

(19) 中华人民共和国国家知识产权局



(12) 发明专利申请

(10) 申请公布号 CN 104508651 A

(43) 申请公布日 2015.04.08

(21) 申请号 201380040826.4

代理人 李晓冬

(22) 申请日 2013.07.30

(51) Int. Cl.

(30) 优先权数据

G06F 15/163(2006.01)

61/677,867 2012.07.31 US

13/828,664 2013.03.14 US

(85) PCT国际申请进入国家阶段日

2015.01.30

(86) PCT国际申请的申请数据

PCT/US2013/052652 2013.07.30

(87) PCT国际申请的公布数据

W02014/022350 EN 2014.02.06

(71) 申请人 F5 网络公司

地址 美国华盛顿州

(72) 发明人 安东尼·金 保罗·I·斯扎伯

威廉·罗斯·鲍曼

(74) 专利代理机构 北京东方亿思知识产权代理

有限责任公司 11258

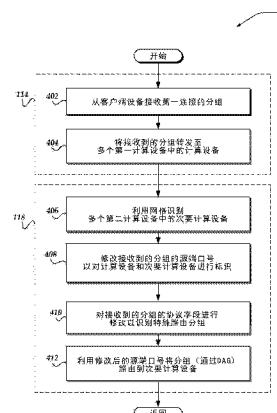
权利要求书2页 说明书9页 附图4页

(54) 发明名称

镜像非对称集群多处理器系统中的连接网

(57) 摘要

实施例针对在主要机架中的多个第一计算设备中的每个第一计算设备与故障转移机架中的多个第二计算设备中的每个计算设备之间建立多个链接。第一计算设备将多个链接作为网格连接来选择要路由关于接收到的分组的信息的第二计算设备。将关于分组的信息路由至所选择的第二计算设备包括将分组中的源端口号修改为包括第一计算设备的标识符和第二计算设备的标识符。该信息可以指示故障转移机架要对修改后的分组执行特殊路由。



1. 一种系统，包括：

主要机架，所述主要机架具有能够被配置为运行指令以执行动作的一个或多个处理器，所述动作包括：

从客户端设备接收分组；

选择所述主要机架的多个第一计算设备中的第一计算设备；以及

将所述分组转发至所选择的第一计算设备；以及

所述第一计算设备具有至少一个处理器，所述至少一个处理器执行如下动作，包括：

在所述第一计算设备与故障转移机架内的多个第二计算设备中的每个第二计算设备之间建立网格连接；

针对所转发的分组来识别所述多个第二计算设备内的镜像计算设备；

对每个分组的字段进行修改以对所述第一计算设备和所述镜像计算设备进行标识，以使得所述故障转移机架基于修改后的字段来将所述分组路由至所述镜像计算设备；以及

将关于修改后的分组的信息转发给所述故障转移机架。

2. 如权利要求 1 所述的系统，其中，所述主要机架包括与所述故障转移机架不同数目的计算设备。

3. 如权利要求 1 所述的系统，其中，所述分组的所述修改后的字段为使用散列将所述第一计算设备的标识符与所述镜像计算设备的标识符进行组合的修改后的源端口号或互联网协议 (IP) 地址。

4. 如权利要求 1 所述的系统，其中，所述分组的头部被修改为包括指示修改后的字段被包括在所述分组内的标记。

5. 如权利要求 1 所述的系统，其中，与所述主要机架相关联的解聚器 (DAG) 被用来基于所述主要机架的健康状态而选择所述第一计算设备。

6. 如权利要求 1 所述的系统，其中，从所述客户端设备接收的所述分组由所述故障转移机架中的所述镜像计算设备转发至目的地服务器设备。

7. 如权利要求 1 所述的系统，其中，所述镜像计算设备被配置为将响应分组提供给所述第一计算设备，其中，所述分组包括修改后的端口号，所述修改后的端口号以与来自所述第一计算设备的分组的端口号中使用的顺序相反的顺序将镜像计算设备标识符与第一计算设备标识符进行组合。

8. 一种非暂态处理器可读存储介质，所述介质存储处理器可读指令，当所述指令被处理器运行时执行如下动作，包括：

在主要机架内的多个第一计算设备内的第一计算设备与故障转移机架内的多个第二计算设备中的每个第二计算设备之间建立网格连接；

针对转发分组来识别所述多个第二计算设备内的镜像计算设备；

对所述分组的端口号进行修改以对所述第一计算设备和所述镜像计算设备进行标识，以使得所述故障转移机架基于修改后的端口号来将所述分组路由至所述镜像计算设备；以及

将关于修改后的分组的信息转发给所述故障转移机架。

9. 如权利要求 8 所述的非暂态处理器可读存储介质，其中，所述修改后的端口号为修改后的目的地端口号，所述修改后的目的地端口号基于所述第一计算设备的标识符与所述

第二计算设备的标识符的组合的散列而被计算。

10. 如权利要求 8 所述的非暂态处理器可读存储介质, 其中, 所述镜像计算设备被配置为将响应分组提供给所述第一计算设备, 其中, 所述分组包括修改后的端口号, 所述修改后的端口号以与来自所述第一计算设备的分组的端口号中使用的顺序相反的顺序将镜像计算设备标识符与第一计算设备标识符进行组合。

11. 如权利要求 8 所述的非暂态处理器可读存储介质, 其中, 与所述主要机架相关联的解聚器 (DAG) 被用来基于所述主要机架的健康状态而选择所述第一计算设备。

12. 如权利要求 8 所述的非暂态处理器可读存储介质, 其中, 故障转移机架中的所述多个第二计算设备中的每个第二计算设备的健康状态被用来识别所述镜像计算设备。

13. 如权利要求 8 所述的非暂态处理器可读存储介质, 其中, 与所述故障转移机架相关联的解聚器 (DAG) 被配置为 :接收所述修改后的分组, 并且基于所述分组的头部内的标记, 利用所述修改后的端口号来确定要路由所述分组的所述镜像计算设备。

14. 如权利要求 13 所述的非暂态处理器可读存储介质, 其中, 所述标记在所述分组头部的协议字段内, 并且其中, 所述协议字段包括指示所述分组是从所述主要机架到所述故障转移机架还是从所述故障转移机架到所述主要机架的附加信息。

15. 一种主要机架, 所述主要机架包括多个第一计算设备, 每个第一计算设备具有至少一个处理器, 所述至少一个处理器被配置为执行如下动作, 包括 :

在所述多个第一计算设备内的每个计算设备与故障转移机架内的多个第二计算设备中的每个第二计算设备之间建立网格连接 ;

从客户端设备在所述多个第一计算设备内的第一计算设备处接收分组 ;

识别所述多个第二计算设备内的镜像计算设备, 以转发接收到的分组 ;

对每个分组头部中的字段进行修改, 以对所述第一计算设备和所述镜像计算设备进行标识, 以使得所述故障转移机架基于修改后的字段来将所述分组路由至所述镜像计算设备 ;以及

