

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

H04L 12/56 (2006.01)

H04L 12/46 (2006.01)

H04L 12/24 (2006.01)



[12] 发明专利申请公开说明书

[21] 申请号 200480020880.3

[43] 公开日 2006年8月30日

[11] 公开号 CN 1826769A

[22] 申请日 2004.9.8

[21] 申请号 200480020880.3

[30] 优先权

[32] 2003.9.18 [33] US [31] 10/666,306

[86] 国际申请 PCT/US2004/029553 2004.9.8

[87] 国际公布 WO2005/029784 英 2005.3.31

[85] 进入国家阶段日期 2006.1.19

[71] 申请人 思科技术公司

地址 美国加利福尼亚州

[72] 发明人 迈克尔·史密斯 艾力·戈尔杉
杰弗里·伊·王 尼利马·梅塔
文卡特斯·扎拉克拉曼

[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
代理人 王 怡

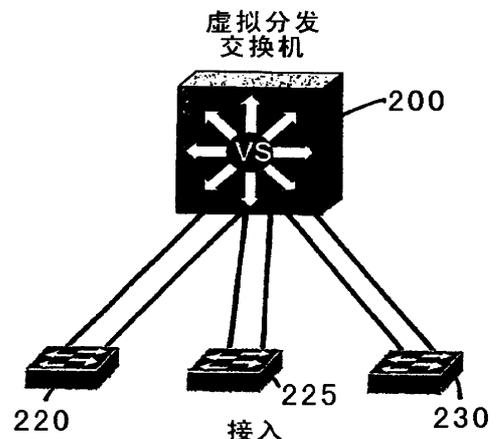
权利要求书 5 页 说明书 15 页 附图 8 页

[54] 发明名称

虚拟网络设备

[57] 摘要

提供了用于形成数据网络的虚拟交换机的方法和设备。如上所述，这里使用的术语“交换机”将应用于交换机、路由器和类似的网络设备。每个虚拟交换机充当单个逻辑单元，同时包含至少两个物理机箱。因此，每个虚拟交换机可以被视为单个管理点。每个虚拟交换机包括主机箱和至少一个从机箱。主机箱被配置用于控制从机箱。主机箱包括至少一个主监管器卡，而从机箱包括至少一个从监管器卡。主机箱和从机箱根据虚拟交换链路协议而经由虚拟交换链路通信。



1. 一种用于网络的虚拟交换机，该虚拟交换机包括：
主机箱，包括：
 - 5 第一多个线路卡；以及
用于控制所述第一多个线路卡的主监管器卡；以及
在所述主监管器卡控制下的从机箱，该从机箱包括：
第二多个线路卡；以及
从监管器卡；以及
- 10 用于在所述主机箱和所述从机箱之间进行通信的虚拟交换链路。
 2. 如权利要求 1 所述的虚拟交换机，其中所述主机箱和所述从机箱根据虚拟交换链路协议进行通信，所述虚拟交换链路协议用于在逻辑上将所述主机箱的数据平面扩展到所述从机箱的数据平面。
 3. 如权利要求 1 所述的虚拟交换机，其中所述虚拟交换链路包括控制
15 虚拟交换链路和数据虚拟交换链路。
 4. 如权利要求 1 所述的虚拟交换机，其中所述虚拟交换链路包括组合构成一个逻辑链路的多个物理链路。
 5. 如权利要求 2 所述的虚拟交换机，其中所述虚拟交换链路协议包括指示分组是否已经穿越所述虚拟交换链路的字段。
 - 20 6. 如权利要求 2 所述的虚拟交换机，其中所述虚拟交换链路被用于同步所述主机箱和所述从机箱的路由表。
 7. 如权利要求 3 所述的虚拟交换机，其中所述控制虚拟交换链路扩展内部的带外信道以在所述主机箱和所述从机箱之间进行通信。
 8. 如权利要求 3 所述的虚拟交换机，其中所述数据虚拟交换链路扩展
25 内部的机箱数据平面以在所述主机箱和所述从机箱之间进行通信。
 9. 如权利要求 3 所述的虚拟交换机，其中所述主监管器经由所述控制虚拟交换链路上的带内消息与所述从监管器通信。
 10. 如权利要求 3 所述的虚拟交换机，其中所述控制虚拟交换链路被首先联机，并被用于确定哪个机箱将是主机箱。

11. 如权利要求 3 所述的虚拟交换机，其中单个物理链路将所述控制虚拟交换链路和所述数据虚拟交换链路组合起来。

12. 如权利要求 3 所述的虚拟交换机，其中所述控制虚拟交换链路和所述数据虚拟交换链路是由分离的物理链路构成的。

5 13. 一种被配置用于控制网络的虚拟交换机的主机箱，该主机箱包括：

第一多个线路卡；以及

10 用于控制所述第一多个线路和从机箱的主监管器卡，该主监管器卡经由虚拟交换链路协议与所述从机箱通信，所述虚拟交换链路协议在逻辑上将所述主机箱的数据平面扩展到所述从机箱的数据平面。

14. 一种用网络中的多个物理交换机构成虚拟交换机的方法，该方法包括：

将第一物理交换机配置为用于控制所述虚拟交换机的主交换机；

将第二物理交换机配置为在所述主交换机的控制下的从交换机；

15 在所述主交换机和所述从交换机之间形成用于通信的虚拟交换链路；
以及

致使所述主交换机和所述从交换机经由虚拟交换链路协议进行通信。

15. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议包括源端口标识符。

20 16. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议包括目的地端口索引。

17. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议包括源泛滥信息。

25 18. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议包括 VLAN 信息。

19. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议指示访问控制列表是否应被应用于帧。

20. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议指示 QoS 指定是否应被应用于帧。

21. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议指示帧是否是 MAC 通知帧。

22. 如权利要求 14 所述的方法，其中所述虚拟交换链路协议包括用于帧的数据平面优先级信息。

5 23. 如权利要求 14 所述的方法，还包括根据所述主交换机和所述从交换机之间的通信经由所述虚拟交换链路协议来扩展所述主交换机的第一数据平面，以包括所述从交换机的第二数据平面。

24. 如权利要求 14 所述的方法，还包括用充当单个逻辑链路的多个物理链路构成所述虚拟交换链路。

10 25. 如权利要求 14 所述的方法，还包括形成所述虚拟交换链路以包括数据虚拟交换链路和控制虚拟交换链路。

26. 如权利要求 14 所述的方法，还包括：

更新所述主机箱中的第 2 层转发表；

更新所述从机箱中的第 2 层转发表；以及

15 校正所述主机箱中的第 2 层转发表和所述从机箱中的第 2 层转发表之间的不一致。

27. 如权利要求 25 所述的方法，其中形成所述虚拟交换链路的步骤包括将所述数据虚拟交换链路和所述控制虚拟交换链路组合在单个物理链路上。

20 28. 如权利要求 25 所述的方法，还包括：

更新所述主机箱中的第 2 层转发表；

更新所述从机箱中的第 2 层转发表；以及

根据在所述数据虚拟交换链路上发送的帧，校正所述主机箱中的第 2 层转发表和所述从机箱中的第 2 层转发表之间的不一致。

25 29. 如权利要求 28 所述的方法，其中所述帧是 MAC 通知帧。

30. 一种用于在网络的分发层或核心层中用多个物理交换机构成虚拟交换机的装置，该装置包括：

用于将第一物理交换机配置为用于控制所述虚拟交换机的主交换机的装置；

用于将第二物理交换机配置为在所述主交换机的控制下的从交换机的装置；

用于在所述主交换机和所述从交换机之间形成用于通信的虚拟交换链路的装置；以及

- 5 用于致使所述主交换机和所述从交换机经由虚拟交换链路协议通信的装置，所述虚拟交换链路协议在逻辑上将所述主交换机的数据平面扩展到所述从交换机的数据平面。

31. 一种包含在机器可读介质中的计算机程序，所述计算机程序包含用于控制网络的多个物理交换机以执行以下步骤的指令：

- 10 将第一物理交换机配置为用于控制所述虚拟交换机的主交换机；
将第二物理交换机配置为在所述主交换机的控制下的从交换机；
在所述主交换机和所述从交换机之间形成用于通信的虚拟交换链路；
以及

