(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 11 月 17 日 (17.11.2016)

WIPO | PCT

(10) 国际公布号
WO 2016/180005 A1

(51) 国际专利分类号:
H04L 12/24 (2006.01) *H04L 12/40* (2006.01)

(21) 国际申请号: PCT/CN2015/095654

(22) 国际申请日: 2015 年 11 月 26 日 (26.11.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510244239.8 2015 年 5 月 14 日 (14.05.2015) 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: **伍湘平 (WU, Xiangping)**; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 **李龙 (LI, Long)**; 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)。

根据细则 4.17 的声明:

— 关于申请人有权申请并被授予专利(细则 4.17(ii))

[见续页]

(54) **Title**: METHOD FOR PROCESSING VIRTUAL MACHINE CLUSTER AND COMPUTER SYSTEM

(54) **发明名称** : 处理虚拟机集群的方法和计算机系统

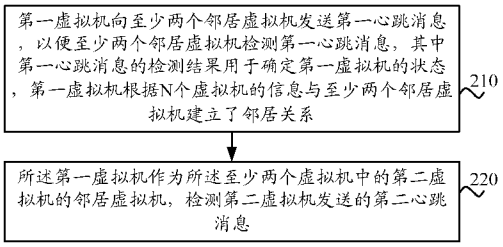

图2

210 A FIRST VIRTUAL MACHINE SENDING A FIRST HEARTBEAT MESSAGE TO AT LEAST TWO NEIGHBOUR VIRTUAL MACHINES, SO THAT THE AT LEAST TWO NEIGHBOUR VIRTUAL MACHINES DETECT THE FIRST HEARTBEAT MESSAGE, WHEREIN A DETECTION RESULT OF THE FIRST HEARTBEAT MESSAGE IS USED FOR DETERMINING THE STATE OF THE FIRST VIRTUAL MACHINE, AND THE FIRST VIRTUAL MACHINE ESTABLISHES A NEIGHBOUR RELATIONSHIP WITH THE AT LEAST TWO NEIGHBOUR VIRTUAL MACHINES ACCORDING TO THE INFORMATION ABOUT THE N VIRTUAL MACHINES

220 THE FIRST VIRTUAL MACHINE, AS A NEIGHBOUR VIRTUAL MACHINE OF A SECOND VIRTUAL MACHINE IN THE AT LEAST TWO VIRTUAL MACHINES, DETECTING A SECOND HEARTBEAT MESSAGE SENT BY THE SECOND VIRTUAL MACHINE

(57) **Abstract**: Provided are a method for processing virtual machine cluster and a computer system. The virtual machine cluster comprises N virtual machines, each of the N virtual machines saves state information about the N virtual machines, and works as a first virtual machine in peer-to-peer mode. A virtual machine list comprises information about the N virtual machines. The method comprises: the first virtual machine sending a first heartbeat message to at least two neighbour virtual machines in the N virtual machines, so that the at least two neighbour virtual machines detect the first heartbeat message, wherein a detection result of the first heartbeat message is used for determining the state of the first virtual machine, and the first virtual machine establishes a neighbour relationship with the at least two neighbour virtual machines according to the information about the N virtual machines; and the first virtual machine, as a neighbour virtual machine of a second virtual machine in the at least two virtual machines, detecting a second heartbeat message sent by the second virtual machine. The technical solution of the embodiments of the present invention can improve the fault-tolerant capability and the performance of a virtual machine cluster.

(57) **摘要**:

[见续页]

WO 2016/180005 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

本发明提供了一种管理虚拟机集群的方法和计算机系统。该虚拟机集群包括 N 个虚拟机，N 个虚拟机中的每个虚拟机保存 N 个虚拟机的状态信息，并以对等方式作为第一虚拟机进行工作，虚拟机列表包括 N 个虚拟机的信息，该方法包括：第一虚拟机向 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，所述第一虚拟机根据所述 N 个虚拟机的信息与所述至少两个邻居虚拟机建立了邻居关系；；第一虚拟机作为至少两个虚拟机中的第二虚拟机的邻居虚拟机，检测第二虚拟机发送的第二心跳消息。本发明实施例的技术方案能够提高虚拟机集群的容错能力和性能。

处理虚拟机集群的方法和计算机系统

本申请要求于 2015 年 5 月 14 日提交中国专利局、申请号为201510244239.8、发明名称为“处理虚拟机集群的方法和计算机系统”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

本发明实施例涉及信息技术领域，并且更具体地，涉及一种处理虚拟机集群的方法和计算机系统。

背景技术

集群（Cluster）通常是由一些互相连接在一起的节点（例如，计算机或虚拟机）构成的一个并行或分布式系统。这些节点一起工作并运行一系列共同的应用程序，同时，为用户和应用程序提供单一的系统映射。例如，对于计算机集群而言，从外部来看，计算机集群是一个系统，对外提供统一的服务，对内部来说，集群内的计算机在物理上通过电缆连接，在逻辑上则通过集群软件连接。服务器集群是把多台服务器通过通信链路连接，从外部看来，这些服务器就像一台服务器在工作，而对内部来说，外来的负载通过一定的机制动态地分配到服务器中去，从而达到超级服务器才有的高性能、高可用。

虚拟机（英文：Virtual Machine，简称“VM”）是在主机（host）上运行的软件，其可以在计算机平台和终端用户之间创造一种环境，而终端用户则是基于这个软件所创造的环境来操作。虚拟机集群是指多个虚拟机相互连接在一起构成的并行或分布式系统。

虚拟机集群通常采用主/从（Master/Slave）架构的方式。Master 主机负责监测所有的 slave 主机，并在 Slave 主机宕机时对 Slave 主机上的虚拟机进行重启。从节点也会接收主节点发送的心跳消息，以便确认主节点是否存活。如果主 Master 主机的主机宕机了，集群中的 Slave 主机会重新选择一个 Master 主机。

Master 主机作为集群的管理中心，负责集群中所有 Slave 主机的监测和管理。当集群中的 Slave 主机过多时，Master 主机的性能会不足以支持维护

大量的 Slave 主机，使得 Master 成为整个集群的瓶颈，降低了虚拟机集群的整体性能。同时，当 Master 主机的主机发生故障时，Slave 主机将重新选出新的 Master 主机。这一过程需要耗费一定的时间，因此会拖延集群的故障恢复时间，降低了虚拟机集群的容错能力。此外，一些 Slave 主机可能失去与 Master 主机的联系。这部分 Slave 主机会重新选举 Master 主机。这就导致了一个集群中出现了两个各自独立的集群分区。由于一个集群中出现了两个独立的集群分区，使得两个分区的 Master 主机均误以为对方出现了故障，从而争抢资源，造成资源不足和数据破坏，降低了虚拟机集群的性能。因此，采用主节点为虚拟机集群的管理中心会影响虚拟机集群的容错能力和性能。

发明内容

本发明实施例提供的处理虚拟机集群的方法和计算机系统，能够提高虚拟机集群的容错能力和性能。

第一方面，提供了一种处理虚拟机集群的方法，虚拟机集群包括 N 个虚拟机，N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，虚拟机列表包括 N 个虚拟机的信息，第一方面的方法包括：第一虚拟机向 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，第一虚拟机根据 N 个虚拟机的信息与至少两个邻居虚拟机建立了邻居关系；第一虚拟机作为至少两个虚拟机中的第二虚拟机的邻居虚拟机，检测第二虚拟机发送的第二心跳消息，其中第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果用于确定第二虚拟机的状态。

结合第一方面，在第一方面的第一种可能的实现方式中，该方法还包括：第一虚拟机向至少两个邻居虚拟机发送第一同步信息，第一同步信息用于指示第一虚拟机中保存的虚拟机列表的更新，以便至少两个邻居虚拟机更新各自保存的虚拟机列表；第一虚拟机接收第二虚拟机发送的第二同步信息，第二同步信息用于指示第二虚拟机中保存的虚拟机列表的更新，第一虚拟机根据第二同步信息更新第一虚拟机保存的虚拟机列表。

结合第一方面或第一种可能的实现方式，在第三种可能的实现方式中，所述至少两个邻居虚拟机包括所述 N 个虚拟机中具备与所述第一虚拟机直

接交互信息的能力的二至六个虚拟机。

结合第二方面或第二方面的上述任一种可能的实现方式，在第三种可能的实现方式中，该方法还包括：第一虚拟机在确定第二虚拟机的状态为故障的情况下，第二虚拟机触发第二虚拟机重启或触发第二虚拟机从第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

结合第一方面或第一方面的上述任一种可能的实现方式，在第四种可能的实现方式中，该第一反馈信息还包括该第二虚拟机的配置信息；该方法还包括：第一虚拟机在确定第二虚拟机的状态为故障且无法重启的情况下，第一虚拟机触发第二虚拟机从第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

结合第一方面上述任一种可能的实现方式，在第五种可能的实现方式中，该方法还包括：第一虚拟机确定第二虚拟机的状态为离开的情况下，第一虚拟机触发第二虚拟机从第二虚报机所在的源主机上删除。

结合第一方面或第一方面的上述任一种可能的实现方式，在第六种可能的实现方式中，第一方面的方法还包括：第一虚拟机向 N 个虚拟机的上层节点发送第二心跳消息的检测结果，以便上层节点根据第二心跳消息的检测结果和第二虚拟机发送给第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定第二虚拟机的状态；第一虚拟机接收上层节点发送的指示消息，指示消息用于指示第二虚拟机的状态。

结合第一方面或第一方面的上述任一种可能的实现方式，在第七种可能的实现方式中，第一方面的方法还包括：第一虚拟机接收第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果；第一虚拟机根据第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果确定第二虚拟机的状态。

结合第一方面或第一方面的上述任一种可能的实现方式，在第八种可能的实现方式中，在第一虚拟机加入虚拟机集群时，第一方面的方法还包括：第一虚拟机向 N 个虚拟机中的其它虚拟机发送第一虚拟机的信息；第一虚拟机接收 N 虚拟机中其它虚拟机发送的各自的信息，以生成虚拟机列表。

结合第一方面或第一方面的上述任一种可能的实现方式，在第九种可能的实现方式中，在第一虚拟机加入虚拟机集群时，第一虚拟机向索引服务器

发送注册信息，注册信息包括第一虚拟机的信息，其中索引服务器为虚拟机集群的注册中心，用于为虚拟机集群中的虚拟机提供注册服务；第一虚拟机接收索引服务器发送的 N 个虚拟机中的其它虚拟机的信息，以生成虚拟机列表。

结合第一方面或第一方面的上述任一种可能的实现方式，在第十种可能的实现方式中，还包括：第一虚拟机采用邻居关系算法，根据虚拟机列表中保存的 N 个虚拟机的信息，从 N 个虚拟机中选择至少两个作为邻居虚拟机。

结合第一方面或第一方面的上述任一种可能的实现方式，在第十一种可能的实现方式中，N 个虚拟机的信息包括：N 个虚拟机中的每个虚拟机的状态信息、N 个虚拟机中的每个虚拟机的邻居关系信息、N 个虚拟机中的每个虚拟机的启动次数、N 个虚拟机中的每个虚拟机的心跳值、N 个虚拟机中的每个虚拟机的配置信息以及虚拟机集群的配置信息中的任意一个或多个的组合。

第二方面，提供了一种计算机系统，其特征在于，计算机系统包括至少一个计算机节点的物理硬件层，在至少一个计算机节点的物理硬件层之上运行虚拟机集群，虚拟机集群包括 N 个虚拟机，N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，虚拟机列表包括 N 个虚拟机的信息，第一虚拟机包括：发送模块，用于向 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，第一虚拟机根据 N 个虚拟机的信息与至少两个邻居虚拟机建立了邻居关系；接收模块，用于检测至少两个虚拟机中的第二虚拟机发送的第二心跳消息，其中第一虚拟机为第二虚拟机的邻居虚拟机，第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果用于确定第二虚拟机的状态。

结合第二方面，在第一种可能的实现方式中，发送模块还向至少两个邻居虚拟机发送第一同步信息，第一同步信息用于指示第一虚拟机中保存的虚拟机列表的更新，以便至少两个邻居虚拟机更新各自保存的虚拟机列表，接收模块还接收第二虚拟机发送的第二同步信息，第二同步信息用于指示第二虚拟机中保存的虚拟机列表的更新，第一虚拟机还包括更新模块，用于根据第二同步信息更新第一虚拟机保存的虚拟机列表。

