(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 107491455 B
(45)授权公告日 2020.11.20

(21)申请号 201610412855.4
(22)申请日 2016.06.13
(65)同一申请的已公布的文献号
申请公布号 CN 107491455 A
(43)申请公布日 2017.12.19
(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱
(72)发明人 刘善阳
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 李红爽 栗若木
(51)Int.Cl.
G06F 16/182(2019.01)
G06F 16/13(2019.01)

(56)对比文件
CN 105635252 A,2016.06.01
CN 105468660 A,2016.04.06
CN 101137089 A,2008.03.05
CN 103685542 A,2014.03.26
CN 104219157 A,2014.12.17
US 2014337539 A1,2014.11.13
CN 101252499 A,2008.08.27
maray.“服务器程序中如何设计backup task功能”.《https://blog.csdn.net/maray/article/details/8616385》.2013,

审查员 倪礼

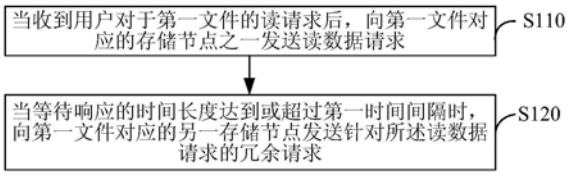
权利要求书2页 说明书10页 附图1页

(54)发明名称

一种分布式系统中的读取方法及装置

(57)摘要

一种分布式系统中的读取方法及装置;所述读取方法包括:当收到用户对于第一文件的读请求后,向所述第一文件对应的存储节点之一发送读数据请求;当等待响应的时间长度达到或超过第一时间间隔时,向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求;其中,所述第一时间间隔是根据读操作的性能指标而动态确定的。本申请能够自适应调整针对读数据请求的冗余请求的发送间隔。



CN 107491455 B

1.一种分布式系统中的读取方法,包括:

当收到用户对于第一文件的读请求后,向所述第一文件对应的存储节点之一发送读数据请求;

当等待响应的时间长度达到或超过第一时间间隔时,向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求;

其中,所述第一时间间隔是根据读操作的性能指标而动态确定的;

所述读操作的性能指标包括:

读操作的延时,和/或,读操作中发送冗余请求的次数。

2.如权利要求1所述的读取方法,其特征在于:

所述读操作的延时是之前一次或之前多次读操作的延时,或者是待确定所述第一时间间隔的时刻之前预定长度的时间内,读操作的延时;

所述读操作中发送冗余请求的次数是之前一次或之前多次读操作中发送冗余请求的次数,或者是待确定所述第一时间间隔的时刻之前预定长度的时间内,读操作中发送冗余请求的次数。

3.如权利要求1所述的读取方法,其特征在于,还包括:

在每次读操作后,记录本次读操作的延时和发送冗余请求的次数。

4.如权利要求1所述的读取方法,其特征在于,所述第一时间间隔根据读操作的性能指标而动态确定包括:

所述第一时间间隔根据L_{avg}以及Qps计算得到;其中,L_{avg}是当前时刻之前预定长度的时间内读操作的延时的平均值;Qps是当前时刻之前预定长度的时间内冗余请求的频率。

5.如权利要求4所述的读取方法,其特征在于,所述第一时间间隔根据L_{avg}以及Qps计算得到包括:

所述第一时间间隔等于预定函数的函数值;所述预定函数的自变量为L_{avg}、Qps,所述预定函数的函数值随L_{avg}单调递增,随Qps单调递增。

6.如权利要求1～5中任一项所述的读取方法,其特征在于,所述第一时间间隔根据读操作的性能指标而动态确定包括:

所述第一时间间隔周期性根据读操作的性能指标动态确定;

或者,所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

7.一种分布式系统中的读取装置,其特征在于,包括:

第一请求模块,用于当收到用户对于第一文件的读请求后,向所述第一文件对应的存储节点之一发送读数据请求;

第二请求模块,用于当等待响应的时间长度达到或超过第一时间间隔时,向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求;其中,所述第一时间间隔是根据读操作的性能指标而动态确定的;

其中,所述读操作的性能指标包括:

读操作的延时,和/或,读操作中发送冗余请求的次数。

8.如权利要求7所述的读取装置,其特征在于:

所述读操作的延时是之前一次或之前多次读操作的延时,或者是待确定所述第一时间间隔的时刻之前预定长度的时间内,读操作的延时;

所述读操作中发送冗余请求的次数是之前一次或之前多次读操作中发送冗余请求的次数，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作中发送冗余请求的次数。

