(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105981347 B
(45)授权公告日 2019.06.04

(21)申请号 201480075525.X

(22)申请日 2014.12.05

(65)同一申请的已公布的文献号
申请公布号 CN 105981347 A

(43)申请公布日 2016.09.28

(30)优先权数据
14/189,442 2014.02.25 US
14/189,403 2014.02.25 US

(85)PCT国际申请进入国家阶段日
2016.08.15

(86)PCT国际申请的申请数据
PCT/US2014/068832 2014.12.05

(87)PCT国际申请的公布数据
WO2015/130372 EN 2015.09.03

(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·D·约翰森 L·霍雷恩
D·G·莫克斯纳斯

(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 李晓芳

(51)Int.Cl.
H04L 29/06(2006.01)

(56)对比文件
CN 101057201 A,2007.10.17,
CN 103125097 A,2013.05.29,
CN 103125102 A,2013.05.29,
WO 2013009846 A1,2013.01.17,
US 2012072562 A1,2012.03.22,

审查员 张洁

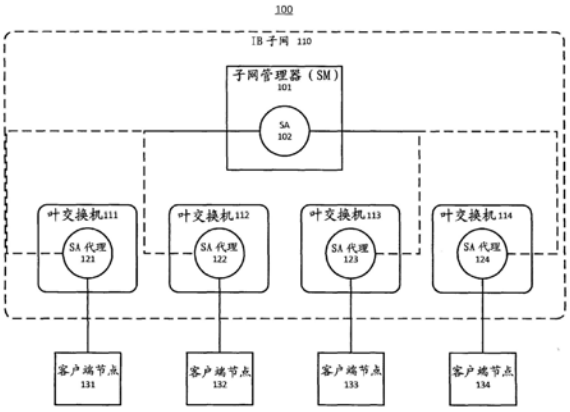
权利要求书4页 说明书10页 附图10页

(54)发明名称

在网络环境中支持子网管理的系统、方法和计算机介质

(57)摘要

一种系统和方法支持子网管理。系统将子网中的子网管理员(SA)与SA代理相关联。所述SA代理接收来自客户端节点的请求。所述SA可以处理从所述SA代理转发的所述一个或多个请求。为每一客户端节点分配专用队列对(QP)编号,以便不需要始终向预先定义的已知的QP编号发送初始请求。子网管理器(SM)检索用于在子网中的子网管理员(SA)和客户端节点之间建立可靠连接(RC)的信息。系统为与SM节点相关联的端口设置连接状态,以在与SM节点相关联的端口和与所述客户端节点相关联的端口之间建立RC连接。SM激活与所述客户端节点相关联的所述端口。



CN 105981347 B

1.一种用于在网络环境中支持子网管理的方法,包括:

使用在与客户端节点相关联的主机信道适配器HCA上工作的子网管理代理SMA来通告能力信息,所述能力信息指示所述客户端节点具有对于基于可靠连接RC的子网管理员SA访问的支持;

利用子网中的子网管理器SM节点上的子网管理器SM,检索用于在所述子网管理员SA和所述客户端节点之间建立可靠连接RC的所述能力信息;

为与所述SM节点相关联的端口,设置一个或多个连接状态,以在与所述SM节点相关联的所述端口和与所述客户端节点相关联的所述HCA的端口之间建立所述RC连接;

通过所述SM,激活与所述客户端节点相关联的所述HCA的所述端口;以及

通过所述RC在与所述客户端节点相关联的所述HCA的所述端口和与所述SM节点相关联的所述端口之间发送和接收通信。

2.根据权利要求1所述的方法,进一步包括:

在所述能力信息中包括可用于基于RC的SA访问的一组队列对QP编号。

3.根据权利要求1或2所述的方法,进一步包括:

为与所述客户端节点相关联的所述端口,建立第一组一个或多个RC队列对QP,其中,所述第一组一个或多个RC QP与为所述基于RC的SA访问分配的一个或多个队列对QP编号相关联。

4.根据权利要求3所述的方法,进一步包括:

为与所述SM节点相关联的所述端口建立第二组一个或多个RC队列对QP,其中,所述第二组一个或多个RC QP的QP编号基于以下各项中的一项:

被分配到与所述客户端节点相关联的所述端口的基本本地标识符(LID)的函数,以及

由所述SM为与所述客户端节点相关联的所述SMA设置的SMA属性。

5.根据权利要求1或2所述的方法,进一步包括:

使用管理消息交换来建立所述RC连接。

6.根据权利要求1或2所述的方法,进一步包括:

使用所述RC连接,通过所述客户端节点,将消息发送到所述SA。

7.根据权利要求1或2所述的方法,进一步包括:

使用多个RC连接来有效利用消息缓冲区存储器。

8.根据权利要求1或2所述的方法,进一步包括:

在所述客户端节点中建立预先定义的存储器区域,以包含通过一个或多个RC连接从所述SA和SA代理中的至少一项传输的状态信息。

9.根据权利要求1或2所述的方法,进一步包括:

通过所述SA,发送一个或多个多播消息以通知至少一个分区中的多个客户端节点关于一个或多个事件。

10.根据权利要求9所述的方法,进一步包括:

向每一所述多播消息分配序列号。

11.一种用于在网络环境中支持子网管理的系统,包括:

一个或多个微处理器;

在与客户端节点相关联的主机信道适配器HCA上工作的子网管理代理SMA,其中所述

SMA操作以：

通告能力信息，所述能力信息指示所述客户端节点具有对于基于可靠连接RC的子网管理员SA访问的支持；

在所述一个或多个微处理器上运行的子网中的子网管理器SM节点上的子网管理器SM，其中，所述SM操作以：

检索用于在所述子网管理员SA和所述客户端节点之间建立可靠连接RC的所述能力信息；

为与所述SM节点相关联的端口设置一个或多个连接状态，以在与所述SM节点相关联的所述端口和与所述客户端节点相关联的所述HCA的端口之间建立所述RC连接；

激活与所述客户端节点相关联的所述HCA的所述端口；以及

通过所述RC在与所述客户端节点相关联的所述HCA的所述端口和与所述SM节点相关联的所述端口之间发送和接收通信。

12.根据权利要求11所述的系统，其中：

与所述客户端节点相关联的子网管理代理SMA在所述能力信息中包括可用于基于RC的SA访问的一组队列对QP编号。

13.根据权利要求11或12所述的系统，其中：

所述客户端节点操作来为与所述客户端节点相关联的所述端口建立第一组一个或多个RC队列对QP，其中，所述第一组一个或多个RC QP与为所述基于RC的SA访问分配的一个或多个队列对QP编号相关联。

14.根据权利要求13所述的系统，其中：

所述SM操作来为与所述SM节点相关联的所述端口建立第二组一个或多个RC队列对QP，其中，所述第二组一个或多个RC QP的QP编号基于以下各项中的一项：

被分配到与所述客户端节点相关联的所述端口的基本本地标识符(LID)的函数，以及

由所述SM为与所述客户端节点相关联的所述SMA设置的SMA属性。

15.根据权利要求11或12所述的系统，其中：

所述SM操作来使用管理消息交换建立所述RC连接。

16.根据权利要求11或12所述的系统，其中：

所述客户端节点操作来使用所述RC连接将消息发送到所述SA。

17.根据权利要求11或12所述的系统，其中：

使用多个RC连接来有效利用消息缓冲区存储器。

18.根据权利要求11或12所述的系统，其中：

在所述客户端节点中设置预先定义的存储器区域，以包含通过RC连接从所述SA和SA代理中的至少一项传输的状态信息。

19.根据权利要求11或12所述的系统，其中：

所述SA操作来发送一个或多个多播消息以通知至少一个分区中的多个客户端节点关于一个或多个事件，以及向每一所述多播消息分配序列号。

20.一种在其上存储了当执行时使得系统执行权利要求1到10中的任一项的方法的非瞬时的机器可读的存储介质。

21.一种在其上存储了当执行时使得系统执行以下步骤的指令的非瞬时的机器可读存

储介质,所述步骤包括:

使用在与客户端节点相关联的主机信道适配器HCA上工作的子网管理代理SMA来通告能力信息,所述能力信息指示所述客户端节点具有对于基于可靠连接RC的子网管理员SA访问的支持;

通过子网中的子网管理器SM节点上的子网管理器SM,检索用于在所述子网管理员SA和所述客户端节点之间建立可靠连接RC的所述能力信息;

为与所述SM节点相关联的端口,设置一个或多个连接状态,以在与所述SM节点相关联的所述端口和与所述客户端节点相关联的所述HCA的端口之间建立所述RC连接;以及

通过所述SM,激活与所述客户端节点相关联的所述HCA的所述端口;以及

通过所述RC在与所述客户端节点相关联的所述HCA的所述端口和与所述SM节点相关联的所述端口之间发送和接收通信。

22.一种子网管理器,包括:

被配置成检索用于在子网管理员SA和客户端节点之间建立可靠连接RC的能力信息的检索单元;

其中所述能力信息由在与客户端节点相关联的主机信道适配器HCA上工作的子网管理代理SMA通告;

设置单元,其被配置成为与子网管理器SM节点相关联的端口设置一个或多个连接状态,以在与所述SM节点相关联的所述端口和与所述客户端节点相关联的所述HCA的端口之间建立所述RC连接;以及

被配置成激活与所述客户端节点相关联的所述HCA的所述端口的激活单元。

23.根据权利要求22所述的子网管理器,其中:

所述能力信息包括可用于基于RC的SA访问的一组队列对QP编号。

24.根据权利要求22所述的子网管理器,其中:

所述客户端节点操作来为与所述客户端节点相关联的所述端口建立第一组一个或多个RC队列对QP,其中,所述第一组一个或多个RC QP与为所述基于RC的SA访问分配的一个或多个队列对QP编号相关联。

25.根据权利要求24所述的子网管理器,进一步包括:

