(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 103503414 A
(43)申请公布日 2014.01.08

(21)申请号 201280005761.5

(22)申请日 2012.12.31

(85)PCT国际申请进入国家阶段日
2013.07.29

(86)PCT国际申请的申请数据
PCT/CN2012/088109 2012.12.31

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 顾炯炯 王道辉 闵小勇

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

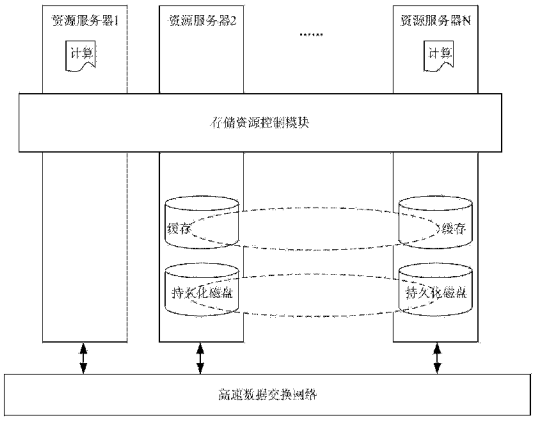
权利要求书2页 说明书11页 附图3页

(54)发明名称

一种计算存储融合的集群系统

(57)摘要

本发明实施例提供一种计算存储融合的集群系统，包括：资源服务器群，该资源服务器群包括至少两个资源服务器，该资源服务器群中的每个资源服务器至少具有计算资源和存储资源中的一种，并且该资源服务器群中的至少一个资源服务器具有计算资源和存储资源，该存储资源包括持久化磁盘以及和所述持久化磁盘对应的缓存；存储资源控制模块，用于将所述资源服务器群中的存储资源构建成共享的存储资源池，以提供存储服务，其中每个所述存储资源池包括的存储资源来自于所述资源服务器群中的至少两个资源服务器。



CN 103503414 A

1. 一种计算存储融合的集群系统，其特征在于，包括：

资源服务器群，所述资源服务器群包括至少两个资源服务器，所述资源服务器群中的每个资源服务器至少具有计算资源和存储资源中的一种，并且所述资源服务器群中的至少一个资源服务器具有计算资源和存储资源，所述存储资源包括持久化磁盘以及和所述持久化磁盘对应的缓存；

存储资源控制模块，用于将所述资源服务器群中的存储资源构建成共享的存储资源池，以提供存储服务，其中每个所述存储资源池包括的存储资源来自于所述资源服务器群中的至少两个资源服务器。

2. 根据权利要求 1 所述的计算存储融合的集群系统，其特征在于，所述存储资源控制模块，包括：

集群视图管理子模块，用于根据用户预设的配置信息对所述资源服务器群中的持久化磁盘进行逻辑划分以得到逻辑分区，所述每个逻辑分区都对应所述持久化磁盘上固定数量的物理存储资源块。

3. 根据权利要求 2 所述的计算存储融合的集群系统，其特征在于，所述存储资源控制模块，还包括数据子模块，所述资源服务器群中的每个具有存储资源的资源服务器都对应一个所述数据子模块；

所述集群视图管理子模块，还用于为所述数据子模块分配所述逻辑分区资源，建立所述数据子模块的逻辑分区拓扑结构，并根据所述数据子模块的请求，将分配的逻辑分区信息分别发给对应的所述数据子模块。

4. 根据权利要求 2 或 3 所述的计算存储融合的集群系统，其特征在于，所述存储资源控制模块，还包括接口子模块，所述资源服务器群中的每个资源服务器都对应一个所述接口子模块；

所述接口子模块，用于接收用户发送的用户卷创建请求，并根据所述用户卷创建请求，在所述共享的存储资源池中为需要创建的用户卷分配与所述用户卷创建请求相应的物理存储资源块，所述物理存储资源块来自于不同的持久化磁盘。

5. 根据权利要求 4 所述的计算存储融合的集群系统，其特征在于，所述接口子模块，还用于接收用户发送的读 / 写数据操作请求，根据所述读 / 写数据操作请求中包含的用户卷标识，计算出读 / 写数据操作对应的物理存储资源块以及所述物理存储资源块所属的逻辑分区；根据计算出的所述物理存储资源块所属的逻辑分区确定需要执行读 / 写数据操作的数据子模块；将所述读 / 写数据操作请求发送给所述需要执行读 / 写数据操作的数据子模块。

6. 根据权利要求 5 所述的计算存储融合的集群系统，其特征在于，所述数据子模块，用于根据所述接口子模块发送的写数据操作请求，通过所述数据子模块对应的资源服务器的缓存写入数据；或者，用于根据所述接口子模块发送的读数据操作请求，通过所述数据子模块对应的资源服务器的缓存读取数据。

7. 根据权利要求 6 所述的计算存储融合的集群系统，其特征在于，所述通过所述数据子模块对应的资源服务器的缓存写入数据，具体包括：

所述数据子模块先将数据写入所述数据子模块对应的资源服务器的缓存中，再由所述缓存将数据写入与所述用户卷标识对应的物理存储资源块中。

8. 根据权利要求 6 所述的计算存储融合的集群系统，其特征在于，所述通过所述数据子模块对应的资源服务器的缓存读取数据，具体包括：

所述数据子模块先从所述数据子模块对应的资源服务器的缓存中读取请求的数据，如果所述缓存中没有所述请求的数据，再从所述用户卷标识对应的物理存储资源块中读取所述请求的数据，并将读取到的所述请求的数据写入缓存中。

9. 根据权利要求 3 所述的计算存储融合的集群系统，其特征在于，如果所述集群系统增加或者删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于根据预设的逻辑分区分配算法重新生成所述数据子模块的逻辑分区拓扑结构，将新的逻辑分区拓扑结构广播给所述集群系统中所有的接口子模块和数据子模块，并通知原有的数据子模块基于所述新的逻辑分区拓扑结构发起数据迁移。

10. 根据权利要求 9 所述的计算存储融合的集群系统，其特征在于，如果所述集群系统删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于在所述数据子模块完成数据迁移后，将所述被删除的资源服务器设置为不可用状态。

11. 根据权利要求 4 或 5 所述的计算存储融合的集群系统，其特征在于，

所述接口子模块的接口符合小型计算机系统接口协议。

12. 根据权利要求 1 所述的计算存储融合的集群系统，其特征在于，所述存储资源控制模块，还用于将所述具有存储资源的资源服务器的单个持久化磁盘作为独立节点，构建所述共享的存储资源池。

