(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 103701916 A
(43)申请公布日 2014.04.02

(21)申请号 201310749353.7

(22)申请日 2013.12.31

(71)申请人 赛凡信息科技(厦门)有限公司
地址 361000 福建省厦门市软件园望海路55号A901-903单元

(72)发明人 吴江

(74)专利代理机构 厦门市精诚新创知识产权代理有限公司 35218
代理人 张伟星

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

权利要求书1页 说明书5页

(54)发明名称

分布式存储系统的动态负载均衡方法

(57)摘要

本发明公开一种分布式存储系统的动态负载均衡方法，该方法用于n个存储节点的非共享型分布式存储系统为m个节点的客户端服务器提供s个数据单位的存储服务，其包括以下步骤：步骤1：统计如下信息：统计各存储节点上对每一个客户端连接的数据存取访问，统计s个数据单位中的每一个数据单位通过不同存储节点的访问次数，统计每个节点的空间使用率；步骤2：根据步骤1中的上述统计数据，预先制定带宽阈值，远程访问阈值和容量阈值，所有的阈值均为百分比，判断分布式存储系统的各节点是否出现网络负载不均衡，跨节点访问次数过多导致的高延迟和容量极度不均衡，并根据判断结果选择迁移数据或者通过路由重定向客户端接入访问点。

CN 103701916 A

1. 分布式存储系统的动态负载均衡方法，该方法用于 n 个存储节点的非共享型分布式存储系统为 m 个节点的客户端服务器提供 s 个数据单位的存储服务，包括以下步骤：

步骤 1：统计如下信息：统计各存储节点上对每一个客户端连接的数据存取访问，统计 s 个数据单位中的每一个数据单位通过不同存储节点访问的次数，统计每个节点的空间使用率；

步骤 2：根据步骤 1 中的上述统计数据，预先制定容量最高阈值 HighWaterMark 和最低阈值 LowWaterMark，最高阈值 HighWaterMark 和最低阈值 LowWaterMark 为百分比，判断分布式存储系统的各节点是否出现容量极度不均衡，并根据判断结果选择迁移数据或者通过路由重定向客户端接入访问点。

2. 根据权利要求 1 所述的分布式存储系统的动态负载均衡方法，其特征在于：所述步骤 1 中，统计各存储节点上对每一个客户端连接的数据存取访问，具体是：统计每一个客户端的数据存取服务的可以由访问点直接满足的次数 LocalAccess，并统计每一个客户端的数据存取服务必须由除访问点以外的节点满足的次数 $RemoteAccess_{[n-1]}$；[n-1] 为 RemoteAccess 的下标值，是除访问点外的其他节点的序号。

3. 根据权利要求 2 所述的分布式存储系统的动态负载均衡方法，其特征在于：所述步骤 1 中，统计 s 个数据单位中的每一个数据单位通过不同存储节点的访问次数，具体是：以文件或块为单位统计该数据单位直接由客户端存取的次数 LocalHit，以及统计对与每一个集群中的其它节点，通过这些节点访问这个数据单位的次数 $RemoteHit_{[n-1]}$；[n-1] 为 RemoteHit 的下标值，是除访问点外的其他节点的序号。

4. 根据权利要求 3 所述的分布式存储系统的动态负载均衡方法，其特征在于：所述步骤 2 中，当判断结果满足：其中一个节点的容量的使用量超过最高阈值 HighWaterMark，一个节点的容量的使用量低于最低阈值 LowWaterMark，则进行如下操作：

迁移数据：对于选定的数据单位首先通过 DHT 算法确定目标节点 nodeDst，如果这个数据单位通过 nodeDst 访问次数远小于通过其中某个节点 nodeY 的访问次数，那么目标节点选择 nodeY；如果 nodeY 就是本地节点，那么不迁移该数据单位。

5. 根据权利要求 4 所述的分布式存储系统的动态负载均衡方法，其特征在于：所述步骤 2 中，令某个客户端跨节点访问数据的次数 $RemoteAccess_{[nodeX]}$ 为 K1，令本地访问次数 LocalAccess 为 K2，当判断结果满足：K1-K2 大于预设阈值 Z，则进行如下操作：

选取客户端跨节点访问次数最多的那个节点即 RemoteAccess 数组中值最大的那个下标 nodeX，通过路由重定向客户端到这个节点 nodeX，如果此目标结点 nodeX 的带宽一直处于满带宽，而有其它节点带宽统计没有达到满带宽，那么暂时不迁移，首先尝试触发数据均衡；同时清除客户端访问统计；如果所有的节点都近似带宽饱和，通过路由重定向客户接入访问点到目标节点。