建立单元,其被配置成为与所述SM节点相关联的所述端口建立第二组一个或多个RC队列对QP,其中,所述第二组一个或多个RC QP的QP编号基于以下各项中的一项:

被分配到与所述客户端节点相关联的所述端口的基本本地标识符(LID)的函数,以及

基于由所述SM为与所述客户端节点相关联的所述SMA设置的SMA属性。

26.根据权利要求22所述的子网管理器,其中:

使用管理消息交换来建立所述RC连接。

27.根据权利要求22所述的子网管理器,其中:

所述客户端节点操作来使用所述RC连接将消息发送到所述SA。

28.根据权利要求22所述的子网管理器,其中:

使用多个RC连接来有效利用消息缓冲区存储器。

29.根据权利要求22所述的子网管理器,其中:

在所述客户端节点中建立预先定义的存储器区域,以包含通过RC连接从所述SA和SA代

理中的至少一项传输的状态信息。

30.根据权利要求22所述的子网管理器,进一步包括:

发送单元,其被配置成发送一个或多个多播消息,以通知至少一个分区中的多个客户端节点关于一个或多个事件,以及

分配单元,其被配置成给每一所述多播消息分配序列号。

在网络环境中支持子网管理的系统、方法和计算机介质

[0001] 版权声明：

[0002] 此专利文献的公开的一部分包含受版权保护的材料。版权所有者不反对任何人影印专利文献或专利说明书，因为它出现在专利商标局专利文件或记录中，但在别的方面却保留所有版权。

技术领域

[0003] 本发明一般涉及计算机系统，更具体地涉及用于中间件以及应用执行的工程系统。

背景技术

[0004] 随着引入较大的云计算架构，与传统的网络和存储相关联的性能和管理瓶颈变得越来越严重。无线带(InfiniBand)(IB)技术已经作为云计算结构的基础越来越多地被部署。这是本发明的各实施例旨在解决的一般领域。

发明内容

[0005] 此处描述了可以支持网络环境(诸如用于中间件和应用执行或中间件机器环境的工程系统)中的子网管理的系统和方法。系统可以将子网中的子网管理员(SA)与一个或多个SA代理相关联。进一步地，所述一个或多个SA代理可以接收来自一个或多个客户端节点的一个或多个请求。然后，所述SA可以处理从所述一个或多个SA代理转发的所述一个或多个请求。另外，还可以为每一客户端节点分配专用队列对(QP)编号，以便不需要始终向预先定义的已知的QP编号发送初始请求。

[0006] 此处描述了可以支持网络环境(诸如用于中间件和应用执行或中间件机器环境的工程系统)中的子网管理的系统和方法。子网管理器(SM)可以检索用于在子网中的子网管理员(SA)和客户端节点之间建立可靠连接(RC)的信息。进一步地，系统还可以为与SM节点相关联的端口设置一个或多个连接状态，以在与SM节点相关联的端口和与所述客户端节点相关联的端口之间建立RC连接。然后，SM可以激活与所述客户端节点相关联的端口。

附图说明

[0007] 图1示出了根据本发明的一实施例的用于在网络环境中支持分布式子网管理员(SA)方案的图示。

[0008] 图2示出了根据本发明的一实施例的防止网络环境中的对SA访问的拒绝服务(DoS)攻击的图示。

[0009] 图3示出了根据本发明的一实施例的用于在网络环境中支持分布式子网管理员(SA)方案的示例性流程图。

[0010] 图4示出了根据本发明的一实施例的用于在网络环境中支持基于可靠连接(RC)的子网管理员(SA)访问的图示。

[0011] 图5示出了根据本发明的一实施例的在网络环境中支持子网管理员(SA)和多个客户端节点之间的通信的图示。

[0012] 图6示出了根据本发明的一实施例的用于在网络环境中支持基于可靠连接(RC)的子网管理员(SA)访问的示例性流程图。

[0013] 图7是根据本发明的一实施例的用于支持网络环境中的子网管理的系统的框图。

[0014] 图8是根据本发明的一实施例的用于支持分布式数据网格中的异步调用的子网管理器的框图。

[0015] 图9示出了本发明的实施例的功能配置。

[0016] 图10示出了用于实现本发明的实施例的计算机系统。

具体实施方式

[0017] 本发明通过示例的方式而不是限制的方式示出在各个附图的图中,在附图中类似的参考编号表示类似的元素。应该注意,在本发明中对“一个”或“某些”实施例的指代不一定是指同一个实施例,而这样的指代表示至少一个。

[0018] 如下的本发明的描述使用Infiniband(IB)网络作为高性能网络的示例。对所属领域的技术人员显而易见的是,可以使用其他类型的高性能网络,而没有任何限制。

[0019] 此处描述了可以支持网络环境(诸如用于中间件和应用执行或中间件机器环境的工程系统)中的子网管理员(SA)访问的系统和方法。

[0020] InfiniBand(IB)架构

[0021] IB架构是串行点对点技术。IB网络、或子网中的每一个都可包括一组使用交换机和点对点链路互连的主机。单一子网可以可扩展到一万以上的节点,而两个或更多子网可以使用IB路由器来互连。子网内的主机和交换机使用本地标识符(LID)来寻址,例如,单一子网可以仅限于49151个单播地址。

[0022] IB子网可以使用至少一个子网管理器(SM),该子网管理器负责初始化并启动包括驻留在子集中的交换机、路由器和主机通道适配器(HCA)上的所有IB端口的配置的子网。SM的职责还包括路由表计算和部署。网络的路由旨在获取完全连接、死锁自由,以及所有源和目的地对之间的负载平衡。路由表可以在网络初始化时计算,每当拓扑变化时,都可以重复此过程,以便更新路由表,并确保最佳性能。

[0023] IB网络中的HCA可以使用队列对(QP)来相互进行通信。队列对是在通信建立过程中创建的,并提供一组初始属性,诸如QP编号、HCA端口、目的地LID、队列大小,以及传输服务。另一方面,当通信结束时,销毁与通信中的HCA相关联的QP。HCA可以处理许多QP,每一QP都包括一对队列,即发送队列(SQ)和接收队列(RQ)。有一个这样的对存在于参与通信的每一个端节点中。发送队列持有要传输到远程节点的工作请求,而接收队列持有有关对从远程节点接收到的数据执行什么的信息。除QP之外,每一个HCA都可以具有一个或多个与一组发送和接收队列相关联的完成队列(CQ)。CQ持有发布到发送和接收队列的工作请求的完成通知。

[0024] IB架构是灵活的架构。配置和维护IB子网可以通过特殊的带内子网管理数据包(SMP)来执行。SM的功能可以原则上从IB子网中的任何节点来实现。IB子网中的每一个末端端口都可以具有相关联的子网管理代理(SMA),该SMA负责处理定向到它的基于SMP的请求

数据包。在IB架构中，相同端口可以表示使用基于SMP的通信的SM实例或其他软件组件。如此，只有定义明确的SMP操作的子集可以由SMA来处理。

[0025] SMP使用结构中的专用数据包缓冲区资源，例如，不受流量控制的特殊虚拟通道(VL15)(即，在缓冲区溢出的情况下，可以丢弃SMP数据包)。此外，SMP可以使用SM基于末端端口本地标识符(LID)设置的路由，或者，SMP可以使用直接路由，其中，路由完全由发送方所定义，并嵌入在数据包中。通过使用直接路由，数据包的路径就HCA和交换机上的端口号的有序序列而言穿过结构。

[0026] SM可以使用每一交换机和/或每一HCA中存在的SMA来监测网络中的变化。SMA使用陷阱和通知，将诸如新连接、断开连接，以及端口状态变化之类的变化传递到SM。陷阱是被发送以警告端节点关于某一事件的消息。陷阱可以包含带有描述事件的细节的通知属性。可以为不同的事件定义不同的陷阱。为了减少陷阱的不必要的分发，IB应用事件转发机制，其中，需要端节点显式地预订它们希望被通知的陷阱。

[0027] IB架构提供分区，作为定义哪些IB末端端口应该被允许与其他IB末端端口进行通信的方式。分区是为IB结构上的所有非SMP数据包定义的。除默认分区以外的分区的使用是可选的。数据包的分区可以由包括15比特分区号码和单一比特成员类型(完全或受限)的16比特P_Key进行定义。

[0028] 主机端口，或HCA端口的分区成员资格可以基于这样的前提：SM设置带有P_Key值的端口的P_Key表，该P_Key表对应于该主机的当前分区成员资格策略。为了补偿主机可能不是完全受信任的可能性，IB体系结构还定义了交换机端口可以可选地建立以进行分区实施。因此，然后，可以建立连接到主机端口的交换机端口的P_Key表以反映主机端口被认为是其成员的相同分区(即，本质上相当于以太网LAN中的交换机实施的VLAN控制)。

[0029] 由于IB架构允许通过SMP对IB子网进行完全带内配置和维护，因此，SMP本身不受任何分区成员资格限制的影响。如此，为了避免IB结构上的任何粗略的或受损的节点能够定义任意结构配置(包括分区成员资格)的可能性，需要其他保护机制。

[0030] 由IB架构所提供的灵活性允许IB结构/子网(例如，HPC群集)的管理员判断是否要使用结构中的一个或多个交换机上的嵌入的SM实例和/或建立IB结构上的一个或多个主机以执行SM功能。此外，还由于SM所使用的SMP所定义的线路协议通过API可用，因此，不同的工具和命令可以基于这样的SMP用于发现、诊断来实现，并独立于任何当前子网管理器操作被控制。

[0031] 从安全角度来看，IB体系结构的灵活性表示，在对连接到IB结构的各种主机的根访问和允许IB结构配置的访问的根访问之间没有根本的差异。这对在物理上是安全的并且稳定的系统是好的。然而，这对于其中IB结构上的不同的主机由不同的系统管理员来控制的以及这样的主机在IB结构中应该在逻辑上彼此隔离的系统配置是会有问题的。