结合第二方面或第二方面的第一种可能的实现方式，在第二种可能的实现方式中，所述至少两个邻居虚拟机包括所述 N 个虚拟机中具备与所述第一虚拟机直接交互信息的能力的二至六个虚拟机。

结合第二方面或第二方面的上述任一种可能的实现方式，在第三种可能的实现方式中，第二方面的计算机系统还包括：触发模块，用于在确定第二虚拟机的状态为故障的情况下，触发第二虚拟机重启或触发第二虚拟机从第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

结合第二方面或第二方面的上述任一种可能的实现方式，在第四种可能的实现方式中，第二方面的计算机系统还包括：触发模块，用于在确定所述第二虚拟机的状态为故障且无法重启的情况下，触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

结合第二方面或第二方面的上述任一种可能的实现方式，在第五种可能的实现方式中，第二方面的计算机系统还包括：触发模块，用于在确定所述第二虚拟机的状态为离开的情况下，触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

结合第二方面或第二方面的上述任一种可能的实现方式，在第六种可能的实现方式中，发送模块还向 N 个虚拟机的上层节点发送第二心跳消息的检测结果，以便上层节点根据第二心跳消息的检测结果和第二虚拟机发送给第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定第二虚拟机的状态，接收模块还接收上层节点发送的指示消息，指示消息用于指示第二虚拟机的状态。

结合第二方面或第二方面的上述任一种可能的实现方式中，在第七种可能的实现方式中，接收模块还接收第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果，接收模块还根据第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果确定第二虚拟机的状态。

结合第二方面或第二方面的上述任一种可能的实现方式，在第八种可能的实现方式中，发送模块还在第一虚拟机加入虚拟机集群时，向 N 个虚拟机中的其它虚拟机发送第一虚拟机的信息，其中索引服务器为虚拟机集群的注

册中心，用于为虚拟机集群中的虚拟机提供注册服务，接收模块还接收 N 虚拟机中其它虚拟机发送的各自的信息，以生成虚拟机列表。

结合第二方面或第二方面的上述任一种可能的实现方式中，在第九种可能的实现方式中，发送模块还在第一虚拟机加入虚拟机集群时，向索引服务器发送注册信息，注册信息包括第一虚拟机的信息，接收模块还接收索引服务器发送的 N 个虚拟机中的其它虚拟机的信息，以生成虚拟机列表。

结合第二方面或第二方面的上述任一种可能的实现方式中，在第十种可能的实现方式中，第二方面的计算机系统还包括：选择模块，用于采用邻居关系算法，根据虚拟机列表中保存的 N 个虚拟机的信息，从 N 个虚拟机中选择至少两个作为邻居虚拟机。

结合第二方面或第二方面的上述任一种可能的实现方式中，在第十一种可能的实现方式中，N 个虚拟机的信息包括：N 个虚拟机中的每个虚拟机的状态信息、N 个虚拟机中的每个虚拟机的邻居关系信息、N 个虚拟机中的每个虚拟机的启动次数、N 个虚拟机中的每个虚拟机的心跳值、N 个虚拟机中的每个虚拟机的配置信息以及虚拟机集群的配置信息中的任意一个或多个的组合。

上述技术方案中，根据本发明的实施例，虚拟机集群中的每个虚拟机可以根据其保存虚拟机列表确定至少两个邻居虚拟机，虚拟机集群中的每个虚拟机可以向其至少两个邻居虚拟机发送心跳消息以便根据至少两个邻居虚拟机的检测结果确定该虚拟机的状态。由于每个虚拟机的状态均可以由其邻居虚拟机检测该虚拟机发送的心跳信息的结果来确定，因此避免了主/从结构存在的 Master 成为整个集群的瓶颈问题，同时由于不会出现重新选举 Master 主机的情况，因此，使得采用这种方案进行状态确定的虚拟机集群不会造成故障恢复时间的延迟和争抢资源的情况，因此，提高了虚拟机集群的容错能力和性能。

附图说明

为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例中所需要使用的附图作简单地介绍，显而易见地，下面所描述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。

图 1 是根据本发明实施例提供的虚拟机集群的架构的示意图。

图 2 是根据本发明实施例的一种处理虚拟机集群的方法的示意性流程图。

图 3 是本发明的实施例的虚拟机集群的邻居关系的示意图。

图 4 是根据本发明的实施例的建立虚拟机集群的过程的示意性流程图。

图 5 是根据本发明的实施例的心跳检测机制的示意图。

图 6 是根据本发明的实施例的虚拟机间同步的示意性流程图。

图 7 是根据本发明的实施例的一种计算机系统的结构示意图。

图 8 是根据本发明的实施例的一种计算机系统 800 的结构示意图。

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清楚、完整地描述，显然，的实施例是本发明的一部分实施例，而不是全部实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都应属于本发明保护的范围。

图 1 是根据本发明实施例提供的虚拟机集群 100 的架构的示意图。

如图 1 所示，虚拟机集群 100 为分布式架构，虚拟机集群中的多个虚拟机从用户端来看整体上作为单一虚拟机为用户设备提供业务。与常规技术中 Master/Slave 虚拟机集群架构相比，虚拟机集群 100 中的虚拟机之间的关系是对等的。虚拟机集群 100 所在的计算机系统例如可以包括物理主机 ESXi-1、ESXi-2 和 ESXi-3。例如，在主机 ESXi-1 上运行虚拟机 VM101 和 VM102，在主机 ESXi-2 上运行虚拟机 VM103 和 VM104，在主机 ESXi-3 上运行虚拟机 VM105 和 VM106。虚拟机集群 100 中的每个虚拟机可以维护一个虚拟机列表，用于保存虚拟机集群 100 中的虚拟机的信息，包括状态信息、配置信息和/或管理信息。虚拟机之间的信息同步可以采用对等（Peer to Peer, P2P）协议来实现。其中状态信息用于指示虚拟机的工作状态，例如，CPU 的使用量或内存的使用量等信息。配置信息用于指示配置虚拟机的相关信息，例如，分配给虚拟机的 IP 地址等信息。管理信息用于指示管理虚拟机的信息，例如，虚拟机的心跳值、虚拟机的启动次数、虚拟机的故障、重启或迁移等信息。

虚拟机 VM101 与虚拟机 VM102、VM103、VM105 和 VM106 建立了邻

居关系；虚拟机 102 与虚拟机 VM101、VM104 和 VM105 建立了邻居关系；虚拟机 103 与虚拟机 VM101、VM106 和 VM104 建立了邻居关系；虚拟机 104 与虚拟机 VM101、VM102 和 VM106 建立了邻居关系；虚拟机 105 与虚拟机 VM101、VM102 和 VM103 建立了邻居关系；虚拟机 106 与虚拟机 VM101、VM104 和 VM105 建立了邻居关系。虚拟机的邻居关系也可以记录在虚拟机列表中。例如，虚拟机列中还可以列出了每个虚拟机的邻居虚拟机。每个虚拟机可以采用特定的邻居关系算法确定各自的邻居虚拟机。

应理解，上述虚拟机集群可以位于一个物理主机上，也可以位于多个物理主机上。虚拟机之间的邻居关系与其所处的物理主机无关，即从分布式集群的角度，只考虑虚拟机，而不考虑其实际位于哪一台物理主机。虚拟机可以将位于同一物理主机上的虚拟机作为邻居，也可以将位于其它物理主机上的虚拟机作为邻居。

还应理解，上述邻居关系可以是指物理上的邻居关系，也可以是指逻辑上的邻居关系。

还应理解，上述物理主机和每个物理主机上虚拟机的数目仅仅是举例说明。本发明的实施例对每个物理主机上的虚拟机的数目和虚拟机集群所在的物理主机的数目不作限定。

图 2 是根据本发明实施例的一种处理虚拟机集群的方法的示意性流程图。图 2 所示的方法由图 1 的虚拟机集群 100 中的每个虚拟机执行，虚拟机集群包括 N 个虚拟机，N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，虚拟机列表包括 N 个虚拟机的信息。图 2 的方法包括如下内容。

210，第一虚拟机向 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，所述第一虚拟机根据所述 N 个虚拟机的信息与所述至少两个邻居虚拟机建立了邻居关系。

220，第一虚拟机作为上述至少两个虚拟机中的第二虚拟机的邻居虚拟机，检测第二虚拟机发送的第二心跳消息，其中第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果用于确定第二虚拟机的状态。

根据本发明的实施例，第一虚拟机作为虚拟机集群中的任一个虚拟机可

以通过向它的邻居虚拟机发送心跳消息，以便第一虚拟机的邻居虚拟机监测第一虚拟机是否存活。同时，该第一虚拟机作为其它虚拟机的邻居虚拟机也可以接收其它虚拟机发送心跳消息，以便监测其它虚拟机是否存活。具体而言，虚拟机集群中的每个虚拟机都可以以对等方式执行上述 210 和 220，即每个虚拟机均可以向其邻居虚拟机发送心跳消息，以便其邻居虚拟机监测该虚拟机的状态，例如，是否存活。同时，该虚拟机作为其它虚拟机的邻居虚拟机也可以接收其它虚拟机发送心跳消息，以便监测其它虚拟机的状态，这样，每个虚拟机的状态可以根据该虚拟机的多个邻居虚拟机的一次心跳检测结果来综合确定。换句话说，该虚拟机集群中的每个虚拟机都可以对其邻居虚拟机进行监测，同时也被其邻居虚拟机监测。因此，该虚拟机集群无需一个用于监测所有虚拟机的心跳消息的主节点。

根据本发明的实施例，虚拟机集群中的每个虚拟机可以根据其保存虚拟机列表确定至少两个邻居虚拟机，虚拟机集群中的每个虚拟机可以向其至少两个邻居虚拟机发送心跳消息以便根据至少两个邻居虚拟机的检测结果确定该虚拟机的状态。由于每个虚拟机的状态均可以由其邻居虚拟机检测该虚拟机发送的心跳信息的结果来确定，因此避免了主/从结构存在的 Master 成为整个集群的瓶颈问题，同时由于不会出现重新选举 Master 主机的情况，因此，使得采用这种方案进行状态确定的虚拟机集群不会造成故障恢复时间的延迟和争抢资源的情况，因此，提高了虚拟机集群的容错能力和性能。

具体地，由于该虚拟机集群中的每个虚拟机与其他虚拟机的地位都是相同的或者对等的，因此该虚拟机集群不会出现主/从架构的虚拟机集群存在的问题。由于本发明的实施例的虚拟机集群中不需要主节点，因此也就不会出现由于主节点性能不足以支持维护大量从节点导致的问题，也不会出现由于主节点发生故障导致的重新选取主节点带来的问题（例如，恢复时间较长、出现集群脑裂等）。

在 220 中，图 1 的方法还包括：第一虚拟机可以采用邻居关系算法，根据虚拟机列表中保存的 N 个虚拟机的信息，从 N 个虚拟机中选择至少两个作为邻居虚拟机。上述至少两个邻居虚拟机可以包括 N 个虚拟机中具备与第一虚拟机直接进行信息交互的能力的二至六个虚拟机。

例如，上述邻居关系算法可以包含用于确定虚拟机的邻居关系的准则或策略。根据本发明的实施例可以采用 P2P 技术中的邻居发现算法来确定邻居

关系，例如，每个虚拟机可以选择距离该虚拟机的物理位置最近且不属于同一主机的 2 至 6 个虚拟机作为该虚拟机的邻居虚拟机。本发明的实施例对邻居关系算法不作限定，例如，每个虚拟机还可以从虚拟机列表中随机选择多个虚拟机作为其邻居虚拟机。

根据本发明的实施例，N 个虚拟机的信息包括：N 个虚拟机中的每个虚拟机的状态信息、N 个虚拟机中的每个虚拟机的邻居关系信息、N 个虚拟机中的每个虚拟机的启动次数、N 个虚拟机中的每个虚拟机的心跳值。配置信息包括：N 个虚拟机中的每个虚拟机的配置信息以及虚拟机集群的配置信息。虚拟机列表可以包括上述管理信息和配置信息中的任意一个或多个的组合。

