(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 101072125 B
(45)授权公告日 2010.09.22

(21)申请号 200710091597.5

(22)申请日 2007.03.29

(30)优先权数据
2006-130037 2006.05.09 JP

(73)专利权人 株式会社日立制作所
地址 日本东京都

(72)发明人 关口知纪 天野光司 大平崇博

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 许静

(51)Int.Cl.
H04L 12/24(2006.01)
H04L 12/46(2006.01)

(56)对比文件
CN 1480863 A,2004.03.10, 全文.
JP 特开2005-260134 A,2005.09.22, 全文.
JP 特开2001-344125 A,2001.12.14, 全文.
JP 特开2001-109642 A,2001.04.20, 全文.
WO 0250678 A1,2002.06.27, 全文.

审查员 采健

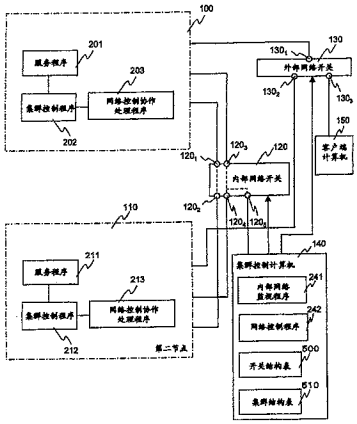
权利要求书 1 页 说明书 8 页 附图 6 页

(54)发明名称

集群结构及其控制单元

(57)摘要

在由两台计算机节点构成的不具有共享存储装置的集群中,存在如下课题:通过网络来监视相互的正常状态和停止状态,但仅通过这些有时会错误地判断对方节点已停止。当根据错误的判断执行了系切换时,在系切换后,对方节点恢复正常状态,两台计算机都作为执行系进行动作。构成集群的两台节点和与集群进行通信的其他计算机通过可以使各计算机所连接的端口无效的开关进行连接。控制这些开关的网络控制程序与节点的系切换同步地对节点所连接的端口是否可以使用进行变更。



CN 101072125 B

1. 一种集群系统，其特征在于，

由以下各部分构成：

形成节点的两个计算机；

内部网络开关，用于两个所述计算机相互交换信息，分别监视另一计算机的正常和停止；

外部网络开关，用于将两个所述计算机和访问两个所述计算机来接收服务的客户端计算机连接；和

集群控制计算机，与所述内部网络开关连接，将两个所述计算机中的一个计算机作为对来自所述客户端计算机的请求进行处理的主系，将另一个计算机作为为了接替主系的处理而进行待机的从系，来控制运转模式，

所述内部网络开关以及外部网络开关和各个所述计算机的连接是通过从外部可以控制连接的无效、有效的端口来进行连接，而且，

所述两个计算机通过经由内部网络开关的信息交换判断是否需要进行运转模式迁移，并且所述集群控制计算机接收所述运转模式迁移的通知，将节点所连接的内部网络开关以及外部网络开关的端口变更为有效或无效。

2. 根据权利要求 1 所述的集群系统，其特征在于，

在将所述节点的计算机的运转模式从待机状态迁移到运转状态时，所述集群控制计算机使连接之前处于运转状态的另一节点的计算机的所述内部网络开关的端口和为了向所述客户端计算机提供服务连接了所述另一节点的计算机的所述外部网络开关的端口无效。

3. 根据权利要求 1 所述的集群系统，其特征在于，

在将所述节点的计算机的运转模式从停止状态迁移到开始状态时，所述集群控制计算机使连接该计算机的所述内部网络开关的端口和为了向所述客户端计算机提供服务连接了所述另一节点的计算机的所述外部网络开关的端口有效。

4. 根据权利要求 1 所述的集群系统，其特征在于，

在将所述节点的计算机的运转模式迁移到停止状态时，所述集群控制计算机使连接该计算机的所述内部网络开关的端口和为了向所述客户端计算机提供服务连接了所述另一节点的计算机的所述外部网络开关的端口无效。

5. 根据权利要求 1 所述的集群系统，其特征在于，

所述集群控制计算机是收集与所述内部网络开关的端口的有效化、无效化有关的数据的计算机，而且，参照该数据判断是否需要对与所述内部网络开关连接的计算机进行运转模式迁移，并且，所述集群控制计算机接收所述运转模式迁移的通知，将节点所连接的内部网络开关以及外部网络开关的端口变更为有效或无效。

集群结构及其控制单元

技术领域

[0001] 本发明涉及一种用于由两台计算机构成的集群（cluster）系统的高可用化的结构以及控制单元。尤其涉及不具有在两台计算机之间共享的外部存储装置的结构的集群系统的高可用化方式。

背景技术

[0002] 作为提高在计算机系统中执行的处理的可用性的方式，具有被称为集群的思想。在集群系统中，在多个计算机中安装同一程序，将其中的几个计算机作为执行实际处理的计算机。剩下的计算机被控制成在检测到执行处理的计算机发生了故障时取代该计算机来执行处理。

[0003] 一般的集群系统由两台计算机构成。一方是进行实际处理的计算机（执行系），剩下的是待机的计算机（待机系），为执行系的异常而准备来接替进行执行系的处理。两台计算机通过网络进行通信，由此定期地监视相互的执行状况。另外，一般在从待机系向执行系进行系切换时，为了使待机系接替数据的处理，配置由两台计算机都可以访问的共享的外部存储装置。该共享存储装置在排他控制下使用，使得只能由当前的执行系访问。作为实现该访问的访问方法一般使用 SCSI 协议。

[0004] 在这样的集群中，当待机系检测到执行系的异常时，待机系就将自己切换为执行系。此时，待机系在争夺到共享存储装置的访问权后开始执行应用程序。应用程序参照存储在共享存储装置中的数据执行用于接替的处理，并开始实际的处理。

[0005] 这样的集群由用于集群控制的软件和与其协作执行的应用程序构成。另外，作为与集群控制软件进行协作的软件的例子，可以列举出数据库。

[0006] 另一方面，在集群系统中，有时直至待机系作为执行系开始执行为止的时间成为问题。在上述的集群系统中，在争夺共享存储装置的访问权的处理和成为执行系的计算机一侧的接替处理的期间，系统无法对外提供服务。特别是共享存储装置的访问权控制，一般需要花费十几秒。