6. 根据权利要求 4 或 5 所述的分布式存储系统的动态负载均衡方法，其特征在于：所述步骤 2 中，令一个数据单位通过某个节点跨节点访问的次数 $RemoteHit_{[nodeY]}$ 为 L1，令本地访问次数 LocalHit 为 L2，当判断结果满足：L1-L2 大于预设阈值 W，则进行如下操作：迁移数据到目标节点 nodeY。

分布式存储系统的动态负载均衡方法

技术领域

[0001] 本发明涉及分布式存储系统，具体涉及非对称分布式存储系统的动态负载均衡方法。

背景技术

[0002] 随着云计算技术的发展，云数据中心中的计算服务器的数量越来越多，保存处理的数据量越来越多，存储的负载越来越重。传统的分布式存储系统如 SAN（存储域网）和单独的 NAS（网络附加存储）已经没有办法轻松的应对云计算和大数据对存储容量，存储带宽的需求了。随之而产生了多种多样的分布式存储系统，共享型分布式存储系统和非共享型分布式存储系统就是其中最为主要的两种分布式存储系统。多样的分布式存储系统一般是要解决两个问题 :1. 容量的扩充，2. 性能的扩充。目前的分布式存储系统，都是依托于硬件的基础上，实现一个硬件集群，但是，这些集群方案仅根据分布式存储系统中的单一资源来分配任务请求，而且不能根据需要实现灵活的配置，造成集群中一些平台的资源浪费，降低了分布式存储系统的处理能力，使得分布式存储系统不能做出正确的分配决策。

[0003] 为此，申请号为 201010002264.2 的发明专利，公开了一种实现负载均衡的方法、负载均衡服务器以及集群系统，其中，实现负载均衡的方法主要包括如下过程 :负载均衡服务器采用 MINA 的点对点的模式来接收任务请求，根据负载均衡策略和至少两个子节点分别发送的各种资源信息确定处理所述任务请求的子节点，其中包括根据配置的最小资源处理数算法和所述各种资源信息确定出处理所述任务请求的子节点，具体而言，包括为所述至少两个子节点中的各个子节点设置节点权值，为所述各种资源信息中的每一种资源信息设置资源值和资源负载权值，对所述各个子节点中的每一个子节点按如下处理得到各个子节点各自的负载值 :分别对该子节点的每一种资源信息的资源值和资源负载权值做乘积处理得到第一结果值，将各种资源信息的第一结果值相加得到第二结果值，将所述第二结果值与该子节点的节点权值做乘积处理得到该子节点的负载值，根据预先配置的规则，将所述各个子节点中负载值符合该规则的子节点确定为处理所述任务请求的子节点，其中预先配置的规则为选择该各个节点中负载值最小的子节点 ;将所述任务请求发送给确定出的子节点，其中所述资源信息包括处理完毕的上一个任务请求的信息。该方法虽然在一定程度上实现了集群的灵活配置和更有效的负载均衡，但是其算法中，要实时获取子节点的每一种资源信息的资源值和资源负载权值并进行实时运算，算法复杂，计算量大，对负载均衡设备能力要求很高，同时，当客户端进行数据存取时，需要等待的时间过久。

[0004] 再例如申请号为 201110418127.1 的发明专利，公开了一种负载均衡系统、装置及方法，所述方法包括 :所述网络交换机用于获取用户数据流中的特征信息，当所述特征信息符合预定的第一负载均衡策略时，将收到的用户数据流中的数据报文按所述第一负载均衡策略转发给与所述网络交换机连接的多个服务器单元以实现低层负载均衡。当所述特征信息符合预定的第二负载均衡策略时，将收到的数据报文发送给所述负载均衡设备，然后根据所述负载均衡设备确定的转发策略将数据报文转发给与所述网络交换机连接的多个

服务器单元以实现高层负载均衡。其中，第一负载均衡策略是将具有大带宽特征的特征信息的数据报文转发给特定服务器单元；第二负载均衡策略是将具有小带宽特征的特征信息的数据报文转发给所述负载均衡设备。由上可知，该发明需要独立的负载均衡设备来完成上述过程；同时，第一负载均衡策略和第二负载均衡策略分别需要标记大带宽特征和小带宽特征的特征信息，对于一般的数据报文，其大带宽特征和小带宽特征区分不是很明显，因此，该发明提供的方法在具体实现上面来说，具有一定的困难；另外，上述发明的方法只考虑到带宽问题，并未考虑到延迟问题，具有一定的缺陷。除此之外，以上的几种负载均衡策略都是由一个中央节点来扮演负载均衡的仲裁者。这种集中式负载均衡策略不适合大规模的集群系统，第一级中的负载均衡仲裁设备是一个单点故障，其次级中的负载均衡设备随着前段设备和后端设备数目的增多，负载的加大很有可能成为整个系统的瓶颈。

