(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 106062717 B
(45)授权公告日 2019.05.03

(21)申请号 201480040590.9

(22)申请日 2014.11.06

(65)同一申请的已公布的文献号
申请公布号 CN 106062717 A

(43)申请公布日 2016.10.26

(85)PCT国际申请进入国家阶段日
2016.01.15

(86)PCT国际申请的申请数据
PCT/CN2014/090445 2014.11.06

(87)PCT国际申请的公布数据
WO2016/070375 ZH 2016.05.12

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

(72)发明人 王道辉 张烽 刘叙友

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
US 2013290249 A1,2013.10.31,
US 2012166390 A1,2012.06.28,
CN 101751284 A,2010.06.23,
CN 103051691 A,2013.04.17,
CN 1474275 A,2004.02.11,
CN 103810047 A,2014.05.21,
CN 102724057 A,2012.10.10,
Philip A. Bernstein 等."Adapting microsoft SQL server for cloud computing".《2011 IEEE 27th International Conference on Data Engineering》.2011,第1255-1263页.

审查员 牛洪波

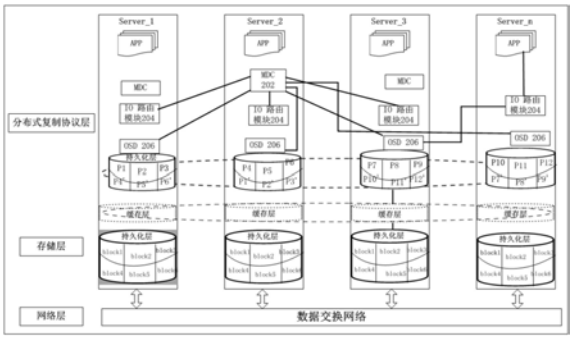
权利要求书12页 说明书33页 附图10页

(54)发明名称

一种分布式存储复制系统和方法

(57)摘要

本发明公开了一种分布式存储复制系统、方法,该系统包括至少一个元数据控制(MDC)模块,多个IO路由模块以及多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源相对应的至少一个逻辑分区(partition);所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求执行IO请求所对应的数据的存储;所述MDC确定所述系统中某个OSD节点为故障OSD节点,确定所述故障OSD节点上的partition,更新所述故障OSD节点上的partition所在的partition群的partition视图,向所述更新的partition视图中的主OSD节点发送更新通知;所述主OSD节点用于接收所述MDC模块发送的所述更新通知后,根据所述更新的partition视图处理IO请求所对应的数据的复制。应用本发明实施例,提高了一致性复制的处理性能,容错能力以及可用性。



CN 106062717 B

1.一种分布式存储复制系统,所述系统包括至少一个元数据控制(MDC)模块,多个IO路由模块以及多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源相对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图,所述partition视图包括一个partition群中的各partition所在的OSD信息;所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求执行IO请求所对应的数据的存储;其特征在于:

所述MDC确定所述系统中某个OSD节点为故障OSD节点,确定所述故障OSD节点上的partition,更新所述故障OSD节点上的partition所在的partition群的partition视图,向所述更新的partition视图中的主OSD节点发送更新通知;以及

所述主OSD节点用于接收所述MDC模块发送的所述更新通知后,根据所述更新的partition视图处理IO请求所对应的数据的复制;所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态;所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图;

所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括:

根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上,或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

2.根据权利要求1所述的系统,其特征在于,所述MDC模块进一步用于生成IO视图,所述IO视图包括一个partition群的主OSD节点的标识,将所述IO视图发送给所述IO路由模块以及所述partition视图中的partition所在的OSD节点;以及

所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的IO视图,并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

3.根据权利要求2所述的系统,其特征在于:

所述的MDC模块进一步用于当确定所述故障OSD节点上的partition包括主partition时,更新所述主partition所在的partition群的IO视图,将所述更新的IO视图通知给所述更新的partition视图中的备OSD节点;以及

所述更新的partition视图中的备OSD节点用于根据所述更新的IO视图更新本地保存的IO视图并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

4.根据权利要求3所述的系统,其特征在于:

所述MDC模块进一步用于将所述更新的IO视图通知给所述IO路由模块;以及

所述IO路由模块根据所述更新的IO视图更新所述IO路由模块本地保存的IO视图并根

据所述更新后本地保存的IO视图处理IO请求的转发。

5.根据权利要求3所述的系统,其特征在于,所述更新所述故障OSD节点上的partition所在的partition群的partition视图具体包括:

当所述故障OSD节点上的partition包括备partition时,将所述备partition所在的partition群的partition视图中所述故障OSD节点的partition状态标记为不一致;以及

当所述故障OSD节点上的partition包括主partition时,将所述主partition所在的partition群的partition视图中作为主OSD节点的所述故障OSD节点降为新的备OSD节点并将所述新的备OSD节点对应的partition状态标记为不一致,并从所述主partition所在的partition群的partition视图中的原备OSD节点中选择一个partition状态为一致的备OSD节点升为新的主OSD节点。

6.根据权利要求5所述的系统,其特征在于:

在所述故障OSD节点故障恢复并完成数据恢复后,所述MDC模块用于进一步更新所述更新的partition视图以及所述更新的IO视图,向所述进一步更新的partition视图中的主OSD节点发送更新通知,向所述进一步更新的partition视图中的备OSD节点发送更新通知;

所述进一步更新的partition视图中的主OSD节点用于根据所述进一步更新的partition视图处理所述IO请求所对应的数据的复制;以及

所述进一步更新的partition视图中的备OSD节点用于根据所述进一步更新的IO视图处理所述IO请求所对应的数据的复制。

7.根据权利要求1-6所述的任一系统,其特征在于,所述系统包括多个主机,所述MDC模块,IO路由模块和OSD节点分别部署在所述多个主机中的至少一个主机上,所述OSD节点用于管理所述主机上的物理存储资源。

8.一种分布式存储复制系统,所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块,多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图以及IO视图,所述partition视图包括一个partition群中的各partition所在的OSD信息,所述IO视图包括一个partition群的主OSD的标识;所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求执行IO数据的存储;其特征在于:

所述IO路由模块用于接收IO请求,所述IO请求包括key,根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点,将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求,并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点;

所述主OSD节点用于接收所述IO请求,在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与所述主OSD节点本地保存的IO视图版本一致后,执行所述IO请求,生成携带有所述IO视图版本信息的复制请求,并将所述的复制请求发送至所述数据所属的

partition群的备OSD节点;以及

所述备OSD节点用于接收所述复制请求,并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后,执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

9.根据权利要求8所述的系统,其特征在于,所述主OSD节点进一步用于在根据所述IO视图版本信息确定所述IO请求中的IO视图版本旧于所述主OSD本地保存的IO视图版本后向所述的IO路由模块返回错误,以及在确定所述IO请求中的IO视图版本新于所述主OSD本地保存的IO视图版本后将所述IO请求加入缓冲队列,并向所述MDC模块查询所述数据所属的partition群的IO视图的IO视图版本信息以确定所述主OSD本地保存的IO视图版本与所述IO请求中的IO视图版本一致后执行所述IO请求;以及

所述IO路由模块用于收到所述主OSD节点返回的错误后向所述MDC模块查询所述数据所属的partition群的IO视图,在获得更新的IO视图版本信息后发送所述携带所述更新的IO视图版本信息的IO请求。

10.根据权利要求8所述的系统,其特征在于,其特征在于,所述IO视图版本信息包括IO视图版本号,所述主OSD节点进一步为所述IO请求生成序列标识,所述序列标识包括所述IO视图版本号和序列号,并将所述序列标识加入到发送至所述备OSD节点的所述复制请求;所述序列号表示在一个IO视图版本内针对所述IO视图中的partition群对应的数据的修改操作的连续编号;

所述备OSD节点进一步用于根据所述序列标识执行所述复制请求。

11.根据权利要求10所述的系统,其特征在于,所述主OSD节点进一步用于在所述IO请求为修改操作时记录一条entry,所述的entry包括操作类型、所述数据所属的partition群的partition群ID、所述序列标识以及所述key。

12.根据权利要求10所述的系统,其特征在于,所述复制请求中进一步携带有所述主OSD节点针对所述partition群发送的上一次复制请求中的序列标识;

所述备OSD节点用于在收到所述的复制请求后并在确定所述上一次复制请求的序列标识与所述备OSD节点本地保存的最大序列标识一致的情况下执行所述复制请求。

13.根据权利要求12所述的系统,其特征在于,所述备OSD节点进一步用于在确定所述上一次复制请求的序列标识大于所述备OSD节点本地保存的最大序列标识时请求所述主OSD节点重发丢失的请求,以保持所述备OSD节点与所述主OSD节点上的数据一致。

14.根据权利要求8所述的系统,其特征在于,所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态;所述MDC模块进一步用于:

在所述IO请求处理过程中检测到所述主OSD节点发生故障时将所述数据所属的partition群的partition视图中的所述主OSD节点降为新的备OSD节点,并将所述新的备OSD的partition状态标记为不一致,将所述数据所属的partition群的partition视图中的所述备OSD节点中任一备OSD节点升为新的主OSD节点,将更新后的所述数据所属的partition群的partition视图通知至所述新的主OSD节点,用所述新的主OSD节点更新所述数据所属的partition群的IO视图,将更新后的所述数据所属的partition的IO视图通知至

所述IO路由模块；

所述IO路由模块进一步用于接收所述MDC模块发送的所述更新后的所述partition群的IO视图，根据所述更新后的所述partition群的IO视图将所述IO请求发送至所述新的主OSD节点；以及

所述新的主OSD节点用于接收所述IO请求，在执行所述IO请求后生成第二复制请求，将所述第二复制请求发送至所述更新后的所述数据所属的partition群的partition视图中的partition状态为一致的备OSD节点。

15.根据权利要求8所述的系统，其特征在于，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及相应的partition状态，所述MDC模块进一步用于在所述IO请求处理过程中检测到所述备OSD节点中任一备OSD节点发生故障时，将所述数据所属的partition群的partition视图中的所述任一备OSD的partition状态标记为不一致，将更新后的所述数据所属的partition群的partition视图通知至所述主OSD节点；以及

所述主OSD节点用于接收到所述更新后的所述数据所属的partition群的partition视图后，将所述复制请求发送至所述更新后的partition视图中的partition状态为一致的备OSD节点，不将所述复制请求发送至partition状态为不一致的备OSD节点。

16.根据权利要求8-15所述的任一系统，其特征在于，所述系统包括多个主机，所述MDC模块，IO路由模块和OSD节点分别部署在所述多个主机中的至少一个主机上，所述OSD节点用于管理所述主机上的物理存储资源。

17.一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于对接收的IO请求进行路由转所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；其特征在于：

所述OSD节点用于在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图，所述OSD节点称为故障恢复OSD节点，所述查询请求中携带所述故障恢复OSD节点的OSD标识，接收所述MDC返回的所述IO视图，向所述IO视图中的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据，接收所述主OSD发送的所述故障期间更新的数据，根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制；

所述MDC模块用于接收所述故障恢复OSD节点的所述查询请求，根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点，并在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图；以及

所述主OSD节点用于接收所述故障恢复OSD节点的数据恢复请求，将所述故障期间更新的数据发送至所述的故障恢复OSD节点，以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

18.根据权利要求17所述的系统，其特征在于：

所述MDC模块进一步用于在所述故障恢复OSD节点的故障发生后以及故障恢复之前更新所述partition群的partition视图和IO视图；

所述根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点具体为将所述更新后的IO视图返回给所述故障恢复OSD节点。

19.根据权利要求17所述的系统，其特征在于：

所述主OSD节点进一步用于在接收所述数据恢复请求后接收所述IO路由模块发送的针对所述故障恢复OSD节点上的partition的IO请求，执行所述IO请求，向所述的故障恢复OSD节点发送携带有IO关键信息以及所述IO请求所对应的数据的复制请求；以及

所述的故障恢复OSD节点将所述IO关键信息以及所述IO请求所对应的数据的复制请求写入日志，在所述数据恢复执行完毕后根据所述日志的记录将所述IO请求所对应的数据更新到所述故障恢复OSD节点所管理的物理存储资源中。

20.根据权利要求17所述的系统，其特征在于，所述数据恢复请求中携带有所述故障恢复OSD节点本地记录的针对所述故障恢复OSD节点上的partition的IO操作的最大序列标识；所述最大序列标识为：所述故障恢复OSD节点上的partition所在的partition群的IO视图的最新IO视图版本号，以及针对所述最新IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最大编号；

所述将所述故障期间更新的数据发送至所述的故障恢复OSD节点包括：

确定所述数据恢复请求中最大的序列标识大于或等于所述主OSD节点本地存储的当前最小的序列标识，向所述故障恢复OSD节点发送所述故障恢复OSD节点在故障期间缺少的entry，接收所述故障恢复OSD节点根据所述entry发起的数据恢复请求，将对应所述entry的数据发送至所述的故障恢复OSD节点；

所述最小序列标识为：所述主OSD节点保存的所述partition群的IO视图的最小IO视图版本号，以及针对所述最小IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最小编号。

21.根据权利要求17所述的系统，其特征在于，所述数据恢复请求中携带有所述故障恢复OSD节点本地记录的针对所述故障恢复OSD节点上的partition的IO操作的最大序列标识；

所述最大序列标识包括：所述故障恢复OSD节点上的partition所在的partition群的IO视图的最新IO视图版本号，以及在所述最新IO视图版本号对应的IO视图内针对所述IO视图中的partition对应的数据的修改操作的最大编号；

所述将所述故障期间更新的数据发送至所述的故障恢复OSD节点包括：

确定所述数据恢复请求中最大的序列标识小于所述主OSD节点本地存储的当前最小的序列标识，向所述故障恢复OSD节点发送所述主OSD本地存储的当前最小的序列标识，接收所述故障恢复OSD节点发起的同步所述主OSD节点上所属所述partition群的主partition

所对应的全部数据的数据恢复请求，将所述主partition所对应的全部数据发送至所述故障恢复OSD节点；

所述最小序列标识为：所述主OSD节点保存的所述partition群的IO视图的最小IO视图版本号，以及针对所述最小IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最小编号。

22.根据权利要求17所述的系统，其特征在于，所述MDC模块进一步用于在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图之前：

接收所述主OSD节点发送的请求更新视图的通知，所述请求更新视图的通知中携带所述主OSD节点本地保存的所述partition群的最新partition视图版本号，确定所述主OSD节点本地保存的所述partition群的最新partition视图版本号与所述MDC本地维护的最新partition视图版本号一致后执行所述更新操作。

23.根据权利要求17所述的系统，其特征在于，所述更新所述partition群的partition视图具体包括：

将所述partition视图中的故障恢复OSD节点的partition状态更新为一致，并在确认所述故障恢复OSD节点在故障前是所述partition群的主OSD节点后将所述的故障恢复OSD节点恢复为所述partition群的新的主OSD节点，将所述主OSD节点降为所述partition群的新的备OSD节点。

24.根据权利要求17-23所述的任一系统，其特征在于，所述系统包括多个主机，所述MDC模块，IO路由模块和OSD节点分别部署在所述多个主机中的至少一个主机上，所述OSD节点用于管理所述主机上的物理存储资源。

25.一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图，所述partition视图包括一个partition群中的各partition所在的OSD信息；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；其特征在于：

所述系统包括存储器以及处理器；

所述的存储器用于存储计算机可读指令，所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能；

所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，根据所述的指令促使所述处理器执行下述操作：

确定所述系统中某个OSD节点为故障OSD节点，确定所述故障OSD节点上的partition，更新所述故障OSD节点上的partition所在的partition群的partition视图，向所述更新的partition视图中的partition群所在的主OSD节点发送更新通知，以便所述主OSD节点根据

所述更新的partition视图处理IO请求所对应的数据的复制；

所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态；所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图；

所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括：

根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上，或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

26.一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；其特征在于：

所述系统包括存储器以及处理器；

所述的存储器用于存储计算机可读指令，所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能；

所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，根据所述的指令促使所述处理器执行下述操作：

促使所述IO路由模块接收IO请求，所述IO请求包括key，根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点，将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求，并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点；

促使所述主OSD节点接收所述IO请求，在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与本地保存的IO视图版本一致后，执行所述IO请求，生成携带有所述IO视图版本信息的复制请求，并将所述的复制请求发送至所述数据所属的partition的备OSD节点；以及

促使所述备OSD节点接收所述复制请求，并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后，执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

27.一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO

路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；其特征在于：

所述系统包括存储器以及处理器；

所述的存储器用于存储计算机可读指令，所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能；

所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，根据所述的指令促使所述处理器执行下述操作：

促使所述OSD节点在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图，所述OSD节点称为故障恢复OSD节点，所述查询请求中携带所述故障恢复OSD节点的OSD标识，接收所述MDC返回的所述IO视图，向所述IO视图中的的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据，接收所述主OSD发送的所述故障期间更新的数据，根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制；

促使所述MDC模块接收所述故障恢复OSD节点的所述查询请求，根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点，并在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图；以及

促使所述主OSD节点接收所述故障恢复OSD节点的数据恢复请求，将所述故障期间更新的数据发送至所述的故障恢复OSD节点，以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

28.一种应用于分布式存储系统中管理数据存储和复制的方法，所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点所管理的物理存储资源相对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图，所述partition视图包括一个partition群中的各partition所在的OSD信息；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求执行

IO请求所对应的数据的存储;其特征在于,所述的方法包括:

确定所述系统中某个OSD节点为故障OSD节点,确定所述故障OSD节点上的partition,更新所述故障OSD节点上的partition所在的partition群的partition视图,向所述更新的partition视图中的主OSD节点发送更新通知;以及

所述主OSD节点用于接收所述MDC模块发送的所述更新通知后,根据所述更新的partition视图处理IO请求所对应的数据的复制;所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态;所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图;

所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括:

根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上,或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

29.根据权利要求28所述的方法,其特征在于,所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态;所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图;

所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括:

根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上,或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

30.根据权利要求28所述的方法,其特征在于,所述的方法进一步包括:

当所述的MDC模块确定所述故障OSD节点上的partition包括主partition时,更新所述主partition所在的partition群的IO视图,将所述更新的IO视图通知给所述更新的partition视图中的备OSD节点;以及

所述更新的partition视图中的备OSD节点根据所述更新的IO视图更新本地保存的IO视图并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

31.根据权利要求30所述的方法,其特征在于,所述更新所述故障OSD节点上的partition所在的partition群的partition视图具体包括:

当所述故障OSD节点上的partition包括备partition时,将所述备partition所在的partition群的partition视图中所述故障OSD节点的partition状态标记为不一致;以及

当所述故障OSD节点上的partition包括主partition时,将所述主partition所在的partition群的partition视图中作为主OSD节点的所述故障OSD节点降为新的备OSD节点并将所述新的备OSD节点对应的partition状态标记为不一致,并从所述主partition所在的partition群的partition视图中的原备OSD节点中选择一个partition状态为一致的备OSD

节点升为新的主OSD节点。

32.根据权利要求31所述的方法,其特征在于,所述方法进一步包括:

所述MDC模块在所述故障OSD节点故障恢复并完成数据恢复后进一步更新所述更新的partition视图以及所述更新的IO视图,向所述进一步更新的partition视图中的主OSD节点发送更新通知,向所述进一步更新的partition视图中的备OSD节点发送更新通知;

所述进一步更新的partition视图中的主OSD节点根据所述进一步更新的partition视图处理所述IO请求所对应的数据的复制;以及

所述进一步更新的partition视图中的备OSD节点根据所述进一步更新的IO视图处理所述IO请求所对应的数据的复制。

33.一种应用于分布式存储系统中管理数据存储和复制的方法,所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块,多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图以及IO视图,所述partition视图包括一个partition群中的各partition所在的OSD信息,所述IO视图包括一个partition群的主OSD的标识;所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求执行IO数据的存储;其特征在于,所述的方法包括:

所述IO路由模块用于接收IO请求,所述IO请求包括key,根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点,将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求,并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点;

所述主OSD节点用于接收所述IO请求,在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与所述主OSD节点本地保存的IO视图版本一致后,执行所述IO请求,生成携带有所述IO视图版本信息的复制请求,并将所述的复制请求发送至所述数据所属的partition群的备OSD节点;以及

所述备OSD节点用于接收所述复制请求,并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后,执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

34.根据权利要求33所述的方法,其特征在于,所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态,所述方法进一步包括:

所述MDC模块在所述IO请求处理过程中检测到所述主OSD节点发生故障时将所述数据所属的partition群的partition视图中的所述主OSD节点降为新的备OSD节点,并将所述新的备OSD的partition状态标记为不一致,将所述数据所属的partition群的partition视图