[0007] 在无法允许十几秒的服务中断的系统中，例如构成一种所谓的作为并行集群而公知的不配置共享存储装置的集群系统。作为这样的例子，具有专利文献 1（特开 2001-109642）。在这里，在执行系对请求进行处理，将其结果发送给待机系，使执行系和待机系的处理状况一致。另外，如专利文献 2（特开 2001-344125）那样，使执行系、待机系之间的协作双重化来提高系切换处理的可靠性。并且，如专利文献 3（特开平 05-260134）那样，使监视装置层次化，进行针对监视装置的异常的处理来提高系统的可靠性。

[0008] 另外，还有执行系、待机系双方的计算机接受处理请求来进行处理的情况。作为执行系的计算机输出处理结果，待机系将处理结果保存在内部，为切换到执行系时而准备。双方计算机还可以一边互相通信来使处理的进展同步，一边进行请求的处理。

[0009] 通过这些方式，在系切换中不需要共享存储装置的访问权的交接，待机系可以作为执行系立即开始执行。如此，控制待机系使其具有与执行系相同的状态，并且始终准备系

切换，由此可以缩短从待机系向执行系的切换时间，可以缩短服务中断时间。

[0010] 在集群系统中，重点是两台计算机正确掌握相互的状态。具有共享存储装置的结构的集群使用基于网络的通信和共享存储装置的访问权控制这两个不同的公共媒体来确认对方的状态。在一方的并行集群中，相互或者通过经由第三者的网络通信来掌握两台计算机的状态。

发明内容

[0011] 在并行集群中，用于使执行系和待机系的两台计算机协作的公共媒体只有相互的基于网络的通信。在基于网络通信的状态监视下，根据无法通信这一状况来判断对方系已停止。

[0012] 但是，仅仅通过基于网络通信的状态监视，在构成集群的计算机中无法区别是由于对方系的不正常而通信中断、还是由于自身系的网络处理或网络设备的不正常而通信中断、还是由于网络自身的不正常而通信中断。因此存在以下的问题：对方系实际上没有停止，但是由于通信中断，一方的计算机误判断为对方系已停止。

[0013] 而且，在由于某种原因通信暂时中断的期间，当待机系由于误判断执行系切换时，存在系切换后对方系恢复正常状态，两台计算机都作为执行系进行动作的可能性。此时，存在集群系统可能会使外部系统发生混乱的问题。

[0014] 作为解决该问题的手段之一，具有如下的方法：要求被判断为已停止的计算机停止、或者发送复位信号等来强制停止计算机。前者的方法是对被认为已停止的计算机发送指示，因此不知道是否能正常接收，存在可靠性欠缺的问题。后者的方法使计算机复位，因此该计算机的故障信息消失，存在故障原因分析困难的问题。

[0015] 通过一台以上的网络开关连接构成并行集群（第一节点、第二节点）的两台计算机以及与各集群的计算机进行通信的其他计算机（例如，客户端计算机），上述网络开关使连接各计算机的端口独立，来控制其有效、无效。在这些网络开关上连接集群控制计算机，由他执行的网络控制程序执行所述网络开关的控制，以便在构成第一节点的计算机以及构成第二节点的计算机所执行的集群控制程序将待机系切换为执行系之前，使原来的执行系的计算机连接的端口无效化。由此，将原来作为执行系的计算机从网络切断。

[0016] 另一方面，构成集群的各节点的计算机所执行的集群控制程序与集群控制计算机所执行的网络控制程序协作，在通过所述网络开关开始进行系切换之前，向集群控制计算机所执行的网络控制程序请求切断执行系。

[0017] 为了集群控制计算机所执行的网络控制程序恰当地执行与集群的节点的状态相符合的控制，构成集群的节点的计算机所执行的集群控制程序向集群控制计算机所执行的网络控制程序通知节点的启动、执行系／待机系的迁移、节点的停止等事件。

[0018] 根据本发明，是一种由两台计算机构成的集群，在为了集群控制没有在计算机之间共享的存储装置的集群结构的情况下，可以防止错误识别对方系的状态来执行系切换，防止双方计算机都作为执行系进行动作。

[0019] 另外，从构成集群的计算机的外部监视计算机之间的相互监视的状况，从集群中隔离被判断为通信中断一侧的计算机，由此可以防止两系都作为执行系进行动作，并且可以可靠地进行执行系的切换。

[0020] 另外，因为可以不强制停止不正常的计算机，因此可以防止删除该计算机的故障分析所需的数据。

附图说明

[0021] 图 1 是表示本发明实施例 1 的系统结构的框图。

[0022] 图 2 是实施例 1 的执行用于实现集群控制的步骤的程序的结构框图。

[0023] 图 3 是表示本发明实施例 1 的集群的系切换步骤的前半部分的处理流程。

[0024] 图 4 是表示本发明实施例 1 的集群的系切换步骤的后半部分的处理流程。

[0025] 图 5(a)、(b) 表示本发明实施例中的集群控制计算机所保存的数据结构的例子。

[0026] 图 6 是表示本发明实施例 2 的内部网络的监视步骤的处理流程。

具体实施方式

[0027] 以下，参照附图对本发明的实施方式进行说明。

[0028] (实施例 1)

[0029] 图 1 是表示本发明实施例 1 的系统结构的框图。本发明的集群由以下部分构成：构成集群的第一节点的计算机 100 和第二节点的计算机 110；形成集群相互的通信网络的内部网络开关 120；对各个集群进行访问的客户端计算机；形成各个集群和客户端计算机相互的通信网络的外部网络开关 130；以及接收来自各个集群的信息，执行控制所述各个网络开关的端口的有效和无效的程序的集群控制计算机 140。

[0030] 第一节点的计算机 100 以及第二节点的计算机 110 是普通的计算机，分别具有：CPU104、114 以及存储器 105、115；控制它们向总线 106、116 连接的总线控制装置 107、117；以及经由盘适配器 108、118 向总线 106、116 连接的存储装置 109、119。这些计算机具有：用于连接总线 106、116 和外部网络开关 130 的外部网络适配器 101、111；用于控制各节点的计算机 100、110 的执行系・待机系的切换，连接各节点的计算机 100、110 和内部网络开关 120 的控制网络适配器 102、112；以及用于进行各节点的计算机的执行系・待机系的评价，并且连接各节点的计算机 100、110 和内部网络开关 120 的内部网络适配器 103、113。

