(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 104679796 A
(43)申请公布日 2015.06.03

(21)申请号 201310642613.0

(22)申请日 2013.12.03

(71)申请人 方正信息产业控股有限公司
地址 100871 北京市海淀区成府路298号中关村方正大厦6层
申请人 上海方正数字出版技术有限公司

(72)发明人 刘慧娟 王浩 郭春庭 郑程光

(74)专利代理机构 北京银龙知识产权代理有限公司 11243
代理人 许静 安利霞

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 11/16(2006.01)
H04L 29/06(2006.01)

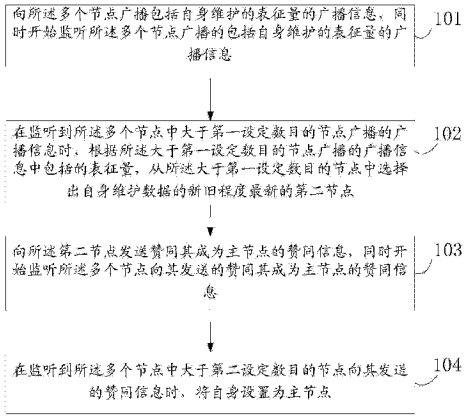
权利要求书3页 说明书10页 附图3页

(54)发明名称

一种选举方法、装置及数据库镜像集群节点

(57)摘要

本发明实施例提供一种选举方法、装置及数据库镜像集群节点。方法包括：向多个节点广播包括自身维护的表征量的广播信息，同时开始监听多个节点广播的包括自身维护的表征量的广播信息；在监听到多个节点中大于第一设定数目的节点广播的广播信息时，根据大于第一设定数目的节点广播的广播信息中包括的表征量，从大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点；向第二节点发送赞同其成为主节点的赞同信息，同时开始监听多个节点向其发送的赞同其成为主节点的赞同信息；在监听到多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点。本发明实施例避免了数据的丢失。



CN 104679796 A

1. 一种选举方法，其特征在于，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述方法包括：

向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

2. 根据权利要求 1 所述的方法，其特征在于，还包括：

向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；

当所述判断结果为否时，进入所述向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息的步骤。

3. 根据权利要求 1 所述的方法，其特征在于，所述多个节点中每个节点自身维护的数据中包括日志，所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

所述在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

4. 根据权利要求 1 所述的方法，其特征在于，所述根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

5. 根据权利要求 4 所述的方法，其特征在于，所述选择优先级由节点的镜像类型来表

征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；或者，

所述选择优先级由节点的 IP 地址和监听端口号组成的字符串来表征，其中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高；或者，

所述选择优先级由节点的镜像类型、和 IP 地址和监听端口号组成的字符串来表征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；镜像类型相同的至少两个节点中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高。

6. 一种选举装置，其特征在于，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述装置包括：

广播及监听模块，用于向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

选择模块，用于在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

发送及监听模块，用于向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

设置模块，用于在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

7. 根据权利要求 6 所述的装置，其特征在于，还包括：

监听模块，用于向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

判断模块，用于判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；当所述判断结果为否时，进入所述广播及监听模块。

8. 根据权利要求 6 所述的装置，其特征在于，所述多个节点中每个节点自身维护的数据中包括日志，所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

所述选择模块包括：

第一选择单元，用于在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节

点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

9. 根据权利要求 6 所述的装置，其特征在于，所述选择模块包括：

第二选择单元，用于根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

第三选择单元，用于在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

10. 一种数据库镜像集群节点，其特征在于，包括如权利要求 6 至 10 中任一权利要求所述的选举装置。

一种选举方法、装置及数据库镜像集群节点

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种选举方法、装置及数据库镜像集群节点。

背景技术

[0002] 随着数据量的不断膨胀，以及商业智能(BI)在企业内的快速发展，BI 用户对信息分析平台的访问频率和查询复杂度也快速提升，因此要求相应的数据库对海量数据高并发查询进行支持，因而采取 MPP (Massive Parallel Processing)架构的数据库系统是非常有益的，在 MPP 中增加节点就可以线性提高系统的数据存储能力和数据处理能力。

[0003] 在 MPP 架构中，为了保证高可用，对每个处理单元提供数据库层的镜像机制保护，当然这个处理单元不仅仅能提供高可用，同时可以提供更多的查询能力，也可以通过镜像获取数据库的备份以便对数据进行分析。当然，每个处理单元也是一个单独的集群(我们称该集群为数据库镜像集群)，也是一个主(Master)节点多个从(Slave)节点的集群，因而需要选取唯一 Master 节点的方法。

