(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102170460 A
(43)申请公布日 2011.08.31

(21)申请号 201110057576.8

(22)申请日 2011.03.10

(71)申请人 浪潮(北京)电子信息产业有限公司
地址 100085 北京市海淀区上地信息路 2 号 2-1 号 C 栋 1 层

(72)发明人 刘家驹 张立强

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 栗若木 王漪

(51)Int.Cl.
H04L 29/08(2006.01)

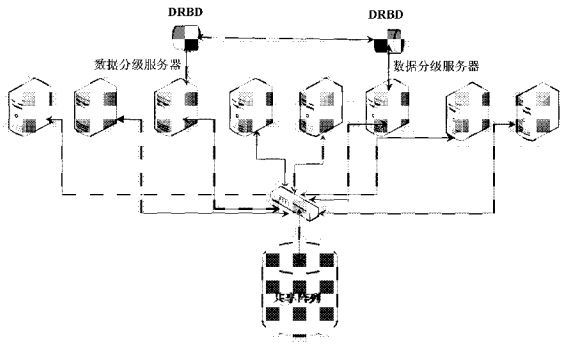
权利要求书 1 页 说明书 4 页 附图 1 页

(54)发明名称

一种集群存储系统及其数据存储方法

(57)摘要

本发明公开了一种集群存储系统及其数据存储方法，涉及集群存储系统。本发明公开的系统，包括共享存储设备，位于元节点的数据分级服务器和 DRBD，其中：所述数据分级服务器，确定所述共享存储设备中各文件的优先值，并将优先值大于设定值的文件的数据上传到所述 DRBD；所述 DRBD，接收所述数据分级服务器上传的文件的数据并存储。本发明实施例采用混合存储架构兼顾集中式存储低沉本大容量的优势和分布式存储高可靠性的优点，同时构建数据提取模型分类安放数据，便于数据管理，提高整个集群的容灾性，为电子信息系统的安全运行提供了有效保障。



CN 102170460 A

1. 一种集群存储系统，包括共享存储设备，其特征在于，该系统还包括位于元节点的数据分级服务器和分布式复制块设备 (DRBD)，其中：

所述数据分级服务器，确定所述共享存储设备中各文件的优先值，并将优先值大于设定值的文件的数据上传到所述 DRBD；

所述 DRBD，接收所述数据分级服务器上传的文件的数据并存储。

2. 如权利要求 1 所述的系统，其特征在于，

所述数据分级服务器确定所述共享存储设备中各文件的优先值指：

所述数据分级服务器将所述共享存储设备中文件的参数值的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

3. 如权利要求 1 所述的系统，其特征在于，

所述数据分级服务器确定所述共享存储设备中各文件的优先值指：

所述数据分级服务器为所述共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

4. 如权利要求 1、2 或 3 所述的系统，其特征在于，

所述集群存储系统中至少两个元节点上具有数据分级服务器和 DRBD。

5. 一种如权利要求 1 所述的集群存储系统的数据存储方法，其特征在于，该方法包括：

所述集群存储系统，确定共享存储设备中各文件的优先值，仅将优先值大于设定值的文件的数据存储到分布式复制块设备 (DRBD) 中。

6. 如权利要求 5 所述的方法，其特征在于，

所述集群存储系统确定所述共享存储设备中各文件的优先值指：

所述集群存储系统将所述共享存储设备中文件的参数值的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

7. 如权利要求 5 所述的方法，其特征在于，

所述集群存储系统确定所述共享存储设备中各文件的优先值指：

所述集群存储系统为所述共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

8. 如权利要求 5、6 或 7 所述的方法，其特征在于，

所述集群存储系统中至少两个元节点上具有 DRBD。

一种集群存储系统及其数据存储方法

技术领域

[0001] 本发明涉及集群存储系统，特别涉及一种集群存储系统及其数据存储方法。

背景技术

[0002] 目前，不少企事业单位所采用的传统 HA 架构存在一定缺陷，比如两台小型机搭配一个磁盘阵列，组成一套集群系统，所有的信息数据都存储在这台磁盘阵列上，存储只有一份，一旦此磁盘阵列发生问题，就面临整个业务系统停顿的危险，而采用分布式存储，虽然可以保证备份，但是磁盘利用率低下，且受成本限制容量受到限制。可见，要实现业务的高可用，必须先保证存储高可用；或者说，缺少高可用性存储的业务系统，不能实现真正的高可用性。针对这种情况，我们提出了存储高可用解决方案。