9.如权利要求7所述的读取装置，其特征在于，还包括：

记录模块，用于在每次读操作后，记录本次读操作的延时和发送冗余请求的次数。

10.如权利要求7所述的读取装置，其特征在于，所述第一时间间隔根据读操作的性能指标而动态确定包括：

所述第一时间间隔根据L_{avg}以及Qps计算得到；其中，L_{avg}是待确定所述第一时间间隔的时刻之前预定长度的时间内读操作的延时的平均值；Qps是待确定所述第一时间间隔的时刻之前预定长度的时间内冗余请求的频率。

11.如权利要求10所述的读取装置，其特征在于，所述第一时间间隔根据L_{avg}以及Qps计算得到包括：

将预定函数的函数值作为所述第一时间间隔；所述预定函数的自变量为L_{avg}、Qps，所述预定函数的函数值随L_{avg}单调递增，随Qps单调递增。

12.如权利要求7～11中任一项所述的读取装置，其特征在于，所述第一时间间隔根据读操作的性能指标而动态确定包括：

所述第一时间间隔周期性根据读操作的性能指标动态确定；

或者，所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

一种分布式系统中的读取方法及装置

技术领域

[0001] 本发明涉及领域，尤其涉及一种分布式系统中的读取方法及装置。

背景技术

[0002] 针对分布式系统中读毛刺的优化，相关技术中提出了一种BackupRequest的策略。BackupRequest是在普通的读流程中添加的冗余请求，在不要求单个请求的读延时降低的基础上，降低用户感知的读延时毛刺率。

[0003] 以Google的谷歌文件系统(Google File System，GFS)为例来描述包含BackupRequest的读取过程，谷歌文件系统具体结构如图1所示：

[0004] 客户端(Client)库：为分布式存储系统的用户提供各种接口；

[0005] 块服务器(Chunkserver)：作为数据管理模块，具体管理用户的数据；

[0006] 主控端(Master)：作为命名空间管理模块，管理分布式存储系统的元(meta)数据。

[0007] 在客户端，主控端，块服务器模式的分布式存储系统中，用户文件的所有元数据都存储在主控端中。文件的数据以多副本的方式保存在不同的块服务器中。

[0008] 客户端发送文件名称(file name)和块索引(Chunk index)到主控端，主控端返回块句柄(Chunk handle)和块位置(Chunk location)；客户端发送块句柄和字节范围(byte range)到块服务器，块服务器返回块数据(Chunk data)。块服务器还向主控端上报块服务器状态(state)，主控端下发对于块服务器的指令(Instructions to Chunkserver)给块服务器。其中，主控端中文件命名空间包含(File namespace)；块服务器使用Linux文件系统(file system)。其中，块服务器返回给客户端的块数据是数据消息，客户端发送给块服务器的块句柄和字节范围、客户端与主控端之间、主控端与块服务器之间交互的均是控制消息。

[0009] 带有BackupRequest的读流程如下：

[0010] 101、客户端接收到用户的对文件F读数据的请求；

[0011] 102、客户端向主控端请求文件F的数据所在块的信息；

[0012] 103、主控端将块的多个副本的块服务器的地址返回给客户端；

[0013] 104、客户端向其中一个副本所在的块服务器发起读数据请求，并等待块服务器返回结果，此时用户后续的写被阻塞；

[0014] 105、如果步骤104的等待超过发送间隔T，客户端会再向另一个副本所在的块服务器发起读数据请求，这个请求被称作BackupRequest；

[0015] 106、步骤104和步骤105中任何一个请求先成功，客户端就会返回给用户表示读成功的消息。

[0016] 用户收到读成功的结果后，如果继续发起下一次读，则回到上述的步骤101。

[0017] 上述的BackupRequest方法用冗余读的方法来实现读毛刺率的降低。在冗余读的方法中，有一个重要的参数决定了方法的性能和稳定性：冗余请求的发送间隔。发送间隔越短，对毛刺率的优化效果越好，但需要的额外的资源越多，相关技术中，BackupRequest方法

中发送间隔设置定义成固定值，在前端压力基本不变的场景可以正常工作，但在压力变化明显的场景下，固定的发送间隔的会有如下问题：

[0018] 压力小的场景发送间隔过大，因为没有及时的发送冗余请求，所以无法得到更好的毛刺率优化效果；

[0019] 压力大的场景发送间隔过小，因为冗余请求的发送本身也耗费资源，在压力大的场景下，这种耗费不但不能优化毛刺率，反而会增加更多的排队等待，造成毛刺率升高和冗余请求发送增多的恶性循环，降低系统稳定性。

