(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 103384267 B
(45)授权公告日 2017.09.01

(21)申请号 201310226210.8

(22)申请日 2013.06.07

(65)同一申请的已公布的文献号
申请公布号 CN 103384267 A

(43)申请公布日 2013.11.06

(73)专利权人 曙光信息产业(北京)有限公司
地址 100193 北京市海淀区东北旺西路8号中关村软件园36号

(72)发明人 刘冠川 秦东明 杨亮 曹振南 王勇 何牧君 张新风 陈飞 刘超 龚超 明立波 王慧 吕永安

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 12/24(2006.01)

(56)对比文件
CN 103095837 A,2013.05.08,
CN 102170460 A,2011.08.31,
陈嘉迅,刘晓洁.分布式块设备复制系统的分析与改进.《计算机工程与设计》.2012,第32卷(第11期),第3599~3601,3806页.
马艳军,等.集群文件系统lustre的介绍及应用.《科技信息》.2012,(第5期),第139~140页.

审查员 彭云

权利要求书1页 说明书3页

(54)发明名称

一种基于分布式块设备的Parastor200并行存储管理节点高可用方法

(57)摘要

本发明涉及一种基于分布式块设备的Parastor200管理节点高可用方法,所述方法通过以下两个方面实现:(1)管理节点存储系统信息的同步;(2)管理节点故障切换。本发明通过实现Parastor200管理节点的高可用使Parastor200实现了完全意义上的全冗余设计,系统中任何部件的损坏不影响存储系统的使用。管理节点任何部件的损坏,都可以在数秒内将服务切换到备用节点上。这样既不影响正常使用,又有充足的时间去修复故障。

CN 103384267 B

1.一种基于分布式块设备的Parastor200并行存储管理节点高可用方法，其特征在于，所述方法通过以下两个方面实现：

(1)管理节点存储系统信息文件的同步：采用分布式块设备实现；

(2)管理节点故障切换；

所述(1)中，所述管理节点存储系统信息同步是实现当管理节点上的存储系统信息发生变更时，主管理节点和备管理节点相应目录下的信息一致；

所述分布式块设备是一个用软件实现的、无共享的、服务器之间镜像块设备内容的存储复制解决方案；

当将数据写入本地主机分布式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远程主机之上，并以相同的形式记录在一个文件系统中；所述文件系统的创建是由分布式块设备的同步来实现的；

当远程主机和本地主机都返回写入成功时，整个数据写的过程才返回成功；当主管理节点出现故障时，备管理节点上保留有一份完全相同的数据；

所述(2)中，采用心跳机制判断故障管理节点，即通过在线管理节点和备用管理节点间连接心跳线发送信息和应答对方的监测，并通过ping第三方节点方式判断故障管理节点并自动实现故障切换；

在进行故障切换时结合资源和服务的迁移实现；所述资源和服务包括：

1)管理节点存储系统信息文件；

2)管理节点管理IP；

3)Parastor200管理服务以及Parastor200图形界面服务；

4)数据同步服务；

所述1)中，管理节点存储系统信息文件资源通过同步备份到备用管理节点上；

所述2)中，所述管理节点管理IP为管理节点向元数据节点、数据节点发送管理命令所经过的IP，所述管理节点管理IP在故障切换时从在线管理节点迁移到备用管理节点上；

所述3)中，所述Parastor200管理服务以及Parastor200图形界面服务在故障切换时，从在线管理节点切换到备用管理节点上；

所述4)中，切换后备用管理节点成为主管理节点，将备用管理节点的信息反过来备份到原来的主管理节点上。

一种基于分布式块设备的Parastor200并行存储管理节点高可用方法

技术领域

[0001] 本发明涉及一种基于分布式块设备的Parastor200并行存储管理节点高可用方法。

背景技术

[0002] ParaStor200并行存储系统采用了代表存储技术、网络通信技术以及数据管理技术发展方向的并行体系架构，是一款面向海量非结构化数据处理、拥有自主知识产权的高端存储系统。它可以提供TB/s级的高速带宽和EB级的海量存储空间，能够满足飞机汽车船舶设计、生物基因研究、材料科学研究、天气预报、地震监测、环境监测分析、能源勘探、电子商务、网络游戏、社交与视频分享网站建设、动漫渲染、视频编辑处理等领域中对于存储容量和I/O性能要求极高的应用，可广泛应用于政府、教育、科研、制造、企业、医疗、石油、广电、互联网等行业。

