[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 13/00

[12] 发明专利说明书

[21] ZL 专利号 99812747.7

[45] 授权公告日 2005 年 1 月 5 日

[11] 授权公告号 CN 1183455C

[22] 申请日 1999.10.8 [21] 申请号 99812747.7
[30] 优先权
[32] 1998.10.27 [33] US [31] 09/181,015
[86] 国际申请 PCT/US1999/023424 1999.10.8
[87] 国际公布 WO2000/025219 英 2000.5.4
[85] 进入国家阶段日期 2001.4.27
[71] 专利权人 松下技术公司
地址 美国新泽西
[72] 发明人 艾布拉西姆·M·卡梅尔
韦利德·G·阿莱夫
萨里特·穆克基
审查员 齐 霁

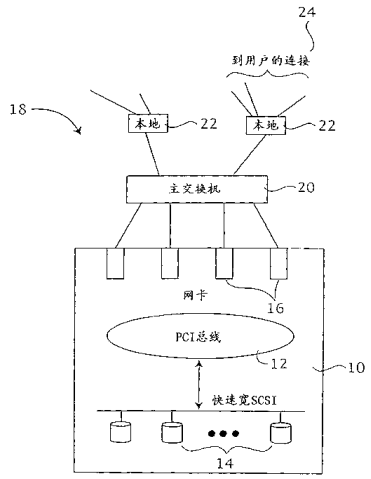
[74] 专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 付建军

权利要求书 1 页 说明书 9 页 附图 5 页

[54] 发明名称 无缝、可扩缩的分布式媒体服务器

[57] 摘要

一种分布式多媒体服务器(18)，它采用通过高带宽的网络(38)与客户系统(40)相连的模块服务器单元(26，28)。 该服务器单元(26，28)以交错方式与交换单元(30，32)相连，而且代表不同媒体选择的数据以条带组方式排列，分布在各服务器单元(26，28)间。 服务器(18)可以小到只是单个服务器单元(10)，或根据需要使其按递增方式增长，通过将额外服务器单元(26，28)连接到该系统中。热媒体选择的多个副本跨分布式架构存储在不同的条带组中，以使存储子系统中的瓶颈效应降到最小。



ISSN1008-4274

知识产权出版社出版

1. 一个可扩缩的分布式媒体服务器，包括：

多个服务器单元，每个服务器单元有多个网络接口电路；

多个交换单元，以交错方式与上述网络接口电路连接；以及

多个被组织成条带组的媒体存储设备，每个媒体存储设备能用来存储一个或多个媒体对象，并与多个服务器单元中的一个相关联；

至少一个被分配给一个给定的条带组的媒体选择，所述媒体选择被存储为在至少两个服务器单元之间分布的媒体存储设备中的至少两个媒体对象；

一个存在于所述媒体存储设备的存储器，该存储器具有每个条带组及其相关联的媒体存储设备的组织数据，并具有媒体选择及其所分配条带组的分配数据。

2. 权利要求 1 的媒体服务器，其中上述多个交换单元包括至少一个主交换机和多个与所述主交换机相连的本地交换机.

3. 权利要求 1 的媒体服务器，其中上述给定的媒体选择被分为多个媒体对象并且所述媒体对象被分配给上述多个服务器单元中的一个共同的条带组.

4. 权利要求 1 的媒体服务器，其中上述多个媒体选择作为媒体对象存储在上述服务器单元的上述媒体存储设备中而且其中当对一个媒体选择的一个单一副本的需求超过了该单一副本被分配给的条带组的带宽时，该媒体选择可以存储为多个副本.

5. 权利要求 4 的媒体服务器，其中所述多个副本包括一个第一副本和一个第二副本，每个副本分配给不同的条带组。

6. 权利要求 1 的媒体服务器，其中媒体选择分为以循环方式分布于上述服务器单元间的块.

7. 权利要求 1 的媒体服务器，其中上述给定的媒体选择被分为多个媒体对象并且所述媒体对象分布于上述服务器单元间，以使每个服务器单元都能够存储构成所述给定的媒体选择的媒体对象的一个基本相等的部分.

无缝、可扩缩的分布式媒体服务器

技术领域

本发明一般地涉及分布的多媒体服务器。更具体地说，本发明涉及可扩缩的媒体服务器架构，该架构可以进行扩展以随着需求的增长支持更多的用户。在支持大负载时该结构是均衡的从而避免瓶颈。

背景技术

“地球村”一词形象地说明，当今的技术已使从地球上的一点向另一点即时发送信息成为可能。使其成为可能的信息基础结构正在构建中。期望有一天，每个说明的多媒体材料都可以按需在地球上的任何地方传送。

上述设想正在按指数发展。如果可以将以前互联网使用的增长看作某种指标的话，那么，随着数字多媒体传送替代当前的模拟传送以及随着地球村成为现实，我们可以期望多媒体服务器基础结构按指数增长。