发明内容

[0005] 因此，针对上述的问题，本发明提出一种非共享型分布式存储系统的动态负载均衡方法，使非共享型分布式存储系统的性能随着负载的变化而动态调整，并充分考虑带宽，延迟和容量的均衡问题，以最大化发挥分布式存储系统的带宽并且保证较低的延迟，同时该方法简单易实现，从而解决现有技术之不足。

[0006] 与共享型(centralized)分布式存储系统相比，非共享型(decentralized)分布式存储系统中所有的存储结点完全对称，能够做到所有的存储节点都可以作为访问点直接对客户端提供存储的存取(I/O)服务。这样可以看起来，非共享型分布式存储系统的性能可以随着存储结点的增加而近似线性增长。相比较传统的服务器负载均衡方法，本发明的负载均衡方法是可以应用于非对称分布式存储系统及这种负载的系统架构，而且该负载均衡方法综合考虑了网络的带宽和存储的延迟以及容量的平衡，可以最大程度的发挥分布式存储系统的聚合性能。

[0007] 为了更好的说明本发明的方法，假设有一个 n 个存储节点的非共享型分布式存储系统为 m 个节点的客户端服务器提供 s 个数据单位的存储服务。分布式存储系统的负载均衡方法需要在每一个节点上统计流量。以网络流量来衡量带宽的负载。如果多个客户端同时连接分布式存储系统，根据动态负载均衡方法，先按照均匀轮流策略加上统计结果来将多个客户端路由到不同的存储结点上。每当一个客户端连接请求来时，动态负载均衡方法需要从上一次纪录的游标(cursor)开始顺序查询带宽统计，找到第一个小于满带宽的节点，将客户端连接路由到这个节点。

[0008] 根据分布式存储系统的特点，大多数情况下每一个客户端在任意时刻都直接入到分布式存储系统的一个访问点上，那么当客户端所想要访问的数据恰好不再访问点存储的时候，分布式存储系统就不得不由访问点代为转发这个数据请求到数据所在的目标节点上做数据的存取服务。这种跨节点的数据访问增加的客户端数据访问的延迟，所以在整个负载均衡策略中要尽可能的避免跨节点的操作降低延迟。默认初始情况下，分布式存储系统采用就近原则，在客户端的连接访问点上创建新的数据，在大多数情况下，用户通过一个存储节点创建文件，有很大的机率他会一直使用这个存储会话连接访问存储，并且很大的机率创建数据的客户端就是今后处理数据的客户端。所以在一开始，所有的客户端基本上都不需要跨节点访问。

[0009] 有了以上的条件和假设，具体的，本发明的分布式存储系统的动态负载均衡方法，该方法用于 n 个存储节点的非共享型分布式存储系统为 m 个节点的客户端服务器提供 s 个数据单位的存储服务，包括以下步骤：

步骤 1：统计如下数据：

a：在各存储节点上对每一个客户端（m 个客户端中的每一个）连接的数据存取访问作统计：统计每一个客户端的数据存取服务的可以由访问点直接满足（本地数据访问）的命中次数 LocalAccess，并统计每一个客户端的数据存取服务必须由除访问点以外的节点满足（跨节点访问的目标结点的访问）的次数 $RemoteAccess_{[n-1]}$；[n-1] 为 RemoteAccess 的下标值，是除访问点外的其他节点的序号。对每一个访问点来讲，这里共有 n-1 个其他节点是远程访问；

b：以文件或块为单位统计 s 个数据单位中的每一个数据单位通过不同存储节点访问的次数：该数据单位直接由客户端存取（本地访问命中）的次数 LocalHit，以及统计对与每一个集群中的其它节点，通过这些节点访问这个数据单位的次数 $RemoteHit_{[n-1]}$；[n-1] 为 RemoteHit 的下标值，是除访问点外的其他节点的序号。对每一个数据单位而言，通过数据单位所在的节点直接访问的值就是 LocalHit；而除了数据单位以外，通过其他的访问点来访问这个数据单位就是 $RemoteHit_{[n-1]}$，这里共有 n-1 个节点是远程访问；

c：统计每个节点的空间使用率；

步骤 2：根据上述统计数据，并预先制定带宽阈值、远程访问阈值和容量阈值，各阈值均为百分比；设容量最高阈值 HighWaterMark 和最低阈值 LowWaterMark；判断分布式存储系统的各节点是否出现容量极度不均衡（一个存储节点的空间近满，而其他的节点的存储空间近似为空），其中一个节点的容量使用了 HighWaterMark 以上，而其他的节点有少于 LowWaterMark 的（HighWaterMark，LowWaterMark 都是相对百分比），如果是则转到步骤 31；