[0004] 数据库镜像集群选取唯一的 Master 节点与一般的集群选取唯一的 Master 节点有不同的地方，因为镜像集群的功能决定了该集群可能有多个节点，为了不影响集群的写数据的性能，只能少部分的节点与 Master 节点是同步的，大部分的节点是异步的，所以一定要选取当前数据最新的节点作为 Master 节点，否则会导致数据丢失。

[0005] 现有技术中作为 Master 的节点通常是随机选择的，这就可能选择出所维护数据最旧的数据库节点作为 Master 节点，在这种情况下，当其它比该 Master 节点维护数据更新的节点以该 Master 节点上的数据为准来维护数据时，就会发生一部分数据的丢失，从而引起用户使用上的不便。这种不便例如：数据丢失了，也就会发生某个用户已经提交的事务，像根本没在这个系统中操作过，在用户看起来就是，他提交的事务，被回滚了。

发明内容

[0006] 有鉴于此，本发明实施例的目的是提供一种选举方法、装置及数据库镜像集群节点，以避免因选择所维护数据最旧的节点作为唯一主节点而造成数据丢失，提升用户体验。

[0007] 为解决上述技术问题，本发明实施例提供方案如下：

[0008] 本发明实施例提供一种选举方法，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述方法包括：

[0009] 向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

[0010] 在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

[0011] 向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

[0012] 在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

[0013] 优选地，还包括：

[0014] 向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

[0015] 判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；

[0016] 当所述判断结果为否时，进入所述向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息的步骤。

[0017] 优选地，所述多个节点中每个节点自身维护的数据中包括日志，所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

[0018] 所述在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

[0019] 在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

[0020] 优选地，所述根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

[0021] 根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

[0022] 在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

[0023] 优选地，所述选择优先级由节点的镜像类型来表征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；或者，

[0024] 所述选择优先级由节点的 IP 地址和监听端口号组成的字符串来表征，其中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高；或者，

[0025] 所述选择优先级由节点的镜像类型、和 IP 地址和监听端口号组成的字符串来表征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；镜像类型相同的至少两个节点中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高。

[0026] 本发明实施例还提供一种选举装置，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述装置包括：

[0027] 广播及监听模块，用于向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

[0028] 选择模块，用于在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

[0029] 发送及监听模块，用于向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

[0030] 设置模块，用于在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

[0031] 优选地，还包括：

[0032] 监听模块，用于向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

[0033] 判断模块，用于判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；当所述判断结果为否时，进入所述广播及监听模块。

[0034] 优选地，所述多个节点中每个节点自身维护的数据中包括日志，所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

[0035] 所述选择模块包括：

[0036] 第一选择单元，用于在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

[0037] 优选地，所述选择模块包括：

[0038] 第二选择单元，用于根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

[0039] 第三选择单元，用于在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

[0040] 本发明实施例还提供一种包括以上所述的选举装置的数据库镜像集群节点。

[0041] 从以上所述可以看出，本发明实施例至少具有如下有益效果：

[0042] 通过数据库镜像集群中的多个节点中的第一节点根据监听到的所述多个节点中大于第一设定数目的节点广播的表征量，从所述至少两个节点中选择出自身维护数据的新旧程度最新的第二节点，向其发送赞同信息，同时开始监听所述多个节点发送的赞同信息，

在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，又第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值，则所述第一节点监听到大于所述第二设定数目的节点向其发送的赞同信息时，也就表明所述多个节点中的其它节点不可能监听到所述设定数目的节点向其发送的赞同信息，从而不可能将自身设置为主节点，即这种情况下所述多个节点中只有所述第一节点将自身设置为主节点，从而实现了选举出唯一主节点，又赞同信息是向选择出的大于设定数目的节点中自身维护数据的新旧程度最新的节点发送的，且大于第一设定数目的节点广播的消息中包括的表征量不完全相同、亦即选择最新的节点所依据的节点所维护数据新旧程度不完全相同，从而所述多个节点中存在比成为主节点的节点所维护数据更旧的节点，从而保证了将所述多个节点中自身维护数据并非最旧的节点选举为唯一主节点，与现有技术相比，避免了因选择所维护数据最旧的节点作为唯一主节点而造成的数据的丢失，提升了用户体验。

附图说明

[0043] 图 1 表示本发明实施例提供的一种选举方法的步骤流程图；

[0044] 图 2 表示本发明实施例的较佳实施方式提供的一种在数据库镜像集群中选取唯一 Master 节点的方法的步骤流程图；

[0045] 图 3 表示本发明实施例提供的一种选举装置的结构示意图。

具体实施方式

[0046] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合附图及具体实施例对本发明实施例进行详细描述。