中的所述备OSD节点中任一备OSD节点升为新的主OSD节点，将更新后的所述数据所属的partition群的partition视图通知至所述新的主OSD节点，用所述新的主OSD节点更新所述数据所属的partition群的IO视图，将更新后的所述数据所属的partition的IO视图通知至所述IO路由模块；

所述IO路由模块进一步用于接收所述MDC模块发送的所述更新后的所述partition群的IO视图，根据所述更新后的所述partition群的IO视图将所述IO请求发送至所述新的主OSD节点；以及

所述新的主OSD节点用于接收所述IO请求，在执行所述IO请求后生成第二复制请求，将所述第二复制请求发送至所述更新后的所述数据所属的partition群的partition视图中的partition状态为一致的备OSD节点。

35.根据权利要求33所述的方法，其特征在于，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及相应的partition状态，所述方法进一步包括：

所述MDC模块在所述IO请求处理过程中检测到所述备OSD节点中任一备OSD节点发生故障时，将所述数据所属的partition群的partition视图中的所述任一备OSD的partition状态标记为不一致，将更新后的所述数据所属的partition群的partition视图通知至所述主OSD节点；以及

所述主OSD节点用于接收到所述更新后的所述数据所属的partition群的partition视图后，将所述复制请求发送至所述更新后的partition视图中的partition状态为一致的备OSD节点，不将所述复制请求发送至partition状态为不一致的备OSD节点。

36.一种应用于分布式存储系统中管理数据存储和复制的方法，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于对接收的IO请求进行路由转所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；其特征在于，所述方法包括：

所述OSD节点用于在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图，所述OSD节点称为故障恢复OSD节点，所述查询请求中携带所述故障恢复OSD节点的OSD标识，接收所述MDC返回的所述IO视图，向所述IO视图中的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据，接收所述主OSD发送的所述故障期间更新的数据，根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制；

所述MDC模块用于接收所述故障恢复OSD节点的所述查询请求，根据查询请求中的OSD

标识将所述IO视图返回给所述故障恢复OSD节点，并在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图；以及

所述主OSD节点用于接收所述故障恢复OSD节点的数据恢复请求，将所述故障期间更新的数据发送至所述的故障恢复OSD节点，以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

一种分布式存储复制系统和方法

技术领域

[0001] 本发明涉及IT(Information Technology)信息技术领域,特别涉及一种分布式存储复制系统和方法。

背景技术

[0002] 随着信息技术的快速发展和互联网的广泛应用,人类产生的数据呈现爆炸式增长,这样对数据存储提出了更高的扩展性要求。相对于传统的存储阵列系统,分布式存储系统有着更好的扩展性和通用硬件设备适应性,能够更好的满足未来对数据存储的需求。

[0003] 分布式存储系统一般是把大量存储节点组织成一个分布式系统,通过不同节点之间的数据复制备份来保证数据的可靠性,这样一份数据在不同的存储节点上都有副本,如何保证多个数据副本的数据一致性,是分布式存储系统长期以来面临的一个问题。在保证数据一致性的情况下,系统的性能和可用性也越来越成为很重要的考虑因素。

[0004] 如图1所示为现有的两阶段提交协议(2 Phase Commit,2PC),其是一种典型的中心化的强一致性副本控制协议,许多分布数据库系统采用该协议来保证副本的一致性。

[0005] 在两阶段提交协议中,系统一般包含两类节点:一类是协调者(coordinator),一类是参与者(participant)。协调者负责发起对数据更新的投票以及通知投票决定的执行,参与者参与数据更新的投票以及执行投票的决定。

[0006] 两阶段提交协议由两个阶段组成:阶段一:请求阶段,协调者通知参与者对数据的修改进行投票,参与者告知协调者自己的投票结果:同意或者否决;阶段二:提交阶段,协调者根据第一阶段的投票结果进行决策:执行或者取消。

[0007] 一次成功的两阶段提交协议中,协调者与每个参与者至少需要两轮交互4个消息,过多的交互次数会降低性能,此外,两阶段提交协议中的节点如果发生故障或者一直没有响应,其他的IO请求便会阻塞,最后导致超时失败而需要进行数据回滚,容错能力及可用性较低。

发明内容

[0008] 本发明实施例提供一种分布式存储复制系统,以及一种应用于分布式存储系统中管理数据存储和复制的方法以解决现有一致性复制协议中低性能以及可用性低的问题。

[0009] 第一方面,一种分布式存储复制系统,所述系统包括至少一个元数据控制(MDC)模块,多个IO路由模块以及多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源相对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition

视图，所述partition视图包括一个partition群中的各partition所在的OSD信息；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求执行IO请求所对应的数据的存储；所述MDC确定所述系统中某个OSD节点为故障OSD节点，确定所述故障OSD节点上的partition，更新所述故障OSD节点上的partition所在的partition群的partition视图，向所述更新的partition视图中的主OSD节点发送更新通知；所述主OSD节点用于接收所述MDC模块发送的所述更新通知后，根据所述更新的partition视图处理IO请求所对应的数据的复制。

[0010] 在第一方面的第一种可能的实现方式中，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态；所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图；所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括：根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图partition状态为一致的备OSD节点上，或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

[0011] 结合第一方面的第一种可能的实现方式，在第二种可能的实现方式中，所述MDC模块进一步用于生成IO视图，所述IO视图包括一个partition群的主OSD节点的标识，将所述IO视图发送给所述IO路由模块以及所述partition视图中的partition所在的OSD节点；以及所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的IO视图，并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

[0012] 结合第一方面的第二种可能的实现方式，在第三种可能的实现方式中，所述的MDC模块进一步用于当确定所述故障OSD节点上的partition包括主partition时，更新所述主partition所在的partition群的IO视图，将所述更新的IO视图通知给所述更新的partition视图中的备OSD节点；以及所述更新的partition视图中的备OSD节点用于根据所述更新的IO视图更新本地保存的IO视图并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

[0013] 结合第一方面的第三种可能的实现方式，在第四种可能的实现方式中，所述更新所述故障OSD节点上的partition所在的partition群的partition视图具体包括：当所述故障OSD节点上的partition包括备partition时，将所述备partition所在的partition群的partition视图中所述故障OSD节点的partition状态标记为不一致；以及，当所述故障OSD节点上的partition包括主partition时，将所述主partition所在的partition群的partition视图中作为主OSD节点的所述故障OSD节点降为新的备OSD节点并将所述新的备OSD节点对应的partition状态标记为不一致，并从所述主partition所在的partition群的partition视图中的原备OSD节点中选择一个partition状态为一致的备OSD节点升为新的主OSD节点。

[0014] 第二方面，一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所

述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求执行IO数据的存储；所述IO路由模块用于接收IO请求，所述IO请求包括key，根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点，将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求，并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点；所述主OSD节点用于接收所述IO请求，在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与所述主OSD节点本地保存的IO视图版本一致后，执行所述IO请求，生成携带有所述IO视图版本信息的复制请求，并将所述的复制请求发送至所述数据所属的partition群的备OSD节点；以及所述备OSD节点用于接收所述复制请求，并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后，执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

[0015] 在第二方面的第一种可能的实现方式中，所述主OSD节点进一步用于在根据所述IO视图版本信息确定所述IO请求中的IO视图版本旧于所述主OSD本地保存的IO视图版本后向所述的IO路由模块返回错误，以及在确定所述IO请求中的IO视图版本新于所述主OSD本地保存的IO视图版本后将所述IO请求加入缓冲队列，并向所述MDC模块查询所述数据所属的partition群的IO视图的IO视图版本信息以确定所述主OSD本地保存的IO视图版本与所述IO请求中的IO视图版本一致后执行所述IO请求；以及所述IO路由模块用于收到所述主OSD节点返回的错误后向所述MDC模块查询所述数据所属的partition群的IO视图，在获得更新的IO视图版本信息后发送所述携带所述更新的IO视图版本信息的IO请求。

[0016] 结合第二方面或第二方面的第一种可能的实现方式，在第二种可能的实现方式中，所述IO视图版本信息包括IO视图版本号，所述主OSD节点进一步为所述IO请求生成序列标识，所述序列标识包括所述IO视图版本号和序列号，并将所述序列标识加入到发送至所述备OSD节点的所述复制请求；所述序列号表示在一个IO视图版本内针对所述IO视图中的partition群对应的数据的修改操作的连续编号；所述备OSD节点进一步用于根据所述序列标识执行所述复制请求。

[0017] 结合第二方面的第二种可能的实现方式，在第三种可能的实现方式中，所述复制请求中进一步携带有所述主OSD节点针对所述partition群发送的上一次复制请求中的序列标识；所述备OSD节点用于在收到所述的复制请求后并在确定所述上一次复制请求的序列标识与所述备OSD节点本地保存的最大序列标识一致的情况下执行所述复制请求。

[0018] 结合第二方面或者第二方面的第一、第二以及第三种可能的实现方式，在第四种可能的实现方式中，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态；所述MDC模块进一步用于：在所述IO请求处理过程

中检测到所述主OSD节点发生故障时将所述数据所属的partition群的partition视图中的所述主OSD节点降为新的备OSD节点，并将所述新的备OSD的partition状态标记为不一致，将所述数据所属的partition群的partition视图中的所述备OSD节点中任一备OSD节点升为新的主OSD节点，将更新后的所述数据所属的partition群的partition视图通知至所述新的主OSD节点，用所述新的主OSD节点更新所述数据所属的partition群的IO视图，将更新后的所述数据所属的partition的IO视图通知至所述IO路由模块；所述IO路由模块进一步用于接收所述MDC模块发送的所述更新后的所述partition群的IO视图，根据所述更新后的所述partition群的IO视图将所述IO请求发送至所述新的主OSD节点；以及所述新的主OSD节点用于接收所述IO请求，在执行所述IO请求后生成第二复制请求，将所述第二复制请求发送至所述更新后的所述数据所属的partition群的partition视图中的partition状态为一致的备OSD节点。

[0019] 结合第二方面或者第二方面的第一、第二以及第三种可能的实现方式，在第五种可能的实现方式中，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及相应的partition状态，所述MDC模块进一步用于在所述IO请求处理过程中检测到所述备OSD节点中任一备OSD节点发生故障时，将所述数据所属的partition群的partition视图中的所述任一备OSD的partition状态标记为不一致，将更新后的所述数据所属的partition群的partition视图通知至所述主OSD节点；以及所述主OSD节点用于接收到所述更新后的所述数据所属的partition群的partition视图后，将所述复制请求发送至所述更新后的partition视图中的partition状态为一致的备OSD节点，不将所述复制请求发送至partition状态为不一致的备OSD节点。

[0020] 第三方面，一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于对接收的IO请求进行路由转所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；所述OSD节点用于在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图，所述OSD节点称为故障恢复OSD节点，所述查询请求中携带所述故障恢复OSD节点的OSD标识，接收所述MDC返回的所述IO视图，向所述IO视图中的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据，接收所述主OSD发送的所述故障期间更新的数据，根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制；所述MDC模块用于接收所述故障恢复OSD节点的所述查询请求，根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点，并在所述故障恢复OSD节点数据恢复执行完毕后更

新所述partition群的partition视图;以及所述主OSD节点用于接收所述故障恢复OSD节点的数据恢复请求,将所述故障期间更新的数据发送至所述的故障恢复OSD节点,以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

[0021] 在第三方面的第一种可能的实现方式中,所述主OSD节点进一步用于在接收所述数据恢复请求后接收所述IO路由模块发送的针对所述故障恢复OSD节点上的partition的IO请求,执行所述IO请求,向所述的故障恢复OSD节点发送携带有IO关键信息以及所述IO请求所对应的数据的复制请求;以及所述的故障恢复OSD节点将所述IO关键信息以及所述IO请求所对应的数据的复制请求写入日志,在所述数据恢复执行完毕后根据所述日志的记录将所述IO请求所对应的数据更新到所述故障恢复OSD节点所管理的物理存储资源中。

[0022] 结合第三方面的第一种可能的实现方式中,在第二种可能的实现方式中,所述数据恢复请求中携带有所述故障恢复OSD节点本地记录的针对所述故障恢复OSD节点上的partition的IO操作的最大序列标识;所述最大序列标识为:所述故障恢复OSD节点上的partition所在的partition群的IO视图的最新IO视图版本号,以及针对所述最新IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最大编号;所述将所述故障期间更新的数据发送至所述的故障恢复OSD节点包括:确定所述数据恢复请求中最大的序列标识大于或等于所述主OSD节点本地存储的当前最小的序列标识,向所述故障恢复OSD节点发送所述故障恢复OSD节点在故障期间缺少的entry,接收所述故障恢复OSD节点根据所述entry发起的数据恢复请求,将对应所述entry的数据发送至所述的故障恢复OSD节点;所述最小序列标识为:所述主OSD节点保存的所述partition群的IO视图的最小IO视图版本号,以及针对所述最小IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最小编号。

[0023] 结合第三方面的第一种可能的实现方式中,在第二种可能的实现方式中,所述数据恢复请求中携带有所述故障恢复OSD节点本地记录的针对所述故障恢复OSD节点上的partition的IO操作的最大序列标识;所述最大序列标识包括:所述故障恢复OSD节点上的partition所在的partition群的IO视图的最新IO视图版本号,以及在所述最新IO视图版本号对应的IO视图内针对所述IO视图中的partition对应的数据的修改操作的最大编号;所述将所述故障期间更新的数据发送至所述的故障恢复OSD节点包括:确定所述数据恢复请求中最大的序列标识小于所述主OSD节点本地存储的当前最小的序列标识,向所述故障恢复OSD节点发送所述主OSD本地存储的当前最小的序列标识,接收所述故障恢复OSD节点发起的同步所述主OSD节点上所属所述partition群的主partition所对应的全部数据的数据恢复请求,将所述主partition所对应的全部数据发送至所述故障恢复OSD节点;所述最小序列标识为:所述主OSD节点保存的所述partition群的IO视图的最小IO视图版本号,以及针对所述最小IO视图版本号对应的IO视图中的partition对应的数据的修改操作的最小编号。

[0024] 第四方面,一种分布式存储复制系统,所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块,多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主

partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图，所述partition视图包括一个partition群中的各partition所在的OSD信息；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；所述系统包括存储器以及处理器；所述的存储器用于存储计算机可读指令，所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能；所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，根据所述的指令促使所述处理器执行下述操作：确定所述系统中某个OSD节点为故障OSD节点，确定所述故障OSD节点上的partition，更新所述故障OSD节点上的partition所在的partition群的partition视图，向所述更新的partition视图中的partition群所在的主OSD节点发送更新通知，以便所述主OSD节点根据所述更新的partition视图处理IO请求所对应的数据的复制。

[0025] 第五方面，一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求所对应的数据的存储；所述系统包括存储器以及处理器；所述的存储器用于存储计算机可读指令，所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能；所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，根据所述的指令促使所述处理器执行下述操作：促使所述IO路由模块接收IO请求，所述IO请求包括key，根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点，将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求，并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点；促使所述主OSD节点接收所述IO请求，在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与本地保存的IO视图版本一致后，执行所述IO请求，生成携带有所述IO视图版本信息的复制请求，并将所述的复制请求发送至所述数据所属的partition的备OSD节点；以及促使所述备OSD节点接收所述复制请求，并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后，执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

[0026] 第六方面，一种分布式存储复制系统，所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配

置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图以及IO视图,所述partition视图包括一个partition群中的各partition所在的OSD信息,所述IO视图包括一个partition群的主OSD的标识;所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求所对应的数据的存储;所述系统包括存储器以及处理器;所述的存储器用于存储计算机可读指令,所述指令用于执行所述MDC模块、IO路由模块以及OSD节点的功能;所述的处理器用于与所述的存储器连接,读取所述存储器中的指令,根据所述的指令促使所述处理器执行下述操作:促使所述OSD节点在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图,所述OSD节点称为故障恢复OSD节点,所述查询请求中携带所述故障恢复OSD节点的OSD标识,接收所述MDC返回的所述IO视图,向所述IO视图中的的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据,接收所述主OSD发送的所述故障期间更新的数据,根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制;促使所述MDC模块接收所述故障恢复OSD节点的所述查询请求,根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点,并在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图;以及促使所述主OSD节点接收所述故障恢复OSD节点的数据恢复请求,将所述故障期间更新的数据发送至所述的故障恢复OSD节点,以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

[0027] 第七方面,一种应用于分布式存储系统中管理数据存储和复制的方法,所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块,多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点所管理的物理存储资源相对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图,所述partition视图包括一个partition群中的各partition所在的OSD信息;所述IO路由模块用于将接收的IO请求路由至所述OSD节点;所述OSD节点用于根据所述IO请求执行IO请求所对应的数据的存储;所述的方法包括:确定所述系统中某个OSD节点为故障OSD节点,确定所述故障OSD节点上的partition,更新所故障OSD节点上的partition所在的partition群的partition视图,向所述更新的partition视图中的主OSD节点发送更新通知;以及所述主OSD节点用于接收所述MDC模块发送的所述更新通知后,根据所述更新的partition视图处理IO请求所对应的数据的复制。

[0028] 在第七方面的第一种可能的实现方式中，所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态；所述主OSD节点进一步用于根据所述更新的partition视图更新所述主OSD节点本地保存的partition视图；所述根据所述更新的partition视图处理IO请求所对应的数据的复制具体包括：根据所述更新后本地保存的partition视图将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上，或者复制到所述更新后本地保存的partition视图中partition状态为一致的备OSD节点上以及所述更新后本地保存的partition视图中partition状态为不一致但正在进行数据恢复的备OSD节点上。

[0029] 结合第七方面的第一种可能的实现方式，在第二种可能的实现方式中，所述的方法进一步包括：当所述的MDC模块确定所述故障OSD节点上的partition包括主partition时，更新所述主partition所在的partition群的IO视图，将所述更新的IO视图通知给所述更新的partition视图中的备OSD节点；以及所述更新的partition视图中的备OSD节点根据所述更新的IO视图更新本地保存的IO视图并根据所述更新后本地保存的IO视图处理所述IO请求所对应的数据的复制。

[0030] 结合第七方面的第二种可能的实现方式，在第三种可能的实现方式中，所述更新所述故障OSD节点上的partition所在的partition群的partition视图具体包括：当所述故障OSD节点上的partition包括备partition时，将所述备partition所在的partition群的partition视图中所述故障OSD节点的partition状态标记为不一致；以及当所述故障OSD节点上的partition包括主partition时，将所述主partition所在的partition群的partition视图中作为主OSD节点的所述故障OSD节点降为新的备OSD节点并将所述新的备OSD节点对应的partition状态标记为不一致，并从所述主partition所在的partition群的partition视图中的原备OSD节点中选择一个partition状态为一致的备OSD节点升为新的主OSD节点。

[0031] 第八方面，一种应用于分布式存储系统中管理数据存储和复制的方法，所述系统包括至少一个元数据控制(MDC)模块以及多个IO路由模块，多个存储对象设备(OSD)节点，所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition)，所述至少一个partition为主partition或备partition或主备partition的任意组合，主partition和所述主partition对应的备partition构成一个partition群，同一个partition群中的主partition和备partition位于不同的OSD节点上，主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点，备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点，根据所述partition生成partition视图以及IO视图，所述partition视图包括一个partition群中的各partition所在的OSD信息，所述IO视图包括一个partition群的主OSD的标识；所述IO路由模块用于将接收的IO请求路由至所述OSD节点；所述OSD节点用于根据所述IO请求执行IO数据的存储；所述的方法包括：所述IO路由模块用于接收IO请求，所述IO请求包括key，根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点，将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求，并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD

节点;所述主OSD节点用于接收所述IO请求,在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与所述主OSD节点本地保存的IO视图版本一致后,执行所述IO请求,生成携带有所述IO视图版本信息的复制请求,并将所述的复制请求发送至所述数据所属的partition群的备OSD节点;以及所述备OSD节点用于接收所述复制请求,并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后,执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

[0032] 在第八方面的第一种可能的实现方式中,所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及对应的partition状态,所述方法进一步包括:所述MDC模块在所述IO请求处理过程中检测到所述主OSD节点发生故障时将所述数据所属的partition群的partition视图中的所述主OSD节点降为新的备OSD节点,并将所述新的备OSD的partition状态标记为不一致,将所述数据所属的partition群的partition视图中的所述备OSD节点中任一备OSD节点升为新的主OSD节点,将更新后的所述数据所属的partition群的partition视图通知至所述新的主OSD节点,用所述新的主OSD节点更新所述数据所属的partition群的IO视图,将更新后的所述数据所属的partition的IO视图通知至所述IO路由模块;所述IO路由模块进一步用于接收所述MDC模块发送的所述更新后的所述partition群的IO视图,根据所述更新后的所述partition群的IO视图将所述IO请求发送至所述新的主OSD节点;以及所述新的主OSD节点用于接收所述IO请求,在执行所述IO请求后生成第二复制请求,将所述第二复制请求发送至所述更新后的所述数据所属的partition群的partition视图中的partition状态为一致的备OSD节点。