致使所述主交换机和所述从交换机经由虚拟交换链路协议通信。

- 15 32. 如权利要求 31 所述的计算机程序，还包括用于控制网络的多个物理交换机以在逻辑上将所述主交换机的数据平面扩展到所述从交换机的数据平面的指令。

33. 一种初始化虚拟网络设备的方法，包括：

- 20 在第一机箱和第二机箱之间执行握手序列，所述第一机箱和所述第二机箱是数据网络的冗余网络设备；以及

判断是所述第一机箱还是所述第二机箱将是主机箱，该主机箱用于控制包含所述第一机箱和所述第二机箱的虚拟网络设备。

- 25 34. 如权利要求 33 所述的方法，其中所述握手序列包括交换从以下群组中选出的信息，所述群组包括：监管器的硬件版本；机箱标识符；机箱号；机箱中的每个监管器的软件版本；机箱中的插槽的硬件值；以及针对所述第一机箱和所述第二机箱之间的特定链路的远程端点的插槽/端口。

35. 如权利要求 33 所述的方法，还包括如下步骤：根据在所述握手序列期间交换的信息来形成所述虚拟网络设备的控制虚拟交换链路。

36. 如权利要求 35 所述的方法，还包括如下步骤：确定将成为数据虚

拟交换链路的物理链路是否既连接到所述第一机箱又连接到所述第二机箱。

37. 如权利要求 36 所述的方法，还包括如下步骤：如果所述确定步骤指示所述物理链路既连接到所述第一机箱又连接到所述第二机箱，则形成
- 5 所述虚拟网络设备的数据虚拟交换链路。

虚拟网络设备

5 技术领域

本发明涉及数据网络。更具体而言，本发明涉及网络拓扑。

背景技术

在大多数企业网络中，采用的是分层（tiered）网络设计，在各个层上
10 具有冗余网络设备。典型的网络涉及如图 1 所示。核心层 105 可被连接到
数据中心 110 和/或因特网 115。核心层 105 一般包括 2 个交换机，其中每
个交换机与分发层 120 中的每个设备相连以用于冗余目的。（这里所使用的
术语“交换机”将用于指实际的交换机、路由器或任何类似网络设备。）
15 类似地，在接入层 125 中的每个配线柜通常连接到分发层 120 的两个
设备。

虽然已证明这种网络拓扑非常健壮（robust）而可靠，但它具有某些
缺点。例如，每对冗余交换机代表两个管理点。这意味着与配置单个设备
所需的相比，配置第 2 层和第 3 层协议、生成树协议等等需要花费双倍
的时间和努力。此外，每个配线柜必须针对分发层中的两个冗余设备中的
20 每个设备的上行链路来配置。

将需要形成这样的网络，该网络保持了传统网络拓扑的健壮性质，但
更易于管理。

发明内容

25 提供了用于形成数据网络的虚拟交换机的方法和设备。如上所述，这
里使用的术语“交换机”将应用于交换机、路由器和类似的网络设备。每
个虚拟交换机充当单个逻辑单元，同时包含至少两个物理机箱。因此，每
个虚拟交换机可以被视为单个管理点。每个虚拟交换机包括主机箱和至少
一个从机箱。主机箱被配置用于控制从机箱。主机箱包括至少一个主监管

器卡，而从机箱包括至少一个从监管器卡。主机箱和从机箱根据虚拟交换链路协议而经由虚拟交换链路通信。

本发明的某些实现方式提供了用于网络的分发层或核心层的虚拟交换机。虚拟交换机包括具有第一多个线路卡和用于控制第一多个线路卡的主
5 监管器卡的主机箱。虚拟交换机包括在主监管器卡的控制下的从机箱，该从机箱具有第二多个线路卡和从监管器卡。虚拟交换机还包括用于在主机箱和从机箱之间通信的虚拟交换链路（“VSL”）。VSL 可以包括组合构成一个逻辑链路的多个物理链路。

主机箱和从机箱可以根据虚拟交换链路协议通信，该虚拟交换链路协
10 议用于在逻辑上将主机箱的数据平面（构架、数据总线等）扩展到从机箱的数据平面。

在优选实施例中，VSL 包括控制虚拟交换链路（“CVSL”）和数据
虚拟交换链路（“DVSL”）。在某些实施例中，CVSL 和 DVSL 在同一物理链路上。CVSL 可以扩展内部的带外信道（“OBC”）以用于主机箱
15 和从机箱之间的通信。DVSL 可以扩展内部的机箱数据平面以用于主机箱和从机箱之间的通信。主监管器可以经由 CVSL 上的带内消息与从监管器通信。CVSL 优选地被首先联机，并且可被用于确定哪个机箱将是主机箱。

VSL 协议优选地包括指示分组是否已经穿越虚拟交换链路的字段。穿
20 越 VSL 的 OBC 可被用于同步主机箱和从机箱的路由表。DVSL 或 CVSL 可被用于同步第 2 层表。在某些实施例中，单个物理链路将控制虚拟交换链路和数据虚拟交换链路组合起来。在替换实施例中，控制虚拟交换链路和数据虚拟交换链路由分离的物理链路构成。

本发明的某些实施例提供了被配置用于控制网络的分发层或核心层的
25 虚拟交换机的主机箱。主机箱包括第一多个线路卡和用于控制第一多个线路卡和从机箱的主监管器卡。主监管器卡经由 VSL 协议与从机箱通信，VSL 协议在逻辑上将主机箱的数据平面扩展到从机箱的数据平面。

本发明的某些方面提供了用网络中的分发层或核心层中的多个物理交
换机构成虚拟交换机的方法。该方法包括以下步骤：将第一物理交换机配

置为用于控制虚拟交换机的主交换机；将第二物理交换机配置为在主交换机的控制下的从交换机；在主交换机和从交换机之间形成用于通信的 VSL；以及致使主交换机和从交换机经由 VSL 协议通信。

5 虚拟交换链路协议可以包括源端口标识符、目的地端口索引、源泛滥信息或 VLAN 信息。虚拟交换链路协议可以指示访问控制列表或 QoS 指定是否应被应用到帧。虚拟交换链路协议可以指示帧是否是 MAC 通知帧，并且可以包括用于帧的数据平面优先级信息。

10 主交换机的第一数据平面可根据经由虚拟交换链路协议在主交换机和从交换机之间的通信被扩展，以包括从交换机的第二数据平面。虚拟交换链路可以由充当单个逻辑链路的多个物理链路构成。虚拟交换链路可以被形成为包括数据虚拟交换链路和控制虚拟交换链路。数据虚拟交换链路和控制虚拟交换链路可以被形成在单个物理链路上。

15 该方法还可以包括以下步骤：更新主机箱和从机箱中的第 2 层转发表，并且校正主机箱和从机箱中的第 2 层转发表之间的不一致。该不一致可以根据在数据虚拟交换链路上发送的帧来校正。所述帧可以是媒体访问控制（“MAC”）通知帧。本发明的某些实现方式提供了包含在机器可读介质中的计算机程序。该计算机程序包含用于控制网络的分发层或核心层中的多个物理交换机以执行以下步骤的指令：将第一物理交换机配置为用于控制所述虚拟交换机的主交换机；将第二物理交换机配置为在所述主交换机的控制下的从交换机；在所述主交换机和所述从交换机之间形成用于通信的 VSL；以及致使所述主交换机和所述从交换机经由 VSL 协议通信。

20 计算机程序可以包括用于控制网络的多个物理交换机以在逻辑上将主交换机的数据平面扩展到从交换机的数据平面的指令。

25 本发明的其他实现方式提供了用于初始化虚拟网络设备的方法。该方法包括以下步骤：在第一机箱和第二机箱之间执行握手序列，所述第一机箱和所述第二机箱是数据网络的冗余网络设备；以及判断是所述第一机箱还是所述第二机箱将是主机箱，该主机箱用于控制包含所述第一机箱和所述第二机箱的虚拟网络设备。该方法还可以包括以下步骤：根据在握手序

