(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 103324577 B
(45)授权公告日 2016.04.06

(21)申请号 201310228642.2

(22)申请日 2013.06.08

(73)专利权人 北京航空航天大学
地址 100191 北京市海淀区学院路37号

(72)发明人 阮利 陈鲲 肖利民 董斌

(74)专利代理机构 北京金恒联合知识产权代理事务所 11324
代理人 李强

(51)Int.Cl.
G06F 12/02(2006.01)

(56)对比文件
CN 102629219 A,2012.08.08,
CN 102882983 A,2013.01.16,
US 7571168 B2,2009.08.04,
Bin Dong.A File Assignment Strategy for Parallel I/ O System with Minimum I / O Contention Probability.《Communication in Computer and Information Science》.2011,
Bin Dong.Self-acting Load Balancing with Parallel Sub File Migration for Parallel File System.《IEEE》.2010,

审查员 杨牛

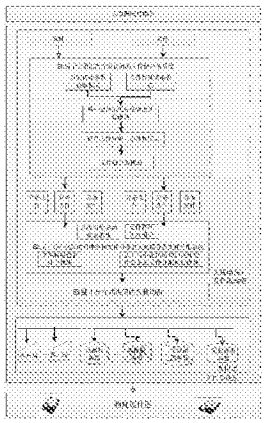
权利要求书1页 说明书5页 附图2页

(54)发明名称

基于最小化IO访问冲突和文件分条的大规模分条文件分配系统

(57)摘要

本发明针对大数据的存储挑战以及需要频繁的进行文件读写的科学计算程序输入输出瓶颈等方面存在的问题,发明了一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统,在模块构成上主要包括系统初始参数获取模块,文件特征读取模块,文件和磁盘的排序模块,基于最小化IO访问冲突和文件分条的文件分配执行模块。由于本发明面向大数据应用和高性能计算机中大规模分条文件,充分考虑大规模文件请求的磁盘I/O冲突概率,为高性能计算机和大数据存储系统提供了适于大数据、最小化IO访问冲突的文件分配支持,进而为高效的大数据并行输入/输出提供有力支撑,故本发明具有广阔的应用前景。



CN 103324577 B

1.一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，主要完成在在多个并行的磁盘内对要被访问的大数据分条文件进行文件的有效分配，具有满足面向大数据分条文件，充分考虑文件请求的磁盘I/O冲突概率需求的技术特征，其特征在于：

在体系结构上，位于适于大数据应用的并行文件系统的体系结构自上而下为，大数据应用层->并行文件系统层->物理硬件层，该体系结构中的并行文件系统层；

在流程上，位于负载均衡处理流程中的最小化I/O访问冲突的文件分配步骤，是基于分布式决策的负载均衡步骤的前提步骤，

其中所述体系结构在模块构成上包括：

(1)系统初始参数获取模块：主要完成系统初始参数获取并将所获得的参数存储到并行文件系统中的配置文件中；

(2)文件特征读取模块：主要完成从文件应用层和文件系统支撑层读取系统输入参数；

(3)文件和磁盘的排序模块：主要完成按照文件的访问密度对文件进行排序以及对并行磁盘按照负载进行排序；

(4)基于最小化IO访问冲突和文件分条的文件分配执行模块：主要完成执行文件在多个并行的磁盘分配。

2.根据权利要求1的大规模分条文件分配系统，其特征在于其系统运行流程包括：

第一步，系统初始参数获取，首先由系统初始参数获取模块获取初始系统参数，获取的五个参数包括并行磁盘的数量、文件的数目、文件分条大小、文件的访问频率和文件的文件请求大小，然后系统初始参数获取模块将所获得的参数存储到并行文件系统中的配置文件中；

第二步，文件特征读取，首先由文件特征读取模块从文件系统应用层和文件系统支撑层读入输入参数，然后由文件特征读取模块对每个磁盘初始化其访问密度矩阵；

第三步，文件和磁盘排序，由第一步和第二步得到的数据，基于最小化IO访问冲突和文件分条对文件和磁盘进行排序，具体计算方法为：首先对每个文件都计算该文件的访问密度，然后对所有文件按照它们的访问密度按降序进行排序，得到一个分条文件按照降序排序得到的文件序列，然后根据磁盘的负载对磁盘按照升序进行排序，得到一个并行磁盘按照负载大小升序排列的磁盘队列；

