

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2017 年 2 月 23 日 (23.02.2017)

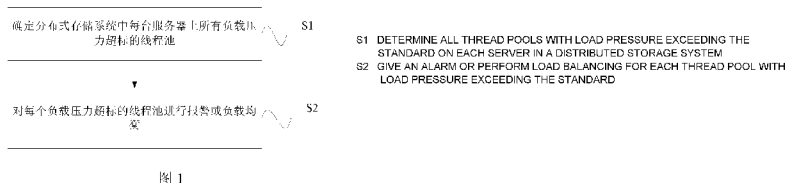


(10) 国际公布号
WO 2017/028696 A1

- (51) 国际专利分类号:
G06F 9/50 (2006.01)
- (21) 国际申请号: PCT/CN2016/093893
- (22) 国际申请日: 2016 年 8 月 8 日 (08.08.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510504654.2 2015 年 8 月 17 日 (17.08.2015) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛英属开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 张潇雨 (ZHANG, Xiaoyu); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。
- (74) 代理人: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (CN)。
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。
- 本国际公布:
— 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD AND DEVICE FOR MONITORING LOAD OF DISTRIBUTED STORAGE SYSTEM

(54) 发明名称: 分布式存储系统的负载监控方法及设备



(57) Abstract: A method and device for monitoring the load of a distributed storage system. The method comprises: determining all thread pools with load pressure exceeding the standard on each server in a distributed storage system (S1); and giving an alarm or performing load balancing for each thread pool with load pressure exceeding the standard (S2). The method and the device are capable of giving an alarm or automatically performing balanced load assignment between servers according to the states of thread pools of the servers, namely, the load excess service ability of the single servers, do not depend on a user request mode, can correctly process a request that is reached by different users at the same time, do not depend on the service capabilities of the servers, and are also capable of correctly giving an alarm or performing load balancing under the condition of inconsistency of the service capabilities of the servers in a cluster of a distributed storage system, thereby preventing the occurrence of a hotspot and improving the service quality of the distributed storage system.

(57) 摘要: 一种分布式存储系统的负载监控方法及设备, 通过确定分布式存储系统中每台服务器上所有负载压力超标的线程池 (S1), 对每个负载压力超标的线程池进行报警或负载均衡 (S2), 能够根据服务器的线程池状态即单台服务器的负载超出服务能力, 来进行报警或自动地在服务器之间均衡分配负载, 不依赖于用户请求模式, 能正确的处理不同用户同时达到的请求, 也不依赖于服务器的服务能力, 在分布式存储系统的集群内部服务器的服务能力不一致的情况下也能正确执行的报警或负载均衡, 从而防止出现热点, 提高分布式存储系统服务质量。

WO 2017/028696 A1

分布式存储系统的负载监控方法及设备

本申请要求 2015 年 08 月 17 日递交的申请号为 201510504654.2、发明名称为“分布式存储系统的负载监控方法及设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及计算机领域，尤其涉及一种分布式存储系统的负载监控方法及设备。

背景技术

分布式存储系统是一种使用集群提供存储服务的分布式系统，用户使用键码（Key）
10 作为索引对相应的键值（Value）进行读写等操作。对于一个键码（Key），用户可以对其进行写入键值（Value）、读取对应的键值（Value）或者删除对应的键值（Value）等不同类型的操作，每个操作称为一个请求。分布式存储系统中的线程池是一种拥有一定数目线程的服务单元，请求先加入线程池的队列中进行等待，线程池中的线程会在空闲的时候依次从队列中取出请求进行处理。分区（Partition）是分布式存储系统调度的基本单元，键码（Key）通过分区（Partition）的开始键码（BeginKey）和结束键码（EndKey）
15 来唯一的确定隶属的分区（Partition），不同的分区（Partition）之间不存在重叠。分布式存储系统中的服务器（Server）是提供服务的基本单元，每个服务器（Server）上拥有若干个分区（Partition），不同键码（Key）的请求根据所属的分区（Partition）的不同从而被不同的服务器（Server）所处理，服务器（Server）内部使用线程池作为实际处理
20 单元来处理不同的请求。

用户的键码（Key）是分成分区（Partition）然后按照顺序存储在分布式文件系统中，由于单个分区只能隶属于一台服务器（Server），因此，单个分区范围内的用户请求增多的时候，会导致服务器的负载增大，用户的延时（Latency）增高等，而且还会对这台服务器上的其他分区造成影响。因此，为了保证充分利用集群内部的所有服务器的服务能力，需要一种负载监控的方案来分散热点，提高服务质量。目前解决请求热点的手段
25 是分区的分裂和迁移。其中，分裂是将分区（Partition）按照不同的键码范围（Keys）分成多个分区（Partition），分裂之后的分区（Partition）会随机分散到其他的服务器（Server）上；迁移是将分区（Partition）从一台服务器（Server）移动到另外一台服务器（Server）上。

30 已有的解决请求热点的手段大致有如下三种：

1. 当单个分区 (Partition) 的大小超过一定的限制之后, 将 Partition 分成平均分裂成若干个分区 (Partition)。但是, 按照单个分区 (Partition) 的大小切分的方案中并不能准确的反映分区 (Partition) 的处理能力, 用户的请求模式不一致, 导致分区 (Partition) 大小对其的影响也不一样, 有时候在很小的分区 (Partition) 大小的情况下也会因为用户请求集中在很小范围内的原因导致热点的出现。

2. 当单个分区 (Partition) 的请求的每秒查询率 (QPS) 大于某个阈值之后, 将分区 (Partition) 按照用户请求的范围进行分裂划分。但是, 按照请求的每秒查询率 (QPS) 作为阈值来划分需要测定不同服务器 (Server) 的处理能力, 因此在不同的服务器 (Server) 上需要配置不同的值, 而且在服务器 (Server) 上还运行有其他的程序情况下有时候并不能达到理论的处理能力。

3. 获取请求执行的一些参数, 比如 IO 操作时间, Cache 命中次数等, 然后配置一定的规则, 当满足预设的条件的时候执行分裂。根据参数配置一定的规则来进行分裂的方案虽然具有很强的灵活性, 但是, 也正是因为这个原因, 在不同的场景下配置的规则也不尽相同, 而且需要根据用户的请求模式更新规则, 不够自动化。

发明内容

本申请的一个目的是提供一种用于分布式存储系统的负载监控方法及设备, 能够解决分布式存储系统的出现热点的问题。

根据本申请的一个方面, 提供了一种分布式存储系统的负载监控方法, 该方法包括:

确定分布式存储系统中每台服务器上所有负载压力超标的线程池;

对每个负载压力超标的线程池进行报警或负载均衡。

进一步的, 上述方法中, 确定分布式存储系统中每台服务器上所有负载压力超标的线程池, 包括:

获取每台服务器上的每个线程池的队列中一个请求的等待时间与逗留时间的比值, 所述逗留时间为每个线程池的队列中一个请求的等待时间与实际处理时间之和;

当所述等待时间与逗留时间的比值超过预设超标阈值时, 确定该请求所在的服务器上的线程池的负载压力超标。

进一步的, 上述方法中, 所述预设超标阈值根据线程池的请求到达率的预设阈值确定, 其中, 所述线程池的请求到达率为请求到达线程池的队列的速率与该线程池的单位时间的服务能力的比值, 当线程池的请求到达率超过线程池的请求到达率的

预设阈值时，对应的等待时间与逗留时间的比值开始急剧上升，所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值。

进一步的，上述方法中，对每个负载压力超标的线程池进行负载均衡，包括：

将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中
5 隶属于不同分区的请求的个数，并将分区按请求的个数降序排列；