例如，某个虚拟机的启动次数可以指该虚拟机加入虚拟机集群以来启动的次数，用于确定该虚拟机故障之后是否重启，例如当启动次数超过预设的阈值之后不再重启该虚拟机，并加入新的虚拟机以保证整个系统的稳定性。某个虚拟机的心跳值指该虚拟机上一次启动以来发送心跳消息的总数，用于确定该虚拟机上次启动以来正常时间。虚拟机的配置信息可以包括该虚拟机所属的虚拟机集群的配置信息（英文：Cluster Configuration，简称：ClusterConf）和虚拟机的节点信息（英文：Node Information，简称：NodeInf）。其中 ClusterConf 的取值越大，表示 ClusterConf 的值越新，NodeInf 的取值越大，表示 NodeInf 的值越新。该虚拟机可以将虚拟机列表中保存的信息发送给其它虚拟机。其它虚拟机在接收到该信息后，根据该信息，对该虚拟机保存的虚拟机的信息进行维护或更新。

应理解，上述信息也可以采用其它形式进行保存，例如，上述信息可以采用数组的形式来保存。还应理解，上述信息还可以包含其它可以用于确定邻居关系的信息，例如其它虚拟机的邻居关系信息。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机向至少两个邻居虚拟机发送第一同步信息，第一同步信息用于指示第一虚拟机中保存的虚拟机列表的更新，以便至少两个邻居虚拟机更新各自保存的虚拟机列表；第一虚拟机接收第二虚拟机发送的第二同步信息，第二同步信息用于指示第二虚拟机中保存的虚拟机列表的更新；第一虚拟机根据第二同步信息更新第一虚拟机保存的虚拟机列表。

例如，当虚拟机集群中的每个虚拟机中保存的状态被更新时，该虚拟机

可以通过同步信息将该虚拟机中保存的 N 个虚拟机的信息发给其邻居虚拟机，或者，该虚拟机可以通过同步信息仅将更新的虚拟机的信息发送给邻居虚拟机，或者该虚拟机可以通过同步信息仅将更新的虚拟机的信息的指示或索引发送给邻居虚拟机，以便其邻居虚拟机根据上述同步信息更新虚拟机的信息。

通过同步信息的交互，每个虚拟机都可以获取该虚拟机的每个邻居虚拟机保存的虚拟机的管理信息和配置信息，并且这些管理信息和配置信息都是最新的。可以理解的是，如果该虚拟机集群中的每个虚拟机都能获取邻居虚拟机的所保存的虚拟机的管理信息和配置信息，并且能够将自己保存的虚拟机的管理信息和配置信息发送给自己的邻居虚拟机，那么每个虚拟机都能够保存有整个虚拟机集群中的所有虚拟机的管理信息和配置信息。这样，当该虚拟机集群中的任一个虚拟机发生故障时，保存有发生故障的虚拟机的管理信息的虚拟机可以利用保存的管理信息进行对该发生故障的虚拟机进行恢复。例如，可以在其他物理主机上重建该发生故障的虚拟机。可以理解的是，在虚拟机发生故障时，首先可以对发生故障的虚拟机进行重启，如果重启不成功，则可以在其他物理主机上重建该虚拟机。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机在确定第二虚拟机的状态为故障的情况下，第二虚拟机触发第二虚拟机重启或触发所述第二虚拟机从第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

例如，虚拟机集群中的每个虚拟机在确定另一虚拟机的状态为故障的情况下，可以指示故障虚拟机所在的物理主机重新启动或迁移该虚拟机。例如，虚拟机可以通过报警方式指示人工完成虚拟机的重启或迁移，或者运行专用的迁移软件来执行重启或迁移。通过重启，有可能能够使故障虚拟机重新正常工作。通过迁移，可以将故障虚拟机的配置文件和磁盘文件从源主机拷贝至目标主机，从而使得故障虚拟机能够在目标主机上重新工作。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机在确定第二虚拟机的状态为故障且无法重启的情况下，所述第一虚拟机在确定所述第二虚拟机的状态为故障且无法重启的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

可选地，第一虚拟机接收迁移后的第二虚拟机发送的信息，以更新迁移后的第二虚拟机的信息，其中，第一虚拟机向至少两个邻居虚拟机发送第一同步信息，包括：第一虚拟机向至少两个邻居虚拟机发送第一同步信息，第一同步信息包括迁移后的第二虚拟机的信息，以便第一虚拟机的至少两个邻居虚拟机更新迁移后的第二虚拟机的信息。

例如，如果虚拟机集群中的每个虚拟机在确定另一虚拟机的状态为故障的情况下，可以指示故障虚拟机所在的物理主机重新启动该虚拟机。如果该虚拟机确定故障虚拟机无法重启，例如，在发出重启命令预设时间（例如，该预设时间可以大于虚拟机的重启时间）后仍未收到故障虚拟机发送的心跳消息，则认为该故障虚拟机无法重启，在这种情况下，该虚拟机发出迁移指示，例如，可以将故障虚拟机迁移至另一物理主机，即在另一物理主机上重启该故障虚拟机。上述预设时间的设置使得故障虚拟机在能够重启成功的情况下保持虚拟机集群的邻居关系不变，从而无需再重新确定邻居关系。

应理解，故障虚拟机在重启后可以仍然保留原来的邻居关系，或者重新确定邻居关系。在重新确定邻居关系的情况下，可以通过同步信息进行信息的同步。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机确定第二虚拟机的状态为离开的情况下，所述第一虚拟机确定所述第二虚拟机的状态为离开的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

例如，虚拟机集群中的每个虚拟机在确定另一虚拟机的状态为离开的情况下，可以指示删除该虚拟机。例如，虚拟机可以通过报警方式指示人工完成虚拟机的删除，或者运行专用的迁移软件来执行删除。例如，通过删除，可以将故障虚拟机的配置文件和磁盘文件从源主机上删除。

可选地，第一虚拟机可以删除虚拟机列表中保存的第二虚拟机的信息，其中，第一虚拟机向至少两个邻居虚拟机发送第一同步信息，包括：第一虚拟机向第一虚拟机的至少两个邻居虚拟机发送第一同步信息，第一同步信息包括用于指示删除第二虚拟机的指示信息，以便第一虚拟机的至少两个邻居虚拟机删除第二虚拟机的信息。

例如，如果某个虚拟机被主动停止运行，则说明该虚拟机不再使用，可以从虚拟机集群中删除，在这种情况下，该离开虚拟机的邻居虚拟机在获知

该虚拟机离开之后，可以从保存虚拟机列表中删除该离开虚拟机的信息，从而触发其它虚拟机从保存的虚拟机列表中删除该离开虚拟机的信息。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机向 N 个虚拟机的上层节点发送第二心跳消息的检测结果，以便上层节点根据第二心跳消息的检测结果和第二虚拟机发送给第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定第二虚拟机的状态；第一虚拟机接收上层节点发送的指示消息，指示消息用于指示第二虚拟机的状态。

根据本发明的实施例，可以通过每个虚拟机的多个邻居虚拟机向上层节点（例如，管理节点）上报心跳消息的检测结果，并综合多个邻居虚拟机的检测结果来确定该虚拟机的状态，以便更准确地确定该虚拟机的状态。例如，每个虚拟机的邻居虚拟机均检测该虚拟机发送的一次心跳消息，并且均将检测结果上报给上层节点。上层节点在确定该虚拟机的每个邻居虚拟机在预设时间内均未检测到心跳消息时，可以确定该虚拟机故障或离开，并将该虚拟机故障或离开的信息通知给该虚拟机的各个邻居虚拟机。这样做的好处于在于可以通过检测虚拟机发送的一次心跳消息确定该虚拟机是否故障或离开，从而准确及时发现故障或离开的虚拟机。而在采用 Master/Slave 结构的常规技术中，Master 节点需要根据 Slave 节点发送的多次心跳消息来准确发现故障或离开的节点，从而无法及时发现故障或离开的节点。

可选地，作为另一实施例，图 2 的方法还包括：第一虚拟机接收第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果；第一虚拟机根据第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果确定第二虚拟机的状态。

例如，每个虚拟机的邻居虚拟机均可以根据该虚拟机发送的心跳消息确定该虚拟机的状态（例如，故障或离开），而且该虚拟机的每个邻居虚拟机可以从该虚拟机的其它邻居虚拟机接收其它邻居虚拟机检测的该虚拟机的状态，并根据这些检测来确定该虚拟机的状态。例如，如果该虚拟机的每个邻居虚拟机检测到该虚拟机故障，同时接收到其它邻居虚拟机发送的检测到该虚拟机故障的消息，则可以判断该虚拟机故障。虚拟机的每个邻居虚拟机可以在预定时间内没有检测到心跳消息时向其它邻居虚拟机主动发送检测结果，或者周期性地向其它邻居虚拟机发送检测结果，本发明的实施例对此不作限定，例如，也可以是在预设定时间内没有检测到心跳消息时向其它邻

居虚拟机主动请求发送检测结果。

在根据本发明的实施例的虚拟机集群的初始建立过程中，每个节点需要获知虚拟机集群中有哪些虚拟机，以及这些虚拟机的相关信息，然后在此基础上确定哪些虚拟机可以生成为自己的邻居。

可选地，作为另一实施例，图 2 的方法还包括：在第一虚拟机加入虚拟机集群时，第一虚拟机向 N 个虚拟机中的其它虚拟机发送第一虚拟机的信息；第一虚拟机接收 N 虚拟机中其它虚拟机发送的各自的信息。

具体而言，虚拟机之间可以采用 P2P 协议，例如，Gossip 协议进行通信。Gossip 进程是一个定时程序，本发明的实施例可以利用该进程，每隔 1s 从本地维护的虚拟机列表中随机选取一定数量（例如，3 个）的其它虚拟机进行通信，以交换各自的信息。这种方式的优点是整个集群的构建过程完全是自组织的，减少了配置的环节。

可选地，作为另一实施例，图 2 的方法还包括：在第一虚拟机加入虚拟机集群时，第一虚拟机向索引服务器发送注册信息，注册信息包括第一虚拟机的信息，其中所述索引服务器为所述虚拟机集群的注册中心，用于为所述虚拟机集群中的虚拟机提供注册服务；第一虚拟机接收索引服务器发送的 N 个虚拟机中的其它虚拟机的信息。

例如，虚拟机集群中所有虚拟机节点在启动和初始加入集群时，都将自身信息注册到该索引服务器中，再从该服务器获取当前集群的其他虚拟机的信息。应理解，索引服务器可以是一个单独的节点，也可以是集群中的一个虚拟机，如第一台加入集群的虚拟机。

应理解，根据本发明的实施例也可以由管理员通过配置工具在每个虚拟机上配置所有虚拟机的信息，然后各虚拟机根据特定的邻居关系算法计算自己的邻居虚拟机并与之建立邻接关系。这种方式的优点是，可以快速获知集群节点信息，避免了广播 hello 报文所存在的诸多弊端。

为了帮助本领域技术人员更好地理解本发明，下面将结合具体实施例对本发明进行进一步描述。可以理解的是，该具体实施例仅是为了帮助更好地理解本发明的技术方案，而并非对本发明的技术方案的限制。

图 3 是本发明的实施例的虚拟机集群 300 的邻居关系的示意图。以图 3 的虚拟机集群 300 为例，图 2 的实施例中的第一虚拟机可以是集群 300 中的任一个虚拟机。

假设本具体实施例中的第一虚拟机为 VM 301。可以看出，VM 301 与 VM 302 和 VM 304 建立了邻居关系。因此，VM 301 的邻居虚拟机为 VM 302 和 VM 304。第二虚拟机可以是 VM 302 和 VM 304 中的任一个虚拟机。假设本具体实施例中的第二虚拟机为 VM 302。

假设第一虚拟机 VM 301 保存有 VM 301 的管理信息、VM 304 的管理信息和 VM 307 的管理信息。该第一节点列表保存的 VM 301 的管理信息中 VM 301 的启动次数为 2，VM 301 的心跳值为 8；该第一节点列表保存的 VM 204 的管理信息中的 VM 304 的启动次数为 3，VM 304 的心跳值为 3；该第一节点列表保存的 VM 307 的管理信息中 VM 307 的启动次数为 1，VM 307 的心跳值为 9。

