[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 12/10 (2006.01)
G06F 15/167 (2006.01)
G06F 13/14 (2006.01)

[12] 发明专利说明书

专利号 ZL 200410042742.7

[45] 授权公告日 2007 年 11 月 14 日

[11] 授权公告号 CN 100349142C

[22] 申请日 2004.5.25
[21] 申请号 200410042742.7
[73] 专利权人 中国科学院计算技术研究所
地址 100080 北京市海淀区中关村科学院南路 6 号
[72] 发明人 史 岗 胡明昌 尹宏达 胡伟武 唐志敏
[56] 参考文献
CN1451114A 2003.10.22
CN1474296A 2004.2.11
US5592625A 1997.1.7
CN1421789A 2003.6.4

VIA(Virtual Interface Architecture)上的软件 DSM 系统实现和性能. 史岗，尹宏达，胡明昌，胡伟武. 计算机学报，第卷号：26 卷第期号：12 期. 2003
审查员 吴翔晖
[74] 专利代理机构 北京泛华伟业知识产权代理有限公司
代理人 王凤华

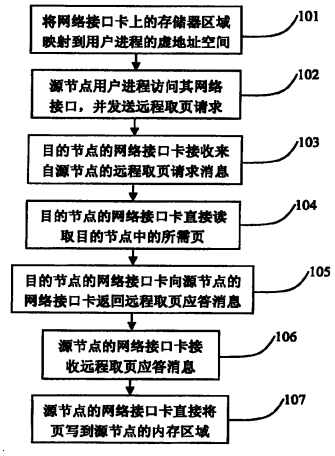
权利要求书 2 页 说明书 5 页 附图 2 页

[54] 发明名称

一种用于虚拟共享存储系统的远程取页方法及网络接口卡

[57] 摘要

本发明公开了一种用于虚拟共享存储系统的远程取页方法及网络接口卡。该方法将网络接口卡上的存储器区域映射到用户进程的虚地址空间；源节点的用户进程直接访问其网络接口卡，产生并向目的节点的网络接口卡发送远程取页请求消息；目的节点的网络接口卡直接读取目的节点中的所需页；目的节点的网络接口卡向源节点的网络接口卡返回远程取页应答消息；源节点的网络接口卡直接将所述页写到源节点的内存区域。该网络接口卡增加了帧头分析逻辑、RDMA 操作逻辑和虚实地址转换逻辑。在本发明中，用网络接口卡上的硬件处理大部分的协议开销，用户进程和网络接口卡可以双向直接访问，可以在不打断远程节点 CPU 的当前工作的情况下实现远程取页操作。



1、一种用于虚拟共享存储系统的远程取页方法，所述虚拟共享存储系统包括用网络接口卡和交换机互联的源节点和目的节点；包括步骤：

1）将网络接口卡上的存储器区域映射到用户进程的虚地址空间；

2）源节点的用户进程直接访问其网络接口卡，产生并向目的节点的网络接口卡发送远程取页请求消息；

3）目的节点的网络接口卡接收来自源节点的远程取页请求消息；

4）目的节点的网络接口卡直接读取目的节点中的所需页；

5）目的节点的网络接口卡向源节点的网络接口卡返回远程取页应答消息，该消息中包含步骤4）中读取的页；

6）源节点的网络接口卡接收所述远程取页应答消息；

7）源节点的网络接口卡直接将所述页写到源节点的内存区域。

2、根据权利要求1所述的用于虚拟共享存储系统的远程取页方法，其特征在于，所述远程取页请求消息包括一个消息类型标志，所述目的节点的网络接口卡在接收到该远程取页请求消息后对该类型标志进行判断；所述远程取页应答消息包括一个消息类型标志，所述源节点的网络接口卡在接收到该远程取页应答消息后对该类型标志进行判断。

3、根据权利要求1所述的用于虚拟共享存储系统的远程取页方法，其特征在于，所述远程取页请求消息中包含要读取页在目的节点上的虚拟地址、数据长度和该页返回源节点的虚拟地址。

4、根据权利要求1所述的用于虚拟共享存储系统的远程取页方法，其特征在于，在步骤 4）中网络接口卡读取所需页时，还要进行虚实地址转换，以便将虚拟地址转换为物理地址；然后目的节点的网络接口卡将获得的物理地址所指向的页以DMA方式读取到该网络接口卡。

