



(12)发明专利

(10)授权公告号 CN 108780386 B

(45)授权公告日 2020.09.04

(21)申请号 201780011483.7

(22)申请日 2017.12.20

(65)同一申请的已公布的文献号
申请公布号 CN 108780386 A

(43)申请公布日 2018.11.09

(85)PCT国际申请进入国家阶段日
2018.08.15

(86)PCT国际申请的申请数据
PCT/CN2017/117532 2017.12.20

(87)PCT国际申请的公布数据
W02019/119311 ZH 2019.06.27

(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 刘小威 伍华涛 殷丽慧

(74)专利代理机构 广州三环专利商标代理有限公司 44202

代理人 熊永强 李稷芳

(51)Int.Cl.
G06F 3/06(2006.01)
G06F 11/10(2006.01)

(56)对比文件
CN 105095013 A,2015.11.25
CN 107391301 A,2017.11.24
CN 106406758 A,2017.02.15
CN 106383665 A,2017.02.08
US 2017206135 A1,2017.07.20
CN 104932953 A,2015.09.23
US 2015067421 A1,2015.03.05
US 2012297154 A1,2012.11.22

审查员 陈雪梅

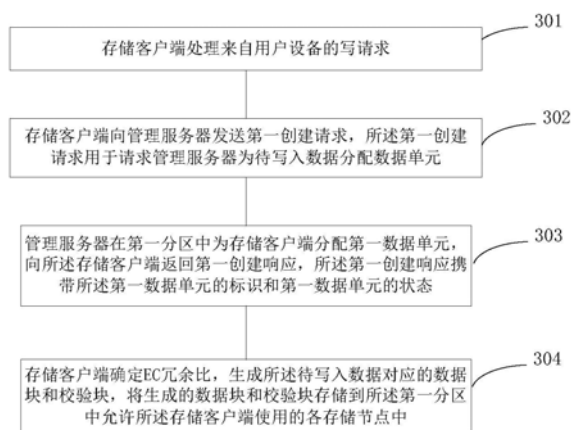
权利要求书7页 说明书17页 附图5页

(54)发明名称

一种数据存储的方法、装置和系统

(57)摘要

一种数据存储的方法、装置和系统,存储客户端需要将待写入数据存储到分布式存储系统中,管理服务器为待写入数据分配的第一数据单元所对应的存储节点仅为存储节点组中的部分节点,当存储客户端收到管理服务器返回的第一数据单元的状态时,即可确定对待写入数据执行EC编码时需要生成的数据块和校验块的数量,存储客户端将生成的数据块和校验块存储到第一数据单元所在的分区中管理服务器指定的部分存储节点中,从而实现了EC冗余比的动态调整,管理服务器可以根据需要将分区中的部分节点排除出待写入数据存储的范围,降低了数据存储的IO量。



1. 一种数据存储系统,其特征在于,所述数据存储系统包括存储客户端和分布式存储系统,所述分布式存储系统采用纠删码EC技术存储数据,所述分布式存储系统包括管理服务器以及多个存储节点, $N+M$ 个存储节点组成的存储节点组对应一个分区,其中, N 个存储节点用来存储数据块, M 个存储节点用于存储根据数据块生成的校验块,所述 N 和 M 为大于2的正整数,

所述存储客户端,用于为待写入数据向所述管理服务器请求分配数据单元;

所述管理服务器,用于为所述待写入数据在第一分区中分配第一数据单元,向所述存储客户端返回所述第一数据单元的标识和第一数据单元的状态,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于 $N+M$;

所述存储客户端,还用于获取所述第一数据单元的标识、所述第一数据单元的状态以及所述第一分区的第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表;

所述存储客户端,还用于根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

所述存储客户端,还用于根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

2. 如权利要求1所述的系统,其特征在于,

所述管理服务器,还用于记录所述第一数据单元的状态。

3. 如权利要求1所述的系统,其特征在于,

所述存储客户端,还用于执行读操作,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

所述存储客户端,还用于根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

所述存储客户端,还用于从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

4. 如权利要求1-3任一所述的系统,其特征在于,在所述存储客户端为待写入数据向所述管理服务器请求分配数据单元之前,

所述存储客户端,还用于根据所述管理服务器返回的第三数据单元的标识,向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块,当写入失败时,向所述管理服务器请求重新分配数据单元。

5. 如权利要求1-3任一所述的系统,其特征在于,

所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单

元时所述第一分区包含的各存储节点的状态;或者,

所述允许所述存储客户端使用的存储节点为所述管理服务器在所述第一分区中为所述存储客户端指定的存储所述待写入数据的存储节点。

6.如权利要求1-3任一所述的系统,其特征在于,

生成的校验块的数量等于M,生成的数据块的数量等于允许所述存储客户端使用的存储节点的数量减去M。

7.如权利要求1-3任一所述的系统,其特征在于,

所述管理服务器,还用于根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则在所述第一分区中为所述待写入数据分配所述第一数据单元。

8.如权利要求7所述的系统,其特征在于,

所述管理服务器,具体用于判断存在故障节点的分区数量与分区总数比值是否大于预设比例,如果是,则确定允许分区降级。

9.如权利要求7所述的系统,其特征在于,所述管理服务器,还用于在不允许分区降级时,在不存在故障节点的第四分区中为所述待写入数据分配第四数据单元,向所述存储客户端返回所述第四数据单元的标识。

10.一种存储客户端,其特征在于,所述存储客户端采用纠删码EC技术向分布式存储系统中存储数据,所述分布式存储系统包括管理服务器以及多个存储节点,N+M个存储节点组成的存储节点组对应一个分区,其中,N个存储节点用来存储数据块,M个存储节点用于存储根据数据块生成的校验块,所述N和M为大于2的正整数,包括:

第一发送单元,用于为待写入数据向所述管理服务器请求分配数据单元;

获取单元,用于获取所述管理服务器返回的第一数据单元的标识、所述第一数据单元的状态以及所述第一数据单元所在的第一分区的第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于N+M;

处理单元,用于根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

第一发送单元,还用于根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

11.如权利要求10所述的存储客户端,其特征在于,

所述获取单元,还用于在所述存储客户端执行读操作时,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

所述处理单元,还用于根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

所述获取单元,还用于从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

12.如权利要求10或11所述的存储客户端,其特征在于,在第一发送单元为待写入数据向所述管理服务器请求分配数据单元之前,

所述获取单元,还用于获取所述管理服务器返回的第三数据单元的标识;

所述第一发送单元,还用于向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块;

当写入失败时,所述第一发送单元向所述管理服务器请求重新分配数据单元。

13.一种分布式存储系统中的管理服务器,其特征在于,所述分布式存储系统采用纠删码EC技术存储数据,所述分布式存储系统还包括多个存储节点,N+M个存储节点组成的存储节点组对应一个分区,其中,N个存储节点用来存储数据块,M个存储节点用于存储根据数据块生成的校验块,所述N和M为大于2的正整数,所述管理服务器包括:

接收单元,用于接收存储客户端发送的为待写入数据分配数据单元的请求;

分配单元,用于为所述待写入数据在第一分区中分配第一数据单元;

第二发送单元,还用于向所述存储客户端返回所述第一数据单元的标识、第一数据单元的状态以及所述第一分区的第一分区视图,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,所述第一分区视图表示所述第一分区包含的存储节点列表,允许所述存储客户端使用的存储节点的数量小于N+M。

14.如权利要求13所述的管理服务器,其特征在于,还包括:

存储单元,用于记录所述第一数据单元的状态。

15.如权利要求14所述的管理服务器,其特征在于,所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单元时所述第一分区包含的各存储节点的状态。

16.如权利要求13-15任一所述的管理服务器,其特征在于,

所述第二发送单元,还用于在存储客户端执行读操作时,向所述存储客户端返回待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障。

17.如权利要求13-15任一所述的管理服务器,其特征在于,

在所述存储客户端为待写入数据向所述管理服务器请求分配数据单元之前,所述第二发送单元,还用于向所述存储客户端返回第三数据单元的标识;

所述接收单元,还用于在所述存储客户端向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块失败时,接收所述存储客户端发送的重新分配数据单元的请求。

18.如权利要求13-15任一所述的管理服务器,其特征在于,还包括:

降级判断单元,用于根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则在所述第一分区中为所述待写入数据分配所述第一数据单元。

19. 如权利要求18所述的管理服务器,其特征在于,

所述降级判断单元,具体用于判断存在故障节点的分区数量与分区总数比值是否大于预设比例,如果是,则确定允许分区降级。

20. 如权利要求18所述的管理服务器,其特征在于,

所述分配单元,还用于在所述降级判断单元确定不允许分区降级时,在不存在故障节点的第三分区中为所述待写入数据分配第三数据单元,向所述存储客户端返回所述第三数据单元的标识。

21. 一种分布式存储系统中数据存储的方法,其特征在于,所述分布式存储系统采用纠删码EC技术存储数据,所述分布式存储系统包括管理服务器以及多个存储节点, $N+M$ 个存储节点组成的存储节点组对应一个分区,其中, N 个存储节点用来存储数据块, M 个存储节点用于存储根据数据块生成的校验块,所述 N 和 M 为大于2的正整数,所述方法包括:

存储客户端为待写入数据向所述管理服务器请求分配数据单元;

所述管理服务器为所述待写入数据在第一分区中分配第一数据单元,向所述存储客户端返回所述第一数据单元的标识和第一数据单元的状态,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于 $N+M$;

所述存储客户端获取所述第一数据单元的标识、所述第一数据单元的状态以及所述第一分区的第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表;

所述存储客户端根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

所述存储客户端根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

22. 如权利要求21所述的方法,其特征在于,在所述管理服务器分配所述第一数据单元后,还包括:

所述管理服务器记录所述第一数据单元的状态。

23. 如权利要求21所述的方法,其特征在于,还包括:

所述存储客户端执行读操作,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

所述存储客户端根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

所述存储客户端从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

24. 如权利要求21-23任一所述的方法,其特征在于,在所述管理服务器分配所述第一数据单元之前,所述方法还包括:

所述管理服务器根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则

在所述第一分区中为所述待写入数据分配所述第一数据单元。

25. 一种分布式存储系统中数据存储的方法, 其特征在于, 所述分布式存储系统采用纠删码EC技术存储数据, 所述分布式存储系统包括管理服务器以及多个存储节点, $N+M$ 个存储节点组成的存储节点组对应一个分区, 其中, N 个存储节点用来存储数据块, M 个存储节点用于存储根据数据块生成的校验块, 所述 N 和 M 为大于2的正整数, 所述方法包括:

存储客户端为待写入数据向所述管理服务器请求分配数据单元;

所述存储客户端获取所述管理服务器返回的第一数据单元的标识、所述第一数据单元的状态以及所述第一数据单元所在的第一分区的第一分区视图, 其中, 所述第一分区视图表示所述第一分区包含的存储节点列表, 所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点, 允许所述存储客户端使用的存储节点的数量小于 $N+M$;

所述存储客户端根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量, 对所述待写入数据执行EC编码, 生成所述待写入数据对应的数据块和校验块, 生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

所述存储客户端根据所述第一分区视图和所述第一数据单元的状态, 将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中, 每个节点存储一个数据块或校验块。

26. 如权利要求25所述的方法, 其特征在于, 还包括:

所述存储客户端在执行读操作时, 获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态, 所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点, 所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

所述存储客户端根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态, 确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

所述存储客户端从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

27. 如权利要求25或26所述的方法, 其特征在于, 在所述存储客户端为待写入数据向所述管理服务器请求分配数据单元之前, 还包括:

所述存储客户端获取所述管理服务器返回的第三数据单元的标识;

所述存储客户端向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块, 当写入失败时, 向所述管理服务器请求重新分配数据单元。

28. 一种分布式存储系统中数据存储的方法, 其特征在于, 所述分布式存储系统采用纠删码EC技术存储数据, 所述分布式存储系统包括管理服务器以及多个存储节点, $N+M$ 个存储节点组成的存储节点组对应一个分区, 其中, N 个存储节点用来存储数据块, M 个存储节点用于存储根据数据块生成的校验块, 所述 N 和 M 为大于2的正整数, 所述方法包括:

所述管理服务器接收存储客户端发送的为待写入数据分配数据单元的请求;

所述管理服务器为所述待写入数据在第一分区中分配第一数据单元;

所述管理服务器向所述存储客户端返回所述第一数据单元的标识、第一数据单元的状态以及所述第一分区的第一分区视图, 其中, 所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点, 所述第一分区视图表示所述第一分区包含的存储

节点列表,允许所述存储客户端使用的存储节点的数量小于 $N+M$ 。

29.如权利要求28所述的方法,其特征在于,在所述管理服务器分配所述第一数据单元后,所述方法还包括:

所述管理服务器记录所述第一数据单元的状态。

30.如权利要求29所述的方法,其特征在于,

所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单元时所述第一分区包含的各存储节点的状态。

31.如权利要求28-30任一所述的方法,其特征在于,还包括:

所述管理服务器在存储客户端执行读操作时,向所述存储客户端返回待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障。

32.如权利要求28-30任一所述的方法,其特征在于,在所述管理服务器分配所述第一数据单元之前,所述方法还包括:

所述管理服务器根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则在所述第一分区中为所述待写入数据分配所述第一数据单元。

33.如权利要求28-30任一所述的方法,其特征在于,还包括:

所述存储客户端根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

34.一种数据存储系统,其特征在于,所述数据存储系统包括存储客户端和分布式存储系统,所述分布式存储系统采用纠删码EC技术存储数据,所述分布式存储系统包括管理服务器以及多个存储节点, $N+M$ 个存储节点组成的存储节点组对应一个分区,其中, N 个存储节点用来存储数据块, M 个存储节点用于存储根据数据块生成的校验块,所述 N 和 M 为大于2的正整数,

所述存储客户端,用于为待写入数据向所述管理服务器请求分配数据单元;

所述管理服务器,用于在第一分区中确定允许所述存储客户端使用的存储节点,为所述待写入数据在第一分区中分配第一数据单元,向所述存储客户端返回所述第一数据单元的标识、第一数据单元的状态以及需要生成的数据块的数量和校验块的数量,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于 $N+M$,需要生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

所述存储客户端,还用于根据所述需要生成的数据块的数量和校验块的数量对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块;

所述存储客户端,还用于根据第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点

存储一个数据块或校验块,其中,所述第一分区视图表示所述第一分区包含的存储节点列表。

35.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机程序,所述计算机程序被计算机设备执行时,能够实现权利要求21至27中由存储客户端执行的方法。

36.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机程序,所述计算机程序被计算机设备执行时,能够实现权利要求28至33中由管理服务器执行的方法。

一种数据存储的方法、装置和系统

技术领域

[0001] 本发明涉及IT技术领域,尤其涉及一种数据存储的方法、装置和系统。

背景技术

[0002] 现阶段的分布式存储系统越来越多的采用纠删码(Erasure Code,EC)技术对数据进行存储。纠删码技术原理是将数据分割成N个数据块,并且采用冗余算法对N个数据块进行编码,生成M个校验块,该N个数据块与M个校验块称为一个EC条带。其中,数据块或校验块也称为EC块,当丢失的EC块不大于M个时都可以恢复出原始数据,EC条带的冗余比 $r=M/N$ 或 $r=M/(N+M)$ 。示例性的,当N等于4,M等于2时,EC条带可以表示为EC 4+2。

[0003] 为了保证存储的可靠性,分布式存储系统会将每个EC条带包含的N+M个EC块分配给N+M个存储节点进行存储,N+M个存储节点构成一个存储节点组。这样,即使某个节点出现了故障,也能够根据存储节点组的其他节点中存储的数据将故障节点中存储的EC块恢复出来。可以理解的,采用EC技术的分布式存储系统能够容忍M个存储节点同时出现故障。

[0004] 当存储节点组中的某个存储节点发生故障时,一个新的存储节点会被加入到存储节点组中,替换发生故障的存储节点,在新的存储节点恢复故障节点的数据。当故障节点恢复正常后,再将新的存储节点的数据写回到恢复的节点中。由于存储节点存储的数据量一般比较大,故障节点中的数据的迁移和回迁将会对存储系统的性能造成较大的影响。

发明内容

[0005] 本文描述了一种数据存储的方法、装置和系统,以避免故障节点中数据的迁移和回迁对分布式存储系统的性能造成的影响。

[0006] 第一方面,本发明实施例提供了一种数据存储系统,所述数据存储系统包括存储客户端和分布式存储系统,所述分布式存储系统采用纠删码EC技术存储数据,所述分布式存储系统包括管理服务器以及多个存储节点,N+M个存储节点组成的存储节点组对应一个分区,其中,N个存储节点用来存储数据块,M个存储节点用于存储根据数据块生成的校验块,所述N和M为大于2的正整数,

[0007] 所述存储客户端处理来自用户设备的写请求,为待写入数据向所述管理服务器请求分配数据单元,所述管理服务器为所述待写入数据在第一分区中分配第一数据单元,向所述存储客户端返回所述第一数据单元的标识和第一数据单元的状态,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于N+M,所述存储客户端根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量,所述存储客户端根据第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

[0008] 存储客户端需要将待写入数据存储到分布式存储系统中,管理服务器为待写入数据分配的第一数据单元所对应的存储节点仅为存储节点组中的部分节点,当存储客户端收到管理服务器返回的第一数据单元的状态时,即可确定对待写入数据执行EC编码时需要生成的数据块和校验块的数量,存储客户端将生成的数据块和校验块存储到第一数据单元所在的分区中管理服务器指定的部分存储节点中,从而实现了EC冗余比的动态调整,减少了EC块的生成数量,管理服务器可以根据需要将分区中的部分节点排除出待写入数据存储的范围,从而降低IO数据量。更进一步的,当分区中存在故障节点时,管理服务器可以主动将故障节点排除,避免存储客户端将EC块写入到故障节点引起的写入失败。

[0009] 存储客户端获取所述第一数据单元的标识,根据所述第一数据单元的标识确定所述第一数据单元所在的分区为所述第一分区,在确定所述第一数据单元所在的分区为第一分区后,查询分区表获得所述第一分区的第一分区视图。

[0010] 在一种可能的实施方式中,所述允许所述存储客户端使用的存储节点为所述管理服务器在所述第一分区中为所述存储客户端指定的存储所述待写入数据的存储节点;或者,

[0011] 在另一种可能的实施方式中,所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单元时所述第一分区包含的各存储节点的状态。

[0012] 进一步的,当所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点时,管理客户端需要生成的数据块的数量等于N减去第一分区中故障节点的数量。

[0013] 管理服务器可以监控分布式存储系统中的各个存储节点的状态,在第一分区中存在故障节点时,可以将故障节点排除出待写入数据存储的范围,降低生成数据块和校验块的数量,避免向第一分区中添加新的存储节点以替代故障节点,也无需将故障节点中的数据进行迁移和回迁。

[0014] 存储客户端根据第一数据单元的状态和第一分区视图,确定第一分区中非故障存储节点列表。

[0015] 存储客户端根据第一数据单元的状态,确定第一数据单元所在的第一分区中非故障存储节点的数量。

[0016] 在一种可能的实施方式中,为了保证数据存储的持久度,需要保证校验块的数量不变,即使分区中存在故障节点,也只是会降低生成数据块的数量,而不会降低生成校验块的数量,此时,生成的数据块的数量等于N减去第一分区中故障节点的数量,生成的校验块的数量等于M。由于校验块的数量M保存不变,存储客户端可以使用第一分区中非故障节点的数量减去M,即可获得需要生成的数据块的数量。校验块M的数量可以由用户预先在数据存储系统中设置,也可以由存储客户端和管理服务器协商确定。

[0017] 在另一种可能的实施方式中,当允许降低数据存储持久度时,可以降低校验块的数量,此时,管理服务器可以在返回给存储客户端的响应消息中携带需要生成的数据块的数量和/或校验块的数量,当响应消息中只携带了需要生成的数据块的数量时,存储客户端使用第一分区中非故障节点的数量减去需要生成的数据块的数量,得到需要生成的校验块的数量;当响应消息中只携带了需要生成的校验块的数量时,存储客户端使用第一分区中非故障节点的数量减去需要生成的校验块的数量,得到需要生成的数据块的数量。