判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，

若是，对该请求个数最多的分区进行分裂操作。

进一步的，上述方法中，对该请求个数最多的分区进行分裂操作，包括：

10 将该分区分成若干个子分区，将子分区分散到其它服务器上，其中，每个子分区对应该分区的键码范围内的一个子键码范围，每个子分区隶属的请求个数基本相等。

进一步的，上述方法中，判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半之后，还包括：

15 若否，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；

对选择的分区进行迁移操作。

进一步的，上述方法中，将选择的分区进行迁移操作，包括：

将每个选择的分区迁移到没有负载压力超标的线程池的服务器上。

20 进一步的，上述方法中，将每个选择的分区迁移到没有负载压力超标的线程池的服务器上，包括：

查找符合条件的没有负载压力超标的线程池的服务器，若查找到，将该选择的分区迁移到该查找到的服务器上。

25 进一步的，上述方法中，所述符合条件的没有负载压力超标的线程池的服务器包括：

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器。

30 进一步的，上述方法中，每个线程池的线程平均使用率通过如下公式

$(\lambda_1 + \lambda) * B / n$ 获取, 其中,

λ_1 表示在迁移前服务器上的某个线程池中的请求到达线程池的队列的速率;

λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达线程池的队列的速率;

5 B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间;

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。

根据本申请的另一个方面, 还提供一种分布式存储系统的负载均衡设备, 该设备包括:

10 负载监控装置, 用于确定分布式存储系统中每台服务器上所有负载压力超标的线程池;

报警或负载均衡装置, 用于对每个负载压力超标的线程池进行报警或负载均衡。

进一步的, 上述设备中, 所述负载监控装置, 用于获取每台服务器上的每个线程池的队列中一个请求的等待时间与逗留时间的比值, 所述逗留时间为每个线程池的
15 队列中一个请求的等待时间与实际处理时间之和; 当所述等待时间与逗留时间的比值超过预设超标阈值时, 确定该请求所在的服务器上的线程池的负载压力超标。

进一步的, 上述设备中, 所述预设超标阈值根据线程池的请求到达率的预设阈值确定, 其中, 所述线程池的请求到达率为请求到达线程池的队列的速率与该线程池的单位时间的服务能力的比值, 当线程池的请求到达率超过线程池的请求到达率的
20 预设阈值时, 对应的等待时间与逗留时间的比值开始急剧上升, 所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值。

进一步的, 上述设备中, 所述报警或负载均衡装置, 用于将经过每个负载压力超标的线程池的请求按照分区进行统计, 统计出该线程池中隶属于不同分区的请求的个数, 并将分区按请求的个数降序排列; 判断请求个数最多的分区的请求数量是否超过
25 该线程池的所有分区的请求的总数的一半, 若是, 对该请求个数最多的分区进行分裂操作。

进一步的, 上述设备中, 所述报警或负载均衡装置, 用于将该分区分成若干个子分区, 将子分区分散到其它服务器上, 其中, 每个子分区对应该分区的键码范围内的一个子键码范围, 每个子分区隶属的请求个数基本相等。

30 进一步的, 上述设备中, 所述报警或负载均衡装置, 用于判断请求个数最多的分区

的请求数量是否超过该线程池的所有分区的请求的总数的一半，若否，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；对选择的分区进行迁移操作。

5 进一步的，上述设备中，所述报警或负载均衡装置，用于将每个选择的分区迁移到没有负载压力超标的线程池的服务器上。

进一步的，上述设备中，所述报警或负载均衡装置，用于查找符合条件的没有负载压力超标的线程池的服务器，若查找到，将该选择的分区迁移到该查找到的服务器上。

10 进一步的，上述设备中，所述符合条件的没有负载压力超标的线程池的服务器包括：

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器。

15 进一步的，上述设备中，每个线程池的线程平均使用率通过如下公式 $(\lambda_1 + \lambda) * B / n$ 获取，其中，

λ_1 表示在迁移前服务器上的某个线程池中的请求到达线程池的队列的速率；

λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达线程池的队列的速率；

B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间；

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。

与现有技术相比，本申请通过确定分布式存储系统中每台服务器上所有负载压力超标的线程池，对每个负载压力超标的线程池进行报警或负载均衡，能够根据服务器的线程池状态即单台服务器的负载超出服务能力，来进行报警或自动地在服务器之间均衡分配负载，不依赖于用户请求模式，能正确的处理不同用户同时达到的请求，也不依赖于服务器的服务能力，在分布式存储系统的集群内部服务器的服务能力不一致的情况下也能正确执行的报警或负载均衡，从而防止出现热点，提高分布式存储系统服务质量。

进一步的，本申请通过等待时间 w_q 与逗留时间 w 的比值和预设超标阈值 th 的比较，能够精确地得到每台服务器上所有负载压力超标的线程池。

进一步的，本申请根据线程池的请求到达率的预设阈值确定精确的预设超标阈值，从而能够更精确地得到每台服务器上所有负载压力超标的线程池。

进一步的，本申请将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列，当请求个数最多的分区的请求数量超过该线程池的所有分区的请求的总数的一半时，对该请求个数最多的分区进行分裂操作，可以精确找到需要进行分裂操作的分区，从而有效地实现负载均衡。

进一步的，本申请中当请求个数最多的分区的请求数量未超过该线程池的所有分区的请求的总数的一半时，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半，对选择的分区进行迁移操作，能够精确地找到一个线程池中需要迁移的分区，从而更好地实现负载均衡。

进一步的，本申请在查找符合条件的没有负载压力超标的线程池的服务器的前提下，才将选择的分区迁移到该查找到的服务器上，从而更好地实现负载均衡。

进一步的，本申请中若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器，能够精确地找到符合条件的没有负载压力超标的线程池的服务器，从而更好地实现负载均衡。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

图 1 示出根据本申请一个方面的一种分布式存储系统的负载监控方法的流程图；

图 2 示出本申请分布式存储系统的负载监控方法一优选的实施例的流程图；

图 3 示出根据本申请一个实施例的预设超标阈值确定原理图；

图 4 示出根据本申请分布式存储系统的负载监控方法另一优选的实施例的流程

图；

图 5 示出根据本申请分布式存储系统的负载监控方法再一优选的实施例的流程图；

图 6 示出根据本申请分布式存储系统的负载监控方法一具体应用实施例的流程图；

图 7 示出根据本申请另一个方面的分布式存储系统的负载监控设备的结构图。

附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

下面结合附图对本发明作进一步详细描述。

10 在本申请一个典型的配置中，终端、服务网络的设备和可信方均包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器 (RAM) 和/或非易失性内存等形式，如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

15 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆
20 体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体 (transitory media)，如调制的数据信号和载波。

如图 1 所示，本申请提供一种分布式存储系统的负载监控方法，该方法包括：

25 步骤 S1，确定分布式存储系统中每台服务器上所有负载压力超标的线程池；

步骤 S2，对每个负载压力超标的线程池进行报警或负载均衡。本实施例中根据服务器 (Server) 的线程池状态即单台服务器 (Server) 的负载超出服务能力，来进行报警或自动地在服务器之间均衡分配负载，不依赖于用户请求模式，能正确的处理不同用户同时达到的请求，也不依赖于服务器的服务能力，在分布式存储系统的
30 集群内部服务器的服务能力不一致的情况下也能正确执行的报警或负载均衡，从而

防止出现热点，提高分布式存储系统服务质量。具体的，可根据服务器（Server）的线程池基本信息，确定分布式存储系统中每台服务器上所有负载压力超标的线程池，或对每个负载压力超标的线程池进行报警或负载均衡。单个服务器（Server）上一般拥有若干个线程池，典型的线程池可以用一个单队列模型来描述，线程池基
5 本信息可体现为队列参数，具体的队列参数可包括如下内容：

- a) W_q ：表示某个线程池的队列中一个请求的等待时间。
- b) B ：表示某个线程池的队列中一个请求的实际处理时间。
- c) W ：表示某个线程池的队列中一个请求的逗留时间，即等待时间 W_q 加上实际处理时间 B 。
- 10 d) λ ：表示某个线程池中的请求到达该线程池的队列的速率。
- e) μ ：表示某个线程池的单位时间的服务能力。

如图 2 所示，本申请的分布式存储系统的负载监控方法一优选的实施例中，步骤 S1，确定分布式存储系统中每台服务器上所有负载压力超标的线程池，包括：

步骤 S11，获取每台服务器上的每个线程池的队列中一个请求的等待时间 W_q 与
15 逗留时间 W 的比值，所述逗留时间 W 为每个线程池的队列中一个请求的等待时间 W_q 与实际处理时间 B 之和；

