(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 108196788 B
(45)授权公告日 2021.05.07

(21)申请号 201711464651.6
(22)申请日 2017.12.28
(65)同一申请的已公布的文献号
申请公布号 CN 108196788 A
(43)申请公布日 2018.06.22
(73)专利权人 新华三技术有限公司
地址 310052 浙江省杭州市滨江区长河路466号
(72)发明人 李洁 丁文彪 刘庆典 杨植
(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371
代理人 王宁宁
(51)Int.Cl.
G06F 3/06(2006.01)

(56)对比文件
US 2017269850 A1,2017.09.21
US 2015199136 A1,2015.07.16
US 2017104663 A1,2017.04.13
CN 107220184 A,2017.09.29
CN 103577115 A,2014.02.12
CN 103701916 A,2014.04.02
CN 105072201 A,2015.11.18
CN 105511799 A,2016.04.20
CN 104331253 A,2015.02.04

审查员 王永贵

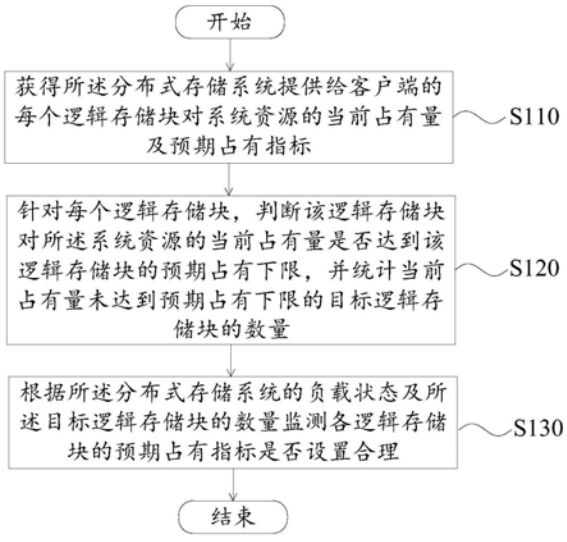
权利要求书4页 说明书15页 附图5页

(54)发明名称

QoS指标监测方法、装置、存储介质

(57)摘要

本申请提供一种QoS指标监测方法、装置、存储介质，应用于包括多个存储设备的分布式存储系统。方法包括：获得分布式存储系统提供的每个逻辑存储块对系统资源的当前占有量及预期占有指标；判断每个逻辑存储块对系统资源的当前占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量；根据系统的当前负载状态及目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理。如此，可以为预期占有指标的调节提供依据，从而使分布式存储系统的服务质量始终达到用户的预期，改善用户体验。



CN 108196788 B

1.一种QoS指标监测方法，其特征在于，应用于包括多个存储设备的分布式存储系统，所述方法包括：

获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标，其中，所述分布式存储系统根据所述预期占有指标为各逻辑存储块分配系统资源，每个逻辑存储块对系统资源的预期占有指标包括该逻辑存储块对系统资源的预期占有下限；

针对每个逻辑存储块，判断该逻辑存储块对所述系统资源的当前占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量；

根据所述分布式存储系统的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理；

其中，各逻辑存储块的预期占有指标的监测结果为设置合理或者设置不合理；所述分布式存储系统为每个逻辑存储块设置有用于限定为所述逻辑存储块分配的系统资源上限的限流值；

则，在确定各逻辑存储块的预期占有指标设置不合理之后，所述方法还包括：

判断所述目标逻辑存储块的数量是否等于所述分布式存储系统提供给客户端的逻辑存储块的数量；

若所述目标逻辑存储块的数量等于所述分布式存储系统提供给客户端的逻辑存储块的数量，则发出降低每个逻辑存储块的预期占有下限的提示信息；

若所述目标逻辑存储块的数量不等于所述分布式存储系统提供给客户端的逻辑存储块的数量，则根据各逻辑存储块的预期占有下限对对应逻辑存储块的限流值进行调整。

2.根据权利要求1所述的方法，其特征在于，所述分布式存储系统的当前负载状态通过以下方式确定：

判断每个存储设备的当前性能数据是否满足预设条件，在均满足所述预设条件中的至少一个时，确定所述分布式存储系统的当前负载状态为繁忙；否则，确定所述分布式存储系统的当前负载状态为空闲。

3.根据权利要求1或2所述的方法，其特征在于，根据所述分布式存储系统的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理的步骤，包括：

当所述目标逻辑存储块的数量大于0且所述分布式存储系统的当前负载状态为繁忙时，则确定各逻辑存储块的预期占有指标设置不合理；

当所述目标逻辑存储块的数量大于0且所述分布式存储系统的当前负载状态为空闲时，或者，当所述目标逻辑存储模块的数量为0时，确定各逻辑存储块的预期占有指标设置合理。

4.根据权利要求1所述的方法，其特征在于，所述根据各逻辑存储块的预期占有下限对对应逻辑存储块的限流值进行调整的步骤，包括：

针对每个当前占有量达到预期占有下限的逻辑存储块，将该逻辑存储块的预期占有下限作为该逻辑存储块的当前限流值。

5.根据权利要求1所述的方法，其特征在于，每个逻辑存储块的预期占有指标还包括该

逻辑存储块对系统资源的预期占有权重；

则，所述方法还包括：

当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块的当前占有量与预期占有下限的差值，得到多个差值；

对所述多个差值求和，得到所述分布式存储系统当前的剩余系统资源；

按照各逻辑存储块的预期占有权重，将所述剩余系统资源分配给每个逻辑存储块；

对每个逻辑存储块分配到的系统资源量及该逻辑存储块的预期占有下限求和，并将求得的和作为该逻辑存储块的当前限流值。

6.根据权利要求3所述的方法，其特征在于，每个逻辑存储块的预期占有指标还包括该逻辑存储块对系统资源的预期占有权重；

则，所述方法还包括：

当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块的当前占有量与预期占有下限的差值，得到多个差值；

对所述多个差值求和，得到所述分布式存储系统当前的剩余系统资源；

按照各逻辑存储块的预期占有权重，将所述剩余系统资源分配给每个逻辑存储块；

对每个逻辑存储块分配到的系统资源量及该逻辑存储块的预期占有下限求和，并将求得的和作为该逻辑存储块的当前限流值。

7.根据权利要求4所述的方法，其特征在于，每个逻辑存储块的预期占有指标还包括该逻辑存储块对系统资源的预期占有权重；

则，所述方法还包括：

当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块的当前占有量与预期占有下限的差值，得到多个差值；

对所述多个差值求和，得到所述分布式存储系统当前的剩余系统资源；

按照各逻辑存储块的预期占有权重，将所述剩余系统资源分配给每个逻辑存储块；

对每个逻辑存储块分配到的系统资源量及该逻辑存储块的预期占有下限求和，并将求得的和作为该逻辑存储块的当前限流值。

8.根据权利要求1或2所述的方法，其特征在于，所述分布式存储系统还包括多个目标进程及与每个逻辑存储块对应的存储块接口，所述分布式存储系统通过所述目标进程及所述存储块接口为客户端提供存储访问服务，每个存储块接口与至少一个所述目标进程绑定；

获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标的步骤，包括：

针对每个目标进程，通过与所述目标进程绑定的各存储块接口获取与所述各存储块接口对应的逻辑存储块对系统资源的当前占有量及预期占有指标。

9.一种QoS指标监测装置，其特征在于，应用于包括多个存储设备的分布式存储系统，所述装置包括：

获得模块，用于获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标，其中，所述分布式存储系统根据所述预期占有指标为各逻辑存储块分配系统资源，每个逻辑存储块对系统资源的预期占有指标包括该逻辑存储块对

系统资源的预期占有下限；

判断模块，用于针对每个逻辑存储块，判断该逻辑存储块对所述系统资源的当前占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量；

监测模块，用于根据所述分布式存储系统的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理；

其中，各逻辑存储块的预期占有指标的监测结果为设置合理或者设置不合理；所述分布式存储系统为每个逻辑存储块设置有用于限定为所述逻辑存储块分配的系统资源上限的限流值；则，所述装置还包括：

处理模块，用于在确定各逻辑存储块的预期占有指标设置不合理之后，判断所述目标逻辑存储块的数量是否等于所述分布式存储系统提供给客户端的逻辑存储块的数量；若等于，则发出降低每个逻辑存储块的预期占有下限的提示信息；若不等于，则根据各逻辑存储块的预期占有下限对对应逻辑存储块的限流值进行调整。

10.根据权利要求9所述的装置，其特征在于，所述分布式存储系统的当前负载状态通过以下方式确定：

判断每个存储设备的当前性能数据是否满足预设条件，在均满足所述预设条件中的至少一个时，确定所述分布式存储系统的当前负载状态为繁忙；否则，确定所述分布式存储系统的当前负载状态为空闲。

