[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 12/00 (2006.01)

[12] 发明专利说明书

专利号 ZL 02814042.7

[45] 授权公告日 2009 年 5 月 6 日

[11] 授权公告号 CN 100486159C

[22] 申请日 2002.7.26 [21] 申请号 02814042.7

[30] 优先权

[32] 2001. 7.27 [33] US [31] 09/917,464

[86] 国际申请 PCT/US2002/023633 2002.7.26

[87] 国际公布 WO2003/013059 英 2003.2.13

[85] 进入国家阶段日期 2004.1.13

[73] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 维维克·卡施雅普

[56] 参考文献

US6108300A 2000.8.22

审查员 曹雅春

[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 吴丽丽

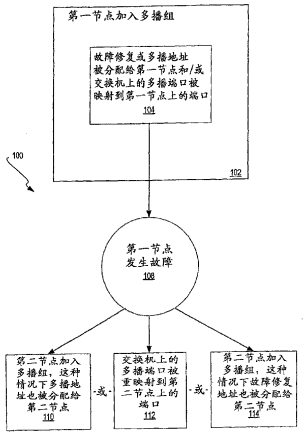
权利要求书 5 页 说明书 16 页 附图 10 页

[54] 发明名称

用于多播地址的网络节点故障修复的方法和装置

[57] 摘要

公开网络节点的故障修复。第一节点加入多播组(102)。通过执行三种操作之一实现加入(104)。首先，使故障修复地址与第一节点相关联，第一节点有效加入把故障修复地址作为多播地址的组。其次，使多播地址与第一节点相关联。第三，交换机的多播端口被映射到第一节点的端口。当第一节点发生故障时(106)，执行三种操作之一。如果加入涉及故障修复地址，则使故障修复地址与第二节点相关联，第二节点有效加入该组(114)。如果加入涉及多播地址，则第二节点加入该组，该地址与第二节点相关(110)。如果加入映射交换机的多播端口，则该端口被重映射到第二节点端口(112)。



1、一种方法，包括:

网络的第一节点加入具有多播地址的多播组（102），这里所述加入选自下述之一:

使故障修复地址与第一节点相关联，从而第一节点有效加入到把故障修复地址作为多播地址的多播组，给故障修复地址的通信通过网络被引向第一节点;

使多播地址与第一节点相关联，从而给多播地址的通信通过网络被引向第一节点；和

把网络的交换机上的多播端口映射到第一节点上的端口，从而给多播地址的通信从交换机上的多播端口被引向第一节点上的端口（104）；和

当第一节点发生故障时（108），通过执行以下三种操作之一实现所述加入;

如果所述加入使故障修复地址与第一节点相关联，则使故障修复地址与第二节点相关联，从而第二节点有效加入多播组，并且给故障修复地址的通信由第二节点处理（114）;

如果所述加入使多播地址与第一节点相关联，网络的第二节点加入多播组，从而多播地址与第二节点相关联，并且给多播地址的通信由第二节点处理（110）；或

如果所述加入把交换机上的多播端口映射到第一节点上的端口，则把交换机上的多播端口重映射到第二节点上的端口，从而给多播地址的通信被引向第二节点上的端口（112）。

2、按照权利要求 1 所述的方法，其中网络是 Infiniband 网络。

3、按照权利要求 1 所述的方法，其中故障修复地址选自下述之一:

值小于故障修复位置标识符阈值的故障修复位置标识符

（LID），网络包括 Infiniband 网络;

有效位置标识符范围内的故障修复位置标识符（LID），网络包括 Infiniband 网络；和

故障修复位置标识符，其作为源位置标识符，不被通过网络的任意传送方法检查，其作为多播目的地位置标识符，被通过网络的任意传送方法接受，其中网络包括 Infiniband 网络。

4、按照权利要求 1、2 或 3 所述的方法，还包括如果加入使多播地址或故障修复地址与第一节点相关联，在网络的第二节点加入多播组之前，通过第二节点代表第一节点向子网管理器（SM）发送离开请求，第一节点离开多播组（706）。

5、按照权利要求 4 所述的方法，还包括如果加入使多播地址或故障修复地址与第一节点相关联，当第一节点消除故障时，使故障修复地址与第一节点相关联，从而给故障修复地址的通信重新由第一节点处理（712）。

6、按照权利要求 4 所述的方法，其中如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组。

7、按照权利要求 1、2 或 3 所述的方法，还包括如果加入使多播地址或故障修复地址与第一节点相关联，当第一节点消除故障时，使故障修复地址与第一节点相关联，从而给故障修复地址的通信重新由第一节点处理（712）。

8、按照权利要求 7 所述的方法，其中如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组。

9、按照权利要求 1、2 或 3 所述的方法，其中如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组。

10、按照权利要求 1 或 2 所述的方法，其中如果加入把交换机

上的多播端口映射到第一节点上的端口，通过第二节点向子网管理器请求把交换机上的多播端口重映射到第二节点，SM 把交换机上的多播端口重映射到第二节点上的端口，交换机上的多播端口被重映射到第二节点上的端口。

11、按照权利要求 10 所述的方法，还包括如果加入把交换机上的多播端口映射到第一节点上的端口，当第一节点故障消除时，把交换机上的多播端口重映射到第一节点的端口上，从而给多播地址的通信再次被引向第一节点上的端口（912）。

12、按照权利要求 1 或 2 所述的方法，还包括如果加入把交换机上的多播端口映射到第一节点上的端口，当第一节点故障消除时，把交换机上的多播端口重映射到第一节点的端口上，从而给多播地址的通信再次被引向第一节点上的端口（912）。

13、一种多播地址的网络内节点的故障修复设备，其中网络第一节点加入具有多播地址的多播组（102），这里所述加入选自下述之一：

使故障修复地址与第一节点相关联，从而第一节点有效加入到把故障修复地址作为多播地址的多播组，给故障修复地址的通信通过网络被引向第一节点；

使多播地址与第一节点相关联，从而给多播地址的通信通过网络被引向第一节点；和

把网络的交换机上的多播端口映射到第一节点上的端口，从而给多播地址的通信从交换机上的多播端口被引向第一节点上的端口（104）；

当第一节点发生故障时（108），还包括以下三个装置之一以实现所述加入：

如果所述加入使故障修复地址与第一节点相关联，则使故障修复地址与第二节点相关联，从而第二节点有效加入多播组，并且给故障修复地址的通信由第二节点处理（114）的装置；

如果所述加入使多播地址与第一节点相关联，网络的第二节点加入多播组，从而多播地址与第二节点相关联，并且给多播地址的通信由第二节点处理（110）的装置；或

如果所述加入把交换机上的多播端口映射到第一节点上的端口，则把交换机上的多播端口重映射到第二节点上的端口，从而给多播地址的通信被引向第二节点上的端口（112）的装置。

14、按照权利要求 13 所述的设备，其中网络是 Infiniband 网络。

15、按照权利要求 13 所述的设备，其中故障修复地址选自下述之一：

值小于故障修复位置标识符阈值的故障修复位置标识符，网络包括 Infiniband 网络；

有效位置标识符范围内的故障修复位置标识符，网络包括 Infiniband 网络；和

故障修复位置标识符，其作为源位置标识符，不被通过网络的任意传送方法检查，其作为多播目的地位置标识符，被通过网络的任意传送方法接受，其中网络包括 Infiniband 网络。

16、按照权利要求 13、14 或 15 所述的设备，还包括如果加入使多播地址或故障修复地址与第一节点相关联，在网络的第二节点加入多播组之前，通过第二节点代表第一节点向子网管理器发送离开请求，第一节点离开多播组（706）的装置。

17、按照权利要求 16 所述的设备，还包括如果加入使多播地址或故障修复地址与第一节点相关联，当第一节点消除故障时，使故障修复地址与第一节点相关联，从而给故障修复地址的通信重新由第一节点处理（712）的装置。

18、按照权利要求 16 所述的设备，其中还包括如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组的装置。

19、按照权利要求 13、14 或 15 所述的设备，还包括如果加入使多播地址或故障修复地址与第一节点相关联，当第一节点消除故障时，使故障修复地址与第一节点相关联，从而给故障修复地址的通信重新由第一节点处理（712）的装置。

20、按照权利要求 19 所述的设备，其中还包括如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组的装置。

21、按照权利要求 13、14 或 15 所述的设备，其中还包括如果加入使多播地址或故障修复地址与第一节点相关联，则网络的第一节点加入多播组包括第一节点向子网管理器请求加入多播组的装置。

22、按照权利要求 13 或 14 所述的设备，其中还包括如果加入把交换机上的多播端口映射到第一节点上的端口，通过第二节点向子网管理器请求把交换机上的多播端口重映射到第二节点，SM 把交换机上的多播端口重映射到第二节点上的端口，交换机上的多播端口被重映射到第二节点上的端口的装置。

23、按照权利要求 22 所述的设备，还包括如果加入把交换机上的多播端口映射到第一节点上的端口，当第一节点故障消除时，把交换机上的多播端口重映射到第一节点的端口上，从而给多播地址的通信再次被引向第一节点上的端口（912）的装置。

24、按照权利要求 13 或 14 所述的设备，还包括如果加入把交换机上的多播端口映射到第一节点上的端口，当第一节点故障消除时，把交换机上的多播端口重映射到第一节点的端口上，从而给多播地址的通信再次被引向第一节点上的端口（912）的设备。

用于多播地址的网络节点故障修复的方法和装置

技术领域

本发明涉及网络，例如 Infiniband 网络，特别涉及这种网络内节点的故障修复（failover）。

背景技术

输入/输出（I/O）网络，例如系统总线，可被用于计算机的处理器，以与诸如网络适配器之类外围设备通信。但是，常见 I/O 网络的结构，例如外设组件接口（PCI）总线方面的约束，限制了计算机的总体性能。于是，提出了新型的 I/O 网络。