13. 根据权利要求 1 至 11 中任一项所述的计算存储融合的集群系统，其特征在于，所述存储资源控制模块，还用于通过分布式算法构建并管理所述共享的存储资源池。

14. 根据权利要求 1 至 13 中任一项所述的计算存储融合的集群系统，其特征在于，所述集群系统中的资源服务器之间通过高速数据交换网络进行通信。

15. 根据权利要求 14 所述的计算存储融合的集群系统，其特征在于，所述高速数据交换网络包括高速外设组件互连标准 PCI-E 网络或光纤网络。

一种计算存储融合的集群系统

技术领域

[0001] 本发明涉及通信技术领域，尤其涉及一种计算存储融合的集群系统。

背景技术

[0002] 企业要进行信息化，购买的信息技术（Information Technology，简称 IT）基础设施一般包括服务器设备、网络设备和存储设备三大部分，然后企业自己需要将这三部分搭建成一个网络，进行各种配置，整个过程复杂，而且后续的扩容等操作也复杂。虚拟化技术的成熟和云概念的出现，促进了一体机诞生。为减少组建网络和后续维护 IT 设备的复杂性，主要 IT 设备厂家（例如 IBM/DELL/HP/CISCO 等）纷纷推出了“一体机”产品，即，在一个机架内预集成服务器、存储和网络。客户购买了这种一体机后，无需进行复杂的硬件安装；更进一步，无需进行复杂的软件安装和配置，插上电以后，简单的进行配置（例如配置外网 IP），就可以使用了（典型的一体机架构示意图如图 1 所示）。

[0003] 从已经存在的各种一体机来看，都有如下两个共同的目标：（1）降低成本，保证优于个人计算机（Personal Computer，简称 PC）的性价比；（2）优化架构，保证不差于 PC 的性能。为实现这两个目标，高效使用存储是个关键。

[0004] 一体机中的存储设备，一般使用的是存储区域网络（Storage AreaNetwork，简称 SAN）或网络附加存储（Network Attached Storage，简称 NAS）等专用存储设备。虽然，SAN 或 NAS 等专用存储设备作为独立的设备已经有很长时间了，但随着云计算、一体机的出现，在一体机中使用 SAN 或 NAS 专用存储设备也暴露了其固有的缺点：专用 SAN 或 NAS 需要进行复杂配置，后续维护困难；专用 SAN 或 NAS 需要控制机头，性价比不高，无法满足用户低成本要求；专用 SAN 或 NAS 受制于控制机头，横向扩展受限，无法线性扩展而满足大量数据突发的查询性能要求。这些缺点使得 SAN 或 NAS 设备成了一体机中的薄弱环节。

发明内容

[0005] 本发明实施例提供一种计算存储融合的集群系统，用以解决现有技术中因为使用专用 SAN 而导致的操作复杂、成本较高以及由于无法线性扩展而不能满足大量数据突发的查询性能要求的问题。

[0006] 第一方面，本发明实施例提供一种计算存储融合的集群系统，包括：

[0007] 资源服务器群，所述资源服务器群包括至少两个资源服务器，所述资源服务器群中的每个资源服务器至少具有计算资源和存储资源中的一种，并且所述资源服务器群中的至少一个资源服务器具有计算资源和存储资源，所述存储资源包括持久化磁盘以及和所述持久化磁盘对应的缓存；

[0008] 存储资源控制模块，用于将所述资源服务器群中的存储资源构建成共享的存储资源池，以提供存储服务，其中每个所述存储资源池包括的存储资源来自于所述资源服务器群中的至少两个资源服务器。

[0009] 结合第一方面，在第一方面的第一种可能的实现方式中，所述存储资源控制模块，

包括：

[0010] 集群视图管理子模块，用于根据用户预设的配置信息对所述资源服务器群中的持久化磁盘进行逻辑划分以得到逻辑分区，所述每个逻辑分区都对应所述持久化磁盘上固定数量的物理存储资源块。

[0011] 结合第一方面的第一种可能的实现方式，在第二种可能的实现方式中，所述存储资源控制模块，还包括数据子模块，所述资源服务器群中的每个具有存储资源的资源服务器都对应一个所述数据子模块；

[0012] 所述集群视图管理子模块，还用于为所述数据子模块分配所述逻辑分区资源，建立所述数据子模块的逻辑分区拓扑结构，并根据所述数据子模块的请求，将分配的逻辑分区信息分别发给对应的所述数据子模块。

[0013] 结合第一方面的第一种可能的实现方式以及第一方面的第二种可能的实现方式，在第三种可能的实现方式中，所述存储资源控制模块，还包括接口子模块，所述资源服务器群中的每个资源服务器都对应一个所述接口子模块；

[0014] 所述接口子模块，用于接收用户发送的用户卷创建请求，并根据所述用户卷创建请求，在所述共享的存储资源池中为需要创建的用户卷分配与所述用户卷创建请求相应的物理存储资源块，所述物理存储资源块来自于不同的持久化磁盘。

[0015] 结合第一方面的第三种可能的实现方式，在第四种可能的实现方式中，所述接口子模块，还用于接收用户发送的读/写数据操作请求，根据所述读/写数据操作请求中包含的用户卷标识，计算出读/写数据操作对应的物理存储资源块以及所述物理存储资源块所属的逻辑分区；根据计算出的所述物理存储资源块所属的逻辑分区确定需要执行读/写数据操作的数据子模块；将所述读/写数据操作请求发送给所述需要执行读/写数据操作的数据子模块。

[0016] 结合第一方面的第四种可能的实现方式，在第五种可能的实现方式中，所述数据子模块，用于根据所述接口子模块发送的写数据操作请求，通过所述数据子模块对应的资源服务器的缓存写入数据；或者，用于根据所述接口子模块发送的读数据操作请求，通过所述数据子模块对应的资源服务器的缓存读取数据。

[0017] 结合第一方面的第五种可能的实现方式，在第六种可能的实现方式中，所述通过所述数据子模块对应的资源服务器的缓存写入数据，具体包括：所述数据子模块先将数据写入所述数据子模块对应的资源服务器的缓存中，再由所述缓存将数据写入与所述用户卷标识对应的物理存储资源块中。

[0018] 结合第一方面的第五种可能的实现方式，在第七种可能的实现方式中，所述通过所述数据子模块对应的资源服务器的缓存读取数据，具体包括：所述数据子模块先从所述数据子模块对应的资源服务器的缓存中读取请求的数据，如果所述缓存中没有所述请求的数据，再从所述用户卷标识对应的物理存储资源块中读取所述请求的数据，并将读取到的所述请求的数据写入缓存中。