发明内容

[0020] 本申请提供一种分布式系统中的读取方法及装置，能够自适应调整针对读数据请求的冗余请求的发送间隔。

[0021] 本申请采用如下技术方案。

[0022] 一种分布式系统中的读取方法，包括：

[0023] 当收到用户对于第一文件的读请求后，向所述第一文件对应的存储节点之一发送读数据请求；

[0024] 当等待响应的时间长度达到或超过第一时间间隔时，向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求；

[0025] 其中，所述第一时间间隔是根据读操作的性能指标而动态确定的。

[0026] 可选地，所述读操作的性能指标包括：

[0027] 读操作的延时，和/或，读操作中发送冗余请求的次数。

[0028] 可选地，所述读操作的延时是之前一次或之前多次读操作的延时，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作的延时；

[0029] 所述读操作中发送冗余请求的次数是之前一次或之前多次读操作中发送冗余请求的次数，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作中发送冗余请求的次数。

[0030] 可选地，所述的读取方法还包括：

[0031] 在每次读操作后，记录本次读操作的延时和发送冗余请求的次数。

[0032] 可选地，所述第一时间间隔根据读操作的性能指标而动态确定包括：

[0033] 所述第一时间间隔根据L_{avg}以及Qps计算得到；其中，L_{avg}是当前时刻之前预定长度的时间内读操作的延时的平均值；Qps是当前时刻之前预定长度的时间内冗余请求的频率。

[0034] 可选地，所述第一时间间隔根据L_{avg}以及Qps计算得到包括：

[0035] 所述第一时间间隔等于预定函数的函数值；所述预定函数的自变量为L_{avg}、Qps，所述预定函数的函数值随L_{avg}单调递增，随Qps单调递增。

[0036] 可选地，所述第一时间间隔根据读操作的性能指标而动态确定包括：

[0037] 所述第一时间间隔周期性根据读操作的性能指标动态确定；；

[0038] 或者，所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

[0039] 一种分布式系统中的读取装置，包括：

[0040] 第一请求模块，用于当收到用户对于第一文件的读请求后，向所述第一文件对应

的存储节点之一发送读数据请求；

[0041] 第二请求模块，用于当等待响应的时间长度达到或超过第一时间间隔时，向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求；其中，所述第一时间间隔是根据读操作的性能指标而动态确定的。

[0042] 可选地，所述读操作的性能指标包括：

[0043] 读操作的延时，和/或，读操作中发送冗余请求的次数。

[0044] 可选地，所述读操作的延时是之前一次或之前多次读操作的延时，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作的延时；

[0045] 所述读操作中发送冗余请求的次数是之前一次或之前多次读操作中发送冗余请求的次数，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作中发送冗余请求的次数。

[0046] 可选地，所述的读取装置还包括：

[0047] 记录模块，用于在每次读操作后，记录本次读操作的延时和发送冗余请求的次数。

[0048] 可选地，所述第一时间间隔根据读操作的性能指标而动态确定包括：

[0049] 所述第一时间间隔根据L_{avg}以及Qps计算得到；其中，L_{avg}是待确定所述第一时间间隔的时刻之前预定长度的时间内读操作的延时的平均值；Qps是待确定所述第一时间间隔的时刻之前预定长度的时间内冗余请求的频率。

[0050] 可选地，所述第一时间间隔根据L_{avg}以及Qps计算得到包括：

[0051] 将预定函数的函数值作为所述第一时间间隔；所述预定函数的自变量为L_{avg}、Qps，所述预定函数的函数值随L_{avg}单调递增，随Qps单调递增。

[0052] 可选地，所述第一时间间隔根据读操作的性能指标而动态确定包括：

[0053] 所述第一时间间隔周期性根据读操作的性能指标动态确定；

[0054] 或者，所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

[0055] 本申请包括以下优点：

[0056] 本申请至少一个备选方案针对读取过程中的冗余请求，提出一种自适应的确定冗余请求的发送间隔的方法，针对读数据请求的冗余请求的发送间隔不再是固定的，而是根据情况动态调整的，可以根据读操作执行过程中实时的性能指标变化而相应变化，从而能够更好地适应当前的实际情况

[0057] 本申请又一个备选方案中，所述读操作的性能指标包括读操作的延时，和/或，读操作中发送冗余请求的次数；这样可以较为准确地反映出前端压力的变化情况；根据读操作的延时，和/或读操作中发送冗余请求的次数来调整冗余请求的发送间隔，可以使冗余请求的发送间隔能较为精确地跟随前端压力的改变而变化。

