

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2016 年 8 月 18 日 (18.08.2016)



(10) 国际公布号
WO 2016/127580 A1

- (51) 国际专利分类号:
H04L 29/08 (2006.01) H04L 12/26 (2006.01)
- (21) 国际申请号: PCT/CN2015/084573
- (22) 国际申请日: 2015 年 7 月 21 日 (21.07.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201510068053.1 2015 年 2 月 10 日 (10.02.2015) CN
- (71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
- (72) 发明人: 袁健清 (YUAN, Jianqing); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 倪绍基 (NI, Shaoji); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。
- (74) 代理人: 北京龙双利达知识产权代理有限公司 (LONGSUN LEAD IP LTD.); 中国北京市海淀区丹棱街 16 号海兴大厦 C 座 1108, Beijing 100080 (CN)。

- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD, DEVICE AND SYSTEM FOR PROCESSING FAULT IN AT LEAST ONE DISTRIBUTED CLUSTER

(54) 发明名称: 处理至少一个分布式集群中的故障的方法、设备和系统



S210 A first backup node receiving a heartbeat message sent by a first reference node, the heartbeat message comprising first indication information used for indicating that the first reference node and a first Master node are in a disconnected state
S220 According to the first indication information, the first backup node determining that the first reference node and the first Master node are in a disconnected state
S230 The first backup node detecting whether the heartbeat message sent to the first backup node by the first Master node is interrupted
S240 When it is detected that the heartbeat message sent to the first backup node by the first Master node is interrupted, the first backup node determining that the first backup node and the first Master node are also in a disconnected state
S250 The first backup node determining that the first Master node disconnected from both the first backup node and the first reference node is faulty

图 2

(57) Abstract: Provided are a method, device and system for processing a fault in at least one distributed cluster. At least one distributed cluster comprises a first distributed cluster, the first distributed cluster comprising a first Master node, a first Slave node, a first reference node, and a first backup node serving as a backup of the first Master node. The first backup node receives a heartbeat message which is sent by the first reference node and comprises first indication information used for indicating that the first reference node and the first Master node are in a disconnected state; according to the first indication information, the first backup node determines that the first reference node and the first Master node are in a disconnected state; when it is detected that the heartbeat message sent to the first backup node by the first Master node is interrupted, the first backup node determines that the first backup node and the first Master node are also in a disconnected state; and the first backup node determines that the first Master node is faulty. In the embodiments of the present invention, fault detection time can be effectively shortened.

(57) 摘要:

[见续页]

WO 2016/127580 A1



本发明实施例提供一种处理至少一个分布式集群中的故障的方法、设备和系统，至少一个分布式集群包括第一分布式集群，第一分布式集群中包括第一 Master 节点和第一 Slave 节点，第一参考节点，以及作为第一 Master 节点的备份的第一备用节点，第一备用节点接收第一参考节点发送的包括用于指示第一参考节点与第一 Master 节点处于失联状态的第一指示信息的心跳消息；第一备用节点根据第一指示信息，确定第一参考节点与第一 Master 节点处于失联状态；第一备用节点在检测到第一 Master 节点向第一备用节点发送的心跳消息中断的情况下，确定第一备用节点与第一 Master 节点也处于失联状态；第一备用节点确定该第一 Master 节点发生故障。在本发明实施例中，能够有效缩短故障检测时间。

处理至少一个分布式集群中的故障的方法、设备和系统

本申请要求于 2015 年 2 月 10 日提交中国专利局、申请号为 201510068053.1、发明名称为“处理至少一个分布式集群中的故障的方法、
5 设备和系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明实施例涉及集群管理领域，并且更具体地，涉及一种处理至少一个
10 个分布式集群中的故障的方法、设备和系统。

背景技术

目前大部分的高可用性（High Available，简称为“HA”）分布式集群通常为一主（Master 节点）多从（Slave 节点）集中式的，其中，Master 节点
15 向集群中所有 Slave 节点发送心跳消息，集群中每个 Slave 节点也向该 Master 节点发送心跳消息。Slave 节点通过检测 Master 节点发送的心跳消息来判断该 Master 节点是否发生故障，Master 节点通过检测 Slave 节点发送的心跳消息来判断该 Slave 节点是否发生故障。

但在上述故障检测中存在以下弊端，Slave 节点通常是在多次判断没有
20 接收到 Master 节点的心跳消息的情况下，才会认为该 Master 节点发生故障，才会发起用于在集群内确定新的 Master 节点的选举策略，因此，在目前技术中，Master 节点的故障检测时间过长。

发明内容

25 本发明实施例提供一种处理至少一个分布式集群中的故障的方法、设备和系统，能够有效缩短故障检测时间。

第一方面提供了一种处理至少一个分布式集群中的故障的方法，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，该分布式集群还包括第一参考节点，以及作为该第一
30 一 Master 节点的备份的第一备用节点，该方法包括：

该第一备用节点接收该第一参考节点发送的心跳消息，该心跳消息包括

用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息;

该第一备用节点根据该第一指示信息, 确定该第一参考节点与该第一 Master 节点处于失联状态;

- 5 该第一备用节点检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断;

在检测到该第一 Master 节点向该第一备用节点发送的心跳消息中断的情况下, 该第一备用节点确定该第一备用节点与该第一 Master 节点也处于失联状态;

- 10 该第一备用节点确定与该第一备用节点和该第一参考节点均失联的该第一 Master 节点发生故障。

结合第一方面, 在第一方面的第一种可能的实现方式中, 在确定该第一 Master 节点发生故障的情况下, 该方法还包括:

- 15 该第一备用节点向该第一分布式集群中所有节点发送广播消息, 该广播消息用于指示该第一备用节点升级为该第一 Master 节点。

结合第一方面或第一方面的第一种可能的实现方式, 在第一方面的第二种可能的实现方式中, 该第一备用节点检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断, 包括:

- 20 该第一备用节点检测在第一时刻与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息, 该第一时刻为该第一备用节点接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻, 该第二时刻早于该第一时刻, 该第一时刻与该第二时刻之间的时间间隔为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

- 25 结合第一方面或第一方面的第一种或第二种可能的实现方式, 在第一方面的第三种可能的实现方式中, 该至少一个分布式集群还包括第二分布式集群, 该第二分布式集群中包括第二 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为该第二 Master 节点的备份的第二备用节点, 该方法还包括:

- 30 该第一备用节点接收该第二参考节点发送的心跳消息, 该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息, 该第二分布式集群与该第一分布

式集群具有相同的集群标识;

该第一备用节点根据该第二指示信息,确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一备用节点的节点属性不相符,并确定该第二参考节点所在的第二分布式集群与该第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

该第一备用节点与该第二分布式集群中的第二备用节点进行协商,以便于整合该第一分布式集群和该第二分布式集群。

结合第一方面的第三种可能的实现方式,在第一方面的第四种可能的实现方式中,该第一备用节点与该第二分布式集群中的第二备用节点进行协商,包括:

该第一备用节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点的节点属性为备用节点的广播消息;

该第一备用节点接收该第二备用节点发送的协商消息,该协商消息中包括用于指示该第二备用节点的权重的信息,该协商消息是该第二备用节点在检测到该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一备用节点发送的;

在确定该第一备用节点的权重大于或等于该第二备用节点的权重的情况下,该第一备用节点向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息;

在确定该第一备用节点的权重小于该第二备用节点的权重的情况下,该第一备用节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点降级为 Slave 节点的广播消息

第二方面提供了一种处理至少一个分布式集群中的故障的方法,该至少一个分布式集群包括第一分布式集群,该第一分布式集群中包括第一 Master 节点和第一 Slave 节点,该第一分布式集群还包括第一参考节点以及作为该第一 Master 节点的备份的第一备用节点,该方法包括:

该第一 Master 节点接收该第一备用节点发送的心跳消息,该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息;

该第一 Master 节点根据该第三指示信息,确定该第一备用节点与该第一 Slave 节点处于失联状态;

该第一 Master 节点检测该第一 Slave 节点向该第一 Master 节点发送的心跳消息是否中断;

在检测到该第一 Slave 节点向该第一 Master 节点发送的心跳消息中断的情况下, 该第一 Master 节点确定该第一 Master 节点与该第一 Slave 节点也处于失联状态;

该第一 Master 节点确定与该第一 Master 节点和该第一备用节点均失联的该第一 Slave 节点发生故障。

结合第二方面, 在第二方面的第一种可能的实现方式中, 该第一 Master 节点检测该第一 Slave 节点向该第一 Master 节点发送的心跳消息是否中断, 包括:

该第一 Master 节点检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息, 该第三时刻为该第一 Master 节点接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻, 该第四时刻早于该第三时刻, 该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

结合第二方面或第二方面的第一种可能的实现方式, 在第二方面的第二种可能的实现方式中, 该方法还包括:

该第一 Master 节点在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息, 该预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息, 也没有接收到该第一参考节点发送的心跳消息的情况下, 该第一 Master 节点确定该第一备用节点和该第一参考节点均发生故障。

结合第二方面或第二方面的第一种或第二种可能的实现方式, 在第二方面的第三种可能的实现方式中, 该至少一个分布式集群还包括第二分布式集群, 该第二分布式集群中包括第二 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为该第二 Master 节点的备份的第二备用节点, 该方法还包括:

该第一 Master 节点接收该第二参考节点发送的心跳消息, 该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息, 该第二分布式集群与该第一分

布式集群具有相同的集群标识;

该第一 Master 节点根据该第四指示信息,确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一 Master 节点的节点属性不相符,并确定该第二参考节点所在的第二分布式集群与该第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

该第一 Master 节点与该第二分布式集群中的第二 Master 节点进行协商,以便于整合该第一分布式集群和该第二分布式集群。

结合第二方面的第三种可能的实现方式,在第二方面的第四种可能的实现方式中,该第一 Master 节点与该第二分布式集群中的第二 Master 节点进行协商,包括:

该第一 Master 节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点的节点属性为 Master 节点的广播消息;

该第一 Master 节点接收该第二 Master 节点发送的协商消息,该协商消息中包括用于指示该第二 Master 节点的权重的信息,该协商消息是该第二 Master 节点在检测到该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一 Master 节点发送的;

在确定该第一 Master 节点的权重大于或等于该第二 Master 节点的权重时,该第一 Master 节点向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息;

在确定该第一 Master 节点的权重小于该第二 Master 节点的权重时,该第一 Master 节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点降级为 Slave 节点的广播消息。

第三方面提供了一种用于处理至少一个分布式集群中的故障的设备,该至少一个分布式集群包括第一分布式集群,该第一分布式集群中包括第一 Master 节点和第一 Slave 节点,该分布式集群还包括第一参考节点,以及作为该第一 Master 节点的备份的第一备用节点,其中,该第一参考节点用于,在检测到该第一 Master 节点向该第一参考节点发送的心跳消息中断的情况下,向该第一备用节点发送心跳消息,该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息,该设备为该第一备用节点,该设备包括:

第一接收模块,用于接收该第一参考节点发送的包括有该第一指示信息的心跳消息;

第一确定模块,用于根据该第一接收模块接收的该第一指示信息,确定该第一参考节点与该第一 Master 节点处于失联状态;

5 检测模块,用于检测该第一 Master 节点向该设备发送的心跳消息是否中断;

第二确定模块,用于在该检测模块检测到该第一 Master 节点向该设备发送的心跳消息中断的情况下,确定该设备与该第一 Master 节点也处于失联状态;

10 第三确定模块,用于根据该第一确定模块和该第二确定模块确定的结果,确定该第一 Master 节点发生故障。

结合第三方面,在第三方面的第一种可能的实现方式中,该设备还包括:

15 发送模块,用于在该第三确定模块确定该第一 Master 节点发生故障的情况下,向该第一分布式集群中所有节点发送广播消息,该广播消息用于指示该设备升级为该第一 Master 节点。

结合第三方面或第三方面的第一种可能的实现方式,在第三方面的第二种可能的实现方式中,该检测模块具体用于,检测在第一时刻与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息,该第一时刻为该设备接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻,该第二
20 时刻早于该第一时刻,该第一时刻与该第二时刻之间的时间间隔为该第一 Master 节点向该设备发送心跳消息的心跳周期的 N 倍, N 为正整数。

结合第三方面或第三方面的第一种或第二种可能的实现方式,在第三方面的第三种可能的实现方式中,该至少一个分布式集群还包括第二分布式集群,该第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考
25 节点,以及作为该第二 Master 节点的备份的第二备用节点,该设备还包括:

第二接收模块,用于接收该第二参考节点发送的心跳消息,该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信息,该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息,该第二分布式集群与该第一
30 分布式集群具有相同的集群标识;

第四确定模块,用于根据该第二接收模块接收的该第二指示信息,确定

该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该设备的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

- 5 协商模块，用于与该第二分布式集群中的第二备用节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

结合第三方面的第三种可能的实现方式，在第三方面的第四种可能的实现方式中，该协商模块包括：

- 10 第一发送单元，用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备的节点属性为备用节点的广播消息；

接收单元，用于接收该第二备用节点发送的协商消息，该协商消息中包括用于指示该第二备用节点的权重的信息，该协商消息是该第二备用节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的；

- 15 第二发送单元，用于在确定该设备的权重大于或等于该第二备用节点的权重的情况下，向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息；

- 20 第三发送单元，用于在确定该设备权重小于该第二备用节点的权重的情况下，向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备降级为 Slave 节点的广播消息。

- 25 第四方面提供了一种用于处理至少一个分布式集群中的故障的设备，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，其特征在于，该第一分布式集群还包括第一参考节点以及作为该第一 Master 节点的备份的第一备用节点，该第一备用节点用于，在检测到该第一 Slave 节点向该第一备用节点发送的心跳消息中断的情况下，向该第一 Master 节点发送心跳消息，该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息，该设备为该第一 Master 节点，该设备包括：

- 30 第一接收模块，用于接收该第一备用节点发送的包括该第三指示信息的心跳消息；

第一确定模块，用于根据该第一接收模块接收的该第三指示信息，确定

该第一备用节点与该第一 Slave 节点处于失联状态;

第一检测模块, 用于检测该第一 Slave 节点向该设备发送的心跳消息是否中断;

第二确定模块, 用于在该第一检测模块检测到该第一 Slave 节点向该设备发送的心跳消息中断的情况下, 确定该设备与该第一 Slave 节点也处于失联状态;

第三确定模块, 用于根据该第一确定模块与该第二确定模块确定的结果, 确定该第一 Slave 节点发生故障。

结合第四方面, 在第四方面的第一种可能的实现方式中, 该第一检测模块具体用于, 检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息, 该第三时刻为该设备接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻, 该第四时刻早于该第三时刻, 该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该设备发送心跳消息的心跳周期的 N 倍, N 为正整数。

结合第四方面或第四方面的第一种可能的实现方式, 在第四方面的第二种可能的实现方式中, 该设备还包括:

第二检测模块, 用于在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息, 该预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

第四确定模块, 用于在该第二检测模块在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息, 也没有接收到该第一参考节点发送的心跳消息的情况下, 确定该第一备用节点和该第一参考节点均发生故障。

结合第四方面或第四方面的第一种或第二种可能的实现方式, 在第四方面的第三种可能的实现方式中, 该至少一个分布式集群还包括第二分布式集群, 该第二分布式集群中包括第二 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为该第二 Master 节点的备份的第二备用节点, 该设备还包括:

第二接收模块, 用于接收该第二参考节点发送的心跳消息, 该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息, 该第二分布式集群与该第一分布式集群具有相同的集群标识;

第五确定模块,用于根据该第二接收模块接收的该第四指示信息,确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该设备的节点属性不相符,并确定该第二参考节点所在的第二分布式集群与该设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

协商模块,用于与该第二分布式集群中的第二 Master 节点进行协商,以便于整合该第一分布式集群和该第二分布式集群。

结合第四方面的第三种可能的实现方式,在第四方面的第四种可能的实现方式中,该协商模块包括:

10 第一发送单元,用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备的节点属性为 Master 节点的广播消息;

接收单元,用于接收该第二 Master 节点发送的协商消息,该协商消息中包括用于指示该第二 Master 节点的权重的信息,该协商消息是该第二 Master 节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的;

第二发送单元,用于在确定该设备的权重大于或等于该第二 Master 节点的权重时,向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息;

20 第三发送单元,用于在确定该设备的权重小于该第二 Master 节点的权重时,向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备降级为 Slave 节点的广播消息。

第五方面,提供了一种分布式集群系统,该分布式集群系统包括第一分布式集群,该第一分布式集群中包括第一 Master 节点和第一 Slave 节点,其特征在于,该分布式集群还包括第一参考节点,以及作为该第一 Master 节点的备份的第一备用节点,该第一 Master 节点、该第一备用节点和该第一参考节点用于彼此之间互发心跳消息,该第一参考节点还用于向该第一 Slave 节点发送心跳消息,该第一 Slave 节点用于分别向该第一 Master 节点和该第一备用节点发送心跳消息,其中,

30 该第一参考节点还用于,在检测到该第一 Master 节点向该第一参考节点发送的心跳消息中断的情况下,向该第一备用节点发送心跳消息,该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一

指示信息,

该第一备用节点包括:

第一接收模块,用于接收该第一参考节点发送的包括该第一指示信息的心跳消息;

5 第一确定模块,用于根据该第一接收模块接收的该第一指示信息,确定该第一参考节点与该第一 Master 节点处于失联状态;

检测模块,用于检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断;

10 第二确定模块,用于在该检测模块检测到该第一 Master 节点向该第一备用节点发送的心跳消息中断的情况下,确定该第一备用节点与该第一 Master 节点也处于失联状态;

第三确定模块,用于根据该第一确定模块和该第二确定模块确定的结果,确定该第一 Master 节点发生故障。

15 结合第五方面,在第五方面的第一种可能的实现方式中,该第一备用设备还包括:

发送模块,用于在该第三确定模块确定该第一 Master 节点发生故障的情况下,向该第一分布式集群中所有节点发送广播消息,该广播消息用于指示该第一备用节点升级为该第一 Master 节点。