由所述第一计算设备将修改后的分组转发给所述故障转移机架。

16. 如权利要求 15 所述的主要机架, 其中, 所述修改后的字段为修改后的源端口号, 所述修改后的源端口号基于所述第一计算设备的标识符与所述第二计算设备的标识符的组合的散列而被计算。

17. 如权利要求 15 所述的主要机架, 其中, 所述镜像计算设备被配置为将响应分组提供给所述第一计算设备, 其中, 所述分组包括修改后的端口号, 所述修改后的端口号以与来自所述第一计算设备的分组的端口号中使用的顺序相反的顺序将镜像计算设备标识符与第一计算设备标识符进行组合。

18. 如权利要求 15 所述的主要机架, 其中, 与所述主要机架相关联的解聚器 (DAG) 被用来基于所述主要机架的健康状态而选择所述第一计算设备。

19. 如权利要求 15 所述的主要机架, 其中, 故障转移机架中的所述多个第二计算设备中的每个第二计算设备的健康状态被用来识别所述镜像计算设备。

20. 如权利要求 15 所述的主要机架, 其中, 所述分组还被修改为在协议字段内包括指示所述字段为修改后的端口号的标记。

镜像非对称集群多处理器系统中的连接网

[0001] 相关申请的交叉引用

[0002] 本申请要求于 2013 年 3 月 14 日申请的、题为“镜像非对称集群多处理器系统中的连接网”的美国专利申请 No. 13/828,664 的优先权，该申请要求于 2012 年 7 月 31 日申请的、题为“镜像非对称集群多处理器系统中的连接网”的美国临时专利申请 No. 61/677,867 的权益，这两个申请通过引用被合并于此。

技术领域

[0003] 本发明实施例总体涉及网络通信，更具体而非排他地，涉及使用连接网将主要机架上的计算设备镜像到故障转移机架上的计算设备。

背景技术

[0004] 对于高可用性计算服务存在持久需求。包括关键任务应用的计算应用日益由数据中心来处理，尤其作为云计算架构而被包含。同时，整体的计算设备被替换为一个或多个机架，每个机架包含并行操作的花费较少的计算设备群组，例如刀片式服务器。

[0005] 机架的可用性通常通过镜像来改善。例如，主要机架可以由故障转移机架进行镜像，以使得在主要机架上出现设备故障（或任何其他错误）的情形中，故障转移机架能够接管主要机架的处理。然而，尽管机架可以作为单元而出现故障，但仍有可能主要机架中的一个或多个单独的计算设备出现故障，而其余的计算设备继续运行。而且，故障转移机架上的一个或多个计算设备可能出现故障。这些场景中的计算设备之间的镜像是不间断的问题。因此，针对这些考虑和其他考虑来对本发明实施例进行描绘。

附图说明

[0006] 参照以下附图对非限制且非穷尽的实施例进行了描述。在附图中，除非另有说明，否则相似的参考标号在各个附图中指代相似的部分。

[0007] 为更好地理解所描述的实施例，将参照下面详细的描述，这些描述将结合附图而被阅读，其中：

[0008] 图 1 示出所描述的实施例可以被实施于的说明性环境的部件；

[0009] 图 2 示出解聚器 (disaggregator) 设备的一个实施例；

[0010] 图 3 示出计算设备的一个实施例；以及

[0011] 图 4 示出一般示出用于使用连接网创建从主要机架至故障转移机架的连接的处理的一个实施例的逻辑流程图。

发明内容

[0012] 在下面示例性实施例的详细描述中，对附图进行参照，这些附图形成其中的一部分，并且通过所描述的实施例可以被实施的示意性示例的方式来示出。提供了充足的细节以使得本领域技术人员能够实施所描述的实施例，并且应当理解，在不背离精神或范围的

情况下,可以使用其他实施例,并且可以做出其他改变。而且,对“一个实施例”的引用不一定涉及相同或单一实施例,尽管其可以。因此,下面详细的描述不是在限制的意义上进行的,并且所描述的实施例的范围仅由所附权利要求进行限定。

[0013] 贯穿说明书和权利要求,以下术语采用与本文相关联的确切含义,除非上下文清楚地规定其他含义。如本文所使用的,除非上下文清楚地规定其他含义,否则术语“或”包括“或”操作符,并且等同于术语“和 / 或”。除非上下文清楚地规定其他含义,否则术语“基于”不是排他的,并且允许基于未描述的另外的因素。此外,贯穿说明书,泛指和特指“该”的含义包括复数引用。“在 … 中”的含义包括“在 … 中”和“在 … 上”。

[0014] 如本文所使用的,术语“网络连接”(也被称为“连接”)指代使得计算设备通过网络与另一计算设备进行通信的链路和 / 或软件元件的连接。一种这样的网络连接可以是传输控制协议 (TCP) 连接。TCP 连接是两个网络节点之间的虚拟连接,并且通常通过 TCP 握手协议来建立。在互联网工程任务组 (IETF) 的请求评论 (RFC) 793 中更加详细地描述了 TCP 协议,并且本文以其全文通过引用将其合并于此。“通过”特定路径或链路的网络连接指代使用具体路径或链路来建立和 / 或维护通信的网络连接。

[0015] 如本文所使用的,机架指代容纳多个物理计算设备(下文称为计算设备)的外壳。在一个实施例中,计算设备可以包括刀片式服务器,然而,同样考虑任意其他类型的计算设备。在一个实施例中,机架可以包括下面所定义的解聚器 (DAG)。

[0016] 如本文所使用的,解聚器 (DAG) 指代将传入 (incoming) 连接路由到多个计算设备之一的计算设备。在一个实施例中,DAG 可以基于散列 (hash) 算法和与传入连接相关联的一个或多个属性将传入连接路由到特定的计算设备。这些属性可以包括但不限于源端口号、目的地端口号、源 IP 地址、目的地 IP 地址、与连接相关联的一个或多个分组头部中的其他连接字段、等等。在一些实施例中,源端口号和目的地端口号可以分别包括 TCP 源端口号和 TCP 目的地端口号。例如,DAG 可以通过将传入连接的源 (远程) 端口和目的地 (本地) 端口进行散列来创建散列值。DAG 然后可以基于散列值到网格连接的预定映射以及网格连接与计算设备之间的关联来将传入连接路由至特定计算设备。将传入网络连接路由至特定计算设备的其他技术包括不同的散列算法、与传入连接相关联的不同属性、将散列值映射到网格连接的不同算法,并且同样考虑将网格连接映射到计算设备的不同技术。

[0017] 简略陈述,实施例针对在主要机架和故障转移机架之间创建网格连接来促进主要机架内的第一计算设备与故障转移机架内的第二计算设备之间的双向通信。主要机架可以包括多个第一计算设备,故障转移机架可以包括多个第二计算设备。在一些实施例中,主要机架和故障转移机架可以是非对称的,以使得主要机架内的多个计算设备可以与故障转移机架内的多个计算设备不同。在一些实施例中,可以在主要机架的每个主要计算设备与故障转移机架内的每个次要计算设备之间建立网格连接。