假设第二虚拟机 VM 302 保存有 VM 302 的管理信息，VM 301 的管理信息和 VM 303 的管理信息。该第二节点列表保存的 VM 301 的管理信息中 VM 301 的启动次数为 2，VM 301 的心跳值为 2；该第二节点列表保存的 VM 302 的管理信息中 VM 302 的启动次数为 3，VM 302 的心跳值为 3；该第二节点列表中保存的 VM 303 的管理信息中的 VM 303 的启动次数为 1，VM 303 的心跳值为 5。

此外，第一虚拟机和第二虚拟机保存的虚拟机的管理信息中还包括虚拟机的标识符。为方便描述，假设在本实施例中每个虚拟机的标识符就是该虚拟机在集群 300 中对应的编号，例如 VM 301 的标识符就是“VM 301”。

进一步，该第一虚拟机的配置信息可以包括该第一虚拟机所属的簇的配置信息（ClusterConf）和该第一虚拟机所属的节点信息（NodeInf）。同理，该第二虚拟机的配置信息包括该第二虚拟机所属的簇的配置信息和该第二邻居虚拟机所属的节点信息。在本实施例中，该第一虚拟机的配置信息中的 ClusterConf 可以为 0，NodeInf 可以为 1。该第二虚拟机的配置信息中的 ClusterConf 可以为 1，NodeInf 可以为 0，其中 ClusterConf 的取值越大，表示 ClusterConf 的值越新，NodeInf 的取值越大，表示 NodeInf 的值越新

该第一虚拟机可以将该第一节点列表中保存的 3 个虚拟机的管理信息以及该第一虚拟机的配置信息发送给该第二虚拟机。该第二虚拟机在接收到该信息后，根据 3 个虚拟机的管理信息，对该第二虚拟机保存的虚拟机的管理信息进行维护。

具体地，该第二虚拟机可以比较 3 个虚拟机的管理信息和该第二虚拟机

保存的 3 个虚拟机的管理信息，确定该第二虚拟机保存的虚拟机的管理信息中是否有需要更新的虚拟机。由于 VM 301 管理信息中的启动次数为 2，心跳值为 8，而该第二虚拟机保存的 VM 301 的管理信息中的启动次数为 2，心跳值为 2。该第二虚拟机可以确定 VM 301 管理信息比在该第二虚拟机保存的管理信息更新。此外该第二虚拟机还可以确定出 VM 304 的管理信息和 VM 307 的管理信息，而该第二虚拟机保存的虚拟机的管理信息中没有 VM3204 的管理信息和 VM 307 的管理信息。

该第二虚拟机在确定出 VM 301 管理信息比在该第二虚拟机保存的管理信息更新的情况下，将该保存的 VM 301 的管理信息更新为 VM 301 的管理信息。同时，该第二虚拟机可以保存 VM 304 和 VM 307 的管理信息。

进一步，该第二虚拟机还可以确定出该第二虚拟机在接收到 VM 302 的管理信息和 VM 303 的管理信息且 VM 302 的管理信息和 VM 303 的管理信息。该第二邻居虚拟机可以将 VM 302 的管理信息和 VM 303 的管理信息发送给该第一虚拟机，以便于该第一虚拟机保存 VM 302 的管理信息和 VM 303 的管理信息。

该第二虚拟机可以根据第一虚拟机的配置信息对该第二虚拟机的配置信息进行维护。在本具体实施例中，该第二虚拟机确定出该第一虚拟机的配置信息中的 ClusterConf 小于该第一虚拟机的配置信息中的 ClusterConf，则该第二虚拟机可以确定该第二虚拟机的配置信息中的 ClusterConf 更新。该第二邻居虚拟机确定该第一虚拟机的配置信息中的 NodeInf 大于该第二虚拟机的配置信息中的 NodeInf，则该第二虚拟机可以确定该第一虚拟机的配置信息中的 NodeInf 更新。在此情况下，该第二虚拟机可以保持该第二虚拟机的配置信息中的 ClusterConf 不变，将该第二虚拟机的配置信息中的 NodeInf 更新为该第一虚拟机的配置信息中的 NodeInf。

该第二虚拟机在确定该第二虚拟机的配置信息比该第一虚拟机的配置信息更新的情况下，可以将该第二虚拟机的配置信息发送给该第一虚拟机。该第一虚拟机在接收到该第二虚拟机的配置信息后，可以根据该第二虚拟机的配置信息对该第一虚拟机的配置信息进行更新。该第一虚拟机更新配置信息的过程与该第二虚拟机更新配置信息的过程类似，在此就不必赘述。

在通过上述过程后，该第一虚拟机所维护的虚拟机的管理信息中的内容与该第二虚拟机所维护的虚拟机的管理信息中的内容完全相同，并且该第一

虚拟机的配置信息也与该第二虚拟机完全相同。类似的，集群 300 中的每一个虚拟机都可以与相应的虚拟机进行上述过程。这样，在不需要主节点的情况下，集群 300 中的每一个虚拟机都保存有其他虚拟机的管理信息，并且所有虚拟机的配置信息也是相同的。因此，可以避免主节点主机发生宕机导致的恢复时间过长以及集群脑裂的问题。

上面详细描述了根据本发明的实施例的处理虚拟机集群的方法。下面结合具体的例子分别描述根据本发明的实施例的虚拟机集群的初始建立、监测和管理的过程。

在根据本发明的实施例的虚拟机集群的初始建立过程中，每个虚拟机需要获知虚拟机集群中有哪些虚拟机，以及这些虚拟机的相关信息，然后在此基础上确定哪些虚拟机可以成为自己的邻居。

在本实施例中，初始建立的虚拟机列表可以包括状态信息（例如，中央处理器（Center Process Unit, CPU）的使用量和内存的使用量等信息）、管理信息（心跳值和启动次数等信息）和配置信息（例如，虚拟机的 IP 地址等信息）。例如，初始建立虚拟机集群后，虚拟机列表如表 1 所示。

表 1

虚拟机 ID	心跳值	启动次数	CPU 的使用量	内存的使用量	IP 地址	…
101	50	1	10%	10%	190.158.101.2	…
102	10	2	15%	5%	190.158.101.10	…
103	45	1	5%	5%	190.158.102.8	…
104	…	…	…	…	…	…
105	…	…	…	…	…	…
106	…	…	…	…	…	…
…	…	…	…	…	…	…

本发明的实施例可以通过广播方式、人工配置方式以及索引服务器方式初始建立虚拟机集群。

在采用广播方式建立虚拟机集群时，每个虚拟机可以广播 Hello 消息，其它虚拟机接收到 Hello 消息后向该虚拟机返回 Hello 报文的确认消息，以便在两个虚拟机之间交互状态信息、管理信息和/或配置信息。

图 4 是根据本发明的实施例的建立虚拟机集群的过程的示意性流程图。

参见图 4，采用广播方式建立虚拟机集群的具体过程如下。

410，虚拟机 1 广播 HelloMessage（你好报文），该 HelloMessage 包含该

虚拟机的的信息，例如，状态信息、管理信息和/或配置信息。

420，虚拟机 2 接收到虚拟机 1 发送的 HelloMessage，从该 HelloMessage 中读取虚拟机 1 的信息，并将虚拟机 1 添加到本地虚拟机列表中，即在本地虚拟机列表中记录虚拟机 1 的 ID 和虚拟机 1 的状态信息、管理信息和/或配置信息。

430，虚拟机 2 向虚拟机 1 返回确认消息 AckMessage（确认报文），该 AckMessage 包含有虚拟机 2 保存的虚拟机的信息。

440，虚拟机 1 接收到来自虚拟机 2 的 AckMessage，从该 AckMessage 中提取虚拟机 2 保存虚拟机的信息，并添加到本地保存的虚拟机列表。

450，虚拟机 1 向虚拟机 2 返回 Ack2Message（确认 2 报文），该 Ack2Message 包含本地添加的虚拟机的信息，以确认与虚拟机 2 间的消息同步。

460，虚拟机 1 可以根据该虚拟机列表，采用特定的邻居关系算法计算出自己的邻居并与之建立邻接关系。

各个虚拟机可以在建立虚拟机集群时在本地生成虚拟机列表。然后，各个虚拟机可以根据该虚拟机列表，采用特定的邻居关系算法计算出自己的邻居并与之建立邻接关系。例如，各个虚拟机可以在初始建立的虚拟机列表中进一步添加其邻居虚拟机的信息。各个虚拟机可以根据表 1 的虚拟机列表，并采用特定的邻居关系算法计算邻居虚拟机，可以是每个虚拟机将自己确定的邻居虚拟机通知给其它虚拟机，也可以是每个虚拟机采用相同的算法根据虚拟机列表直接计算出所有虚拟机的邻居虚拟机。例如，确定邻居关系之后，虚拟机列表如表 2 所示。

表 2

虚拟机 ID	心跳值	启动次数	CPU 的使用量	内存的使用量	IP 地址	邻居虚拟机
101	50	1	10%	10%	190.158.101.2	102、105、106、103
102	10	2	15%	5%	190.158.101.10	101、104、105
103	45	1	5%	5%	190.158.102.8	101、104、105
104	…	…	…	…	…	…
105	…	…	…	…	…	…
106	…	…	…	…	…	…
…	…	…	…	…	…	…

通过以上的过程，虚拟机集群中每个虚拟机都可以获知集群中其它虚拟

机的存在，以及它们的状态信息、管理信息和/或配置信息，从而在每个虚拟机本地形成一个虚拟机集群的全局信息列表。这种方式的优点是整个集群的构建过程完全是自组织的，减少了配置的环节。

应理解，当采用 P2P 协议构建虚拟机集群时，可以采用 P2P 协议中的 Gossip 协议。通过 Gossip 协议进行交互，每个 P2P 节点可以知道所有其他节点，也可能仅知道几个邻居节点，只要这些节点可以通过网络连通，最终他们的状态都是一致的。

可替代地，作为另一实施例，在采用人工配置方式建立虚拟机集群时，可以由管理员通过配置工具在每个虚拟机上配置所有虚拟机的状态信息、管理信息和/或配置信息。这种方式的优点是，虚拟机集群中的每个虚拟机可以快速获知虚拟机集群的状态信息、管理信息和/或配置信息。

可替代地，作为另一实施例，在采用索引服务器方式建立虚拟机集群时，索引服务器相当于一个注册中心，虚拟机集群中所有虚拟机在启动和初始加入集群时，都将自身的状态信息、管理信息和/或配置信息注册到该索引服务器中，再从该索引服务器获取虚拟机集群中的其他虚拟机的状态信息、管理信息和/或配置信息。应理解，索引服务器可以是一个单独的物理节点或主机，也可以是虚拟机集群中的一个虚拟机，如第一台加入集群的虚拟机。

应理解，在建立虚拟机集群的过程中，在虚拟机集群中的每个虚拟机逐个加入虚拟机集群的情况下，每个虚拟机的邻居关系可以是动态变化的，例如，第二个加入的虚拟机可以将第一个加入的虚拟机作为邻居，而第三个加入虚拟机可以将第二个加入的虚拟机作为邻居，而随着加入虚拟机的增多，第三个加入的虚拟机可能选择其它虚拟机作为邻居。

图 5 是根据本发明的实施例的心跳检测机制的示意图。

采用上述实施例的方法构建虚拟机集群后，每个虚拟机也就确立了各自的邻居关系，通常一个虚拟机会有 2~6 个邻居虚拟机，虚拟机之间会通过快速心跳机制相互进行监测。这样，对于其中任一虚拟机，都会有多个邻居虚拟机对其同时进行检测，一旦该虚拟机发生故障，其多个邻居虚拟机会同时检测到，即所谓的多点检测机制。多点检测的好处在于可以通过空间换时间的理念，缩短故障检测的时间。例如，参见图 5，VM305 的邻居虚拟机 302、304、306、308 均在预定的时间（例如，可以大于一个发送心跳消息的周期且小于两次发送心跳消息的周期）之后未能检测到 VM305 发送的心跳

