[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
H04L 29/00 (2006.01)
G06F 11/20 (2006.01)

[12] 发明专利说明书

专利号 ZL 200410042886.2

[45] 授权公告日 2009 年 9 月 23 日

[11] 授权公告号 CN 100544342C

[22] 申请日 2004.5.27
[21] 申请号 200410042886.2
[30] 优先权
[32] 2004. 3.19 [33] JP [31] 2004 - 079882
[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 中谷洋司 中野隆裕
[56] 参考文献
US2003/0018927A1 2003.1.23
CN1404277A 2003.3.9
US2003/0217030A1 2003.11.20
CN1469253A 2004.1.21
审查员 刘心蕾

[74] 专利代理机构 北京银龙知识产权代理有限公司
代理人 郝庆芬

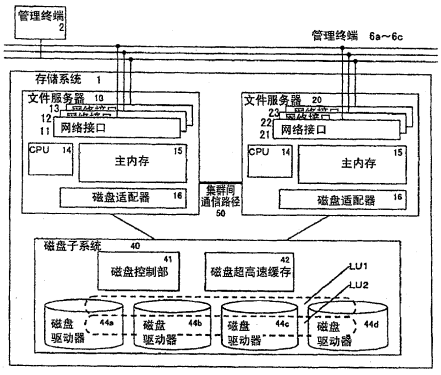
权利要求书 5 页 说明书 12 页 附图 4 页

[54] 发明名称

存储系统

[57] 摘要

在由被设定了虚拟文件服务器的多个文件服务器构成的集群内、使虚拟文件服务器动态地移动。是配备有第 1 文件服务器、第 2 文件服务器和磁盘子系统的存储系统；在各文件服务器中，设有：设定虚拟文件服务器的虚拟文件服务器控制处理部；在虚拟文件服务器中，设有：存储在通信中所需要的路径信息的路由表；在将上述第 1 文件服务器的虚拟文件服务器、故障切换到了上述第 2 文件服务器后，在上述第 2 文件服务器中被起动的虚拟文件服务器，使用在上述第 1 文件服务器中相应虚拟文件服务器使用过的路由表、来决定通信路径。



1. 一种存储系统，其特征为：

是配备有第 1 文件服务器、第 2 文件服务器和磁盘子系统的存储系统；

上述各文件服务器，具有：

针对网络输入输出数据的网络接口，

和进行针对上述磁盘子系统进行输入输出处理的磁盘适配器，

和进行运算处理的 CPU，

和存储上述 CPU 运算处理中所需数据的存储部；

在上述各文件服务器中，设有：

多个虚拟文件服务器，

和控制虚拟文件服务器的起动、终止，并在上述文件服务器中设定多个虚拟文件服务器的虚拟文件服务器控制处理部，

和接收来自上述虚拟文件服务器的请求，对磁盘访问部发出指示来处理对文件的访问的文件系统处理部，

和接收来自上述文件系统处理部的指示，进行针对上述磁盘子系统的输入输出处理的磁盘访问部，

和监视在上述文件服务器中所设定的虚拟文件服务器的动作状态，检测发生了故障的虚拟文件服务器的虚拟文件服务器故障监视处理部，

和监视上述文件服务器的故障，根据相应故障监视结果进行文件服务器间的故障切换的装置间故障监视处理部，

通过上述文件服务器间的通信，同步上述虚拟文件服务器的起动和终止的装置间同步处理部，

进行上述文件服务器的管理的装置管理处理部；

在上述多个虚拟文件服务器的每个中，设有：

利用上述网络接口的设定和上述网络接口，在与网络之间接收发送信号的网络处理部，

和存储有上述网络接口的控制信息的网络接口信息存储部，

和存储在与经由上述网络接口所连接的设备间的通信中所需路径信息的路由表，

和根据来自上述存储系统之外的管理终端的指示，进行每个虚拟文件服务器的设定的虚拟文件服务器管理处理部，

和存储有由上述虚拟文件服务器可访问的文件系统信息的安装表，

和记录有设备名和设备 ID 的设备文件；

上述磁盘子系统，具有：

存储由上述虚拟文件服务器访问的数据的磁盘驱动器，

和控制针对上述磁盘驱动器的数据的输入输出等磁盘子系统动作的磁盘控制部，

和临时存储对于上述磁盘驱动器所输入输出的数据的磁盘超高速缓存；

在上述磁盘驱动器中，存储有服务状况文件的副本和构成信息，并设定了由上述各文件服务器共同可访问的公用卷，该构成信息记录对于各虚拟文件系统的资源的分配，该服务状况文件记录上述安装表、上述路由表以及上述设备文件；

上述第 1 文件服务器的虚拟文件服务器控制处理部，为了将在上述第 1 文件服务器上运行的虚拟文件服务器故障切换到上述第 2 文件服务器，发布在上述第 1 文件服务器上运行的虚拟文件服务器的终止指示，

上述第 1 文件服务器的装置间同步处理部，对上述第 2 文件服务器请求故障切换，

上述第 2 文件服务器的虚拟文件服务器控制处理部，从上述磁盘驱动器读出在上述第 1 文件服务器上运行的上述虚拟文件服务器的构成信息、上述安装表、上述路由表以及上述设备文件，