发明内容

[0003] 本发明所要解决的技术问题是，如何提高集群系统的容灾性。因此，提供一种集群存储系统及其数据存储方法。

[0004] 为了解决上述问题，本发明公开了一种集群存储系统，包括共享存储设备，位于元节点的数据分级服务器和 DRBD，其中：

[0005] 所述数据分级服务器，确定所述共享存储设备中各文件的优先值，并将优先值大于设定值的文件的数据上传到所述 DRBD；

[0006] 所述 DRBD，接收所述数据分级服务器上传的文件的数据并存储。

[0007] 较佳地，上述系统中，所述数据分级服务器确定所述共享存储设备中各文件的优先值指：

[0008] 所述数据分级服务器将所述共享存储设备中文件的参数值的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

[0009] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0010] 或者，上述系统中，所述数据分级服务器确定所述共享存储设备中各文件的优先值指：

[0011] 所述数据分级服务器为所述共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

[0012] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0013] 较佳地，上述集群存储系统中至少两个元节点上具有数据分级服务器和 DRBD。

[0014] 本发明还公开了一种如上所述的集群存储系统的数据存储方法，包括：

[0015] 所述集群存储系统，确定共享存储设备中各文件的优先值，仅将优先值大于设定值的文件的数据存储到分布式复制块设备 (DRBD) 中。

[0016] 较佳地，上述方法中，所述集群存储系统确定所述共享存储设备中各文件的优先值指：

[0017] 所述集群存储系统将所述共享存储设备中文件的参数值的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

[0018] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0019] 或者，上述方法中，所述集群存储系统确定所述共享存储设备中各文件的优先值指：

[0020] 所述集群存储系统为所述共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

[0021] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0022] 较佳地，上述集群存储系统中至少两个元节点上具有 DRBD。

[0023] 本发明实施例采用混合存储架构兼顾集中式存储低沉本大容量的优势和分布式存储高可靠性的优点，同时构建数据提取模型分类安放数据，便于数据管理，提高整个集群的容灾性，为电子信息系统的安全运行提供了有效保障。

附图说明

[0024] 图 1 为本实施例 1 中集群存储系统结构示意图。

具体实施方式

[0025] 下面结合附图及具体实施例对本发明技术方案做进一步详细说明。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互任意组合。

[0026] 目前，集群存储系统中有两种广泛采用的存储方式。其一是集中式存储方式，采用该方式，存储器成为单一失效节点。其二是分布式存储方式，采用该方式，存储器磁盘利用率太低，并且数据安放策略单一，不能进行有效管理。基于此，本发明申请人考虑到可采用混合存储架构兼顾集中式存储低沉本大容量的优势和分布式存储高可靠性的优点，同时构建数据提取模型，以便于数据管理，提高整个集群的容灾性。

[0027] 具体地，通过修改 /etc/multipath.conf 配置文件，实现集群中的各节点对共享存储设备的多路径访问和故障切换。即至少在两个以上的元节点上安装 DRBD 设备，实现通过网络通信来同步镜像整个设备，有点类似于一个网络 RAID 的功能。也就是说当用户将数据写入本地的 DRBD 设备上的文件系统时，数据会同时被发送到网络中的另外一台主机之上，并以完全相同的形式记录在一个文件系统中，从而达到分布式存储的效果。这样既可以满足海量数据的存储要求，也可以部分满足数据安全的要求，提高磁盘利用率并且平衡成本。

