(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102413156 A
(43)申请公布日 2012.04.11

(21)申请号 201010291566.6

(22)申请日 2010.09.21

(71)申请人 腾讯科技(深圳)有限公司
地址 518044 广东省深圳市福田区振兴路赛格科技园2栋东403室

(72)发明人 田明 舒军 陈伟华 庄泗华 熊欢

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 罗正云 王琦

(51)Int.Cl.
H04L 29/08(2006.01)

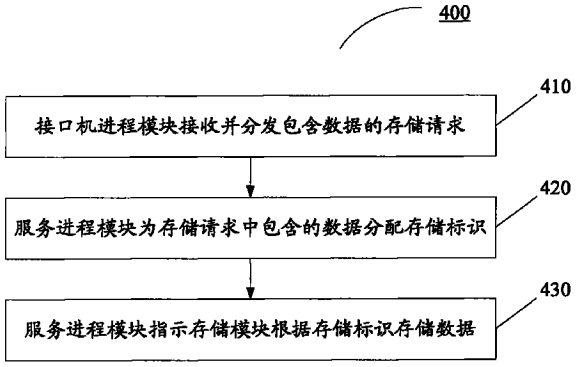
权利要求书 3 页 说明书 9 页 附图 5 页

(54)发明名称

一种分布式数据存储系统和方法

(57)摘要

本发明提出了一种分布式数据存储系统和方法。该分布式数据存储系统包括:接口机进程模块,用于从存储请求方接收存储请求并分发所述存储请求,所述存储请求包含待存储的数据;至少一个服务进程模块,用于分配存储标识并提供数据存储服务,所述至少一个服务进程模块之一从所述接口机进程模块接收所述存储请求,并为所述存储请求中包含的所述数据分配存储标识;和至少一个存储模块,用于根据所述至少一个服务进程模块之一分配的所述存储标识存储所述数据。本发明的分布式存储方法和系统可以为用户提供稳定、高并发的海量数据存储、读写服务。



CN 102413156 A

1. 一种分布式数据存储系统，包括：

接口机进程模块，用于从存储请求方接收存储请求并分发所述存储请求，所述存储请求包含待存储的数据；

至少一个服务进程模块，用于分配存储标识并提供数据存储服务，所述至少一个服务进程模块之一从所述接口机进程模块接收所述存储请求，并为所述存储请求中包含的所述数据分配存储标识；和

至少一个存储模块，用于根据所述至少一个服务进程模块之一分配的所述存储标识存储所述数据。

2. 根据权利要求 1 所述的分布式数据存储系统，其中所述接口机进程模块进一步维护可用服务进程模块列表，并且所述接口机进程模块从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一，并将所述存储请求转发给所述至少一个服务进程模块之一。

3. 根据权利要求 2 所述的分布式数据存储系统，其中所述接口机进程模块通过确定并删除所述可用服务进程模块列表中不能够提供新增存储服务的服务进程模块来维护可用服务进程模块列表。

4. 根据权利要求 3 所述的分布式数据存储系统，其中所述接口机进程模块进一步记录上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，并基于所述上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

5. 根据权利要求 2 所述的分布式数据存储系统，其中所述接口机进程模块进一步随机从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

6. 根据权利要求 1 所述的分布式数据存储系统，其中所述至少一个服务进程模块之一进一步在将所述数据存储在所述至少一个存储模块之后，向所述存储请求方响应成功应答，并且所述成功应答包括所述存储标识。

7. 根据权利要求 6 所述的分布式数据存储系统，其中所述接口机进程模块进一步从所述存储请求方接收包含所述存储标识的读取请求，并根据包含在所述存储请求中的所述存储标识将所述读取请求转发给所述至少一个服务进程模块之一，所述至少一个服务进程模块之一根据包含在所述读取请求中的所述存储标识指示所述至少一个存储模块将存储在所述至少一个存储模块中的所述数据返回给所述存储请求方。

8. 根据权利要求 6 所述的分布式数据存储系统，其中所述接口机进程模块进一步从所述存储请求方接收包含所述存储标识和待修改的内容的写入请求，并根据包含在所述写入请求中的所述存储标识将所述写入请求转发给所述至少一个服务进程模块之一，所述至少一个服务进程模块之一根据包含在所述写入请求中的所述存储标识和待改的内容指示所述至少一个存储模块修改存储在所述至少一个存储模块中的所述数据。

9. 根据权利要求 1 所述的分布式数据存储系统，其中所述至少一个服务进程模块之一在至少一段连续标识内按照存储标识的顺序或随机为所述存储请求中包含的所述数据分配所述存储标识。

10. 根据权利要求 1 所述的分布式数据存储系统，进一步包括：另一接口机进程模块，用于与所述接口机进程模块协同接收和分发所述存储请求。

11. 一种分布式数据存储方法，包括：

接口机进程模块从存储请求方接收存储请求，并分发所述存储请求，所述存储请求包含待存储的数据；

至少一个服务进程模块之一从所述接口机进程模块接收所述存储请求，并为所述存储请求中包含的所述数据分配存储标识；并且

至少一个存储模块根据所述至少一个服务进程模块之一分配的所述存储标识存储所述数据。

12. 根据权利要求 11 所述的分布式数据存储方法，进一步包括：

所述接口机进程模块维护可用服务进程模块列表，其中分发所述存储请求包括：

从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一，并将所述存储请求转发给所述至少一个服务进程模块之一。

13. 根据权利要求 12 所述的分布式数据存储方法，其中维护可用服务模块列表包括：

确定并删除所述可用服务进程模块列表中不能提供新增存储服务的服务进程模块。

14. 根据权利要求 13 所述的分布式数据存储方法，进一步包括：

所述接口机进程模块记录上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，其中从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一包括：

基于上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

15. 根据权利要求 12 所述的分布式数据存储方法，其中从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一包括：

随机从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

16. 根据权利要求 11 所述的分布式数据存储方法，进一步包括：

所述至少一个服务进程模块之一在将所述数据存储在所述至少一个存储模块之后，向所述存储请求方响应成功应答，并且所述成功应答包含所述存储标识。

17. 根据权利要求 16 所述的分布式数据存储方法，进一步包括：

所述接口机进程模块从所述存储请求方接收包含所述存储标识的读取请求，并根据包含在所述读取请求中的所述存储标识将所述读取请求转发给所述至少一个服务进程模块之一；并且

所述至少一个服务进程模块之一根据包含在所述读取请求中的所述存储标识指示所述至少一个存储模块将存储在所述至少一个存储模块中的所述数据返回给所述存储请求方。

18. 根据权利要求 16 所述的分布式数据存储方法，进一步包括：

所述至少一个接口机进程模块之一从所述存储请求方接收包含所述存储标识和待修改的内容的写入请求，并根据包含在所述写入请求中的所述存储标识将所述写入请求转发给所述至少一个服务进程模块之一；并且

所述至少一个服务进程模块之一根据包含在所述写入请求中的所述存储标识和待修改的内容指示所述存储至少一个存储模块对存储在所述至少一个存储模块中的所述数据进行修改。

19. 根据权利要求 11 所述的分布式数据存储方法，其中为所述存储请求中包含的所述数据分配存储标识包括：

在至少一段标识内，按照存储标识的顺序或随机为所述存储请求中包含的所述数据分配所述存储标识。

20. 根据权利要求 11 所述的分布式数据存储方法，进一步包括：

所述接口机进程模块与另一接口机进程模块协同接收和分发所述存储请求。

一种分布式数据存储系统和方法

技术领域

[0001] 本发明一般涉及计算机软件及互联网技术领域，尤其是涉及一种分布式数据存储系统和方法。

背景技术

[0002] UGC(用户生成内容,Users Generate Content)是一种用户使用互联网的新方式，即由原来的以下载为主变成下载和上传并重。例如，社区网络、视频分享和博客等都是UGC的主要应用形式。随着全球互联网业务的不断发展，UGC业务正在日渐崛起，引起了业界的广泛关注。

