



# [12] 发明专利说明书

专利号 ZL 200410058622.6

[45] 授权公告日 2006 年 10 月 11 日

[11] 授权公告号 CN 1279732C

[22] 申请日 2004.7.23

[21] 申请号 200410058622.6

[30] 优先权

[32] 2003.7.31 [33] US [31] 10/631,053

[71] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 维维科·卡斯雅普

格雷戈里·弗朗西斯·费斯特

审查员 毛习文

[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

代理人 康建忠

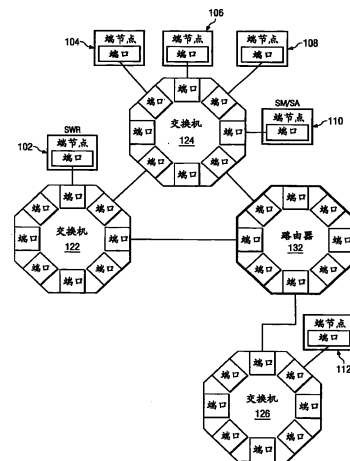
权利要求书 3 页 说明书 9 页 附图 3 页

## [54] 发明名称

多播组管理的方法和设备

## [57] 摘要

提供一种在产生和删除多播组时，不使用俘获的情况下，管理具有发送不接收(SWR)加入者的多播组的机制。当 SWR 成员存在时，多播组信息被连续保持。当尝试 SWR 加入，并且多播组还未存在时，组信息(MLID)被标记成已用，路由 SWR 分组被发往的第一交换机丢弃发送给该组的所有分组。当接收成员加入该组时，路由被更新，从而 SWR 成员开始向接收成员发送。当最后的接收成员离开该组时，再次为第一交换机提供路由以丢弃分组。



- 1、一种管理系统区域网中的多播组的方法，所述方法包括：  
接收来自一节点的加入多播组的加入请求，其中所述节点与第一交换机连接，并且其中加入请求是发送-不接收请求；  
确定多播组是否存在；和  
如果多播组不存在，则产生多播组，并为第一交换机提供路由以丢弃给该多播组的所有分组。
- 2、按照权利要求1所述的方法，其中产生多播组的步骤包括向多播组分配多播标识符。
- 3、按照权利要求1所述的方法，其中为第一交换机提供路由的步骤包括在第一交换机的多播路由数据结构中，插入关于多播组的一个条目。
- 4、按照权利要求3所述的方法，其中依据多播标识符索引多播路由数据结构。
- 5、按照权利要求3所述的方法，其中关于多播组的条目包括分组将被丢弃的指示。
- 6、按照权利要求1所述的方法，还包括：  
响应来自接收节点的加入请求，更新系统区域网中至少一个交换机的至少一个多播路由表，以便把给多播组的分组发送给接收节点。
- 7、按照权利要求6所述的方法，还包括：  
接收来自另一节点的离开多播组的离开请求；  
确定单个节点是否留在多播组中；和  
如果单个节点留在多播组中，则路由和单个节点最接近的交换机丢弃给多播组的所有分组。
- 8、一种管理系统区域网中的多播组的方法，所述方法包括：  
接收来自第二节点的离开多播组的离开请求，其中多播组具有位于与第一交换机连接的第一节点的第一成员；  
确定单个节点是否留在多播组中；和  
如果单个节点留在多播组中，则为第一交换机提供路由丢弃给多播