[0047] 图 1 表示本发明实施例提供的一种选举方法的步骤流程图，参照图 1，本发明实施例提供一种选举方法，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述方法包括如下步骤：

[0048] 步骤 101，向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

[0049] 步骤 102，在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

[0050] 步骤 103，向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

[0051] 步骤 104，在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

[0052] 可见，通过数据库镜像集群中的多个节点中的第一节点根据监听到的所述多个节点中大于第一设定数目的节点广播的表征量，从所述至少两个节点中选择出自身维护数据

的新旧程度最新的第二节点，向其发送赞同信息，同时开始监听所述多个节点发送的赞同信息，在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，又第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值，则所述第一节点监听到大于所述设定数目的节点向其发送的赞同信息时，也就表明所述多个节点中的其它节点不可能监听到所述设定数目的节点向其发送的赞同信息，从而不可能将自身设置为主节点，即这种情况下所述多个节点中只有所述第一节点将自身设置为主节点，从而实现了选举出唯一主节点，又赞同信息是向选择出的大于设定数目的节点中自身维护数据的新旧程度最新的节点发送的，且大于第一设定数目的节点广播的消息中包括的表征量不完全相同、亦即选择最新的节点所依据的节点所维护数据新旧程度不完全相同，从而所述多个节点中存在比成为主节点的节点所维护数据更旧的节点，从而保证了将所述多个节点中自身维护数据并非最旧的节点选举为唯一主节点，与现有技术相比，避免了因选择所维护数据最旧的节点作为唯一主节点而造成的数据的丢失，提升了用户体验。

[0053] 其中，节点自身维护数据的新旧程度可以通过各种方式来表征，例如：自身维护的最新日志的序号，自身的镜像类型等。其中，自身维护的最新日志的序号越大表明自身维护数据越新。自身的镜像类型为同步镜像的节点自身维护数据比自身的镜像类型为异步镜像的节点新。

[0054] 〈方式一〉

[0055] 对于通过最新日志的序号来表征的情况，所述多个节点中每个节点自身维护的数据中可以包括日志，相应地，可以有：

[0056] 所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

[0057] 所述在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

[0058] 在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

[0059] 〈方式二〉

[0060] 对于通过镜像类型来表征的情况，所述多个节点中只包括一个同步镜像节点，相应地，可以有：

[0061] 所述每个节点自身维护的表征量为该节点的镜像类型；

[0062] 所述在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

[0063] 从所述大于第一设定数目的节点中选择出镜像类型为同步镜像的所述第二节点。

[0064] 这种情况下，由所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同，也即所述大于第一设定数目的节点广播的消息中包括的镜像类型不完全相同，于是选择出的节点的镜像类型就是同步镜像，从而选举出的唯一主节点是镜像类型为同步镜像的节点，也就是说，选举出的唯一主节点所维护数据在所述多个节点中是最新的。

[0065] 在本发明实施例中，数目的二分之一的向下取整值是指，对数目的二分之一进行

向下取整后得到的值。例如：10 的二分之一的向下取整值为 5，17 的二分之一的向下取整值为 8。

[0066] 在本发明实施例中，考虑到所述大于第一设定数目的节点中自身维护数据的新旧程度最新的节点的数目可能大于一个，于是可以有：

[0067] 所述根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点包括：

[0068] 根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

[0069] 在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

[0070] 其中，所述选择优先级由节点的镜像类型来表征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；或者，

[0071] 所述选择优先级由节点的 IP 地址和监听端口号组成的字符串来表征，其中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高；或者，

[0072] 所述选择优先级由节点的镜像类型、和 IP 地址和监听端口号组成的字符串来表征，其中，镜像类型为同步镜像的节点的选择优先级比镜像类型为异步镜像的节点的选择优先级高；镜像类型相同的至少两个节点中，节点的 IP 地址和监听端口号组成的字符串映射成的数值越大表明节点的选择优先级越高，或者，节点的 IP 地址和监听端口号组成的字符串映射成的数值越小表明节点的选择优先级越高。

[0073] 此外，所述选择优先级也可以由用户预设，例如，硬件配置较好的节点对应的选择优先级较高。

[0074] 所述多个节点各自对应的选择优先级均可以预先配置给所述多个节点中的每个节点；或者，所述多个节点中每个节点对应的选择优先级可以预先配置给该节点，再由该节点广播给其它节点，例如，可以有：

[0075] 所述多个节点中每个节点广播的包括自身维护的表征量的广播信息中包括该节点对应的选择优先级；

[0076] 所述所有节点各自对应的选择优先级由所述第一节点从所述所有节点广播的广播信息中分别解析出来。

[0077] 在本发明实施例中，所述第一节点可以通过与所述多个节点中其它每个节点分别建立的单个 TCP 连接来向该节点发送自身维护的表征量，通过自身内部的模块调用来向自身发送自身维护的表征量，通过这种方式，来实现向所述多个节点广播包括自身维护的表征量的广播信息。