11.根据权利要求9或10所述的装置，其特征在于，

若所述目标逻辑存储块的数量大于0且所述分布式存储系统的当前负载状态为繁忙，所述监测模块确定各逻辑存储块的预期占有指标设置不合理；

若所述目标逻辑存储块的数量大于0且所述分布式存储系统的当前负载状态为空闲，所述监测模块确定各逻辑存储块的预期占有指标设置合理；

若所述目标逻辑存储块的数量为0，所述监测模块确定各逻辑存储块的预期占有指标设置合理。

12.根据权利要求9所述的装置，其特征在于，所述处理模块根据各逻辑存储块的预期占有下限对对应逻辑存储块的限流值进行调整的方式，为：

针对每个当前占有量达到预期占有下限的逻辑存储块，将该逻辑存储块的预期占有下限作为该逻辑存储块的当前限流值。

13.根据权利要求9所述的装置，其特征在于，每个逻辑存储块的预期占有指标还包括该逻辑存储块对系统资源的预期占有权重；

则，所述装置还包括：

第一限流更新模块，用于当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块的当前占有量与预期占有下限的差值，得到多个差值；对所述多个差值求和，得到所述分布式存储系统当前的剩余系统资源；按照各逻辑存储块的预期占有权重，将所述剩余系统资源分配给每个逻辑存储块；对每个逻辑存储块分配到的系统资源量及及该逻辑存储块的预期占有下限求和，并将求得的和作为该逻辑存储块的当前限流值。

14.根据权利要求11所述的装置，其特征在于，每个逻辑存储块的预期占有指标还包括该逻辑存储块对系统资源的预期占有权重；

则,所述装置还包括:

第二限流更新模块,用于当所述目标逻辑存储块的数量为0时,计算每个逻辑存储块的当前占有量与预期占有下限的差值,得到多个差值;对所述多个差值求和,得到所述分布式存储系统当前的剩余系统资源;按照各逻辑存储块的预期占有权重,将所述剩余系统资源分配给每个逻辑存储块;对每个逻辑存储块分配到的系统资源量及及该逻辑存储块的预期占有下限求和,并将求得的和作为该逻辑存储块的当前限流值。

15.根据权利要求12所述的装置,其特征在于,每个逻辑存储块的预期占有指标还包括该逻辑存储块对系统资源的预期占有权重;

则,所述装置还包括:

第三限流更新模块,用于当所述目标逻辑存储块的数量为0时,计算每个逻辑存储块的当前占有量与预期占有下限的差值,得到多个差值;对所述多个差值求和,得到所述分布式存储系统当前的剩余系统资源;按照各逻辑存储块的预期占有权重,将所述剩余系统资源分配给每个逻辑存储块;对每个逻辑存储块分配到的系统资源量及及该逻辑存储块的预期占有下限求和,并将求得的和作为该逻辑存储块的当前限流值。

16.根据权利要求9或10所述的装置,其特征在于,所述分布式存储系统还包括多个目标进程及与每个逻辑存储块对应的存储块接口,所述分布式存储系统通过所述目标进程及所述存储块接口为客户端提供存储访问服务,每个存储块接口与至少一个所述目标进程绑定;

针对每个目标进程,所述获得模块通过与所述目标进程绑定的各存储块接口获取与所述各存储块接口对应的逻辑存储块对系统资源的当前占有量及预期占有指标。

17.一种存储介质,其上存储有计算机可读指令,其特征在于,所述计算机可读指令被执行时,实现权利要求1-8任意一项所述的方法。

QoS指标监测方法、装置、存储介质

技术领域

[0001] 本申请涉及分布式存储技术领域，具体而言，涉及一种基于分布式存储系统的QoS指标监测方法、装置、存储介质。

背景技术

[0002] 分布式存储系统将用户逻辑上的文件、存储块或存储对象的数据分散地存储到不同的存储设备中。同一个存储设备可能被多个用户使用，从而导致了不同用户(或业务)对同一存储资源的竞争。为了确保关键业务不中断，通常会通过存储QoS(Quality of Service，服务质量)控制来为不同的用户(或业务)分配不同的IOPS(Input/Output Per Second，每秒进行读操作或写操作的次数)或带宽，以使重要业务的存储资源不被过度抢占。

[0003] 现有的分布式存储系统通常允许用户(通常是系统的管理员)根据系统实际性能及业务情况配置相应的QoS指标，该QoS指标体现了用户希望该分布式存储系统达到的服务质量。然而，当分布式存储系统中的硬件发生变化，进而导致整个系统的性能发生变化时，用户配置的QoS指标将不再适用。若分布式存储系统继续根据该QoS指标进行存储QoS控制，将始终无法达到用户预期的服务质量，导致用户体验较差。

发明内容

[0004] 为了改善现有技术中的上述不足，本申请的目的在于提供一种QoS指标监测方法，应用于包括多个存储设备的分布式存储系统，所述方法包括：

[0005] 获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标，其中，所述分布式存储系统根据所述预期占有指标为各逻辑存储块分配系统资源，每个逻辑存储块对系统资源的预期占有指标包括该逻辑存储块对系统资源的预期占有下限；

[0006] 针对每个逻辑存储块，判断该逻辑存储块对所述系统资源的当前占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量；

[0007] 根据所述分布式存储系统的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理。

[0008] 本申请的另一目的在于提供一种QoS指标监测装置，应用于分布式存储系统，所述装置包括：

[0009] 获得模块，用于获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标，其中，所述分布式存储系统根据所述预期占有指标为各逻辑存储块分配系统资源，每个逻辑存储块对系统资源的预期占有指标包括该逻辑存储块对系统资源的预期占有下限；

[0010] 判断模块，用于针对每个逻辑存储块，判断该逻辑存储块对所述系统资源的当前

占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量；

[0011] 监测模块，用于根据所述分布式存储系统的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理。

[0012] 本申请的另一目的在于提供一种存储介质，其上存储有计算机可读指令，所述计算机可读指令被执行时，实现本申请实施例提供的QoS指标监测方法。

[0013] 相对于现有技术而言，本申请实施例具有以下有益效果：

[0014] 本申请实施例提供的QoS指标监测方法、装置、存储介质，通过获得的每个逻辑存储块对系统资源的当前占有量和预期占有指标，得到对系统资源的当前占有量未达到预期占有下限的目标逻辑存储块的数量；并根据分布式存储系统的当前负载状态及目标逻辑存储块的数量，监测各逻辑存储块的预期占有指标是否合理。如此，可以为预期占有指标的调节提供依据，进而使分布式存储系统的服务质量达到用户的预期，改善用户体验。

附图说明

[0015] 为了更清楚地说明本申请实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本申请的某些实施例，因此不应被看作是对范围的限定，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他相关的附图。

[0016] 图1为本申请实施例提供的一种分布式存储系统与客户端的交互示意图；

[0017] 图2为本申请实施例提供的一种QoS指标监测方法的流程示意图；

[0018] 图3为本申请实施例提供的一种Ceph RADOS系统与客户端的交互示意图；

[0019] 图4为本申请实施例提供的一种基于RADOS的Ceph集群与客户端的交互示意图；

[0020] 图5为本申请实施例提供的一种QoS指标监测装置的功能模块框图。

[0021] 图标：100-分布式存储系统；110-存储设备；120-逻辑存储块；200-客户端；300-QoS指标监测装置；310-获得模块；320-判断模块；330-监测模块；340-处理模块；350-限流更新模块。

具体实施方式

[0022] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请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申请一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本申请实施例的组件可以以各种不同的配置来布置和设计。

[0023] 因此，以下对在附图中提供的本申请的实施例的详细描述并非旨在限制要求保护的本申请的范围，而是仅仅表示本申请的选定实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0024] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

[0025] 请参照图1，图1是本申请实施例提供的一种分布式存储系统100与至少一个客户

端200(图1中仅示出一个)的交互示意图。其中,所述客户端200是指使用所述分布式存储系统100提供的存储资源的用户设备,所述客户端200可以是任意能够连接到网络的数据处理设备,例如,个人计算机(Personal Compute,PC)、服务器、移动电子设备等。

[0026] 所述分布式存储系统100包括多个存储设备110,所述存储设备110可以是任意具有数据处理能力和数据存储能力的硬件设备。所述多个存储设备110通过网络连接构成所述分布式存储系统100的存储资源池。所述分布式存储系统100可以将所述存储资源池中的存储资源提供给客户端200使用。

[0027] 在本实施例中,所述分布式存储系统100提供给客户端200的每一存储资源是一个逻辑上的存储空间,也即,逻辑存储块120。每个逻辑存储块120所对应的存储资源可以位于一个存储设备110上,也可以分布在多个存储设备110上。

[0028] 可选地,为了让客户端200能够访问所述分布式存储系统100提供的逻辑存储块120,进而对所述逻辑存储块120中的数据进行I/O操作,所述分布式存储系统100中可以包括与每个逻辑存储块120对应的存储块接口,通过所述存储块接口,客户端200可以对相应的逻辑存储块120中的数据进行I/O操作。需要说明的是,在本实施例中,I/O操作包括对数据的读操作或写操作。

[0029] 在本实施例中,所述存储块接口可以运行在所述分布式存储系统100中专门提供存储访问的服务器上,也可以运行在所述存储设备110上。

[0030] 为了避免非关键业务占用过多的系统资源,在实际应用中,分布式存储系统100通常允许系统管理员为系统提供的各个逻辑存储块120配置QoS指标,以对各个逻辑存储块120分配的系统资源的多少进行限制。其中,所述系统资源可以是IOPS或带宽。所述QoS指标通常包括预留(Reservation)指标、上限(limit)指标以及权重(weight)指标。

