(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 107105032 B
(45)授权公告日 2019.08.06

(21)申请号 201710262463.9

(22)申请日 2017.04.20

(65)同一申请的已公布的文献号
申请公布号 CN 107105032 A

(43)申请公布日 2017.08.29

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 郭锐 李茂材 梁军 屠海涛 赵琦 王宗友 张建俊 朱大卫 刘斌华

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 朱雅男

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 12/24(2006.01)

(56)对比文件
CN 104933132 A,2015.09.23,
CN 105511987 A,2016.04.20,
CN 105512266 A,2016.04.20,
CN 103152434 A,2013.06.12,
CN 103152434 A,2013.06.12,

审查员 文娟

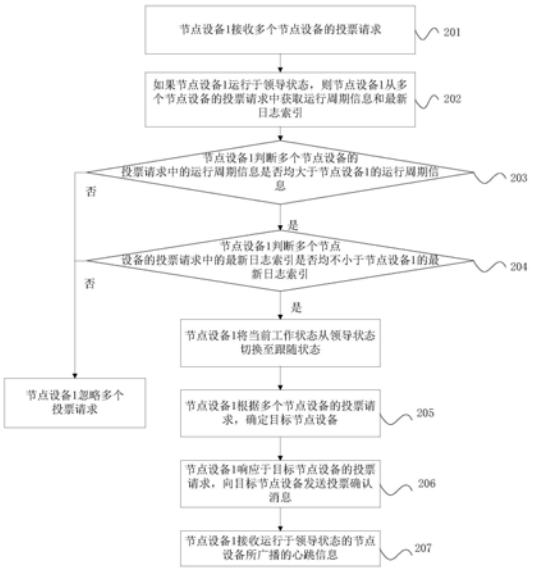
权利要求书2页 说明书10页 附图5页

(54)发明名称

节点设备运行方法及节点设备

(57)摘要

本发明公开了一种节点设备运行方法及节点设备，属于网络技术领域。该方法包括：接收多个节点设备的投票请求，多个节点设备的数量大于系统中节点设备数量的半数；如果当前节点设备运行于领导状态，则从多个节点设备的投票请求中获取运行周期信息和最新日志索引；如果多个节点设备的投票请求中的运行周期信息均大于当前节点设备的运行周期信息，且多个节点设备的投票请求中的最新日志索引均不小于当前节点设备的最新日志索引，将当前工作状态从领导状态切换至跟随状态或候选状态。本发明使得第一子集群可以和第二子集群合为一个系统共同工作，提高了系统的工作可靠性。



CN 107105032 B

1.一种节点设备运行方法,其特征在于,所述方法包括:

当前节点设备接收多个节点设备的投票请求,所述多个节点设备的数量大于系统中节点设备数量的半数;

如果所述当前节点设备运行于领导状态,则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引;

如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节点设备的最新日志索引,将当前工作状态从所述领导状态切换至跟随状态或候选状态。

2.根据权利要求1所述的方法,其特征在于,所述接收多个节点设备的投票请求包括:

当接收到第一个投票请求后,启动定时器进行计时;

在所述定时器的运行过程中,继续接收其他节点设备的投票请求,直到所述定时器超时后,停止接收其他节点设备的投票请求。

3.根据权利要求1所述的方法,其特征在于,所述接收多个节点设备的投票请求之后,所述方法还包括:

如果所述当前节点设备运行于跟随状态,则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引;

如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节点设备的最新日志索引,将当前工作状态从所述跟随状态切换至候选状态或保持跟随状态。

4.根据权利要求1所述的方法,其特征在于,所述如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节点设备的最新日志索引,将当前工作状态从所述领导状态切换至跟随状态或候选状态之后,所述方法还包括:

根据所述多个节点设备的投票请求,确定目标节点设备;

响应于所述目标节点设备的投票请求,向所述目标节点设备发送投票确认消息。

5.根据权利要求1所述的方法,其特征在于,所述如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节点设备的最新日志索引,将当前工作状态从所述领导状态切换至跟随状态或候选状态之后,所述方法还包括:

接收运行于领导状态的节点设备所广播的心跳信息;或,

接收运行于领导状态的节点设备所广播的日志复制指令,基于所述日志复制指令复制日志。

6.一种节点设备,其特征在于,所述节点设备包括:

接收模块,用于接收多个节点设备的投票请求,所述多个节点设备的数量大于系统中节点设备数量的半数;

获取模块,用于如果所述节点设备运行于领导状态,则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引;

运行模块,用于如果所述多个节点设备的投票请求中的运行周期信息均大于所述节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述节

点设备的最新日志索引，将当前工作状态从所述领导状态切换至跟随状态或候选状态。

7.根据权利要求6所述的节点设备，其特征在于，所述接收模块用于：

当接收到第一个投票请求后，启动定时器进行计时；

在所述定时器的运行过程中，继续接收其他节点设备的投票请求，直到所述定时器超时后，停止接收其他节点设备的投票请求。

8.根据权利要求6所述的节点设备，其特征在于，

所述获取模块还用于：如果所述节点设备运行于跟随状态，则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引；

所述运行模块还用于：如果所述多个节点设备的投票请求中的运行周期信息均大于所述节点设备的运行周期信息，且所述多个节点设备的投票请求中的最新日志索引均不小于所述节点设备的最新日志索引，将当前工作状态从所述跟随状态切换至候选状态或保持跟随状态。

9.根据权利要求6所述的节点设备，其特征在于，所述节点设备还包括：

确定模块，用于根据所述多个节点设备的投票请求，确定目标节点设备；

发送模块，用于响应于所述目标节点设备的投票请求，向所述目标节点设备发送投票确认消息。

10.根据权利要求6所述的节点设备，其特征在于，

所述接收模块，还用于接收运行于领导状态的节点设备所广播的心跳信息；或，

所述接收模块，还用于接收运行于领导状态的节点设备所广播的日志复制指令，基于所述日志复制指令复制日志。

节点设备运行方法及节点设备

技术领域

[0001] 本发明涉及网络技术领域,特别涉及一种节点设备运行方法及节点设备。

背景技术

[0002] 随着网络技术的发展,基于集群为客户端提供服务的方式越来越普遍。为了保证集群中各个节点设备保持一致性,节点设备运行时一般可以应用BFT-Raft(Byzantine Fault Tolerance algorithm-Raft,拜占庭容错筏算法)。