第四步，基于最小化IO访问冲突和文件分条的文件分配执行，由文件分配模块对所有的文件按照其访问密度的降序采用贪心算法的方式在磁盘上进行分配，即具有最大访问密度的分条文件放到负载最小的硬盘上。

基于最小化IO访问冲突和文件分条的大规模分条文件分配系统

技术领域

[0001] 本发明公开了一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，尤其涉及一种面向大数据应用的、面向分条文件的、充分考虑最小化大数据文件请求的磁盘I/O冲突概率、采用启发式方法的大规模分条文件分配系统，属于计算机技术领域。

背景技术

[0002] 大数据(Big Data)指大规模数据。自2010年来，大数据成为了学术界、工业界的研究热点，原因在于一方面，在于互联网、移动互联网、物联网，以及云计算的规模和应用的激增，大量的用户和应用的交互导致产生巨量的数据；另一方面，随着数据采集技术的进步，诸如卫星遥感、传感器、GPS等，也导致每时每刻都在产生巨量的数据；最后，在科研和工业等领域的复杂的新技术、新仪器的使用也导致数据量的产生与日俱增，例如，欧洲核子研究中心的大型粒子对撞机(Large Hardon Collider)每年产生约15PB的数据。据<<经济学人>>杂志分析称，全世界的数据量，在2005年约为150EB，2010年约为1200EB，到2020年，则预期为35000EB。数据规模的飞速发展对传统的数据存储、处理、共享等方式提出了更高的要求，而且为了充分发挥长期积累的巨量数据的效能，学术界和工业界再一次将目光转向大数据的研究，并成为学术和工业等领域的热点技术。

[0003] 最早提出“大数据”时代已经到来的机构是全球知名咨询公司麦肯锡(McKinsey)。麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。麦肯锡将大数据定义为“规模超过典型的数据库软件工具的采集、存储、管理和分析能力的数据集”。IBM则从四个维度定义了大数据的特征(4V)：容量(Volume)，速度(Velocity)，多样性(Variety)，真实性(Veracity)。Wikipedia。给出的定义是：大数据是一个大而复杂，以至于难以用现有数据库管理工具或传统数据分析程序来处理的数据集，包括在采集(capture)、管理(curation)、存储(storage)、搜索(search)、共享(sharing)、分析(analysis)，以及可视化(visualization)等多方面的挑战。

[0004] 计算机软件系统优化是缓解计算机系统“输入/输出性能瓶颈”、解决适于大数据计算的高性能计算机系统规模的扩大和科学计算程序中数据敏感性计算的增加所带来的新问题的至关重要的方法之一。由于适于大数据应用的并行I/O系统软件能够将独立的资源(如磁盘、服务器、网络带宽等)整合在一起来为并行程序提供高速的聚合I/O，因此并行I/O系统软件作为适于大数据应用的高性能计算机整机系统软件的重要部分，是有效解决高性能计算机I/O性能问题行之有效的方法。并行文件系统作为并行I/O系统软件堆栈的基础层，是适于大数据应用的并行I/O系统重要组成部分，其不仅向机群提供单一的存储镜像，而且还扩展了传统文件系统对并行I/O的语义和接口限制。并行文件系统还提供了文件分条方法、分条文件在服务器间的分配方法、I/O服务器间并行访问的协调机制(例如动态负载均衡)来保证高速的聚合I/O速度。有效的、面向分条文件的分配算法是并行I/O性能的

有力保障。然而现有研究中目前仍然缺乏一种面向分条文件的、充分考虑文件请求的磁盘I/O冲突概率、采用启发式的大数据文件分配系统，本发明即公开一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统。