上述第 2 文件服务器的虚拟文件服务器控制处理部，在上述第 2 文件服务器中起动虚拟文件服务器，

上述第 2 文件服务器的虚拟文件服务器，用相应读出的路由表来决定通信路径。

2. 一种存储系统，其特征为：

是配备有第 1 文件服务器、第 2 文件服务器和磁盘子系统的存储系统；

上述磁盘子系统，具有：

存储由虚拟文件服务器访问的数据的磁盘驱动器，

和将上述磁盘驱动器中所存储的数据发送给上述文件服务器的磁盘控制部；

上述各文件服务器，具有针对网络输入输出数据的网络接口；

在上述各文件服务器中，设有：

多个虚拟文件服务器，

和控制虚拟文件服务器的起动、终止，并在上述文件服务器中设定虚拟文件服务器的虚拟文件服务器控制处理部；

在上述虚拟文件服务器中，设有：

利用上述网络接口的设定和上述网络接口，在与网络之间接收发送信号的网络处理部，

和存储在与经由上述网络接口所连接的设备间的通信中所需路径信息的路由表；

将上述第 1 文件服务器的虚拟文件服务器故障切换到上述第 2 文件服务器后，在上述第 2 文件服务器中所起动的虚拟文件服务器，使用在上述第 1 文件服务器中相应虚拟文件服务器使用过的路由表来决定通信路径。

3. 权利要求 1 所述的存储系统，其特征在于：

在将上述第 1 文件服务器的虚拟文件服务器故障切换到上述第 2 文件服务器时，上述第 2 文件服务器的虚拟文件服务器控制处理部，通过从在上述磁盘子系统内所设的公用卷读出路由表，取得在上述第 1 文件服务器中相应虚拟文件服务器所使用过的路由表来起动虚拟文件服务器，

在上述第 2 文件服务器中所起动的虚拟文件服务器，使用相应取得的路由表来决定通信路径。

4. 权利要求 2 所述的存储系统，其特征在于：

在将上述第1文件服务器的虚拟文件服务器故障切换到上述第2文件服务器时，上述第2文件服务器的虚拟文件服务器控制处理部，起动虚拟文件服务器，

被起动的虚拟文件服务器，使用通过与上述第1文件服务器进行通信取得的、在上述第1文件服务器中相应虚拟文件服务器使用过的路由表来决定通信路径。

5. 权利要求2所述的存储系统，其特征在于：

上述虚拟文件服务器控制处理部，在其所在的文件服务器中，设定被连接到第1网络的第1虚拟文件服务器、和被连接到第2网络的第2虚拟文件服务器，

上述第1网络和上述第2网络属于不同的网段。

6. 权利要求2所述的存储系统，其特征在于：

上述第1文件服务器，配备有：

检测在上述第1文件服务器中所设定的虚拟文件服务器的故障的虚拟文件服务器故障监视处理部，

和控制上述虚拟文件服务器的起动和终止的定时的装置间同步处理部；

由上述虚拟文件服务器故障监视处理部一旦检测到虚拟文件服务器的故障，上述装置间同步处理部对上述第2文件服务器请求故障切换。

7. 权利要求6所述的存储系统，其特征在于：

上述虚拟文件服务器故障监视处理部，特别指定发生故障的虚拟文件服务器，

上述装置间同步处理部，对于上述第2文件服务器发送包含特别指定发生故障的虚拟文件服务器的信息的故障切换请求。

8. 权利要求7所述的存储系统，其特征在于：

上述虚拟文件服务器故障监视处理部，根据来自上述虚拟文件服务器的通知识别上述虚拟文件服务器的故障。

9. 权利要求2所述的存储系统，其特征在于：

上述第1文件服务器，配备有控制上述虚拟文件服务器的起动和终

止的定时的装置间同步处理部，

上述装置间同步处理部，在终止了上述第 1 文件服务器的虚拟文件服务器之后，起动上述第 2 文件服务器的虚拟文件服务器。

存储系统

技术领域

本发明，涉及由设定了虚拟文件服务器的多个文件服务器构成了集群的存储系统，特别是涉及虚拟文件服务器的切换技术。

背景技术

在传统的计算机中的逻辑分割技术中，是逻辑分割计算机内的处理器、内存等资源将其分配给各个虚拟计算机。

提出了这样一种技术：利用这一技术，通过将网络资源等分割给每个虚拟文件服务器，可以设定在一个文件服务器上动作的、作为虚拟服务单位的虚拟文件服务器，来使各虚拟文件服务器从属于不同的网络。依据这个技术，就可以用一个文件服务器对具有相同专用地址的多个网段提供个别服务（例如，参照专利文献1）。

另外，已知的故障切换功能是：在多个文件服务器之间，利用通信电路或共享的磁盘，定期将运行中的状态发送给对方、来相互监视对方的状态，当检测到了对方的故障时，接替对方的服务（例如，参照专利文献2）。