[0031] 外部网络适配器 101、111 通过端口 130_1、130_2 连接到外部网络开关 130。另外，客户端计算机 150 通过端口 130_3 连接到外部网络开关 130。如果第一节点的计算机 100 为执行系，则只有端口 130_1、130_3 被有效化，第一节点的计算机 100 和客户端计算机 150 相连接。如果第二节点的计算机 110 为执行系，则只有 130_2、130_3 被有效化，第二节点的计算机 110 和客户端计算机 150 相连接。

[0032] 另外，内部网络适配器 103、113 通过端口 120_1、120_2 连接到内部网络开关 120，互相传递有关自身节点的计算机 100、110 的状态的信息。

[0033] 控制网络适配器 102、112 通过端口 120_3、120_4 连接到内部网络开关 120。另外，集群控制计算机 140 通过端口 120_5 连接到内部网络开关 120。控制网络适配器 102、112 互相交换经由所述内部网络适配器 103、113 得到的有关其他节点的计算机 110、100 的状态的信息以及与自身节点的计算机 100、110 的状态相对应的控制信号，并且还与集群控制计算机 140 交换控制信号。集群控制计算机 140 以收集到的信息为基础，向内部网络开关 120 以及外部网络开关 130 的各端口发送有效化或无效化的信号。

[0034] 为了第一节点的计算机 100 的内部网络适配器 103 和第二节点的计算机 110 的内部网络适配器 113 经由内部网络开关 120 互相进行通信而构成的网络以及为了第一节点的计算机 100、第二节点的计算机 110、集群控制计算机 140 经由内部网络开关 120 进行有关集群控制的通信而构成的网络通过内部网络开关 120 的设定来实现。

[0035] 图 2 是实施例 1 的执行用于实现集群控制的步骤的程序的结构框图。各节点的计算机 100、110 的各程序被存储在执行各程序的计算机的存储装置 108、118 中，执行时载入存储器 105、115，然后由 CPU104、114 执行程序，这简明地表现了简单的程序执行。关于集群控制计算机 140，没有图示存储装置、存储器、CPU 以及与内部网络适配器 103、113、外部网络适配器 101、111 对应的适配器，但不言而喻，与各节点的计算机 100、110 相同，具有存储装置、存储器、CPU 以及适配器。另外，有关所保存的程序的执行也相同。

[0036] 构成集群的各节点的计算机 100、110 具备并执行：向集群的外部，即向客户端计算机 150 提供实际服务的服务程序 201、211；执行集群结构的控制的集群控制程序 202、212；向集群控制计算机 140 联络节点的执行状态的变更的网络控制协作程序 203、213。

[0037] 集群控制计算机 140 具备并执行：对内部网络开关 120 的各集群的连接端口的有效、无效的网络状况进行监视的内部网络监视程序 241；以及对外部网络开关 130 的各集群的连接端口的有效、无效的设定进行变更的网络控制程序 242。另外，具有对这些程序参照的设定数据进行保存的开关结构表 500 以及集群结构表 510。对这些将在后面进行叙述。

[0038] 下面，对实施例 1 的各程序的动作进行说明。

[0039] 各节点的集群控制程序 202、212 是管理各节点的运转模式的程序。集群控制程序 202、212 经由内部网络开关 120 互相监视对方节点的执行状态。例如，由第一节点的计算机 100 执行的集群控制程序 202 和第二节点的计算机 110 执行的集群控制程序 212 经由连接控制网络适配器 102 的内部网络开关 120 的端口 120_3、以及连接控制网络适配器 112 的端口 120_4，互相在一定周期持续发送消息。各个集群控制程序 202、212 确认在一定的周期持续接收到来自对方节点的消息。通过该相互通信，各节点的计算机 100、110 互相监视执行状态。

[0040] 各节点的计算机的运转模式为：没有执行集群控制程序 202、212 的停止状态、正在执行集群控制程序 202、212 但没有执行服务程序 201、211 的开始状态、服务程序 201、211 正在提供服务的执行状态、正在执行服务程序 201、211 但没有输出处理结果的待机状态中的某一种模式。

[0041] 对各节点的计算机的运转模式的迁移进行说明。当启动节点的计算机时，运转模式从停止状态迁移到开始状态。从开始状态向执行状态、或者向待机状态的迁移通常是根据集群的操作员的指示来执行。在自身节点的计算机处于待机状态时，对方节点的计算机成为待机状态或者处于执行状态的对方节点的计算机的运转状态不明的情况下，集群控制程序 202、212 使自身节点的计算机的运转模式从待机状态迁移到运转状态。在根据操作员的指示对执行状态的节点和待机状态的节点进行转换时，使执行状态的节点迁移到待机状态。由此，执行处于待机状态的对方节点的集群控制程序，来对处于执行状态的节点迁移到待机状态的情况进行检测。

[0042] 服务程序 201、211 与集群控制程序 202、212 进行协作，来处理经由连接外部网络适配器 101、111 的外部网络开关 130 的端口 130_1、130_2 以及连接客户端计算机 150 的端口

1303，从客户端计算机 150 发送来的服务请求。集群控制程序 202、212 和服务程序 201、211 的协作包括：取得正在执行服务程序 201、211 的节点的计算机 100、110 的执行状态。

[0043] 在第一节点的计算机 100 的运转模式为执行状态时，服务程序 201 输出请求的处理结果。此时，在处于待机状态的第二节点的计算机 110 中，服务程序 211 不把处理结果输出到外部，而是记录在计算机 110 的内部，例如记录在盘 119 中。记录的数据内容是在第二节点的计算机 110 变成执行状态时，服务程序 211 作为执行状态，足够作为服务请求处理的处理结果输出的数据。另外，执行系和待机系的服务程序之间也可以进行协作，使请求处理的进展同步。