[0005] 下面从本发明相关的学术研究及已发表论文分析、以及相关的专利分析两方面对本专利的创新性进行分析。首先大数据是现有的最优的文件分配模型能够很准确的描述整个文件分配的问题，并且提供最优的文件分配结果。然而，最优的文件分配问题被证明为一个NP-完全问题，基于最优文件分配模型的文件分配系统计算复杂度高，实现难度大；另一方面，基于启发式思想的文件分配方法具有低的计算复杂度，因此变成了有效且实用的文件分条策略。典型的启发式文件分条方法包括排序分区(Sort Partition(SP))，混合分区(Hybrid Partition(HP))，静态循环分配(Static Round-robin(SOR))，平衡排序分配(Balanced Allocation with Sort(BAS))，和平衡队列排序分配(Balanced Allocation with Sort for Batch(BASB))等。如表1所示，现有的文件典型的启发式文件分配方法通过平衡磁盘间的负载或者最小化单一磁盘上的文件大小的方差等方法来优化文件请求的平均响应时间等指标。尽管这些解决方案的优势或者可信性已经通过大量的实验获得了证明，但是这些方法可能会具有下面两个不足：首先，现有的文件分配方法及系统不能处理分条的文件。在并行I/O系统中，一般的文件都是首先按照固定的分条大小分成多个子文件，然后这些子文件被分配到多个磁盘上以提供文件内数据的并行读取。其次，这些文件分配方法及系统往往忽略了动态文件访问特性—文件请求的磁盘I/O冲突概率。磁盘I/O冲突概率在适于大数据应用的并行I/O系统的性能优化中扮演着重要的角色。这主要的原因是磁盘的冲突访问会把并行I/O变成顺序I/O，从而导致整个并行I/O系统内的磁盘并发度得不到充分的利用。因此，通过最小化磁盘的I/O冲突访问概率可以进一步的提高并行I/O系统的性能。然而现有的方案和系统总体缺乏一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统。

[0006]

算法名	静态/动态[注1]	文件分配指标	是否面向分条文件	是否考虑磁盘I/O冲突概率	是否考虑负载均衡
Greedy[23,24]	静态/动态	热量[注2]	分条/非分条文件	否	否
SP[26]	静态	文件服务时间	非分条文件	否	否
HP[26]	动态	文件服务时间	非分条文件	否	否
SOR[68]	静态	文件服务时间	非分条文件	否	是
BAS[73]	静态	文件服务时间	非分条文件	否	是
BASB[73]	动态	文件服务时间	非分条文件	否	是
SEA[24]	静态	文件服务时间	分条文件	否	否
PVFS[81]	静态/动态	—	分条文件	一定程度	否

[0007] 表1文件分配方法对比

发明内容

[0008] 1、目的

[0009] 本发明的目的是针对大数据应用挑战，以及现有文件分配方法缺乏一种面向大数据应用的、面向分条文件的、充分考虑最小化大数据文件请求的磁盘I/O冲突概率、采用启发式方法的大数据应用中大规模文件分配系统的问题，发明一种基于最小化IO访问冲突和

文件分条的大规模分条文件分配系统，该分条系统能够处理分条的文件，和最小化文件请求I/O请求冲突概率，最终达到提高整个大数据存储系统的性能的目的。

[0010] 2、技术方案

[0011] 首先给出本发明中所涉及的数学符号说明，$\{d_1,d_2,\cdots,d_n\}$表示n个磁盘，$F=f_1,f_2,\cdots,f_m$表示m个待分配的分块文件。对任意一个文件fi而言，文件的访问信息包括该文件的访问频率λ_i和该文件的大小S_i。q_i表示所有的文件采用的相同的分条宽度。第i个子文件的访问密度为$d_i=\lambda_i/s_i*q_i$。

[0012] 本发明的技术方案如下：

[0013] 一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，主要完成在在多个并行的磁盘内对要被访问的大数据分条文件进行文件的有效分配，具有满足面向大数据分条文件，充分考虑文件请求的磁盘I/O冲突概率需求的技术特征。其具体的特征包括：

[0014] 一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，在体系结构上，位于适于大数据应用的并行文件系统的体系结构(自上而下为，应用层->并行文件系统层->物理硬件层)中的并行文件系统层。在流程上，位于负载均衡处理流程中的最小化I/O访问冲突的文件分配步骤，是基于分布式决策的负载均衡步骤的前提步骤。

[0015] 基于上述体系结构，一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，在模块构成上，该系统主要包括：