【专利文献1】美国专利申请公开第2003／0135578号说明书

【专利文献2】美国专利第6317844号说明书

在由具有上述的虚拟文件服务器的多个文件服务器构成了集群的场合，未曾考虑在文件服务器间移动虚拟文件服务器的情况。

另外，由于不能在文件服务器间移动虚拟文件服务器，所以，就不能进行以虚拟文件服务器为单位的负荷分散，在特定的文件服务器中往往会造成负荷集中。

本发明，其目的是：让虚拟文件服务器在由设定了虚拟文件服务器的多个文件服务器所构成的集群内动态移动。

发明内容

（用来解决课题的装置）

本发明，是配备有第 1 文件服务器、第 2 文件服务器和磁盘子系统的存储系统；上述磁盘子系统，具有存储由虚拟文件服务器访问的数据的磁盘驱动器、和将在上述磁盘驱动器中所存储的数据发送给上述文件服务器的磁盘控制部；上述各文件服务器，具有针对网络输入输出数据的网络接口；在上述各文件服务器中，设有：多个虚拟文件服务器、和控制虚拟文件服务器的起动、终止，并在上述文件服务器中设定虚拟文件服务器的虚拟文件服务器控制处理部；在上述虚拟文件服务器中，设有：利用上述网络接口的设定和上述网络接口，在与网络之间发送接收信号的网络处理部，和存储在与经由上述网络接口所连接的设备的通信中所需要的路由信息的路由表；在将上述第 1 文件服务器的虚拟文件服务器故障切换到了上述第 2 文件服务器后，在上述第 2 文件服务器中被起动的虚拟文件服务器，使用在上述第 1 文件服务器中相应虚拟文件服务器使用过的路由表来决定通信路径。

（发明的效果）

依据本发明，通过在由设定了虚拟文件服务器的多个装置构成的集群内使虚拟文件服务器动态地移动，可以只对发生了故障的虚拟文件服务器进行故障切换，可以进行虚拟文件服务器为单位的负荷分散。

附图说明

【图 1】是表示本发明的实施方式的存储系统的构成框图。

【图 2】是本发明的实施方式的存储系统的功能框图。

【图 3】是在本发明的实施方式的存储系统中所用的各种表的说明图。

【图 4】是本发明的实施方式的存储系统中的故障切换处理步骤说明图。

【图 5】是本发明的实施方式的存储系统中的虚拟文件服务器配置设定画面的说明图。

具体实施方式

下面，参照附图来说明本发明的实施方式。

图 1，是表示本发明的实施方式的存储系统的构成框图。

本发明的实施方式的存储系统 1，是由多个文件服务器 10、20 以及

磁盘子系统 40 构成，并构成了 NAS（Network Attached Storage：网络附加存储）。另外，由文件服务器 10 和文件服务器 20 构成了集群。

文件服务器 10，具有网络接口 11～13、CPU14、主内存 15、以及磁盘适配器 16，是由这些硬件构成的装置。

另外，文件服务器 10 中配备的资源（网络接口 11～13、CPU14、主内存 15、磁盘适配器 16），通过构成虚拟文件服务器 10a~10c 的程序在 CPU14 中动作，在文件服务器 10 内构成了独立动作的虚拟文件服务器。即，CPU14，通过执行主内存 15 中所存储的、构成虚拟文件服务器的程序，在文件服务器 10 内构筑多个虚拟文件服务器，并在虚拟文件服务器间共享资源（CPU14、主内存 15、磁盘适配器 16 等）。

网络接口 11～13，是针对客户（图中省略）的接口，例如，用 TCP / IP 等协议进行通信。再者，也可以是由光纤通道、或 iSCSI（internet SCSI）能通信的接口。另外，网络接口 11～13，还可以是被连接到在网络内所设定的虚拟分组的网络（VLAN：Virtual LAN）的接口。另外，网络接口 11～13，被连接到各个不同的网络 6a～6c（不同的网段）。

磁盘适配器 16，进行对于光纤通道等磁盘子系统 40 的协议处理。

文件服务器 10，通过磁盘适配器 16，可以访问磁盘子系统 40 内的 LU（逻辑单元），可以读写磁盘驱动器中所存储的数据。

再者，虽然就其文件服务器 10 进行了说明，但文件服务器 20 也具有同样的构成，文件服务器 10 和文件服务器 20 在物理上是由不同的硬件构成的。

文件服务器 10 和文件服务器 20，由集群间的通信路径 50 连接起来。集群间的通信路径 50，既可以是存储系统 1 内所设的 LAN 和无限带宽等通信路径，也可以是经由外部网络的通信路径。

文件服务器 10 和文件服务器 20，通过互相通知相互的状态，来监视相互的状态，并由文件服务器 10 和文件服务器 20 构成了集群。再者，不具有特定的集群间的通信路径、文件服务器 10 和文件服务器 20 共同使用磁盘超高速缓存和磁盘驱动器的特定区域，定期相互读写规定的数据，这样，也可以相互监视状态。