20 结合第五方面或第五方面的第一种可能的实现方式,在第五方面的第二种可能的实现方式中,该检测模块具体用于,检测在第一时刻与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息,该第一时刻为该第一备用节点接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻,该第二时刻早于该第一时刻,该第一时刻与该第二时刻之间的时间间隔为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的 N 倍, N
25 为正整数。

结合第五方面或第五方面的第一种或第二种可能的实现方式,在第五方面的第三种可能的实现方式中,该分布式集群系统还包括第二分布式集群,该第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为该第二 Master 节点的备份的第二备用节点,该第一备用节点还包
30 括:

第二接收模块,用于接收该第二参考节点发送的心跳消息,该心跳消息

包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息, 该第二分布式集群与该第一分布式集群具有相同的集群标识;

5 第四确定模块, 用于根据该第二接收模块接收的该第二指示信息, 确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一备用节点的节点属性不相符, 并确定该第二参考节点所在的第二分布式集群与该第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

10 协商模块, 用于与该第二分布式集群中的第二备用节点进行协商, 以便于整合该第一分布式集群和该第二分布式集群。

结合第五方面的第三种可能的实现方式, 在第五方面的第四种可能的实现方式中, 该协商模块包括:

15 第一发送单元, 用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点的节点属性为备用节点的广播消息;

接收单元, 用于接收该第二备用节点发送的协商消息, 该协商消息中包括用于指示该第二备用节点的权重的信息, 该协商消息是该第二备用节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的;

20 第二发送单元, 用于在确定该第一备用节点的权重大于或等于该第二备用节点的权重的情况下, 向该第二备用节点发送用于指示该第二备用节点降为 Slave 节点的协商响应消息;

25 第三发送单元, 用于在确定该第一备用节点权重小于该第二备用节点的权重的情况下, 向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点降为 Slave 节点的广播消息。

30 第六方面提供了一种分布式集群系统, 其特征不在于, 该分布式集群系统包括第一分布式集群, 该第一分布式集群中包括第一 Master 节点和第一 Slave 节点, 其特征不在于, 该分布式集群还包括第一参考节点, 以及作为该第一 Master 节点的备份的第一备用节点, 该第一 Master 节点、该第一备用节点和该第一参考节点用于彼此之间互发心跳消息, 该第一参考节点还用于向该第一 Slave 节点发送心跳消息, 该第一 Slave 节点用于分别向该第一

Master 节点和该第一备用节点发送心跳消息，其中，

该第一备用节点还用于，在检测到该第一 Slave 节点向该第一备用节点发送的心跳消息中断的情况下，向该第一 Master 节点发送心跳消息，该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息，

该第一 Master 节点包括：

第一接收模块，用于接收该第一备用节点发送的包括该第三指示信息的心跳消息；

第一确定模块，用于根据该第一接收模块接收的该第三指示信息，确定该第一备用节点与该第一 Slave 节点处于失联状态；

第一检测模块，用于检测该第一 Slave 节点向该第一 Master 节点发送的心跳消息是否中断；

第二确定模块，用于在该第一检测模块检测到该第一 Slave 节点向该第一 Master 节点发送的心跳消息中断的情况下，确定该第一 Master 节点与该第一 Slave 节点也处于失联状态；

第三确定模块，用于根据该第一确定模块与该第二确定模块确定的结果，确定该第一 Slave 节点发生故障。

结合第六方面，在第六方面的第一种可能的实现方式中，该第一检测模块具体用于，检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息，该第三时刻为该第一 Master 节点接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻，该第四时刻早于该第三时刻，该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周期的 N 倍， N 为正整数。

结合第六方面或第六方面的第一种可能的实现方式，在第六方面的第二种可能的实现方式中，该第一 Master 节点还包括：

第二检测模块，用于在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息，该预设检测周期为发送心跳消息的心跳周期的 M 倍， M 为正整数；

第四确定模块，用于在该第二检测模块在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息，也没有接收到该第一参考节点发送的心跳消息的情况下，确定该第一备用节点和该第一参考节点均发生故障。

结合第六方面或第六方面的第一种或第二种可能的实现方式，在第六方面的第三种可能的实现方式中，该分布式集群系统还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该第一 Master 节点还包括：

第二接收模块，用于接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；

第五确定模块，用于根据该第二接收模块接收的该第四指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一 Master 节点的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

协商模块，用于与该第二分布式集群中的第二 Master 节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

结合第六方面的第三种可能的实现方式，在第六方面的第四种可能的实现方式中，该协商模块包括：

第一发送单元，用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点的节点属性为 Master 节点的广播消息；

接收单元，用于接收该第二 Master 节点发送的协商消息，该协商消息中包括用于指示该第二 Master 节点的权重的信息，该协商消息是该第二 Master 节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一 Master 节点发送的；

第二发送单元，用于在确定该第一 Master 节点的权重大于或等于该第二 Master 节点的权重时，向该第二 Master 节点发送用于指示该第二 Master 节点降级的 Slave 节点的协商响应消息；

第三发送单元，用于在确定该第一 Master 节点的权重小于该第二 Master 节点的权重时，向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点降级的 Slave 节点的广播消息。

基于上述技术方案，在本发明实施例中，该分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点，该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态，并通过检测该 Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态，从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。在本发明实施例中，备用节点通过参考节点发送的用于指示 Master 节点失联的心跳消息，以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Master 节点发生故障，相比于现有技术，能够有效缩短 Master 节点的故障检测时间。

附图说明

为了更清楚地说明本发明实施例的技术方案，下面将对实施例或现有技术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性的前提下，还可以根据这些附图获得其他的附图。

图 1 示出了本发明实施例涉及的分布式集群的示意图。

图 2 示出了本发明实施例提供的处理至少一个分布式集群中的故障的方法的示意性流程图。

图 3 示出了本发明实施例提供的处理至少一个分布式集群中的故障的方法的另一示意性流程图。

图 4 示出了本发明实施例提供的处理至少一个分布式集群中的故障的方法的再一示意性流程图。

图 5(a)，(b)，(c)和(d)示出了本发明实施例提供的处理至少一个分布式集群中的故障的方法的示意图。

图 6 示出了本发明实施例提供的处理至少一个分布式集群中的故障的方法的再一示意性流程图。

图 7 示出了本发明实施例提供的用于处理至少一个分布式集群中的故障的设备的示意性框图。

图 8 示出了本发明实施例提供的用于处理至少一个分布式集群中的故障的设备的再一示意性框图。

图 9 示出了本发明实施例提供的用于处理至少一个分布式集群中的故障的设备的再一示意性框图。

图 10 示出了本发明实施例提供的用于处理至少一个分布式集群中的故障的设备的再一示意性框图。

5 图 11 示出了本发明实施例提供的分布式集群系统的示意性框图。

图 12 示出了本发明实施例提供的分布式集群系统的另一示意性框图。

具体实施方式

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
10 清楚、完整地描述，显然，所描述的实施例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。

为了便于理解本发明实施例提供的数据处理方案，首先介绍以下概念：

15 心跳消息

心跳消息（Heartbeat Message）是一种发送源发送到接收方的消息，这种消息可以让接收方确定发送源是否以及何时出现故障或终止。通常，心跳消息从发送源启动时开始发送，直到发送源关闭，期间发送源会不间断的发送周期性或重复的心跳消息。当接收方在某个心跳消息接收周期内没有接收到心跳消息，接收方可能会认为发送源已经关闭、或出现故障，或者当前不可用。
20

心跳消息可能常常被用于高可用性或容错处理的目的。

为了便于本领域技术人员理解本发明实施例提出的处理至少一个分布式集群中的故障的方法、设备和系统，下面首先结合图 1 介绍本发明实施例
25 涉及到的分布式集群。

如图 1 所示，分布式集群 100 包括 Master 节点（主节点）110，备用节点 120，参考节点 130 和 Slave 节点（从属节点）140。其中，Master 节点 110，备用节点 120 和参考节点 130 三者之间互发心跳消息；参考节点 130 还向 Slave 节点 140 发送心跳消息；Slave 节点 140 分别向主节点 110 和备用节点
30 120 发送心跳消息。

如图 1 所示，Master 节点 110，备用节点 120 和参考节点 130 三者之中

的两两之间发送心跳消息，具体地，Master 节点 110 分别向备用节点 120 和参考节点 130 发送心跳消息，备用节点 120 分别向 Master 节点 110 和参考节点 130 发送心跳消息，参考节点 130 分别向 Master 节点 110 和备用节点 120 发送心跳消息。即 Master 节点 110，备用节点 120 和参考节点 130 三者互相
5 监控彼此是否存在故障。应理解，如果一个节点存在故障，则该节点就无法向其他节点发送心跳消息，也无法接收到其他节点发送的心跳消息。

如图 1 所示，由于 Slave 节点 140 分别向主节点 110 和备用节点 120 发送心跳消息，则 Master 节点 110 和备用节点 120 均可以监控 Slave 节点 130 的故障与否。

10 由于参考节点 130 向 Slave 节点 140 发送心跳消息，则 Slave 节点 140 也可以监控该参考节点 130。

应理解，分布式集群 100 中包括的所有节点，均具有基本的本地业务计算对象故障监控功能，该业务计算对象例如为进程等。其中，

Master 节点 110 用于，负责业务计算对象（例如进程）的调度，例如为
15 应用层的业务集群提供仲裁，具体地包括主备仲裁，集群一致性协调控制等。该 Master 节点监控分布式集群 100 中所有节点。

备用节点 120 用于，作为 Master 节点 110 的备用节点，即作为 Master 节点的备份，例如，在 Master 节点 110 发生故障时，该备用节点 120 快速升主接替该 Master 节点 110 业务，该备用节点 120 也监控分布式集群 100 中所
20 有节点。

参考节点 130 用于，实时监控 Master 节点 110 和备用节点 120 发送的心跳消息，以便于及时地发现 Master 节点 110 或备用节点 120 发生故障。具体地，例如，在参考节点 130 检测到 Master 节点 110 向参考节点 130 发送的心跳消息中断的情况下，认为与该 Master 节点 110 失联，然后向备用节点 120
25 发送携带有指示信息的心跳消息，该指示信息用于指示该参考节点 130 与该 Master 节点 110 处于失联状态，以便于该备用节点 120 根据该指示信息，能够有效地确定出该 Master 节点 110 发生故障。例如，当备用节点 120 也检测到 Master 节点 110 向备用节点 120 发送的心跳消息中断的情况下，该备用节点 120 可以综合考虑自身的检测结果以及参考节点 130 发送的心跳消息中的
30 该指示信息，较为快捷、准确地确定该 Master 节点 110 确实发生故障。

类似的，参考节点 130 也可以协助 Master 节点 110 快捷、准确地确定备

用节点 120 发生故障。

应理解，集群中各个节点发送的心跳消息，可以基于相同的心跳周期来发送，换句话说，每个节点都是按照预设的心跳周期来发送心跳消息，相应地，也按照该预设的心跳周期来接收心跳消息。因此，例如，当参考节点 130 在一个心跳周期内没有接收到 Master 节点 110 发送的心跳消息，就可以认为该 Master 节点 110 的心跳消息中断，进而认为与该 Master 节点 110 失联。

应理解，该参考节点 130 也可以根据预设的预设检测周期来判断 Master 节点 110 的心跳消息是否中断，该预设检测周期例如设置为心跳周期的 3 倍或其他整数倍。

综上所述，该参考节点 130 作为 Master 节点 110 和备用节点 120 的参考方，用于协助 Master 节点 110 和备用节点 120 快速有效地检测彼此是否存在故障，可以有效增强对 Master 节点 110 的故障判断的准确性。

Slave 节点 140，为集群 100 中除上述 Master 节点 110、备用节点 120 和参考节点 130 以外的其它所有节点。该 Slave 节点 140 用于，负责搜集本地业务计算对象（例如进程）的状态信息，并将该状态信息上报给 Master 节点 110，并执行 Master 节点 110 下发的指令。

可选地，在本发明实施例涉及的分布式集群 100 包括一个 Master 节点 110 和一个备用节点 120，可以包括多个参考节点 130 和多个 Slave 节点 140。

具体地，当该分布式集群 100 为虚拟机集群时，该集群 100 中的所有节点均为虚拟机。再例如，当该分布式集群 100 为物理组机集群时，该集群 100 中的所有节点均为物理组设备。此外，上述各个节点也可以为逻辑站点等。

图 2 示出了本发明实施例的处理至少一个分布式集群中的故障的方法 200，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，该分布式集群还包括第一参考节点，以及作为该第一 Master 节点的备份的第一备用节点，该方法 200 包括：

S210，该第一备用节点接收该第一参考节点发送的心跳消息，该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息；

S220，该第一备用节点根据该第一指示信息，确定该第一参考节点与该第一 Master 节点处于失联状态；

S230，该第一备用节点检测该第一 Master 节点向该第一备用节点发送

的心跳消息是否中断;

S240, 在检测到该第一 Master 节点向该第一备用节点发送的心跳消息中断的情况下, 该第一备用节点确定该第一备用节点与该第一 Master 节点也处于失联状态;

- 5 S250, 该第一备用节点确定与该第一备用节点和该第一参考节点均失联的该第一 Master 节点发生故障。

在本发明实施例中, 该分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点, 该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态, 并通过检测该
10 Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态, 从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。

因此, 本发明实施例提供的处理至少一个分布式集群中的故障的方法, 备用节点通过参考节点发送的用于指示 Master 节点失联的心跳消息, 以及自
15 身检测到的该 Master 节点发送的心跳消息中断的检测结果, 可以快速、有效地确定该 Master 节点发生故障, 相比于现有技术, 能够有效缩短 Master 节点的故障检测时间。

可选地, 本发明实施例提供的处理至少一个分布式集群中的故障的方法
200 所涉及的第一分布式集群可以是如图 1 所示的分布式集群 100, 本发明
20 实施例中的第一 Master 节点、第一备用节点、第一参考节点和第一 Slave 节点分别对应于图 1 所示的分布式集群 100 中的 Master 节点 110, 备用节点 120, 参考节点 130 和 Slave 节点 140。

在本发明实施例中, 各个节点发送的心跳消息, 可以基于相同的心跳周期来发送, 换句话说, 每个节点都是按照预设的心跳周期来发送心跳消息,
25 相应地, 也按照该预设的心跳周期来接收心跳消息。该心跳周期可以是系统预设的。

在本发明实施例中, 可选地, 第一参考节点当在预设检测周期内没有接收到第一 Master 节点的心跳消息时, 确定与第一 Master 节点处于失联状态, 然后在下一个心跳周期, 向该第一备用节点发送包括用于指示该第一参考节点
30 点与该第一 Master 节点处于失联状态的第一指示信息的心跳消息。其中, 该预设检测周期可以等于 N 倍的心跳周期, N 可以为 1 或大于 1 的整数, 具体

地, 例如 N 为 4~7。

还应理解, 一个分布式集群中所有节点的节点属性(即节点的角色信息, 例如 Master 节点, Slave 节点, 参考节点和备用节点)都是公知的, 例如, 每一个 Slave 节点都知道该分布式集群中哪个是 Master 节点, 哪个是备用节点, 哪个是参考节点。

可选地, 在本发明实施例中, 各个节点发送的心跳消息中均包括该心跳消息的发送端节点所认为的该心跳消息的接收端节点的节点属性信息(Message Type)。

例如, 在图 1 所示的分布式集群 100 中, Master 节点 110 向备用节点 120 发送的心跳消息中包括的心跳消息的接收端的节点属性为备用节点; 再例如, Master 节点 110 发送给参考节点 130 的心跳消息中包括用于表示参考节点的节点属性信息。

可选地, 在本发明实施例中, S230 该第一备用节点检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断, 包括:

该第一备用节点检测在第一时间与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息, 该第一时间为该第一备用节点接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻, 该第二时刻早于该第一时间, 该第一时间与该第二时刻之间的时间间隔为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

具体地, 例如在第一时间 t1 该第一备用节点接收到第一参考节点发送的包括有该第一指示信息的心跳消息, 该第一备用节点检测在该第一时间 t1 与该第一时间 t1 之前的第二时刻 t2 之间, 是否接收到该第一 Master 节点发送的心跳消息。

该第一时间与该第二时刻之间的时间间隔可以为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的整数倍, 例如 1 倍、4 倍或 7 倍。

应理解, 在 S230 中, 第一备用节点还可以采用其他可行的方案, 来检测自己与第一 Master 节点是否处于失联状态, 例如, 该第一备用节点还可以在接收到第一参考节点的包括该第一指示信息的心跳消息之前, 自主地判断自己是否与第一 Master 节点处于失联状态, 本发明实施例对此不作限定。

因此, 本发明实施例提供的处理至少一个分布式集群中的故障的方法, 备用节点通过参考节点发送的包括第一指示信息的心跳消息以及自己的检

测结果，确定 Master 节点与备用节点和参考节点均失联，从而确定出该 Master 节点发生故障。能够较为高效、准确地判断分布式集群中的 Master 节点是否发生故障，相对于现有技术，能够有效缩短故障检测时间，也能够提高故障检测准确性。

5 可选地，如图 3 所示，在本发明实施例中，在确定该第一 Master 节点发生故障的情况下，该方法 200 还包括：

S260，该第一备用节点向该第一分布式集群中所有节点发送广播消息，该广播消息用于指示该第一备用节点升级为该第一 Master 节点。

10 具体地，该第一备用节点更新为该第一分布式集群内新的第一 Master 节点，即接替了原来第一 Master 节点（即发生故障的第一 Master 节点）的相关业务。

可选地，在本发明实施中，该方法 200 还包括：

S270，由原有第一备用节点升级后的第一 Master 节点重新确定该第一分布式集群中的新的第一备用节点。

15 具体地，例如该由原有第一备用节点升级后的第一 Master 节点从该第一分布式集群的第一 Slave 节点中选出新的第一备用节点，该新的第一备用节点与原有的第一备用节点具有相同功能，即用于监控第一分布式集群内的第一 Master 节点和第一参考节点，并在第一 Master 节点发生故障时，接替其业务。