步骤 S12，当所述等待时间 W_q 与逗留时间 W 的比值超过预设超标阈值 th 时，确定该请求所在的服务器上的线程池的负载压力超标。对每台服务上的每个线程池参数进行分析，找出负载压力超标的线程池，关于如何判断超标负载的线程池，可以
20 使用如下公式判断：

$$W_q / W > th$$

该公式的含义是当所述等待时间 W_q 与逗留时间 W 的比值超过预设超标阈值 th 时，确定该请求所在的服务器上的线程池的负载压力超标。本实施例通过等待时间 W_q 与逗留时间 W 的比值和预设超标阈值 th 的比较，能够精确地得到每台服务器上所有
25 有负载压力超标的线程池。

本申请的分布式存储系统的负载监控方法一优选的实施例中，步骤 S12 中所述预设超标阈值根据线程池的请求到达率的预设阈值确定，其中，所述线程池的请求到达率为请求到达线程池的队列的速率 λ 与该线程池的单位时间的服务能力 μ 的比值，当线程池的请求到达率 λ / μ 超过线程池的请求到达率的预设阈值时，对应的等

待时间与逗留时间的比值 W_q/W 开始急剧上升, 所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值 W_q/W 。具体的, 线程池的请求到达率的预设阈值可以通过图 3 来判断, 图 3 中, 每一条线表一个线程池, n 分别表示对应的线程池中的线程个数, 其中, 第一个线程池中的线程有 1 个, 第二个线程池中的线程有 2 个, 第三个线程池中的线程有 3 个, 第四个线程池中的线程有 10 个, 第五个线程池中的线程有 24 个, 图 3 中横坐标表示请求到达率 λ/μ , 纵坐标表示等待时间 W_q 与逗留时间 W 的比值 W_q/W 。从图 3 中, 可以看出 W_q/W 在请求到达率 λ/μ 超过一定的数值 (线程池的请求到达率的预设阈值) 之后就开始急剧上升, W_q/W 的这一开始急剧上升的点的值即为拐点, 因此 W_q/W 的预设超标阈值只要超过拐点即可, 例如, 实际中可将拐点的值设定为 0.5, 则所述预设超标阈值只要大于 0.5 即可。本实施例根据线程池的请求到达率的预设阈值确定精确的预设超标阈值, 从而能够更精确地得到每台服务器上所有负载压力超标的线程池。

如图 4 所示, 本申请的分布式存储系统的负载监控方法一优选的实施例中, 步骤 S2 中的对每个负载压力超标的线程池进行负载均衡, 包括:

步骤 S21, 将经过每个负载压力超标的线程池的请求按照分区进行统计, 统计出该线程池中隶属于不同分区的请求的个数, 并将分区按请求的个数降序排列;

步骤 S22, 判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半, 若是, 转到步骤 S23,

步骤 S23, 对该请求个数最多的分区进行分裂操作。具体的, 例如, 某个负载压力超标的线程池中有三个分区分别为分区 A、分区 B 和分区 C, 其中, 隶属于分区 A 的请求数为 100, 隶属于分区 B 的请求数为 20, 隶属于分区 C 的请求数为 10, 则求个数最多的分区 A 的请求数量 100 超过了该线程池的所有分区的请求的总数的一半 $65 = (100 + 20 + 10) / 2$ 。本实施例中, 对于每个负载压力超标的线程池提取请求信息进行分析, 因为线程池和分区 (Partition) 并不是一一对应的, 分区 (Partition) 只是一个逻辑单元, 隶属于该分区 (Partition) 的请求可能会使用多个线程池进行处理。将经过该线程池的请求按照分区 (Partition) 进行统计, 统计出隶属于不同分区 (Partition) 的请求个数, 然后按照请求个数降序排列。如果一个线程池中请求数目最多的某个分区 (Partition) 请求数目超过了该线程池的所有分区的请求的总数的一半, 那么选择将该分区 (Partition) 并跳转到步骤 S23, 步骤 S23 中, 因为隶属于该

选择的分区 (Partition) 的请求数目占了该线程池的所有分区的请求的总数的一半或一半以上, 因此, 需要对该选择的分区 (Partition) 进行分裂, 分裂之后处理结束。本实施例中通过找到超过该线程池的所有分区的请求的总数的一半的请求个数最多的分区, 该分区可以认为是该线程池中的对负载压力超标有显著影响的分区, 因此, 需要将其选择出来, 对其进行分裂, 从而有效地实现负载均衡。

本申请的分布式存储系统的负载监控方法一优选的实施例中, 步骤 S23, 对该请求个数最多的分区进行分裂操作, 包括:

将该分区分成若干个子分区, 将子分区分散到其它服务器上, 其中, 每个子分区对应该分区的键码范围内的一个子键码范围, 每个子分区隶属的请求个数基本相等。具体的, 分裂点是从该分区 (Partition) 落在该线程池上的请求排序之后按照分区范围内的请求数平均选取, 例如, 某个分区范围的键码范围是 0.1~0.4, 其中, 0.1~0.2 范围的请求个数有 200 个, 0.2~0.3 范围的请求个数有 200 个, 0.3~0.4 范围的请求个数有 200 个, 则可以将该分区分成三个子分区, 对应的子键码范围分别为 0.1~0.2、0.2~0.3 和 0.3~0.4, 从而更好地实现负载均衡。

如图 5 所示, 本申请的分布式存储系统的负载监控方法一优选的实施例中, 步骤 S22, 判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半之后, 还包括:

若否, 则转到步骤 S24,

步骤 S24, 在所述降序排列的分区中从第一分区开始依次选择一个或多个分区, 直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半;

步骤 S25, 对选择的分区进行迁移操作。在此, 若请求个数最多的分区未超过该线程池的所有分区的请求的总数的一半, 则将从该线程池中降序排列的第一个分区 (Partition) 开始依次往后选择分区 (Partition) 直到剩下的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半, 并跳转到步骤 S25, 步骤 S25 中, 因为, 请求个数最多的分区未超过该线程池的所有分区的请求的总数的一半, 所以该线程池的请求中没有具有显著影响的分区 (Partition), 因此将从第一分区开始依次选择出来的分区 (Partition) 进行逐个迁移。例如, 某个负载压力超标的线程池中有五个分区分别为分区 D、分区 E、分区 F、分区 G 和分区 H, 其中, 隶属于分区 D 的请求数为 100, 隶属于分区 E 的请求数为 100, 隶属于分区 F 的请求数为

100, 隶属于分区 G 的请求数为 100, 隶属于分区 H 的请求数为 100, 则需要选择前三个分区 D、E 和 F 进行逐个迁移, 从而剩下的分区 G 和 H 所隶属的请求的总数 $200=100+100$ 小于该线程池的所有分区的请求的总数的一半 $250=(100+100+100+100+100)/2$ 。本实施例能够精确地找到一个线程池中需要迁移的

5 分区, 从而更好地实现负载均衡。

本申请的分布式存储系统的负载监控方法一优选的实施例中, 将选择的分区进行迁移操作, 包括:

将每个选择的分区迁移到没有负载压力超标的线程池的服务器上, 从而实现负载均衡。

10 本申请的分布式存储系统的负载监控方法一优选的实施例中, 将每个选择的分区迁移到没有负载压力超标的线程池的服务器上, 包括:

查找符合条件的没有负载压力超标的线程池的服务器, 若查找到, 将该选择的分区迁移到该查找到的服务器上。本实施例在查找符合条件的没有负载压力超标的线程池的服务器的前提下, 才将选择的分区迁移到该查找到的服务器上, 从而更好地实现负载均衡。

15

本申请的分布式存储系统的负载监控方法一优选的实施例中, 所述符合条件的没有负载压力超标的线程池的服务器包括:

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后, 当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有

20 超过预设使用率阈值, 则该服务器为符合条件的没有负载压力超标的线程池的服务器。具体的, 从待迁移集合即所有选择的分区中取一个分区 (Partition), 获取其在目前服务器上所有线程池上的请求数目, 例如, 服务器 M 上某个选择的分区使用两个线程池处理隶属于其的请求, 分别为读请求线程池 Q1 和写请求线程池 Q2, 在没有超负载线程池的服务器集合中随机选择一个, 例如选择了服务器 N, 然后计算该