判断某个客户端跨节点访问数据的次数 RemoteAccess[nodeX]=K1 大于本地访问次数 LocalAccess=K2 是否到达一个预设的阈值 Z，即 K1-K2>Z，如果是则转到步骤 32；

判断一个数据单位（文件或数据块）通过某个节点跨节点访问的次数 RemoteHit[nodeY]=L1 大于本地访问次数 LocalHit=L2 是否到达一个设定的阈值 W，即 L1-L2>W，如果是则转到步骤 33；

步骤 31：迁移数据：对于选定的数据单位首先通过 DHT（分布式哈希）算法确定目标节点 nodeDst，如果这个数据单位通过 nodeDst 访问次数远小于通过其中某个节点 nodeY 的访问次数，那么目标节点选择 nodeY；如果 nodeY 就是本地节点，那么暂时不迁移本数据单位；

步骤 32：选取客户端跨节点访问次数最多的那个节点即 RemoteAccess 数组中值最大的那个下标 nodeX，通过路由重定向客户端到这个节点 nodeX，如果此目标结点 nodeX 的带宽一直处于满带宽，而有其它节点带宽统计没有达到满带宽，那么暂时不迁移，首先尝试触发数据均衡；同时清除客户端访问统计；如果所有的节点都近似带宽饱和，通过路由重定向客户接入访问点到目标节点；

步骤 33：迁移数据到目标节点 nodeY。

[0010] 负载均衡的手段一般有两种：一种是迁移数据，数据的均衡是带宽均衡的前提。用分布式散列表算法对数据单位重新分布。数据尽可能分布在频繁访问点上，以此降低延迟。

另一种是通过路由重定向客户端接入访问点。本发明通过采用上述方法，其融合了访问路径均衡和数据均衡的负载均衡方法。其方法简单易于实现，且整个分布式存储系统根据如上所述的均衡策略，以尽可能小的延迟达到分布式存储系统负载的均衡。

具体实施方式

[0011] 现结合具体实施方式对本发明进一步说明。

[0012] 分布式存储系统性能的衡量指标中最主要的是：带宽和延迟。这两个性能指标在某种情况下是互相作用和互相影响的。对于一个非共享型分布式存储系统，数据分散存放于多个存储结点上，如果任意一个节点都可以作为数据访问点的话，那么意味着当用户通过一个节点访问数据时，有可能他所想要访问的数据在其他的存储结点上，这时候作为入口点的节点必须代表客户端从数据所在的目标节点上将数据取到访问点上，然后将数据返回给客户端，那么这种跨节点访问势必带来的是将对较高的延迟，在数据并发访问相对较少的情况下，高的延迟限制了带宽。那么为了降低延迟和提高带宽是否路由所有的客户端都直接去访问数据所在的目标节点就可以降低延迟，达到好的性能呢？结论不是，因为随着访问同一组数据的客户端数目的增多，并发访问量的变大，整个分布式存储系统的带宽一定受限于那一个存储结点的带宽。所以共享型分布式存储系统最好是能通过一种高级的负载均衡调度策略来最大化发挥分布式存储系统的带宽并且保证较低的延迟。

[0013] 这种调度策略是一种融合了访问路径均衡和数据均衡的负载均衡方法。如下策略假设有一个 n 个存储节点的非共享型分布式存储系统为 m 个节点的客户端服务器提供 s 个数据单位的存储服务。

[0014] 分布式存储系统的负载均衡方法需要在每一个节点上统计流量。以网络流量来衡量带宽的负载。如果多个客户端同时连接分布式存储系统本策略先按照均匀轮流策略加上统计结果来将多个客户端路由到不同的存储结点上。每当一个客户端连接请求来时，均衡策略需要从上一次纪录的游标(cursor)开始顺序查询带宽统计，找到第一个小于满带宽的节点，将客户端连接路由到这个节点。

[0015] 分布式存储系统采用就近原则，在客户端的连接访问点上创建新的数据，在大多数情况下，用户通过一个存储节点创建文件，有很大的几率他会一直使用这个存储会话连接访问存储，并且很大的几率创建数据的客户端就是今后处理数据的客户端。所以在一开始，所有的客户端基本上都不需要跨节点访问。

