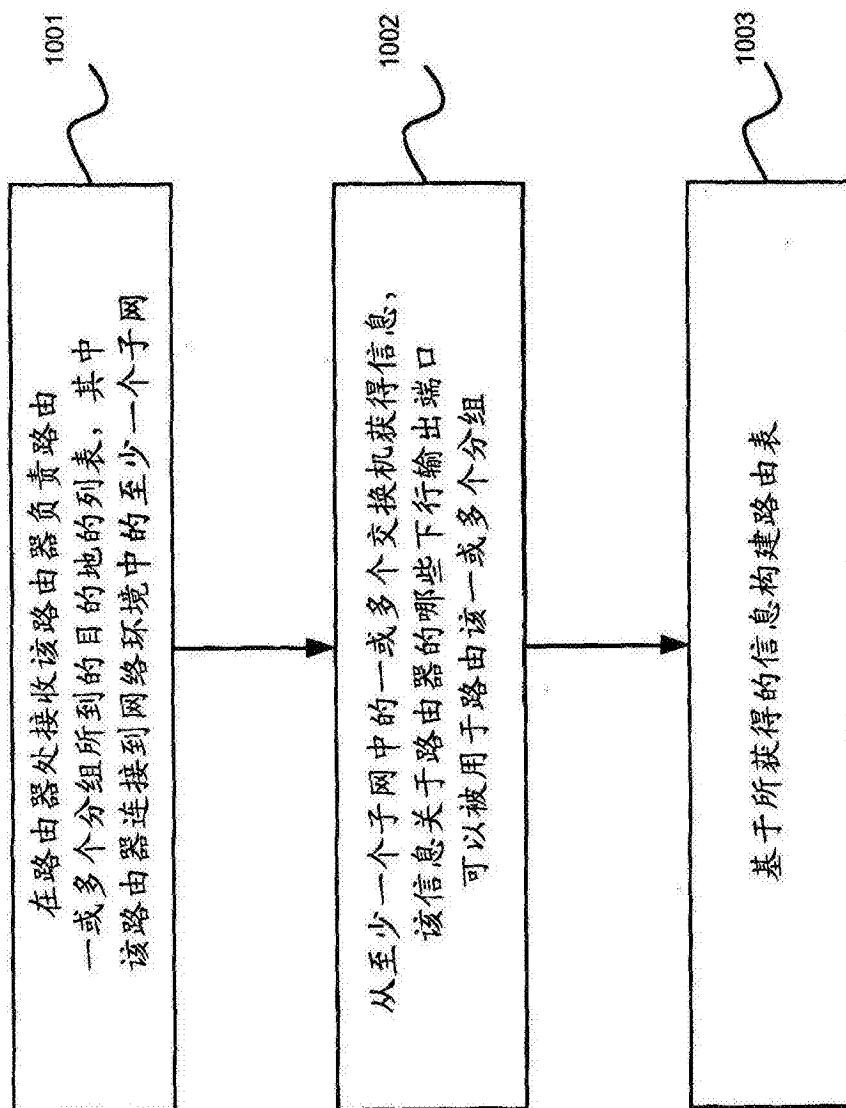


图10

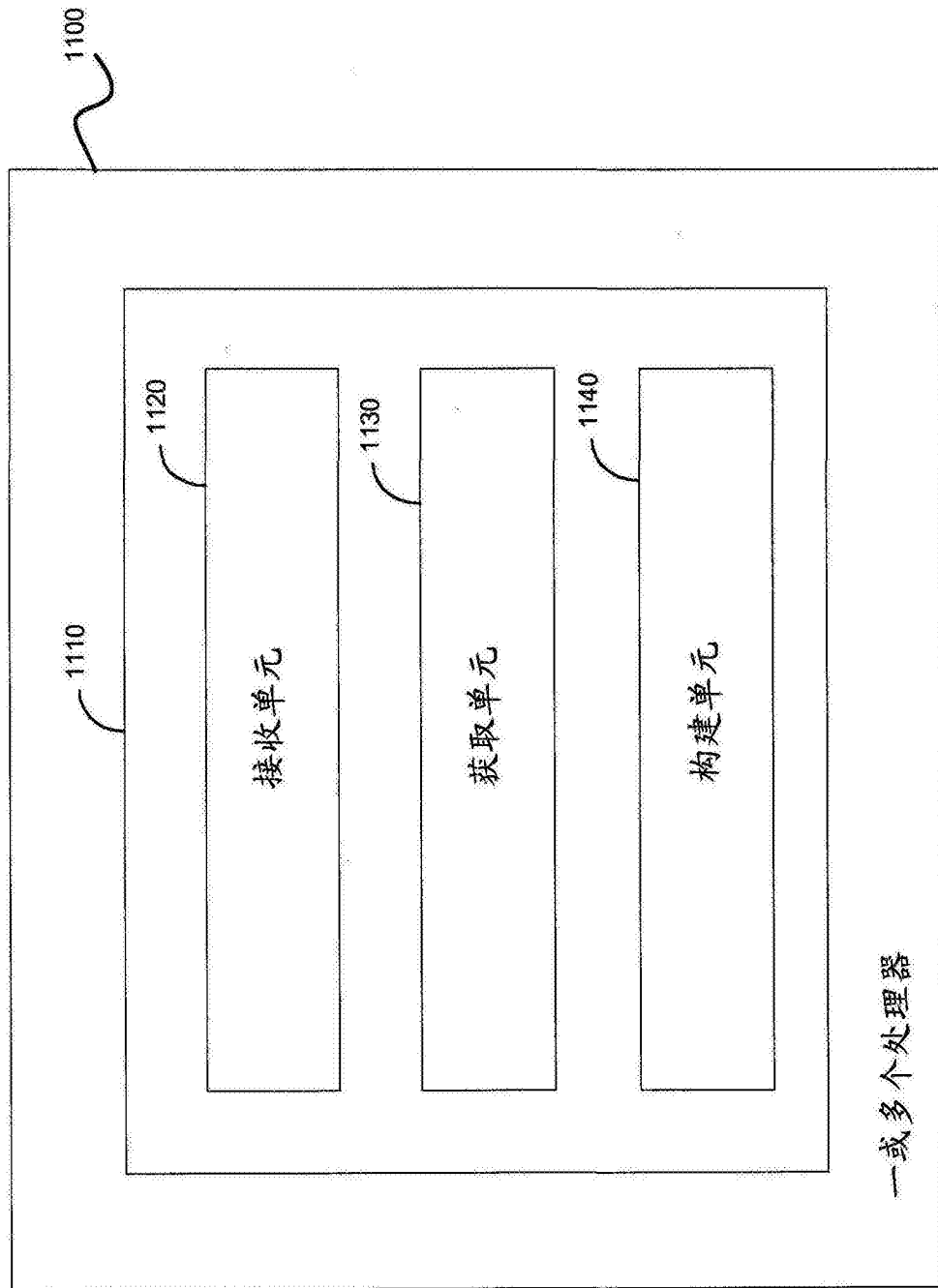


图11