[0019] 结合第一方面的第二种可能的实现方式，在第八种可能的实现方式中，如果所述集群系统增加或者删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于根据预设的逻辑分区分配算法重新生成所述数据子模块的逻辑分区拓扑结构，将新的逻辑分区拓扑结构广播给所述集群系统中所有的接口子模块和数据子模块，并通知原有的数据子

模块基于所述新的逻辑分区拓扑结构发起数据迁移。

[0020] 结合第一方面的第八种可能的实现方式，在第九种可能的实现方式中，如果所述集群系统删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于在所述数据子模块完成数据迁移后，将所述被删除的资源服务器设置为不可用状态。

[0021] 结合第一方面的第三种可能的实现方式以及第一方面的第四种可能的实现方式，在第十种可能的实现方式中，所述接口子模块的接口符合小型计算机系统接口协议。

[0022] 结合第一方面，在第十一种可能的实现方式中，所述存储资源控制模块，还用于将所述具有存储资源的资源服务器的单个持久化磁盘作为独立节点，构建所述共享的存储资源池。

[0023] 结合第一方面，以及第一方面的任意一种可能的实现方式，在第十二种可能的实现方式中，所述存储资源控制模块，还用于通过分布式算法构建并管理所述共享的存储资源池。

[0024] 结合第一方面，以及第一方面的任意一种可能的实现方式，在第十三种可能的实现方式中，所述集群系统中的资源服务器之间通过高速数据交换网络进行通信。

[0025] 结合第一方面的第十三种可能的实现方式，在第十四种可能的实现方式中，所述高速交换网络包括高速外设组件互连标准 PCI-E 网络或光纤网络。

[0026] 由上述技术方案可知，通过本发明实施例提供计算存储融合的集群系统，由于不存在专用的 SAN，省略了对 SAN 存储系统的复杂管理，在硬件上解决了现有技术中因为使用专用 SAN 而导致的操作复杂、成本较高的问题；存储设备可以有多个，每个存储设备上都可以部署缓存，在硬件上极大的提升了存储端缓存的扩展能力；存储资源不依赖于计算资源，存储资源可以独立的增加和减少，增强了系统的可扩展性；将系统中的持久化磁盘、缓存资源虚拟化为共享资源池并被所有计算共享，数据读写时所有计算和存储都可以参与，通过并发性的提高而提升了系统的存储性能。另外，由于本发明实施例提供计算存储融合的集群系统采用高速数据交换网络进行通信，进一步加快了数据的交换速度。

附图说明

[0027] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例。

[0028] 图 1 为现有技术中典型的一体机架构示意图；

[0029] 图 2 为本发明一实施例提供的计算存储融合的集群系统的示意性框图；

[0030] 图 3 为本发明一实施例提供的计算存储融合的集群系统的结构示意图；

[0031] 图 4 为本发明一实施例提供的磁盘分区和用户卷构成示意图；

[0032] 图 5 为本发明一实施例提供的计算存储融合的集群系统中的存储资源控制模块的结构示意图；

[0033] 图 6 为本发明一实施例提供的 MDC 模块为 OSD 节点分配分区资源的拓扑示意图。

具体实施方式

[0034] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例

中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。

[0035] 另外，本文中术语“和 / 或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和 / 或 B，可以表示 ：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0036] 本发明实施例提供的技术方案将计算资源与存储资源垂直整合，将一个集群系统中资源服务器的存储资源整合起来，通过存储资源控制模块，使用分布式算法，将各个资源服务器中的存储资源（持久化存储资源、缓存资源）虚拟化为资源池，供所有资源服务器的计算共享使用。资源服务器之间通过高速数据交换网络进行数据通信。

[0037] 图 2 示出了根据本发明实施例的计算存储融合的集群系统 100 的示意性框图。如图 1 所示，该计算存储融合的集群系统 100 包括 ：

[0038] 资源服务器群 110，该资源服务器群包括至少两个资源服务器，该资源服务器群中的每个资源服务器至少具有计算资源和存储资源中的一种，并且该资源服务器群中的至少一个资源服务器具有计算资源和存储资源，所述存储资源包括持久化磁盘以及和所述持久化磁盘对应的缓存 ；

[0039] 存储资源控制模块 120，用于将所述资源服务器群中的存储资源构建成共享的存储资源池，以提供存储服务，其中每个所述存储资源池包括的存储资源来自于所述资源服务器群中的至少两个资源服务器。

[0040] 在本发明实施例中，计算存储融合的集群系统中提供计算资源和存储资源的设备来自于资源服务器群，该资源服务器群由两个或两个以上的资源服务器组成，每个资源服务器能够提供计算资源和 / 或存储资源，并且至少有一个资源服务器能够同时提供计算资源和存储资源 ；计算存储融合的集群系统中的存储资源控制模块通过对共享的一个或多个存储资源池进行管理，从而提供虚拟存储服务，该共享的存储资源池由该资源服务器群中的存储资源构建成，并且每个存储资源池包括的存储资源来自于该资源服务器群中的两个或两个以上的资源服务器。

[0041] 在本发明实施例中，由于资源服务器群中的每个资源服务器具有计算资源和存储资源中的至少一种资源，并且该资源服务器群中的至少一个资源服务器具有计算资源和存储资源，即系统的存储资源和计算资源联合部署，因此能够节省设备投入成本、设备占用空间成本以及设备能源消耗成本 ；并且由于存储资源池由资源服务器群中的至少两个资源服务器的存储资源构成，因此能够均衡各资源服务器的存储资源，提高存储资源的利用效率和可靠性。

[0042] 因此，本发明实施例提供的计算存储融合的集群系统，由于不存在专用的 SAN，省略了对 SAN 存储系统的复杂管理，在硬件上解决了现有技术中因为使用专用 SAN 而导致的操作复杂、成本较高的问题 ；又由于存储设备可以有多个，每个存储设备上都包括持久化磁盘和缓存，在硬件上极大的提升了存储端缓存的扩展能力 ；存储资源不依赖于计算资源，存储资源可以独立的增加和减少，增强了系统的可扩展性 ；将系统中的持久化磁盘、缓存资源虚拟化为共享资源池并被所有计算共享，数据读写时所有计算和存储都可以参与，通过并发性的提高而显著提升了系统的存储性能。