如果我们要满足这个按指数增长的需求，那么可扩缩性是非常重要的。如果在添加了额外硬件的情况下，系统能够支持更多的数据流(stream)，则该系统是可扩缩的。这里，数据流可以包括视频内容、音频内容、代表其他类型的信息的数据以及这些内容的组合。因而，可扩缩的服务器是一种服务器，它可以用额外的硬件升级，以便在用户增多时支持更多的数据流。如果硬件容量增加一倍时使服务器能够支持两倍数量的数据流，便获得了线性扩大。

开发一个可扩缩的分布式媒体服务器会遇到几个必须解决的问题。带宽瓶颈是常见的问题来源。系统成本是个竞争因素。使用昂贵的组件来增加组件带宽在经济上并不实用。

如果考虑到用户，则会在实现可扩缩设计时带来进一步的问题。并非所有多媒体内容生来都是平等的。例如，某些电影非常受欢迎并且整天经常会有人请求，会在特定的最佳时间出现需求高峰。其他电影的需求相对较少，但也需要使其可用，以供偶然需要它们的用户来

选择。随着时间变化，媒体选择的受欢迎程度会发生变化，于是可扩缩架构在这方面也必须具有相应的灵活性。同时还会不断添加新的媒体选择，因而问题就更加复杂。

发明内容

本发明提供了一个可扩缩的分布式媒体服务器架构，它通过模块方式解决上述问题。各个服务器单元(element)，每个都有多个网络接口电路和多个媒体存储设备，构成了该架构的信息存储的组成部分(component)。多个交换单元以交错方式与服务器单元相连，从而定义了一个分布式的网络。

本发明所提供的一个可扩缩的分布式媒体服务器，包括：多个服务器单元，每个服务器单元有多个网络接口电路；多个交换单元，以交错方式与上述网络接口电路连接；以及多个被组织成条带组的媒体存储设备，每个媒体存储设备能用来存储一个或多个媒体对象，并与多个服务器单元中的一个相关联；至少一个被分配给一个给定的条带组的媒体选择，所述媒体选择被存储为在至少两个服务器单元之间分布的媒体存储设备中的至少两个媒体对象；一个与所述媒体存储设备相关的存储器，该存储器具有每个条带组及其相关联的媒体存储设备的组织数据，并具有媒体选择及其所分配条带组的分配数据。

单个服务器单元可用于向少量用户提供媒体。但是，只要通过添加更多的服务器单元和相关联的交换单元，便可随时对该架构进行扩展，从而容纳更多的用户。服务器单元与交换单元间的交错连接支持均衡的分布式服务器系统。

为进一步均衡系统和避免瓶颈，与媒体存储设备相关联的数据结构将这些设备组织为条带组。会对条带组进行排列，以使一个给定的媒体选择分配给一个条带组，并因而以分布方式存储在服务器单元集合间。需求量大的选择则存储为多个副本，每个副本均分配给不同的条带组。结果便可以得到一个高度均衡、可扩缩的媒体服务器系统，该系统可以充分利用其各组成部分的可用带宽。

为了更深入地了解本发明以及其目标和优点，请参阅下列说明和

附图。

附图说明

图 1 是一个客户-服务器网络图，表示使服务器单元与多个客户系统相连的当前优选方式；

图 2 表示替换的客户-服务器网络配置；

图 3 表示依照本发明服务器单元及交换单元如何进行交错连接；

图 4 是数据结构图表，表示依照本发明如何将媒体存储设备组织为条带组以及如何在不同的服务器单元间分布数据；以及

图 5 是软件架构图表，表示使用本发明的系统用于数据传送的消息交换。

具体实施方式

图 1 表示用于实施本发明的可扩缩服务器的当前优选架构。可扩

缩的媒体服务器是分布式媒体服务器，由称为服务器单元的模块单元构成。图 1 显示了单个服务器单元 10。单个服务器单元可用于向少量用户提供媒体内容。实际上，会将多个服务器单元结合在一起，如下所述，来构建可扩缩的分布式媒体服务器系统。

服务器单元 10 可以使用现货供应的计算机部件进行构建。所示的实施例采用了 PCI 总线 12，有多个媒体存储设备 14 通过适当的接口，如既快又宽的 SCSI 接口，挂接到该总线。媒体存储设备 14 可以是磁盘驱动器或类似设备。还有多个网络接口电路 16 挂接到 PCI 总线 12 上。这些可以是现货供应的网卡，如以太网卡。为了说明目的，图 1 中显示了四个网络接口电路。根据应用以及各组成部分使用的带宽，网卡的个数也可以不同。

服务器单元 10 连到在 18 处显示的交换电路阵列。在当前的优选实施例中，网卡 16 与主交换机或集线器 20 相连，主交换机或集线器又与多个本地交换机或本地集线器 22 相连。本地交换机或本地集线器再与多个单个客户系统 24 相连。为本公开的目的，交换机和集线器这两个词可以互换。主交换机和本地交换机可以是来自各种货源的现货供应以太网交换电路。