[0018] 所述管理服务器,还用于记录所述第一数据单元的状态。

[0019] 在所述存储客户端执行读操作时,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

[0020] 所述存储客户端根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

[0021] 所述存储客户端从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

[0022] 上述存储节点当前是否故障表示存储节点是故障节点还是非故障的正常节点。第二数据单元为写入该待读取数据到分布式存储系统时,管理服务器为所述待读取数据分配的数据单元,第二数据单元的状态指示在第二分区中存储有所述待读取数据对应的EC块的存储节点。所述存储客户端处理用户设备发起的读请求,所述读请求中携带数据单元的标识,存储客户端根据所述数据单元的标识确定所述待读取数据所在的第二数据单元的状态和第二分区视图,从而找到存储有待读取数据对应的EC块的非故障存储节点。上述数据读取过程与前述的数据写入过程对应,实现了分区中存在故障节点时,采用分区降级的情况下读取数据的过程,无需向第一分区中添加新的存储节点以替代故障节点,也无需将故障节点中的数据进行迁移和回迁。

[0023] 所述管理服务器,还用于根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则在所述第一分区中为所述待写入数据分配所述第一数据单元。

[0024] 在本发明实施例中,分区降级是指,管理服务器分配给存储客户端存储待写入数据使用的存储节点的数量少于分区包含的节点数量(N+M)。

[0025] 具体的,所述管理服务器判断存在故障节点的分区数量与分区总数比值是否大于预设比例,如果是,则确定允许分区降级。其中,分区总数为DHT环包含的分区数量。

[0026] 在不允许分区降级时,所述管理服务器在不存在故障节点的第四分区中为所述待写入数据分配第四数据单元,向所述存储客户端返回所述第四数据单元的标识。当不存在故障节点的分区数量较多时,无需进行分区降级,此时,管理服务器可以执行分区重定向操作,向存储客户端返回在无故障节点的分区中分配的数据单元的标识,从而充分利用已有的存储资源。

[0027] 在一种可能的实现方式中,分区降级时,也可以通过保持校验块的数量与数据块的数量比值不变小。

[0028] 管理服务器确定需要分区降级的触发条件可以是:存储客户端在向管理服务器分配的数据单元所在的分区包含的存储节点写入待写入数据对应的EC条带时,一个或多个存储节点故障导致写入失败。相对应的,在所述存储客户端为待写入数据向所述管理服务器请求分配数据单元之前,所述存储客户端根据所述管理服务器返回的第三数据单元的标识,向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块,当写入失败时,向所述管理服务器请求重新分配数据单元。

[0029] 管理服务器可以向所述存储客户端返回数据单元的状态和分区视图;或者,管理服务器在数据库中记录数据单元的状态和分区视图,由所述存储客户端根据数据单元的标识从数据库中读取数据单元的状态和分区视图。

[0030] 在本发明实施例中,EC冗余比表示执行EC编码时需要生成的校验块的数量比需要生成的数据块的数量。通过EC冗余比即可确定EC编码时N和M的值。

[0031] 本发明实施例中的各分区属于同一DHT环。

[0032] 第二方面,本发明实施例提供了一种存储客户端,所述存储客户端为第一方面中所述的存储客户端,执行第一方面定义的存储客户端的功能。

[0033] 具体的,所述存储客户端包括第一发送单元、获取单元、以及处理单元,

[0034] 第一发送单元,用于为待写入数据向所述管理服务器请求分配数据单元;

[0035] 获取单元,用于获取所述管理服务器返回的第一数据单元的标识、所述第一数据单元的状态以及所述第一数据单元所在的第一分区的第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于N+M;

[0036] 处理单元,用于根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

[0037] 第一发送单元,还用于根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

[0038] 所述获取单元,还用于在所述存储客户端执行读操作时,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

[0039] 所述处理单元,还用于根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

[0040] 所述获取单元,还用于从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

[0041] 在第一发送单元为待写入数据向所述管理服务器请求分配数据单元之前,

[0042] 所述获取单元,还用于获取所述管理服务器返回的第三数据单元的标识;

[0043] 所述第一发送单元,还用于向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块;

[0044] 当写入失败时,所述第一发送单元向所述管理服务器请求重新分配数据单元。

[0045] 第三方面,本发明实施例提供了一种管理服务器,所述管理服务器为第一方面中所述的管理服务器,执行第一方面定义的管理服务器的功能。

[0046] 所述管理服务器包括:

[0047] 接收单元,用于接收存储客户端发送的为待写入数据分配数据单元的请求;

[0048] 分配单元,用于为所述待写入数据在第一分区中分配第一数据单元;

[0049] 所述第二发送单元,还用于向所述存储客户端返回所述第一数据单元的标识、第

一数据单元的状态以及所述第一分区的所述第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表,所述第一数据单元的状态指示所述存储客户端确定所述第一分区中允许所述存储客户端使用的存储节点,确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,允许所述存储客户端使用的存储节点的数量小于 $N+M$ 。

[0050] 第四方面,本发明实施例还提供了与前述数据存储系统、存储客户端和管理服务器对应的方法,所述方法描述了前述各方面中存储客户端和管理服务器实现各自的功能所执行的步骤。在此不再赘述。

[0051] 第五方面,与前述方法、装置以及系统相对应,本发明实施例提供了一种物理服务器,该物理服务器具有实现第一方面中定义的存储客户端或管理服务器的功能。所述功能可以通过硬件实现,也可以通过硬件执行相应的软件实现。所述硬件或软件包括前述装置中定义的与上述功能相对应的模块。具体的,物理服务器的结构包括处理器、存储器、系统总线以及输入输出接口,所述处理器被配置为支持所述系统中存储客户端或管理服务器的相应的功能。所述输入输出接口用于与所述数据存储系统中的其他部件进行通信,处理器执行所述存储器中存储的指令。

[0052] 第六方面,本发明实施例提供了一种计算机存储介质,用于储存为上述存储客户端所用的计算机软件指令,其包含用于执行上述方面所设计的程序。

[0053] 第七方面,本发明实施例提供了另一种计算机存储介质,用于储存为上述物理服务器所用的计算机软件指令,其包含用于执行上述方面所设计的程序。

附图说明

[0054] 如图1所示,为本发明实施例提供的一种数据存储系统的架构示意图;

[0055] 如图2所示,为现有技术中基于DHT的数据分布路由的结构示意图;

[0056] 如图3所示,为本发明实施例提供的使用动态调整EC冗余比进行数据存储的方法流程示意图;

[0057] 如图4所示,为本发明实施例提供的分区视图的示意图;

[0058] 如图5所示,为本发明实施例提供的一种数据存储方法流程示意图;

[0059] 如图6所示,为为本发明实施例提供的另一种数据存储的方法流程示意图;

[0060] 如图7所示,为本发明实施例提供的一种数据读取的流程示意图;

[0061] 如图8所示,为本发明实施例提供的存储客户端的结构示意图;

[0062] 如图9所示,为本发明实施例提供的管理服务器的结构示意图;

[0063] 如图10所示,为本发明实施例提供的物理服务器的结构示意图。

具体实施方式

[0064] 如图1所示,为本发明实施例提供的一种数据存储系统的架构示意图,所述数据存储系统包括存储客户端100以及分布式存储系统200,所述分布式存储系统200包括管理服务器201以及多个存储节点202。用户通过持久化层中的存储客户端100将数据写入到分布式存储系统200的分区包含的存储节点中。所述存储客户端100处理来自用户设备的读写请

求,向分布式存储系统200读写数据。存储客户端使用EC技术将待存储的数据生成多个EC块(block),然后下发给分布式存储系统的一个分区进行存储。数据分布式存储能够降低数据丢失的风险,提高分布式存储系统的可靠性。

[0065] 数据存储系统可以采用EC技术存储数据。EC技术原理是将待存储的数据分割成N个数据块,并且采用冗余算法对N个数据块进行编码,生成M个校验块,该N个数据块与M个校验块组成一条EC条带,相对应的分布式存储系统中的一个分区可以包含N+M个存储节点。其中,每个数据块或校验块也可以称为一个EC块,该N个数据块和M个校验块组成每个EC条带的所有EC块。每个EC条带能容忍M个EC块的丢失。分布式存储系统使用EC技术进行数据存储时,存储客户端会根据待存储的数据生成一个或多个EC条带,并将每个EC条带的N+M个EC块下发给分布式存储系统的一个分区中的N+M个存储节点进行存储。这样,当分布式存储系统中有节点故障时,可以根据非故障节点上的EC块将故障节点上存储的EC块恢复出来。除非存储某条EC条带的N+M个存储节点中有多于M个存储节点发生故障,否则该条EC条带的数据可以被数据下发装置成功读出。在本发明各实施例中,每个存储节点存储EC条带的一个EC块。冗余比用于表示EC条带中的校验块的数量和数据块的数量对比,具体可以采用各种表达方式,例如, M/N 、 $M/(N+M)$ 或者 $N+M$ 等等。本发明后续实施例以 $N+M$ 的形式为例。

[0066] 分布式存储系统中包含若干存储节点,在具体的实施场景中, $N+M$ 个存储节点组成一个存储节点组,不同的存储节点组可以包含相同的存储节点。更进一步的,分布式存储系统可以分为多个存储资源池,存储资源池相互独立,不同存储资源池中的存储节点组不包含相同的存储节点。

[0067] 如图2所示,为现有技术中基于分布式哈希表(Distributed Hash Table,DHT)的数据分布路由的结构示意图。一个DHT环对应一个存储资源池,DHT环由多个分区组成,每个分区中包含存储资源池中的 $N+M$ 个存储节点。 $plog$ 为数据单元,每个数据单元对应于一个分区,每个数据单元映射到分区中包含的存储节点上的一段存储空间。当存储客户端100存储待写入数据时,管理服务器为待写入数据分配数据单元($plog$),将数据单元标识发送给存储客户端,存储客户端计算数据单元标识的哈希值,根据计算出的哈希值确定数据单元对应的分区,查询分区视图确定分区对应的存储节点列表,对待写入数据进行EC编码,将生成的EC块存储到存储节点列表包含的存储节点中。所述存储节点列表又称为DHT分区表。