[0044] 图 3 是表示本发明实施例 1 的集群的系切换步骤的前半部分的处理。参照该图，以第一节点的计算机 100 的动作为主对运转模式的迁移进行说明。

[0045] 在第一节点的计算机 100 中，集群控制程序 202 的监视处理准备接收来自第二节点的计算机 110 的一定周期的消息而进行待机（步骤 301）。在一定时间消息没有到达与内部网络开关 120 的端口 120_1 连接的内部网络适配器 103 时，该接收处理失败。在内部网络适配器 103 正常接收到消息时（步骤 302 的判断为 Yes），反复执行消息待机。在接收来自第二节点的计算机 110 的消息失败的情况下（步骤 302 的判断为 No），判断第二节点的计算机 110 是否停止（步骤 303）。该判断方法具有各种方法，一般在预先规定的期间消息的正常接收连续失败的情况下判断为第二节点的计算机 110 已停止。在无法判断为停止时，回到消息的接收处理（步骤 301）。

[0046] 当在步骤 303 中判断为第二节点的计算机 110 已停止时，判断是否需要状态迁移（系切换处理）（步骤 304）。在判断为需要状态迁移时，判断第一节点的计算机 100 的运转模式是否为待机状态（步骤 305）。在判断为 No，即第一节点的计算机 100 的运转模式为执行状态时，关于系切换不进行任何处理，但如果是待机状态时，则执行状态迁移开始处理（步骤 306）。此时，步骤 306 是启动系切换处理的处理。

[0047] 以上是并行集群的基本动作。下面，对用于实现本发明而追加的步骤进行说明。

[0048] 一般，由集群的节点的计算机 100、110 执行的集群控制程序 202、212 具有一种接口，该接口在开始进行节点的计算机的运转模式的变更时，可以加入与该节点的计算机所提供的服务相符合的处理。在本发明中，以此为前提。在本发明中，使用该接口加入网络控制协作程序 203、213。这些网络控制协作程序 203、213 在集群控制程序 202、212 启动时、停止时以及节点的计算机的运转模式迁移时执行。

[0049] 下面，对本发明的系切换处理进行说明。图 3 所示流程的状态迁移开始处理（步骤 306）是启动系切换处理的处理。系切换处理被状态迁移开始处理（步骤 306）触发，启动所加入的网络控制协作程序 203（步骤 311）。此时，将当前的运转模式和新设定的运转模式作为参数交给网络控制协作程序 203。系切换处理在网络控制协作程序 203 启动后，等待其结束（步骤 312）。步骤 312 的结束待机处理也可以根据预先定义的时间而暂停（time out）。

[0050] 网络控制协作程序 203 向由集群控制计算机 140 执行的网络控制程序 242 联络在第一节点的计算机 100 中已开始了运转模式迁移（步骤 321），等待网络控制程序 242 的处理（网络切断处理，即外部网络开关 130 的端口 1301 的无效化）的完成（步骤 322），在处理完成后结束。步骤 322 的待机处理也可以根据预先定义的时间而暂停。

[0051] 收到网络控制协作程序 203 的结束后，集群控制程序 202 的系切换处理执行节点的计算机的运转模式的变更处理（步骤 313）。

[0052] 集群控制程序 202 的启动处理和停止处理也同样包括启动网络控制协作程序 203 的处理。其与从图 3 的步骤 306 开始的处理为相同的处理。即，启动时是从停止向开始的迁移，停止时是从此时的模式向停止的迁移。对于这些处理流程，省略其说明。

[0053] 图 4 是表示本发明实施例 1 的集群的系切换步骤的后半部分的处理流程。参照该图，对与节点的计算机的运转模式的迁移进行协作，来变更集群的网络结构的集群控制计算机 140 的网络控制程序 242 的处理流程进行说明。在这里，也以第一节点的计算机 100 的动作为主进行说明。

[0054] 网络控制程序 242 等待来自集群的节点的计算机的运转模式迁移通知（步骤 401）。迁移通知经由连接第一节点的计算机 100 的控制网络适配器 102、第二节点的计算机 110 的控制网络适配器 112 的端口 120_3、120_4 被导入内部网络开关 120，并在步骤 313 中通过端口 120_5 传递给集群控制计算机 140。

[0055] 当接收到运转模式迁移通知时，根据得到的迁移内容对处理进行分支（步骤 402）。例如，在由所述对方节点的计算机异常引起的系切换处理中，将第二节点的计算机 110 判断为停止的第一节点的计算机 100 的集群控制程序 202 在第一节点的计算机 100 的运转模式为待机模式时，从待机模式变更为执行模式。网络控制程序 242 根据该迁移内容将处理移动到步骤 403。在步骤 403 中，将发送了对运转模式进行迁移的通知的第一节点的计算机 100 的对象的第二节点的计算机 110 从内部网络开关 120 和外部网络开关 130 切断。具体而言，网络控制程序 242 指示内部网络开关 120 和外部网络开关 130 使第二节点的计算机 110 的内部网络适配器 113 和外部网络适配器 111 所连接的端口 120_2 和 130_2 无效。

[0056] 在网络控制协作程序 203 的通知（步骤 401）为集群控制程序 202 的启动处理时，即作为集群节点的计算机从停止向开始的迁移的启动时，指示使运转模式迁移通知源的第一节点的计算机 100 连接的内部网络开关 120 的端口 120_1 和外部网络开关 130 的端口 130_1 有效（步骤 404）。相反，在停止集群节点的计算机时，即在停止集群控制程序 202 时，使这些端口无效（步骤 405）。在除此之外的迁移、执行→待机、执行·待机→开始的情况下，不进行任何处理（在图 4 的流程中没有记载）。

[0057] 在进行这些处理后，向通知的发送源发送网络结构变更的完成通知（步骤 406）。

[0058] 下面，关于集群控制计算机 140 保存的数据结构，参照图 5(a)、(b) 对实施例 1 的数据结构进行说明。该数据结构例如在集群控制计算机 140 内的设定文件中以集群控制计算机 140 所执行的程序可以解析的形式被进行记录，且这些程序可参照该数据结构。也可以在集群控制计算机 140 中具有生成这样的设定文件的步骤。

[0059] 图 5(a) 所示的 500 是开关结构表。该表 500 保存构成集群的网络的内部网络开关 120、外部网络开关 130 的信息。例如，存储控制用网络地址、控制程序的路径等。所述控制用网络地址是对内部网络开关 120、外部网络开关 130 的设定进行变更的请求的发送源，所述控制程序安装实际进行端口的有效化、无效化的控制或取得统计信息的处理。

[0060] 图 5(b) 所示的 510 是集群结构表。在该表 510 中保存集群的各节点的计算机与开关的哪一个端口连接。例如记录内部网络开关 120 和其端口号码、外部网络开关 130 和