[0032] 子网管理员(SA)和拒绝服务(DoS)攻击

[0033] 子网管理员(SA)是与主控SM相关联的可以存储有关子网的不同的信息的子网数据库。可以帮助各种端节点建立QP的与SA的通信，可以基于通过指定的QP，例如，QP1，发送一般服务管理数据报(MAD)。另外，发送方和接收方两者都可能需要诸如源/目的地LID、服务级别(SL)，以及最大传输单元(MTU)之类的信息，来通过QP建立通信。可以从由SA所提供的被称为路径记录的数据结构中检索这样的信息。为了获取路径记录，端节点可以向SA执

行路径记录查询，例如，使用SubnAdmGet/SubnAdmGetable操作。然后，SA可以将请求的路径记录返回到端节点。

[0034] 例如，IB规范定义，到子网管理员(SA)的所有请求都被发送到由本地IB端口信息中的SM_LID的值所定义的(即，由子网管理器(SM)设置的)目的地LID。此外，请求还可以使用由SM_SL值所定义的SL值，该值也可以由SM在本地IB端口信息中设置。

[0035] 另外，IB规范还定义，用于与SA建立通信的目的地QP是一般服务接口QP(即，QP1)。进一步地，当不能基于每个客户端端口，以在作出初始SA请求之前允许客户端节点观察到此信息的方式，设置对于这些值的替代值时，与SA的通信可以在默认分区上执行。在初始请求之后，SA可以向客户端发送重定向响应，并指示客户端使用不同的地址来访问SA。地址的这种改变可以涉及不同的目的地端口，不同的SL、不同的分区以及不同的目的地QP编号。

[0036] 如此，只要初始请求需要使用默认分区被发送到QP1，就有会在子网中发生对SA访问的拒绝服务(DoS)攻击的可能性，例如，当流氓客户端，或处于错误状态的客户端使SA端口的QP1过载时，和/或当SA端口本身由于来自多个源的巨大的业务而被过载时。

[0037] 分布式子网管理员(SA)方案

[0038] 图1示出了根据本发明的实施例的用于在网络环境中支持分布式子网管理员(SA)方案的图示。如图1所示，网络环境100中的IB子网110可包括多个交换机(例如，胖树拓扑或从主机端口的连接是通过叶交换机的任何拓扑中的叶交换机111-114)。进一步地，IB子网110可包括子网管理器(SM)101和子网管理员(SA)102。

[0039] 为了避免利用SA请求使SA端口过载，系统可以使用分布式SA方案来通过SA代理121-124，将系统中的总的SA 102客户端负载分发到不同的SA端口。例如，直接连接到叶交换机111-114的客户端节点131-134的SA代理121-124可以是相应的叶交换机111-114上的管理处理器。

[0040] 另外，本地叶交换机111-114上的交换机实现可以确保本地叶交换机111-114可以只获得到SA 102代理端口的输入带宽的其公平/合理的份额。如此，通过将SA客户端节点131-134限制为只使用本地叶交换机121-124上的本地SA代理121-124，没有客户端节点可以防止其他本机客户端的向前进展。

[0041] 进一步地，SA代理实施方式可以确保对来自不同的本地SA客户端131-134的请求的实际处理可以在不同的本地客户端131-134之间带有合理的交错而发生。如此，每一表现好的客户端节点131-134都可以观察，或实现，可接受的响应时间和合理的向前进展。

[0042] 根据本发明的实施例，与SA客户端节点131-134不同，SA代理121-124可以表示被总的结构实施方式完全控制并且仅仅由带有结构的所有者特权的管理实体(人或软件)管理的受信任的软件组件。如此，在总的结构状态被中央实体(例如，SM 101)保持的情况下，各种SA代理实例121-124和中央实体之间的协议可以确保所有SA代理实例121-124的公平性和向前进展。

[0043] 图2示出了根据本发明的实施例的在网络环境中防止对SA访问的拒绝服务(DoS)攻击的图示。如图2所示，网络环境200可包括多个交换机(例如，叶交换机211-213)。进一步地，叶交换机211-213中的每一个都可以被配置成支持可以与SA 202进行通信的SA代理221-223。

[0044] 另外，叶交换机211-213中的每一个都可以连接到一个或多个客户端节点或主机

节点。例如，客户端节点231-232通过SA端口241连接到SA代理221，客户端节点233-235通过SA端口242连接到SA代理222，客户端节点236-238通过SA端口243连接到SA代理223。

[0045] 进一步地，可以在网络环境200中设置用于SA访问的各种专用的分区。如图2所示，可以为结构中的客户端节点的物理子集(例如，直接连接到叶交换机213的一组节点236-238)设置用于SA访问的专用分区250。此外，还可以为逻辑子集(例如，属于特定的承租人或系统以及可以已经一开始就共享一个分区的一组节点232-234)设置用于SA访问的专用的分区252。如此，SM 201可以防止客户端节点231-238访问不在相关分区里的任何SA端口，并由此防止这样的业务被发出。

[0046] 此外，还可以为每一客户端节点231-238分配专用QP编号，以便不需要始终向QP1发送初始请求。然后，在指定的SA端口，对于指定的QP编号，只有指定的客户端节点才可以被接受。例如，SA实施方式可以确保每一客户端节点231-238只能使用显式地与每一指定的QP相关联的专用的一组数据包接收缓冲区。如此，系统可以确保在端口级别，没有客户端节点能够消耗所有可用接收缓冲区，并由此防止其他客户端向前进展。

[0047] 根据本发明的实施例，系统可以扩展IB端口信息以包括定义各种专用分区的SA_Partition字段，以及定义各种专用QP编号的SA_QP领域。如此，系统可以防止任何本地客户端节点在本地SA端口上过载QP1，并可以防止任何节点向非本地SA端口发送请求，甚至在所有SA端口都是默认分区的完全会员的情况下。

[0048] 另外，系统还可以为客户端端口促进分布式SA方案，甚至在相关联的SMA实施方式不允许对由SM设置为IB端口信息的信息的任何修改的情况下。例如，用于选择用于访问SA/SA-代理的分区号码(P_Key)的一个约定可以是使用本地端口P_Key表中的第一P_Key值。

[0049] 此外，为了使用在没有任何显式的额外的本地IB端口信息的情况下先验地已知的专用QP编号，一个约定可以是使用本地端口LID值作为目的地QP编号。可另选地，系统可以使用本地LID编号为用于SA访问的目的地QP编号内的比特字段中的值。在任一种情况下，目的地QP编号的其余部分可以是预先定义的已知的编号(即，诸如当前QP1值)，或者可以被定义为各种IB客户端节点231-238上的IB软件栈的通用配置参数的一部分的配置参数。

[0050] 根据本发明的实施例，系统可以基于为每一个客户端端口提供受信任的HCA实施方式。例如，受信任的HCA实施方式可以确保每一SA请求中的源LID始终是由SM为对应的客户端端口进行定义的正确的值。进一步，SA代理实施方式可以确保客户端源LID值是每个客户端端口目的地QP编号的一部分，例如，作为特定的比特字段。可另选地，可以设置目的地QP上下文以定义用于允许传入的数据包以类似于连接模式QP(例如，如在IB规范中定义的)的方式被输送到QP的特定的数据包源LID。如此，SA端口硬件实施方式可以允许在QP级别的直接访问控制，无需相对于旧式实施方式的显著的增强，也无需取决于每一客户端端口使用专用分区用于访问SA代理实例。

[0051] 图3示出了根据本发明的实施例的用于在网络环境中支持分布式子网管理员(SA)方案的示例性流程图。如图3所示，在步骤301中，系统可以将子网中的子网管理员(SA)与一个或多个SA代理相关联。然后，在步骤302中，所述一个或多个SA代理可以接收来自一个或多个客户端节点的一个或多个请求。进一步地，在步骤303中，所述SA可以处理从所述一个或多个SA代理转发的所述一个或多个请求。

[0052] 用于子网管理员(SA)访问的可靠连接(RC)

[0053] 根据本发明的实施例，系统可以提供基于可靠连接(RC)的对子网管理员(SA)的访问(例如，支持端对端可靠性、完整传输分摊、大消息和远程直接存储器访问的IB RC传输服务)。

[0054] 图4示出了根据本发明的实施例的用于在网络环境中支持基于可靠连接(RC)的子网管理员(SA)访问的图示。如图4所示，网络环境400可包括连接到主机信道适配器(HCA)404的客户端节点402和SM节点403上的子网管理器(SM)401。

[0055] 根据本发明的实施例，用于客户端节点402的HCA 404上的子网管理代理(SMA)406可以通告表示客户端节点402支持基于RC的SA访问的能力405。

[0056] 另外，SMA 406还可以通过通过一组可用于基于RC的SA访问的QP编号415，来公开一个或多个队列对(QP)411。例如，SMA406可以确保它具有为基于RC的SA访问分配的已定义的QP编号集合(例如，QP编号2-N)，或者SMA 406可以将基于RC的SA访问分配的QP编号报告为一个或多个节点特定的SMA属性值(例如，n到m，其中，n和m两者都可以由客户端节点402进行定义)。

[0057] 进一步地，与子网管理员(SA)422相关联的SM 401能够检索有关IB子网中的每一客户端节点的类型的信息以及有关这些客户端节点的不同的能力的信息。SM 401可以在发现过程中记录能力405以及QP编号415信息。如此，SM 401可以获得用于在其一端建立RC连接的必要的信息。

[0058] 然后，SM 401和/或SA 422可以在SM节点403上建立本地RC QP 412，例如，在端口413后面。这里，可以使用SM 401用来设置所有端口的SM LID属性来定义端口413。

[0059] 如图4所示，在利用特定的客户端节点402的连接状态配置RC QP 412之后，SM 401可以激活客户端节点402上的端口414。例如，SM 401可以通过将端口414设置为活动状态，使客户端节点402上的端口414运转。