20 因此，在本发明实施例中，当备用节点确定 Master 节点发生故障后，快速升级为新的 Master 节点，跳过了目前技术中临时选举新的 Master 节点的环节，节省了选举新的 Master 节点所消耗的时间，从而能够较为快捷地恢复集群中存在的故障，有利于业务的延续性。

25 应理解，在本发明实施例中，如果第一备用节点在收到第一参考节点发送的包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息的心跳消息后，检测到自己与该第一 Master 节点没有失联，则没有上述 S250、S260 和 S270 的动作。

可选地，在本发明实施例中，第一参考节点的数量可以有多个。

30 具体地，结合图 1 所示，例如分布式集群 100 中包括 3 个参考节点 131、132 和 133，且每个参考节点都与 Master 节点 110 和备用节点 120 构成一个由三个节点组成的监控环，即在该监控环内，三个节点之间互发心跳消息，利用

方法 200 中描述的方法，互相监控彼此是否为故障节点。

可选地，在本发明实施例中，备用节点 120 分别接收到参考节点 131,132 和 133 发送的心跳消息 A, B 和 C，其中心跳消息 A 中包括有用于指示参考节点 131 与 Master 节点 110 处于失联状态的指示信息，心跳消息 B 中包括
5 有用于指示参考节点 132 与 Master 节点 110 处于失联状态的指示信息，心跳消息 C 中包括有用于指示参考节点 133 与 Master 节点 110 处于失联状态的指示信息；备用节点 120 检测到自身与 Master 节点 110 也处于失联状态；该备用节点 120 根据心跳消息 A, B 和 C 中的指示信息，以及自己的检测信息，推断出，该 Master 节点 110 与参考节点 131,132,133 以及备用节点 120 均处
10 于失联状态，则可以较为准确地确定该 Master 节点 110 发生故障。

因此，在本发明实施例中，分布式集群中包括多个参考节点，可以向备用节点提供较多的参考信息，从而有助于备用节点快速且较为准确地判断出 Master 节点是否发生故障。

还应理解，在上述包括多个参考节点的实施例中，备用节点可以自主地
15 检测自身是否与 Master 节点失联，也可以在接收到心跳消息 A 和/或心跳消息 B 和/或心跳消息 C 的情况下，再去检测自身是否与 Master 节点失联。

结合图 1，在分布式集群 100 中，Master 节点 110 也可以根据方法 200 描述的两方共同监控一方的方法，监控 Slave 节点 140 的故障情况。

图 4 示出了本发明另一实施例提供的处理至少一个分布式集群中的故障
20 的方法 300 的示意性流程图，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，该第一分布式集群还包括第一参考节点以及作为该第一 Master 节点的备份的第一备用节点，该方法 300 包括：

S310，该第一 Master 节点接收该第一备用节点发送的心跳消息，该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息；
25

S320，该第一 Master 节点根据该第三指示信息，确定该第一备用节点与该第一 Slave 节点处于失联状态；

S330，该第一 Master 节点检测该第一 Slave 节点向该第一 Master 节点
30 发送的心跳消息是否中断；

S340，在检测到该第一 Slave 节点向该第一 Master 节点发送的心跳消息

中断的情况下,该第一 Master 节点确定该第一 Master 节点与该第一 Slave 节点也处于失联状态;

S350,该第一 Master 节点确定与该第一 Master 节点和该第一备用节点均失联的该第一 Slave 节点发生故障。

5 在本发明实施例中,分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点,该 Master 节点根据备用节点发送的心跳消息确定备用节点与该 Slave 处于失联状态,并通过检测该 Slave 节点向该 Master 节点发送的心跳消息也中断来确定该 Slave 节点与该 Master 节点也处于失联状态,从而可以有效地确定与该 Master 节点和该备用节点均
10 失联的该 Slave 节点发生故障。

因此,本发明实施例提供的处理至少一个分布式集群中的故障的方法,Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息,以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果,可以快速、有效地确定该 Slave 节点发生故障,相比于现有技术,能够有效缩短 Slave 节点
15 的故障检测时间,同时也能够提高故障检测的准确性。

在本发明实施例中,Master 节点无需像现有技术中一样向所有 Slave 节点发送心跳消息,只需通过检测 Slave 节点发送的心跳消息,实现对 Slave 节点的监控,并结合备用节点发送的心跳消息中的用于指示 Slave 节点失联的指示信息,可以确定出存在故障的 Slave 节点。相对于现有方案,Master
20 节点的心跳消息处理量可以减半,有效降低了 Master 节点的负载。

在本发明实施例中,可选地,第一备用节点当在预设检测周期内没有接收到第一 Slave 节点的心跳消息时,确定与该第一 Slave 节点失联,然后在下一个心跳周期,向第一 Master 节点发送包括用于指示第一备用节点与该第一 Slave 节点失联的第三指示信息的心跳消息。其中,该预设检测周期可以
25 等于 N 倍的心跳消息发送周期,N 可以为 1 或大于 1 的整数。

可选地,在本发明实施例中,该备用节点发送的包括第三指示信息的心跳消息中还包括用于指示该心跳消息的接收端的节点属性的指示信息,其中节点属性也可称之为角色信息(Message Type)。

可选地,在本发明实施例中,S330 该第一 Master 节点检测该第一 Slave
30 节点向该第一 Master 节点发送的心跳消息是否中断,包括:

该第一 Master 节点检测在第三时刻与第四时刻之间是否接收到该第一

Slave 节点发送的心跳消息，该第三时刻为该第一 Master 节点接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻，该第四时刻早于该第三时刻，该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周期的 N 倍，N 为正整数。

5 具体地，例如在第三时刻 t3 该第一 Master 节点接收到第一备用节点发送的包括有该第三指示信息的心跳消息，该第一 Master 节点检测在该第三时刻 t2 与该第三时刻 t3 之前的第四时刻 t4 之间，是否接收到该第一 Slave 节点发送的心跳消息。

10 该第三时刻与该第四时刻之间的时间间隔可以为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周期的整数倍，例如，1 倍、4 倍或 7 倍。

应理解，在 S330 中，该第一 Master 节点还可以采用其他可行的方案，来检测与第一 Slave 节点是否处于失联状态，例如，该第一 Master 节点还可以在接收到第一备用节点发送的包括该第三指示信息的心跳消息之前，自主地判断自己是否与第一 Slave 节点处于失联状态，本发明实施例对此不作限定。

20 因此，在本发明实施例中，Master 节点根据备用节点发送的包括第三指示信息的心跳消息以及自己对 Slave 节点的检测信息，通过确定 Master 节点和备用节点均与该 Slave 节点均处于失联状态，从而能够较为高效、准确地确定该 Slave 节点发生故障，相比于现有的处理集群中故障的方案，有效缩短了故障检测的时间，同时也能够降低在故障检测过程中 Master 节点的能耗。

25 现有的分布式集群中包括两种角色的节点，即 Master 节点和 Slave 节点，通过互发心跳消息来监控彼此是否发生故障，由于监控信息的单一，一般需要通过多次检测心跳消息，例如，Slave 节点在多个预设检测周期（心跳周期的整数倍）内都接收不到 Master 节点发送的心跳消息时，才会认为 Master 节点发生故障，导致故障检测时间较长。

30 在本发明实施例提供的分布式集群中包括四种角色的节点，例如图 1 所示的，Master 节点，Slave 节点，参考节点和备用节点，其中，Master 节点，备用节点和参考节点互发心跳消息，能够互相监控；Slave 节点向 Master 节点和备用节点发送心跳消息，即 Master 节点和备用节点可以共同监控 Slave

节点。参考节点还向 Slave 节点发送心跳消息，用于让 Slave 节点感知到自己所在的集群。例如，Master 节点通过参考备用节点发送的心跳消息和自己对 Slave 节点的监控信息，可以快速、准确地确定 Slave 节点是否发生故障。在本发明实施例中，如果 Master 节点发现备用节点和/或参考节点发生故障，可以在 Slave 节点中重新确定出新的备用节点和/或参考节点；如果备用节点发现 Master 节点发生故障，会快速更新为新的 Master 节点。即在本发明实施例中，集群中四个角色的节点实时地维护这集群的完整性和有效性。

此外，在本发明实施例中，Master 节点只接收 Slave 节点的心跳消息，而不向各个 Slave 节点发送心跳消息，根据备用节点发送的心跳消息和自己的检测结果，就可以有效地监控 Slave 节点的故障情况，同时相比于现有技术，一方面能够提高检测故障的效率，另一方面有效地降低了 Master 节点的负载。

可选地，在本发明实施例中，该方法 300 还包括：

S360，该第一 Master 节点在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息，该预设检测周期为发送心跳消息的心跳周期的 M 倍，M 为正整数；

S370，在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息，也没有接收到该第一参考节点发送的心跳消息的情况下，该第一 Master 节点确定该第一备用节点和该第一参考节点均发生故障。

具体地，该预设检测周期可以是系统预设的用于判断故障的时间段，例如可以直接等于心跳周期，也可以为心跳周期的 4 倍或 7 倍。

可选地，在本发明实施例中，该方法 300 还包括：

S380，当该第一 Master 节点确定该第一备用节点和该第一参考节点均发生故障的情况下，重新确定该第一分布式集群中的新的第一备用节点和新的第一参考节点。

具体地，例如利用分布式选举算法、从满足某一原则（如心跳正常）的节点集合里面随机指定新的备用节点和新的参考节点。

在本发明实施例中，Master 节点利用多路径检测结果，综合判断 Slave 节点的故障情况，一方面可以缩短检测故障的时间，另一方面可以提高故障检测的可靠性。

一个网络集群因交换节点或路由节点发生故障而导致分裂成多个子集

群是一种典型的集群故障场景。本发明实施例涉及的包括四个角色的集群，可以有效避免或者解决上述集群分裂的问题。

可选地，在本发明实施例中，

该方法 300 还包括：该至少一个分布式集群还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该方法 300 还包括：

S3010，该第一 Master 节点接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；

具体地，该第一分布式集群与该第二分布式集群具有相同的集群标识，所以，第二分布式集群中的第二参考节点能够感知到第一分布式集群中的节点，并将第一分布式集群中的所有节点都看作为 Slave 节点，因此，第二参考节点分别向第二分布式集群中的第二 Slave 节点和第一分布式集群中的所有节点发送包括用于指示接收端的节点属性为 Slave 节点的心跳消息。

S3020，该第一 Master 节点根据该第四指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一 Master 节点的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

S3030，该第一 Master 节点与该第二分布式集群中的第二 Master 节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

可选地，在本发明实施例中，S3030 第一 Master 节点与该第二分布式集群中的第二 Master 节点进行协商，包括：

S3031，该第一 Master 节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点的节点属性为 Master 节点的广播消息；

S3032，该第一 Master 节点接收该第二 Master 节点发送的协商消息，该协商消息中包括用于指示该第二 Master 节点的权重的信息，该协商消息是该第二 Master 节点在检测到该广播消息指示的节点属性与自身的节点属性相

同的情况下向该第一 Master 节点发送的;

S3033, 在确定该第一 Master 节点的权重大于或等于该第二 Master 节点的权重时, 该第一 Master 节点向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息;

- 5 S3034, 在确定该第一 Master 节点的权重小于该第二 Master 节点的权重时, 该第一 Master 节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点降级为 Slave 节点的广播消息。

具体地, 如图 5(a)所示, 一个网络集群分裂为两个子集群 A 和 B, 子集群 A 和 B 是由一个网络集群分裂而来, 因此该子集群 A 和 B 具有相同的集群标识。记子集群 A 包括 Master 节点 A, 备用节点 A, 参考节点 A 和 Slave 节点 A; 子集群 B 包括 Master 节点 B, 备用节点 B, 参考节点 B 和 Slave 节点 B。子集群 A 和子集群 B 中的各个节点都具有如图 1 所示的分布式集群 100 中的对应节点的特性。

15 在子集群 B 中的参考节点 B 看来, 它能够接收到 Master 节点 B 和备用节点 B 的心跳消息, 所以, 在参考节点 B 看来, 除了 Master 节点 B 和备用节点 B 之外, 其他节点都是 Slave 节点, 即子集群 A 中的所有节点在参考节点 B 看来, 都是处于失联状态下的 Slave 节点。

20 如图 5(b)所示, 子集群 B 中的参考节点 B 除了向子集群 B 中的 Master 节点 B 发送心跳消息(接收端的节点属性为 Master)、向子集群 B 中的备用节点 B 发送心跳消息(接收端的节点属性为备用)、向子集群 B 中的 Slave 节点 B 发送心跳消息(接收端的节点属性为 Slave)之外, 所以还会向子集群 A 中的所有节点发送心跳消息, 其该心跳消息中指示的接收端的节点属性为 Slave 节点。

25 子集群 A 中的所有节点都接收到子集群 B 中的参考节点 B 发送的包括用于指示接收端的节点属性为 Slave 节点的指示信息的心跳消息, 子集群 A 中的 Master 节点 A, 备用节点 A 和参考节点 A 接收到参考节点 B 发送的该心跳消息后, 会感知到该心跳消息携带的节点属性与自身的节点属性信息不相符。

30 以子集群 A 中的 Master 节点 A 为例, 子集群 A 中的 Master 节点 A 根据参考节点 B 发送的心跳消息, 确定该心跳消息携带的节点属性与自身的节点属性信息不相符, 子集群 A 中的 Master 节点 A 向所有节点发送用于指示

自己的节点属性的广播消息，如图 5(c)所示。

子集群 B 中 Master 节点 B 接收到 Master 节点 A 发送的广播消息后，检测到角色冲突，向子集群 A 中的 Master 节点 A 发送协商消息，该协商消息中携带有 Master 节点 B 的权重信息。

5 Master 节点 A 接收到 Master 节点 B 发送的协商消息后，比较自身与 Master 节点 B 的权重，如果判断 Master 节点 A 的权重高于 Master 节点 B 的权重，则 Master 节点 A 向 Master 节点 B 发送用于指示 Master 节点 B 降级为 Slave 节点的协商响应消息；如果判断 Master 节点 B 的权重高于 Master 节点 A 的权重，则该 Master 节点 A 主动降级为 Slave 节点。

10 具体地，节点的权重例如为：节点存活时长，节点本地缓存的新旧程度（一般用版本号来表征），或者所管理的集群的规模等等。

应理解，图 5(c)只示出了以 Master 节点 A 为例的示意图，实际中，接收到参考节点 B 发送的用于指示接收端节点属性为 Slave 节点的心跳消息后，子集群 A 中的备用节点 B 和参考节点 B 也会判断出该心跳消息携带的
15 节点属性与自身的节点属性信息不相符，也会发送包括自己节点属性的广播消息。备用节点 A 和备用节点 B 也会采用类似图 5(c)和图 5(d)所述的示意图，协商最终的备用节点，参考节点 A 和参考节点 B 也会采用类似图 5(c)和图 5(d)所述的示意图，协商最终的参考节点。

可选地，在图 2 所示的实施例提供的方法 200 中，该至少一个分布式集
20 群还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该方法 200 还包括：

S2010，该第一备用节点接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示
25 信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；

S2020，该第一备用节点根据该第二指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一备用节
30 点的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两

个子集群;

S2030, 该第一备用节点与该第二分布式集群中的第二备用节点进行协商, 以便于整合该第一分布式集群和该第二分布式集群。

可选地, 在本发明实施例中, S2030 该第一备用节点与该第二分布式集群中的第二备用节点进行协商, 包括:

S2031, 该第一备用节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点的节点属性为备用节点的广播消息;

S2032, 该第一备用节点接收该第二备用节点发送的协商消息, 该协商消息中包括用于指示该第二备用节点的权重的信息, 该协商消息是该第二备用节点在检测到该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一备用节点发送的;

S2033, 在确定该第一备用节点的权重大于或等于该第二备用节点的权重的情况下, 该第一备用节点向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息;

S2034, 在确定该第一备用节点的权重小于该第二备用节点的权重的情况下, 该第一备用节点向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点降级为 Slave 节点的广播消息。

具体流程与图 5 所示一致, 这里不再赘述。

应理解, 本发明实施例中的第一分布式集群可以对应于图 5 所示的子集群 A, 第二分布式集群对应于图 5 中的子集群 B。

还应理解, 图 5(a)只示出了, 子集群 B 中的参考节点 B 将子集群 A 中的所有节点看作 Slave 节点, 而向其发送包括有指示接收端的节点属性为 Slave 节点的心跳消息。还应理解, 在子集群 A 与子集群 B 合并为一个集群之前, 子集群 A 中的参考节点 A 也会将子集群 B 中的所有节点看作为 Slave 节点, 也会向其发送包括有指示接收端的节点属性为 Slave 节点的心跳消息。子集群 B 中的 Master 节点 B, 参考节点 B, 以及备用节点 B 也会执行图 5 (b), 图 5(c)和图 5 (d)所示的动作, 直至子集群 A 和子集群 B 整合为一个集群为止。

还应理解, 上面结合图 5 只描述了一个分布式集群分裂为两个子集群后, 将其整合恢复的过程, 本发明实施例提供的集群整合恢复的方案同样适应于, 将两个以上分裂的分布式子集群整合为一个分布式集群的场合, 本发明

实施例对此不作限定。

在本发明实施例中，各个节点发送的心跳消息中都携带该心跳消息的发送端节点所认为的心跳消息的接收端节点的节点属性信息（角色信息），这样能够让接收心跳消息的节点，感知到集群的分裂，从而触发后续的动作，

5 实现分裂的集群的整合。

综上所述，本发明实施例提供的处理至少一个分布式集群中的故障的方法，能够有效缩短集群中故障检测的时间，并能够有效提高故障检测的准确性，同时也能够降低 Master 节点处理心跳消息的负担。

10 上文结合图 1 至图 3，描述了在本发明实施例提出的分布式集群中，备用节点确定 Master 节点发生故障的过程，结合图 1 和图 4，描述了在本发明实施例提出的分布式集群中，Master 节点确定 Slave 节点发生故障的过程。下面从参考节点的角度描述本发明实施例提供的用于处理分布式集群的故障的方法。

15 图 6 示出了本发明另一实施例提供的处理至少一个分布式集群中的故障的方法 400，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，该分布式集群还包括第一参考节点，以及作为该第一 Master 节点的备份的第一备用节点，该方法 400 包括：