其端口号码。

[0061] 网络控制程序 242 可以参照这些表 500、510 来变更集群的网络结构。

[0062] 集群控制计算机 140 还具有在表内存储上述设定内容的步骤。

[0063] 另外，在表 510 中也可以记录与有关过去取得的统计信息的记录有关的数据。关于这些，在实施例 2 中进行说明。

[0064] 由此，可以与集群的运转模式迁移进行协作，在系切换时对构成集群的网络结构进行变更。由此，可以从集群中断开通过相互监视判定为已停止的节点的计算机，可以切实隔断发生了故障的节点的计算机的影响。除此之外，即使在对方节点的计算机暂时停止的情况下，也可以切实地防止两个节点的计算机的运转模式都变成执行状态。

[0065] （实施例 2）

[0066] 在实施例 2 中，除了实施例 1 的控制，还执行以下的控制。由集群控制计算机 140 执行的网络控制程序 242 参照内部网络开关 120 的端口收发的统计信息，在判断为来自对方节点的计算机的通信中断时，通知集群控制程序 202、212，并请求系切换，上述内部网络开关 120 构成用于节点的计算机相互监视的网络。或者，网络控制程序 242 实施开关的控制，使判断为通信中断的对方节点的计算机所连接的端口无效。

[0067] 下面，具体说明本发明的实施例 2。在实施例 2 中，集群控制计算机 140 参照与内部网络开关 120 所收集的内部网络的通信状况有关的统计信息，变更集群的网络结构，由此实现对怀疑发生了故障的节点的计算机进行隔离的方式。

[0068] 一般，构成网络的网络开关以连接计算机的各端口为单位对数据包收发数等统计信息进行记录。另外，可以从外部参照这些统计信息。

[0069] 在实施例 2 中，由集群控制计算机 140 执行的内部网络监视程序 241 取得构成内部网络的内部网络开关 120 所取得的统计信息。具体而言，取得第一节点的计算机 100 的内部网络适配器 103 以及第二节点的计算机 110 的内部网络适配器 113 分别连接的内部网络开关 120 的端口 120_1 以及端口 120_2 的网络统计信息。

[0070] 图 6 表示内部网络监视程序 241 的处理的流程。内部网络监视程序 241 在一定的周期执行步骤 601 至 602 的处理。首先，参照开关结构表 500 和集群结构表 510，取得构成内部网络的内部网络开关 120 的端口的网络统计信息（步骤 601）。具体而言，参照集群结构表 510 的内部网络的定义，求出该开关和端口的号码，取得并记录其统计信息。

[0071] 在图 5(b) 所示的表 510 中，将第一节点的内部网络开关端口记载为 120_1-120_3，意味着第一节点通过内部网络开关 120 的第一端口 120_1、第三端口 120_3 与内部网络连接。这意味着在图 1 的结构中，在内部网络开关 120 的端口 120_1 上连接内部网络适配器 103，在内部网络开关 120 的端口 120_3 上连接控制网络适配器 102。同样，将第二节点的内部网络开关端口记载为 120_2-120_4，意味着第二节点通过内部网络开关 120 的第二端口 120_2、第四端口 120_4 与内部网络开关 120 连接。另一方面，将第一节点的外部网络开关端口记载为 130_1，意味着第一节点通过外部网络开关 130 的第一端口 130_1 与外部网络连接。这意味着在图 1 的结构中，在外部网络开关 130 的端口 130_1 上连接了外部网络适配器 101。同样，意味着第二节点通过外部网络开关 130 的端口 130_2 与外部网络开关 130 连接。而且，如果参照表 500，则可以取得从内部网络开关 120 取得统计信息所需要的管理网络的地址、开关控制程序。通过这些，可以取得与构成内部网络的端口有关的统计信息。

[0072] 然后，根据所取得的统计信息，判断集群的节点的运转状态（步骤602）。判断的条件多种多样，例如，可以在节点一定时间以上没有对内部网络开关120发送数据时判断为该节点已停止。

[0073] 当存在判断为异常的节点时，使该节点为了与内部网络、外部网络连接而使用的端口无效（步骤603）。如果在这里参照表510，则也可以取得必须无效化的开关以及其端口号码。如果被判断为异常的节点的运转模式为执行状态，对方节点为待机状态，则对方节点的集群控制程序202、212执行系切换，将运转模式从待机状态迁移到执行状态。

[0074] 根据以上，可以通过开关构成集群的内部网络，从集群中隔离根据在此处收集到的统计信息被判断为异常的节点。由此，与在节点执行的集群控制程序202或212独立地使发生了故障的节点从集群断开。例如，在由于集群控制程序或某种原因，节点的运转模式无法变更时，可以断开该节点，抑制对外部的影响。

[0075] 除此之外，除了使异常节点的计算机所连接的端口无效之外，还可以由集群控制计算机140指示对剩余节点的计算机执行系切换（步骤604）。如果被指示的节点的计算机在该时刻的运转模式为待机状态，则可以开始进行启动系切换来迁移到执行状态的处理。由此，可以不等待节点的计算机的集群控制程序检测异常，来开始系切换处理。

[0076] 在实施例2中，通过一个内部网络开关120来构成集群的内部网络，但也可以通过多个开关来构成。此时，可以在节点的计算机上搭载多个用于连接到内部网络的网络适配器，并在集群结构表510的内部端口记载多个端口。网络控制程序242执行记载在表510中的所有端口的有效化／无效化。另外，内部网络监视程序241也可以取得表510中记载的所有内部端口的统计信息来判断节点的计算机的运转状态。由此，即使构成内部网络的内部网络开关120中的一个发生了故障，也可以继续进行作为集群的动作。

