[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 15/16 (2006.01)

[12] 发明专利说明书

专利号 ZL 02805445.8

[45] 授权公告日 2006 年 12 月 6 日

[11] 授权公告号 CN 1288576C

[22] 申请日 2002.2.25 [21] 申请号 02805445.8

[30] 优先权

[32] 2001. 2.24 [33] US [31] 60/271,124

[86] 国际申请 PCT/US2002/005570 2002.2.25

[87] 国际公布 WO2002/069096 英 2002.9.6

[85] 进入国家阶段日期 2003.8.22

[73] 专利权人 国际商业机器公司

地址 美国纽约州

[72] 发明人 陈 东 保罗 W·科特尤斯

艾伦 G·加拉 马克 E·贾姆帕帕

托德 E·塔肯

审查员 陈晓华

[74] 专利代理机构 中国专利代理(香港)有限公司

代理人 吴立明 王 勇

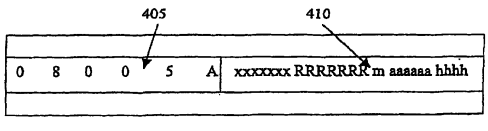
权利要求书 2 页 说明书 6 页 附图 5 页

[54] 发明名称

用于大规模并行系统的经由物理位置的以太网寻址

[57] 摘要

在一个大规模并行系统中，一种用于唯一地向一个设备分配一个 MAC 地址(400)的方法和装置用该设备(410)的物理位置编码 MAC 地址。 该方法和装置包含：用诸如机架号、中平面号、卡号和芯片号的物理拓扑信息配置该并行系统的设备互连。一个具有物理位置编码的 MAC 地址的设备或者节点然后可以通过位置查询用于测试、诊断、以及程序加载目的。



MAC地址 400

1.在一个包含以三个维度配置的多个节点的大规模并行计算系统中，每个节点包括一个计算设备，一种用于唯一地分配一个 MAC 地址到该计算设备的方法，包含:

编程该计算设备以把该 MAC 地址编码为该计算设备的一个物理位置;

对于上述编码步骤，使用 MAC 地址的预定数目的位，其中该计算设备的物理位置被唯一地描述。

2.如权利要求 1 所述的用于 MAC 地址分配的方法,其特征在于: 该 MAC 地址唯一地与一个以太网地址相关联。

3.如权利要求 1 所述的用于分配 MAC 地址的方法，其特征在于: 该计算设备的编程基于一个唯一机架、中平面、和包含该计算设备的卡的预定连线配置。

4.如权利要求 1 所述的用于分配 MAC 地址的方法，其特征在于: 该计算设备的编程基于来自一个主计算机的指令。

5.如权利要求 4 所述的用于分配 MAC 地址的方法,其特征在于: 该主计算机指令包含 IEEE 1149.1 JTAG 信号。

6.如权利要求 1 所述的用于分配 MAC 地址的方法，其特征在于: 该 MAC 地址的预定数目的位包含该 MAC 地址的最低有效部分。

7.如权利要求 6 所述的用于分配 MAC 地址的方法，其特征在于: 该 MAC 地址的最低有效部分包括一个物理位置描述符，其包含:

一个计算机架字段;

一个中平面字段;

一个卡字段; 以及

一个计算设备字段。

8.如权利要求 2 所述的用于 MAC 地址分配的方法，其特征在于: 该以太网地址唯一地与一个 TCP / IP 地址相关联。

9.如权利要求 1 所述的用于分配 MAC 地址的方法，进一步包含:

使用该 MAC 地址来管理该并行计算系统;

使用该 MAC 地址来诊断该并行计算系统; 以及

使用该 MAC 地址来调试该并行计算系统的功能。

10.在一个包含以三个维度配置的多个节点的大规模并行计算系

统中，每个节点包括一个计算设备，一种用于唯一地分配一个 MAC 地址到该计算设备的装置，包含:

a）一个系统互连配置装置，创建相对于在该大规模并行计算系统中的多个计算机架的一个计算机架位置的一个计算机架编码的位置，其中该计算机架编码的位置用来在该计算设备的 MAC 地址的计算机架字段中编程一个预定数目的位，以便唯一地描述该计算设备的计算机架位置;

b）一个计算机架互连配置装置，创建相对于连接到该计算机架的多个中平面的一个中平面位置的一个中平面编码的位置，其中该中平面编码的位置用来在该计算设备 MAC 地址的中平面字段中编程一个预定数目的位，以便唯一地描述该计算设备的中平面位置;

c）一个中平面互连配置装置，创建相对于连接到该中平面的多个卡的一个卡位置的一个卡编码的位置，其中该卡编码的位置用来在该计算设备 MAC 地址的卡字段中编程预定数目的位，以便唯一地描述该计算设备的卡位置;

d）一个卡互连配置装置，创建相对于连接到该卡的多个计算设备的一个计算设备位置的一个计算设备编码的位置，其中该计算设备编码的位置用来在该计算设备 MAC 地址的计算设备字段中编程一个预定数目的位，以便唯一地描述该计算设备在该卡上的位置。

