(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103125102 B
(45)授权公告日 2016.02.24

(21)申请号 201180040064.9
(22)申请日 2011.09.16
(30)优先权数据
61/384,228 2010.09.17 US
61/484,390 2011.05.10 US
61/493,330 2011.06.03 US
61/493,347 2011.06.03 US
61/498,329 2011.06.17 US
(85)PCT国际申请进入国家阶段日
2013.02.18
(86)PCT国际申请的申请数据
PCT/US2011/052033 2011.09.16
(87)PCT国际申请的公布数据
WO2012/037520 EN 2012.03.22
(73)专利权人 甲骨文国际公司
地址 美国加利福尼亚
(72)发明人 B-D·约翰逊
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 李晓芳
(51)Int.Cl.
H04L 12/46(2006.01)
(56)对比文件
JP 特开2002-247089 A,2002.08.30,
US 20080031266 A1,2008.02.07,
JP 特开2008-54214 A,2008.03.06,
审查员 吴荻

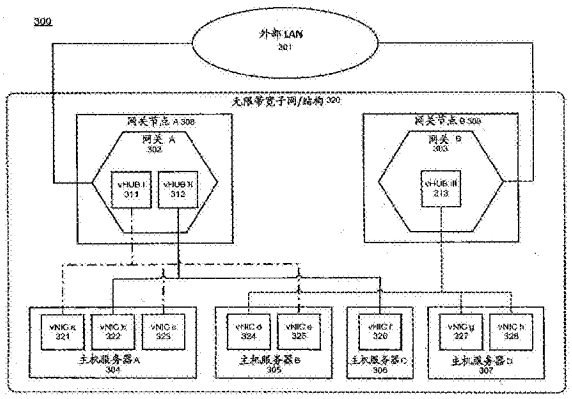
权利要求书5页 说明书9页 附图5页

(54)发明名称

用于在中间件机器环境中提供基于无限带宽的以太网虚拟集线器可伸缩性的系统和方法

(57)摘要

一种系统和方法能够支持包括一个或多个网关节点的中间件机器环境。在具有多个主机服务器的子网中提供驻留在一个或多个网关节点上的一个或多个网关，其中每个主机服务器与一个或多个虚拟网络接口卡(vNIC)相关联。中间件机器环境还包括一个或多个网关上的多个虚拟集线器(vHUB)，其中每个vHUB与一个或多个所述vNIC相关联。所述网关被适配为与外部网络连接，并且操作来经由多个vHUB将多播分组转发到外部网络和多个主机服务器二者，并且防止子网和外部网络之间的多播分组业务循环。



CN 103125102 B

1. 一种用于支持中间件机器环境的系统，包括：

一个或多个网关节点，其在一个或多个微处理器上执行，该一个或多个网关节点包括一个或多个网关和驻留在所述一个或多个网关上的多个虚拟集线器 vHUB，每个网关处于具有多个主机服务器的子网中，其中所述主机服务器和 vHUB 中的每一个与一个或多个虚拟网络接口卡 vNIC 相关联；

其中所述一个或多个网关被适配为与外部网络连接，并且操作来：

经由所述多个 vHUB 将多播分组转发到外部网络和多个主机服务器二者，以及

防止所述子网和所述外部网络之间的多播分组业务循环。

2. 根据权利要求 1 所述的系统，其中：

所述一个或多个网关节点包括一个或多个网络交换机，以及其中所述一个或多个网关驻留在所述一个或多个网络交换机上。

3. 根据权利要求 2 所述的系统，其中：

每个所述网络交换机提供用于与外部网络连接的一个或多个外部端口和用于与多个主机服务器连接的一个或多个内部端口。

4. 根据权利要求 2 或 3 所述的系统，还包括：

单独的存储系统，其通过所述一个或多个网络交换机连接到多个主机服务器。

5. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述子网是无限带宽 (IB) 子网。

6. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

至少一个网关包括多个 vHUB。

7. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

属于相同的 vHUB 的 vNIC 和主机服务器能够彼此通信而不涉及相关的网关实例。

8. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

属于不同的 vHUB 的 vNIC 和主机服务器能够通过对应的网关外部端口和外部网络来彼此通信。

9. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来防止内部 vNIC 或网关端口接收相同的逻辑分组的多个版本。

10. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来将一个或多个输入的多播分组转发到表示私有 vHUB 的多播群。

11. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来检测输出的多播分组是否来源于相关的 vHUB 中的 vNIC。

12. 根据权利要求 11 所述的系统，其中：

所述一个或多个网关操作来在输出的多播分组来源于相关的 vHUB 中的 vNIC 时，仅仅将输出的多播分组转发到外部网络。

13. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来使用一个或多个集合的范围寄存器定义与所述 vNIC 相关联的不同的源 MAC 地址。

14. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来使用源媒体访问控制 (MAC) 地址过滤方法管理属于不同的 vHUB 的 vNIC 之间的通信。

15. 根据权利要求 14 所述的系统，其中：

所述一个或多个网关操作来防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

16. 根据权利要求 1 到 3 中的任何一个所述的系统，还包括：

与所述一个或多个 vNIC 相关联的一个或多个主机 vNIC 驱动器，其操作来使用 MAC 地址过滤方法管理属于不同的 vHUB 的 vNIC 之间的通信。

17. 根据权利要求 16 所述的系统，其中：

所述一个或多个主机 vNIC 驱动器操作来防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

18. 根据权利要求 1 到 3 中的任何一个所述的系统，其中：

所述一个或多个网关操作来通过基于源 MAC 地址过滤来自于外部网络的多播分组，来防止重复的多播分组上的网络带宽的浪费。

19. 一种用于在中间件机器环境中提供可伸缩性的方法，包括：

在具有多个主机服务器的子网中提供一个或多个网关，其中所述一个或多个网关驻留在包含一个或多个微处理器的一个或多个网关节点上，其中每个主机服务器与一个或多个虚拟网络接口卡 (vNIC) 相关联；

在所述一个或多个网关上提供多个虚拟集线器 (vHUB)，其中每个 vHUB 与一个或多个所述 vNIC 相关联；

使得所述一个或多个网关与外部网络连接；

经由多个 vHUB 将多播分组转发到外部网络和所述多个主机服务器二者；以及

防止所述子网和所述外部网络之间的多播分组业务循环。

20. 根据权利要求 19 所述的方法，其中：

所述一个或多个网关节点包括一个或多个网络交换机，以及其中所述一个或多个网关驻留在所述一个或多个网络交换机上。

21. 根据权利要求 20 所述的方法，其中：

每个所述网络交换机提供用于与外部网络连接的一个或多个外部端口和用于与多个主机服务器连接的一个或多个内部端口。

22. 根据权利要求 20 或 21 所述的方法，还包括：

提供单独的存储系统，所述存储系统通过所述一个或多个网络交换机连接到所述多个主机服务器。

23. 根据权利要求 19 到 21 中的任何一个所述的方法，其中：

所述子网是无限带宽 (IB) 子网。

24. 根据权利要求 19 到 21 中的任何一个所述的方法，其中：

至少一个网关包括多个 vHUB。

25. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

属于相同的 vHUB 的 vNIC 和主机服务器彼此通信而不涉及相关的网关实例。

26. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

属于不同的 vHUB 的 vNIC 和主机服务器通过外部网络彼此通信。

27. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来防止内部 vNIC 或网关端口接收相同的逻辑分组的多个版本。

28. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来将一个或多个输入的多播分组转发到表示私有 vHUB 的多播群。

29. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来检测输出的多播分组是否来源于相关的 vHUB 中的 vNIC。

30. 根据权利要求 29 所述的方法，还包括：

所述一个或多个网关操作来在输出的多播分组来源于相关的 vHUB 中的 vNIC 时，仅仅将输出的多播分组转发到外部网络。

31. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来使用一个或多个集合的范围寄存器定义与所述 vNIC 相关联的不同的源 MAC 地址。

32. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来使用源媒体访问控制 (MAC) 地址过滤方法管理属于不同的 vHUB 的 vNIC 之间的通信。

33. 根据权利要求 32 所述的方法，还包括：

所述一个或多个网关操作来防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

34. 根据权利要求 19 到 21 中的任何一个所述的方法，还包括：

所述一个或多个网关操作来通过基于源 MAC 地址过滤来自于外部网络的多播分组，来防止重复的多播分组上的网络带宽的浪费。

35. 一种用于在中间件机器环境中提供可伸缩性的系统，包括：

用于在具有多个主机服务器的子网中提供运行在一个或多个微处理器上的一个或多个网关的装置，其中每个主机服务器与一个或多个虚拟网络接口卡 (vNIC) 相关联；

用于在所述一个或多个网关上提供多个虚拟集线器 (vHUB) 的装置，其中每个 vHUB 与一个或多个所述 vNIC 相关联；

用于使得所述一个或多个网关与外部网络连接的装置；

用于经由多个 vHUB 将多播分组转发到外部网络和所述多个主机服务器二者的装置；以及

用于防止所述子网和外部网络之间的多播分组业务循环的装置。

36. 根据权利要求 35 所述的系统，还包括：

用于提供一个或多个网络交换机的装置，其中所述一个或多个网关驻留在所述一个或

多个网络交换机上。

37. 根据权利要求 36 所述的系统，其中：

每个所述网络交换机提供用于与外部网络连接的一个或多个外部端口和用于与多个主机服务器连接的一个或多个内部端口。

38. 根据权利要求 35 或 36 所述的系统，还包括：

用于提供单独的存储系统的装置，所述存储系统通过所述一个或多个网络交换机连接到所述多个主机服务器。

39. 一种在包括具有多个主机服务器的子网的中间件机器环境中的网关，其中每个主机服务器与一个或多个虚拟网络接口卡 (vNIC) 相关联，所述网关被适配为与外部网络连接，所述网关包括：

一个或多个虚拟集线器 (vHUB)，其中每个 vHUB 与一个或多个所述 vNIC 相关联；

转发模块，被配置为经由所述一个或多个 vHUB 将多播分组转发到外部网络和所述多个主机服务器二者；以及

防止模块，被配置为防止所述子网和外部网络之间的多播分组业务循环。

40. 根据权利要求 39 所述的网关，其中：

所述子网是无限带宽 (IB) 子网。

41. 根据权利要求 39 所述的网关，其中：

所述网关包括多个 vHUB。

42. 根据权利要求 39 所述的网关，其中：

属于相同的 vHUB 的 vNIC 和主机服务器能够彼此通信而不涉及所述网关。

43. 根据权利要求 39 所述的网关，其中：

属于不同的 vHUB 的 vNIC 和主机服务器能够通过外部网络彼此通信。

44. 根据权利要求 39 所述的网关，其中：

所述防止模块进一步被配置为防止内部 vNIC 或网关端口接收相同的逻辑分组的多个版本。

45. 根据权利要求 39 所述的网关，其中：

所述转发模块进一步被配置为将一个或多个输入的多播分组转发到表示私有 vHUB 的多播群。

46. 根据权利要求 39 所述的网关，还包括：

检测模块，被配置为检测输出的多播分组是否来源于相关的 vHUB 中的 vNIC。

47. 根据权利要求 39 所述的网关，其中：

所述转发模块进一步被配置为，当检测模块检测到输出的多播分组来源于相关的 vHUB 中的 vNIC 时，仅仅将输出的多播分组转发到外部网络。

48. 根据权利要求 39 所述的网关，还包括：

定义模块，被配置为使用一个或多个集合的范围寄存器定义与所述 vNIC 相关联的不同的源 MAC 地址。

49. 根据权利要求 39 所述的网关，还包括：

管理模块，被配置为使用源媒体访问控制 (MAC) 地址过滤方法来管理属于不同的 vHUB 的 vNIC 之间的通信。

50. 根据权利要求 39 所述的网关，其中：

所述防止模块进一步被配置为防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

51. 根据权利要求 39 所述的网关，其中：

所述防止模块进一步被配置为通过基于源 MAC 地址过滤来自于外部网络的多播分组来防止重复的多播分组上的网络带宽的浪费。

52. 一种包括根据权利要求 39-51 中的任何一个所述的一个网关的网络交换机。

53. 根据权利要求 52 所述的网络交换机，还包括：

用于与外部网络连接的一个或多个外部端口，和

用于与多个主机服务器连接的一个或多个内部端口。

54. 一种用于支持中间件机器环境的系统，包括根据权利要求 52 或 53 所述的一个或多个网络交换机。

55. 根据权利要求 54 所述的系统，还包括：

单独的存储系统，其通过所述一个或多个网络交换机连接到多个主机服务器。

56. 一种用于支持中间件机器环境的系统，包括：

一个或多个网关，在其上驻留有多个虚拟集线器 vHUB，其中每个网关处于具有多个主机服务器的子网中，其中所述主机服务器和 vHUB 中的每一个与一个或多个虚拟网络接口卡 (vNIC) 相关联；

其中所述一个或多个网关被适配为与外部网络连接，并且操作来：

经由所述多个 vHUB 将多播分组转发到外部网络和所述多个主机服务器二者，以及

防止所述子网和外部网络之间的多播分组业务循环。

用于在中间件机器环境中提供基于无限带宽的以太网虚拟集线器可伸缩性的系统和方法

[0001] 版权通知

[0002] 本专利文件的公开的一部分包含受版权保护的材料。版权所有者不反对任何人对专利文件或专利公开的传真复制，因为它出现在专利商标局专利文档或记录中，但在别的方面保留所有任何版权。

技术领域

[0003] 本发明一般涉及计算机系统和诸如中间件之类的软件，并且特别涉及支持中间件机器环境。

背景技术

[0004] 无限带宽（Infiniband）（IB）架构是支持用于一个或多个计算机系统的 I/O 和处理器间通信二者的通信和管理基础设施。IB 架构系统的规模可以从具有几个处理器和几个 I/O 装置的小服务器到具有数百处理器和数千 I/O 装置的大规模并行设施。

[0005] IB 架构定义交换式通信结构，使得许多装置在受保护的、远程管理的环境中以高带宽和低延迟同时通信。端结点可以通过多个 IB 架构端口通信并且可以通过 IB 架构结构利用多个路径。提供通过网络的许多 IB 架构端口和路径以用于容错和增大的数据传送带宽二者。

[0006] 这些一般是本发明的实施例预期针对的领域。

发明内容

[0007] 这里描述的是用于支持包括一个或多个网关节点的中间件机器环境的系统和方法。在具有多个主机服务器的子网中提供驻留在一个或多个网关节点上的一个或多个网关，其中每个主机服务器与一个或多个虚拟网络接口卡（vNIC）相关联。中间件机器环境还包括一个或多个网关上的多个虚拟集线器（vHUB），其中每个 vHUB 与一个或多个所述 vNIC 相关联。网关被适配为与外部网络连接，并且操作来经由多个 vHUB 将多播分组转发到外部网络和多个主机服务器二者，并且防止子网和外部网络之间的多播分组业务循环。

[0008] 在一个方面中，提供一种在包括具有多个主机服务器的子网的中间件机器环境中的网关，其中每个主机服务器与一个或多个虚拟网络接口卡（vNIC）相关联，所述网关被适配为与外部网络连接，所述网关包括：一个或多个虚拟集线器（vHUB），其中每个 vHUB 与一个或多个所述 vNIC 相关联；转发模块，被配置为经由所述一个或多个 vHUB 将多播分组转发到外部网络和多个主机服务器二者；以及防止模块，被配置为防止所述子网和外部网络之间的多播分组业务循环。