25 选择的分区 (Partition) 的读请求线程池 Q1 和写请求线程池 Q2 分别迁移到服务器 N 上对应的读请求线程池 Q1+和写请求线程池 Q2+之后, 即将读请求线程池 Q1 迁移到读请求线程池 Q1+, 及将写请求线程池 Q2 迁移到写请求线程池 Q2+之后, 服务器 N 上读请求线程池 Q1+和写请求线程池 Q2+的线程平均使用率均没有超过预设使用率阈值, 预设使用率阈值可以是一经验值, 即如果迁移之后所有的对应线程池

30 的使用率都不超过预设使用率阈值, 则判定允许迁移, 否则选择另外一台没有负载

压力超标的线程池的服务器重复本过程，直到所有的没有负载压力超标的线程池的服务器都已经检查完毕，没有找到符合条件的没有负载压力超标的线程池的服务器，即放弃迁移该分区（Partition）。本实施例能够精确地找到符合条件的没有负载压力超标的线程池的服务器，从而更好地实现负载均衡。

5 本申请的分布式存储系统的负载监控方法一优选的实施例中，每个线程池的线程平均使用率通过如下公式 $(\lambda_1 + \lambda) * B / n$ 获取，其中，

λ_1 表示在迁移前服务器上的某个线程池中的请求到达该线程池的队列的速率；

λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达该线程池的队列的速率；

10 B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间；

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。具体的，以上述迁移前的服务器 M 和待迁移到的目标服务器 N 为例，需要分别计算读请求线程池 $Q1+$ 和写请求线程池 $Q2+$ 的线程平均使用率是否均没有超过预设使用率阈值，其中，读请求线程池 $Q1+$ 的线程平均使用率计算公式为 $(\lambda_{Q1} + \lambda_{Q1+}) * B_{Q1+} / n_{Q1+}$ ，写请求线程池 $Q2+$ 的线程平均使用率的计算公式为 $(\lambda_{Q2} + \lambda_{Q2+}) * B_{Q2+} / n_{Q2+}$ ，本实施例可以精确地计算得到每个线程池的线程平均使用率，从而更好地实现负载均衡。

如图 6 所示，本申请一具体的应用实例中，分布式存储系统的负载监控方法包括如下步骤：

20 步骤 S61，获取分布式存储系统中未处理的一个线程池，并判断是否获取到，若未获取到，则转到步骤 S62，若获取到，则转到步骤 S63，

步骤 S62，结束；

步骤 S63，获取该线程池的队列中一个请求的等待时间 w_q 与逗留时间 W 的比值，所述逗留时间 W 为该线程池的队列中一个请求的等待时间 w_q 与实际处理时间 B

25 之和；

步骤 S64，判断所述等待时间 w_q 与逗留时间 W 的比值是否超过预设超标阈值 th ，若否，转到步骤 S61，若是，转到步骤 S65，

步骤 S65，将经过该线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列，判断请求个数最多的

分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，若是，转到步骤 S66，若否，转到步骤 S67，

步骤 S66，对该请求个数最多的分区进行分裂操作；

步骤 S67，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；

步骤 S68，判断所述选择的分区中是否有未处理的分区，若有，则转到步骤 S69，若无，则转到步骤 S61，

步骤 S69，取下一个未处理的分区；

10 步骤 S70，查找符合条件的没有负载压力超标的线程池的服务器，并判断是否查找到，若未查找到，则转到步骤 S68，以从所述选择的分区中取下一个未处理的分区并进行后续处理，直到所有选择的分区（Partition）都处理完毕，若查找到，则转到步骤 S71，

步骤 S71，将该选择的分区迁移到该查找到的服务器上后，转到步骤 S68。

15 如图 7 所示，根据本申请的另一面，还提供一种分布式存储系统的负载监控设备，该设备 100 包括：

负载监控装置 1，用于确定分布式存储系统中每台服务器上所有负载压力超标的线程池；

20 报警或负载均衡装置 2，用于对每个负载压力超标的线程池进行报警或负载均衡。本实施例中根据服务器（Server）的线程池状态即单台服务器（Server）的负载超出服务能力，来进行报警或自动地在服务器之间均衡分配负载，不依赖于用户请求模式，能正确的处理不同用户同时达到的请求，也不依赖于服务器的服务能力，在分布式存储系统的集群内部服务器的服务能力不一致的情况下也能正确执行的报警或负载均衡，从而防止出现热点，提高分布式存储系统服务质量。具体的，可根据服务器（Server）的线程池基本信息，确定分布式存储系统中每台服务器上所有负载压力超标的线程池，或对每个负载压力超标的线程池进行报警或负载均衡。
25 单个服务器（Server）上一般拥有若干个线程池，典型的线程池可以用一个单队列模型来描述，线程池基本信息可体现为队列参数，具体的队列参数可包括如下内容：

a) W_q ：表示某个线程池的队列中一个请求的等待时间。

30 b) B ：表示某个线程池的队列中一个请求的实际处理时间。

- c) W ：表示某个线程池的队列中一个请求的逗留时间，即等待时间 W_q 加上实际处理时间 B 。
- d) λ ：表示某个线程池中的请求到达该线程池的队列的速率。
- e) μ ：表示某个线程池的单位时间的服务能力。

5 本申请分布式存储系统的负载监控设备一优选的实施例中，所述负载监控装置 1，用于获取每台服务器上的每个线程池的队列中一个请求的等待时间 W_q 与逗留时间 W 的比值，所述逗留时间 W 为每个线程池的队列中一个请求的等待时间 W_q 与实际处理时间 B 之和；当所述等待时间 W_q 与逗留时间 W 的比值超过预设超标阈值 th 时，确定该请求所在的服务器上的线程池的负载压力超标。对每台服务上的每个线程池
10 参数进行分析，找出负载压力超标的线程池，关于如何判断超标负载的线程池，可以使用如下公式判断：

$$W_q / W > th$$

该公式的含义是当所述等待时间 W_q 与逗留时间 W 的比值超过预设超标阈值 th 时，确定该请求所在的服务器上的线程池的负载压力超标。本实施例通过等待时间
15 W_q 与逗留时间 W 的比值和预设超标阈值 th 的比较，能够精确地得到每台服务器上所有负载压力超标的线程池。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述预设超标阈值根据线程池的请求到达率的预设阈值确定，其中，所述线程池的请求到达率为请求到达线程池的队列的速率与该线程池的单位时间的服务能力 μ 的比值，当线程池的
20 请求到达率超过线程池的请求到达率 λ / μ 的预设阈值时，对应的等待时间与逗留时间的比值 W_q / W 开始急剧上升，所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值 W_q / W 。每一条线表一个线程池， n 分别表示对应的线程池中的线程个数，其中，第一个线程池中的线程有 1 个，第二个线程池中的线程有 2 个，第三个线程池中的线程有 5 个，第四个线程池中的线程有 10 个，第五个线程池中的
25 线程有 24 个，图 3 中横坐标表示请求到达率 λ / μ ，纵坐标表示等待时间 W_q 与逗留时间 W 的比值 W_q / W 。从图 3 中，可以看出 W_q / W 在请求到达率 λ / μ 超过一定的数值（线程池的请求到达率的预设阈值）之后就开始急剧上升， W_q / W 的这一开始急剧上升的点的值即为拐点，因此 W_q / W 的预设超标阈值只要超过拐点即可，例如，