11.如权利要求 10 所述的用于分配 MAC 地址的装置，其特征在于:该 MAC 地址唯一地与一个以太网地址相关联。

12.如权利要求 10 所述的用于分配 MAC 地址的装置，其特征在于:该 MAC 地址的最低有效部分包含计算机架字段、中平面字段、卡字段、和计算设备字段。

13.如权利要求 11 所述的用于分配 MAC 地址的装置，其特征在于:该以太网地址唯一地与一个 TCP / IP 地址相关联。

14.如权利要求 10 所述的用于分配 MAC 地址的装置，包含:

用于使用该 MAC 地址来管理该并行计算系统的装置;

用于使用该 MAC 地址来诊断该并行计算系统的装置; 以及

用于使用该 MAC 地址来调试该并行计算系统的功能的装置。

用于大规模并行系统的经由物理位置的以太网寻址

交叉引用

本发明要求享受于 2001 年 2 月 24 号提出的、标题为 MASSIVELY PARALLEL SUPERCOMPUTER 的共同拥有的、待决美国临时专利申请 60 / 271,124，其全部内容和公开就好像在此被充分阐述的那样通过引用被明确地包含在此。这个专利申请另外涉及以下在同一日期提出的、共同拥有的待决美国专利申请，其中这些申请中每一个的全部内容和公开就好像在此充分阐述的那样通过引用被明确地包含在此。美国专利申请（YOR920020027US1、YOR920020044US1（15270）），“Class Networking Routing”；美国专利申请（YOR920020028US1（15271）），“A Global Tree Network for Computing Structures”；美国专利申请（YOR920020029US1（15272）），“Global Interrupt and Barrier Networks”；美国专利申请（YOR920020030US1 （15273）），“Optimized Scalable Network Switch”；美国专利申请（YOR920020031US1、YOR920020032US1（15258）），“Arithmetic Functions in Torus and Tree Networks”；美国专利申请（YOR920020033US1、YOR920020034US1（15259））、“Data Capture Technique for High Speed Signaling”；美国专利申请（YOR920020035US1（15260））、“Managing Coherence Via Put / Get Windows”；美国专利申请（YOR920020036US1、YOR920020037US1（15261））、“Low Latency Memory Access And Synchronization”；美国专利申请（YOR920020038US1（15276））、“Twin - Tailed Fail - Over for Fileservers Maintaining Full Performance in the Presence of Failure”；美国专利申请（YOR920020039US1（15277））、“Fault Isolation Through No - Overhead Link Level Checksums”；美国专利申请（YOR920020040US1（15278））、“Ethernet Addressing Via Physical Location for Massively Parallel Systems”；美国专利申请（YOR920020041US1（15274））、“Fault Tolerance in a Supercomputer Through Dynamic Repartitioning”；美国专利申请（YOR920020042US1（15279））、“Checkpointing Filesystem”；美国专利申请（YOR920020043US1（15262））、“Efficient Implementation

of Multidimensional Fast Fourier Transform on a Distributed - Memory Parallel Multi Node Computer"; 美国专利申请(YOR920010211US2 (15275))、"A Novel Massively Parallel Supercomputer"; 以及美国专利申请(YOR920020045US1 (15263))、"Smart Fan Modules and System"。

技术领域

申请人要求依据 35 U. S. C. 119 (e)享受于 2001 年 2 月 24 日提出的美国临时申请 60 /271,124 的优先权，该临时申请的公开通过引用包含在此。

本发明概括地说涉及一种向电子设备分配地址的方法。它尤其涉及一种向一个计算设备节点分配一个编码的唯一硬件地址的方法，其中编码表示该计算设备节点的物理地址。

背景技术

一个用于计算机数据网络的众所周知标准，开放系统互连（OSI）标准，为了兼容的数据通信系统设计规定了几层互连。一个这样的层是数据链路层。这个层表示这样的传输介质，通过它网络设备在它下面的层、硬件进行连接的物理层，和紧挨着在它上面的层、网络层之间进行通信。