[0016] (1)系统初始参数获取模块：主要完成系统初始参数获取并将所获得的参数存储到并行文件系统中的配置文件中。

[0017] (2)文件特征读取模块：主要完成从文件应用层和文件系统支撑层读取系统输入参数。

[0018] (3)文件和磁盘的排序模块：主要完成按照文件的访问密度对文件进行排序以及对并行磁盘按照负载进行排序。

[0019] (4)基于最小化IO访问冲突和文件分条的文件分配执行模块：主要完成执行文件在多个并行的磁盘分配。

[0020] 基于上述体系结构和模块构成，一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，其系统运行流程如下所示：

[0021] 第一步，系统初始参数获取。由系统初始参数获取模块获取初始系统参数。首先，获取的五个参数包括并行磁盘的数量、文件的数目、文件分条大小、文件的访问频率和文件的文件请求大小。然后，系统初始参数获取模块将所获得的参数存储到并行文件系统中的配置文件中。

[0022] 第二步，文件特征读取。首先由文件特征读取模块从文件系统应用层和文件系统支撑层读入输入参数。然后由文件特征读取模块对每个磁盘初始化其访问密度矩阵。

[0023] 第三步，文件和磁盘排序。由第一步和第二步得到的数据，基于最小化IO访问冲突和文件分条对文件和磁盘进行排序。具体计算方法为：首先对每个文件都计算该文件的访问密度。然后对所有文件按照它们的访问密度按降序进行排序，得到一个分条文件按照降序排序得到的文件序列。然后，根据磁盘的负载对磁盘按照升序进行排序，得到一个并行磁盘按照负载大小升序排列的磁盘队列。

[0024] 第四步，基于最小化IO访问冲突和文件分条的文件分配执行。由文件分配模块对所有的文件按照其访问密度的降序采用贪心算法的方式在磁盘上进行分配，即具有最大访问密度的分条文件放到负载最小的硬盘上。

附图说明

[0025] 图1基于最小化IO访问冲突和文件分条的大规模分条文件分配系统架构图

[0026] 图2基于最小化IO访问冲突和文件分条的大规模分条文件分配系统运行流程图

具体实施方式

[0027] 为使本发明的目的、技术方案和优点表达得清楚明白，以PVFS2(一种典型的并行文件系统)和支持分布式负载均衡的应用实例为例，下面结合附图及具体实例对本发明再作进一步详细的说明，但不构成对本发明的限制。具体实施方法如下：

[0028] 如附图1所示，本发明所实施的一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，在体系结构上，位于适于大数据高并发访问的文件级分条系统(PVFS2)的体系结构(自上而下为，应用层->并行文件系统层->物理硬件层)中的并行文件系统层。在流程上，位于PVFS2的负载均衡处理流程中的最小化I/O访问冲突的文件分配(S2)步骤，是基于分布式决策的负载均衡(S3)步骤的前提步骤。

[0029] 基于上述体系结构，一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，如附图1所示，其模块构成的实施方法如下所示：该系统主要包括：

[0030] (1)系统初始参数获取模块：主要完成系统初始参数获取并将所获得的参数存储到并行文件系统中的配置文件中。

[0031] (2)文件特征读取模块：主要完成从文件应用层和文件系统支撑层读取系统输入参数。

[0032] (3)文件和磁盘的排序模块：主要完成按照文件的访问密度对文件进行排序以及对并行磁盘按照负载进行排序。

[0033] (4)基于最小化IO访问冲突和文件分条的文件分配执行模块：主要完成执行文件在多个并行的磁盘分配。

[0034] 基于上述体系结构和模块实施方法，一种基于最小化IO访问冲突和文件分条的大规模分条文件分配系统，如附图2所示，基于PVFS2的运行流程的实施方法如下所示：

[0035] 第一步，系统初始参数获取。由系统初始参数获取模块获取初始系统参数。获取的五个参数包括并行磁盘的数量m、文件的数目n、文件分条大小q、文件的访问频率λ和文件的文件请求大小l。由于本实例中，所实施的系统为已经建成的存储系统，磁盘的数量m是个定值，即当前系统中的磁盘数。文件分条大小q采用PVFS2文件系统通用的默认设置值。文件的数目n由系统初始参数获取模块动态跟踪PVFS2当前文件数目获得，文件的文件请求大小l和文件访问频率λ从PVFS2中所记录的该文件访问的历史信息log文件中获得并输入系统初始参数获取模块。然后，系统初始参数获取模块将所获得的这五个参数存储到并行文件系统中的配置文件中。

[0036] 第二步，文件特征读取。首先由文件特征读取模块从文件系统应用层和文件系统支撑层读入三个输入参数，三个参数为该文件的文件大小s_i、该文件的访问频率λ_i、该文件

访问的分条大小q_i。然后由文件特征读取模块对每个磁盘i初始化其访问密度矩阵$D_i=0$.