[0060] 进一步地，SM 401可以显式地向客户端节点402发送信号，通知基于RC的SA访问可用。例如，由SM 401和/或SA 422代表客户端节点402建立的RC QP 412的QP编号，可以是被分配到客户端节点端口414的基本LID的函数。换言之，客户端节点402能够先验地知道RC QP 412的QP编号。

[0061] 可另选地，可以独立地并动态地分配用于与客户端节点402进行通信的RC QP 412的QP编号。例如，RC QP 412的QP编号可以是SMA属性，SM 401可以为客户端节点402上的SMA 406与其他属性(诸如SM LID)一起设置该SMA属性。

[0062] 然后，在客户端节点402一侧的端口411由SM 401激活之后，客户端节点402可以利用用于SA访问的连接信息，更新本地RC QP 411。可以从一个或多个SMA属性中检索这样的连接信息，或者可以基于无先验已知的信息，获得这样的连接信息。随后，客户端节点402可以开始向SA 422发送请求，例如，在没有与SA 422的任何进一步的连接管理协议通信的情况下。

[0063] 根据本发明的实施例，系统可以利用如下约定：使用将客户端节点一侧的端口设置为活动状态作为SA准备接收RC QP上的传入的消息的指示。如图4所示，在客户端节点402上的端口414被激活之后，客户端节点402可以意识到SA 422准备接收RC QP 412上的传入的消息(即，RC QP 412已经被配置为操作状态)。

[0064] 进一步地，SA 422可以假设在从客户端节点402接收到第一传入的消息之后，已经

由客户端节点402建立了RC连接410。相应地，SA 403可以向客户端节点402发送一个或多个消息，可包括响应和各种事件的通知。

[0065] 可替换地，客户端节点402和SA 422可以基于通信管理消息交换420，独立于QP编号是如何建立的，来建立必需的RC连接。

[0066] 另外，为了有效利用消息缓冲区存储器，系统可以扩展上述协议以支持多个连接。根据本发明的实施例，系统允许为不同的连接使用不同的最大消息大小。例如，系统可以使用一个RC连接来支持简单小消息请求/响应协议，并使用另一RC连接来支持较大的消息，诸如多路径响应。

[0067] 如图4所示，系统可以在客户端节点402中设置一个或多个预先定义的存储器区域421，用于包含通过一个或多个RC连接从SA 422，或SA代理传输的状态信息。例如，预先定义的存储器区域421可以包含从本地客户端节点402到客户端节点402可以到达的所有其他客户端节点的所有可能的路径。

[0068] 进一步地，SA 422可以使用RDMA写入，更新预先定义的存储器区域421，例如，在使用大量的路径的情况下。例如，任何随后的更新都可以使用选择的RDMA写入来实现，选择的RDMA写入指向需要更新的准确的部分。另外，这样的选择的更新可以伴随有通过用于RC连接的专用事件实现的或者作为与RDMA写入操作相关联的直接数据实现的事件消息。

[0069] 根据本发明的实施例，客户端节点402可以使用基于管理数据报(MAD)的协议来支持SA访问，例如，在SM 401不显式地表明用于SA访问的RC能力的情况下。例如，可以基于MAD(这是256字节不可靠数据报(UD))，定义IB子网中的客户端节点和SA之间的交互和通信协议。这里，基于UD的MAD表示所有IB端节点能够支持的最小数据包格式和协议。

[0070] 进一步地，为了提供其大小大于单一MAD数据包的消息的可靠传输，可以在MAD协议上面支持诸如可靠多数据包协议(RMPP)之类的额外的基于软件的协议。另一方面，为可靠消息传送使用基于软件的协议可能没有为可靠连接使用HCA能力那样有效率，特别是对于SA功能在启用的HCA后面实现并且大多数客户端从启用的HCA操作的情况。

[0071] 图5示出了根据本发明的实施例的在网络环境中支持子网管理员(SA)和多个客户端节点之间的通信的图示。如图5所示，网络环境500可包括子网管理器(SM)501和多个客户端节点(例如，客户端节点511-514)。进一步地，SM 501可以与子网管理员(SA)502相关联，而客户端节点511-514可以连接到HCA 521-524，其中每一个HCA都支持SMA 531-534。

[0072] 根据本发明的实施例，系统允许SA 502发送不同的多播消息，用于与大量的客户端节点进行通信。例如，SA 502可以在IB子网的不同的分区(例如，一个或多个SA访问分区)发送一个或多个多播消息，以便通知客户端节点511-514关于相关的事件。

[0073] 进一步地，系统可以在相关的分区中用不同的序列号配置多播消息。如此，相关的分区中的客户端节点511-514可以检测缺少的事件，随后，可以请求SA 503提供缺少的事件消息。如图5所示，客户端节点511-514可以使用基于RC的消息(即，单播消息)来确认接收到基于多播的事件消息。

[0074] 如图5所示，当SA 502判断没有从客户端节点514接收到对于基于多播的事件消息的确认时，SA 502可以再次使用单播消息，将事件消息发送到客户端节点514，例如，通过从SA 502到客户端节点514的专用RC连接。

[0075] 图6示出了根据本发明的实施例的用于在网络环境中支持基于可靠连接(RC)的子

网管理员(SA)访问的示例性流程图。如图6所示，在步骤601中，子网中的子网管理器(SM)可以检索用于在子网管理员(SA)和客户端节点之间建立可靠连接(RC)的信息。然后，在步骤602中，SM可以为与SM节点相关联的端口设置一个或多个连接状态，以在与SM节点相关联的端口和与所述客户端节点相关联的端口之间建立RC连接。进一步地，在步骤603中，SM可以激活与所述客户端节点相关联的端口。

[0076] 图7是根据本发明的一实施例的用于支持网络环境中的子网管理的系统700的框图。系统700的块可以通过硬件、软件或硬件和软件的组合来实现以执行本发明的原理。所属领域的技术人员可以理解，在图7中所描述的块可以被组合或分成子块以实现如上文所描述的本发明的原理。因此，此处的描述可以支持任何可能的组合或分离或此处所描述的功能块的进一步的定义。

[0077] 如图7所示，用于支持网络环境中的子网管理的系统700包括关联单元710、接收单元720以及处理单元730。关联单元710可以将子网中的子网管理员(SA)与一个或多个SA代理相关联。接收单元720可以通过所述一个或多个SA代理，接收来自一个或多个客户端节点的一个或多个请求。处理单元730可以通过所述SA，处理从所述一个或多个SA代理转发的所述一个或多个请求。

[0078] 根据本发明的实施例，子网是实现为胖树拓扑或其中来自主机端口的连接是通过叶交换机的拓扑中的至少一项的InfiniBand(IB)子网。

[0079] 根据本发明的实施例，可以使用直接连接到客户端节点的叶交换机上的管理处理器作为SA代理。

[0080] 根据本发明的实施例，可以为所述一个或多个客户端节点设置用于SA访问的专用分区。一个或多个客户端节点可包括网络结构中的客户端节点的物理子集，以及网络结构中的客户端节点的逻辑子集中的至少一项。

[0081] 根据本发明的实施例，可以为每一客户端节点分配专用队列对(QP)编号。

[0082] 根据本发明的实施例，每一客户端节点只能使用显式地与每一指定的QP编号相关联的专用的一组数据包接收缓冲区。

[0083] 根据本发明的实施例，对从不同的客户端节点接收到的一个或多个请求的实际处理可以带有合理的交错地发生。

[0084] 根据本发明的实施例，系统700还可以包括管理结构状态的中央实体，该中央实体确保一个或多个SA代理的公平性和向前进展。

[0085] 根据本发明的实施例，可以使用一个或多个约定来为一个或多个客户端端口选择专用的分区号码以及专用的QP编号，和/或，受信任的主机信道适配器实现可以用于与所述一个或多个客户端节点相关联的客户端端口，并依赖于SA请求数据包中的LID信息，以便将对特定的客户端QP编号的访问限制到所述指定的特定的客户端端口。

[0086] 图8是根据本发明的实施例的用于支持网络环境中的子网管理的子网管理器800的框图。子网管理器800可以包括检索单元810、设置单元820以及激活单元830。检索单元810可以检索用于在子网管理员(SA)和客户端节点之间建立可靠连接(RC)的信息。设置单元820可以为与SM节点相关联的端口设置一个或多个连接状态，以在与SM节点相关联的端口和与客户端节点相关联的端口之间建立RC连接。激活单元830可以激活与所述客户端节点相关联的端口。

[0087] 根据本发明的实施例，与客户端节点相关联的子网管理代理(SMA)可以操作以通告表示客户端节点具有对于基于RC的SA访问的支持的能力。

[0088] 根据本发明的实施例，客户端节点可以操作来为与客户端节点相关联的端口，建立第一组一个或多个RC队列对(QP)。第一组一个或多个RC QP可以与为基于RC的SA访问分配的一个或多个队列对(QP)编号相关联。

[0089] 根据本发明的实施例，子网管理器800还可以包括建立单元(未示出)。建立单元可以为与SM节点相关联的端口，建立第二组一个或多个RC队列对(QP)。第二组一个或多个RC QP的QP编号可以基于被分配到与客户端节点相关联的端口的基本本地标识符(LID)的函数，以及由SM为与客户端节点相关联的SMA设置的SMA属性中的一项。

[0090] 根据本发明的实施例，可以使用管理消息交换来建立RC连接。

[0091] 根据本发明的实施例，客户端节点可以操作以使用RC连接，向SA发送消息。

[0092] 根据本发明的实施例，可以使用多个RC连接来有效利用消息缓冲区存储器。

[0093] 根据本发明的实施例，可以在客户端节点中建立预先定义的存储器区域，以包含通过RC连接从SA和SA代理中的至少一项传输的状态信息。

[0094] 根据本发明的实施例，子网管理器800还可以包括发送单元(未示出)和分配单元(未示出)。发送单元可以发送一个或多个多播消息，以通知至少一个分区中的多个客户端节点关于一个或多个事件。分配单元可以给每一多播消息分配序列号。

