



1. 一种服务器系统，包括：

多个节点，每个节点具有用于通过一个或多个网络从服务器系统的客户机接收数据请求的多个物理端口，每个节点还包括一个或多个虚拟服务器，每个虚拟服务器具有与节点的物理端口相关联的至少一个虚拟接口，每个虚拟服务器具有至少一个文件系统，而不管其他虚拟服务器；

通过存储网络与节点连接的多个存储系统；

所述系统还包括：

用于检测第一节点上的物理端口的故障的装置；

用于确定第一节点上的任何其他物理端口是否良好的装置；

用于如果第一节点上有良好的物理端口，则将与故障物理端口相关联的所有虚拟接口迁移到这样的良好物理端口的装置。

2. 如权利要求 1 所述的系统，其特征在于，所述虚拟接口包括一个虚拟 IP 地址。

3. 如权利要求 1 所述的系统，其特征在于，客户机使用 NFS 或 CIFS 协议访问所述文件系统。

4. 如权利要求 1 所述的系统，进一步包括：

用于检测第一节点的故障的装置；以及

用于使在所述第一节点上的每个虚拟服务器迁移到系统中的不同节点的装置。

5. 如权利要求 4 所述的系统，其特征在于，每个虚拟服务器具有相关的故障转移优先级，并且所述系统进一步包括：

用于按它们各自的优先级顺序迁移虚拟服务器的装置。

6. 如权利要求 4 所述的系统，进一步包括：

用于识别标记为在第一节点故障的情况下不进行迁移的虚拟服务器，并在第一节点出故障时阻止迁移这样识别的虚拟服务器的装置。

7. 如权利要求 1 所述的系统，进一步包括：

用于检测连接到第一节点的第一子网中的故障的装置，所述第一节点具有到第一客户机的网络连接；

用于识别具有到所述第一客户机的网络连接以及通过一个不同的第二子网的到所述第一节点的连接的第二节点的装置；

用于响应所检测的故障，将所述第二节点用作路由器以便在所述第一客户机和所述第一节点间路由数据的装置。

8. 如权利要求 7 所述的系统，其特征在于，在所述第一子网中出故障之前，所述第一客户机和所述第一节点间的连接是通过分配给所述第一节点上的物理端口的第一虚拟接口进行的，所述系统进一步包括：

用于将所述第一虚拟接口迁移到与所述第二子网连接的第二节点上的物理端口的装置。

9. 如权利要求 1 所述的系统，进一步包括：

用于如果所述第一节点上没有良好的物理端口，则将与故障物理端口相关的所有虚拟接口连同连接到这种虚拟接口上的所有虚拟服务器一起迁移到一个不同的第二节点上的装置。

10. 如权利要求 9 所述的系统,其特征在于,所述故障物理端口处于第一子网上以及所述良好的物理端口处于一个不同的第二子网上。

11. 如权利要求 1 所述的系统,其特征在于,所述系统进一步包括:

用于计算虚拟服务器负载在系统中除任何故障节点外的节点上的平衡分布的装置;以及

用于通过将一个或多个虚拟服务器从负载重的节点迁移到负载较轻的节点来执行负载平衡的装置。

12. 如权利要求 1 所述的系统,进一步包括在第一节点上:

用于确定所述第一节点上的每个物理端口上的负载的装置;以及

用于在所述第一节点的物理端口中重新分布所述第一节点上的虚拟接口,以进行所述物理端口上的负载平衡的装置。

13. 如权利要求 1 所述的系统,进一步包括:

用于检测第一节点不能访问共享存储系统的装置;以及

用于响应于检测到不能访问所述共享存储系统,如果在所述系统中存在能访问所述共享存储系统的替代节点,则将包含所述共享存储系统上的文件系统的所有虚拟服务器迁移到这样的替代节点上的装置。

14. 如权利要求 11 所述的系统,其特征在于,所述用于计算虚拟服务器负载在系统中除任何故障节点外的节点上的平衡分布的装置以及用于通过将一个或多个虚拟服务器从负载重的节点迁移到负载较轻的节点来执行负载平衡的装置进一步用来确定每个虚拟服务器上的负载。

15. 如权利要求 11 所述的系统,其特征在于,所述用于计算虚拟服务器负载在系统中除任何故障节点外的节点上的平衡分布的装置以及用于通过将一个或多个虚拟服务器从负载重的节点迁移到负载较轻的节点来执行负载平衡的装置用来从第一节点迁移第一虚拟服务器和第二虚拟服务器,所述第一虚拟服务器被迁移到所述系统的第二节点上并且所述第二虚拟服务器被迁移到所述系统的一个不同的第三节点上。

16. 如权利要求 11 所述的系统,其特征在于,所述用于计算虚拟服务器负载在系统中除任何故障节点外的节点上的平衡分布的装置以及用于通过将一个或多个虚拟服务器从负载重的节点迁移到负载较轻的节点来执行负载平衡的装置用来平衡系统负载,作为故障转移过程的一部分。

17. 一种在服务器系统中执行的方法,所述服务器系统包括多个存储系统和通过存储网络与所述存储系统相连的多个节点,每个节点具有用于通过一个或多个网络从群集的客户机接收数据请求的多个物理端口,每个节点还包括多个虚拟服务器,每个虚拟服务器具有分配给节点的物理端口的至少一个虚拟 IP 地址和至少一个文件系统,而不管其他虚拟服务器,所述方法包括:

检测服务器系统的第一节点上的物理端口的故障;

确定第一节点上的任何其他物理端口是否良好;

如果第一节点上有良好的物理端口,则将与故障物理端口相关联的所有虚拟 IP 地址迁移到这样的良好端口;以及

如果所述第一节点上没有这样的良好端口,则将与故障物理端口相关的所有虚拟 IP

地址连同连接到这种虚拟 IP 地址上的所有虚拟服务器一起迁移到服务器系统的一个不同的第二节点上。

18. 如权利要求 17 所述的方法,包括 :

检测服务器系统的第一节点的故障;以及

使所述第一节点上的每个虚拟服务器迁移到所述服务器系统中的不同节点上。

19. 如权利要求 18 的方法,进一步包括 :

按它们各自的优先级顺序迁移第一节点的虚拟服务器。

20. 如权利要求 18 的方法,进一步包括 :

识别标识为在第一节点故障的情况下不进行迁移的虚拟服务器,并且在第一节点出故障的情况下阻止迁移这样识别的虚拟服务器。

21. 如权利要求 18 的方法,进一步包括 :

检测连接到第一节点的第一子网中的故障,所述第一节点具有到第一客户机的网络连接;

识别具有到所述第一客户机的网络连接以及通过一个不同的第二子网的到所述第一节点的连接的第二节点;

响应所检测的故障,将所述第二节点用作路由器以便在所述第一客户机和所述第一节点间路由数据。

22. 如权利要求 18 的方法,进一步包括 :

确定由每个虚拟服务器产生的负载;

计算虚拟服务器负载在服务器系统的除任何故障节点外的节点上的平衡分布;以及

通过将一个或多个虚拟服务器从负载重的节点迁移到负载较轻的节点来执行负载平衡。

## 高可用性集群虚拟服务器系统

### 技术领域

[0001] 本发明涉及俗称为文件服务器的高可用性文件服务器系统。

### 背景技术

[0002] 高可用性服务器系统是即使在系统硬件或软件故障后,也能继续运行的系统。提供高可用性的通用方法是使系统部件加倍。如果一些部件不可用,可以使用另一个代替。强壮的高可用性系统不具有单点故障。单点故障是其故障致使系统不可用的部件。高可用性文件服务器系统通常由两个或多个服务器(节点)的集群(cluster)构成。集群的节点具有它们自身与客户端间的网络连接,并且每个节点直接或间接连接到一个或多个磁盘存储单元。