[0037] 第三步，文件和磁盘排序。由第一步和第二步得到的数据，对文件和磁盘进行排序。具体计算方法为：首先计算对每个文件根据公式$d_i=\lambda_i/s_i*q_i$(d_i表示第i个文件的访问密度)计算该文件的访问密度。然后对所有文件按照它们的访问密度按降序进行排序，得到一个分条文件按照降序排序得到的文件序列FileQuence=<f_k,f_m,…,f_{fq}>(其中k,m,fq为文件编号)。然后，根据磁盘的负载对磁盘按照升序进行排序，得到一个并行磁盘按照负载大小升序排列的磁盘队列DiskQuence=<d_p,d_i,…,d_{fq}>(其中p,I,fq为磁盘编号)。第四步，基于最小化IO访问冲突和文件分条的文件分配执行。由文件分配模块对所有的文件按照其访问密度的降序采用贪心算法的方式在磁盘上进行分配，即具有最大访问密度的分条文件放到负载最小的硬盘上。更具体的实施措施为：首先将文件队列FileQuence的第一个文件放入磁盘队列DiskQuence第一个磁盘中；然后对磁盘队列DiskQuence进行排序，再选出负载最少的磁盘去存放文件队列FileQuence中第二个文件，以此类推直到所有文件分配完成。该步骤的效果是会尽可能的把属于同一个文件的各个子文件分配到不同的磁盘上。至此，完成基于最小化IO访问冲突和文件分条的文件分配。

[0038] 应说明的是：以上实施例仅用以说明而非限制本发明的技术方案，尽管参照上述实施例对本发明进行了详细说明，本领域的普通技术人员应当理解：依然可以对本发明进行修改或者等同替换，而不脱离本发明的精神和范围的任何修改或局部替换，其均应涵盖在本发明的权利要求范围当中。