尽管交换电路阵列采用主交换机 20 是当前优选的，图 2 表示了一个替换配置。在替换配置中，未使用主交换机，各网卡 16 均直接与专用的本地交换机 22 相连合。本地交换机又支持与客户系统 24 的多组连接。在替换实施例中，数据流在 PCI 总线 12 中交换给客户系统。该替换方案相对便宜，因为它不使用主交换机。但是，由于各连接的带宽受一个网卡的最大带宽限制，因而在某些应用中该替换实施例会带来一些限制。相比之下，图 1 中的优选实施例中则可以拥有整套网卡最大的带宽。

为了进行说明，假设两个客户系统与一个服务器单元中的相同网卡相连，且每个系统要求 100Mbs（每秒兆字节）。还假设每个网络接口电路的最大带宽是 150Mbs。图 1 的实施例可以满足两个请求因为每个本地交换机/集线器均可从多个网卡获得数据。相反，图 2 的实施

例，即使服务器单元有足够的带宽来服务两个客户系统，也不能满足两个请求。在图 2 所示的实施例中，由于两个客户与同一个网卡相连，最大带宽 (150Mbs) 不足以同时支持两个 100Mbs 的数据流。

尽管一个服务器单元可以用作独立服务器，但多个服务器单元也可随时连接在一起形成一个功能更强大的服务器，以满足增长的需求。本发明的可扩缩架构可以线性扩展。也就是说，两个服务器单元提供的数据流是一个服务器单元提供的数据流的两倍。

图 3 显示多个服务器单元相互连接形成更大的服务器的方式。服务器单元 26 和 28 以交错的方式与主交换机对 30 和 32 相连。特别地，网络接口电路排列为互相交错的组（例如，偶数组 34，奇数组 36）。偶数组与主交换机 30 相连，奇数组则与主交换机 32 相连。主交换机 30 和 32 又与由本地交换机与客户系统共同构成的外部网络 38 相连。多个单元服务器的用户 40 只与外部网络 38 的客户系统之一相连。以下文所示的方式，从两个服务器单元向用户提供媒体流，而用户并不知晓多个服务器单元参与了。

当前的优选实施例在用户和服务器单元间提供了对称的连接。目的是避免在服务器单元及相关联的网络中可能的瓶颈。在最实际的应用中，不同的媒体选择，如不同的电影，可以存储在不同的服务器单元上，且并非每个电影的请求频率都相同。

为了理解如何实现负载平衡，请参看如图 3 所示都两个服务器单元系统。在这方面，应认识到一个系统可以使用更多的服务器单元进行配置；因此，本说明应视为是对权力要求中所阐明的发明范围的限制。在本说明中，每个服务器单元均有四个网络接口电路（但也可以使用不同的数量）。主交换机的数量（这里是两个）等于服务器单元的数量。因此，所示的交错连接的优点在于，无论服务器单元 26 和 28 经历的负载的不同，交换机 30 和 32 都具有相同的通信量(traffic)。在这个配置中，当网络接口电路的数量等于 4 且服务器单元的数量也等于 4，则可以实现最佳负载均衡。

在更为一般的情形中，当系统使用 S 个服务器单元，每个组件包

含 N_k 个网络接口电路，且当有 W 个主交换机（W 不一定等于 S），则有：

（1） K =1

（2） 对于 SE_k 中的 NICj

（3） 以循环(round-robin)方式使每个 NIC 与一个不同交换机中的端口相连，直到 SE_k 中的所有 NIC 都被连接

（4） 使 K 递增，转至（2）

每个交换机 SW_i 有 P_i 个端口可与服务器单元相连。

当然，并非 N 与 S 的所有组合都能实现最佳负载平衡。例如，如果服务器包括三个服务器单元和四个网络接口电路（S=3;N=4），则只通过交错方式便无法保证最佳均衡。然而，如果能在不同的服务器单元间谨慎地分配数据，则可以显著改善各交换机与服务器单元间的负载平衡情况。

当前的优选实施例在各服务器单元间分配数据。系统采用与媒体存储设备相关联的数据结构，以此将媒体存储设备组织为条带组。对条带组进行排列，以便能将给定的媒体选择分配给一个条带组并且以分布方式存储在集合服务器单元间。

每个媒体选择均由连续的数据块（例如，视频帧）构成。这些块会分配给条带组并以特定的方式（例如，循环、随机等）分布于所有服务器间如图 4 所示。在典型实施例中，每个块可以与一个表示不足一秒钟的节目材料的视频帧相对应。在典型应用中，视频帧可以按每秒三十帧的速率表现，其中每个块代表 1/30 秒的节目材料。于是，当向用户提供媒体选择时，在用户看来，就如同所有服务器单元都在同时提供媒体选择一样。例如，媒体服务器可以按循环方式提供媒体选择（例如，每个服务器均按快速顺序提供不同的帧）。当然，在不同的服务器应用程序中可以使用不同的数据类型。因此，根据应用程序，块大小也有所不同。