列期间交换的信息来形成虚拟网络设备的控制虚拟交换链路。

握手序列可以包括交换以下信息中的任意信息：监管器的硬件版本；机箱标识符；机箱号；机箱中的每个监管器的软件版本；机箱中的插槽的硬件值；以及针对所述第一机箱和所述第二机箱之间的特定链路的远程端
5 点的插槽/端口。

该方法可以包括以下步骤：确定将成为数据虚拟交换链路的物理链路是否既连接到所述第一机箱又连接到所述第二机箱。该方法还可以包括以下步骤：如果所述确定步骤指示所述物理链路既连接到所述第一机箱又连接到所述第二机箱，则形成所述虚拟网络设备的数据虚拟交换链路。

10

附图说明

图 1 是示出传统网络拓扑的网络图。

图 2A 和 2B 提供了构成虚拟交换机的简化图示。

图 2C 是示出根据某些实现方式，在接入层设备和分发层设备之间发
15 送的帧的简化版本的框图。

图 3 示出了虚拟交换机的最小连接配置。

图 4 示出了虚拟交换机的更健壮连接配置。

图 5 示出了示例性的数据虚拟交换链路和虚拟交换机的中等健壮的连接配置。

20 图 6 是示出根据虚拟交换链路协议的一种实现方式的帧头部的简化版本的框图。

图 7 示出了可被配置用于实现本发明的某些方法的网络设备的简化版本。

25 具体实施方式

在以下描述中，提出了多个具体细节以提供对本发明的全面理解。但是，本领域技术人员将会发现，无需这些具体细节中的某些或全部也可以实现本发明。在其他实例中，没有详细描述公知的过程步骤，以免不必要地模糊本发明。

虚拟交换机概述

在虚拟交换机内，只有 1 个主监管器。主监管器向用户提供单个管理点。包含主监管器的机箱（chassis）被称为主机箱。构成虚拟交换机的其他机箱被称为从机箱。从机箱中的活动监管器将充当主监管器的从属，并且在主监管器发生故障时用作主监管器的备用。如果在机箱中只有 1 个监管器，则整个机箱将在发生故障时失效，但是虚拟交换机将继续工作，就好像在发生故障的机箱上的那些端口在经历热插拔（“OIR”）事件一样。位于两个机箱上的所有接口将像一个大交换机一样呈现给用户。端口寻址是单个全局空间，其中虚拟交换机内的每个第 2 层（“L2”）接口具有唯一的端口索引。

虚拟交换机的软件图像应该被配置在主监管器上，并被下载到所有其他的监管器。这确保了整个虚拟交换机将一直运行相同的软件图像版本。

虚拟交换机的示例性实施例

图 2A 是示出根据本发明某些实施例的虚拟交换机 200 的高层物理视图的网络图。在本实施例中，虚拟交换机 200 包括分发层交换机 205 和 210，它们经由虚拟交换链路 215 进行通信。在某些优选实施例中，虚拟交换链路 215 是被组合构成一个逻辑链路的多个物理链路。根据某些这样的实施例，虚拟交换链路 215 是根据专门的虚拟交换链路协议运行的以太网信道端口束。接入层设备 220、225 和 230 物理上连接到分发层交换机 205 和 210 中的每一个。

图 2B 示出了虚拟交换机 200 的逻辑图。接入层设备 220、225 和 230 虽然在物理上连接到分发层交换机 205 和 210 中的每一个，但它们与虚拟交换机 200 之间的交互就好像虚拟交换机 200 是单个网络设备一样。虚拟交换机 200 外部的所有设备都将其视为单个网络设备。在第 3 层，虚拟交换机 200 充当到接入层 125 和核心层 105 的单个路由器。类似地，在第 2 层，虚拟交换机 200 充当到接入层 125 和核心层 105 的单个交换机。分发层交换机 205 和 210 的两个配置点可以被看作单个配置点。

虽然在上述示例中虚拟交换机 200 被形成在分发层上，但是虚拟交换机 200 也可以形成在网络的其他部分中，例如在核心层上。此外，根据某

些实施例，形成在分发层上的虚拟交换机 200 还包括接入层 125 中的设备。根据某些这样的实施例，与虚拟交换机 200 通信的接入层设备不执行独立的转发判决（关于访问控制列表（“ACL”）、服务质量（“QoS”）等）。在这样的实施例中，接入层设备充当对分发层设备的
5 远程静默线路卡（有时被称为“卫星”）。因此，虚拟交换机 200 既可以包括分发层中的设备，也可以包括接入层中的设备。虚拟交换机 200 的这种实施例创建了用于接入层和分发层两者的单个管理点。

图 2C 示出了在卫星和分发层中的设备之间的通信所使用的帧的示例性格式。这里使用的术语“分组”和“帧”具有相同的意义。每个字段的大小不一定与每个字段中的字节数目相对应。这里示出和描述的字段类型纯粹是示例性的。此外，每个字段可以包含一个或多个子字段，其中某些子字段可以被预留。

字段 235 包括目的地信息，该目的地信息例如可以指示由端口索引表使用的目的地索引、泛滥（flooding）是否已被使能、VLAN 信息等等。
15 字段 240 包括源信息，例如由端口索引表使用的源索引。字段 245 包括状态、控制和帧类型信息。字段 250 包括服务种类和服务类型信息。字段 255 可以包括诸如束散列信息和转发信息之类的信息。

字段 260 包括各种控制信息。字段 260 例如可以包括关于输入或输出 QoS 或 ACL 是否应该被应用到帧的信息。字段 260 的某些子字段优选地被预留。字段 265 包括有效载荷，而字段 270 是 CRC 字段。
20

卫星交换机通过向简单交换机提供非本地转发能力来执行对配线柜的管理。所有转发智能性都由上游交换机提供。卫星需要到 2 个上游交换机的上行链路链接以用于冗余目的。

如果没有虚拟交换机，则需要将某些卫星端口分配给一个分发交换机，而将某些卫星端口分配给另一分发交换机。如果这 2 个分发交换机是
25 独立的，则配置可以不同，并且使卫星端口的配置同步而不是使本地端口的配置同步引入了大量复杂性。另外，对两个交换机的全局配置可以不同，因此同步将导致本地端口出现问题，而全局信息不同步将导致卫星端口在切换时出现问题。虚拟交换机模型允许配线柜卫星连接到 1 个虚拟交

交换机。由于整个虚拟交换机只存在一种配置，因此解决了连接到多个分发交换机时遇到的所有问题。

虚拟交换链路

5 虚拟交换机的一个重要特征是虚拟交换链路（“VSL”）。VSL 是作为同一虚拟交换机的一部分的至少两个机箱之间的点到点连接。穿过 VSL 的分组例如以预置头部的形式提供附加信息。在 VSL 中传递的最重要的几条信息的其中一条是源端口索引。这使对等机箱上的转发引擎能够得知原始进入端口。它还向主监管器指示当分组被提供到软件时从机箱上的进入物理端口。

10 在优选实施方式中，源于 VSL 的分组不包含 VSL 端口索引。因此，这种实现方式需要附加的硬件机制，以确保分组不会在构成 VSL 的多个物理链路之间循环。在某些这样的实现方式中，每个进入 VSL 端口的端口 ASIC 将在分组头部标记一个 VSL 位。在每个外出 VSL 端口上，端口 ASIC 检查分组头部，如果设置了 VSL 位，则不将分组重传出 VSL。

15 VSL 不仅用于数据流量，也用于内部控制流量。通常不会被发送到机箱外部的分组被准许到 VSL 外部去，以便使主机箱和从机箱能够构成单个虚拟交换机。主监管器 CPU 和在从机箱上的 CPU 之间的通信通过 VSL 而发生在带内。分组根据 VSL 协议被发送。示例性的 VSL 协议将在下面参考图 6 描述。

20 软件图像通过这种带内通信被从主监管器分发到从机箱。诸如 OIR 事件和端口事件（例如链路启动/关闭）之类的附加信息也经由这种带内机制传播。通常，执行某些机制（即头部中的“不转发”位）来确保这样的信息不会被传播到机箱外。在 VSL 上，这些机制被禁止（即在 VSL 上，设置了“不转发”位的分组允许到机箱外）。