[0078] 在本发明实施例中，所述向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息的步骤，可以由所述第一节点在初始化配置完成时执行，或者，也可以由所述第一节点在检测到所述数据库镜像集群中主节点运行出现问题时执行。对于后者，具体可以有：

[0079] 所述方法还包括：

[0080] 向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

[0081] 判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；

[0082] 当所述判断结果为否时，进入所述向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息的步骤。

[0083] 为将本发明实施例阐述得更加清楚明白，下面提供本发明实施例的较佳实施方式。

[0084] 本较佳实施方式提供一种在数据库镜像集群中选取唯一、数据最新、性能好的Master 节点的方法。

[0085] 在 MPP 架构中，数据库镜像集群是一个 Master 节点多个 Slave 的集群，对于需要选取唯一 Master 节点的场景，数据库镜像集群选取唯一的 Master 节点与一般的集群选取唯一的 Master 节点有不同的地方，因为镜像集群的功能决定了该集群可能有多个节点，为了不影响集群的写数据的性能，只能少部分的节点与 Master 节点是同步的，大部分的节点是异步的，所以一定要选取当前数据最新的节点作为 Master 节点，否则会导致数据丢失。另外，数据库集群环境中主节点的性能，决定集群写数据的性能，因而尽可能的保证主节点的性能好。本较佳实施方式就是针对上述数据库复制集群的这些特点，设计的一种专门的选举方法。这里的同步是指同步镜像，异步是指异步镜像。

[0086] 集群中的 Master 节点，在没有信息需要同步给 Slave 节点时，则每隔一定时间间隔 T0 向所有的节点广播心跳信息，以便集群中的所有节点能够知道 Master 节点的运行状态良好。

[0087] 集群中的 Slave 节点，在给定的时间 T1（>T0）内没有收到 Master 节点的心跳信息，则开始广播自己的一些信息，包括数据的新旧程度，机器硬件配置(配置好的可以用 3 表示，配置中等的用 2 表示，配置差的用 1 表示)，IP 地址和监听端口字符串。同时也会收到其它 Slave 节点广播的自己的一些信息，等待一定的时间 T2，收到 >Max（N-N1-1，1）(Max 是取二者中较大的值，N 是集群中节点的总数，N1 是集群中宕机的节点数目，这是为了让这些节点能够选出相同的数据最新的节点）个节点的信息，则根据相应的算法(该算法一定要保证，相同的信息无论在哪个节点上运行都一定能选出的相同的最优节点）选出最优的节点，并向该节点发送信息，表示赞同该节点成为新的Master节点。在一定的时间T3内，收到 >N/2 赞同信息的 Slave 节点(这是为了保证选举出来的 Master 节点是唯一的，并且选取最优节点算法的特性也保证了正常情况下，一定有某一 Slave 节点会收到 >N/2 的赞同信息)，将自己的角色切换为 Master 节点。详细流程见图 2。

[0088] 参见图 2，该在数据库镜像集群中选取唯一 Master 节点的方法包括如下步骤：

[0089] 步骤 201，节点当前状态为从节点，在接收到主节点周期性广播的心跳信息之后，进入步骤 202；

[0090] 步骤 202，等待时间 T1；

[0091] 步骤 203，判断是否收到主节点广播的心跳信息，如果是，则返回步骤 201，即保持从节点的角色；否则，进入步骤 204；

[0092] 步骤 201 ～ 202 可以替换为 :节点初始化完成后,等待时间 T1 ;相应地,步骤 203 可以替换为 :判断是否收到主节点广播的心跳信息,如果是,则将自身状态设置为从节点,然后进入步骤 202。

[0093] 步骤 204,该节点广播自身的相关信息并等待时间 T2 ;

[0094] 步骤 205,判断是否收到数目大于 Max (N-N1-1,1) 的广播信息,如果是,则进入步骤 206 ;否则,进入步骤 210 ;

[0095] 步骤 206,计算最优节点,并向最优节点发送赞同信息,等待时间 T3 ;

[0096] 步骤 207,判断是否收取了数目大于 N/2 的赞同信息,如果是,则进入步骤 208 ;否则,进入步骤 209 ;

[0097] 步骤 208,该节点设置自己的角色为 Master 节点,并每隔时间 T0 广播心跳信息 ;

[0098] 步骤 209,等待时间 T4,返回步骤 203 ;