[0031] 所述预留指标是指对给定逻辑存储块120的最低系统资源分配保证,例如,假设某个逻辑存储块120配置的IOPS值预留指标为300,那么当用户对该逻辑存储块120的IOPS值超过300时,系统能够保证处理针对该逻辑存储块120的I/O请求的实际IOPS不低于300。

[0032] 所述上限指标是用户希望该给定逻辑存储块120能够分配到的最高系统资源。通过上限指标的设置,可以避免该给定逻辑存储块120占用过多的系统资源。

[0033] 所述权重指标,通常是一个具体的权重比例或权重等级,用户希望系统在满足所有逻辑存储块120的预留指标的前提下,在上限指标之下,按照该权重指标来为该给定逻辑存储块120分配系统资源。

[0034] 分布式存储系统100的系统管理员可以根据系统性能预估整个系统能够提供的总系统资源,并根据所述总系统资源及业务重要程度给系统中的各个逻辑存储块120配置相应的预留指标、上限指标以及权重指标。

[0035] 经发明人研究发现,在正常情况下,分布式存储系统100可以通过相应的QoS控制算法,依照用户配置的QoS指标处理用户对各个逻辑存储块120的I/O请求,从而达到用户在各个逻辑存储块120配置的所述QoS指标。其中,所述QoS控制算法可以在各逻辑存储块120的存储块接口中通过限流器实现,所述限流器可以依据所述QoS指标对各逻辑存储块120的I/O操作进行控制。

[0036] 然而,当分布式存储系统100中的硬件发生改变进而导致系统性能改变时,系统实际能够提供的系统资源也会发生变化,系统管理员配置的QoS指标将不再合理,换言之,无

论系统如何进行控制，也无法满足用户的预期，导致用户体验较差。

[0037] 发明人仔细分析后发现，当系统性能发生改变时，往往会导致系统无法满足用户设置的预留指标，而不会对上限指标造成影响。例如，假设存在某个逻辑存储块A，用户为其配置的带宽上限指标为2MB/s，表示用户不希望所述逻辑存储块A分配到的带宽超过2MB/s。那么，当系统性能降低，导致所述逻辑存储块A分配到的系统资源变少，从而无法达到2MB/s，这本身就是符合用户的预期的。当系统性能上升，导致所述逻辑存储块A分配到的系统资源变多时，由于上限指标的限定作用，所述逻辑存储块A分配到的系统资源也不会超过2MB/s，这也是符合用户的预期的。由此可见，分布式存储系统100总是能够满足用户配置的上限指标的。

[0038] 而预留指标作为用户希望给定逻辑存储块120分配到的最低系统资源，一旦系统性能下降，导致无法达到各逻辑存储块120的预留指标时，与用户的预期不相符，需要进行相应的调节。

[0039] 然而，导致分布式存储系统100无法达到用户设置的预留指标的原因不止QoS指标设置不合理这一种，即便在检测到分布式存储系统100无法达到预留指标的情况下，也不能直接对各逻辑存储块120的QoS指标进行调节。也就是说，在调节QoS指标之前，还需对QoS指标进行监测，以判断其设置是否合理，从而确定是否需要对QoS指标进行调整。

[0040] 为解决上述问题，本申请实施例提供一种QoS指标监测方法及装置，用于监测分布式存储系统100中各个逻辑存储块120上配置的QoS指标是否设置合理，以为之后可能进行的指标调整提供依据。

[0041] 其中，所述QoS指标监测装置可以运行在所述分布式存储系统100中的任意存储设备110或其他用于管理的服务器(如，元数据服务器)上，所述QoS指标监测装置也可以运行在与所述分布式存储系统100通信且独立于所述分布式存储系统100的设备上，本实施例对此不做限制。

[0042] 如图2所示，是本申请实施例提供的一种QoS指标监测方法的流程示意图，所述QoS指标监测方法应用于图1所示的分布式存储系统100。下面对所述方法的具体流程及步骤做详细阐述。

[0043] 步骤S110，获得所述分布式存储系统100提供给客户端200的每个逻辑存储块120对系统资源的当前占有量及预期占有指标。

[0044] 其中，每个逻辑存储块120对系统资源的当前占有量可以是该逻辑存储块120当前实际占用的带宽，或是该逻辑存储块120当前实际的IOPS。每个逻辑存储块120的预期占有指标是指系统管理员为该逻辑存储块120配置的QoS指标，所述QoS指标通常记录在该逻辑存储块120的元数据中，通过该逻辑存储块120所对应的存储块接口访问元数据，以获取所述Qos指标。

[0045] 在本实施例中，每个逻辑存储块120的预期占有指标包括该逻辑存储块120对系统资源的预期占有下限，所述预期占有下限是指所述QoS指标中的预留指标。

[0046] 在现有技术中，通常是在逻辑存储块120与客户端200之间增加一SVM(存储虚拟机)，客户端200发送的I/O请求先到达SVM，再由SVM写入到分布式存储系统中。基于这种系统架构，SVM可以很容易地获取到所连接的各逻辑存储块120对系统资源的当前占有量，从而在所述当前占有量超过预期占有上限时进行控制。然而，对预期占有下限的监测需要考

虑到所述分布式存储系统中所有逻辑存储块120的情况，但在分布式存储系统中，为了避免引起不必要的开销，每个SVM只与Ceph集群中的一组逻辑存储块120连接，因此，无法通过某个SVM获取到所有的逻辑存储块120的数据。

[0047] 为了解决上述问题，在本实施例中，可以通过向每个逻辑存储块120的存储块接口发送数据获取请求或是由每个逻辑存储块120的存储块接口主动上报的方式获得每个逻辑存储块120对系统资源的当前占有量及预期占有指标。

[0048] 下面将以基于iSCSI(Internet Small ComputerSystem Interface，互联网小型计算机系统接口)协议的Ceph集群存储系统为例，对步骤S110做进一步的阐述。在对步骤S110做进一步阐述之前，对基于iSCSI协议的Ceph集群存储系统进行说明。

[0049] Ceph集群存储系统是一种运行在普通商用硬件(commodity hardware)之上的分布式存储系统，图3为一种实例中该系统与客户端200的交互示意图。

[0050] 如图3所示，所述Ceph集群存储系统包括由多个OSD(Object Storage Device，对象存储设备)构成的RADOS(Reliable Autonomic Distributed Object Storage，分布式对象存储)系统及其他用于提供管理、存储访问等服务的服务器。其中，所述OSD相当于本实施例中的存储设备110，每个OSD包括用于管理其上存储的数据的管理进程，所述管理进程通常被称作“OSD守护进程”。客户端200可以直接与需要操作的存储设备110的管理进程进行通信。

[0051] 所述Ceph集群存储系统提供给客户端200的逻辑存储块120通常被称作RBD(RADOS Block Device)块设备，每个RBD块设备具有对应的RBD客户端接口(存储块接口)，客户端200通过RBD客户端接口可以访问对应的RBD块设备。系统管理员为每个RBD块设备配置的QoS指标通常设置在该RBD块设备的元数据中。

[0052] 所述iSCSI协议是集成了SCSI(SmallComputerSystemInterface)协议和TCP/IP协议的新的协议，它在SCSI基础上扩展了网络功能，可以让SCSI命令通过IP网络传送到远程的SCSI设备上，而SCSI协议只能访问本地的SCSI设备。

[0053] 通过上述的iSCSI协议，可以将所述分布式存储系统100的存储资源共享给客户端200使用，相应地，客户端200可以通过iSCSI协议对所述存储资源中的数据进行I/O操作。

[0054] 详细地，图3所示的Ceph集群存储系统与客户端200通过iSCSI协议实现存储访问。在这一情形下，所述Ceph集群存储系统中包括多个目标进程(Target)，Target通常被称作iSCSI服务端。每个iSCSI服务端(Target)通过RBD客户端接口访问包括所述RBD客户端接口的RBD块设备。通常，每个RBD客户端接口可以被至少一个iSCSI服务端绑定(bind)并访问。

[0055] 实施时，所述Target可以接收客户端200发送的iSCSI命令和数据，并根据所述iSCSI命令和数据，在自己绑定的RBD客户端接口中确定所述客户端200要操作的RBD块设备(逻辑存储块120)。由于每个RBD客户端接口与其对应的RBD块设备相关联，因而，根据所述RBD客户端接口可以查找到所述客户端200要操作的存储资源(即，RBD块设备)，并对所述要操作的RBD块设备执行相应的I/O操作。

[0056] 当所述分布式存储系统100是基于iSCSI协议的Ceph集群存储系统时，作为一种实施方式，所述步骤S110可以包括如下子步骤：

[0057] 针对每个目标进程，通过与所述目标进程绑定的各存储块接口获取所述各存储块接口对应的逻辑存储块120对系统资源的当前占有量及预期占有指标。

[0058] 详细地，可以通过一查询模块与每个iSCSI服务端(Target)建立socket连接，并在建立所述socket连接后，向每个iSCSI服务端发送请求，在接收到该请求时，iSCSI服务端通过绑定的每个RBD客户端接口，获取该RBD客户端接口对应的RBD块设备对系统资源的当前占有量及预期占有指标，从而获得所述Ceph集群存储系统中的各RBD块设备对系统资源的当前占有量及预期占有指标。