[0077] 此外，在上述实施例中，将内部网络开关120和外部网络开关130作为两个开关来构成，但不言而喻也可以将他们做成一个网络开关。

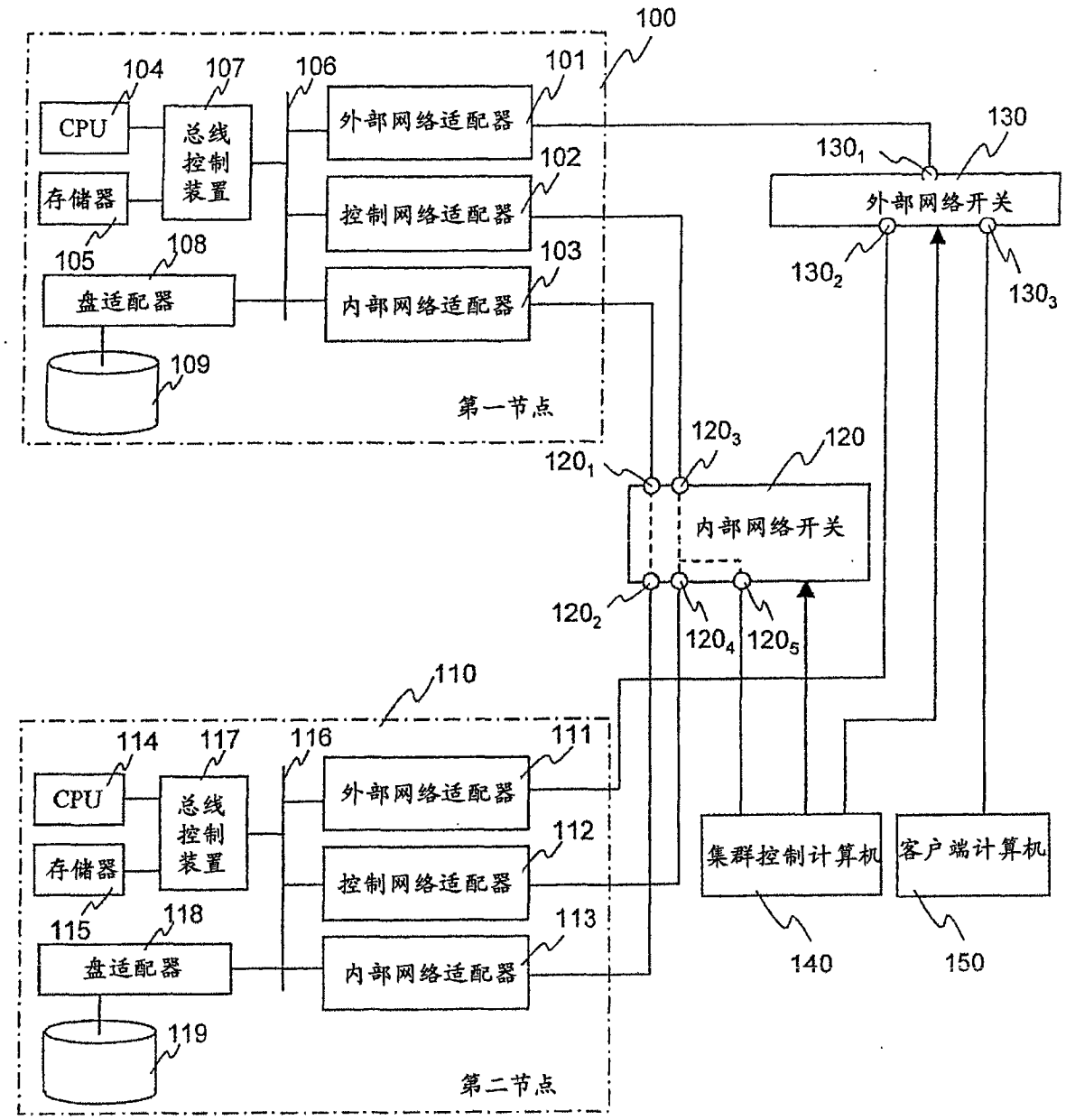

100
104
107
106
101
CPU
总线控制装置
外部网络适配器
130
130₁
102
存储器
控制网络适配器
外部网络开关
105
108
103
130₂
130₃
盘适配器
内部网络适配器
109
第一节点
120₃
120
120₁
内部网络开关
120₂
120₄
120₅
110
114
117
116
111
CPU
总线控制装置
外部网络适配器
集群控制计算机
客户端计算机
112
存储器
控制网络适配器
115
118
113
盘适配器
内部网络适配器
140
150
119
第二节点