[0003] 根据BFT-Raft,节点设备的工作状态可以分为三种:跟随状态follower、候选状态candidate和领导状态leader。当任一节点设备a处于跟随状态时,可以根据该集群中运行于领导状态的节点设备b所广播的心跳信息,确定该节点设备b运行正常,并基于节点设备b的指示复制日志。当节点设备a在一段时间内未接收到节点设备b的心跳信息,可以确定节点设备b运行故障,并切换为候选状态运行,将投票请求广播至集群中的各个节点设备,一旦接收到该集群中半数以上的节点设备的投票,节点设备a可以切换为领导状态运行,并将心跳信息广播至集群中的各个节点设备、基于和客户端的交互存储日志、指示各个节点设备复制日志。需要说明的是,在运行于领导状态的节点设备a运行正常的情况下,如果接收到投票请求或心跳信息,会自动忽略。

[0004] 在实现本发明的过程中,发明人发现现有技术至少存在以下问题:

[0005] 由于集群可能分裂成网络相隔离的两个子集群,如,子集群A和子集群B,该子集群A中包括该集群中运行于领导状态的节点设备a,且子集群A的节点设备数量小于子集群B的节点设备数量,则子集群B中的节点设备可以通过投票选出一个新的运行于领导状态的节点设备b,而当节点设备b运行故障时,子集群B中处于候选状态的节点设备会再次广播投票请求,如果子集群A与子集群B此时恢复网络连接,由于节点设备a运行正常,会忽略投票请求,即使子集群B中的某一节点设备b切换为领导状态运行,节点设备a也会忽略该节点设备b的心跳信息,导致节点设备a无法与子集群b合为一个系统共同工作,系统的工作可靠性低。

发明内容

[0006] 为了解决现有技术的问题,本发明实施例提供了一种节点设备运行方法及节点设备。所述技术方案如下:

[0007] 一方面,提供了一种节点设备运行方法,所述方法包括:

[0008] 接收多个节点设备的投票请求,所述多个节点设备的数量大于系统中节点设备数量的半数;

[0009] 如果当前节点设备运行于领导状态,则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引;

[0010] 如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息,且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节

点设备的最新日志索引，将当前工作状态从所述领导状态切换至跟随状态或候选状态。

[0011] 另一方面，提供了一种节点设备，所述节点设备包括：

[0012] 接收模块，用于接收多个节点设备的投票请求，所述多个节点设备的数量大于系统中节点设备数量的半数；

[0013] 获取模块，用于如果当前节点设备运行于领导状态，则从所述多个节点设备的投票请求中获取运行周期信息和最新日志索引；

[0014] 运行模块，用于如果所述多个节点设备的投票请求中的运行周期信息均大于所述当前节点设备的运行周期信息，且所述多个节点设备的投票请求中的最新日志索引均不小于所述当前节点设备的最新日志索引，将当前工作状态从所述领导状态切换至跟随状态或候选状态。

[0015] 本发明实施例通过在接收到多个投票请求时，获取投票请求中的运行周期信息和最新日志索引，如果获取的运行周期信息均大于当前节点设备的运行周期信息，且获取的最新日志索均不小于当前节点设备的最新日志索引，则以跟随状态运行或候选状态运行，使得第一子集群中运行于领导状态的节点设备可以降级为跟随状态或候选状态，进而使得第一子集群中的节点设备均可以与第二子集群中的节点设备共同参与选举，直到新的领导状态的节点设备的出现时，该第一子集群可以和第二子集群合为一个系统共同工作，提高了系统的工作可靠性。

附图说明

[0016] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述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。

[0017] 图1A是本发明实施例提供的一种节点设备运行的实施环境示意图；

[0018] 图1B是本发明实施例提供的一种节点设备工作状态的切换示意图；

[0019] 图2是本发明实施例提供的一种节点设备运行方法的流程图；

[0020] 图3是本发明实施例提供的一种节点设备运行方法的流程图；

[0021] 图4是本发明实施例提供的一种节点设备的模块示意图；

[0022] 图5是本发明实施例提供的一种节点设备的模块示意图；

[0023] 图6是本发明实施例提供的一种节点设备结构示意图。

具体实施方式

[0024] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

[0025] 图1A是本发明实施例提供的一种节点设备运行的实施环境示意图。参见图1A，该实施环境为一个由多个节点设备构成的系统，该系统也相当于一个集群，节点设备1为该系统中运行于领导状态的节点设备，在节点设备1运行正常时，可以定时地向各个运行于跟随状态的节点设备广播心跳信息，如，节点设备2、节点设备3和节点设备4，每个运行于跟随状态的节点设备在接收到心跳信息时可以确定节点设备1运行正常，并重置定时器(一般为

0.5-1秒之间的随机值，这样可以避免各个节点设备的定时器的计时时长相同可能造成反复选举的情况)，等待下一次心跳信息。

[0026] 事实上，系统中各个节点设备的工作状态是可以动态切换的，参见图1B，本发明实施例提供了一种节点设备工作状态的切换示意图。一旦运行于跟随状态(follower)的节点设备在定时器超时的情况下没有接收到心跳信息，可以确定运行于领导状态的节点设备运行故障，并切换为候选状态(candidate)运行；进而，节点设备可以重置定时器，并广播投票请求，直到接收到该系统中半数以上的投票确认消息切换为领导状态(leader)运行，或者接收到运行于领导状态的节点设备的心跳信息时切换为跟随状态运行，或者定时器超时的情况下保持候选状态开始新一轮选举；运行于领导状态的节点设备可以在发现比自身具有更高运行周期信息(term)的节点设备时切换为跟随状态运行。

[0027] 在该系统为客户端提供服务时，当该系统中的任一节点设备接收到客户端的服务命令时，可以将该服务命令重定向至节点设备1，由节点设备1向各个节点设备广播日志添加请求，该日志添加请求用于请求将该服务命令添加到日志中，如果节点设备1可以接收到各个节点设备对日志添加请求的确认消息，可以响应该客户端的服务命令，将该服务命令添加到日志中，并向各个节点设备广播日志复制指令，使得各个节点设备将该服务命令复制到日志中。在实际的应用场景中，该系统可以是底层基于区块链技术的交易系统，该服务命令可以为客户端的交易信息，每个节点设备所存储的日志可以对应一条区块链，当添加交易信息到日志中时，实际是将该交易信息存储到当前区块的下一区块中，由于已存储至区块链中的数据不可更改，可以有效地防止交易信息被篡改，提高交易信息的安全性。