消息，则认为虚拟机 VM305 故障或离开。

为了提高检测的准确性，避免误报，传统的心跳检测机制，检测节点需要多次确认机制才能判定被检测节点是否故障，例如，检测节点连续丢失 3 个心跳才认为被检测节点故障。而根据本发明的实施例的采用多点同时检测和空间换时间的机制，则可以仅通过一次心跳的丢失即可判定故障。由于是多点同时检测，所以可以有效避免误报，缩短检测时间的同时还可以确保高的准确性。

图 6 是根据本发明的实施例的虚拟机间同步的示意性流程图。

在本发明实施例的虚拟机集群中，没有中心节点的概念，虚拟机集群中的虚拟机列表及其它有关状态信息、管理信息和/或配置信息需要虚拟机集群中的每台虚拟机来维护，并通过虚拟机相互间的同步来达到数据的一致性。而虚拟机间的状态信息、管理信息和/或配置信息的同步可以有两种方式。

1）周期性定时同步

虚拟机集群中的虚拟机按照规定的时间间隔或周期进行状态信息、管理信息和/或配置信息的同步。

2）事件触发式同步

虚拟机集群中的虚拟机只有在虚拟机的状态信息、管理信息和/或配置信息有变化时才触发状态信息、管理信息和/或配置信息的同步。

例如，虚拟机 1 发起的与虚拟机 2 的消息同步过程包括如下内容。

610，虚拟机 1 向虚拟机 2 发送同步消息，

具体而言，该同步消息可以为 SynMessage（同步报文）。SynMessage 中可以包含虚拟机 1 中保存的虚拟机列表的更新内容的信息，例如，可以将虚拟机列表中更新内容通过 SynMessage 发送给虚拟机 2，也可以将更新内容对应的索引通过 SynMessage 发送给虚拟机 2。在本发明的实施例中，虚拟机 2 可以为虚拟机 1 的邻居虚拟机。虚拟机 1 可以向其所有邻居虚拟机发送 SynMessage。

620，虚拟机 2 根据同步消息更新本地的虚拟机列表。

例如，如果同步消息中包含的是更新的状态信息、管理信息和/或配置信息，则虚拟机 2 在接收到同步消息后，将同步消息中包含的更新的状态信息、管理信息和/或配置信息与本地保存的状态信息、管理信息和/或配置信息进行比较，确定哪个虚拟机上保存的状态信息、管理信息和/或配置信息更新，

如果确定虚拟机 1 保存的状态信息、管理信息和/或配置信息更新，则根据该更新的状态信息、管理信息和/或配置信息更新本地保存的虚拟机列表。该同步消息可以为 SynMessage。

630，虚拟机 2 向虚拟机 1 返回确定消息。

如果本地保存的状态信息、管理信息和/或配置信息更新，则将本地保存的状态信息、管理信息和/或配置信息再通过确认消息反馈给虚拟机 1。该确认消息可以为 AckMessage。

应理解，也可以是两个虚拟机的虚拟机列表中各有一部分虚拟机的状态信息、管理信息和/或配置信息更新。

640，虚拟机 1 根据确定消息更新本地的虚拟机列表。

当虚拟机 1 接收到该 AckMessage 后，以根据 Ackmessage 中包含的虚拟机 2 保存的状态信息、管理信息和/或配置信息，更新本地的虚拟机列表。

650，虚拟机 1 向虚拟机 2 发送确定消息。

可替代地，如果同步消息中包含的是更新的状态信息、管理信息和/或配置信息的指示信息(例如，更新的状态信息、管理信息和/或配置信息的索引)，则虚拟机 2 在接收到同步消息后，会将接收到的更新的状态信息、管理信息和/或配置信息的指示信息与本地保存的状态信息、管理信息和/或配置信息的指示信息进行比较，确定哪个虚拟机上保存的状态信息、管理信息和/或配置信息更新，如果确定虚拟机 1 保存的状态信息、管理信息和/或配置信息更新，则向虚拟机 1 发送确认消息，该确认消息包含虚拟机 2 请求的虚拟机 1 上更新的状态信息、管理信息和/或配置信息的指示信息，用于请求虚拟机 1 向虚拟机 2 发送更新的状态信息、管理信息和/或配置信息。该确认消息是 Ack2Message，虚拟机 2 接收到 Ack2Message 后，可以将虚拟机 1 上保存的更新的状态信息、管理信息和/或配置信息在本地更新。

应理解，当本发明的实施例采用 P2P 协议时，可以由一个 Gossip 进程来实现消息同步功能。例如，两个虚拟机可以通过 GossipSynMessage、GossipAckMessage 和 GossipAck2Message 交互状态信息、管理信息和/或配置信息。

上面的例子详细描述了根据本发明的实施例的虚拟机集群的初始建立和监测的过程，下面详细描述根据本发明的实施例的虚拟机集群的其它管理的过程。

虚拟机集群的管理包括虚拟机集群初始建、新虚拟机的加入，虚拟机的故障或重启、虚拟机的迁移、虚拟机的离开等等。

新虚拟机的加入过程与虚拟机集群初始建立过程类似，新加入的虚拟机也可以通过广播、人工配置和索引服务器三种方式获得虚拟机集群的全局信息，即全部虚拟机的状态信息、管理信息和/或配置信息，然后通过新加入虚拟机内部的邻居关系算法找到自己的邻居，并将自己加入到虚拟机集群中。

当某个虚拟机故障时，可以首先重启该虚拟机，如果重启不成功，则可以在其它物理主机上重建该虚拟机，即迁移该虚拟机。

因此，在检测到某个虚拟机故障时，为了防止该虚拟机重启成功后再次加入虚拟机集群而造成的虚拟机集群结构的震荡问题，需要在检测到该虚拟机故障时等待一段时间(例如，虚拟机重启的时间)，如果该故障虚拟机重启成功，则仍然回到虚拟机集群中的原来位置，这个整个集群不需要重新配置或同步。即当虚拟机故障或重启时，不会将该虚拟机从虚拟机列表删除。此时虚拟机集群中的虚拟机间的故障检测机制(如心跳)，仍然会对该虚拟机进行持续的监测，以便确定该虚拟机是否已恢复。只有当该虚拟机重启不成功时，才会触发整个集群的重新配置或同步。

当某个虚拟机离开虚拟机集群时，其它虚拟机通过同步过程可以获知该虚拟机离开，从而更新各自的本地虚拟机列表。

进一步，为了便于虚拟机管理和提高虚拟机的通讯效率，可以考虑将虚拟机集群中的虚拟机进行分类，通常可以分为下面的三类：种子虚拟机、普通虚拟机和不可达虚拟机。

种子虚拟机的作用主要是为新加入虚拟机集群的虚拟机提供一个初始的虚拟机列表。不可达虚拟机是指通过虚拟机间的检测机制发现临时不可达的那些虚拟机，包括虚拟机故障、重启中等。普通虚拟机为虚拟机集群中除种子虚拟机和不可达虚拟机之外的虚拟机。

如果采用上述虚拟机分类方式，通讯方式也可以随之改变，即由原来随机选取虚拟机（默认为 3 个），改为从上述三类虚拟机中各随机选取一个。这样做的好处是，可以确保种子虚拟机的存活性，以便于新加入集群的虚拟机获得初始的虚拟机列表，同时又可以对那些由于各种原因而临时不可达的虚拟机保持监测，以便在它们恢复后能被其它虚拟机及时获知。例如，当某个虚拟机新加入集群，可以首先从种子虚拟机获取一个虚拟机列表，此后该

虚拟机可以根据该虚拟机列表与其它虚拟机同步状态信息、管理信息和/或配置信息。

在不对虚拟机集群中的虚拟机进行上述分类的情况下，新加入集群的虚拟机可以通过发送一个广播消息从其它虚拟机获取节点信息，重新建立自己本地的节点列表。虚拟机集群中其它的虚拟机收到此广播消息后，可以向新加入的虚拟机返回自己的状态信息、管理信息和/或配置信息，同时更新本地的状态信息、管理信息和/或配置信息。

图 7 是根据本发明的实施例的一种计算机系统 700 的结构示意图。计算机系统包括至少一个计算机节点的物理硬件层，在至少一个计算机节点的物理硬件层之上运行虚拟机集群，虚拟机集群包括 N 个虚拟机，所述 N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，所述虚拟机列表包括所述 N 个虚拟机的信息，所述第一虚拟机包括:

发送模块 710，用于向所述 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便所述至少两个邻居虚拟机检测所述第一心跳消息，其中所述第一心跳消息的检测结果用于确定所述第一虚拟机的状态，所述第一虚拟机根据所述 N 个虚拟机的信息与所述至少两个邻居虚拟机建立了邻居关系;

接收模块 720，用于检测所述至少两个虚拟机中的第二虚拟机发送的第二心跳消息，其中所述第一虚拟机为所述第二虚拟机的邻居虚拟机，所述第二心跳消息的检测结果和所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果用于确定所述第二虚拟机的状态。

根据本发明的实施例，虚拟机集群中的每个虚拟机可以根据其保存虚拟机列表确定至少两个邻居虚拟机，虚拟机集群中的每个虚拟机可以向其至少两个邻居虚拟机发送心跳消息以便根据至少两个邻居虚拟机的检测结果确定该虚拟机的状态。由于每个虚拟机的状态均可以由其邻居虚拟机检测该虚拟机发送的心跳信息的结果来确定，因此避免了主/从结构存在的 Master 成为整个集群的瓶颈问题，同时由于不会出现重新选举 Master 主机的情况，因此，使得采用这种方案进行状态确定的虚拟机集群不会造成故障恢复时间的延迟和争抢资源的情况，因此，提高了虚拟机集群的容错能力和性能。

可选地，作为另一实施例，发送模块 710 还向至少两个邻居虚拟机发送第一同步信息，第一同步信息用于指示第一虚拟机中保存的虚拟机列表的更新，以便至少两个邻居虚拟机更新各自保存的虚拟机列表，接收模块 720 还

接收第二虚拟机发送的第二同步信息，第二同步信息用于指示第二虚拟机中保存的虚拟机列表的更新，第一虚拟机还包括更新模块 730，用于根据第二同步信息更新第一虚拟机保存的虚拟机列表。

可选地，作为另一实施例，至少两个邻居虚拟机包括 N 个虚拟机中具备与第一虚拟机直接交互信息的能力的二至六个虚拟机。

可选地，作为另一实施例，计算机系统 700 包括：触发模块 740，用于在确定所述第二虚拟机的状态为故障的情况下，触发所述第二虚拟机重启或触发所述第二虚拟机从所述第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

可选地，作为另一实施例，计算机系统 700 包括：触发模块 740，用于在确定所述第二虚拟机的状态为故障且无法重启的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

可选地，作为另一实施例，计算机系统 700 包括：触发模块 740，用于在确定所述第二虚拟机的状态为离开的情况下，触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

可选地，作为另一实施例，发送模块 710 还向 N 个虚拟机的上层节点发送第二心跳消息的检测结果，以便上层节点根据第二心跳消息的检测结果和第二虚拟机发送给第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定第二虚拟机的状态，接收模块 720 还接收上层节点发送的指示消息，指示消息用于指示第二虚拟机的状态。

可选地，作为另一实施例，接收模块 720 还接收第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果，接收模块 720 还根据第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果确定第二虚拟机的状态。

可选地，作为另一实施例，发送模块 710 还在第一虚拟机加入虚拟机集群时，向 N 个虚拟机中的其它虚拟机发送第一虚拟机的信息，接收模块 720 还接收 N 虚拟机中其它虚拟机发送的各自的信息，以生成虚拟机列表。

可选地，作为另一实施例，发送模块 710 还在第一虚拟机加入虚拟机集群时，向索引服务器发送注册信息，注册信息包括第一虚拟机的信息，其中所述索引服务器为所述虚拟机集群的注册中心，用于为所述虚拟机集群中的

虚拟机提供注册服务，接收模块 720 还接收索引服务器发送的 N 个虚拟机中的其它虚拟机的信息，以生成虚拟机列表。

根据本发明的实施例，选择模块 750 采用邻居关系算法，根据虚拟机列表中保存的 N 个虚拟机的信息，从 N 个虚拟机中选择至少两个作为邻居虚拟机。

根据本发明的实施例，N 个虚拟机的信息包括：N 个虚拟机中的每个虚拟机的状态信息、N 个虚拟机中的每个虚拟机的邻居关系信息、N 个虚拟机中的每个虚拟机的启动次数、N 个虚拟机中的每个虚拟机的心跳值、N 个虚拟机中的每个虚拟机的配置信息以及虚拟机集群的配置信息中的任意一个或多个的组合。