组的所有分组。

9、按照权利要求 8 所述的方法，其中第一成员是发送-不接收成员。

10、按照权利要求 8 所述的方法，其中为第一交换机提供路由的步骤包括在第一交换机的多播路由数据结构中，插入关于多播组的一个条目。

11、按照权利要求 10 所述的方法，其中依据多播标识符索引多播路由数据结构。

12、按照权利要求 10 所述的方法，其中关于多播组的条目包括分组将被丢弃的指示。

13、按照权利要求 8 所述的方法，还包括：

响应来自接收节点的加入请求，更新系统区域网中至少一个交换机的至少一个多播路由表，以便把给多播组的分组发送给接收节点。

14、一种管理系统区域网中的多播组的设备，所述设备包括：

接收来自某一节点的加入多播组的加入请求的接收装置，其中所述节点与第一交换机连接，其中加入请求是发送-不接收请求；

确定多播组是否存在的确定装置；和

产生多播组的产生装置；和

如果多播组不存在，则为第一交换机提供路由以丢弃给该多播组的所有分组的路由装置。

15、按照权利要求 14 所述的设备，其中产生装置包括向多播组分配多播标识符的装置。

16、按照权利要求 14 所述的设备，其中路由装置包括在第一交换机的多播路由数据结构中，插入关于多播组的一个条目的装置。

17、按照权利要求 14 所述的设备，还包括：

响应来自接收节点的加入请求，更新系统区域网中至少一个交换机的至少一个多播路由表，以便把给多播组的分组发送给接收节点的装置。

18、一种管理系统区域网中的多播组的设备，所述设备包括：

接收来自第二节点的离开多播组的离开请求的接收装置，其中多播组具有位于与第一交换机连接的第一节点的第一成员；

确定单个节点是否留在多播组中的确定装置；和

如果单个节点留在多播组中，则为第一交换机提供路由以丢弃给多播组的所有分组的路由装置。

19、按照权利要求 18 所述的设备，其中第一成员是发送-不接收成员。

20、按照权利要求 18 所述的设备，其中路由装置包括在第一交换机的多播路由数据结构中，插入关于多播组的一个条目的装置。

21、按照权利要求 18 所述的设备，还包括：

响应来自接收节点的加入请求，更新系统区域网中至少一个交换机的至少一个多播路由表，以便把给多播组的分组发送给接收节点的装置。

## 多播组管理的方法和设备

### 技术领域

本发明涉及改进的数据处理系统，具体地说，涉及系统区域网。更具体地说，本发明提供一种具有发送不接受组成员的多播组管理的方法和设备。

### 背景技术

InfiniBand (IB) (这是系统区域网 (SAN) 的一种形式) 定义一种多播设施，所述多播设施允许信道适配器 (CA) 向单个地址发送分组，并把分组传送给多个端口。在 InfiniBand 标准中描述了 InfiniBand 体系结构，InfiniBand 标准作为参考包含于此。

单播分组从一个节点发送给另一节点。单播分组包括报头中目标节点的唯一地址。路由器和交换机根据所述唯一地址或标识符，把分组发送给目标节点。

相反，多播分组被发送给称为多播组的多个端口的所有端口。这些端口可在 SAN 中的相同或不同节点上。每个多播组由唯一的多播本地标识符 (MLID) 识别。MLID 用于在子网内引导分组。MLID 是 IB 分组的报头。

当节点加入多播组时，使用借助子网管理分组 (SMP) 的 IB 管理操作，此时，节点上端口的 LID 被链接到多播组。子网的子网管理器 (SM) 随后利用 SMP，把该信息保存在其子网的交换机中。借助 SMP，SM 把各个多播组的路由信息告知交换机，交换机保存该信息，从而交换机能够把多播分组发送给正确的节点。

当节点将要向多播组发送分组时，它使用它希望分组被传送的那个多播组的 MLID。子网中的交换机检测分组的目的地本地标识符 (DLID) 字段中的 MLID，并复制该分组，将其发送给先前 SM 确定的恰当端口。

多播组成员可在不接收的情况下发送分组。例如通常需要称为发送不接收 (SWR, send-without-receive) 成员的这些组成员流出数据多播，或者与其它常见多播实现例如网际协议 (IP) 多播的兼容性。

诸如 InfiniBand 之类交换媒介并不自动允许参与者在未加入组的情况下进行发送。所有通信必须由交换部件明确路由，包括发送数据而不接收数据。当发送加入请求时，SM 程控交换机，把多播分组转发给已请