[0009] 在一些实施例中，子网是无限带宽（IB）子网。

[0010] 在一些实施例中，网关包括多个 vHUB。

[0011] 在一些实施例中，属于相同的 vHUB 的 vNIC 和主机服务器可以彼此通信而不涉及

网关。

[0012] 在一些实施例中，属于不同的 vHUB 的 vNIC 和主机服务器可以通过外部网络彼此通信。

[0013] 在一些实施例中，防止模块还被配置为防止内部 vNIC 或网关端口接收相同的逻辑分组的多个版本。

[0014] 在一些实施例中，转发模块还被配置为将一个或多个输入的多播分组转发到表示私有 vHUB 的多播群。

[0015] 在一些实施例中，网关还包括检测模块，被配置为检测输出的多播分组是否来源于相关的 vHUB 中的 vNIC。

[0016] 在一些实施例中，转发模块还被配置为，当检测模块检测到输出的多播分组来源于相关的 vHUB 中的 vNIC 时，仅仅将输出的多播分组转发到外部网络。

[0017] 在一些实施例中，网关还包括定义模块，被配置为使用一个或多个集合的范围寄存器来定义与所述 vNIC 相关联的不同的源 MAC 地址。

[0018] 在一些实施例中，网关还包括管理模块，被配置为使用源媒体访问控制（MAC）地址过滤方法来管理属于不同的 vHUB 的 vNIC 之间的通信。

[0019] 在一些实施例中，防止模块还被配置为防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

[0020] 在一些实施例中，防止模块还被配置为通过基于源 MAC 地址过滤来自于外部网络的多播分组来防止重复的多播分组上的网络带宽的浪费。

[0021] 在另一个方面中，提供一种包括根据本公开的一个方面的一个网关的网络交换机。

[0022] 在一些实施例中，该网络交换机还包括：一个或多个外部端口，用于与外部网络连接；和一个或多个内部端口，用于与多个主机服务器连接。

[0023] 在另一个方面中，提供一种用于支持中间件机器环境的系统，包括根据本公开的另一个方面的一个或多个网络交换机。

[0024] 在一些实施例中，该系统还包括通过所述一个或多个网络交换机连接到多个主机服务器的单独的存储系统。

附图说明

[0025] 图 1 示出了根据本发明的实施例的用于中间件机器的示范性配置的例示。

[0026] 图 2 示出了根据本发明的实施例的中间件机器环境的例示。

[0027] 图 3 示出了根据本发明的实施例的提供基于 IB 的以太网（EOIB）vHUB 可伸缩性的中间件机器环境的例示。

[0028] 图 4 示出了根据本发明的实施例的用于在中间件机器环境中提供 EOIB vHUB 可伸缩性的示范性流程图。

[0029] 图 5 是根据本发明的一些实施例的中间件机器环境中的网关的功能框图。

[0030] 图 6 是根据本发明的一些实施例的中间件机器环境中的网络交换机的功能框图。

具体实施方式

[0031] 这里描述的是用于提供中间件机器或相似的平台的系统和方法。根据本发明的实施例，该系统包括高性能硬件(例如，64 位处理器技术、高性能大容量存储器和冗余无限带宽和以太网联网)与诸如 WebLogic 套件之类的应用服务器或中间件环境的组合，以提供完整的 Java EE 应用服务器联合体，该 Java EE 应用服务器联合体包括整体上并行的存储器内网格，可以被快速提供，并且可以根据需要缩放。根据本发明的实施例，该系统可以被布置为完全的、一半或四分之一机架，或其它配置，这些配置提供应用服务器网格、存储区域网络和无限带宽(IB)网络。中间件机器软件可以提供应用服务器、中间件和诸如例如 WebLogic Server、JRockit 或 Hotspot JVM、Oracle Linux 或 Solaris 和 Oracle VM 之类的其它功能。根据本发明的实施例，该系统可以包括经由 IB 网络彼此通信的多个计算节点、一个或多个 IB 交换机网关和存储节点或单元。当被实现为机架配置时，机架的不使用的部分可以保持为空或被填充物占据。

[0032] 根据本发明的实施例，这里称为“Sun Oracle Exalogic”或“Exalogic”的系统是对于容纳诸如 Oracle Middleware SW 套件或 Weblogic 之类的中间件或应用服务器软件的容易布置的解决方案。如这里所述，根据实施例，系统是“盒装网格(grid in a box)”，其包括一个或多个服务器、存储单元、用于存储联网的 IB 结构、以及容纳中间件应用所需的所有其它组件。显著的性能可以通过使用例如 Real Application Clusters 和 Exalogic Open 存储器来平衡整体上并行的网格架构而对于所有类型的中间件应用分发。该系统利用线性 I/O 可伸缩性分发改善的性能、使用和管理起来简单、并且分发对任务关键的可用性和可靠性。

[0033] 图 1 示出了根据本发明的实施例的用于中间件机器的示范性配置的例示。如图 1 所示，中间件机器 100 使用单个机架配置，其包括两个网关网络交换机或连接到二十八个服务器节点的叶子网络交换机 102 和 103。另外，可以存在对于中间件机器的不同的配置。例如，可以存在一半机架配置，其包含服务器节点的一部分，并且也可以存在多机架配置，其包含大量服务器。

[0034] 如图 1 所示，服务器节点可以连接到由网关网络交换机提供的端口。如图 1 所示，每个服务器机器可以具有单独地到两个网关网络交换机 102 和 103 的连接。例如，网关网络交换机 102 连接到服务器 1-14106 的端口 1 和服务器 15-28107 的端口 2，并且网关网络交换机 103 连接到服务器 1-14108 的端口 2 和服务器 15-28109 的端口 1。

[0035] 根据本发明的实施例，每个网关网络交换机可以具有用于与不同的服务器连接的多个内部端口，并且网关网络交换机也可以具有用于与诸如现有数据中心服务网络之类的外部网络连接的外部端口。

[0036] 根据本发明的实施例，中间件机器可以包括通过网关网络交换机连接到服务器的单独的存储系统 110。另外，中间件机器可以包括连接到两个网关网络交换机 102 和 103 的干线网络交换机 101。如图 1 所示，可以可选地存在从存储系统到干线网络交换机的两个链路。

[0037] IB 结构 / 子网

[0038] 根据本发明的实施例，中间件机器环境中的 IB 结构 / 子网可以包含以胖树状拓扑互连的大量物理主机或服务器、交换机实例和网关实例。

[0039] 图 2 示出了根据本发明的实施例的中间件机器环境的例示。如图 2 所示，中间件机器环境 200 包括与多个末端节点连接的 IB 子网或结构 220。IB 子网包括多个子网管理器 211-214，每个子网管理器驻留在多个网络交换机 201-204 中的一个上。子网管理器可以使用带内通信协议 210 彼此通信，带内通信协议诸如基于管理数据报（MAD）/ 子网管理分组（SMP）的协议或诸如基于 IB 的互联网协议（IPoIB）之类的其它协议。

[0040] 根据本发明的实施例，可以在 IB 结构上构造单个 IP 子网，IB 结构使得交换机在相同的 IB 结构中彼此安全地通信（即，所有交换机之间的完全连接）。当在两个交换机之间存在具有操作链路的至少一个路线时，基于该结构的 IP 子网可以提供任何一对交换机之间的连接。如果通过重新路由而存在可替换的路线，则可以实现从链路故障的恢复。

[0041] 交换机的管理以太网接口可以连接到提供所有交换机之间的 IP 级别的连接的单个网络。每个交换机可以由两个主要 IP 地址标识：一个用于外部管理以太网并且一个用于基于该结构的 IP 子网。每个交换机可以使用两个 IP 地址监视到所有其它交换机的连接，并且可以使用任一操作地址用于通信。另外，每个交换机可以具有到该结构上的每个直接连接的交换机的点对点 IP 链路。因此，可以存在至少一个附加的 IP 地址。