5、根据权利要求1所述的用于虚拟共享存储系统的远程取页方法，其特征在于，所述远程取页应答消息中还包含要读取的页返回源节点的虚拟地址。

6、根据权利要求5所述的用于虚拟共享存储系统的远程取页方法，其特征在于，在步骤 7）中网络接口卡向所述内存区域写页时，还要进行虚实地址转换，以便将虚拟地址转换为物理地址；网络接口卡根据该物理地址向所述内存区域写页。

7、根据权利要求6所述的用于虚拟共享存储系统的远程取页方法，其特征在于，在源节点的网络接口卡向所述内存区域写页时，将源节点用户进程的虚页面锁定在

其物理内存中。

8、根据权利要求 4 或 6 所述的用于虚拟共享存储系统的远程取页方法，其特征在于，所述虚实地址转换由在网络接口卡上提供的虚实地址转换缓冲器完成，所述虚实地址转换缓冲器提供了虚拟地址与物理地址的匹配表项。

9、根据权利要求 8 所述的用于虚拟共享存储系统的远程取页方法，其特征在于，在进行虚实地址转换时，如果在虚实地址转换缓冲器没有找到与虚拟地址匹配的物理地址，则产生中断，由节点主机的网卡驱动程序进行虚实地址转换；网卡驱动程序完成虚实地址转换后将获得的物理地址写入网络接口卡，并更新虚实地址转换缓冲器内的匹配表项，同时返回物理地址。

10、一种实施权利要求 1 所述方法的网络接口卡，包括 PCI 接口控制器通过本地总线分别与消息传递操作逻辑、控制寄存器逻辑、发送 FIFO 和接收 FIFO 电连接；所述发送 FIFO 的一端通过发送逻辑与发送驱动电连接；所述接收 FIFO 的一端通过接收逻辑和接收驱动电连接；其特征在于，还包括：

帧头分析逻辑，用于分析远程取页请求消息和远程取页应答消息的帧头；所述帧头分析逻辑一端通过本地总线与 PCI 接口控制器电连接，另一端与接收 FIFO 电连接；

RDMA 操作逻辑，用于控制对内存读写操作，以便将页面从内存中直接读取或者写入；所述 RDMA 操作逻辑通过本地总线与 PCI 接口控制器电连接，并分别与所述帧头分析逻辑和虚实地址转换逻辑电连接；

虚实地址转换逻辑，用于虚拟地址和物理地址的转换；所述虚实地址转换逻辑通过本地总线与 PCI 接口控制器电连接。

一种用于虚拟共享存储系统的远程取页方法及网络接口卡

技术领域

本发明涉及计算机系统，更具体地说，涉及一种用于虚拟共享存储系统的远程取页方法。

背景技术

由于共享存储系统（下同）的易编程性，研究人员考虑在用消息传递进行通信的多台机器上实现共享存储的编程模式，这便是虚拟共享存储。虚拟共享存储同时具有易编程性而且由于大部分虚拟共享存储系统在机群上实现，具备良好的性能价格比。虚拟共享存储系统多以页为共享粒度。

远程取页指源节点发生缺页时产生缺页信号，当信号处理程序发现共享的虚拟地址不在源节点地址范围内时产生远程取页请求，目的节点接收到取页请求后根据虚拟地址查找对应的物理地址，然后将物理地址对应的页发送回源节点。传统的虚拟共享存储系统关于远程取页的做法是：源节点发现缺页中断后向目的节点的相应进程（而不是网络接口卡硬件）发送取页请求的消息，目的节点的 CPU 的正常工作将被取页请求打断，把所要读取的页面内容以消息形式封装并发送消息，源节点接收返回的读页面返回消息，源节点的 CPU 再次被打断来解封消息并将消息体（页面内容）送到对应的页面。

如上所述，传统的远程取页方法需要对远程主机产生中断，远程主机接收中断后停止正在进行的计算进行中断服务，使计算和通信不能重叠，降低了并行计算的效率。同时；将页的内容以消息的形式进行封装和拷贝，增加了额外的开销。

因此需要有一种方法，能够减小远程取页开销，使计算和通信重叠。

发明内容

本发明的目的之一在于提供一种远程取页方法，减小远程取页的协议处理开销；本发明的目的之二在于提供一种远程取页方法，在进行远程取页时不打断远程主机的计算，获得计算与通信的重叠；本发明的目的之三在于提供一种远程取页方法，使远程取页能快速返回。