实际中可将拐点的值设定为 0.5，则所述预设超标阈值只要大于 0.5 即可。本实施例根据线程池的请求到达率的预设阈值确定精确的预设超标阈值，从而能够更精确地得到每台服务器上所有负载压力超标的线程池。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述报警或负载均衡装置 2，用于将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列；判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，若是，对该请求个数最多的分区进行分裂操作。具体的，例如，某个负载压力超标的线程池中有三个分区分别为分区 A、分区 B 和分区 C，其中，隶属于分区 A 的请求数为 100，隶属于分区 B 的请求数为 20，隶属于分区 C 的请求数为 10，则请求个数最多的分区 A 的请求数量 100 超过了该线程池的所有分区的请求的总数的一半 $65 = (100 + 20 + 10) / 2$ 。本实施例中，对于每个负载压力超标的线程池提取请求信息进行分析，因为线程池和分区 (Partition) 并不是一一对应的，分区 (Partition) 只是一个逻辑单元，隶属于该分区 (Partition) 的请求可能会使用多个线程池进行处理。将经过该线程池的请求按照分区 (Partition) 进行统计，统计出隶属于不同分区 (Partition) 的请求个数，然后按照请求个数降序排列。如果一个线程池中请求数目最多的某个分区 (Partition) 请求数目超过了该线程池的所有分区的请求的总数的一半，那么选择将该分区 (Partition) 并跳转到步骤 S23，步骤 S23 中，因为隶属于该选择的分区 (Partition) 的请求数目占了该线程池的所有分区的请求的总数的一半或一半以上，因此，需要对该选择的分区 (Partition) 进行分裂，分裂之后处理结束。本实施例中通过找到超过该线程池的所有分区的请求的总数的一半的请求个数最多的分区，该分区可以认为是该线程池中的对负载压力超标有显著影响的分区，因此，需要将其选择出来，对其进行分裂，从而有效地实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述报警或负载均衡装置 2，用于将该分区分成若干个子分区，将子分区分散到其它服务器上，其中，每个子分区对应该分区的键码范围内的一个子键码范围，每个子分区隶属的请求个数基本相等。具体的，分裂点是从该分区 (Partition) 落在该线程池上的请求排序之后按照分区范围内的请求数平均选取，例如，某个分区范围的键码范围是 0.1~0.4，其中，0.1~0.2 范围的请求个数有 200 个，0.2~0.3 范围的请求个数有 200 个，0.3~0.4 范围的请求个数有 200 个，则可以将该分区分成三个子分区，对应的子键码范

围分别为 0.1~0.2、0.2~0.3 和 0.3~0.4，从而更好地实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述报警或负载均衡装置 2，用于判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，若否，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；对选择的分区进行迁移操作。在此，若请求个数最多的分区未超过该线程池的所有分区的请求的总数的一半，则将从该线程池中降序排列的第一个分区（Partition）开始依次往后选择分区（Partition）直到剩下的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半，并对选择的分区进行迁移操作，因为，请求个数最多的分区未超过该线程池的所有分区的请求的总数的一半，所以该线程池的请求中没有具有显著影响的分区（Partition），因此将从第一分区开始依次选择出来的分区（Partition）进行逐个迁移。例如，某个负载压力超标的线程池中有五个分区分别为分区 D、分区 E、分区 F、分区 G 和分区 H，其中，隶属于分区 D 的请求数为 100，隶属于分区 E 的请求数为 100，隶属于分区 F 的请求数为 100，隶属于分区 G 的请求数为 100，隶属于分区 H 的请求数为 100，则需要选择前三个分区 D、E 和 F 进行逐个迁移，从而剩下的分区 G 和 H 所隶属的请求的总数 $200=100+100$ 小于该线程池的所有分区的请求的总数的一半 $250=(100+100+100+100+100)/2$ 。本实施例能够精确地找到一个线程池中需要迁移的分区，从而更好地实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述报警或负载均衡装置 2，用于将每个选择的分区迁移到没有负载压力超标的线程池的服务器上，从而实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述报警或负载均衡装置 2，用于查找符合条件的没有负载压力超标的线程池的服务器，若查找到，将该选择的分区迁移到该查找到的服务器上。本实施例在查找符合条件的没有负载压力超标的线程池的服务器的前提下，才将选择的分区迁移到该查找到的服务器上，从而更好地实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，所述符合条件的没有负载压力超标的线程池的服务器包括：

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应

线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器。具体的，从待迁移集合即所有选择的分区中取一个分区（Partition），获取其在目前服务器上所有线程池上的请求数目，例如，服务器 M 上某个选择的分区使用两个线程池处理隶属于其的请求，分别为读请求线程池 Q1 和写请求线程池 Q2，在没

5 有超负载线程池的服务器集合中随机选择一个，例如选择了服务器 N，然后计算该选择的分区（Partition）的读请求线程池 Q1 和写请求线程池 Q2 分别迁移到服务器 N 上对应的读请求线程池 Q1+和写请求线程池 Q2+之后，即将读请求线程池 Q1 迁移到读请求线程池 Q1+，及将写请求线程池 Q2 迁移到写请求线程池 Q2+之后，服

10 务器 N 上读请求线程池 Q1+和写请求线程池 Q2+的线程平均使用率均没有超过预设使用率阈值，预设使用率阈值可以是一经验值，即如果迁移之后所有的对应线程池的使用率都不超过预设使用率阈值，则判定允许迁移，否则选择另外一台没有负载压力超标的线程池的服务器重复本过程，直到所有的没有负载压力超标的线程池的服务器都已经检查完毕，没有找到符合条件的没有负载压力超标的线程池的服

15 器，即放弃迁移该分区（Partition）。本实施例能够精确地找到符合条件的没有负载压力超标的线程池的服务器，从而更好地实现负载均衡。

本申请分布式存储系统的负载监控设备一优选的实施例中，每个线程池的线程平均使用率通过如下公式 $(\lambda_1 + \lambda) * B / n$ 获取，其中，

λ_1 表示在迁移前服务器上的某个线程池中的请求到达线程池的队列的速率；

20 λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达线程池的队列的速率；

B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间；

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。具体的，以上

25 述迁移前的服务器 M 和待迁移到的目标服务器 N 为例，需要分别计算读请求线程池 Q1+和写请求线程池 Q2+的线程平均使用率是否均没有没有超过预设使用率阈值，其中，读请求线程池 Q1+的线程平均使用率计算公式为 $(\lambda_{Q1} + \lambda_{Q1+}) * B_{Q1+} / n_{Q1+}$ ，写请求线程池 Q2+的线程平均使用率的计算公式为 $(\lambda_{Q2} + \lambda_{Q2+}) * B_{Q2+} / n_{Q2+}$ ，本实施例可以精确地计算得到每个线程池的线程平均使用率，从而更好地实现负载均衡。

综上所述，本申请通过确定分布式存储系统中每台服务器上所有负载压力超标的线程池，对每个负载压力超标的线程池进行报警或负载均衡，能够根据服务器的线程池状态即单台服务器的负载超出服务能力，来进行报警或自动地在服务器之间均衡分配负载，不依赖于用户请求模式，能正确的处理不同用户同时达到的请求，
5 也不依赖于服务器的服务能力，在分布式存储系统的集群内部服务器的服务能力不一致的情况下也能正确执行的报警或负载均衡，从而防止出现热点，提高分布式存储系统服务质量。

进一步的，本申请通过等待时间 w_q 与逗留时间 w 的比值和预设超标阈值 th 的比较，能够精确地得到每台服务器上所有负载压力超标的线程池。

10 进一步的，本申请根据线程池的请求到达率的预设阈值确定精确的预设超标阈值，从而能够更精确地得到每台服务器上所有负载压力超标的线程池。

进一步的，本申请将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列，当请求个数最多的分区的请求数量超过该线程池的所有分区的请求的总数的
15 一半时，对该请求个数最多的分区进行分裂操作，可以精确找到需要进行分裂操作的分区，从而有效地实现负载均衡。

进一步的，本申请中当请求个数最多的分区的请求数量未超过该线程池的所有分区的请求的总数的一半时，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有
20 分区的请求的总数的一半，对选择的分区进行迁移操作，能够精确地找到一个线程池中需要迁移的分区，从而更好地实现负载均衡。

进一步的，本申请在查找符合条件的没有负载压力超标的线程池的服务器的前提下，才将选择的分区迁移到该查找到的服务器上，从而更好地实现负载均衡。

进一步的，本申请中若将某个选择的分区迁移到某个没有负载压力超标的线程
25 池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器，能够精确地找到符合条件的没有负载压力超标的线程池的服务器，从而更好地实现负载均衡。

显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技
30

术的范围之内，则本申请也意图包含这些改动和变型在内。

需要注意的是，本发明可在软件和/或软件与硬件的组合体中被实施，例如，可采用专用集成电路（ASIC）、通用目的计算机或任何其他类似硬件设备来实现。在一个实施例中，本发明的软件程序可以通过处理器执行以实现上文所述步骤或功能。同样地，本发明的软件程序（包括相关的数据结构）可以被存储到计算机可读记录介质中，例如，RAM 存储器，磁或光驱动器或软磁盘及类似设备。另外，本发明的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