[0018] 在一些实施例中,来自客户端设备的第一连接的分组可以使用网格连接之一通过主要机架内的第一计算设备被路由到故障转移机架内镜像的第二计算设备。在一个实施例中,第一计算设备和第二计算设备可以使用修改后的分组头部来回地转发分组。在一个实施例中,修改后的分组头部可以包括修改后的源端口号,该修改后的源端口号标识了第一计算设备和第二计算设备。在一些实施例中,修改后的源端口号使用散列将第一计算设备标识符与第二计算设备标识符进行结合。第一计算设备可以将分组和 / 或关于分组的信息

转发给故障转移机架，故障转移机架可以使用修改后的源端口号来将分组转发给镜像第二计算设备。在一些实施例中，第二计算设备可以通过使用在第一修改后的源端口号的位置（或顺序）处的另一修改后的源端口号将分组和/或关于分组的信息转发回第一计算设备，其中，另一修改后的源端口号包括第二计算设备标识符和第一计算设备标识符。在一些其他实施例中，分组头部可以被修改为包括指示第一计算设备与第二计算设备之间的分组/信息流方向和/或修改后的源端口地址的标记。应当注意，除了或并非源端口号，分组头部中的其他信息可以被修改。例如，分组头部内的互联网协议（IP）地址（源和/或目的地）、（七层开放系统互连（OSI）模型的）二层、三层和/或四层数据可以被修改。

[0019] 在其他实施例中，并非将标记提供给 DAG 来指示要对分组执行特殊处理，而是可以通过使用用户数据报协议（UDP）帧针对镜像通道的每个方向封装 TCP 帧来创建网格连接。在这些实施例中，UDP 帧具有使得分组被发送至故障转移机架上的特定第二计算设备的源/目的地端口。连接上的返回分组也被封装有针对散列至该连接的另一端的源/目的地端口，在至少一个实施例中，这些源/目的地端口不一定与起初发送的端口集相同。仍可以使用其他机制，包括明确指定端口信息、使用针对从故障转移机架返回的流量的暂时端口号，其中，从初始暂时端口号来计算这些暂时端口号，如下面进一步所讨论的。

[0020] 当第二计算设备出现故障时，第一计算设备使用现有的且可用的网格连接选择故障转移机架内的另一次要（可用）计算设备。使用主要机架与故障转移机架中的计算机设备之间的现有且可用的网格连接是针对用于维护连接备份的快速故障转移操作的。

[0021] 说明性操作环境

[0022] 图 1 示出所描述的实施例可以被实施于的说明性环境 100 的部件。不需要所有的部件来实施所描述的实施例，并且可以在不背离所描述的实施例的精神或范围的情况下在部件的安排和类型方面做出变化。图 1 示出客户端设备 102-104、网络 108、服务器设备 105 以及主要机架 110 和次要机架 112。

[0023] 一般，客户端设备 102-104 可以虚拟地包括能够连接到另一计算设备并且发送和/或接收信息的任意计算设备。例如，客户端设备 102-104 可以包括个人计算机、多处理器系统、基于多处理器或可编程的消费类电子设备、网络设备、服务器设备、虚拟机等。客户端设备 102-104 还可以包括便携式设备（例如，蜂窝电话、智能手机、显示寻呼机、射频（RF）设备、红外（IR）设备、个人数字助理（PDA）、手持计算机、可穿戴计算机、平板计算机、将前面设备中的一个或多个进行组合的集成设备，等等）。客户端设备 102-104 还可以包括在超级管理器或一些其他虚拟化环境中运行的虚拟计算设备。据此，客户端设备 102-104 在能力和特征方面可以具有广泛范围。

[0024] 网络 108 被配置为将网络使能设备（例如，客户端设备 102-104 以及机架 110 和 112）与其他网络使能设备进行耦合。网络 108 能够实现使用任意形式的计算机可读介质以将信息从一个电子设备传输至另一电子设备。在一个实施例中，网络 108 可以包括互联网，并且可以包括局域网（LAN）、广域网（WAN）、直接连接（例如，通过通用串行总线（USB）端口、其他形式的计算机可读介质或其任意组合）。在互连的 LAN 集（包括基于不同架构和协议的那些 LAN）上，路由器可以作为 LAN 之间的链路来使得能够将消息从一个 LAN 发送至另一 LAN。另外，LAN 内的通信链路通常包括光纤、双绞线或同轴电缆，而网络之间的通信链路可以使用模拟电话线、包括 T1、T2、T3 和 T4 的全部或部分专用数字线、综合服务数字网络

(ISDN)、数字用户线 (DSL)、无线链路 (包括卫星链路) 或本领域技术人员熟知的其他通信链路。

[0025] 网络 108 还可以使用多种无线接入技术 (包括但不限于蜂窝系统的第二代 (2G)、第三代 (3G)、第四代 (4G) 无线电接入、无线 LAN、无线路由器 (WR) 网格, 等等)。诸如 2G、3G、4G 之类的接入技术以及未来的接入网络可以使得网络设备 (例如, 客户端设备 102–104 等) 能够实现具有各种程度的移动性的广域覆盖。例如, 网络 108 可以通过无线电网络接入 (例如, 全球移动通信系统 (GSM)、通用分组无线业务 (GPRS)、增强型数据 GSM 环境 (EDGE)、宽带码分多址接入 (WCDM), 等等) 实现无线电连接。

[0026] 而且, 远程计算机和其他相关的电子设备可以经由调制解调器和临时电话线、DSL 调制解调器、电缆调制解调器、光纤调制解调器、802.11(Wi-Fi) 接收机等被远程连接到 LAN 或 WAN。本质上, 网络 108 包括信息可以在一个网络设备与另一网络设备之间穿越的任意通信方法。

[0027] 服务器设备 105 可以包括能够将分组传输至另一网络设备 (例如但不限于, 机架设备 110 和 / 或 112、以及客户端设备 102–104 中的至少一个) 的任意计算设备。在一个实施例中, 服务器设备 105 可以被配置为作为网站服务器进行操作。然而, 服务器设备不限于 web 服务器设备, 并且还可以作为消息发送服务器、文件传输协议 (FTP) 服务器、数据库服务器、内容服务器等进行操作。尽管图 1 将服务器设备 105 示出为单个设备, 但本发明的实施例不限于此。例如, 服务器设备 105 可以包括多个不同的网络设备。在一些实施例中, 每个不同的网络设备可以被配置为执行不同的操作, 例如, 一个网络设备被配置为消息发送服务器, 而另一网络设备被配置为数据库服务器, 等等。

[0028] 可以作为服务器设备 105 进行操作的设备包括个人计算机、台式计算机、多处理器系统、基于多处理器或可编程消费类电子设备、网络 PC、服务器设备, 等等。