[0003] 由于数据是用户产生的，海量的用户催生出海量的数据，同时又会带来海量的读写量。如何存储这些数据，如何提供高并发的读写服务，是技术领域必然面临的问题。

[0004] 图1示出了现有的分布式数据存储系统的系统架构100，包括存储标识(ID)分配系统（或者说ID放号系统）120和数据存储系统130。

[0005] 存储标识分配系统120系统负责在存储请求方请求存储数据时为待存储的数据分配存储标识。存储标识分配系统120保证存储标识的全局唯一性，并让存储标识在某个或某些存储标识段（在某些号段范围）内有一定的随机性，一定程度上保证了数据存储系统130的负载均衡。

[0006] 数据存储系统130负责数据的存储并提供读写服务，其包括接口机进程模块131、多个服务进程模块132和多个存储模块133。接口机进程模块用于接收存储请求方110发送的包含存储标识的读写、存储请求，并把读写、存储请求分发到对应的服务进程模块132上，同时把服务进程模块132的部署细节对外屏蔽掉；每个服务进程模块132负责某个或某些存储标识段内的数这些数据的读写服务，并在将数据成功存储后向存储请求110响应成功应答；存储模块133，用于根据服务进程模块132的指示存储、读写数据。

[0007] 图2是现有技术的分布式数据存储方法200的示意性流程图。

[0008] 参见图2，当增加一条新数据时，上述分布式数据存储方法包括如下步骤：

[0009] 步骤210：存储标识分配系统为待存储的数据分配唯一的存储标识；

[0010] 步骤220：存储请求方根据分配的存储标识提交包含数据的请求至接口机进程模块；

[0011] 步骤230：接口机进程模块根据存储标识所属的存储标识段转发存储请求至对应的服务进程模块；

[0012] 步骤240：服务进程模块根据存储标识指示存储模块存储数据，并向存储请求方响应成功应答。

[0013] 另外，当读取一条数据时，上述分布式数据存储方法还可以包括如下步骤：存储请求方提交包含存储标识的读取请求至接口机进程模块，接口机进程模块根据存储标识所属的存储标识段分发写入请求至对应的服务进程模块，服务进程模块根据存储标识指示存储模块向存储请求方返回该数据。

[0014] 进一步，当写入一条数据时，上述分布式数据存储方法还可以包括如下步骤：存储请求方提交包含存储标识和待修改内容的写入请求至接口机进程模块，接口机进程模块根据存储标识所属的存储标识段分发写入请求至对应的服务进程模块，服务进程模块指示存储模块写入修改的内容。

[0015] 以上分布式数据存储系统具有如下不足之处：

[0016] 1. 耦合性高。数据存储系统对存储标识分配系统有依赖关系。首先，存储标识分配系统需要保证存储标识的均匀性和随机性，一旦存储标识分配系统的随机性被打破，可能导致某个服务进程模块所执行的进程因写请求量突增而被压垮；而且当存储标识分配系统出现单点故障时，整个分布式数据存储系统的存储请求都无法完成。

[0017] 2. 设计复杂。两个系统同等重要，为了保证对外的正常服务，两者都需要进行各种容灾设计。

[0018] 3. 耦合性和设计的复杂度直接导致运维成本增加。

[0019] 4. 针对新增请求，存在单点故障。当某个服务进程模块所执行的进程挂掉时，其针对所负责的存储标识段的新增请求会失败。

[0020] 5. 增加了带宽成本。每次新增数据时，都要先获取存储标识后才能进行实际存储，比直接存储多了一次交互，带宽成本倍增。

[0021] 可见，需要有一种简单、高效、低成本的存储服务模型来解决上述技术问题，以便为用户提供稳定、高并发的海量数据存储、读写服务。而这样的存储服务模型也将为该技术领域带来意义深远的变革。

发明内容

[0022] 有鉴于此，本发明提供了一种新分布式数据存储系统和方法，可以为用户提供稳定、高并发的海量数据存储、读写服务。

[0023] 本发明的技术方案具体是这样实现的：

[0024] 根据本发明的实施例的一种分布式数据存储系统，包括：接口机进程模块，用于从存储请求方接收存储请求并分发所述存储请求，所述存储请求包含待存储的数据；至少一个服务进程模块，用于分配存储标识并提供数据存储服务，所述至少一个服务进程模块之一从所述接口机进程模块接收所述存储请求，并为所述存储请求中包含的所述数据分配存储标识；和至少一个存储模块，用于根据所述至少一个服务进程模块之一分配的所述存储标识存储所述数据。

[0025] 所述接口机进程模块可以进一步维护可用服务进程模块列表，并且所述接口机进程模块从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一，并将所述存储请求转发给所述至少一个服务进程模块之一。

[0026] 所述接口机进程模块可以通过确定并删除所述可用服务进程模块列表中不能够提供新增存储服务的服务进程模块来维护可用服务进程模块列表。

[0027] 所述接口机进程模块可以进一步记录上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，并基于所述上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

[0028] 所述接口机进程模块可以进一步随机从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

[0029] 所述至少一个服务进程模块之一可以进一步在将所述数据存储在所述至少一个存储模块之后，向所述存储请求方响应成功应答，并且所述成功应答包括所述存储标识。

[0030] 所述接口机进程模块可以进一步从所述存储请求方接收包含所述存储标识的读取请求，并根据包含在所述存储请求中的所述存储标识将所述读取请求转发给所述至少一个服务进程模块之一，所述至少一个服务进程模块之一根据包含在所述读取请求中的所述存储标识指示所述至少一个存储模块将存储在所述至少一个存储模块中的所述数据返回给所述存储请求方。

[0031] 所述接口机进程模块可以进一步从所述存储请求方接收包含所述存储标识和待修改的内容的写入请求，并根据包含在所述写入请求中的所述存储标识将所述写入请求转发给所述至少一个服务进程模块之一，所述至少一个服务进程模块之一根据包含在所述写入请求中的所述存储标识和待修改的内容指示所述至少一个存储模块修改存储在所述至少一个存储模块中的所述数据。

[0032] 所述至少一个服务进程模块之一可以在至少一段连续标识内按照存储标识的顺序或随机为所述存储请求中包含的所述数据分配所述存储标识。

[0033] 所述分布式数据存储系统可以进一步包括：另一接口机进程模块，用于与所述接口机进程模块协同接收和分发所述存储请求。

[0034] 根据本发明的另一实施例的一种分布式数据存储方法，包括：接口机进程模块从存储请求方接收存储请求，并分发所述存储请求，所述存储请求包含待存储的数据；至少一个服务进程模块之一从所述接口机进程模块接收所述存储请求，并为所述存储请求中包含的所述数据分配存储标识；并且至少一个存储模块根据所述至少一个服务进程模块之一分配的所述存储标识存储所述数据。

[0035] 所述分布式数据存储方法可以进一步包括：所述接口机进程模块维护可用服务进程模块列表，其中分发所述存储请求包括：从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一，并将所述存储请求转发给所述至少一个服务进程模块之一。