[0059] 作为另一种实施方式，每个iSCSI服务端可以通过绑定的RBD客户端接口，获取该RBD客户端接口对应的RBD块设备对系统资源的当前占有量及预期占有指标，并将获取到的数据主动上报给所述查询模块。

[0060] 步骤S120，针对每个逻辑存储块120，判断该逻辑存储块120对所述系统资源的当前占有量是否达到该逻辑存储块120的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量。

[0061] 其中，所述目标逻辑存储块是指：所述分布式存储系统100包括的各逻辑存储块120中对系统资源的当前占有量未达到预期占有下限的逻辑存储块120。相应地，所述目标逻辑存储块的数量为对所述系统资源的当前占有量未达到预期占有下限(预留指标)的逻辑存储块120的数量。

[0062] 步骤S130，根据所述分布式存储系统100的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块120的预期占有指标是否设置合理。

[0063] 发明人研究发现，当所述分布式存储系统100无法达到用户配置的预留指标时，用户配置的QoS指标可能不合理。然而，当用户的业务量小，I/O操作少的时候，所述分布式存储系统100也无法达到用户配置的预留指标(即，预期占有下限)，这种情况属于正常现象。因此，需要根据各逻辑存储块120达到预期占有指标的情况以及所述分布式存储系统100的当前负载状态对各逻辑存储块120的预期占有指标的合理性进行判断。

[0064] 在本实施例中，所述分布式存储系统100的当前负载状态可以包括繁忙和空闲两种，在一种具体实施方式中，所述分布式存储系统100的当前负载状态可以根据所述分布式存储系统100中的每个存储设备110的当前性能数据进行判断。

[0065] 其中，所述存储设备110的当前性能数据可以包括所述存储设备110当前的队列深度、CPU利用率、磁盘利用率及内存占用率中的至少一个。

[0066] 所述队列深度是指队列中等待响应的请求的数量。在本实施例中，所述存储设备110中通常存在多种队列，如消息收发队列、恢复队列以及与其他业务对应的队列等。

[0067] 可选地，在本实施例中，可以通过向每个存储设备110发送数据获取请求或是每个存储设备110主动上报的方式获得每个存储设备110的当前性能数据。

[0068] 仍旧以图3所示的基于iSCSI协议的Ceph集群存储系统为例，对步骤S120做进一步阐述。

[0069] 当所述分布式存储系统100是图3所示的基于iSCSI协议的Ceph集群存储系统时，作为一种实施方式，所述步骤S110可以通过如下子步骤实现：

[0070] 向每个存储设备110(即OSD)上的管理进程请求该存储设备110的当前性能数据。

[0071] 详细地，可以通过所述查询模块与每个OSD上的管理进程建立socket(套接字)连接，并在建立所述socket连接后向该OSD上的管理进程发送性能数据获取请求，以获取该OSD的当前性能数据。其中，所述当前性能数据可以包括OSD当前的队列深度、CPU利用率、磁

盘利用率及内存占用率中的至少一种。

[0072] 作为另一种实施方式，每个OSD上的管理进程可以与所述查询模块建立socket连接，并主动向所述查询模块发送自己的当前性能数据。

[0073] 在获得各存储设备110的当前性能数据之后，可以判断所述当前性能数据是否满足预设条件，在满足所述预设条件中的至少一个时，确定所述分布式存储系统100的当前负载状态为繁忙；否则，确定所述分布式存储系统100的当前负载状态为空闲。其中，所述预设条件包括以下至少一项：

[0074] 每个存储设备110的队列深度均达到设定的深度阈值；

[0075] 每个存储设备110的CPU利用率均达到设定的CPU利用率阈值；

[0076] 每个存储设备110的磁盘利用率均达到设定的磁盘利用率阈值；

[0077] 每个存储设备110的内存占用率均达到设定的内存占用率阈值。

[0078] 在本实施例中，所述存储设备110中的每个队列具有相应的深度阈值，所述深度阈值可以通过测试确定。如，测试所述分布式存储系统100的系统资源占用达到极限时每个队列的深度，可以将该深度或该深度预设范围内的值作为该队列的深度阈值。所述CPU利用率阈值、磁盘利用率阈值、内存占用率阈值也可以通过测试进行确定，测试方式与测试深度阈值的方式类似。例如，所述CPU利用率阈值可以是90％，所述磁盘利用率阈值可以是75％，所述内存占用率阈值可以是90％。应当理解的是，上述各阈值可以根据经验或根据实际情况设定，本申请在此并不做特别限定。

[0079] 在本实施例中，所述预设条件的内容与所述当前性能数据相对应。例如，当所述存储设备110的当前性能数据包括所述存储设备110当前的的队列深度、CPU利用率、磁盘利用率及内存占用率中，所述预设条件包括上述的四个条件。当所述存储设备110的当前性能数据包括所述存储设备110当前的队列深度、CPU利用率、磁盘利用率及内存占用率中的一个、两个或三个时，也可以依照上述判断方法相应地调整所述预设条件的个数，从而对所述分布式存储系统100的当前负载状态进行判断。

[0080] 例如，当所述存储设备110的当前性能数据包括所述存储设备110当前的CPU利用率和磁盘利用率时，所述预设条件包括：每个存储设备110的CPU利用率均达到设定的CPU利用率阈值；以及，每个存储设备110的磁盘利用率均达到设定的磁盘利用率阈值。

[0081] 在确定所述分布式存储系统100的当前负载状态及目标逻辑存储块的数量之后，可以通过如下方式判断各逻辑存储块的预期占有指标(即QoS指标)是否合理，即步骤S130可以包括如下子步骤：

[0082] 当所述目标逻辑存储块的数量大于0且所述分布式存储系统100的当前负载状态为繁忙时，确定各逻辑存储块120的预期占有指标设置不合理；

[0083] 当所述目标逻辑存储块的数量大于0且所述分布式存储系统100的当前负载状态为空闲时，或者，当所述目标逻辑存储块的数量为0，确定各逻辑存储块120的预期占有指标设置合理。

[0084] 在本实施例中，当所述目标逻辑存储块的数量大于0时，表明所述分布式存储系统100中存在当前占有量未达到预期占有下限的逻辑存储块120。

[0085] 此时，若所述分布式存储系统100处于繁忙状态，表明所述分布式存储系统100达不到设置的预期占有下限的原因是系统资源不足，在这种情形下，无论如何控制，都无法满

足用户设置的预期占有下限，因而，可以确定用户设置的预期占有指标不合理。

[0086] 若所述分布式存储系统100处于空闲状态，表明所述分布式存储系统100达不到设置的预期占有下限的原因是用户业务量小，属于正常现象，可以确定各逻辑存储块120的预期占有指标是合理的。当所述目标逻辑存储块的数量为0时，表明所述分布式存储系统100中不存在对系统资源的当前占有量未达到预期占有下限的逻辑存储块120，因此，各逻辑存储块120的预期占有指标是合理的。

[0087] 在本实施例中，当确定各逻辑存储块120的预期占有指标设置不合理之后，还可以根据所述目标逻辑存储块的数量采取相应的措施以确保各逻辑存储块120的预期占有指标的合理性。

[0088] 详细地，在确定各逻辑存储块120的预期占有指标设置不合理之后，所述QoS指标监测方法还可以包括如下步骤。

[0089] 首先，判断所述目标逻辑存储块的数量是否等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量。

[0090] 也即，判断是否所有的逻辑存储块120对系统资源的当前占有量均未达到各自对系统资源的预期占有下限。

[0091] 然后，若所述目标逻辑存储块的数量等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量，则发出降低每个逻辑存储块120的预期占有下限的提示信息，以提示客户端200降低每个逻辑存储块120的预期占有下限。若所述目标逻辑存储块的数量不等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量，则根据各逻辑存储块120的预期占有下限对对应逻辑存储块120的限流值进行调整，以使所述分布式存储系统100为每个逻辑存储块120分配的系统资源达到该逻辑存储块120的预期占有下限。

[0092] 在本实施例中，当所述目标逻辑存储块的数量等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量时，表示所有的逻辑存储块120对系统资源的当前占有量均未达到各自的预期占有下限。此时，无法在不改变用户预期的情况下对用户设置的预期占有指标进行调整，进而使所述分布式存储系统100满足用户的预期。因而，此时可以对用户进行提示，告知用户当前系统资源不足，已无法满足其配置的预期占有下限，从而使用户根据实际情况降低各个逻辑存储块120的预期占有下限。

[0093] 通常情况下，所述提示信息被输出到系统管理员的客户端200。

[0094] 在此处值得说明的是，基于预期占有上限(也即，上限指标)的含义可以得知，将用户设置的预期占有上限的值调整为大于该用户设置的预期占有下限的值，仍旧是符合该用户的预期的。

[0095] 通过上述设计，用户可以根据实际需求进行调整，例如，降低某些用于存储非关键业务数据的逻辑存储块120的预期占有下限调低。其中，所述用户是指所述分布式存储系统100的系统管理员。相应地，所述提示信息会被发送到所述系统管理员对应的客户端200。

[0096] 其中，所述分布式存储系统100可以根据每个逻辑存储块120对系统资源的预期占有上限确定该逻辑存储块120的初始限流值，并将该初始限流值设置在该逻辑存储块120的元数据中，并在分配系统资源时限制该逻辑存储块120分配的系统资源不超过该初始限流值。通常情况下，每个逻辑存储块120的初始限流值等于该逻辑存储块120的预期占有上限。