求加入组并接收分组的节点。

但是，当 SWR 成员最初加入某一组，并且该组未存在时，则存在 SWR 成员在无任何接收器的情况下进行发送的问题。现在，IB 体系结构不产生该组。相反，SWR 加入者必须签字参加，以接收每当产生任意组时发出的俘获 (trap) 消息。SWR 随后检查每个俘获消息，了解已产生哪一组。当它发现产生了所关心的一组时，SWR 加入者可重复其请求，以便具有一定成功希望地加入该组。通过向与 SM 相联系的称为“子网管理机构”(SA) 的实体发送消息，完成接收俘获消息的“签字参加”(signing up)。当已成功加入该组时，SWR 加入者通常通过发送请求删除其对这些俘获消息的预定的另一消息，消除其对这些俘获消息的预定。

另外，当最后的接收成员离开该组时，IB 体系结构通常删除该组，即使 SWR 仍在发送。于是，SWR 必须签名以接收额外的俘获消息，该俘获消息用信号通知任意组的删除，并不断检查这些俘获消息，查看它所关心的组是否已被删除。在发现该删除之后，SWR 随后必须清除其关于该组的 MLID 信息，因为 SM 可能把相同的 MLID 值重新用于不同的组。否则 SWR 可能把分组发送给错误的组。

当 SWR 正在向其发送的组被删除时，SWR 必须再次签字参加，以便每当产生一个组时接收俘获消息，重复该过程，直到 SWR 停止向该组发送为止。这样，只有当存在接收者时，SWR 才加入组中，当不存在接收者时，SWR 被强制等待。

但是，对于 SM 和 SWR 加入者来说，这种过程导致很大的开销。SWR 接收所产生的每个组的一条消息，不论该组是否是它所关心的组。SWR 还必须接收每个被删除组的消息，不仅仅只在关心的特定组被删除时。每当 SWR 试图向组发送时，SM 产生这些消息，SWR 加入者接收这些消息。

于是，有利的是提供一种 InfiniBand 中多播组管理的改进方法和设备。

## 发明内容

本发明提供了一种管理系统区域网中的多播组的方法，所述方法包括：接收来自一节点的加入多播组的加入请求，其中所述节点与第一交换机连接，并且其中加入请求是发送-不接收请求；确定多播组是否存在；和如果多播组不存在，则产生多播组，并为第一交换机提供路由以丢弃给该多播组的所有分组。

本发明还提供了一种管理系统区域网中的多播组的方法，所述方法包括：接收来自第二节点的离开多播组的离开请求，其中多播组具有位

于与第一交换机连接的第一节点的第一成员；确定单个节点是否留在多播组中；和如果单个节点留在多播组中，则为第一交换机提供路由丢弃给多播组的所有分组。

本发明还提供了一种管理系统区域网中的多播组的设备，所述设备包括：接收来自某一节点的加入多播组的加入请求的接收装置，其中所述节点与第一交换机连接，其中加入请求是发送-不接收请求；确定多播组是否存在的确定装置；产生多播组的产生装置；和如果多播组不存在，则为第一交换机提供路由以丢弃给该多播组的所有分组的的路由装置。

本发明还提供了一种管理系统区域网中的多播组的设备，所述设备包括：接收来自第二节点的离开多播组的离开请求的接收装置，其中多播组具有位于与第一交换机连接的第一节点的第一成员；确定单个节点是否留在多播组中的确定装置；和如果单个节点留在多播组中，则为第一交换机提供路由以丢弃给多播组的所有分组的的路由装置。

本发明提供一种在产生和删除多播组时，不使用俘获的情况下，管理具有发送不接收（SWR）加入者的多播组的方法和设备。当 SWR 成员存在时，本发明的机制连续保持组信息。当尝试 SWR 加入，并且多播组还未存在时，组信息（MLID）被标记成已用，路由 SWR 分组被发往的第一交换机以丢弃发送给该组的所有分组。当接收成员加入该组时，路由选择被更新，从而 SWR 成员开始向接收成员发送。当最后的接收成员离开该组时，再次为第一交换机提供路由以丢弃分组。

