[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
H04L 12/56
G06F 12/02

[12] 发明专利申请公开说明书

[21] 申请号 200410078707.0

[43] 公开日 2005 年 4 月 6 日

[11] 公开号 CN 1604565A

[22] 申请日 2004.9.17
[21] 申请号 200410078707.0
[30] 优先权
[32] 2003.10.2 [33] US [31] 10/677,425
[71] 申请人 国际商业机器公司
地址 美国纽约州
[72] 发明人 罗伯特·A·希勒

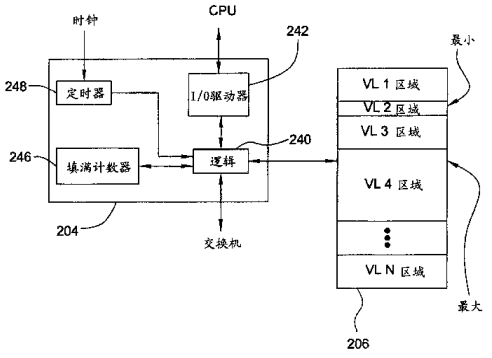
[74] 专利代理机构 北京市柳沈律师事务所
代理人 郭定辉 黄小临

权利要求书 3 页 说明书 6 页 附图 3 页

[54] 发明名称 具有硬件控制缓冲区的共享缓冲器

[57] 摘要

具有硬件控制的缓冲空间区域的缓冲存储器，其中所述硬件控制多个缓冲空间区域的尺寸，以满足特定系统的需求。 所述硬件监视缓冲数据区域随时间的使用情况，并且根据这些缓冲区的使用情况，相应地自动调整缓冲空间区域的尺寸。



ISSN1008-4274

1. 一种存储设备，包括:

缓冲存储器，其具有多个可寻址的存储器寄存器;

计数器，其具有多个存储寄存器;

逻辑网络，用来将数据写入所述缓冲存储器和从所述缓冲存储器读取数据，所述逻辑网络用来将所述缓冲存储器分区为多个缓冲区，其中所述逻辑网络写入和读取来自多个数据类别的数据到所述多个缓冲区，使得每个数据类别被写入不同的缓冲区和从不同的缓冲区中读取，并且每次缓冲区达到预定使用水平时，所述逻辑网络递增与该缓冲区相关联的存储寄存器；以及

定时器，用来定期发送定时信号给所述逻辑网络;

其中，作为所述定时信号的响应，所述逻辑网络从所述计数器寄存器中调出数据，并且重新分区所述缓冲存储器，以使得被更充分利用的缓冲区被分配更多的可寻址存储器寄存器。

2. 根据权利要求 1 的存储设备，其中所述逻辑网络向较不经常被使用的缓冲区分配更少的可寻址存储器寄存器。

3. 根据权利要求 1 的存储设备，其中每个缓冲区总是被分配至少最少数目的可寻址存储器寄存器。

4. 根据权利要求 1 的存储设备，其中所述预定使用水平为满。

5. 根据权利要求 1 的存储设备，其中当最少被使用的缓冲区被分配最少数目的可寻址存储器寄存器时，所述逻辑网络分配较不经常被充分利用、但具有多于最少数目的可寻址存储器寄存器的缓冲区。

6. 根据权利要求 1 的存储设备，其中所述数据类别表示虚拟通路。

7. 根据权利要求 1 的存储设备，其中所述定时信号启动为对于所述多个存储寄存器的重置。

8. 一种交换网络，包括:

网络交换机;

卡适配器，用来发送并接收来自所述网络交换机的数据；以及

存储设备，用来存储用于和来自所述卡适配器的数据，所述存储设备包括:

缓冲存储器，其具有多个可寻址的存储器寄存器;

计数器，其具有多个存储寄存器；

逻辑网络，用来将数据写入所述缓冲存储器和从所述缓冲存储器读取数据，所述逻辑网络用来将所述缓冲存储器分区为多个缓冲区，其中所述逻辑网络写入和读取来自多个数据类别的数据到所述多个缓冲区，使得每个数据类别被写入不同的缓冲区和从不同的缓冲区中读取，并且每次缓冲区达到预定使用水平时，所述逻辑网络递增与该缓冲区相关联的存储寄存器；以及