[0029] 在一些实施例中, 客户端设备 (例如, 客户端设备 102) 可以从服务器设备 105 请求内容或其他动作。如本文所公开的, 从客户端设备的这样的连接然后将通过主要机架 110 和 / 或故障转移机架 112 内的计算设备进行路由, 并且被转发至服务器设备 105。来自服务器设备 105 的响应将类似地通过主要机架 110 和 / 或故障转移机架 112 内的计算设备进行路由, 并且被转发至请求客户端设备。

[0030] 机架设备 110 和 112 中的每个机架设备可以包括 DAG 和多个计算设备。主要机架 110 包括 DAG 114 以及计算设备 118、120、122 和 124, 而故障转移机架 112 包括 DAG 116 以及计算设备 126、128 和 130。尽管图 1 示出故障转移机架 112 具有少于主要机架 110 的计算设备, 但也可以设想其他配置。例如, 在其他实施例中, 主要机架 110 和故障转移机架 112 可以包括相同数目的计算机设备, 或者主要机架 110 可以包括比故障转移机架 112 少的计算设备。因此, 考虑了各种配置和安排。

[0031] 如图所示, 机架 110 内的计算设备可以打开并维护与机架 112 内的每个可用计算设备的连接。这样的连接可以被配置以形成连接网。例如, 所示出的网格连接 158 示出了从计算设备 118 与计算设备 126、128 和 130 中的每个的连接。类似地, 网格连接 154 示出了从计算设备 124 与计算设备 126、128 和 130 中的每个的连接。未示出的 (为附图的简单起见) 计算设备 120 和 122 可以包括类似的网格连接。在一些实施例中, 这些网格连接是双向的, 以使得消息和其他信息可以由主要机架 110 或故障转移机架 112 中的计算设备

进行发送。

[0032] 如下面进一步所讨论的,计算机设备 128 被示出为灰色以表示故障转移状况。在该情景中,从主要机架 110 中的每个计算设备至出现故障的计算设备 128 的连接将变为不可操作(通过连接上的“X”来示出)。

[0033] 尽管图 1 示出遮盖 DAG 和多个计算设备的每个机架,但在另一实施例中,机架和 / 或机架内的部件之一可以是虚拟设备。例如,虚拟机架可以将物理 DAG 与多个物理计算设备进行关联。替代地,多个计算设备中的一个或多个可以是与物理 DAG 进行通信的虚拟机,并且由虚拟机架进行关联。在一些实施例中,DAG 114 和 DAG 116 的功能可以由 L2 交换硬件、网络处理单元 (NPU) 或其他计算设备(例如,图 2 的 DAG 设备 200) 中的现场可编程门阵列 (FPGA)、专用集成电路 (ASIC) 来实现和 / 或在其上被运行。

[0034] 计算设备 118、120、122、124、126、128 和 130 中的每个可以包括一个或多个处理器核(未示出)。在一个实施例中,每个处理器核作为独立的计算设备进行操作。例如,包括 4 核的计算设备可以作为 4 个独立的计算设备进行操作,并且由 DAG 看作是 4 个独立的计算设备。因此,贯穿该公开,对计算设备的任何引用还指代在计算设备上运行的许多核之一。在一个实施例中,计算设备可以被设计为作为单元而出现故障。在该实施例中,特定计算设备的故障可以使得被包括在该计算设备中的所有处理器核出现故障。

[0035] 在一些其他实施例中,计算设备 118、120、122、124、126、128 和 130 中的每个可以包括独立的 DAG。在一个这样的实施例中,每个 DAG 可以与一个或多个计算设备相对应。在一些实施例中,组合的计算设备和 DAG 可以共享处理器核或使用独立的处理器核来执行下面将更加详细描述的计算设备和 DAG 的动作。

说明性解聚器设备环境

[0037] 图 2 示出解聚器 (DAG) 设备的一个实施例。DAG 设备 200 可以包括多于或少于所示出的部件的部件。然而,所示出的部件足以公开说明性实施例。举例来说,DAG 设备 200 可以表示图 1 的 DAG 114 或 DAG116。然而,本发明不限于此,并且 FPGA、ASIC、L2 交换硬件、NPU 等可以用于 DAG(例如,图 1 的 DAG 114 或 DAG 116) 的功能。

[0038] DAG 设备 200 包括中央处理单元 212、视频显示适配器 214 以及大型存储器,所有这些经由总线 222 进行相互通信。大型存储器一般包括随机存取存储器 (RAM) 216、只读存储器 (ROM) 232 以及一个或多个永久大型存储设备(例如,硬盘驱动器 228、磁带驱动器、压缩盘 ROM (CD-ROM) / 数字通用盘 ROM (DVD-ROM) 驱动器 226 和 / 或软盘驱动器)。硬盘驱动器 228 可以被用来存储(除了其他之外)由 DAG 路由的连接的状态、DAG 被遮盖在其中或者与其相关联的机架的健康状态等。大型存储器存储用于控制 DAG 设备 200 的操作的操作系统 220。还提供基本的输入 / 输出系统 (“BIOS”) 218 来控制 DAG 设备 200 的低层操作。DAG 设备 200 还包括解聚模块 252。

[0039] 如图 2 所示,DAG 设备 200 还可以经由网络接口单元 210 与互联网或一些其他通信网络进行通信,网络接口单元 210 被构建为利用各种通信协议(包括 TCP/IP 协议)进行使用。网络接口单元 210 有时被认为是收发器、收发设备或网络接口卡 (NIC)。

[0040] DAG 设备 200 还可以包括用于与外部设备(例如,鼠标、键盘、扫描仪或图 2 中未示出的其他输入 / 输出设备)进行通信的输入 / 输出接口 224。

[0041] 上面所描述的大型存储器说明了另一种计算机可读介质,被称为计算机存储介

质。计算机存储介质可以包括以用于存储信息的任意方法或技术（例如，计算机可读指令、数据结构、程序模块或其他数据）来实现的易失性介质、非易失性介质、可移除介质以及不可移除介质。计算机存储介质的示例包括 RAM、ROM、电可擦除可编程只读存储器 (EEPROM)、闪存或其他存储器技术、CD-ROM、DVD 或其他光存储设备、磁盒存储设备、磁带存储设备、磁盘存储设备或其他磁存储设备、或者可以被用来存储预期的信息并且可以由计算设备进行访问的任意其他非暂态介质。

[0042] 大型存储器还存储程序代码和数据。解聚模块 252 被加载到大型存储器中并且在操作系统 220 上运行。在一个实施例中，解聚模块 252 可以通过与主要计算设备的连接来接收分组，并且使用包括故障转移机架地址的修改后的目的地地址和修改后的源端口号将分组转发给次要计算设备。下面将结合图 4 对解聚模块 252 的进一步细节进行讨论。

[0043] 说明性计算机设备环境

[0044] 图 3 示出计算设备的一个实施例。计算设备 300 可以包括比所示的部件更多的部件。然而，所示出的部件足以公开用于实施这些实施例的说明性实施例。举例来说，计算设备 300 可以表示图 1 的计算设备 118、120、122、124、126、128 和 130 之一。