[0043] 在本发明实施例中，存储资源控制模块可以部署在资源服务器上，也可以单独部

署。

[0044] 在本发明实施例中，可选地，该存储资源控制模块 120 还用于通过分布式算法创建并管理该存储资源池。例如，存储资源控制模块通过分布式哈希表(Distributed Hash Table，简称为“DHT”)方法、伪随机算法等分布式算法，创建并管理存储资源池，使得系统能够提供统一共享的存储服务。

[0045] 可选的，所述存储资源控制模块 120 还可以将所述具有存储资源的资源服务器的单个持久化磁盘作为独立节点，构建所述共享的存储资源池。

[0046] 在本发明实施例中，计算资源和存储资源采用统一的硬件服务器架构，合一部署，从而能够充分地利用物理服务器的计算能力和存储能力。即系统包括的每个资源服务器可同时提供计算资源和 / 或存储资源。在部署形态上，提供存储资源和 / 或计算资源的资源服务器，在物理部署形态上为同一物理设备，即存储资源和计算资源合一部署在一台服务器上，而计算资源和存储资源在逻辑上相互独立，可以分别构建自己的资源池。

[0047] 进一步，为了使计算和存储深度融合，本发明实施例提供的计算存储融合的集群系统可以采用纯高速外设组件互连标准(Peripheral Component Interconnection Express，简称 PCI-E) 交换架构。基于该 PCI-E 交换架构，包括计算、存储所有的资源服务器都可以通过 PCI-E 接口直接互连进行无阻塞交换，使集群系统中的计算和存储之间的通道更快速。

[0048] 具体地，由于本发明实施例提供的计算存储融合的集群系统的各部分都可以采用纯 PCI-E 交换，不管是计算设备、存储设备，还是直接的磁盘组，所有的单元都可以直接挂接到 PCI-E 交换网络下，无需进行不同协议之间的适配，因而进一步加快了数据的交换速度。

[0049] 应理解，上述 PCI-E 网络只是本发明实施例采用的高速数据交换网络的一种可能实现方式，本发明实施例并不对此进行限定。例如，实际应用中，也可以使用光纤网络作为该集群系统的高速数据交换网络。

[0050] 图 3 为本发明一实施例提供的计算存储融合的集群系统的结构示意图。如图 3 所示，本实施例的计算存储融合的集群系统由高速数据交换网络、资源服务器群以及存储资源控制模块组成，该资源服务器群包括三个资源服务器，即资源服务器 1、资源服务器 2 和资源服务器 3。其中，资源服务器 1 只具有计算资源，资源服务器 2 只具有存储资源，资源服务器 3 既具有计算资源又具有存储资源。资源服务器 2 和资源服务器 3 的存储资源被构建成统一的共享存储资源池，该共享存储资源池包括共享的缓存资源和共享的持久化磁盘。

[0051] 需要说明的是：图 3 仅仅为该计算存储融合的集群系统部署的一种示例，实际应用中，既可以部署更多的资源服务器，也可以在每个资源服务器上同时部署计算资源和存储资源。持久化磁盘可以为硬盘驱动器(Hard Disk Driver，简称 HDD) 或者固态硬盘(Solid State Disk，简称 SSD)。

[0052] 通过这种计算存储融合的集群系统，可以带来以下好处：存储设备可以有多个，每个存储设备上都可以部署缓存，在硬件上极大的提升了存储端缓存的扩展能力；存储资源不依赖于计算资源，存储资源可以独立的增加和减少，增强了系统的可扩展性。

[0053] 进一步，基于图 3 所示的计算存储融合的集群系统架构，由于去掉了专用 SAN，将存储资源分散到了各个资源服务器上，需要通过存储资源控制模块对这些分散的存储资源

进行统一的管理。具体包括：

[0054] (1)集群系统初始化时，将系统中的磁盘按照固定大小块 Block (例如 1M)进行划分(如图 4 中对每个 DISK 的划分)；

[0055] (2)通过存储资源控制模块将一定数量的 Block 组成一个分区 (Partition，简称为 PT，参见图 4 中标有 P 的方块)，如图 4 所示，每个分区由 4 个 Block 组成，该分区为逻辑分区，对用户而言是不可见的；

[0056] (3)当用户创建卷(Volume)时，资源存储控制模块负责指定所属该卷的 Block，如图 4 示例：Block2、Block4、Block5、Block7 属于 Volume1，Block6、Block8、Block9、Block12 属于 Volume2，Block11、Block13、Block14、Block16 属于 Volume3，卷对用户而言是可见的；

[0057] (4)用户对卷进行数据读写时，资源存储控制模块根据自身的算法，计算出所要读写操作的 Block，并根据某种算法(如 Hash 算法)计算出每个 Block 所对应的 Partition，通过 Partition 完成对数据的读写操作。

[0058] 这样，对于一个用户卷来说，其实际的存储物理资源是分布在很多磁盘上的。这样就达到了将不同的磁盘资源共享给一个用户卷，供用户使用，即达到存储资源共享使用的目的。当用户对卷进行数据读写时，读写请求经过存储资源控制模块的处理后，将转换为对卷中不同 Partition 进行分布式读写。

[0059] 需要说明的是：在集群系统启动前，用户会根据自身的需要制作集群系统的配置文件，该配置文件规划了集群系统中存储资源控制模块的部署、集群系统的分区规格(如 Partition 个数)以及不同资源服务器间的互相通信地址等信息。

[0060] 如图 5 所示，为了实现上述功能，所述存储资源控制模块包括：

[0061] 集群视图管理子模块，用于根据用户预设的配置信息对所述资源服务器群中的持久化磁盘进行逻辑划分以得到逻辑分区，所述每个逻辑分区都对应所述持久化磁盘上固定数量的物理存储资源块。

[0062] 进一步的，所述存储资源控制模块，还包括数据子模块，所述资源服务器群中的每个具有存储资源的资源服务器都对应一个所述数据子模块；

[0063] 应理解，在实际应用中，可以在每一个具有存储资源的资源服务器上部署所述数据子模块。

[0064] 相应的，所述集群视图管理子模块，还用于为所述数据子模块分配所述逻辑分区资源，建立所述数据子模块的逻辑分区拓扑结构，并根据所述数据子模块的请求，将分配的逻辑分区信息分别发给对应的所述数据子模块。

[0065] 进一步的，所述存储资源控制模块，还包括接口子模块，所述资源服务器群中的每个资源服务器都对应一个所述接口子模块；

[0066] 应理解，在实际应用中，可以在每一个资源服务器上部署所述接口子模块。其中，所述接口子模块可以采用符合小型计算机系统接口协议的接口。

[0067] 所述接口子模块，用于接收用户发送的用户卷创建请求，并根据所述用户卷创建请求，在所述共享的存储资源池中为需要创建的用户卷分配与所述用户卷创建请求相应的物理存储资源块，所述物理存储资源块来自于不同的持久化磁盘。