[0058] 本申请又一个备选方案采用最近一段时间内的读操作延时和冗余请求频率作为调整依据，可以反映出变化趋势，使调整结果更为准确。

[0059] 本申请又一个备选方案中，冗余请求的发送间隔随延时平均值和冗余请求频率单调递增，在压力变小的时候可以尽可能地发挥冗余请求对毛刺率的优化效果，在压力变大的时候可以抑制冗余请求对系统稳定性的负面影响；从而在不同压力下兼顾性能和稳定性。

[0060] 当然,实施本申请的任一产品必不一定需要同时达到以上所述的所有优点。

附图说明

[0061] 图1是谷歌文件系统的结构示意图;

[0062] 图2是实施例一的分布式系统中的读取方法的流程图;

[0063] 图3是实施例二的分布式系统中的读取装置的示意图。

具体实施方式

[0064] 下面将结合附图及实施例对本申请的技术方案进行更详细的说明。

[0065] 需要说明的是,如果不冲突,本申请实施例以及实施例中的各个特征可以相互结合,均在本申请的保护范围之内。另外,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0066] 在一个典型的配置中,客户端或服务器的计算设备可包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存(memory)。

[0067] 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。内存可能包括模块1,模块2,……,模块N(N为大于2的整数)。

[0068] 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,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,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。

[0069] 实施例一、一种分布式系统中的读取方法,如图2所示,包括步骤S110～S120:

[0070] S110、当收到用户对于第一文件的读请求后,向所述第一文件对应的存储节点之一发送读数据请求;

[0071] S120、当等待响应的时间长度达到或超过第一时间间隔时,向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求;

[0072] 其中,所述第一时间间隔是根据读操作的性能指标而动态确定的。

[0073] 本实施例中,针对读数据请求的冗余请求的发送间隔(即所述第一时间间隔)不再是固定的,而是根据情况动态调整的,动态调整意味着第一时间间隔会跟随读操作的性能指标变化而发生变化;由于读操作的性能指标可以反应出前端压力的变化情况,因此可以根据读操作的性能指标变化而相应变化第一时间间隔,从而改变发送冗余请求的频率,比如可以在压力小的时候积极发送冗余请求,压力大的时候消极发送冗余请求,从而能够更好地适应当前的实际情况。

[0074] 本实施例中,所述步骤S110、步骤S120可以但不限于均由客户端或设置于客户端中的装置/模块实施。所述第一时间间隔的动态确定可以但不限于由客户端完成,此时客户

端可用于记录和统计写操作的性能指标，以及根据写操作的性能指标设置所述第一时间间隔；也不排除采用另外的设备完成第一时间间隔的动态确定并通知给客户端的做法。

[0075] 本实施例中，所述第一文件对应的存储节点，即第一文件的副本所在的存储节点，一般为两个或两个以上，可以通过向主控端请求获知一个文件对应的存储节点。所述存储节点包括但不限于块服务器、扩展节点等。当存储节点是块服务器时，所述第一文件对应的存储节点是指：所述第一文件可以读数据的块对应的块服务器。

[0076] 本实施例中，所述第一时间间隔表示时间的长度，单位可以但不限于是秒、毫秒、微秒等；当向所述第一文件对应的存储节点之一发送读数据请求后，从发送时刻开始计时，如果在未达到或超过所述第一时间间隔时就收到了对于所述读数据请求的响应，则可以停止计时，该情况下也可以不发送针对所述读数据请求的冗余请求。

[0077] 本实施例中，如果第一文件对应的存储节点超过两个时，当发送针对所述读数据请求的冗余请求后，也同样可以从发送时刻开始计时，如果达到或超过所述第一时间间隔时还未收到对于所述冗余请求的响应，则向第一文件对应的再一个存储节点发送另一个针对读数据请求的冗余请求；以此类推，直到不存在第一文件对应的、未发送过读数据请求或冗余请求的存储节点为止。

[0078] 比如假设第一文件有三个副本，分别在存储节点A、B、C上，收到用户读请求后先向存储节点A发送读数据请求，如果等待存储节点A响应的时间长度达到或超过所述第一时间间隔，则向存储节点B发送针对所述读数据请求的冗余请求。如果等待存储节点B响应的时间长度也达到或超过所述第一时间间隔，则还可以向存储节点C发送针对所述读数据请求的冗余请求。副本为三个以上的情况可以类推。当然，也可以设置成一次读操作中(即针对同一个读数据请求)只许发送一次或预定次数的冗余请求。