[0033] 在第八方面的第二种可能的实现方式中,所述partition视图具体包括一个partition群中各partition所在的OSD的主备身份及相应的partition状态,所述方法进一步包括:所述MDC模块在所述IO请求处理过程中检测到所述备OSD节点中任一备OSD节点发生故障时,将所述数据所属的partition群的partition视图中的所述任一备OSD的partition状态标记为不一致,将更新后的所述数据所属的partition群的partition视图通知至所述主OSD节点;以及所述主OSD节点用于接收到所述更新后的所述数据所属的partition群的partition视图后,将所述复制请求发送至所述更新后的partition视图中的partition状态为一致的备OSD节点,不将所述复制请求发送至partition状态为不一致的备OSD节点。

[0034] 第九方面,一种应用于分布式存储系统中管理数据存储和复制的方法,所述系统包括至少一个元数据控制(MDC)模以及多个IO路由模块,多个存储对象设备(OSD)节点,所述MDC模块用于为所述各OSD节点配置与所述各OSD节点所管理的物理存储资源所对应的至少一个逻辑分区(partition),所述至少一个partition为主partition或备partition或主备partition的任意组合,主partition和所述主partition对应的备partition构成一个partition群,同一个partition群中的主partition和备partition位于不同的OSD节点上,主partition所在的OSD节点为所述主partition所在的partition群的主OSD节点,备partition所在的OSD节点为所述备partition所在的partition群的备OSD节点,根据所述partition生成partition视图以及IO视图,所述partition视图包括一个partition群中的各partition所在的OSD信息,所述IO视图包括一个partition群的主OSD的标识;所述IO路

由模块用于对接收的IO请求进行路由转所述OSD节点;所述OSD节点用于根据所述IO请求所对应的数据的存储;所述方法包括:所述OSD节点用于在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图,所述OSD节点称为故障恢复OSD节点,所述查询请求中携带所述故障恢复OSD节点的OSD标识,接收所述MDC返回的所述IO视图,向所述IO视图中的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据,接收所述主OSD发送的所述故障期间更新的数据,根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制;所述MDC模块用于接收所述故障恢复OSD节点的所述查询请求,根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点,并在所述故障恢复OSD节点数据恢复执行完毕后更新所述partition群的partition视图;以及所述主OSD节点用于接收所述故障恢复OSD节点的数据恢复请求,将所述故障期间更新的数据发送至所述的故障恢复OSD节点,以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

附图说明

[0035] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,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的附图。

[0036] 图1是现有技术中的两阶段提交协议流程图;

[0037] 图2A是本发明实施例提供的分布式存储复制系统架构示意图;

[0038] 图2B是本发明另一实施例提供的分布式存储复制系统结构示意图;

[0039] 图2C是本发明另一实施例提供的分布式存储复制系统结构示意图;

[0040] 图3是本发明实施例提供的集群视图示意图;

[0041] 图4是本发明实施例提供的OSD视图的状态变迁示意图;

[0042] 图5是本发明另一实施例提供的分布式存储复制系统结构示意图;

[0043] 图6是本发明实施例提供的视图初始化流程图;

[0044] 图7是本发明实施例提供的IO请求处理的流程图;

[0045] 图8是本发明实施例提供的OSD故障处理流程图;

[0046] 图9是本发明实施例提供的OSD故障恢复处理流程图;

[0047] 图10是本发明实施例提供的OSD故障恢复处理过程中数据恢复流程图;

[0048] 图11是本发明实施例提供的OSD退出集群后的处理流程图;

[0049] 图12是本发明实施例提供的OSD扩容后的处理流程图。

具体实施方式

[0050] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,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,都属于本发明保护的范围。

[0051] 如图2A所示,本发明的一个具体实施例提供了一种分布式存储复制控制系统用于

管理控制本发明实施例中提及的数据存储和复制。该分布式存储复制控制系统主要包括状态层、接口层以及数据层三个子层，其中状态层包括元数据控制Meta Data Controller(MDC)模块202，实际应用中可以根据需要确定是否需要配置MDC的备MDC用于在主MDC模块发生故障时承担主MDC的角色以及备MDC数量；接口层包括多个IO路由模块(Input/Output输入/输出路由模块)204(也可以称为client，本发明实施中该两个概念可以互用)；数据层包括多个对象存储设备Object Storage Device(OSD)节点206。其中所述的状态层通过状态视图消息与所述接口层和所述的数据层进行通信，例如MDC模块202通过状态视图消息发送更新通知给所述IO路由模块204以及OSD节点206告知更新本地的集群视图(也可以称为视图，本发明实施中该两个概念可以互用)，或者直接发送MDC模块202生成或者更新的集群视图给所述IO路由模块204以及OSD节点206。接口层和数据层之间通过业务消息进行通信，例如IO路由模块204向OSD节点206发送IO请求消息以请求IO数据的存储和复制。

[0052] 其中，所述MDC模块202作为集群配置信息下发的入口，用于将应用存储空间中的逻辑存储资源的逻辑分区(partition)分配给各OSD节点，根据该partition生成集群视图，维护并更新所述集群视图，并向相应的IO路由模块204和OSD节点206通知所述集群视图的更新。

[0053] 所述IO路由模块204用于根据所述集群视图将上层应用的IO请求路由转发至相应的ODS节点。

[0054] 所述OSD节点206用于根据所述的集群视图对所述IO请求执行相关IO操作，主要包括执行数据的存储和复制，实现数据的一致性备份，以及组织对其管理的物理存储资源(例如本地磁盘或者外部存储资源)的数据操作。

[0055] 可以理解的是，上述MDC模块，IO路由模块以及OSD节点可以由硬件、固件、软件或者他们的组合予以实施，实际应用中，具体用什么方式实现是根据产品设计要求或者制造成本考虑所确定，本发明不应被某种特定的实现方式限定。

[0056] 在本发明的一个具体实施例中，所述的分布式存储复制系统可以统一部署在一个独立的平台或者服务器上(如上述图2B中的分布式存储复制控制平台)用于管理与该平台或者服务器连接的分布式存储系统的数据复制和存储。

[0057] 在本发明的另一个具体实施例中，该分布式存储复制控制系统可以分布式部署在如图2C所示的分布式存储系统中，所述分布式存储系统包括多个服务器server或者主机，本实施例中的主机或服务器为物理意义上的主机或服务器，即包括处理器以及存储器等硬件。上述MDC模块202可以只部署在所述分布式存储系统中的一个(没有备份MDC)、两个(一主一备)或三个(一主两备)服务器或主机上；IO路由模块204部署在所述分布式存储系统中的每个服务器或主机上；OSD节点206部署在所述分布式存储系统中的每个具有存储资源的服务器或主机上，用于管理和控制本地存储资源或外部存储资源。实际应用中，也可以在一台主机上可以只部署IO路由模块，或者只部署OSD节点，或二者都部署，具体的部署方式可以根据实际具体的情况确定，本发明不做限制。图2C中的MDC模块202、IO路由模块204以及OSD节点206构成分布式存储控制系统在图2B所示的分布式存储系统中称为分布式复制协议层，所述分布式存储系统通过该分布式复制协议层控制IO数据存储和复制到所述存储层中的存储资源。所述存储层由所述多个服务器或主机上的本地存储资源构成，所述分布式复制协议层分布在服务器或主机中的各模块通过所述网络层的数据交换网络进行交互，一

个具体的实施方式中可以采用以太网或者infiniband，应该理解上述太网或者infiniband只是本发明实施例采用的高速数据交换网络的示例性实现方式，本发明实施例并不对此进行限定。

[0058] 以下通过具体实施例以及实施方式详细说明上述分布式存储复制控制系统中所述MDC模块202、IO路由模块204，以及OSD节点206相互之间的连接交互以及具体功能等情况。

[0059] 在本发明的一个具体实施例中，所述MDC模块的分区功能具体可以包括：所述MDC模块根据各OSD节点所管理的物理存储资源情况在各OSD节点上配置与所述各OSD节点所管理的物理存储资源对应的逻辑分区(partition)。其中，该partition是由应用存储空间中的一定数量的数据block组成的。其中该应用层的应用存储空间是相对于存储层的物理存储空间而言的，是在应用层为用户分配的一定数量的逻辑存储空间，是存储层的物理层出空间的逻辑映射。即，此处的partition的概念区别于物理存储空间分区的概念，在数据进行存储时，可以把应用存储空间上的一个partition的空间映射到物理存储空间的一个或者多个分区，该partition具体粒度可以从集群配置信息中获取，也可以由该MDC模块根据一定的规则确定，以及其他方式确定，本发明对此不作限制。

[0060] 具体实施方式中，所述MDC模块可以根分区大小配置信息，本地存储资源情况，外部存储资源情况(例如正常接入的SAN(Storage Area Network，存储区域网络)的LUN(Logical Unit Number，逻辑单元数)信息等信息生成所述partition的集群视图。

[0061] 通常为了保证数据的可靠性和可用性，一个partition在不同的OSD节点上存储有副本，partition的副本数目可以通过配置文件进行配置，也可以由MDC根据一定的算法确定，partition分为主partition和备partition，在partition的多个副本中，选取其中一个副本为主副本，称之为主partition，一个partition主副本之外的副本，称之为备partition，主partition和该主partition对应的备partition构成一个partition群，主partition所在的OSD节点，称为该主partition所在partition群的主OSD节点，本实施例中描述的主OSD都是指某一个partition群的主OSD，备partition所在的OSD节点，称之为该备partition所在的partition群的备OSD节点，本实施例中描述的备OSD都是指某一个partition群的备OSD。

[0062] 为了便于理解，下面结合图2B提供的实施例进行进一步说明，如图2B所示，主机或者服务器(本发明实施例中主机和服务器的概念可以互换)server_1上的OSD中的所管理的存储资源被划分为partition 1，partition 2和partition 3(简称P1，P2和P3)，以及partition 4’，partition 5’，和partition 6’(简称P4’，P5’和P6’)。其中，P4’，P5’和P6’分别为服务器server_2上的OSD节点上的partition4，partition 5和partition 6(简称P4，P5和P6)的副本。server_1上的OSD中的partition与存储层中的物理存储资源之间具有对应的映射关系，例如，该OSD上的一个partition的空间映射到物理存储空间的一个或者多个Block。

[0063] 主机或者服务器server_1中的OSD管理主partition(P1，P2和P3)以及备partition(P4’，P5’和P6’)，该ODS分别为P1和P1’组成的partition群主OSD节点，P2和P2’组成的partition群主OSD节点以及P3和P3’组成的partition群的主OSD节点，同时该OSD分别为P4和P4’组成的partition的备OSD节点，P5和P5’组成的partition的备OSD节点，以及

P6和P6'组成的partition的备OSD节点。可以看出，同一OSD节点，针对不同的partition群，可以既作为主OSD节点同时又是作为备OSD节点而存在。

[0064] 上述分区partition以及对应副本的设置可以根据基于下述因素进行，具体实际应用中可以加入其他的考虑因素设置规划磁盘分区。

[0065] 其一，数据安全性，每个partition的副本应尽量分布到不同主机或服务器中。数据安全的底线是，某个partition的多个副本不允许放在同一个主机或服务器中。其二，数据平衡性，每个OSD上的partition数量尽量相同，每个OSD上的主patition、备patition 1、备partition 2数量尽量相同。这样，每个OSD上处理的业务是均衡的，不会出现热点。其三，数据分散性，每个OSD上的partition，其副本应尽量平均地分布到其他不同的OSD上，对于更高级别的物理部件也有同样的要求。

[0066] 如图4所示，本发明一个具体的实施例中，所述MDC模块生成集群视图信息具体可以包括MDC根据管理员下发的集群配置信息，以及所述分区情况生成集群视图信息。具体而言，该集群视图信息包括三个维度的集群视图，OSD视图(OSD view)、IO视图(IO view)以及Partition视图(partition view)。

[0067] 所述OSD view包括在集群中的OSD节点的状态信息。具体实施方式中所述OSD视图可以包括OSD节点的ID以及OSD节点的状态信息。其中OSD ID为OSD标示或者编号，如图5所示的本发明的一个实施例中，OSD状态具体可以包括根据所述OSD是故障分为“有效(UP)”状态和“失效(DOWN)”状态，以及根据所述OSD是否退出集群分为“退出集群(OUT)”和“在集群中(IN)”状态。如图5所示，具体的状态变迁包括，当某个OSD节点于故障回复后经过初始化或者重启后由“在集群中(IN)”以及“不失效(DOWN)”状态变更为“在集群中(IN)”以及“有效(UP)”。当某个OSD故障发生超过一定阈值(例如超过5分钟时间)，则所述OSD节点将被提出集群，相应的其状态由“在集群中(IN)”以及“失效(DOWN)”状态变更为“退出集群(OUT)”以及“失效(DOWN)”状态。本发明的具体实施例中，所述OSD视图还可以进一步包括OSD视图版本信息，例如OSD视图版本号、OSD视图ID或者其他标记视图版本的任何信息。

[0068] 所述IO view包括标识一个partition群的主OSD节点的标识。具体实施方式中，所述IO view可以包括partition群ID以及所述partition群ID所对应的partition群的主OSD节点的标识。其中每一个IO view有标识该IO view的IO视图版本信息，该IO视图版本信息可以是IO view ID(也可以称为IO视图版本号)，用于标识IO view的版本，方便不同模块进行IO view版本新旧的对比。具体实施例中，该IO视图版本信息可以包括在IO视图中，也可以在IO视图之外。

[0069] partition view包括一个partition群中的各partition所在的OSD信息。具体实施方式中，所述partition view可以包括partition群ID，所述partition群ID对应的partition群中各partition所在的OSD以及所述OSD的主备身份，以及对应所述各partition的所述OSD的partition状态。该partition view中包括主partition所在的OSD节点信息(例如OSD节点ID、该OSD节点的主备身份以及对应所述主partition的该OSD节点的partition状态)、该主partition对应的备partition(可以为一个或者多个)所在的OSD节点信息(例如OSD节点ID、所述OSD节点的主备身份以及所述OSD对应所述备partition的partition状态)。具体实施例中，所述partition状态可以分为“一致”和“不一致”两种，“一致”表示备partition的数据与主partition是一致的，“不一致”表示备partition的数据可

能与主partition不一致;每一个partition view有标识该partition view的partition视图版本信息,该partition视图版本信息可以是partition view ID(也可以称为partition视图版本号),用于模块之间比较view的新旧。具体实施例中,该partition视图版本信息可以包括在partition视图中,也可以在partition视图之外。具体实施例中,由于IO视图是partition视图的一个子集,即partition视图包括了IO视图信息,该partition视图可以进一步包括IO视图版本信息。

[0070] 所述MDC进一步用于维护管理和更新所述集群视图,根据OSD节点的状态、如故障、故障恢复、故障后退出集群、和故障恢复后恢复加入集群、新加入集群等进行集群视图的更新并通知到相关模块,以便相关模块根据该更新的集群视图处理相应的IO请求所对应的数据的复制。

[0071] 在具体的实施方式中,为了减少交互以及节省管理和存储资源,所述OSDview可以只存在于MDC中,所述partition view可以只存在于MDC模块和主OSD节点中,所述IO view存在所述MDC模块、IO路由模块、主OSD节点以及备OSD节点中。所述MDC模块仅将所述partition view发给所述partition view中的partition所在的主OSD节点,或者仅通知所述partition view中的partition所在的主OSD节点进行本地partition view更新;而将构成partition view一部分的IO view(即IO view可以看成是partition view的子视图)发给IO路由模块、主OSD节点以及备OSD节点中,或者通知相应的模块进行本地保存的IO视图的更新。具体的实施过程可以参考下述各具体流程,以及OSD加入集群流程。实际应用中,MDC模块可以根据配置信息或者一定的策略根据集群视图的基本功能设置不同形式的集群视图,本发明实施例不做限制。

[0072] 本发明的一个具体实施例中,所述IO路由模块主要用于实现IO请求路由的功能。所述IO路由模块从所述MDC模块获取集群中所有partition的IO view并进行缓存,当业务IO请求到达所述IO路由模块时,IO路由模块通过IO请求中的key计算出IO所属的partition群(计算法可以采用哈希算法或者其他算法);然后查找本地保存的IO view,找到partition群对应的主OSD节点,把IO请求发送给该主OSD节点。处理从所述MDC模块接收的IO view更新通知,所述的更新通知可以包括所述的更新后的IO view或者相应的更新知识信息,例如指示更新哪些内容,根据更新通知更新本地保存的IO view,并根据本地保存的更新后的IO view进行IO请求的路由。具体的实施过程可以参考下述各具体流程。

[0073] 本发明的一个具体实施例中,所述OSD节点根据所述集群视图处理IO请求执行IO操作具体包括:

[0074] 当所述OSD节点作为主OSD节点时,所述的主OSD节点主要用于接收所述IO路由模块发送的IO请求,执行IO请求以及向相应的备OSD节点发送复制请求,以执行IO数据的存储和复制。所述主OSD节点从所述MDC模块接收其作为主OSD节点上的partition的partition view,并进行保存,所述主OSD节点根据所述partition view对所述IO请求的复制进行处理。所述主OSD节点进一步接收MDC模块关于partition view的更新通知,根据所述更新通知更新本地保存的partition view,并根据更新后的partition view处理所述IO请求所对应的数据的复制,所述更新通知可以包括更新的partition view或者相应的更新信息以便所述OSD节点根据所述更新的partition view或者所述更新信息更新本地保存的partition view和IO view。当所述OSD节点作为备OSD节点时,所述备OSD节点用于接收主

OSD节点的复制请求，根据所述复制请求进行数据的复制备份，从所述MDC模块接收其该备OSD节点上该数据所属的partition的IO view，并进行保存，根据IO view处理所述的IO请求所对应的数据的复制，并进一步接收MDC模块关于IO view的更新通知，根据所述更新通知更新本地保存的IO view，以及根据所述更新的IO view处理所述IO请求所对应的数据的复制。具体的实施过程可以参考下述各具体流程。

[0075] 本发明的一个具体实施例中，上述实施例中的分布式存储复制系统(如图2A、2B以及2C)可以基于如图5所示的系统实施，如图5所述，该系统可以包一个或者多个括存储器502、一个或者多个通信接口504，以及一个或者多个处理器506，或者其他数据交互网络(用于多个处理器以及存储器之间的交互，图中对此未示出)。

[0076] 其中，存储器502可以是只读存储器(Read Only Memory，ROM)，静态存储设备，动态存储设备或者随机存取存储器(Random Access Memory，RAM)等各种类型的存储器。存储器502可以存储操作系统和其他应用程序的指令以及应用数据，所述指令包括用于执行本发明各实施例中的MDC模块、IO路由模块以及OSD节点的功能的指令。

[0077] 存储器502中存储的指令由处理器506来运行执行。

[0078] 通信接口504用来实现存储器502与处理器506之间的通信，以及处理器之间和存储器之间的通信，以及该系统与其他设备或通信网络之间的通信。

[0079] 处理器506可以采用通用的中央处理器(Central Processing Unit，CPU)，微处理器，应用专用集成电路(Application Specific Integrated Circuit，ASIC)，或者一个或多个集成电路，用于执行相关程序，以执行本发明各实施例中所描述的MDC模块、IO路由模块以及OSD节点之间的交互流程以及实现的功能。

[0080] 为了便于理解同时避免不必要的重复描述，以下具体的实施例用于说明图5所示的系统具体如何执行本发明各实施例中所描述的MDC模块、IO路由模块以及OSD节点之间的交互流程以及实现的功能。本领域技术人员在阅读本发明所有实施例的基础上，可以理解图5所示的系统可以用于实施其他各实施例中描述的情况，这些都在本发明的记载以及公开的范围内。

[0081] 实施例一：

[0082] 所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，所述指令包括用于执行所述MDC模块、IO路由模块以及OSD节点的功能的指令，根据所述的指令促使所述处理器执行下述操作：

[0083] 促使所述MDC模块确定所述系统中某个OSD节点为故障OSD节点，确定所述故障OSD节点上的partition，更新所述故障OSD节点上的partition所在的partition群的partition视图，向所述更新的partition视图中的partition群所在的主OSD节点发送更新通知，以便所述主OSD节点根据所述更新的partition视图处理IO请求所对应的数据的复制。

[0084] 需要说明的是，上述实施例中，上述MDC模块、IO路由模块以及OSD节点功能可以由一个主机实现，在这种情况下，实现上述MDC模块、IO路由模块以及OSD节点功能的指令可以存在于该主机的存储器中，由该主机的处理器读取该存储器中实现上述MDC模块、IO路由模块以及OSD节点功能的指令。在另外的实施例中，上述MDC模块、IO路由模块以及OSD节点功能可以由多个主机交互实现，在这种情况下，上述MDC模块、IO路由模块以及OSD节点分布式