[0068] 进一步的，所述接口子模块，还用于接收用户发送的读 / 写数据操作请求，根据所述读 / 写数据操作请求中包含的用户卷标识，计算出读 / 写数据操作对应的物理存储资源

块以及所述物理存储资源块所属的逻辑分区；根据计算出的所述物理存储资源块所属的逻辑分区确定需要执行读 / 写数据操作的数据子模块；将所述读 / 写数据操作请求发送给所述需要执行读 / 写数据操作的数据子模块。

[0069] 相应的，所述数据子模块，用于根据所述接口子模块发送的写数据操作请求，通过所述数据子模块对应的资源服务器的缓存写入数据；或者，用于根据所述接口子模块发送的读数据操作请求，通过所述数据子模块对应的资源服务器的缓存读取数据。

[0070] 具体地，所述通过所述数据子模块对应的资源服务器的缓存写入数据，包括：

[0071] 所述数据子模块先将数据写入所述数据子模块对应的资源服务器的缓存中，再由所述缓存将数据写入与所述用户卷标识对应的物理存储资源块中。

[0072] 具体地，所述通过所述数据子模块对应的资源服务器的缓存读取数据，包括：

[0073] 所述数据子模块先从所述数据子模块对应的资源服务器的缓存中读取请求的数据，如果所述缓存中没有所述请求的数据，再从所述用户卷标识对应的物理存储资源块中读取所述请求的数据，并将读取到的所述请求的数据写入缓存中。

[0074] 可选的，如果所述集群系统增加或者删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于根据预设的逻辑分区分配算法重新生成所述数据子模块的逻辑分区拓扑结构，将新的逻辑分区拓扑结构广播给所述集群系统中所有的接口子模块和数据子模块，并通知原有的数据子模块基于所述新的逻辑分区拓扑结构发起数据迁移。

[0075] 如果所述集群系统删除具有存储资源的资源服务器，所述集群视图管理子模块，还用于在所述数据子模块完成数据迁移后，将所述被删除的资源服务器设置为不可用状态。

[0076] 基于本发明实施例提供计算存储融合的集群系统，由于不存在专用的 SAN，省略了对 SAN 存储系统的复杂管理，在硬件上解决了现有技术中因为使用专用 SAN 而导致的操作复杂、成本较高的问题；存储节点可以有多个，每个存储节点上都可以部署 Cache，在硬件上极大的提升了存储端 Cache 的扩展能力；存储节点不依赖于计算节点，存储节点可以独立的增加和减少，增强了系统的可扩展性。另外，由于本发明实施例提供的一体机系统的各部分都是采用纯 PCI-E 交换，不管是计算节点、存储节点，还是直接的磁盘组，所有的单元都是直接挂接到 PCI-E 交换下，因而无需进行不同协议之间的适配，进一步加快了数据的交换速度。

[0077] 作为本发明实施例提供的计算存储融合的集群系统的一种具体应用，所述系统中的存储资源控制模块的功能可以通过一种分布式存储控制软件实现。

[0078] 为了更清楚地理解本发明实施例的技术方案，下面将以该分布式存储控制软件为例对本发明实施例的技术方案做进一步的说明。

[0079] 具体地，该分布式存储控制软件主要包括三个功能模块：Meta Data Controller（简称 MDC）模块，Virtualization Block Service（简称 VBS）模块和 Object Storage Device（简称 OSD）模块。其中：

[0080] MDC 模块，主要用于实现分布式集群系统的状态视图控制，以及当资源服务器加入、退出集群系统时进行的输入 / 输出（Input/Output，简称 I/O）视图、分区分配视图、节点视图的更新；同时，还对数据分布式规则和数据重建规则进行控制；

[0081] VBS 模块，主要用于实现基于小型计算机系统接口（Small Computer System

Interface，简称 SCSI）的块设备的访问接口，同时完成块存储元数据的保存和访问逻辑；另外，VBS 模块还接受 MDC 模块下发的 I/O 视图，然后根据视图规则，将数据转发到相应的 OSD 节点（即，部署了 OSD 模块的资源服务器）上，其中，I/O 视图可以由 MDC 模块主动下发给 VBS 模块，也可以由 VBS 模块主动从 MDC 模块上获取；

[0082] OSD 模块，主要用于实现读写缓存功能，以及数据的一致性备份，组织磁盘数据访问等；另外，OSD 模块主要接受 MDC 模块下发的 OSD 视图，然后接受 VBS 模块的读写命令，完成数据的存放与获取。

[0083] 在实际应用中，上述 MDC 模块可以只部署在集群系统的两个（一主一备）或三个（一主两备）资源服务器上，VBS 模块部署在集群系统中的每个资源服务器上，作为驱动；OSD 模块部署在集群系统中的每个具有存储资源的资源服务器上，用于控制本地存储资源。具体的部署方式可以依据用户提供的配置文件执行，该配置文件中包括上述功能模块的部署、集群系统的分区规格（即，把每个硬盘分为多少份）以及不同资源服务器间的互相通信地址信息（包括 MDC 模块、VBS 模块和 OSD 模块的地址信息）等。

[0084] 当系统启动后，用户通过系统的管理端将配置信息导入系统，MDC 模块根据导入的配置信息建立系统的分区信息，然后根据系统下发的添加 OSD 节点请求建立逻辑 OSD 节点（是真实 OSD 节点在 MDC 模块侧的映射），并且为每个 OSD 节点分配分区等资源（即，每个资源服务器上的 OSD 模块对应的分区信息）。如图 6 所示，为 MDC 模块为 OSD 节点分配分区资源的拓扑示意图。当资源服务器上的 OSD 模块被激活后，所述 OSD 模块向 MDC 模块请求分区信息，根据该请求，MDC 模块将已经分配好的分区资源信息发送给对应的 OSD 模块。

[0085] 当系统中的 VBS 模块也被激活时，系统中的 MDC 模块、VBS 模块和 OSD 模块就都处于激活状态了，并彼此之间建立了连接。同时，MDC 模块也建立了全局的分区信息，以及完成了对每个 OSD 节点的分区资源分配和同步。

[0086] 当系统完成初始化进程后，VBS 模块会根据用户发起的创建用户卷命令创建一个用户卷，该用户卷包含卷 ID 信息、卷大小以及确定了哪些 Block 构成该用户卷，其中，不同的 Block 可能属于同一个分区，也可能属于不同的分区。