25 优选地，第 3 层（“L3”）转发表经由来自主监管器的带内 CPU 通信来填充。L2 转发表优选地通过硬件学习来填充。泛滥的分组将被泛滥到属于虚拟交换机内的 VLAN 的所有端口。组播分组将被发送到已经加入该组的所有端口，无论该端口位于哪个机箱。

由于学习是通过分布式硬件完成的，因此分布式转发引擎（FE）在

丢失同步时需要 L2 转发表的校正。这是以 MAC 通知帧的形式执行的，该 MAC 通信帧从外出 FE 发送到进入 FE。利用 VSL 的引入，MAC 通知被扩展到机箱外部。这只在 VSL 链路上是允许的。在某些实现方式中，帧大小被扩展到 64 字节以符合最小以太网标准大小。通常，MAC 通知只由外出 FE 生成。在虚拟交换机中，如果进入端口是 VSL 的一部分，则 MAC 通知也可以在进入 FE 查找时生成。MAC 通知允许遍及整个虚拟交换机的硬件学习。

VSL 在两个机箱之间运载控制和数据流量。它用于为数据流量在对等机箱之间扩展内部数据平面（构架、数据总线等等），并且为了控制流量（即 IPC、SCP 流量）而扩展内部 OBC。在本发明的某些优选实现方式中，VSL 被分离成控制 VSL（“CVSL”）和数据 VSL（“DVSL”）。CVSL 和 DVSL 可以是分离的物理链路，或者可以被组合在相同的物理链路上。

由于它们的设计和实现方式的本质，大多数路由协议被设计为工作在单个 CPU 内。在虚拟交换机内，可以存在多个监管器卡。在多个监管器之间，将选择出一个监管器来运行整个 VS 的所有路由协议。

需要被发送到从机箱本地端口外部的 L3 和 L2 控制分组（即 OSPF LSA、生成树 BPDU 等等）将经由数据虚拟交换链路（“DVSL”）被发送到从机箱，如下面详细描述。到从机箱的所有控制信息都将经由控制虚拟交换链路（“CVSL”）被发送到从机箱。与 L2 和 L3 控制协议一道，软件数据路径可以以集中方式运行在活动的主监管器上。在替换实施例中，软件数据路径可以以分布的方式运行。

根据某些优选实现方式，内部的带外信道（OBC）被用于机箱内的卡之间的控制软件通信。在这样的实现方式中，CVSL 被用于将内部 OBC 扩展到远程机箱。DVSL 被用于将内部机箱数据平面扩展到远程机箱。

CVSL 被用于 CPU 到 CPU 的通信。另外，CVSL 被用于与对等机箱通信以确定其主人角色（mastership role）。机箱的主人角色极大地影响软件的行为。因此，主人角色应该在机箱启动引导时大多数软件应用开始运行之前尽早确定。为了避免初始化任意线路卡并缩短引导时间，应该将

CVSL 限制于物理上位于监管器卡上的端口。

CVSL 在确定主人身份之前被使用。这推断出，由于主人是未知的，因此用于启动引导的适当配置文件是未知的。由于正确的配置文件是未知的，因此 CVSL 端口应该是公知端口。因此，每个监管器卡的前 2 个上行
5 链路被优选地用作控制 VSL。当机箱上存在 2 个监管器时，本地备用监管器上的端口将可被本地的活动监管器所访问。

主机箱监管器将经由带内消息与其在从机箱上的对等体通信。这种带内消息通过 CVSL 发送。带内消息可以被引导至从机箱监管器或从机箱上的线路卡。

10 当消息正被发送到从机箱上的线路卡时，它将被从监管器所代理。从监管器代表主机箱监管器将消息经由 OBC 发送到线路卡。

在内部，CVSL 优选地被实现为端口束，并且在全网状系统中将易于从单个链路的故障中恢复。但是，由于散列算法优选地将基于源和目的地 MAC 地址，而对于带内 CPU 通信来说，源和目的地 MAC 地址总是相同的，因此只需要一条链路。
15

到远程机箱的线路卡的通信将经由 CVSL 来发送，并被远程机箱上的活动监管器所终止。然后，该监管器将消息经由 OBC 发送到线路卡，并代理发回原始机箱的响应。

在某些实施例中，每个机箱可以具有一个或两个监管器。由于只有一条 CVSL 物理链路是强制性的，因此导致 CVSL 存在多种硬件部署场景。
20

图 3 示出了最小硬件部署场景。在场景 300 中，主机箱 305 具有单个监管器 315 和多个线路卡 325。类似地，从机箱 310 具有单个监管器 320 和多个线路卡 330。链路 335 形成在监管器 315 的端口 333 和监管器 320 的端口 340 之间。在本实施例中，链路 335 将 CVSL 和数据虚拟交换链路
25 (“DVSL”) 组合在一条物理链路中。以下将描述 DVSL 的示例性属性。

场景 300 的主要优点在于最小成本和最小配置时间。缺点包括缺乏冗余性，这是由于任何组件的故障都会导致断供 (outage)。因此，场景 300 不能像根据本发明的虚拟交换机的其他配置一样健壮。

在图 4 中示出了更健壮的部署场景 400。主机箱 405 包括主监管器 415、备用监管器 417 和线路卡 425。从机箱 410 包括从监管器 420、备用从监管器 422 和线路卡 430。

在此场景中，两个机箱都包含两个监管器，并且在监管器之间存在 4 条物理链路：链路 435 连接端口 440 和 445；链路 450 连接端口 455 和 460；链路 465 连接端口 470 和 475；链路 480 连接端口 485 和 490。

部署场景 400 的优点在于其具有比部署场景 300 更大的冗余性。由于在两个机箱上都具有监管器冗余，因此利用 4 个监管器中的每一个之间的物理链路，可以创建健壮得多的物理链路场景：部署场景 400 使虚拟交换机即使在 3 个监管器发生故障之后也能够工作。部署场景 400 的缺点包括更高的成本和更长的配置时间。

控制 VSL 初始化

在虚拟交换机可以变得活动之前，必须使 CVSL 联机。根据一个优选实施例，以下各项作为初始握手序列的一部分被传递：监管器的硬件版本；机箱标识符（例如来自机箱背板的 MAC 地址）；机箱号；机箱中的每个监管器的软件版本；机箱中的每个插槽的硬件（例如 EEPROM）值；以及针对特定 CVSL 链路的远程端点的插槽/端口。

以上各项中的大多数项被用于确定主人身份。直接影响 CVSL 的两项是机箱标识符和针对特定 CVSL 链路的远程端点的插槽/端口。

插槽/端口

插槽/端口被用于确定多种可能的 CVSL 硬件部署场景中的哪种部署场景正在使用。在优选实现方式中，该信息还被用于在必要时生成警报。

机箱标识符

机箱标识符被用于确定是否所有 CVSL 链路都被连接到同一机箱。如果该配置发生错误，机箱将优选地选择对等机箱中与其构成虚拟交换机的那个机箱。优选地，机箱随后将在管理上切断未连接到所选中的对等机箱的所有 CVSL 链路。初始化序列优选地将通过剩余的 CVSL 与被选中的对等机箱进行协商，以避免当存在多个错误配置时无法形成虚拟交换机的情况发生。

通过 CVSL 执行的初始化握手利用了公知的目的地 MAC 地址。这允许分组被重定向到 CPU，并确保初始化握手分组不会被视作数据分组并被转发到虚拟交换机外部。初始化握手分组被重定向到主监管器（如果可获得的话）或被重定向到本地监管器，主监管器和本地监管器两者中的任意一个都可以基于源端口索引来确定错误场景。在检测到配置错误之后，5 优选地向用户发布警报，并在管理上切断链路的两端。初始握手序列将在链路启动/关闭事件发生时被优选地重新触发。

数据 VSL

在将 CVSL 联机之后，主人身份被确定并形成了虚拟交换机，则必须10 使数据 VSL (DVSL) 联机。DVSL 是内部数据平面的扩展，并被用于在虚拟交换机的机箱之间转发分组。CVSL 优选地不被用于任意用户数据流量。