为了实现上述目的，本发明提供一种用于虚拟共享存储系统的远程取页方法，所述虚拟共享存储系统包括用网络接口卡和交换机互联的源节点和目的节点；包括步骤：

1）将网络接口卡上的存储器区域映射到用户进程的虚地址空间；

2）源节点的用户进程直接访问其网络接口卡，产生并向目的节点的网络接口卡发送远程取页请求消息；

3）目的节点的网络接口卡接收来自源节点的远程取页请求消息；

4）目的节点的网络接口卡直接读取目的节点中的所需页；

5）目的节点的网络接口卡向源节点的网络接口卡返回远程取页应答消息，该消息中包含步骤4）中读取的页；

6）源节点的网络接口卡接收所述远程取页应答消息；

7）源节点的网络接口卡直接将所述页写到源节点的内存区域。

为了实施上述方法，本发明还提供一种用于虚拟共享存储系统的网络接口卡，包括PCI接口控制器通过本地总线分别与消息传递操作逻辑、控制寄存器逻辑、发送FIFO和接收FIFO电连接；所述发送FIFO的一端通过发送逻辑与发送驱动电连接；所述接收FIFO的一端通过接收逻辑和接收驱动电连接；其特征在于，还包括：

帧头分析逻辑，用于分析远程取页请求消息和远程取页应答消息的帧头；所述帧头分析逻辑一端通过本地总线与PCI接口控制器电连接，另一端与接收FIFO电连接；

RDMA操作逻辑，用于控制对内存读写操作，以便将页面从内存中直接读取或者写入；所述RDMA（“RDMA”是远程内存直接访问Remote Direct Memory Access英文缩写）操作逻辑通过本地总线与PCI接口控制器电连接，并分别与所述帧头分析逻辑和虚实地址转换逻辑电连接；

虚实地址转换逻辑，用于虚拟地址和物理地址的转换；所述虚实地址转换逻辑通过本地总线与PCI接口控制器电连接。

本发明具有如下优点：

1）在本发明中，通过将网络接口卡上的存储器区域映射到用户进程的虚地址空间，使得用户进程和网络接口卡可以双向直接访问，可以在不打断远程节点CPU的当前工作的情况下实现取页操作。

2）本发明用网络接口卡上的硬件处理大部分的协议开销，减少了用户进程的接收消息和解封消息等协议处理开销，提高了远程取页的效率。

3）一般的 DMA 和 RDMA 操作都是直接采用物理地址，而在本发明中采用虚拟地址，为实现远程取页和虚拟共享存储提供了非常大的便利。

附图说明

图 1 是一个示例性的虚拟共享系统；

图 2 是实现本发明的远程取页方法的网络接口卡；

图 3 是本发明的远程取页方法的方法流程框图；

图 4 是实现本发明远程取页方法的地址映射示意图。

具体实施方式

下面结合附图和具体实施方式对本发明作进一步详细描述。

如图 1 所示的一个示例性的虚拟共享系统，包括源节点 10 和目的节点 20，它们分别具有网络接口卡 30 和 30'，网络接口卡 30 和 30' 通过交换机 40 互联。在具体应用时，交换机 40 由 Altera 的 FPGA 芯片 EP20K200EFC484-2X 实现。

在本发明中，远程取页操作主要由节点上的网络接口卡完成。网络接口卡 30 和 30' 的逻辑结构如图 2 所示。在图 2 中，网络接口卡用 PCI 接口控制器 301 与主机的 PCI 总线连接，帧头分析逻辑 302、消息传递操作逻辑 303、RDMA 操作逻辑 304、虚实地址转换逻辑 305、控制寄存器逻辑 306、接收 FIFO307 和发送 FIFO308 通过本地总线连接，发送逻辑 310 和发送驱动 312 顺序与发送 FIFO308 连接，接收逻辑 309 和接收驱动 311 顺序与接收 FIFO307 连接。其中，帧头分析逻辑 302 根据帧的类型而采取不同的动作；消息传递操作逻辑 303 负责普通的消息收发操作；RDMA 操作逻辑 304 控制内存读写操作；虚实地址转换逻辑 305 用于虚拟地址和物理地址的转换，包括 TLB（Translation Lookaside Buffer）匹配逻辑和一个虚实地址对照表；控制寄存器逻辑 306 的读写操作。在具体应用时，图 2 所示的网络接口卡的逻辑功能由 Altera 的 FPGA 芯片 EP1K100FC484-1 实现。