如上所述，不同的媒体对象有不同的访问频率。*称为热对象(hot*

*object)*的对象有高的访问频率，反之，称为*冷对象(cold object)*的对象则很少被访问。与不存储*热对象*的服务器单元相比，存储*热对象*的服务器单元会遇到更多的通信量。这会在总体系统性能方面造成瓶颈。条带组方案便可以避免出现这种瓶颈。

如图 4 所示，当前的优选数据结构定义了多个条带组，并将一个或多个媒体存储设备分配给每个条带组。条带组分布在服务器单元集合间（图 4 显示了三个服务器单元）。构成一个条带组的磁盘按循环方式分布于多个不同的服务器单元间。

图 4 显示了三个条带组，组 A、组 B、组 C。与两个媒体选择相对应的媒体对象表示为 X_1、X_2、X_3...;Y_1、Y_2、Y_3....。媒体对象 X_1 按循环方式存储于服务器单元 SE_1 上的条带组 A 中。媒体对象 X_2 存储于服务器单元 SE_2 上的条带组 A 中，依此类推。于是，在进行下一轮分配时，媒体对象 X_4 将分配给服务器单元 SE1 上的条带组 A。包含媒体选择 Y 的媒体对象以类似方式分配给条带组 B。在上述两种情形中，注意媒体对象如何在保留同样的条带组分配的同时在所有服务器间进行平均分配。

通过在多个服务器单元间存储每个媒体选择的均等部分，不同对象的不对称的访问频率并不会阻碍服务器单元上的负载平衡。尽管这里给出了三个服务器单元和三个条带组，但此架构可以轻松地扩展为使用其它数量的条带组和其它数量的服务器。对于一个特定媒体对象中不同的数据流的数量依据下列方程式指定条带组的数量（条带化大小 g），其中，DBW 指单个磁盘带宽；BW 指来自服务器单元的可用带宽，H 表示同时访问同一媒体对象的数据流的最大数量。

$$H * BW_i <= DBW_1 + DBW_2 + ... + DBW_g$$

对于最大同时数据流数量较大的热对象而言，会再为其制作一个副本并存储在不同的条带组中。在图 4 中，X' 表示对象 X 的第二个副本。第二个副本 X' 存储在条带组 C 上。这样，对于对象 X，服务器可以为其支持两倍数量的不同的数据流。

客户系统通过一个接纳控制过程与媒体服务器通信，该接纳控制

过程以图 5 所示的软件系统为中介。该软件系统定义了主协调器(coordinator) 40，它负责接纳控制以及生成客户协调器模块，以为给定的客户系统服务客户流（即，写、读、删除以及其它类似操作）。

主协调器是分布式媒体服务器中唯一的集中实体。其位置被所有客户系统知道。主协调器可以由其中一个服务器单元作为宿主。在图 5 中，服务器单元 SE2 是主协调器 40 的宿主。主协调器 40 包含有关构成媒体对象的每个块的物理位置（服务器单元 id、媒体存储设备 id、物理块地址）的信息。每次接纳控制过程成功完成后，主协调器都会选择一个服务器单元作为新客户协调器模块的宿主。在分布式服务器系统中，有多少个客户协调器模块，就有多少个正在服务的不同的流。

客户协调器模块统一地分布在不同的服务器单元间，以避免出现瓶颈。在图 5 中，客户协调器模块在每个服务器单元中显示为 42。与客户 Y 相关联的客户协调器模块被进一步命名为 CC_y。主协调器包括一个规则集，以确保不同服务器单元间的负载均衡。一个客户协调器模块在负载最少时被分配给一个服务器单元。

应记住，一个媒体对象的块是在多个服务器单元（不一定是所有组件）间统一条带化的。与给定的数据流相对应的客户协调器模块会将控制消息发送给适当的服务器单元，以使该服务器单元将被请求的数据块直接发送给相关联的客户。控制消息包括有关媒体存储设备上的块的物理位置的信息以及有关数据必须发送给客户的最后时间的信息。有利的是，数据块不直接经过客户协调器模块。这有助于避免出现瓶颈。

图 5 显示了在数据传输时交换的消息的顺序。消息按照如图 5 中括号所示的号的顺序出现。因此，消息如下进行交换:

（1） 客户 Y 向主协调器 40，服务器单元 SE2 是其宿主，发送读请求。

（2） 如果客户被接纳，主协调器生成客户协调器模块 CC_Y，这里将其分配给服务器单元 SE_x。主协调器还将媒体

块表（media block table, MBT）发送给服务器单元 SE_x。

（3） 客户 Y 直接将对数据块的读请求发送给客户协调器 CC_Y.