一种已知的新型 I/O 网络称为 Infiniband 网络。Infiniband 网络用拥有一个或多个路由器的分组交换网络替换目前计算机中的 PCI 或其它总线。主通道适配器（HCA）耦接处理器和子网，而目标通道适配器（TCA）耦接外设和子网。子网包括至少一个交换机，和使 HCA 和 TCA 与交换机连接的链路。例如，简单的 Infiniband 网络可具有一个交换机，HCA 和 TCA 通过链路与其连接。更复杂的布局也是可能的和可预期的。

Infiniband 网络的每个端节点包括一个或多个通道适配器（CA），每个 CA 包含一个或多个端口。每个端口具有由本地子网管理器（SM）分配的本地标识符（LID）。在子网内，LID 是唯一的。 交换机使用 LID 在子网内路由分组。数据的每个分组包含源 LID（SLID）和目的地 LID（DLID），源 LID 识别把分组注入子网的端口，目的地 LID 识别 Infiniband 结构或网络将向该处传送分组的端口。

Infiniband 网络方法通过定义 LID 掩码计数（LMC），提供物理端口内的多个虚拟端口。LMC 规定当证实分组 DLID 与其分配的 LID 相符时，物理端口掩蔽或忽略的 LID 的最低有效位的数目。但是交换机不忽略这些位。于是，SM 能够根据最低有效位，对通过 Infiniband 结构的不同路径编程。从而，该端口可认为是用于在

Infiniband 结构内路由目的的 2^{LMC} 个端口。

对于需要无故障的持续可用性的关键应用程序来说，通常要求单个应用程序的故障修复，从而要求通信端点或者端节点的故障修复。Infiniband 网络环境中的通信端点与 CA 端口相关。应用程序使用端点在 Infiniband 网络内通信，例如与其它应用程序等通信。端点的透明故障修复意味着另一端点按照不干扰网络自身内的通信的方式，接管故障端点的责任。

但是，由于对端点寻址的方式的缘故，端点或 Infiniband 网络内的其它节点的透明故障修复较困难。故障修复要求 LID 被重新分配给接管故障端口的新端口。但是，新端口通常已具有分配给它的 LID。于是，分配额外 LID 的唯一方式是扩展该端口上的 LMC 范围，从而确保新的 LID 落入该范围之内。

但是实际上难以扩展端口上的 LMC 范围，有时需要相当大的开销来确保接管端口能够具有分配给它们的故障端口的 LID。于是，LID 故障修复被认为是需要透明故障修复的 Infiniband 网络的成功转出（rollout）的问题和障碍。由于上述原因，需要本发明。

发明内容

本发明提供一种方法，包括：网络的第一节点加入具有多播地址的多播组，这里所述加入选自下述之一：使故障修复地址与第一节点相关联，从而第一节点有效加入到把故障修复地址作为多播地址的多播组，给故障修复地址的通信通过网络被引向第一节点；使多播地址与第一节点相关联，从而给多播地址的通信通过网络被引向第一节点；和把网络的交换机上的多播端口映射到第一节点上的端口，从而给多播地址的通信从交换机上的多播端口被引向第一节点上的端口；和当第一节点发生故障时，通过执行以下三种操作之一实现所述加入；如果所述加入使故障修复地址与第一节点相关联，则使故障修复地址与第二节点相关联，从而第二节点有效加入多播组，并且给故障修复地址的通信由第二节点处理；如果所述加入使多播地址与第一节点相关联，网络的第二节点加入多播组，从而多播地址与第二节点相关联，并且给多播地址的通信由第二节点处理；或如果所述加入把交换机上的多播端口映射到第一节点上的端口，则把交换机上的多播端

口重映射到第二节点上的端口，从而给多播地址的通信被引向第二节点上的端口。

本发明提供一种多播地址的网络内节点的故障修复设备，其中网络第一节点加入具有多播地址的多播组，这里所述加入选自下述之一：使故障修复地址与第一节点相关联，从而第一节点有效加入到把故障修复地址作为多播地址的多播组，给故障修复地址的通信通过网络被引向第一节点；使多播地址与第一节点相关联，从而给多播地址的通信通过网络被引向第一节点；和把网络的交换机上的多播端口映射到第一节点上的端口，从而给多播地址的通信从交换机上的多播端口被引向第一节点上的端口；当第一节点发生故障时，还包括以下三个装置之一以实现所述加入：如果所述加入使故障修复地址与第一节点相关联，则使故障修复地址与第二节点相关联，从而第二节点有效加入多播组，并且给故障修复地址的通信由第二节点处理的装置；如果所述加入使多播地址与第一节点相关联，网络的第二节点加入多播组，从而多播地址与第二节点相关联，并且给多播地址的通信由第二节点处理的装置；或如果所述加入把交换机上的多播端口映射到第一节点上的端口，则把交换机上的多播端口重映射到第二节点上的端口，从而给多播地址的通信被引向第二节点上的端口的装置。

本发明涉及使用故障修复或多播地址的网络内节点的故障修复。在本发明的一种方法中，网络的第一节点加入具有多播地址的多播组。通过执行三种操作之一实现所述加入。首先，可使故障修复地址与第一节点相关联，从而第一节点有效地加入把故障修复地址作为多播地址的多播组。给故障修复地址的通信通过网络被引向第一节点。其次，可使多播地址与第一节点相关联，从而给多播地址的通信通过网络被引向第一节点。第三，网络的交换机上的多播端口可被映射到第一节点上的端口。给多播地址的通信从交换机上的多播端口被引向第一节点上的端口。

当第一节点发生故障时，对应于第一节点加入网络的方法，执行三种操作之一。如果加入使故障修复地址与第一节点关联起来，故障修复地址与第二节点相关，从而第二节点有效地加入多播组，给故障修复地址的通信由第二节点处理。如果加入使多播地址与第一节点相关联，第二节点加入多播组，从而多播地址与第二节点相关，给多

播地址的通信由第二节点处理。如果加入把交换机上的多播端口映射到第一节点上的端口，则交换机上的多播端口被重映射到第二节点上的端口。从而给多播地址的通信被引向第二节点上的端口。

本发明还包括故障修复节点和制造产品。故障修复节点是实现本发明方法的节点，而制造产品具有计算机可读介质和所述介质中的实现本发明方法的装置。结合附图，根据下面的本发明的优选实施例的详细说明，本发明的其它特征和优点将是显而易见的。

附图说明

图 1 是根据本发明的优选实施例的方法的流程图，并被建议打印在颁发专利的第一页上。

图 2 是本发明的实施例可结合其实现的 **Inifiniband** 网络的图解。

图 3 是本发明的实施例可结合其实现的例证 **Inifiniband** 系统区网络（**SAN**）的图解。

图 4 是 **Inifiniband** 网络的例证端节点的通信接口的图解。

图 5 和 6 是表示 **Inifiniband** 寻址如何进行的 **Inifiniband** 网络的图解。

图 7 是表示本发明的实施例如何能够通过使多播组的故障修复地址和/或多播组的多播地址与另一节点相关联，实现网络节点故障修复的方法的流程图。

图 8 是表示图 7 的实施例的性能的图解。

图 9 是表示本发明的实施例如何能够通过把交换机多播端口重映射到另一节点上的端口，实现网络节点故障修复的方法的流程图。

图 10 是表示图 9 的实施例的性能的图解。

具体实施方式

概述

图 1 表示了根据本发明的优选实施例的方法 **100**。网络的第一节点最初有效加入多播组（**102**）。多播组具有多播地址或者故障修复地址。执行三种操作中的至少一种（**104**）。在第一种模式下，多播

地址被分配给第一节点。对多播地址的通信随后可被自动导向第一节点，这里先前可能已手动或自动建立了网络，以便实现这种通信。在第二种模式下，网络的交换机上的多播端口被映射到或者与第一节点上的端口相关联。对多播地址的通信随后可从交换机上的多播端口被导向第一节点上的端口，这里交换机不支持多播。在第三种模式下，故障修复地址被分配给该节点。对故障修复地址的通信随后被自动导向第一节点，这里先前已手动或自动建立了网络，以便实现这种通信。网络最好是 Infiniband 网络。第一和第二节点可以是这种网络上具有通道适配器（CA）和端口的主机。

第一节点随后发生故障（108），从而最好由网络的第二节点实现第一节点的透明故障修复。这可涉及执行三种操作之一。首先，第二节点可加入多播组，从而多播地址也被分配给第二节点（110）。从而给多播地址的通信被导向第二节点以及被导向第一节点（出故障节点），以致第二节点从第一节点接管这种通信的处理。其次，交换机上的多播端口可被重新映射到第二节点上的端口（112）。从而给多播地址的通信被导向第二节点上的端口，以致第二节点接管这种通信的处理。第三，使第二节点与故障修复地址相关联，从而第二节点有效加入多播组（114）。给故障修复地址的通信从而被导向第二节点以及被导向第一节点（出故障的节点），以致第二节点从第一节点接管这种通信的处理。

诸如 Infiniband 子网的子网管理器（SM）之类管理组件可把初始分配给第一节点的多播组的多播地址分配给第二节点。管理组件还可把最初映射到第一节点上端口的交换机的多播端口重新映射到第二节点上的端口。制造品的计算机可读介质中的装置也可实现这种功能。该装置可以是可记录的数据存储介质，调制的载波信号或者另一类型的介质或信号。

于是在第一模式中，多播地址被用于单播通信。多播地址允许发生本地标识符（LID）的故障修复，因为只有多播 LID 可被一个以上的端口共享。在第二模式下，所讨论的节点与交换机的主多播端口连接。当进行故障修复时，通过重新分配主端口，修改交换机配置，从而分组传播到故障修复节点。在第三种模式下，允许故障修复 LID 与任意多播组相关联。此外，允许故障修复 LID 不包括多播组地