[0079] 本实施例中，无论是收到针对所述读数据请求的表示读成功的响应，还是收到针对读数据的任一冗余请求的表示读成功的响应，都说明本次读操作成功，可以向用户返回表示读成功的消息。如果对于读数据请求和任一冗余请求都没有收到表示读成功的响应(没收到响应，或响应表示读失败)，则说明本次读操作失败，向用户返回表示读失败的消息。

[0080] 本实施例的一种备选方案中，所述读操作的性能指标可以包括：

[0081] 读操作的延时，和/或，读操作中发送冗余请求的次数。

[0082] 本备选方案中，所述读操作的延时可以定义为接收到用户读请求后第一次发送读数据请求的时刻，与接收到表示读数据成功的响应的时刻之间所间隔的时间长度；如果第一次发送的读数据请求和冗余请求都没有成功，则读操作的延时也可以是预定的超时时间长度(即，从第一次发起读数据请求开始，到达该超时时间长度时仍未接收到表示读数据成功的响应，则判断读操作失败)。所述读操作的延时也可以定义为从接收到用户读请求开始，到完成该读请求(包括向用户返回表示读成功或读失败的消息)为止整个过程的耗时。其中，如果对于读数据请求和针对所述读数据请求的冗余请求都收到表示读成功的响应，则接收到表示读成功的响应的时刻是指接收到的第一个表示读成功的响应的时刻。对于读操作的延时的定义不限于上面的示例，还可以根据需要自行设置。

[0083] 本备选方案中，所述读操作的延时、读操作中发送冗余请求的次数可以较为准确地反映出前端压力的变化情况；根据读操作的延时，和/或，读操作中发送冗余请求的次数

来调整第一时间间隔，可以使第一时间间隔能较为精确地跟随前端压力的改变而变化。

[0084] 本备选方案中，所述读操作的延时可以是之前一次或之前多次读操作的延时，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作的延时；

[0085] 所述读操作中发送冗余请求的次数可以是之前一次或之前多次读操作中发送冗余请求的次数，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作中发送冗余请求的次数。

[0086] 本备选方案中，所述方法还可以包括：

[0087] 在每次读操作后，记录下来本次读操作的延时和发送冗余请求的次数。

[0088] 本备选方案的一种实施方式中，可以只根据读操作的延时、或者只根据读操作中发送冗余请求的次数来确定所述第一时间间隔；可以设置成该第一时间间隔随读操作的延时或发送冗余请求的次数变大而变大。本实施方式可以只记录一种性能指标，节省存储资源，而且确定所述第一时间间隔时的处理也会相对简单，可以少占用处理资源，提高处理效率。

[0089] 本实施方式一种可选的方案是根据以读操作的延时或发送冗余请求的次数作为自变量，以第一时间间隔作为应变量的函数，来确定所述第一时间间隔。另一种可选的方案是建立读操作的延时或发送冗余请求的次数的数值范围和第一时间间隔之间的对应关系，根据该对应关系来确定第一时间间隔；比如当读操作的延时或发送冗余请求的次数属于第一数值范围时，第一时间间隔为第一时间长度T1，当读操作的延时或发送冗余请求的次数属于第二数值范围时，第一时间间隔为第二时间长度T2，以此类推。当然，也可以使用其它根据读操作的延时或发送冗余请求的次数来确定第一时间间隔的方案。

[0090] 本备选方案的另一种实施方式中，可以根据读操作的延时和发送冗余请求的次数共同来确定所述第一时间间隔；可以设置成该第一时间间隔随读操作的延时和发送冗余请求的次数变大而变大。本实施方式两个性能指标可以互相作为参考和修正，使结果更加准确。

[0091] 本实施方式一种可选的方案是根据以读操作的延时以及发送冗余请求的次数作为自变量，以第一时间间隔作为应变量的函数，来确定所述第一时间间隔。另一种可选的方案是对读操作的延时和发送冗余请求的次数这两者进行预定计算，建立计算得到的结果的数值范围和第一时间间隔之间的对应关系，根据该对应关系确定第一时间间隔。当然，也可以使用其它根据读操作的延时和发送冗余请求的次数来确定第一时间间隔的方案。

[0092] 本备选方案的一种实施方式中，利用前一次读操作的性能指标来确定所述第一时间间隔，本实施方式可以节省存储空间，计算量也相对较小，计算速度快，前端压力的变化可以立刻体现在第一时间间隔的变化上，实时性比较好。