[0016] 同时本负载均衡方法要求作如下 3 个统计：

1：在各存储节点上对每一个客户端(m 个客户端中的每一个)的连接作统计：统计每一个客户端的数据存取服务的本地数据访问的命中次数 LocalAccess，并统计每一个跨节点访问的目标结点的访问次数 $RemoteAccess_{[n-1]}$；[n-1] 为 RemoteAccess 的下标值，是除访问点外的其他节点的序号。对每一个访问点来讲，这里共有 n-1 个其他节点是远程访问；

2：以文件或块为单位统计 s 个数据单位中的每一个数据单位的访问次数：本地访问命中次数 LocalHit，以及统计对与每一个集群中的其它节点，通过这些节点访问这个数据单位的次数 $RemoteHit_{[n-1]}$；[n-1] 为 RemoteHit 的下标值，是除访问点外的其他节点的序号。对每一个数据单位而言，通过数据单位所在的节点直接访问的值就是 LocalHit；而除了数据单位以外，通过其他的访问点来访问这个数据单位就是 $RemoteHit_{[n-1]}$，这里共有 n-1 个

节点是远程访问；

3：统计每个节点的空间使用率。

[0017] 负载均衡的手段有两种：

1. 迁移数据：数据的均衡是带宽均衡的前提。用 DHT 算法对数据单位重新分布。数据尽可能分布在频繁访问点上，以此降低延迟；

2. 通过路由重定向客户端接入访问点。

[0018] 预先制定容量最高阈值 HighWaterMark 和最低阈值 LowWaterMark，最高阈值 HighWaterMark 和最低阈值 LowWaterMark 为百分比，判断分布式存储系统的各节点是否出现容量极度不均衡，并根据判断结果选择迁移数据或者通过路由重定向客户端接入访问点。其中，触发负载均衡的条件有三个：

条件 1：容量极度不均衡，其中一个节点的容量的使用量超过最高阈值 HighWaterMark，其他有节点的容量的使用量低于最低阈值 LowWaterMark；

条件 2：对某个客户端跨节点访问数据的次数大于本地访问次数到达一个设定的阈值 Z；令某个客户端跨节点访问数据的次数 $RemoteAccess_{[nodeX]}$ 为 K1，令本地访问次数 LocalAccess 为 K2，如果 K1-K2 大于预设阈值 Z；

条件 3：令一个数据单位（文件或数据块）通过某个节点跨节点访问的次数 $RemoteHit_{[nodeY]}$ 为 L1，令本地访问次数 LocalHit 为 L2，如果 L1-L2 大于预设阈值 W；

均衡判定规则：

当条件 1 触发时，迁移数据：对于选定的数据单位首先通过 DHT（分布式哈希）算法确定目标节点 nodeDst 进行数据迁移，如果这个数据单位通过 nodeDst 访问次数远小于通过其中某个节点 nodeY 的访问次数，那么目标节点选择 nodeY；如果 nodeY 就是本地节点，那么不迁移该数据单位。

[0019] 当条件 2 触发时，选取客户端跨节点访问次数最多的那个节点即 RemoteAccess 数组中值最大的那个下标 nodeX，通过路由重定向客户端到这个节点 nodeX，如果此目标结点 nodeX 的带宽一直处于满带宽，而有其它节点带宽统计没有达到满带宽，那么暂时不迁移，首先尝试触发数据均衡；同时清除客户端访问统计；如果所有的节点都近似带宽饱和，通过路由重定向客户接入访问点到目标节点。

[0020] 当条件 3 触发时，迁移数据到目标节点 nodeY。

[0021] 本发明结合负载均衡的两种手段（迁移数据、通过路由重定向客户端接入访问点），在不同的条件触发选择不同的数据迁移策略。现有技术中的负载均衡策略一般都由一个中央节点来扮演负载均衡的仲裁者，这种集中式负载均衡策略不适合大规模的集群系统，第一级中的负载均衡仲裁设备是一个单点故障，其次级中的负载均衡设备随着前段设备和后端设备数目的增多，负载的加大很有可能成为整个系统的瓶颈。而本发明提供的上述均衡策略，尤其适合大规模的集群系统。整个分布式存储系统可以根据如上所述的均衡策略，以尽可能小的延迟达到分布式存储系统负载的均衡。

[0022] 尽管结合优选实施方案具体展示和介绍了本发明，但所属领域的技术人员应该明白，在不脱离所附权利要求书所限定的本发明的精神和范围内，在形式上和细节上可以对本发明做出各种变化，均为本发明的保护范围。