[0003] MGR表示Parastor200的管理节点，提供统一的控制管理界面，管理员通过该节点管理整个存储系统。

[0004] oPara表示Parastor200元数据节点，用于管理存储系统的所有索引数据和命名空间，对外提供单一的全局映像，支持多个节点以Active-Active集群模式工作。

[0005] oStor表示Parastor200数据节点，用于提供数据存储空间，内嵌高性能数据存取引擎，并行处理所有客户端的数据访问请求，支持多个oStor以副本方式(1-3个副本)容错。

[0006] Parastor200的管理节点，提供统一的控制管理界面，它保存着整个系统重要的拓扑结构及配置信息，管理员通过该节点管理整个存储系统。在整个存储系统中，管理节点的使用频度相对较低，只有当挂载客户端、查看存储系统状态、添加存储单元、删除存储单元等管理操作时才会用到管理节点。在小规模集群中通常管理较为简单，管理操作也比较少，此时管理节点的重要性相对较低，即使管理节点出现故障，我们也有充分的时间去修复管理节点，就算出现管理节点磁盘永久损坏也不至于出现灾难性后果，因为我们可以通过元数据节点、数据节点上的配置信息来重构管理节点上的重要信息。而丢失的只是一些历史数据和客户端授权信息，不会对存储系统造成太大的影响。目前，针对这一问题的解决办法是通过管理界面定期备份管理节点配置信息，当管理节点出现故障时，可以使用备用节点安装管理节点图形界面程序，然后导入备份的信息来完成。另外还有一种技术就是使用共享盘阵通过光纤交换机挂载到主、备管理节点上。主管理节点发生故障时，备管理节点通过挂载保存存储系统信息的分区获得存储管理节点的所有信息。

[0007] 现有方案有几个潜在的风险。首先，即便备份频率较高，但还是无法避免两次备份间系统配置被更改的可能。特别是进行了增加或者减少存储单元、更改客户端授权信息等操作，恢复后的信息和真实信息不一样，将会影响系统的正常运行。其次，即便没有任何信息丢失，重构一台管理节点耗费的时间还是比较长的，对于那些规模较大，用户较多，需要经常进行管理操作的系统显然是无法接受的。使用共享盘阵可以解决以上问题，但共享盘

阵的成本太高。

发明内容

[0008] 针对现有技术的不足，本发明提供一种基于分布式块设备的Parastor200并行存储管理节点高可用方法；本发明通过实现Parastor200管理节点的高可用使Parastor200实现了完全意义上的全冗余设计，系统中任何部件的损坏不影响存储系统的使用。管理节点任何部件的损坏，都可以在数秒内将服务切换到备管理节点上。这样既不影响正常使用，又有充足的时间去修复故障。使用分布式块设备技术能够在很小的成本的情况下实现真正的实时同步，保证主、备管理节点存储系统信息完全一致。

[0009] 本发明的目的是采用下述技术方案实现的：

[0010] 一种基于分布式块设备的Parastor200并行存储管理节点高可用方法，其改进之处在于，所述方法通过以下两个方面实现：

[0011] (1)管理节点存储系统信息文件的同步：采用分布式块设备实现。

[0012] (2)管理节点故障切换。

[0013] 其中，所述(1)中，所述管理节点存储系统信息同步是实现当管理节点上的存储系统信息发生变更时，主管理节点和备管理节点相应目录下的信息一致。

[0014] 其中，所述分布式块设备是一个用软件实现的、无共享的、服务器之间镜像块设备内容的存储复制解决方案；

[0015] 当将数据写入本地主机分布式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远程主机之上，并以相同的形式记录在一个文件系统中；所述文件系统的创建是由分布式块设备的同步来实现的；

[0016] 其中于，当远程主机和本地主机都返回写入成功时，整个数据写的过程才返回成功；当主管理节点出现故障时，备管理节点上保留有一份完全相同的数据。

[0017] 其中，所述(2)中，采用心跳机制判断故障管理节点，即通过在线管理节点和备用管理节点间连接心跳线发送信息和应答对方的监测，并通过ping第三方节点方式判断故障管理节点并自动实现故障切换。