存储于不同主机的存储器中，例如，主机1中的处理器执行上述MDC模块的功能，主机2的处理器执行主OSD节点的功能，主机3执行IO路由模块功能。

[0085] 实施例二：

[0086] 所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，所述指令包括用于执行所述MDC模块、IO路由模块以及OSD节点的功能的指令，根据所述的指令促使所述处理器执行下述操作：

[0087] 促使所述IO路由模块接收IO请求，所述IO请求包括key，根据所述key确定所述IO请求所对应的数据所属的partition群并确定所述数据所属的partition群的主OSD节点，将所述数据所属的partition群的IO视图的IO视图版本信息加入所述IO请求，并将携带有所述IO视图版本信息的所述IO请求发送至所述确定出的主OSD节点；

[0088] 促使所述主OSD节点接收所述IO请求，在根据所述IO视图版本信息确定所述IO请求中的IO视图版本与本地保存的IO视图版本一致后，执行所述IO请求，生成携带有所述IO视图版本信息的复制请求，并将所述的复制请求发送至所述数据所属的partition的备OSD节点；以及

[0089] 促使所述备OSD节点接收所述复制请求，并在根据所述IO视图版本信息确定所述复制请求中的IO视图版本与所述备OSD节点本地保存的IO视图版本一致后，执行所述复制请求以使得所述备OSD节点上的所述IO请求所对应的数据与所述主OSD节点上的所述IO请求所对应的数据保持一致。

[0090] 与上面一样，上述实施例中，上述MDC模块、IO路由模块以及OSD节点功能可以由一个主机实现，在这种情况下，实现上述MDC模块、IO路由模块以及OSD节点功能的指令可以存在于该主机的存储器中，由该主机的处理器读取该存储器中实现上述MDC模块、IO路由模块以及OSD节点功能的指令。在另外的实施例中，上述MDC模块、IO路由模块以及OSD节点功能可以由多个主机交互实现，在这种情况下，上述MDC模块、IO路由模块以及OSD节点分布式存储于不同主机的存储器中，例如，主机1中的处理器执行上述IO路由模块的功能，主机2的处理器执行主OSD节点的功能，主机3处理器执行备OSD节点的功能，主机4处理器执行MDC模块的功能。

[0091] 实施例三：

[0092] 所述的处理器用于与所述的存储器连接，读取所述存储器中的指令，所述指令包括用于执行所述MDC模块、IO路由模块以及OSD节点的功能的指令，根据所述的指令促使所述处理器执行下述操作：

[0093] 促使所述OSD节点在自身故障恢复后向所述MDC模块发送查询请求以请求所述OSD节点上的partition所在的partition群的IO视图，所述OSD节点称为故障恢复OSD节点，所述查询请求中携带所述故障恢复OSD节点的OSD标识，接收所述MDC返回的所述IO视图，向所述IO视图中的的主OSD发起数据恢复请求以请求恢复所述故障恢复OSD节点在故障期间更新的数据，接收所述主OSD发送的所述故障期间更新的数据，根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理IO请求的复制；

[0094] 促使所述MDC模块接收所述故障恢复OSD节点的所述查询请求，根据查询请求中的OSD标识将所述IO视图返回给所述故障恢复OSD节点，并在所述故障恢复OSD节点数据恢复

执行完毕后更新所述partition群的partition视图;以及

[0095] 促使所述主OSD节点接收所述故障恢复OSD节点的数据恢复请求,将所述故障期间更新的数据发送至所述的故障恢复OSD节点,以及根据所述MDC模块在所述故障恢复OSD节点数据恢复执行完毕后更新的所述partition群的partition视图处理所述IO请求所对应的数据的复制。

[0096] 同理,上述实施例中,上述MDC模块、IO路由模块以及OSD节点功能可以由一个主机实现,在这种情况下,实现上述MDC模块、IO路由模块以及OSD节点功能的指令可以存在于该主机的存储器中,由该主机的处理器读取该存储器中实现上述MDC模块、IO路由模块以及OSD节点功能的指令。在另外的实施例中,上述MDC模块、IO路由模块以及OSD节点功能可以由多个主机交互实现,在这种情况下,上述MDC模块、IO路由模块以及OSD节点分布式存储于不同主机的存储器中,例如,主机1中的处理器执行上述故障恢复OSD节点的功能,主机2的处理器执行主OSD节点的功能,主机3处理器执行所述MDC模块的功能,主机4处理器执行IO路由模块的功能。

[0097] 下面通过多个具体流程实施例进一步详细说明上述分布式存储复制控制系统中所述MDC模块202、IO路由模块204以及OSD节点206相互之间的连接交互以及具体功能等情况。这些具体的流程实施例包括:集群视图初始化生成和获取流程、IO请求处理流程、OSD故障的处理流程、OSD节点故障恢复流程、数据恢复流程、OSD节点故障后退出集群的流程以及OSD节点扩容流程。以下一一进行详细说明。

[0098] 需要说明的是,以下实施例中各步骤或功能并不都是必须的,各步骤之间的顺序仅仅是为了描述的方便,除了根据本发明方案的原理必然要求的顺序之外,各步骤之间的顺序没有特别限定,而且各步骤中具体的实现或实施方式仅仅是举例说明,并不对本发明的保护范围构成特定限定。本领域技术人员在阅读本发明整个说明书的基础上,可以根据实际情况对上述各步骤进行相应的增减或者进行非创造性的变更或者替换,将各步骤中不同的实施方式与其他步骤的实施方式进行组合形成不同的实施例等,都属于本发明记载以及公开的范围内。

[0099] 集群视图初始化生成和获取流程

[0100] 如图6所示,为本发明提供的一个集群视图初始化生成和获取的实施例,在该实施例中MDC模块根据管理员下发的集群配置信息,生成初始的集群视图。IO路由模块和OSD节点初始化时,向所述MDC模块查询视图。本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的MDC模块,IO路由模块以及OSD节点。

[0101] 系统初始化时,先启动所述MDC模块,然后所述IO路由模块和OSD节点从所述MDC获取视图启动,具体过程包括:

[0102] 602:用户或者管理员下发集群配置信息至所述MDC模块。所述集群配置信息可以包括集群的拓扑信息以及partition数目、副本数目等系统配置参数,其中所述集群拓扑信息主要包括服务器的数目及其IP地址、机架信息、每个服务器上的OSD节点数目以及其所管理的物理存储资源信息(例如其对应的本地磁盘信息)等。

[0103] 604:所述MDC根据下发的所述集群配置生成初始的集群视图,上文已经介绍了集群的三种视图(OSD view、partition view、IO view);根据配置的OSD信息生成OSD view,通过partition分配算法以及配置的partition数目、副本数目、OSD节点的数目生成

partition view,IO view是partition view的一个子集,不需要额外去生成。在生成partition视图时,通常需要考虑partition分配的均衡性(每个OSD节点上的partition数目尽量一致),安全性(partition的副本所在的OSD节点位于不同的服务器或者不同的机架)。

[0104] 606:IO路由模块初始化启动,向所述MDC模块查询IO view,所述IO路由模块启动时需要从MDC获取到相关的视图才能正常工作。

[0105] 608:OSD节点初始化启动,向所述MDC模块查询partition view和IO view,所述OSD节点需要获取分布在该OSD节点上的主partition所在的partition群的的partition view和分布在该OSD节点上的备partition所在的partition群的IO view。

[0106] 610:所述MDC模块向所述IO路由模块返回所有partition群的IO view。

[0107] 612:所述MDC模块向所述OSD节点返回分布在该OSD上主partition所在的partition群的的partition view和备partition所在的partition群的的IO view。

[0108] IO请求处理流程

[0109] 如图7所示,为本发明提供的IO请求流程的实施例,本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的IO路由模块和OSD节点。为了理解的方便本实施例以某个partition群(Partition X)为例说明,Partition X可以是本发明实施例中的分布式复制协议层的管理维护的任一分区Partition,Partition X所在的主OSD节点(简称Partition X主OSD节点),Partition X的备partition所在的备OSD1节点(简称Partition X备OSD1节点),Partition X的备partition所在的备OSD2节点(简称Partition X备OSD2节点),后续具体描述中将以partition X为例描述。该实施例中的IO操作流程(例如写操作或修改操作)具体包括:

[0110] 702:所述IO路由模块接收主机(例如,图3中所示的该IO路由模块所在的server)发送的IO请求。

[0111] 704:所述IO路由模块根据该接收的IO请求获得该IO请求所对应的数据(也可以称为IO数据)的partition,并获取该partition所在的partition群的主OSD节点。

[0112] 在具体实施方式中,IO路由模块可以根据该IO请求中携带的Key,通过哈希算法计算出该IO请求所对应的数据所在的partition群的partition ID,然后通过该partition ID查找IO view,获取该partition群的主OSD节点。如上所述,本实施例中此处partition ID所对应的partition群为partition X。所述的Key是上层业务定义的一个数字或者字符串,用于标识一个数据块。

[0113] 706:所述IO路由模块将该IO请求发送给所述partition群的主OSD节点,所述请求中携带IO视图版本信息(例如IO view ID),IO关键信息以及该IO数据。

[0114] 根据前述描述本实施例中的partition群为partition X,相应的,该IO请求被发送给所述partition X的主OSD节点,即partition X主OSD节点。在具体实施方式中,所述IO view ID也可以称为IO视图版本号,其主要用来标识一个视图的版本,单调递增,IO view ID小说明其维护的视图版本是一个过时的版本,保证一致性的一个要求是IO处理流程中所有模块看到的视图是一致的。所述的IO关键信息可以包括key、offset以及length,其中offset是指IO数据在key所标识的数据块中相对起始位置的偏移量,length是指IO数据的长度大小。

[0115] 708:根据所述IO视图版本信息确定所述IO请求中携带的视图版本与本地保存的视图版本一致后,为所述IO请求生成一个sequence(Seq)ID(也可以称为序列标识)。具体而言,所述主OSD节点确定所述IO请求中携带的IO view ID与本地保存的IO view ID一致后,为所述的IO请求生成一个序列标识。

[0116] 在具体实施方式中,所述的Seq ID由所述视图版本号以及序列号sequence number(Seq NO)组成。其中视图版本号随着所述IO view的变更单调递增,所述Seq NO表示在一个IO view版本内,针对该IO视图中的partition对应的数据的修改操作(例如写和删除)的连续编号。IO view变更后,所述Seq ID中的Seq NO重新从0开始递增。其中,所述Partition X主OSD节点比较所述IO请求中携带的IO view ID与本地保存的IO view ID是否一致具体可以为先比较其中的IO view ID,IO view ID大则认为其seqID大,如果IO view ID相等,再比较其中的seq NO,seq NO大的seq ID大,只有两者都相同则表示所述的Seq ID一致。

[0117] 710:如果本地IO view大,则向所述IO路由模块返回错误,如果本地的IO view小,则将所述IO请求加入缓冲队列,向所述MDC模块查询partition view。

[0118] 具体实施方式中,所述Partition X主OSD节点在确定所述IO请求中的IO视图ID小于本地保存的IO视图ID后向所述的IO路由模块返回错误,由所述IO路由模块向所述MDC模块查询所述partition群的IO视图,在获得更新的IO视图ID后重新发送所述IO请求;以及在确定所述IO请求中的IO视图ID大于本地保存的IO视图ID后将所述IO请求加入缓冲队列,并向所述MDC模块查询所述partition群的IO视图的IO视图ID以确定本地保存的IO视图ID与所述IO请求中的IO视图ID一致后执行所述IO请求。

[0119] 712:记录一条entry,所述的entry包括操作类型,Partition群ID,Seq ID,以及key。

[0120] 具体实施方式中,所述的操作类型可以包括写或者删除等。如果是写操作,则所述的entry可以进一步包括前述的offset以及length。另外,针对各种类型的操作,所述的entry中还可以进一步包括状态信息,用于描述所述的操作是否成功。一般而言,针对同一个partition群的修改操作(如写和删除操作)进行统一连续编号的。

[0121] 714:执行所述的IO请求,查询所述partition群的partition view获得所述partition群备OSD节点信息。

[0122] 具体实施方式中,如果所述执行所述IO请求为写IO请求,则将所述IO数据写入所述Partition X主OSD节点所管理的本地物理存储资源(例如图3中所示的缓存层或者持久化层,如磁盘,或者前面提及的外部物理存储资源SAN);如果所述IO请求为删除请求,则将所述Partition X主OSD节点所管理的本地物理存储资源上的相应数据删除。具体实施例中,所述Partition X主OSD节点进一步的生成复制请求。具体实现中,所述的复制请求可以为将IO请求的控制部分另外组装以生成复制请求,从而使得Partition X的备OSD节点上的所述IO请求所对应的数据与所述Partition X主OSD节点上的所述IO请求所对应的数据保持一致。

[0123] 716(1)/(2):分别向所述partition X备OSD1节点以及所述partitionX备OSD2节点发送复制请求,所述复制请求携带Seq ID。

[0124] 具体实施例方式中,所述的复制请求可以进一步包括原始请求中的key、offset、

length以及IO数据等信息。当所述复制请求为写入复制请求时，所述复制请求中携带有key、offset、length以及IO数据。当所述的复制请求为删除复制请求，所述复制请求中仅携带有所述的key。

[0125] 718:确定所述复制请求中的IO view ID与本地保存的IO view ID一致后，处理所述的复制请求。

[0126] 在具体的实施方式中，在上述718步骤之外可以进一步包括以下步骤：

[0127] 如果所述Partition X备OSD1节点或备OSD2在确定所述复制请求中的IOview ID小于本地保存的IOview ID后向所述的Partition X主OSD节点返回错误，由所述Partition X主OSD节点向所述MDC模块查询所述partition群的IO视图，在获得更新的IO视图ID后重新发送所述复制请求；或者在确定所述复制请求中的IO视图ID大于本地保存的IO视图ID后将所述复制请求加入缓冲队列，并向所述MDC模块查询所述partition群的IO视图版本号以确定本地保存的IO视图ID与所述IO请求中的IO视图ID一致后执行所述复制请求。

[0128] 相比现有两阶段提交协议中，如果参与者拒绝了提议，整个IO流程需要进行回退，带来了很大的开销，而本实施例中如果备OSD节点拒绝了请求，而是先查询最新的视图后继续处理，不需要进行回退，从而提高了整个系统的容错能力以及可用性。

[0129] 现有的两阶段提交协议在正常的一次IO过程中，协调者与参与者之间需要两次消息交互，而本实施例中只需要进行一次IO过程，主节点与备节点只有一次消息交互，降低了消息交互带来的IO时延，提高了整个系统的效率和性能。

[0130] 在具体的实施方式中，所述的复制请求进一步可以携带有所述Partition X主OSD节点针对所述partition X发送的上一次复制请求中的Seq ID，所述IO流程可以进一步包括步骤720.

[0131] 720:比较复制请求中携带携带的上一次复制请求的SeqID与本地最大的SeqID进行比较，如果比本地的大，则要求所述Partition X主OSD节点发送丢失的请求，如果一致则继续后续步骤的处理。

[0132] 通过携带上一次复制请求的Seq ID可以防止IO view版本号变更导致的数据遗漏，而且通过发送丢失的数据，可以保证主备OSD节点对每个IO的执行顺序是一致的，从而进一步提高数据备份的一致性。

[0133] 具体实施方式中，如果复制请求中携带携带的上一次复制请求的Seq ID比本地最大的Seq ID小，则向所述partition X主OSD节点返回错误，由所述partition X主OSD节点重发复制请求；或者通过进一步的查询确认Seq ID，并在确认后获取更新seq ID后继续进行处理，而不直接返回错误。两种情况都不需要终止处理，进而进行回退，进一步提高了系统的容错能力和可用性，以及整个系统的性能。

[0134] 当IO在主备OSD节点上都需要执行时，需要保证主备OSD节点对每个IO操作的执行顺序是一致的，这时保证partition的多个副本一致性的重点

[0135] 722、记录一条entry，执行所述的复制操作。

[0136] 具体实施方式中，所述的entry包括操作类型，Partition ID，Seq ID，以及key，所述的操作类型可以包括写或者删除等。如果是写操作，则所述的entry可以进一步包括前述的offset以及length。另外，针对各种类型的操作，所述的entry中还可以进一步包括状态信息，用于描述所述的操作是否成功。执行所述的复制请求具体包括：当所述复制请求为写

入复制请求时，所述复制请求中携带有key、offset、length以及IO数据。当所述的复制请求为删除复制请求，所述复制请求中仅携带有所述的key。

[0137] 724：所述partition X备OSD1节点和partition X备OSD2节点分别向所述partition X主OSD节点发送响应请求成功消息。

[0138] 726：所述partition X主OSD节点向所述的IO路由模块发送响应请求成功消息。

[0139] 在上述IO请求实施例的基础上，假设在所述IO请求处理过程中，所述的partition X主OSD节点或者partition X备OSD节点发生故障(例如在IO请求达到partition X主OSD节点之后，所述partition X主OSD节点发生故障)，或者所述partition X备OSD节点发生故障，或者系统有新的OSD节点加入作为所述partition X的备OSD节点等，在这些情况下，前述实施例中的IO请求处理流程进一步可以包括以下实施例中所描述的处理过程。

[0140] 在IO请求达到partition X主OSD节点之后，所述partition X主OSD节点发生故障后IO处理流程包括如下：

[0141] 当所述partition X主OSD节点发生故障后，系统中的MDC模块在所述IO请求处理过程中检测到所述partition X主OSD节点发生故障后将所述partition群的partition view中的所述partition X主OSD节点降为新的partition X备OSD节点，并将所述新的partition X备OSD节点的partition状态标记为不一致，将所述partition群的partition view中的所述partition X备OSD1节点升为新的partition X主OSD节点，将更新后的所述partition群的partition view发送至所述新的partition X主OSD节点，将所述partition群的IO view中的所述partition X备OSD1节点升为新的partition X主OSD节点，将更新后的所述partition群的IO视图发送至所述IO路由模块。

[0142] 所述IO路由模块接收所述MDC模块发送的所述更新后的所述partition群的IO视图，根据所述更新后的所述partition群的IO视图将所述IO请求发送至所述新的partition X主OSD节点；以及

[0143] 所述新的partition X主OSD节点用于接收所述IO请求，在执行所述IO请求后生成复制请求，将所述新的复制请求发送至所述更新后的所述partition的partition视图中的其他备OSD节点。这里生成复制请求与发送复制请求的步骤与上述步骤714和716步骤一样。

[0144] 在IO请求达到partition X主OSD节点之后，所述partition X备OSD节点发生故障后IO处理流程包括如下：

[0145] 当所述partition X备OSD节点发生故障后，所述MDC模块进一步用于在所述IO请求处理过程中检测到所述partition X备OSD节点发生故障时，将所述partition视图中的所述partition X备OSD节点的partition状态标记为不一致，将更新后的所述partition群的partition视图发送至所述partition X主OSD节点；以及

[0146] 所述主OSD节点用于接收到所述更新后的所述partition群的partition视图后，将所述复制请求发送至所述更新后的partition视图中的partition状态为一致的其他备OSD节点，不将所述复制请求发送至所述partition状态为不一致的partition X备OSD节点。

[0147] 在两阶段提交协议中，协调者发生故障，IO处理中断，需要协调者恢复后才能继续处理，在本实施例中，主OSD节点发生故障后，MDC节点能够快速选举新的主OSD节点，快速的恢复IO的处理，可用性高，容错能力强。

[0148] 另外，在两阶段端提交协议中，如果参与者发生故障或者一直没有响应，其他的IO请求一直阻塞，最后超时失败，需要进行回退；本实施例中如果备节点发生故障后，MDC会通知主节点进行视图变更，隔离或忽略该故障OSD节点，继续IO请求的处理，不会阻塞其他的IO请求处理，具有更好的容错能力，能够快速处理节点故障和进行节点故障恢复，例如，N+1个副本可以容忍N个副本故障，进而提高了存储系统的性能和可用性。可用性差的系统决定了其扩展性不会太好，因为在大规模分布式存储系统中，存储节点的故障是一件经常发生的事情，复杂且大量的协议交互也会使得系统的扩展能力减低。

[0149] 此外，可以针对partition粒度的集群视图控制，可以极大的缩小存储节点故障的影响范围，使得存储系统能够进行大规模的扩展，提高了系统的扩展性。

[0150] 在IO请求达到partition X主OSD节点之后，所述MDC发现有新的OSD节点加入集群，并且作为所述partition X的备OSD节点，所述IO处理流程包括如下：