[0036] 维护可用服务模块列表可以包括：确定并删除所述可用服务进程模块列表中不能提供新增存储服务的服务进程模块。

[0037] 所述分布式数据存储方法可以进一步包括：所述接口机进程模块记录上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，其中从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一包括：基于上一次存储数据时从所述可用服务进程列表中选择的服务进程模块的标识，从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

[0038] 从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一包括：随机从所述可用服务进程模块列表中选择所述至少一个服务进程模块之一。

[0039] 所述分布式数据存储方法可以进一步包括：所述至少一个服务进程模块之一在将所述数据存储在所述至少一个存储模块之后，向所述存储请求方响应成功应答，并且所述成功应答包含所述存储标识。

[0040] 所述分布式数据存储方法可以进一步包括：所述接口机进程模块从所述存储请求

方接收包含所述存储标识的读取请求，并根据包含在所述读取请求中的所述存储标识将所述读取请求转发给所述至少一个服务进程模块之一；并且所述至少一个服务进程模块之一根据包含在所述读取请求中的所述存储标识指示所述至少一个存储模块将存储在所述至少一个存储模块中的所述数据返回给所述存储请求方。

[0041] 所述分布式数据存储方法可以进一步包括：所述至少一个接口机进程模块之一从所述存储请求方接收包含所述存储标识和待修改的内容的写入请求，并根据包含在所述写入请求中的所述存储标识将所述写入请求转发给所述至少一个服务进程模块之一；并且所述至少一个服务进程模块之一根据包含在所述写入请求中的所述存储标识和待修改的内容指示所述存储至少一个存储模块对存储在所述至少一个存储模块中的所述数据进行修改。

[0042] 为所述存储请求中包含的所述数据分配存储标识可以包括：在至少一段标识内，按照存储标识的顺序或随机为所述存储请求中包含的所述数据分配所述存储标识。

[0043] 所述分布式数据存储方法可以进一步包括：所述接口机进程模块与另一接口机进程模块协同接收和分发所述存储请求。

[0044] 从上述方案可以看出，本发明通过在各个服务进程模块中分配存储标识避免了使用专门的存储标识分配系统，从而因为存储标识分配系统的故障而使整个分布式存储系统无法正常工作，并且为用户提供了简单、高效、低成本的存储服务。

附图说明

[0045] 附图与说明书一起示出本发明的示例性实施例，并且与描述一起用于解释本发明的原理；

[0046] 图 1 示出了现有的分布式数据存储系统的系统架构；

[0047] 图 2 是现有的分布式数据存储方法的示意性流程图；

[0048] 图 3 示出了根据本发明实施例的分布式数据存储系统的系统架构；

[0049] 图 4 是示出根据本发明另一实施例的分布式数据存储方法的示意性流程图；

[0050] 图 5 是示出根据本发明另一实施例的分布式数据存储方法的示例性流程图；

[0051] 图 6 是示出根据本发明另一实施例的选择可用服务进程模块的方法的示例性流程图；和

[0052] 图 7 是示出根据本发明另一实施例的选择可用服务进程模块的方法的示例性流程图。

具体实施方式

[0053] 为使本发明的目的、技术方案和优点更加清楚，以下举实施例对本发明进一步详细说明。

[0054] 在本发明的分布式数据存储方案中，提出了一种新的分配存储标识的方法，并设计了一种分布式的后台服务架构模型。在本发明的分配存储标识方案中，不再使用存储标识分配系统，改由各个服务进程模块直接负责存储标识的分配，在这种方案下，服务性能没有丝毫降低，但系统的整体复杂度、耦合性、带宽成本都大幅降低。

[0055] 以上是本发明的总体构思，下面将详细给出示例性的实施方案，以便于本领域人

员理解本发明。

[0056] 图 3 示出了根据本发明实施例的分布式数据存储系统的系统架构 300。

[0057] 下面将参见图 3 描述根据本发明的实施例的分布式数据存储系统 330 的结构。分布式数据存储系统 330 包括 :接口机进程模块 331、至少一个服务进程模块 332 和至少一个存储模块 333。接口机进程模块 331 用于从存储请求方 310 接收包含待存储的数据的存储请求,并向服务进程模块 332 分发存储请求。服务进程模块 332 用于分配存储标识并提供数据存储服务,其从接口机进程模块 331 接收存储请求,并为存储请求中包含的数据分配存储标识。存储模块 333 用于根据服务进程模块 332 分配的存储标识存储数据。

[0058] 下面详细说明根据本发明的示例性实施例的接口机进程模块 331、服务进程模块 332 和存储模块 333 的功能。

[0059] 为了使得分布式数据存储系统 330 在增加新数据时不会因为服务进程模块 332 不能接收新数据而存储失败,可以在接口机进程模块 331 中维护可用服务进程模块列表。例如,接口机进程模块 331 通过确定并删除可用服务进程模块列表中不能够提供新增存储服务的服务进程模块 332 来维护可用服务进程模块列表。实际上,接口机进程模块 331 可以通过多种方法判断服务进程模块 332 是否可用并更新可用服务进程模块列表。例如,接口机进程模块 331 可以维护一个状态为“可追加”的服务进程模块的列表,在该列表里列出的每个服务进程模块的“可追加”状态都是有效的。可以在接口机进程模块 332 中采用轮循方式周期性地向各服务进程模块确认其“可追加″ 状态是否是有效的。可替代地,可以从服务进程模块 332 周期性地向接口机进程模块报 331 告其“可追加”状态是否是有效的。当然,可替代地,也可以不删除该列表中不能提供新增储服务的服务进程模块 332,而是根据该列表中记录的服务进程模块 332 的状态来确定该模块是否可用。本发明并不限于上述维护可用服务进程模块列表的方式。

[0060] 接口机进程模块 331 在接收到存储请求时,可以从上述可用服务进程模块列表中选择一个可用的服务进程模块 332,并将存储请求转发给该可用的服务进程模块 332。例如,当某个服务进程模块 332 执行的服务进程挂掉时,该服务进程模块的“可追加”状态不再有效,接口机进程模块 331 便不再分发新增请求至该服务进程模块,并把存储请求均摊到其余可用的服务进程模块 332 上,以保证新增服务的可用性。再例如,当某个服务进程模块 332 分配完毕其所负责的所有存储标识时,接口机进程模块认为该服务进程模块不再可用,并且不再向其分发存储请求。另外,当某个服务进程模块 332 分配完其所负责的存储标识时,其“可追加”状态也不再有效。

[0061] 进一步,接口机进程模块 331 可以采用多种方式从“可追加”服务进程模块列表中选择可用服务进程模块。

[0062] 例如,为了使得各服务进程模块 332 的负载均衡,根据本发明的示例性实施例可以进一步记录上一次存储数据时从可用服务进程列表中选择的服务进程模块的标识,并基于上一次存储数据时从可用服务进程列表中选择的服务进程模块的标识,从可用服务进程模块列表中选择可用的服务进程模块。例如,记录上次被分发存储请求的服务进程模块的下标(即上一次存储请求分发给了哪个服务进程模块),从而实现在所有“可追加”状态有效的服务进程模块中逐次分发新增存储请求,以保证各服务进程模块的负载均衡。另外,接口机进程模块 331 还可以随机从可用服务进程模块列表中选择可用的服务进程模块。本发