[0018] 其中，在进行故障切换时结合资源和服务的迁移实现；所述资源和服务包括：

[0019] 1)管理节点存储系统信息文件；

[0020] 2)管理节点管理IP；

[0021] 3)Parastor200管理服务以及Parastor200图形界面服务；

[0022] 4)数据同步服务。

[0023] 其中，所述1)中，管理节点存储系统信息文件资源通过同步备份到备用管理节点上。

[0024] 其中，所述2)中，所述管理节点管理IP为管理节点向元数据节点、数据节点发送管理命令所走的IP，所述管理节点管理IP在故障切换时从在线管理节点迁移到备用管理节点上。

[0025] 其中，所述3)中，所述Parastor200管理服务以及Parastor200图形界面服务在故障切换时，从在线管理节点切换到备用管理节点上。

[0026] 其中，所述4)中，切换后备用管理节点成为主管理节点(主管理节点和备管理节点

是相对的概念，在线的管理节点即是主管理节点)，将备用管理节点的信息反过来备份到原来的主管理节点上。

[0027] 与现有技术比，本发明达到的有益效果是：

[0028] 本发明提供基于分布式块设备的Parastor200并行存储管理节点高可用方法，通过实现Parastor200管理节点的高可用使Parastor200实现了完全意义上的全冗余设计，系统中任何部件的损坏不影响存储系统的使用。管理节点任何部件的损坏，都可以在数秒内将服务切换到备管理节点上。这样既不影响正常使用，又有充足的时间去修复故障。使用分布式块设备技术能够在很小的成本的情况下实现真正的实时同步，保证主、备管理节点存储系统信息完全一致。

具体实施方式

[0029] 下面对本发明的具体实施方式作进一步的详细说明。

[0030] 本发明是要实现Parastor200管理节点的高可用。通过分析现有技术存在的问题我们便知道，本发明是要解决以下两个问题：(1)管理节点存储系统信息的同步；(2)管理节点故障切换。

[0031] 解决管理节点存储系统信息文件同步，就是要实现管理节点存储信息文件发生任何变更时，备用管理节点对应目录下的存储系统信息文件同时也发生变更，主备节点相应目录下的信息完全一致。本专利采用分布式块设备来解决这个问题。分布式块设备是一个用软件实现的、无共享的、服务器之间镜像块设备内容的存储复制解决方案。当你将数据写入本地的分布式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(实际上文件系统的创建也是由分布式块设备的同步来实现的)。当远程主机和本地主机都返回写入成功时整个写的过程才返回成功。因此本地节点与远程节点的数据可以保证实时的同步，并保证IO的一致性。所以当主管理节点出现故障时，备管理节点上还会保留有一份完全相同的数据，可以继续使用，以达到高可用目的。

[0032] 管理节点故障切换时，故障切换首先需要解决的问题就是如何判断故障，这里我们采用心跳机制，通过管理节点和备用管理节点间连接心跳线发送信息和应答对方的监测，并通过ping第三方节点等方式判断故障节点并自动实现故障切换。进行故障切换还需要解决一个重要的问题就是服务、资源的迁移。在本本发明中资源和服务包括：1)管理节点存储系统信息文件，这些资源已通过同步备份到备用节点上。2)管理节点管理IP，这个IP不同于两个节点间同步文件所走的网络的IP。它是管理节点向元数据节点、数据节点发送管理命令所走的IP。这个IP需要在故障切换时从主管理节点迁移到备用管理节点上。3)Parastor200管理服务以及Parastor200图形界面服务，这两个服务也在故障切换时，从管理节点切换到备用节点上。4)数据同步服务，即切换后备用管理节点成为了主管理节点，它需要将它上面的信息反过来备份到原来的主管理节点上。

[0033] 最后应当说明的是：以上实施例仅用以说明本发明的技术方案而非对其限制，尽管参照上述实施例对本发明进行了详细的说明，所属领域的普通技术人员应当理解：依然可以对本发明的具体实施方式进行修改或者等同替换，而未脱离本发明精神和范围的任何修改或者等同替换，其均应涵盖在本发明的权利要求范围当中。