[0003] 高可用性实现能基于共享磁盘模型或非共享磁盘模型。在共享磁盘模型中,由集群节点同时共享数据,并且锁定管理器用于访问控制。在非共享磁盘模型中,共享数据访问;但是随时地,由一个节点永久地拥有每个磁盘卷。共享磁盘模型是最通用的方法。当未共享磁盘时,必须在两个未共享磁盘集间复制数据,这增加了一些风险和复杂性。

[0004] 高可用性系统中的节点通常由一个或多个指令处理器(通常称为CPUs)、磁盘、存储器、电源、主板、扩展插槽以及接口板组成。在主从设计中,系统集群的一个节点被称为主要或主服务器,以及其他节点被称为辅助、接管或从服务器。主要和辅助节点具有相似的硬件,运行相同的操作系统,具有所安装的相同的补丁,支持相同的二进制可执行程序以及具有相同或非常相似的结构。主要和辅助节点均连接到相同的网络,通过该网络,它们彼此通信以及与客户端通信。两种节点运行兼容的故障转移(failover)软件版本。在一些结构中,除共享磁盘外,每个节点具有它自己的专用磁盘。专用磁盘通常包含引导信息、操作系统、联网软件以及故障转移软件。在一些实现中,专用磁盘是镜像的,或提供冗余磁盘。

[0005] 系统节点不断地彼此监视以便每个节点知道另一个的状态。这种监视能使用称为心跳网络(heartbeat network)的通信链路来实现。能在任一可靠连接上实现心跳网络。在许多实现中,心跳是基于以太网连接。心跳网络还能使用诸如运行串行协议的串行线路之类的来实现,串行协议诸如PPP(点对点协议)或SLIP(串行线因特网协议)。还能通过共享磁盘提供心跳,其中磁盘、或磁盘片专用于基于磁盘的心跳的交换。当心跳停止时,服务器了解心跳伙伴的故障。为避免单点故障,可实现多个心跳网络。一些实现在专用网络(即仅用于心跳通信的网络)上运行心跳,其他的则在公用网络上运行心跳。当心跳停止时,在续存的(surviving)节点上运行的故障转移软件能使自动故障转移透明地发生。

[0006] 在故障转移后,健康节点(healthy node)可以访问与故障节点访问的相同的数据并提供相同的服务。这是通过使健康节点采用与故障节点相同的网络身份以及准许健康节点访问共享磁盘中的数据同时切断(lock out)故障节点来实现的。

[0007] NIC(网络接口卡)有时出故障。一些高可用性系统通过提供备用NIC,具有冗余的网络连接性。NIC能具有一个或多个网络端口。在网络端口出故障的情况下,由故障网络端口提供的网络服务被迁移到备用端口。在这种情况下,不需要故障转移到另一节点。能

为公用和专用心跳网络提供冗余网络连接性。

[0008] 一些高可用性系统支持虚拟网络接口,其中,将多个 IP(因特网协议)指定到相同的物理端口。服务与网络身份(虚拟网络接口)和文件系统(存储器)关联。节点(物理服务器)中的硬件提供网络连接和文件服务器所需的计算资源。虚拟 IP 地址不将客户端与特定的物理服务器连接,其将客户端与在特定物理服务器上运行的特定的服务连接。磁盘和存储设备不与特定物理服务器关联。它们与文件服务器关联。当在节点中有故障时,虚拟网络接口和文件系统被迁移到健康节点。因为这些服务不与物理服务器关联,关于哪个物理服务器提供服务,对客户端来说无关紧要。当设置虚拟 IP 地址或将虚拟 IP 地址从一个物理端口移动到另一个时,生成免费 ARP(地址解析协议)包。这允许客户端、集线器以及交换机更新它们的高速缓存中的 MAC(媒体访问控制)地址,该地址对应于虚拟 IP 地址的位置。

[0009] 所有故障转移导致一些客户端中断。在一些情况下,在完成故障转移后,系统具有比故障转移前更低的性能。当健康节点除提供它自己的服务外,还负责提供由故障节点提供的服务时,会发生这种情形。

## 发明内容

[0010] 一般来说,在一个方面,本发明提供具有两个或多个自主服务器(autonomous server)(称为节点或物理服务器、连接到存储设备)的集群的高可用性集群服务器系统以及计算机程序产品和用于操作这些系统的方法。节点之一是主节点以及其余的为从节点。每个节点运行一个或多个虚拟服务器。虚拟服务器由网络资源和文件系统组成。当一个节点失败时,透明地将其虚拟服务器传送到一个或多个其他节点。这是通过提供两组无缝连接性来实现的。第一组位于节点和客户端之间。第二组位于节点和存储系统之间。第一连接性基于客户端和节点间的虚拟 IP 技术。第二连接性 - 后端(backend)连接性 - 能使用光纤信道、SCSI(小型计算机系统接口)、iSCSI(IP 上的小型计算机系统接口)、InfiniBand™ 体系结构或任何其他这些技术,或使用它们的组合来实现。

[0011] 节点通过心跳网络彼此通信以便确定彼此的健康状况。心跳能在 IP 或 SAN(存储区网络)基础结构,或在两者上操作以便确定节点的可用性。如果节点之一或其部件之一出故障以致在那个节点上运行的虚拟服务器失败,发生故障转移。

[0012] 在故障转移中,使故障节点的虚拟服务器迁移到另一节点上。在某些故障条件下,无缝连接性和冗余硬件和软件部件允许继续访问文件系统而不调用故障转移过程。能为虚拟服务器指定优先级,并且在故障转移后,能在较低优先级虚拟服务器之前调用(bring up)较高优先级虚拟服务器。能通过将虚拟服务器从故障节点分配到多个不同节点来提供负载平衡。

[0013] 一般来说,在另一个方面中,本发明提供系统、程序和方法,其中多个虚拟服务器驻留在单个物理服务器上。每个虚拟服务器独占拥有一个或多个文件系统以及一个或多个虚拟 IP 地址,并且它看不到其它虚拟服务器独占拥有的资源。将虚拟服务器管理为单独的实体并且它们共享物理服务器上的物理资源。

[0014] 一般来说,在另一方面中,本发明提供系统、程序和方法,其中,能可选地不从故障节点迁移不重要的服务。能通过管理器配置来完成设置虚拟服务器的优先级以及防止迁移

不重要的虚拟服务器。

[0015] 一般来说,在另一方面中,本发明提供系统、程序和方法,其中监视节点的加载以便识别加载比其他节点少的节点。使用这一信息来执行负载平衡。在故障转移后,优先于负载较重的节点,将虚拟服务器迁移到负载少的节点。因为节点能支持多个虚拟服务器,即使在没有故障时,在正常操作期间,也能用这种方法实现负载平衡。

[0016] 一般来说,在另一方面中,本发明提供系统、程序和方法,其中,为最小化出现故障转移,每个节点在单个子网或不同子网内具有多个网络端口。(子网是通过提供具有相同前缀的 IP 地址,共享共用地址部件的网络的一部分)。如果端口之一失败,将服务移动到一个续存的端口。这允许多个网络端口故障发生而不调用故障转移,以便仅当没有续存的端口时才出现故障转移。