磁盘子系统 40，被连接到文件服务器 10，配备有：磁盘控制部 41、磁盘超高速缓存 42、以及磁盘驱动器 44a～44d。

磁盘控制部 41，接受来自文件服务器 10 的磁盘适配器 16 的数据输入输出请求，控制对磁盘驱动器 44a～44d 的数据的输入输出。

磁盘超高速缓存 42，临时存储从磁盘驱动器 44a～44d 读出的数据、和被写入到磁盘驱动器 44a～44d 的数据，提高针对客户的、存储系统 1 的访问性能。

在磁盘驱动器 44a～44d 中，设定有 OS 作为一个磁盘可以识别的单位的逻辑单元（LU）。另外，逻辑单元用 RAID（Redundant Array of Independent Disks：冗余独立磁盘阵列）构成，使所存储的数据具有冗余性。因此，即使磁盘驱动器 44a～44d 中一部发生了故障，所存储的数据也不会消失。

管理终端 2，是配备有 CPU、内存、存储装置以及网络接口的计算机装置，运行着用来进行集群和文件服务器的设定等管理程序。再者，也可以在每个虚拟服务器中设置管理终端。

网络 6a～6c，例如，是用 TCP / IP 等协议进行通信的网络。

图 2，是本发明的实施方式的存储系统的功能框图。

在文件服务器 10 中，设定有虚拟文件服务器 10a、虚拟文件服务器 10b、以及虚拟文件服务器 10c。下面，就虚拟文件服务器 10a 进行说明，而虚拟文件服务器 10b、10c 也具有同样的构成。

在虚拟文件服务器 10a 中，设置有网络处理部 105 以及虚拟文件服务器管理处理部 106。

网络处理部 105，用属于虚拟文件服务器的网络接口 11～13 的设定、以及这些网络接口在与网络之间接收发送数据和控制信号。另外，还进行有关网络文件系统（NFS：Network File System、CIFS：Common Internet File System 等）的处理。

虚拟文件服务器管理处理部 106，根据来自管理终端 2 的指示，进行每个虚拟文件服务器的设定（例如，网络的设定，文件系统的安装、用户管理等）。

另外，在虚拟文件服务器 10a 中，设有：网络接口信息 101、路由表 102、安装表 103、以及设备文件 104。

在网络接口信息 101 中，存储有由虚拟文件服务器 10a 可访问的网络接口 11～13 的控制信息（例如，决定所谓通信传送长度的通信协议的协议文件等）。

在路由表 102 中，存储有在与经由网络接口 11～13 被连接到网络 6a～6c 的设备的通信中所必要的路径信息。路由表 102，是被分离设在每个虚拟文件服务器中，所以，可以将同一文件服务器 10 的不同的虚拟文件服务器连接到使用相同 IP 地址的不同网段。

在安装表 103 中，存储有由虚拟文件服务器 10a 可访问的文件系统信息（安装点、设备名等）。

设备文件 104，是用来访问 LU 的文件。在请求向磁盘子系统 40 输入输出数据时，通过访问设备文件 104，起动在 OS 内核中所组装入的设备驱动程序，来实现对磁盘子系统 40 上的 LU 的访问。安装表 103 和设备文件 104，是被分离设在每个虚拟文件服务器中，所以，可以在每个虚拟文件服务器中进行起动、终止。

上面，就虚拟文件服务器 1 进行了说明，虚拟文件服务器 2 和虚拟文件服务器 3 也具有同样的构成。

再者，为了分离由各虚拟文件服务器所提供的文件系统，安装表 103 和设备文件 104 被分离设置到每个虚拟文件服务器中。但是，如果无此必要，也未必非要将安装表 103 和设备文件 104 分离设置到每个虚拟文件服务器中。

另外，安装表 103、设备文件 104 以及网络处理部 105 被分离设置到每个虚拟文件服务器中，但在不同的虚拟文件服务器可以访问相同的 LU 的场合，也可以设置由文件服务器能共通使用这些部分的公共处理部。

另外，在文件服务器 10 中，作为在各虚拟文件服务器 10a～10c 中共通的处理，设有：文件系统处理部 111、磁盘访问部 112、装置间故障监视处理部 113、虚拟文件服务器控制处理部 114、虚拟文件服务器

故障监视处理部 115、装置间同步处理部 116 以及装置管理处理部 117。这些各个部分，通过在 CPU14 中执行在主存储器 15 中所存储的程序来实现。

文件系统处理部 111，接收来自虚拟文件服务器 10a～10c 的请求，对磁盘访问部 112 等发出指示，处理对文件的访问。

磁盘访问部 112，接收来自文件系统处理部 111 等的请求，进行针对磁盘子系统的数据输入输出处理。

装置间故障监视处理部 113，定期地监视集群内的其他装置（文件服务器 20）的动作状态。而后，当检测出了其他文件服务器故障的场合，进行用来接替由该文件服务器正在进行的服务的处理，来进行文件服务器间的故障切换。

虚拟文件服务器控制处理部 114，进行虚拟文件服务器的起动和终止、定义和削除、资源的分配和削除等的、虚拟文件服务器 10a~10c 的控制。亦即，虚拟文件服务器控制处理部 114，在文件服务器 10 中设定虚拟文件服务器。