[0151] 所述MDC模块在所述IO请求处理过程中确定有新的OSD节点加入集群时，通知所述partition X主OSD节点所述新的OSD节点作为所述partition X所在的新的备OSD节点，并在partition的数据同步完毕后更新所述partition群的partition视图以及IO视图，通知所述的partition X主OSD节点更新所述partition X主OSD节点本地保存的partition视图；以及

[0152] 所述partition X主OSD节点将其上面的主partition的数据同步至所述新的备OSD节点，并根据更新后的本地保存的partition将所述复制请求发送至新的备OSD节点。

[0153] OSD故障的处理流程

[0154] 在大规模的分布式存储系统中，存储节点的故障是一种常态。当出现部分OSD节点故障时，系统需要能够正常提供IO服务。在本发明实施例中，所有的IO请求的处理都依赖于MDC模块维护的集群视图，当集群中的OSD节点发生故障，集群视图也需要进行相应的更新，以便于IO请求的处理能够正常有效的进行。

[0155] 为不让所述的OSD节点的故障影响IO请求的复制的正常处理，通常通常需要进行以下处理：第一，MDC模块对OSD节点的状态进行检测，当某个OSD节点发生了故障，MDC模块能够及时发现该故障；第二，MDC模块发现某OSD故障后，需要及时进行处理以对视图做出正确的变更，并通知到相关的IO路由模块和OSD节点；第三，相关的IO路由模块和OSD节点根据所述MDC的更新通知后根据更新后的视图处理相应的IO请求，以便所述的模块和节点能够及时获得更新的视图，从而保证IO请求顺利高效的处理。

[0156] MDC模块可以通过下述两种模式对OSD节点的故障进行检测：1)由MDC模块统一负责，每个OSD节点都定时向MDC模块发送心跳消息，如果某个OSD节点在规定的时间周期内没有向MDC模块发送心跳消息，则MDC模块判定该OSD节点发生了故障；2)OSD节点之间定时互相发送心跳消息进行故障检测，检测者在规定的时间周期内没有收到被检测者的心跳消息，则上报MDC模块对应的OSD节点发生了故障。

[0157] 如前所述，partition view和IO view都是针对某个partition群来说的，一般情况下，一个OSD节点上有多个partition；一个OSD节点故障，涉及到多个partition所在的多个partition群的视图的更新，每个partition群的视图更新是相互独立的：

[0158] 1)当故障的OSD节点上的partition包括备partition时，将所述备partition所在的partition群的partition视图中所述故障OSD节点的partition状态标记为不一致，同时

把更新后的partition视图通知给所述partition群的主OSD节点，所述partition的主OSD节点根据更新的partition视图进行IO请求所对应的数据的复制处理；以及

[0159] 2)当故障OSD节点上的partition包括主partition时，将所述主partition所在的partition群的partition视图中作为主OSD节点的所述故障OSD节点降为新的备OSD节点并将所述新的备OSD节点对应的partition状态标记为不一致，并从所述主partition所在的partition群的partition视图中的原备OSD节点中选择一个partition状态为一致的备OSD节点升为新的主OSD节点。之后，通知新主的OSD节点partition视图更新，通知其他的备OSD节点IO view更新。若partition群的主OSD故障并且所有的备partition所在的OSD节点的partition状态都处于不一致状态，则不进行partition视图和IO视图的变更，因为需要保证partition的主副本拥有最新的完整数据，从而保证数据复制的一致性。

[0160] 相关的IO路由模块和OSD节点根据所述MDC的更新通知后根据更新后的视图处理相应的IO请求具体可以包括：

[0161] 新主的OSD节点根据所述更新后本地保存的partition view将来自所述IO路由模块的IO请求所对应的数据复制到所述更新后本地保存的partition view中的partition状态为一致的partition所在的备OSD节点上，或复制到所述更新后本地保存的partition view中的partition所在的partition状态为不一致但正在进行数据恢复的备OSD节点上，从而进行故障隔离，保障IO请求的正常不受干扰的处理，从而提高了系统的容错能力，相应的提高了系统的性能和可用性。此外，可以通过对partition粒度的集群视图控制，可以缩小OSD节点故障的影响范围，从而使得系统能够进行大规模的扩展，提高了系统的扩展性。

[0162] 最后，在所述故障OSD节点故障恢复并完成数据恢复后，MDC模块进一步更新所述partition view以及IO view，将所述进一步更新的partitionview通知给所述进一步更新的partition view中的partition所在的主OSD节点，将所述进一步更新的IOview发送给所述进一步更新的partition view中的partition所在的备OSD节点，以便接收到该进一步更新的partition view或者IO view的模块或者OSD节点，更新本地保存的partition view或者IO view，并根据该进一步更新的partition view或者IO view处理IO请求所对应数据的复制。

[0163] 通过及时的更新视图，故障恢复节点可以快速加入集群处理IO请求，提高了系统的性能和效率。

[0164] 为了便于理解，以下通过一个具体的实施例进行说明，如图8所示，为本发明提供的OSD节点故障处理流程实施例，本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的MDC模块，IO路由模块以及OSD节点。为了理解的方便，本实施例以OSD节点中的OSDx节点、OSDy节点以及OSDz节点为例说明，OSDx节点、OSDy节点以及OSDz节点可以是本发明实施例中的分布式复制协议层的中多个OSD节点的任一OSD节点，并且，为了便于理解，本实施例中假设OSDx节点为partition 1群(简称P1)的主OSD节点，为partition n群(简称Pn)的备OSD节点；OSDy节点为所述Pn的主OSD节点，为所述P1的备OSD节点；OSDz节点为所述Pn的备OSD节点，为所述P1的备OSD节点。该实施例中的OSD节点故障处理流程具体包括：

[0165] 802/804/806：OSDx节点、OSDy节点、OSDz节点分别定时向主MDC模块发送心跳消息。

[0166] 808:OSDx节点发生故障。

[0167] 实际应用中,该故障可以包括各种软、硬件或者网络的故障,如软件BUG导致的程序进程重启、网络短暂不通、服务器重启等导致OSD节点不能处理所述的IO请求,不能完成数据的存储和复制功能。

[0168] 810:MDC模块检测到所述OSDx节点在预定的时间内没有发送心跳消息,则判断其发生了故障。

[0169] 实际应用中,如前所述,MDC还可以通过其他方式,比如由没有故障的OSD节点通知某一OSD节点发生故障。

[0170] 812:MDC模块根据故障情况进行视图更新。

[0171] 具体实施方式中,MDC根据确定后的故障OSD节点上的partition所在的partition群,更新相应partition群的集群视图,本实施例中,故障OSDx节点上的partition所在的partition包括P1和Pn,因此,MDC需要更新P1和Pn的集群视图,具体可以包括:

[0172] 1)更新OSD view:将OSDx节点的状态由“在集群中(IN)”以及“有效(UP)”更新为“在集群中(IN)”以及“失效(DOWN)”状态。

[0173] 2)变更partition view:针对P1,将OSDy节点升为P1的主(在P1的备OSD节点列表中,选择第一个partition状态为一致的备OSD节点升为partition群的主OSD节点),将OSDx降为P1的备,且对应的partition状态更新为“不一致”;针对Pn,将Pn的备OSDx的partition状态变更为不一致。

[0174] 3)更新IO view:针对P1,将P1的IO view中的主OSD节点由原来的OSDx节点变更为OSDy节点;针对Pn,由于故障OSDx节点仅作为Pn的备OSD节点,Pn的主OSDy节点并没有发生故障,所以Pn的IO view不发生更新。

[0175] 814:向OSDy节点(作为更新后的P1的主OSD节点,以及继续保持Pn的主OSD节点)通知P1和Pn的partition视图的更新,包括OSDy节点升为P1的主OSD节点,OSDx降为P1的备OSD节点且对应的partition状态为“不一致”,Pn的备OSDx节点的partition状态变更为“不一致”。

[0176] 816:向OSDz节点通知P1的IOview更新,告知P1的主OSD节点变更为OSDy节点。

[0177] 818:向IO路由模块通知P1的IO view更新,告知P1的主OSD节点变更为OSDy节点。

[0178] 820:OSDy节点处理MDC模块的通知,更新本地保存的视图信息(partition view和IO view),并根据MDC模块通知的最新的视图处理IO请求所对应的数据的复制。

[0179] 具体实施方式中,OSDy节点作为更新后的P1的主OSD节点,更新P1的partition view以及IO view,作为原Pn的主OSD节点,更新针对Pn的partition view。

[0180] 针对P1的IO操作,当所述OSDy节点接收到所述IO路由模块转发的IO请求之后,执行IO请求,产生复制请求,将所述的复制请求发送至更新后的partition view中的P1的备OSD节点OSDz节点,且对应的partition状态为“一致”,由于OSDx节点作为P1的新的备OSD节点,其partition状态为“不一致”,OSDy节点不再将所述的复制请求发给OSDx节点,实现故障隔离,从而不影响所述针对P1的IO请求处理的继续进行。

[0181] 针对Pn的IO操作,当所述OSDy节点接收到所述IO路由模块转发的IO请求之后,执行IO请求,产生复制请求,将所述的复制请求发送至更新后的partition view中的Pn的备OSD节点OSDz节点,且对应的partition状态为“一致”,由于OSDx节点作为Pn的备OSD节点,

其partition状态为“不一致”,OSDy节点不再将所述的复制请求发给OSDx节点,实现故障隔离,从而不影响所述针对P1的IO请求处理的继续进行。

[0182] 在两阶段端提交协议中,如果参与者发生故障或者一直没有响应,其他的IO请求一直阻塞,最后超时失败,需要进行回退;本实施例中如果备OSD节点发生故障后,MDC会通知主节点进行视图变更,忽略该故障节点,对该故障节点进行隔离,继续IO请求的处理,不会阻塞其他的IO请求处理,具有更好的容错能力和可用性。

[0183] 822:处理MDC模块的视图通知,更新本地保存的IO view信息。

[0184] 824:处理MDC模块的视图通知,更新本地保存的IO view信息,并根据MDC模块通知的最新IO view视图进行IO请求的路由转发。

[0185] 具体实施方式中,针对P1的IO处理,如果在所述P1的IO处理过程中,P1所在的原主OSDx节点故障,MDC会及时更新P1的partition view,IO路由模块根据更新后的partition view中新选出的P1的主OSDy节点,将所述的IO请求重新转发至新选出的OSDy节点。

[0186] 在本发明实施例中,主OSD节点发生故障后,MDC节点能够快速选举新的主节点,快速的恢复IO的处理,如果备节点发生故障后,MDC会通知主节点进行视图变更,隔离或忽略该故障OSD节点,继续IO请求的处理,不会阻塞其他的IO请求处理,具有更好的容错能力,能够快速处理节点故障,例如,N+1个副本可以容忍N个副本故障,进而提高了存储系统的性能和可用性。可用性差的系统决定了其扩展性不会太好,因为在大规模分布式存储系统中,存储节点的故障是一件经常发生的事情,复杂且大量的协议交互也会使得系统的扩展能力减低。此外,可以针对partition粒度的集群视图控制,可以极大的缩小存储节点故障的影响范围,使得存储系统能够进行大规模的扩展,提高了系统的扩展性。

[0187] OSD节点故障恢复流程

[0188] 在OSD节点故障期间,可能发生了新的数据修改操作,因此,故障的OSD节点恢复重新加入集群提供服务前,需要先进行数据的恢复同步,使自己的状态恢复到主副本一致。

[0189] 本发明的一个实施例中,OSD节点故障恢复流程可以分为以下三个阶段:

[0190] 1)备OSD节点向主OSD节点同步所述主OSD节点在故障期间所做的数据修改,这是一个增量同步的过程,当然实际应用中,也可以根据实际情况同步某个partition的所有数据;

[0191] 2)备OSD节点恢复到与主OSD节点一致的状态后,MDC模块进行集群视图的变更;

[0192] 3)MDC模块通知各模块和节点更新后的集群视图,以便各模块和节点根据通知所述更新的集群视图处理IO请求的复制或转发。

[0193] 实际应用中,如果在数据恢复过程中,主OSD节点收到IO请求后,将复制请求发送至该故障恢复节点,则该OSD节点故障恢复流程可以进一步包括以下阶段:

[0194] 备OSD节点回放在向主OSD同步数据的过程中在收到主OSD节点的复制请求后记录的log,将log中的数据写入。这样可以保证,故障恢复OSD节点在故障恢复过程中确保所有数据同主OSD节点保持完全一致,进而提高主备OSD节点的数据一致性。

[0195] 其中的数据同步过程的具体流程可以参见下述图10的给出的一个具体实施例。

[0196] 其中的集群视图更新与更新通知过程可以包括以下情况:

[0197] 1)若OSD节点在故障前是partition群的备OSD节点,则可以只变更partition view,把partition view中对应的备OSD节点的partition状态修改为“一致”状态,并把变

更后的partition view通知给主OSD节点；

[0198] 2)若OSD节点在故障前是partition群的主OSD,则同时变更partition view和IO view,该前OSD节点升为该partition群的新的主OSD节点,当前的主OSD降为该partition群的备OSD节点,通知新的主OSD节点进行partition view变更,通知IO路由模块和备OSD节点进行IO view变更。

[0199] 实际应用中,如果故障的OSD节点在规定的时间周期内没有恢复,MDC模块则把该OSD节点踢出集群,把分布在该OSD节点的partition迁移到其他的OSD节点上。具体流程参见下述图11给出的关于OSD节点退出集群的一个具体实施例。

[0200] 为了便于理解,以下通过一个具体的实施例进行说明,如图9所示,为本发明提供的OSD节点故障恢复处理流程实施例,本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的MDC模块,IO路由模块以及OSD节点。为了理解的方便本实施例假设所述故障恢复OSD节点上的partition所在的partition群包括partition 1群(简称P1)以及partition n群(简称Pn)为例进行说明,该实施例中的OSD节点故障处理流程具体包括：

[0201] 902:故障恢复OSD节点向MDC模块请求该故障恢复OSD节点上的集群视图信息,请求中携带该OSD节点的OSD ID。

[0202] 具体的实施方式中,可以参考上述图6中步骤608所描述的过程。

[0203] 904:MDC模块根据OSD ID查询partition view获取该OSD上的partition信息。

[0204] 具体实施方式中,MDC根据所述故障恢复OSD的OSD ID分别查询所述故障恢复节点OSD上的P1和Pn所对应的partition view分别获得P1的partition信息和Pn的partition信息。该partition信息可以包括IO view,也可以进一步包括partition状态。

[0205] 906:向OSD节点返回partition信息。

[0206] 908/910:根据MDC返回的Partition信息向partition群的主OSD节点发起数据恢复过程,分别向P1和Pn的主OSD节点获取故障期间缺少的entry信息,携带SeqID。

[0207] 在本发明的具体实施例中,如果在所述故障恢复过程中,某个主OSD节点(如Pn的主OSD节点)收到IO写请求,需要向该Pn的所有备OSD节点发送复制请求,则该故障恢复流程可以进一步包括以下步骤912-916以及918：

[0208] 912:Pn的主节OSD点在故障恢复期间接收主机IO写请求。

[0209] 914:向Partition所在备OSD节点发送复制IO信息。

[0210] 具体实施方式中,上述步骤912-914可以参考上述图7中的IO操作流程。

[0211] 916:将复制IO关键信息写入日志(log)。

[0212] 具体实施方式中,也可以将所述IO请求所对应的数据写入日志,该步骤中的IO关键信息可以参考上述图7中的IO操作流程的描述。

[0213] 918:根据主OSD节点返回的差异entry数目信息,反复发送根据entry请求数据的消息,确保数据恢复过程执行完毕。

[0214] 具体实施方式参考下属图11中关于数据恢复的流程。

[0215] 920:根据日志记录的IO信息,按日志记录将IO写入。

[0216] 具体实施方式中,在所故障恢复OSD节点与所述主OSD节点完成数据恢复同步后,再根据日志中记录的IO信息将来自该Pn的主OSD的复制请求对应的数据写入该故障恢复

OSD节点所管理的物理存储资源。

[0217] 通过将故障恢复过程中的IO请求写入日志,然后在故障期间遗漏的数据恢复完成后写入故障期间发生的IO请求,可以保证主备OSD节点对每个IO的执行顺序是一致的,从而进一步提高数据备份的一致性。

[0218] 922/924:数据恢复完成,分别通知P1和Pn的主OSD节点。

[0219] 具体实施方式中,可以包括两种方式通知MDC模块实执行集群视图的更新,方式之一是数据恢复完毕后,由所述故障恢复OSD节点通知P1和Pn的主OSD节点,以便由该主OSD节点通知所述MDC节点进行集群视图更新。方式二参见下面的步骤930。

[0220] 实际应用中,所述故障恢复OSD节点在进行集群视图更新的通知之前可以进一步确认该故障恢复OSD节点上的partition的partition状态,并在确认partition状态不一致后触发集群式图更新流程。该故障恢复OSD节点可以通过上述步骤906中返回的partition信息汇总进一步获取partition状态信息。

[0221] 926/928:P1和Pn的主OSD节点分别通知MDC变更partition View,携带partition群ID、备OSD ID以及视图版本。

[0222] 具体实施方式中,向MDC模块发送通知,要求其将故障恢复OSD节点上的partition的partition的状态更新为一致,通知中携带OSD节点上partition所在的partition群的partition群ID、备OSD节点ID(即所述故障恢复OSD节点),以及视图版本。这里的partition群ID用于标示待更新的partition群的视图,OSD ID用于标示故障OSD节点,这里视图版本为该P1和Pn所在的主OSD节点本地保存的最新partition view的视图版本,其作用在于所述MDC收到该通知后,确认通知中的partition view的视图版本与该MDC本地维护的最新的partition view的视图版本一致,则进行集群视图更新处理,从而保证IO处理流程中所有模块或者节点看到的集群视图是一致的,从而提高数据备份的一致性。

[0223] 实际应用中,如果确认通知中的partition view的视图版本与该MDC本地维护的最新的partition view的视图版本不一致,则进一步将最新的partition view发送给所述P1和Pn所在的主OSD节点,并在确认针对该P1和Pn的主备数据一致后,进行集群视图的更新。

[0224] 928:通知MDC变更Partition View,将故障恢复OSD节点的状态更新为一致,携带的Partition群ID,备OSD ID,视图版本。

[0225] 930:数据恢复完成,通知主MDC变更Partition View,携带的Partition群ID,备OSD ID,视图版本。

[0226] 具体实施方式中,该故障恢复OSD节点在数据恢复完成之后,向MDC模块发送通知,告知MDC模块将故障恢复OSD节点上的partition的partition状态更新为一致,与上述方式一不同点在于,此处的视图版本为所述故障恢复OSD节点本地保存的最新的partition view的partition viw版本或者最新IO view的IO view版本(具体看该故障恢复OSD节点在故障前是主OSD节点还是备OSD节点确定是partition view还是IO view),MDC模块根据通知中的视图版本,确定该MDC模块于本地维护的相应的视图版本一致后,进行集群视图更新处理。

[0227] 932:将partition View中的故障恢复OSD节点的partition状态更新为一致。

[0228] 具体实施方式中,分别将故障恢复OSD节点上的P1和Pn的partition view中的所

述故障恢复OSD节点对应的partition状态更新为“一致”。

[0229] 934:通过对视图比较,确认故障恢复OSD节点在故障前是否是某些partition的主OSD节点。

[0230] 具体实施方式中,通过比较所述MDC模块本地维护的针对所述故障恢复OSD节点上P1和Pn的最新的partition view与该P1和Pn的初始化partition view,确认所述故障恢复OSD节点是P1的主OSD节点;或者通过比较所述MDC模块本地维护的针对所述故障恢复OSD节点上P1和Pn的最新的IO view与该P1和Pn的初始化IO view,确认所述故障恢复OSD节点是P1的主OSD节点。

[0231] 936:将故障恢复节点恢复为主OSD节点,更新Partition View。

[0232] 具体实施方式中,针对P1,所述MDC将所述故障恢复OSD节点恢复为所述P1的新的主OSD节点,并将所述P1所在的当前主OSD节点降为P1的新的备OSD节点,并更新所述P1的partition视图,针对Pn,由于所述故障恢复OSD节点原来就是Pn的备OSD节点,所以针对Pn不涉及所述故障恢复OSD节点的主备身份变更问题。

[0233] 938:如果故障恢复OSD节点原先是主节点,则将最新的Partition View发送至所述故障恢复OSD节点。

[0234] 具体实施方式中,由于该故障恢复OSD节点原先是P1的主节点,按照步骤937将其上升为所述故障恢复OSD节点的新的主OSD节点,因此,需要将更新后P1的最新的Partition View发送给所述故障恢复OSD节点的新的主OSD节点,即所述故障恢复OSD节点。由于故障恢复OSD节点原先不是Pn的主节点,所以Pn的最新partition view可不发给所述故障恢复OSD节点。