定时器，用来定期发送定时信号给所述逻辑网络；

其中，作为所述定时信号的响应，所述逻辑网络从所述计数器寄存器中调出数据，并且重新分区所述缓冲存储器，以使得被更充分利用的缓冲区被分配更多的可寻址存储器寄存器。

9. 根据权利要求 8 的交换网络，其中所述逻辑网络向较不经常被使用的缓冲区分配更少的可寻址存储器寄存器。

10. 根据权利要求 8 的交换网络，其中每个缓冲区总是被分配至少最少数目的可寻址存储器寄存器。

11. 根据权利要求 8 的交换网络，其中所述预定使用水平为满。

12. 根据权利要求 8 的交换网络，其中当最少被使用的缓冲区被分配最少数目的可寻址存储器寄存器时，所述逻辑网络分配较不经常被充分利用、但具有多于最少数目的可寻址存储器寄存器的缓冲区。

13. 根据权利要求 8 的交换网络，其中所述数据类别表示虚拟通路。

14. 根据权利要求 8 的交换网络，其中所述定时信号启动为对于所述多个存储寄存器的重置。

15. 根据权利要求 8 的交换网络，其中所述卡适配器为主机通道适配器。

16. 根据权利要求 15 的交换网络，其中所述主机通道适配器为 Infiniband 主机通道适配器。

17. 根据权利要求 8 的交换网络，其中所述卡适配器为目标通道适配器。

18. 根据权利要求 17 的交换网络，其中所述目标通道适配器为 Infiniband 目标通道适配器。

19. 根据权利要求 8 的交换网络，还包括：中央处理单元，用来发送数据给所述存储设备以及从所述存储设备接收数据。

20. 一种用来管理包含多个可寻址存储器寄存器的缓冲器的方法，包含：

将所述缓冲器分区为由硬件控制的多个缓冲区；

在一时间段内使用所述硬件监视每个缓冲区的使用情况；以及

根据所监视的使用情况，使用所述硬件在所述缓冲区中重新分配所述存储器寄存器。

21. 根据权利要求 20 的方法，还包含：将每个缓冲区与数据类别相关联。

22. 根据权利要求 20 的方法，其中至少一个缓冲区与表示虚拟通路的数据类别相关联。

具有硬件控制缓冲区的共享缓冲器

技术领域

一般地讲，本发明的实施方式涉及缓冲存储器。更具体地讲，本发明涉及具有对于缓冲区的基于硬件的自我调整的共享缓冲器。

背景技术

在处理器与输入/输出设备之间传送数据一般通过根据预定的规格（例如PCI（外围组件互连）总线体系结构）并行传送数据来进行。然而，因为市场对于提高的数据传送速度的需求，并行数据传送的速度限制已经变得很明显。由于这些限制，所以在诸如高端服务器等应用中使用串行数据传送，例如Infiniband体系结构，已经开始替代并行数据传送体系结构。

在原理上，串行数据传送逐比特地串行发送数据。然而，高性能串行数据传送体系结构，例如Infiniband，一般使用允许高速传送速度的多个数据虚拟通路（通道）。通过将主机通道适配器（HCA）以及目标通道适配器（TCA）通过Infiniband交换机连接在一起，形成Infiniband虚拟通路，其中主机通道适配器为服务器内的I/O引擎，目标通道适配器为外部I/O引擎。每个HCA与TCA都支持多达16个虚拟通路（VL）。Infiniband互连方案被称为构造（fabric）。由于其速度与多个VL，Infiniband体系结构在单独一个子网中可以在铜线上（长达17米）以及在光缆上（长达10公里）以2.5GBps以上的传送速度支持成千上万的节点。

Infiniband规格要求充足的缓冲空间，以使每个虚拟通路存储完整的最大发送单元（MTU），其中MTU为Infiniband构造中的最大数据包大小。然而，该缓冲空间要求表示了最小值，并且附加的缓冲空间可以提高系统性能。因此，大缓冲空间是有利的。