虚拟文件服务器故障监视处理部 115，监视在装置（文件服务器 10、20）内运行中的虚拟文件服务器 10a~10c 的运行状态，并检测发生了故障的虚拟文件服务器。

装置间同步处理部 116，由集群内的装置（文件服务器 10）间的通信、来控制虚拟文件服务器的起动、终止的定时，同步虚拟文件服务器的起动、终止。这个同步处理，进行被设置在硬件上不同的文件服务器中的虚拟文件服务器间的同步处理。

装置管理处理部 117，根据来自管理终端 2 的指示，进行文件服务器的管理。例如，通过集群的设定和网络（包括 VLAN）的构成设定来管理存储系统 1 的动作。另外，进行虚拟文件服务器的设定的变更、和装置间的虚拟文件服务器的移动指示等。进而，当发出了对虚拟文件服务器的操作指示时，指示虚拟文件服务器控制处理部 114 进行处理。

在磁盘子系统 40 中，设有逻辑单元（LU），并将每个虚拟文件服务器中使用的 LU 分离开来。例如，LU1 由虚拟文件服务器 1 访问，LU2

由虚拟文件服务器 2 访问，LU3 由虚拟文件服务器 3 访问。

再者，为了分离由各虚拟文件服务器所提供的文件系统，LU 被分离到每个虚拟文件服务器，但若无此必要，也未必非要分设到每个虚拟文件服务器。

在磁盘子系统 40 中，设有在文件服务器间所共享的系统 LU。在这个系统 LU 中，存储由各文件服务器的虚拟文件服务器可访问的、由虚拟文件服务器所使用的各种处理程序和数据。因此，系统 LU 是作为公用卷来工作的。

在系统 LU 中，存储有构成信息和服务状况文件。在构成信息中，记录着针对各虚拟文件系统的物理资源的分配信息。例如，网络接口 11～13 的信息、和针对虚拟文件服务器的 LU 的分配等。另外，在服务状况文件中，复制存储有：在虚拟文件服务器 10a 中所存储的信息中，安装表 103、路由表 102 以及设备文件 104 的信息。例如，记录有：用虚拟文件服务器进行什么样的服务，磁盘的安装、拆卸的信息，对磁盘访问的限制的信息等。

在共享这个构成信息和服务状况的方法中，除了使用上述的公用卷的方法之外，还可以使用这样的方法：当在内存中所存储的构成信息或服务状况中一旦有变更、通过集群间通信路径 50 将变更后的构成信息或服务状况通知给其他的文件服务器，这样，在文件服务器间就具有共通的信息。

图 3，是本发明的实施方式的存储系统中所使用的各种表的说明图。

安装表 103，被设在每个虚拟文件服务器中，记录有文件系统 ID、inode#、父文件系统、安装点、父 inode#以及设备名。

文件系统 ID，是在文件系统中被唯一决定的识别符。inode#，是从相应文件系统观察到该文件系统的根目录的编号。父文件系统，规定将该文件系统设在哪个文件的下级。安装点，是该文件系统被设置的场所，是该文件系统的根目录的路径名。父 inode#，是从上级文件系统观察到该文件系统的根目录的编号。亦即，在父目录 fs0 的 inode#＝20 中安装 fs1，在 fs1 的根目录中分配有 inode#＝200。就是说，fs1 的安装点，若

从 fs0 侧来看是 inode#＝20，若从 fs1 侧来看是 inode#＝200。设备名，是在该文件系统中所分配的设备名，是用来特别指定文件系统的。

在安装表 103 中，每当安装虚拟文件服务器 10a 中所分配的设备（被登录到了设备文件中的设备）时追加对应的表项。例如，每当在虚拟文件服务器中安装文件系统时追加表项。

安装表 103，被分离设置在每个虚拟文件服务器中。通过在每个虚拟文件服务器中分离设置安装表 103，可以让每个虚拟文件服务器有不同的目录结构。再者，当在虚拟文件服务器之间具有共同的目录结构的场合，就无需在每个虚拟文件服务器中分离设置安装表。

在设备文件 104 中，登录有设备名和设备 ID。设备 ID，是在对应于由该文件系统所使用的 LU 的存储装置内分配的唯一的编号。

设备文件 104，针对虚拟文件服务器，每当分配设备（磁盘）时，做成对应的表项。在安装设备时，通过使用在每个虚拟文件服务器中所分配的表，将来自虚拟文件服务器的访问、限制为只对该表中所分配的设备。再者，在虚拟文件服务器之间共享设备文件的场合，无需在每个虚拟文件服务器中分离设置设备文件。

路由表 102，用来决定经由网络进行通信时的数据包的传送路径，记录有：目的地址、网关（gateway）、网络掩码（mask）以及网络接口。亦即，对于 192.168.1.0～192.168.1.225 的目的地址，由网络接口 eth0 来传送数据包。另外，对于 192.168.2.0～192.168.2.225 的目的地，由网络接口 eth1 来传送数据包。另外，对于除此以外的目的地址，由网络接口 eth0、对于具有 192.168.1.1 的地址的网关传送数据包。