S410，该第一参考节点检测该第一 Master 节点向该第一参考节点发送的心跳消息是否中断；

20 S420，当确定该第一 Master 向该第一参考节点发送的心跳消息中断时，该第一参考节点确定与该第一 Master 节点处于失联状态；

S430，该第一参考节点向第一备用节点发送心跳消息，该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息，以便于所述第一备用节点判断所述第一 Master 节点是否为故障节点。

25 在本发明实施例中，参考节点在确定与 Master 节点处于失联状态的情况下，向备用节点发送包括用于指示该状况的指示信息的心跳消息，从而以便于备用节点，通过参考该指示信息，进一步判断该 Master 节点是否发生故障。有助于备用节点根据多路径的检测信息综合确定分布式集群中的故障情况，能够有效提高故障检测的效率和准确性。

30 可选地，在本发明实施例中，S410 该第一参考节点检测该第一 Master 节点向该第一参考节点发送的心跳消息是否中断，包括：

该第一参考节点在预设检测周期内,检测是否接收到该第一 Master 发送的心跳消息。

具体地,该预设检测周期可以为该第一 Master 节点向该第一参考节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

5 可选地,在本发明实施例中,该至少一个分布式集群还包括第二分布式集群,该第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为该第二 Master 节点的备份的第二备用节点,该第二分布式集群与第一分布式集群为一个网络集群分裂后形成的两个子集群,该方法 400 还包括:

10 该第一参考节点分别向该第一分布式集群中的第一 Slave 节点,以及该第二分布式集群中的所有节点发送心跳消息,该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的指示信息,以便于该第二分布式集群中的第二 Master 节点,第二备用节点和第二参考节点能够通过该指示信息,感知到该第二分布式集群。

15 具体地,如图 5(b)所示,子集群 B 中的参考节点 B 向 Slave 节点 B,以及子集群 A 中的所有节点发送包括用于指示接收端为 Slave 节点的指示信息的心跳消息。

应理解,本发明实施例中的第一参考节点可对应于方法 200 和方法 300 中涉及的第二参考节点,还可以对应于图 5(b)所示的参考节点 B。

20 在本发明实施例中,参考节点发送的心跳消息携带用于指示心跳消息的接收端的节点属性的指示信息,这样能够让心跳消息的接收端端节点,感知到集群的分裂,从而触发后续的动作,实现分裂的集群的整合。

综上所述,在本发明实施例提供的处理至少一个分布式集群中的故障的方法中,该分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点
25 和该 Master 节点的备用节点。该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态,并通过检测该 Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态,从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。该 Master 节点根据备用节点发送的心跳消息确定备用节点与
30 该 Slave 处于失联状态,并通过检测该 Slave 节点向该 Master 节点发送的心跳消息也中断来确定该 Slave 节点与该 Master 节点也处于失联状态,从而可

以有效地确定与该 Master 节点和该备用节点均失联的该 Slave 节点发生故障。即本发明实施例提供的分布式集群中包括 Master 节点, Slave 节点, 备用节点和参考节点, 可以通过两方节点的检测结果综合判断第三方节点是否发生故障, 从而能够较为高效、准确地检测故障, 相对于现有技术, 能够有效缩短故障检测时间, 同时也能够降低 Master 节点处理心跳消息的负担。

上文中结合图 1 至图 6, 描述了本发明实施例提供的处理至少一个分布式集群中的故障的方法, 下面将结合图 7 和图 8, 描述本发明实施例提供的用于处理至少一个分布式集群中的故障的设备。

图 7 示出了根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 500 的示意性框图。该至少一个分布式集群包括第一分布式集群, 该第一分布式集群中包括第一 Master 节点和第一 Slave 节点, 该分布式集群还包括第一参考节点, 以及作为该第一 Master 节点的备份的第一备用节点, 其中, 该设备为该第一备用节点, 该设备 500 包括:

第一接收模块 510, 用于接收该第一参考节点发送的包括有该第一指示信息的心跳消息;

第一确定模块 520, 用于根据该第一接收模块接收的该第一指示信息, 确定该第一参考节点与该第一 Master 节点处于失联状态;

检测模块 530, 用于检测该第一 Master 节点向该设备发送的心跳消息是否中断;

第二确定模块 540, 用于在该检测模块检测到该第一 Master 节点向该设备发送的心跳消息中断的情况下, 确定该设备与该第一 Master 节点也处于失联状态;

第三确定模块 550, 用于根据该第一确定模块和该第二确定模块确定的结果, 确定该第一 Master 节点发生故障。

在该第一分布式集群中, 该第一参考节点用于, 在检测到该第一 Master 节点向该第一参考节点发送的心跳消息中断的情况下, 向该第一备用节点发送心跳消息, 该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息。

在本发明实施例中, 该分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点, 该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态, 并通过检测该

Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态,从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。

因此,本发明实施例提供的设备,通过参考节点发送的用于指示 Master 节点失联的心跳消息,以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果,可以快速、有效地确定该 Master 节点发生故障,相比于现有技术,能够有效缩短 Master 节点的故障检测时间。

可选地,本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 500 中涉及的分布式集群可以是如图 1 所示的分布式集群 100。

10 可选地,在本发明实施例中,该设备 500 还包括:

发送模块 560,用于在该第三确定模块确定该第一 Master 节点发生故障的情况下,向该第一分布式集群中所有节点发送广播消息,该广播消息用于指示该设备升级为该第一 Master 节点。

15 可选地,在本发明实施例中,该检测模块具体用于,检测在第一时间与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息,该第一时间为该设备接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻,该第二时刻早于该第一时间,该第一时间与该第二时刻之间的时间间隔为该第一 Master 节点向该设备发送心跳消息的心跳周期的 N 倍, N 为正整数。

20 可选地,在本发明实施例中,该至少一个分布式集群还包括第二分布式集群,该第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为该第二 Master 节点的备份的第二备用节点,该设备 500 还包括:

25 第二接收模块 570,用于接收该第二参考节点发送的心跳消息,该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信息,该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息,该第二分布式集群与该第一分布式集群具有相同的集群标识;

30 第四确定模块 580,用于根据该第二接收模块接收的该第二指示信息,确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该设备的节点属性不相符,并确定该第二参考节点所在的第二分布

式集群与该设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

协商模块 590, 用于与该第二分布式集群中的第二备用节点进行协商, 以便于整合该第一分布式集群和该第二分布式集群。

5 可选地, 在本发明实施例中, 该协商模块 590 包括:

第一发送单元 591, 用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备的节点属性为备用节点的广播消息;

10 接收单元 592, 用于接收该第二备用节点发送的协商消息, 该协商消息中包括用于指示该第二备用节点的权重的信息, 该协商消息是该第二备用节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的;

第二发送单元 593, 用于在确定该设备的权重大于或等于该第二备用节点的权重的情况下, 向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息;

15 第三发送单元 594, 用于在确定该设备权重小于该第二备用节点的权重的情况下, 向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备降级为 Slave 节点的广播消息。

20 因此, 本发明实施例提供的设备, 通过参考节点发送的用于指示 Master 节点失联的心跳消息, 以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果, 可以快速、有效地确定该 Master 节点发生故障, 相比于现有技术, 能够有效缩短 Master 节点的故障检测时间。

25 应理解, 根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 500 可对应于本发明实施例的处理至少一个分布式集群中的故障的方法中的备用节点, 并且设备 500 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程, 为了简洁, 在此不再赘述。

30 图 8 示出了根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 600 的示意性框图。该至少一个分布式集群包括第一分布式集群, 该第一分布式集群中包括第一 Master 节点和第一 Slave 节点, 该第一分布式集群还包括第一参考节点以及作为该第一 Master 节点的备份的第一备用节点, 该设备 600 为该第一 Master 节点, 该设备 600 包括:

第一接收模块 610, 用于接收该第一备用节点发送的包括该第三指示信息的心跳消息;

第一确定模块 620, 用于根据该第一接收模块接收的该第三指示信息, 确定该第一备用节点与该第一 Slave 节点处于失联状态;

5 第一检测模块 630, 用于检测该第一 Slave 节点向该设备发送的心跳消息是否中断;

第二确定模块 640, 用于在该第一检测模块检测到该第一 Slave 节点向该设备发送的心跳消息中断的情况下, 确定该设备与该第一 Slave 节点也处于失联状态;

10 第三确定模块 650, 用于根据该第一确定模块与该第二确定模块确定的结果, 确定该第一 Slave 节点发生故障。

在该第一分布式集群中, 该第一备用节点用于, 在检测到该第一 Slave 节点向该第一备用节点发送的心跳消息中断的情况下, 向该第一 Master 节点发送心跳消息, 该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息。

在本发明实施例中, 分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点, 该 Master 节点根据备用节点发送的心跳消息确定备用节点与该 Slave 处于失联状态, 并通过检测该 Slave 节点向该 Master 节点发送的心跳消息也中断来确定该 Slave 节点与该 Master 节点也处于失联状态, 从而可以有效地确定与该 Master 节点和该备用节点均失联的该 Slave 节点发生故障。

因此, 在本发明实施例中, Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息, 以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果, 可以快速、有效地确定该 Slave 节点发生故障, 相比于现有的处理集群中故障的方案, 有效缩短了故障检测的时间, 同时也能够降低在故障检测过程中 Master 节点的能耗。

可选地, 在本发明实施例中, 该第一检测模块具体用于, 检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息, 该第三时刻为该设备接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻, 该第四时刻早于该第三时刻, 该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该设备发送心跳消息的心跳周期的 N 倍, N 为

正整数。

可选地，在本发明实施例中，该设备 600 还包括：

第二检测模块，用于在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息，该预设检测周期为发送心跳消息的心跳周期的 M 倍， M 为正整数；

第四确定模块，用于在该第二检测模块在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息，也没有接收到该第一参考节点发送的心跳消息的情况下，确定该第一备用节点和该第一参考节点均发生故障。

可选地，在本发明实施例中，该至少一个分布式集群还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该设备 600 还包括：

第二接收模块，用于接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；

第五确定模块，用于根据该第二接收模块接收的该第四指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该设备的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

协商模块，用于与该第二分布式集群中的第二 Master 节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

可选地，在本发明实施例中，该协商模块包括：

第一发送单元，用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备的节点属性为 Master 节点的广播消息；

接收单元，用于接收该第二 Master 节点发送的协商消息，该协商消息中包括用于指示该第二 Master 节点的权重的信息，该协商消息是该第二 Master 节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的；

第二发送单元,用于在确定该设备的权重大于或等于该第二 Master 节点的权重时,向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息;

5 第三发送单元,用于在确定该设备的权重小于该第二 Master 节点的权重时,向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该设备降级为 Slave 节点的广播消息。

10 因此,在本发明实施例中,Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息,以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果,可以快速、有效地确定该 Slave 节点发生故障,相比于现有的处理集群中故障的方案,有效缩短了故障检测的时间,同时也能够降低在故障检测过程中 Master 节点的能耗。

15 应理解,根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 600 可对应于本发明实施例的处理至少一个分布式集群中的故障的方法中的 Master 节点,并且设备 600 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程,为了简洁,在此不再赘述。

20 如图 9 所示,本发明实施例还提供了一种用于处理至少一个分布式集群中的故障的设备 700,该至少一个分布式集群包括第一分布式集群,该第一分布式集群中包括第一 Master 节点和第一 Slave 节点,该分布式集群还包括第一参考节点,以及作为该第一 Master 节点的备份的第一备用节点,该设备 700 为该第一备用节点,该设备 700 包括处理器 710、存储器 720、总线系统 730、接收器 740 和发送器 750。其中,处理器 710、存储器 720、接收器 740 和发送器 750 通过总线系统 730 相连,该存储器 720 用于存储指令,该处理器 710 用于执行该存储器 720 存储的指令,以控制接收器 740 接收信号,并
25 控制发送器 750 发送信号。其中,接收器 740 用于,接收该第一参考节点发送的心跳消息,该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息;该处理器 710 用于,根据该第一指示信息,确定该第一参考节点与该第一 Master 节点处于失联状态;检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断;在检测到该第一
30 Master 节点向该第一备用节点发送的心跳消息中断的情况下,该第一备用节点确定该第一备用节点与该第一 Master 节点也处于失联状态;确定与该第一

备用节点和该第一参考节点均失联的该第一 Master 节点发生故障。

因此，本发明实施例提供的设备，通过参考节点发送的用于指示 Master 节点失联的心跳消息，以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Master 节点发生故障，相比于现有技术，能够有效缩短 Master 节点的故障检测时间。

可选地，在本发明实施例中，发送器 750 还用于，在处理器 710 确定该第一 Master 节点发生故障之后，向该第一分布式集群中所有节点发送广播消息，该广播消息用于指示该第一备用节点升级为该第一 Master 节点。

可选地，在本发明实施例中，处理器 710 具体用于，检测在第一时间与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息，该第一时间为该第一备用节点接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻，该第二时刻早于该第一时间，该第一时间与该第二时刻之间的时间间隔为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的 N 倍，N 为正整数。

可选地，在本发明实施例中，该至少一个分布式集群还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该接收器 740 用于，接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；该处理器 710 用于，根据该第二指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一备用节点的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；与该第二分布式集群中的第二备用节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

可选地，在本发明实施例中，该发送器 750 用于，向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点的节点属性为备用节点的广播消息；该接收器 740 用于，接收该第二备用节点发送的协商消息，该协商消息中包括用于指示该第二备用节点的权重的信息，该协

商消息是该第二备用节点在检测到该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一备用节点发送的；该发送器 750 用于，在确定该第一备用节点的权重大于或等于该第二备用节点的权重的情况下，向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息；在确定该第一备用节点的权重小于该第二备用节点的权重的情况下，向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点降级为 Slave 节点的广播消息。

应理解，在本发明实施例中，该处理器 710 可以是中央处理单元 (Central Processing Unit，简称为“CPU”)，该处理器 710 还可以是其他通用处理器、数字信号处理器 (DSP)、专用集成电路 (ASIC)、现成可编程门阵列 (FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

该存储器 720 可以包括只读存储器和随机存取存储器，并向处理器 710 提供指令和数据。存储器 720 的一部分还可以包括非易失性随机存取存储器。

例如，存储器 720 还可以存储设备类型的信息。

该总线系统 730 除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见，在图中将各种总线都标为总线系统 730。

在实现过程中，上述方法的各步骤可以通过处理器 710 中的硬件的集成逻辑电路或者软件形式的指令完成。结合本发明实施例所公开的方法的步骤可以直接体现为硬件处理器执行完成，或者用处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器 720，处理器 710 读取存储器 720 中的信息，结合其硬件完成上述方法的步骤。为避免重复，这里不再详细描述。

因此，本发明实施例提供的设备，通过参考节点发送的用于指示 Master 节点失联的心跳消息，以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Master 节点发生故障，相比于现有技术，能够有效缩短 Master 节点的故障检测时间。

应理解，根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 700 可对应于本发明实施例的处理至少一个分布式集群中的故障的

方法中的备用节点，也对应于根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 500，并且设备 700 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程，为了简洁，在此不再赘述。

- 5 如图 10 所示，本发明实施例还提供了用于处理至少一个分布式集群中的故障的设备 800，该至少一个分布式集群包括第一分布式集群，该第一分布式集群中包括第一 Master 节点和第一 Slave 节点，该第一分布式集群还包括第一参考节点以及作为该第一 Master 节点的备份的第一备用节点，该设备 800 为该第一 Master 节点，该设备 800 包括处理器 810、存储器 820、总线系统 830、接收器 840 和发送器 850。其中，处理器 810、存储器 820、接收器 840 和发送器 850 通过总线系统 830 相连，该存储器 820 用于存储指令，该处理器 810 用于执行该存储器 820 存储的指令，以控制接收器 840 接收信号，并控制发送器 850 发送信号。其中，该接收器 840 用于，接收该第一备用节点发送的心跳消息，该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息；该处理器 810 用于，根据该第三指示信息，确定该第一备用节点与该第一 Slave 节点处于失联状态；检测该第一 Slave 节点向该第一 Master 节点发送的心跳消息是否中断；在检测到该第一 Slave 节点向该第一 Master 节点发送的心跳消息中断的情况下，该第一 Master 节点确定该第一 Master 节点与该第一 Slave 节点也处于失联状态；点
15 确定与该第一 Master 节点和该第一备用节点均失联的该第一 Slave 节点发生故障。

因此，在本发明实施例中，Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息，以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Slave 节点发生故障，相比于现
25 有的处理集群中故障的方案，有效缩短了故障检测的时间，同时也能够降低在故障检测过程中 Master 节点的能耗。

可选地，在本发明实施例中，该处理器 810 用于，检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息，该第三时刻为该第一 Master 节点接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻，该第四时刻早于该第三时刻，该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周
30

期的 N 倍, N 为正整数。

可选地, 在本发明实施例中, 该处理器 810 用于, 在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息, 该预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

5 在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息, 也没有接收到该第一参考节点发送的心跳消息的情况下, 确定该第一备用节点和该第一参考节点均发生故障。

可选地, 在本发明实施例中, 该至少一个分布式集群还包括第二分布式集群, 该第二分布式集群中包括第二 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为该第二 Master 节点的备份的第二备用节点, 该接收器 840
10 用于, 接收该第二参考节点发送的心跳消息, 该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第四指示信息的心跳消息, 该第二分布式集群与该第一分布式集群具有相
15 同的集群标识; 该处理器 810 用于, 根据该第四指示信息, 确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一 Master 节点的节点属性不相符, 并确定该第二参考节点所在的第二分布式集群与该第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群; 与该第二分布式集群中的第二 Master 节点进行协商,
20 以便于整合该第一分布式集群和该第二分布式集群。

可选地, 在本发明实施例中, 发送器 850 用于, 向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点的节点属性为 Master 节点的广播消息; 接收器 840 用于, 接收该第二 Master 节点发送的协商消息, 该协商消息中包括用于指示该第二 Master 节点的权重的信息,
25 该协商消息是该第二 Master 节点在检测到该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一 Master 节点发送的; 发送器 850 用于, 在确定该第一 Master 节点的权重大于或等于该第二 Master 节点的权重时, 向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息; 在确定该第一 Master 节点的权重小于该第二 Master 节点的权重时, 向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示
30 该第一 Master 节点降级为 Slave 节点的广播消息。