因为与许多较小存储器相比，大存储器具有性能与成本优势，所以在Infiniband通道适配器（HCA或TCA）中使用数目减少了的、理想为1个的缓冲存储器设备是非常有利的。因此，形成具有至少包含对于每个虚拟通路所需最小缓冲空间的尺寸的大型共享缓冲存储器是有利的。为了提高性能，

希望有对于虚拟通路的附加缓冲空间。简单地为每个 VL 添加附加存储器可能大大提高共享缓冲存储器的大小与成本。因为有些虚拟通路很少使用多于其最小所需的缓冲空间，所以给这些虚拟通路分配更多的缓冲空间是一种浪费。然而，其他虚拟通路可以从增加了的缓冲空间中得到很大的好处。因此，一种动态增大对于能够从增大了的缓冲空间获益的虚拟通路的缓冲空间的大小的方法将是有用的。

共享缓冲存储器是公知的。对于缓冲空间尺寸的软件控制已经使许多应用受益。不幸的是，不容易设计并实现允许如在 Infiniband 构造中所发生的不同并且变化的流量负载的软件配置的缓冲空间。常见的是，即使软件配置的缓冲空间可用，这些软件空间也是或者没有被配置或者配置不良，这是因为控制软件没有很好地理解系统的性质、其特定应用以及缓冲空间要求的变化来分配适当的缓冲空间大小。

相同的缓冲存储器芯片可能以非常不同的方式用于非常不同的应用，这使软件控制缓冲空间的问题复杂化。考虑到对于大型缓冲存储器在经济上与性能上的需要以及软件控制的缓冲空间的局限性，具有硬件控制的缓冲空间尺寸、并且该硬件调整缓冲空间尺寸以满足特定系统需求的缓冲存储器将是有用的。

发明内容

本发明提供了具有硬件控制的缓冲空间区域的共享缓冲存储器，其中所述硬件根据特定应用的要求控制缓冲空间区域的尺寸。所述硬件监视一段时间上多个缓冲区的使用情况，并且根据各个缓冲区的使用情况，相应地自动调整多个缓冲区的尺寸。

根据本发明，由至少两个不同的数据类别所共享的大型缓冲存储器内的边界由硬件分配。数据类别为只与一个网络相关联的数据。该边界将缓冲存储器分为两个部分，每个数据类别占一个部分。每个数据类别可以被存储到数据缓冲存储器空间其自己的部分中。该硬件监视这两个部分的使用情况，以确定在一时间段内哪个数据类别使用其部分更多。如果一个数据类别使用其部分比另一个多，则该硬件动态地重新调整所述边界，以为更多需要的数据类别提供较大的部分，并为较少需要的数据类别提供较小的部分。确定哪个数据类别使用其部分更多的一种方法是对给定时间段内每个部分被填满的

次数进行计数。

可替换地，大型缓冲存储器可以被多个数据类别共享，如在 Infiniband 虚拟通路中。给每个数据类别分配缓冲存储器的一个区域供其使用。硬件监视各种区域以确定在给定时间段内各个区域如何频繁地被充分利用。如果一个区域比其他区域利用得更多，则该硬件动态地调整区域边界，使得最经常填满其区域的数据类别分配较大的区域，其中从较少充分使用其区域的数据类别取得附加的存储器。

在某些应用中，每个数据类别都保留了最小大小。在这种情况下，最经常填满其区域的数据类别分配较大的区域，其中从较不经常充分使用其区域、但大于所保留的最小大小的数据类别取得附加的存储器。

在 Infiniband 体系结构的特定情况下，如果与特定 VL 相关联的一个数据类别需要较大的存储区，则该 VL 的区域被硬件动态调整，以满足该特定 VL 的需要。所有 VL 的数据缓冲器使用情况由中央单元监视，该中央单元对可配置时间段内每个缓冲区的总充满状态计数。在该时间段结束时，比较各计数值，并且调整 VL 缓冲区，通过减少与最不经常使用的 VL 缓冲区相关联的缓冲区的尺寸，使得最经常被使用的 VL 缓冲区可以使用较大的缓冲区。然后，清除计数，并且再次开始计数过程。一旦 VL 的缓冲区被减少到预定的最小大小，则不再减少该缓冲区的尺寸，但是下一个最不忙的缓冲区被重新确定尺寸，以具有较小的缓冲区。