图 3 是本发明的远程取页方法的流程图。如下所述：

在步骤 101 中，为了能够让用户进程直接访问网络接口卡上的存储器，需要在通信协议中通过地址映射的方法将网络接口卡（NIC）上的存储器区域映射到用户进程（User）的虚地址空间。这里的地址映射如图 4 所示，通过两次映射完成：

首先，NIC 寄存器区域位于 PCI 空间，在基于 x86 体系结构的 PC 机上，PCI 空间与主机内存物理地址空间是一致的，这一阶段的映射不需要通信协议做任何干预；

第二，在本发明中，为了使用户进程能直接访问 NIC 上的这一寄存器区域，在

通信协议中将NIC寄存器在物理地址空间的这部分区域再映射到用户进程的虚地址空间，这样用户进程就可以直接访问网络接口了。

在步骤102中，源节点的用户进程直接访问其网络接口卡，产生并向目的节点的网络接口卡发送远程取页请求消息。具体包括如下子步骤：

步骤 102a)：源节点上的用户进程通过通信库把要读取的页在目的节点上虚拟地址、数据长度（或者说页的大小）、该页返回源节点的虚拟地址填入一个可用的描述符；

步骤 102b)：由通信线程把已经填写完毕的描述符写到网络接口卡的寄存器上（发送一次写一个描述符）；

步骤 102c)：网络接口卡根据用户进程所写的内容来组帧，并在帧头置一类型标志，以表明该帧为远程取页请求消息，并将此帧发送至目的节点的网络接口卡。如步骤102a）所述，该远程取页请求消息的帧体内还包括：要读取的页在目的节点上虚拟地址域、数据长度域、该页返回源节点的虚拟地址域。该步骤由图2所示的网络接口卡中的消息传递操作逻辑303实现。

在步骤103中，目的节点的网络接口卡接收来自源节点的远程取页请求消息。目的节点从其网络接口卡的接收FIFO缓冲区中取出帧头数据并进行分析，确定是远程取页读操作，进行帧头转换——转换为返回的帧头。该步骤由图2所示的网络接口卡中的帧头分析逻辑实现。

在步骤104中，目的节点的网络接口卡直接读取目的节点中的所需页。具体包括如下子步骤：

步骤104a)：网络接口卡申请网络接口本地总线；

步骤104b)：网络接口卡申请进行虚实地址转换，将虚拟地址转换为物理地址。在这里，虚实地址转换由图2所示的网络接口卡上的虚实地址转换逻辑完成。该虚实地址转换逻辑包括一个虚实地址转换缓冲器（TLB，Translation Lookaside Buffer)，其中存储由虚拟地址和物理地址的匹配表项。在进行虚实地址转换时，在网络接口卡上的TLB表项中查找与虚拟地址匹配的项，若找到则直接返回对应的物理地址，若没有找到则产生中断，由节点主机的网卡驱动程序接收中断后根据中断类型进行虚实地址转换，将获得的物理地址写入网络接口卡，虚实地址转换逻辑接收到数据后对TLB表项进行更新同时返回物理地址。

步骤 104c)：网络接口卡将返回帧头写入网络接口发送 FIFO 缓冲区；

步骤 104d)：将 TLB 转换得到的物理地址指定的内存区域以 DMA 方式读到发送 FIFO 缓冲区。该步骤由图 2 所示的网络接口卡上的 RDMA 操作逻辑完成。

在步骤 105 中，目的节点的网络接口卡向源节点的网络接口卡返回远程取页应答消息，该消息中包含已经读取的页。并在远程取页应答消息的帧头中设置一类型标志，以表明该帧为远程取页应答消息。

在步骤 106 中，源节点的网络接口卡接收远程取页应答消息。源节点从其网络接口卡的接收 FIFO 缓冲区中取出帧头数据并进行分析，根据类型标志，确定是远程取页的读返回操作。

在步骤 107 中，源节点的网络接口卡直接将所述页写到源节点的内存区域。具体包括如下子步骤：

步骤 107a)：网络接口卡申请进行虚实地址转换，该步骤的具体操作与步骤 104b）基本相同；由于网络接口卡只能根据 PCI 总线地址进行工作，而 PCI 空间与物理地址空间又是一致的，所以只要知道需要访问的进程空间所对应的物理地址，网络接口卡就可以通过该地址访问对应的内存区域；