[0017] 本发明的实现能实现一个或多个下述优点。故障转移仅用作最后一个手段,因此限制由故障转移引起的对服务的可访问性的中断。通过负载平衡,提高总的系统性能。当出现故障时,通过可选消除低优先级服务来提高总的系统性能。

[0018] 在下述附图和说明书中阐述了本发明的一个或多个实现的细节。从说明书、附图和权利要求书,本发明的其他特征和优点将变得显而易见。

## 附图说明

- [0019] 图 1 是根据本发明的一个方面的高可用性服务器系统的框图。
- [0020] 图 2 是示例说明在虚拟服务器故障转移之前,如何使用网络故障转移的图。
- [0021] 图 3 是在虚拟服务器故障转移前,基于网络失败的本发明的实施例。
- [0022] 图 4 是在故障转移后,在图 3 中示例说明的相同实施例。
- [0023] 图 5 示例说明用于高可用性服务器集群的存储体系结构。
- [0024] 图 6 是示例说明高可用性服务器集群的初始化的流程图。
- [0025] 图 7 是示例说明网络端口故障恢复的流程图。
- [0026] 图 8 是示例说明移出虚拟服务器的流程图。
- [0027] 图 9 是示例说明调用虚拟服务器的流程图。
- [0028] 在不同图中相同的标记表示相同的元件。

## 具体实施方式

[0029] 图 1 示例说明根据本发明的高可用性服务器的部件。服务器具有节点集群,节点 A101、节点 B102、...、节点 J103。每个节点具有一个或多个虚拟服务器。节点 A 具有标记为 VSA1、VAS2、... VSAn1 的 n1 个虚拟服务器。节点 B 具有标记为 VSB1、VSB2、... VSBn2 的 n2 个虚拟服务器。节点 J 具有 n3 个虚拟服务器,标记为 VSJ1、VSJ2、...、VSJn3。每个节点通过存储网络 110 连接到一个或多个存储系统。服务器具有多个存储系统 121、122、123。如图 1 所示,每个虚拟服务器拥有一个或多个文件系统 121a、121b、121c、122a、122b、122c、123a、123b 和 123c。有一个可由所有节点访问的共享磁盘 124。该共享磁盘被称为涂写磁盘 (scribble disk);它包含状态和配置数据。该存储网络 110 能是光纤信道、SCSI、iSCSI、InfiniBand 或任何其他这种技术。客户端 105、106 和 107 通过一个或多个网络 104,诸如网络 1、网络 2、...、网络 N 连接到节点上。每个节点具有至少一个物理端口,并且多个虚拟

地址能驻留在相同的物理端口上。RAID 存储接口 112 向所有节点提供逻辑卷支持。每个逻辑卷能由多个磁盘组成 - 例如，在 RAID0、1、5、1+0 或 5+0 配置中。

[0030] 虚拟服务器拥有不包括其他虚拟服务器在内的文件系统和虚拟 IP 地址。它们共享物理服务器上的其他物理资源。虚拟服务器不能看见由其他虚拟服务器独有的资源，并且作为单独的实体被管理。使用虚拟服务器来集合资源（虚拟 IP 地址和文件系统）便于在故障转移期间移动资源并且比单独处理每个资源更有效。

[0031] 每个节点能具有多个网络端口，也称为物理 IP 端口 (PIP)。如果一个端口失败，节点将恢复，只要在该节点上有健康的网络端口。节点上最后一个端口故障导致故障转移到健康的节点。

[0032] 集群中的节点能充当主或从节点。仅有一个主节点，其余节点是从节点（或者处于转移状态，例如均不是）。主节点协调从节点的活动。从节点向主节点报告它们控制的资源。从服务器仅知道它们自己的资源和状态。主服务器维持用于整个集群的状态信息。还维持在负载平衡期间使用的有关服务器负载的信息，其中系统试图在健康节点中大约平均地划分它的工作。

[0033] 在正常操作期间，每个节点测量它的 CPU 使用率以及它的 IOPS（“每秒的 I/O 操作”）的总量。IOPS 的数量表示当由客户端访问时，节点上的总负载。通过共享磁盘或网络，将这一信息传送到主服务器。当特定节点上 CPU 利用率和 / 或 IOPS 上的数量超出阈值时，主节点将检查其他节点上的负载。如果在系统中有能处理更多工作的节点，主节点将一些虚拟服务器迁移到它们上。目的是在健康节点中或多或少平均地划分工作。用于触发负载平衡的 CPU 和 / 或 IOPS 负载的阈值是能通过系统的管理接口控制的可配置参数。

[0034] 在相同节点中，能通过在健康网络端口中重新分配虚拟接口来可选地执行网络端口上的负载平衡。节点中的软件监视节点的物理端口上的负载。如果一个端口实际上处理比其他端口更多的网络通信量，将它的一些虚拟接口移动到不太忙的端口中。选择移动哪一个虚拟接口或哪些虚拟接口能基于每个虚拟接口传送多少通信量而定。

[0035] 在集群中，通过在节点间的网络连接上和在共享磁盘上操作的心跳协议监视资源以便确定每个服务器的可用性。当节点停止接收心跳消息时，它知道另一节点出故障，网络连接上的心跳基于使用 Ping 和 / 或 RPC（远程过程调用）调用探查从节点的主节点而定。能在专用或公用网络上实现 Ping。基于 RPC 的心跳能使用公用网络发送。

[0036] 如果主节点没有在特定时间（例如 3 秒）内从从节点接收响应，那么不能到达从节点，或者从节点有其他的问题。如果主节点停止发送 Ping 或 RPC，从节点假定不能到达主节点或主节点有其他问题。当集群中一个续存的节点确定与一个节点具有连接性或其他问题时，续存的节点必须仍然确定另一节点是真正失效还是简单的不可到达。

[0037] 如果通过 Ping 和 / 或 RPC 的心跳检测到节点故障，使用通过共享磁盘的心跳来找出出故障节点是真正失效还是仅不可到达。如果失效节点是主节点，从节点之一变为新的主节点。为处理所有网络连接损失的可能性，实现通过共享磁盘（涂写磁盘）的心跳。节点通过涂写，换句话说，通过写入和读取涂写磁盘来交换有关它们的状态的信息。用于主从节点的涂写周期随集群的状态而改变。在正常操作期间，主节点缓慢地，例如以每 60 秒一次涂写的速率涂写。当主节点丢失从节点时，它更快，例如以每 3 秒一次涂写的速率涂写。由主节点控制的从节点不涂写。最近丢失主节点的从节点快速地，例如以每 3 秒一次涂写

的速率涂写。既不是主节点也不是从节点的节点缓慢地，例如，以每 60 秒的速率涂写。

[0038] 图 2 示例说明一种实现如何处理网络故障。如果节点具有多个网络端口，以及如果端口之一故障，节点恢复而没有故障转移。图 2 表示节点 1、节点 2、…、节点 N。节点 1 具有 n1 个网络端口，标记为 1PIP1、1PIP2、1PIP3、…、1PIPn1。节点 2 具有 n2 个网络端口，标记为 2PIP1、2PIP2、2PIP3、…、2PIPn2。节点 N 具有 nn 个端口，标记为 NPIP1、NPIP2、NPIP3、…、NPIPnn。例如，假定节点 1 具有附加在虚拟服务器上的虚拟 IP 地址 -VIP1。当端口 1PIP1 失败时，将 VIP1 移动到 1PIP2，如箭头所示。这将不会导致故障转移，因为它处于相同的节点 1 中。当 1PIP2、1PIP3、…、1PIPn-1 故障时，会发生相同的情形。然而，当 1PIPn1 出故障时，在节点 1 上的所有其他 PIP 已经出故障后，故障转移发生以及将 VIP1 移动到节点 2 中的 2PIP1。对其他节点也发生相同的情形；即，虚拟 IP 地址移动到相同节点内的另一物理端口以及仅当当前节点中的所有物理端口失败时，才发生故障转移。在节点内或以其它方式，能将虚拟 IP 地址移动到与故障端口相同的子网内的端口或不同子网中的端口。在一种实现中，优先于不同子网中的端口，将选择相同子网中的端口。