附图说明

为了详细地理解上述本发明的特征，通过参照实施方式（其中的某些部分在附图中示出），具体描述上面所概括的本发明。然而，应该注意附图只是显示了本发明的典型实施方式，因此不能被认为是对本发明范围的限制，因为本发明可允许其他等效的实施方式。

图 1 显示 Infiniband 构造。

图 2 显示主机通道适配器与相关主机设备之间 Infiniband 连接，并且还显示了多个虚拟通路；

图 3 显示存储器控制器以及被分为多个缓冲区的大型缓冲存储器，其中在启动时每个虚拟通路占一个缓冲区；以及

图 4 显示重新配置之后的存储器控制器与缓冲存储器。

为了帮助理解，在可能时，使用相同标号来表示这些附图所共有的相同元素。

具体实施方式

本发明提供了缓冲存储器及其应用，其具有硬件控制的缓冲区，其中硬件控制缓冲区的尺寸以满足特定系统的需求。该硬件在一段时间上监视缓冲空间的使用情况，然后根据缓冲区的使用情况自动调整缓冲区的尺寸。本发明尤其适用于具有相当恒定、定义完备的数据流量模式（data traffic pattern）的应用。在这些情况下，数据缓冲空间将保持非常高的利用率，同时具有非常少的缓冲管理花费。

图 1 显示一般的 Infiniband 构造。通过铜线或光缆网络，多个交换网络 102、104 以及 106 串行互连。在有些应用中，这些交换网络可以在控制上连接到可以（例如）连接到互连网的网络链接。互连与交换网络从和/或向一个或多个节点 110 传送信息包，这些节点可以是 CPU、网络链接、打印机、和/或其他类型的 I/O 驱动器。

图 2 显示一般交换网络 198，其被显示分为交换机 200、主机卡适配器（HCA）202、存储器控制器 204 以及缓冲存储器 206。缓冲存储器 206 包含以地址组织并可寻址的多个存储器寄存器。图 2 还显示多个节点 210，其被显示为与交换网络 198 交互的 CPU。节点 210 不是交换网络 198 的部分，但也被显示以帮助理解交换网络 198 的操作。图 2 还显示多个虚拟通路 VL1-VLN，其不是物理实体，而是在理论上表示数据类与通道，其中每个都与缓冲存储器 206 中的缓冲区相关联。虚拟通路数据被交换进入或离开主机卡适配器 202。一般交换网络 198 串行连接到外部环境。

图 3 更详细地显示存储器控制器 204 与缓冲存储器 206。虽然图 3 显示了两个离散的设备，但是在实践中存储器控制器 204 与缓冲存储器 206 最好一起制造在单一芯片设备中。存储器控制器 204 以硬件实现（其应该被理解为包含在硬件中执行的固件）。

存储器控制器 204 包含写数据到缓冲存储器 206 和从缓冲存储器 206 读取数据的逻辑网络 240。逻辑网络 240 还传递数据到 I/O 驱动器 242 和从 I/O 驱动器 242 传递数据，其中 I/O 驱动器 242 连接到节点 210（参看图 2），并传递数据到主机连接适配器 202 和从主机连接适配器 202 传递数据（参看图

2)。逻辑网络 240 连接到填满计数器 246 与定时器 248。随后将讨论填满计数器 246 与定时器 248。

在原理上，缓冲存储器 206 分区为多个虚拟通路缓冲区，显示为 VL1 区至 VLN 区。虚拟通路缓冲区作用为虚拟通路的短期存储器。因此，当要在虚拟通路 VL1 上发送数据时，VL1 区存储该数据，直至主机连接器适配器 HCA 202 能够处理该数据。类似地，当正在 VL1 上接收数据时，VL1 区存储所接收的数据，直至节点 210 能够接受该数据。当虚拟通路缓冲区变满时，即当其不能处理更多信息时，存储器控制器 204 启动系统延迟，以使 HCA 202 或节点 210 能够接受来自满虚拟通路区的数据。因此，填满虚拟通路缓冲区造成系统延迟。