（4） 客户协调器模块 CC_Y 使用媒体块表中的信息将控制消息发送给适当的服务器单元，该服务器单元中可能找到被请求的媒体选择的数据对象。

（5） 作出响应的服务器单元 SE_n 直接将被请求的数据块发送给客户 Y。

（6） 重复执行步骤 3-5，直到被请求的媒体选择的所有块都发送到客户 Y 。

通过上述内容应当理解到，本发明的媒体服务器非常适用于要求可扩缩性（最好是线性可扩缩性）的分布式应用。该服务器由多个相互连接且通过高带宽网络与客户系统连接的服务器单元构成。条带组数据布局方案保证了用户与分布式服务器间的无缝连接。每个用户对任何服务器单元都具有相同的连接带宽。由于请求频率不平衡而可能在存储子系统中出现的瓶颈通过条带组数据布局而得以避免。由于条带组数据可以方便地接受热媒体对象的额外副本，因而系统可以方便地支持热对象和冷对象的组合。

尽管在当前的优选实施例中对本方面进行了说明，但应理解，对上述发明所做的特定更改将不背离由所附的权力要求所限定的本发明精神。

图1

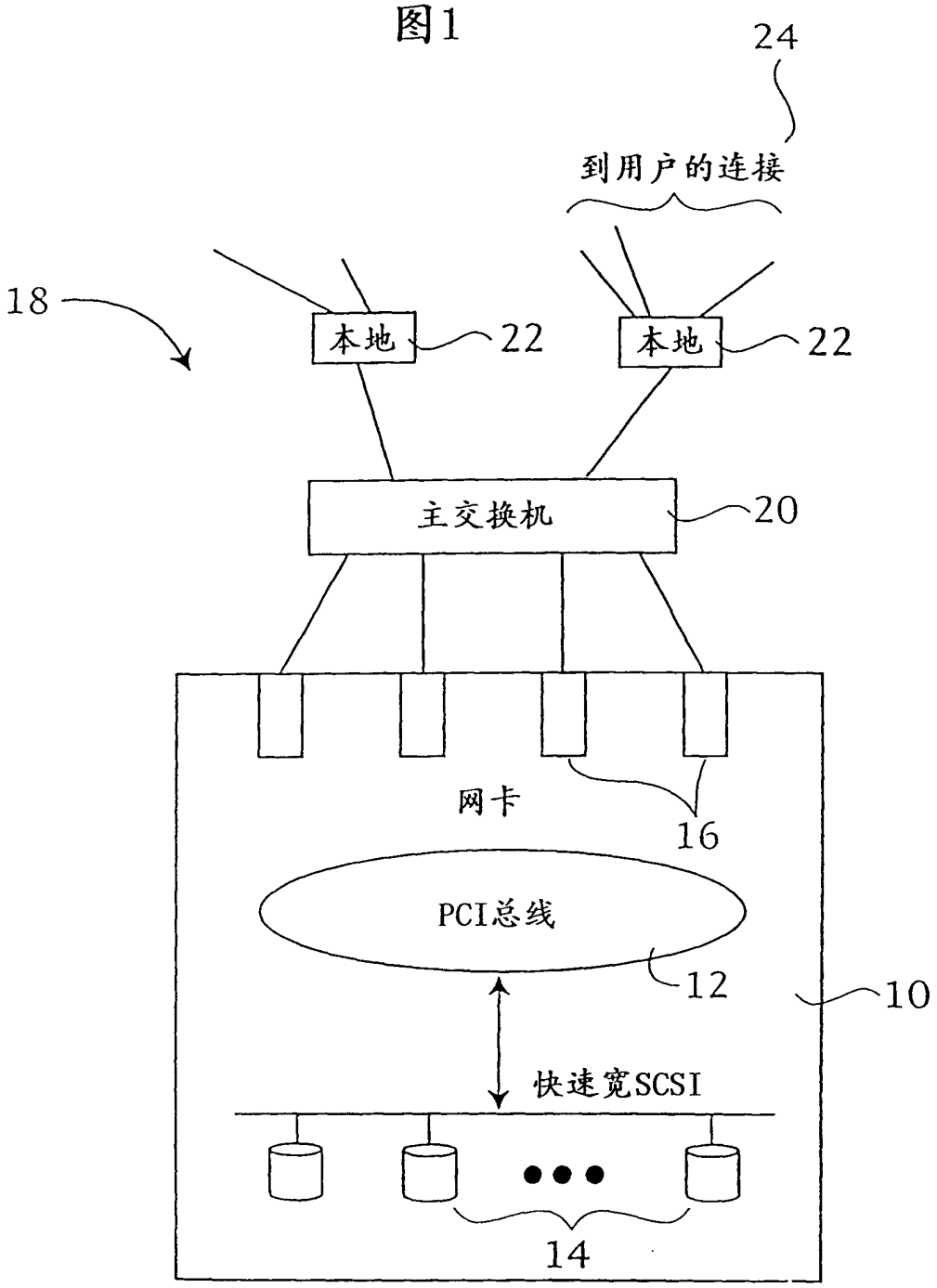


图2

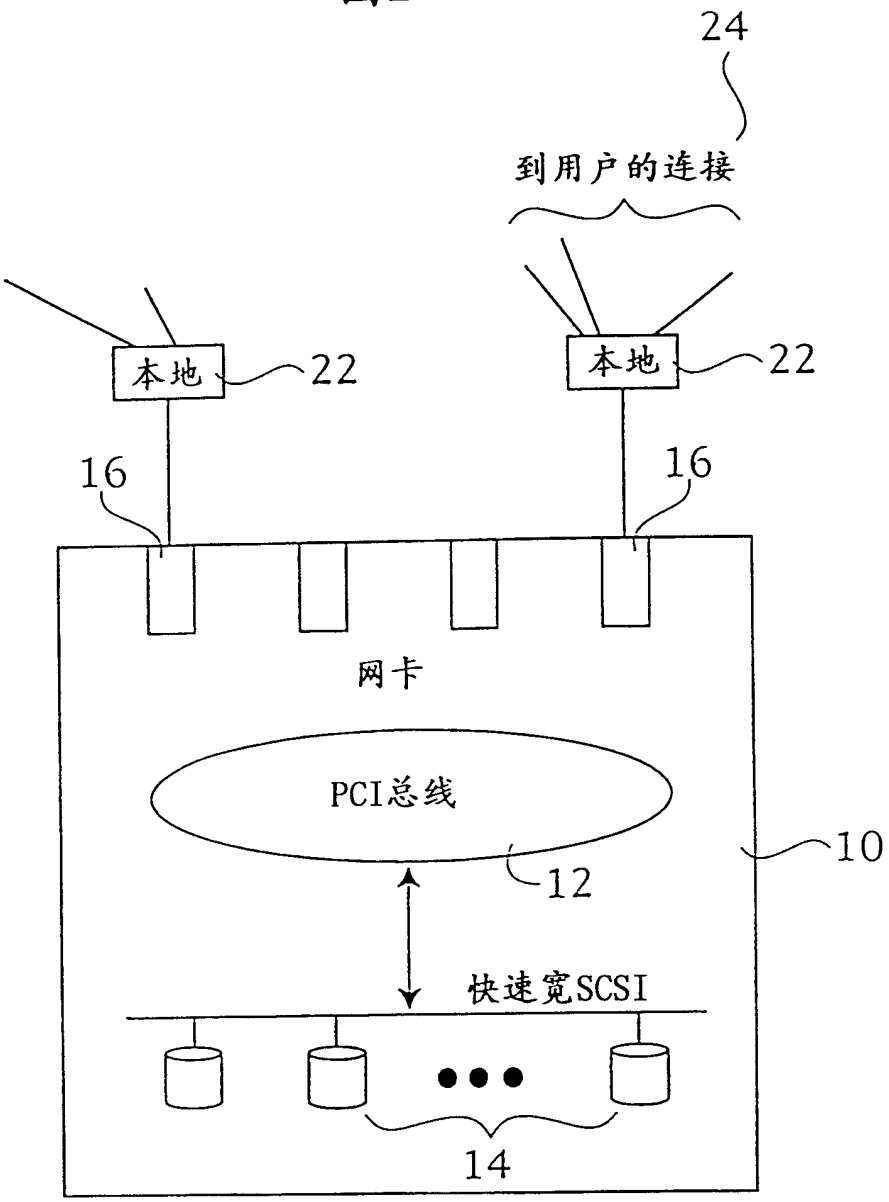


图 3

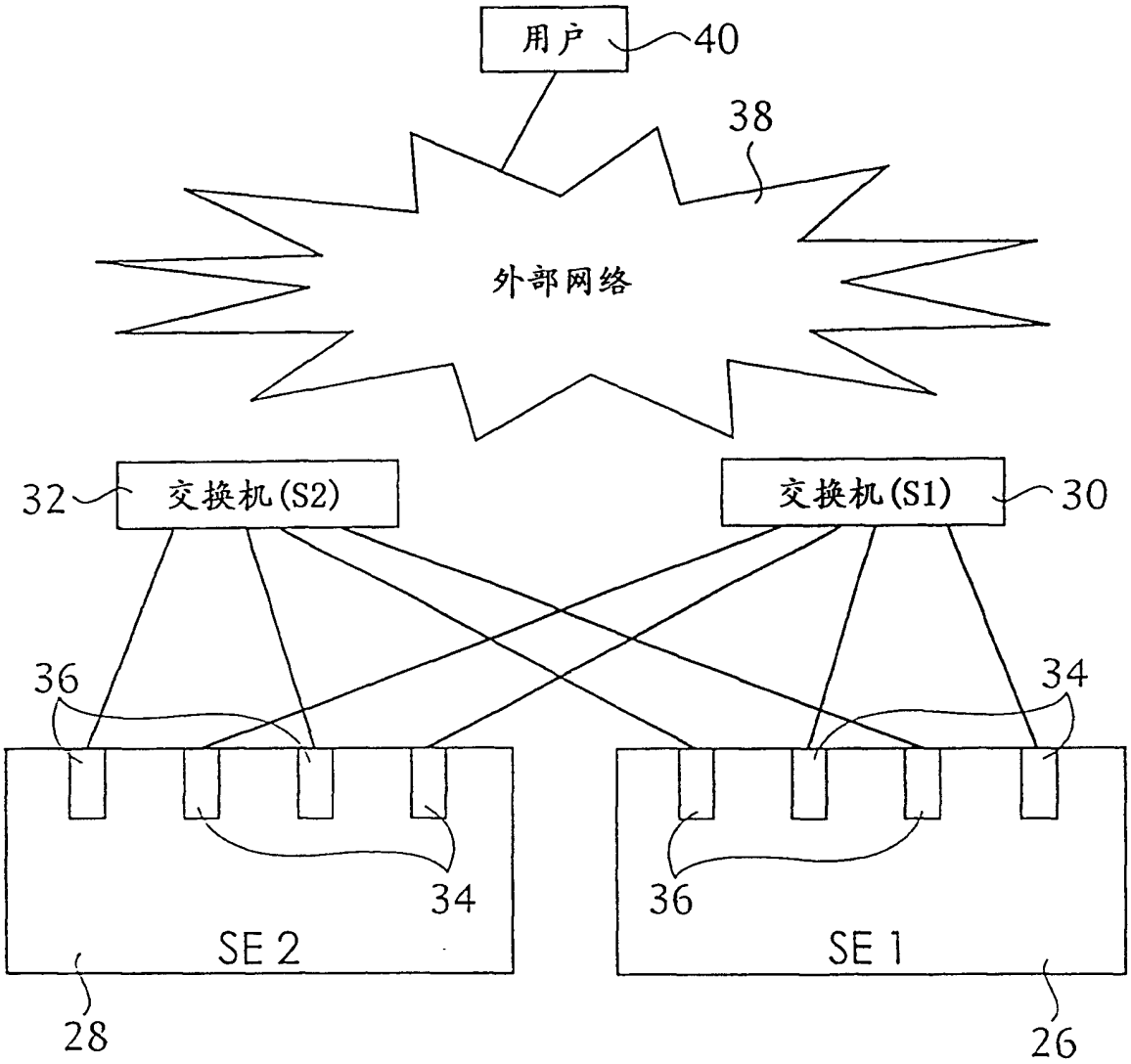


图4