因此,在本发明实施例中,Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息,以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果,可以快速、有效地确定该 Slave 节点发生故障,相比于现有的处理集群中故障的方案,有效缩短了故障检测的时间,同时也能够降低在故障检测过程中 Master 节点的能耗。

应理解,根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 800 可对应于本发明实施例的处理至少一个分布式集群中的故障的方法中的 Master 节点,也可以对应于根据本发明实施例提供的用于处理至少一个分布式集群中的故障的设备 600,并且设备 800 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程,为了简洁,在此不再赘述。

图 11 示出了本发明实施例提供的分布式集群系统 900 的示意性框图,该分布式集群系统包括第一分布式集群 901,该第一分布式集群中包括第一 Master 节点 910 和第一 Slave 节点 920,该分布式集群还包括第一参考节点 930,以及作为该第一 Master 节点的备份的 941,该第一 Master 节点 910、该第一备用节点 940 和该第一参考节点 930 用于彼此之间互发心跳消息,该第一参考节点 930 还用于向该第一 Slave 节点 920 发送心跳消息,该第一 Slave 节点 920 用于分别向该第一 Master 节点 910 和该第一备用节点 930 发送心跳消息,其中,

该第一参考节点 930 还用于,在检测到该第一 Master 节点向该第一参考节点发送的心跳消息中断的情况下,向该第一备用节点发送心跳消息,该心跳消息包括用于指示该第一参考节点与该第一 Master 节点处于失联状态的第一指示信息,

该第一备用节点 940 包括:

第一接收模块 941,用于接收该第一参考节点发送的包括该第一指示信息的心跳消息;

第一确定模块 942,用于根据该第一接收模块接收的该第一指示信息,确定该第一参考节点与该第一 Master 节点处于失联状态;

检测模块 943,用于检测该第一 Master 节点向该第一备用节点发送的心跳消息是否中断;

第二确定模块 944,用于在该检测模块检测到该第一 Master 节点向该第

一备用节点发送的心跳消息中断的情况下，确定该第一备用节点与该第一 Master 节点也处于失联状态；

第三确定模块 945，用于根据该第一确定模块和该第二确定模块确定的结果，确定该第一 Master 节点发生故障。

5 具体地，该第一分布式集群可对应于图 1 所示的分布式集群 100，具体描述可参见上文，为了简洁，这里不再赘述。

在本发明实施例中，该分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点，该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态，并通过检测该 Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态，从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。在本发明实施例中，备用节点通过参考节点发送的用于指示 Master 节点失联的心跳消息，以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Master 节点发生故障，相比于现有技术，能够有效缩短 Master 节点的故障检测时间。

可选地，作为一个实施例，该第一备用设备 940 还包括：

发送模块 946，用于在该第三确定模块确定该第一 Master 节点发生故障的情况下，向该第一分布式集群中所有节点发送广播消息，该广播消息用于指示该第一备用节点升级为该第一 Master 节点。

20 可选地，作为一个实施例，该检测模块具体用于，检测在第一时刻与第二时刻之间是否接收到该第一 Master 节点发送的心跳消息，该第一时刻为该第一备用节点接收到该第一参考节点发送的包括有该第一指示信息的心跳消息的时刻，该第二时刻早于该第一时刻，该第一时刻与该第二时刻之间的时间间隔为该第一 Master 节点向该第一备用节点发送心跳消息的心跳周期的 N 倍，N 为正整数。

可选地，作为一个实施例，该分布式集群系统还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该第一备用节点还包括：

30 第二接收模块，用于接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第二指示信

息, 该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中的所有节点发送包括该第二指示信息的心跳消息, 该第二分布式集群与该第一分布式集群具有相同的集群标识;

5 第四确定模块, 用于根据该第二接收模块接收的该第二指示信息, 确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一备用节点的节点属性不相符, 并确定该第二参考节点所在的第二分布式集群与该第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

10 协商模块, 用于与该第二分布式集群中的第二备用节点进行协商, 以便于整合该第一分布式集群和该第二分布式集群。

可选地, 作为一个实施例, 该协商模块包括:

第一发送单元, 用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点的节点属性为备用节点的广播消息;

15 接收单元, 用于接收该第二备用节点发送的协商消息, 该协商消息中包括用于指示该第二备用节点的权重的信息, 该协商消息是该第二备用节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该设备发送的;

20 第二发送单元, 用于在确定该第一备用节点的权重大于或等于该第二备用节点的权重的情况下, 向该第二备用节点发送用于指示该第二备用节点降级为 Slave 节点的协商响应消息;

第三发送单元, 用于在确定该第一备用节点权重小于该第二备用节点的权重的情况下, 向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一备用节点降级为 Slave 节点的广播消息。

25 因此, 在本发明实施例中, 该分布式集群包括 Master 节点、Slave 节点参考节点和该 Master 节点的备用节点, 该备用节点根据参考节点发送的心跳消息确定 Master 节点与该参考节点处于失联状态, 并通过检测该 Master 节点向该备用节点发送的心跳消息也中断来确定该 Master 节点与该备用节点也处于失联状态, 从而可以有效地确定与该备用节点和该参考节点均失联的该 Master 节点发生故障。在本发明实施例中, 备用节点通过参考节点发送的
30 用于指示 Master 节点失联的心跳消息, 以及自身检测到的该 Master 节点发送的心跳消息中断的检测结果, 可以快速、有效地确定该 Master 节点发生故

障，相比于现有技术，能够有效缩短 Master 节点的故障检测时间。

应理解，根据本发明实施例提供的分布式集群系统 900 中的第一分布式集群 901 可对应于本发明实施例提供的用于处理分布式集群中的故障的方法中的第一分布式集群，且该分布式集群系统 900 中的第一备用节点 940 也对应于本发明实施例提供的用于处理分布式集群中的故障的设备 700，并且分布式集群系统 900 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程，为了简洁，在此不再赘述。

图 12 示出了本发明实施例提供的分布式集群系统 1000，该分布式集群系统 1000 包括第一分布式集群 1001，该第一分布式集群中包括第一 Master 节点 1010 和第一 Slave 节点 1020，该分布式集群还包括第一参考节点 1030，以及作为该第一 Master 节点的备份的第一备用节点 1040，该第一 Master 节点 1010、该第一备用节点 1040 和该第一参考节点 1030 用于彼此之间互发心跳消息，该第一参考节点 1030 还用于向该第一 Slave 节点 1020 发送心跳消息，该第一 Slave 节点 1020 用于分别向该第一 Master 节点 1010 和该第一备用节点 1030 发送心跳消息，其中，

该第一备用节点 1030 还用于，在检测到该第一 Slave 节点向该第一备用节点发送的心跳消息中断的情况下，向该第一 Master 节点发送心跳消息，该心跳消息包括用于指示该第一备用节点与该第一 Slave 节点处于失联状态的第三指示信息，

20 该第一 Master 节点 1010 包括：

第一接收模块 1011，用于接收该第一备用节点发送的包括该第三指示信息的心跳消息；

第一确定模块 1012，用于根据该第一接收模块接收的该第三指示信息，确定该第一备用节点与该第一 Slave 节点处于失联状态；

25 第一检测模块 1013，用于检测该第一 Slave 节点向该第一 Master 节点发送的心跳消息是否中断；

第二确定模块 1014，用于在该第一检测模块检测到该第一 Slave 节点向该第一 Master 节点发送的心跳消息中断的情况下，确定该第一 Master 节点与该第一 Slave 节点也处于失联状态；

30 第三确定模块 1015，用于根据该第一确定模块与该第二确定模块确定的结果，确定该第一 Slave 节点发生故障。

在本发明实施例中，分布式集群中除了 Master 节点和 Slave 节点之外还包括参考节点和该 Master 节点的备用节点，该 Master 节点根据备用节点发送的心跳消息确定备用节点与该 Slave 处于失联状态，并通过检测该 Slave 节点向该 Master 节点发送的心跳消息也中断来确定该 Slave 节点与该 Master 节点也处于失联状态，从而可以有效地确定与该 Master 节点和该备用节点均失联的该 Slave 节点发生故障。

因此，在本发明实施例提供的分布式集群系统中，Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息，以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Slave 节点发生故障，相比于现有技术，能够有效缩短 Slave 节点的故障检测时间，同时也能够提高故障检测的准确性。

可选地，作为一个实施例，该第一检测模块 1013 具体用于，检测在第三时刻与第四时刻之间是否接收到该第一 Slave 节点发送的心跳消息，该第三时刻为该第一 Master 节点接收到该第一备用节点发送的包括有该第三指示信息的心跳消息的时刻，该第四时刻早于该第三时刻，该第三时刻与该第四时刻之间的时间间隔为该第一 Slave 节点向该第一 Master 节点发送心跳消息的心跳周期的 N 倍，N 为正整数。

可选地，作为一个实施例，该第一 Master 节点 1010 还包括：

第二检测模块，用于在预设检测周期内检测是否接收到该第一备用节点发送的心跳消息和该第一参考节点发送的心跳消息，该预设检测周期为发送心跳消息的心跳周期的 M 倍，M 为正整数；

第四确定模块，用于在该第二检测模块在该预设检测周期内既没有接收到该第一备用节点发送的心跳消息，也没有接收到该第一参考节点发送的心跳消息的情况下，确定该第一备用节点和该第一参考节点均发生故障。

可选地，作为一个实施例，该分布式集群系统 1000 还包括第二分布式集群，该第二分布式集群中包括第二 Master 节点，第二 Slave 节点，第二参考节点，以及作为该第二 Master 节点的备份的第二备用节点，该第一 Master 节点 1010 还包括：

第二接收模块，用于接收该第二参考节点发送的心跳消息，该心跳消息包括用于指示该心跳消息的接收端的节点属性为 Slave 节点的第四指示信息，该第二参考节点用于分别向该第二 Slave 节点以及该第一分布式集群中

的所有节点发送包括该第四指示信息的心跳消息，该第二分布式集群与该第一分布式集群具有相同的集群标识；

5 第五确定模块，用于根据该第二接收模块接收的该第四指示信息，确定该第二参考节点发送的心跳消息中所指示的该心跳消息的接收端的节点属性与该第一 Master 节点的节点属性不相符，并确定该第二参考节点所在的第二分布式集群与该第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

协商模块，用于与该第二分布式集群中的第二 Master 节点进行协商，以便于整合该第一分布式集群和该第二分布式集群。

10 可选地，作为一个实施例，该协商模块包括：

第一发送单元，用于向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点的节点属性为 Master 节点的广播消息；

15 接收单元，用于接收该第二 Master 节点发送的协商消息，该协商消息中包括用于指示该第二 Master 节点的权重的信息，该协商消息是该第二 Master 节点在检测到该第一发送单元发送的该广播消息指示的节点属性与自身的节点属性相同的情况下向该第一 Master 节点发送的；

第二发送单元，用于在确定该第一 Master 节点的权重大于或等于该第二 Master 节点的权重时，向该第二 Master 节点发送用于指示该第二 Master 节点降级为 Slave 节点的协商响应消息；

20 第三发送单元，用于在确定该第一 Master 节点的权重小于该第二 Master 节点的权重时，向该第一分布式集群和该第二分布式集群内所有的节点发送用于指示该第一 Master 节点降级为 Slave 节点的广播消息。

25 因此，在本发明实施例提供的分布式集群系统中，Master 节点通过备用节点发送的用于指示 Slave 节点失联的心跳消息，以及自身检测到的该 Slave 节点发送的心跳消息中断的检测结果，可以快速、有效地确定该 Slave 节点发生故障，相比于现有技术，能够有效缩短 Slave 节点的故障检测时间，同时也能够提高故障检测的准确性。

30 应理解，根据本发明实施例提供的分布式集群系统 1000 中的第一分布式集群 1001 可对应于本发明实施例提供的用于处理分布式集群中的故障的方法中的第一分布式集群，且该分布式集群系统 1000 中的第一 Master 节点 1010 也对应于本发明实施例提供的用于处理分布式集群中的故障的设备

800, 并且分布式集群系统 1000 中的各个模块的上述和其它操作和/或功能分别为了实现图 1 至图 5 中的各个方法的相应流程, 为了简洁, 在此不再赘述。

应理解, 在本发明的各种实施例中, 上述各过程的序号的大小并不意味着执行顺序的先后, 各过程的执行顺序应以其功能和内在逻辑确定, 而不应

5 对本发明实施例的实施过程构成任何限定。

本领域普通技术人员可以意识到, 结合本文中所公开的实施例描述的各示例的单元及算法步骤, 能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行, 取决于技术方案的特
10 定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能, 但是这种实现不应认为超出本发明的范围。

所属领域的技术人员可以清楚地了解到, 为描述的方便和简洁, 上述描述的系统、装置和单元的具体工作过程, 可以参考前述方法实施例中的对应过程, 在此不再赘述。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的系统、装置和方法, 可以通过其它的方式实现。例如, 以上所描述的装置实施例仅仅是示
15 意性的, 例如, 所述单元的划分, 仅仅为一种逻辑功能划分, 实际实现时可以有另外的划分方式, 例如多个单元或组件可以结合或者可以集成到另一个系统, 或一些特征可以忽略, 或不执行。另一点, 所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口, 装置或单元的间接耦合
20 或通信连接, 可以是电性, 机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的, 作为单元显示的部件可以是或者也可以不是物理单元, 即可以位于一个地方, 或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

25 另外, 在本发明各个实施例中的各功能单元可以集成在一个处理单元中, 也可以是各个单元单独物理存在, 也可以两个或两个以上单元集成在一个单元中。

所述功能如果以软件功能单元的形式实现并作为独立的产品销售或使用
30 时, 可以存储在一个计算机可读取存储介质中。基于这样的理解, 本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来, 该计算机软件产品存储在一个存储介质

中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM, Read-Only Memory）、随机存取存储器（RAM, Random Access Memory）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，仅为本发明的具体实施方式，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所述权利要求的保护范围为准。

权利要求

1、一种处理至少一个分布式集群中的故障的方法，所述至少一个分布式集群包括第一分布式集群，所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点，其特征在于，所述分布式集群还包括第一参考节点，以及
5 作为所述第一 Master 节点的备份的第一备用节点，所述方法包括：

所述第一备用节点接收所述第一参考节点发送的心跳消息，所述心跳消息包括用于指示所述第一参考节点与所述第一 Master 节点处于失联状态的第一指示信息；

所述第一备用节点根据所述第一指示信息，确定所述第一参考节点与所述第一 Master 节点处于失联状态；
10

所述第一备用节点检测所述第一 Master 节点向所述第一备用节点发送的心跳消息是否中断；

在检测到所述第一 Master 节点向所述第一备用节点发送的心跳消息中断的情况下，所述第一备用节点确定所述第一备用节点与所述第一 Master 节点也处于失联状态；
15

所述第一备用节点确定与所述第一备用节点和所述第一参考节点均失联的所述第一 Master 节点发生故障。

2、根据权利要求 1 所述的方法，其特征在于，在确定所述第一 Master 节点发生故障的情况下，所述方法还包括：

所述第一备用节点向所述第一分布式集群中所有节点发送广播消息，所述广播消息用于指示所述第一备用节点升级为所述第一 Master 节点。
20

3、根据权利要求 1 或 2 所述的方法，其特征在于，所述第一备用节点检测所述第一 Master 节点向所述第一备用节点发送的心跳消息是否中断，包括：

所述第一备用节点检测在第一时刻与第二时刻之间是否接收到所述第一 Master 节点发送的心跳消息，所述第一时刻为所述第一备用节点接收到所述第一参考节点发送的包括所述第一指示信息的心跳消息的时刻，所述第二时刻早于所述第一时刻，所述第一时刻与所述第二时刻之间的时间间隔为所述第一 Master 节点向所述第一备用节点发送心跳消息的心跳周期的 N 倍，
25
30 N 为正整数。

4、根据权利要求 1 至 3 中任一项所述的方法，其特征在于，所述至少一个分布式集群还包括第二分布式集群，所述第二分布式集群中包括第二

Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为所述第二 Master 节点的备份的第二备用节点, 所述方法还包括:

所述第一备用节点接收所述第二参考节点发送的心跳消息, 所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第二指示信息, 所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分布式集群中的所有节点发送包括所述第二指示信息的心跳消息, 所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

所述第一备用节点根据所述第二指示信息, 确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述第一备用节点的节点属性不相符, 并确定所述第二参考节点所在的第二分布式集群与所述第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

所述第一备用节点与所述第二分布式集群中的第二备用节点进行协商, 以便于整合所述第一分布式集群和所述第二分布式集群。

5、根据权利要求 4 所述的方法, 其特征在于, 所述第一备用节点与所述第二分布式集群中的第二备用节点进行协商, 包括:

所述第一备用节点向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一备用节点的节点属性为备用节点的广播消息;

所述第一备用节点接收所述第二备用节点发送的协商消息, 所述协商消息中包括用于指示所述第二备用节点的权重的信息, 所述协商消息是所述第二备用节点在检测到所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述第一备用节点发送的;

在确定所述第一备用节点的权重大于或等于所述第二备用节点的权重的情况下, 所述第一备用节点向所述第二备用节点发送用于指示所述第二备用节点降级的 Slave 节点的协商响应消息;

在确定所述第一备用节点的权重小于所述第二备用节点的权重的情况下, 所述第一备用节点向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一备用节点降级的 Slave 节点的广播消息。

6、一种处理至少一个分布式集群中的故障的方法, 所述至少一个分布式集群包括第一分布式集群, 所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点, 其特征在于, 所述第一分布式集群还包括第一参考节点以

及作为所述第一 Master 节点的备份的第一备用节点, 所述方法包括:

所述第一 Master 节点接收所述第一备用节点发送的心跳消息, 所述心跳消息包括用于指示所述第一备用节点与所述第一 Slave 节点处于失联状态的第三指示信息;