[0099] 步骤 210,等待时间 T3,进入步骤 207。

[0100] 对于上节中描述的选取唯一 Master 节点的方法,整个选举方法的流程是很清晰的,但是选举方法中涉及到的一些难点,在具体实施过程中需要留意,具体如下 :

[0101] 1. 节点之间如何通信 ;

[0102] 2. 节点上的数据的新旧程度如何确定 ;

[0103] 3. 选出最优节点的算法如何实现 ;

[0104] 下面将对这三点进行详细的描述。

[0105] 对于难点 1,不同节点之间的通信可以通过 TCP 协议来完成,首先在每个节点上有一个监听其它节点连接的端口,还要与其它所有节点分别建立连接。当需要向某节点发送信息时,通过先前建立的连接去发送即可 ;当需要发送广播信息时,则遍历所有的有效连接,通过所有的连接向所有的节点发送信息即可完成广播的功能。系统中,还需要维护一个专门监听这些连接状态的模块,当连接断开时,间隔固定的时间主动进行重连,重连如果失败,则间隔二倍于之前的时间(直到达到某个最大值)再进行重连。如果这两个节点同时与对方重连,并且建立了两个连接(认为发生冲突),则关闭这两个连接,各自分别等待随机的时间再重连,直至冲突解决,因为随机时间不会一直相同,并且当一个节点发现与另一个节点之间的连接已经存在了,则不再重连直到该连接断开,所以可以保持节点之间两两保持一个 TCP 的长连接。同一节点自身的通信可以通过内部模块调用来实现。

[0106] 对于难点 2,集群中的 Master 节点,对数据库中的每条日志用一个 128bits 的数字来进行编号,且该编号是单调递增的(该数字足够大,目前不考虑用完的情形),将该编号称为日志序号。镜像集群中 Master 节点向所有 Slave 节点同步日志和日志序号,Slave 节点根据日志(redo)完成镜像并保留与 Master 节点一致的日志序号。此时便可以通过当前节点上最新的日志序号来描述当前节点的数据新旧程度,最新的日志序号越大,说明数据越新。

[0107] 对于难点 3,如果数据最新的 Slave 节点只有一个,那么认为该节点是最优节点。如果不止一个的话,那么认为硬件配置最优(其中,“配置好”优于“配置中”,而“配置中”优于“配置差”)的节点是最优节点。若此时还是能选出多个节点,那么根据 Slave 节点的 IP 地址和监听选举信息的端口号来选取最优节点,因为这些节点的 IP 地址和端口号连接的字符串在集群内部一定是唯一的,所以按字符串比较的方式,一定能选出唯一的最优节点,

而且无论该算法运行在任何的 Slave 节点上，只要收到的广播信息是一致的，那么选出来的最优节点也一定是相同的。

[0108] 其中，IP 地址和端口号连接的字符串，举例来说，假设 IP 地址是 192.168.0.1，端口号是 2222，那么 IP 地址和端口号连接的字符串就是“192.168.0.1:2222”，这个字符串在集群内部一定是唯一的。字符串比较的方法，比如有两个 IP 地址和端口号连接的字符串“192.168.0.1:2222”和“192.168.0.2:2222”，就是按照普通的字符串比较的方法来进行比较，认为“192.168.0.1:2222”小于“192.168.0.2:2222”。IP 地址和端口连接的字符串越小，认为该节点越优。

[0109] “只要收到的广播信息是一致的”具体是指收到的广播信息的集合是一致的，比如节点 1 收到来自于节点 1,2,3 的广播信息，节点 2 也收到来自于节点 1,2,3 的广播信息，那么节点 1 上收到的广播信息的集合和节点 2 上收到的广播信息的集合就是一样的。那么，节点 1 和节点 2 根据各自收到的广播信息的集合，再利用上面描述的算法，节点 1 和节点 2 选出的最优节点是一样的，它们都会认为某一个节点(例如，节点 3) 是最优的。

[0110] 本较佳实施方式不仅提供了一种选举唯一 Master 节点的方法，而且可以选举出一个数据最新的硬件配置好数据库节点作为 Master 节点，这样不仅避免了普通的选举方法在集群中所有的活着的节点中随机选择一个节点作为 Master 节点，带来的数据丢失，已经提交成功的事务所做的操作被回滚了之类的现象，还选取出硬件配置好的节点作为 Master 节点，从而提升了整个数据库镜像集群的写性能。