图 5 示出了 CVSL 和 DVSL 的一种示例性配置。CVSL 链路 515 将主机箱 501 中的主监管器 505 的端口 520 与从机箱 502 中的从监管器 510 的15 端口 525 相连。CVSL 链路 530 将主机箱 501 中的主监管器 505 的端口 535 与从机箱 502 中的备用从监管器 542 的端口 540 相连。DVSL 链路 545 将主机箱 501 中的线路卡 550 与从机箱 502 中的线路卡 555 相连。类似地，DVSL 链路 560 将主机箱 501 中的线路卡 565 与从机箱 502 中的线路卡 570 相连。

DVSL 的数据流量使用

根据本发明的某些优选实现方式，从非 DVSL 端口到达的分组将在以下情形中被发送出 DVSL 端口：（1）分组在 VLAN 上泛滥，并且在对等25 机箱上存在一个或多个承载该特定 VLAN 的端口；（2）分组想去往如下组播组，该组播组中的成员已经加入到对等机箱上的一个或多个端口上；（3）分组想去往如下 MAC 地址，该 MAC 地址已在对等机箱的端口上获知；或者（4）分组是想去往对等机箱上的端口的 MAC 通知帧。

在情形 1、2 和 3 中，分组通过 VSL 被发送，这是因为 VSL 是到达外出端口的唯一途径。在情形 4 中分组通过 VSL 被发送，因为它们是想去往对等机箱 EARL 的内部控制分组。注意，对于情形 1，给定分组可以通

过 VSL 被发送。

穿越 DVSL 的所有分组都将被封装以带内头部。在优选实现方式中，该头部将由外出端口（例如外出端口的 ASIC）附加到分组，并在 DVSL 的另一侧被进入端口剥离。带内头部携带了诸如进入端口索引、目的地端口索引、VLAN、CoS 等的信息。

图 6 示出了一种示例性头部格式。每个字段的大小不一定与每个字段中的字节数相对应。示出并描述的字段类型纯粹是示例性的。此外，每个字段可以包含一个或多个子字段，其中某些子字段可以被预留。这里，字段 605 包括服务种类信息，字段 610 指示帧的类型（例如以太网、ATM 等等）。字段 615 包括控制信息，例如端口索引表是否应该被更新，以及端口是否是“可信的”。字段 620 包括源 VLAN 信息，并指示输入或输出 ACL 或 QoS 是否应该被应用到该帧。

字段 625 包括源信息，例如源索引、源泛滥信息等等。字段 630 是帧长度字段。字段 635 是状态字段，该字段包括关于例如封装类型以及是否已正确接收到 CRC 的信息。字段 640 包括接收端口从帧中提取出的第 3 层信息。字段 645 指示该帧是否是 MAC 通知帧等等。字段 650 包括构架优先级位和退出端口（port-of-exit）位。字段 655 包括目的地信息，例如目的地索引和目的地泛滥信息。字段 655 中的信息可以由地址转发逻辑提供。字段 660 是 CRC 字段。

20 DVSL 的 MAC 通知的使用

除了一般的网络流量之外，DVSL 还被用于传输在某些优选实现方式中使用的 MAC 通知（MN）帧。MAC 地址表管理将优选地在每个机箱上独立发生。诸如静态 MAC 条目之类的项的配置将从主机箱提供，作为配置同步的一部分。

25 MAC 地址表管理中的大部分工作优选地以硬件形式执行，例如动态 MAC 条目的学习。由于某些网络设备的转发机制（例如 Cisco 的 Catalyst 6000™ 的 L2 转发）的分布式本质，在 ASIC 之间存在针对 MAC 地址表管理的内部通信。MAC 通知（MN）帧被用于校正在 L2 表中发现的任何不一致。

DVSL 被用户所配置，并且可以在支持带内头部的任意线路卡上。虽然 DVSL 可以仅操作单个物理链路，但是推荐的做法是，DVSL 被实现为多个物理链路，这些物理链路被组合以构成一个逻辑链路，例如作为多模块以太网信道。

- 5 DVSL 可以经由接口级配置命令来配置。在某些实现方式中，该命令可以仅在以太网信道接口上被配置。根据某些这样的实现方式，最多可以将 2 个以太网信道接口配置为 VSL。2 个被配置的以太网信道接口用于 VSL 的每一端。当命令被输入时，应该检查以太网信道成员端口以确保它们都是同一物理机箱的一部分。如果它们不是，则应该发布警报，并且该
- 10 命令应该被拒绝。DVSL 应该仅在 CVSL 启动时被配置，因为它要求 DVSL 的两端都被配置在活动的主监管器上。

在存在 2 条以太网信道被配置作为虚拟交换机中的 DVSL 之前，DVSL 初始化序列不应该开始。每个以太网信道代表 DVSL 的一端。在配置了 2 个 DVSL 端点之前，已配置的第一 DVSL 将优选地保持管理上切断

15 的状态。这将有助于确保初始化握手分组在从常规以太网信道到 DVSL 的短暂配置期间不会被发送并被理解为数据分组。

DVSL 初始化

DVSL 初始化序列优选地是 CVSL 初始化序列的子集。在优选实现方式中，初始化握手使用公知的目的地 MAC 地址以确保即使在错误配置

20 时，握手分组也将被发送到本地 CPU，而不会被交换到虚拟交换机外部。在初始化握手序列期间，所有分组都被重定向到主监管器 CPU。在从机箱上，这经由 CVSL 完成。这是速率高度受限的，并通过对等机箱中的本地活动监管器来代理。DVSL 初始化握手的目的主要在于检查各种 DVSL 特有的配置错误。

- 25 大多数配置错误已通过 CVSL 初始化握手被检查过。DVSL 初始化握手交换机箱标识符以确保所有 DVSL 链路都被连接在同样的两个机箱之间。常规的数据流量将穿越该 DVSL，直到初始化握手已成功完成为止。如果链路由于任何原因关闭，则初始化握手序列将再次开始，并且数据流量将不会穿越该链路，直到握手完成为止。

图 7 示出了可被配置用于实现本发明的某些方法的网络设备的示例。网络设备 760 包括主中央处理单元 (CPU) 762、接口 768 和总线 767 (例如 PCI 总线)。一般而言, 接口 768 包括适合于与适当媒体通信的端口 769。在某些实施例中, 一个或多个接口 768 包括至少一个独立的处理器 774, 并且在某些示例中还包括易失性 RAM。独立处理器 774 例如可以是 ASIC 或任意其他适当的处理器。根据某些这样的实施例, 这些独立处理器 774 执行这里描述的逻辑的至少某些功能。在某些实施例中, 一个或多个接口 768 控制诸如媒体控制和管理之类的通信密集型任务。通过为通信密集型任务提供分离的处理器, 接口 768 使主微处理器 762 能够高效地执行其他功能, 例如路由计算、网络诊断、安全性功能等等。

接口 768 通常提供为接口卡 (有时被称为“线路卡”)。一般而言, 接口 768 控制网络上数据分组的发送和接收, 并且有时支持网络设备 760 所使用的其他外设。可以提供的接口有 FC 接口、以太网接口、帧中继接口、线缆接口、DSL 接口、令牌环接口等等。另外, 可以提供各种甚高速接口, 例如快速以太网接口、千兆位以太网接口、ATM 接口、HSSI 接口、POS 接口、FDDI 接口、ASI 接口、DHEI 接口等等。

在本发明的某些实现方式中, 当在适当软件或固件的控制下工作时, CPU 762 可以负责执行与所需网络设备的功能相关联的特定功能。根据某些实施例, CPU 762 在软件的控制下完成所有这些功能, 所述软件包括操作系统 (例如 Windows NT) 和任意适当的应用软件。

CPU 762 可以包括一个或多个处理器 763, 例如来自 Motorola 微处理器族的处理器或来自 MIPS 微处理器族的处理器。在替换实施例中, 处理器 763 是专门设计的用于控制网络设备 760 的操作的硬件。在特定实施例中, 存储器 761 (例如非易失性 RAM 和/或 ROM) 也构成 CPU 762 的一部分。但是, 存在很多不同方式可以使存储器耦合到系统。存储器块 761 可以被用于各种目的, 例如缓存和/或存储数据、编程指令等等。