步骤 107b)：网络接口卡申请获得本地总线；

步骤 107c)：将接收 FIFO 缓冲区中的数据送到 TLB 转换得到的物理地址所指定的内存区域。为了实现这一点，源节点用户进程的虚页面锁定(pin)在其物理内存中，以防止网络接口在访问这一区域时，操作系统将对应页面交换出去。并且，如图 4 所示，对于涉及 RDMA 的页面（即图 4 中的 RDMA 区域），建立用户进程虚地址和物理地址的映射，再将该 RDMA 区域在物理地址空间的区域映射到 PCI 空间，这样一来，网络接口通过该物理地址就可以可靠地访问用户空间；

步骤 107d)：网络接口卡向源节点主机内存中对应的描述符的完成标志置位。

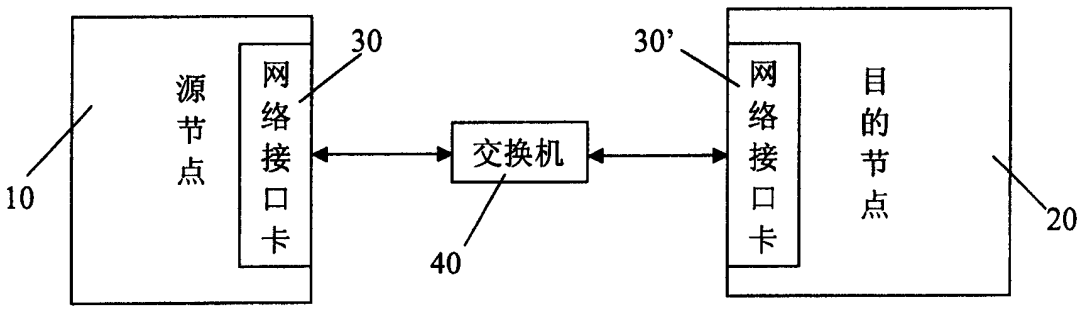

源节点
网络接口卡
30
交换机
40
网络接口卡
30'
目的节点
10
20

图 1

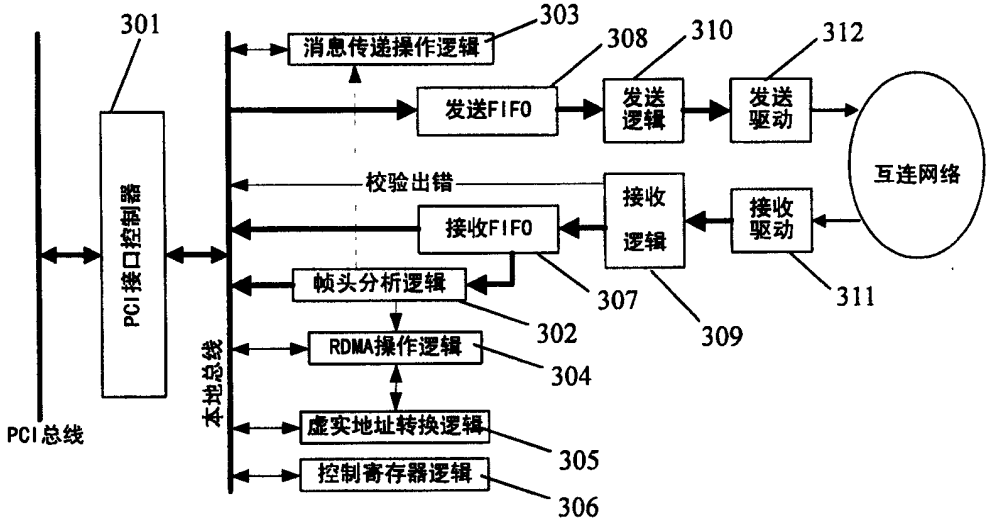

301
303
消息传递操作逻辑
308
310
312
发送FIFO
发送逻辑
发送驱动
互连网络
校验出错
接收逻辑
接收驱动
接收FIFO
帧头分析逻辑
307
309
311
302
RDMA操作逻辑
304
PCI接口控制器
本地总线
PCI总线
虚实地址转换逻辑
305
控制寄存器逻辑
306