[0042] IP 路由设置使得网络交换机能够使用该结构的 IP 子网、外部管理以太网网络和交换机对之间的一个或多个结构级别的点对点 IP 链路的组合经由中间交换机将业务路由到另一个交换机。IP 路由使得对网络交换机的外部管理访问能够经由网络交换机上的外部以太网端口以及通过该结构上的专用路由服务来路由。

[0043] IB 结构包括具有对管理网络的管理以太网访问的多个网络交换机。存在该结构中的交换机之间的带内物理连接。在一个示例中，当 IB 结构不退化时，在每一对交换机之间存在一个或多个跳的至少一个带内路由。对于 IB 结构的管理节点包括连接到 IB 结构的网络交换机和管理主机。

[0044] 子网管理器可以经由它的私有 IP 地址中的任何一个被访问。子网管理器也可以经由浮动 IP 地址访问，当子网管理器起主子网管理器的作用时，浮动 IP 地址被配置用于主子网管理器，并且当子网管理器被从该角色明确地释放时，子网管理器被解配置。可以对于外部管理网络以及对于基于该结构的管理 IP 网络二者定义主 IP 地址。对于点对点 IP 链路，不需要定义特殊的主 IP 地址。

[0045] 根据本发明的实施例，可以使用基于虚拟机的客户机将每个物理主机虚拟化。可以每个物理主机同时存在多个客户机，例如每个 CPU 核一个客户机。另外，每个物理主机可以具有至少一个双端口主机通道适配器（HCA），其可以被虚拟化并且在客户机之间共享，以使得虚拟化的 HCA 的结构图是单个双端口 HCA，正如非虚拟化的 / 共享的 HCA 一样。

[0046] IB 结构可以分为由 IB 分区实现的动态集的资源域。IB 结构中的每个物理主机和每个网关实例可以是多个分区的成员。此外，相同的或不同的物理主机上的多个客户机可以是相同的或不同的分区的成员。对于 IB 结构的 IB 分区的数目可以由 P_Key 表大小限制。

[0047] 根据本发明的实施例，客户机可以开启直接从客户机中的 vNIC 驱动器访问的两个或更多个网关实例上的一组虚拟网络接口卡（vNIC）。客户机可以在物理主机之间迁移同时保持或具有更新的 vNIC 伙伴。

[0048] 根据本发明的实施例，交换机可以按照任何顺序启动并且可以根据例如 IB 指定的协商协议之类的不同的协商协议动态地选择主子网管理器。如果没有指定分区策略，则

可以使用默认的使能分区的策略。另外，可以独立于任何附加的策略信息并且独立于主子网管理器是否知道完整的结构策略，来建立管理节点分区和基于结构的管理 IP 子网。为了使得使用基于该结构的 IP 子网同步结构级别的配置策略信息，子网管理器可以最初使用默认分区策略启动。当已经实现结构级别的同步时，对于结构当前的分区配置可以由主子网管理器安装。

[0049] 提供基于 IB 的以太网（EOIB）vHUB 可伸缩性

[0050] 根据本发明的实施例，该系统可以提供基于 IB 的以太网（EOIB）级别 2（L2）子网实施方式，其在 IB 结构上的成员主机端口的数目方面以及在连接到外部以太网结构上的对应 L2 子网的网关端口的数目方面进行伸缩。

[0051] 图 3 示出了根据本发明的实施例的提供 EOIB vHUB 可伸缩性的中间件机器环境的例示。如图 3 所示，中间件机器环境 300 包括 IB 子网 / 结构 320，其可以与多个主机服务器 304-307 以及外部局域网（LAN）301 连接。IB 结构包括几个网关 302-303，其与不同的 vNIC321-328 相关联。每个网关实例可以驻留在包含一个或多个微处理器的网关节点 308-309 上，网关实例中的核心网关功能可以使用“数据路径”操作在硬件中实现。

[0052] 根据本发明的实施例，系统中的网关可以与不同的 vHUB 相关联。每个 vHUB 定义包含与相同的网关实例相关联的 vNIC 的 IB 结构侧上的逻辑级别 2 链路。属于相同的 vHUB 的 vNIC 和主机可以在不涉及相关的网关实例的情况下彼此通信。

[0053] 在如图 3 所示的示例中，网关 A 上的 vHUB I311 与主机服务器 A 上的 vNIC a321 和 vNIC c323 以及主机服务器 B 上的 vNIC e325 相关联。因此，vNIC a、vNIC c 和 vNIC e 可以在不涉及相关的网关 A 的情况下彼此通信。此外如图 3 所示，网关 A 上的 vHUB II312 与主机服务器 A 上的 vNIC b322 和主机服务器 C 上的 vNIC f326 相关联；并且网关 B 上的 vHUB III313 与主机服务器 B 上的 vNIC d324 以及主机服务器 D 上的 vNIC g327 和 vNIC h328 相关联。

[0054] 根据本发明的实施例，几个 vHUB 可以表示相同的逻辑以太网 L2 链路和 / 或相同的级别 3（L3）IP 子网。在如图 3 所示的示例中，网关 A302 和网关 B303 二者经由多个网关端口连接到相同的外部 LAN301。属于不同的 vHUB311-313 的各个 vNIC321-328 可以通过外部 LAN301 或者可替换地通过由 IB 结构上的主机实现的路由逻辑连接并且可以彼此通信。

[0055] 根据本发明的实施例，可以在 IB 结构侧提供单独的 IP 子网（IPoIB 或 EoIB），用于在 IB 结构上的主机之间以及在主机和外部 LAN 之间处理高带宽 IP 业务。IB 结构侧的此单独的 IP 子网可以避免浪费用于 IB 结构上的主机之间的业务的网关带宽。另外，当不期望多个 IP 子网时，允许属于内部 IB 结构上的不同的 vHUB 的 vNIC 之间的通信是有用的。

[0056] 在如图 3 所示的示例中，当多个 vHUB311-313 连接在一起时，可以将多播分组通过网关实例 302-303 转发给外部 LAN301 并且转发给 IB 结构中的主机服务器 304-307 上的成员主机端口二者。例如，IB 结构 302 中的第一网关端口可以向外部 LAN301 发送多播分组，外部 LAN301 可以将多播分组发送回 IB 结构 320 中的第二网关端口。第二网关端口又可以在 IB 结构上再次转发多播分组。这可能导致 IB 结构 320 和外部网络 301 之间的业务循环。另外，内部 vNIC321-328 和网关端口可能接收相同逻辑分组的多个版本。

[0057] 为了解决与从外部网络到 IB 结构的业务有关的进入分组循环问题，网关实例可以将输入的多播分组转发到表示私有 vHUB 的 IB 多播群，IB 多播群具有单个网关成员并且

允许每个 vNIC 成为单个私有 vHUB 的成员。

[0058] 为了解决与从 IB 结构到外部网络的业务有关的输出分组循环问题，每个网关实例可以检测从本地 IB 结构接收到的多播分组是否来源于它的私有 vHUB 中的 vNIC。如果多播分组来源于它的私有 vHUB 中的 vNIC，则网关实例可以进行以将多播分组转发到外部网络。专用硬件逻辑可以用于做出这样的决定。在一个示例中，专用硬件逻辑可以使用关于本地 vNIC 的信息，或者使用私有 vHUB 中的源媒体访问控制（MAC）地址的一个或多个集合的范围寄存器。

[0059] 根据本发明的实施例，可替换的方法可以要求在不同的私有 vHUB 之间转发的多播分组必须涉及外部 LAN，以使得下层 IB 多播群总是局限于单个私有 vHUB。同时，单播业务可以跨越多个 vHUB。此可替换的方法可以解决进入循环问题并且也可以保证从 IB 结构向外部 LAN 转发多播分组的仅仅单个版本。