[0235] 940/942:通知备OSD更新本地的IO view。

[0236] 具体实施方式中,根据所述P1和Pn的最新的partition View或者IO view,分别获取P1和Pn的新的主OSD节点,将所述所述P1和Pn的最新的IO View分别发送给所述P1的新的主OSD节点和Pn的主OSD节点(由于Pn不涉及主OSD节点的变更,所以还是发给Pn原来的主OSD节点)。

[0237] 944:所述故障恢复OSD节点判断Partition View与本地保存的IO view中主OSD是否一致,判断自己是否升主,更新IO view。

[0238] 所述故障恢复OSD节点接收到所述最新的partition视图后,判断该最新的partition View与本地保存的IO view中的主OSD是否一致,判断自己是否升主,如果是,则更新IO view,以及本地保存的partition view。并根据所述更新的partition视图和IO view处理IO请求所涉及的数据的复制。

[0239] 946:更新IO view,如果是主OSD降备,则删除Partition View。

[0240] 具体实施方式中,所述P1和Pn的当前的主OSD节点(即数据恢复完毕后更新前的OSD节点)接收到所述更新的IO视图,更新各自的本地保存的IO视图。针对P1的当前主OSD节点,由于其已经由MDC模块降为给P1的新的备OSD节点,所以P1的当前主OSD节点将本地保存的partition视图删除,并根据更新后的本地保存的IO视图处理IO请求的复制。

[0241] 在本发明实施例中,OSD节点发生故障后,MDC会通知其他相关节点进行视图更新,隔离或忽略该故障OSD节点,继续IO请求的处理,不会阻塞其他的IO请求处理,并在节点故障恢复后更新视图,通知各相关节点,是的该故障恢复节点可以快速的重新加入集群工作,

能够快速处理节点故障和故障的恢复，具有更好的容错能力，提高了存储系统的性能和可用性。可用性差的系统决定了其扩展性不会太好，因为在大规模分布式存储系统中，存储节点的故障是一件经常发生的事情，复杂且大量的协议交互也会使得系统的扩展能力减低。此外，可以针对partition粒度的集群视图控制，可以极大的缩小存储节点故障的影响范围，使得存储系统能够进行大规模的扩展，提高了系统的扩展性。

[0242] 数据恢复流程

[0243] 下面通过一个具体实施例说明上述图9的OSD故障恢复处理过程中的数据恢复处理流程，为了理解的方便本实施例假设所述故障恢复OSD节点上的partition所在的partition群包括partition 1(简称P1)以及partition n(简称Pn)为例进行说明，如图10所示，该实施例中的OSD节点故障处理流程具体包括：

[0244] 1002：故障恢复OSD节点从本地获取记录的每个partition的entry中最大的SeqID。

[0245] 如上述图7所述的实施例中步骤712以及722所示，系统中的OSD节点在IO处理过程中针对某个partition的每一IO操作会记录一条entry，如前所述，所述的entry包括IO操作类型，Partition ID，Seq ID，以及key，所述的entry中还可以进一步包括状态信息，用于描述所述的操作是否成功，另外针对写IO操作还可以进一步包括前述的offset以及length。例如，本实施例中，分别获取针对P1以及Pn的写IO操作的最大Seq ID。

[0246] 1004/1006：分别向P1和Pn的主OSD节点请求获取该OSD缺少的entry，请求中携带partition群ID，各自的最大的SeqID。

[0247] 场景一：备OSD最大的SeqID在主OSD记录的entry范围内。

[0248] 1008/1010：P1和Pn的主OSD节点分别向所述故障恢复OSD节点发送该故障恢复OSD节点OSD缺少的entry。

[0249] 具体实施方式中，如果所述故障恢复OSD节点最大的SeqID为1.6，主OSD当前最大的SeqID为1.16，则把SeqID为1.7到1.16对应的10条entry发送给备OSD。这里仅仅是为了理解的方便而给出的一个例子，实际应用中，Seq ID的编号规则或者方式可能不一样，而且P1和Pn的主OSD节点上可能会有不同的缺少的entry。

[0250] 1012/1014：所述故障恢复OSD节点根据上一步获取的entry反复进行数据同步：请求中携带key、offset、length等IO关键信息。

[0251] 具体实施方式中，所述故障恢复OSD节点根据获取的entry分批逐条向所述P1和Pn所在的主OSD节点进行数据同步请求，请求中包括key、offset、length等IO关键信息。

[0252] 1016：发送对应的数据。

[0253] 具体实施方式中，所述P1和Pn的主OSD节点分别根据获取的所述数据同步请求中的信息将每条entry对应的数据发送至所述故障恢复OSD节点。

[0254] 场景二：备OSD最大的SeqID不在主OSD记录的entry范围内，比主OSD最小SeqID小。

[0255] 1018：发送主OSD最小SeqID和零条entry。

[0256] 具体实施方式中，所述Pn的主OSD节点在确定所述故障恢复OSD节点的最大的SeqID不在主OSD记录的entry范围内，即比所述Pn的主OSD的最小SeqID还小的情况下，向所述故障恢复OSD节点发送所述Pn的主OSD节点的最小SeqID和零条entry，方便所述故障恢复OSD节点进行判断所述Pn的主OSD节点是没有写过数据还是写的数据太多，以至于不能通过

增量同步完成数据恢复。

[0257] 1020:反复请求同步partition的数据,请求中携带partition群ID。

[0258] 所述故障恢复OSD节点向所述Pn的主OSD节点请求同步整个partition的数据,如本实施例中Pn的主OSD上的主partition数据,与所述故障恢复OSD节点上的Pn备partition上的数据进行同步,请求中携带partition群ID。具体实施方式中,由于通常一个partition的数据量很大,一个请求无法传输完,同时主节点不知道所述故障恢复节点的IO能力,一直往所述故障恢复节点发送数据,所述故障恢复节点可能无法处理,所以只有等所述故障恢复节点来请求同步数据时才会把数据发送给所述故障恢复节点;所以所述故障恢复OSD节点会根据情况反复向Pn的主OSD节点发送的同步请求,同步整个partition中的数据,直至所述partition中的所有数据都同步完毕。实际应用中,也可以有其他的整体partition同步的方式,本发明不做限制。

[0259] 1022:发送一个key或者多个key对应的数据。

[0260] 所述Pn的主OSD节点根据所述故障恢复OSD节点每次发送的同步请求发送一个或多个Key对应的数据。

[0261] OSD节点故障后退出集群的流程

[0262] OSD节点故障后超过预定的时间阈值(例如5分钟)仍不能恢复加入集群,或者OSD节点发生硬件故障,为了保证数据的可靠性,需要把故障的OSD节点踢出集群。

[0263] OSD节点退出集群是一个partition重新分布和数据迁移的过程,partition的重新分布要考虑各个节点的均衡性和副本的安全性;数据迁移过程的IO处理与故障恢复流程以及数据恢复流程中的处理一致;数据迁移完成,主备副本到达一致状态,MDC进行视图更新与通知的过程与故障恢复完成后视图更新的处理一致。各相关OSD节点根据更新后的视图进行IO请求的复制或转发处理过程也故障恢复完成后各OSD节点根据最新的视图进行IO的处理一致。

[0264] 为了便于理解,以下通过一个具体的实施例进行说明,如图11所示,为本发明提供的OSD节点故障恢复处理流程实施例,本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的MDC模块,IO路由模块以及OSD节点。为了理解的方便本实施例假设所涉及的OSD节点为OSD1节点,OSD2节点,以及OSDn节点,其中OSD1节点为partition1群(简称P1)的主OSD节点,以及partition 2群(简称P2)的备OSD节点,OSD2节点为P1的备OSD节点,OSDn节点为P2的主OSD节点。该实施例中的OSD节点故障退出处理流程具体包括:

[0265] 1100:OSD1节点发生故障。

[0266] 1102:MDC模块发现节点OSD1故障超过预定的阈值或者发生了硬件故障,MDC把OSD1踢出集群,进行视图变更,把OSD1节点上的partition(此处为P1的主partition和P2的备partition)迁移到其他OSD节点,例如本实施例中的OSD2节点和OSDn节点。

[0267] 1104:MDC模块向OSD2节点通知视图更新:OSD2节点升为P1的主,成为P2的备。

[0268] 1106:MDC模块向OSDn节点通知视图更新:OSDn节点为P2的主,成为P1的备。

[0269] 1108:OSD2节点向OSDn节点请求同步P2的数据。

[0270] 既然OSDn节点为P2的主OSD节点,所以OSD2节点向OSDn节点请求同步OSDn上的P2的主partition的数据,使得OSD2上的P2的备partition的数据与OSDn上的P2的主partition的数据一致。具体的同步流程与上述图10所示的数据恢复流程中整个partition

的数据同步流程相似,此处不再赘述。

[0271] 1110:OSDn节点向OSD2节点请求同步P1的数据。

[0272] 由于OSD2节点原来就是P1的备OSD节点,OSD1节点故障后,OSDn节点作为P1新的备OSD节点,只能向OSD2同步P1的数据,OSDn节点向OSD2节点请求同步OSD2上的P1的主partition的数据,使得OSDn上的P1的备patition的数据与OSD2上的P1的主partition的数据一致。

[0273] 1112:partition数据同步完成。

[0274] 1114:OSD2节点通知MDC模块已经完成P2的数据迁移。

[0275] 1116:OSDn节点通知MDC模块已经完成P1的数据迁移。

[0276] 1118:MDC模块根据相应的通知进行视图更新。

[0277] 具体的视图更新原理与上述各流程中所描述的一样,更新过程与上述各流程中所描述的一样,此处不再赘述。

[0278] 1120:通知视图更新:P1的备OSDn节点的partition状态为一致。

[0279] 1122:通知视图更新:P2的备OSD2节点的partition状态为一致。

[0280] 1124:所述OSD2节点和OSDn节点根据最新的视图进行IO请求所对应的数据的复制处理。

[0281] OSD节点扩容流程

[0282] 新节点扩容加入集群,需要把原来分布在其他OSD节点的partition迁移到新扩容加入集群的OSD节点,以保证数据分布的均衡性。节点扩容的流程主要涉及数据迁移,数据迁移后的视图更新,视图更新的通知,相关OSD节点根据更新后的视图处理IO请求的复制。

[0283] 数据迁移过程的IO处理与故障恢复流程以及数据恢复流程中的处理一致;数据迁移完成,MDC进行视图更新与通知的过程与故障恢复完成后视图更新的处理一致。各相关OSD节点根据更新后的视图进行IO请求所对应的数据的复制处理过程也故障恢复完成后各OSD节点根据最新的视图进行IO的处理一致。

[0284] 具体的实施例中,在数据迁移完成后进行视图更新可以包括:1)新扩容节点仍然是部分partition群的备OSD节点,partition状态为一致状态,该patition群的原来的一个备OSD不再是partition的备节点;2)新扩容节点升为部分partition群的主OSD节点,该partition群不再属于原来的主OSD节点(该partition群不再分布在原来的主OSD)。

[0285] 为了便于理解,以下通过一个具体的实施例进行说明。本实施例中涉及的执行主体为图2A至图2C以及图5中所描述的实施例中提及的MDC模块,IO路由模块以及OSD节点。为了理解的方便本实施例假设所涉及的OSD节点为OSD1节点、OSDn节点以及新扩容OSD节点,其中OSD1为partition群P1(简称P1)的主OSD节点,partition群Pn(简称Pn)的备OSD节点,OSDn为Pn的主OSD节点,P1的备OSD节点,新扩容OSD节点为P1的备OSD节点,以及Pn的备OSD节点。如图12所示,该实施例中的OSD节点扩容流程具体包括:

[0286] 1202:向MDC模块下发OSD节点扩容配置和命令。

[0287] 具体实施方式中,可以是系统管理员通过配置命令向MDC模块通知OSD节点扩容。

[0288] 1202:MDC模块进行视图更新,把部分OSD节点上的partition迁移到新扩容OSD节点。

[0289] 本实施例中,MDC模块将OSD1节点上的P1备和OSDn节点上的Pn备partition迁移到

所述新扩容OSD节点，使得新扩容OSD节点作为新的P1的备OSD节点和新的Pn的备OSD节点。

[0290] 1206：向OSD1节点通知视图更新：增加新的备OSD。

[0291] 具体实施方式中，MDC模块向OSD1节点通知视图更新，针对P1的新的partition view中增加新扩容OSD节点作为P1的备OSD节点，且对应的partition状态为“不一致”(因为新扩容OSD节点和OSD1节点还没有对P1的数据进行同步)。

[0292] 1208：向OSDn节点通知视图更新：增加新的备OSD。

[0293] 具体实施方式中，针对Pn的新的partition view中增加新扩容OSD节点作为Pn的备OSD节点，且对应的partition状态为“不一致”(因为新扩容OSD节点和OSDn节点还没有对Pn的数据进行同步)。

[0294] 1210：扩容新OSD节点启动。

[0295] 具体实施方式中，新扩容OSD节点作为新的OSD节点加入集群后，进行初始化过程。具体过程与上述图6中的集群视图初始化生成和获取流程一样，此处不再赘述。

[0296] 1212：返回该OSD节点的partition信息。

[0297] 具体实施方式中，MDC模块向所述新扩容OSD节点返回该新扩容OSD节点上的partition的视图，本实施例中为P1的IO view和Pn的IO view。

[0298] 1214：扩容新OSD节点向主OSD1节点请求同步partition的数据。

[0299] 具体实施方式中，扩容新OSD节点根据MDC返回的P1的IO view向P1的主OSD节点OSD1请求同步P1的数据，即同步OSD1节点上P1的主partition的数据，以使得该扩容新OSD节点上P1的备partition的数据与OSD1节点上P1的主partition的数据一致。

[0300] 具体的同步流程与上述图10所示的数据恢复流程中整个partition的数据同步流程相似，此处不再赘述。

[0301] 1216：扩容新OSD节点向主OSDn节点请求同步partition的数据。

[0302] 具体实施方式中，扩容新OSD节点根据MDC返回的Pn的IO view向Pn的主OSD节点OSDn请求同步Pn的数据，，即同步OSDn节点上Pn的主partition的数据，以使得该扩容新OSD节点上Pn的备partition的数据与OSDn节点上Pn的主partition的数据一致。

[0303] 具体的同步流程与上述图10所示的数据恢复流程中整个partition的数据同步流程相似，此处不再赘述。

[0304] 1218：partition数据同步完成。

[0305] 1220：通知MDC备节点已经完成partition数据迁移。

[0306] 具体实施方式中，OSD1节点向MDC模块通知，新扩容OSD节点已经完成P1的数据同步。

[0307] 1222：通知MDC备节点已经完成partition数据迁移。

[0308] 具体实施方式中，OSDn节点向MDC模块通知，新扩容OSD节点已经完成Pn的数据同步。

[0309] 1224：MDC模块进行视图更新。

[0310] 具体的视图更新原理与上述各流程中所描述的一样，更新过程与上述各流程中所描述的一样，此处不再赘述。

[0311] 1226-1230：分别通知所述OSD1节点、OSDn节点和新扩容OSD节点通知视图更新。

[0312] 1232：所述OSD1节点、OSDn节点和新扩容OSD节点根据更新后的视图进行IO请求所

对应的数据的复制处理。

[0313] 通过以上的实施例的描述，本领域普通技术人员可以理解：实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，所述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，包括如上述方法实施例的步骤，所述的存储介质，如：ROM/RAM、磁碟、光盘等。

[0314] 以上所述，仅为本发明的具体实施例，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

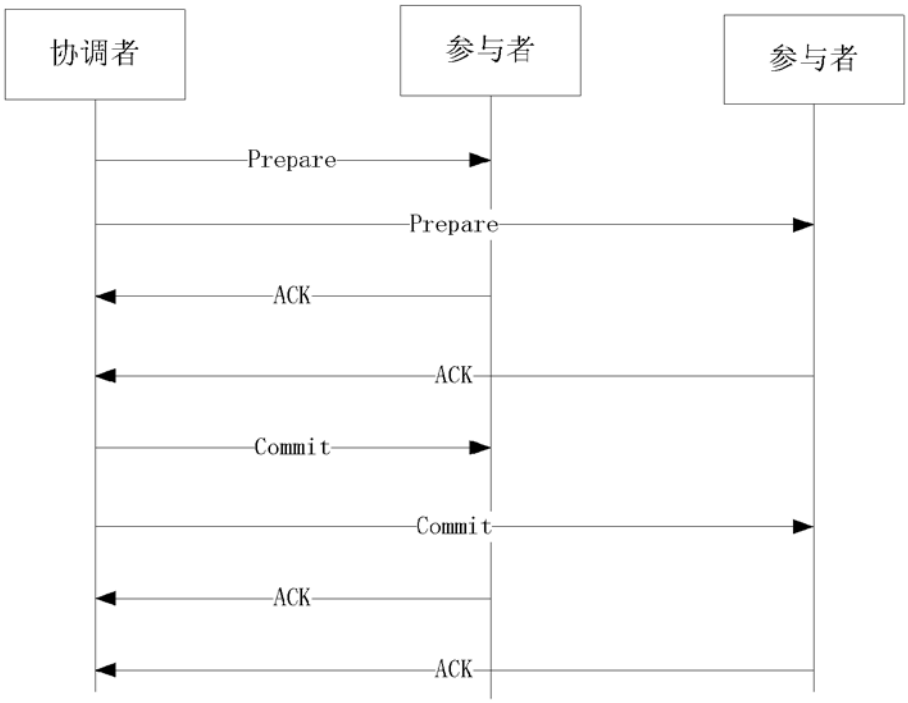

协调者
参与者
参与者
Prepare
Prepare
ACK
ACK
Commit
Commit
ACK
ACK

图1

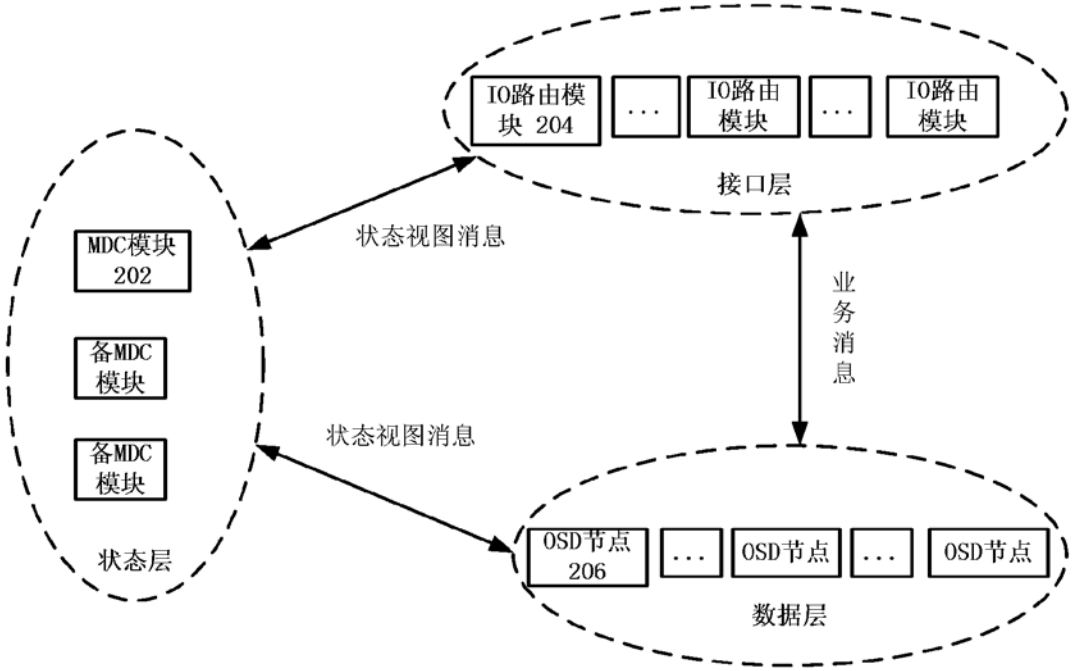

IO路由模块 204
...
IO路由模块
...
IO路由模块
接口层
状态视图消息
业务消息
MDC模块 202
备MDC模块
备MDC模块
状态视图消息
状态层
OSD节点 206
...
OSD节点
...
OSD节点
数据层