[0068] 如表1所示,示出了DHT分区表的一个举例

[0069]

Partition ID	Node List
P0	Node1, Node2, Node3, Node4, Node5, Node6
P1	Node2, Node4, Node5, Node7, Node1, Node8
P2	Node3, Node5, Node6, Node8, Node9, Node10
P3	Node4, Node6, Node7, Node9, Node10, Node1
...	...
Pn	Node10, Node1, Node4, Node9, Node6, Node1

[0070] 表1

[0071] DHT分区表中的每个分区包含 $N+M$ 个存储节点,存储到各分区中的EC条带的冗余比 $r=N+M$ 。与现有技术中每个DHT环只具备固定EC冗余比不同的是,本发明实施例提供了一种使用动态调整EC冗余比进行数据存储的方法,可以动态调整存储客户端将待写入数据存储到DHT环中各分区包含的存储节点时使用的EC冗余比。

[0072] 如图3所示,使用动态调整EC冗余比进行数据存储的方法包括:

[0073] 步骤301:存储客户端处理来自用户设备的写请求,对待写入数据执行写操作。

[0074] 步骤302:存储客户端向管理服务器发送第一创建请求,所述第一创建请求用于请求管理服务器为待写入数据分配数据单元。

[0075] 步骤303:管理服务器在第一分区中为存储客户端分配第一数据单元,向所述存储客户端返回第一创建响应,所述第一创建响应携带所述第一数据单元的标识和第一数据单元的状态,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于 $N+M$ 。

[0076] 管理服务器可以依据第一分区中节点的状态或者待写入数据的大小等因素,确定第一分区中允许所述存储客户端使用的存储节点列表。具体的:

[0077] 在一种可能的实施方式中,所述允许所述存储客户端使用的存储节点为所述管理服务器在所述第一分区中为所述存储客户端指定的存储所述待写入数据的存储节点。即管理服务器在待写入数据较小时,仅选择分区中的部分存储节点为存储客户端提供存储服务。或者,

[0078] 在另一种可能的实施方式中,所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单元时所述第一分区包含的各存储节点的状态。即,为避免写入失败,管理服务器将第一分区中的故障节点排除出待写入数据存储的范围。

[0079] 上述两种实施方式均会导致存储客户端执行EC编码时采用的EC冗余比发生变化,生成的数据块的数量和校验块的数量之和不再等于 $N+M$,而是等于允许所述存储客户端使用的存储节点的数量。

[0080] 管理服务器在确定分区降级后,通过第一数据单元的状态将允许所述存储客户端使用的存储节点通知给存储客户端。

[0081] 分区降级是指,管理服务器分配给存储客户端存储待写入数据使用的存储节点的数量少于分区包含的存储节点数量($N+M$)。

[0082] 步骤304:存储客户端确定EC冗余比,生成所述待写入数据对应的数据块和校验块,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中。

[0083] 具体的,存储客户端获取所述第一数据单元的标识、所述第一数据单元的状态以及所述第一分区的第一分区视图,根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

[0084] 存储客户端需要将待写入数据存储到分布式存储系统中,管理服务器为待写入数

据分配的第一数据单元所对应的存储节点仅为存储节点组中的部分节点,当存储客户端收到管理服务器返回的第一数据单元的状态时,即可确定对待写入数据执行EC编码时需要生成的数据块和校验块的数量,存储客户端将生成的数据块和校验块存储到第一数据单元所在的分区中管理服务器指定的部分存储节点中,从而实现了EC冗余比的动态调整,管理服务器可以根据需要将分区中的部分节点排除出待写入数据存储的范围。

[0085] 需要说明的是,需要生成的数据块的数量和校验块的数量也可以由管理服务器确定后,通知给存储客户端。具体的,所述管理服务器在第一分区中确定允许所述存储客户端使用的存储节点,为所述待写入数据在第一分区中分配第一数据单元,向所述存储客户端返回所述第一数据单元的标识、第一数据单元的状态以及需要生成的数据块的数量和校验块的数量,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小于 $N+M$,需要生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量。

[0086] 基于上面所述的数据存储系统,本发明实施例提出了一种数据存储方法,用于在存储故障存储节点情况下,向分布式存储系统读写数据。

[0087] 在一种可能的实施方式中,当某个分区的存储节点发生故障时,通过分区重定向,将待写入的数据重定向到其他的可用分区。

[0088] 示例性的,如表2所示,为Node7故障时DHT分区表的举例,此时,分区P1和P3故障。

[0089]	Partition ID	Node List
	P0	Node1, Node2, Node3, Node4, Node5, Node6
	P1 (fault)	Node2, Node4, Node5, Node7 (fault), Node1, Node 8
	P2	Node3, Node5, Node6, Node8, Node9, Node10
	P3 (fault)	Node4, Node6, Node7 (fault), Node9, Node10, Node11
	...	...
	Pn	Node10, Node1, Node4, Node9, Node6, Node11

[0090] 表2

[0091] 当存储客户端向P1包含的存储节点写入数据时,P1中node7发生故障,此时,分配给node7存储的EC块写入失败,从而导致整个写入过程失败。存储客户端可以向管理服务器请求重新分配数据单元,由管理服务器在不存在故障节点的可用分区中为待写入数据重新分配数据单元,从而使得存储客户端将数据写入到可用分区包含的存储节点中。

[0092] 管理服务器记录分区元数据,所述分区元数据包括分区视图和分区状态,分区视图中记录有分区所包含的存储节点列表,分区状态记录存储节点列表中各存储节点的状态。

[0093] 在另一种可能的实施方式中,除了分区重定向外,本发明实施例还提供了一种通过动态调整EC冗余比在分布式存储系统中存储数据的方法。如图4所示,为本发明实施例提供的分区视图的示意图,以 $N=4, M=2$ 为例进行说明,此时,根据待写入数据生成的EC条带需要包含4个数据块和2个校验块。当分区中的6个存储节点均正常时,4个存储节点(n1-n4)

存储数据块,2个存储节点(p和q)存储校验块。当其中一个存储节点(例如,n3)故障时,存储数据块的存储节点数量调整为3个(n1、n2和n4),存储校验块的存储节点保持不变,此时,EC冗余比N+M为3+2,存储客户端根据新的EC冗余比生成3个数据块和2个校验块,并将生成的5个EC块存储到状态正常的5个存储节点。

[0094] 具体的,每个分区的分区元数据包括分区视图以及分区状态,其中,分区视图用来标识分区中包含的存储节点列表,分区状态用于标识分区包含的各个存储节点的状态。示例性的,分区P的分区视图为 $P = \{n1, n2, n3, n4, p, q\}$,当分区中的所有存储节点均正常时,分区P的分区状态为{111111},其中,取值1表示对应的存储节点状态为正常,取值0表示对应的存储节点状态为故障。当存储节点n3故障时,存储数据块的存储节点数量调整为3个,此时,分区状态为{110111},第3个属性的取值为0,表示存储节点n3故障。

[0095] 当故障节点数量比较多时,可能导致大部分的分区中都包含故障节点,此时,不存在故障节点的少量分区会承担所有的数据写入,导致数据存储不均衡,且在不存在故障节点的少量分区无空闲空间后,后续的数据写入就会由于无存储空间而导致写入失败。因此,上述动态EC调整方案可以针对故障节点较多的场景,将存储故障节点的分区充分利用,此时,管理服务器在返回给存储客户端的响应消息中携带分配的数据单元的状态(即创建数据单元时该数据单元所在的分区包含的各存储节点的状态),使得存储客户端可以根据数据单元的状态获得数据单元所在的分区的冗余比(即获取EC编码需要生成的数据块的数量和校验块的数量对比),存储客户端对待写入数据执行EC编码,生成与分区中非故障节点数量匹配的数据块和校验块,将生成的数据块和校验块存储到分区的非故障节点中。通过上述分区降级的方式,使得存在故障节点的分区可以被充分利用,提高了系统存储的性能。

[0096] 在一种可能的实施方式中,为了保证数据存储的持久度,需要保证校验块的数量不变,即使分区中存在故障节点,也只是会降低生成数据块的数量,而不会降低生成校验块的数量,此时,生成的数据块的数量等于N减去第二分区中故障节点的数量,生成的校验块的数量等于M。由于校验块的数量M保存不变,存储客户端可以使用第二分区中非故障节点的数量减去M,即可获得需要生成的数据块的数量。校验块M的数量可以由用户预先在数据存储系统中设置,也可以由存储客户端和管理服务器协商确定。

[0097] 通过上述方式,在存储节点组中出现故障节点时,对故障节点影响的分区采用动态EC,减少EC条带中数据块的数量,提高分区冗余比。例如EC冗余比N+M为6+3时,当分区中分别有1、2或3个节点故障时,分区冗余比可以调整为5+3、4+3或者3+3。

[0098] 需要说明的是,当存储客户端向分布式存储系统写入数据时,管理服务器为待写入数据分配数据单元,向存储客户端返回数据单元标识以及数据单元状态。数据单元状态中记录的是管理服务器创建数据单元时,数据单元所在的分区中各存储节点的状态。数据单元创建后,数据单元状态则固定,且不会随着存储节点状态的变化而变化。以EC冗余比为4+2进行举例说明,在管理服务器在第一分区中为待写入数据创建数据单元时,假设第一分区中故障节点为第3个节点,则此时分区状态为{110111},创建的数据单元的状态为{110111};在管理服务器在第一分区中为另一待写入数据创建另一数据单元时,假设第一分区中第3个节点故障恢复,第4个节点故障,此时,第一分区状态为{111011},所述另一数据单元的状态为{111011}。因此可以看出,同一分区的分区状态可以随着分区包含的节点