OSI 规定了几个在数据链路层处的候选介质，一个这样的介质是以太网。任何一个在数据链路层处使用的介质都必须包含一个用于在该网络上的每个设备的唯一硬件地址。这个唯一的硬件地址、亦称为媒体存取控制（MAC）地址与一个用于使用的介质的唯一地址、例如一个以太网地址相同。因此，一个设备的 MAC 地址和它的以太网地址是相同的唯一数字。作为当前通常的实现，对于以太网，MAC 地址是一个通常被表示为 12 个十六进制数字的 48 位数字。在众所周知的当前地址映射方案下，最重要的 6 个十六进制数字编码硬件设备生产商，例如，08005A 用于 IBM。最不重要的 6 个十六进制数字编码一个用于由该硬件设备生产商制造的设备的序列号。

在一个相关的美国临时申请 60 / 271,124、"A Novel Massively Parallel Supercomputer" 的公开中，其中描述了一个具有两个电子处理器在一个多计算机的每个节点内的半导体设备。在该多计算机

内，有多个高速度内部网络，以及一个使用以太网的外部网络。

在如上所述的大规模并行计算机系统中，预计会使用 162,000 个不同的以太网地址。这个大数量的以太网地址对一个主机，以及中间网络路由器和转换器产生一个重要的问题，为了包括测试、诊断、初始程序装入、等等的各种目的，所有这些设备都必须跟踪记载 MAC 地址。例如，如果一个特定设备的 MAC 地址在一次测试期间没有响应，则为了进一步的测试和诊断必须确定该设备的物理位置。当如在一个大规模并行计算机系统中那样，有许多节点布置在许多不同的位置时，这个查找设备的问题被放大了。例如，要被分配 MAC 地址的巨型计算机节点是物理上驻留在卡上的计算机芯片。该卡被安装在称作中平面的底板上。中平面本身又安装在机架中。因此，当已知有关一个失败设备的唯一东西是它的 MAC 地址时，必须某种程度上隔离机架、中平面、底板、卡和芯片。而目前没有已知的、把一个物理位置和一个设备的 MAC 地址相关的现有技术，所以通过创建这样一个关联来解决这个问题是合符需要的。

发明内容

因此，本发明的一个目的是提供一种为向一个设备唯一地分配一个物理位置编码的 MAC 地址的方法和设备。

本发明的一个进一步目的是提供一种用于唯一地向设备分配一个物理位置编码的 MAC 地址的方法和设备，其中该 MAC 地址通过一个到该设备的外部接口编码。

当前发明的还有另一个目的是提供一种用于唯一地向该设备分配一个物理位置编码的 MAC 地址的方法和设备，其中一个数据链接介质是以太网，以及一个相应的以太网地址与编码的 MAC 地址相同。

当前发明的一个进一步目的是提供一种用于唯一地向设备分配一个物理位置编码的 MAC 地址的方法和设备，其中该数据链接介质是当前存在或者可以为在数据链路层处的通信开发的任何介质，以及相应的数据链接介质地址与编码的 MAC 地址相同。

当前发明的一个更进一步目的是提供一种用于为了测试、诊断、程序载入和监控在一个大规模并行系统中的设备而确定多个互连设备中的任何一个的物理位置的方法和设备。

可以在本发明中，通过提供一种把一个物理位置编码成为一个 MAC 地址并且把该物理位置编码的 MAC 地址唯一地分配到一个设备的方法和设备，来获得这些及其它目的和优点。

具体地说，提供了一种用于唯一地向一个设备分配一个 MAC 地址的方法，其包含：配置设备互连以把 MAC 地址编码为该设备的一个物理位置；把该编码的 MAC 地址作为一个唯一的以太网地址使用；使用线路来在该 MAC 地址中编码一个预定数目的唯一位；把唯一位中的预定数目分配给一个表示硬件设备坐标到该设备物理位置的值，诸如机架号、中平面号、卡号、以及芯片号。

附图说明

现在将通过参考伴随着本申请的附图更详细地描述本发明。要注意到：在附图中类似的参考数字用来描述它的类似以及对应单元。

图 1 显示了本发明中的硬件环境的物理布局；

图 2 显示了通过一个以太网转换器互连的计算节点；

图 3 显示了现有技术的 MAC 地址字节结构；

图 4 显示了本发明中的 MAC 地址字节结构；以及

图 5 显示了在本发明的一个安装表面上编码的物理地址的一个示例。

具体实施方式

这个发明的一个方面应用于一个基于以太网的外部网络。这个发明的一个最佳实施例以以太网“MAC”硬件地址的形式编码一个节点的一个物理位置，其通过一个包含该节点的特定机架、包含该节点的特定中平面、以及包含该节点的特定节点-卡的组合进行分配。