仍然参照图 3，每次虚拟通路缓冲区变满时，逻辑 240 递增在与该满虚拟通路相关联的填满计数器 246 中的寄存器。因此，填满计数器 246 跟踪每个虚拟通路缓冲区造成了多少延迟。虽然上述描述了当虚拟通路缓冲区变满时递增寄存器，但在实践中可以使用其他条件。例如，当达到特定水平（例如 80%满）时，或者在每次从虚拟通路区写入和/或读取时，递增寄存器。重要的是：使用虚拟通路缓冲区使用情况的某种表示，从而可以调整虚拟通路缓冲区的尺寸以改进系统操作。

在预定时间段之后，定时器 248 向逻辑 240 发信号。作为响应，逻辑 240 查询填满计数器 246 以确定在该预定时间段内每个虚拟通路缓冲区被填满（或者达到某些其他使用标志）多少次。如果一个（或多个）虚拟通路缓冲区比其他虚拟通路缓冲区被更充分地利用，则逻辑 240 重新分配虚拟通路缓冲区尺寸（分区），使得被充分利用的虚拟通路缓冲区分配更多的缓冲存储器 206，并且使得较不充分利用的另一虚拟通路缓冲区分配较小部分的缓冲存储器 206。这在图 4 中示出。如图所示，虚拟通路缓冲区 2 与 3 现在具有较少的缓冲存储器 206，而虚拟通路缓冲区 4 具有大得多的缓冲存储器 206。

虚拟通路缓冲区的尺寸可以由逻辑 240 控制。不需要物理分区，这是因为逻辑 240 启动所有读取与写入，并且因此分配缓冲存储器 206 的所有区域。

因为在 Infiniband 系统中使用存储器控制器 204，并且 Infiniband 体系结构对于每个虚拟通路要求最小大小的缓冲存储器，所以配置存储器控制器以使得没有虚拟通路缓冲区会减少到最小要求之下。