图 2

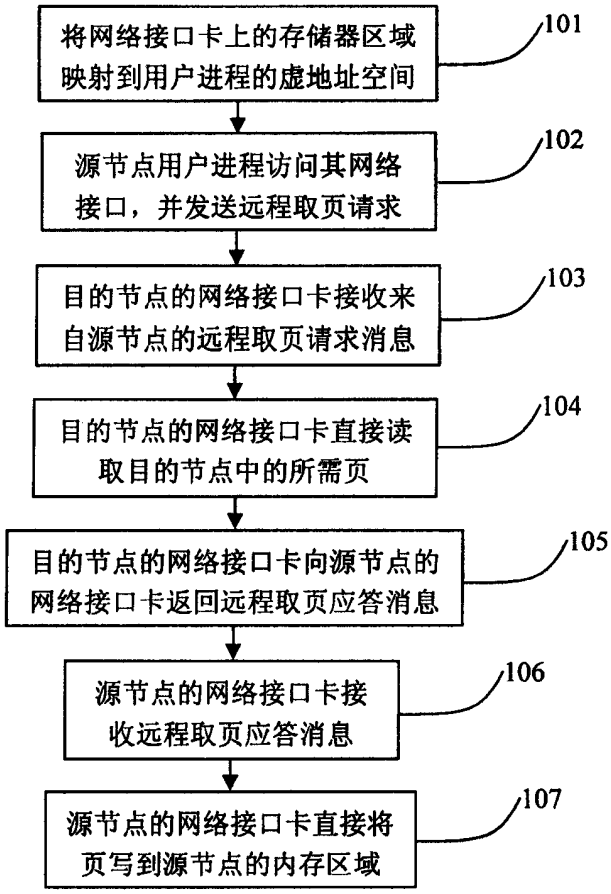

将网络接口卡上的存储器区域映射到用户进程的虚地址空间
101
源节点用户进程访问其网络接口，并发送远程取页请求
102
目的节点的网络接口卡接收来自源节点的远程取页请求消息
103
目的节点的网络接口卡直接读取目的节点中的所需页
104
目的节点的网络接口卡向源节点的网络接口卡返回远程取页应答消息
105
源节点的网络接口卡接收远程取页应答消息
106
源节点的网络接口卡直接将页写到源节点的内存区域
107

图 3

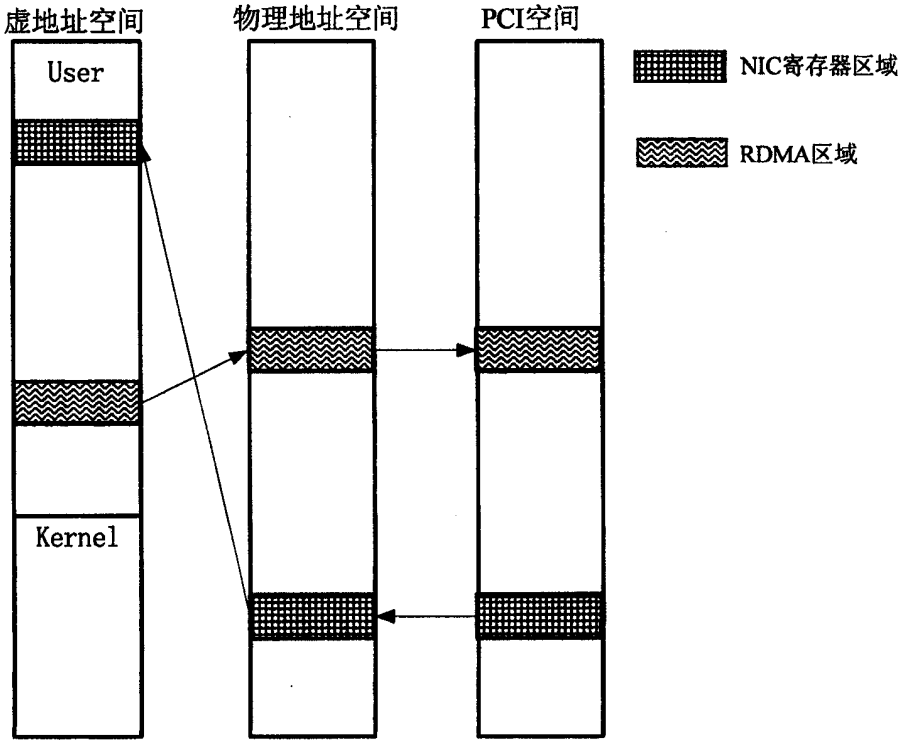

虚地址空间
物理地址空间
PCI空间
User
Kernel
NIC寄存器区域
RDMA区域

图 4