[0028] 实施例 1

[0029] 本实施例基于上述思想，提供一种集群存储系统，该系统架构如图 1 所示，包括位数据分级服务器、分布式复制块设备 (DRBD，Distributed ReplicatedBlock Device) 以及

共享存储设备，本实施例中共享存储设备选用共享阵列，共享阵列用来满足业务级 HA 的需求保证节点出现故障时服务不中断，DRBD 则用于满足存储级 HA 的要求，保证重要数据不丢失。从图 1 可以看到，所有节点都与共享阵列整列相连，两个元节点除与共享阵列相连外还安装有 DRBD。

[0030] 其中，数据分级服务器，位于两个元节点上，其主要负责为共享阵列中的文件构建数据提取模型以确定各文件的优先值，并将共享阵列中优先值大于设定值的文件的数据上传到 DRBD ；

[0031] 具体地，在客户调研的基础上，数据分级服务器将文件的参数值的和作为文件的优先值，其中，文件的参数值包括如下一种或几种 ：

[0032] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0033] 在优选方案中，数据分级服务器除了将文件的各参数值的和作为优先值时，还要考虑到各参数的权重，即为共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值。例如，将文件数据的大小值记为 x、文件数据的读取频率值记为 y、文件数据的修改频率值记为 z，文件对应的用户的等级值记为 v，之后建立数据提取模型，即确定文件的优先值如下 ：

[0034] ax+by+cz+dv ＝ f

[0035] 其中，a、b、c 和 d 为各参数的权值，可通过样本训练确定各参数的权值 ；

[0036] f 即为文件的优先值。

[0037] 另外，事关整个集群运行的关键信息也认为是大于设定值的文件的数据，也要上传到 DRBD。这样一旦集群崩溃或者阵列损坏，可以将数据损失的代价减少到最少，同时使集群在最短时间内恢复运转，达到提高容灾性的目标。

[0038] 而对于优先值小于设定值的文件的数据仍保留在共享阵列中。

[0039] DRBD，存储数据分级服务器上传的文件的数据。

[0040] 其中，为了提高集群存储系统的容灾性，一般 DRBD 位于元节点上。

[0041] 这样，光纤交换机可将共享存储设备（即本实施例中的共享阵列）和每个节点相连，设置 /etc/corosync/corosync.conf ；由 Pacemaker 建立起 active/active 模式的高可用集群，这样每个节点都成为潜在的备源节点，选择两台大内存服务器作为元节点，通过设置 DRBD 和配置文件，建立起 active/passive 模式的高可用集群，这样在一个集群里既有 active/active 模式又有 active/passive 模式，从而实现混合架构。

[0042] 共享阵列，存储优先值小于设定值的文件的数据。

[0043] 本实施例，在大量实验和抽样统计的基础之上，将文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值作为参数值，设计数据提取模型，编写程序，实现数据自动安置，从而达到提高容灾能力的要求，同时维护一份索引，便于查找数据，记录数据迁移状况。此外，将集群中重要的配置文件，设备信息备份在 DRBD 中，一旦集群崩溃可以迅速恢复。通过对用户授权，限制用户访问 DRBD 的权限，可提高这个集群的安全性。

[0044] 实施例 2

[0045] 本实施例基于上述集群存储系统，提出一种集群存储系统的数据存储方法，其核

心是对数据进行筛选，将重要数据（即优先值大于设定值的文件的数据）放在分布式复制块设备(DRBD,Distributed Replicated Block Device)中，将一般数据（即优先值小于设定值的文件的数据）置在共享存储设备（本实施例中即为共享阵列）中，这样即使共享阵列损坏，也可以将数据丢失的损失降到最低，并且DRBD中还会备份重要的系统信息（如服务器的配置文件，管理员信息，日志信息等等由管理员确定）当集群崩溃时即可快速恢复。

[0046] 具体地，该方法包括：集群存储系统确定共享存储设备中各文件的优先值，仅将优先值大于设定值的文件的数据存储到DRBD中。其中，DRBD一般位于各元节点上。