路由表 102，被分离设置在每个虚拟文件服务器中。通过在每个虚拟文件服务器中设置路由表 102，可以对每个虚拟文件服务器中不同的网络进行服务。亦即，在每个虚拟文件服务器中，由虚拟文件服务器进行发送的网络接口 11 等不同，路由表 102 不同，由此，可以与具有在不同网段中可能存在相同地址的装置进行通信。

下面，就本发明实施方式的存储系统的动作进行说明。

首先，在对文件进行访问之前，要在虚拟文件服务器中安装文件系

统。这个安装处理，是依据来自管理终端 2 的指示，由虚拟文件服务器管理处理部 106 来进行。这时，使用设备文件，在安装表中做成针对该设备的表项。

在对文件的访问中，必须对文件的名字进行解析。在对某个文件进行访问的场合，首先要从文件的名字、变换成可以特别指定该文件的识别符。例如，在文件系统内部，用 inode#识别文件。但是，在 NFS（Network File System）中，来自客户的请求使用被称之为文件称号（句柄）的识别符，但它们是一一对应的，所以，被视之为与 inode#是相同的。

在名字解析中，提供父目录的 inode#和文件名，求得针对该文件的 inode#。在网络处理部 105 中，在超出文件系统的安装点时，用安装表、将父目录的安装点的父 inode#变换成所安装的文件系统的 inode#，将该 inode#和文件名转交给文件系统处理部 111，这样，来进行名字的解析。

而后，用由名字解析所得到的识别符指定文件、来执行对文件的请求。网络处理部 105，一旦接收进行访问的 inode#和访问的种类，就将其发送到文件系统处理部 111。文件系统处理部 111，根据需要将请求发送给磁盘访问部 112、来访问磁盘。

在文件服务器 10 的场合，是经由网络从客户接受对文件的访问请求。究竟网络接口 11～13 中哪个接受对文件的访问请求，这要由虚拟文件服务器决定，所以，要将访问请求送到该虚拟文件服务器的网络处理部 105。而后，也要求将请求送到文件系统处理部 111。文件系统处理部 111，根据需要进行磁盘访问后生成响应，并将该响应送到网络处理部 105。而后，参照路由表 102 将该响应发送给客户。

图 4，是本发明的实施方式的存储系统中的故障切换处理步骤的说明图。

首先，虚拟文件服务器控制处理部 114，接受虚拟文件服务器 1 的故障切换指示。这个故障切换指示，当在虚拟文件服务器 1 发生了故障时，由虚拟文件服务器故障监视处理部 115 发布（1a）。作为这个虚拟文件服务器的故障，可以考虑：虚拟文件服务器的网络接口 11 等故障、或由网络的连接端的故障造成不能通信、或不能访问磁盘子系统 40 等

原因。

另外，也有以负荷分散为目的由管理终端 2 发布故障切换指示的。这种场合，虚拟文件服务器控制处理部 114，通过装置管理处理部 117，接受故障切换指示（1b）。

而后，虚拟文件服务器控制处理部 114，发布虚拟文件服务器 1 的终止指示（2）。而后，装置间同步处理部 116，将故障切换请求发送给构成集群的其他文件服务器（文件服务器 2）（3）。

而后，文件服务器 2 的虚拟文件服务器控制处理部 114，从系统 LU 读出构成信息和服务状况文件（4）。而后，虚拟文件服务器控制处理部 114，在文件服务器 2 中起动虚拟文件服务器 1（5）。

一旦起动虚拟文件服务器 1，文件服务器 2 的虚拟文件服务器管理处理部 106，依据来自虚拟文件服务器控制处理部 114 的指示，安装文件系统，开始对客户的服务。亦即，在文件服务器 2 中为了进行与文件服务器 1 同样的服务，在文件服务器 2 中所起动的虚拟文件服务器 1，必须要接替文件服务器 1 中的虚拟文件服务器 1 拥有的安装表 103、设备文件 104、和路由表 102 这些所谓的服务状况文件。为此，文件服务器 2，读出系统 LU 中所存储的这些表的内容。再者，也可以是这样的结构：不是如上所述那样、在故障切换时从磁盘子系统 40 的系统 LU 读出构成信息和服务状况，而是每当构成信息和服务状况有变更时、通过集群间通信路径 50 从文件服务器 1 将构成信息和服务状况发送到文件服务器 2，在故障切换时，使用文件服务器 2 通过集群间通信路径 50、预先从文件服务器 1 取得的构成信息和服务状况。

由此，在文件服务器 2 中所起动的虚拟文件服务器 1，就可以进行与在文件服务器 1 中动作的虚拟文件服务器 1 同样的业务。例如，在文件服务器 2 中所起动的虚拟文件服务器 1，使用与在文件服务器 1 中动作的虚拟文件服务器 1 同样的路由表，所以，可以在与以前同样的网段中决定同样的通信路径。

再者，文件服务器 1 中的虚拟文件服务器 1 的终结指示的发布（2），也可以是在文件服务器 2 中的虚拟文件服务器 1 的起动之后。但是，装

置间同步处理部 116，为了要在文件服务器 1 的服务终结之后才开始文件服务器 2 的服务，必须要让文件服务器 1 的动作和文件服务器 2 的动作同步。

图 5，是本发明的实施方式的存储系统中的虚拟文件服务器配置设定画面的说明图。

虚拟文件服务器的设定和移动的指示，由管理终端 2 通过装置管理处理部 117 来进行。针对用户的界面，是 GUI（Graphical User Interface：图形用户界面）或 CLI（Command Line Interface：命令行界面），哪个都行。