另外，本发明的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本发明的方法和/或技术方案。而调用本发明的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据所述程序指令运行的计算机设备的工作存储器中。在此，根据本发明的一个实施例包括一个装置，该装置包括用于存储计算机程序指令的存储器和用于执行程序指令的处理器，其中，当该计算机程序指令被该处理器执行时，触发该装置运行基于前述根据本发明的多个实施例的方法和/或技术方案。

对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。装置权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

权 利 要 求 书

1. 一种分布式存储系统的负载监控方法，其中，该方法包括：

确定分布式存储系统中每台服务器上所有负载压力超标的线程池；

对每个负载压力超标的线程池进行报警或负载均衡。

5 2. 如权利要求 1 所述的方法，其中，确定分布式存储系统中每台服务器上所有负载压力超标的线程池，包括：

获取每台服务器上的每个线程池的队列中一个请求的等待时间与逗留时间的比值，所述逗留时间为每个线程池的队列中一个请求的等待时间与实际处理时间之和；

10 当所述等待时间与逗留时间的比值超过预设超标阈值时，确定该请求所在的服务器上的线程池的负载压力超标。

3. 如权利要求 2 所述的方法，其中，所述预设超标阈值根据线程池的请求到达率的预设阈值确定，其中，所述线程池的请求到达率为请求到达线程池的队列的速率与该线程池的单位时间的服务能力的比值，当线程池的请求到达率超过线程池的请求到达率的预设阈值时，对应的等待时间与逗留时间的比值开始急剧上升，所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值。

4. 如权利要求 1 至 3 任一项所述的方法，其中，对每个负载压力超标的线程池进行负载均衡，包括：

20 将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列；

判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，

若是，对该请求个数最多的分区进行分裂操作。

25 5. 如权利要求 4 所述的方法，其中，对该请求个数最多的分区进行分裂操作，包括：

将该分区分成若干个子分区，将子分区分散到其它服务器上，其中，每个子分区对应该分区的键码范围内的一个子键码范围，每个子分区隶属的请求个数基本相等。

6. 如权利要求 4 所述的方法，其中，判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半之后，还包括：

30 若否，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直

至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；

对选择的分区进行迁移操作。

7. 如权利要求 6 所述的方法，其中，将选择的分区进行迁移操作，包括：

5 将每个选择的分区迁移到没有负载压力超标的线程池的服务器上。

8. 如权利要求 7 所述的方法，其中，将每个选择的分区迁移到没有负载压力超标的线程池的服务器上，包括：

查找符合条件的没有负载压力超标的线程池的服务器，若查找到，将该选择的分区迁移到该查找到的服务器上。

10 9. 如权利要求 8 所述的方法，其中，所述符合条件的没有负载压力超标的线程池的服务器包括：

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务器。

10. 如权利要求 9 所述的方法，其中，每个线程池的线程平均使用率通过如下公式 $(\lambda_1 + \lambda) * B / n$ 获取，其中，

λ_1 表示在迁移前服务器上的某个线程池中的请求到达线程池的队列的速率；

20 λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达线程池的队列的速率；

B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间；

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。

11. 一种设备分布式存储系统的负载监控设备，其中，该设备包括：

25 负载监控装置，用于确定分布式存储系统中每台服务器上所有负载压力超标的线程池；

报警或负载均衡装置，用于对每个负载压力超标的线程池进行报警或负载均衡。

12. 如权利要求 11 所述的设备，其中，所述负载监控装置，用于获取每台服务器上的每个线程池的队列中一个请求的等待时间与逗留时间的比值，所述逗留时间为
30 每个线程池的队列中一个请求的等待时间与实际处理时间之和；当所述等待时间与

逗留时间的比值超过预设超标阈值时，确定该请求所在的服务器上的线程池的负载压力超标。

13. 如权利要求 12 所述的设备，其中，所述预设超标阈值根据线程池的请求到达率的预设阈值确定，其中，所述线程池的请求到达率为请求到达线程池的队列的速率与该线程池的单位时间的服务能力的比值，当线程池的请求到达率超过线程池的请求到达率的预设阈值时，对应的等待时间与逗留时间的比值开始急剧上升，所述预设超标阈值超过所述开始急剧上升时等待时间与逗留时间的比值。

14. 如权利要求 11 至 13 任一项所述的设备，其中，所述报警或负载均衡装置，用于将经过每个负载压力超标的线程池的请求按照分区进行统计，统计出该线程池中隶属于不同分区的请求的个数，并将分区按请求的个数降序排列；判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，若是，对该请求个数最多的分区进行分裂操作。

15. 如权利要求 14 所述的设备，其中，所述报警或负载均衡装置，用于将该分区分成若干个子分区，将子分区分散到其它服务器上，其中，每个子分区对应该分区的键码范围内的一个子键码范围，每个子分区隶属的请求个数基本相等。

16. 如权利要求 14 所述的设备，其中，所述报警或负载均衡装置，用于判断请求个数最多的分区的请求数量是否超过该线程池的所有分区的请求的总数的一半，若否，在所述降序排列的分区中从第一分区开始依次选择一个或多个分区，直至未选择的剩余的分区所隶属的请求的总数小于该线程池的所有分区的请求的总数的一半；对选择的分区进行迁移操作。

17. 如权利要求 16 所述的设备，其中，所述报警或负载均衡装置，用于将每个选择的分区迁移到没有负载压力超标的线程池的服务器上。

18. 如权利要求 17 所述的设备，其中，所述报警或负载均衡装置，用于查找符合条件的没有负载压力超标的线程池的服务器，若查找到，将该选择的分区迁移到该查找到的服务器上。

19. 如权利要求 18 所述的设备，其中，所述符合条件的没有负载压力超标的线程池的服务器包括：

若将某个选择的分区迁移到某个没有负载压力超标的线程池的服务器的对应线程池上后，当该迁移到的目标服务器的每个对应线程池的线程平均使用率均没有超过预设使用率阈值，则该服务器为符合条件的没有负载压力超标的线程池的服务

器。

20. 如权利要求 19 所述的设备, 其中, 每个线程池的线程平均使用率通过如下公式 $(\lambda_1 + \lambda) * B / n$ 获取, 其中,

λ_1 表示在迁移前服务器上的某个线程池中的请求到达线程池的队列的速率;

5 λ 表示待迁移到的目标服务器在迁移前其上的某个对应线程池中的请求到达线程池的队列的速率;

B 表示待迁移到的目标服务器的某个对应线程池中的每个线程对一个请求的实际处理时间;