计算机系统 700 的各个部分的操作和功能可以参考上述图 3 的方法，为了避免重复，在此不再赘述。

图 8 是根据本发明的实施例的一种计算机系统 800 的结构示意图。计算机系统包括至少一个计算机节点的物理硬件层，在至少一个计算机节点的物理硬件层之上运行虚拟机集群，虚拟机集群包括 N 个虚拟机，N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，虚拟机列表包括 N 个虚拟机的信息，第一虚拟机包括：处理器 810，通过总线 840 调用和执行存储在存储器 850 中的代码；发送器 820，用于向 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，第一虚拟机根据 N 个虚拟机的信息与至少两个邻居虚拟机建立了邻居关系；接收器 830，用于检测至少两个虚拟机中的第二虚拟机发送的第二心跳消息，其中第一虚拟机为第二虚拟机的邻居虚拟机，第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果用于确定第二虚拟机的状态，第二虚拟机为 N 个虚拟机之一。

根据本发明的实施例，虚拟机集群中的每个虚拟机可以根据其保存虚拟机列表确定至少两个邻居虚拟机，虚拟机集群中的每个虚拟机可以向其至少两个邻居虚拟机发送心跳消息以便根据至少两个邻居虚拟机的检测结果确定该虚拟机的状态。由于每个虚拟机的状态均可以由其邻居虚拟机检测该虚拟机发送的心跳信息的结果来确定，因此避免了主/从结构存在的 Master 成为整个集群的瓶颈问题，同时由于不会出现重新选举 Master 主机的情况，因

此，使得采用这种方案进行状态确定的虚拟机集群不会造成故障恢复时间的延迟和争抢资源的情况，因此，提高了虚拟机集群的容错能力和性能。

可选地，作为另一实施例，发送器820还向至少两个邻居虚拟机发送第一同步信息，第一同步信息用于指示第一虚拟机中保存的虚拟机列表的更新，以便至少两个邻居虚拟机更新各自保存的虚拟机列表，接收器830还接收第二虚拟机发送的第二同步信息，第二同步信息用于指示第二虚拟机中保存的虚拟机列表的更新，处理器810还根据第二同步信息更新第一虚拟机保存的虚拟机列表。

可选地，作为另一实施例，至少两个邻居虚拟机包括N个虚拟机中具备与第一虚拟机直接交互信息的能力的二至六个虚拟机。

可选地，作为另一实施例，处理器810还用于在确定所述第二虚拟机的状态为故障的情况下，所述第二虚拟机触发所述第二虚拟机重启或触发所述第二虚拟机从所述第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

可选地，作为另一实施例，处理器810还用于所述第一虚拟机在确定所述第二虚拟机的状态为故障且无法重启的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

可选地，作为另一实施例，处理器810还在确定所述第二虚拟机的状态为离开的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

可选地，作为另一实施例，发送器820还向N个虚拟机的上层节点发送第二心跳消息的检测结果，以便上层节点根据第二心跳消息的检测结果和第二虚拟机发送给第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定第二虚拟机的状态，接收器830还接收上层节点发送的指示消息，指示消息用于指示第二虚拟机的状态。

可选地，作为另一实施例，接收器830还接收第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果，接收模块730还根据第二心跳消息的检测结果和第二虚拟机的其它邻居虚拟机检测第二虚拟机发送的心跳消息得到的检测结果确定第二虚拟机的状态。

可选地，作为另一实施例，发送器820还在第一虚拟机加入虚拟机集群

时，向 N 个虚拟机中的其它虚拟机发送第一虚拟机的信息，接收器 830 还接收 N 虚拟机中其它虚拟机发送的各自的信息，以生成虚拟机列表。

可选地，作为另一实施例，发送器 820 在第一虚拟机加入虚拟机集群时，向索引服务器发送注册信息，注册信息包括第一虚拟机的信息，其中所述索引服务器为所述虚拟机集群的注册中心，用于为所述虚拟机集群中的虚拟机提供注册服务，接收器 830 还接收索引服务器发送的 N 个虚拟机中的其它虚拟机的信息，以生成虚拟机列表。

根据本发明的实施例，处理器 810 采用邻居关系算法，根据虚拟机列表中保存的 N 个虚拟机的信息，从 N 个虚拟机中选择至少两个作为邻居虚拟机。

根据本发明的实施例，N 个虚拟机的信息包括：N 个虚拟机中的每个虚拟机的状态信息、N 个虚拟机中的每个虚拟机的邻居关系信息、N 个虚拟机中的每个虚拟机的启动次数、N 个虚拟机中的每个虚拟机的心跳值、N 个虚拟机中的每个虚拟机的配置信息以及虚拟机集群的配置信息中的任意一个或多个的组合。

计算机系统 800 的各个部分的操作和功能可以参考上述图 3 的方法，为了避免重复，在此不再赘述。

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

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。

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）或处理器（processor）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，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内，因此本发明的保护范围应以权利要求的保护范围为准。

权利要求

1、一种处理虚拟机集群的方法，其特征在于，所述虚拟机集群包括 N 个虚拟机，所述 N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，所述虚拟机列表包括所述 N 个虚拟机的信息，所述方法包括:

所述第一虚拟机向所述 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便所述至少两个邻居虚拟机检测所述第一心跳消息，其中所述第一心跳消息的检测结果用于确定所述第一虚拟机的状态，所述第一虚拟机根据所述 N 个虚拟机的信息与所述至少两个邻居虚拟机建立了邻居关系;

所述第一虚拟机作为所述至少两个虚拟机中的第二虚拟机的邻居虚拟机，检测所述第二虚拟机发送的第二心跳消息，其中所述第二心跳消息的检测结果和所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果用于确定所述第二虚拟机的状态。

2、如权利要求 1 所述的方法，其特征在于，所述方法还包括:

所述第一虚拟机向所述至少两个邻居虚拟机发送第一同步信息，所述第一同步信息用于指示所述第一虚拟机中保存的所述虚拟机列表的更新，以便所述至少两个邻居虚拟机更新各自保存的虚拟机列表;

所述第一虚拟机接收所述第二虚拟机发送的第二同步信息，所述第二同步信息用于指示所述第二虚拟机中保存的虚拟机列表的更新;

所述第一虚拟机根据所述第二同步信息更新所述第一虚拟机保存的所述虚拟机列表。

3、如权利要求 1 所述的方法，其特征在于，所述至少两个邻居虚拟机包括所述 N 个虚拟机中具备与所述第一虚拟机直接交互信息的能力的二至六个虚拟机。

4、根据权利要求 1 至 3 中的任一个所述方法，其特征在于，还包括:

所述第一虚拟机在确定所述第二虚拟机的状态为故障的情况下，所述第二虚拟机触发所述第二虚拟机重启或触发所述第二虚拟机从所述第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

5、根据权利要求 1 至 3 中的任一项所述方法，其特征在于，还包括:

所述第一虚拟机在确定所述第二虚拟机的状态为故障且无法重启的情

况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

6、根据权利要求 2 至 5 中的任一项所述的方法，其特征在于，还包括:

所述第一虚拟机在确定所述第二虚拟机的状态为离开的情况下，所述第一虚拟机触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

7、根据权利要求 1 至 6 中的任一项所述的方法，其特征在于，所述方法还包括:

所述第一虚拟机向所述 N 个虚拟机的上层节点发送所述第二心跳消息的检测结果，以便所述上层节点根据所述第二心跳消息的检测结果和所述第二虚拟机发送给所述第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定所述第二虚拟机的状态;

所述第一虚拟机接收所述上层节点发送的指示消息，所述指示消息用于指示第二虚拟机的状态。

8、根据权利要求 1 至 6 中的任一项所述的方法，其特征在于，所述方法还包括:

所述第一虚拟机接收所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果;

所述第一虚拟机根据所述第二心跳消息的检测结果和所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果确定所述第二虚拟机的状态。

9、如权利要求 1 至 8 中任一项所述的方法，其特征在于，在所述第一虚拟机加入所述虚拟机集群时，所述方法还包括:

所述第一虚拟机向所述 N 个虚拟机中的其它虚拟机发送所述第一虚拟机的信息;

所述第一虚拟机接收所述 N 虚拟机中其它虚拟机发送的各自的信息，以生成所述虚拟机列表。

10、如权利要求 1 至 8 中的任一项所述的方法，其特征在于，在所述第一虚拟机加入所述虚拟机集群时，所述方法还包括:

所述第一虚拟机向索引服务器发送注册信息，所述注册信息包括所述第一虚拟机的信息，其中所述索引服务器为所述虚拟机集群的注册中心，用于为所述虚拟机集群中的虚拟机提供注册服务;

所述第一虚拟机接收所述索引服务器发送的所述 N 个虚拟机中的其它虚拟机的信息，以生成所述虚拟机列表。

11、如权利要求 1 至 10 中的任一项所述的方法，其特征在于，还包括:

所述第一虚拟机采用邻居关系算法，根据所述虚拟机列表中保存的所述 N 个虚拟机的信息，从所述 N 个虚拟机中选择至少两个作为邻居虚拟机。

12、如权利要求 1 至 11 中的任一项所述的方法，其特征在于，所述 N 个虚拟机的信息包括: 所述 N 个虚拟机中的每个虚拟机的状态信息、所述 N 个虚拟机中的每个虚拟机的邻居关系信息、所述 N 个虚拟机中的每个虚拟机的启动次数、所述 N 个虚拟机中的每个虚拟机的心跳值、所述 N 个虚拟机中的每个虚拟机的配置信息以及所述虚拟机集群的配置信息中的任意一个或多个的组合。

13、一种计算机系统，其特征在于，所述计算机系统包括至少一个计算机节点的物理硬件层，在所述至少一个计算机节点的物理硬件层之上运行虚拟机集群，所述虚拟机集群包括 N 个虚拟机，所述 N 个虚拟机中的每个虚拟机保存虚拟机列表，并以对等方式作为第一虚拟机进行工作，所述虚拟机列表包括所述 N 个虚拟机的信息，所述第一虚拟机包括:

发送模块，用于向所述 N 个虚拟机中的至少两个邻居虚拟机发送第一心跳消息，以便所述至少两个邻居虚拟机检测所述第一心跳消息，其中所述第一心跳消息的检测结果用于确定所述第一虚拟机的状态，所述第一虚拟机根据所述 N 个虚拟机的信息与所述至少两个邻居虚拟机建立了邻居关系;

接收模块，用于检测所述至少两个虚拟机中的第二虚拟机发送的第二心跳消息，其中所述第一虚拟机为所述第二虚拟机的邻居虚拟机，所述第二心跳消息的检测结果和所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果用于确定所述第二虚拟机的状态。

14、如权利要求 13 所述的计算机系统，其特征在于，所述发送模块还向所述至少两个邻居虚拟机发送第一同步信息，所述第一同步信息用于指示所述第一虚拟机中保存的所述虚拟机列表的更新，以便所述至少两个邻居虚拟机更新各自保存的虚拟机列表，所述接收模块还接收所述第二虚拟机发送的第二同步信息，所述第二同步信息用于指示所述第二虚拟机中保存的虚拟机列表的更新，所述第一虚拟机还包括更新模块，用于根据所述第二同步信息更新所述第一虚拟机保存的所述虚拟机列表。

15、如权利要求 13 或 14 所述的方法，其特征在于，所述至少两个邻居虚拟机包括所述 N 个虚拟机中具备与所述第一虚拟机直接交互信息的能力的二至六个虚拟机。

16、根据权利要求 13 至 15 中的任一项所述计算机系统，其特征在于，还包括:

触发模块，用于在确定所述第二虚拟机的状态为故障的情况下，触发所述第二虚拟机重启或触发所述第二虚拟机从所述第二虚拟机的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

17、根据权利要求 13 至 15 中的任一项所述计算机系统，其特征在于，还包括:

触发模块，用于在确定所述第二虚拟机的状态为故障且无法重启的情况下，触发所述第二虚拟机从所述第二虚拟机所在的源主机迁移至目标主机，其中所述源主机为故障主机，所述目标主机为正常主机。

18、根据权利要求 13 至 17 中的任一项所述的计算机系统，其特征在于，还包括:

触发模块，用于在确定所述第二虚拟机的状态为离开的情况下，触发所述第二虚拟机从所述第二虚报机所在的源主机上删除。

19、根据权利要求 13 至 18 中的任一项所述的计算机系统，其特征在于，所述发送模块还向所述 N 个虚拟机的上层节点发送所述第二心跳消息的检测结果，以便所述上层节点根据所述第二心跳消息的检测结果和所述第二虚拟机发送给所述第二虚拟机的其它邻居虚拟机的心跳消息的检测结果确定所述第二虚拟机的状态，所述接收模块还接收所述上层节点发送的指示消息，所述指示消息用于指示第二虚拟机的状态。

20、根据权利要求 13 至 18 中的任一项所述的计算机系统，其特征在于，所述接收模块还接收所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果，所述接收模块还根据所述第二心跳消息的检测结果和所述第二虚拟机的其它邻居虚拟机检测所述第二虚拟机发送的心跳消息得到的检测结果确定所述第二虚拟机的状态。

21、如权利要求 13 至 20 中任一项所述的计算机系统，其特征在于，所述发送模块还在所述第一虚拟机加入所述虚拟机集群时，向所述 N 个虚拟机中的其它虚拟机发送所述第一虚拟机的信息，所述接收模块还接收所述 N 虚

拟机中其它虚拟机发送的各自的信息，以生成所述虚拟机列表。

22、如权利要求 13 至 20 中的任一项所述的计算机系统，其特征在于，所述发送模块还在所述第一虚拟机加入所述虚拟机集群时，向索引服务器发送注册信息，所述注册信息包括所述第一虚拟机的信息，其中所述索引服务器为所述虚拟机集群的注册中心，用于为所述虚拟机集群中的虚拟机提供注册服务，所述接收模块还接收所述索引服务器发送的所述 N 个虚拟机中的其它虚拟机的信息，以生成所述虚拟机列表。

23、如权利要求 13 至 22 中的任一项所述的计算机系统，其特征在于，还包括:

选择模块，用于采用邻居关系算法，根据所述虚拟机列表中保存的所述 N 个虚拟机的信息，从所述 N 个虚拟机中选择至少两个作为邻居虚拟机。

24、如权利要求 13 至 23 中的任一项所述的计算机系统，其特征在于，所述 N 个虚拟机的信息包括: 所述 N 个虚拟机中的每个虚拟机的状态信息、所述 N 个虚拟机中的每个虚拟机的邻居关系信息、所述 N 个虚拟机中的每个虚拟机的启动次数、所述 N 个虚拟机中的每个虚拟机的心跳值、所述 N 个虚拟机中的每个虚拟机的配置信息以及所述虚拟机集群的配置信息中的任意一个或多个的组合。

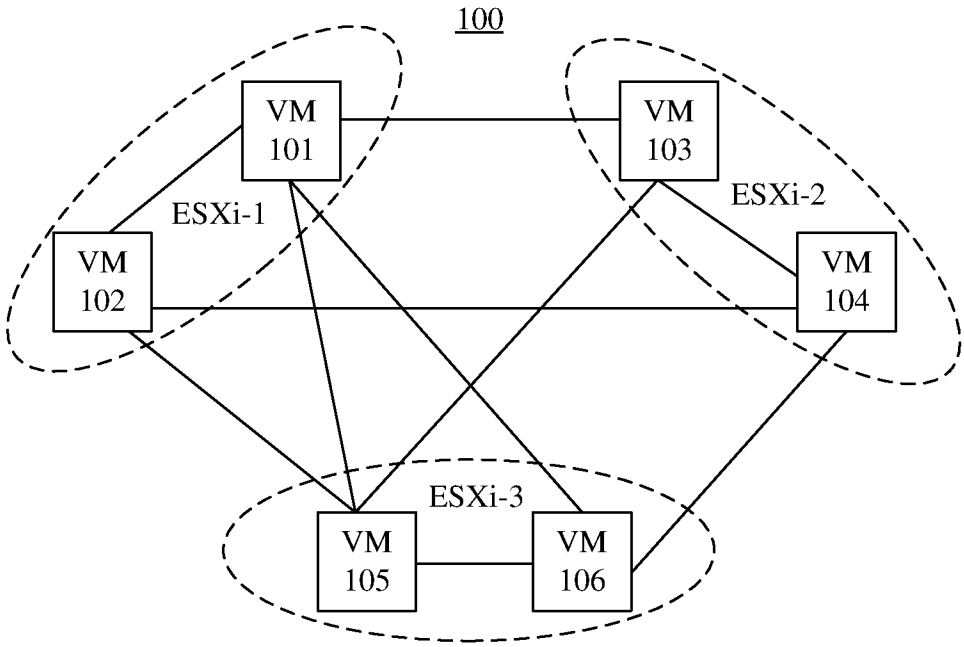
100
VM
101
VM
103
ESXi-1
ESXi-2
VM
102
VM
104
ESXi-3
VM
105
VM
106

图 1

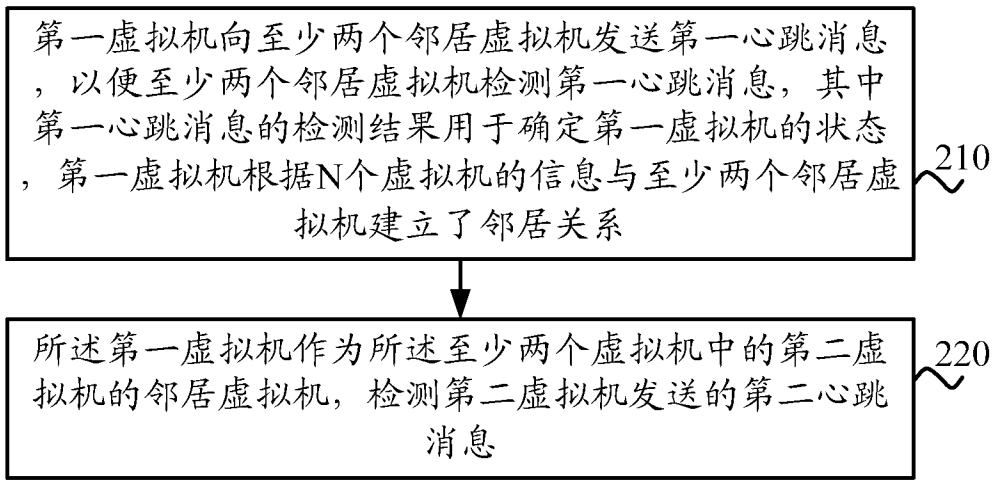
第一虚拟机向至少两个邻居虚拟机发送第一心跳消息，以便至少两个邻居虚拟机检测第一心跳消息，其中第一心跳消息的检测结果用于确定第一虚拟机的状态，第一虚拟机根据N个虚拟机的信息与至少两个邻居虚拟机建立了邻居关系
210
所述第一虚拟机作为所述至少两个虚拟机中的第二虚拟机的邻居虚拟机，检测第二虚拟机发送的第二心跳消息
220

图2

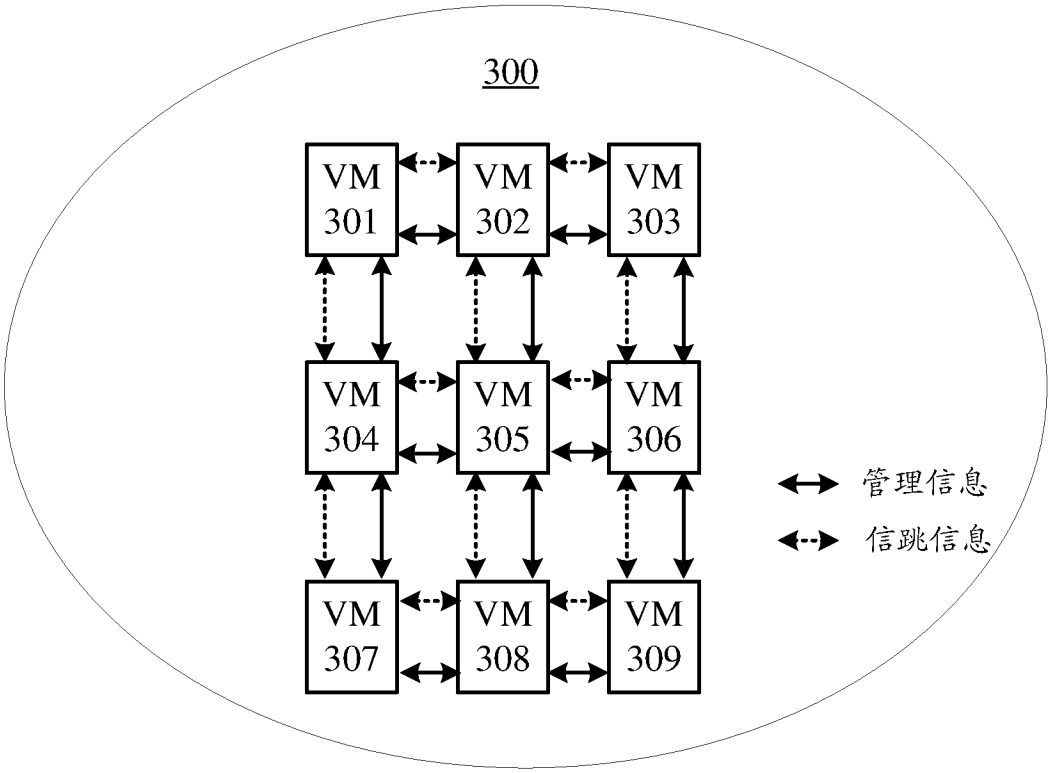
300
VM 301
VM 302
VM 303
VM 304
VM 305
VM 306
VM 307
VM 308
VM 309
管理信息
信跳信息

图3

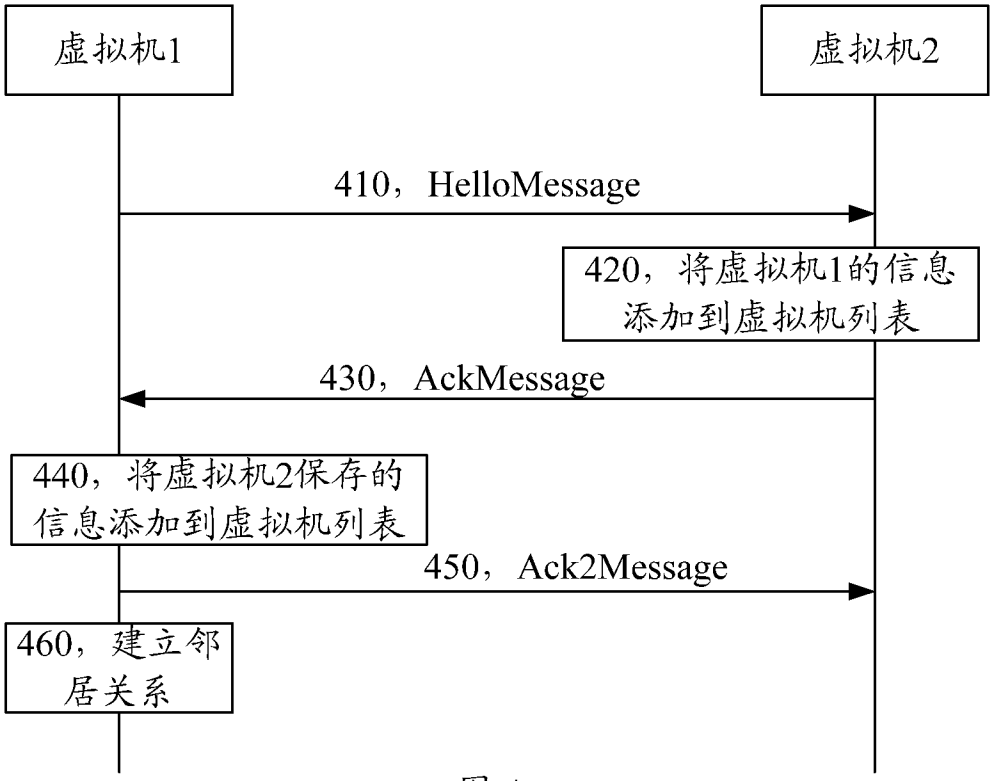
虚拟机1
虚拟机2
410，HelloMessage
420，将虚拟机1的信息添加到虚拟机列表
430，AckMessage
440，将虚拟机2保存的信息添加到虚拟机列表
450，Ack2Message
460，建立邻居关系