址。

技术背景

图 2 表示了结合其可实现本发明的实施例的例证 Infiniband 网络结构 200。Infiniband 网络是一种网络。本发明也可和其它类型的网络一起实现。处理器 202 与主机互连 204 耦接，存储器控制器 206 也与主机互连 204 耦接。存储器控制器 206 管理系统存储器 208。存储器控制器 206 还与主机通道适配器（HCA）210 连接。HCA 210 允许处理器和存储器子系统通过 Infiniband 网络通信，处理器和存储器子系统包括处理器 202，主机互连 204，存储器控制器 206 和系统存储器 208。

图 2 中的 Infiniband 网络被称为子网，子网包含 Infiniband 链路 212、216、224 和 230，和一个 Infiniband 交换机 214。可存在一个以上的 Infiniband 交换机，但是图 2 中只表示了交换机 214。链路 212、216、224 和 230 使 HCA 和目标通道适配器（TCA）218 及 226 能够相互通信，还使 Infiniband 网络能够通过路由器 232 与其它 Infiniband 网络通信。具体地说，链路 212 连接 HCA 210 和交换机 214。链路 216 和 224 分别使 TCA 218 和 226 与交换机 224 连接。链路 230 连接路由器 232 和交换机 214。

TCA 218 是特定外设，这种情况下，是以太网适配器 220 的目标通道适配器。TCA 可容纳多个外设，例如多个网络适配器，SCSI 适配器等。TCA 218 使网络适配器 220 能够通过 Infiniband 网络发送和接收数据。适配器 220 本身允许通过通信网络，尤其是以太网进行通信，如线条 222 所示。其它通信网络也适合于本发明。TCA 226 是另一外设，目标外设 228 的目标通道适配器，图 2 中没有详细说明目标外设 228。路由器 232 允许图 2 的 Infiniband 网络与其它 Infiniband 网络连接，线条 234 表示了该连接。

Infiniband 网络是分组交换输入/输出（I/O）网络。从而，通过互连 204 和存储器控制器 206，处理器 202 经 HCA 210 发送和接收数据分组。类似地，目标外设 228 和网络适配器 220 分别通过 TCA 226 和 218 发送和接收数据分组。也可通过路由器 232 发送和接收数据分组，路由器 232 连接交换机 214 和其它 Infiniband 网络。链路

212、216、224 和 230 可具有变化的容量，取决于它们与交换机 214 连接的特定 HCA、TCA 等所需的带宽。

Infiniband 网络按照这里简要说明的不同方式提供 TCA 和 HCA 之间的通信。类似于其它类型的网络，Infiniband 网络具有物理层，链路层，网络层，传输层和高级协议。如同在其它类型的分组交换网络中，在 Infiniband 网络中，特定的事务被分成消息，消息本身被分成分组以便通过 Infiniband 网络传送。当被预定的接收者接收时，分组被记录到指定事务的组成消息中。Infiniband 网络提供队列和通道，在所述队列和通道接收和发送分组。

此外，Infiniband 网络允许许多不同的传送服务，包括可靠和不可靠的连接，可靠和不可靠的数据报，和原始分组支持。在可靠的连接和数据报中，产生确认和保证分组排序的分组序列号。重复的分组被拒绝，检测丢失的分组。在不可靠的连接和数据报中，不产生确认，不保证分组排序。不拒绝重复的分组，不检测丢失的分组。

Infiniband 网络也可被用于定义系统区网络（SAN），用于连接多个独立的处理器平台，或者主处理器节点，I/O 平台，和 I/O 装置。图 3 表示了结合其可实现本发明的实施例的例证 SAN 300。SAN 300 是支持一个或多个计算机系统的 I/O 和处理器间通信的通信和管理基础结构。Infiniband 系统包括小型服务器到大型并行超级计算机中心站不等。此外，Infiniband 网络的因特网协议（IP）-友好本性允许桥接到因特网，企业内部网，或者连接到远程计算机系统。

SAN 300 具有交换的通信结构 301，或者子网，通信结构 301 或者子网允许许多装置在受保护的远程管理环境中，高带宽低等待时间地同时通信。端节点可通过多个 Infiniband 端口通信，并且能够利用通过结构 301 的多个路径。端口和通过网络 300 的路径的多样性被用于容错和增大的数据传送带宽。Infiniband 硬件卸下多数处理器和 I/O 通信操作。这允许多个同时进行的通信，而不存在与通信协议相关的传统开销。

结构 301 具体包括若干交换机 302、304、306、310 和 312，允许结构 301 与其它 Infiniband 子网、广域网络（WAN）、局域网（LAN）和主机链接（如箭头 303 所示）的路由器 308。结构 301 允许若干主机 318、320 和 322 相互通信，以及与不同的子系统，管理

控制台，驱动器和 I/O 机架通信。在图 3 中，这些不同的子系统、管理控制台、驱动器和 I/O 机架被表示为信息磁盘冗余阵列（RAID）子系统 324、管理控制台 326、I/O 机架 328 和 330、驱动器 332 和存储子系统 334。

图 4 表示了 Infiniband 网络的例证端节点 400 的通信接口。端节点可以是图 3 的主机 318、320 和 322 之一。端节点 400 具有运行于其上的过程 402 和 404。每个过程具有与之相关的一个或多个队列对（QP），每个 QP 与节点 400 的通道适配器（CA）418 通信，以便与 Infiniband 结构链接，如箭头 420 所示。例如，过程 402 具有 QP 406 和 408，而过程 404 具有 QP 410。

在 HCA 和 TCA 之间定义 QP。链路的每一端具有要传送给另一端的消息队列。QP 包括成对的发送工作队列和接收工作队列。一般来说，发送工作队列保存导致在客户机的存储器和另一过程的存储器之间传送数据的指令，接收工作队列保存关于把从另一过程接收的数据置于何处的指令。

QP 代表与 Infiniband 客户机过程的虚拟通信接口，并为该客户机提供虚拟通信端口。CA 可提供多达 2^{24} 个 QP，并且每个 QP 上的操作彼此独立。客户机通过分配 QP，产生虚拟通信端口。客户机启动把该 QP 和另一 QP 连接在一起所需的任意通信建立，并利用某些信息，例如目的地地址、服务级别、协议工作极限等配置 QP 环境。

图 5 和 6 表示在 Infiniband 网络内如何寻址。在图 5 中，表示了简单的 Infiniband 网络 500，Infiniband 网络 500 包括一个端节点 502 和一个交换机 504。端节点 502 具有运行于其上的过程 504，过程 504 具有相关的 QP 506、508 和 510。端节点 502 还包括一个或多个 CA，例如 CA 512。CA 512 包括一个或多个通信端口，例如端口 514 和 516。QP 506、508 和 510 具有由 CA 分配的队列对编号（QPN），队列对编号唯一地识别 CA 512 内的 QP。除原始数据报之外的数据分组包含目的地工作队列的 QPN。当 CA 512 接收一个分组时，它使用目的地 QPN 的环境恰当地处理该分组。

本地子网管理器（SM）向每个端口分配一个本地标识符（LID）。SM 是连接到子网上的管理组件，负责配置和管理交换机、路由器和 CA。可利用其它设备，例如 CA 或交换机嵌入 SM。

例如 SM 可被嵌入端节点 502 的 CA 512 内。作为另一例子，SM 可被嵌入交换机 504 内。

在 Infiniband 子网内，LID 是唯一的。诸如交换机 504 之类的交换机使用 LID 在子网内发送分组。每个分组包含源 LID（SLID）和目的地 LID（DLID），源 LID 识别把分组注入子网中的端口，目的地 LID 识别结构将向该处传送分组的端口。诸如交换机 504 之类的交换机还分别具有许多端口。交换机 504 上的每个端口可与端节点 502 上的端口关联。例如，交换机 504 的端口 518 与端节点 502 的端口 516 关联，如箭头 520 所示。交换机 504 接收的预定给节点 502 的端口 516 的数据分组从而从端口 518 发送给端口 516。更具体地说，当交换机 504 接收具有 DLID 的分组时，该交换机只检查该 DLID 是否非零。否则，交换机按照 SM 设计的表格发送该分组。

除了分别识别 Infiniband 子网内的特定端口的 DLID 之外，还可规定多播 DLID 或者多播地址。通常，一组端节点可加入一个多播组，从而 SM 向每个节点的一个端口分配多播组的一个多播 DLID。发送给多播 DLID 的数据分组被发送给加入多播组的每个节点。每个交换机，例如交换机 504 具有默认的主多播端口和默认的非主要多播端口。主/非主要多播端口用于所有的多播分组，并不和任意特定 DLID 相关。加入多播组的每个节点的一个端口或者与交换机的主多播端口关联，或者与交换机的非主要多播端口关联。

当收到具有多播 DLID 的数据分组时，检测该多播 DLID，并根据 SM 计划的表格转发数据分组。如果多播 DLID 不在该表格中，或者交换机不保存表格，则交换机在主默认多播端口和非主要默认多播端口上转发分组。如果在主端口上被接收，则分组从非主要多播端口出去，而如果在交换机的任意其它端口上被接收，则分组从主多播端口出去。交换机 504 接收的指定多播 DLID 的数据分组从而从这些多播端口之一被发送给多播组节点的相关端口。可利用关于多播通信的路由信息配置交换机 504，所述路由信息指定分组应送往的端口。

此外，虽然任意 Infiniband 节点可向任意多播组传送分组，但是如果交换机，例如交换机 504 不正确转发分组，则不能保证数据分组将被多播组成员正确接收。于是，应设置交换机，以致多播数据分组被组成员接收。这可通过确保多播数据分组总是鱼贯通过被预编程

或者专门编程的一个或多个交换机，从而确保多播数据到达它们正确目的地来实现。另一方面，如果所有交换机都完全支持多播，则端节点加入多播组会导致 SM 对交换机编程，从而分组被多播组的所有成员正确接收。也可执行其它方法。