[0028] 由于网络中断等原因，该系统中的各个节点设备可能分裂形成两个网络相隔的子集群，即第一子集群和第二子集群，且第一子集群的节点设备数量小于第二子集群的节点设备数量，该第一子集群中包括该系统中运行于领导状态的节点1。进而，该第一子集群中运行于跟随状态的节点设备可以依据该节点设备1定时广播的心跳信息继续正常工作；第二子集群由于和节点设备1的网络中断，其中运行于跟随状态的节点设备在定时器超时的情况下也不能接收到节点设备1的心跳信息，依据bft-raft的超时选举机制，运行于跟随状态的节点设备会切换为候选状态运行，将自身的运行周期信息加一，并广播投票请求，该第二子集群中接收到大于该系统中节点数量的一半的投票请求的节点设备可以切换为领导状态运行，并广播心跳信息，该心跳信息携带该运行于领导状态的节点设备的运行周期信息，当运行于候选状态的节点接收到心跳信息时，可以切换为跟随状态运行，并将自身的运行周期信息同步为心跳信息所携带的运行周期信息；当第二子集群中运行于领导状态的节点设备运行故障时，该第二子集群中的各个节点设备将切换为候选状态运行，并再次进行选举，如果此时第一子集群和第二子集群恢复网络连接，依照现有技术，由于第一子集群中的节点设备1运行正常，该第一子集群中的各个节点设备均会忽略来自于第二子集群中的节点设备的投票请求，即使该第二子集群选出新的领导状态的节点设备，且该第二子集群中的节点设备可以按照该新的领导状态的节点设备的心跳信息工作，但该第一子集群中的各个节点设备会忽略新的领导状态的节点设备的心跳信息，并继续按照节点设备1的心跳信息继续工作，导致第一子集群和第二子集群无法恢复成一个系统共同工作，系统的工作可靠性差。

[0029] 图2是本发明实施例提供的一种节点设备运行方法的流程图。参见图2，该方法可

以应用于图1A所示实施例的节点设备1，包括以下步骤：

[0030] 201、节点设备1接收多个节点设备的投票请求，多个节点设备的数量大于系统中节点设备数量的半数。

[0031] 其中，该多个节点设备可以为图1A所示实施例中第二子集群中的节点设备。由于该第二子集群中原有的领导状态的节点设备运行故障，该多个节点设备在自身的定时器超时的情况下也没有接收到心跳信息，因此正在以候选状态运行，并基于自身的运行周期信息、最新日志索引(last log index)和节点设备标识等信息生成投票请求，将投票请求广播至该系统中的各个节点。一般地，接收到投票请求的节点设备会判断是否在定时器未超时的情况下接收到心跳信息，如果是，则确定领导状态的节点设备运行正常，并忽略该投票请求，如果否，则提取投票请求中的运行周期信息和最新日志索引，并将提取的信息分别与自身的信息进行比较，如果二者分别大于等于自身的信息，则向该投票请求中的节点设备标识对应的节点设备发送投票确认消息，否则也会忽略该投票请求，一旦已经为某一节点设备投票，则在该运行周期内都不会再为其他节点设备投票。当然，如果是运行于领导状态的节点设备接收到投票请求，则会自动忽略该投票请求。本发明实施例为使第一子集群中运行于领导状态的节点设备能够与第二子集群合为一个系统工作，提高集群的工作可靠性，进行以下步骤。

[0032] 考虑到接收到一个或少数个投票请求的情况可能为该投票请求来自于伪装成候选状态的节点设备，因此需要排除这种情况。并且，为了初步印证该系统目前处于分裂后的子集群之间已恢复网络连接，且第二子集群内的节点设备正在进行选举的情况，对接收到该投票请求的数量进行限制，也即是，理应接收到大于该系统中的节点设备数量的半数的投票请求。

[0033] 在实际的场景中，由于每一轮选举都有时限，因此该步骤也可以具体为：当节点设备1接收到第一个投票请求后，启动定时器进行计时；在定时器的运行过程中，继续接收其他节点设备的投票请求，直到定时器超时后，停止接收其他节点设备的投票请求。也就是说，节点设备1可以接收定时器计时时长内的投票请求，该计时时长可以为一轮选举的时长，如果节点设备1在该计时时长内接收到的投票请求的数量大于该系统中节点设备的数量的半数，说明该系统发生过分裂，且第一子集群和第二子集群已恢复网络连接，且第二子集群正在进行选举，则进行步骤202，否则，上述情况不能得到印证，可以忽略接收到的投票请求，并继续广播心跳信息。

[0034] 202、如果节点设备1运行于领导状态，则节点设备1从多个节点设备的投票请求中获取运行周期信息和最新日志索引。

[0035] 其中，运行周期信息是指发送该投票请求的节点设备当前所处的运行周期号。每次进行选举时，由跟随状态切换为候选状态的节点设备的运行周期信息会加一，最终成为领导状态的节点设备可以将运行周期信息携带在心跳信息中并广播给其他节点设备，接收到心跳信息的候选状态的节点设备可以切换为跟随状态运行，并将自身的运行周期信息同步为该心跳信息中的运行周期信息，而且可以根据自身的最新日志索引和心跳信息中的最新日志索引，确定自身缺少的日志，并请求领导状态的节点设备返回自身缺少的日志。因此，该运行周期信息可以表征一个节点设备是否始终与运行于领导状态的节点设备保持同步且运行正常。最新日志索引是指发送该投票请求的节点设备最新存储的日志的索引，每

次运行于领导状态的节点设备添加新的日志后，该最新日志索引加一，且该运行于领导状态的节点设备可以将日志复制指令广播给其他节点设备，使得接收到日志复制指令的节点设备可以同步该领导状态的节点设备的日志和最新日志索引，因此，该最新日志索引可以表征一个节点设备的日志完整性，显然，运行于领导状态的节点设备为在其系统中日志完整性最好的节点设备。

[0036] 该步骤中，节点设备1可以分别按照运行周期信息和最新日志索引在投票请求中的协议位置，从投票请求中分别提取出对应协议位置的运行周期信息和最新日志索引。

[0037] 203、节点设备1判断多个节点设备的投票请求中的运行周期信息是否均大于节点设备1的运行周期信息，如果是，执行步骤204，如果否，忽略多个投票请求。

[0038] 该步骤中，为了进一步印证该系统发生过分裂，且第二子集群在分裂后曾经选出过领导状态的节点设备并已运行故障，且第一子集群和第二子集群已恢复网络连接，且第二子集群正在进行选举的实施场景，考虑到第二子集群在选出过的领导状态的节点设备时该第二子集群中的节点设备的运行周期信息已相比第一子集群的运行周期多一，因此运行周期信息可以作为上述实施场景的印证依据之一，如果投票请求中的运行周期信息均大于节点设备1的运行周期信息，上述实施场景得到印证，则继续执行步骤204，如果该投票请求中的运行周期信息不大于自身的运行周期信息，说明该投票请求对应的节点设备很可能运行故障，且不符合上述实施场景，则节点设备1可以忽略多个投票请求，并继续广播心跳信息。