[0093] 本备选方案的另一种实施方式中，利用多次读操作(可以是待确定所述第一时间间隔的时刻之前一段时间内的多次读操作，也可以是之前多次读操作)的性能指标来确定所述第一时间间隔，本实施方式可以体现出前端压力的变化趋势，能够减缓前端压力暂时性的突变带来的影响，更为客观和准确。其它备选方案中，也可以选用读操作的其它性能指标或性能指标的组合作为调整所述第一时间间隔的依据；或者选用读操作的其它性能指标或性能指标的组合，与读操作的延时，和/或，读操作中发送冗余请求的次数进行组合，作为调整所述第一时间间隔的依据。所述其它性能指标比如但不限于包括：数据吞吐率、读操作

的成功率、收到用户的读请求后发送冗余请求的概率或比例等。

[0094] 本实施例的一种备选方案中，所述第一时间间隔根据读操作的性能指标而动态确定可以包括：

[0095] 所述第一时间间隔根据L_{avg}以及Qps计算得到；其中，L_{avg}是待确定所述第一时间间隔的时刻之前预定长度的时间内读操作的延时的平均值；Qps是待确定所述第一时间间隔的时刻之前预定长度的时间内冗余请求的频率。

[0096] 本备选方案采用最近一段时间内的读操作延时和冗余请求次数作为调整依据，根据延时的平均值和单位时间中发送冗余请求的次数进行计算，可以得到一段时间内延时和发送次数的变化趋势，从而可以更加准确地反映出前端压力的变化趋势，使调整后的第一时间间隔对于当前的实际情况更为合适。

[0097] 本备选方案中，所述预定长度可以根据经验值或试验自行确定和修改。所述待确定所述第一时间间隔的时刻可以但不限于是指为了调整所述第一时间间隔而计算L_{avg}以及Qps的时刻、或者触发进行第一时间间隔调整的时刻(比如到达调整周期的时刻、调整的触发条件被满足的时刻等)。所述Qps是用待确定所述第一时间间隔的时刻之前预定长度的时间内冗余请求的总次数(即，这段时间内针对用户一次或多次读请求，总共发送的冗余请求的次数)除以所述预定长度的时间。

[0098] 其它实施方式中，也可以采用另外的方式实现对第一时间间隔的动态确定，比如采用待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作的延时及发送冗余请求的次数的的累加值来计算所述第一时间间隔；再比如根据待确定所述第一时间间隔的时刻之前预定长度的时间内发送冗余请求的次数以及读操作的次数，求出每次读操作的平均发送次数，然后和延时的平均值一起用于计算所述第一时间间隔。

[0099] 本实施方式中，所述第一时间间隔根据L_{avg}以及Qps计算得到可以包括：

[0100] 所述第一时间间隔等于预定函数的函数值；所述预定函数的自变量为L_{avg}、Qps，所述预定函数的函数值随L_{avg}单调递增，随Qps单调递增。

[0101] 当前端压力变小的时候，延时和发送冗余请求的频率会相应降低，因此第一时间间隔也会变短，可以尽可能地发挥冗余请求对毛刺率的优化效果；当前端压力变大的时候，延时和发送冗余请求的频率会相应增加，因此第一时间间隔也会变长，可以抑制冗余请求对系统稳定性的负面影响。将L_{avg}以及Qps代入预定函数，可以非常便捷地计算出相应的第一时间间隔。

[0102] 本实施方式中，所述预定函数可以但不限于为：

[0103] $a1\times L_{avg}+a2\times Qps+a3$；

[0104] 或者，$b1\times L_{avg}\times Qps+b2$；

[0105] 其中，a1、a2、a3、b1、b2是预定值。

[0106] 实际应用中，也可以对上述函数式进行变换，比如改变计算符号等；还可以采用其它不同的函数作为所述预定函数。

[0107] 当然，也可以采用其它方式计算所述第一时间间隔，比如将L_{avg}以及Qps代入预定的第一计算式；将得到的计算结果代入第二计算式得到第一时间间隔；再比如根据L_{avg}以及Qps计算出第一时间间隔的调整方向(增加或减少)，根据调整方向、预定步长或根据L_{avg}和/或Qps计算出的步长，在当前的第一时间间隔的基础上得到调整后的第一时间间隔。

[0108] 本实施例的一种备选方案中，所述第一时间间隔根据读操作的性能指标而动态确定包括：

[0109] 所述第一时间间隔周期性根据读操作的性能指标动态确定；

[0110] 或者，所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

[0111] 本备选方案中，所述触发条件可以但不限于包括以下任一个：收到用户的读请求、生成或发送读数据请求、完成一次读请求、读操作的性能指标发生改变或变动幅度超过阈值等。