的状态变化而变化,但数据单元的状态是创建数据单元时分区的状态,且不会发生变化。

[0099] 另一方面,如果存储校验块的存储节点故障,则可以使用存储数据块的存储节点替代存储校验块的存储节点,从而使得存储数据块的存储节点数量减少,保持存储校验块的存储节点数量不变。以EC冗余比为4+2进行举例说明,如果故障节点为存储校验块的节点p,此时可以维持存储校验块的存储节点的数量不变,存储数据块的节点数量减小1,此时分区P的分区视图更新为{1,2,3,p,0,q},对应的分区状态为

[0100] {111101}。

[0101] 在一种可能的实施方式中,对于存在故障节点的分区,当允许降低数据存储持久度时,可以降低校验块的数量,此时,管理服务器可以在返回给存储客户端的响应消息中携带需要生成的数据块的数量和/或校验块的数量,当响应消息中只携带了需要生成的数据块的数量时,存储客户端使用第二分区中非故障节点的数量减去需要生成的数据块的数量,得到需要生成的校验块的数量;当响应消息中只携带了需要生成的校验块的数量时,存储客户端使用第二分区中非故障节点的数量减去需要生成的校验块的数量,得到需要生成的数据块的数量。

[0102] 对于分区中不存在故障节点的情况下,存储客户端读写数据,可以采用现有技术的方式,通过计算数据单元的标识的哈希值,确定数据单元所在的分区,向分区中包含的存储节点读写数据。当分区中存在故障节点时,在确定分区后,根据数据单元状态,向分区中状态为正常的存储节点读写数据。从而既可以维持分区计算的高性能,又可以在存在故障的情况下基于数据单元的状态进行灵活的数据读写,提高了分布式存储系统的可用性。

[0103] 分区重定向和动态调整EC冗余比可以独立使用,也可以组合使用。即,可以优先使用分区重定向,当多个存储节点故障,管理服务器根据可用分区数量判断允许分区降级时,再动态调整EC冗余比。当存在故障节点的分区数量与分区总数比值大于预设比例时,则确定允许分区降级。上述预设比例可以由管理人员灵活配置,例如,当存在故障节点的分区超过分区总数的90%时,则允许分区降级。

[0104] 下面结合具体方法流程对本发明实施方式进行详细介绍。

[0105] 结合图1所示的系统结构示意图,如图5所示,为本发明实施例提供的一种数据存储方法流程示意图,基于分区重定向技术,所述方法包括:

[0106] 步骤501:存储客户端处理来自用户设备的写请求,对待写入数据执行写操作。

[0107] 步骤502:存储客户端向管理服务器发起第一创建请求,所述第一创建请求用于请求管理服务器为待写入数据分配数据单元。

[0108] 步骤503:管理服务器在可用分区中为存储客户端分配第一数据单元。

[0109] 管理服务器记录有分布式存储系统中所有存储节点的状态,从而可以根据存储节点的状态确定每个分区的状态。管理服务器根据分区的状态、分区之间的数据均衡性以及存储节点的可用空间,在合适的分区中为所述待写入数据分配数据单元。具体分配方法本发明实施例并不进行限定。

[0110] 示例性的,当DHT环包含10个分区时,管理服务器选择在第x个分区分配数据单元,此时,分配的数据单元标识为 $10*n+x$,n为正整数。即当管理服务器选择在第2个分区分配数据单元时,分配的数据单元标识为12、22、32等。在存储客户端获得数据单元标识后,即可采用对10取模的方式,得到余数为2,则可以确定数据单元所在的分区为第2分区。

[0111] 步骤504:管理服务器向存储客户端返回第一创建响应,所述第一创建响应携带第一数据单元标识(plog ID)。

[0112] 步骤505:存储客户端接收所述第一创建响应,根据所述第一数据单元标识确定存储所述待写入数据的第一分区,存储客户端获取第一数据单元的状态以及第一分区视图,示例性的,以上述表1中的P1为例进行说明,此时,分区P1的分区视图为{Node2,Node4,Node5,Node7,Node1,Node8},分区状态为{111111},则第一数据单元状态为{111111}。

[0113] 示例性的,存储客户端可以计算第一数据单元标识的哈希值,获取哈希值对应的第一分区,或者,存储客户端对第一数据单元标识按分区数量取模,根据得到的余数确定第一数据单元所在的第一分区。

[0114] 具体的,存储客户端可以主动从管理服务器或数据库中获取第一数据单元状态以及第一分区视图,也可以被动接收管理服务器在第一创建响应中携带的第一数据单元状态和第一分区视图。

[0115] 所述管理服务器在创建第一数据单元,记录第一数据单元标识后,记录第一数据单元的状态。

[0116] 步骤506:存储客户端对待写入数据执行EC编码,生成待写入数据对应的N个数据块和M个校验块。

[0117] 在一种可能的实施方式中,N和M的值可以为约定数值。在另一种可能的方式中,管理服务器可以在第一响应消息中携带M的值,存储客户端即可以使用第一数据单元状态中包含的1的个数减去M的值,获得的差值即为N的值。确定N和M的值后,存储客户端则可以对待写入数据进行EC编码获得N个数据块以及M个校验块。

[0118] 步骤507:存储客户端将生成的数据块和校验块发送给第一分区包含的各个存储节点。

[0119] 存储客户端可以将所有的数据块和校验块发送给第一分区对应的存储节点组中的主节点,由主节点向每个存储节点下发一个EC块。也可以直接由存储客户端向存储节点组中的每个存储节点下发一个EC块。

[0120] 步骤508:在第一分区节点组中的存储节点写入数据块或校验块的过程中,某个存储节点发生故障,导致EC块写入失败。

[0121] 当任一存储节点写入EC块失败时,则表示存储客户端写入所述待写入数据失败。

[0122] 步骤509:数据写入失败,存储客户端接收来自第一分区对应的存储节点组的失败响应。

[0123] 在一种可能的实施方式中,所述失败响应可以由存储节点组中的主节点发送。

[0124] 步骤510:存储客户端向管理服务器发送第二创建请求,所述第二创建请求用于请求管理服务器为所述待写入数据重新分配数据单元。

[0125] 具体的,所述第二创建请求中可以携带写入失败的第一分区标识,以使得管理服务器获取存在故障节点的分区标识。管理服务器也可以实时感知到存储节点的状态,所以重新分配数据单元的时候,就可以选择其他的无故障节点的分区。

[0126] 步骤511:管理服务器接收所述第二创建请求,为所述待写入数据分配第二分区,在第二分区中分配第二数据单元。

[0127] 所述管理服务器在创建第二数据单元,记录第二数据单元标识后,记录第二数据

单元状态。

[0128] 步骤512:管理服务器向存储客户端返回第二创建响应,所述创建响应携带第二数据单元的标识。

[0129] 步骤513:存储客户端接收所述第二创建响应,根据所述第二数据单元标识确定所述第二分区的标识,存储客户端获取第二数据单元的状态以及第二分区视图。

[0130] 示例性的,以上述表1中的P2为例进行说明,此时,分区P2的分区视图为{Node3, Node5, Node6, Node8, Node9, Node10},分区P2的分区状态为{111111},第二数据单元状态为{111111}。

[0131] 步骤514:存储客户端对待写入数据执行EC编码,生成待写入数据对应的数据块和校验块。

[0132] 可选的,存储客户端可以使用步骤506的计算结果,不再执行步骤514。

[0133] 步骤515:存储客户端将待写入数据对应的数据块和校验块存储到第二分区包含的各个存储节点中。

[0134] 需要说明的是,图5对应的实施例为写入失败时管理服务器将待写入数据重定向到另一分区的方法。在分区不降级的情况下,管理服务器为存储客户端分配的分区中的N+M个节点均为正常节点,因此,图5对应的实施例可以不使用数据单元的状态(由于分区中的N+M个节点均为正常节点,管理服务器分配的数据单元的状态必然是表示N+M个节点都正常),存储客户端也不需要确定需要生成的数据块的数量和校验块的数量,由于分区不降级,数据块的数量必然是N,校验块的数量必然是M。

[0135] 如图6所示,为本发明实施例提供的另一种数据存储的方法流程示意图,在数据存储过程中动态调整EC冗余比。所述方法包括:

[0136] 步骤601-610与步骤501-510相同,本发明实施例不再赘述。

[0137] 步骤611:管理服务器接收所述第二创建请求,在第三分区中为待写入数据分配第三数据单元。其中,所述第三分区包含的存储节点中存在故障节点。

[0138] 管理服务器根据可用分区数量判断是否允许分区降级,当允许分区降级时,在包含故障节点的所述第三分区为待写入数据分配数据单元。

[0139] 示例性的,当第一分区中的第3个存储节点故障时,更新后的第一分区状态为{110111},则第三数据单元状态为{110111}。

[0140] 管理服务器在创建第三数据单元后,记录第三数据单元状态。

[0141] 步骤612:管理服务器向存储客户端返回第二创建响应,所述第二创建响应携带第三数据单元标识。可选的,第三创建响应还可以携带第三数据单元状态和/或第三分区视图。

[0142] 步骤613:存储客户端接收所述存储客户端接收所述第三创建响应,根据所述第三数据单元标识确定所述第三数据单元所在的第三分区,存储客户端获取第三数据单元状态以及第三分区视图,根据所述第三数据单元的状态和第三分区视图确定第三分区中的非故障节点列表。

[0143] 示例性的,以上述表1中的Pn为例进行说明,此时,分区Pn的分区视图为{Node10, Node1, Node4, Node9, Node6, Node11},假设分区Pn中Node4故障,则第三分区状态为{110111},第三数据单元状态为{110111},非故障节点包括Node10, Node1, Node9, Node6,

Node11。

[0144] 步骤614:存储客户端根据所述第三数据单元的状态确定所述第三分区的冗余比,对待写入数据执行EC编码,生成待写入数据对应的数据块和校验块。

[0145] 在一种可能的实施方式中,为了保证数据存储的持久度,需要保证校验块的数量不变,即使分区中存在故障节点,也只是会降低生成数据块的数量,而不会降低生成校验块的数量,此时,生成的数据块的数量等于N减去第二分区中故障节点的数量,生成的校验块的数量等于M。由于校验块的数量M保存不变,存储客户端可以使用第二分区中非故障节点的数量减去M,即可获得需要生成的数据块的数量。校验块M的数量可以由用户预先在数据存储系统中设置,也可以由存储客户端和管理服务器协商确定。