[0045] 计算设备 300 包括中央处理单元 312、视频显示适配器 314 以及大型存储器，所有这些经由总线 322 进行相互通信。大型存储器一般包括 RAM 316、ROM 332 以及一个或多个永久大型存储设备（例如，硬盘驱动器 328、磁带驱动器、CD-ROM/DVD-ROM 驱动器 326 和 / 或软盘驱动器）。大型存储器存储用于控制服务器设备 300 的操作的操作系统 320。还提供 BIOS 318 来控制计算设备 300 的低层操作。如图 3 所示，计算设备 300 还可以经由网络接口单元 310 与互联网或一些其他通信网络进行通信，网络接口单元 310 被构建为利用各种通信协议（包括 TCP/IP 协议）进行使用。网络接口单元 310 有时被认为是收发器、收发设备或网络接口卡 (NIC)。

[0046] 计算设备 300 还可以包括用于与外部设备（例如，鼠标、键盘、扫描仪或图 3 中未示出的其他输入设备）进行通信的输入 / 输出接口 324。

[0047] 上面所描述的大型存储器说明了另一种计算机可读介质，被称为计算机存储介质。计算机存储介质可以包括以用于存储信息的任意方法或技术（例如，计算机可读指令、数据结构、程序模块或其他数据）来实现的易失性介质、非易失性介质、可移除介质以及不可移除介质。计算机存储介质的示例包括 RAM、ROM、电可擦除可编程只读存储器 (EEPROM)、闪存或其他存储器技术、CD-ROM、DVD 或其他光存储设备、磁盒存储设备、磁带存储设备、磁盘存储设备或其他磁存储设备、或者可以被用来存储预期的信息并且可以由计算设备进行访问的任意其他非暂态介质。

[0048] 连接创建模块 350 可以被加载到大型存储器中，并且在操作系统 320 上运行。在一个实施例中，连接创建模块 350 可以创建至另一机架（例如，故障转移机架）的连接。在一个实施例中，连接创建模块 350 可以创建具有属性的网格连接，以使得其他机架的 DAG 将连接路由到与特定网格连接相关联的计算设备。参照图 4 对连接创建进行更加详细的讨论。

[0049] 在一个实施例中，计算设备 300 包括至少一个耦合于总线 322 的专用集成电路 (ASIC) 芯片（未示出）。ASIC 芯片可以包括执行计算设备 300 的一些动作的逻辑。例如，在一个实施例中，ASIC 芯片可以对传入和 / 或传出 (outgoing) 分组执行多个分组处理功能。在一个实施例中，ASIC 芯片可以执行该逻辑的至少一部分以实现连接创建模块 350 的

操作。

[0050] 在一个实施例中,计算设备 300 还可以包括一个或多个现场可编程门阵列 (FPGA) (未示出),而非 ASIC 芯片或除了 ASIC 芯片。计算设备的多个功能可以由 ASIC 芯片、FPGA、具有存储于存储器中的指令的 CPU 312 或者由 ASIC 芯片、FPGA 和 CPU 的任意组合来执行。

[0051] 一般操作

[0052] 现将参照图 4 对某些方面的操作进行描述。图 4 示出一般示出用于管理从主要机架到故障转移机架的网格连接的处理的一个实施例的逻辑流程图。在一个实施例中,处理 400 可以由图 1 的机架 110 来实现。在另一实施例中,框 402 和 404 可以由图 1 的 DAG 114 来实现,而框 406、408 和 410 可以由图 1 的计算设备 118、120、122 或 124 之一来实现,并且框 412 可以由图 1 的 DAG 116 来实现,尽管处理 400 和框 402、404、406、408、410 和 412 中的一个或多个框可以由 DAG 114、116、计算设备 118、120、122、124、126 和 130 的不同组合来执行。

[0053] 在开始框之后,处理 400 在框 402 处开始,在框 402 处,在一个实施例中,从客户端设备(例如,图 1 的客户端设备 102-104 之一)接收网络分组。该网络分组可以指向服务器设备(例如,图 1 的服务器设备 105)。

[0054] 在框 404 处,DAG 选择主要机架中的主要计算设备之一来管理接收到的分组。管理接收到的分组包括主要计算设备将分组路由到故障转移机架中的计算设备以备用,尽管 DAG 可以将分组路由到故障转移机架中的计算设备。DAG 还将分组转发给服务器设备,尽管在其他实施例中,主要机架或故障转移机架中的计算设备可以将分组转发给服务器设备。

[0055] 在一个实施例中,每个 DAG 可以维护相关联的机架的健康状态。在一个实施例中,健康状态是位串,其中,每位表示多个计算设备之一的健康状态。在一个实施例中,DAG 将健康状态位串用作进入针对给定健康状态将连接映射到计算设备的表格的索引。在一个实施例中,如果四个计算设备(如图 1 所示)全部正在操作,则机架的健康状态可以为 1111(假设 1 意思是可操作,0 意思是不可操作)。在一个实施例中,健康状态可以包括所有的解聚状态,例如包括刀片健康状态和所使用的解聚算法。而且,尽管健康状态信息可以是 1111,但在其他实施例中,其还可以更加复杂地指示暂时状态或永久状态。而且,健康状态可以包括表格或其他结构化信息,所述表格或其他结构化信息还提供相关联的机架及其计算设备的状态。在一些实施例中,在任何情况下,可以使用该健康状态信息来选择选择主要计算设备。DAG 然后可以将接收到的分组转发给所选择的主要计算设备。

[0056] 处理流动至下一框 406,框 406 可以由所选择的主要计算设备来执行。应当注意,在进行处理 400 之前和/或继续处理 400,每个主要计算设备建立并且维护与每个可用次要计算设备的网格连接。在一个实施例中,确定次要计算设备是否可用可以基于从相应 DAG 接收到的信息(例如,从故障转移机架的健康状态信息)而进行。在其他实施例中,可以在与次要计算设备的连接出现故障、超时或不能被建立时确定次要计算设备的可用性。

[0057] 因此,在框 406 处,主要计算设备知道哪些网格连接是可用的。然后主要计算设备在框 406 处识别多个第二计算设备中的哪个镜像计算设备来路由分组。

[0058] 流动至下一框 408,主要计算设备对接收到的分组的源端口号进行修改以标识主要计算设备和次要计算设备,尽管主要计算设备还可以或者代替地对接收到的分组的目的地端口号进行修改以标识主要计算设备和次要计算设备和/或对其他分组字段(例如,源

IP 地址和 / 或目的地 IP 地址、MAC 地址等) 进行修改。在一个实施例中, 修改后的源端口号可以是散列, 该散列包括主要计算设备标识符和次要计算设备标识符。这些标识符可以标识机架内的特定刀片和 / 或处理器和 / 或该机架上针对该刀片 / 处理器的特定端口。

[0059] 而且, 在一些实施例中, 主要计算设备可以对目的地地址 / 端口号进行修改, 以指示分组被指向故障转移 DAG。

[0060] 流动至下一框 410, 在一个实施例中, 分组头部内的字段可以被修改为指示接收 DAG 将基于修改后的源端口号而识别出用于特殊处理的分组。在一个实施例中, 该字段可以是分组头部中的协议字段。然而, 也可以使用其他字段或字段的组合。