明并不限于上述选择可用服务进程模块的方式。

[0063] 每个服务进程模块 332 可以各自负责某个或某些存储标识段（或号段）的存储标识的分配和数据服务，各个服务进程模块之间相互独立。例如，规定十万个连续存储标识为一个存储标识段，则存储标识段和存储标识的计算公式可以表示为 :UnitID = [ID/100000]+1，例如 1 至 99999 的存储标识属于存储标识段 1。可以为每个服务进程模块定义三种子状态 :可读、可写、可追加。“可读”表明该服务进程模块可以提供数据读取服务，“可写”表明该服务进程模块可以提供数据修改服务，“可追加”表名该服务进程模块可以提供新增数据服务。服务进程模块的状态可以由这三种子状态任意组合而成（“和 / 或”的关系）。

[0064] 具体来说，服务进程模块 332 除了提供正常的读写功能外，还在至少一段连续存储标识内按照存储标识的顺序或随机地为存储请求中包含的数据分配存储标识。例如，服务进程模块 332 在自己负责的存储标识段内为数据分配存储标识。具体分配规则可以采用从小到大逐次分配存储标识的方式。假设十万的连续存储标识为一个存储标识段，具体存储标识分配举例如下 :假设服务进程模块 i 负责的存储标识段为 6 和 8，则该进程放出的第一个存储标识为 500000，第二个存储标识为 500001，依次类推，当放到 599999 时，存储标识段 6 内的存储标识分配完毕，下一个存储标识为 800000，依次类推，当所有的存储标识段都分配完毕时，该服务进程模块不能再新增数据，或者说其“可追加”状态失效。

[0065] 存储模块 333 可以是各种计算机存储介质，例如固态硬盘 (SSD，Solid State Disk) 和硬盘驱动器 (Hard Disk Drive，HDD) 等等。本领域技术人员知道存储模块 333 是如何根据存储标识存储数据的，因此这里不再详细描述其具体的技术细节。

[0066] 进一步，根据本发明的另一示例性实施例，服务进程模块 332 进一步用于在将数据存储在存储模块 333 之后，向存储请求方 310 响应成功应答消息，并且该成功应答消息包含为该数据分配的存储标识。

[0067] 根据本发明的另一实施例，接口机进程模块 331 可以进一步从存储请求方 310 接收包含存储标识的读取请求，并根据包含在存储请求中的存储标识将读取请求转发给之前分配该存储标识的服务进程模块 332，服务进程模块 332 根据包含在读取请求中的存储标识指示存储模块 333 将存储在存储模块 333 中的数据返回给数据存储请求方。

[0068] 根据本发明的另一实施例，接口机进程模块 331 进一步从存储请求方 310 接收包含存储标识及待修改的内容的写入请求，并根据包含在写入请求中的存储标识将写入请求转发给之前分配该存储标识的服务进程模块 332，服务进程模块 332 根据包含在写入请求中的存储标识指示存储模块 333 对存储在存储模块 333 中的数据进行修改。

[0069] 在上述方案中，虽然图 3 仅仅示出了一个接口机进程模块 331，但本发明并不限于此。根据本发明的另一实施例，分布式数据存储系统还可以包括多个接口机进程模块。多个接口机进程模块可以协同工作，以完成存储请求的接收和分发。例如，多个接口机进程模块可以采用容灾设计，也可以分别负责向一部分可用服务进程模块分发存储请求，或者协同起来向所有可用服务进程模块分发存储请求，以增加系统的可靠性和灵活性。

[0070] 另外，图 3 仅仅示意性地示出了根据本发明的分布式存储系统的上述各个模块在逻辑上的连接关系，实际上，上述模块可以位于在物理上相同或不同的计算机或网络系统中。一个服务进程模块可以将数据存储在一个或多个存储模块上，一个存储模块也可以接

收多个服务进程模块分配的数据。

[0071] 还需要说明的是，存储请求方可以是各种形式的需要数据服务的一方，例如可以是网络、服务器或客户端等等发出数据服务请求的设备或系统。在读取数据时，存储请求方也可以称为读取请求方，而在写入数据时，存储请求方也可以称为写入请求方。

[0072] 图 4 是示出根据本发明另一实施例的分布式数据存储方法 400 的示意性流程图。

[0073] 参见图 4，根据本发明的实施例的分布式数据存储方法 400 包括如下步骤：

[0074] S410：接口机进程模块从存储请求方接收存储请求，并向至少一个服务进程模块中的一个服务进程模块分发存储请求，其中在存储请求中包含待存储的数据。

[0075] S420：该服务进程模块从接口机进程模块接收存储请求，并为存储请求中包含的数据分配存储标识。

[0076] S430：至少一个存储模块根据该服务进程模块分配的存储标识存储数据。

[0077] 在本发明的实施例中，接口机进程模块可以采用多种方式向服务进程模块分发存储请求，而服务进程模块也可以采用多种方式分配存储标识。下文中将结合具体的示例性实施例进行详细的说明。

[0078] 图 5 是示出根据本发明另一实施例的分布式数据存储方法 500 的示例性流程图。

[0079] 参见图 5，根据本发明的示例性实施例的分布式数据存储方法 500 包括如下步骤：

[0080] S510：接口机进程模块从存储请求方接收包含待存储的数据的存储请求。该存储请求的实现类似于图 2 中的存储请求，所不同的是该存储请求中不包含存储标识。

[0081] S520：接口机进程模块从可用服务进程模块列表中选择一个可用服务进程模块。如上述在根据本发明的分布式数据存储系统的实施例中描述的那样，接口机进程模块可以通过多种方式维护可用服务进程模块列表，例如，可以确定并删除所述可用服务进程模块列表中不能提供新增存储服务的服务进程模块来维护可用服务进程查模块列表。进一步，接口机进程模块可以记录上一次存储数据时从可用服务进程列表中选择的服务进程模块的标识，并且基于上一次存储数据时从可用服务进程列表中选择的服务进程模块的标识，从可用服务进程模块列表中选择可用于分发当前存储请求的服务进程模块。另外，接口机进程模块也可以随机地从可用服务进程模块列表中选择可用的服务进程模块。根据本发明的实施例的选择可用服务进程模块的具体方法流程图详见对图 6 和图 7 的描述。

[0082] S530：接口机进程模块将存储请求转发给选中的该可用服务进程模块。

[0083] S540：被选中的可用服务进程模块为该存储请求中包含的数据分配存储标识。服务进程模块可以在至少一段连续存储标识内按照存储标识的顺序或随机地为存储请求中包含的数据分配存储标识。具体的分配方法详见根据本发明的系统实施例中关于存储标识的分配的描述。

[0084] S550：存储模块根据该服务进程模块分配的存储标识存储数据。本领域技术人员可以知道如何根据存储标识存储数据，这里不再详细描述。

[0085] 以上是实现根据本发明实施例的分布式存储方法的基本流程。为了完善数据的存储、读写服务，根据本发明的另一实施例可以进一步包括如下步骤：

[0086] S560：服务进程模块在将数据存储在存储模块之后，向存储请求方响应成功应答消息。该成功应答消息包含可以告诉存储请求方已经成功存储数据的信息，并且可以包括存储标识，以供将来读取和写入数据使用。