n 表示待迁移到的目标服务器的某个对应线程池中的线程个数。

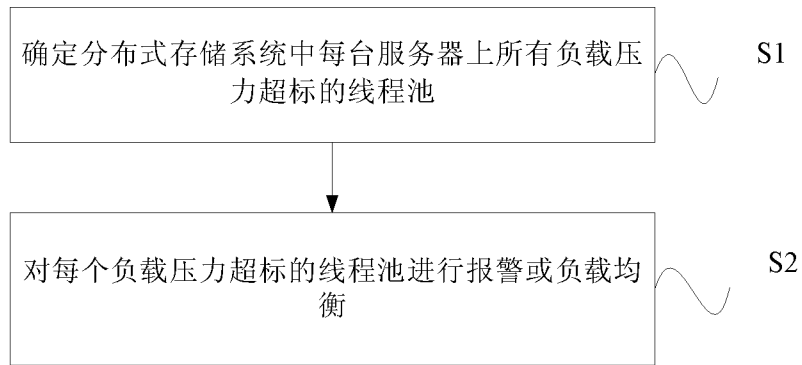


图 1

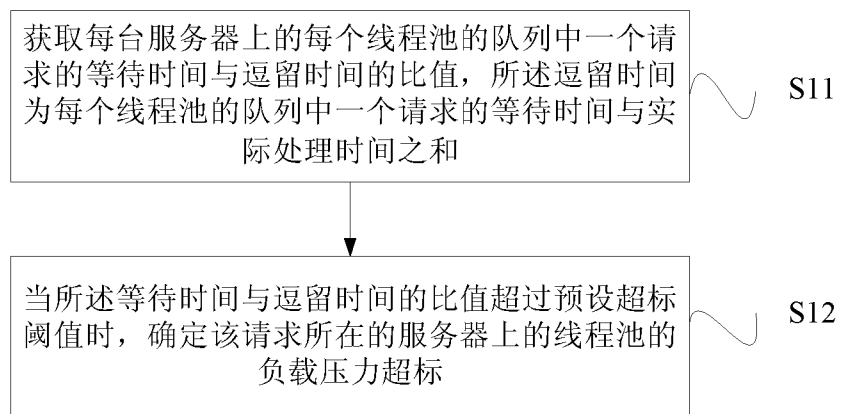


图 2

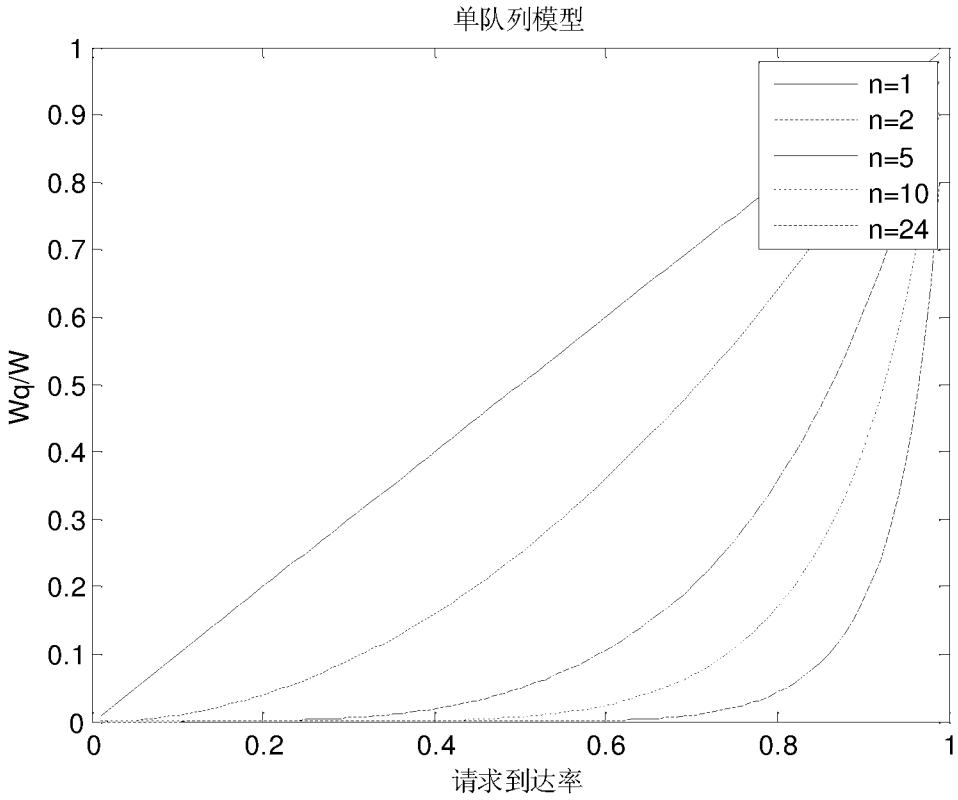


图 3

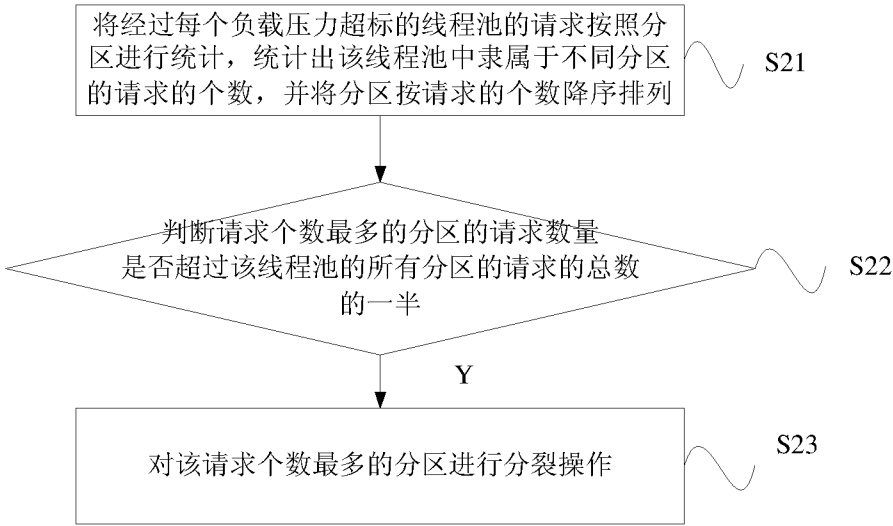


图 4

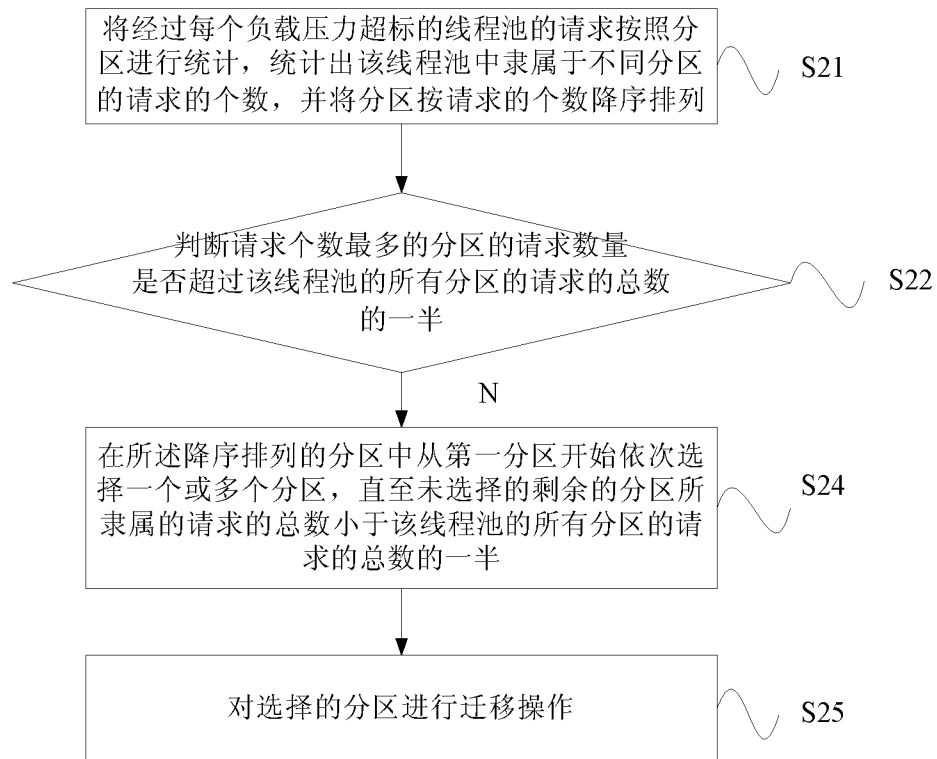


图 5

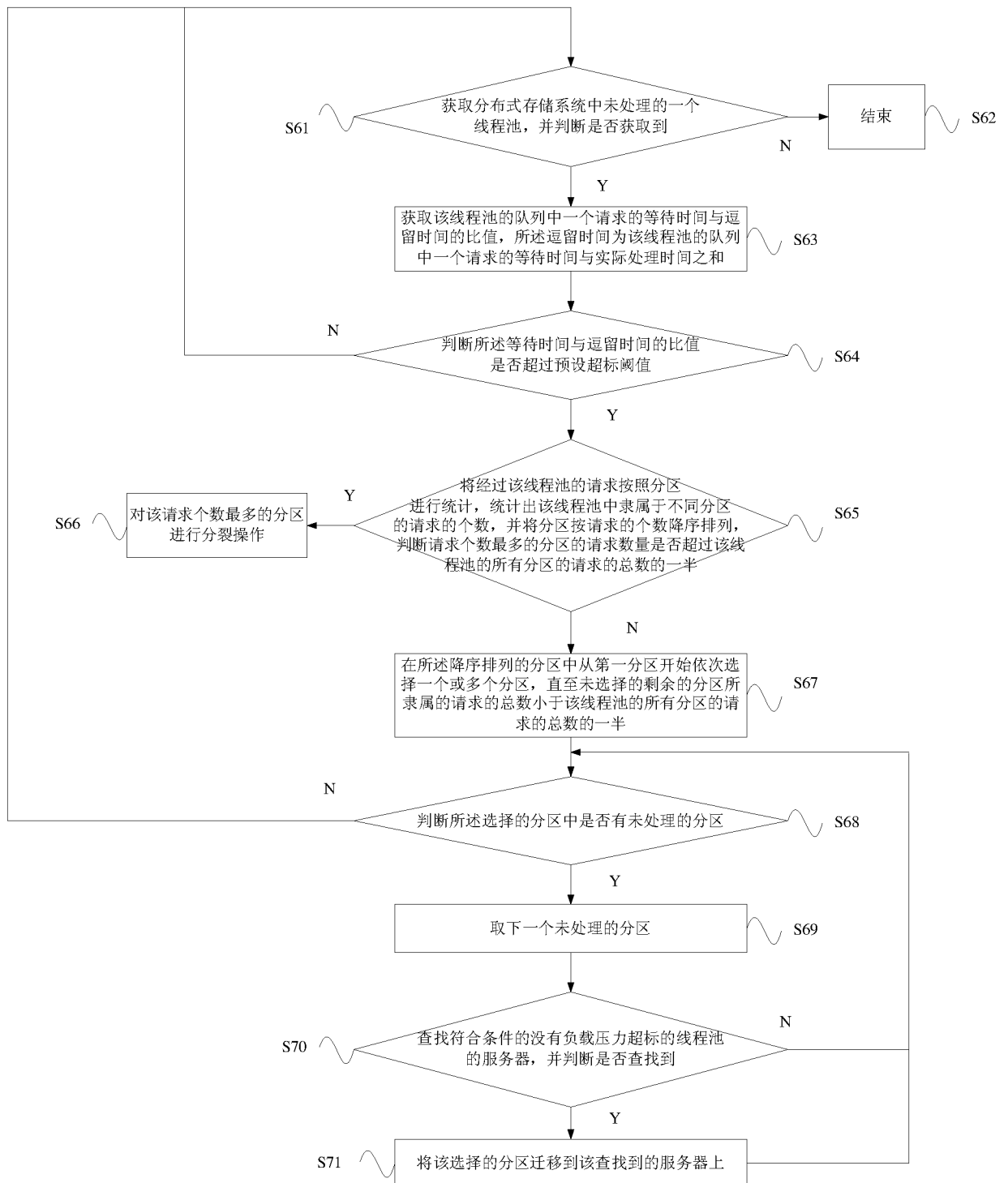


图 6

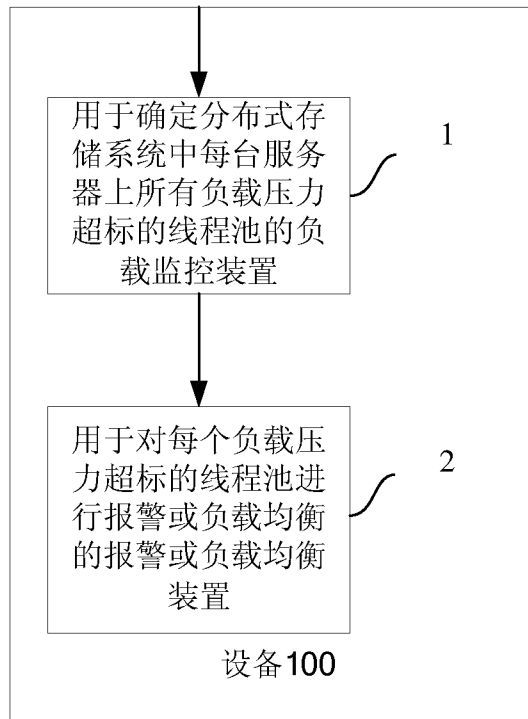


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/093893

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, H04L, H04W

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, IEEE, CNKI, GOOGLE: thread pool, ratio, distributed, server, resource, load, thread, pool, balancing, wait, time, threshold, value, alarm, alert, warning, memory, storage, queue

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102594861 A (HANGZHOU DIANZI UNIVERSITY), 18 July 2012 (18.07.2012), description, paragraphs [0013] and [0023]-[0027]	1-3, 11-13
X	CN 102207891 A (ZHEJIANG UNIVERSITY), 05 October 2011 (05.10.2011), description, paragraphs [0009]-[0028] and [0049]-[0079]	1-3, 11-13
A	US 2009157969 A1 (HARDING, M.J. et al.), 18 June 2009 (18.06.2009), the whole document	1-20
A	US 2014143789 A1 (NETAPP, INC.), 22 May 2014 (22.05.2014), the whole document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 20 September 2016 (20.09.2016)	Date of mailing of the international search report 08 November 2016 (08.11.2016)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer TAN, Jingchan Telephone No.: (86-10) 53318982

International application No.
PCT/CN2016/093893

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类 G06F 9/50 (2006.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F, H04L, H04W 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) CNPAT, WPI, EPDOC, IEEE, CNKI, GOOGLE: 分布式, 服务器, 资源, 负载, 线程池, 均衡, 等待, 时间, 阈值, 比值, 报警, 警报, 存储, 储存, 队列, distributed, server, resource, load, thread, pool, balancing, wait, time, threshold, value, alarm, alert, warning, memory, storage, queue																