[0060] 根据的实施例，可以实现属于不同的 vHUB 的 vNIC 之间的通信而不取决于外部网关端口之间经由外部以太网 LAN 的连接。对于单播业务，主机 vNIC 驱动器可以将单个分组明确地发送到 IB 结构上的目的地 MAC/vNIC 与之相关联的目的地端口。此外，主机 vNIC 驱动器可以向属于不同的 vHUB 的目的地 vNIC 发送单个分组。对于多播业务，系统可以保证仅仅单个网关实例将特定的多播分组从特定的主机或 vNIC 转发到外部 LAN。并且接收相同的逻辑多播分组的多个网关实例可以不将相同的逻辑多播分组从外部 LAN 转发到 IB 结构上的相同的主机或 vNIC。

[0061] 根据本发明的实施例，特殊的全局 vHUB 多播群可以避免对用于在私有 vHUB 之间进行多播转发的外部 LAN 的依赖性，并且还保证从 IB 结构到外部 LAN 转发仅仅单个版本的多播分组。此特殊的全局 vHUB 多播群可以包括仅仅单个网关实例和 IB 结构上的所有相关的 vNIC。使用此方法，系统可以避免对于 IB 结构和外部 LAN 之间的业务创建循环。此外，为了避免接收来源于另一个私有 vHUB 中的 vNIC 的多播分组的多个副本，EoIB 驱动器可以包括如下逻辑：当多播分组的源 MAC 来自于对应全局 vHUB 中的 vNIC 时，丢弃从它的私有 vHUB 中的网关到达的多播分组。

[0062] 此方法的局限性是，由于在全局 vHUB 以及从外部 LAN 接收分组的各个私有 vHUB 二者中的分组转发，在 IB 结构上可能消耗更多的带宽。此外，仅仅单个用于发送多播分组的输出网关实例的限制可能要求结合和其它基于主机的 IP 多连接方案取决于全局 vHUB 重新配置以便在当前输出网关实例故障之后恢复外部多播发送能力。因而，从 IB 结构到外部 LAN 的多播带宽也可以由单个输出网关实例限制。

[0063] 根据本发明的实施例，系统可以使用源 MAC 地址过滤方法管理属于不同的 vHUB 的 vNIC 之间的通信。在如图 3 所示的示例中，IB 结构 320 中的每个网关实例 302-303 可以检测从外部 LAN301 接收到的多播分组的源 MAC 地址是否属于本地 IB 结构 320 上的相关的 vNIC321-328。

[0064] 根据本发明的实施例，该系统可以使用定义属于本地 IB 结构上的相关的 vNIC 的源 MAC 地址的一个或多个集合的范围寄存器，以便防止浪费网络带宽的、在 IB 结构和外部 LAN 之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不会接收到相同的逻辑多播分组的重复的版本。

[0065] 根据本发明的实施例，该系统可以使用专用多播群来将多播分组分布到 IB 结构

上的 vNIC，而不是分布到任何网关实例。为了保证多播分组的刚好一个副本经由刚好一个网关外部端口发出到外部 LAN，vNIC 驱动器可以向单个网关（例如，与本地 vNIC 所属的 vHUB 相关联的一个网关）发送分组。此多播分组然后可以经由外部以太网 LAN 由其它网关外部端口接收到并且被复制到所属其它 vHUB 的 vNIC。

[0066] 进行接收的 vNIC 驱动器可以确定分组中的源 MAC 属于作为相同的逻辑子网的一部分的 IB 结构上的 vNIC。然后，由于该分组已经或将要经由 IB 结构上的所述专用多播群被接收到，因此 vNIC 驱动器可以丢弃该分组。另一方面，如果确定接收的多播分组源地址属于外部以太网 LAN 上的站（即，与本地 IB 结构上的 vNIC 相反），则可以相应地处理多播分组。

[0067] 根据本发明的实施例，利用滤出多播分组的副本的方案，可以使用一个全局多播群来在 IB 结构上的全局 vHUB 之内转发多播分组，并且还可以允许进行发送的 EoIB 驱动器例如经由单播明确地发送要由它的私有 vHUB 中的网关实例转发的另一个分组副本。此方案允许许多网关实例能够被用作用于输出 MC 业务的输出端口（即，每个私有 vHUB 一个激活的输出网关）。

[0068] 根据本发明的实施例，在 IB 结构中可以使用混合 L2 链路，在这种情况下，类似于 IPoIB 链路的单个 IB 结构内部 EoIB vHUB 可以与仅仅用于外部业务的一个或多个 vHUB 组合。网关的基于源地址过滤来自于外部 LAN 的多播分组的能力可以防止在重复的多播分组上的 IB 结构带宽的浪费。

[0069] 根据本发明的实施例，对于 IPoIB 连接模式（CM）比基于 EoIB 的 IP 业务提供更好的带宽的情况，IPoIB CM 可以与 EoIB 合并，以便允许单个 IP 子网跨越 IB 结构上的任何数目的节点以及外部 LAN 上的任何数目的网关实例和节点，同时仍然在任何对端点之间提供最佳可能的带宽。

[0070] 图 4 示出了根据本发明的实施例的用于在中间件机器环境中提供可伸缩性的示范性流程图。如图 4 所示，在步骤 401，可以在具有多个主机服务器的子网中提供一个或多个网关，其中每个主机服务器与一个或多个虚拟网络接口卡（vNIC）相关联。然后，在步骤 402，可以在一个或多个网关上提供多个虚拟集线器（vHUB），其中每个 vHUB 与一个或多个所述所述 VNIC 相关联。另外，在步骤 403，一个或多个网关可以与外部网络连接。此外，在步骤 404，一个或多个网关可以将多播分组经由多个 vHUB 转发到外部网络和多个主机服务器二者。最后，在步骤 405，一个或多个网关可以防止子网和外部网络之间的多播分组业务循环。

[0071] 根据一些实施例，图 5 示出了根据如上所述的本发明的原理配置的网关 500 的功能框图，并且图 6 示出了根据如上所述的本发明的原理配置的网络交换机 600 的功能框图，包括如图 5 所示的网关 500。网关和网络交换机的功能块可以由硬件、软件或硬件和软件的组合实现以执行本发明的原理。本领域技术人员将理解，图 5 和 6 中描述的功能块可以被组合或分成子块以实现如上所述的本发明的原理。因此，这里的描述可以支持这里描述的功能块的任何可能的组合或分离或者进一步定义。

[0072] 网关 500 操作在图 3 所示的包括具有多个主机服务器的子网的中间件机器环境中。每个主机服务器与一个或多个虚拟网络接口卡（vNIC）相关联。网关 500 被适配为与外部网络连接。

[0073] 如图 5 所示，网关 500 可以包括一个或多个虚拟集线器（vHUB）502、转发模块 504 和防止模块 506。

[0074] 在一些实施例中，每个 vHUB502 与一个或多个 VNIC 相关联。转发模块 504 被配置为经由多个 vHUB502 将多播分组转发到外部网络和多个主机服务器二者。防止模块 506 被配置为防止子网和外部网络之间的多播分组业务循环。

[0075] 在一些实施例中，子网是无限带宽（IB）子网。在一些实施例中，网关 500 包括多个 vHUB502。在一些实施例中，属于相同的 vHUB504 的 vNIC 和主机服务器可以彼此通信而不涉及相关的网关实例。在一些实施例中，属于不同的 vHUB504 的 vNIC 和主机服务器可以通过外部网络彼此通信。

[0076] 在一些实施例中，防止模块 506 可以进一步被配置为防止内部 vNIC 或网关端口接收相同的逻辑分组的多个版本。在一些实施例中，转发模块 504 可以进一步被配置为将一个或多个输入的多播分组转发到表示私有 vHUB502 的多播群。

[0077] 在一些实施例中，可替换地，网关 500 可以进一步包括检测模块 508，检测模块 508 被配置为检测输出的多播分组是否来源于相关的 vHUB502 中的 vNIC。转发模块 504 可以进一步被配置为，当检测模块 508 检测到输出多播分组来源于相关的 vHUB502 中的 vNIC 时，仅仅将输出的多播分组转发到外部网络。