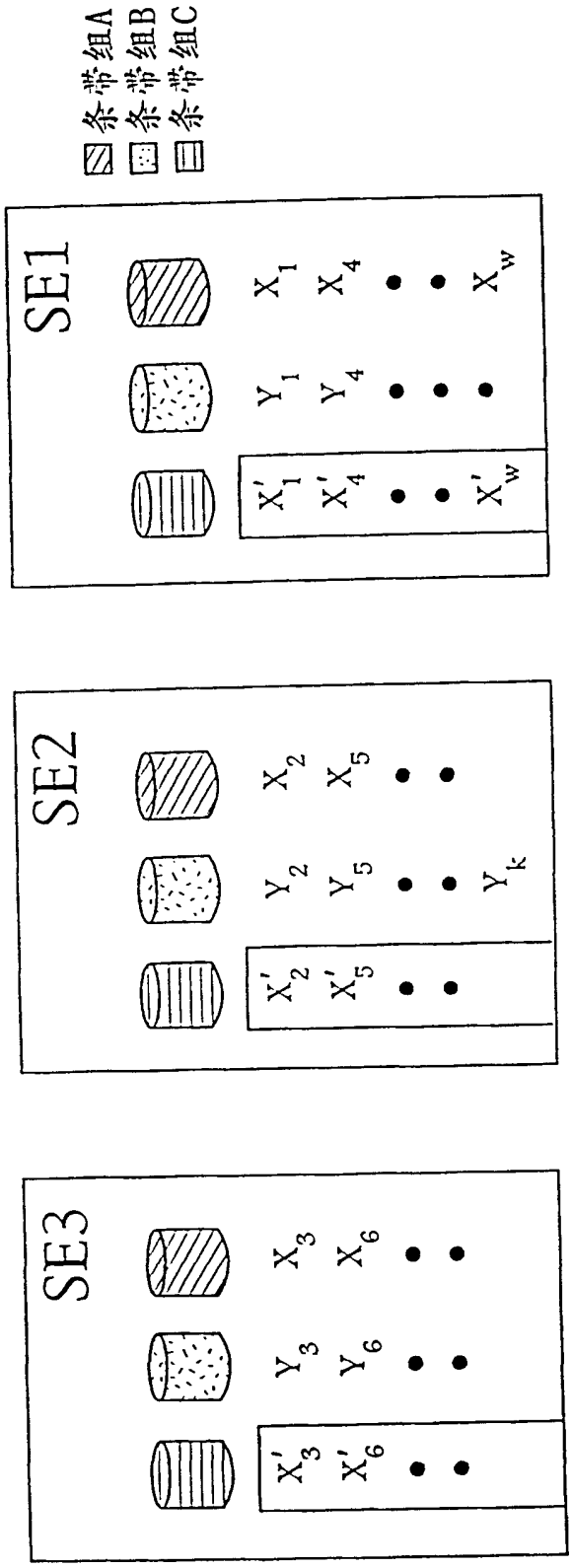
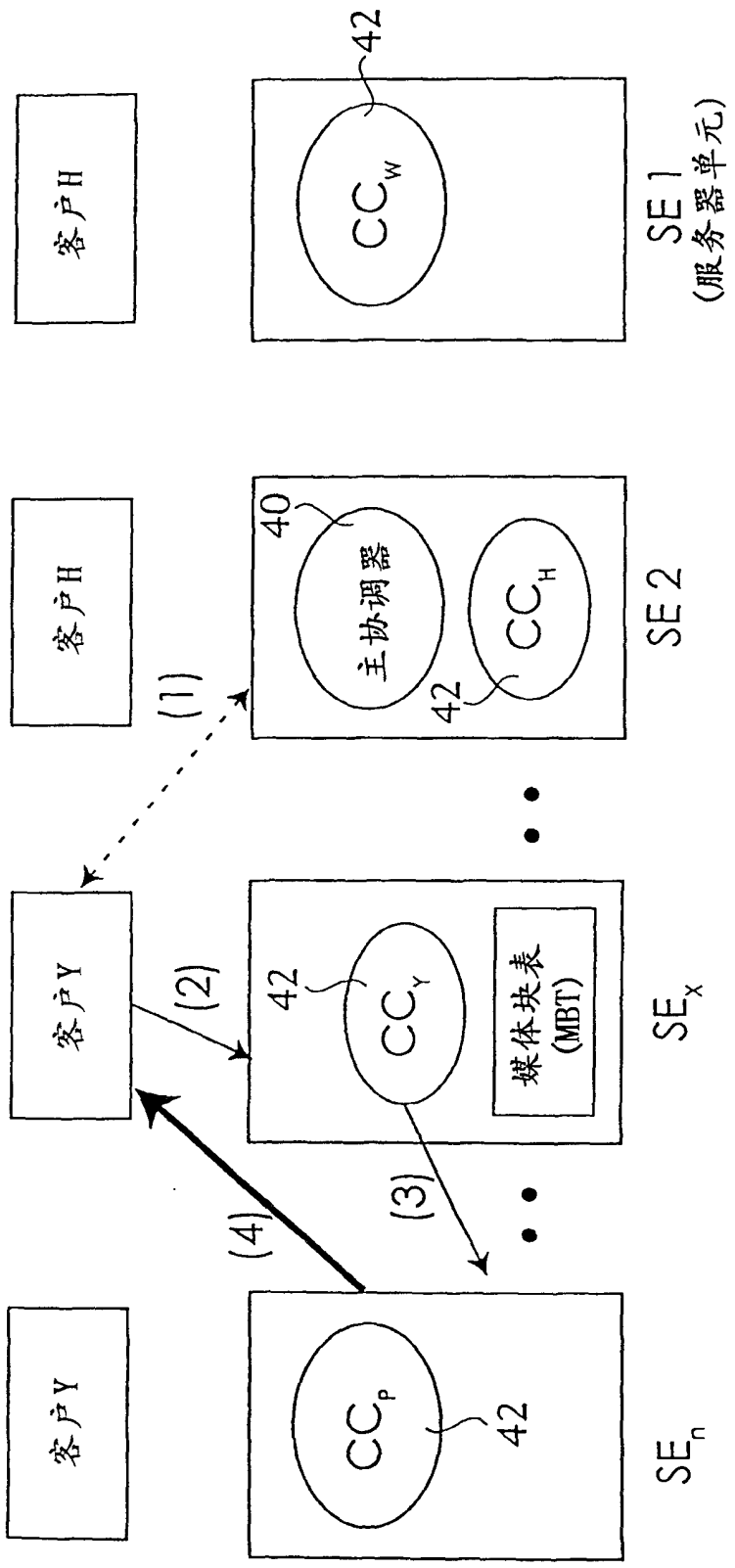


图5


客户H
42
CC_W
SE 1
(服务器单元)
客户H
40
主协调器
42
CC_H
SE 2
(1)
客户Y
(2)
42
CC_Y
媒体块表
(MBT)
SE_X
(3)
(4)
客户Y
CC_P
42
SE_n