[0087] 因此，对于一个用户卷来说，其实际的存储物理资源是分布在很多磁盘上的。这样就达到了将不同的磁盘资源共享给一个用户卷，供用户使用，即达到存储资源共享使用的目的。

[0088] 可选的，当本发明实施例提供的集群系统需要增加资源服务器时，用户通过系统的管理端进行操作，将添加资源服务器的消息发给系统中的 MDC 模块（为系统中的主 MDC 模块）；所述 MDC 模块根据接收到的消息对新增的资源服务器进行参数校验，包括拓扑结果、IP 地址等；校验正确后，MDC 模块再进行（1）拓扑结构计算，将新增的资源服务器加入到 OSD 视图的拓扑结构图中，以及（2）分区结构计算，根据分区分配算法重新生成新的分区视图；完成上述计算后，MDC 模块将新的分区视图信息广播给系统中所有的 VBS 模块和 OSD 模块；随后，MDC 模块通知系统中的原有 OSD 模块发起数据迁移过程（根据新的分区拓扑结构，将原 OSD 节点下的数据迁移到新加入的节点）。

[0089] 可选的，当本发明实施例提供的集群系统需要删除资源服务器时，类似与上述增加资源服务器的流程，也是由用户通过系统的管理端进行操作，MDC 模块收到消息后重新进行分区计算，计算后再通知系统中各个 OSD 模块和 VBS 模块，然后再通知 OSD 模块发起数据

迁移。

[0090] 与增加资源服务器不同的是：在删除资源服务器的流程中，当数据迁移完成后，MDC模块将被删除的资源服务器设置为不可用状态。之后，用户才可以将该被删除的资源服务器撤离集群系统。

[0091] 在本发明实施例提供的计算存储融合的集群系统中，在资源服务器增删的过程中，MDC 模块根据节点的变化情况进行分区的分配计算并将变化情况通知到各个 OSD 模块和 VBS 模块。

[0092] 下面，我们以客户端用户发起对资源服务器 2 中的卷进行写数据操作请求为例，来说明在本发明实施例提供的计算存储融合的集群系统中，用户是如何在写数据过程中，实现对存储资源的共享使用的。

[0093] 首先，该写数据操作请求经过资源服务器 2 中对应的应用程序处理后，以标准的数据访问接口（可以是文件接口，也可以是块接口）要求写数据，并将该写数据操作请求发送给资源服务器 2 中的 VBS 模块；

[0094] 其次，VBS 模块根据自身的算法（如 DHT 等分布式算法）分析计算出需要写数据的逻辑 OSD 节点（包括计算出需要写入数据的 Block，以及所述 Block 所属的分区），VBS 模块通过自身算法，将需要写入的数据进行拆分（通过算法尽量均衡拆分），分别向逻辑 OSD 节点对应的资源服务器中的 OSD 模块发送写数据操作请求；

[0095] 例如，VBS 模块通过计算分析出需要写入数据的 Block 分别属于分区 P6 和 P7，而 P6 和 P7 分别归属资源服务器 2 中的 OSD 节点和资源服务器 3 中的 OSD 节点，VBS 模块通过自身算法，将需要写入的数据进行拆分（通过算法尽量均衡拆分），分别向资源服务器 2 中的 OSD 模块和资源服务器 3 中的 OSD 模块发送写数据操作请求。

[0096] 再次，资源服务器 2 中的 OSD 模块和资源服务器 3 中的 OSD 模块接收到写数据操作请求后，分别将数据写入本资源服务器的缓存中，后续再分别由资源服务器 2 的缓存和资源服务器 3 的缓存写入各自持久化磁盘的指定物理空间中；

[0097] 最后，为了保证写数据的可靠性，资源服务器 2 的缓存和资源服务器 3 的缓存再分别将数据写入到本资源服务器的持久化磁盘的其他空闲物理空间中，从而完成数据写入流程。

[0098] 需要说明的是：最后一步可以由缓存异步并行执行。

[0099] 这样，一个数据写入请求，经过上述步骤处理后，达到了分布式并行写入不同资源服务器的缓存中，再由各资源服务器的缓存写入本地的持久化磁盘中，提高了写数据的效率，实现了磁盘的共享使用。当数据写入完成后，OSD 模块中的逻辑分区与实际的磁盘物理分区建立起了对应关系。

[0100] 进一步，我们以客户端用户发起对资源服务器 2 中的卷进行读数据操作请求为例，来说明在本发明实施例提供的计算存储融合的集群系统中，用户是如何在读数据过程中，实现对存储资源的共享使用的。

[0101] 首先，该读数据操作请求经过资源服务器 2 中对应的应用程序处理后，以标准的数据访问接口（可以是文件接口，也可以是块接口）要求读数据，并将该读数据操作请求发送给资源服务器 2 中的 VBS 模块；

[0102] 其次，VBS 模块根据自身的算法（如 DHT 等分布式算法）分析计算出需要读取数据

的逻辑 OSD 节点(包括计算出需要读数据的 Block,以及所述 Block 所属的分区),分别向逻辑 OSD 节点对应的资源服务器中的 OSD 模块发送读数据操作请求 ;

[0103] 例如,VBS 模块通过计算分析出需要读取数据的 Block 分别属于分区 P6 和 P7,而 P6 和 P7 分别归属资源服务器 2 中的 OSD 节点和资源服务器 3 中的 OSD 节点,则 VBS 模块分别向资源服务器 2 中的 OSD 模块和资源服务器 3 中的 OSD 模块发送读数据操作请求。

[0104] 最后,资源服务器 2 中的 OSD 模块和资源服务器 3 中的 OSD 模块接收到读数据操作请求后,先到本资源服务器中的缓存读取数据,如果缓存中没有所需数据,再到本地持久化磁盘读取数据,从本地持久化磁盘读出数据后,先存入缓存,以便下次从缓存中读取。

[0105] 因此,一个数据读取请求,经过上述步骤处理后,可以分布式并行从不同资源服务器的缓存中读取数据(当缓存中没有所需数据时,再从持久化磁盘读取),从而提高了读数据的效率,实现了磁盘的共享使用。

[0106] 在本发明实施例提供的计算存储融合的集群系统中,随着集群系统中持久化存储资源的增加,整个集群系统的缓存随之线性增加,缓存的增加,意味着系统读写数据时,同一个任务的分布式并发处理会越多,效率会更高,系统整体性能随着系统的扩容不断提高。