[0112] 本备选方案中，在周期性调整第一时间间隔的情况下，当收到用户读请求并发送读数据请求后，根据当前的第一时间间隔判断是否发送针对所述读数据请求的冗余请求。

[0113] 下面用一个例子说明本实施例。该例子中，存储节点是块服务器。

[0114] 该例子中，客户端统计整个进程在最近一段时间内读操作的平均延时L_{avg}、和实际发送冗余请求的频率Qps，根据第一时间间隔T的自适应函数F得到第一时间间隔T：T＝F(Lavg,Qps)，其中函数F的值随L_{avg}单调递增、随Qps单调递增。

[0115] 数据读取流程包括如下步骤201～206：

[0116] 201、客户端接收到用户的对文件F的读请求；

[0117] 202、客户端向主控端请求文件F的数据所在的Chunk的信息；

[0118] 203、主控端将文件F的数据所在的Chunk的多个副本的块服务器(即文件F对应的块服务器)地址返回给客户端；

[0119] 204、客户端向文件F一个副本所在的块服务器发起读数据请求，并等待块服务器响应，此时用户后续的写被阻塞。

[0120] 205、客户端根据F(L_{avg},Qps)计算读冗余的第一时间间隔T，等待T之后如果步骤204发送的读数据请求还没有收到响应，客户端会再向文件F另一个副本所在的块服务器发起针对所述读数据请求的冗余请求；

[0121] 如果这是客户端启动后第一次针对收到的读请求发送针对所述读数据请求的冗余请求，则第一时间间隔可采用默认或用户设置的初始的第一时间间隔，或者根据上次启动时的记录进行计算。

[0122] 当然，计算第一时间间隔T的操作可以不在步骤205中完成，比如可以在步骤201～204的任一步骤中或任一个步骤后完成，还可以周期性完成。

[0123] 206、针对步骤204和步骤205中任何一个请求先收到表示读成功的响应，客户端就会返回给用户表示读成功的消息，记录本次读操作的耗时和冗余请求的个数。

[0124] 用户收到表示读成功的消息后，可以继续发起下一次读，返回到上述的步骤201。

[0125] 实施例二、一种分布式系统中的读取装置，如图3所示，包括：

[0126] 第一请求模块21，用于当收到用户对于第一文件的读请求后，向所述第一文件对应的存储节点之一发送读数据请求；

[0127] 第二请求模块22，用于当等待响应的时间长度达到或超过第一时间间隔时，向所述第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求；其中，所述第一时间间隔是根据读操作的性能指标而动态确定的。

[0128] 本实施例中，所述第一请求模块21是上述装置中负责发起读数据请求的部分，可

以是软件、硬件或两者的结合。

[0129] 本实施例中，所述第二请求模块22是上述装置中负责发起冗余请求的部分，可以是软件、硬件或两者的结合。

[0130] 本实施例的装置可以但不限于设置于客户端中，也可以采用分布式方式设置在不同设备中。

[0131] 本实施例的一种备选方案中，所述读操作的性能指标可以包括：

[0132] 读操作的延时，和/或，读操作中发送冗余请求的次数。

[0133] 本备选方案的一种实施方式中，所述读操作的延时可以是之前一次或之前多次读操作的延时，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作的延时；

[0134] 所述读操作中发送冗余请求的次数可以是之前一次或之前多次读操作中发送冗余请求的次数，或者是待确定所述第一时间间隔的时刻之前预定长度的时间内，读操作中发送冗余请求的次数。

[0135] 本备选方案的一种实施方式中，所述装置还可以包括：

[0136] 记录模块，用于在每次读操作后，记录本次读操作的延时和发送冗余请求的次数。

[0137] 本实施例的一种备选方案中，所述第一时间间隔根据读操作的性能指标而动态确定可以包括：

[0138] 所述第一时间间隔根据L_{avg}以及Qps计算得到；其中，L_{avg}是待确定所述第一时间间隔的时刻之前预定长度的时间内读操作的延时的平均值；Qps是待确定所述第一时间间隔的时刻之前预定长度的时间内冗余请求的频率。

[0139] 本备选方案中，所述第一时间间隔根据L_{avg}以及Qps计算得到可以包括：

[0140] 将预定函数的函数值作为所述第一时间间隔；所述预定函数的自变量为L_{avg}、Qps，所述预定函数的函数值随L_{avg}单调递增，随Qps单调递增。