[0111] 如果选举的结点不是最新的数据库节点，那个这个节点上的数据是比最新的数据库上的节点是滞后的，但是一旦新选举出来的节点，成为 Master 节点，那么它将会把自己的数据作为标准，其它节点的都要以它为参照，也就是整个集群的数据是以新的 Master 节点的数据为准的，而新的 Master 节点的数据是滞后的，自然会有一部分数据丢失了。数据丢失了，也就会发生某个用户已经提交的事务，像根本没在这个系统中操作过，在用户看起来就是，他提交的事务，被回滚了。

[0112] 图 3 表示本发明实施例提供的一种选举装置的结构示意图，参照图 3，本发明实施例还提供一种选举装置，用于数据库镜像集群中的多个节点中的第一节点，所述多个节点中每个节点自身均维护一用于表征自身维护数据的新旧程度的表征量，所述装置包括：

[0113] 广播及监听模块 301，用于向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息；

[0114] 选择模块 302，用于在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点，其中，所述第一设定数目大于或等于一个节点，所述大于第一设定数目的节点广播的消息中包括的表征量不完全相同；

[0115] 发送及监听模块 303，用于向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息；

[0116] 设置模块 304，用于在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，其中，所述第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值。

[0117] 可见，通过数据库镜像集群中的多个节点中的第一节点根据监听到的所述多个节点中大于第一设定数目的节点广播的表征量，从所述至少两个节点中选择出自身维护数据的新旧程度最新的第二节点，向其发送赞同信息，同时开始监听所述多个节点发送的赞同信息，在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点，又第二设定数目大于或等于所述多个节点的数目的二分之一的向下取整值，则所述第一节点监听到大于所述设定数目的节点向其发送的赞同信息时，也就表明所述多个节点中的其它节点不可能监听到所述设定数目的节点向其发送的赞同信息，从而不可能将自身设置为主节点，即这种情况下所述多个节点中只有所述第一节点将自身设置为主节点，从而实现了选举出唯一主节点，又赞同信息是向选择出的大于设定数目的节点中自身维护数据的新旧程度最新的节点发送的，且大于第一设定数目的节点广播的消息中包括的表征量不完全相同、亦即选择最新的节点所依据的节点所维护数据新旧程度不完全相同，从而所述多个节点中存在比成为主节点的节点所维护数据更旧的节点，从而保证了将所述多个节点中自身维护数据并非最旧的节点选举为唯一主节点，与现有技术相比，避免了因选择所维护数据最旧的节点作为唯一主节点而造成的数据的丢失，提升了用户体验。

[0118] 进一步地，还可以包括：

[0119] 监听模块，用于向所述多个节点广播包括自身维护的表征量的广播信息之前，监听所述数据库镜像集群中主节点以设定周期、周期性地向所述多个节点广播的心跳信息；

[0120] 判断模块，用于判断在监听到所述数据库镜像集群中主节点广播的第一心跳信息之后的设定时长时间内，是否监听到所述数据库镜像集群中主节点广播的下一心跳信息，获取一判断结果，其中，所述设定时长大于所述设定周期；当所述判断结果为否时，进入所述广播及监听模块 301。

[0121] 在本发明实施例中，可以有：

[0122] 所述多个节点中每个节点自身维护的数据中包括日志，所述多个节点中每个节点自身维护的表征量包括该节点自身维护的最新日志的序号；

[0123] 所述选择模块 302 包括：

[0124] 第一选择单元，用于在监听到所述多个节点中大于第一设定数目的节点广播的表征量时，根据所述大于第一设定数目的节点广播的广播信息中包括的序号，从所述至少两个节点中选择出所广播的广播信息中包括的序号最大的所述第二节点。

[0125] 在本发明实施例中，所述选择模块 302 可以包括：

[0126] 第二选择单元，用于根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的所有节点；

[0127] 第三选择单元，用于在所述所有节点的数目大于一个时，从所述所有节点中选择出对应选择优先级最高的所述第二节点。

[0128] 本发明实施例还提供一种数据库镜像集群节点，所述数据库镜像集群节点包括以上所述的选举装置。

[0129] 以上所述仅是本发明实施例的实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明实施例原理的前提下，还可以作出若干改进和润饰，这些改进和润饰也应视为本发明实施例的保护范围。