- 5 所述第一 Master 节点根据所述第三指示信息, 确定所述第一备用节点与所述第一 Slave 节点处于失联状态;

所述第一 Master 节点检测所述第一 Slave 节点向所述第一 Master 节点发送的心跳消息是否中断;

- 10 在检测到所述第一 Slave 节点向所述第一 Master 节点发送的心跳消息中断的情况下, 所述第一 Master 节点确定所述第一 Master 节点与所述第一 Slave 节点也处于失联状态;

所述第一 Master 节点确定与所述第一 Master 节点和所述第一备用节点均失联的所述第一 Slave 节点发生故障。

- 15 7、根据权利要求 6 所述的方法, 其特征在于, 所述第一 Master 节点检测所述第一 Slave 节点向所述第一 Master 节点发送的心跳消息是否中断, 包括:

- 20 所述第一 Master 节点检测在第三时刻与第四时刻之间是否接收到所述第一 Slave 节点发送的心跳消息, 所述第三时刻为所述第一 Master 节点接收到所述第一备用节点发送的包括有所述第三指示信息的心跳消息的时刻, 所述第四时刻早于所述第三时刻, 所述第三时刻与所述第四时刻之间的时间间隔为所述第一 Slave 节点向所述第一 Master 节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

8、根据权利要求 6 或 7 所述的方法, 其特征在于, 所述方法还包括:

- 25 所述第一 Master 节点在预设检测周期内检测是否接收到所述第一备用节点发送的心跳消息和所述第一参考节点发送的心跳消息, 所述预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

在所述预设检测周期内既没有接收到所述第一备用节点发送的心跳消息, 也没有接收到所述第一参考节点发送的心跳消息的情况下, 所述第一 Master 节点确定所述第一备用节点和所述第一参考节点均发生故障。

- 30 9、根据权利要求 6 至 8 中任一项所述的方法, 其特征在于, 所述至少一个分布式集群还包括第二分布式集群, 所述第二分布式集群中包括第二 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为所述第二 Master 节

点的备份的第二备用节点,所述方法还包括:

所述第一 Master 节点接收所述第二参考节点发送的心跳消息,所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第四指示信息,所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分布式集群中的所有节点发送包括所述第四指示信息的心跳消息,所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

所述第一 Master 节点根据所述第四指示信息,确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述第一 Master 节点的节点属性不相符,并确定所述第二参考节点所在的第二分布式集群与所述第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

所述第一 Master 节点与所述第二分布式集群中的第二 Master 节点进行协商,以便于整合所述第一分布式集群和所述第二分布式集群。

10、根据权利要求 9 所述的方法,其特征在于,所述第一 Master 节点与所述第二分布式集群中的第二 Master 节点进行协商,包括:

所述第一 Master 节点向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一 Master 节点的节点属性为 Master 节点的广播消息;

所述第一 Master 节点接收所述第二 Master 节点发送的协商消息,所述协商消息中包括用于指示所述第二 Master 节点的权重的信息,所述协商消息是所述第二 Master 节点在检测到所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述第一 Master 节点发送的;

在确定所述第一 Master 节点的权重大于或等于所述第二 Master 节点的权重时,所述第一 Master 节点向所述第二 Master 节点发送用于指示所述第二 Master 节点降级为 Slave 节点的协商响应消息;

在确定所述第一 Master 节点的权重小于所述第二 Master 节点的权重时,所述第一 Master 节点向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一 Master 节点降级为 Slave 节点的广播消息。

11、一种用于处理至少一个分布式集群中的故障的设备,所述至少一个分布式集群包括第一分布式集群,所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点,其特征在于,所述分布式集群还包括第一参考节点,以及作为所述第一 Master 节点的备份的第一备用节点,其中,所述设备为所

述第一备用节点, 所述设备包括:

第一接收模块, 用于接收所述第一参考节点发送的包括有所述第一指示信息的心跳消息;

第一确定模块, 用于根据所述第一接收模块接收的所述第一指示信息,

5 确定所述第一参考节点与所述第一 Master 节点处于失联状态;

检测模块, 用于检测所述第一 Master 节点向所述设备发送的心跳消息是否中断;

第二确定模块, 用于在所述检测模块检测到所述第一 Master 节点向所述设备发送的心跳消息中断的情况下, 确定所述设备与所述第一 Master 节点也

10 处于失联状态;

第三确定模块, 用于根据所述第一确定模块和所述第二确定模块确定的结果, 确定所述第一 Master 节点发生故障。

12、根据权利要求 11 所述的设备, 其特征在于, 所述设备还包括:

发送模块, 用于在所述第三确定模块确定所述第一 Master 节点发生故障
15 的情况下, 向所述第一分布式集群中所有节点发送广播消息, 所述广播消息用于指示所述设备升级为所述第一 Master 节点。

13、根据权利要求 11 或 12 所述的设备, 其特征在于, 所述检测模块具体用于, 检测在第一时刻与第二时刻之间是否接收到所述第一 Master 节点发送的心跳消息, 所述第一时刻为所述设备接收到所述第一参考节点发送的包
20 括有所述第一指示信息的心跳消息的时刻, 所述第二时刻早于所述第一时刻, 所述第一时刻与所述第二时刻之间的时间间隔为所述第一 Master 节点向所述设备发送心跳消息的心跳周期的 N 倍, N 为正整数。

14、根据权利要求 11 至 13 中任一项所述的设备, 其特征在于, 所述至少一个分布式集群还包括第二分布式集群, 所述第二分布式集群中包括第二
25 Master 节点, 第二 Slave 节点, 第二参考节点, 以及作为所述第二 Master 节点的备份的第二备用节点, 所述设备还包括:

第二接收模块, 用于接收所述第二参考节点发送的心跳消息, 所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第二指示信息, 所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分
30 布式集群中的所有节点发送包括所述第二指示信息的心跳消息, 所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

第四确定模块, 用于根据所述第二接收模块接收的所述第二指示信息,

确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述设备的节点属性不相符，并确定所述第二参考节点所在的第二分布式集群与所述设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

- 5 协商模块，用于与所述第二分布式集群中的第二备用节点进行协商，以便于整合所述第一分布式集群和所述第二分布式集群。

15、根据权利要求 14 所述的设备，其特征在于，所述协商模块包括：

第一发送单元，用于向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述设备的节点属性为备用节点的广播消息；

- 10 接收单元，用于接收所述第二备用节点发送的协商消息，所述协商消息中包括用于指示所述第二备用节点的权重的信息，所述协商消息是所述第二备用节点在检测到所述第一发送单元发送的所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述设备发送的；

- 15 第二发送单元，用于在确定所述设备的权重大于或等于所述第二备用节点的权重的情况下，向所述第二备用节点发送用于指示所述第二备用节点降级为 Slave 节点的协商响应消息；

第三发送单元，用于在确定所述设备权重小于所述第二备用节点的权重的情况下，向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述设备降级为 Slave 节点的广播消息。

- 20 16、一种用于处理至少一个分布式集群中的故障的设备，所述至少一个分布式集群包括第一分布式集群，所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点，其特征在于，所述第一分布式集群还包括第一参考节点以及作为所述第一 Master 节点的备份的第一备用节点，所述设备为所述第一 Master 节点，所述设备包括：

- 25 第一接收模块，用于接收所述第一备用节点发送的包括所述第三指示信息的心跳消息；

第一确定模块，用于根据所述第一接收模块接收的所述第三指示信息，确定所述第一备用节点与所述第一 Slave 节点处于失联状态；

- 30 第一检测模块，用于检测所述第一 Slave 节点向所述设备发送的心跳消息是否中断；

第二确定模块，用于在所述第一检测模块检测到所述第一 Slave 节点向所述设备发送的心跳消息中断的情况下，确定所述设备与所述第一 Slave 节

点也处于失联状态;

第三确定模块,用于根据所述第一确定模块与所述第二确定模块确定的结果,确定所述第一 Slave 节点发生故障。

17、根据权利要求 16 所述的设备,其特征在于,所述第一检测模块具体用于,检测在第三时刻与第四时刻之间是否接收到所述第一 Slave 节点发送的心跳消息,所述第三时刻为所述设备接收到所述第一备用节点发送的包括有所述第三指示信息的心跳消息的时刻,所述第四时刻早于所述第三时刻,所述第三时刻与所述第四时刻之间的时间间隔为所述第一 Slave 节点向所述设备发送心跳消息的心跳周期的 N 倍, N 为正整数。

18、根据权利要求 16 或 17 所述的设备,其特征在于,所述设备还包括:第二检测模块,用于在预设检测周期内检测是否接收到所述第一备用节点发送的心跳消息和所述第一参考节点发送的心跳消息,所述预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

第四确定模块,用于在所述第二检测模块在所述预设检测周期内既没有接收到所述第一备用节点发送的心跳消息,也没有接收到所述第一参考节点发送的心跳消息的情况下,确定所述第一备用节点和所述第一参考节点均发生故障。

19、根据权利要求 16 至 18 中任一项所述的设备,其特征在于,所述至少一个分布式集群还包括第二分布式集群,所述第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为所述第二 Master 节点的备份的第二备用节点,所述设备还包括:

第二接收模块,用于接收所述第二参考节点发送的心跳消息,所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第四指示信息,所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分布式集群中的所有节点发送包括所述第四指示信息的心跳消息,所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

第五确定模块,用于根据所述第二接收模块接收的所述第四指示信息,确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述设备的节点属性不相符,并确定所述第二参考节点所在的第二分布式集群与所述设备所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

协商模块,用于与所述第二分布式集群中的第二 Master 节点进行协商,

以便于整合所述第一分布式集群和所述第二分布式集群。

20、根据权利要求 19 所述的设备，其特征在于，所述协商模块包括：

第一发送单元，用于向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述设备的节点属性为 Master 节点的广播消息；

5 接收单元，用于接收所述第二 Master 节点发送的协商消息，所述协商消息中包括用于指示所述第二 Master 节点的权重的信息，所述协商消息是所述第二 Master 节点在检测到所述第一发送单元发送的所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述设备发送的；

10 第二发送单元，用于在确定所述设备的权重大于或等于所述第二 Master 节点的权重时，向所述第二 Master 节点发送用于指示所述第二 Master 节点降级为 Slave 节点的协商响应消息；

第三发送单元，用于在确定所述设备的权重小于所述第二 Master 节点的权重时，向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述设备降级为 Slave 节点的广播消息。

15 21、一种分布式集群系统，其特征在于，所述分布式集群系统包括第一分布式集群，所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点，其特征在于，所述分布式集群还包括第一参考节点，以及作为所述第一 Master 节点的备份的第一备用节点，所述第一 Master 节点、所述第一备用节点和所述第一参考节点用于彼此之间互发心跳消息，所述第一参考节点还
20 用于向所述第一 Slave 节点发送心跳消息，所述第一 Slave 节点用于分别向所述第一 Master 节点和所述第一备用节点发送心跳消息，其中，

所述第一参考节点还用于，在检测到所述第一 Master 节点向所述第一参考节点发送的心跳消息中断的情况下，向所述第一备用节点发送心跳消息，所述心跳消息包括用于指示所述第一参考节点与所述第一 Master 节点处于
25 失联状态的第一指示信息，

所述第一备用节点包括：

第一接收模块，用于接收所述第一参考节点发送的包括所述第一指示信息的心跳消息；

30 第一确定模块，用于根据所述第一接收模块接收的所述第一指示信息，确定所述第一参考节点与所述第一 Master 节点处于失联状态；

检测模块，用于检测所述第一 Master 节点向所述第一备用节点发送的心跳消息是否中断；

第二确定模块,用于在所述检测模块检测到所述第一 Master 节点向所述第一备用节点发送的心跳消息中断的情况下,确定所述第一备用节点与所述第一 Master 节点也处于失联状态;

第三确定模块,用于根据所述第一确定模块和所述第二确定模块确定的结果,确定所述第一 Master 节点发生故障。

22、根据权利要求 21 所述的分布式集群系统,其特征在于,所述第一备用设备还包括:

发送模块,用于在所述第三确定模块确定所述第一 Master 节点发生故障的情况下,向所述第一分布式集群中所有节点发送广播消息,所述广播消息用于指示所述第一备用节点升级为所述第一 Master 节点。

23、根据权利要求 21 或 22 所述的分布式集群系统,其特征在于,所述检测模块具体用于,检测在第一时刻与第二时刻之间是否接收到所述第一 Master 节点发送的心跳消息,所述第一时刻为所述第一备用节点接收到所述第一参考节点发送的包括有所述第一指示信息的心跳消息的时刻,所述第二时刻早于所述第一时刻,所述第一时刻与所述第二时刻之间的时间间隔为所述第一 Master 节点向所述第一备用节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

24、根据权利要求 21 至 23 中任一项所述的分布式集群系统,其特征在于,所述分布式集群系统还包括第二分布式集群,所述第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为所述第二 Master 节点的备份的第二备用节点,所述第一备用节点还包括:

第二接收模块,用于接收所述第二参考节点发送的心跳消息,所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第二指示信息,所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分布式集群中的所有节点发送包括所述第二指示信息的心跳消息,所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

第四确定模块,用于根据所述第二接收模块接收的所述第二指示信息,确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述第一备用节点的节点属性不相符,并确定所述第二参考节点所在的第二分布式集群与所述第一备用节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群;

协商模块,用于与所述第二分布式集群中的第二备用节点进行协商,以

便于整合所述第一分布式集群和所述第二分布式集群。

25、根据权利要求 24 所述的分布式集群系统，其特征在于，所述协商模块包括：

5 第一发送单元，用于向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一备用节点的节点属性为备用节点的广播消息；

10 接收单元，用于接收所述第二备用节点发送的协商消息，所述协商消息中包括用于指示所述第二备用节点的权重的信息，所述协商消息是所述第二备用节点在检测到所述第一发送单元发送的所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述设备发送的；

第二发送单元，用于在确定所述第一备用节点的权重大于或等于所述第二备用节点的权重的情况下，向所述第二备用节点发送用于指示所述第二备用节点降级为 Slave 节点的协商响应消息；

15 第三发送单元，用于在确定所述第一备用节点权重小于所述第二备用节点的权重的情况下，向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一备用节点降级为 Slave 节点的广播消息。

20 26、一种分布式集群系统，其特征在于，所述分布式集群系统包括第一分布式集群，所述第一分布式集群中包括第一 Master 节点和第一 Slave 节点，其特征在于，所述分布式集群还包括第一参考节点，以及作为所述第一 Master 节点的备份的第一备用节点，所述第一 Master 节点、所述第一备用节点和所述第一参考节点用于彼此之间互发心跳消息，所述第一参考节点还用于向所述第一 Slave 节点发送心跳消息，所述第一 Slave 节点用于分别向所述第一 Master 节点和所述第一备用节点发送心跳消息，其中，

25 所述第一备用节点还用于，在检测到所述第一 Slave 节点向所述第一备用节点发送的心跳消息中断的情况下，向所述第一 Master 节点发送心跳消息，所述心跳消息包括用于指示所述第一备用节点与所述第一 Slave 节点处于失联状态的第三指示信息，

所述第一 Master 节点包括：

30 第一接收模块，用于接收所述第一备用节点发送的包括所述第三指示信息的心跳消息；

第一确定模块，用于根据所述第一接收模块接收的所述第三指示信息，确定所述第一备用节点与所述第一 Slave 节点处于失联状态；

第一检测模块,用于检测所述第一 Slave 节点向所述第一 Master 节点发送的心跳消息是否中断;

第二确定模块,用于在所述第一检测模块检测到所述第一 Slave 节点向所述第一 Master 节点发送的心跳消息中断的情况下,确定所述第一 Master 节点与所述第一 Slave 节点也处于失联状态;

第三确定模块,用于根据所述第一确定模块与所述第二确定模块确定的结果,确定所述第一 Slave 节点发生故障。

27、根据权利要求 26 所述的分布式集群系统,其特征在于,所述第一检测模块具体用于,检测在第三时刻与第四时刻之间是否接收到所述第一 Slave 节点发送的心跳消息,所述第三时刻为所述第一 Master 节点接收到所述第一备用节点发送的包括有所述第三指示信息的心跳消息的时刻,所述第四时刻早于所述第三时刻,所述第三时刻与所述第四时刻之间的时间间隔为所述第一 Slave 节点向所述第一 Master 节点发送心跳消息的心跳周期的 N 倍, N 为正整数。

28、根据权利要求 26 或 27 所述的分布式集群系统,其特征在于,所述第一 Master 节点还包括:

第二检测模块,用于在预设检测周期内检测是否接收到所述第一备用节点发送的心跳消息和所述第一参考节点发送的心跳消息,所述预设检测周期为发送心跳消息的心跳周期的 M 倍, M 为正整数;

第四确定模块,用于在所述第二检测模块在所述预设检测周期内既没有接收到所述第一备用节点发送的心跳消息,也没有接收到所述第一参考节点发送的心跳消息的情况下,确定所述第一备用节点和所述第一参考节点均发生故障。

29、根据权利要求 26 至 28 中任一项所述的分布式集群系统,其特征在于,所述分布式集群系统还包括第二分布式集群,所述第二分布式集群中包括第二 Master 节点,第二 Slave 节点,第二参考节点,以及作为所述第二 Master 节点的备份的第二备用节点,所述第一 Master 节点还包括:

第二接收模块,用于接收所述第二参考节点发送的心跳消息,所述心跳消息包括用于指示所述心跳消息的接收端的节点属性为 Slave 节点的第四指示信息,所述第二参考节点用于分别向所述第二 Slave 节点以及所述第一分布式集群中的所有节点发送包括所述第四指示信息的心跳消息,所述第二分布式集群与所述第一分布式集群具有相同的集群标识;

第五确定模块，用于根据所述第二接收模块接收的所述第四指示信息，确定所述第二参考节点发送的心跳消息中所指示的所述心跳消息的接收端的节点属性与所述第一 Master 节点的节点属性不相符，并确定所述第二参考节点所在的第二分布式集群与所述第一 Master 节点所在的第一分布式集群为一个网络分布式集群分裂后形成的两个子集群；