[0141] 本备选方案的一种实施方式中，所述预定函数为：

[0142] $a1\times L_{avg}+a2\times Qps+a3$；

[0143] 或者，$b1\times L_{avg}\times Qps+b2$；

[0144] 其中，a1、a2、a3、b1、b2是预定值。

[0145] 实际应用中，也可以对上述函数式进行变换，比如改变计算符号等；还可以采用其它不同的函数作为所述预定函数。

[0146] 本实施例的一种备选方案中，所述第一时间间隔根据读操作的性能指标而动态确定可以包括：

[0147] 所述第一时间间隔周期性根据读操作的性能指标动态确定；

[0148] 或者，所述第一时间间隔当满足预定的触发条件时根据读操作的性能指标动态确定。

[0149] 本实施例的其它实现细节可参考实施例一。

[0150] 本领域普通技术人员可以理解上述方法中的全部或部分步骤可通过程序来指令相关硬件完成，所述程序可以存储于计算机可读存储介质中，如只读存储器、磁盘或光盘等。可选地，上述实施例的全部或部分步骤也可以使用一个或多个集成电路来实现。相应地，上述实施例中的各模块/单元可以采用硬件的形式实现，也可以采用软件功能模块的形

式实现。本申请不限制于任何特定形式的硬件和软件的结合。

[0151] 当然，本申请还可有其他多种实施例，在不背离本申请精神及其实质的情况下，熟悉本领域的技术人员当可根据本申请作出各种相应的改变和变形，但这些相应的改变和变形都应属于本申请的权利要求的保护范围。

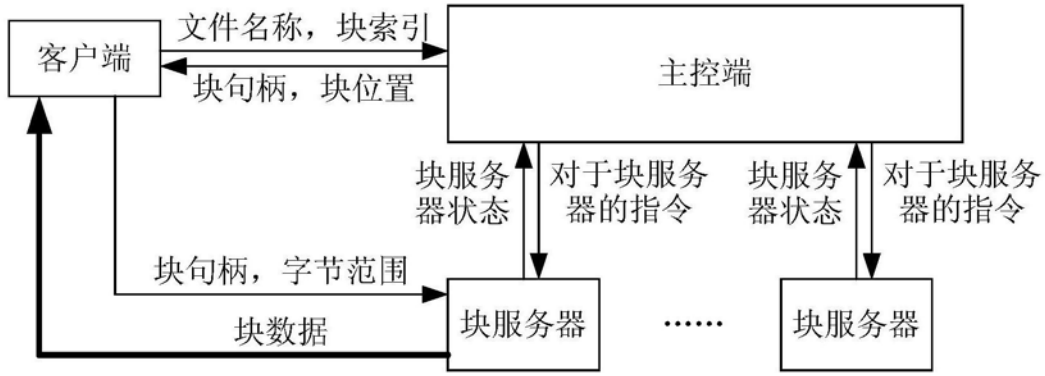

客户端
文件名称，块索引
块句柄，块位置
主控端
块服务器状态
对于块服务器的指令
块服务器状态
对于块服务器的指令
块句柄，字节范围
块数据
块服务器
……
块服务器

图1

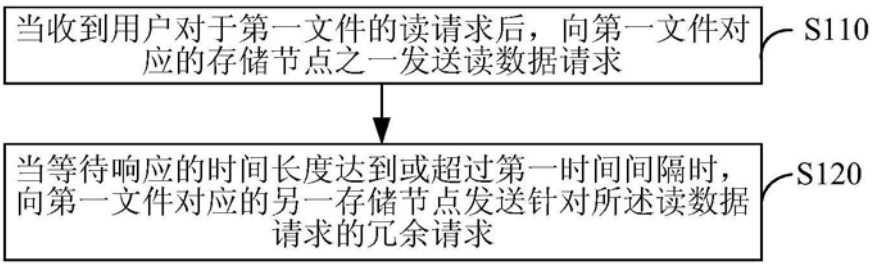

当收到用户对于第一文件的读请求后，向第一文件对应的存储节点之一发送读数据请求
S110
当等待响应的时间长度达到或超过第一时间间隔时，向第一文件对应的另一存储节点发送针对所述读数据请求的冗余请求
S120

图2

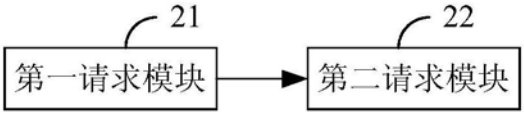

21
22
第一请求模块
第二请求模块

图3