[0039] 在前例子中，虚拟服务器被描述为仅具有一个虚拟 IP 地址。然而，能将单个虚拟服务器连接到多个虚拟 IP 地址上，以及节点能具有许多物理和虚拟 IP 地址。

[0040] 图 3 和图 4 示例说明用于移动虚拟网络接口而不强制故障转移的另一技术。这些图表表示运行两组虚拟服务器：VSA1、…、VSAn1 和 VSB1、…、VSBn2 的两个节点。在图 3 中，两个虚拟 IP 地址 VA11 和 VA12 被连接到虚拟服务器 VSA1。为简化该图，未示出连接到其它虚拟服务器上的虚拟 IP 地址。Net1 和 Net2 是不同的子网。客户端 305 是连接到 Net1 的客户端以及客户端 306 是连接到 Net2 的客户端。HB1 和 HB2 是网络集线器或交换机。客户端 306 在 Net2 上与节点 A 中的虚拟服务器通信。

[0041] 图 4 表示当在 Net2 上的通信故障时所发生的情形。虚拟 IP 地址 VA12 被从节点 A310 迁移到节点 B320 中的物理端口 PIP3。使用网络故障摆脱（failthrough）而不是虚拟服务器故障转移，因为它对客户端损害较少。如前所述，无论何时将虚拟 IP 地址连接到物理接口上以及当将虚拟地址迁移到另一接口上时，生成免费 ARP 包。

[0042] 如图 4 所示，在 Net2 故障后，由节点 B 通过 VA12 所迁移到的 PIP3 接收来自客户端 306 的数据。节点 B 中的路由软件 322 通过 PIP4，将数据转发到节点 A。通过节点 B 中的 PIP4 和 PIP3，通过 PIP1 将来自节点 A 的数据转发到客户端 306。

[0043] 在支持 NFS 文件系统的一种实现中，将 NFS 文件锁定存储在共享磁盘中。每个虚拟服务器拥有相应的 NFS 文件锁定。在故障转移期间，锁的所有权遵循虚拟服务器。因此，虚拟服务器和相应的 NFS 锁被迁移到健康节点。因此，不需要客户端管理 NFS 锁。

[0044] 图 5 详细阐述在其上构建集群的存储基础结构。节点 700、702、…、770 是集群的节点。这些节点能部署适当协议的总线适配器以便连接到共享存储总线或结构 704 上，诸如 SCSI、光纤信道判优环、光纤信道结构、InfinBand、iSCSI 或其他适当的总线或结构上。多个链路 706 和 708、710 和 712、720 和 722 将每个节点连接到共享总线或结构 704 上。这种多个链路允许系统容忍一个链路故障。能提供另外的链路。共享存储单元（多个存储系统）718 能是通过至少两个链路 714 和 716，连接到总线或结构 704 上的一个或多个容错共享存储单元（诸如 RAID5 或 RAI1 阵列）。这种基础结构能经受单点故障。多个故障能导致完全丧失到共享存储单元 718 的访问。

[0045] 在一种有利实现中,集群中的双光纤信道判优环主机总线适配器连接到双光纤信道判优环。这允许光纤信道目标诸如 FC-AL(光纤信道 - 判优环)RAID(独立磁盘冗余阵列)盒附加在光纤信道判优环主机上。在 RAID 盒上定义共享存储单元,诸如 RAID5(奇偶性)或 RAID1(镜像)阵列。

[0046] 可以从每个集群节点访问共享存储单元 718,但通常由用于不同节点的不同路径访问。因此,通过集群范围内的名称 (cluster-widename) 识别每个节点上的每个共享存储单元是很有利的。这消除了当使用本地设备名时,将设备名绑定到共享存储空间上的困难,本地设备名反映路由信息,因为到相同存储空间的路由在不同集群节点上是不同的。为实现此,使用与每个共享存储单元 718 有关的独特标识符。适当的标识符是 FC RAID 控制器的万维 ID(WWID),在该万维 ID 的基础上,定义共享存储单元 718。使用全球可访问名称服务器数据库来将管理器选定的名称与每个共享存储单元的独特标识符相关联。能将数据库存储在任何方便的、全球可访问位置,诸如涂写磁盘或集群外的服务器、但可由所有集群节点访问的位置上。名称服务器由集群节点参考,在它们已经发现共享存储单元并了解共享存储单元的独特标识符后。通过参考名称服务器,集群节点将(能是并且通常是多个)共享存储单元分解成集群范围内的设备名。

[0047] 因为集群节点具有到共享存储单元的多个路径,通过将 I/O(即,输入 / 输出或数据传送)请求更改到相同的共享存储单元,但通过不同路由,执行负载平衡是很有利的。例如,集群节点 700 能通过更改链路 706 和 708 间的数据传送请求来负载平衡。这通过增加可用来访问共享存储单元的整个带宽,有利于集群节点。

[0048] 能将该设计配置成能经受单点或多点故障。设计的强壮性由三个因素来定。第一是每个节点和共享存储总线或结构 704 间的链路数量。第二因素是共享存储总线或结构 704 和数据存储单元 718 间的链路数量。对于每对元件间仅两条链路,如图 5 所示,设计能容许单点故障。对于集群节点中的多个总线适配器,总线适配器能出故障并且到共享存储单元的数据传送请求能以半个带宽性能继续。相关物理接口(诸如电缆)也能出故障。类似地容许电缆的任何单点故障。由于链路数量为 2,通过增加链路的数量,能将单点故障容限提高到更高容限。共享存储单元是能容许成员驱动器(member drive)故障的容错 RAID 阵列。如果使用多个 RAID 控制器来控制相同的共享存储单元,那么能容许 RAID 控制器故障。

[0049] 通过节点所有权锁定保护共享存储单元以便保证专用节点使用率。每个节点知道其它节点的共享存储单元所有权。如果它确定共享存储单元由某个其他节点拥有,其将共享存储单元标识为在那个节点上不可用。

[0050] 存储抽象,诸如虚拟存储技术允许节点使虚拟存储单元跨越多个共享存储单元。这提高容错以及性能。使用多个共享存储单元,在节点上创建虚拟存储单元。这些虚拟存储设备能跨越由不同存储控制器控制的多个共享存储单元,以及支持有效数据保护和数据传送性能特征。虚拟存储设备能是级联、镜像或带状的多个共享存储单元。

[0051] 级联提供的优点是容量扩展。当共享存储单元与另一共享存储单元级联时,当第一共享存储单元满时,使用第二共享存储单元。

[0052] 通过带状的共享存储单元,在各种成员共享存储单元中交替顺序 I/O 请求。带状虚拟存储设备提供扩展和性能。因为在不同共享存储单元上并行分配数据传送请求,与使