#### 附图说明

在附加权利要求中陈述了本发明特有的新特征。但是，结合附图，参考例证实施例的下述详细说明，将更好地理解发明本身，及其优选模式，其它目的和优点，其中：

图 1 是根据本发明的优选实施例的系统区域网的一个例子；

图 2 根据本发明的优选实施例，图解说明了交换机；

图 3A-3D 根据本发明的优选实施例，图解说明了例证的多播路由数据结构；

图 4A 是根据本发明的优选实施例，图解说明多播组加入请求的处理的流程图；和

图 4B 是根据本发明的优选实施例，图解说明多播组离开请求的处理的流程图。

#### 具体实施方式

参见图 1, 根据本发明的优选实施例, 图解说明了系统区域网 (SAN) 的一个例子。系统区域网由多个端节点 102-112 构成。这些端节点通过通信链路, 一个或多个交换机 122、124、126, 和一个或多个路由器 132 相互耦接。交换机是把分组从一个链路发送给同一子网的另一链路的设备。路由器是在网络子网之间发送分组的设备。端节点是网络中为分组最终目的地的节点。

在图 1 中所示的网络中, 端节点 110 被表示成包含子网管理器 (SM) 和子网管理机构 (SA)。这些对应于 (1) SM, 只发送和接收能够影响路由和网络硬件配置的特殊消息的实体; 和 (2) SA, 只发送和接收不能影响网络配置的常规通信消息的实体之间 SAN 管理功能的 InfiniBand 体系结构的划分。SA 被用作利用常规消息与 SM 通信的装置。这样做只是出于说明的目的; 讨论的本发明可使用其它设施管理子网。

在图 1 中所示的网络中, 端节点之一可请求加入多播组。这是通过向节点 110 的 SA 发送加入请求来实现的。SA 随后可产生多播组, 向该多播组分配一个多播本地标识符 (MLID), 并使 SM 更新交换机, 从而把分组发送给多播组的成员。

多播组的成员也可在不接收的情况下发送分组。这些多播组成员被称为发送不接受 (SWR) 成员。例如, 端节点 102 可向 SA 节点 110 发送加入请求, 其中该请求规定节点 102 将成为多播组的一个 SWR 成员。从而, 子网中的交换机被更新, 从而把来自节点 102 的分组发送给多播组的其它成员, 但是不把任意分组发送给节点 102。

但是, 当 SWR 成员最初加入多播组, 并且该多播组还未存在时, 则存在 SWR 成员在无任何接收者的情况下, 进行发送的问题。根据本发明的优选实施例, 当 SWR 成员请求产生多播组时, SA 产生该多播组, 分配 MLID, 并更新第一交换机 (这种情况下是交换机 122), 以便丢弃来自 SWR 节点 102 的多播分组。这在 IB 交换机硬件中规定得有。

当接收节点加入多播组时, SA 随后更新交换机, 从而 SWR 成员开始向接收成员发送分组。类似地, 当最后的接收成员离开多播组, 但是 SWR 成员留下时, SA 再次为第一交换机提供路由 (route) (图 1 中所



示例子中的交换机 122) 以丢弃来自 SWR 节点 102 的多播分组。

本发明还包含(没有变化)跨越多个子网的多个多播组的情况。例如,如果节点 112 是一个子网中的一个多播组的接收成员,节点 102 是另一子网中的一个多播组的 SWR 成员(如图 1 中所示),则来自节点 102 的分组将通过交换机 122 被发送给路由器 132,随后通过交换机 126 被发送给节点 112。如果节点 112 离开多播组,该多播组未留下任何成员,则 SM 更新交换机 122 的路由,从而丢弃从节点 102 发送的分组。这些分组随后不再被发送给节点 112。

现在参见图 2,图 2 根据本发明的优选实施例,图解说明了交换机。本例中,交换机 200 包括八个端口,端口 0~端口 7。在本发明的范围内,交换机可具有更多或更少的端口,取决于具体实现。例如,常见的 IB 交换机只具有四个端口。端口编号惯例也根据使用的具体硬件或特定的具体实现而变化。

交换机 200 还包括多播本地标识符(MLID)表 210。MLID 表被用于把多播分组发送给多播组的接收成员。例如,交换机 200 可在端口 5 接收多播分组。根据 MLID 表 210,交换机可复制该分组,并把该分组转发给端口 1、端口 3 和端口 7。但是,在任意这种实现中,交换机不把分组回送出接收该分组的端口;否则,多播分组会永不停止循环。