[0087] 当读取数据时，根据本发明的示例性实施例进一步包括如下步骤：

[0088] S570：接口机进程模块从存储请求方接收包含存储标识的读取请求。

[0089] S572：接口机进程模块根据包含在存储请求中的存储标识将读取请求转发给之前分配该存储标识的服务进程模块。

[0090] S574：接口机进程模块根据包含在读取请求中的存储标识向存储模块读取数据。

[0091] S576：存储模块根据存储标识将数据返回给存储请求方。

[0092] 当写入数据时，根据本发明的示例性实施例进一步包括如下步骤：

[0093] S580：接口机进程模块从存储请求方接收包含存储标识和待修改的数据的写入请求。

[0094] S582：接口机进程模块根据包含在存储请求中的存储标识将写入请求转发给之前分配该存储标识的服务进程模块。

[0095] S584：该服务进程模块根据包含在读取请求中的存储标识和待修改的内容向存储模块写入数据，以使得存储模块可以修改存储在其上的数据。

[0096] 图 6 是示出根据本发明另一实施例的选择可用服务进程模块的方法 600 的示例性流程图。

[0097] 参见图 6，图 5 中的选择可用服务进程模块的方法 520 可以包括如下步骤：

[0098] S610：接口机进程模块获取下一服务进程模块的状态。接口机进程模块可以按照服务进程模块的标识的顺序，以一定的周期，循环地查询各个服务进程模块的状态。也可以由各个服务进程模块向接口机进程模块报告其当前的状态。

[0099] S620：接口机进程模块确定服务器进程模块是否可用？例如，是否可追加新数据？如果是，则执行步骤 S610，继续查询下一个服务进程模块是否可用。如果否，则执行步骤 S630。

[0100] S630：从可用服务进程列表中删除不可用的服务进程模块。

[0101] 步骤 S610 至步骤 630 可以周期性地执行，以维护可用服务进程模块列表，保证可用服务进程列表中的服务进程模块是可用的，或者说是可追加新数据的。

[0102] 根据本发明的实施例，当存储请求方发出存储请求时，接口进程模块可以执行下列步骤：

[0103] S640：接口机进程模块从数据请求方接收存储请求。

[0104] S650：接口机进程模块给变量 ID 赋值，使它等于上一次提供存储数据服务的服务进程模块的标识 (ID)。例如，可以为服务进程模块分配连续的 ID，例如可以是 1 到 n 的自然数，在这种情况下，当第一次执行数据存储时，上一次存储数据的服务进程模块的 ID 是 0。

[0105] S660：使得 ID ＝ ID+1。执行这一步的目的是为了能够按照顺序依次选择可用服务进程模块。

[0106] S670：如果 ID ＞ IDmax，则执行步骤 S695，否则继续执行步骤 680。在这里，IDmax 表示服务进程模块中 ID 值最大的服务进程模块的 ID。

[0107] S680：接口机进程模块确认标识为 ID 的服务进程模块是否在列表中？如果不在，则执行步骤 660，继续查看下一个（标识为 ID+1）的服务进程模块是否在列表中。如果在，则继续执行步骤 690。

[0108] S690：选择标识为 ID 的服务进程模块作为接收存储请求的服务进程模块。

[0109] S695：使得 ID ＝ 0 并执行步骤 660。如果标识变量 ID ＞ IDmax，说明接口机进程模块已经完成一轮向标识为从 1 至 n 的服务进程模块分发存储请求，因此，可以开始下一轮向标识为从 1 至 n 的服务进程模块分发存储请求。

[0110] 上述选择可用服务进程模块的方法可以保证各服务进程模块的负载均衡，但本发明选择可用服务进程模块的方法并不限于此。

[0111] 图 7 是示出根据本发明另一实施例的选择可用服务进程模块的方法 700 的示例性流程图。

[0112] 参见图 7，在图 5 中描述的选择可用服务进程模块的方法 520 可以具体包括如下步骤：

[0113] S710：接口机进程模块获取下一服务进程模块的状态。

[0114] S720：接口机进程模块确定服务器进程模块是否可用？例如，是否可追加新数据？如果是，则执行步骤 S710，继续查询下一个服务进程模块是否可用。如果否，则执行步骤 S730。

[0115] S730：从可用服务进程列表中删除不可用的服务进程模块。

[0116] 步骤 S710 至步骤 730 与步骤 S610 至步骤 630 类似，可以周期性地执行，以维护可用服务进程模块列表，保证可用服务进程列表中的服务进程模块是可用的，或者说是可追加新数据的。

[0117] 根据本发明的实施例，当存储请求方发出存储请求时，接口进程模块可以执行下列步骤：

[0118] S740：接口模块从数据请求方接收存储请求。

[0119] S750：接口模块随机地从可用服务进程模块列表中选择可用的服务进程模块。

[0120] 另外，根据本发明的另一实施例，在存在多个接口机进程模块时，各个接口机进程模块可以与其它接口机进程模块协同接收和分发存储请求。多个接口机进程模块可以采用容灾设计，可以分别负责维护一部分可用服务进程模块的列表，还可以共同维护同一可用服务进程模块列表，以增加系统的可靠性和灵活性。

[0121] 本发明的上述实施例采用的这种分存式数据存储方案不再使用专门的存储标识分配系统，这样分布式数据存储系统不再有外部依赖模块，彻底解耦合，从而不会因为存储标识分配系统故障而使整个分布式存储系统无法正常工作。

[0122] 另外，运用本发明提出的这种分存式数据存储系统或方法的方案还具有如下优点：1) 因为不再使用专门的存储标识分配系统，进而也不再需要为其进行容灾设计，所以设计简单；2) 由于不存在与专门的存储标识分配系统的耦合性并且设计简单，因此运行维护成本降低；3) 由于可以仅将新增存储请求分发给可用的服务进程模块，因此不存在服务进程模块单点故障的问题；以及 4) 由于不再进行与存储标识分配系统的交互，所以带宽成本减降低。