图2A

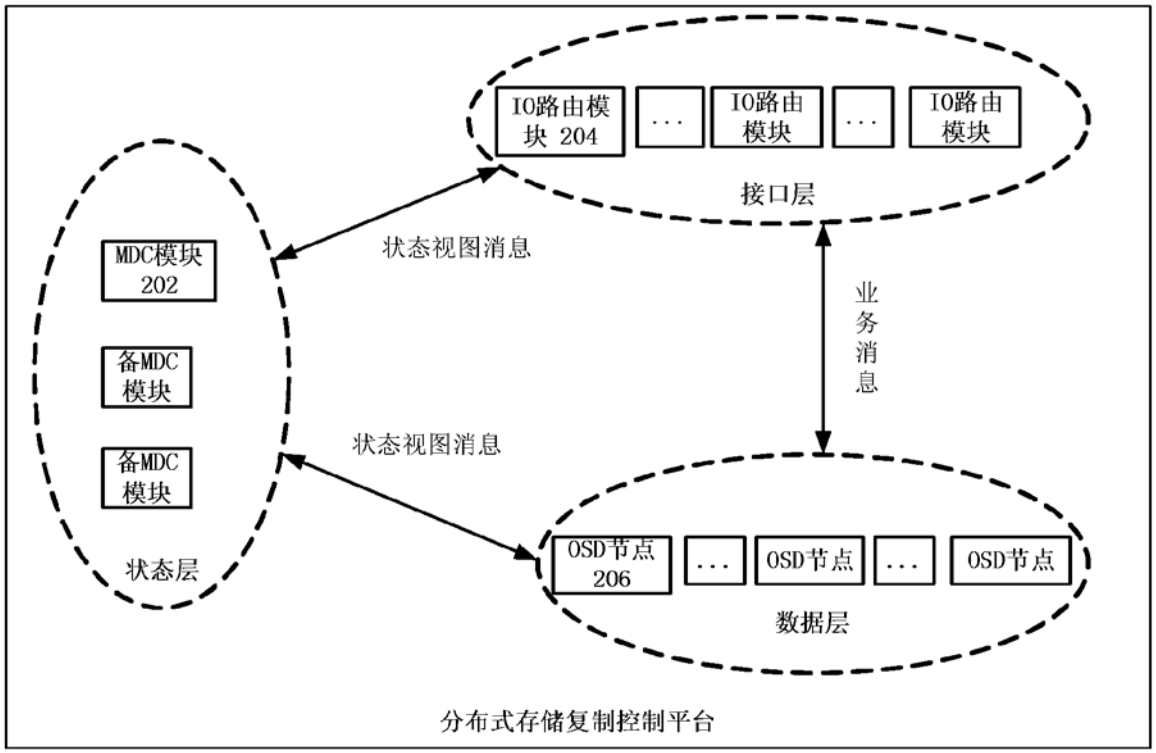

IO路由模块 204
IO路由模块
IO路由模块
接口层
状态视图消息
MDC模块 202
备MDC模块
备MDC模块
状态层
业务消息
状态视图消息
OSD节点 206
OSD节点
OSD节点
数据层
分布式存储复制控制平台

图2B

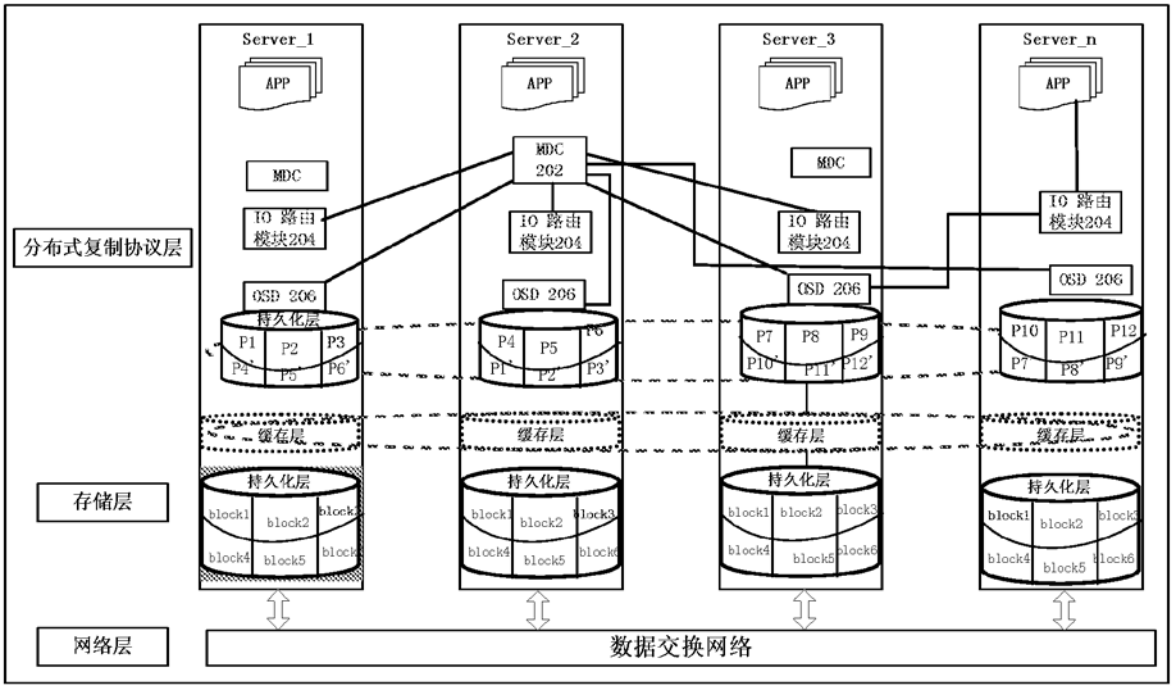

Server_1
Server_2
Server_3
Server_n
APP
MDC
MDC 202
IO 路由模块204
OSD 206
持久化层
缓存层
分布式复制协议层
存储层
网络层
数据交换网络

图2C

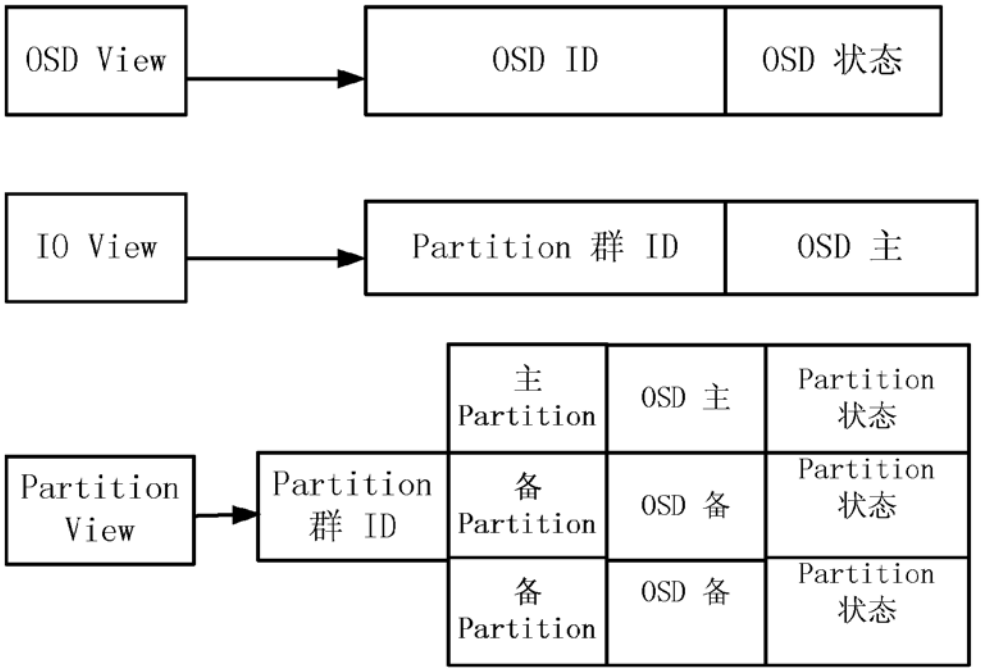

OSD View
OSD ID
OSD 状态
IO View
Partition 群 ID
OSD 主
Partition View
Partition 群 ID
主 Partition
OSD 主
Partition 状态
备 Partition
OSD 备
Partition 状态
备 Partition
OSD 备
Partition 状态

图3

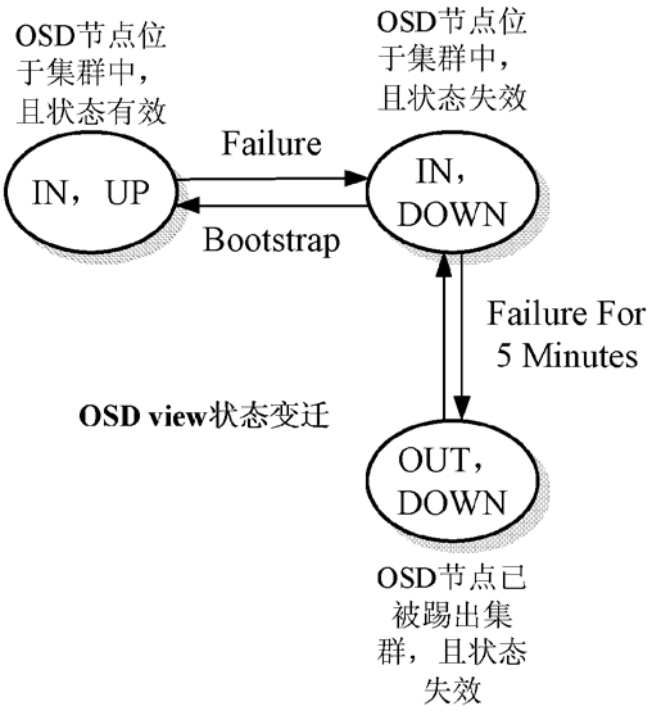

OSD节点位于集群中，且状态有效
OSD节点位于集群中，且状态失效
Failure
IN，UP
IN，DOWN
Bootstrap
Failure For 5 Minutes
OSD view状态变迁
OUT，DOWN
OSD节点已被踢出集群，且状态失效

图4

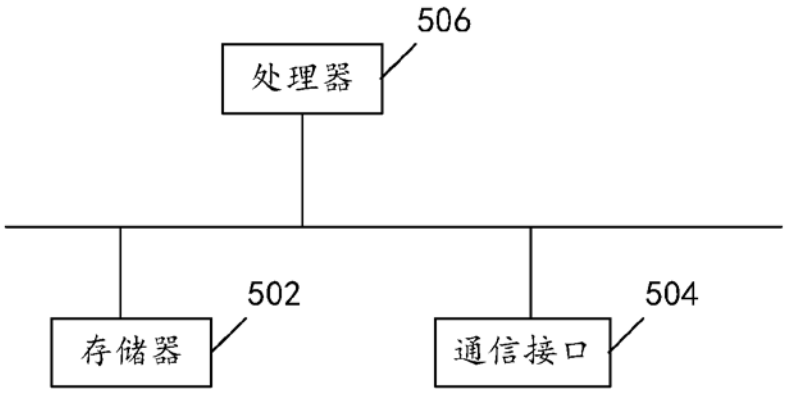

506
处理器
502
存储器
504
通信接口

图5

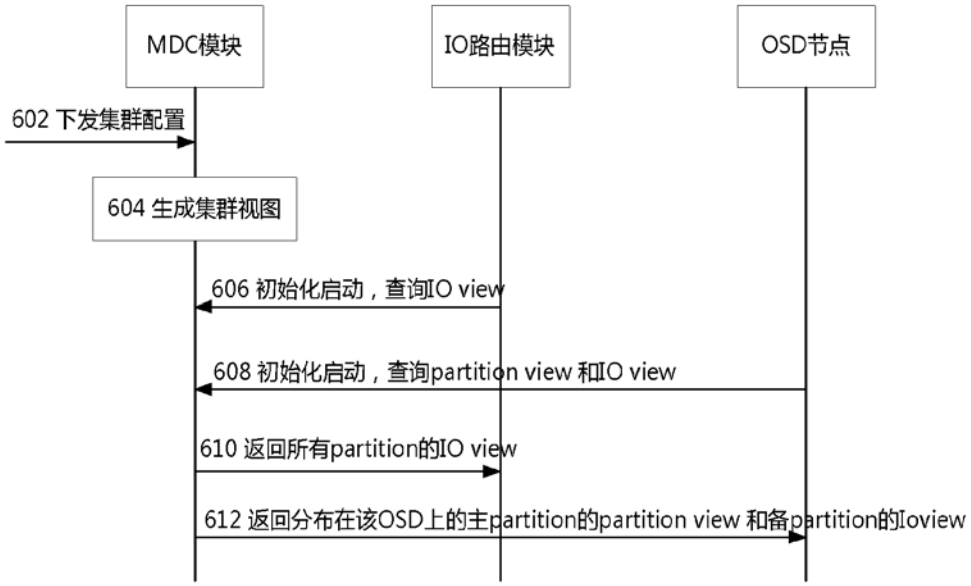

MDC模块
IO路由模块
OSD节点
602 下发集群配置
604 生成集群视图
606 初始化启动，查询IO view
608 初始化启动，查询partition view 和IO view
610 返回所有partition的IO view
612 返回分布在该OSD上的主partition的partition view 和备partition的Ioview

图6

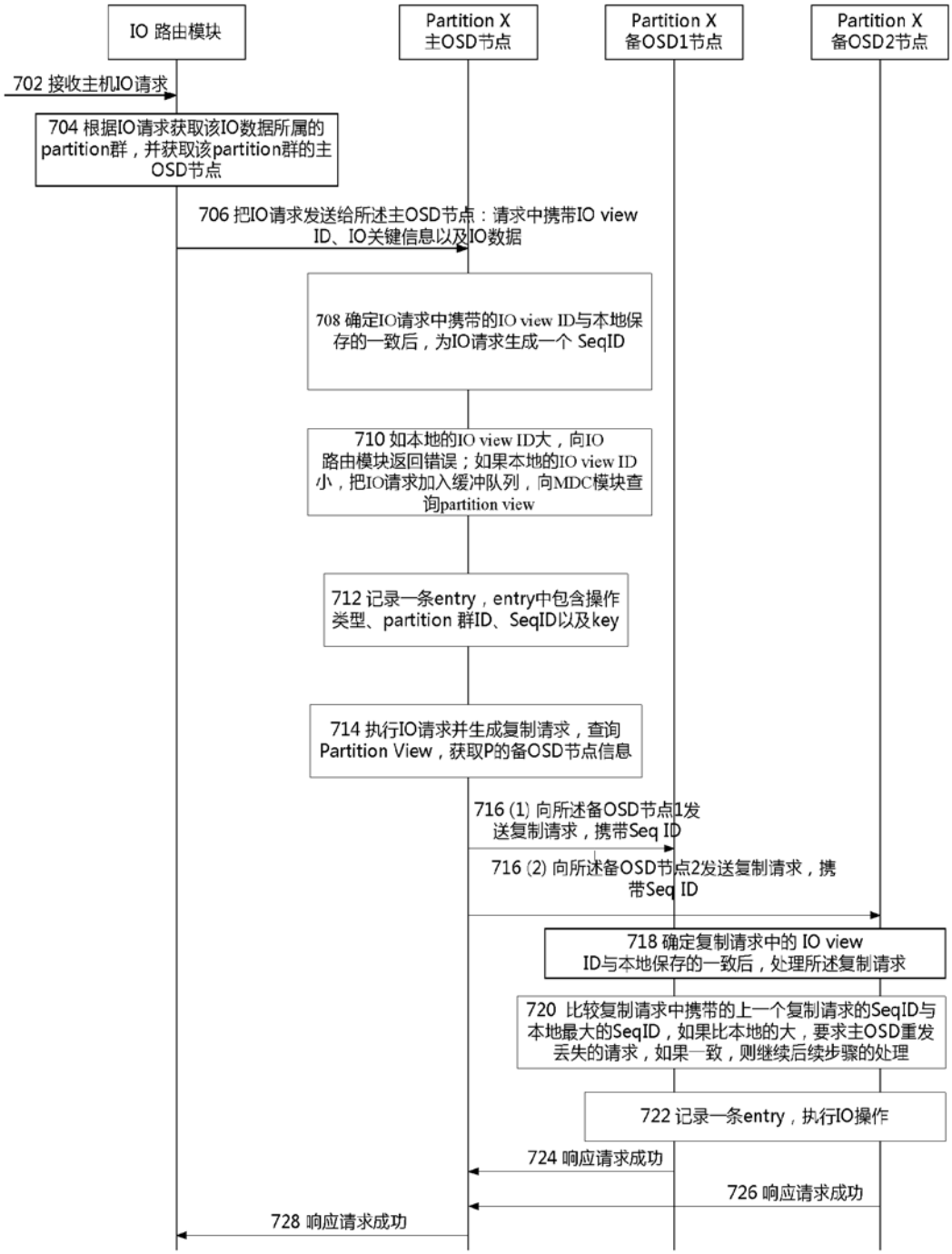

IO 路由模块
Partition X 主OSD节点
Partition X 备OSD1节点
Partition X 备OSD2节点
702 接收主机IO请求
704 根据IO请求获取该IO数据所属的partition群，并获取该partition群的主OSD节点
706 把IO请求发送给所述主OSD节点：请求中携带IO view ID、IO关键信息以及IO数据
708 确定IO请求中携带的IO view ID与本地保存的一致后，为IO请求生成一个 SeqID
710 如本地的IO view ID大，向IO路由模块返回错误；如果本地的IO view ID小，把IO请求加入缓冲队列，向MDC模块查询partition view
712 记录一条entry，entry中包含操作类型、partition 群ID、SeqID以及key
714 执行IO请求并生成复制请求，查询Partition View，获取P的备OSD节点信息
716 (1) 向所述备OSD节点1发送复制请求，携带Seq ID
716 (2) 向所述备OSD节点2发送复制请求，携带Seq ID
718 确定复制请求中的 IO view ID与本地保存的一致后，处理所述复制请求
720 比较复制请求中携带的上一个复制请求的SeqID与本地最大的SeqID，如果比本地的大，要求主OSD重发丢失的请求，如果一致，则继续后续步骤的处理
722 记录一条entry，执行IO操作
724 响应请求成功
726 响应请求成功
728 响应请求成功

图7

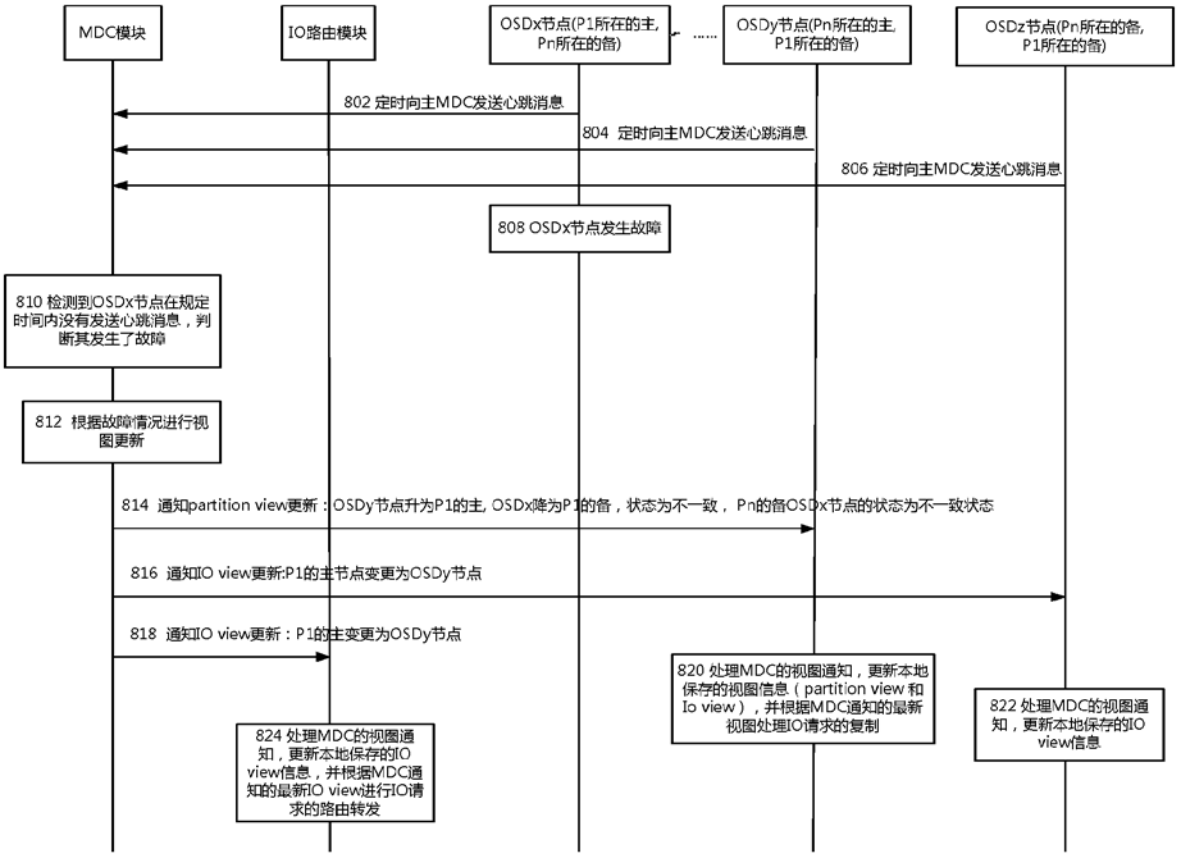

MDC模块
IO路由模块
OSDx节点(P1所在的主, Pn所在的备)
OSDy节点(Pn所在的主, P1所在的备)
OSDz节点(Pn所在的备, P1所在的备)
802 定时向主MDC发送心跳消息
804 定时向主MDC发送心跳消息
806 定时向主MDC发送心跳消息
808 OSDx节点发生故障
810 检测到OSDx节点在规定时间内没有发送心跳消息，判断其发生了故障
812 根据故障情况进行视图更新
814 通知partition view更新：OSDy节点升为P1的主，OSDx降为P1的备，状态为不一致， Pn的备OSDx节点的状态为不一致状态
816 通知IO view更新:P1的主节点变更为OSDy节点
818 通知IO view更新：P1的主变更为OSDy节点
820 处理MDC的视图通知，更新本地保存的视图信息（partition view 和 Io view），并根据MDC通知的最新视图处理IO请求的复制
822 处理MDC的视图通知，更新本地保存的IO view信息
824 处理MDC的视图通知，更新本地保存的IO view信息，并根据MDC通知的最新IO view进行IO请求的路由转发