MLID 表可指示特定 MLID 的分组将被丢弃。根据本发明的优选实施例,交换机 200 还被配置成当需要时,丢弃分组。例如,交换机 200 (从任意端口)接收具有特定值的 MLID 的多播分组。MLID 表 210 指示该 MLID 的分组将被丢弃。交换机 200 只是丢弃该分组,而不是复制并转发该分组。

图 3A-3D 根据本发明的优选实施例,图解说明例证的多播路由数据结构。更具体地说,根据图 3A,MLID 表 300 包括 MLID 列和端口列。MLID 表 300 是根据本发明的多播路由数据结构的一个例子。当子网管理员(administrator)产生多播组时,向该多播组分配一个 MLID,并且 MLID 的一条记录,一行或一个条目被加入恰当的多播路由数据结构中。可使用其它方法。例如,每个 MLID 可隐含地与其在表格中的索引相联

系。从而，MLID 列不会明确存在，可提供一些机制来指示某一条目没有使用。

根据本发明的优选实施例，当 SWR 节点加入还未存在的多播组时，SA 将产生该多播组，并更新第一交换机的多播路由表，以便丢弃该分组。图 3B 图解说明了具有关于多播组的一个条目的例证多播路由表，所述多播组具有一个 SWR 成员。该例子中，为“1”的 MLID 被分配给该多播组，在 MLID 表 310 中保存一个条目。交换机被设置成只是丢弃给该多播组的分组，而不是把分组转发给特定的一个或多个端口。可使用多种机制来指示该分组将被丢弃，包括（但不限于）指示不存在的端口号；或者包含某个二进制位，当为“1”时，所述二进制位指示分组将被丢弃。

例如，如果图 1 中的 SWR 节点 102 加入还未存在的多播组时，则节点 110 的 SA 产生该多播组，并向该多播组分配一个 MLID。SM 随后更新交换机 122 的多播路由表，以便丢弃给该多播组的分组。图 3B 中表示了交换机 122 的多播路由表的一个例子。

下面，参见图 3C，图中表示了接收成员加入多播组之后，例证的多播路由数据结构。本例中，利用图 2 中所示的端口编号惯例，MLID 表 320 指示给 MLID 为“1”的多播组的分组将被转发给端口 7。

例如，如果 SWR 节点 102 是 MLID 为“1”的多播组的成员，并且一个或多个节点 104、106、108 是接收成员，则在交换机 122 从节点 102 接收的分组将被转发给交换机 124。SM 随后更新交换机 122 的多播路由表，以便据此转发这些分组。图 3C 中表示了交换机 122 的这种多播路由表的一个例子。

现在参见图 3C，关于多个接收成员，表示了例证的多播路由数据结构。本例中，利用图 2 中所示的端口编号惯例，MLID 表 330 指示给 MLID 为“1”的多播组的分组将被转发给端口 1、端口 3 和端口 7。

例如，如果图 1 的节点 104 和 108 是 MLID 为“1”的多播组的成员，则利用图 2 中所示的端口编号惯例，在交换机 124 接收的分组被转发给端口 1 和端口 7（除非这些分组是从端口 1 或 7 收到的）。随后，SM 更新交换机 124 的多播路由表，从而据此更新这些分组。如果在另一子网

上存在接收成员，则还更新交换机 124，以便通过端口 3，把分组转发给路由器 132。图 3D 中表示了交换机 124 的这种多播路由表的一个例子。

类似地，当最后的接收成员离开多播组，但是 SWR 成员留下时，SA 再次为第一交换机提供路由，丢弃来自 SWR 节点的多播分组。继续图 1 中所示的例子，如果接收节点 104、108 和所有其它接收节点离开多播组，则 SA 更新交换机 122 的多播路由表，丢弃给该多播组的分组。交换机 122 的这种多播路由表的一个例子同样示于图 3B 中。