[0095] 参考图9，示出了根据本发明的一实施例的系统900。图9示出了由系统900实现的功能配置的图示。系统900包括检索模块910、存储器模块920、设置模块930、与子网管理器(SM)节点相关联的端口940、与客户端节点相关联的端口950，以及激活模块。

[0096] 检索模块910检索用于在子网管理员(SA)和客户端节点之间建立可靠连接(RC)的信息。信息存储在存储器模块920中。设置模块930为与SM节点相关联的端口940设置一个或多个连接状态，以在与SM节点相关联的端口940和与客户端节点相关联的端口950之间建立RC连接。激活模块960激活与客户端节点相关联的端口950。

[0097] 图10示出了包括已知的硬件元件的计算机系统1000的图示。即，计算机系统1000包括中央处理单元(CPU)1010、鼠标1020、键盘1030、随机存取存储器(RAM)1040、硬盘1050、盘驱动器1060、通信接口(I/F)1070，以及监视器1080。计算机系统1000可以充当构成系统900的节点。

[0098] 根据本发明的实施例，检索模块910、设置模块930、与SM节点相关联的端口940、与客户端节点950相关联的端口950，以及激活模块960由一个或多个计算机系统900提供。检索模块910、设置模块930、与SM节点相关联的端口940、与客户端节点950相关联的端口950，以及激活模块960由CPU 1010实现。另一方面，可以使用一个以上的处理器，以便实现检索模块910、设置模块930、与SM节点相关联的端口940、与客户端节点950相关联的端口950，以及激活模块960。即，检索模块910、存储器模块920、设置模块930、与SM节点相关联的端口940、与客户端节点950相关联的端口950，以及激活模块960可以彼此物理地分离。

[0099] 在又一方面，系统900可以通过使用充当检索模块910、设置模块930、与SM节点相关联的端口940、与客户端节点950相关联的端口950，以及激活模块960的多个硬连线电路来实现。

[0100] 本发明可以使用一个或多个常规通用或专门数字计算机、计算设备、机器，或微处

理器，包括一个或多个处理器、存储器和/或根据本发明的原理编程的计算机可读存储介质，来方便地实现。对于那些精通软件技术的人来说显而易见的是，可以由熟练的程序员基于本发明的原理轻松地编制适当的软件代码。

[0101] 在某些实施例中，本发明包括计算机程序产品，该产品是其中存储了指令的存储介质或计算机可读介质，这些指令可以用来对计算机进行编程，以执行本发明的任何一个进程。存储介质可以包括，但不仅限于，任何类型的盘，包括软盘、光盘、DVD、CD-ROM、微驱动，以及磁光盘、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、FLASH存储器设备、磁卡或光卡，纳米系统(包括分子存储器IC)，适于存储指令和/或数据的任何类型的介质或设备。

[0102] 前面的对本发明的描述只是为了说明和描述。它不是详尽的说明或将本发明限于所公开的准确的形式。那些精通本技术的专业人员将认识到，可以进行许多修改。修改以及变化包括所描述的特征的任何相应的组合。所选择和描述的实施例只是为了最好地说明本发明的原理以及其实际应用，并使精通本技术的其他人懂得，带有适合于特定用途的各种修改的各实施例的本发明也是可以接受的。本发明的范围由下面的权利要求以及它们的等效内容进行定义。

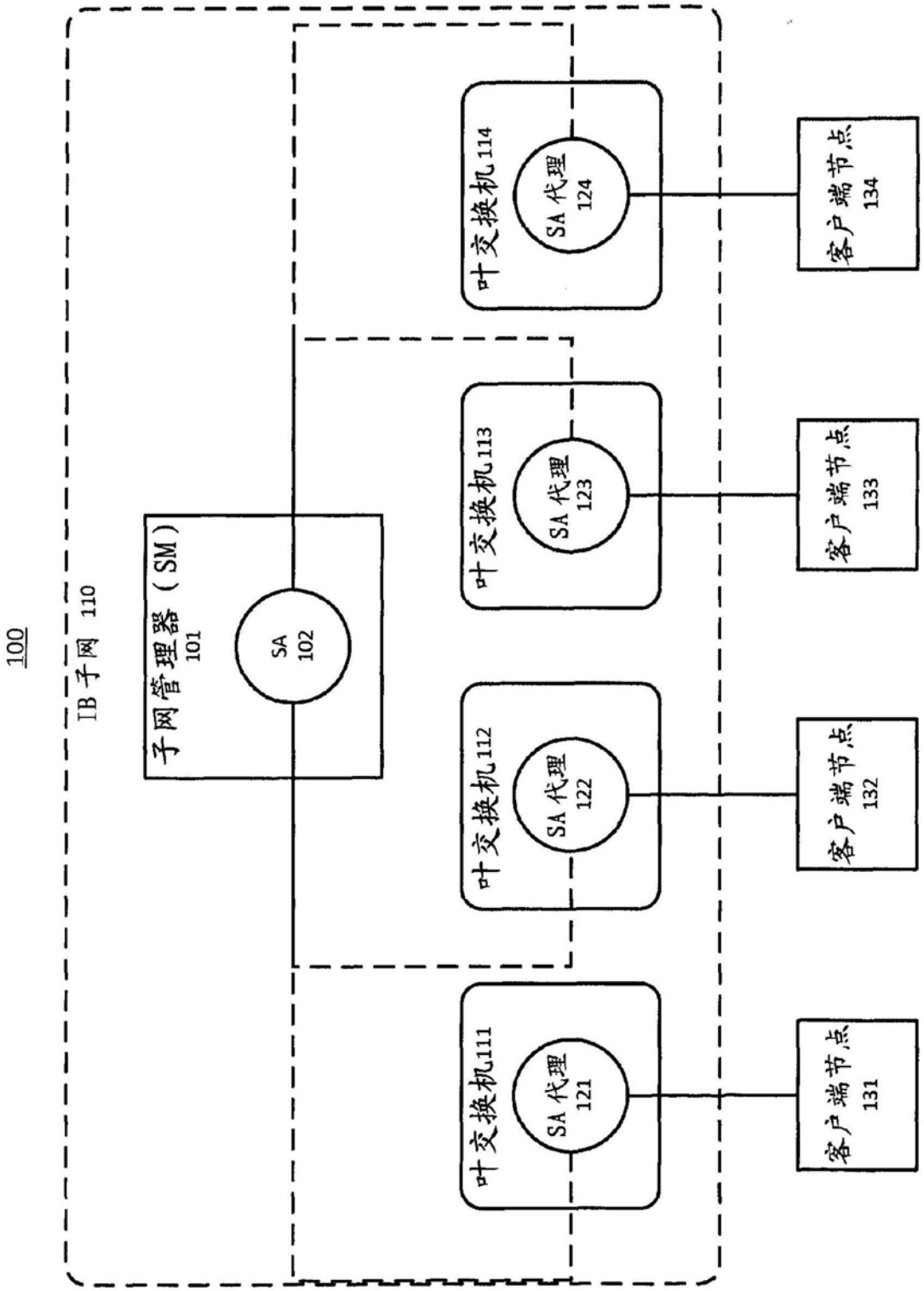

100
IB子网 110
子网管理器（SM）
101
SA
102
叶交换机111
SA 代理
121
叶交换机112
SA 代理
122
叶交换机113
SA 代理
123
叶交换机114
SA 代理
124
客户端节点
131
客户端节点
132
客户端节点
133
客户端节点
134