[0039] 204、节点设备1判断多个节点设备的投票请求中的最新日志索引是否均不小于节点设备1的最新日志索引，如果是，将当前工作状态从领导状态切换至跟随状态，如果否，忽略多个投票请求。

[0040] 考虑到在系统分裂之前，该系统中的各个节点设备的日志理应与节点设备1的日志同步，因此第二子集群经过为客户端服务的一段时间，在两个子集群恢复网络连接之后，第二子集群中的节点设备所存储的日志应该不少于该第一子集群中的节点设备所存储的日志，也因此可以将最新日志索引作为印证上述实施场景的依据之一，如果多个节点设备的投票请求中的最新日志索引是否均不小于节点设备1的最新日志索引，说明该投票请求对应的节点设备已存储的日志量等于或多于节点设备1的日志量，上述实施场景最终得到各项印证，因此节点设备1切换为跟随状态运行，并停止广播心跳信息，如果该投票请求中的最新日志索引小于自身的最新日志索引，上述实施场景没有得到印证，则可以忽略该多个投票请求，并继续广播心跳信息。

[0041] 当节点设备1切换为跟随状态运行时，可以停止广播心跳信息，重置定时器，并等待新的领导状态的节点设备的心跳信息，如果在定时器超时的情况下也没有接收到心跳信息，则可以再切换为候选状态运行，并广播投票请求，直到自身成为领导状态的节点设备，或者接收到新的领导状态的节点设备的心跳信息时切换为跟随状态运行。

[0042] 事实上，节点设备1也可以将当前工作状态切换为候选状态运行，停止广播心跳信息，且广播投票请求，直到自身成为领导状态的节点设备，或者接收到新的领导状态的节点设备的心跳信息时切换为跟随状态运行。

[0043] 需要说明的是，一旦节点设备1停止广播心跳信息，第一子集群中运行于跟随状态的节点设备可以在定时器超时后主动切换为候选状态运行，直到自身成为该系统中领导状

态的节点设备，或者接收到该系统中运行于领导状态的节点设备的心跳信息时切换为跟随状态运行。因此，上述节点设备运行方法还可以使得第一子集群和第二子集群恢复为原来的系统进行工作，提高该系统的工作可靠性。

[0044] 另外，需要说明的是，本发明实施例对节点设备1执行步骤203和204的时序不做具体限定，事实上，节点设备1也可以先对最新日志索引进行判断，再对运行周期信息进行判断，或者，为了提高判断效率，并尽快使得第一子集群和第二子集群合为一个系统工作，节点设备1也可以同时对最新日志索引和运行周期信息进行判断，只要二者分别满足上述各自的判断条件，节点设备1即可将当前工作状态切换至跟随状态(或候选状态)。

[0045] 本发明实施例通过在接收到多个投票请求时，获取投票请求中的运行周期信息和最新日志索引，如果获取的运行周期信息均大于当前节点设备的运行周期信息，且获取的最新日志索均不小于当前节点设备的最新日志索引，则以跟随状态运行或候选状态运行，使得第一子集群中运行于领导状态的节点设备可以降级为跟随状态或候选状态，进而使得第一子集群中的节点设备均可以与第二子集群中的节点设备共同参与选举，直到新的领导状态的节点设备的出现时，该第一子集群可以和第二子集群合为一个系统共同工作，提高了系统的工作可靠性。

[0046] 205、节点设备1根据多个节点设备的投票请求，确定目标节点设备。

[0047] 其中，目标节点设备是指该节点设备1趋于投票的节点设备。该步骤中，经过步骤203和步骤204的判断过程，对该节点设备1来说，任一投票请求对应的节点设备均满足成为领导状态的节点设备的资格，因此节点设备1可以按照投票请求的接收顺序，将接收顺序在前的投票请求对应的节点设备作为目标节点设备。

[0048] 206、节点设备1响应于目标节点设备的投票请求，向目标节点设备发送投票确认消息。

[0049] 该步骤中，节点设备1可以基于自身的节点设备标识，生成投票确认消息，并按照目标节点设备的节点设备标识将投票确认消息发送至目标节点设备。

[0050] 当然，为使目标节点设备能够验证投票者的身份，提高系统安全性，节点设备1可以携带有签名的投票确认消息发送至目标节点设备。该系统中的每个节点设备可以配置有自身的私钥以及各个节点设备的公钥。因此，当目标节点设备接收到该投票确认消息时，可以提取出该节点设备1的签名，采用已配置的该节点设备1的公钥对该节点设备的签名进行验证。

[0051] 需要说明的是，步骤205和206是本发明实施例的可选步骤。事实上，由于第二子集群的节点设备数量大于该系统的节点设备数量的半数，则节点设备1也可以不对任一投票请求进行响应，则该系统中也能选出一个领导状态的节点设备，并且在接收到该领导状态的节点设备的心跳信息时，将自身的运行周期信息同步为该心跳信息携带的运行周期信息，从而与该第二子集群合为一个系统工作。

[0052] 207、节点设备1接收运行于领导状态的节点设备所广播的心跳信息。

[0053] 一旦该系统中任一候选状态的节点设备接收到大于该系统中节点设备的半数的投票确认消息时，可以切换为领导状态运行，并广播自身的心跳信息，使得节点设备1可以接收到该心跳信息。

[0054] 其中，为了避免有的节点设备伪装成领导状态的节点设备，提高系统的安全性，该

心跳信息可以携带该系统中的各个节点设备在响应该切换为领导状态的节点设备的投票请求时的签名。因此，当节点设备1接收到该心跳信息时，可以提取出各个节点设备的签名，采用已配置的任一节点设备的公钥对该节点设备的签名进行验证，如果各个节点设备的签名均验证通过，且验证通过的签名数量大于该系统中节点设备数量的半数，说明该心跳信息确实来自运行于领导状态的节点设备，则可以重置定时器，并等待下一次心跳信息。

[0055] 事实上，为了保证系统的一致性，该运行于领导状态的节点设备可以广播日志复制指令，使得该节点设备1可以接收运行于领导状态的节点设备所广播的日志复制指令，基于日志复制指令复制日志，从而将该系统最新接收到的服务指令添加到日志中。当然，基于bft-raft不仅解决节点设备一致性而且解决了节点设备欺诈，数据被篡改、丢失或顺序错乱的问题，该日志复制指令需携带系统中的各个节点设备在响应该运行于领导状态的节点设备的投票请求时的签名，使得节点设备1可以对该日志复制指令进行验证，并在验证通过后进行日志复制。