[0146] 示例性的,需要生成的校验块的数量 $M=2$ 保持不变,由于Node4故障,需要生成的数据块的数量为非故障节点数量(5)减去校验块数量(2),所有,需要生成的数据块的数量为3。因此确定 $N=3, M=2$ 。存储客户端对待写入数据执行EC编码,生成3个数据块以及2个校验块。

[0147] 步骤615:存储客户端将待写入数据对应的数据块和校验块存储到第三分区中的非故障存储节点中,每个存储节点存储一个EC块。

[0148] 存储客户端将生成的3个数据块和2个校验块发送给第三分区的非故障节点。

[0149] 步骤511或步骤611后,前述图5和图6的实施例进一步包括:管理服务器确定第一分区存在故障节点,更新第一分区的分区状态。

[0150] 存储客户端向分布式存储系统写入数据时,管理服务器可以在存在故障节点的第一分区为待写入数据分配第一数据单元,并将第一数据单元的标识和第一数据单元的状态通知给存储客户端,存储客户端根据第一数据单元的状态确定第一分区的冗余比(需要生成的数据块的数量和校验块的数量),对待写入数据执行EC编码生成数据块和校验块,将生成的数据块和校验块发送给第二分区中的非故障节点进行存储。通过上述方式,当分区中存在故障节点时,管理服务器可以执行分区降级,将在分区中分配的数据单元的状态通知给存储客户端,使得存储客户端获得存在故障节点的第二分区的冗余比,根据确定的第二分区的冗余比,减少生成数据块的数量,将待写入数据存储到第二分区中的非故障节点中,通过将故障节点排除在数据存储范围,避免将故障节点上的数据迁移到新的存储节点以替代故障节点,降低了节点故障对存储系统读写性能的影响。

[0151] 在上述过程中,管理服务器将第一数据单元的状态通知给存储客户端,存储客户端根据第一数据单元的状态即可确定第一分区中非故障节点列表以及需要生成的数据块的数量和校验块的数量。

[0152] 管理服务器确定需要分区降级的触发条件可以是:存储客户端在向管理服务器分配的数据单元所在的分区包含的存储节点写入待写入数据对应的EC条带时,一个或多个存储节点故障导致写入失败。在管理服务器确定需要分区降级之前,存储客户端从管理服务器获取为待写入数据分配的第二数据单元的标识,当第二数据单元所在的第二分区中的某个存储节点发生故障,导致所述存储客户端向第二分区写入数据失败。

[0153] 需要说明的是,在分区降级的情况下,存储客户端如果仍然写入失败,管理服务器可以进一步的为待写入数据分配另一分区,所述另一分区可以为降级分区,即在所述另一分区中允许所述存储客户端使用的节点为所述另一分区的部分节点,所述管理服务器将在

所述另一分区中分配的数据单元的状态通知给所述存储客户端。

[0154] 在本发明各实施例中,同一实施例中的不同分区均属于同一DHT环。

[0155] 如图7所示,为本发明实施例提供的一种数据读取的流程示意图,包括:

[0156] 步骤701:存储客户端处理用户设备的读请求,执行读操作,获取待读取数据所在数据单元的状态,以及获取待读取数据所在分区的分区视图和分区状态,确定待读取的存储节点列表。

[0157] 示例性的,以 $N=4$ 、 $M=2$ 、数据单元状态为{110111}、分区状态{111011}、分区视图为{Node4,Node6,Node7,Node9,Node10,Node11}进行举例说明。其中,数据单元状态为{110111}表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,数据单元的状态{110111}表示所述管理服务器允许存储客户端使用第1、2、4、5和6个存储节点,第3个存储节点不允许使用(第3个存储节点可能故障);分区状态{111011}表示:当前分区包含的第4个存储节点故障,其他存储节点状态正常。存储客户端根据数据单元状态和分区状态确定,存储有待读取数据对应的EC块的节点为第1、2、5、6存储节点,结合分区视图,则可以确定记录有待读取数据对应的EC块的存储节点为Node4,Node6,Node10以及Node11。

[0158] 步骤702:存储客户端从待读取的存储节点列表包含的存储节点中读取数据对应的EC块。

[0159] 存储客户端确定待读取数据所在的数据单元的方式可以采用现有技术中描述的方式,本发明实施例并不进行限定。示例性的,所述读请求中携带待读取数据所在的数据单元的标识,当所述读请求中不携带数据单元的标识时,可以根据读请求携带的卷标识和读取长度查找数据单元元数据,根据卷标识和读取长度确定待读取数据对应的数据单元的标识。

[0160] 需要说明的是,在找到待读取数据对应的EC块所在的存储节点列表后,从列表中找到对应EC块存储的位置可以采用本领域常用技术手段,例如,存储客户端向存储节点发送读请求,所述读请求携带起始地址和读取长度,存储节点根据起始地址和读取场景即可读取EC块。

[0161] 根据图7所述的实施例,当数据写入是采用分区降级的方式时,在读取该数据时,存储客户端可以获取数据写入的数据单元的状态,从而可以确定数据写入到存储节点列表,再结合当前状态为正常的存储节点列表,即可确定存储待读取数据对应的EC块的存储节点列表,从而完成数据读取过程。

[0162] 如图8所示,为本发明实施例提供的存储客户端的结构示意图,所述存储客户端800为上述各实施例中的存储客户端,执行前述各实施例中存储客户端的功能。

[0163] 具体的,所述存储客户端800包括第一发送单元801、获取单元802、以及处理单元803,

[0164] 第一发送单元801,用于为待写入数据向所述管理服务器请求分配数据单元;

[0165] 获取单元802,用于获取所述管理服务器返回的第一数据单元的标识、所述第一数据单元的状态以及所述第一数据单元所在的第一分区的第一分区视图,其中,所述第一分区视图表示所述第一分区包含的存储节点列表,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,允许所述存储客户端使用的存储节点的数量小

于N+M;

[0166] 处理单元803,用于根据所述第一数据单元的状态确定需要生成的数据块的数量和校验块的数量,对所述待写入数据执行EC编码,生成所述待写入数据对应的数据块和校验块,生成的数据块的数量和校验块的数量之和等于所述允许所述存储客户端使用的存储节点的数量;

[0167] 第一发送单元801,还用于根据所述第一分区视图和所述第一数据单元的状态,将生成的数据块和校验块存储到所述第一分区中允许所述存储客户端使用的各存储节点中,每个节点存储一个数据块或校验块。

[0168] 所述获取单元802,还用于在所述存储客户端执行读操作时,获取待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包含的各存储节点当前是否故障;

[0169] 所述处理单元803,还用于根据所述第二数据单元的状态、所述第二分区视图和所述第二分区状态,确定存储有所述待读取数据对应的EC块且状态为正常的存储节点列表;

[0170] 所述获取单元802,还用于从确定的所述存储节点列表中读取所述待读取数据对应的EC块。

[0171] 在第一发送单元801为待写入数据向所述管理服务器请求分配数据单元之前,

[0172] 所述获取单元802,还用于获取所述管理服务器返回的第三数据单元的标识;

[0173] 所述第一发送单元801,还用于向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块;

[0174] 当写入失败时,所述第一发送单元801向所述管理服务器请求重新分配数据单元。

[0175] 如图9所示,为本发明实施例提供的管理服务器的结构示意图,所述管理服务器900为上述各实施例中的管理服务器,执行前述各实施例中管理服务器的功能。

[0176] 所述管理服务900器包括:

[0177] 接收单元901,用于接收存储客户端发送的为待写入数据分配数据单元的请求;

[0178] 分配单元902,用于为所述待写入数据在第一分区中分配第一数据单元;

[0179] 所述第二发送单元903,还用于向所述存储客户端返回所述第一数据单元的标识、第一数据单元的状态以及所述第一分区的所述第一分区视图,其中,所述第一数据单元的状态指示所述第一分区中允许所述存储客户端使用的存储节点,所述第一分区视图表示所述第一分区包含的存储节点列表,允许所述存储客户端使用的存储节点的数量小于N+M。

[0180] 所述管理服务器900还包括存储单元904,用于记录所述第一数据单元的状态。

[0181] 所述第一分区中存在故障节点,所述允许所述存储客户端使用的存储节点为所述第一分区中的非故障节点,所述第一数据单元的状态表示所述管理服务器创建所述第一数据单元时所述第一分区包含的各存储节点的状态。

[0182] 所述第二发送单元903,还用于在存储客户端执行读操作时,向所述存储客户端返回待读取数据对应的第二数据单元的状态、所述第二数据单元所在的第二分区的第二分区视图和第二分区状态,所述第二数据单元的状态表示所述管理服务器创建所述第二数据单元时允许所述存储客户端使用的第二分区中存储节点,所述第二分区状态表示第二分区包

含的各存储节点当前是否故障。

[0183] 在所述存储客户端为待写入数据向所述管理服务器900请求分配数据单元之前,所述第二发送单元903,还用于向所述存储客户端返回第三数据单元的标识;

[0184] 所述接收单元901,还用于在所述存储客户端向所述第三数据单元所在的第三分区中包含的各存储节点写入根据待写入数据生成的EC块失败时,接收所述存储客户端发送的重新分配数据单元的请求。

[0185] 所述管理服务器900还包括降级判断单元905,用于根据存在故障节点的分区数量判断是否允许分区降级,如果允许,则在所述第一分区中为所述待写入数据分配所述第一数据单元。

[0186] 所述降级判断单元905,具体用于判断存在故障节点的分区数量与分区总数比值是否大于预设比例,如果是,则确定允许分区降级。

[0187] 所述分配单元902,还用于在所述降级判断单元确定不允许分区降级时,在不存在故障节点的第三分区中为所述待写入数据分配第三数据单元,向所述存储客户端返回所述第三数据单元的标识。