图1

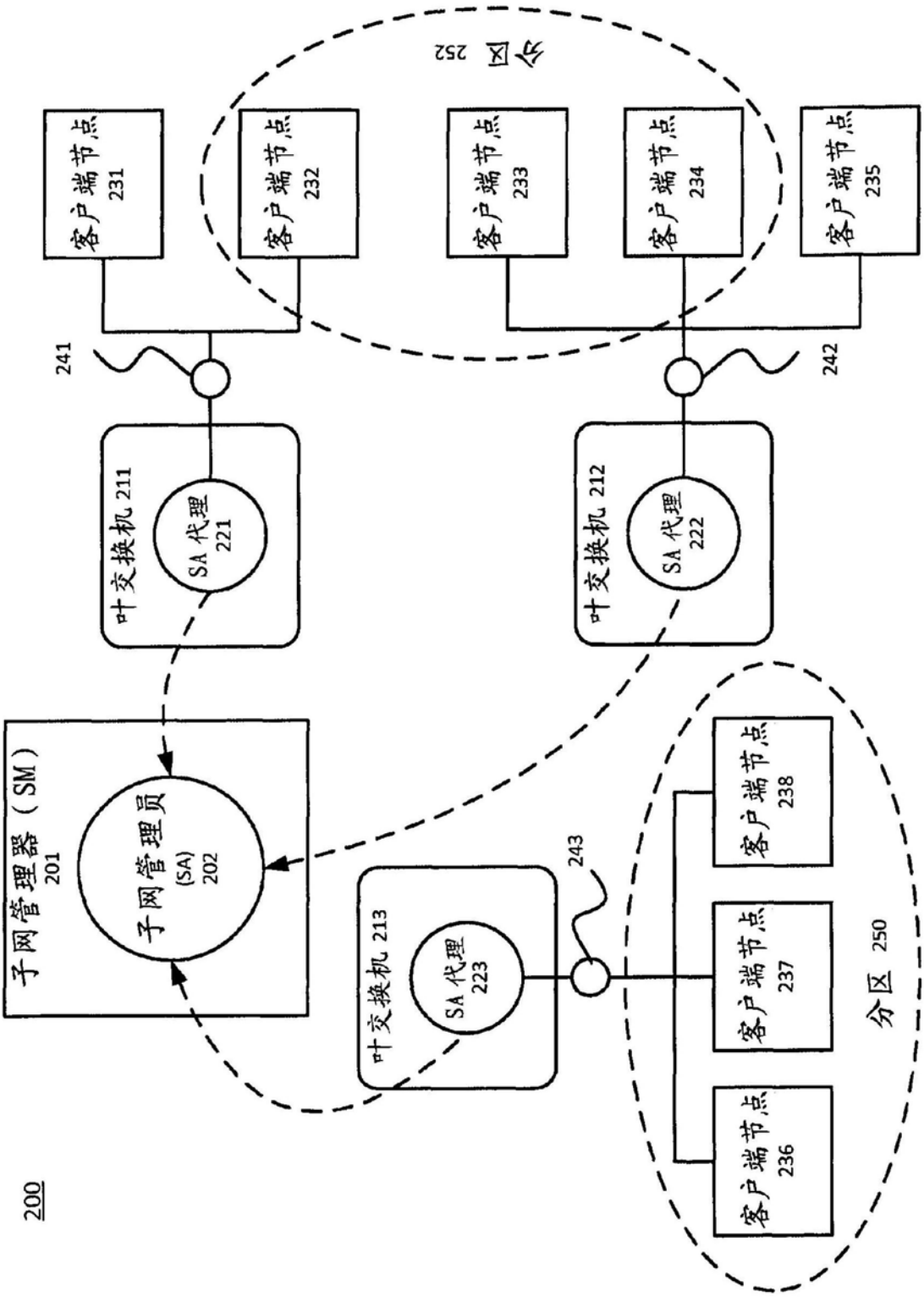

200
子网管理器（SM）
201
子网管理员
(SA)
202
叶交换机 211
SA 代理
221
叶交换机 212
SA 代理
222
叶交换机 213
SA 代理
223
241
242
243
客户端节点
231
客户端节点
232
客户端节点
233
客户端节点
234
客户端节点
235
客户端节点
236
客户端节点
237
客户端节点
238
分区 250
252

图2

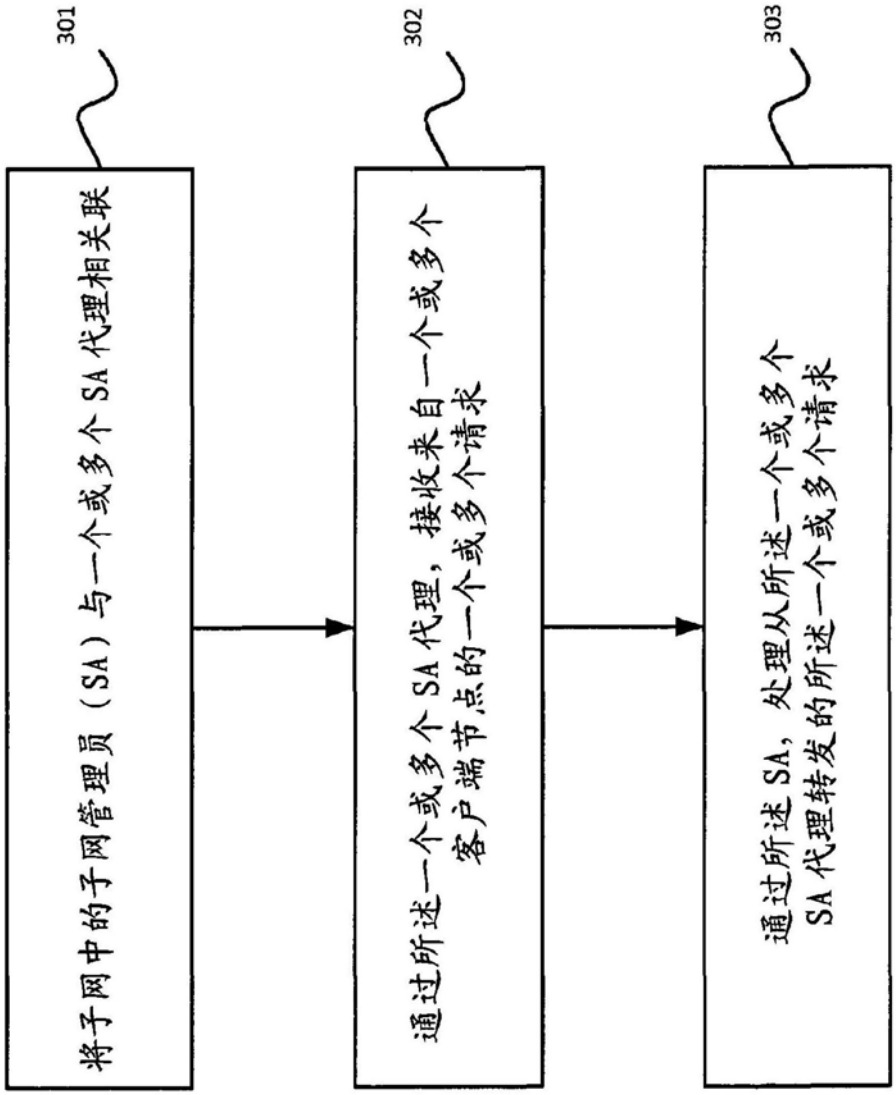

301
将子网中的子网管理员（SA）与一个或多个 SA 代理相关联
302
通过所述一个或多个 SA 代理，接收来自一个或多个客户端节点的一个或多个请求
303
通过所述 SA，处理从所述一个或多个 SA 代理转发的所述一个或多个请求

图3

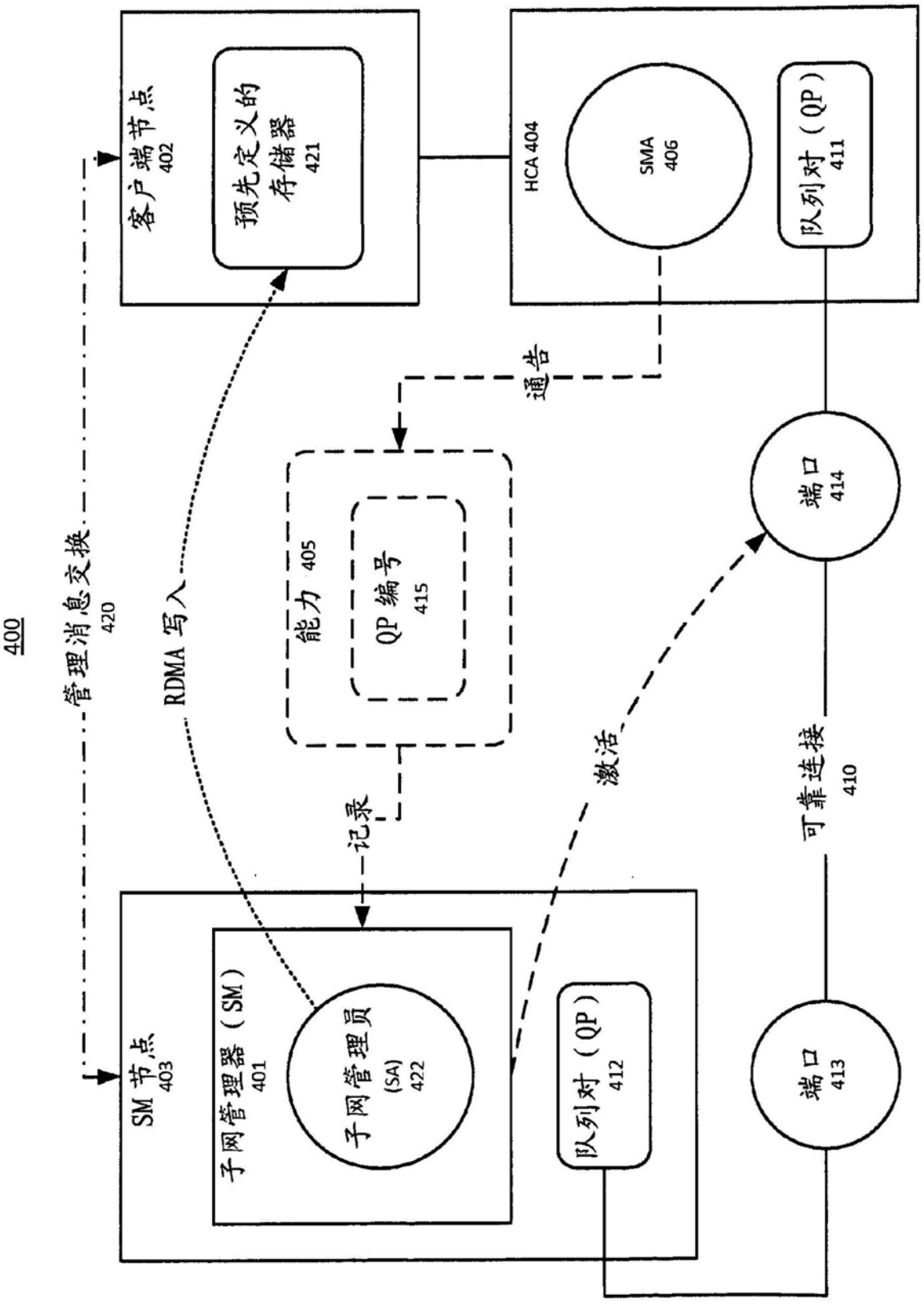

400
管理消息交换
420
客户端节点
402
预先定义的
存储器
421
HCA 404
SMA
406
队列对（QP）
411
RDMA 写入
通告
能力 405
QP 编号
415
端口
414
记录
激活
可靠连接
410
SM 节点
403
子网管理器（SM）
401
子网管理员
(SA)
422
队列对（QP）
412
端口
413