[0097] 在本实施例中，当所述目标逻辑存储块的数量不等于所述分布式存储系统100提

供给客户端200的逻辑存储块120的数量时，表明所述分布式存储系统100所包括的各逻辑存储块120中只有一部分逻辑存储块120对系统资源的当前占有量达到了各自的预期占有下限。

[0098] 针对这一情形，可以根据各逻辑存储块120的预期占有下限对对应逻辑存储块120的限流值进行调整，以使所述分布式存储系统100为每个逻辑存储块120分配的系统资源能够达到该逻辑存储块120的预期占有下限。

[0099] 在本实施例中，根据各逻辑存储块120的预期占有下限对对应逻辑存储块120的限流值进行调整的具体方式可以根据所述分布式存储系统100所采用的QoS控制算法确定。

[0100] 具体通过如下QoS控制算法对各逻辑存储块120进行QoS控制：

[0101] 所述分布式存储系统100将每个逻辑存储块120的预期占有上限作为该逻辑存储块120的初始限流值，在控制过程中，优先将系统资源分配给当前占有量未达到预期占有下限的逻辑存储块120，在每个逻辑存储块120的当前占有量达到预期占有下限时，优先将系统资源分配给预期占有权重大的逻辑存储块120，并确保每个逻辑存储块120分配到的系统资源不超过该逻辑存储块120的当前限流值。

[0102] 其中，所述预期占有权重相当于上述内容中的权重指标。

[0103] 在所述分布式存储系统100基于上述QoS控制算法进行QoS控制的基础上，当所述目标逻辑存储块的数量不等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量时，也即，当所述分布式存储系统100中只有部分逻辑存储块120未达到各自的预期占有下限时，可以通过如下步骤对各逻辑存储块120的限流值进行调整：

[0104] 针对每个当前占有量达到预期占有下限的逻辑存储块120，将该逻辑存储块120的预期占有下限作为该逻辑存储块120的当前限流值。

[0105] 例如，当该逻辑存储块B对IOPS的预期占有上限为500，对IOPS的预期占有下限为300时，分布式存储系统100为该逻辑存储块B的IOPS设置的限流值为500。在实施时，若该逻辑存储块B的当前IOPS达到300，则将300设置为该逻辑存储块B的IOPS的当前限流值。

[0106] 如此，当所述分布式存储系统100给当前占有量达到预期占有下限的逻辑存储块120分配的系统资源达到其预期占有下限(此时也是限流值)后，将不再继续为其分配系统资源，基于上述的QoS控制算法，所述分布式存储系统100将会为其他未达到预期占有下限的逻辑存储块120分配系统资源，从而使其他无法达到预期占有下限的逻辑存储块120对系统资源的占有量达到其预期占有下限，进而使所述分布式存储系统100达到用户预期，提高用户体验。

[0107] 此外，发明人还发现，在做出上述调整的情况下，当所述分布式存储系统100的性能发生改变时，还是可能无法按照各个逻辑存储块120的预期占有权重(权重指标)对所述分布式存储系统100剩余的系统资源进行分配。例如，当系统性能提高时，可能出现某些逻辑存储块120在所分配到的系统资源所占的比例低于自己的预期占有权重时，已经达到了各自的限流值(通常等于预期占有上限)，从而导致这些逻辑存储块120实际分配到的系统资源与各自的预期占有权重不符合。

[0108] 为解决这一问题，所述QoS指标监测方法还可以包括如下步骤：

[0109] 首先，当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块120的当前占有量与预期占有下限的差值，得到多个差值；对所述多个差值求和，得到所述分布式存储系统

100当前的剩余系统资源。

[0110] 所述目标逻辑存储块的数量为0,表示所有的逻辑存储块120对系统资源的当前占有量均达到了各自的预期占有下限。由于所述分布式存储系统100的总体性能处在变化之中,无法实时获取到其系统资源总量,而各逻辑存储块120对系统资源的当前占有量之和,是所述分布式存储系统100当前一定能够提供的系统资源总量。因此,可以直接将各逻辑存储块120对系统资源的当前占有量之和作为所述分布式存储系统100当前的系统资源总量。

[0111] 所述系统资源总量扣除各逻辑存储块120需要达到的预期占有下限,得到的是所述分布式存储系统100可以按照各逻辑存储块120的预期占有权重进行分配的系统资源,即所述分布式存储系统100当前的剩余系统资源。

[0112] 然后,按照各逻辑存储块120的预期占有权重,将所述剩余系统资源分配给每个逻辑存储块120;对每个逻辑存储块120分配到的系统资源量及及该逻辑存储块120的预期占有下限求和,并将求得的和作为该逻辑存储块120的当前限流值。

[0113] 通过上述过程中对限流值的动态调整,在用户业务量足够,且各逻辑存储块120对系统资源的当前占有量达到各自的预期占有下限的情况下,可以确保各逻辑存储块120后续分配到的系统资源一定与各自的预期占有权重相匹配。下面给出一种实例,以对上述过程做进一步阐述。

[0114] 假设所述分布式存储系统100包括逻辑存储块B1及逻辑存储块B2。其中,逻辑存储块B1的预期占有下限为x1,预期占有上限为y1,预期占有权重为40%;逻辑存储块B2的预期占有下限为x2,预期占有上限为y2,预期占有权重为60%。当逻辑存储块B1对系统资源的占有量达到x1,且逻辑存储块B2对系统资源的占有量达到x2时,剩余系统资源为T,则按照预期占有权重将剩余系统资源T分配给逻辑存储块B1和B2。其中,逻辑存储块B1分配到0.4T,逻辑存储块B2分配到0.6T。

[0115] 在进行上述分配后,将逻辑存储块B1分配到系统资源0.4T与其预期占有下限x1求和,得到逻辑存储块B1的限流值为x1+0.4T;同理可以求得,逻辑存储块B2的限流值为x2+0.6T。

[0116] 由于在逻辑存储块B1和B2都达到预期占有下限后,分布式存储系统100会按照预期占有权重进行系统资源的分配,而逻辑存储块B1还可再分配的系统资源为0.4T,逻辑存储块B2还可再分配的系统资源为0.6T,只要用户业务量足够,就不会出现逻辑存储块B1和B2尚未达到各自的预期占有权重就已经达到相应限流值的情况。至于用户业务量不够的情形导致无法达到各自的预期占有权重的情况,属于正常现象,在此不做赘述。可选地,在本实施例中,所述QoS指标监测方法可以每间隔预设时长(如,1-10秒)执行一次,如此,可以当系统性能发生变化导致配置的QoS指标不合理时能够及时地监测到,进而可以根据监测结果采取相应的措施,设置合理的QoS指标。

[0117] 需要说明的是,在本实施例中,合理的QoS指标是指所述分布式存储系统100通过相应的QoS控制算法能够达到的QoS指标,不合理的QoS指标是指所述分布式存储系统100通过相应的QoS控制算法无法达到的QoS指标。

[0118] 下面结合图4,以所述系统资源是IOPS为例,对本申请实施例提供的QoS指标监测方法在基于RADOS的Ceph集群中的具体应用做详细阐述。

[0119] 如图4所示,所述基于RADOS的Ceph集群包括RADOS、多个Target(目标进程)以及查

询模块。其中，所述RADOS包括多个OSD，构成了所述Ceph集群的存储资源池，所述Ceph集群可将所述存储资源池中的存储空间提供给客户端200使用。所述Ceph集群提供的每个存储空间被称作RBD块设备，每个RBD块设备相当于一个逻辑存储块120。

[0120] 客户端200可通过iscsi-tgt来访问所述Ceph集群中的RBD块设备。iscsi-tgt是一种提供存储访问服务的应用程序，包括iSCSI客户端(tgt-initiator)及iSCSI服务端(Target)。其中，所述tgt-initiator运行在客户端200上，所述Target运行在所述Ceph集群中提供访问服务的服务器(后称“存储访问服务器”)上。

[0121] 当用户首次使用所述Ceph集群提供的存储空间时，可以通过客户端200上安装的tgt-initiator在所述Ceph集群中的存储访问服务器上创建RBD块设备，该RBD块设备包括一RBD客户端接口。在创建所述RBD块设备时，用户可以设定该RBD块设备的存储空间的大小。

[0122] 在创建所述RBD块设备之后，可以将所述RBD块设备与所述存储访问服务器中已经创建的Target绑定，也可以重新创建一个Target，并将所述RBD块设备与所述重新创建的Target绑定。

[0123] 在执行上述操作后，可以通过tgt-initiator发现所述存储访问服务器上的所有Target，并登陆到与所需访问的RBD客户端接口绑定的Target上，然后，即可通过tgt-initiator发起访问请求或操作请求，该访问请求或操作请求会经由Target转发给相应的RBD客户端接口。

[0124] 在本实施例中，每个Target可以绑定一个或多个RBD客户端接口。每个Target绑定的RBD客户端接口具有不同的逻辑单元号(Logical Unit number，LUN)。当一个Target绑定有一个RBD客户端接口时，该RBD客户端接口的LUN默认为LUN0；当一个Target绑定有n个RBD客户端接口时，所述n个RBD客户端接口分别为LUN0-LUNn。