在这个发明的一个最佳实施例中，由该巨型计算机发送到主机的每个以太网包唯一地标识产生该包的节点的物理位置并且允许那个信息被使用来跟踪到在该机器中的具体节点的问题。这个发明的另一个方面还能够把一个地理位置唯一地标识为该物理位置中的一部分。

在这个发明的一个方面中，如图 1 中的示例所示，物理上有 80 个系统计算机架 105、110。如上讨论，多个中平面占据每个机架，例如每个机架 2 个中平面。另外有多个卡，例如 64 个卡占据每个中平面。

每个卡具有多个网络可寻址的芯片，例如，9个芯片。并且，在这个发明的一个最佳方面中，在该卡上的每个网络可寻址的芯片表示多个计算节点205中的一个。

依据以上示例，表示任何节点物理位置的需要位的预定数目是18位。位的数目通过如下所述把位置相乘：9芯片×64卡×2中平面×80机架= 92,160个在一个系统内的唯一位置来导出。其数字然后转换为十六进制是16800h，表示18位信息。

图2显示了其中计算节点205使用用于以太网数据链接215的转换器210进行通信的网络环境。在这些条件下，48位以太网MAC地址非常适合用于承载物理位置信息。如图3所示，48位MAC地址被分成一个最有效部分（MSP）305和一个最不有效部分（LSP）310。

现有技术的方法把MSP分配给一个诸如IBM的生产商，如图所示，MSP 305是用于IBM的08005A。在现有技术方法下，LSP 310被分配用于序列号。

在本发明的方法下，MSP 405仍然保留用于生产商标识符，例如，IBM。然而，现在LSP被分配作为一个物理位置描述符410。该物理位置描述符可以通过如上所述的机架、中平面、卡和芯片定义一个诸如计算节点205位置的设备位置。显示的示例物理位置描述符410具有一个7 R位字段来标识一个机架号、一个1 m位字段来标识一个中平面、一个6 a位字段来标识一个卡号、以及一个4 h位字段来标识一个计算设备号。因此，如图所示，一个节点的物理位置被完全描述了。此外，在图4 LSP中显示的x位是额外的位，其可用于描述例如在一个更大物理拓扑结构中的节点物理位置的设备。

本发明的一个最佳方面使用一个硬布线的编程技术来编码物理位置，诸如在图5中的示例所示。应当注意到虽然在此讨论和显示了连线，但是任何配置设备互连的装置，诸如光电子装置，例如可以在本发明的范围内使用。一个安装面510，例如一个中平面，具有一个槽连接器515，其具有到一个正电压、Vcc 511或者地512的连接513。用这样的方式，电压电平可以用来编码对应于该接口的物理拓扑结构的一个预定数目的位。以一种类似的方式，该卡能够被连线以为每个芯片，即在卡上的节点，编码一个计算设备号。此外，把机架连接在一起的系统级连线能够被配置来编码一个通过中平面传播，并且到达卡

上的一个机架号码。类似地，机架级别的连线被配置编码一个中平面号，而中平面连线被配置编码一个卡号。最后，卡级别的连线能够被配置为标识，即编码一个计算设备号。当电能被施加到该系统上时，一个电可擦可编程只读存储器（EEPROM）（没有显示）能够用来存储用于为连接的设备，例如节点，配置 MAC 地址的编码位。

一种用于输入物理位置编码位到该设备或节点中的替换技术将是通过使用每个节点的 IEEE 1149.1 JTAG 接口来为那个节点编程物理位置编码的 MAC 地址。在本技术领域已知的是：和一个 JTAG 兼容的设备，诸如任何计算节点 205 的通信，是通过使用一个主机，诸如例如，一个具有一个到包含该计算节点 205 的 JTAG 兼容卡的连接的硬件控制器，来实现。JTAG 兼容的设备，例如计算节点，必须连接到所有的闪速存储器地址、数据和控制信号。对于这个要起作用的编程方法，闪速存储器不需要是 JTAG 兼容的。该主机发送命令和数据到 JTAG 兼容的设备，例如任何计算节点 205，然后把该数据传送到闪速存储器用于编程。用这样的方式，主机提供一条连接任何计算节点 205 的通信链路用于完成 MAC 地址的物理位置编码。这个发明的一个最佳环境的 JTAG 性能在临时申请 60 / 271,124 中进行了讨论，该申请已经通过引用包含在此。