[0056] 以下对节点设备1切换至跟随状态(或候选状态)后该系统的工作情况进行具体说明：

[0057] 当节点设备1切换为跟随状态(或候选状态)运行时，由于停止广播心跳信息，该第一子集群中运行于跟随状态的节点设备在定时器超时后没有接收到心跳信息，因此切换为候选状态运行。

[0058] 如果此时第二子集群的选举尚未结束，则第一子集群中切换为候选状态的节点设备相当于与第二子集群中的节点设备共同参与选举；当其中任一节点设备接收到大于该系统的节点设备的半数的投票时，可以切换为领导状态运行，并广播心跳信息，当该系统中的其他节点设备接收到该心跳信息时，可以确认选举结束，切换为跟随状态，并将自身的运行周期信息与该心跳信息中的运行周期信息同步，后续可以基于该领导状态的节点设备的心跳信息或日志复制指令等进行工作。

[0059] 如果第一子集群中运行于跟随状态的节点设备切换为候选状态后，第二子集群的选举已结束，该第二子集群中成为领导状态的节点设备可以定时广播心跳信息，第二子集群中曾运行于候选状态的节点设备在首次接收到该心跳信息时可以切换为跟随状态，并将自身的运行周期信息与该心跳信息中的运行周期信息同步；该第一子集群由于没有领导状态的节点设备，运行于跟随状态的节点设备可以在首次接收到该心跳信息时保持跟随状态，但将自身的运行周期信息与该心跳信息中的运行周期信息同步，运行于候选状态的节点设备可以在首次接收到该心跳信息时切换为跟随状态，并将自身的运行周期信息与该心跳信息中的运行周期信息同步。

[0060] 以上图2实施例是以第一子集群中运行于领导状态的节点设备1为执行主体为例进行说明，在节点设备1停止广播心跳信息后，使得该第一子集群中运行于跟随状态的节点设备(命名为节点设备5)可以被动地与第二子集群合为一个系统，事实上，为使节点设备5可以高效地和第二子集群合为一个系统工作，提高集群的可靠性，该节点设备5也可以应用本发明实施例提供的节点设备运行方法，例如，图3是本发明实施例提供的一种节点设备运行方法的流程图。参见图3，该方法包括：

[0061] 301、节点设备5接收多个节点设备的投票请求，多个节点设备的数量大于系统中节点设备数量的半数。

[0062] 与步骤201同理，在此不做赘述。

[0063] 302、如果节点设备5运行于跟随状态，则节点设备5从多个节点设备的投票请求中获取运行周期信息和最新日志索引。

[0064] 与步骤202同理，在此不做赘述。

[0065] 303、节点设备5判断多个节点设备的投票请求中的运行周期信息是否均大于节点设备5的运行周期信息，如果是，执行步骤204，如果否，忽略多个投票请求。

[0066] 与步骤203同理，在此不做赘述。

[0067] 304、节点设备5判断多个节点设备的投票请求中的最新日志索引是否均不小于节点设备5的最新日志索引，如果是，将当前工作状态从所述跟随状态切换至候选状态，如果否，忽略多个投票请求。

[0068] 与步骤204同理。但该节点设备5需切换至候选状态，并广播投票请求，直到接收到新的领导状态的节点设备的心跳信息时切换为跟随状态，或者直到接收到大于该系统中节点设备的半数的投票请求时切换为领导状态。

[0069] 当然，该节点设备5也可以保持跟随状态，当定时器超时，可以自动切换为候选状态，直到接收到新的领导状态的节点设备的心跳信息时切换为跟随状态，或者直到接收到大于该系统中节点设备的半数的投票请求时切换为领导状态。

[0070] 本发明实施例通过在接收到多个投票请求时，获取投票请求中的运行周期信息和最新日志索引，如果获取的运行周期信息均大于当前节点设备的运行周期信息，且获取的最新日志索均不小于当前节点设备的最新日志索引，则以跟随状态运行或候选状态运行，使得第一子集群中运行于领导状态的节点设备可以降级为跟随状态或候选状态，进而使得第一子集群中的节点设备均可以与第二子集群中的节点设备共同参与选举，直到新的领导状态的节点设备的出现时，该第一子集群可以和第二子集群合为一个系统共同工作，提高了系统的工作可靠性。

[0071] 305、节点设备5根据多个节点设备的投票请求，确定目标节点设备。

[0072] 与步骤205同理，在此不做赘述。

[0073] 306、节点设备5响应于目标节点设备的投票请求，向目标节点设备发送投票确认消息。

[0074] 与步骤206同理，在此不做赘述。

[0075] 307、节点设备5接收运行于领导状态的节点设备所广播的心跳信息。

[0076] 与步骤207同理，在此不做赘述。

[0077] 当然，该节点设备还可以继续参与选举，以保证系统整体选举的公正性。

[0078] 图4是本发明实施例提供的一种节点设备的模块示意图。参见图4，该节点设备包括：

[0079] 接收模块401，用于接收多个节点设备的投票请求，多个节点设备的数量大于系统中节点设备数量的半数；

[0080] 获取模块402，用于如果当前节点设备运行于领导状态，则从多个节点设备的投票请求中获取运行周期信息和最新日志索引；

[0081] 运行模块403，用于如果多个节点设备的投票请求中的运行周期信息均大于当前节点设备的运行周期信息，且多个节点设备的投票请求中的最新日志索引均不小于当前节

点设备的最新日志索引，将当前工作状态从领导状态切换至跟随状态或候选状态。

[0082] 本发明实施例通过在接收到多个投票请求时，获取投票请求中的运行周期信息和最新日志索引，如果获取的运行周期信息均大于当前节点设备的运行周期信息，且获取的最新日志索均不小于当前节点设备的最新日志索引，则以跟随状态运行或候选状态运行，使得第一子集群中运行于领导状态的节点设备可以降级为跟随状态或候选状态，进而使得第一子集群中的节点设备均可以与第二子集群中的节点设备共同参与选举，直到新的领导状态的节点设备的出现时，该第一子集群可以和第二子集群合为一个系统共同工作，提高了系统的工作可靠性。

[0083] 在一种可能实现方式中，接收模块401用于：当接收到第一个投票请求后，启动定时器进行计时；在定时器的运行过程中，继续接收其他节点设备的投票请求，直到定时器超时后，停止接收其他节点设备的投票请求。