[0123] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

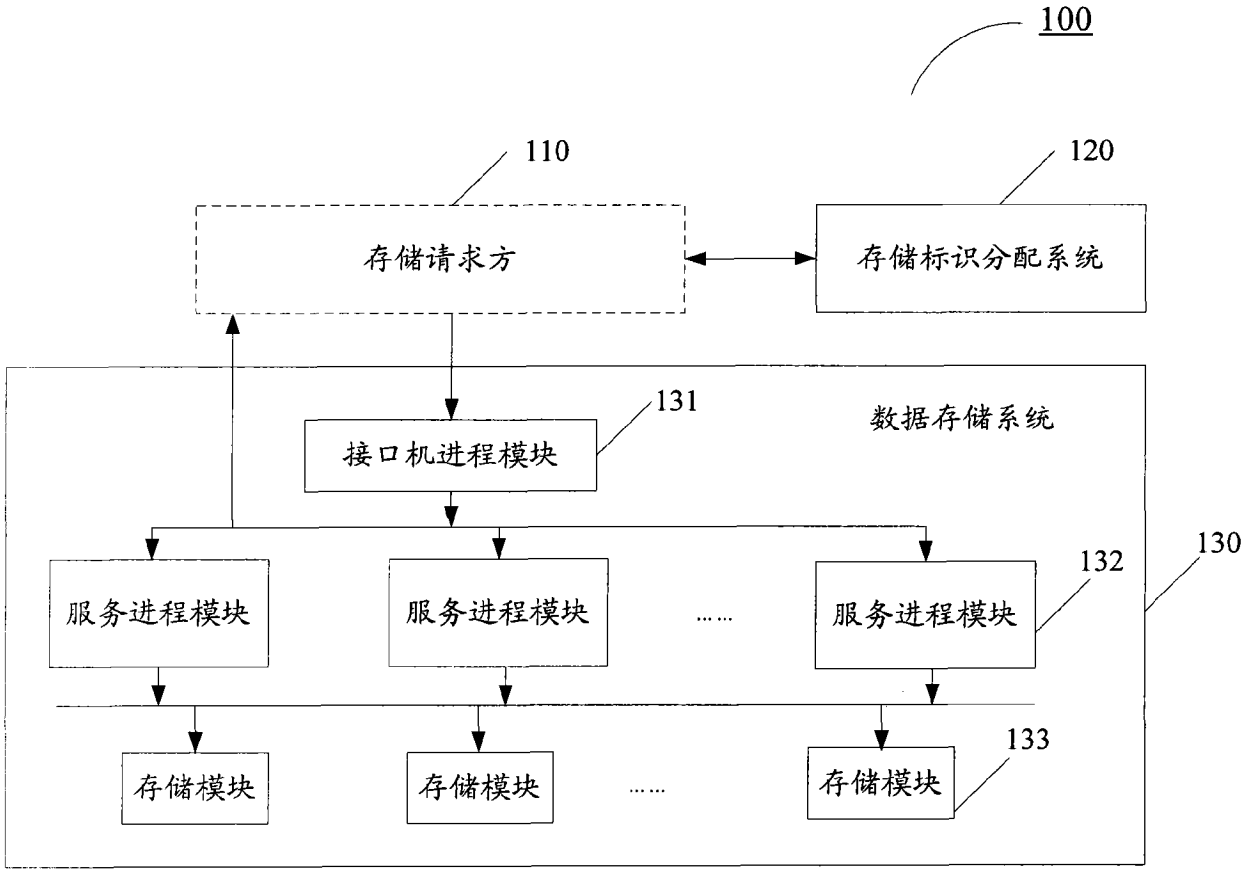


图 1（现有技术）

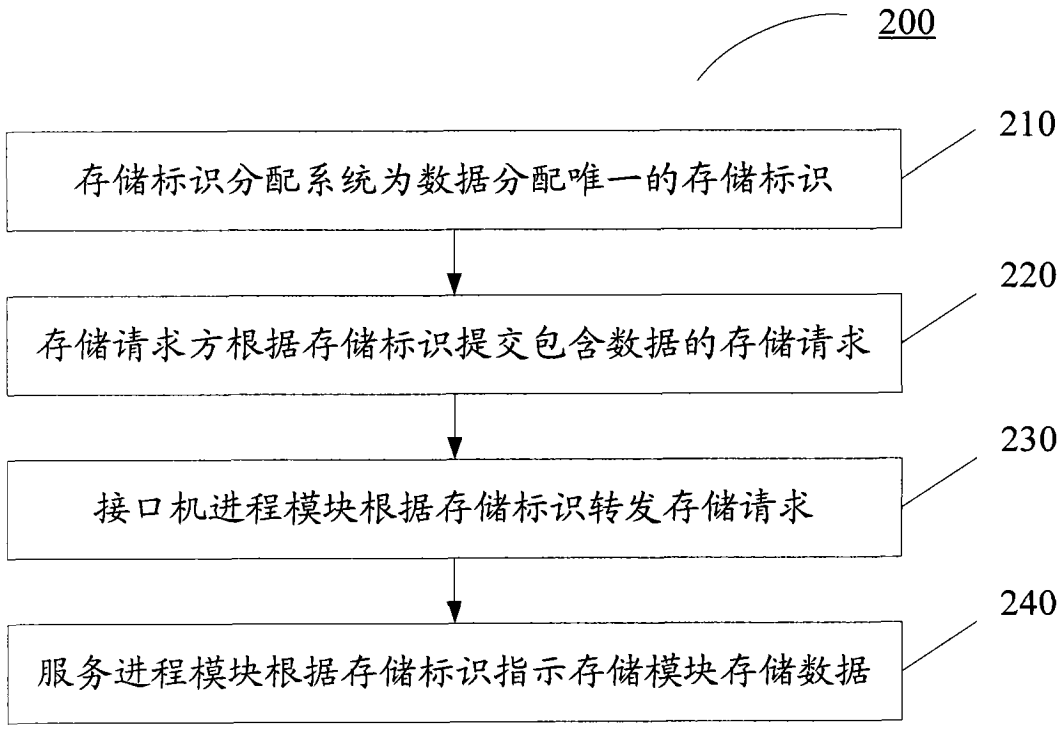


图 2（现有技术）

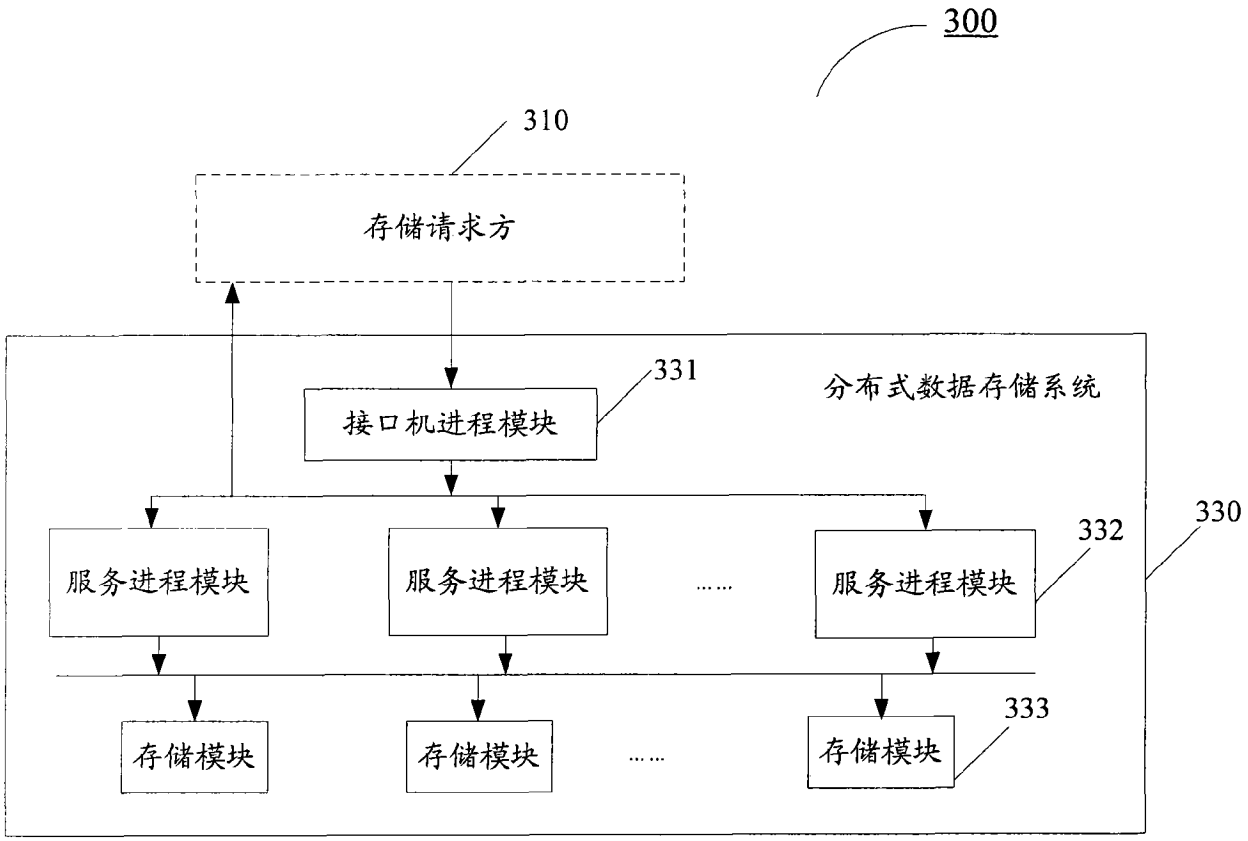

300
310
存储请求方
331
接口机进程模块
分布式数据存储系统
330
332
服务进程模块
服务进程模块
……
服务进程模块
333
存储模块
存储模块
……
存储模块