[0125] 系统管理员为所述Ceph集群提供的各个RBD块设备设置了相应的预期占有指标(QoS指标)，所述QoS指标设置在各个RBD块设备的元数据中，通过每个RBD块设备对应的RBD客户端接口可以读取到该元数据中的内容。图4所示的限流器用于实现QoS控制算法，该QoS控制算法通过设置在RBD块设备中的QoS指标对分配给所述RBD块设备的系统资源进行限制。

[0126] QoS指标包括预留指标(预期占有下限)、上限指标(预期占有上限)、权重指标(预期占有权重)及根据上限指标确定的限流值，每个RBD块设备的初始限流值通常等于该RBD块设备中配置的上限指标的值。

[0127] 其中，所述权重指标被划分为低、中、高三个级别，每个级别对应不同的比例，如，低级对应20％，中级对应30％，高级对应50％。进一步地，针对权重指标为低级的RBD块设备，禁止系统管理员在该RBD块设备中设置预留指标，即自动将该RBD块设备中的预留指标设置为0。针对权重指标为高级的RBD块设备，必须在该RBD块设备中设置预留指标，即该RBD块设备中设置的预留指标必须大于0。

[0128] 在进行QoS控制时，图4所示的Ceph集群会优先处理未达到预留指标的RBD块设备对应的RBD客户端接口接收到的I/O请求，在所有RBD块设备均达到各自的预留指标时，所述Ceph集群会按照各RBD块设备中配置的权重指标，将系统剩余资源分配给各RBD块设备，并确保每个RBD块设备的IOPS不超过其初始限流值，即确保每个RBD块设备的RBD客户端接口

的IOPS不超过其初始限流值。

[0129] 针对图4所示的Ceph集群，可以通过本申请实施例提供的QoS指标监测方法对所述Ceph集群中的各RBD块设备中配置的QoS指标进行实时监测。其中，实时监测是指每间隔预设时长(如，1-10毫秒)采用所述QoS指标监测方法判断各RBD块设备中设置的QoS指标是否合理。所述QoS指标监测方法具体可以包括如下步骤。

[0130] 第一，查询模块每间隔预设时长与每个Target中的socket文件建立socket连接，通过该Target绑定的各个RBD客户端接口获取该RBD客户端接口对应的RBD块设备当前的QOS指标及当前的IOPS，并统计当前的IOPS没有达到预留指标的RBD块设备的数量。所述数量实际是当前的IOPS没有达到预留指标的RBD块设备(目标逻辑存储块)的数量。

[0131] 第二，在执行第一步骤时，查询模块与RADOS中的每个OSD设备建立socket连接，获取该OSD当前的队列深度、CPU利用率、磁盘利用率及内存占用率四个性能数据，并根据所述四个性能数据确定所述Ceph集群当前的当前负载状态。具体通过如下方式确定：

[0132] 当所述RADOS系统中的所有的OSD的队列深度达到相应的深度阈值，或是所有的OSD的CPU利用率达到相应的CPU利用率阈值，或是所有的OSD的磁盘利用率达到相应的磁盘利用率，或是所有的OSD的内存占用率达到相应的内存占用率阈值时，确定所述Ceph集群繁忙；否则，确定所述Ceph集群空闲。

[0133] 第三，在得到所述Ceph集群中目标逻辑存储块的数量及所述Ceph集群当前的当前负载状态后，通过如下条件判断所述Ceph集群中的各个RBD块设备配置的QoS指标是否合理。具体通过如下条件判断：

[0134] 当所述Ceph集群中存在当前的IOPS没有达到各自的预留指标的RBD块设备，且所述Ceph集群的当前负载状态为繁忙时，确定所述Ceph集群的各RBD块设备中的QoS指标设置不合理。

[0135] 当所述Ceph集群中所有RBD块设备当前的IOPS均达到各自的预留指标，或是当所述Ceph集群中存在当前的IOPS未达到各自的预留指标的RBD块设备，且所述Ceph集群的当前负载状态为空闲时，确定各RBD块设备中的QoS指标设置合理。

[0136] 针对QoS指标设置不合理的情形，可以进行如下处理：当所述Ceph集群中所有RBD块设备当前的IOPS都未达到各自的预留指标，且所述Ceph集群的当前负载状态为繁忙时，向系统管理员的客户端200发送提示信息，以提示系统管理员降低各RBD块设备中配置的预留指标；

[0137] 当所述Ceph集群中部分RBD块设备当前的IOPS没有达到各自的预留指标，且所述Ceph集群的当前负载状态为繁忙时，将达到预留指标的RBD块设备的预留指标作为该RBD块设备的当前限流值。

[0138] 针对QoS指标设置合理的情形，可以不做任何处理，但是，当所述Ceph集群中所有RBD块设备当前的IOPS均达到各自的预留指标时，可以根据各RBD块设备中配置的权重指标，调整各RBD块设备中设置的初始限流值。具体可以通过如下步骤实现。

[0139] 针对每个RBD块设备，计算该RBD块设备当前的IOPS与该RBD块设备中配置的预留指标的差值，从而得到多个差值，并对所述多个差值求和得到所述Ceph集群当前的剩余系统资源，并将所述剩余系统资源按各RBD块设备中配置的权重指标分配给所述各RBD块设备，并计算每个RBD块设备分配到的系统资源与该RBD块设备中配置的预留指标的和，将得

到的和作为该RBD块设备的当前限流值。

[0140] 如图5所示，本申请实施例还提供一种QoS指标监测装置300，应用于图1所示的分布式存储系统100。所述QoS指标监测装置300可以包括获得模块310、判断模块320以及监测模块330。

[0141] 其中，所述获得模块310用于获得所述分布式存储系统100提供给客户端200的每个逻辑存储块120对系统资源的当前占有量及预期占有指标。

[0142] 其中，所述分布式存储系统100根据所述预期占有指标为各逻辑存储块120分配系统资源，每个逻辑存储块120对系统资源的预期占有指标包括该逻辑存储块120对系统资源的预期占有下限。

[0143] 在本实施例中，关于所述获得模块310的描述具体可参考对图2所示步骤S110的详细描述，也即，所述步骤S110可以由所述获得模块310执行。

[0144] 可选地，针对每个目标进程，所述获得模块310可以通过与所述目标进程连接的各逻辑存储块120获取所述各逻辑存储块120对系统资源的当前占有量及预期占有指标。

[0145] 所述判断模块320用于针对每个逻辑存储块120，判断该逻辑存储块120对所述系统资源的当前占有量是否达到该逻辑存储块120的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量。

[0146] 在本实施例中，关于所述判断模块320的描述具体可参考对图2所示步骤S120的详细描述，也即，所述步骤S120可以由所述判断模块320执行。

[0147] 所述监测模块330用于根据所述分布式存储系统100的当前负载状态及所述目标逻辑存储块的数量监测各逻辑存储块120的预期占有指标是否设置合理。

[0148] 其中，所述分布式存储系统100的当前负载状态包括繁忙和空闲两种。

[0149] 若所述目标逻辑存储块的数量大于0且所述分布式存储系统100的当前负载状态为繁忙，所述监测模块330可以确定各逻辑存储块120的预期占有指标设置不合理。

[0150] 若所述目标逻辑存储块的数量大于0且所述分布式存储系统100的当前负载状态为空闲，所述监测模块330可以确定各逻辑存储块120的预期占有指标设置合理。

[0151] 若所述目标逻辑存储块的数量为0，所述监测模块330可以确定各逻辑存储块120的预期占有指标设置合理。

[0152] 可选地，所述分布式存储系统100的当前负载状态可以根据所述分布式存储系统100中每个存储设备110的当前性能数据确定。每个存储设备110的当前性能数据可以通过向每个存储设备110上的管理进程发送请求获得该存储设备110的当前性能数据。

[0153] 所述存储设备110的当前性能数据可以包括当前的队列深度、CPU利用率、磁盘利用率及内存占用率。

[0154] 可选地，所述分布式存储系统100的当前负载状态可以通过如下方式确定：

[0155] 判断每个存储设备110的当前性能数据是否满足预设条件，在均满足所述预设条件中的至少一个时，确定所述分布式存储系统100的当前负载状态为繁忙；否则，确定所述分布式存储系统100的当前负载状态为空闲。

[0156] 其中，所述预设条件可以包括以下至少一项：

[0157] 每个存储设备110的队列深度均达到设定的深度阈值；

[0158] 每个存储设备110的CPU利用率均达到设定的CPU利用率阈值；

[0159] 每个存储设备110的磁盘利用率均达到设定的磁盘利用率阈值；

[0160] 每个存储设备110的内存占用率均达到设定的内存占用率阈值。

[0161] 在本实施例中，关于所述监测模块330的描述具体可参考对图2所示步骤S130的详细描述，也即，所述步骤S130可以由所述监测模块330描述。

[0162] 可选地，所述QoS指标监测装置300还可以包括处理模块340。

[0163] 所述处理模块340用于在确定各逻辑存储块120的预期占有指标设置不合理之后，判断所述目标逻辑存储块的数量是否等于所述分布式存储系统100提供给客户端200的逻辑存储块120的数量；若等于，则发出降低每个逻辑存储块120的预期占有下限的提示信息；若不等于，则根据各逻辑存储块120的预期占有下限对对应逻辑存储块120的限流值进行调整。