[0061] 在一个实施例中, 该字段可以包括指示分组要流向哪个方向 (例如, 从主要机架到故障转移机架, 或者从故障转移机架到主要机架) 的信息。

[0062] 处理移至框 412, 其中, 在一个实施例中, 修改后的分组被路由到故障转移 DAG, 其中, 故障转移 DAG 基于修改后的协议字段识别出用于特殊处理的分组。故障转移 DAG 然后使用修改后的源端口号中的信息将修改后的分组路由至次要计算设备。然而, 在另一实施例中, 关于分组的信息而非分组本身可以被路由至故障转移 DAG。在又一实施例中, 修改后的分组和关于这些分组的信息均被提供给故障转移 DAG。

[0063] 来自次要计算设备的响应基于修改后的源端口信息被返回至发起主要计算设备。在一个实施例中, 次要计算设备可以将协议字段 (或一个或多个其他字段) 修改为指示分组将被特殊处理的另一标识符。在一些实施例中, 原始的源端口信息被维护在例如数据存储设备中。在一些实施例中, 原始的源端口信息可以通过将字节插入分组中进行维护, 或者通过利用原始的源端口信息来覆写分组内未使用的字段。

[0064] 在任何情况下, 处理 400 可以返回至另一处理。

[0065] 尽管上面的处理 400 公开了基于修改后的协议字段来使用特殊处理, 但其他实现方式也被考虑。例如, 在另一实施例中, 可以通过利用 UDP 帧对镜像通道的每个方向封装 TCP 帧来创建网格连接, 其中, UDP 帧包括使得分组被发送至故障转移机架内的特定次级计算设备的源 / 目的地端口信息。还可以对来自次级计算设备的返回分组进行封装, 并且可以对将要散列至主要机架中预期的计算设备的源 / 目的地端口信息进行选择。

[0066] 然而, 可以采用其他实现方式。例如, 在其他实施例中, 并非采用修改后的协议字段, 而是上面讨论的 DAG 的特殊规则可以由特殊配置的虚拟局域网 (VLAN) 或其他网络字段 (包括 magic 端口、IP 地址, 等等) 来触发。

[0067] 在又一实施例中, 并非使用 UDP 来在计算设备之间建立两个 (或多个) 单向管道, 而是可以对 TCP 端口号进行修改以允许执行路由。例如, {源 / 暂时端口号, 目的地端口号} 可以最初由主要计算设备进行选择, 以将分组发送至次要计算设备。来自次要计算设备的返回分组然后可以包括修改后的目的地 (或源) TCP 端口号, 以允许分组基于 TCP 端口号的散列而去往主要计算设备。返回端口数据可以由主要计算设备嵌入到同步 (SYN) 分组中或者被发送到带外。将端口信息嵌入到 SYN 分组中可以使用 TCP 序列号、可选的时间戳字段或分组内的一些其他约定字段来完成。主要计算设备然后可以使用 (一个或多个) 返回端口号来创建流, 以从次级计算设备接收分组。类似地, 次级计算设备将使用 (一个或多个) 约定返回端口号在该流上发送分组。

[0068] 在其他实施例中, 并非明确指定 (一个或多个) 端口号, 而是计算设备被配置为同

意返回流量将使用从来自主要计算设备的初始暂时端口号计算得到的（一个或多个）临时端口号。例如，可以对 SYN 的源端口号进行选择，以使得：

[0069] 正确的初始目的地 = DAG_hash(源端口号) ,

[0070] 返回端口号 = F(源端口号) ,

[0071] 正确的返回目的地 = DAG_hash(返回端口号) ,

[0072] 此处，返回端口号可以未被初始主要计算设备使用，从而将返回端口号看作是暂时端口号。

[0073] 在上面，F 表示函数 F()，该函数被配置为调整 (swizzle) 位、添加已知偏移、或将源端口号转换为返回端口号。在一些实施例中，不同的源端口号可以被迭代以识别满足以上标准的号码。而且，根据所选择的 DAG_hash() 函数，另一函数 G() 可以被用来引导对源端口号的选择，从而加速对匹配标准的搜索。因此，其他机制可以被用来实现对次级计算设备的选择，并且控制 DAG 被采用的两个设备之间的分组的目的地。

[0074] 应当理解，附图以及类似于流程的示意图中的步骤的组合可以由计算机程序指令来实现。这些程序指令可以被提供给处理器以产生机器，以使得在处理器上运行的指令创建用于实现一个或多个流程框中指定的动作的装置。计算机程序指令可以由处理器来运行，以使得由处理器执行的一系列操作步骤产生计算机实现的处理，从而在处理器上运行的指令提供用于实现一个或多个流程框中指定的动作的步骤。这些程序指令可以被存储于计算机可读介质或机器可读介质上，例如，计算机可读存储介质。

[0075] 因此，示意图支持用于执行指定动作的装置的组合、用于执行指定动作的步骤以及用于执行指定动作的程序指令装置的组合。还应当理解，流程示意图中的每个框以及流程示意图中的框的组合可以由模块（例如，执行指定动作或步骤的基于专用硬件的系统、或者专用硬件和计算机指令的组合）来实现。

[0076] 上面的说明、示例以及数据提供了对所描述的实施例的组成的制造及使用的完整描述。因为许多实施例可以在不背离该说明书、所附权利要求中的实施例的精神和范围的情况下做出。