在图 5 所示的虚拟文件服务器配置设定画面中，在每个虚拟文件服务器中可以设定虚拟文件服务器（VS#）在集群内的哪个装置中动作。

这个画面中，装置＃表示构成集群的文件服务器（装置），VS#表示被分配的虚拟文件服务器。用户，在这个画面中，选择使虚拟文件服务器动作的文件服务器，每个虚拟文件服务器附加一个标记。而后，通过操作更新按钮反映变更，并将所设定的信息作为构成信息存储到系统 LU 中，虚拟文件服务器就动态地移动。

另外，在使用 CLI 的场合，设置将虚拟文件服务器和装置编号作为变量的命令，来规定虚拟文件服务器和文件服务器间的对应关系。例如，定义「vnasalloc（name）{装置＃}」这样的命令。

如上所说明过的那样，在本发明的实施方式中，

在将文件服务器 10 的虚拟文件服务器，故障切换到文件服务器 20 的时候，文件服务器 10 的虚拟文件服务器控制处理部 114，从系统 LU 读出构成信息、安装表的副本、路由表的副本、以及设备文件的副本，在文件服务器 20 中起动虚拟文件服务器，所以，可以在文件服务器 20 中起动与文件服务器 10 同样的虚拟文件服务器。另外，通过接替构成信息和服务信息、以及在文件服务器间取得同步来起动、终止虚拟文件服务器，即使是文件服务器在动作过程中，也可以变更使虚拟文件服务器动作的装置。

特别是，从系统 UL 读出路由表的副本，在文件服务器 20 中来起

动虚拟文件服务器，所以，在文件服务器 20 中所起动的虚拟文件服务器，可以使用相应读出的路由表来决定通信路径，可以在与以前同样的网段中决定同样的通信路径。

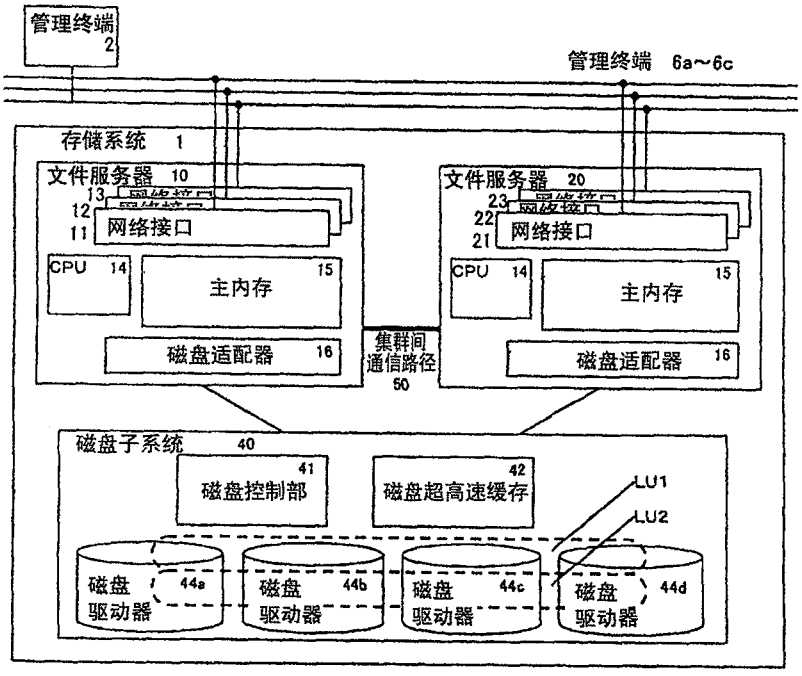

管理终端
2
管理终端 6a～6c
存储系统 1
文件服务器 10
13
12
11
网络接口
CPU 14
主内存 15
磁盘适配器 16
集群间
通信路径
50
文件服务器 20
23
22
21
网络接口
CPU 14
主内存 15
磁盘适配器 16
磁盘子系统 40
41
磁盘控制部
42
磁盘超高速缓存
LU1
LU2
磁盘
驱动器
44a
磁盘
驱动器
44b
磁盘
驱动器
44c
磁盘
驱动器
44d