虽然图 1-4 直接有关于 Infiniband 体系结构，但在实践中，本发明可以应

用于存在不同数据类的其他系统。

虽然上述针对于本发明的实施方式，但是在不脱离权利要求所确定的本发明的范围的前提下，可以设想本发明的其他与进一步的实施方式。

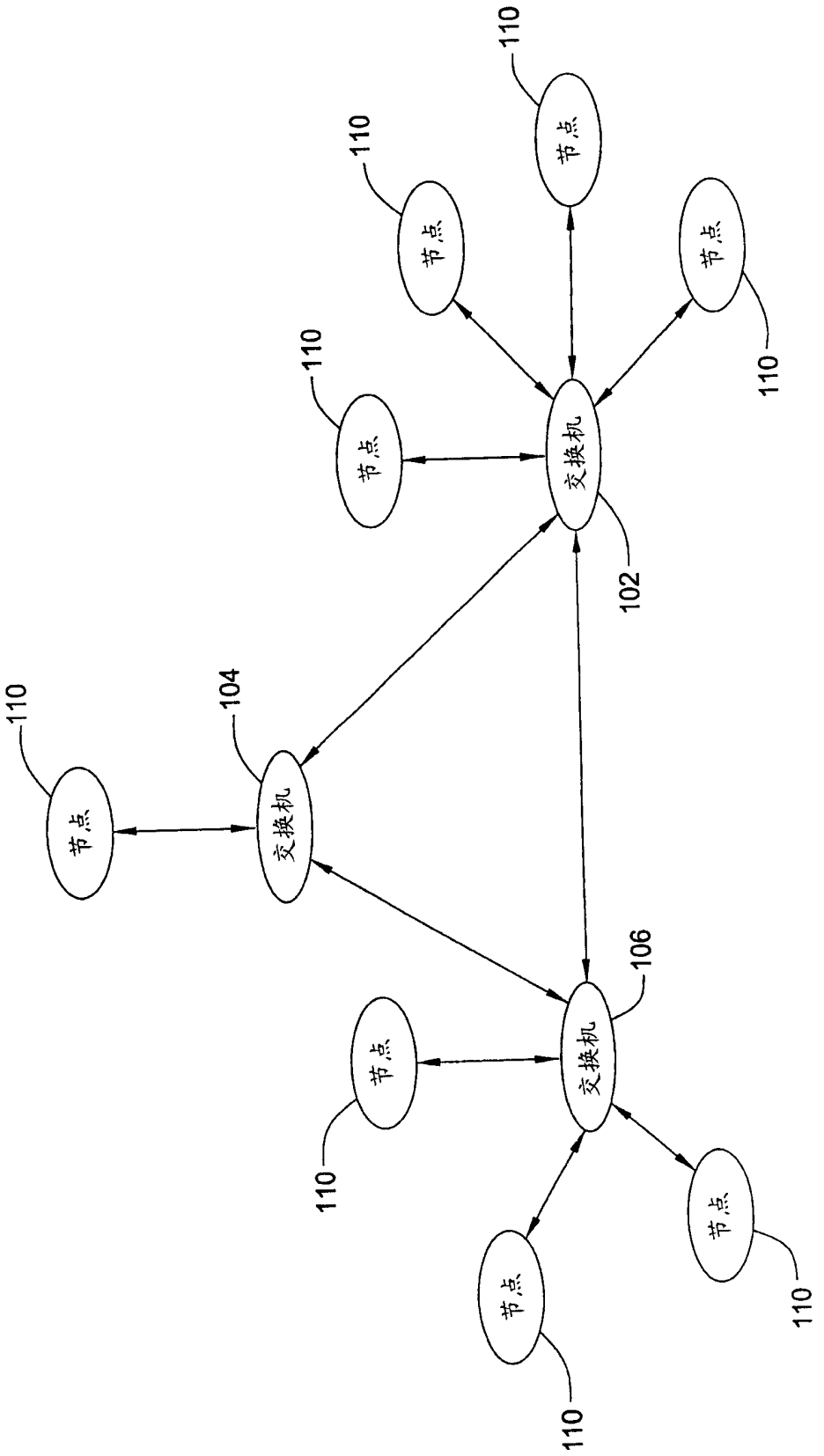