[0078] 在一些实施例中，网关 500 可以进一步包括定义模块 510，定义模块 510 被配置为使用一个或多个集合的范围寄存器来定义与该 vNIC 相关联的不同的源 MAC 地址。

[0079] 在一些实施例中，网关 500 可以进一步包括管理模块 512，管理模块 512 被配置为使用源媒体访问控制（MAC）地址过滤方法来管理属于不同的 vHUB502 的 vNIC 之间的通信。防止模块 506 可以进一步被配置为防止浪费网络带宽资源的、在子网和外部网络之间的循环中发送多播分组，并且保证内部 vNIC 或网关端口不接收相同的逻辑多播分组的重复的版本。

[0080] 在一些实施例中，防止模块 506 可以进一步被配置为通过基于源 MAC 地址过滤来自于外部网络的多播分组来防止重复的多播分组上的网络带宽的浪费。

[0081] 如图 6 所示，网络交换机 600 可以包括如图 5 所示的一个网关 500。在一些实施例中，网络交换机 600 可以进一步包括一个或多个外部端口 604 和一个或多个内部端口 608。外部端口 604 用于与外部网络连接。内部端口 608 用于与中间件机器环境中的多个主机服务器连接。

[0082] 在一些实施例中，提供一种用于支持中间件机器环境的系统，包括如图 6 所示的一个或多个网络交换机 600。该系统可以进一步包括通过所述一个或多个网络交换机 600 连接到多个主机服务器的单独的存储系统。

[0083] 可以使用一个或多个传统的通用或专用数字计算机、计算设备、机器或微处理器方便地实现本发明，包括一个或多个处理器、存储器和 / 或根据本公开的教导编程的计算机可读存储介质。合适的软件编码能够容易地由熟练的程序员基于本公开的教导来准备，这对软件领域的技术人员将是明显的。

[0084] 在一些实施例中，本发明包括计算机程序产品，其是上面或其中存储指令的存储介质或计算机可读介质（介质），指令可以用于编程计算机来执行本发明的处理中的任何一个。存储介质能够包括以下中的一个或多个：任何类型的盘，包括软盘、光盘、DVD、CD-ROM、

微驱动器、和磁光盘、ROM、RAM、EPROM、EEPROM、DRAM、VRAM、闪速存储器件、磁或光卡、纳米系统(包括分子记忆 IC)或适合于存储指令和 / 或数据的任何类型的介质或设备,但是不局限于此。

[0085] 本发明的上面的描述是为了说明和描述的目的而提供的。它不意欲是穷举的或将本发明限制于公开的精确的形式。许多修改和变化能够对本领域的实践者是明显的。选择并且描述实施例以便最佳地说明本发明的原理和它的实际应用,从而使得本领域技术人员能够理解用于各个实施例中的并且具有适合于预期的特定使用的各种修改的本发明。预期本发明的范围由以下权利要求书和它们的等价物定义。

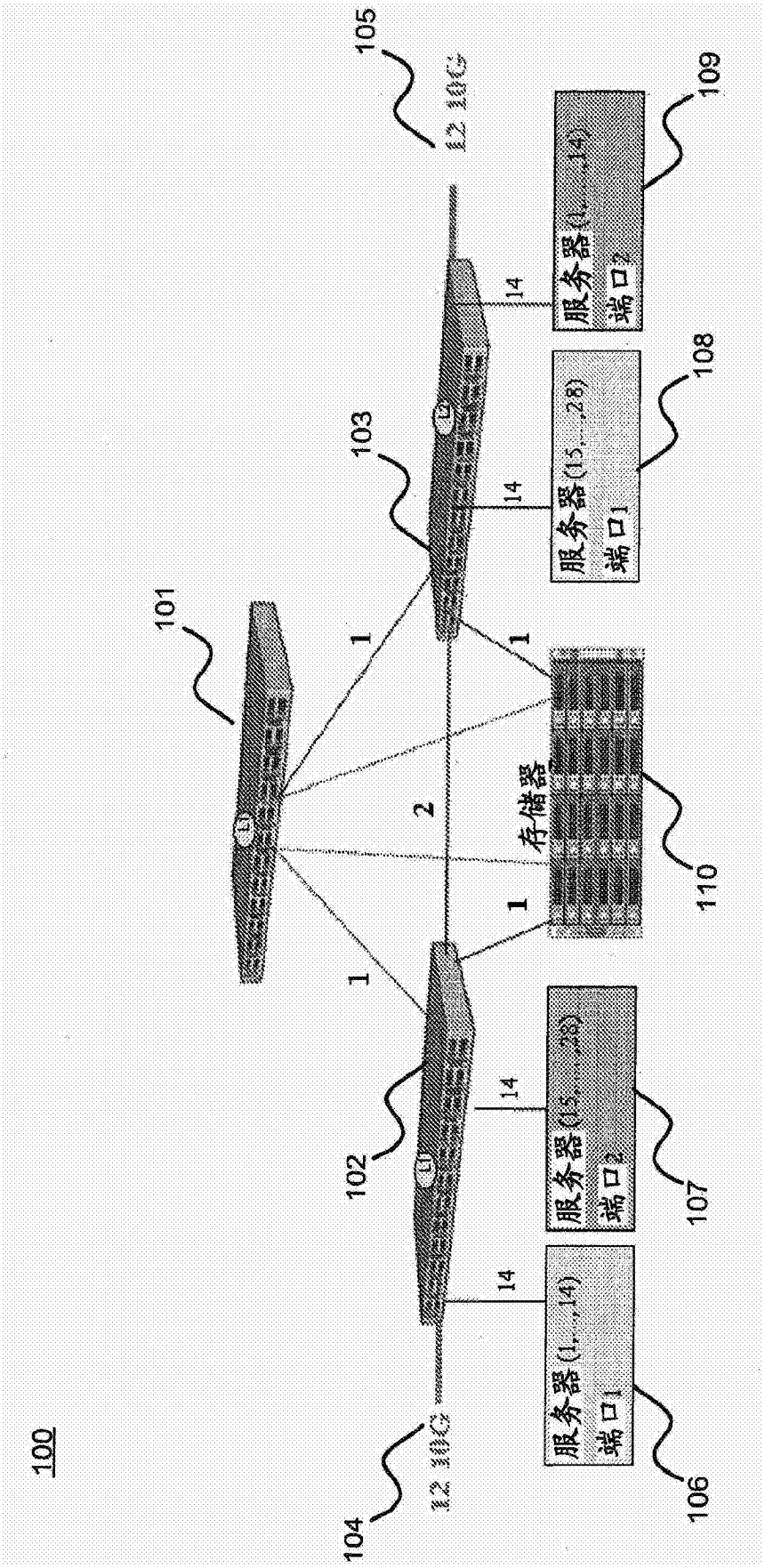

100
101
102
103
104
105
106
107
108
109
110
12 10G
12 10G
存储器
服务器(1,...,14)端口2
服务器(15,...,28)端口1
服务器(15,...,28)端口2
服务器(1,...,14)端口1
1
1
2
1
1
14
14
14
14