无论网络设备的配置怎样, 它都可以采用一个或多个存储器或存储器模块 (例如存储器块 765), 所述存储器或存储器模块被配置用于存储数据、用于通用网络操作的程序指令和/或与这里所述技术的功能相关的其

他信息。程序指令例如可以控制操作系统和/或一个或多个应用程序的操作。

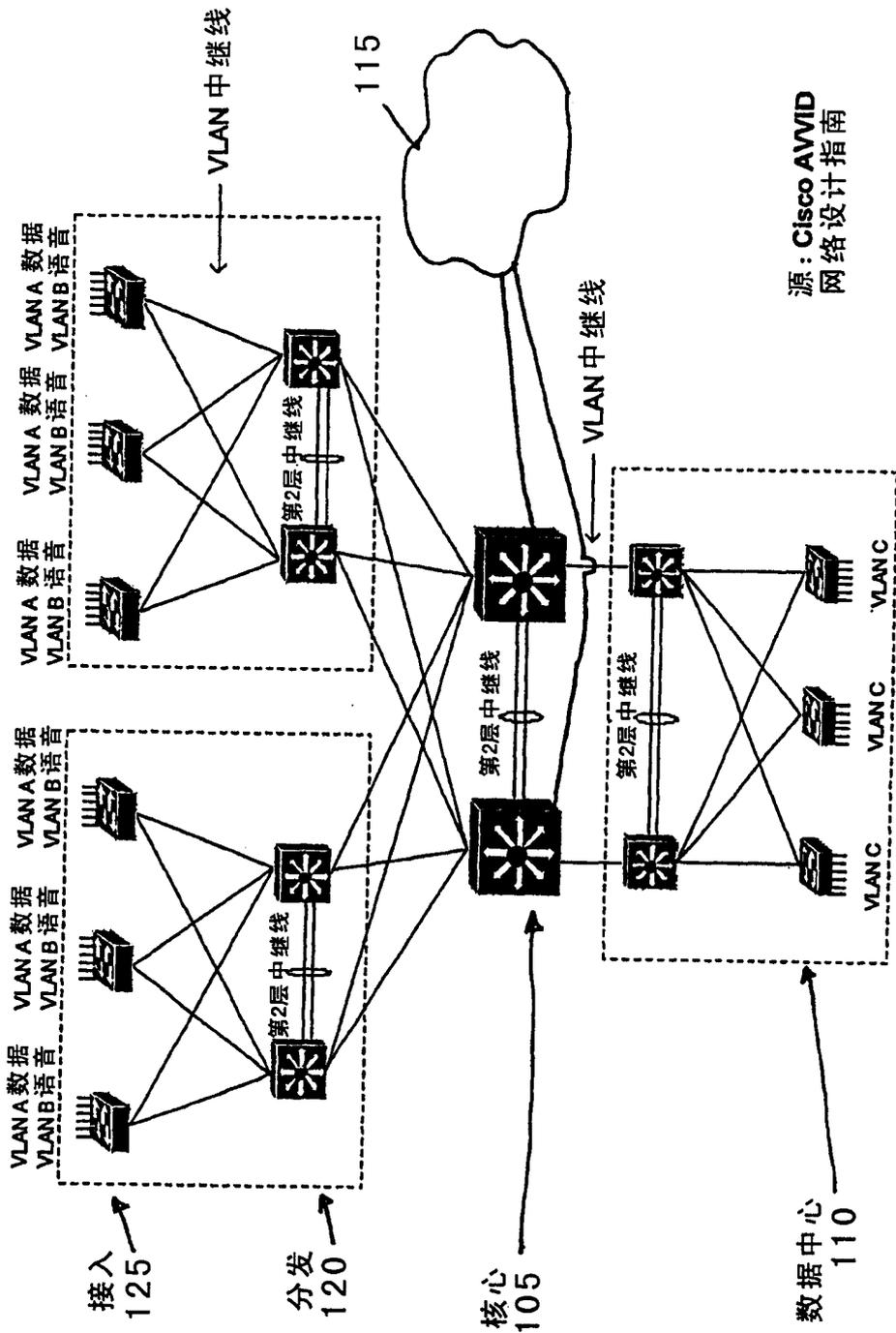
由于这样的信息和程序指令可被用于实现这里所述的系统/方法，因此本发明涉及包括了用于执行这里所述各种操作的程序指令、状态信息等
5 的机器可读介质。机器可读介质的示例包括但不限于：诸如硬盘、软盘和磁带之类的磁介质；诸如 CD-ROM 盘之类的光介质；磁光介质；以及专门配置用于存储和执行程序指令的硬件设备，例如只读存储器设备（ROM）和随机访问存储器（RAM）。本发明还可以被包含在通过适当的介质传播的载波中，例如无线电波、光线路、电线路等等。程序指令的
10 示例既包括机器代码（例如由编译器产生的代码），也包括包含了可以由计算机使用解释器执行的更高级代码的文件。

虽然图 7 所示的系统示出了本发明的一种特定网络设备，但是这并不意味着本发明只能在该网络设备体系结构上实现。例如，也经常使用具有用于处理通信和路由计算的单个处理器等等的体系结构。此外，该网络设备
15 也可以使用其他类型的接口和介质。接口/线路卡之间的通信路径可以是基于总线的（如图 7 所示），也可以是基于交换构架的（例如交叉开关）。

其他实施例

虽然这里示出并描述了本发明的示例性实施例和应用，但是在本发明的
20 的概念、范围和精神内可以存在很多变化和修改，并且本领域普通技术人员在阅读了本申请之后将明白这些变化。例如，虽然本发明的虚拟交换机主要在网络的分发层方面进行了描述，但是它们同样可应用于核心层和数据中心。