图 1

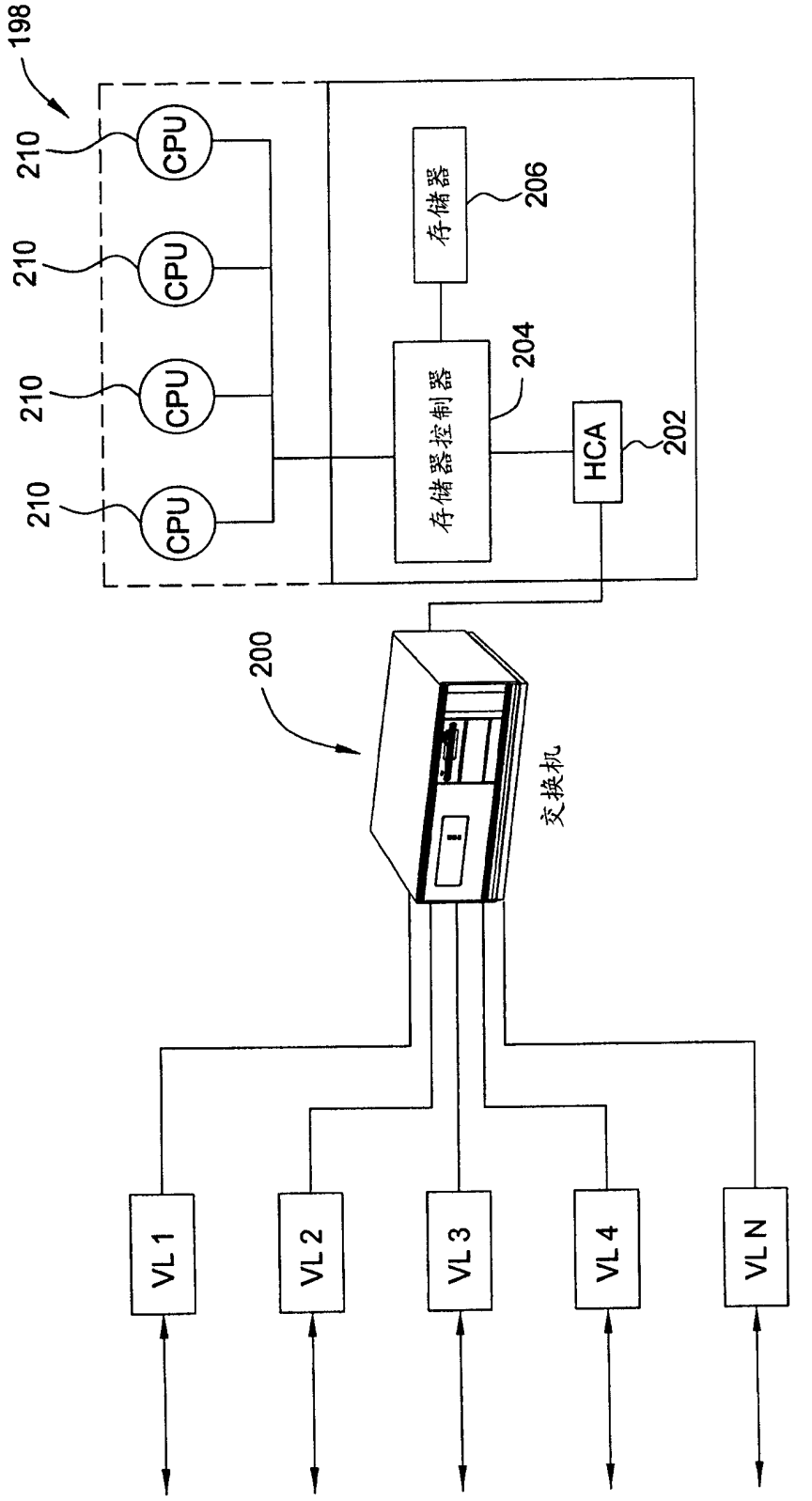

198
210
CPU
210
CPU
210
CPU
210
CPU
存储器
206
存储器控制器
204
HCA
202
200
交换机
VL 1
VL 2
VL 3
VL 4
VL N