图 6 中，表示了更复杂的 Infiniband 网络 600，Infiniband 网络 600 具有两个子网 602 和 604。子网 602 具有不同地与交换机 610 和 612 连接的端节点 604、606 和 608。类似地，子网 604 具有不同地与交换机 622 和 624 连接的端节点 614、616、618 和 620。子网 602 的交换机 610 和 612 通过路由器 626 和 628 不同地与子网 604 的交换机 622 和 624 连接，路由器 626 和 628 能够实现子网间通信。这种情况下，不同地连接意味着一个实体的一个或多个端口与另一实体的一个或多个端口关联。例如，节点 604 可具有两个端口，一个与交换机 610 关联，另一个与交换机 612 关联。

为了第一节点的故障修复，使故障修复（多播）地址与第二节点关联

通过使多播组的故障修复地址与另一节点关联，本发明的实施例能够实现网络节点故障修复。图 7 表示了根据本发明的这种实施例的方法 700。该实施例最好把位置标识符（LID）的 Inifiniband 规范重新如下定义：

LID 地址或地址范围	应用
0x0000	无效
0x0001 ~ ThLID-1	单播端口
ThLID ~ 0xFFFE	故障修复 LID
0xFFFF	许可（只用于管理分组）

ThLID 是管理员规定的阈值，从而最好只有高于 ThLID 的 LID 才可以是故障修复 LID，也是有效地多播 LID。此外，Inifiniband 规范最好被增强，以便允许故障修复 LID 与多播组标识符（GID）关联起来。允许在存在或不存在 GID 的情况下使用这种故障修复 LID。在 ThLID 等于 0XC000（目前的 Inifiniband 规范中，多播范

围的起始值）的情况下，则该实施例和目前的规范一致。

在本发明的另一实施例中，除了许可的 LID 之外，任意有效 LID 都可与多播组关联，从而能够有效地起故障修复 LID 的作用。子网管理器（SM）被增强，以便允许除许可 LID 之外的任意这种有效 LID 与多播组相关。即，Infiniband 规范被修改，从而 SM 可允许除许可 LID 之外的任意有效 LID 与多播组关联，以便允许节点故障修复。最后，在本发明的一个备选实施例中，不对 Infiniband 规范进行任何改变，从而与有效地也是一个多播组 LID 的故障修复 LID 相反，下面关于图 7 的方法 700 的说明只与多播组 LID 相关。

现在参见图 7 的方法 700，Infiniband 网络的第一节点与故障修复 LID（或者多播 LID）相关，故障修复 LID 有效地是多播组 LID，从而第一节点有效地加入多播组（702）。第一节点是其一部分的子网的 SM 使故障修复 LID 与多播组关联起来，例如响应第一节点的加入多播组的请求。第一节点可以是 Infiniband 网络的子网上的主机的通道适配器（CA）。第一节点随后出故障（704），这通常由子网的另一节点检测。通过子网的第二节点代表第一节点向 SM 发送离开请求，第一节点可以可选地脱离多播组（706）。

第二节点随后加入多播组，从而将接收发送给故障修复 LID（或者发送给多播 LID）的分组（708）。更具体地说，响应来自第二节点的加入请求，SM 对交换机编程，从而发送给多播组的分组将被第二节点接收。第二节点也可以是 Infiniband 网络的子网上的主机的 CA。第二节点的主机可以是和第一节点的主机相同的主机。计划给故障修复 LID 的通信由第二节点，而不是由第一节点处理，从而第一节点无缝地向第二节点交接（fail over）。

在某一时刻，第一节点可能消除故障（failback）（710），恢复在线。随后再次使故障修复 LID（或多播 LID）和第一节点关联起来（712），从而第一节点能够重新处理计划给故障修复 LID 的通信。在第一节点重新加入多播组之前，子网的第二节点可离开多播组。从而在第一节点向 SM 发送加入请求，以使 SM 把故障修复 LID 和第一节点关联起来之前，第二节点可向 SM 发送离开请求。故障消

除（failback）还可包括第一节点从第二节点获得状态转储，这里第二节点冻结所有连接，直到完成故障消除为止。另一方面，第二节点可以不离开多播组，直到与第二节点的现有连接到期为止。

从而，原先与第一节点通信的第三节点将不知道已向第二节点进行了故障修复。即，它将继续与故障修复地址通信，而不必知道故障修复地址是与第一节点相关联还是与第二节点相关联。通常，即使利用多播故障修复地址实现故障修复，第三节点和第一节点或第二节点之间的通信实际上也是单播通信。第三节点不知道故障修复地址事实上是多播地址，从而导致认为它和故障修复地址之间的通信实际上是单播通信。即，当事实上正在利用多播地址完成通信时，使得在第三节点看来通信正在正常进行。

图 8 表示了相对于第二节点的第一节点故障修复。多播组被表示为多播组 802A，以便表示第一节点 804 的故障前状态。于是，具有故障修复 LID 的分组 806 被发送给第一节点 804。多播组被表示为多播组 802B， 以便表示第一节点 804 的故障后状态，从而在第一节点 804 发生故障之后，组 802A 变成组 802B，如箭头 808 所示。组 802A 的第一节点 804 变成组 802B 的第一节点 804'，以便指出故障。第二节点 810 加入多播组 802B。第一节点 804'被表示为在组 802B 中，但是可能已离开组 802B。于是，除了第一节点 804'之外，分组 806 现在被发送给第二节点 810。

为第一节点故障修复，把交换机多播端口重映射到第二节点上的端口，通过把交换机多播端口重映射到另一节点上的端口，本发明的实施例也可实现网络节点故障修复。图 9 表示了根据本发明的这种实施例的方法 900。Infiniband 网络的第一节点加入多播组，这种情况下交换机上的主多播端口被映射到第一节点上的端口（902）。响应第一节点的加入请求，第一节点和交换机为其一部分的子网的子网管理器（SM）实现这种映射。第一节点可以是在网络子网上的主机的通道适配器（CA）。

第一节点随后发生故障（904），这通常由子网的另一节点检

测。通过子网的第二节点代表第一节点向 SM 发送脱离请求，第一节点可以可选地脱离多播组（906）。交换机上的主多播端口随后被重映射到第二节点上的端口（708）。更具体地说，响应第二节点的相应请求（可选的是专有请求），SM 把交换机上的主多播端口重映射到第二节点上的端口。第二节点也可以是 Inifiniband 网络的子网上的主机的 CA。第二节点的主机可以是和第一节点的主机相同的主机。给该多播地址的通信被导向第二节点上的端口，而不是第一节点上的端口，从而第一节点无缝地向第二节点交接。

在某一时刻，第一节点可能消除故障（910），从而返回在线。交换机上的主多播端口随后被重映射到第一节点上的端口（912），从而第一节点能够再次处理计划给多播地址的通信，它可以是多播目的地位置标识符（DLID）。子网的第二节点可能不得不初始离开多播组，从而在主多播端口被重映射到第一节点上的端口之前，可向 SM 发送脱离请求。故障消除也可包括第一节点从第二节点获得状态转储，这种情况下第二节点冻结所有连接，直到故障消除完成为止。此外，第二节点可不脱离多播组，直到与第二节点的现有连接到期为止。

图 10 表示了相对于第二节点的第一节点故障修复。子网的一部分被表示为部分 1002A，以便表示第一节点 1004 的故障前状态。第一节点 1004 具有端口 1006。交换机 1008 具有主多播端口 1010。交换机 1008 的主多播端口 1010 被映射到第一节点 1004 的端口 1006，如线条 1012 所示。指向交换机 1008 的多播通信从而被发送给端口 1006。该部分子网被表示为部分 1002B，以便表示第一节点 1004 的故障后状态，从而在第一节点 1004 的故障之后，部分 1002A 变成部分 1002B，如箭头 1014 所示。第二节点 1016 具有端口 1018。交换机 1008 的多播端口 1030 现在成为主多播端口，并被映射到第二节点 1016 的端口 1018，如线条 1020 所示。通过交换机 1008 导引的多播通信现在被发送给端口 1018。

交换机、数据报和连接服务类型

Infiniband 网络采用通常只检查目的地位置标识符（DLID）不为零的交换机，并根据子网管理器（SM）设计的表格发送数据分组。最好利用多播通信的路由信息配置各个交换机，所述路由信息指定多播数据分组需要通过的所有端口。这确保多播分组被发送给它们正确的目的地。

此外，Infiniband 网络可采用不同类型的数据报和连接服务。在和发送分组的顺序相比，接收分组的顺序无关紧要的情况下使用数据报。可和发送数据报分组的顺序相比无序地接收数据报分组。数据报可以是原始的，这意味着它们和非 Inifiniband 规范，例如 Ethertype、因特网协议（IP）版本 6 等相符。相反，在和发送分组的顺序相比，接收分组的顺序至关重要的情况下使用连接服务。按照发送分组的相同顺序接收连接服务分组。

数据报和连接的服务都可以是可靠的或者不可靠的。可靠性通常涉及是否保持分组的序列号，是否关于接收的分组发送确认消息，和/或是否执行其它验证措施，以确保发送的分组被它们预定的接收者接收。不可靠的数据报和不可靠的连接服务不进行这样的验证措施，而可靠的数据报和不可靠的连接服务执行这样的验证措施。

对于不可靠的原始数据报，第一节点使用多播位置标识符（LID）作为其源 LID（SLID）。在第二节点是故障修复节点的情况下，第三节点能够接收这样的分组，因为它们被发送给其单播 DLID，并且因为分组的 SLID 未被检查。第三节点应当应答第一节点的多播 LID。为此，客户机可被发送一个多播 LID 关联，该多播 LID 关联被客户机记录。在不可靠数据报的 SIDR 协议的情况下，第三节点可被发送多播 LID，和/或第二节点可从接收的分组拾取多播 LID。在 Inifiniband 规范中，不存在关于不可靠数据报模式规定的有效性检查。

如果第三节点根据 SM 保持的路径记录确定 LID，则在启动与第三节点的通信之前，LID 的恰当值可被置于路径记录中。当第一节

点，或者故障修复第二节点接收来自客户机的答复分组时，分组具有非多播队列对（QP），但是具有多播 DLID。在可靠数据报和连接模式传送的情况下，连接管理器交换将用于通信的 LID。在该阶段可交换多播或故障修复 LID。该 LID 可在不进行任何有效性检查的情况下被第二节点记录，并被用作所有通信中的单播 LID。