协商模块，用于与所述第二分布式集群中的第二 Master 节点进行协商，以便于整合所述第一分布式集群和所述第二分布式集群。

30、根据权利要求 29 所述的分布式集群系统，其特征在于，所述协商模块包括：

10 第一发送单元，用于向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一 Master 节点的节点属性为 Master 节点的广播消息；

接收单元，用于接收所述第二 Master 节点发送的协商消息，所述协商消息中包括用于指示所述第二 Master 节点的权重的信息，所述协商消息是所述
15 第二 Master 节点在检测到所述第一发送单元发送的所述广播消息指示的节点属性与自身的节点属性相同的情况下向所述第一 Master 节点发送的；

第二发送单元，用于在确定所述第一 Master 节点的权重大于或等于所述第二 Master 节点的权重时，向所述第二 Master 节点发送用于指示所述第二 Master 节点降级为 Slave 节点的协商响应消息；

20 第三发送单元，用于在确定所述第一 Master 节点的权重小于所述第二 Master 节点的权重时，向所述第一分布式集群和所述第二分布式集群内所有的节点发送用于指示所述第一 Master 节点降级为 Slave 节点的广播消息。

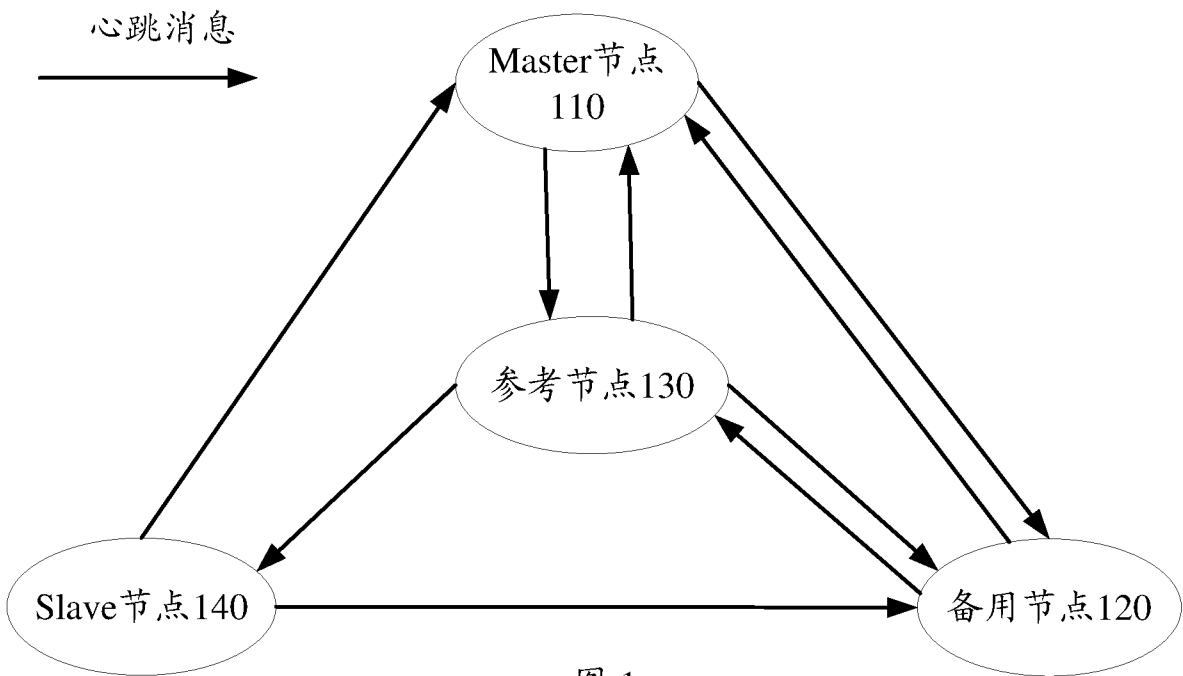


图 1

200

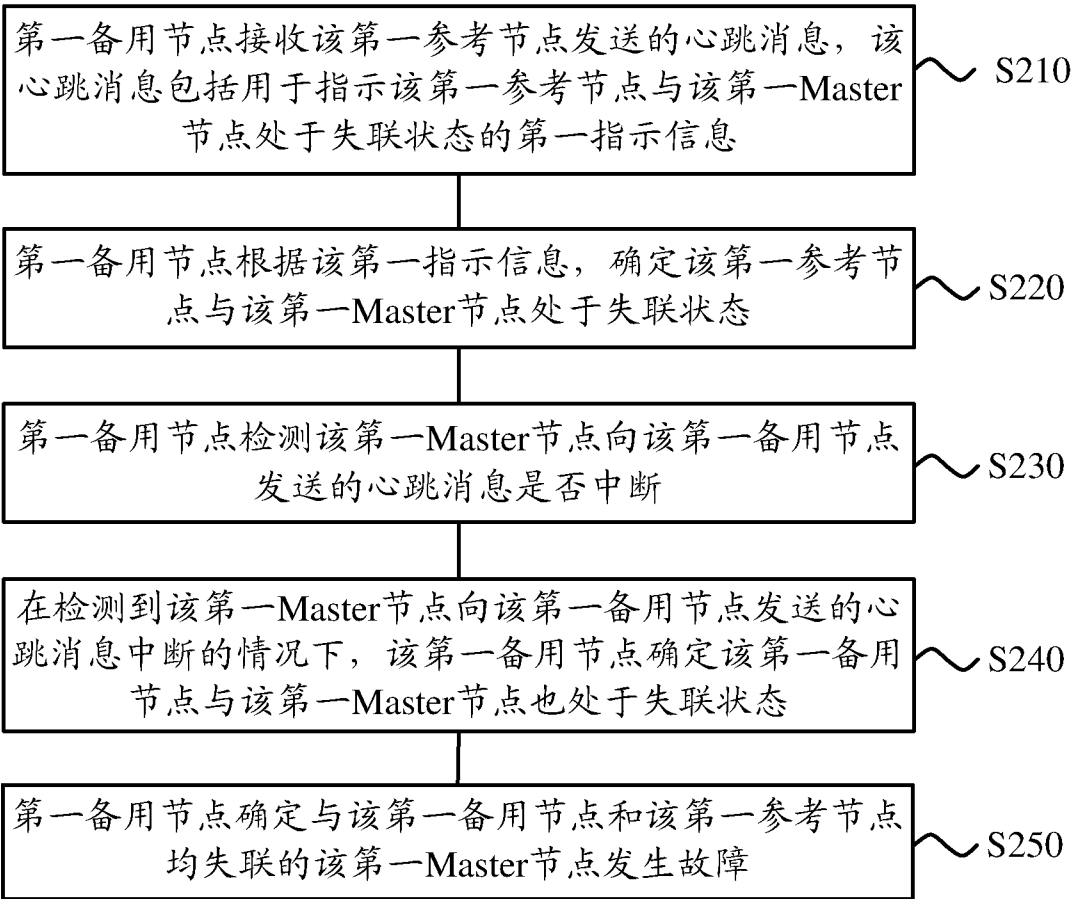


图 2

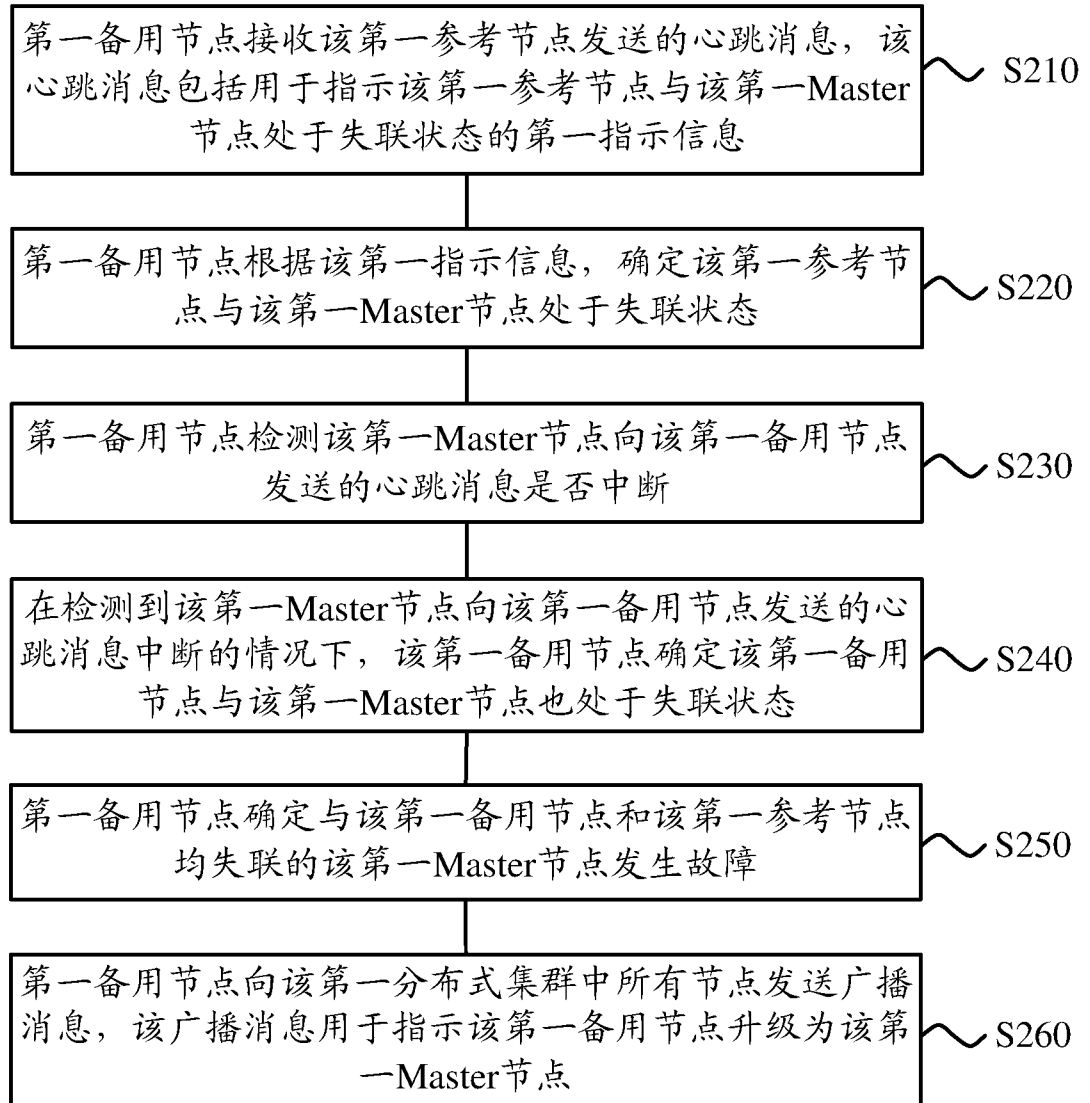
200

图 3

300

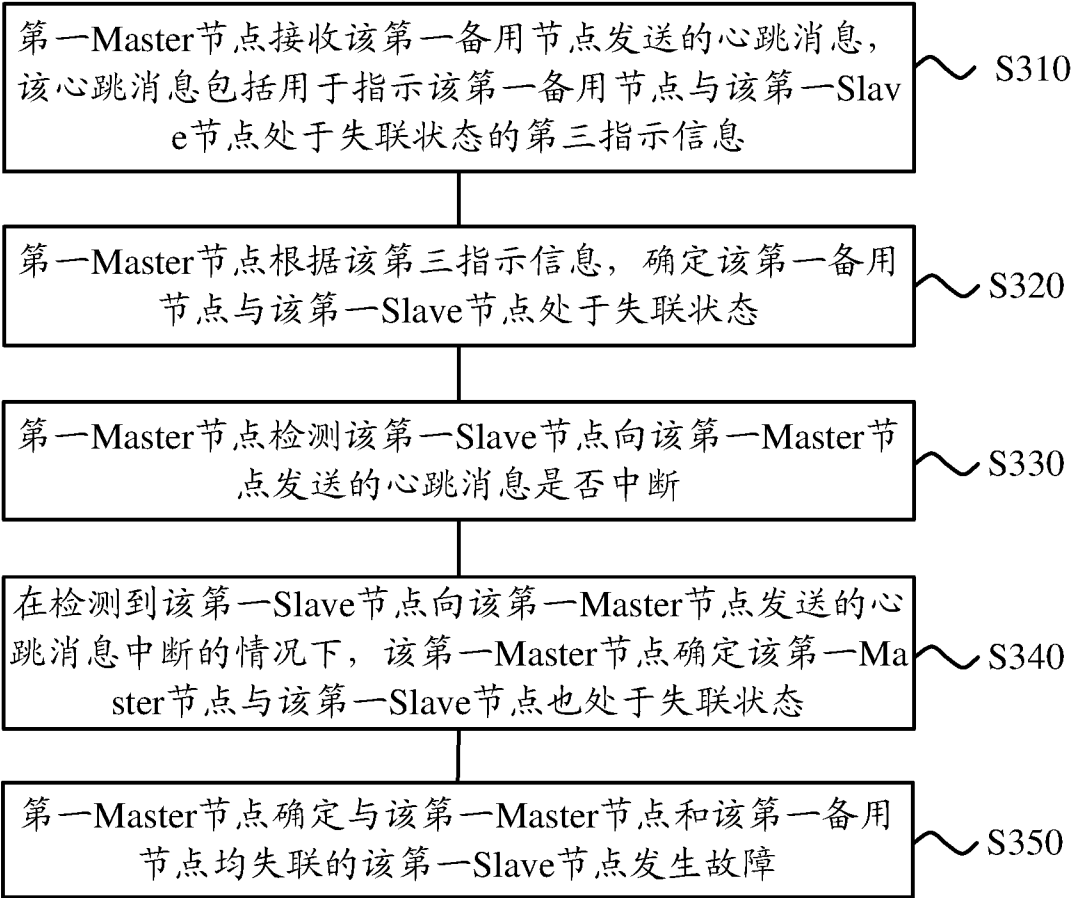


图 4

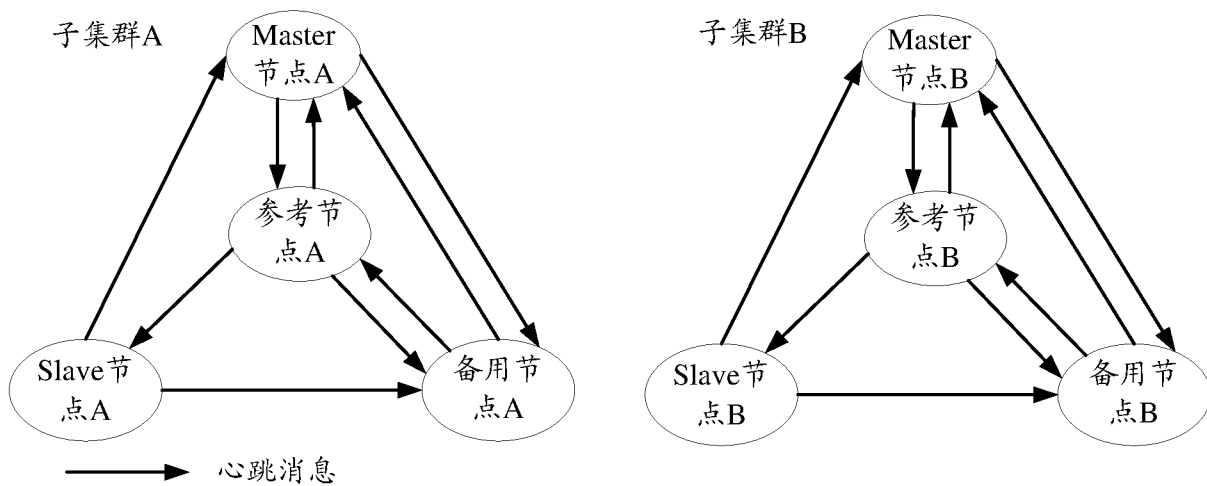


图 5(a)

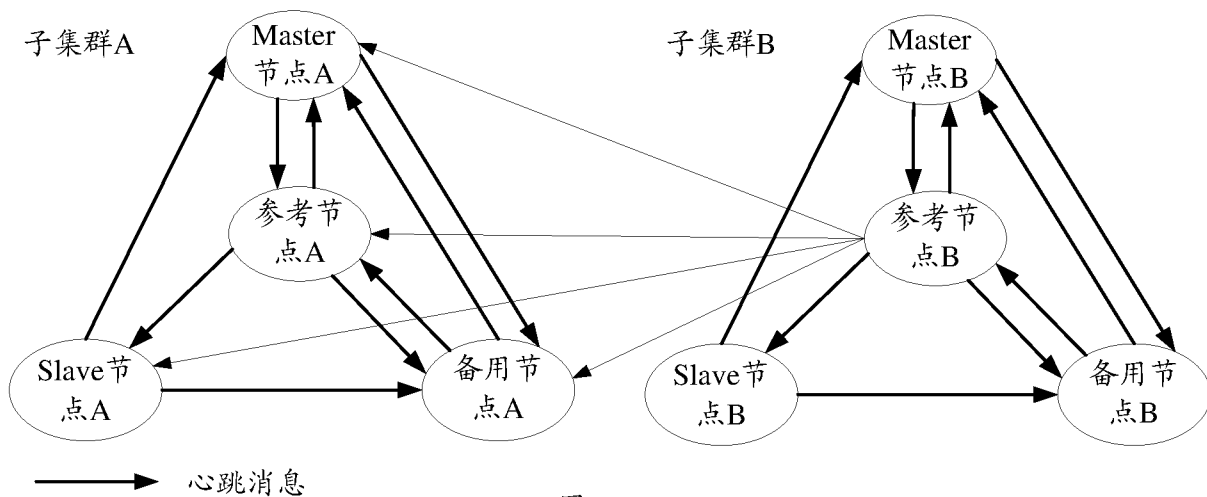


图 5(b)