图8

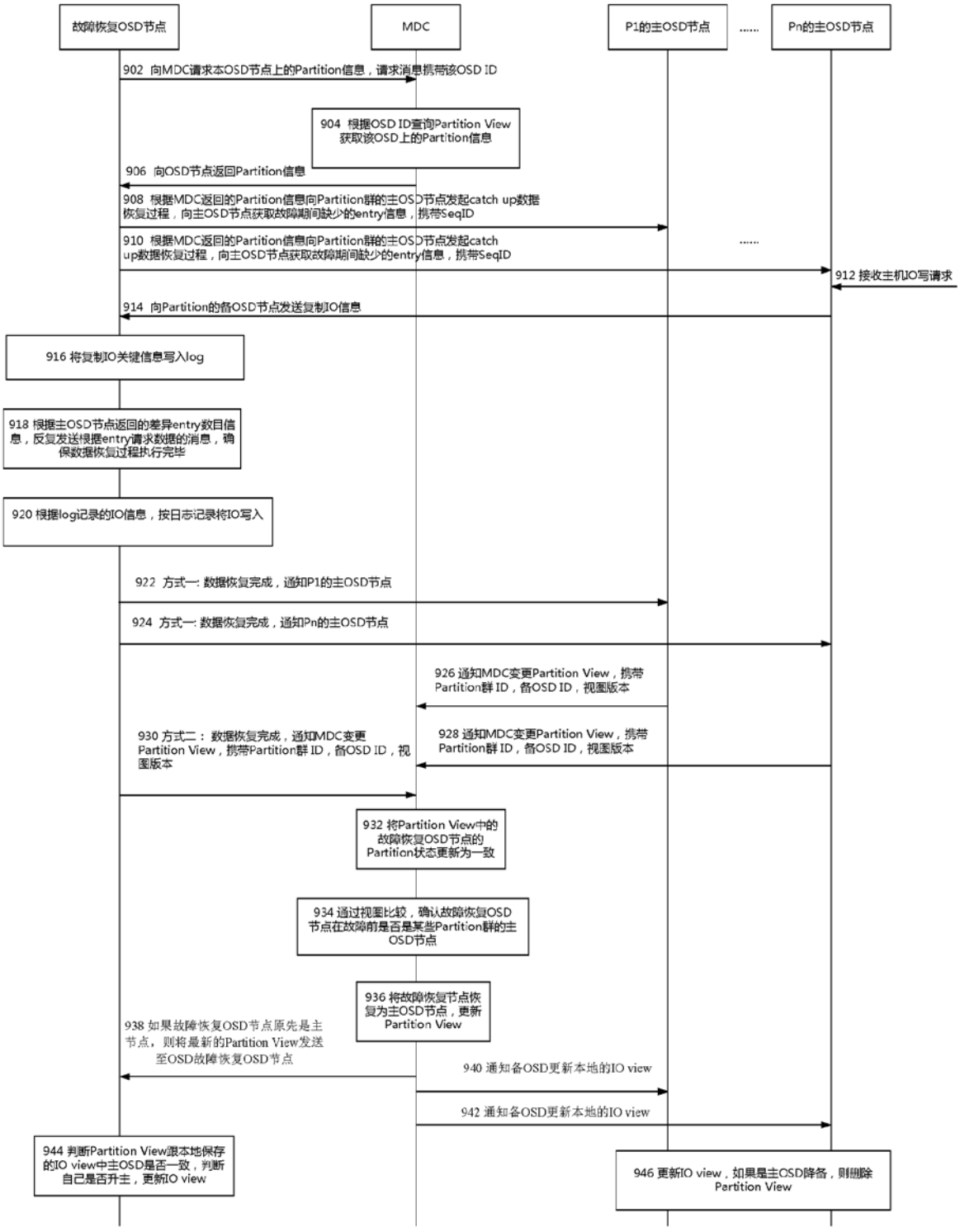

故障恢复OSD节点
MDC
P1的主OSD节点
.....
Pn的主OSD节点
902 向MDC请求本OSD节点上的Partition信息，请求消息携带该OSD ID
904 根据OSD ID查询Partition View 获取该OSD上的Partition信息
906 向OSD节点返回Partition信息
908 根据MDC返回的Partition信息向Partition群的主OSD节点发起catch up数据恢复过程，向主OSD节点获取故障期间缺少的entry信息，携带SeqID
910 根据MDC返回的Partition信息向Partition群的主OSD节点发起catch up数据恢复过程，向主OSD节点获取故障期间缺少的entry信息，携带SeqID
.....
912 接收主机IO写请求
914 向Partition的备OSD节点发送复制IO信息
916 将复制IO关键信息写入log
918 根据主OSD节点返回的差异entry数目信息，反复发送根据entry请求数据的消息，确保数据恢复过程执行完毕
920 根据log记录的IO信息，按日志记录将IO写入
922 方式一: 数据恢复完成，通知P1的主OSD节点
924 方式一: 数据恢复完成，通知Pn的主OSD节点
926 通知MDC变更Partition View，携带Partition群 ID，备OSD ID，视图版本
928 通知MDC变更Partition View，携带Partition群 ID，备OSD ID，视图版本
930 方式二： 数据恢复完成，通知MDC变更Partition View，携带Partition群 ID，备OSD ID，视图版本
932 将Partition View中的故障恢复OSD节点的Partition状态更新为一致
934 通过视图比较，确认故障恢复OSD节点在故障前是否是某些Partition群的主OSD节点
936 将故障恢复节点恢复为主OSD节点，更新Partition View
938 如果故障恢复OSD节点原先是主节点，则将最新的Partition View发送至OSD故障恢复OSD节点
940 通知备OSD更新本地的IO view
942 通知备OSD更新本地的IO view
944 判断Partition View跟本地保存的IO view中主OSD是否一致，判断自己是否升主，更新IO view
946 更新IO view，如果是主OSD降备，则删除Partition View

图9

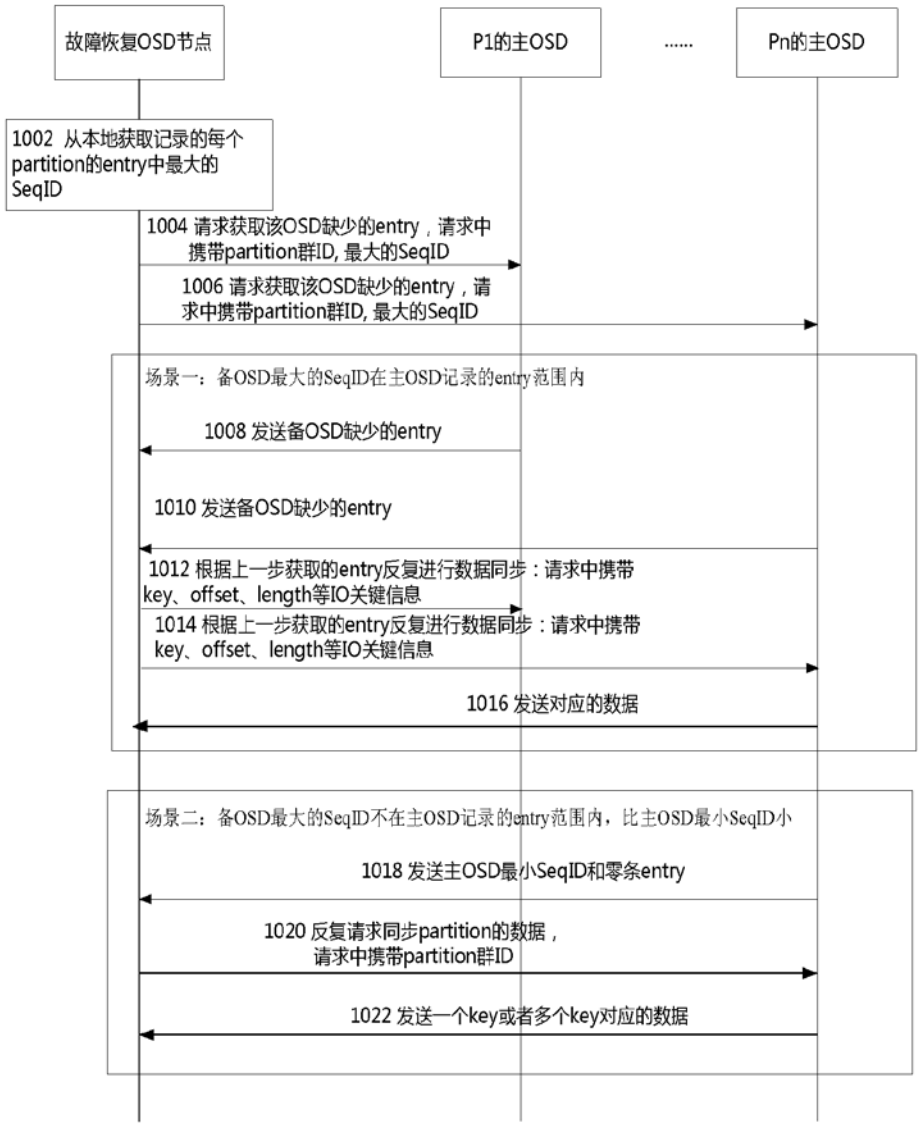

故障恢复OSD节点
P1的主OSD
.....
Pn的主OSD
1002 从本地获取记录的每个partition的entry中最大的SeqID
1004 请求获取该OSD缺少的entry，请求中携带partition群ID, 最大的SeqID
1006 请求获取该OSD缺少的entry，请求中携带partition群ID, 最大的SeqID
场景一：备OSD最大的SeqID在主OSD记录的entry范围内
1008 发送备OSD缺少的entry
1010 发送备OSD缺少的entry
1012 根据上一步获取的entry反复进行数据同步：请求中携带key、offset、length等IO关键信息
1014 根据上一步获取的entry反复进行数据同步：请求中携带key、offset、length等IO关键信息
1016 发送对应的数据
场景二：备OSD最大的SeqID不在主OSD记录的entry范围内，比主OSD最小SeqID小
1018 发送主OSD最小SeqID和零条entry
1020 反复请求同步partition的数据，请求中携带partition群ID
1022 发送一个key或者多个key对应的数据

图10

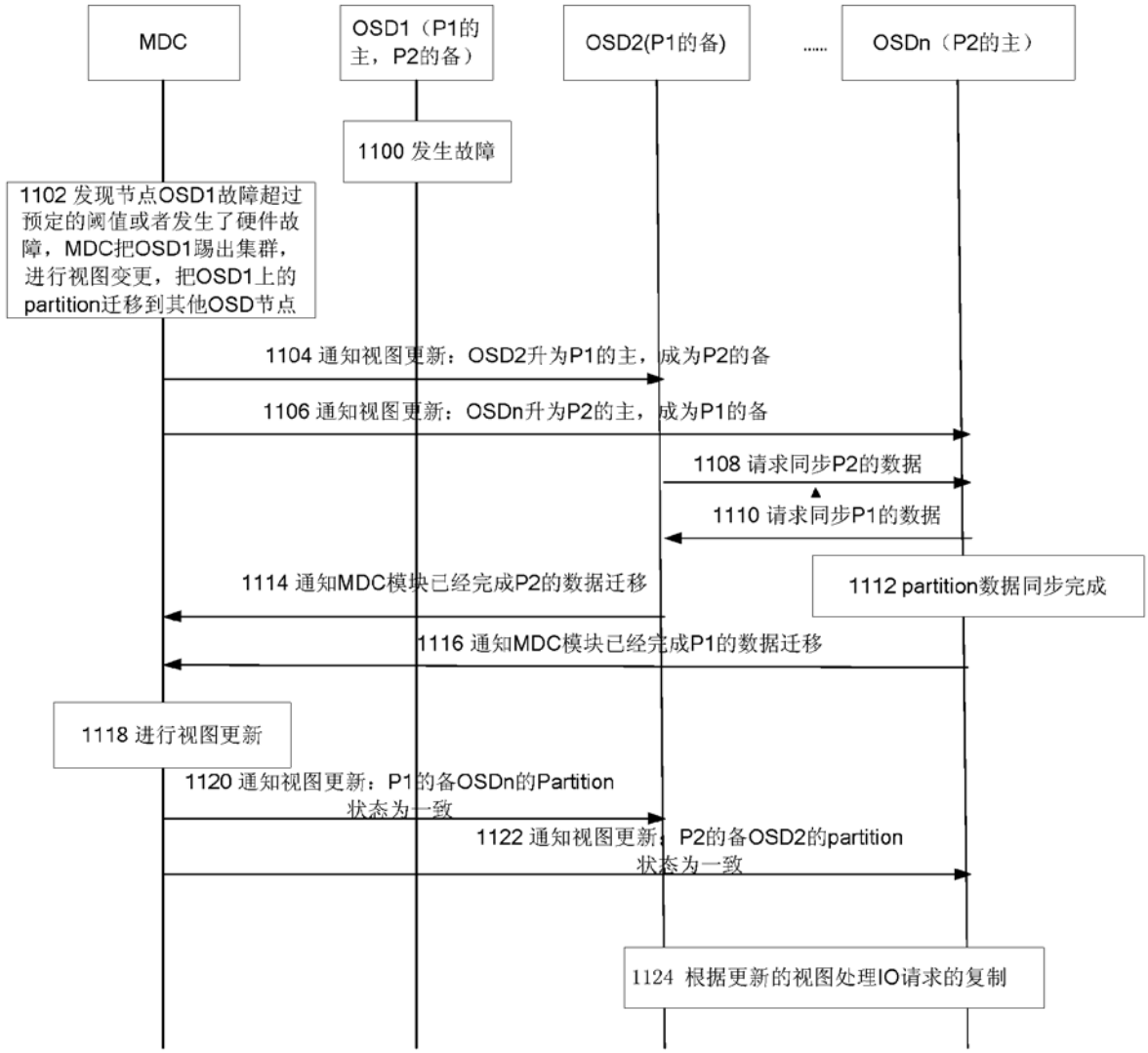

MDC
OSD1（P1的主，P2的备）
OSD2(P1的备)
.....
OSDn（P2的主）
1100 发生故障
1102 发现节点OSD1故障超过预定的阈值或者发生了硬件故障，MDC把OSD1踢出集群，进行视图变更，把OSD1上的partition迁移到其他OSD节点
1104 通知视图更新：OSD2升为P1的主，成为P2的备
1106 通知视图更新：OSDn升为P2的主，成为P1的备
1108 请求同步P2的数据
1110 请求同步P1的数据
1112 partition数据同步完成
1114 通知MDC模块已经完成P2的数据迁移
1116 通知MDC模块已经完成P1的数据迁移
1118 进行视图更新
1120 通知视图更新：P1的备OSDn的Partition状态为一致
1122 通知视图更新：P2的备OSD2的partition状态为一致
1124 根据更新的视图处理IO请求的复制

图11

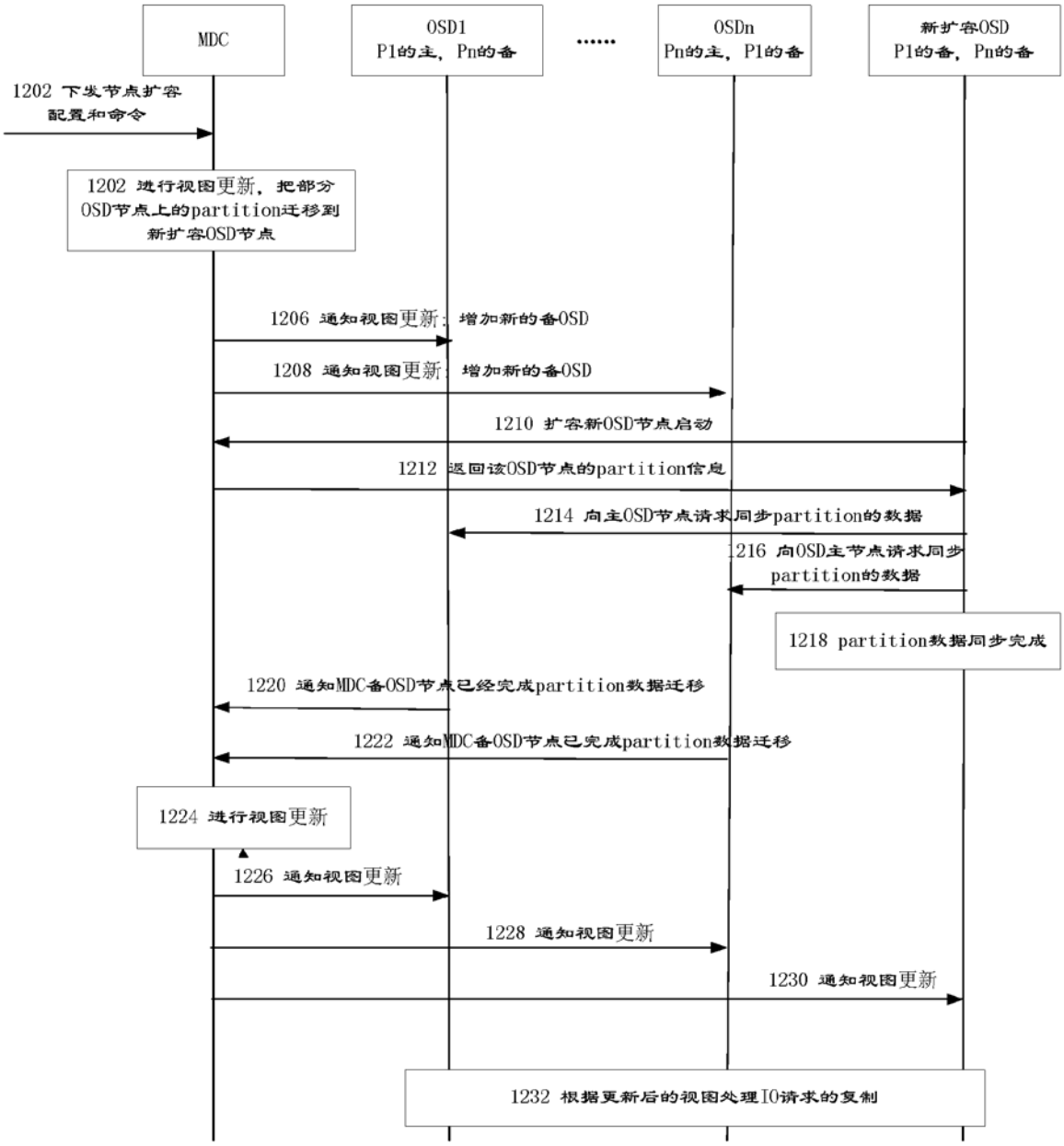

MDC
OSD1
P1的主，Pn的备
……
OSDn
Pn的主，P1的备
新扩容OSD
P1的备，Pn的备
1202 下发节点扩容配置和命令
1202 进行视图更新，把部分OSD节点上的partition迁移到新扩容OSD节点
1206 通知视图更新：增加新的备OSD
1208 通知视图更新：增加新的备OSD
1210 扩容新OSD节点启动
1212 返回该OSD节点的partition信息
1214 向主OSD节点请求同步partition的数据
1216 向OSD主节点请求同步partition的数据
1218 partition数据同步完成
1220 通知MDC备OSD节点已经完成partition数据迁移
1222 通知MDC备OSD节点已完成partition数据迁移
1224 进行视图更新
1226 通知视图更新
1228 通知视图更新
1230 通知视图更新
1232 根据更新后的视图处理IO请求的复制

图12