[0084] 在一种可能实现方式中，获取模块402还用于：如果当前节点设备运行于跟随状态，则从多个节点设备的投票请求中获取运行周期信息和最新日志索引；

[0085] 运行模块403还用于：如果多个节点设备的投票请求中的运行周期信息均大于当前节点设备的运行周期信息，且多个节点设备的投票请求中的最新日志索引均不小于当前节点设备的最新日志索引，将当前工作状态从跟随状态切换至候选状态或保持跟随状态。

[0086] 在一种可能实现方式中，基于图4的节点设备组成，参见图5，节点设备还包括：

[0087] 确定模块404，用于根据多个节点设备的投票请求，确定目标节点设备；

[0088] 发送模块405，用于响应于目标节点设备的投票请求，向目标节点设备发送投票确认消息。

[0089] 在一种可能实现方式中，接收模块401，还用于接收运行于领导状态的节点设备所广播的心跳信息；或，

[0090] 接收模块401，还用于接收运行于领导状态的节点设备所广播的日志复制指令，基于日志复制指令复制日志。

[0091] 上述所有可选技术方案，可以采用任意结合形成本发明的可选实施例，在此不再一一赘述。

[0092] 需要说明的是：上述实施例提供的节点设备在执行节点设备运行方法时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将节点设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的节点设备与节点设备运行方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0093] 图6是本发明实施例提供的一种节点设备结构示意图。参见图6，该节点设备600可以被提供为一服务器。节点设备600包括处理组件622，其进一步包括一个或多个处理器，以及由存储器632所代表的存储器资源，用于存储可由处理部件622的执行的指令，例如应用程序。存储器632中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件622被配置为执行指令，以执行上述节点设备运行方法。

[0094] 节点设备600还可以包括一个电源组件626被配置为执行节点设备600的电源管理，一个有线或无线网络接口650被配置为将节点设备600连接到网络，和一个输入输出(I/O)接口658。节点设备600可以操作基于存储在存储器632的操作系统，例如Windows

Server™,Mac OS X™,Unix™,Linux™,FreeBSD™或类似。

[0095] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0096] 以上所述仅为本发明的较佳实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