用单个共享存储单元相比,节点经历更高吞吐量。

[0053] 通过 2 个不同共享存储单元的虚拟存储镜像 (RAID1),在每个成员共享存储单元上加倍 I/O 操作。通过以预定最低寻道时间从成员进行读取,增强来自镜像的读取操作。当确定镜像损坏以及正确地替换损坏的成员时,镜像同步是自动的。镜像虚拟存储设备通过容许完全损失共享存储单元来提供额外的容错层。通过部署镜像虚拟存储设备,集群的容错性能增加两倍。

[0054] 图 6 示例说明根据本发明的高可用性系统的初始化。在步骤 1100,系统集群中的所有节点被配置成指向将用作涂写磁盘的相同的共享存储单元。在步骤 1101,分配一个节点以便初始化涂写磁盘。初始化涉及从配置文件提取数据。在步骤 1102,在一个节点中启动高可用性软件。该节点变为集群的主服务器。在步骤 1103,在所有其他节点上启动高可用性软件。这些节点是集群中的从节点。在步骤 1104,主节点将虚拟服务器分配给从节点。如果需要的话,能手动完成这一步骤。

[0055] 图 7 表示具有多个网络端口的节点如何检测和处理网络故障。如现在将描述的,其通过测试其每一个端口来完成。在步骤 1200,节点使用正测试的端口,以频繁间隔(诸如每 3 秒)向先前可达的外部端口发送 Ping 包。进行 ping 的频率是可配置的。在判定步骤 1202,节点确定在预定等待时间(诸如 250msec(毫秒))内是否接收到对 Ping 的响应。等待时间是可配置的。如果接收到响应,在步骤 1201,将正测试的端口标记为好的。否则,在步骤 1203,节点已知的可达外部 IP 地址被划分成组。组中的地址总数是可配置的。在步骤 1204,一次一组地将 Ping 消息发送到每个组的地址。这不是使用广播完成的,因为广播更昂贵。在判定步骤 1205 中,节点确定在等待时间内是否达到组中的任一地址。如果到达一个,在步骤 1201,将正测试的端口标记为好的并继续执行。如果没有到达所有组中的任何地址,在步骤 1206 中继续执行。在步骤 1206,发送广播消息。在判定步骤 1207,如果在等待时间内接收到任一响应,将正测试的端口标记为好的以及在步骤 1201 继续执行。否则,节点断定正测试的端口状况不良,以及在步骤 1208,将该端口标记为损坏。

[0056] 在判定步骤 1302,节点确定在节点中是否有健康的网络端口。如果有,在步骤 1304,故障节点的虚拟地址被迁移到健康网络端口。否则,在步骤 1303,请求故障转移到集群中的另一节点。

[0057] 对节点已经标记为好的每个物理端口,执行图 7 的过程。

[0058] 网络端口的故障仅是调用故障转移的可能理由的一个。能导致故障转移的其他事件包括节点或存储系统的一个中的硬件故障、电源故障、节点和存储系统间的链路中的故障、存储总线或结构中的不可恢复故障,以及共享存储单元和存储总线或结构间的链路中的故障。也能手动地启动故障转移。在校正导致故障转移的问题后,能执行手动故障恢复命令以便将虚拟服务器迁移到它们的原始节点。

[0059] 例如,如果由于任何原因,包含文件系统的共享存储单元不能由节点访问(例如,关于在图 5 中示例说明的特定结构中的节点 700,由于节点和单元间的连接完全断开,诸如链路 706 和 708 的故障),那么使包含不可访问文件系统的虚拟服务器迁移到能访问存储单元,因此能访问文件系统的另一物理节点,如果这种替换节点存在的话。

[0060] 图 8 表示当节点中的虚拟服务器在其迁移到另一节点前关闭时执行的步骤。在这一例子中,虚拟服务器具有 NFS 文件系统和 CIFS 文件系统。在步骤 1401,断开属于虚拟服

务器的所有虚拟接口。在步骤 1402, 取消初始化 (de-initialize) 所有 NFS 共享。在步骤 1403, 执行 NFS 锁定清除。在步骤 1404, 取消初始化虚拟 CIFS (公用因特网文件系统) 服务器和共享。在步骤 1405, 取消安装 (un-mount) 属于虚拟服务器的所有文件系统。

[0061] 图 9 示例说明调用虚拟服务器所需的步骤。再次, 在这一例子中, 虚拟服务器具有 NFS 文件系统和 CIFS 文件系统。在步骤 1501, 节点安装属于故障虚拟服务器的所有文件系统。在步骤 1502, 调用属于虚拟服务器的虚拟接口。在步骤 1503, 初始化 NFS 共享。在步骤 1504, 执行 NFS 锁定恢复。在步骤 1505, 初始化虚拟 CIFS 服务器和共享。

[0062] 系统能同时服务于各种文件系统。文件系统由于内部文件系统元数据不一致性会出故障, 有时称为文件系统退化。在系统的一种实现中, 当检测到退化 (通常由文件系统本身来完成), 节点中的软件处理文件系统的修复而不完全中断客户机使用 NFS 协议访问文件系统。在文件系统退化的情况下, 暂时阻塞 NFS 客户机对文件系统的访问。按其属性, NFS 协议会继续向服务器发送请求。在阻塞用于 NFS 访问的文件系统后, 软件防止客户机访问文件系统, 然后修复它 (例如, 通过运行实用程序诸如 fsck)。在修复文件系统后, 软件使得客户机可再次访问。然后, 去除 NFS 阻塞, 以便能再次服务于来自客户机的 NFS 请求。结果, 客户机上的应用会暂时冻结而不会失效, 但只要文件系统回到在线, 即可恢复。

[0063] 可以任何方式实现系统的管理配置。例如, 在系统节点上或单独个人计算机上运行的应用程序能定义和修改用来控制系统配置和操作的参数。在如上所述实现中, 这些参数存储在位于涂写磁盘的配置文件中。然而, 配置数据能被存储在任意数量的文件中、数据库中, 或者以其它方式, 并通过任何适当的装置提供给系统。

[0064] 在某些方面中, 本发明能以实际地嵌入计算机可读存储设备中的、用于由可编程处理器执行的计算机程序产品实现; 以及本发明的方法步骤可由执行指令程序的可编程处理器执行, 以便通过操作输入数据和生成输出来执行本发明的功能。适当的处理器包括, 举例来说, 通用和专用微处理器。通常, 处理器将从只读存储器和 / 或随机存取存储器接收指令和数据。适合于实际嵌入计算机程序指令和数据的存储设备包括所有形式的非易失性存储器, 包括例如半导体存储器件、磁盘诸如内部硬盘和可移动磁盘、磁光盘以及 CD-ROM 盘。上述的任何一个可通过 ASIC (专用集成电路) 来补充或纳入 ASIC 中。

[0065] 为提供与用户的交互作用, 能在具有用于向用户显示信息的显示设备 (诸如监视器或 LCD 屏) 和用户能通过其向计算机系统提供输入的键盘和定点设备 (诸如鼠标或跟踪球) 的计算机系统上实现本发明的方面。计算机系统能编程为提供图形用户界面, 通过该图形用户界面, 计算机程序与用户交互作用。