图 1

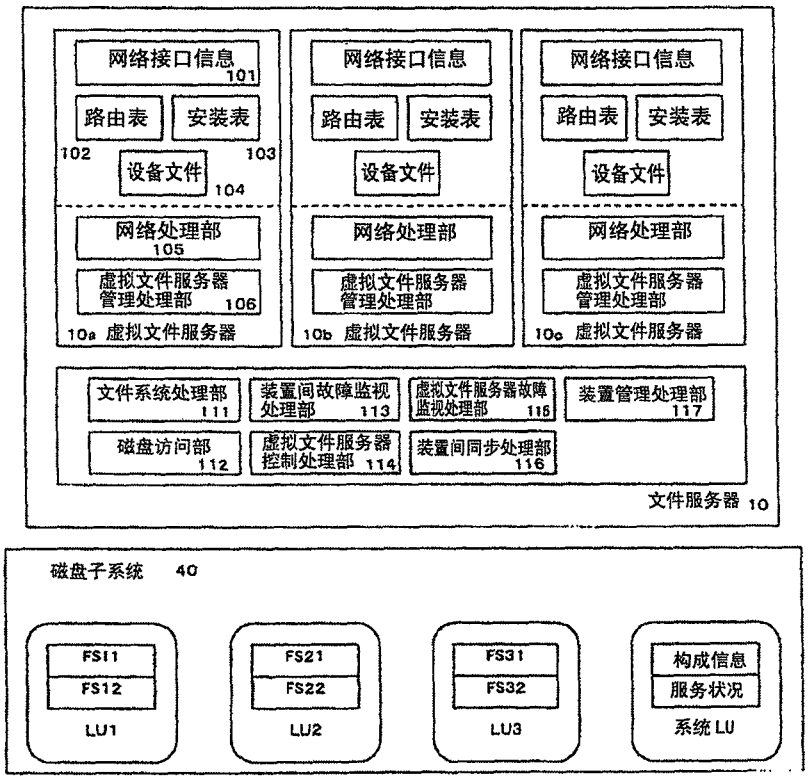

网络接口信息
101
路由表
安装表
102
103
设备文件
104
网络处理部
105
虚拟文件服务器
管理处理部
106
10a 虚拟文件服务器
网络接口信息
路由表
安装表
设备文件
网络处理部
虚拟文件服务器
管理处理部
10b 虚拟文件服务器
网络接口信息
路由表
安装表
设备文件
网络处理部
虚拟文件服务器
管理处理部
10c 虚拟文件服务器
文件系统处理部
111
装置间故障监视
处理部
113
虚拟文件服务器故障
监视处理部
115
装置管理处理部
117
磁盘访问部
112
虚拟文件服务器
控制处理部
114
装置间同步处理部
116
文件服务器 10
磁盘子系统 40
FS11
FS12
LU1
FS21
FS22
LU2
FS31
FS32
LU3
构成信息
服务状况
系统 LU

图 2

安装表 103

文件系统 ID	inode#	父文件系统	安装点	inode#	设备名
fs0	2	-	/	-	dev0
fs1	200	fs0	/export/fs1	20	dev1
fs2	500	fs0	/export/fs2	30	dev2

· 设备文件 104

设备名	设备 ID
dev0	8000
dev1	8001
dev2	8002

· 路由表 102

目的地	网关	掩码	网络接口
192.168.1.0	-	255.255.255.0	eth0
192.168.2.0	-	255.255.255.0	eth1
default	192.168.1.1	0.0.0.0	eth0

图 3

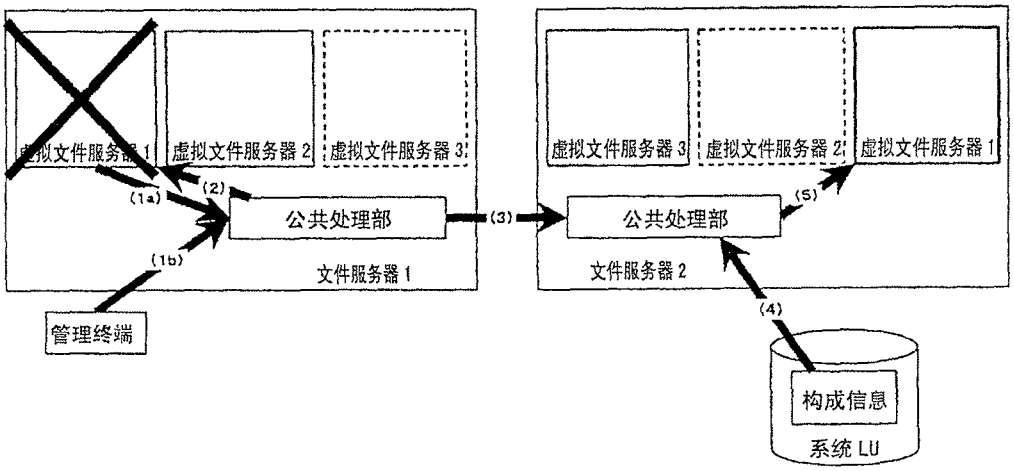


图 4

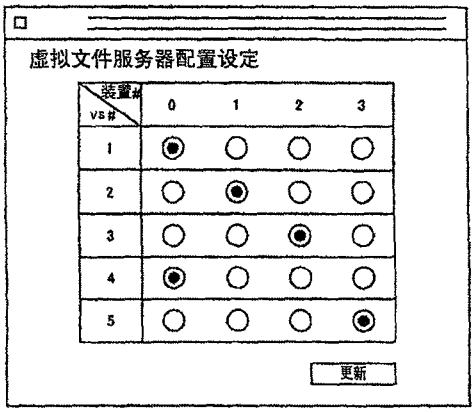


图 5