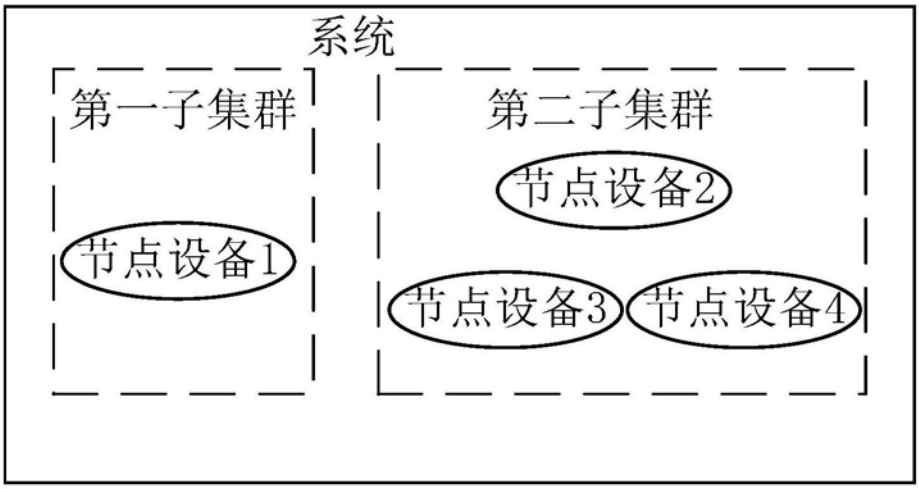

系统
第一子集群
第二子集群
节点设备2
节点设备1
节点设备3
节点设备4

图1A

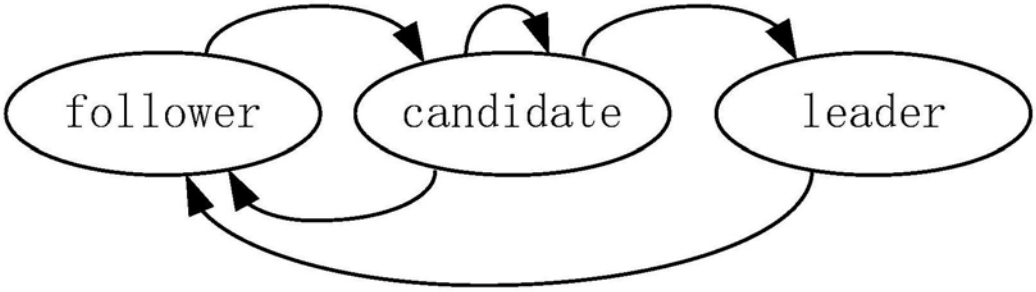

follower
candidate
leader

图1B

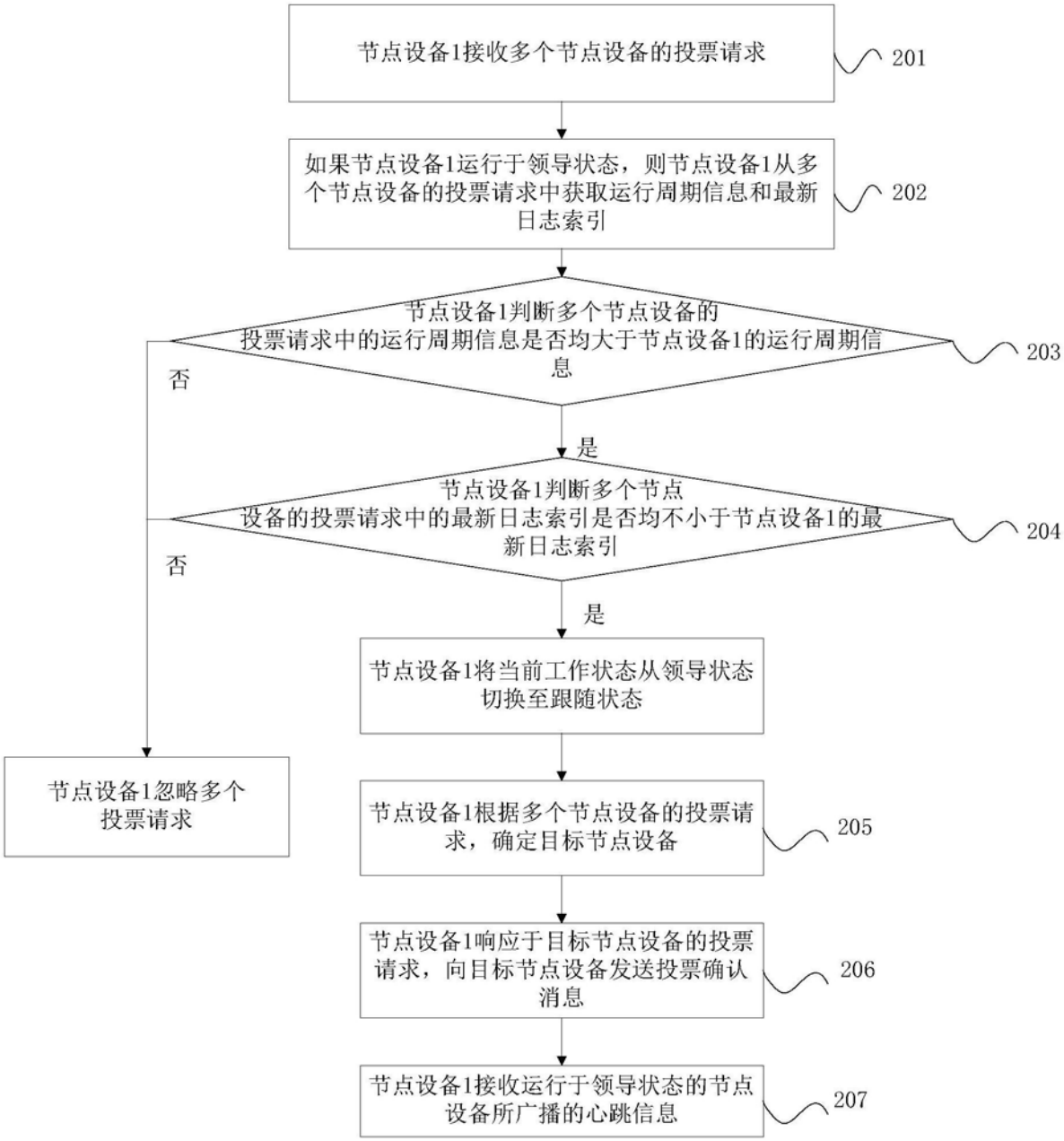

节点设备1接收多个节点设备的投票请求
201
如果节点设备1运行于领导状态，则节点设备1从多个节点设备的投票请求中获取运行周期信息和最新日志索引
202
节点设备1判断多个节点设备的投票请求中的运行周期信息是否均大于节点设备1的运行周期信息
203
否
是
节点设备1判断多个节点设备的投票请求中的最新日志索引是否均不小于节点设备1的最新日志索引
204
否
是
节点设备1将当前工作状态从领导状态切换至跟随状态
节点设备1忽略多个投票请求
节点设备1根据多个节点设备的投票请求，确定目标节点设备
205
节点设备1响应于目标节点设备的投票请求，向目标节点设备发送投票确认消息
206
节点设备1接收运行于领导状态的节点设备所广播的心跳信息
207

图2

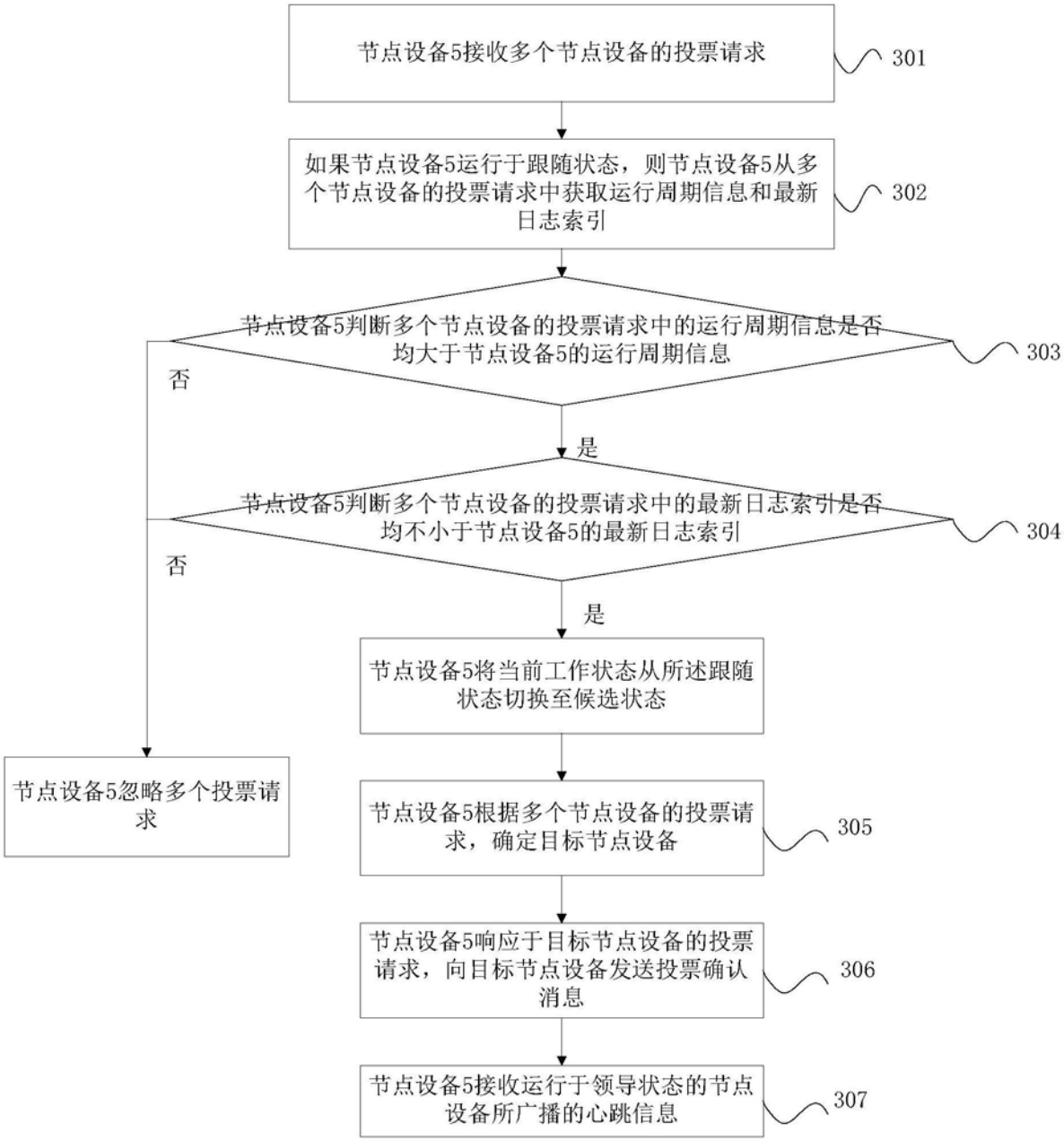

节点设备5接收多个节点设备的投票请求
301
如果节点设备5运行于跟随状态，则节点设备5从多个节点设备的投票请求中获取运行周期信息和最新日志索引
302
节点设备5判断多个节点设备的投票请求中的运行周期信息是否均大于节点设备5的运行周期信息
303
否
是
节点设备5判断多个节点设备的投票请求中的最新日志索引是否均不小于节点设备5的最新日志索引
304
否
是
节点设备5将当前工作状态从所述跟随状态切换至候选状态
节点设备5忽略多个投票请求
节点设备5根据多个节点设备的投票请求，确定目标节点设备
305
节点设备5响应于目标节点设备的投票请求，向目标节点设备发送投票确认消息
306
节点设备5接收运行于领导状态的节点设备所广播的心跳信息
307