链路层和传输层检查也都被核实。链路层检查只核实客户机的 LID，或者为多播 LID 或者为单播 LID。在传输层检查中，接收 QP 首先被核实为有效，因为发送者设置该 QP。最后，QP 被核实为不是 0xFFFFFFF（十六进制），于是，数据分组不被认为是多播分组，从而不检查多播全局路由报头（GRH）的存在。

但是，在本发明的一个实施例中，Infiniband 规范被重新定义，以便通过不严格执行这些传输层检查，提供节点故障修复。本实施例中，对于任意传输方法不检查源 LID（SLID），并且对于任意传输方法接受多播目的地 LID（DLID）。从而，Inifiniband 规范被修改，从而不象以前那么严格地执行 SLID 和 DLID 检查。

在本发明的另一实施例中，另一方面通过把 QP 设置为特定值（例如 0xFFFFFFE），表示多播通信，提供节点故障修复。本发明的该实施例只适用于不可靠的连接服务。特定的 QP 值是可配置的，并且可由 SM 保持。

对于可靠的数据报和可靠及不可靠的连接服务来说，不允许多播，因为它未被定义。但是，如果两个端节点按照单播的方式工作，则可克服这种限制。服务器把分组发送给使用多播 LID 的客户机。远程客户机检查 SLID 是否是多播 LID。如果是，则可修改客户机的主通道适配器（HCA），以便接收多播 SLID，否则可修改 SM，使单播 LID 与多播组关联起来。

即，只有当其大于 0xC000（十六进制）时，未修改的接收器断定该 SLID 是多播的。于是，SM 被修改，从而它把低于 0xC000（十六进制）的值分配给多播组，从而接收器不断定 SLID 是多播。客户机应答服务器，服务器接收规定 DLID 的分组。服务器检查 DLID 是

否是多播 LID。如果是，则服务器的 HCA 可被修改以便接收多播 DLID，或者 SM 可被修改，以使单址通信 LID 与多播组关联起来。

优于现有技术的优点

本发明的实施例提供优于现有技术的优点。通过利用 Infiniband 网络的多播地址和端口，实现节点故障修复。即使指定的 Infiniband 结构不允许多播，在故障节点于另一节点加入多播组之前离开多播组的情况下，仍然可以使用本发明的实施例，从而每次多播组中只存在一个节点。故障节点的故障修复不要求涉及故障节点一直与之通信的远程节点。

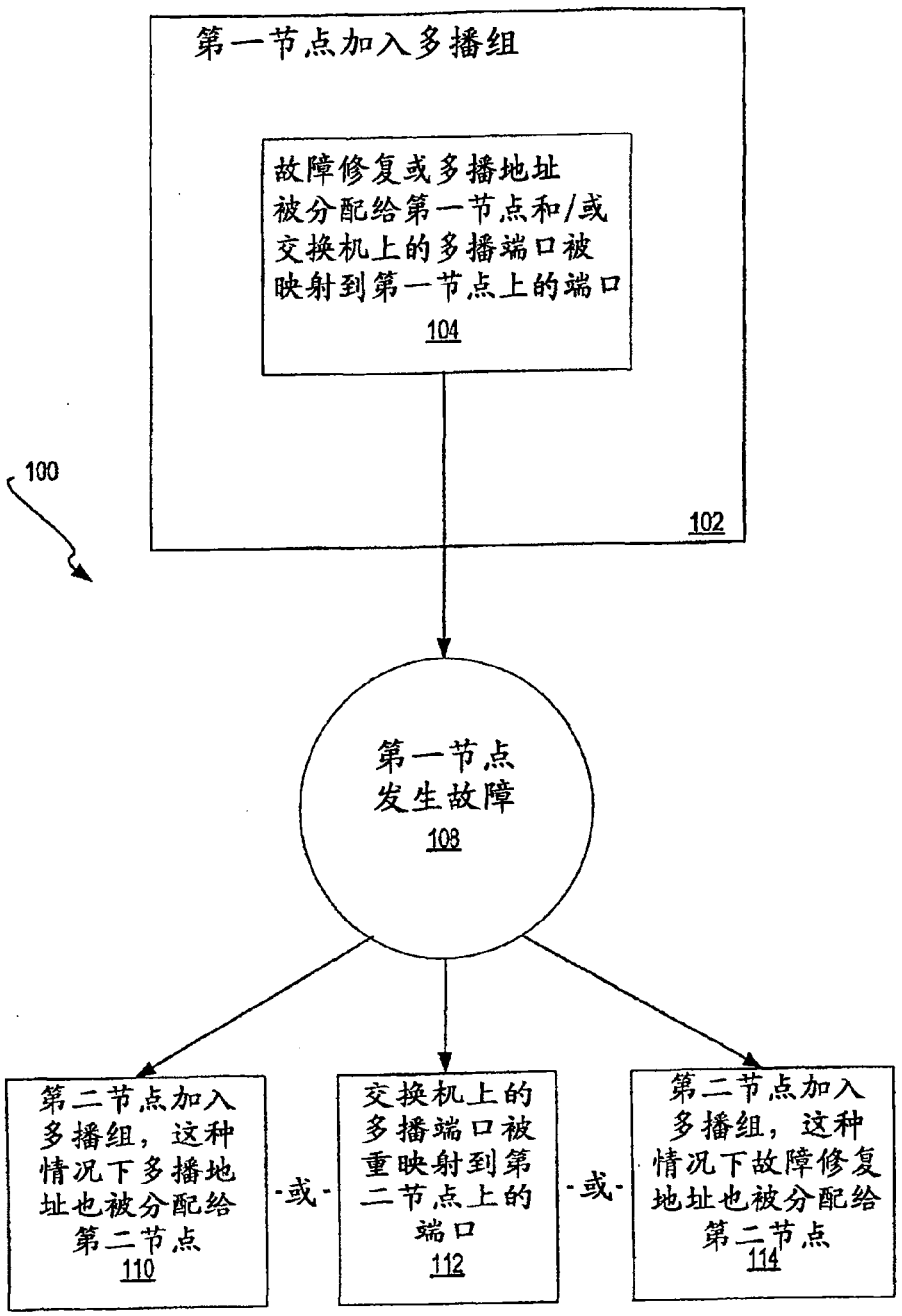
从而，故障修复节点透明地承担故障节点的责任，并且通常不通知远程节点。更可取的是，任意主机能够接管故障主机的职责。本发明的实施例也适用于所有 Infiniband 传送类型。实现本发明的实施例通常不需要对 Infiniband 规范的非专有扩展，从而实施例在 Infiniband 规范的支持下工作。

此外，在本发明的其它实施例中，通过利用本发明规定的 Infiniband 网络的故障修复地址，实现节点故障修复。故障节点的故障修复不需要涉及故障节点一直与之通信的远程节点。相反，接管（takeover）节点透明地承担故障节点的责任，并且通常不通知远程节点。更可取的是，任意主机能够接管故障主机的职责。本发明的实施例也适用于所有 Infiniband 传送类型。

备选实施例

虽然出于举例说明的目的，说明了本发明的具体实施例，不过在不脱离本发明的精神和范围的情况下，可做出各种修改。例如，虽然主要关于 Inifiniband 网络说明本发明，不过本发明也适用于其它类型的网络。因此，本发明的范围只受下述权利要求及其等同物限定。

图1


第一节点加入多播组
故障修复或多播地址被分配给第一节点和/或交换机上的多播端口被映射到第一节点上的端口
104
102
100
第一节点发生故障
108
第二节点加入多播组，这种情况下多播地址也被分配给第二节点
110
-或-
交换机上的多播端口被重映射到第二节点上的端口
112
-或-
第二节点加入多播组，这种情况下故障修复地址也被分配给第二节点
114