[0066] 根据特定的实施例描述了本发明。其他实施例在下述权利要求的范围内。例如, 可按不同顺序执行本发明的步骤, 但仍然能获得所需结果。

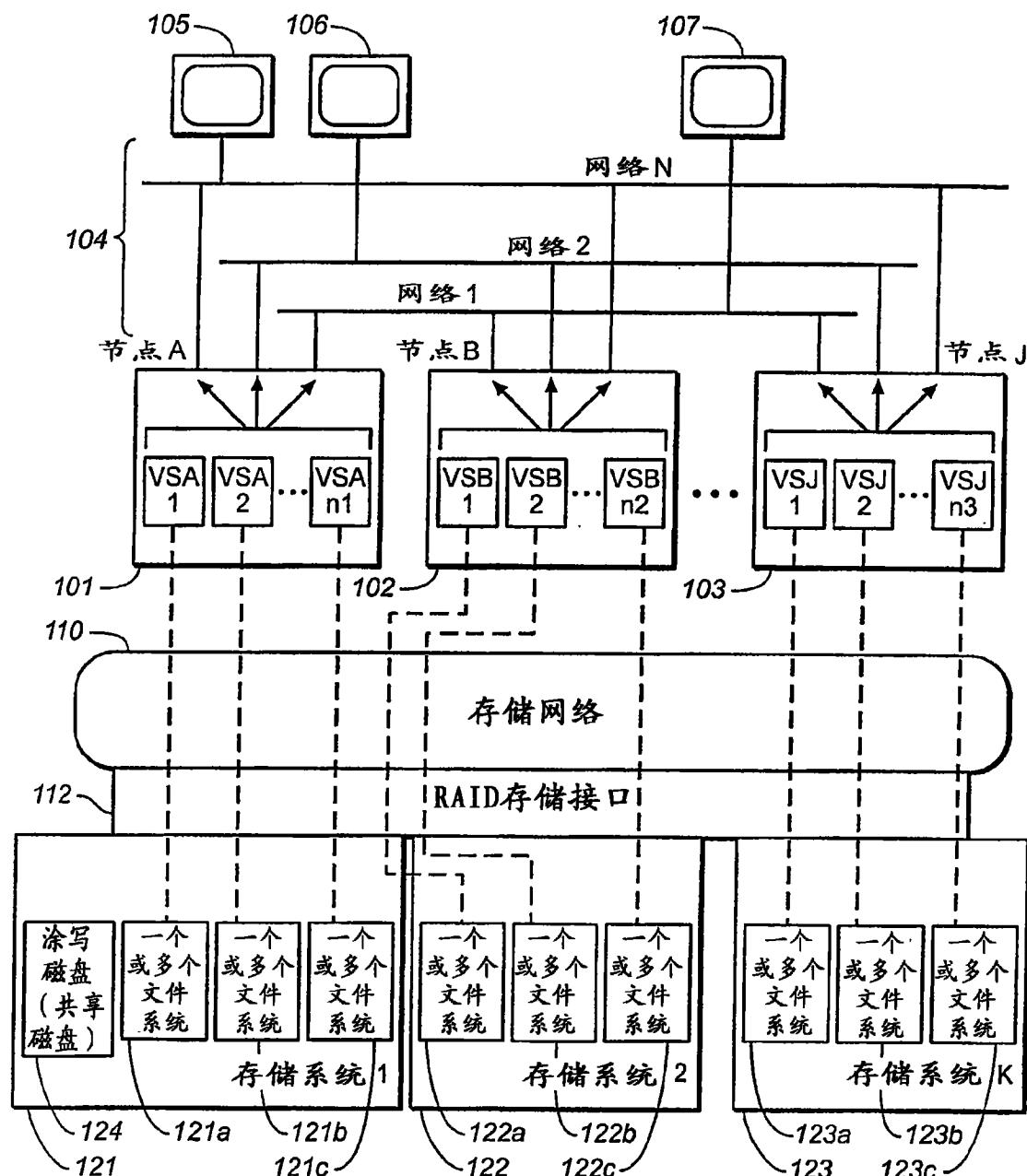


图 1

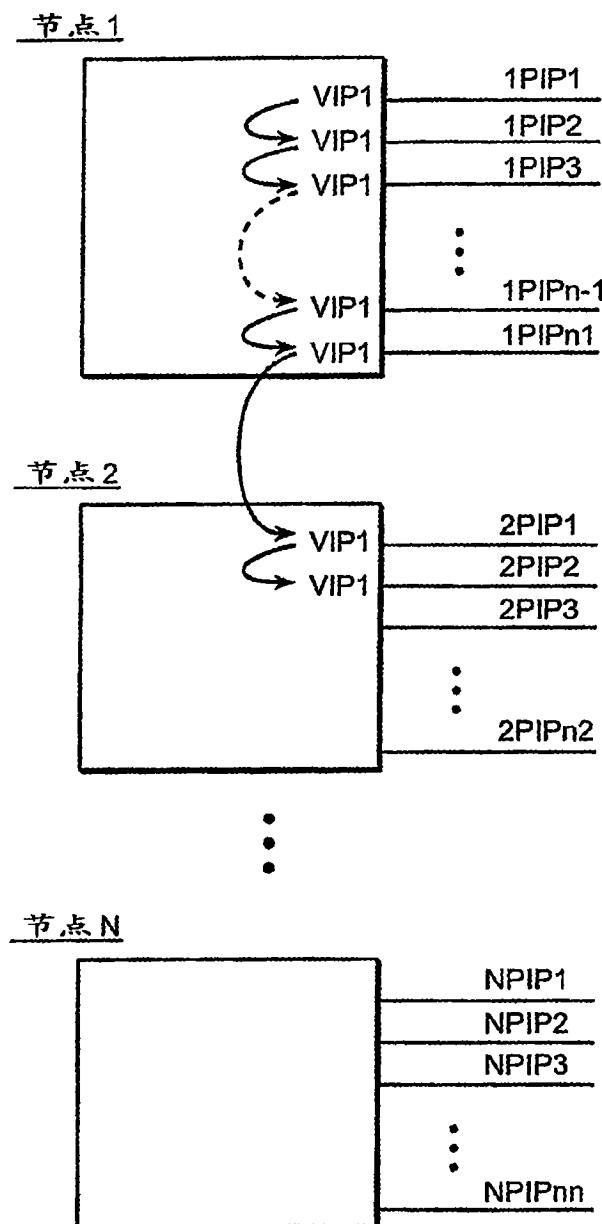


图 2