图3

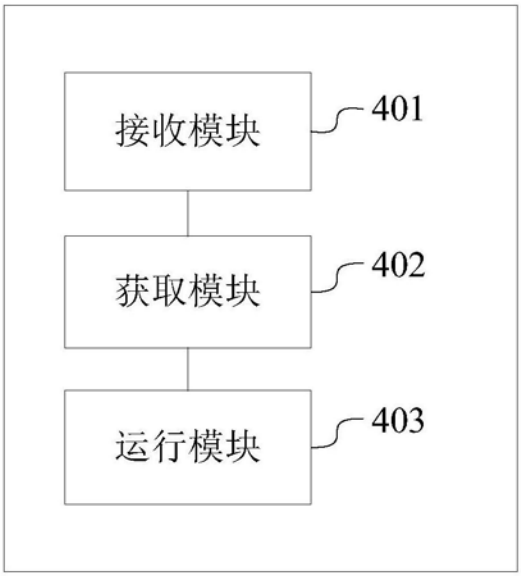

接收模块
401
获取模块
402
运行模块
403

图4

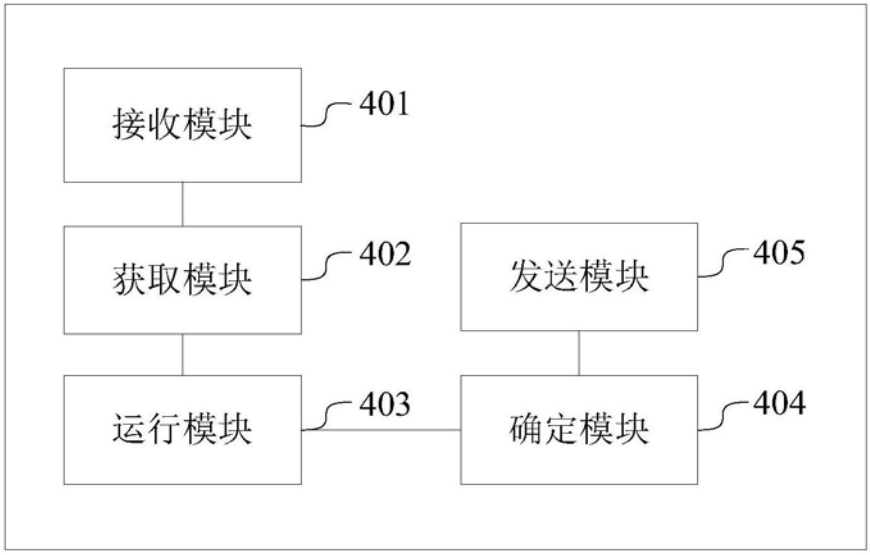

接收模块
401
获取模块
402
发送模块
405
运行模块
403
确定模块
404

图5

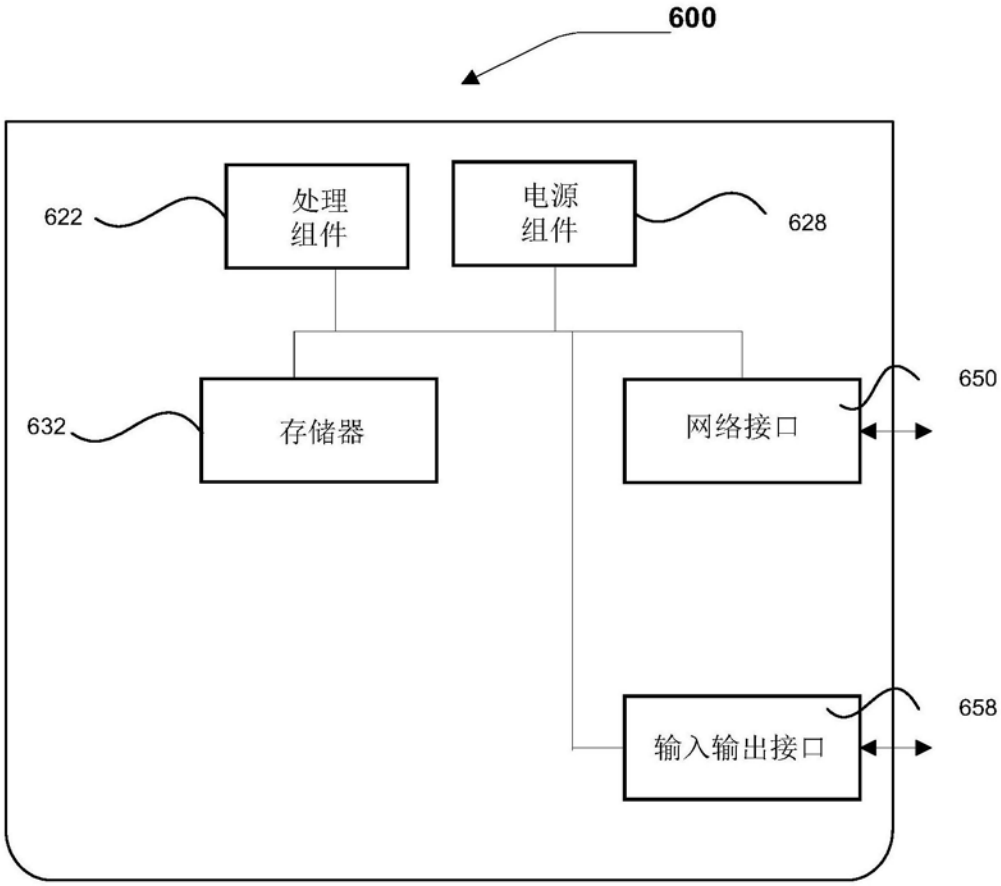


图6