图 3

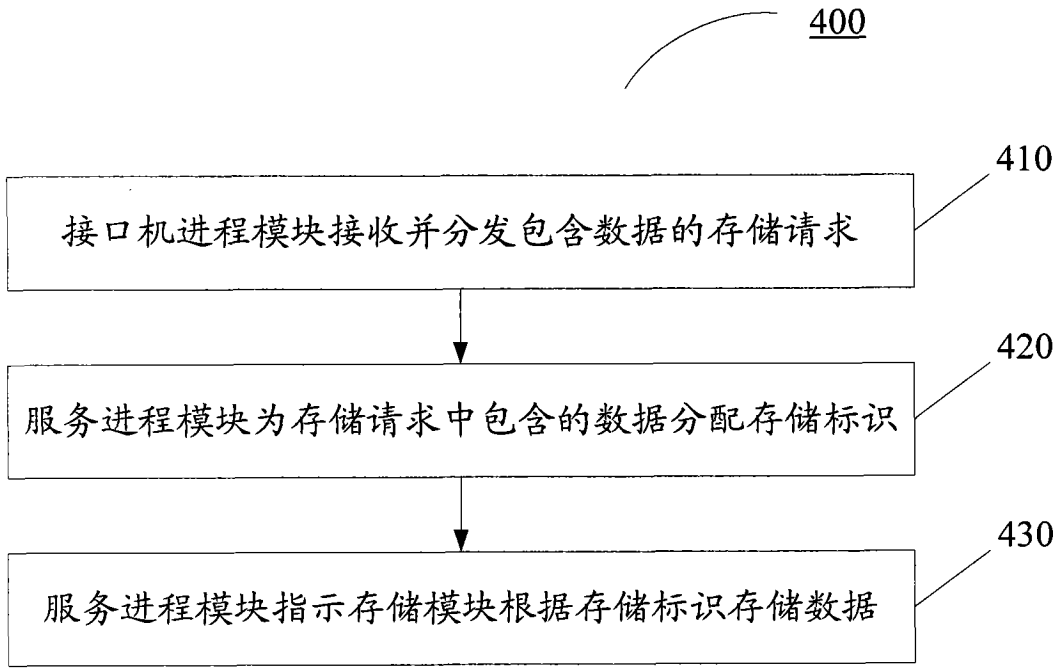

400
410
接口机进程模块接收并分发包含数据的存储请求
420
服务进程模块为存储请求中包含的数据分配存储标识
430
服务进程模块指示存储模块根据存储标识存储数据

图 4

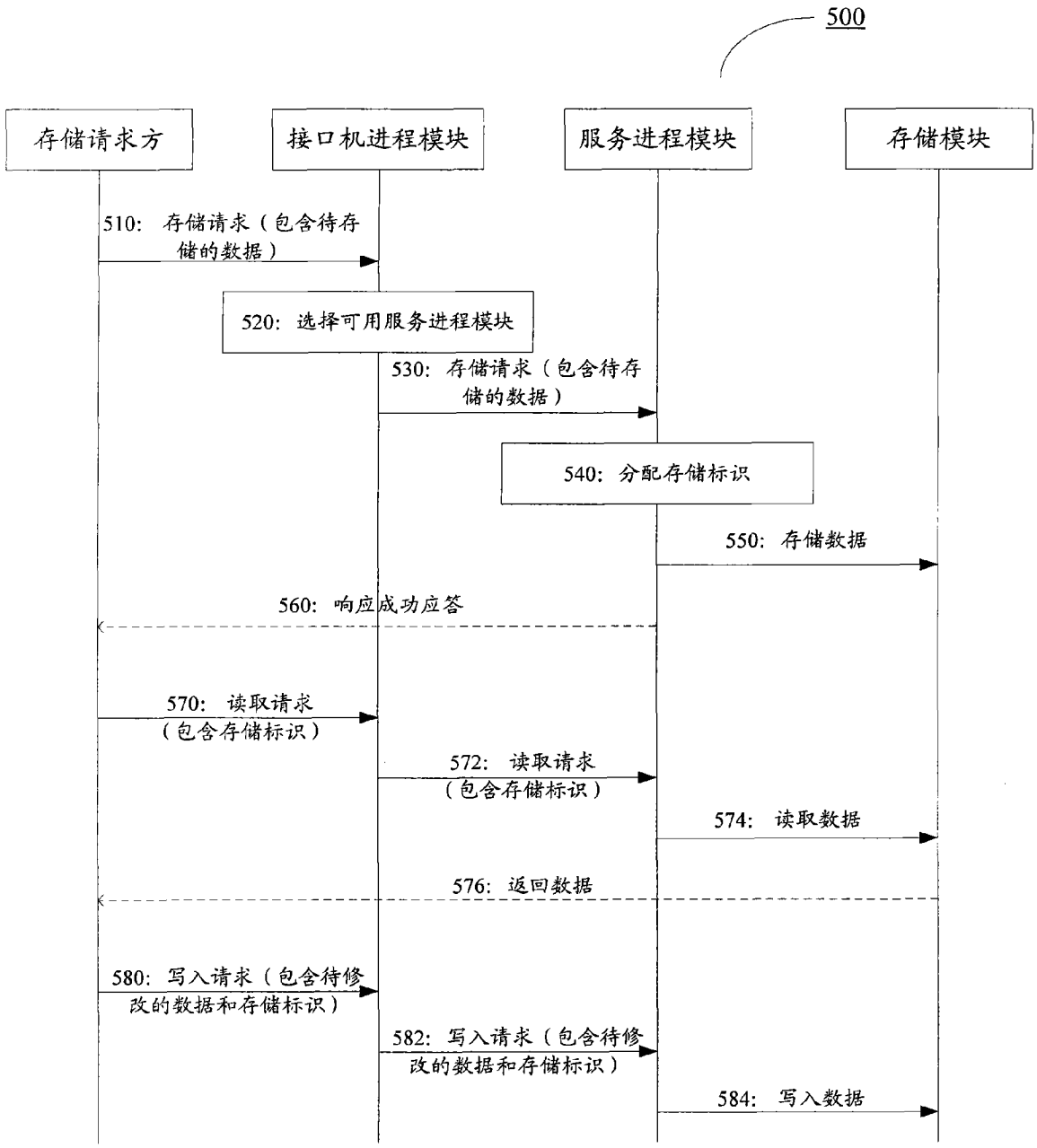

500
存储请求方
接口机进程模块
服务进程模块
存储模块
510: 存储请求（包含待存储的数据）
520: 选择可用服务进程模块
530: 存储请求（包含待存储的数据）
540: 分配存储标识
550: 存储数据
560: 响应成功应答
570: 读取请求（包含存储标识）
572: 读取请求（包含存储标识）
574: 读取数据
576: 返回数据
580: 写入请求（包含待修改的数据和存储标识）
582: 写入请求（包含待修改的数据和存储标识）
584: 写入数据