[0107] 通过本发明实施例提供的计算存储融合的集群系统,由于不存在专用的 SAN,省略了对 SAN 存储系统的复杂管理,在硬件上解决了现有技术中因为使用专用 SAN 而导致的操作复杂、成本较高的问题 ;存储设备可以有多个,每个存储设备上都可以部署缓存,在硬件上极大的提升了存储端缓存的扩展能力 ;存储资源不依赖于计算资源,存储设备可以独立的增加和减少,增强了系统的可扩展性。另外,由于本发明实施例提供计算存储融合的集群系统采用高速数据交换网络进行通信,进一步加快了数据的交换速度。

[0108] 应理解,在本发明实施例中,“与 A 相应的 B”表示 B 与 A 相关联,根据 A 可以确定 B。但还应理解,根据 A 确定 B 并不意味着仅仅根据 A 确定 B,还可以根据 A 和 / 或其它信息确定 B。

[0109] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0110] 在本申请所提供的实施例中,应该理解到,所揭露的系统,可以通过其它的方式实现。例如,以上所描述的系统实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另外,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接,也可以是电的,机械的或其它的形式连接。

[0111] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本发明实施例方案的目的。

[0112] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0113] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0114] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

图 1

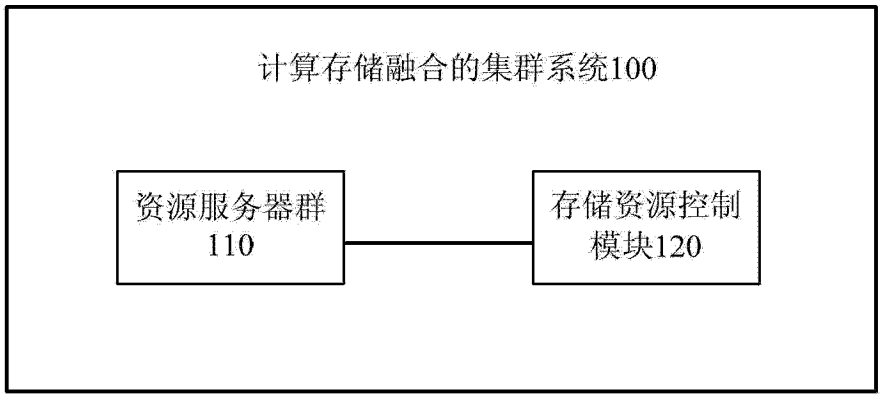


图 2

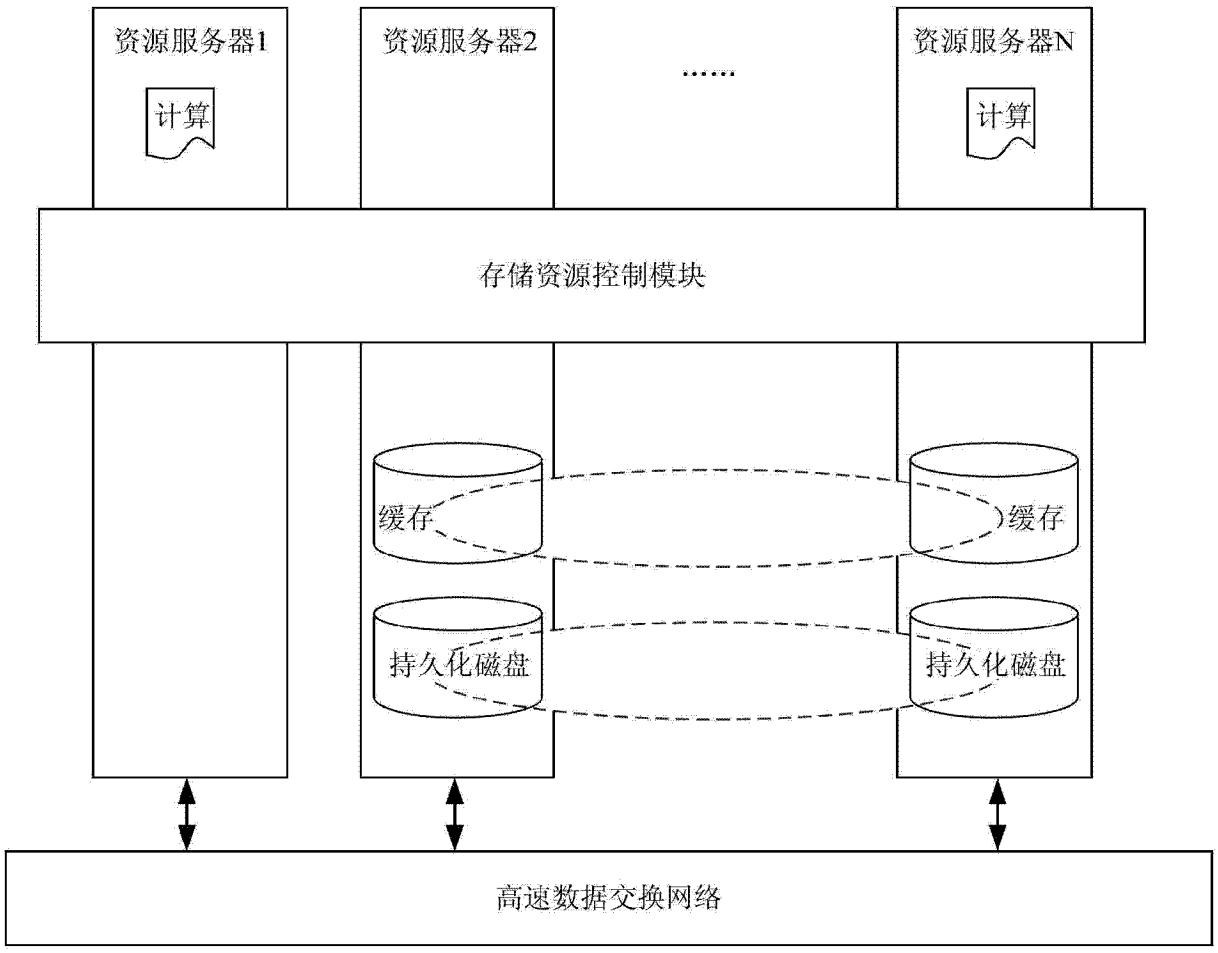

资源服务器1
计算
资源服务器2
.....
资源服务器N
计算
存储资源控制模块
缓存
缓存
持久化磁盘
持久化磁盘
高速数据交换网络

图 3

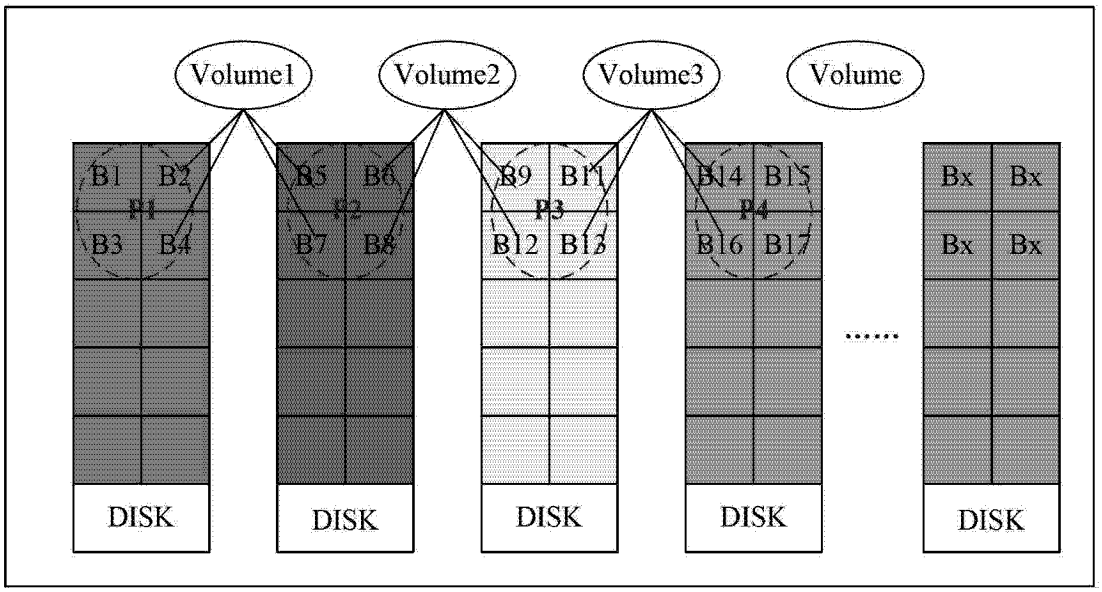

Volume1
Volume2
Volume3
Volume
B1
B2
B3
B4
P1
B5
B6
B7
B8
P2
B9
B11
B12
B13
P3
B14
B15
B16
B17
P4
Bx
Bx
Bx
Bx
.....
DISK
DISK
DISK
DISK
DISK

图 4

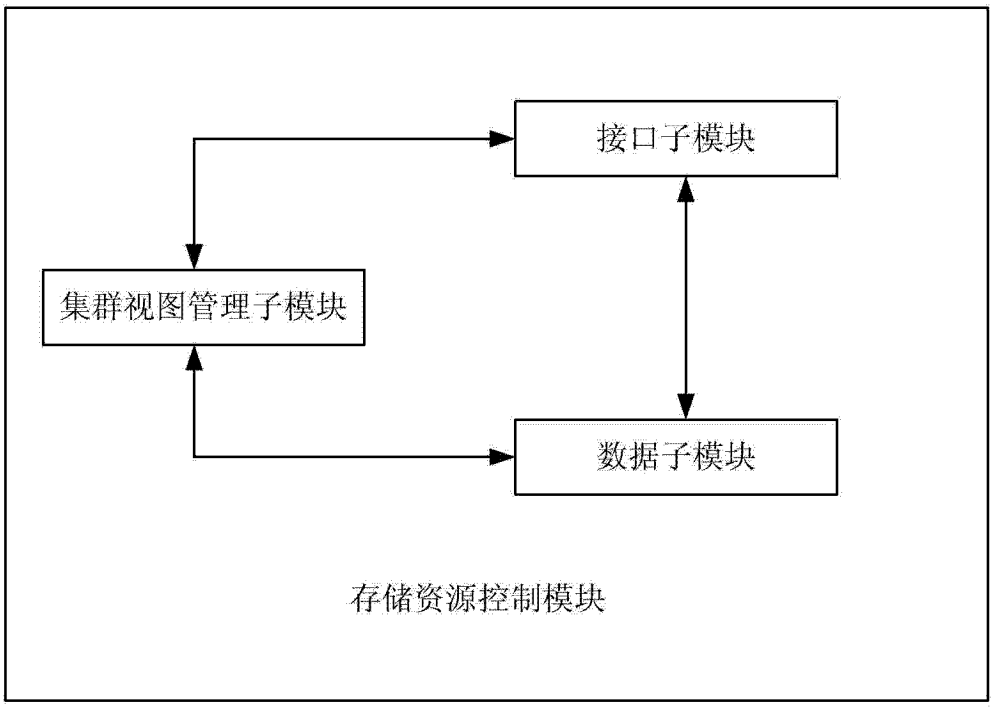

接口子模块
集群视图管理子模块
数据子模块
存储资源控制模块

图 5

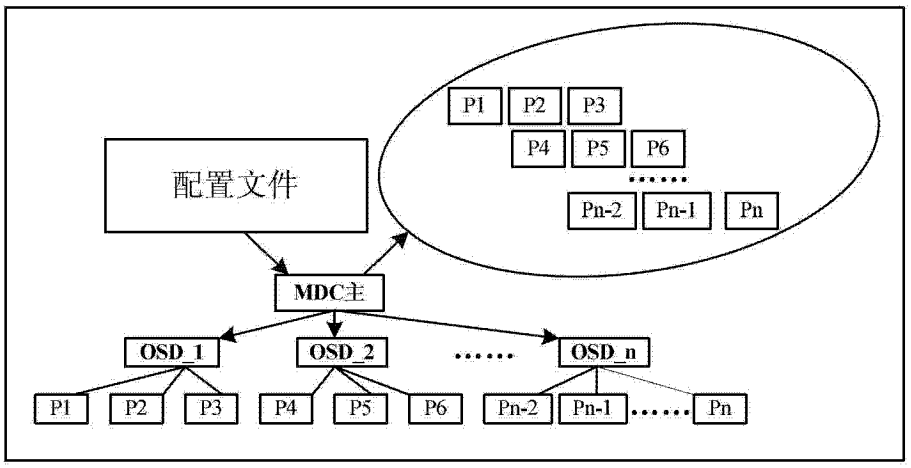

P1
P2
P3
P4
P5
P6
Pn-2
Pn-1
Pn
配置文件
MDC主
OSD_1
OSD_2
OSD_n
P1
P2
P3
P4
P5
P6
Pn-2
Pn-1
Pn

图 6