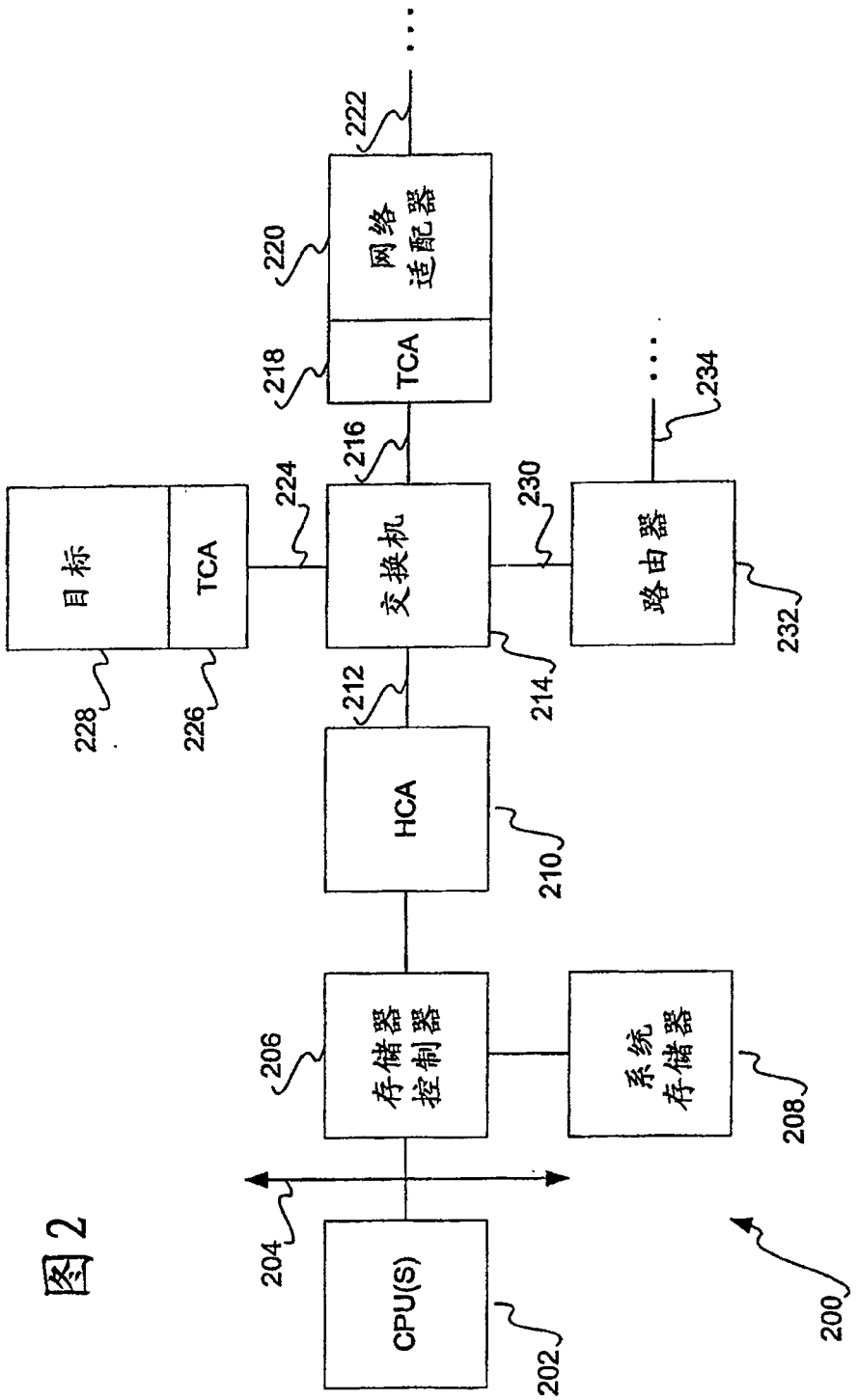

CPU(S)
HCA
存储器控制器
系统存储器
交换机
路由器
网络适配器
TCA
TCA
目标
200
202
204
206
208
210
212
214
216
218
220
222
224
226
228
230
232
234

图 2

图 3

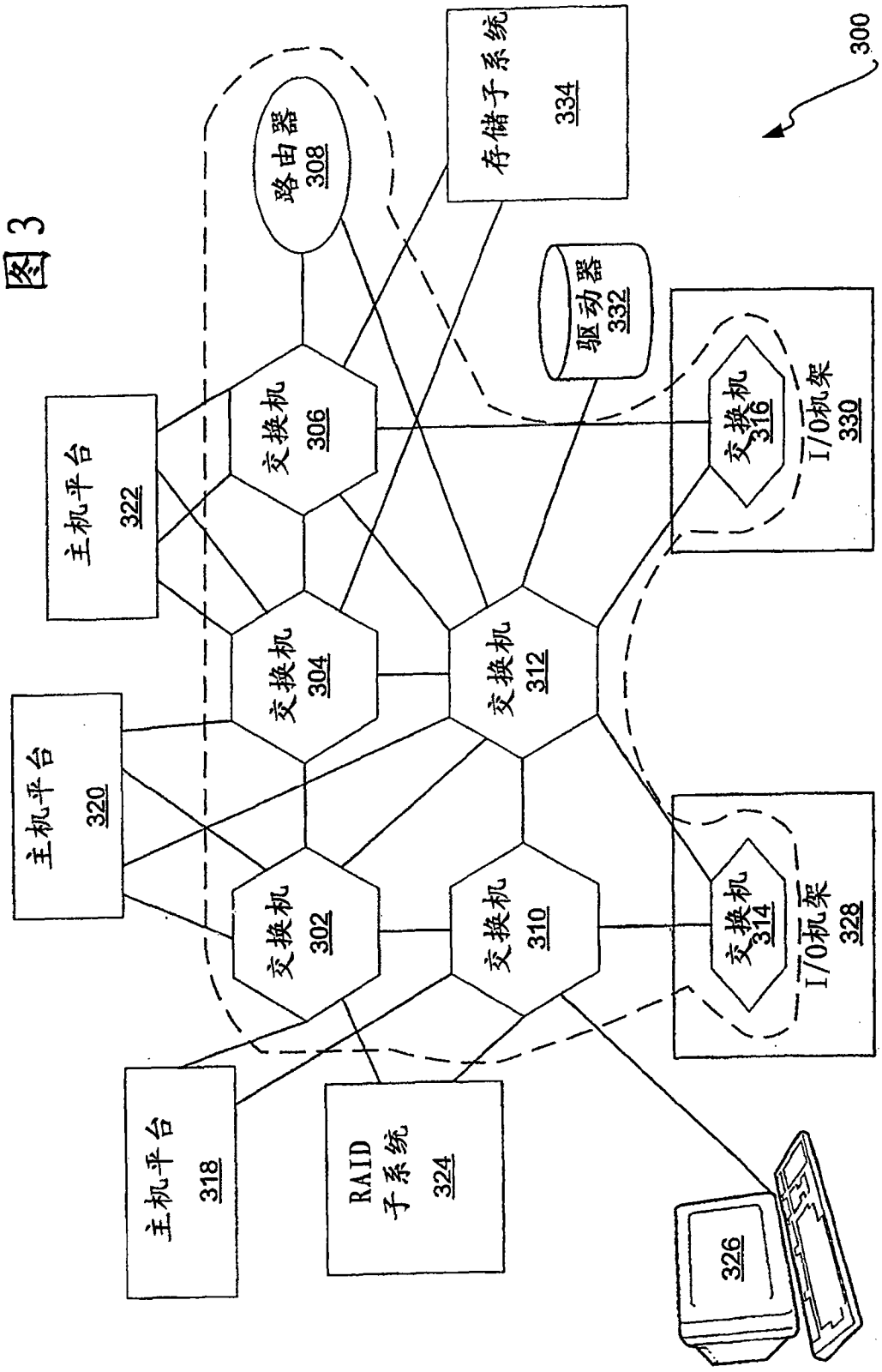

300
存储子系统
334
路由器
308
驱动器
332
主机平台
322
交换机
306
交换机
316
I/O机架
330
主机平台
320
交换机
304
交换机
312
交换机
302
交换机
310
交换机
314
I/O机架
328
主机平台
318
RAID
子系统
324
326

图4

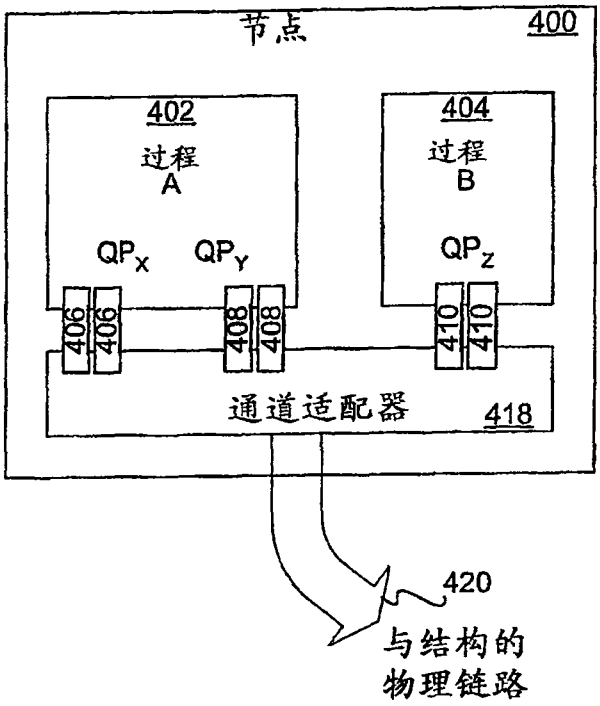

节点
400
402
过程
A
404
过程
B
QP_X
QP_Y
QP_Z
406
406
408
408
410
410
通道适配器
418
420
与结构的
物理链路

图5

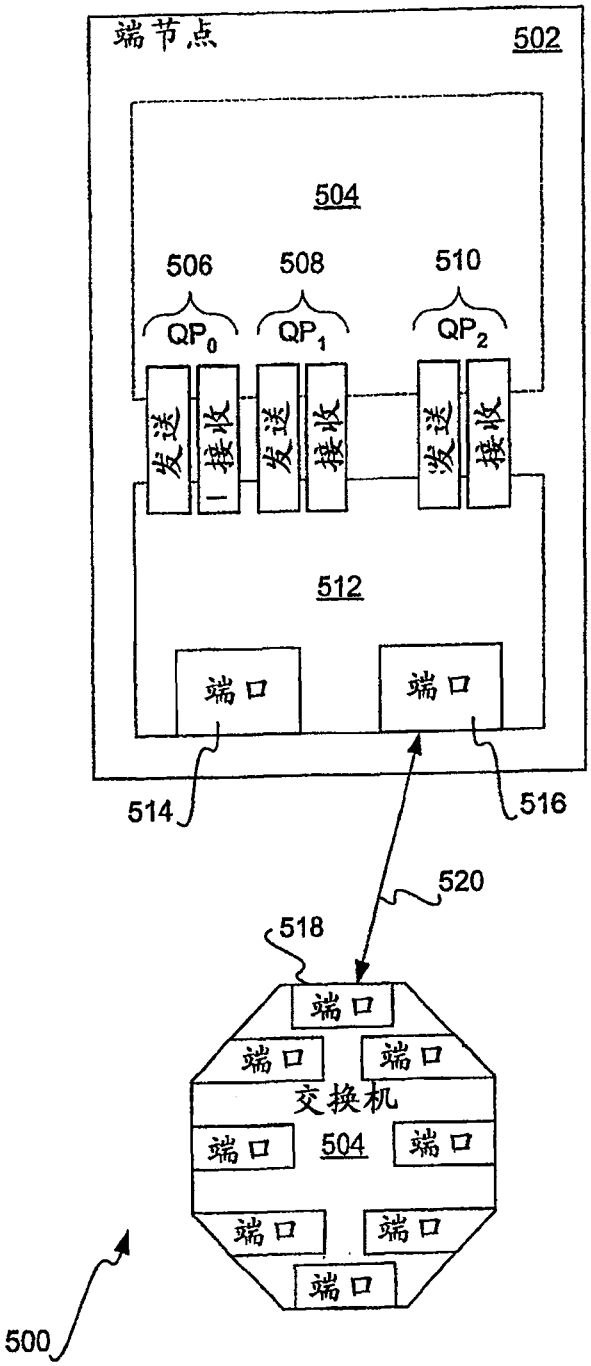

端节点
502
504
506
508
510
QP_0
QP_1
QP_2
发送
接收
发送
接收
发送
接收
512
端口
端口
514
516
520
518
端口
端口
端口
交换机
端口
504
端口
端口
端口
端口
500