图 1

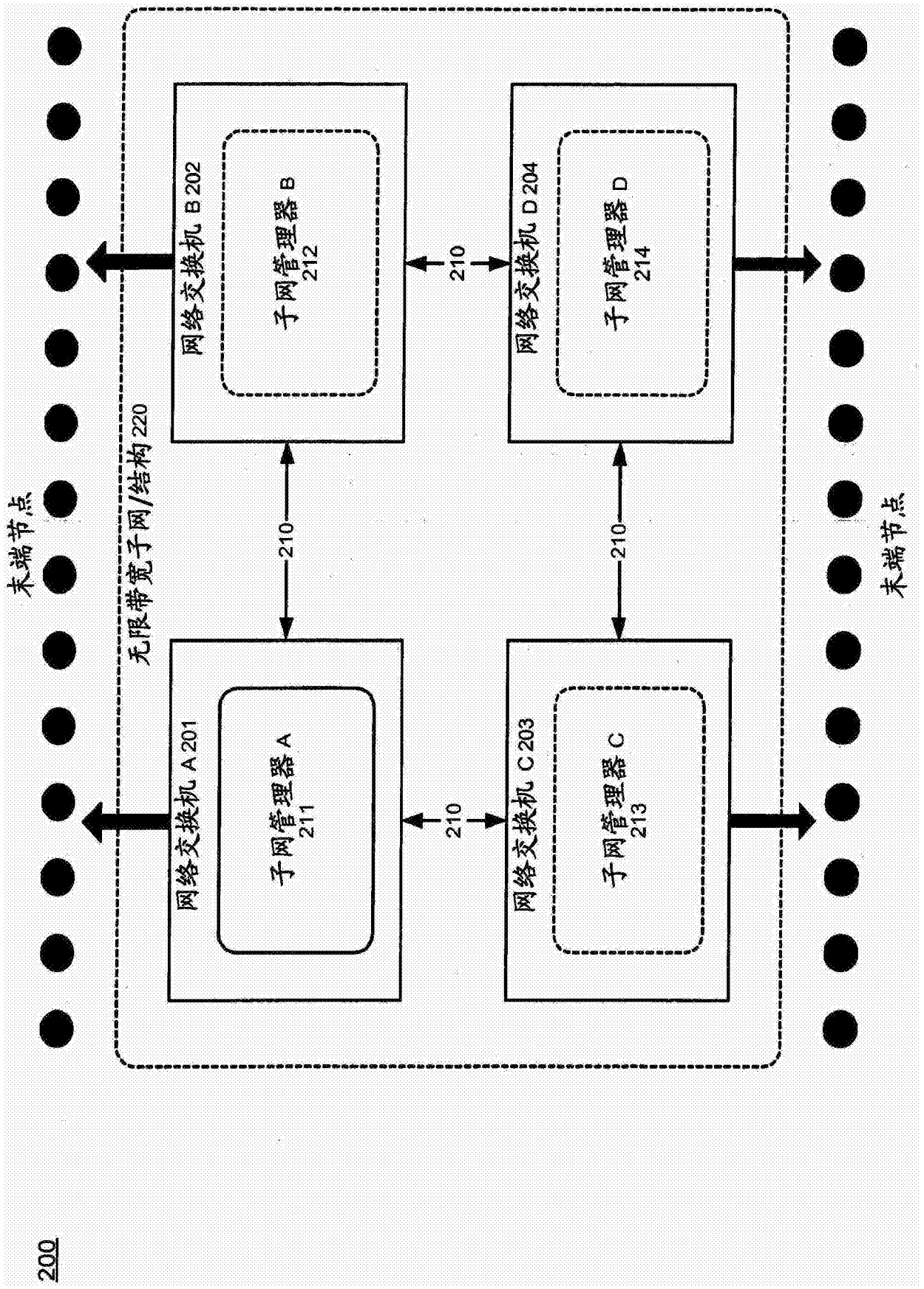

200
无限带宽子网/结构 220
网络交换机 A 201
子网管理器 A 211
网络交换机 B 202
子网管理器 B 212
网络交换机 C 203
子网管理器 C 213
网络交换机 D 204
子网管理器 D 214
210
210
210
210
末端节点
末端节点

图 2

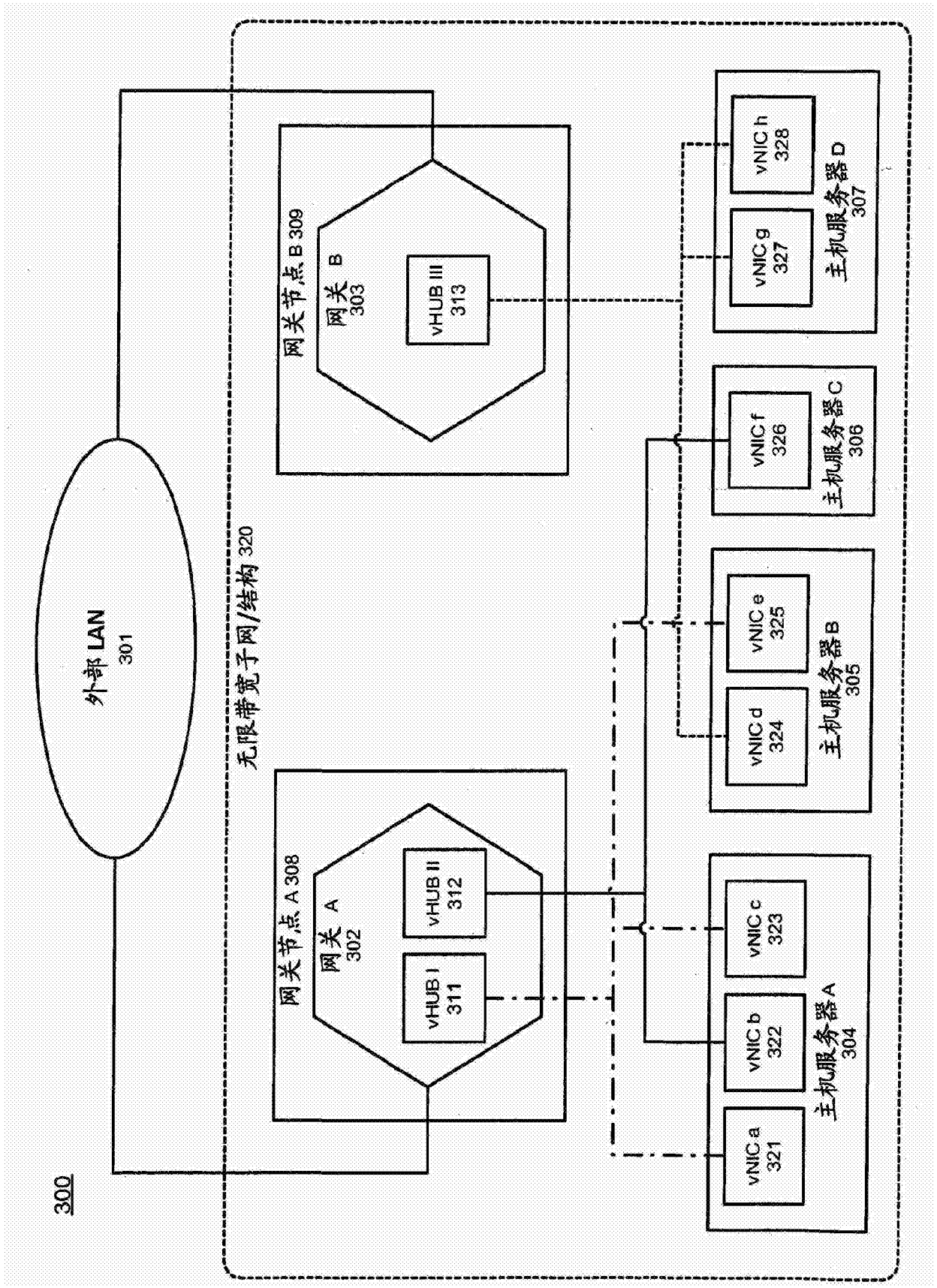

300
外部 LAN
301
无限带宽子网/结构 320
网关节点 A 308
网关 A
302
vHUB I
311
vHUB II
312
网关节点 B 309
网关 B
303
vHUB III
313
主机服务器 A
304
vNIC a
321
vNIC b
322
vNIC c
323
主机服务器 B
305
vNIC d
324
vNIC e
325
主机服务器 C
306
vNIC f
326
主机服务器 D
307
vNIC g
327
vNIC h
328