图4

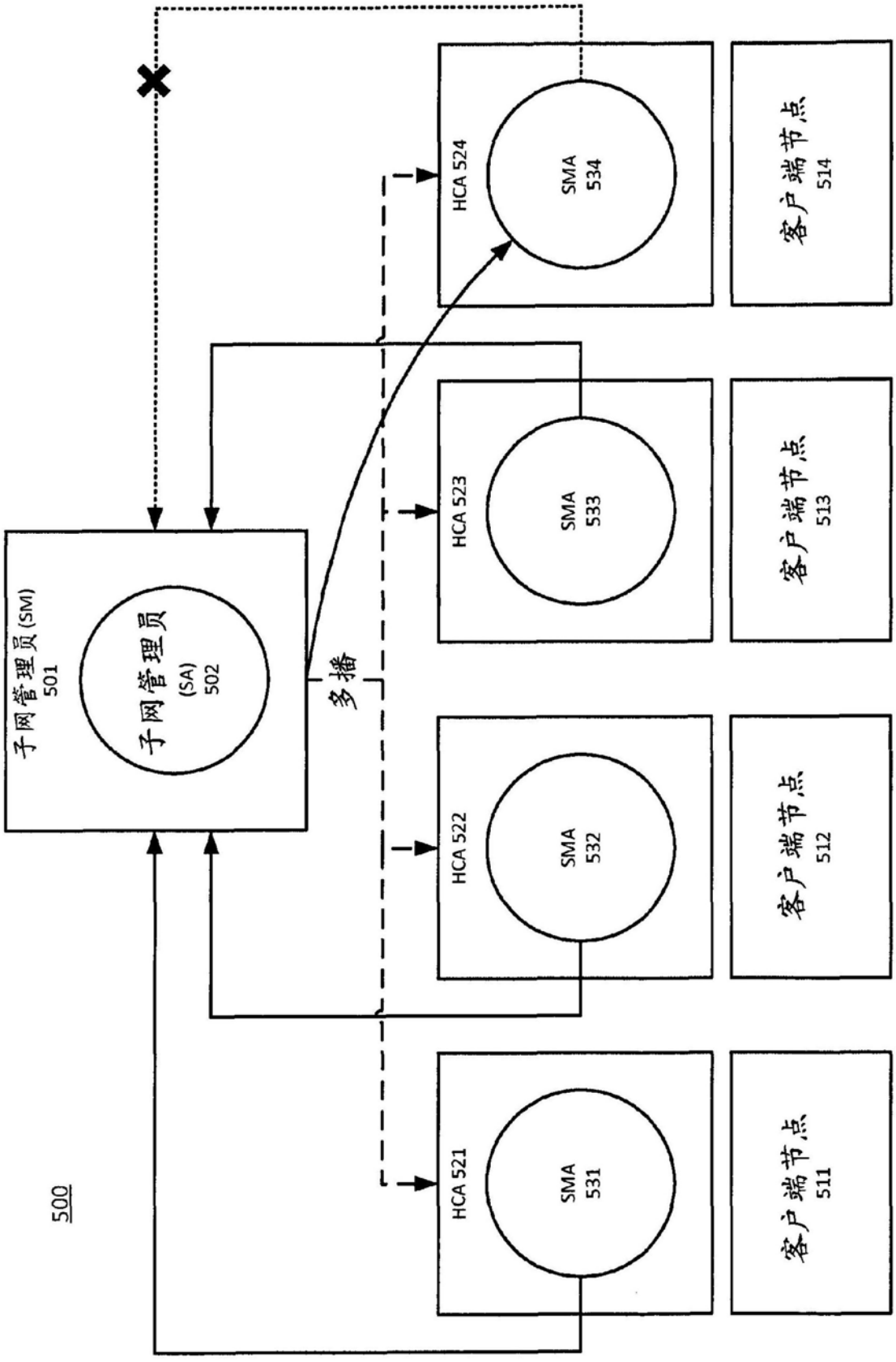

500
子网管理员(SM)
501
子网管理员
(SA)
502
多播
HCA 521
SMA
531
客户端节点
511
HCA 522
SMA
532
客户端节点
512
HCA 523
SMA
533
客户端节点
513
HCA 524
SMA
534
客户端节点
514

图5

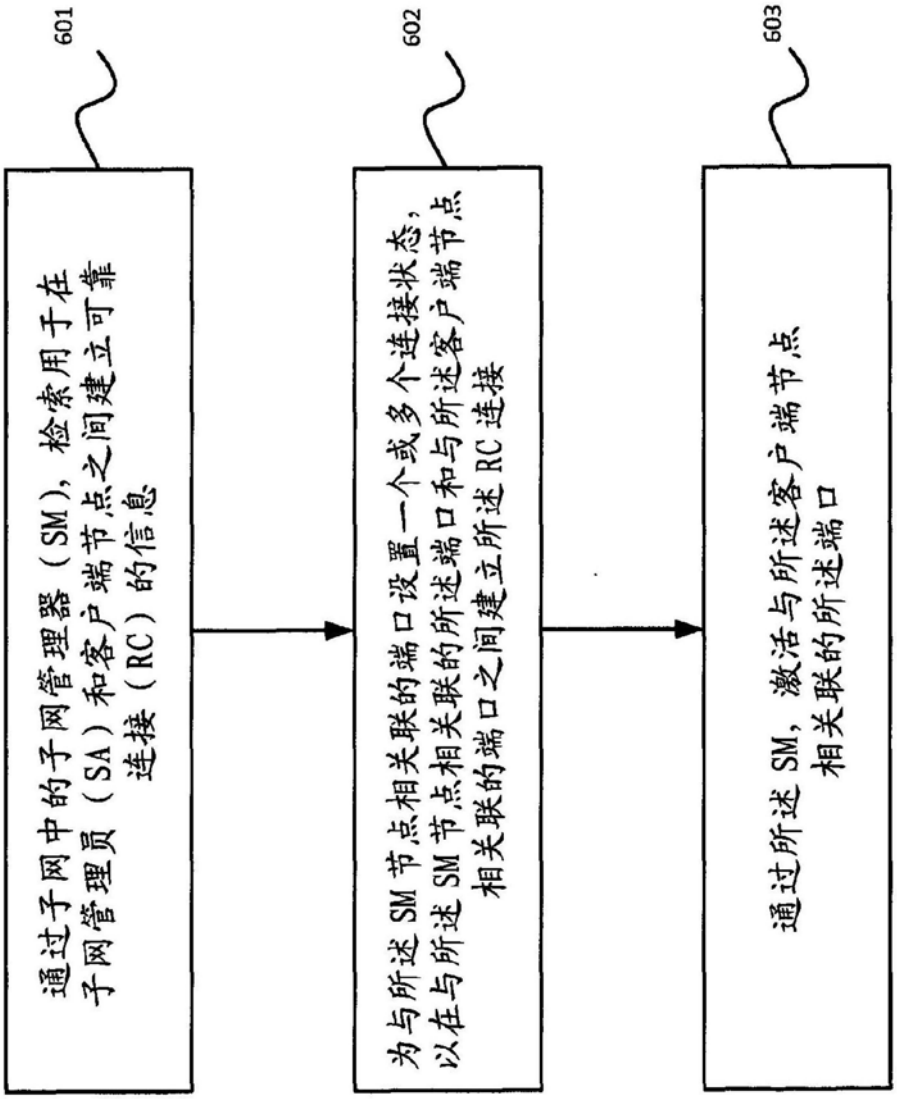

601
通过子网中的子网管理器（SM），检索用于在子网管理员（SA）和客户端节点之间建立可靠连接（RC）的信息
602
为与所述 SM 节点相关联的端口设置一个或多个连接状态，以在与所述 SM 节点相关联的所述端口和与所述客户端节点相关联的端口之间建立所述 RC 连接
603
通过所述 SM，激活与所述客户端节点相关联的所述端口