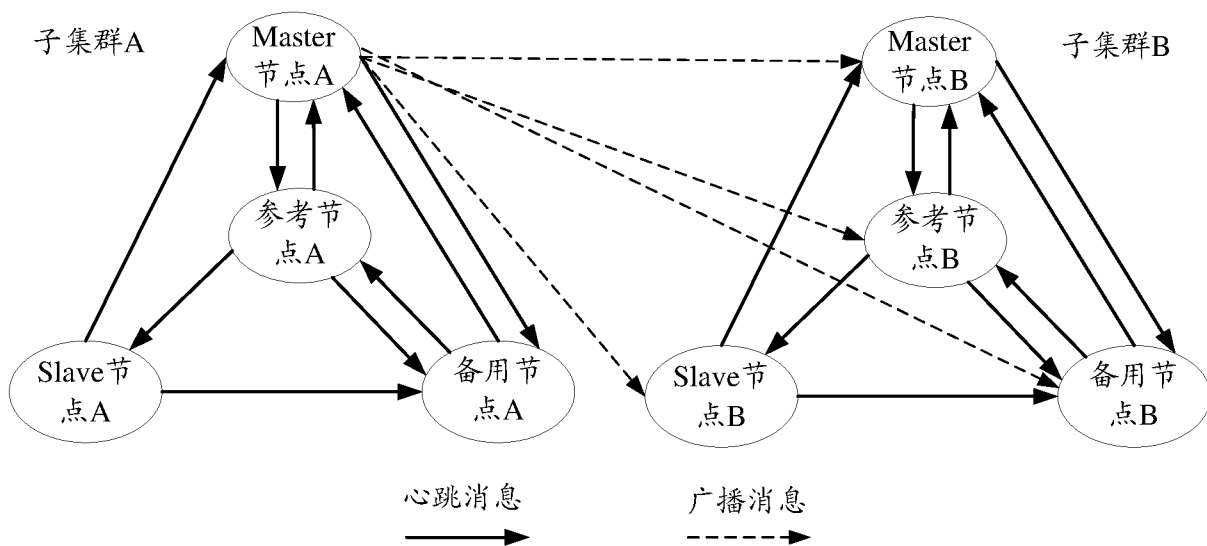


图 5(c)

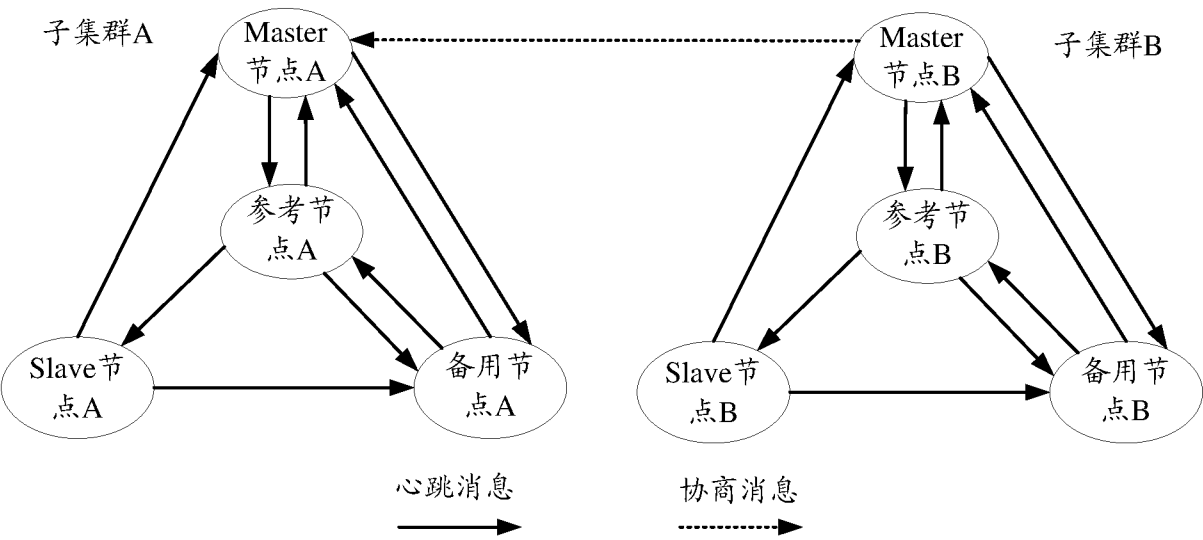


图 5(d)

400



图 6

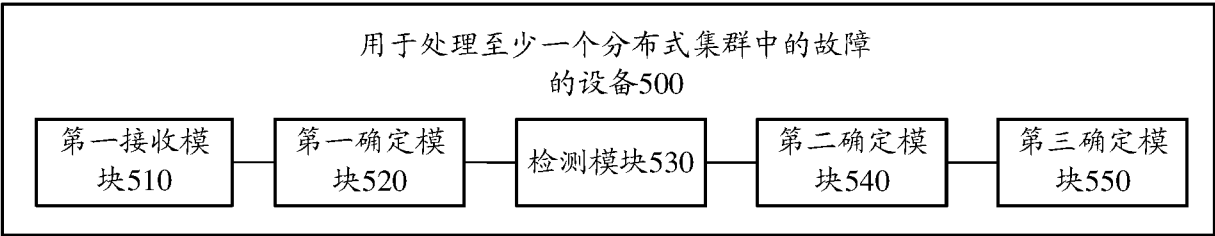


图 7

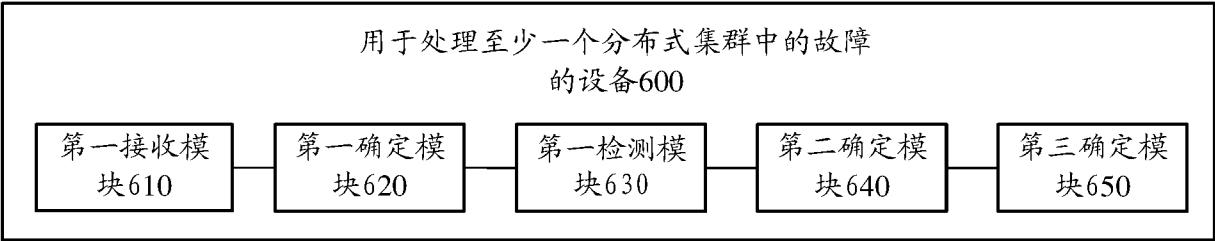


图 8

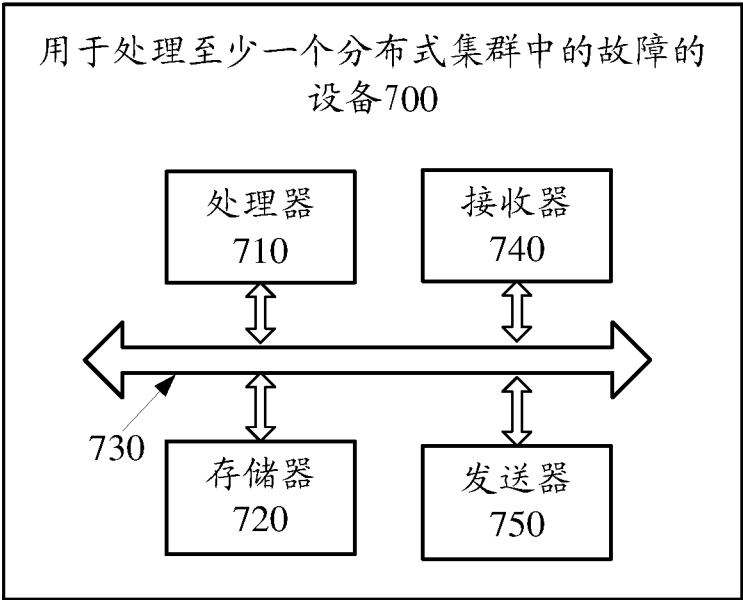


图 9

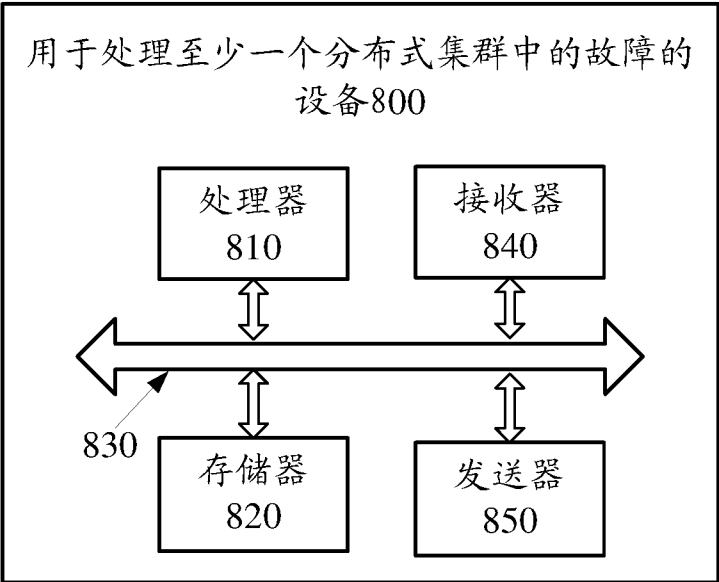


图 10

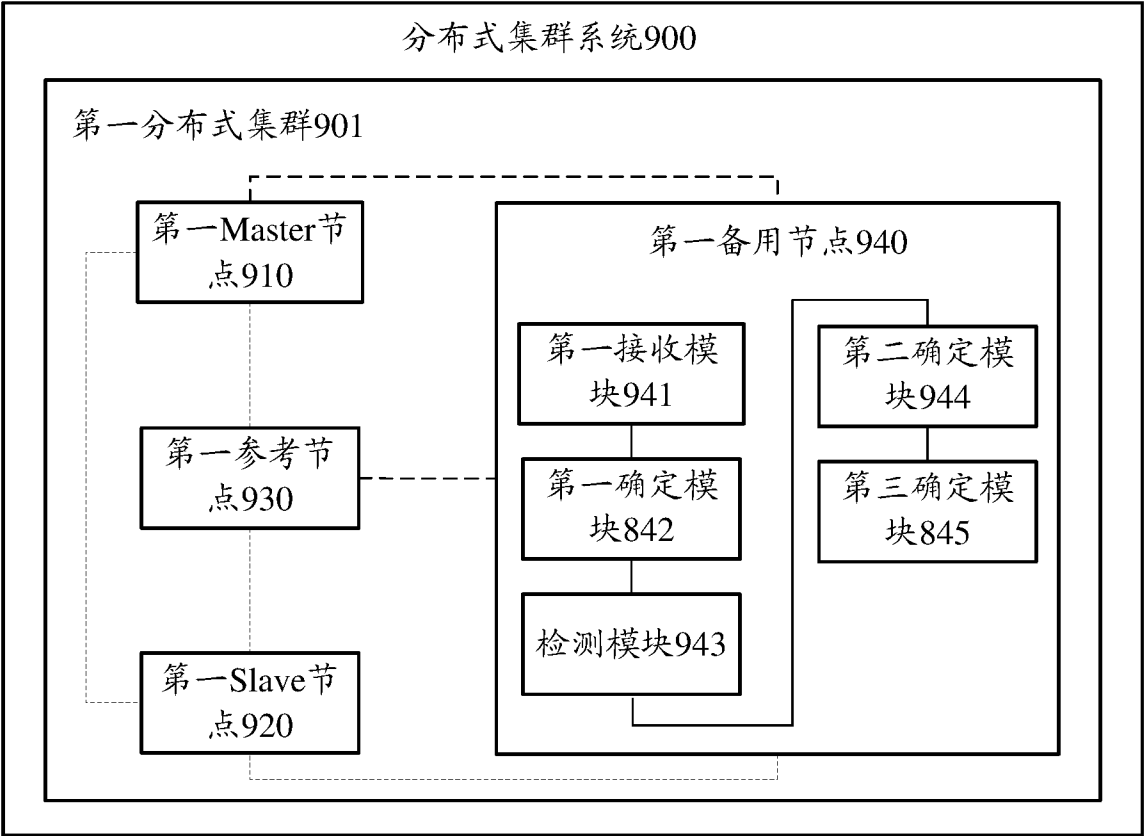


图 11

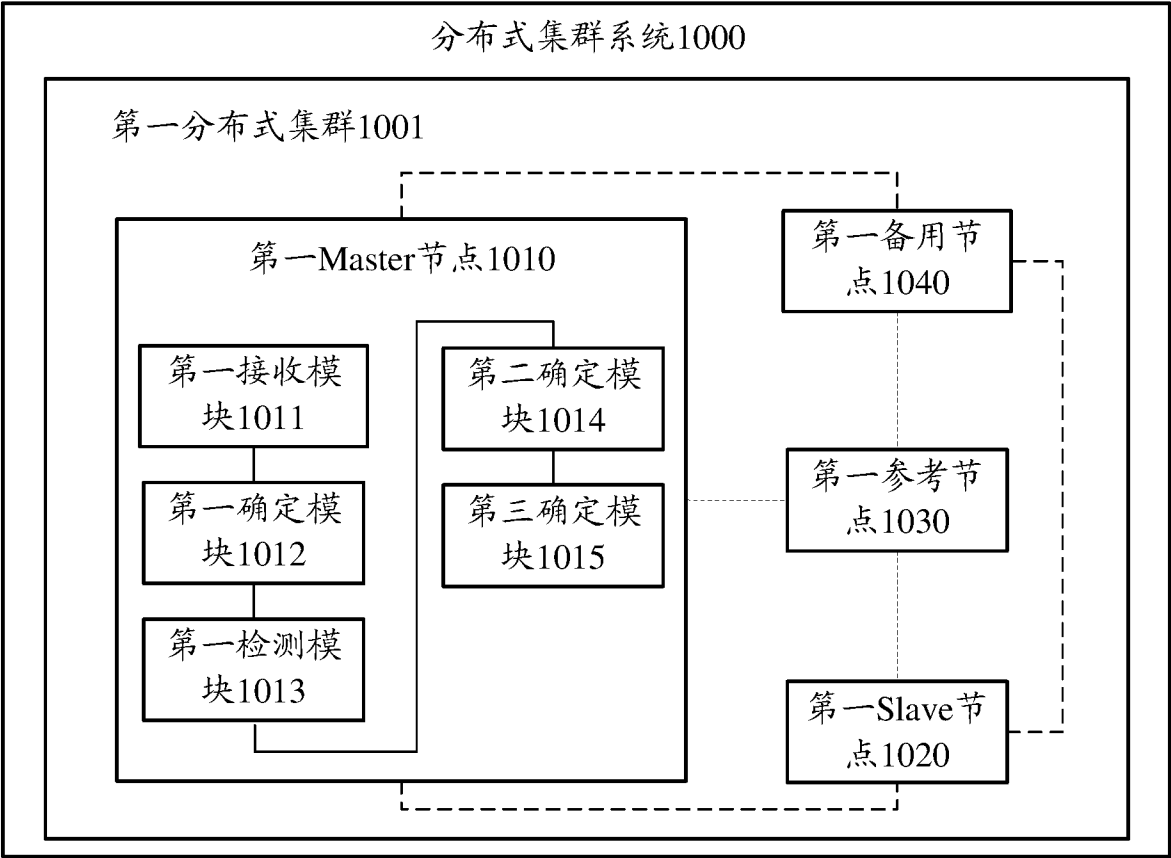


图 12

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/084573

A. CLASSIFICATION OF SUBJECT MATTER

H04L 29/08 (2006.01) i; H04L 12/26 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L, H04B, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CPRSABS, CNTXT, CNKI, VEN: distributed, cluster, failure, node, NB, master, slave, standby, backup, reference, heartbeat message, state, disconnect, interrupt

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 101494565 A (CHINA MOBILE COMMUNICATIONS CORPORATION), 29 July 2009 (29.07.2009), the whole document	1-30
A	CN 102752093 A (CHINA UNITED NETWORK COMMUNICATIONS CORPORATION LIMITED), 24 October 2012 (24.10.2012), the whole document	1-30
A	CN 103500140 A (INSUR ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 08 January 2014 (08.01.2014), the whole document	1-30
A	CN 103117901 A (HUAWEI TECHNOLOGIES CO., LTD.), 22 May 2013 (22.05.2013), the whole document	1-30
A	US 2014181035 A1 (FUJITSU CO., LTD.), 26 June 2014 (26.06.2014), the whole document	1-30

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 27 October 2015 (27.10.2015)	Date of mailing of the international search report 10 November 2015 (10.11.2015)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer LI, Zhenhua Telephone No.: (86-10) 62087678

International application No.
PCT/CN2015/084573

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2015/084573

A. 主题的分类

H04L 29/08(2006.01)i; H04L 12/26(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L, H04B, G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CPRSABS, CNTXT, CNKI, VEN: 分布式, 集群, 故障, 节点, 主, 从, 备用, 备份, 参考, 心跳消息, 状态, 失联, 中断; distributed, cluster, failure, node, NB, master, slave, standby, backup, reference, heartbeat message, state, disconnect, interrupt

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 101494565 A (中国移动通信集团公司) 2009年 7月 29日 (2009 - 07 - 29) 全文	1-30
A	CN 102752093 A (中国联合网络通信集团有限公司) 2012年 10月 24日 (2012 - 10 - 24) 全文	1-30
A	CN 103500140 A (浪潮电子信息产业股份有限公司) 2014年 1月 8日 (2014 - 01 - 08) 全文	1-30
A	CN 103117901 A (华为技术有限公司) 2013年 5月 22日 (2013 - 05 - 22) 全文	1-30
A	US 2014181035 A1 (FUJITSU CO., LTD.) 2014年 6月 26日 (2014 - 06 - 26) 全文	1-30

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2015年 10月 27日

国际检索报告邮寄日期

2015年 11月 10日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
北京市海淀区蓟门桥西土城路6号
100088 中国

传真号 (86-10)62019451

授权官员

李振华

电话号码 (86-10)62087678

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/084573

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	101494565	A	2009年 7月 29日	CN	101494565	B	2011年 1月 26日
CN	102752093	A	2012年 10月 24日	无			
CN	103500140	A	2014年 1月 8日	无			
CN	103117901	A	2013年 5月 22日	无			
US	2014181035	A1	2014年 6月 26日	JP	2014123218	A	2014年 7月 3日

表 PCT/ISA/210 (同族专利附件) (2009年7月)