图6

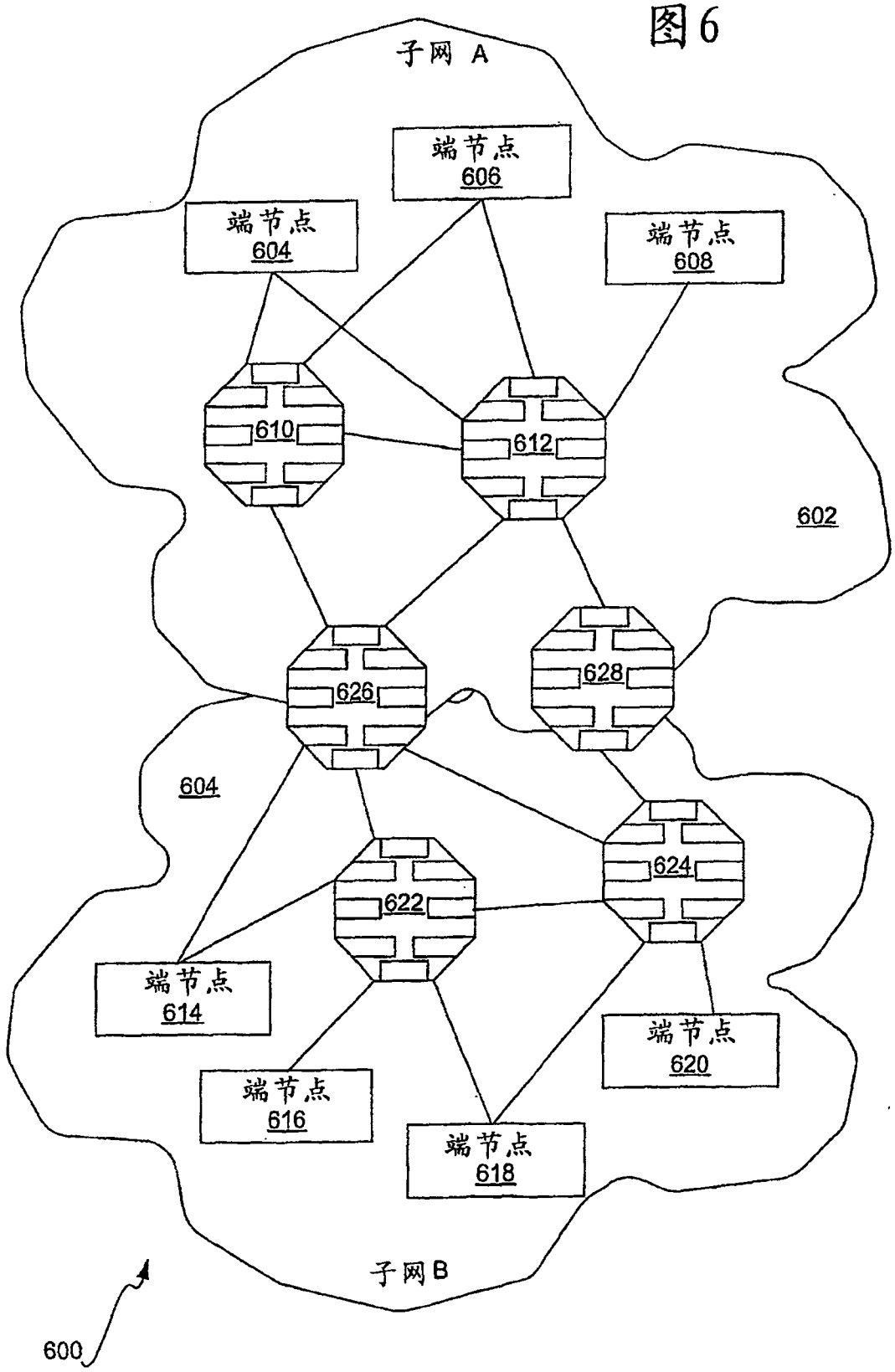
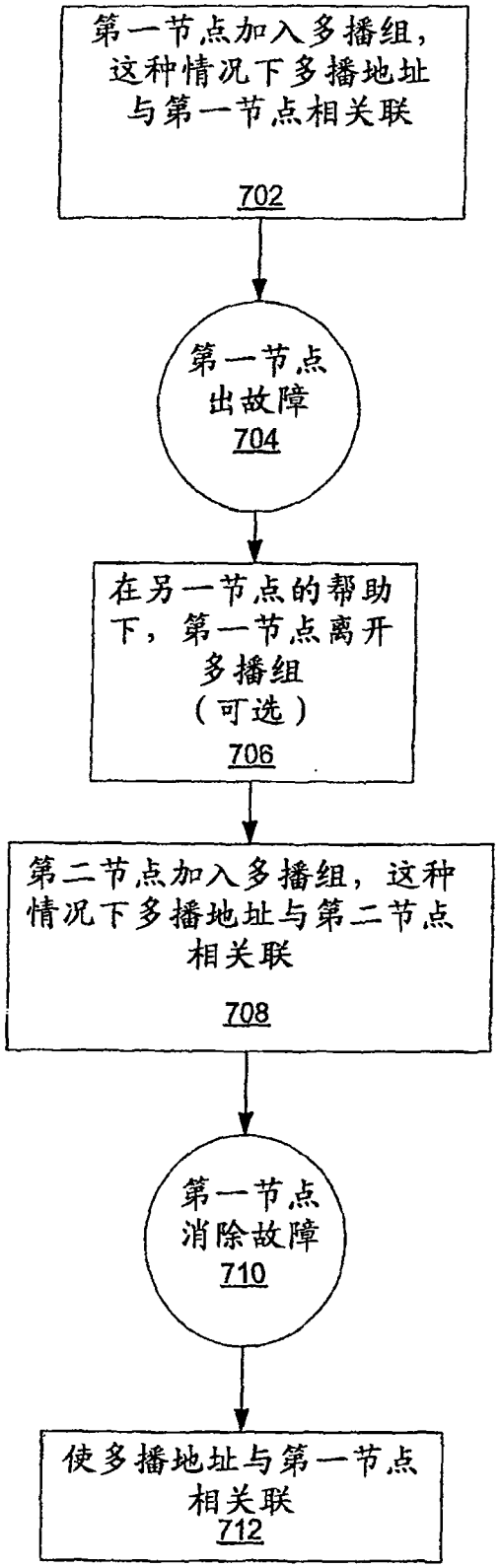

子网 A
端节点
606
端节点
604
端节点
608
610
612
602
626
628
604
622
624
端节点
614
端节点
620
端节点
616
端节点
618
子网B
600

图7


第一节点加入多播组，这种情况下多播地址与第一节点相关联
702
第一节点出故障
704
在另一节点的帮助下，第一节点离开多播组（可选）
706
第二节点加入多播组，这种情况下多播地址与第二节点相关联
708
第一节点消除故障
710
使多播地址与第一节点相关联
712