[0164] 其中，所述分布式存储系统100为每个逻辑存储块120设置有用于限定为所述逻辑存储块120分配的系统资源上限的限流值，该限流值的初始值根据所述逻辑存储块120对系统资源的预期占有上限确定。该限流值被设置在该逻辑存储块120的元数据中，可通过该逻辑存储块120对应的存储块接口获取元数据中的限流值，分布式存储系统100在分配系统资源时，会确保分配给该存储块接口对应的逻辑存储块120的系统资源不超过该限流值。

[0165] 在本实施例中，关于所述处理模块340的描述具体可参考上述内容中对相关步骤的详细描述。

[0166] 现有的分布式存储系统100通常通过如下QoS控制算法进行QoS控制：

[0167] 优先将系统资源分配给当前占有量未达到预期占有下限的逻辑存储块120，在每个逻辑存储块120的当前占有量达到预期占有下限时，优先将系统资源分配给预期占有权重大的逻辑存储块120，并确保每个逻辑存储块120分配到的系统资源不超过该逻辑存储块120的当前限流值。

[0168] 在上述QoS控制算法的基础上，所述处理模块340对各逻辑存储块120的限流值进行调整的方式可以为：

[0169] 针对每个当前占有量达到预期占有下限的逻辑存储块120，将该逻辑存储块120的预期占有下限作为该逻辑存储块120的当前限流值。

[0170] 可选地，在所述分布式存储系统100按照上述方式进行QoS控制的基础上，所述QoS指标监测装置300还可以包括限流更新模块350。

[0171] 所述限流更新模块350用于当所述目标逻辑存储块的数量为0时，计算每个逻辑存储块120的当前占有量与预期占有下限的差值，得到多个差值；对所述多个差值求和，得到所述分布式存储系统100当前的剩余系统资源；按照各逻辑存储块120的预期占有权重，将所述剩余系统资源分配给每个逻辑存储块120；对每个逻辑存储块120分配到的系统资源量及及该逻辑存储块120的预期占有下限求和，并将求得的和作为该逻辑存储块120的当前限流值。

[0172] 本申请实施例还提供一种存储介质，其上存储有计算机可读指令，所述计算机可读指令被执行时实现本申请实施例提供的所述QoS指标监测方法。

[0173] 综上所述，本申请实施例提供的QoS指标监测方法、装置、存储介质，通过获得的每个存储设备110的当前性能数据，判断分布式存储系统100的当前负载状态；通过获得的分布式存储系统100提供给客户端200的每个逻辑存储块120对系统资源的当前占有量和预期

占有指标，得到对系统资源的当前占有量未达到预期占有下限的目标逻辑块的数量；并根据分布式存储系统100的当前负载状态及目标逻辑存储块的数量，监测各逻辑存储块120的预期占有指标是否合理。如此，可以为预期占有指标的动态调节提供依据，从而使分布式存储系统100的服务质量始终达到用户的预期，改善用户体验。

[0174] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

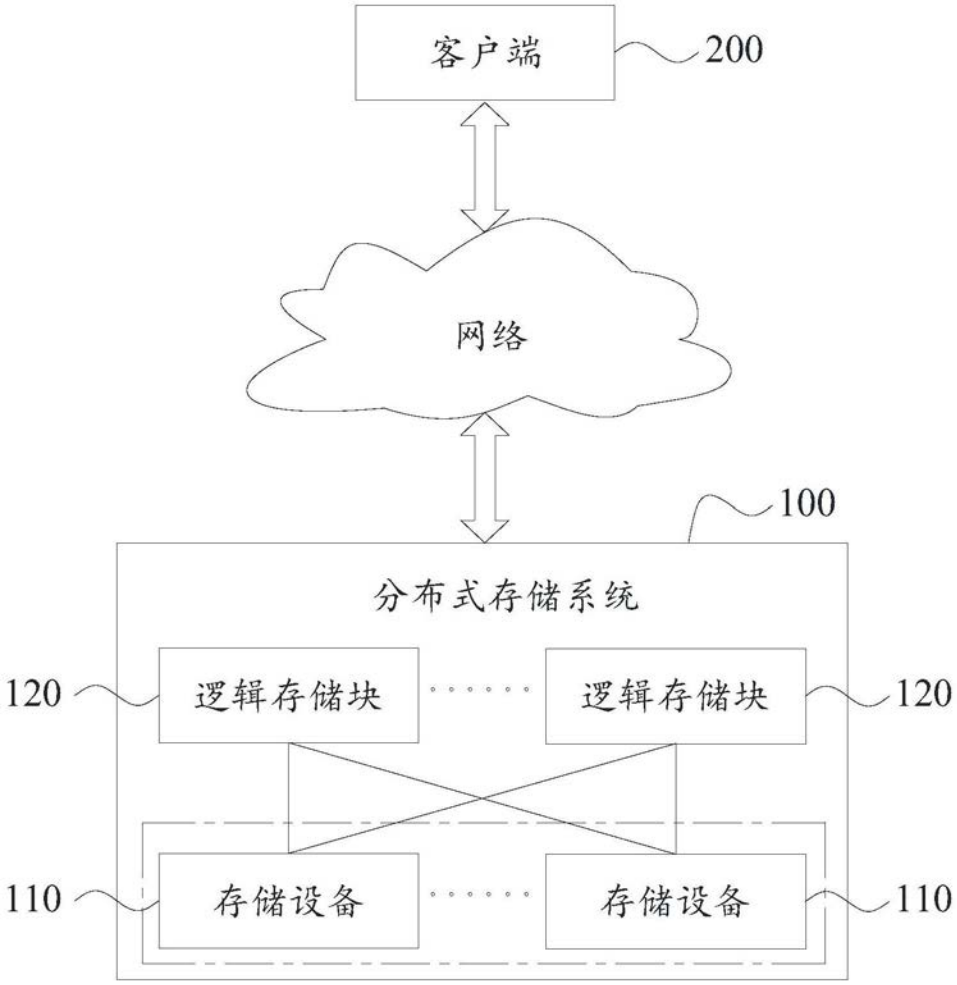


图1

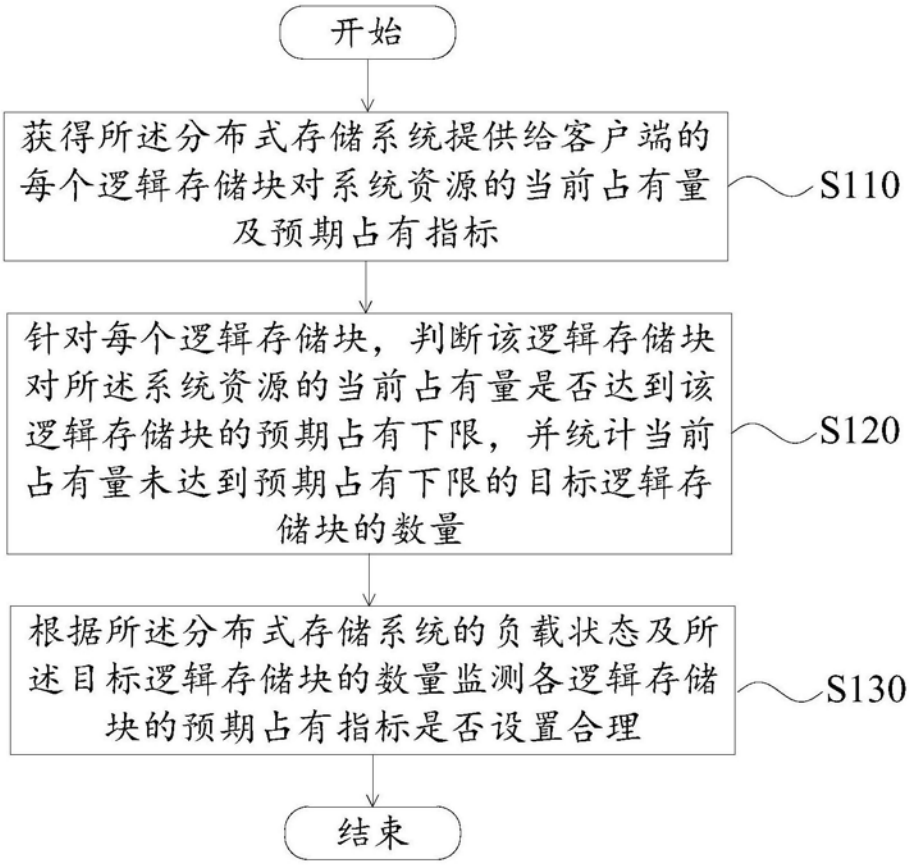

开始
获得所述分布式存储系统提供给客户端的每个逻辑存储块对系统资源的当前占有量及预期占有指标
S110
针对每个逻辑存储块，判断该逻辑存储块对所述系统资源的当前占有量是否达到该逻辑存储块的预期占有下限，并统计当前占有量未达到预期占有下限的目标逻辑存储块的数量
S120
根据所述分布式存储系统的负载状态及所述目标逻辑存储块的数量监测各逻辑存储块的预期占有指标是否设置合理
S130
结束