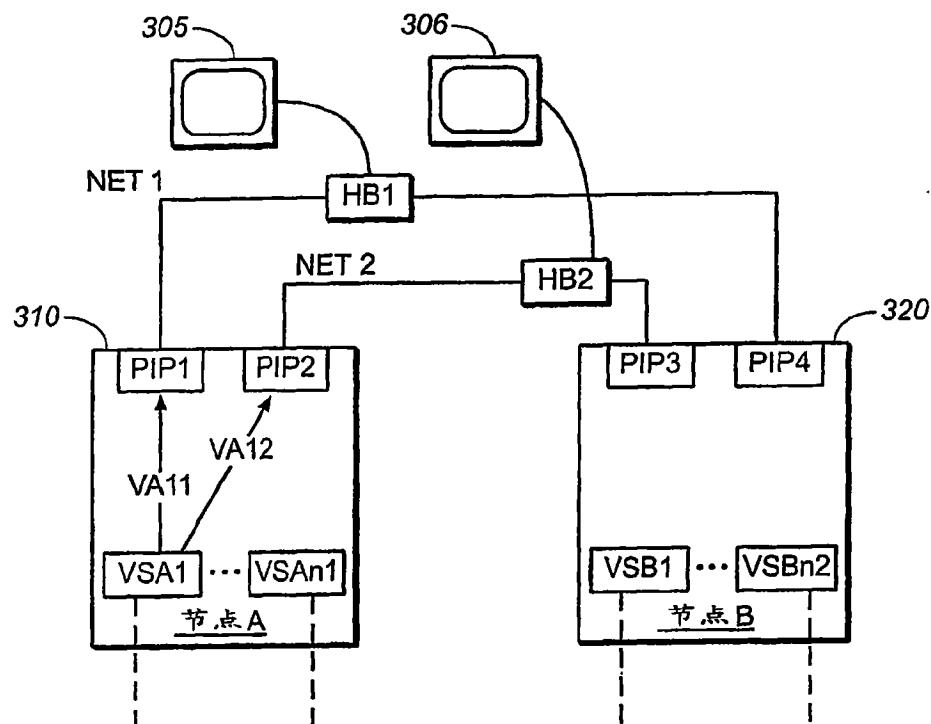


图 3

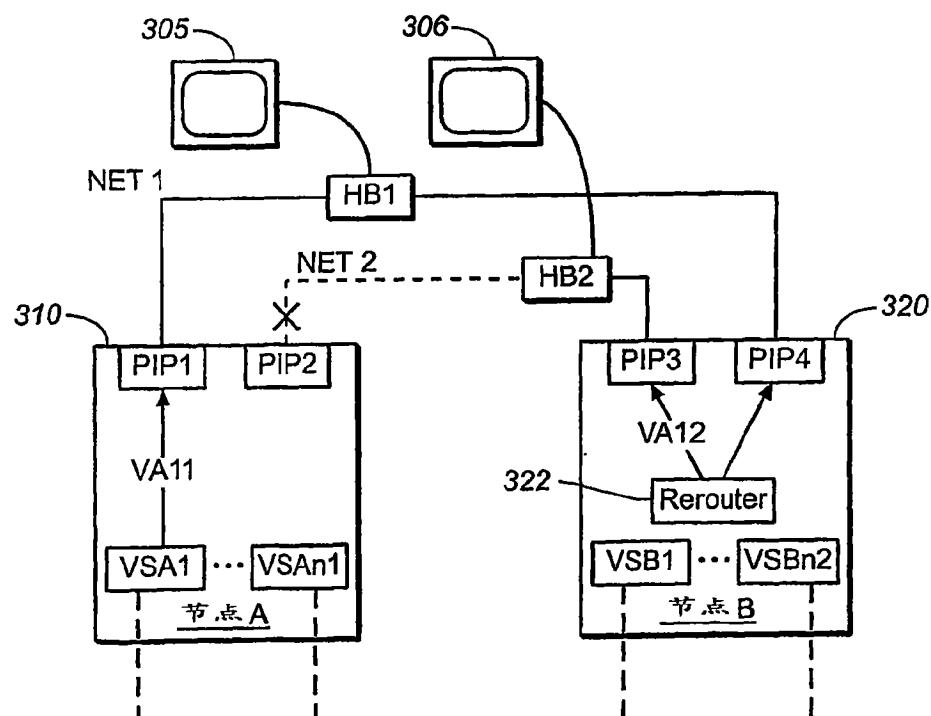


图 4

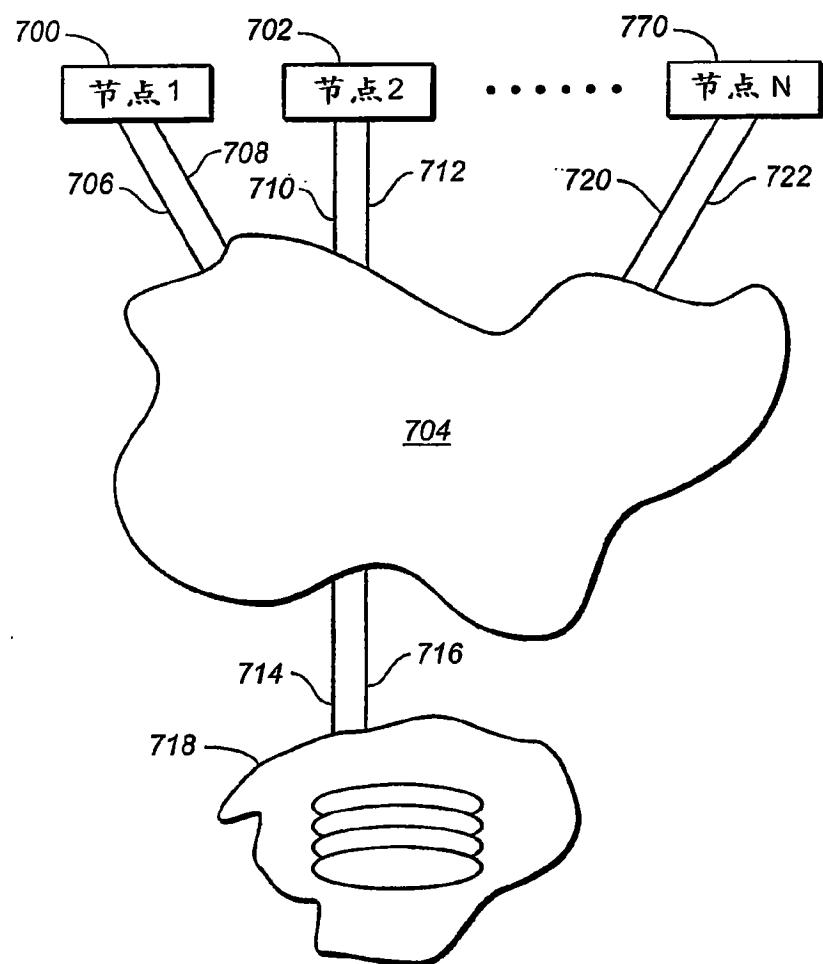


图 5

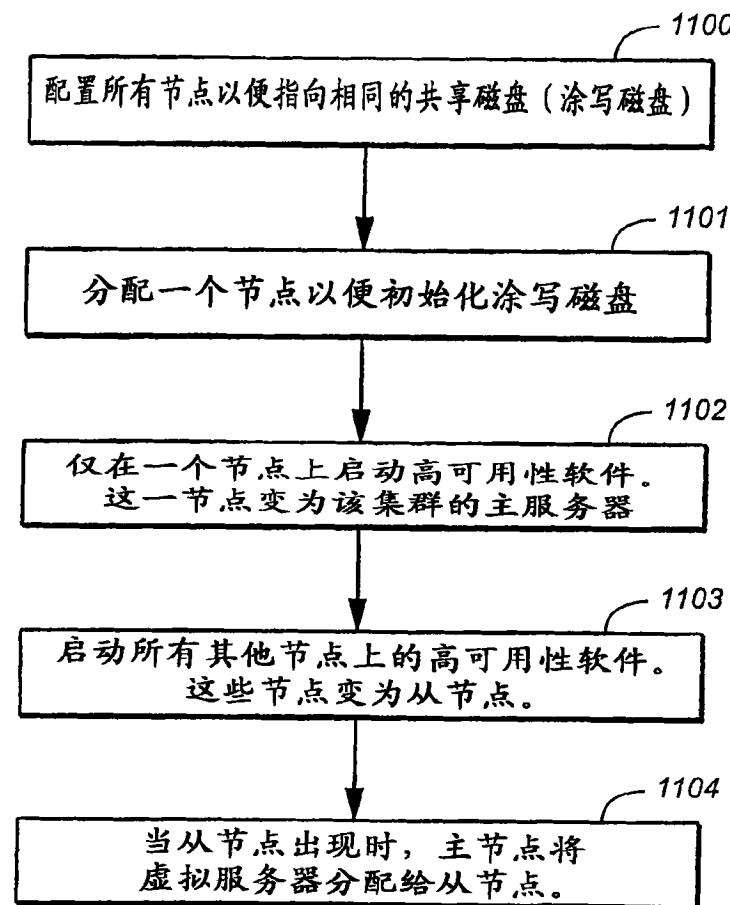