因此，本实施例将被看作示例性的，而非限制性的，并且本发明不局
25 限于这里给出的细节，而是可以在所附权利要求书的范围和等同物内进行修改。



源: Cisco AVID
网络设计指南

图1

逻辑视图

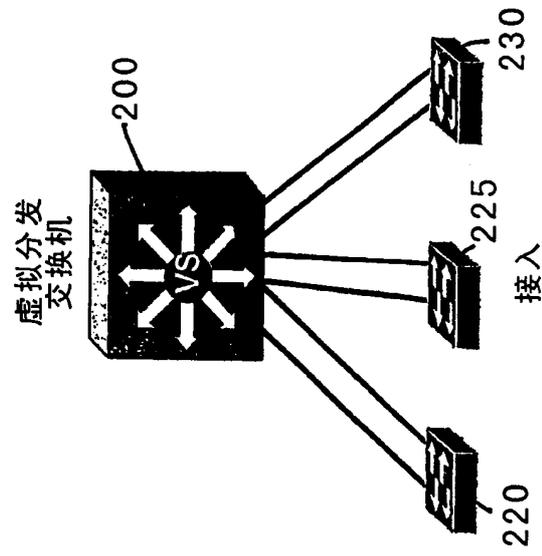


图2B

物理视图

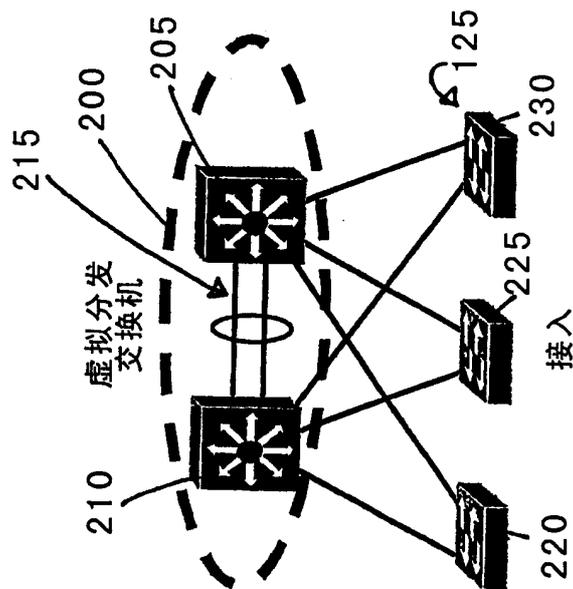


图2A

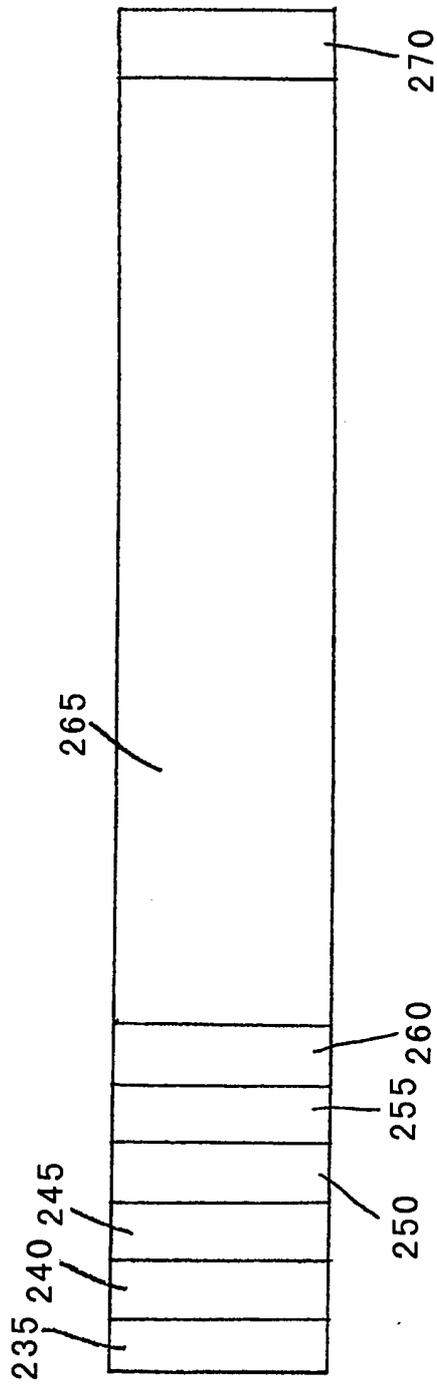