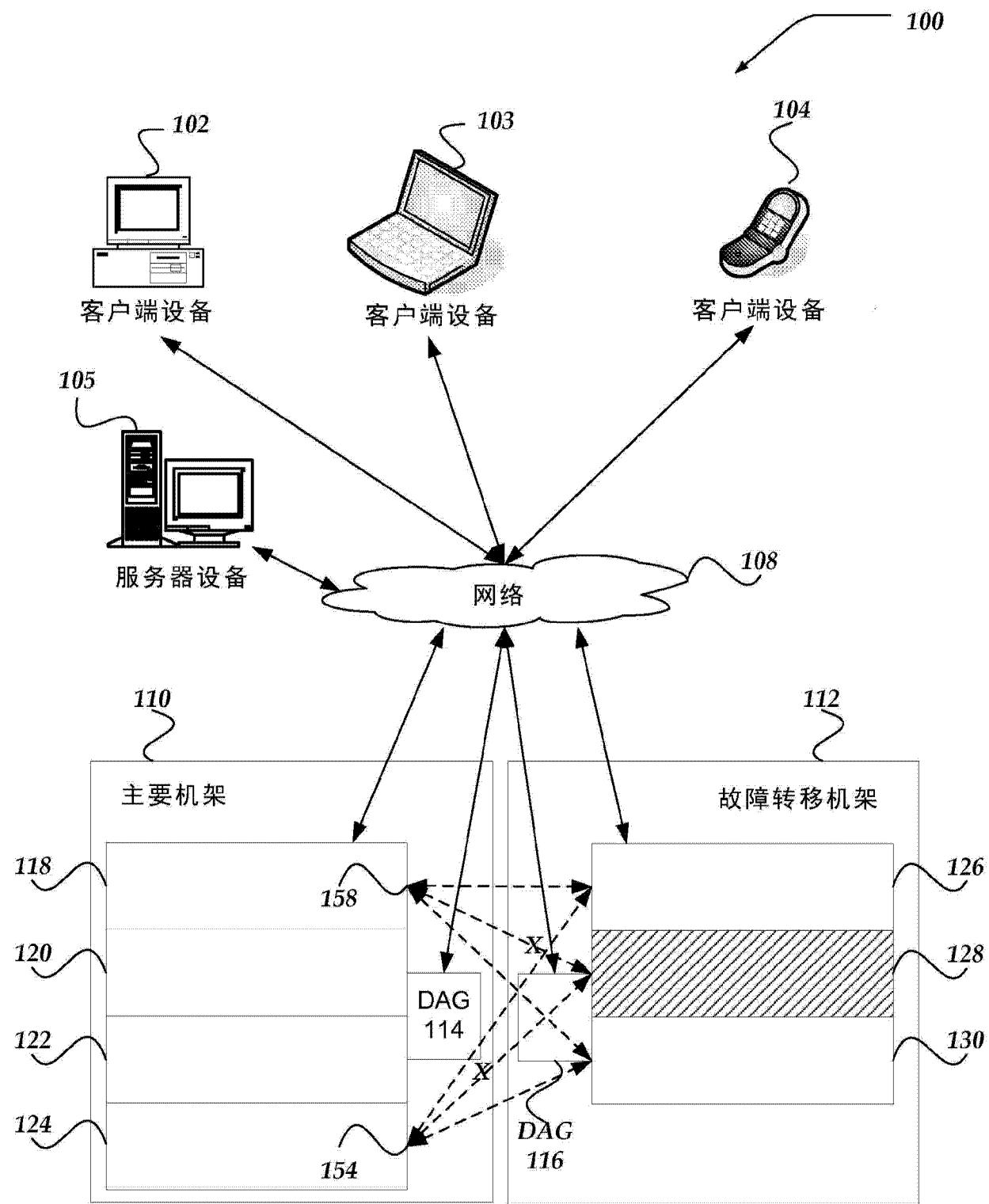


图 1

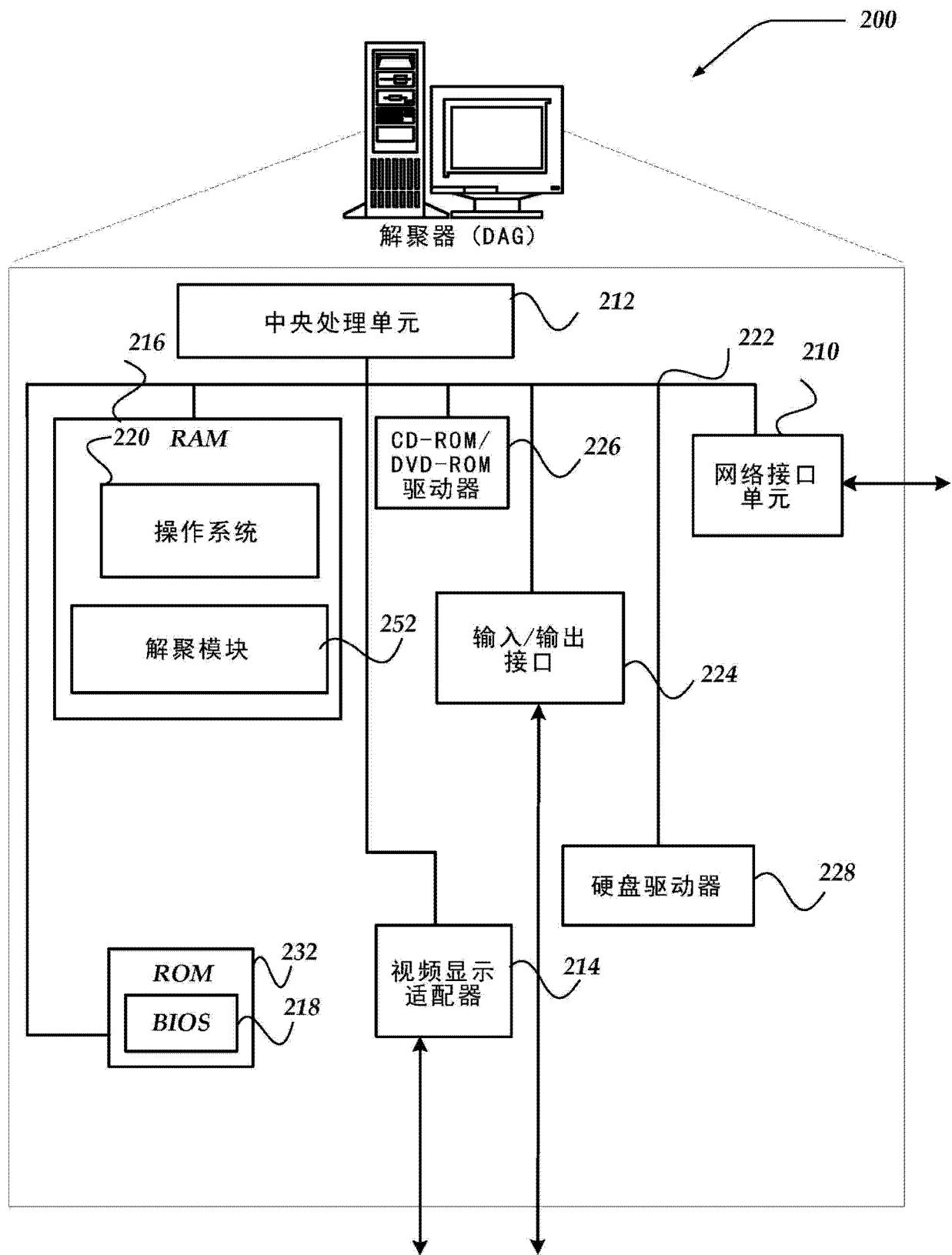


图 2