图 4

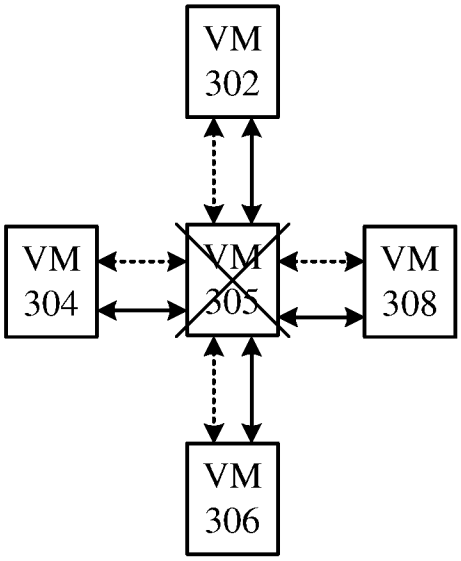
VM
302
VM
304
VM
305
VM
308
VM
306

图5

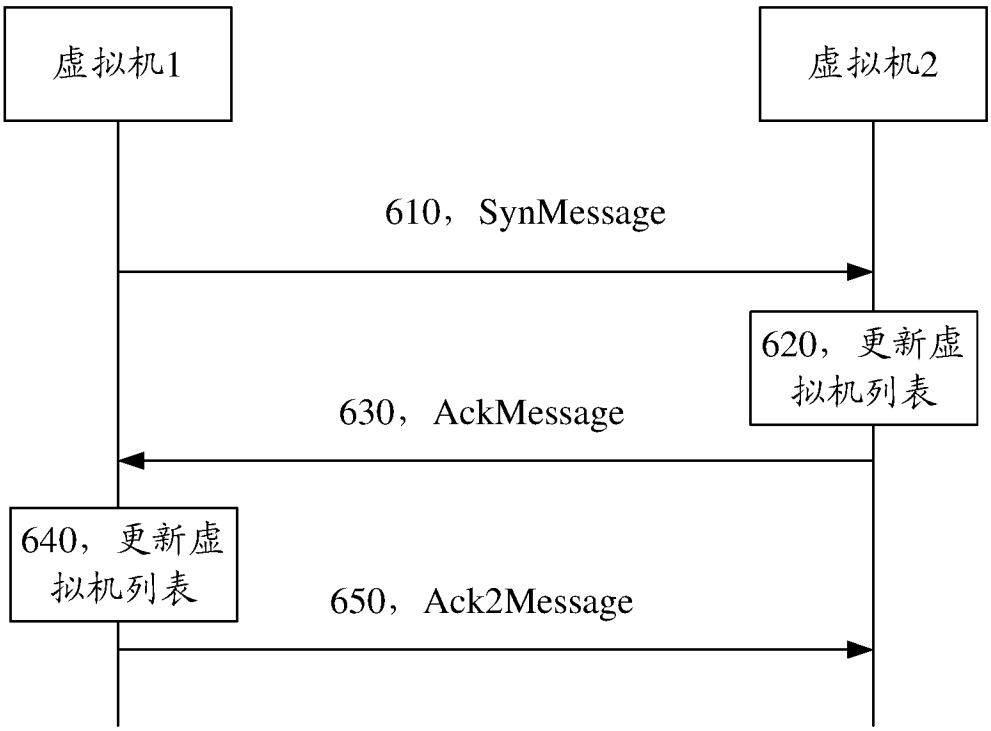
虚拟机1
虚拟机2
610，SynMessage
620，更新虚拟机列表
630，AckMessage
640，更新虚拟机列表
650，Ack2Message

图 6

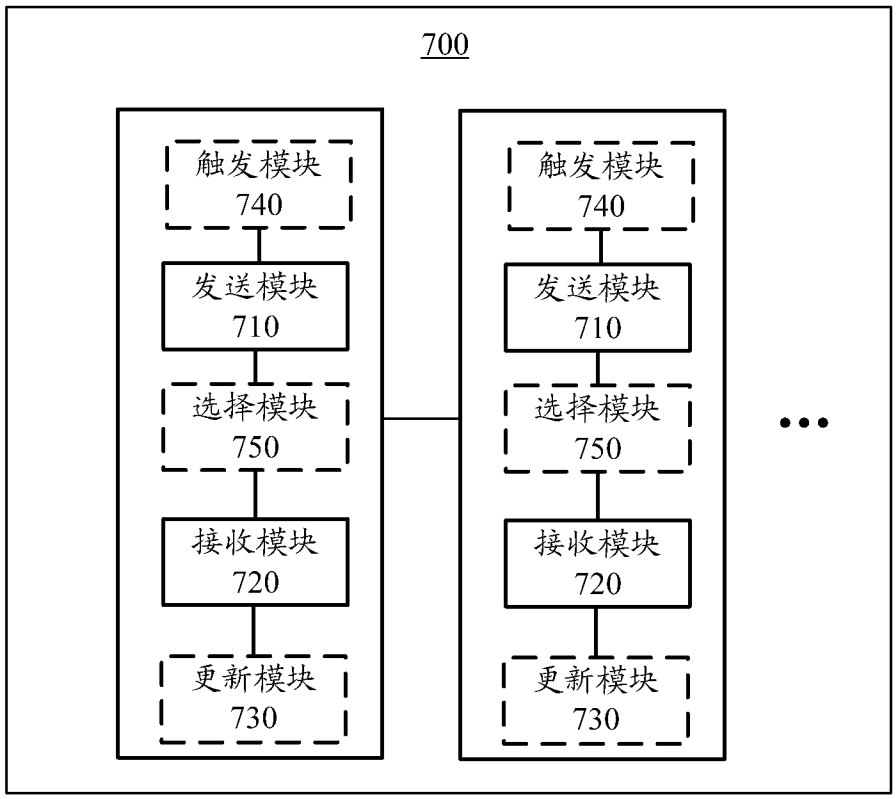
700
触发模块
740
发送模块
710
选择模块
750
接收模块
720
更新模块
730
触发模块
740
发送模块
710
选择模块
750
接收模块
720
更新模块
730

图 7

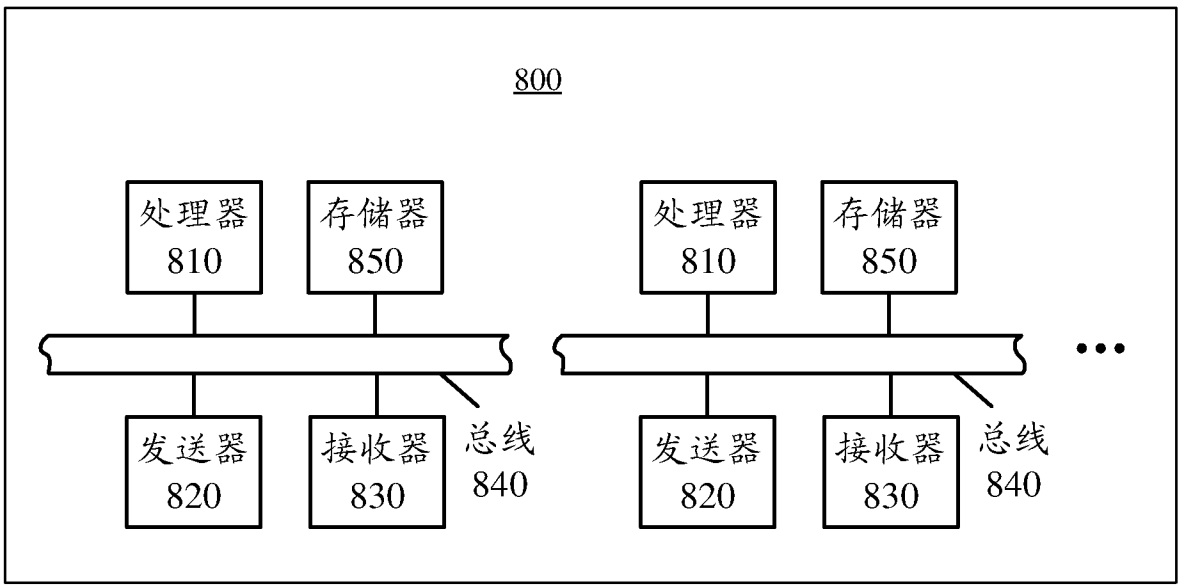
800
处理器
810
存储器
850
发送器
820
接收器
830
总线
840
处理器
810
存储器
850
发送器
820
接收器
830
总线
840

图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/095654

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/24 (2006.01) i; H04L 12/40 (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

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)

CNABS; CPRSABS; CNTXT; VEN; CNKI: cluster, virtual machine, group, each, multiple, monitor, detect, poll, heartbeat, list, update, move, transfer, restart

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104601622 A (INTERNATIONAL BUSINESS MACHINES CORPORATION), 06 May 2015 (06.05.2015), description, paragraph [0004], and figure 4	1-24
A	CN 102110071 A (INSPUR (BEIJING) ELECTRONIC INFORMATION INDUSTRY CO., LTD.), 29 June 2011 (29.06.2011), the whole document	1-24
A	CN 102455951 A (CHINA STANDARD SOFTWARE CO., LTD.), 16 May 2012 (16.05.2012), the whole document	1-24
A	CN 103201724 A (SYMANTEC CORPORATION), 10 July 2013 (10.07.2013), the whole document	1-24

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"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)

"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

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 December 2015 (24.12.2015)	**26 January 2016 (26.01.2016)**
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 **DENG, Lu** Telephone No.: (86-10) **62089138**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/095654

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104601622 A	06 May 2015	US 2015120887 A1	30 April 2015
CN 102110071 A	29 June 2011	CN 102110071 B	17 April 2013
CN 102455951 A	16 May 2012	None	
CN 103201724 A	10 July 2013	EP 2598993 A1	05 June 2013
		US 8424000 B2	16 April 2013
		US 2012030670 A1	02 February 2012
		WO 2012016175 A1	02 February 2012
		JP 2013535745 A	12 September 2013

国际检索报告

国际申请号
PCT/CN2015/095654

A. 主题的分类

H04L 12/24(2006.01)i; H04L 12/40(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

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））

CNABS;CPRSABS;CNTXT;VEN;CNKI:虚拟机，集群，每个，多个，监测，检测，轮询，心跳，列表，更新，迁移，重启，virtual machine, group, each, multiple, monitor, detect, poll, heartbeat, list, update, move, transfer, restart

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
X	CN 104601622 A (国际商业机器公司) 2015年 5月 6日 (2015-05-06) 说明书第[0004]段，图4	1-24
A	CN 102110071 A (浪潮北京电子信息产业有限公司) 2011年 6月 29日 (2011-06-29) 全文	1-24
A	CN 102455951 A (中标软件有限公司) 2012年 5月 16日 (2012-05-16) 全文	1-24
A	CN 103201724 A (赛门铁克公司) 2013年 7月 10日 (2013-07-10) 全文	1-24

☐ 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	国际检索报告邮寄日期
2015年 12月 24日	2016年 1月 26日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	邓璐
传真号 (86-10)62019451	电话号码 (86-10)62089138

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/095654

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	104601622	A	2015年 5月 6日	US	2015120887	A1	2015年 4月 30日
CN	102110071	A	2011年 6月 29日	CN	102110071	B	2013年 4月 17日
CN	102455951	A	2012年 5月 16日		无		
CN	103201724	A	2013年 7月 10日	EP	2598993	A1	2013年 6月 5日
				US	8424000	B2	2013年 4月 16日
				US	2012030670	A1	2012年 2月 2日
				WO	2012016175	A1	2012年 2月 2日
				JP	2013535745	A	2013年 9月 12日