图 3

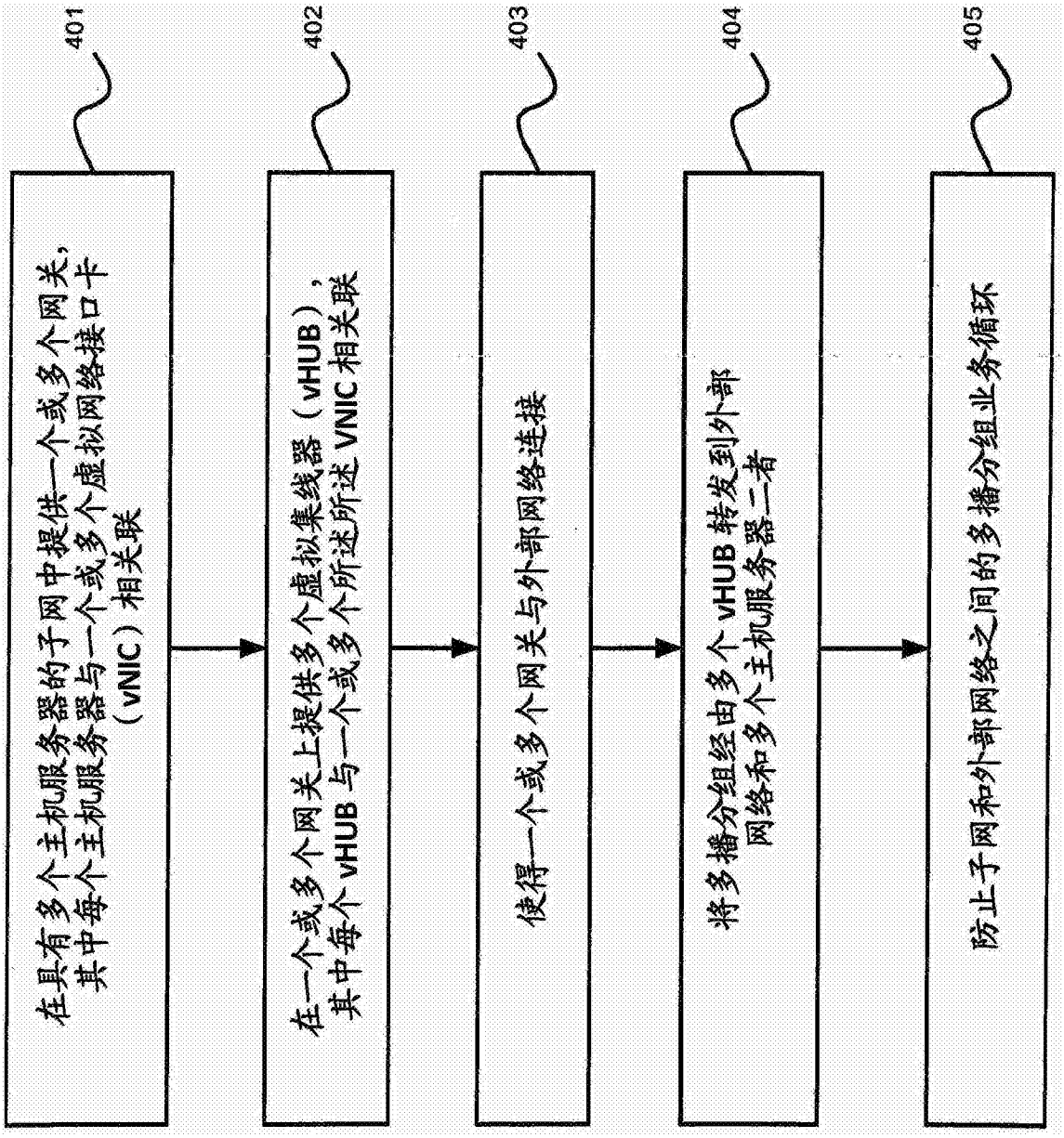

401
在具有多个主机服务器的子网中提供一个或多个网关，其中每个主机服务器与一个或多个虚拟网络接口卡（vNIC）相关联
402
在一个或多个网关上提供多个虚拟集线器（vHUB），其中每个 vHUB 与一个或多个所述 VNIC 相关联
403
使得一个或多个网关与外部网络连接
404
将多播分组经由多个 vHUB 转发到外部网络和多个主机服务器二者
405
防止子网和外部网络之间的多播分组业务循环

图 4

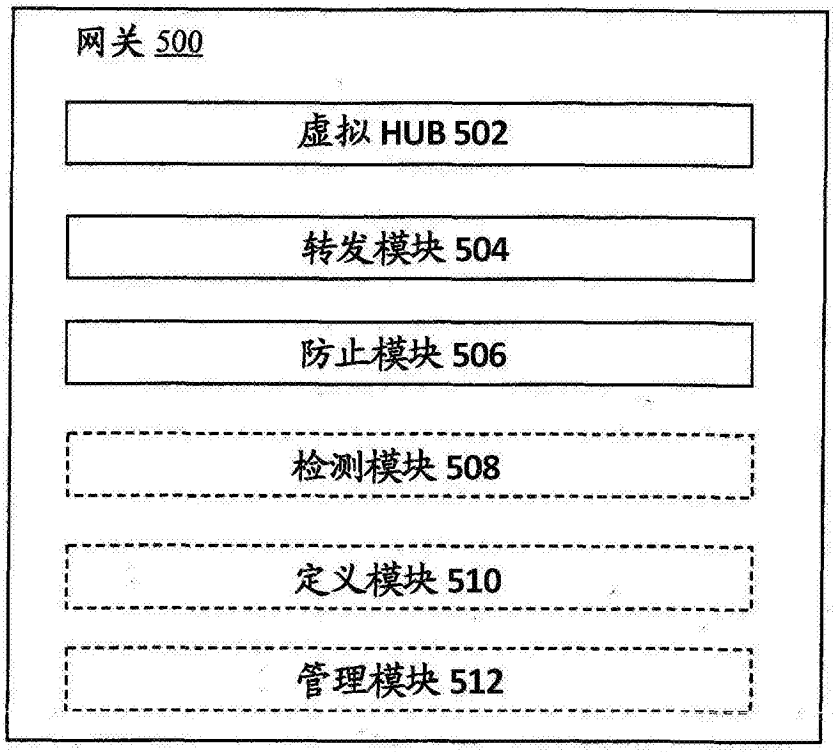

网关 500
虚拟 HUB 502
转发模块 504
防止模块 506
检测模块 508
定义模块 510
管理模块 512

图 5

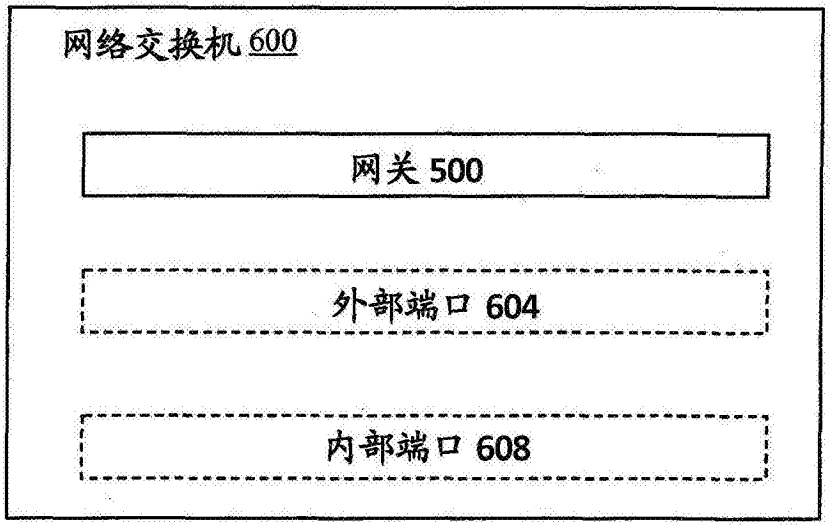

网络交换机 600
网关 500
外部端口 604
内部端口 608

图 6