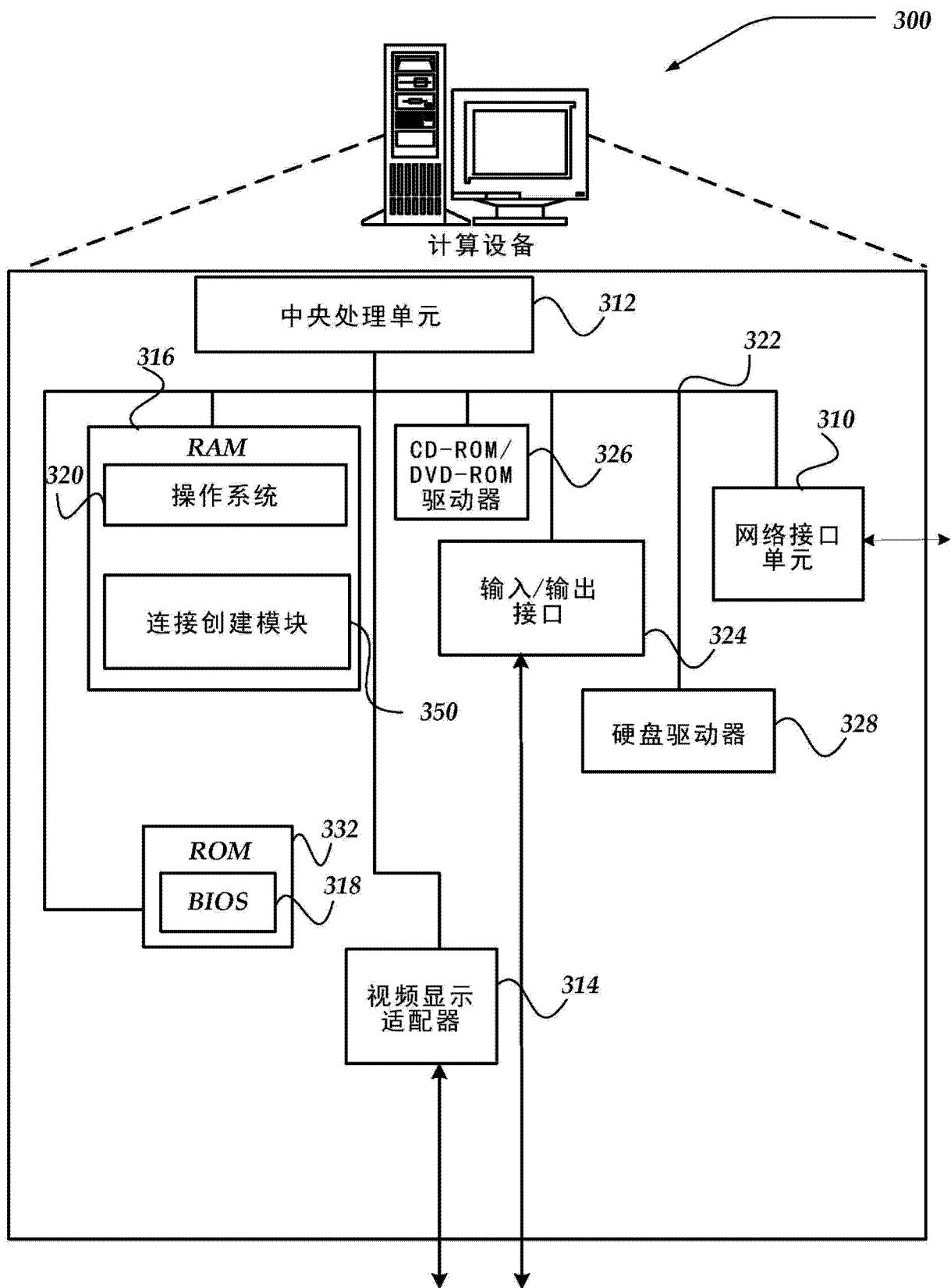


图 3

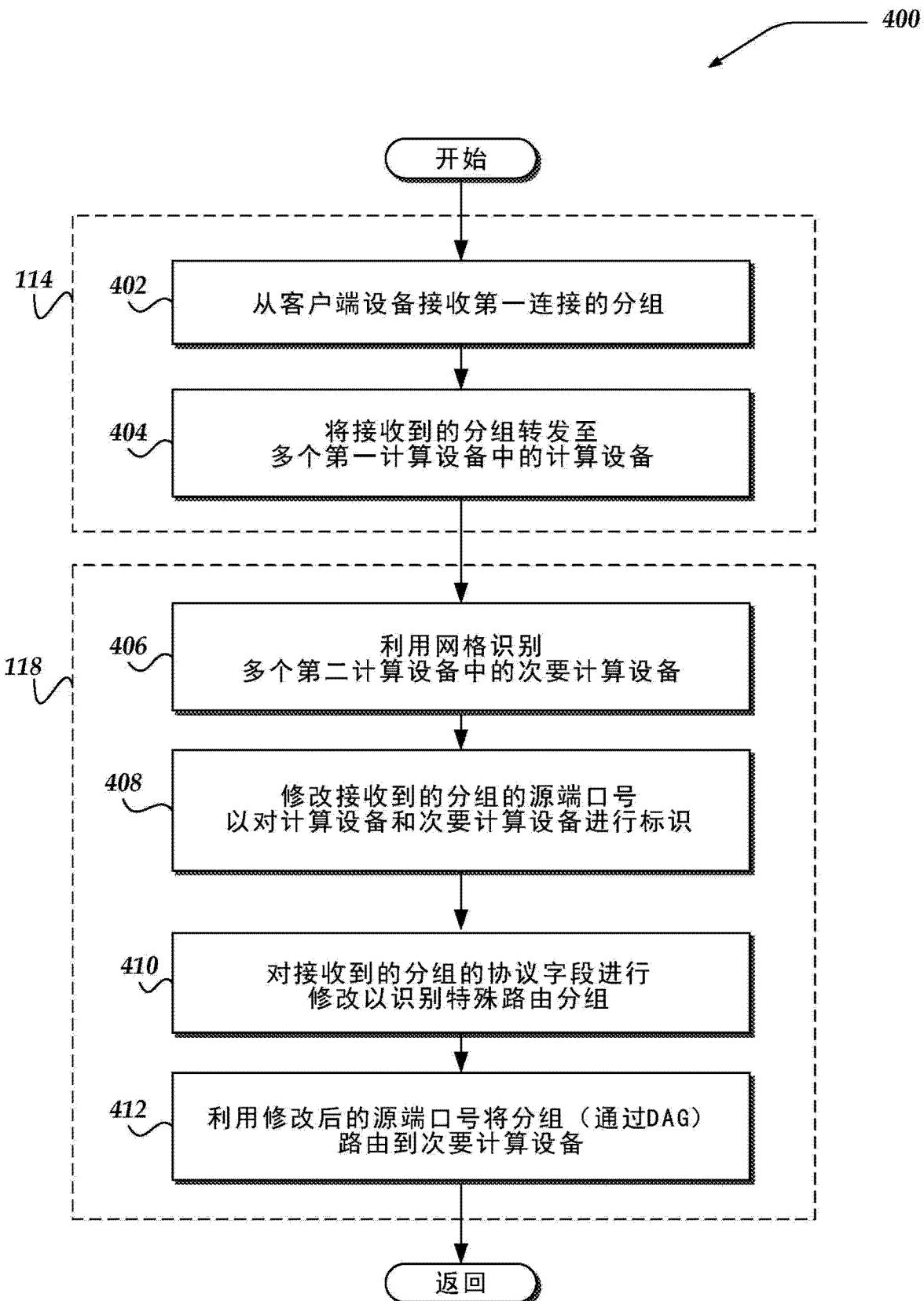


图 4