图6

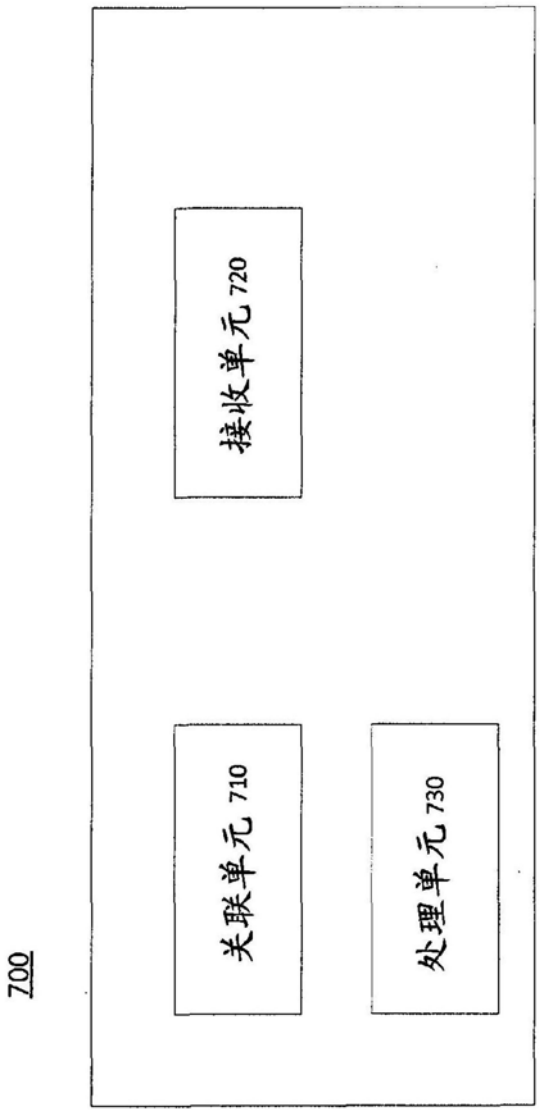


图7

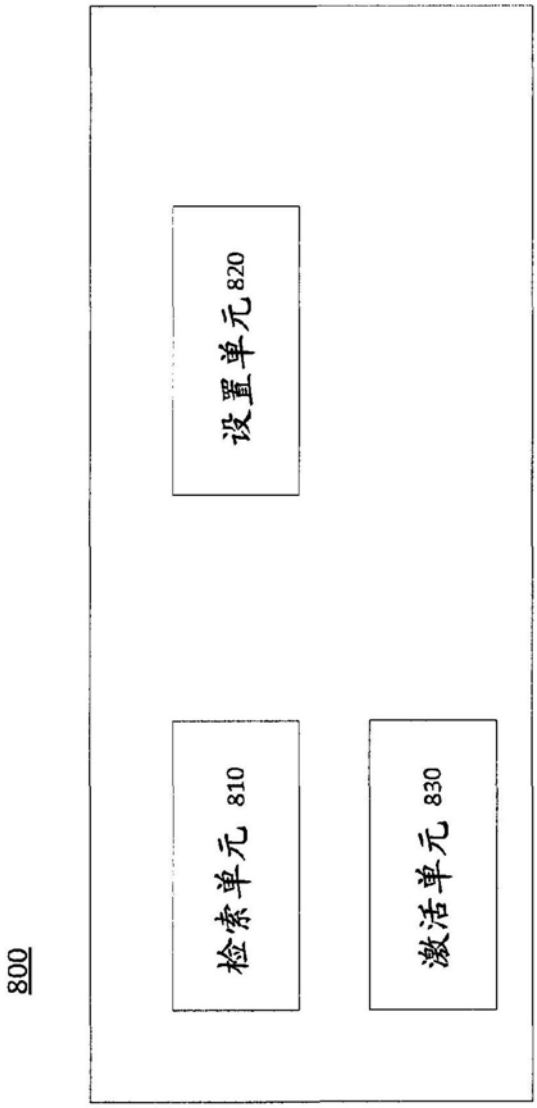


图8

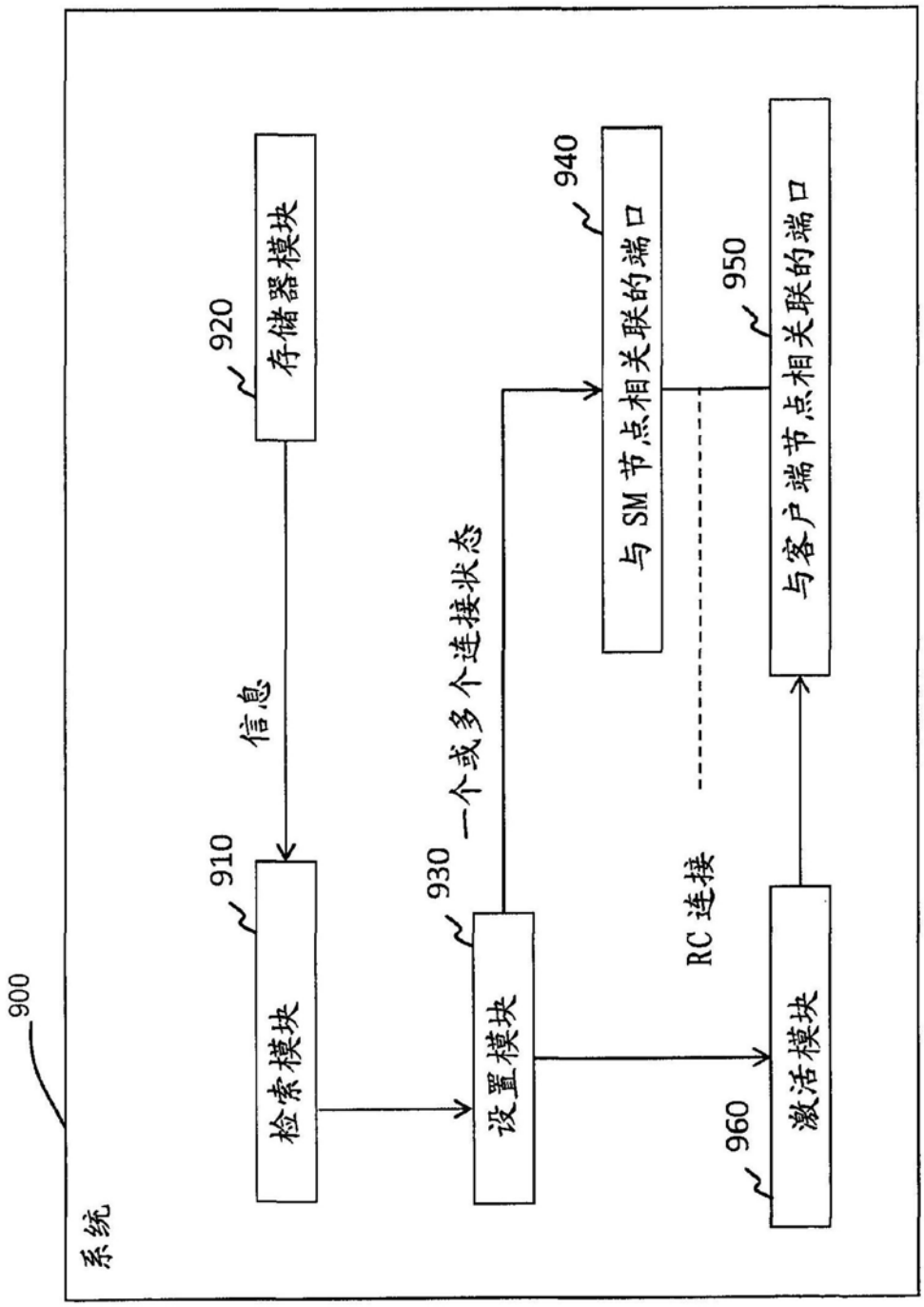

900
系统
检索模块
910
信息
存储器模块
920
设置模块
930
一个或多个连接状态
与 SM 节点相关联的端口
940
RC 连接
与客户端节点相关联的端口
950
激活模块
960

图9

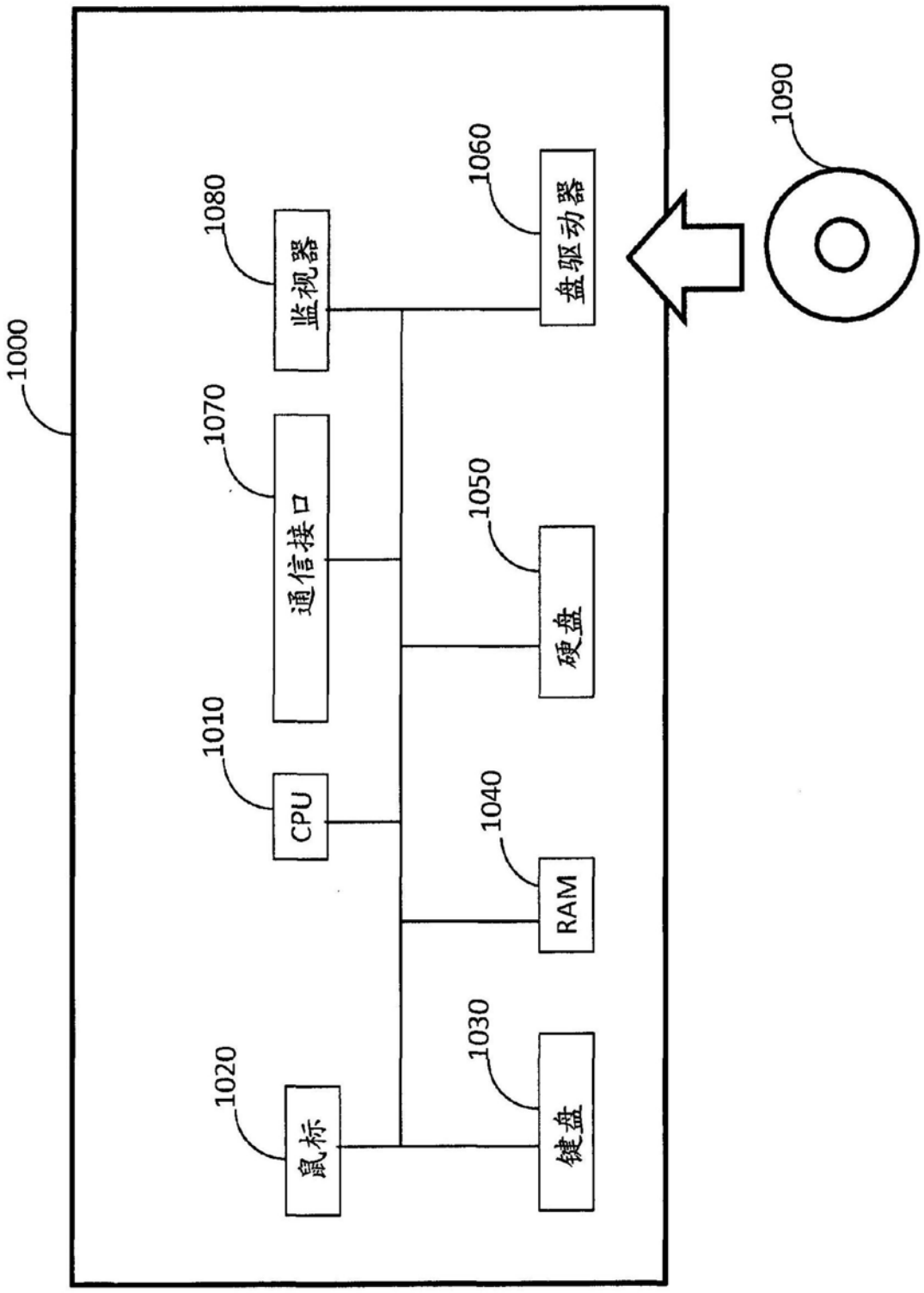

1000
1090
1060
盘驱动器
1080
监视器
1070
通信接口
1050
硬盘
1010
CPU
1040
RAM
1020
鼠标
1030
键盘

图10