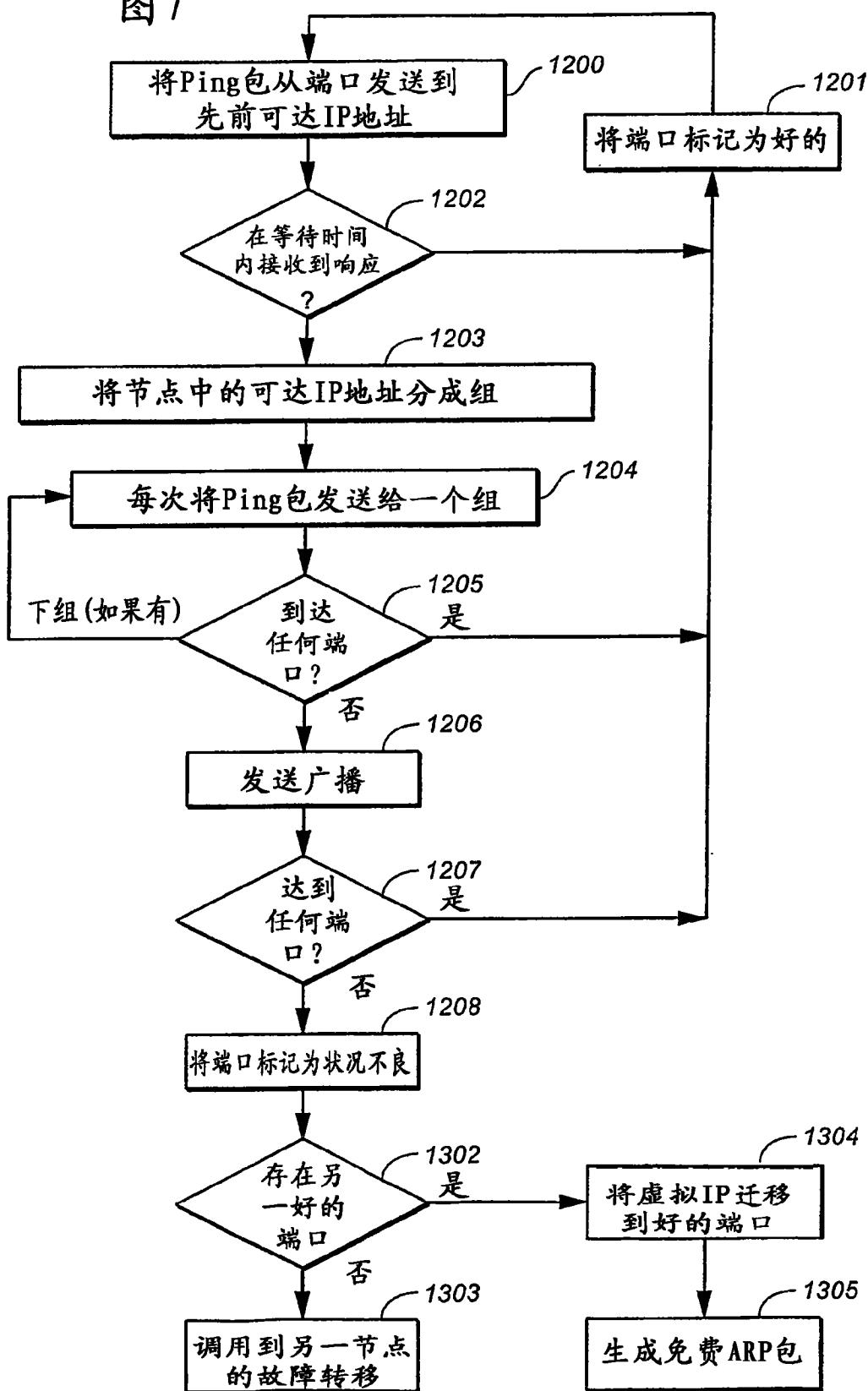


图 6

图 7



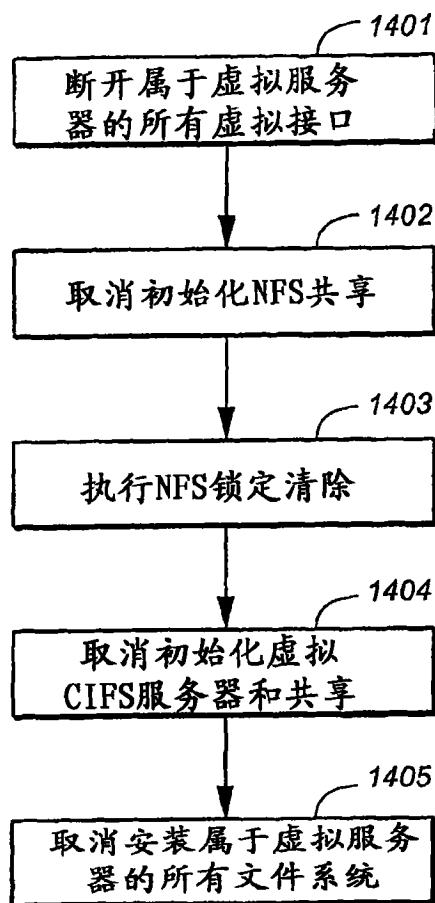


图 8

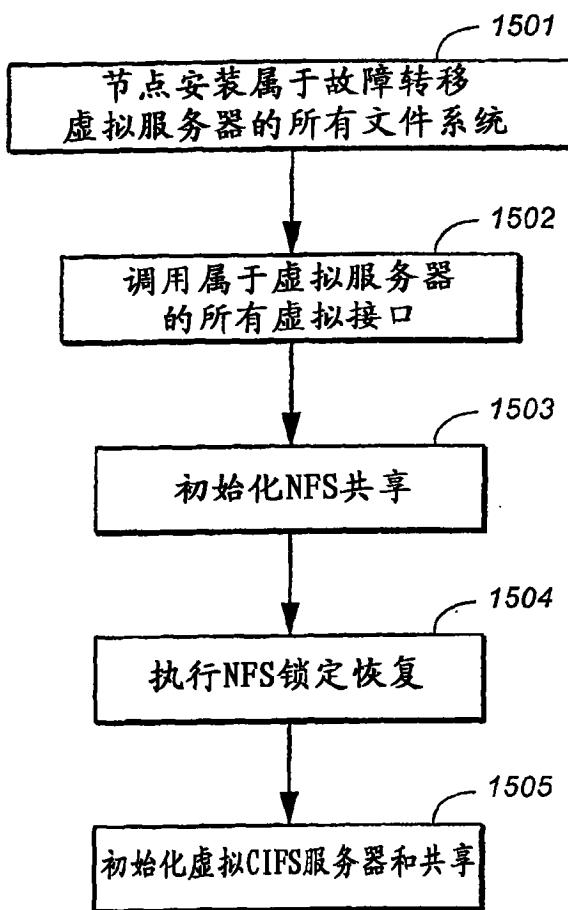


图 9