虽然在图 3A-3D 中，MLID 路由数据结构被表示成表格，但是这些表格只是对本发明的举例说明，而不是对本发明的限制。实际上，可称为 MLID 表的 MLID 路由数据结构可被实现成由一系列二进制位组成的多个条目。如果某一端口的二进制位为“1”，则分组被发送给该端口，如果所述二进制位为“0”，则分组不被发送给该端口。

此外，MLID 路由数据结构可能不包括“MLID”列。相反，可依据 MLID 索引数据结构。换句话说，MLID 数据结构内的位置表示出 MLID 值。从而，所有 MLID 表固有包括介于 0 和表条目的数目减 1 之间的 MLID 值的条目。可向每个 MLID 提供一个二进制位，所述二进制位指示对于多播组，分组是否将被丢弃。从而，如果对于特定的 MLID，该二进制位具有为“1”的值，则关于该 MLID 接收的所有分组将被丢弃。

图 4A 是根据本发明的优选实施例，图解说明多播组加入请求的处理的流程图。当收到多播组加入请求，开始该过程，并确定多播组是否已存在（步骤 402）。如果多播组已存在，则该过程更新 MLID 表（步骤 404）。

如果在步骤 402 中，多播组不存在，则该过程产生该多播组（步骤 408），向该多播组分配 MLID。随后，该过程为第一交换机提供路由，从而给该多播组的所有分组被丢弃（步骤 410）。之后，该过程结束。从而，当产生只具有一个成员的多播组时，分配 MLID，并且允许存在的单个节点向该多播组发送分组。节点不必接收和产生的及删除的多播组无关的分组。根据上面描述的过程，当接收成员加入该多播组时，MLID 表被更新，从而把分组发送给接收成员节点。

现在参见图 4B，图中根据本发明的优选实施例，表示了图解说明多播组离开请求的处理的流程图。当收到多播组离开请求，开始该过程，并确定请求者是否是最后的多播组成员（步骤 452）。如果请求者是最后的多播组成员，该过程把 MLID 标记成未用（步骤 454），从交换机中的 MLID 表中清除 MLID（步骤 456），并结束。

如果在步骤 452 中，请求者不是最后的多播组成员，则确定是否单个成员留在多播组中（步骤 458）。如果一个以上的成员留在多播组中，则过程更新 MLID 表（步骤 460）并结束。

否则，在步骤 458 中，如果单个成员留在多播组中，则该过程给与剩余成员相连的第一交换机提供路由，从而丢弃给该多播组的所有分组（步骤 462）。之后，该过程结束。从而，当接收成员离开多播组，以致只有一个成员留下时，仍然允许留下的节点向该多播组发送分组。留下的节点不必接收和产生的及删除的多播组无关的分组，即使该节点是 SWR 节点。根据上面描述的过程，当接收成员加入多播组时，MLID 表再次被更新，从而把分组发送给接收成员节点。

于是，通过提供一种在产生和删除多播组时，不使用俘获的情况下，管理具有发送不接收（SWR）加入者的多播组的方法和设备，本发明解决了现有技术的缺陷。现有技术避免向无接收者的多播组分配 MLID。当可分配的 MLID 的数目有限时，这是所关心的。但是，本发明认识到可能 MLID 的数目不是问题。此外，当 IB 交换机中存储器的数量增大时，可保存的 MLID 条目的数目也增大。事实上，目前的交换机可包括支持一千或更多条目（这多于通常存在的多播组）的 MLID 表。

当存在 SWR 成员时，本发明的机制连续保持多播组信息。SWR 节点不必接收和产生的或删除的每个多播组无关的消息。从而，本发明减轻了 SWR 节点的负担，子网管理节点的负担，以及中间的所有交换机的负担。另外，只要 SWR 是成员，则 MLID 仍然被分配给该多播组。于是，减小了 SWR 节点向错误多播组发送分组的可能性。