图2C

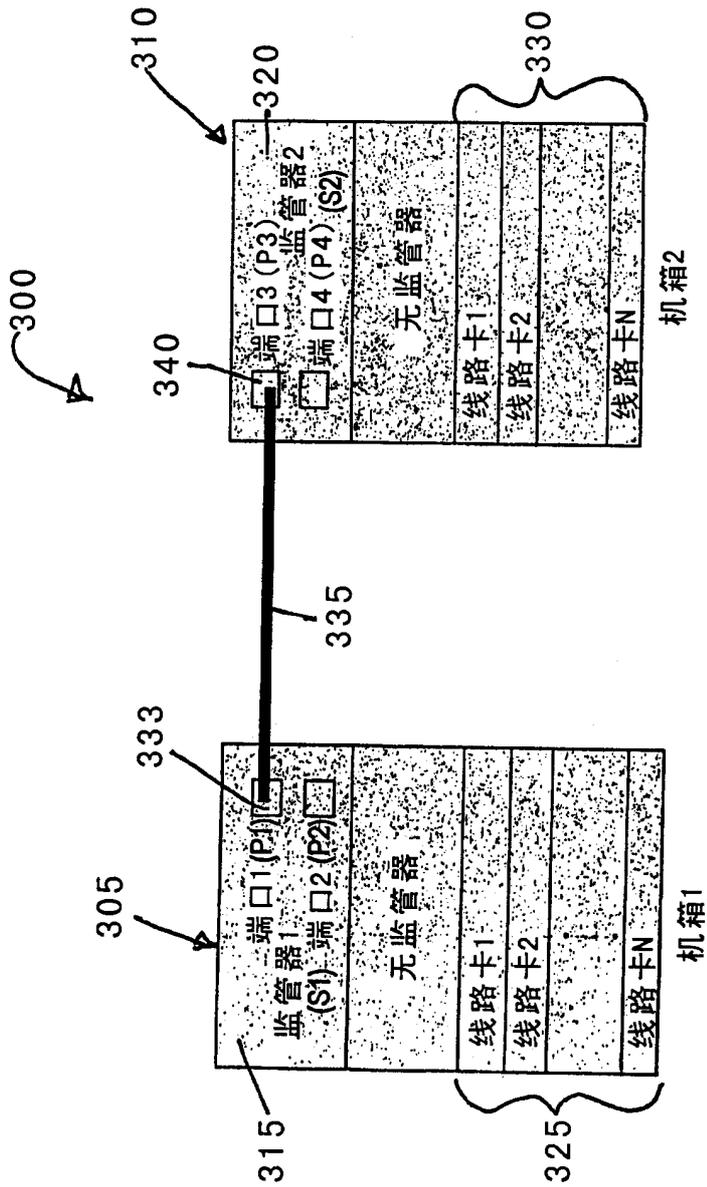


图3

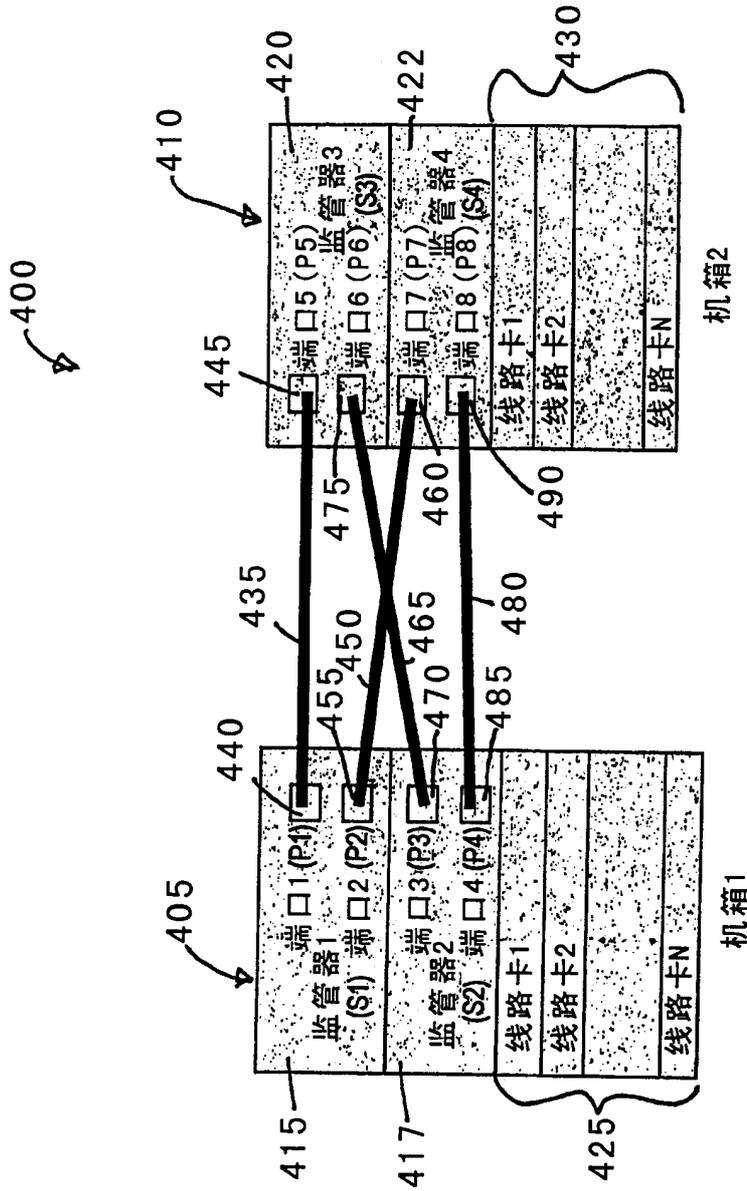


图4

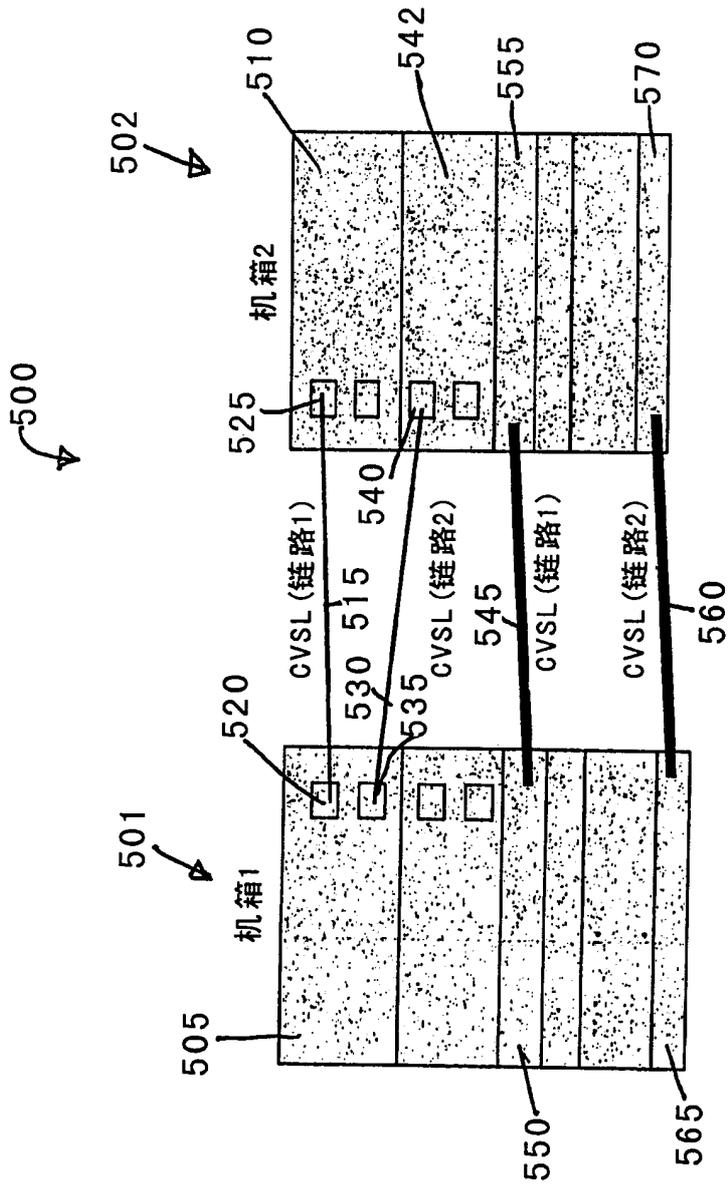


图5

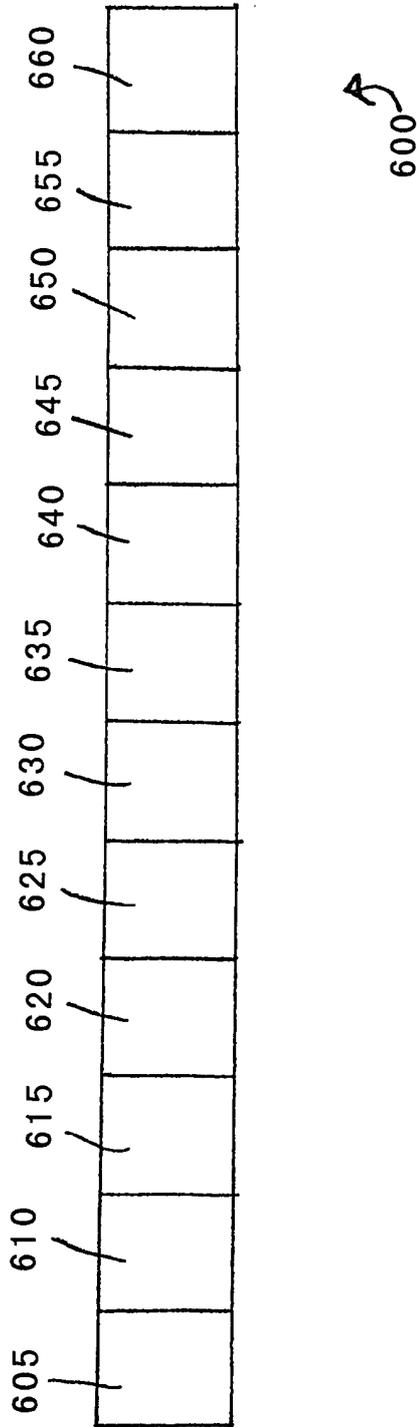


图6

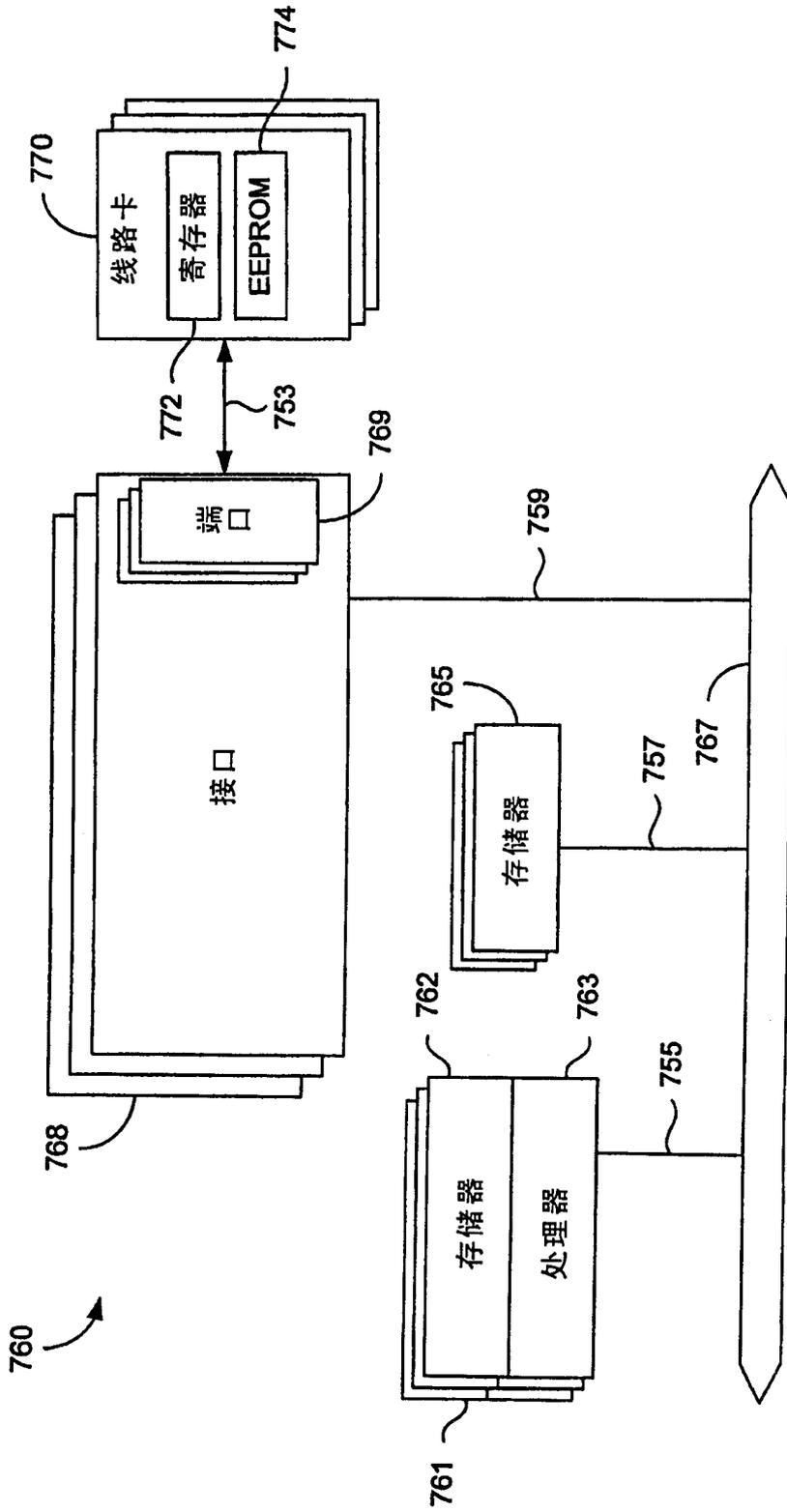


图7