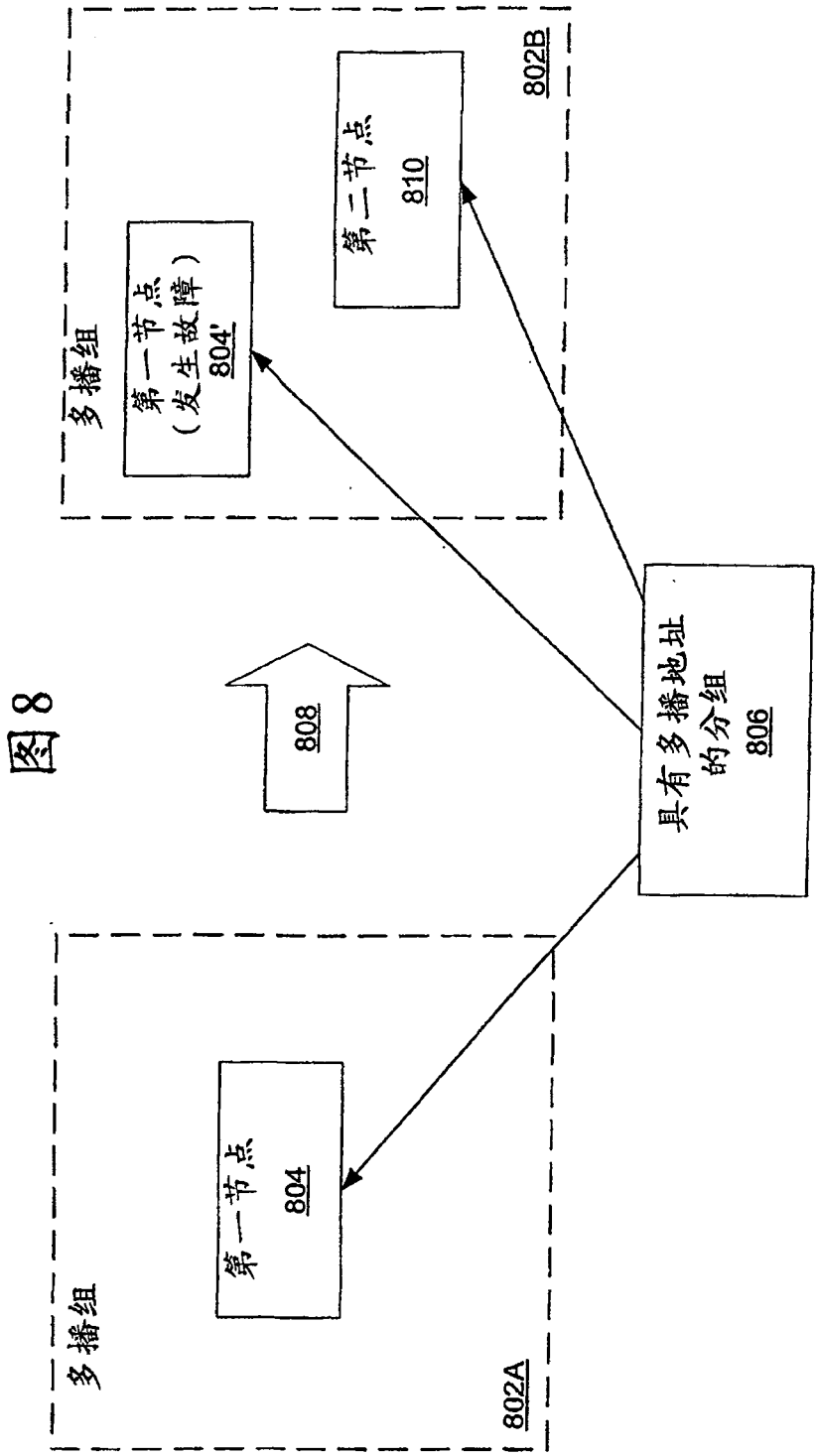

多播组
802B
第一节点
（发生故障）
804'
第二节点
810
808
具有多播地址
的分组
806
多播组
802A
第一节点
804

图8

图9

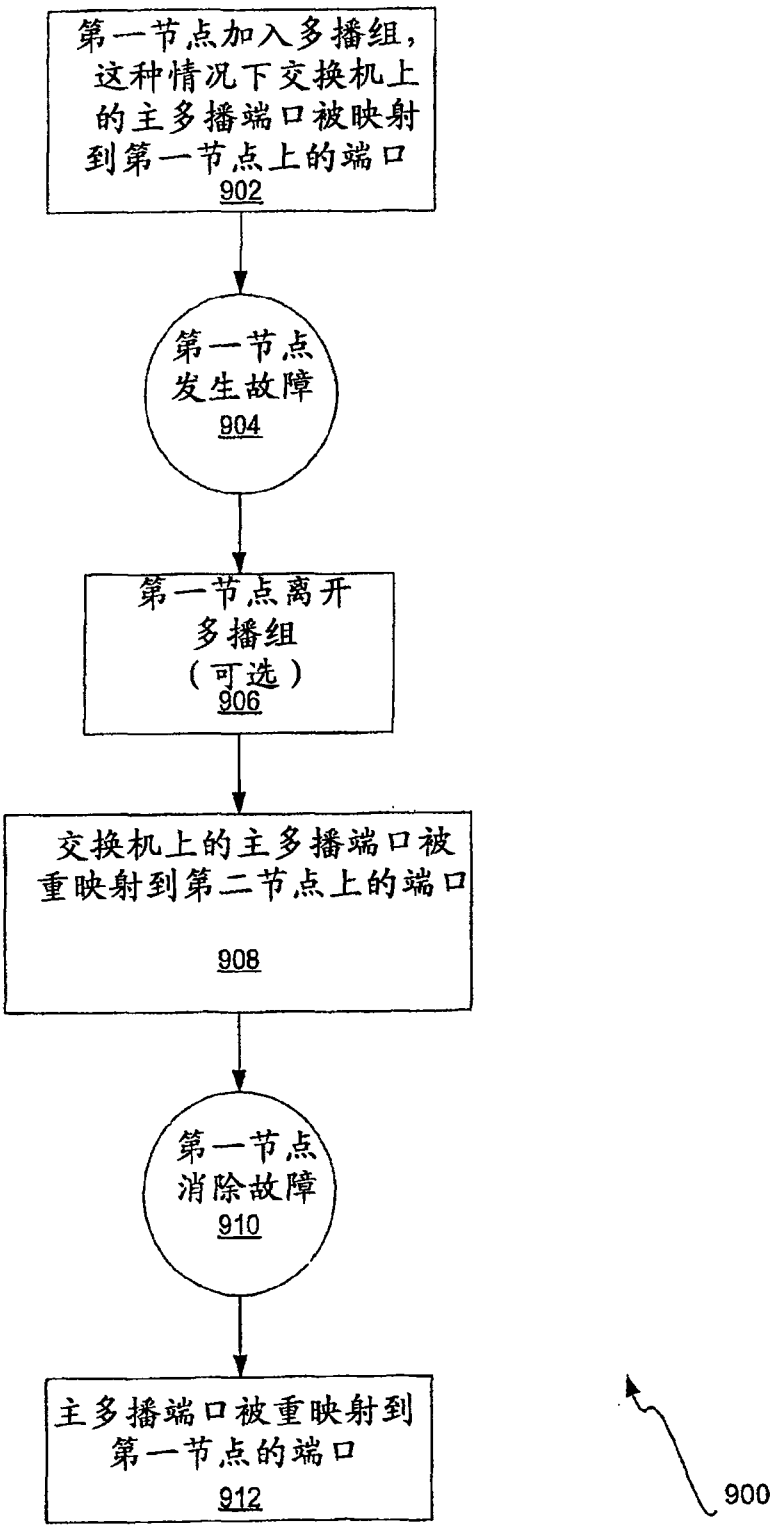
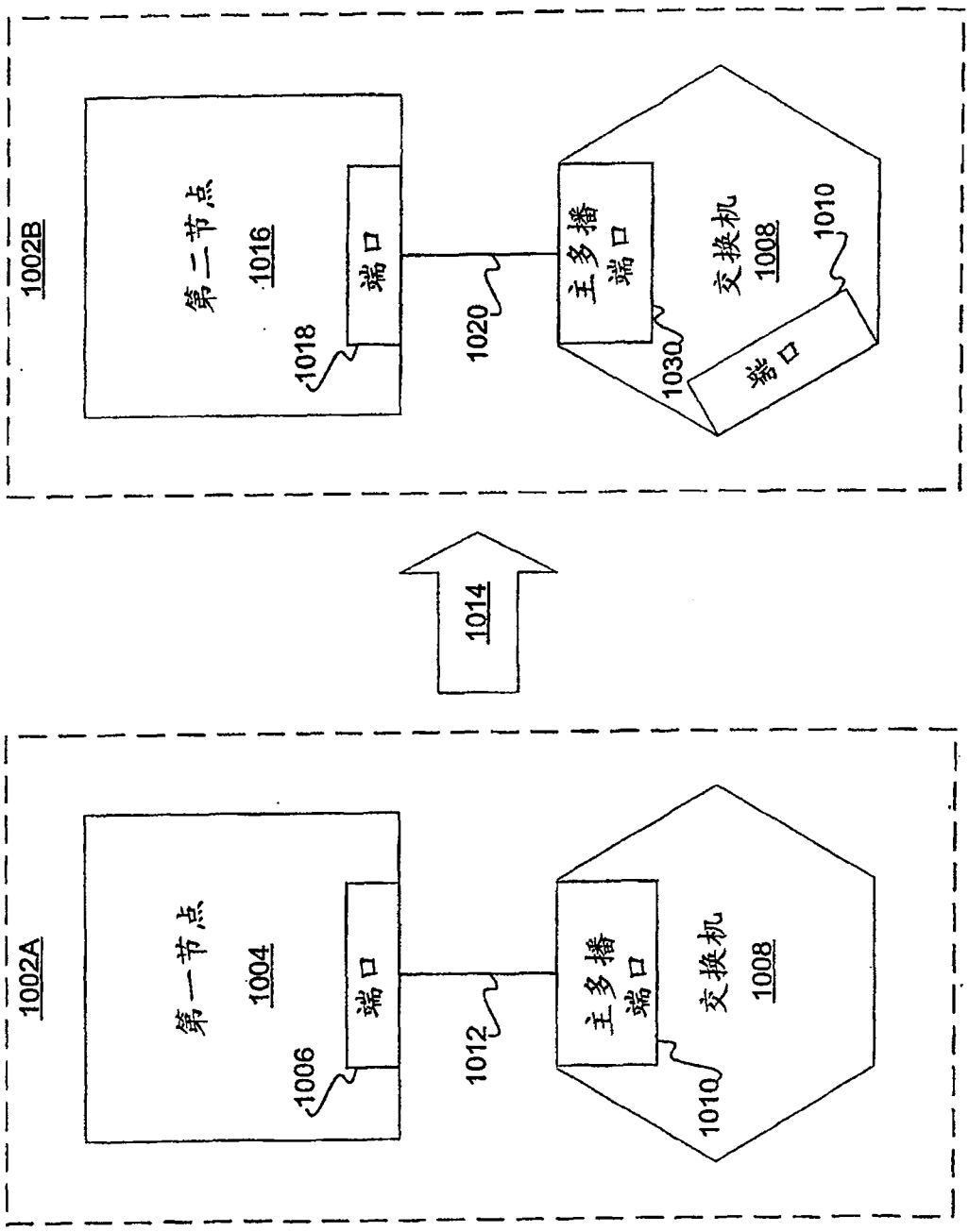

第一节点加入多播组，这种情况下交换机上的主多播端口被映射到第一节点上的端口
902
第一节点发生故障
904
第一节点离开多播组（可选）
906
交换机上的主多播端口被重映射到第二节点上的端口
908
第一节点消除故障
910
主多播端口被重映射到第一节点的端口
912
900

图10


1002B
第二节点
1016
端口
1018
1020
主多播端口
交换机
1008
1010
端口
1030
1014
1002A
第一节点
1004
端口
1006
1012
主多播端口
交换机
1008
1010