图 2

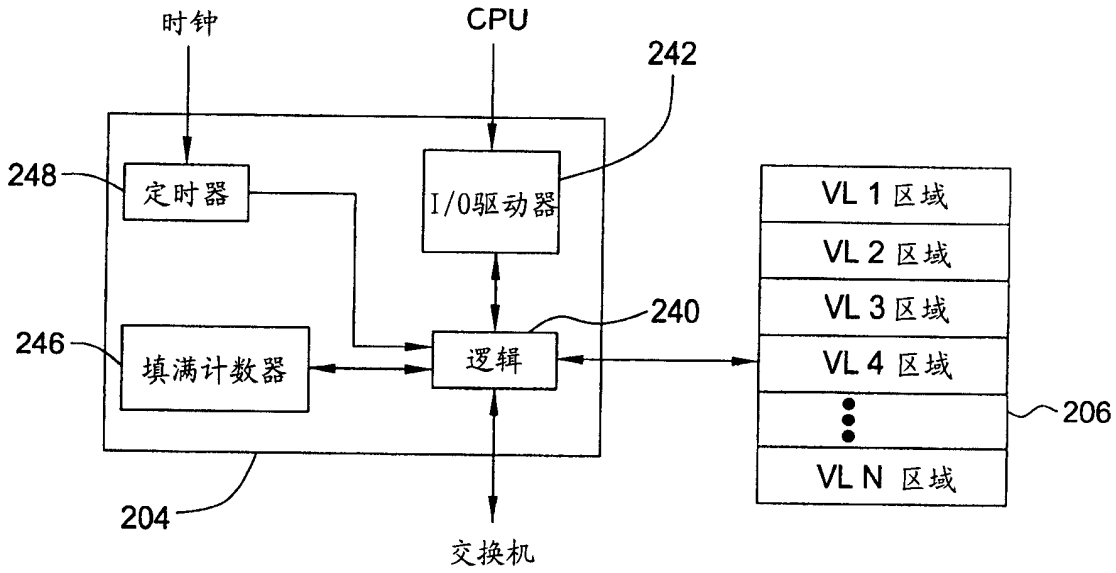

时钟
CPU
242
248
定时器
I/O驱动器
VL 1 区域
VL 2 区域
VL 3 区域
VL 4 区域
VL N 区域
206
240
246
填满计数器
逻辑
204
交换机

图 3

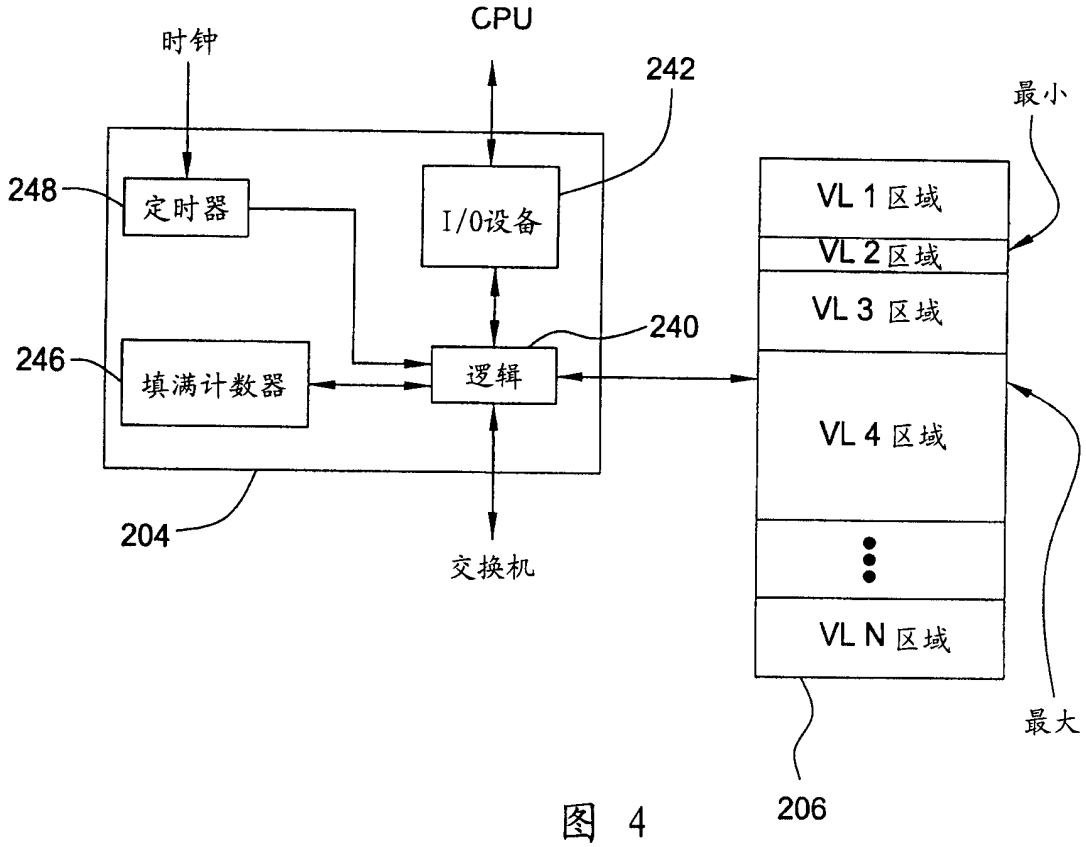

CPU
时钟
242
最小
248
定时器
I/O设备
VL 1 区域
VL 2 区域
VL 3 区域
240
246
填满计数器
逻辑
VL 4 区域
204
交换机
VL N 区域
最大
206

图 4