[0047] 具体地，集群存储系统确定共享存储设备中各文件的优先值指：

[0048] 将共享存储设备中文件的参数值的总和作为该文件的优先值，其中，文件的参数值包括如下一种或几种：

[0049] 文件数据的大小值、文件数据的读取频率值、文件数据的修改频率值、文件对应的用户的等级值。

[0050] 还有一些优选方案中，集群存储系统为共享存储设备中文件的参数值分别确定一权值，将各参数值与其对应的权值的乘积作为优先值计算参数，并将所有优先值计算参数的总和作为该文件的优先值。例如，将文件数据的大小值记为x、文件数据的读取频率值记为y、文件数据的修改频率值记为z，文件对应的用户的等级值记为v，之后建立数据提取模型，即确定文件的优先值如下：

[0051] $ax+by+cz+dv = f$

[0052] 其中，a、b、c和d为各参数的权值，可通过样本训练确定各参数的权值；

[0053] f即为文件的优先值。

[0054] 从上述实施例可以看出，本发明的实施例通过搭配使用DRBD和共享存储设备，对数据分类，分开存放，提高了整个系统的容灾能力。同时达到兼顾存储安全性和降低成本的目的。

[0055] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

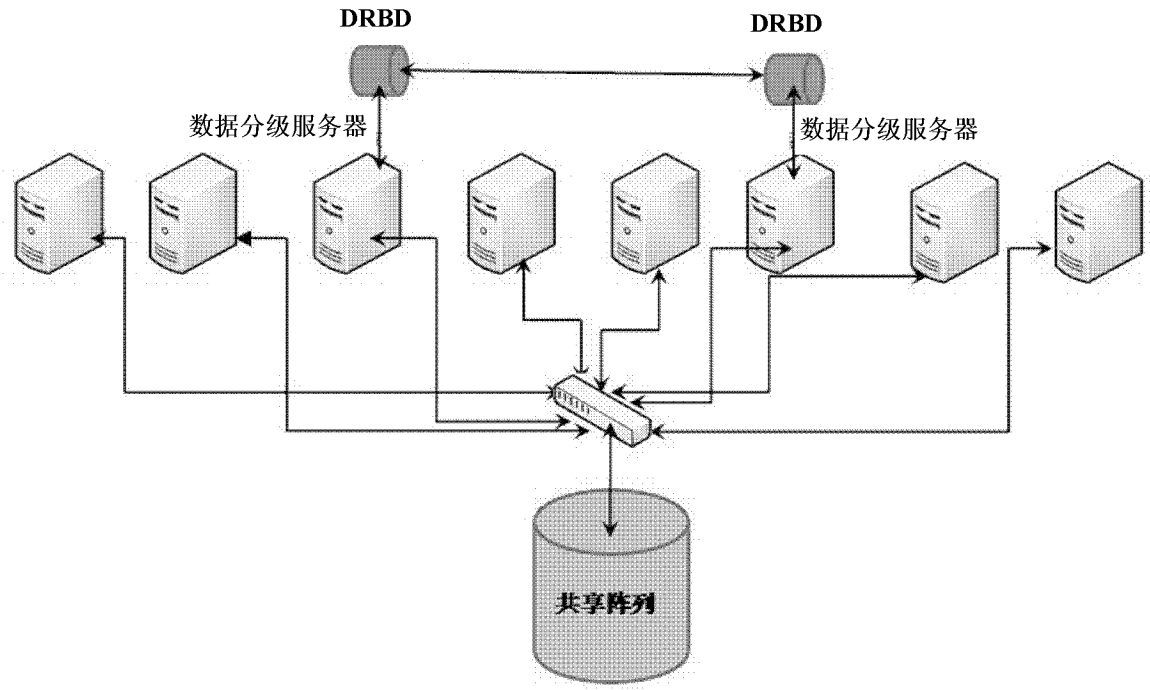


图 1