图2

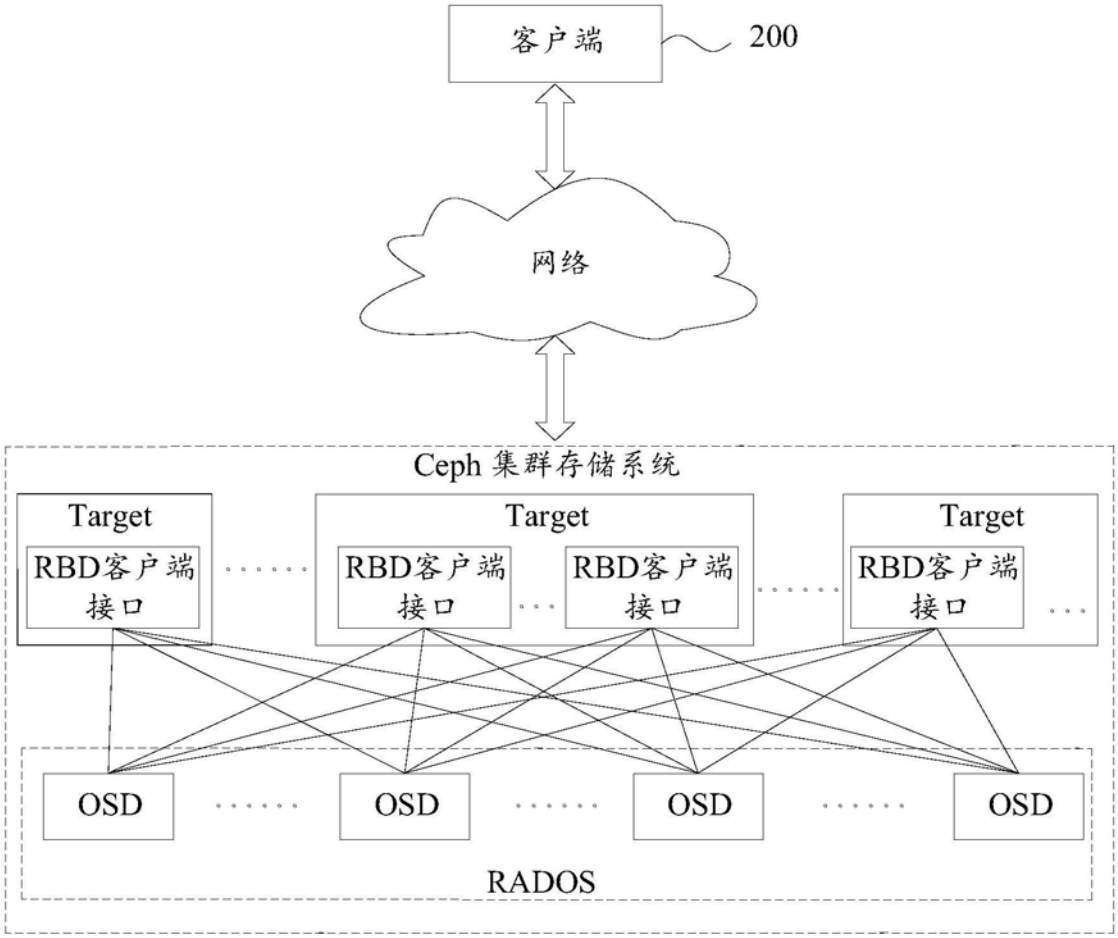

客户端
200
网络
Ceph 集群存储系统
Target
RBD客户端接口
Target
RBD客户端接口
RBD客户端接口
Target
RBD客户端接口
OSD
OSD
OSD
OSD
RADOS

图3

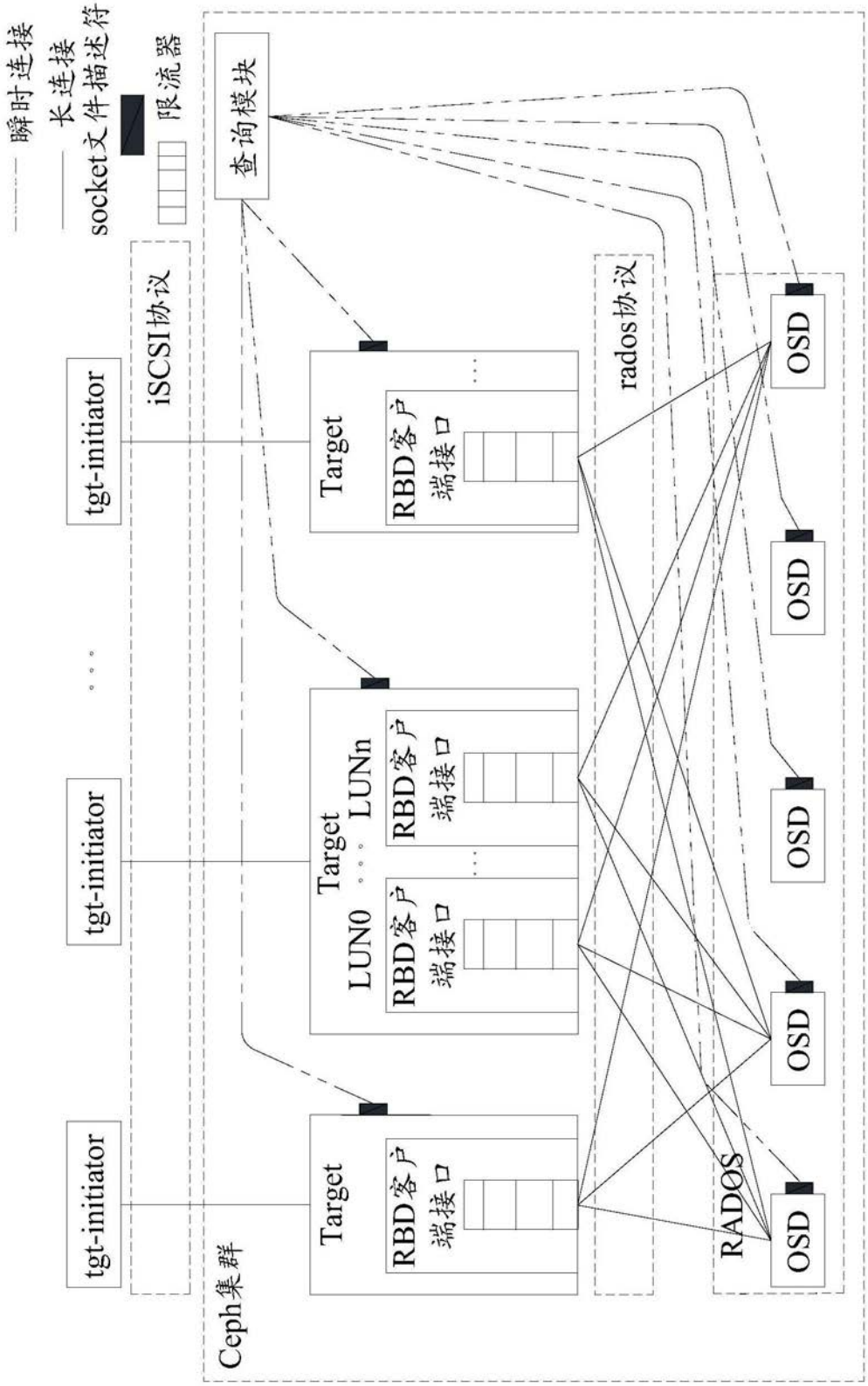

瞬时连接
长连接
socket文件描述符
限流器
查询模块
iSCSI协议
rados协议
tgt-initiator
tgt-initiator
tgt-initiator
Target
RBD客户端接口
Target
LUN0 ... LUNn
RBD客户端接口
RBD客户端接口
Target
RBD客户端接口
OSD
OSD
OSD
OSD
OSD
OSD
RADOS
Ceph集群

图4

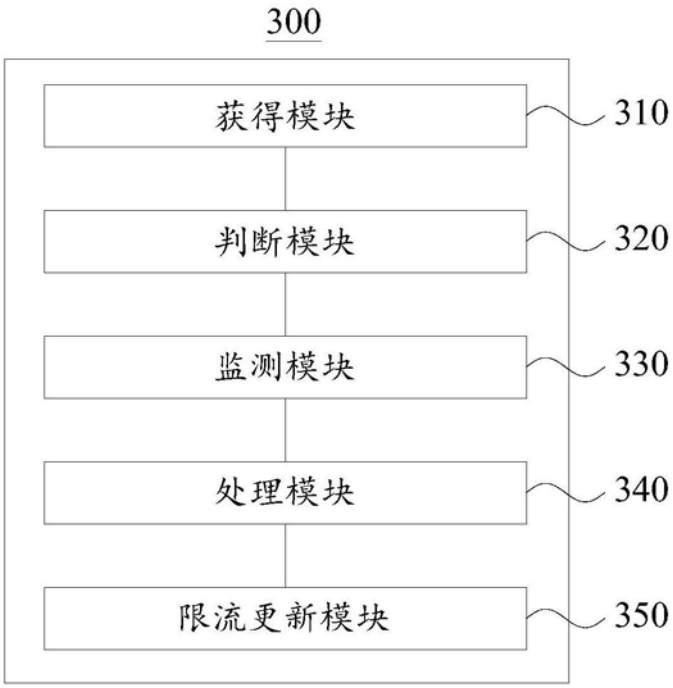


图5