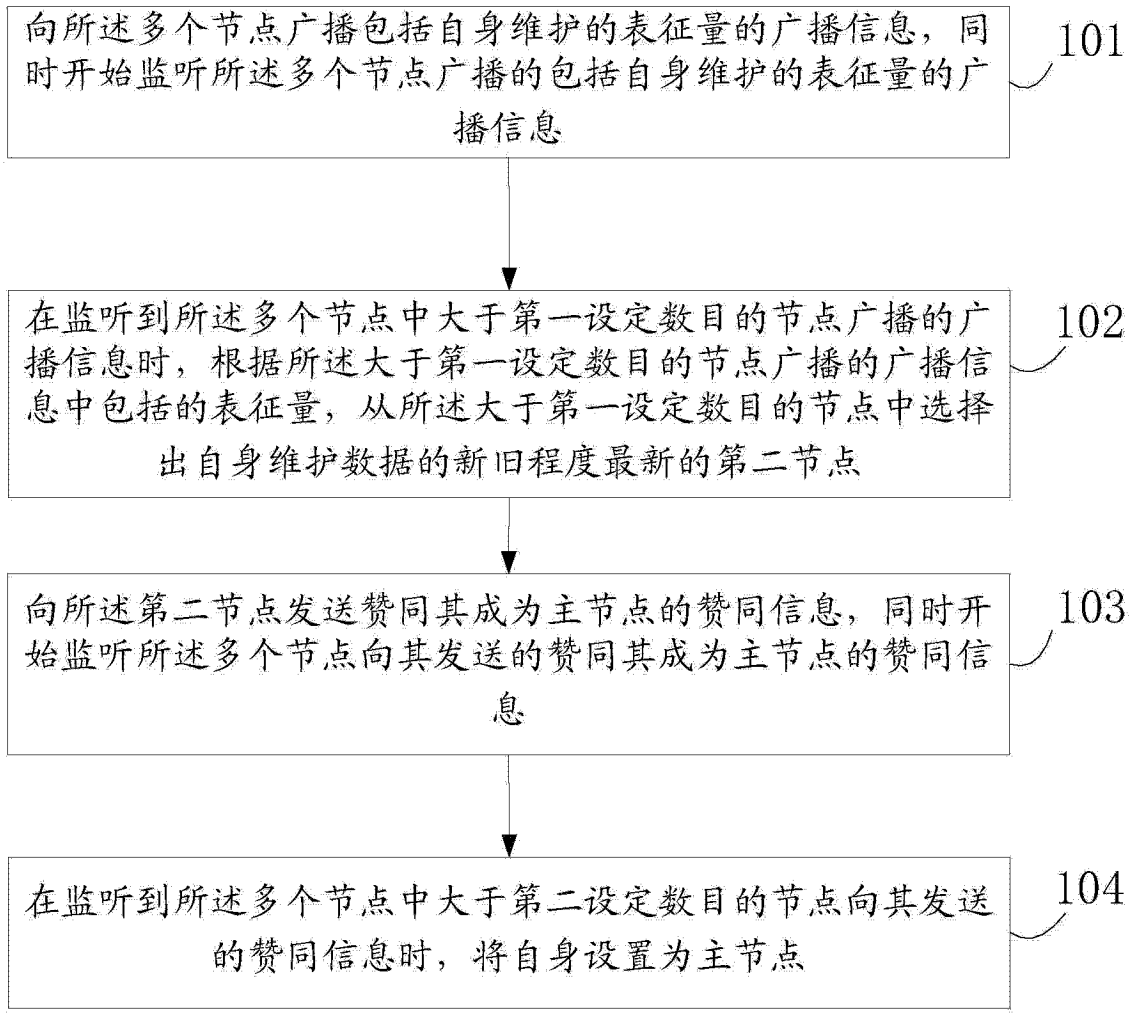

向所述多个节点广播包括自身维护的表征量的广播信息，同时开始监听所述多个节点广播的包括自身维护的表征量的广播信息
101
在监听到所述多个节点中大于第一设定数目的节点广播的广播信息时，根据所述大于第一设定数目的节点广播的广播信息中包括的表征量，从所述大于第一设定数目的节点中选择出自身维护数据的新旧程度最新的第二节点
102
向所述第二节点发送赞同其成为主节点的赞同信息，同时开始监听所述多个节点向其发送的赞同其成为主节点的赞同信息
103
在监听到所述多个节点中大于第二设定数目的节点向其发送的赞同信息时，将自身设置为主节点
104