在系统操作期间，一个由一个如上所述的连接设备传输的 MAC 地址可以由转换器、网络监控器、和主机查询以确定确切的设备物理位置。这个性能提供了该并行计算系统改进的管理、诊断和调试功能。另外，当分配了 TCP/IP 地址，诸如在一个运行动态主机配置协议（DHCP）的系统中时，TCP/IP 地址变为该设备位置的一个同样有效的指示符。

现在已经通过一个最佳实施例对本发明进行了描述，对于本领域的那些技术人员来说，可以发生各种修改和改进。 因此，要理解：该最佳实施例作为一个示例提供而不是作为一个限制。本发明的范围由附加权利要求定义。

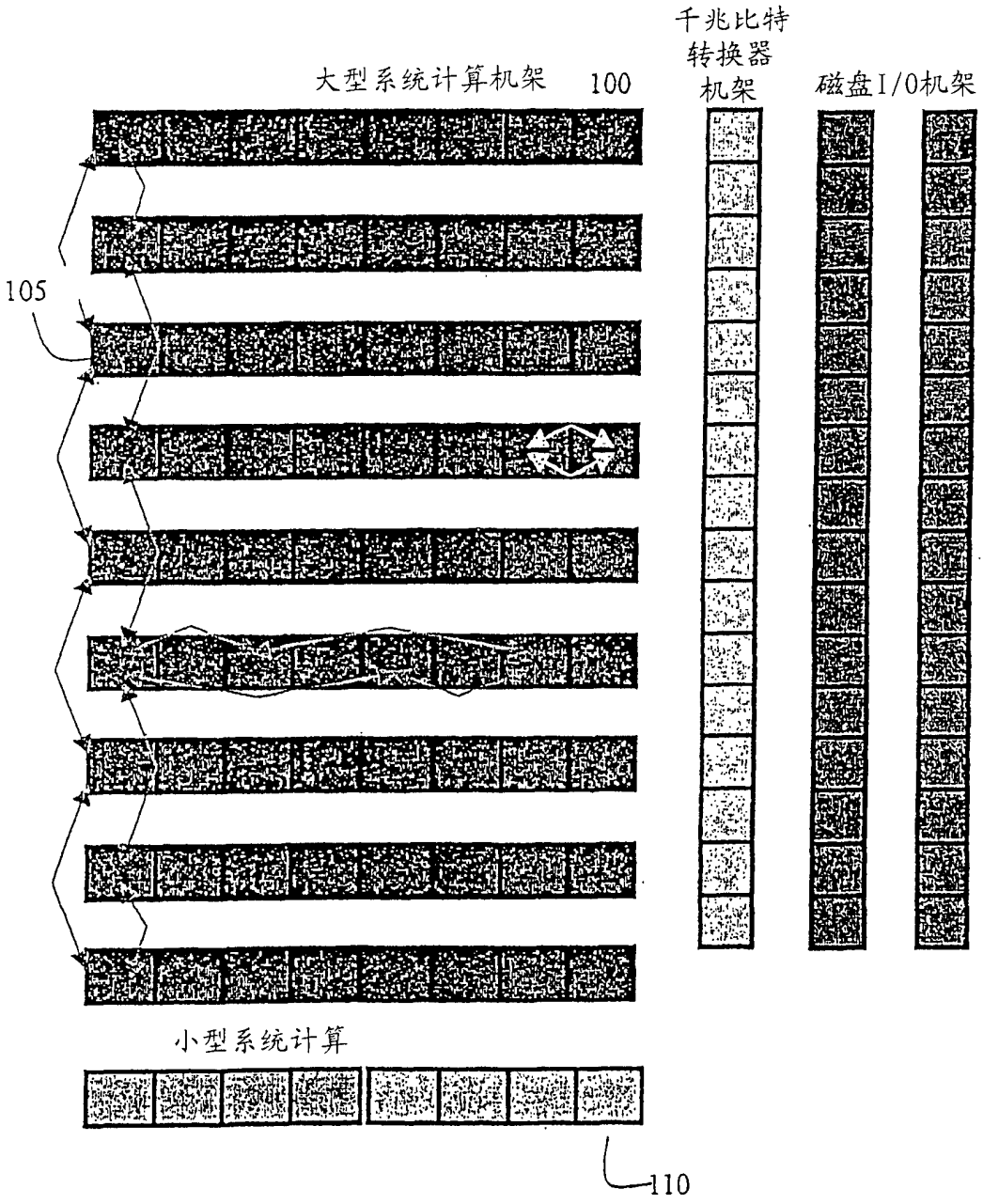


图 1

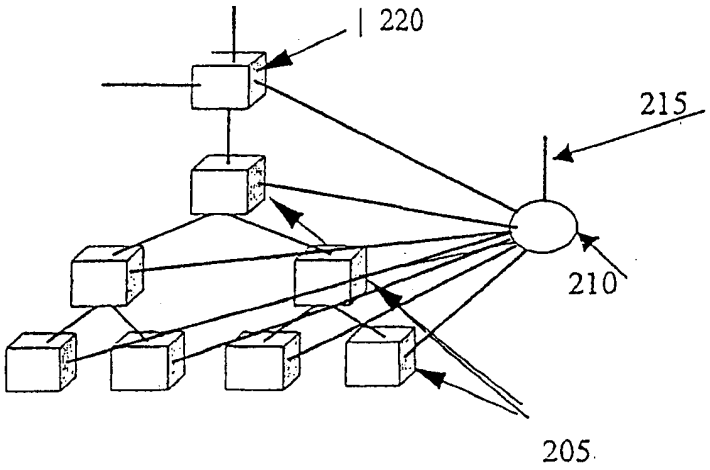


图 2

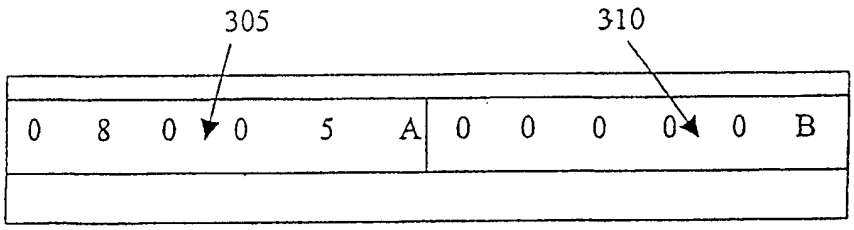
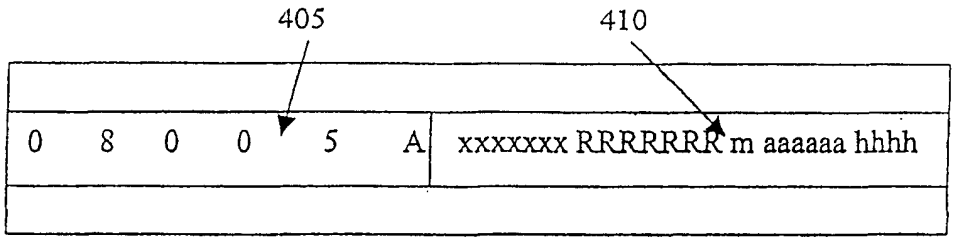


图 3（现有技术）



MAC地址 400

图 4（当前发明）

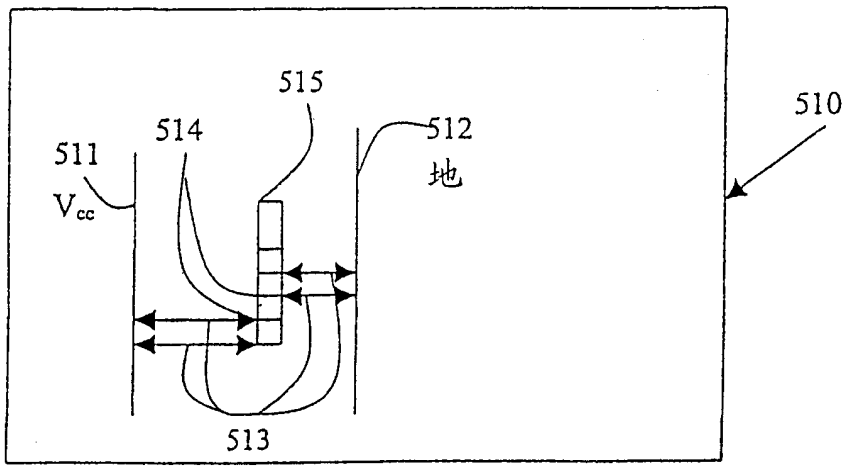


图 5