[0188] 存储客户端和管理服务器可以采用通用的计算机设备,示例性的,如图10所示,为本发明实施例提供的物理服务器硬件结构示意图。所述物理服务器执行前述存储客户端或管理服务器的功能。物理服务器1000包括至少一个处理器1001,通信总线1002,存储器1003以及至少一个通信接口1004。

[0189] 处理器1001可以为通用中央处理器(CPU)、微处理器、特定应用集成电路(application-specific integrated circuit,ASIC),或一个或多个用于控制本发明方案程序执行的集成电路。

[0190] 通信总线1002可包括一通路,在上述组件之间传送信息。所述通信接口1004,使用任何收发器一类的装置,用于与其他设备或通信网络通信,如以太网,无线接入网(RAN),无线局域网(Wireless Local Area Networks,WLAN)等。

[0191] 存储器1003可以是只读存储器(read-only memory,ROM)或可存储静态信息和指令的其他类型的静态存储设备,随机存取存储器(random access memory,RAM)或者可存储信息和指令的其他类型的动态存储设备,也可以是电可擦可编程只读存储器(Electrically Erasable Programmable Read-Only Memory,EEPROM)、只读光盘(Compact Disc Read-Only Memory,CD-ROM)或其他光盘存储、光碟存储(包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等)、存储节点存储介质或者其他磁存储设备、或者能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。存储器1003可以是独立存在,通过总线与处理器相连接。存储器也可以和处理器集成在一起。

[0192] 其中,所述存储器1003用于存储执行本发明方案的应用程序代码,并由处理器1001来控制执行。所述处理器1001用于执行所述存储器1003中存储的应用程序代码,以实施前述存储客户端或者管理服务器的功能。

[0193] 本发明实施例涉及到的计算机程序指令可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,通过执行计算机程序指令可以实现前述各实施例中各部件的功能。

[0194] 需要说明的是,在权利要求书和说明书中,第一、第二、第三、第四等词仅仅是为了描述的便利而增加的,主要用于对同类名称进行进行描述上的区分,这些词本身不具有特定化的含义,也不表示时序、程度或重要性等方面的不同。例如,权利要求书中的“第一分区”与说明书中的“第一分区”,以及说明书中不同实施例之间的“第一分区”可以相同,也可以不同。

[0195] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的保护范围。

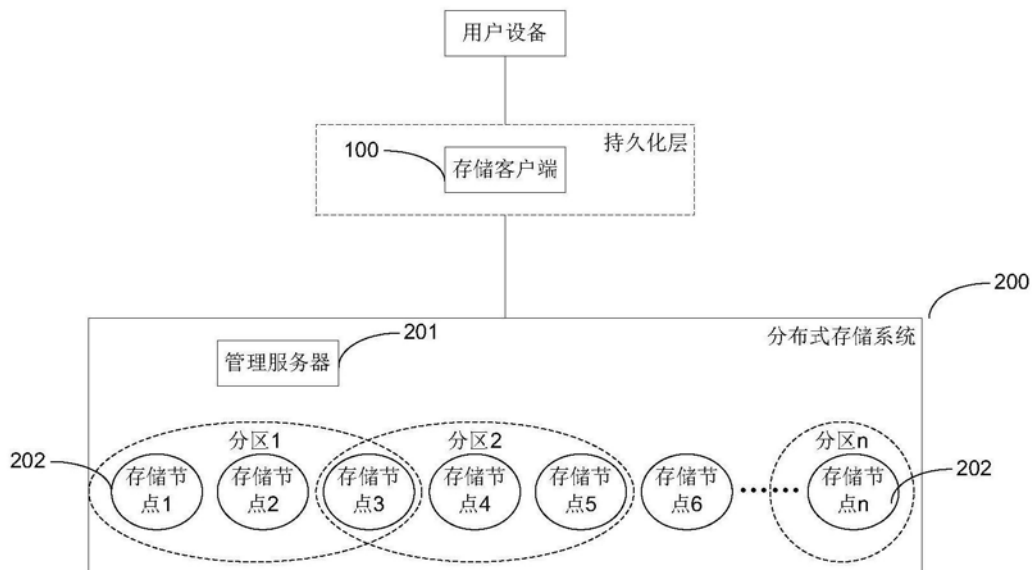


图1

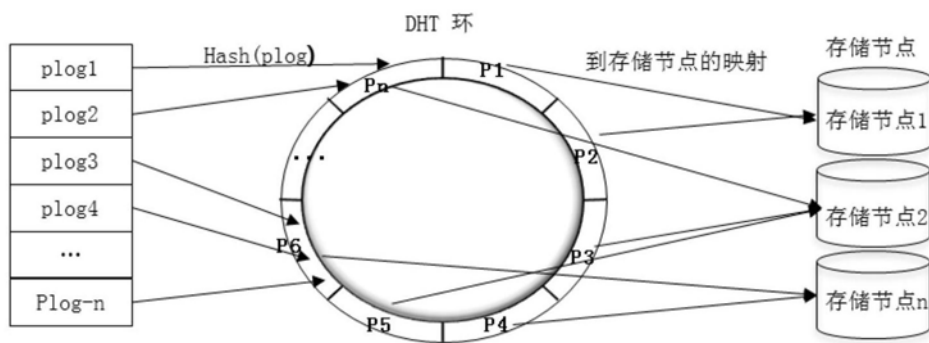


图2

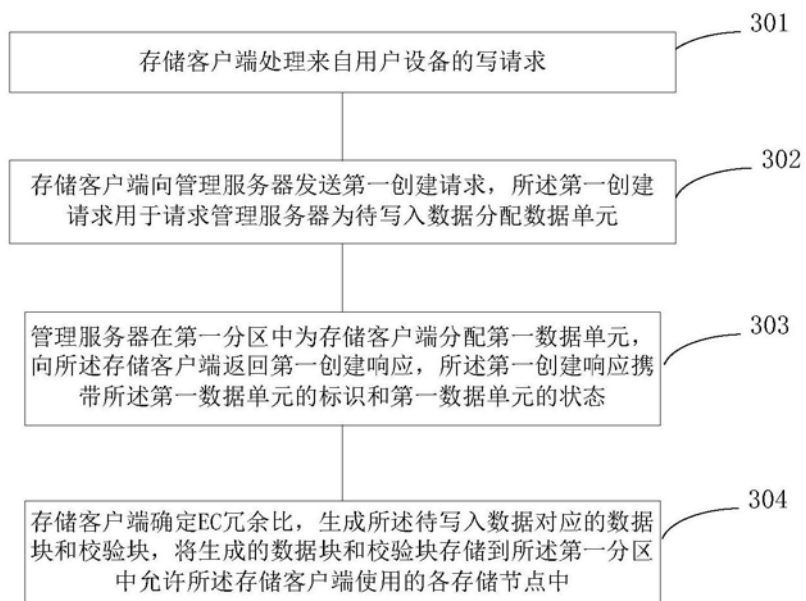
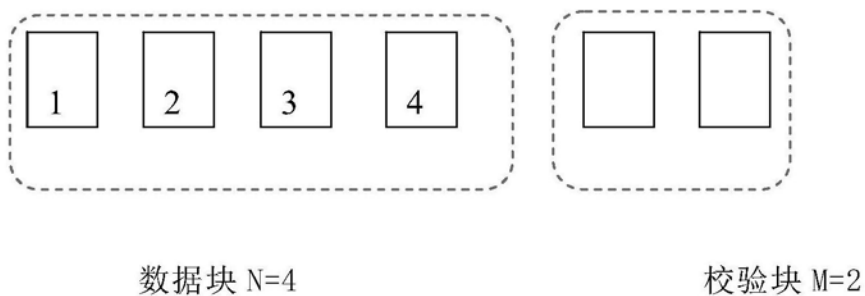