图 1

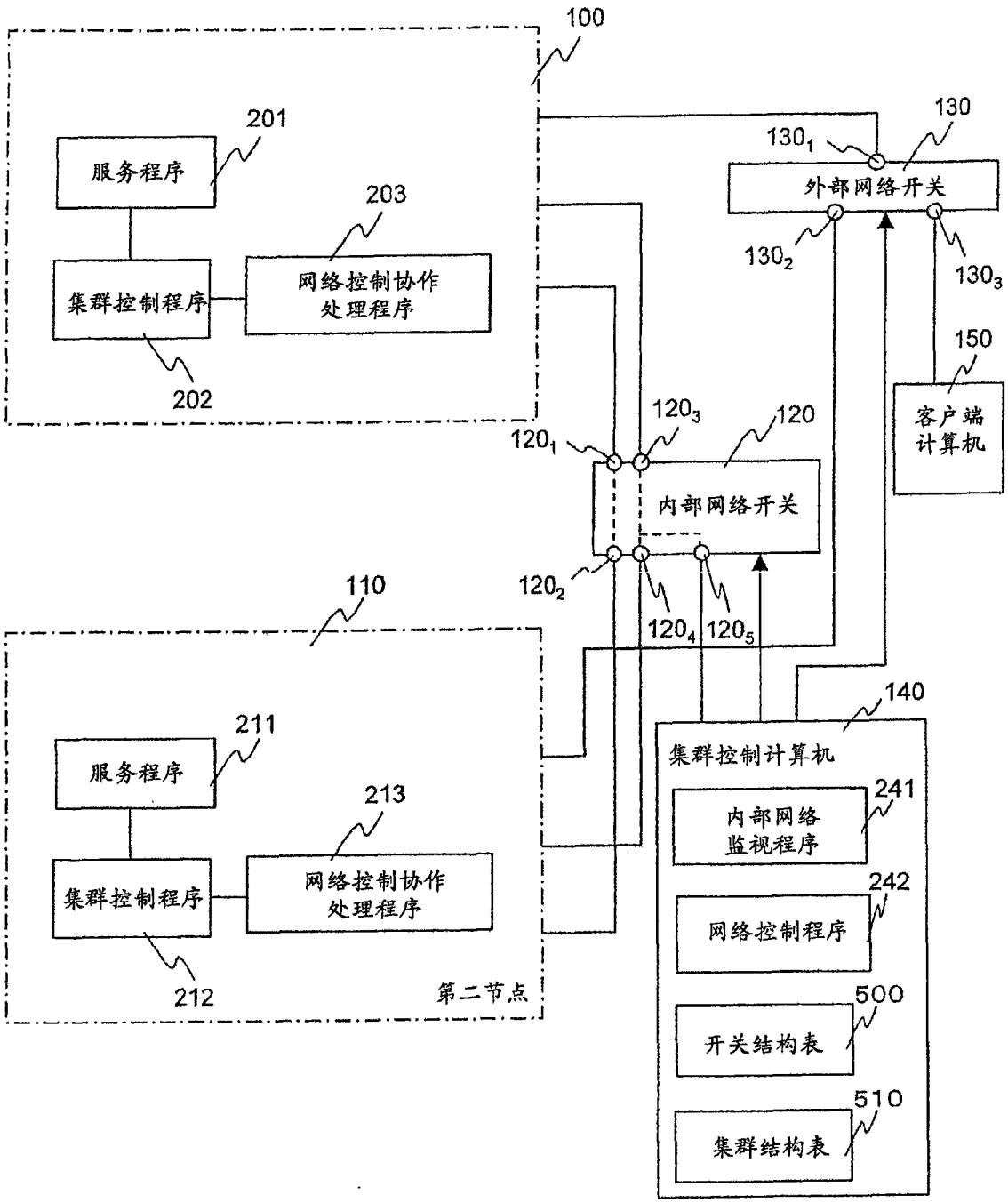

100
201
服务程序
203
网络控制协作
处理程序
集群控制程序
202
130
130_1
外部网络开关
130_2
130_3
150
客户端
计算机
120_3
120
120_1
内部网络开关
120_2
120_4
120_5
110
211
服务程序
213
网络控制协作
处理程序
集群控制程序
212
第二节点
140
集群控制计算机
241
内部网络
监视程序
242
网络控制程序
500
开关结构表
510
集群结构表

图 2

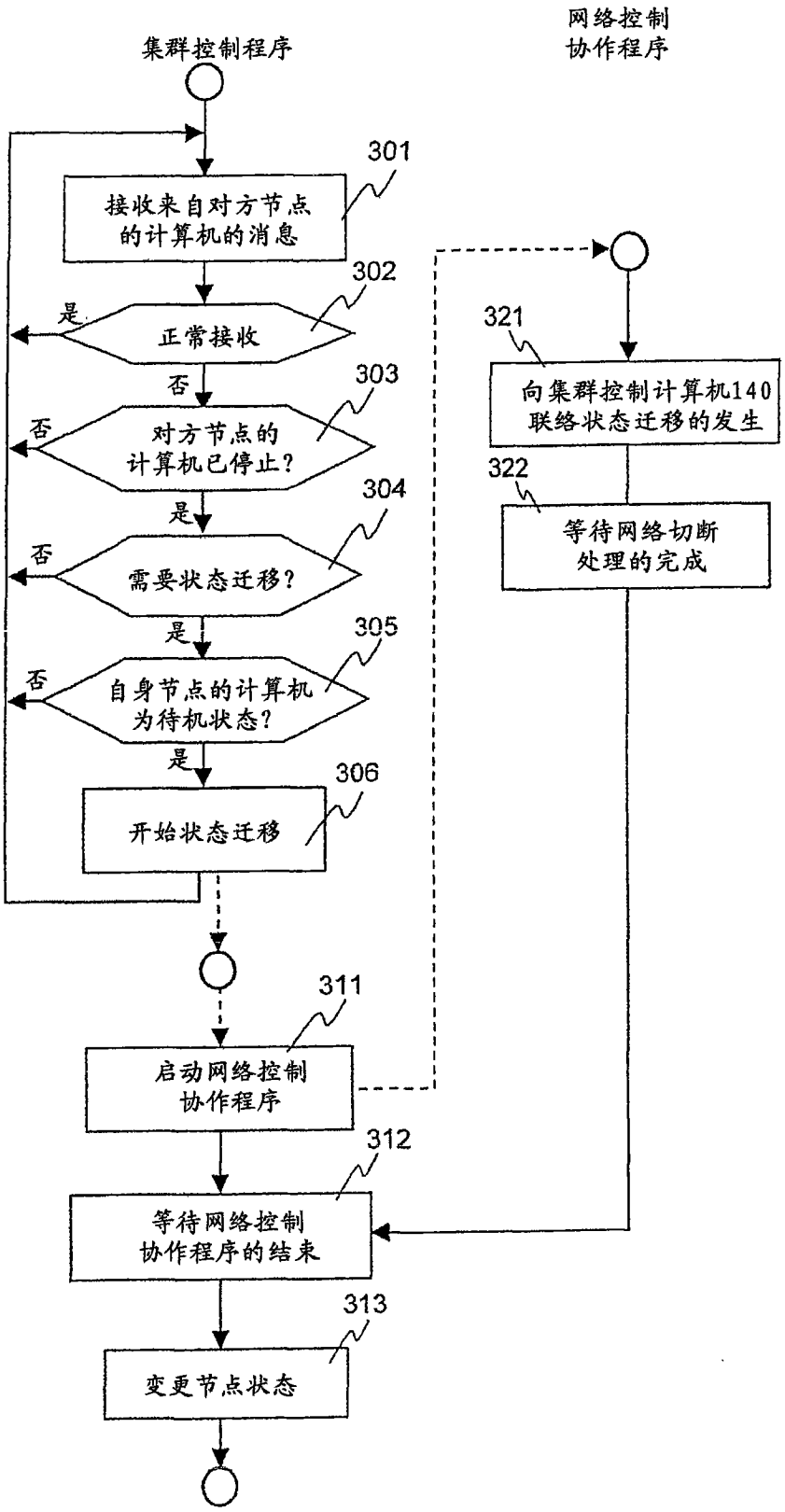

集群控制程序
网络控制
协作程序
301
接收来自对方节点
的计算机的消息
302
是
正常接收
否
303
否
对方节点的
计算机已停止?
是
304
否
需要状态迁移?
是
305
否
自身节点的计算机
为待机状态?
是
306
开始状态迁移
311
启动网络控制
协作程序
312
等待网络控制
协作程序的结束
313
变更节点状态
321
向集群控制计算机140
联络状态迁移的发生
322
等待网络切断
处理的完成