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 102594861 A (杭州电子科技大学) 2012年 7月 18日 (2012 - 07 - 18) 说明书第[0013], [0023]-[0027]段</td> <td>1-3, 11-13</td> </tr> <tr> <td>X</td> <td>CN 102207891 A (浙江大学) 2011年 10月 5日 (2011 - 10 - 05) 说明书第[0009]-[0028], [0049]-[0079]段</td> <td>1-3, 11-13</td> </tr> <tr> <td>A</td> <td>US 2009157969 A1 (HARDING, MATTHEW J. 等) 2009年 6月 18日 (2009 - 06 - 18) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2014143789 A1 (NETAPP, INC.) 2014年 5月 22日 (2014 - 05 - 22) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 102594861 A (杭州电子科技大学) 2012年 7月 18日 (2012 - 07 - 18) 说明书第[0013], [0023]-[0027]段	1-3, 11-13	X	CN 102207891 A (浙江大学) 2011年 10月 5日 (2011 - 10 - 05) 说明书第[0009]-[0028], [0049]-[0079]段	1-3, 11-13	A	US 2009157969 A1 (HARDING, MATTHEW J. 等) 2009年 6月 18日 (2009 - 06 - 18) 全文	1-20	A	US 2014143789 A1 (NETAPP, INC.) 2014年 5月 22日 (2014 - 05 - 22) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	CN 102594861 A (杭州电子科技大学) 2012年 7月 18日 (2012 - 07 - 18) 说明书第[0013], [0023]-[0027]段	1-3, 11-13															
X	CN 102207891 A (浙江大学) 2011年 10月 5日 (2011 - 10 - 05) 说明书第[0009]-[0028], [0049]-[0079]段	1-3, 11-13															
A	US 2009157969 A1 (HARDING, MATTHEW J. 等) 2009年 6月 18日 (2009 - 06 - 18) 全文	1-20															
A	US 2014143789 A1 (NETAPP, INC.) 2014年 5月 22日 (2014 - 05 - 22) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。 <table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件													
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																
国际检索实际完成的日期 2016年 9月 20日		国际检索报告邮寄日期 2016年 11月 8日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		授权官员 覃婧婵 电话号码 (86-10) 53318982															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/093893

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	102594861	A	2012年 7月 18日	无	
CN	102207891	A	2011年 10月 5日	无	
US	2009157969	A1	2009年 6月 18日	无	
US	2014143789	A1	2014年 5月 22日	US 8631415 B1	2014年 1月 14日