图 1

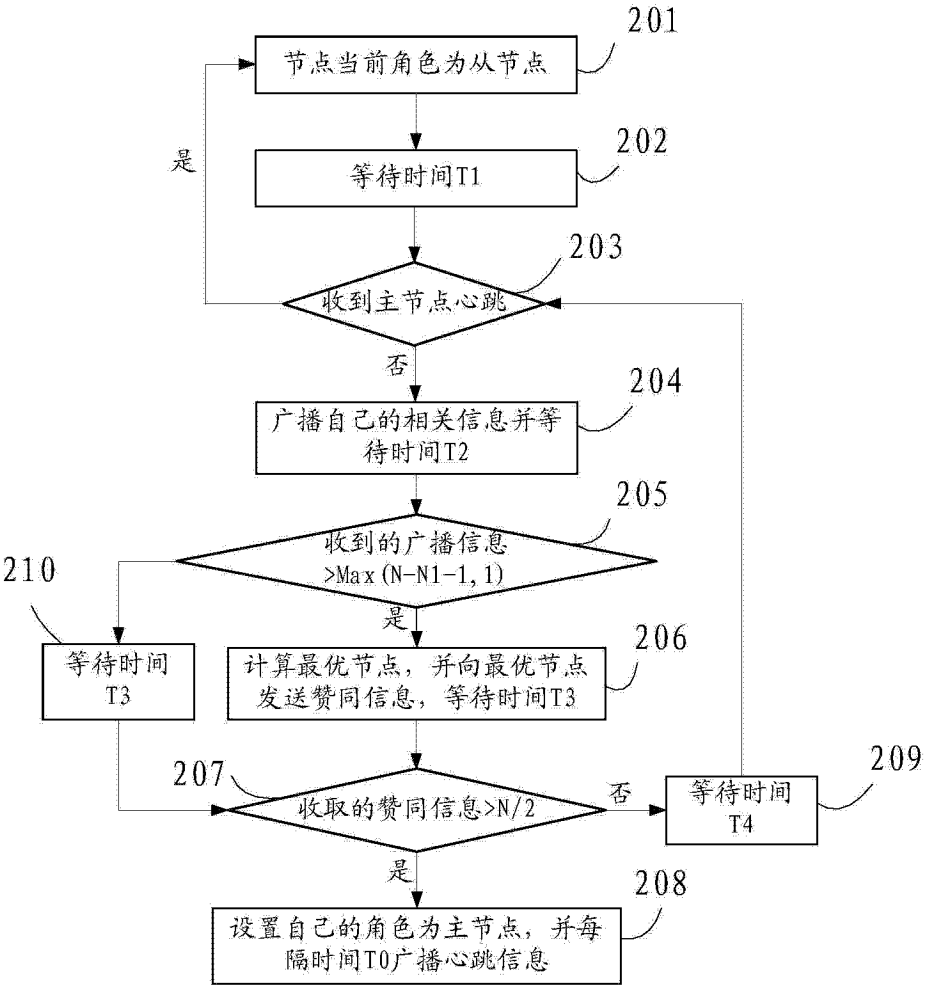

201
节点当前角色为从节点
202
等待时间T1
203
收到主节点心跳
是
否
204
广播自己的相关信息并等待时间T2
205
收到的广播信息>Max(N-N1-1,1)
210
是
等待时间T3
206
计算最优节点，并向最优节点发送赞同信息，等待时间T3
207
收取的赞同信息>N/2
否
209
等待时间T4
是
208
设置自己的角色为主节点，并每隔时间T0广播心跳信息

图 2

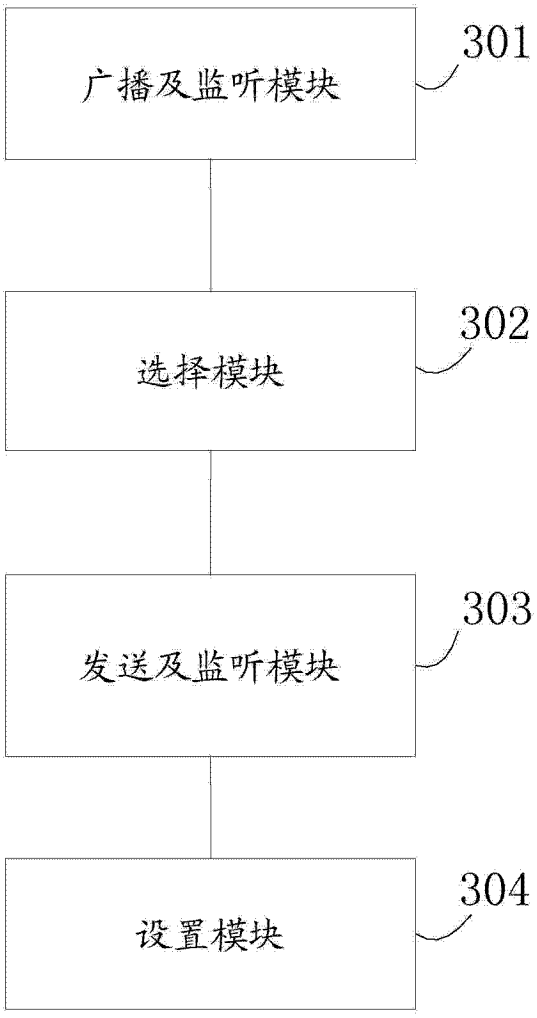


图 3