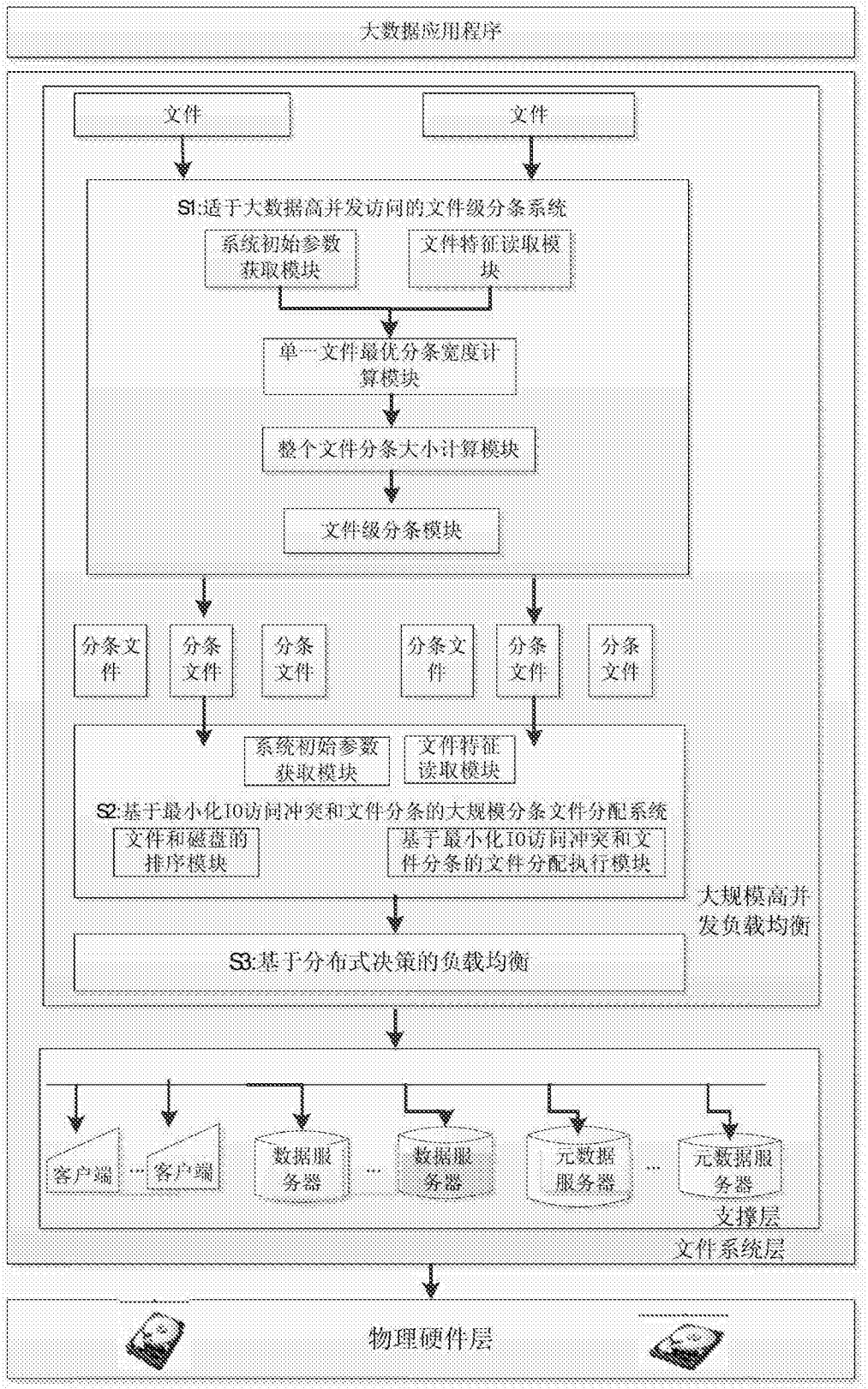

大数据应用程序
文件
文件
S1:适于大数据高并发访问的文件级分条系统
系统初始参数获取模块
文件特征读取模块
单一文件最优分条宽度计算模块
整个文件分条大小计算模块
文件级分条模块
分条文件
分条文件
分条文件
分条文件
分条文件
分条文件
系统初始参数获取模块
文件特征读取模块
S2:基于最小化IO访问冲突和文件分条的大规模分条文件分配系统
文件和磁盘的排序模块
基于最小化IO访问冲突和文件分条的文件分配执行模块
大规模高并发负载均衡
S3:基于分布式决策的负载均衡
客户端
客户端
数据服务器
数据服务器
元数据服务器
元数据服务器
支撑层
文件系统层
物理硬件层

图1

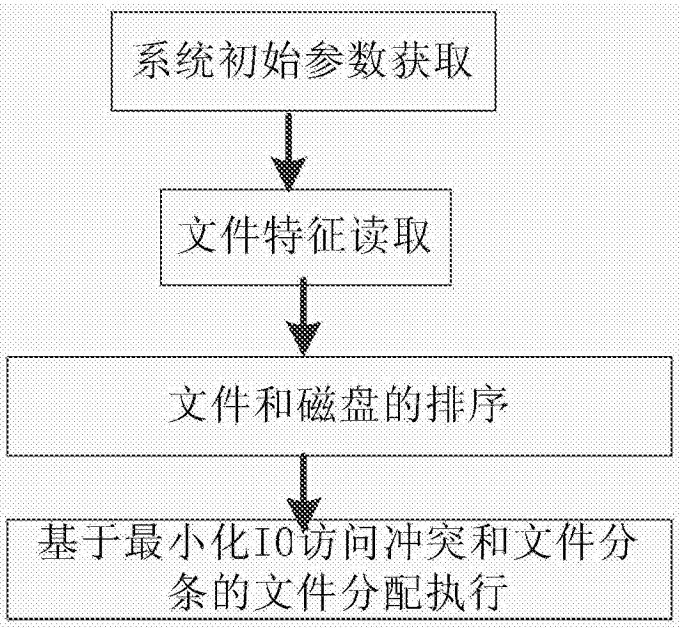


图2