图 5

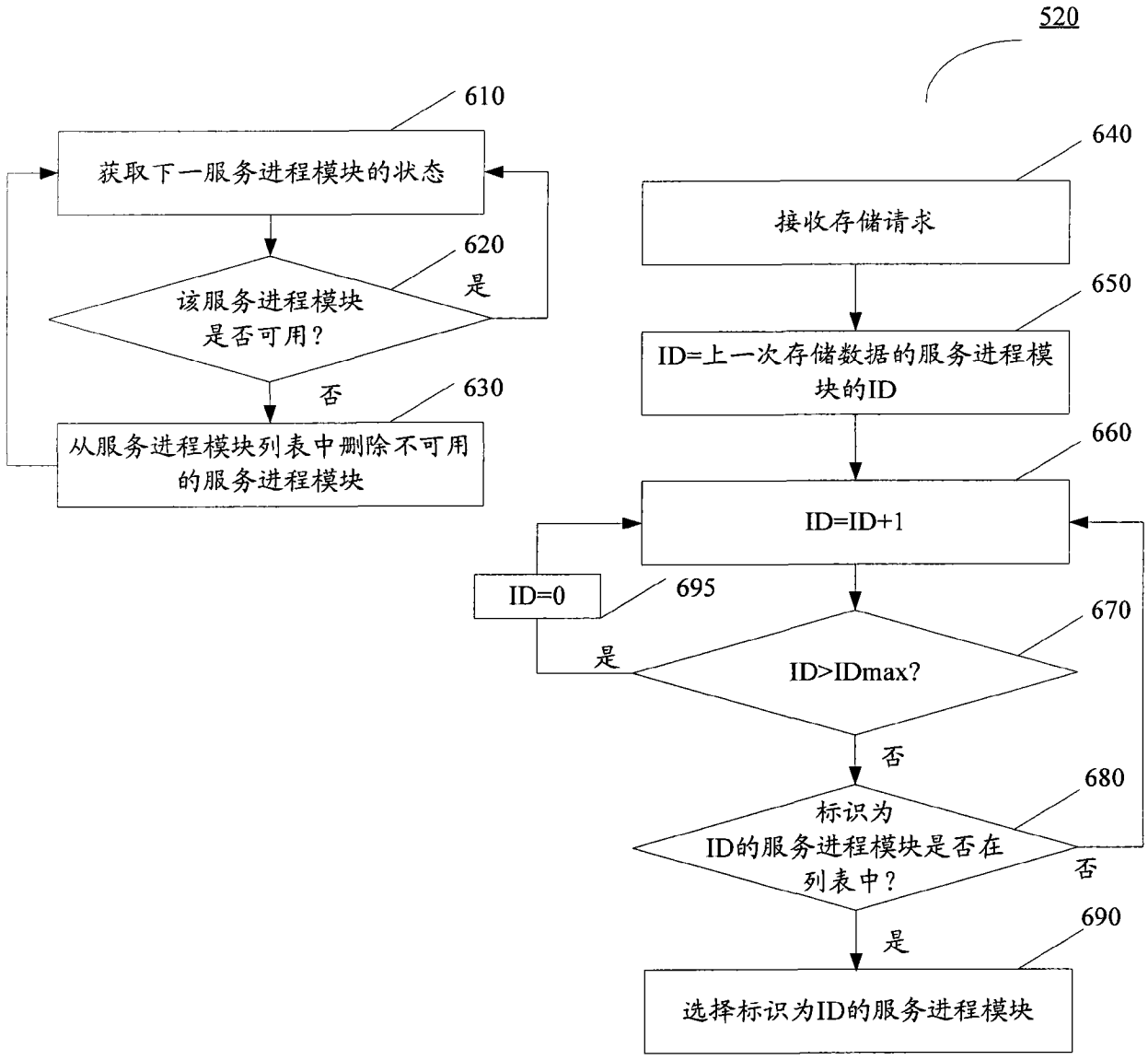

520
610
获取下一服务进程模块的状态
620
该服务进程模块是否可用?
是
否
630
从服务进程模块列表中删除不可用的服务进程模块
640
接收存储请求
650
ID=上一次存储数据的服务进程模块的ID
660
ID=ID+1
695
ID=0
是
670
ID>IDmax?
否
680
标识为ID的服务进程模块是否在列表中?
否
是
690
选择标识为ID的服务进程模块

图 6

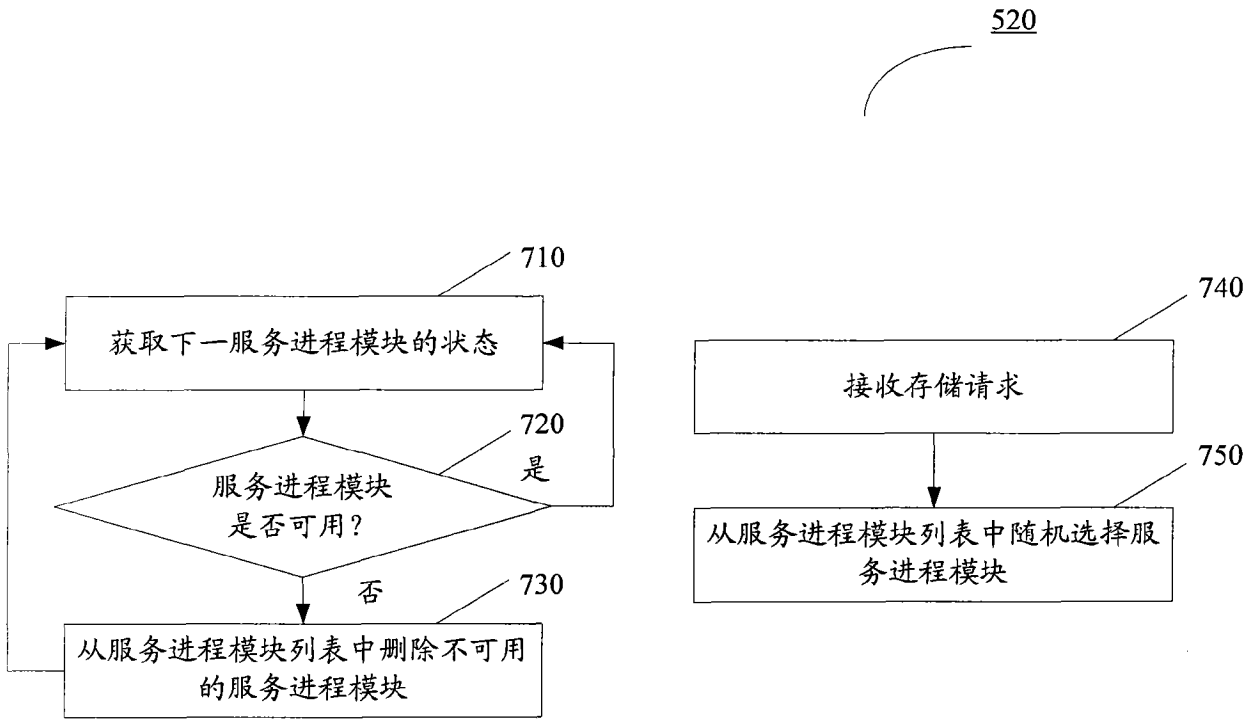

520
710
获取下一服务进程模块的状态
720
服务进程模块是否可用？
是
否
730
从服务进程模块列表中删除不可用的服务进程模块
740
接收存储请求
750
从服务进程模块列表中随机选择服务进程模块

图 7