图 3

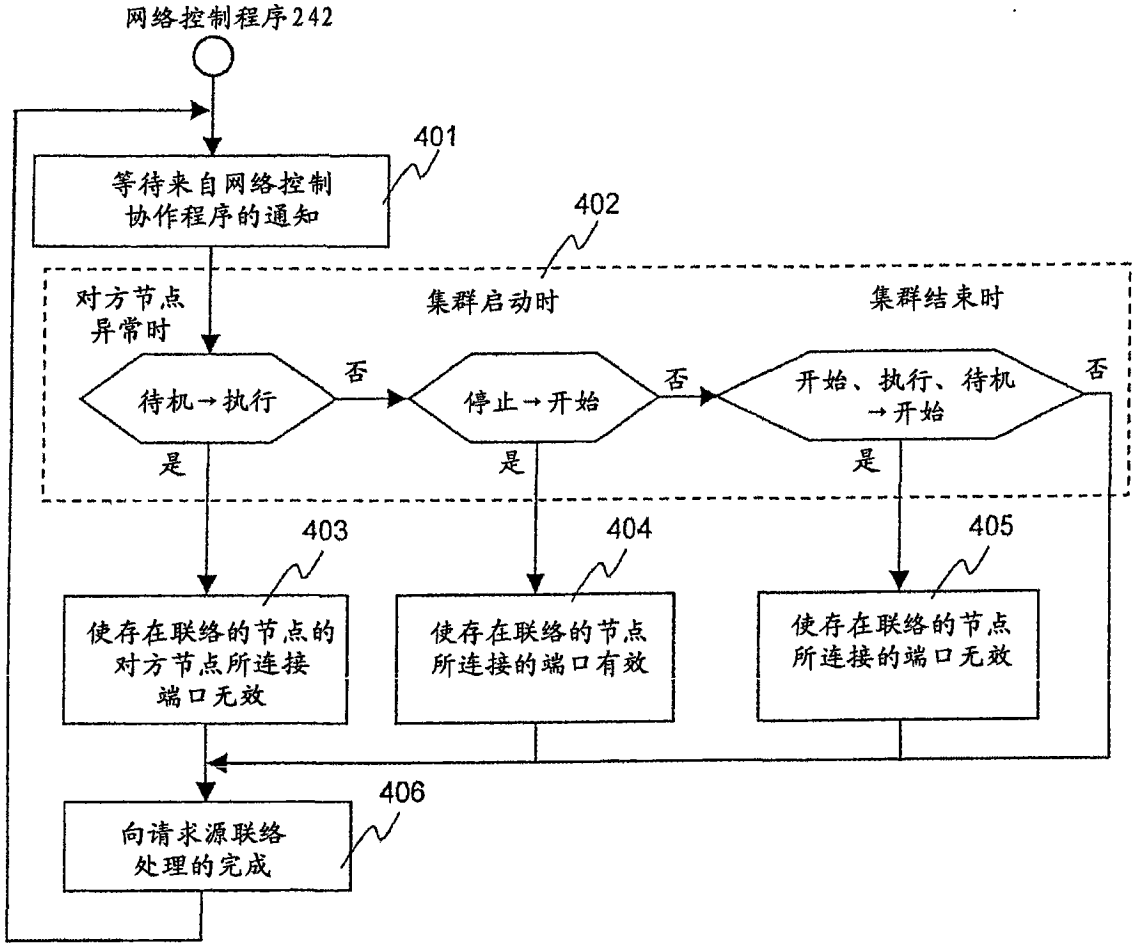

网络控制程序242
401
等待来自网络控制协作程序的通知
402
对方节点异常时
集群启动时
集群结束时
待机→执行
否
停止→开始
否
开始、执行、待机→开始
否
是
是
是
403
404
405
使存在联络的节点的对方节点所连接端口无效
使存在联络的节点所连接的端口有效
使存在联络的节点所连接的端口无效
406
向请求源联络处理的完成

图 4

(a) 开关结构表

500

网络开关号码	控制用网络地址	控制程序
120	10. 0. 0. 1	/usr/bin/sw1ctr1
130	10. 0. 0. 2	/usr/bin/sw2ctr1

(b) 组结构表

510

节点号码	管理用网络地址	内部网络开关端口	外部网络开关端口	开关统计
1	10. 1. 0. 1	120_1—120_3	130_1	正常
2	10. 1. 0. 2	120_2—120_4	130_2	过去两次异常

图 5

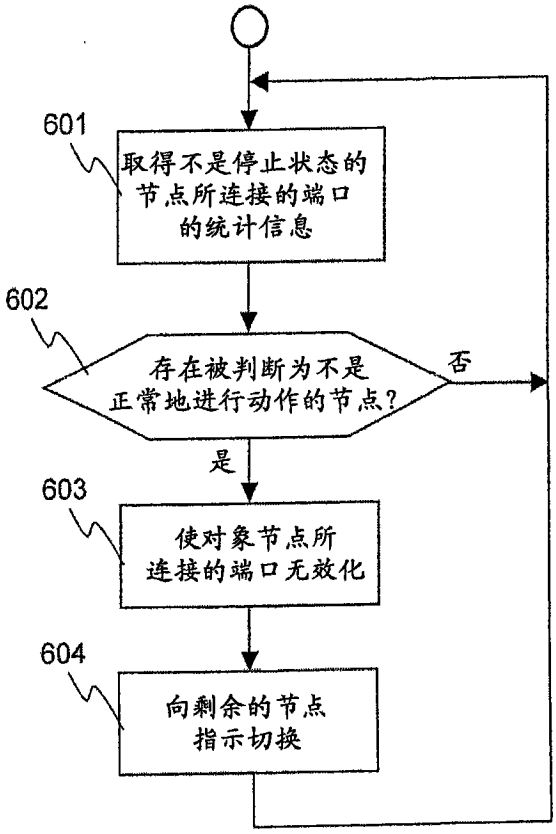

601
取得不是停止状态的节点所连接的端口的统计信息
602
存在被判断为不是正常地进行动作的节点?
否
是
603
使对象节点所连接的端口无效化
604
向剩余的节点指示切换

图 6