图3

分区 P 中各节点均正常时：



节点 n3 故障时：

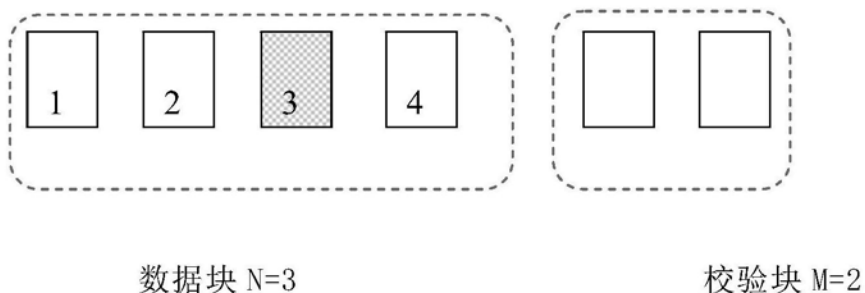


图4

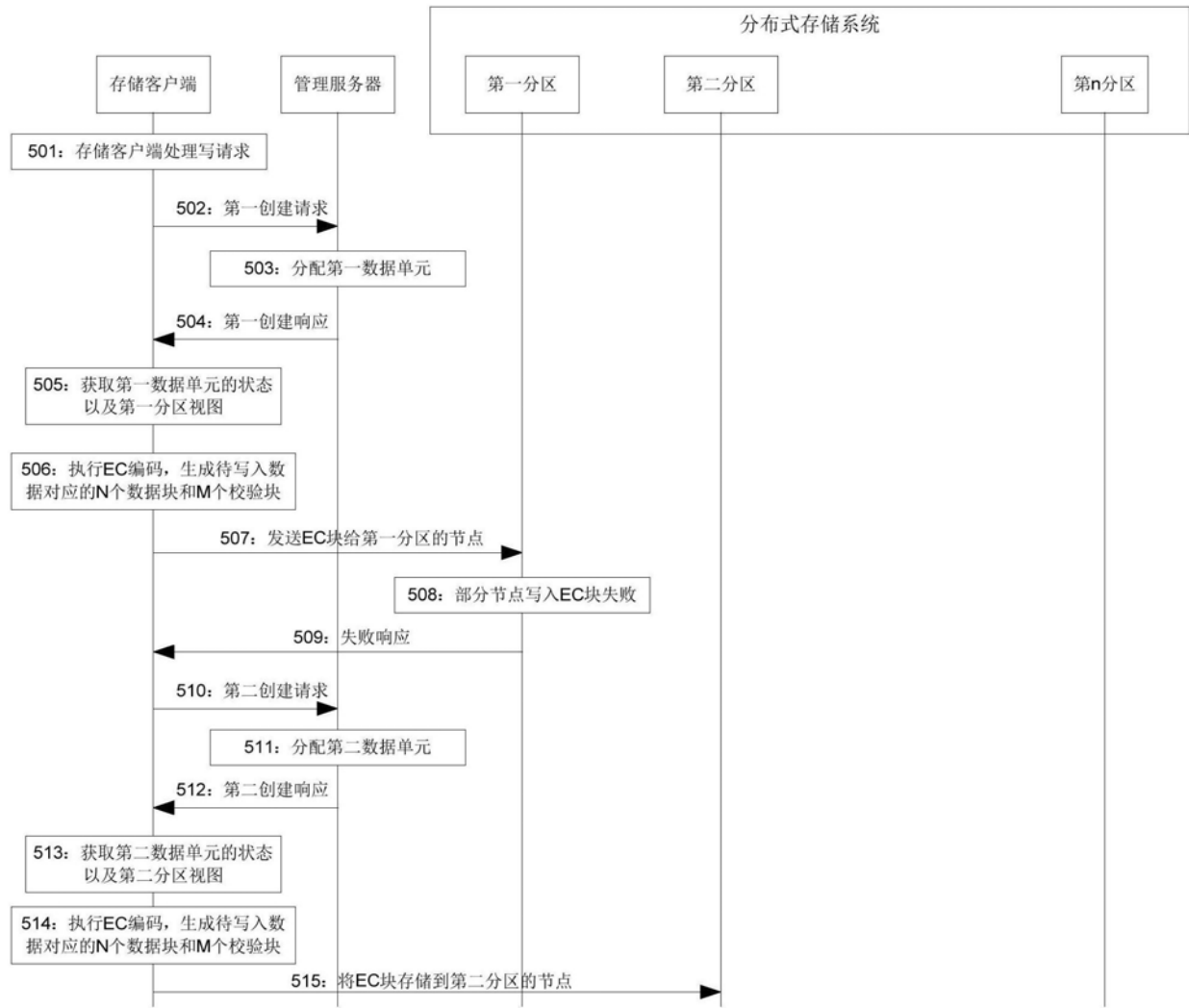


图5

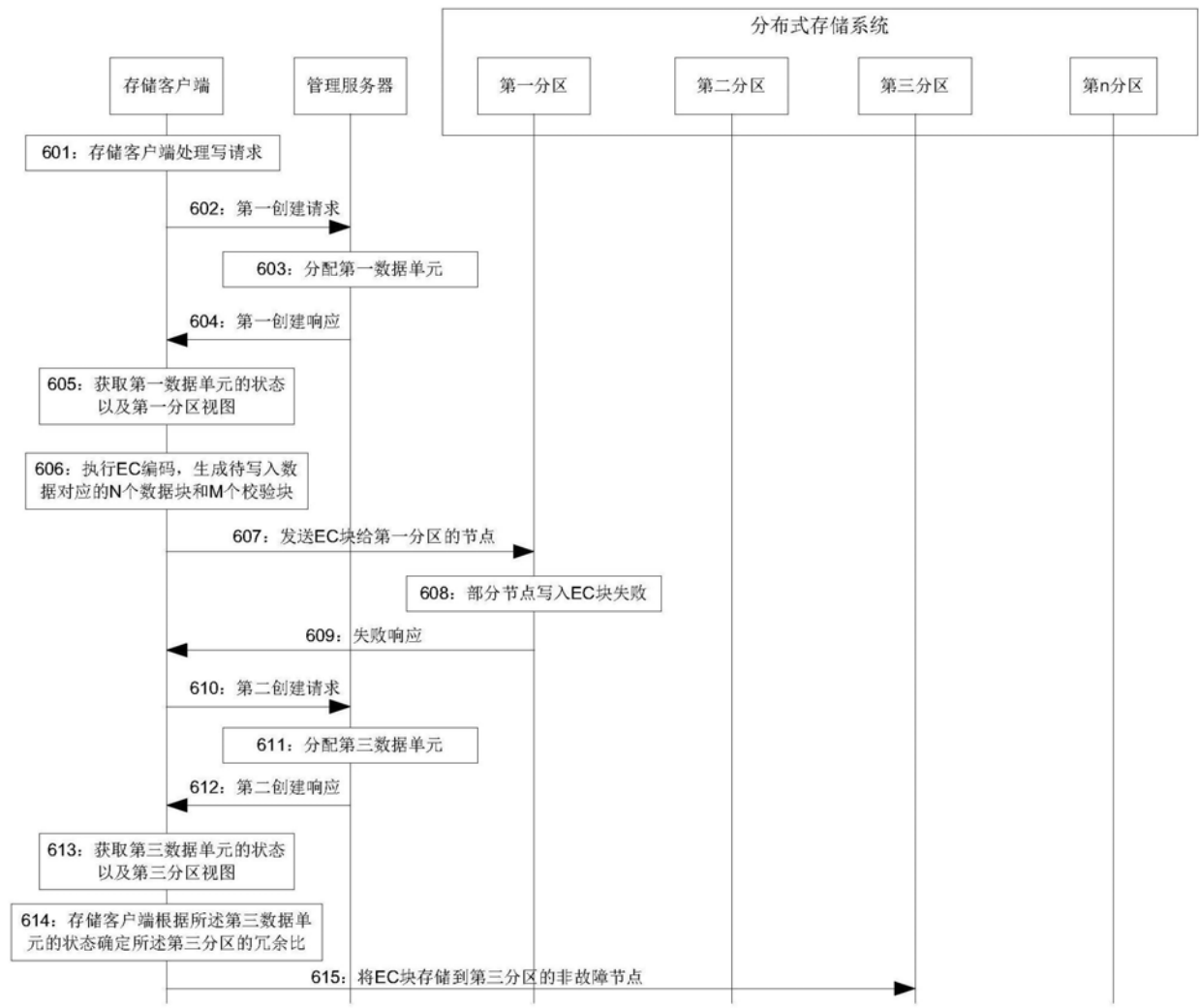


图6

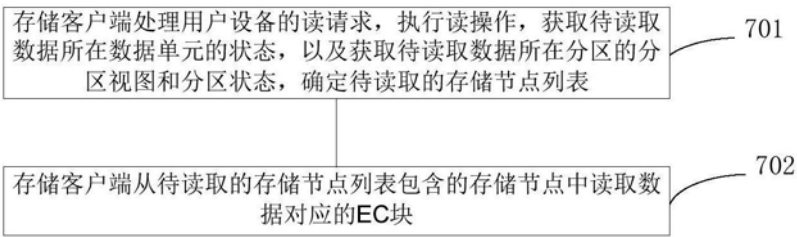


图7

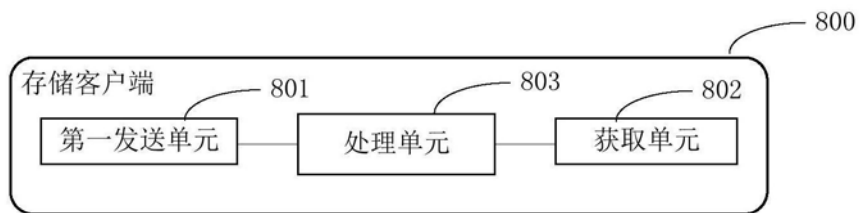


图8

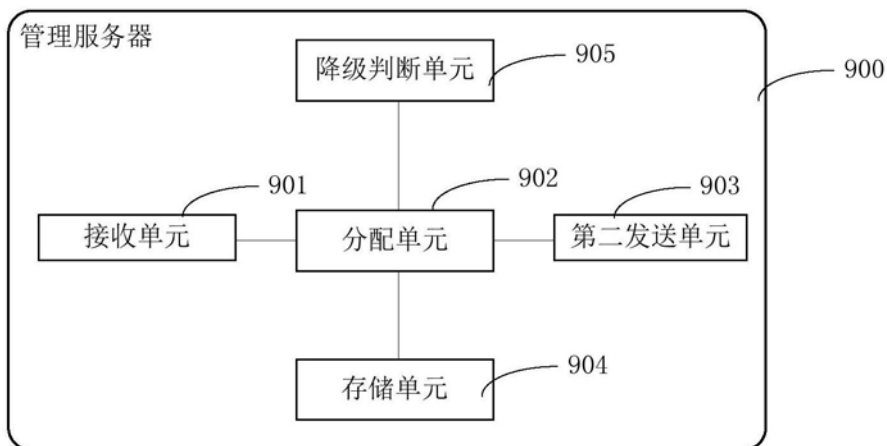


图9

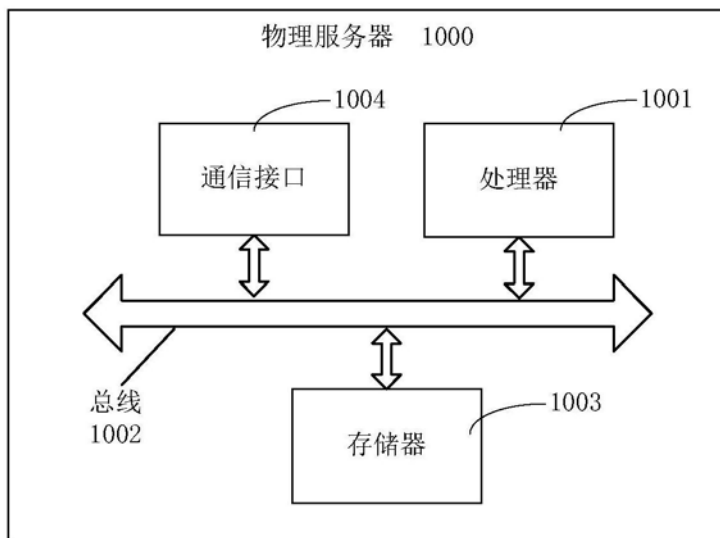


图10