重要的是注意虽然在全功能数据处理系统的环境下说明了本发明，但是本领域的普通技术人员会认识到，本发明的过程能够以指令的计算

机可读媒介的形式，以及各种形式被分布，并且本发明同样适用，与实际用于实现所述分布的信号承载媒介的特定类型无关。计算机可读媒介的例子包括可记录型媒介，例如软盘、硬盘驱动器、RAM、CD-ROM、DVD-ROM，和传输型媒介，例如数字和模拟通信链路，使用传输形式的有线或无线通信链路，例如射频和光波传输。计算机可读媒介可采取编码格式，所述编码格式被解码，以便在特定的数据处理系统中实际使用。

出于举例说明的目的，给出了本发明的说明，但是本发明并不局限于公开的形式。对本领域的普通技术人员来说，许多修改和变化是显而易见的。选择并描述了实施例，以便更好地解释本发明的原理，实际应用，并使本领域的普通技术人员能够理解具有各种修改的各种实施例适合于所考虑的特定应用。

图1

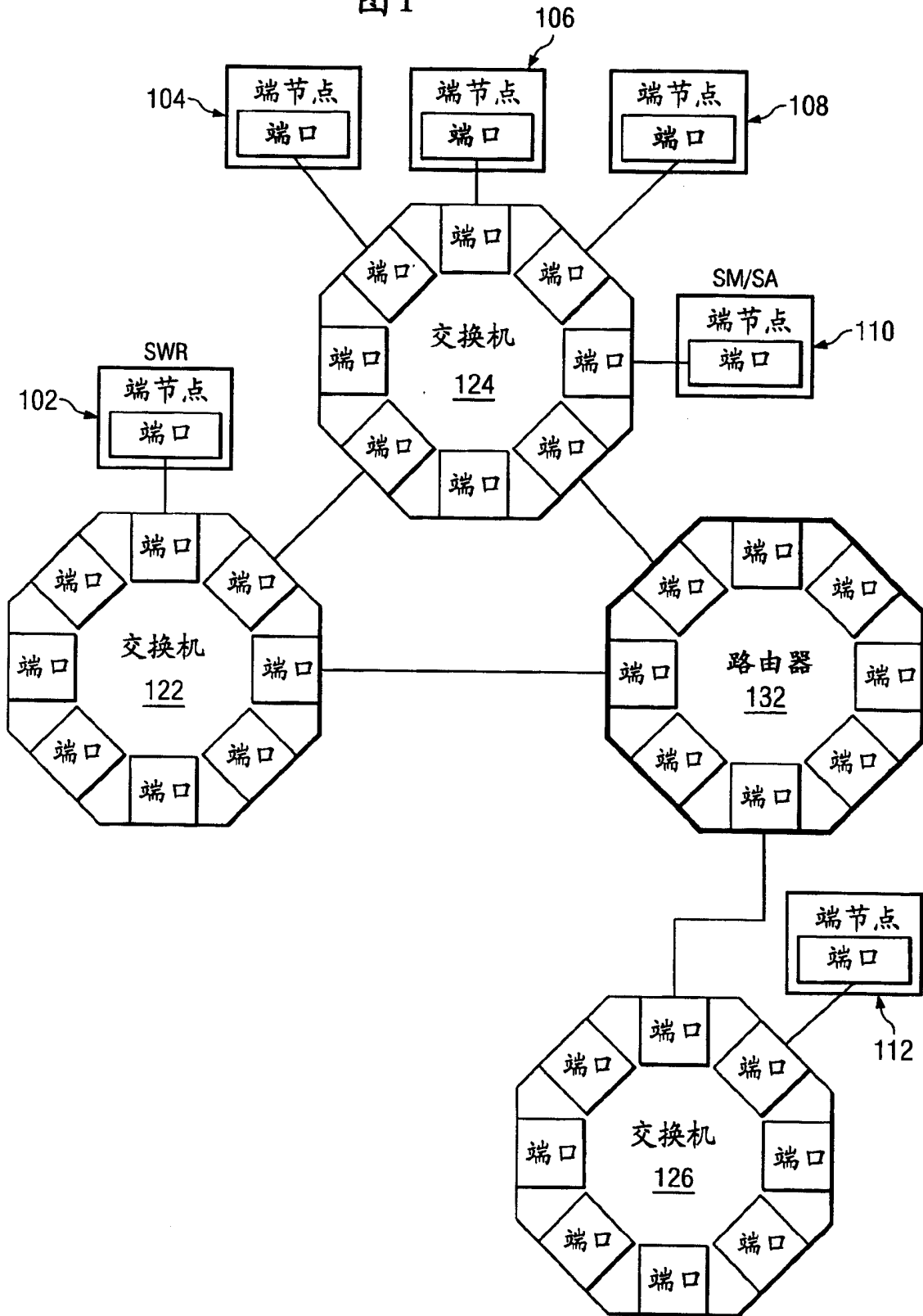


图 2

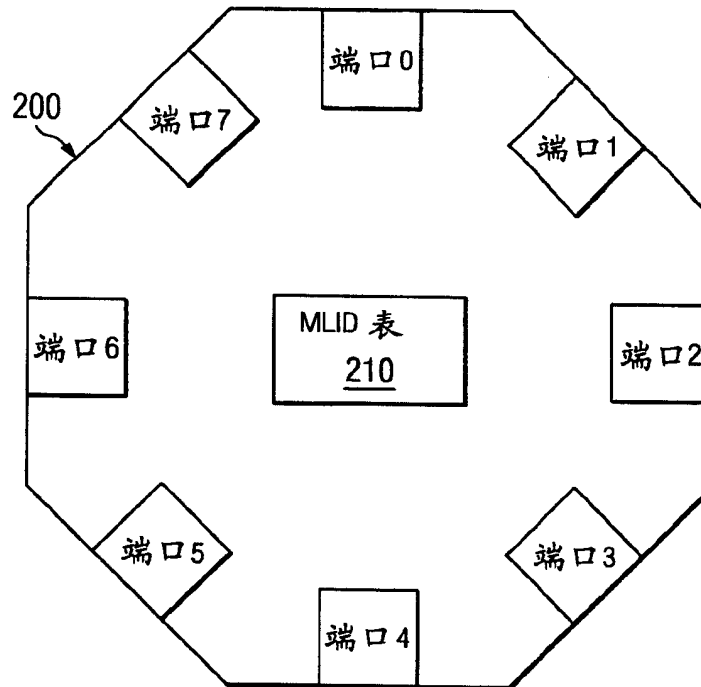


图 3A

300

MLID	端口

图 3B

310

MLID	端口
1	丢弃

图 3C

320

MLID	端口
1	7

图 3D

330

MLID	端口
1	1,3,7

图 4A

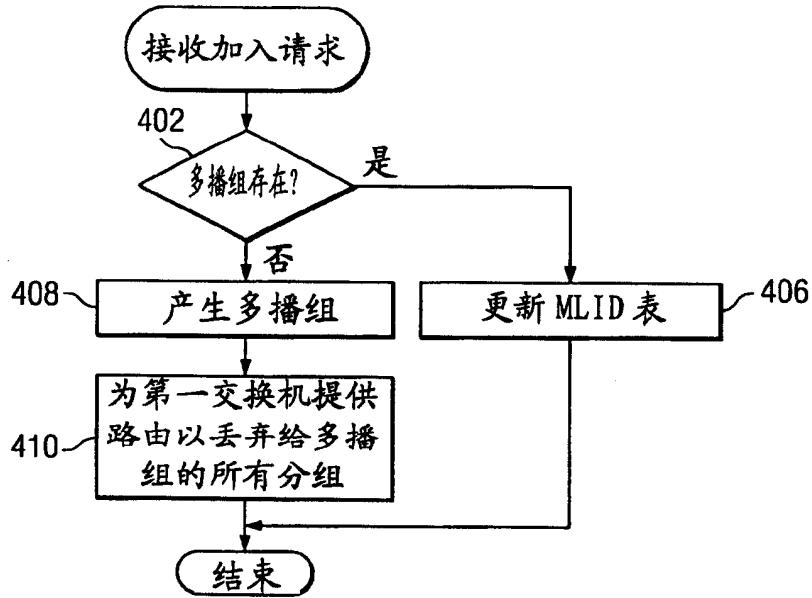


图 4B

