(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106933775 B
(45)授权公告日 2021.08.20

(21)申请号 201710044364.3
(22)申请日 2013.10.29
(65)同一申请的已公布的文献号
申请公布号 CN 106933775 A
(43)申请公布日 2017.07.07
(62)分案原申请数据
201380002065.3 2013.10.29
(73)专利权人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 何剑 施广宇 倪小珂 诺伯特·埃吉 李显才 刘毓 刘华伟
(51)Int.Cl.
G06F 15/173(2006.01)
H04L 29/08(2006.01)
(56)对比文件
CN 102439983 A,2012.05.02
CN 103946828 A,2014.07.23
CN 102439983 A,2012.05.02
CN 101751371 A,2010.06.23
CN 101819543 A,2010.09.01
审查员 吴瑶裔

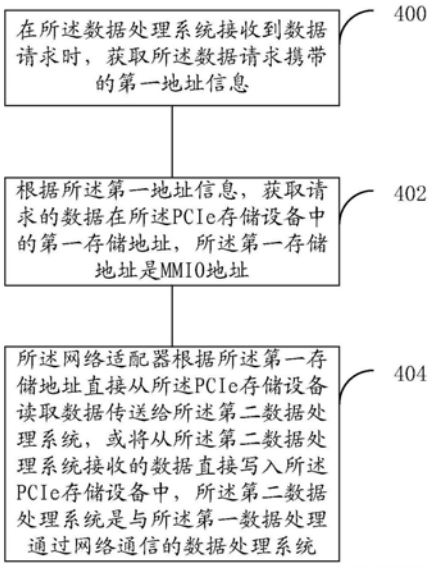
权利要求书3页 说明书15页 附图4页

(54)发明名称

数据处理系统和数据处理的方法

(57)摘要

本发明实施例提供了一种数据处理系统和数据处理方法。通过获取数据请求的MMIO地址，该MMIO地址能够直接获取PCIe存储设备中存储的数据，网络适配器根据该MMIO地址，能够直接从所述数据处理系统的PCIe存储设备中读取数据并传送给第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中。使得所述处理系统能够实现在网络通信的两个数据处理系统间传递数据时，数据的传输从PCIe存储设备直接到网络适配器之间传递，不需要经过内存。降低了在两个数据处理系统件传递数据时对内存、CPU等资源的占用率，并提高了数据传输的效率。



CN 106933775 B

1.一种数据处理系统,包括中央处理器CPU、内存、快捷外围部件互连标准PCIe控制器、网络适配器和至少一个PCIe存储设备,其特征在于,还包括:

管理单元,用于在所述数据处理系统接收到数据请求时,根据所述数据请求中携带的第一地址信息,获取请求的数据在所述PCIe存储设备中的第一存储地址,所述第一存储地址是内存映射输入输出MMIO地址;

所述网络适配器,根据所述第一存储地址,从所述PCIe存储设备直接读取数据传送给第二数据处理系统,或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中,所述第二数据处理系统是与所述数据处理系统通过网络通信的数据处理系统;

所述PCIe存储设备包括地址转换单元,用于根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址;所述第二存储地址是逻辑地址,所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

2.根据权利要求1所述的数据处理系统,其特征在于,

所述地址转换单元还用于配置第一BAR地址寄存器,所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系,所述第二存储地址是线性连续的存储地址。

3.根据权利要求1所述的数据处理系统,其特征在于,

所述地址转换单元还用于配置第二BAR地址寄存器,所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系,所述第二存储地址是非线性连续的存储地址,所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

4.根据权利要求1-3任意一项所述的数据处理系统,其特征在于,

所述CPU为每个所述PCIe存储设备分配一个唯一标识,所述唯一标识用于标识每个所述PCIe存储设备。

5.根据权利要求4所述的数据处理系统,其特征在于,

所述第一地址信息包括所述PCIe存储设备的唯一标识和逻辑区块地址LBA。

6.根据权利要求4所述的数据处理系统,其特征在于,

所述管理单元还用于建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。

7.根据权利要求4所述的数据处理系统,其特征在于,

所述唯一标识包括供应商识别码Vender ID、设备识别码Device ID或硬盘序列号中的至少一个;或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

8.根据权利要求1所述的数据处理系统,其特征在于,所述CPU将获取到的所述第一存储地址注册在所述网络适配器中。

9.根据权利要求1所述的数据处理系统,其特征在于,

所述数据处理系统还包括发送单元,所述发送单元用于将所述管理单元获取的所述第一存储地址发送给所述第二数据处理系统。

10.根据权利要求1所述的数据处理系统,其特征在于,所述PCIe控制器获取所述网络适配器发出的携带所述第一存储地址的数据请求,所述地址转换单元获取第二存储地址的

数据并将获取到的数据返回给所述网络适配器，或将网络适配器发送的数据写入所述第二存储地址。

11.根据权利要求6所述的数据处理系统，其特征在于，所述管理单元还包括全局基地址获取单元和全局存储地址获取单元：

所述全局基地址获取单元，用于在所述数据处理系统接收到第二数据请求时，根据所述第二数据请求中携带的PCIe存储设备的唯一标识，获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，所述第二数据请求是向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据的请求；

所述全局存储地址获取单元，用于根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的LBA地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

12.根据权利要求11所述的数据处理系统，其特征在于，所述第二数据处理系统的PCIe控制器中的数据转换单元，根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址，所述物理地址是能够直接读取数据的线性连续地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

13.一种数据处理方法，所述方法应用于包括中央处理器CPU、内存、快捷外围部件互连标准PCIe控制器、网络适配器和至少一个PCIe存储设备的数据处理系统中，其特征在于，所述方法包括：

在所述数据处理系统接收到第一数据请求时，获取所述第一数据请求携带的第一地址信息；

根据所述第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是内存映射输入输出MMIO地址；

所述网络适配器根据所述第一存储地址，直接从所述PCIe存储设备读取数据传送给第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中，所述第二数据处理系统是与所述数据处理系统通过网络通信的数据处理系统；所述PCIe存储设备根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址；所述第二存储地址是逻辑地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

14.根据权利要求13所述的数据处理方法，其特征在于，所述方法还包括：

所述PCIe控制器配置第一BAR地址寄存器，所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系，所述第二存储地址是线性连续的存储地址。

15.根据权利要求13所述的数据处理方法，其特征在于，所述方法还包括：

所述PCIe控制器配置第二BAR地址寄存器，所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系，所述第二存储地址是非线性连续的存储地址，所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

16.根据权利要求13-15任意一项所述的数据处理方法，其特征在于，所述CPU为每个所述PCIe存储设备分配一个唯一标识，所述唯一标识用于标识每个所述PCIe存储设备。

17.根据权利要求16所述的数据处理方法，其特征在于，

所述第一地址信息包括所述PCIe存储设备的唯一标识和逻辑区块地址LBA。

18.根据权利要求16所述的数据处理方法,其特征在于,所述方法还包括:

建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。

19.根据权利要求16所述的数据处理方法,其特征在于,所述唯一标识包括供应商识别码Vender ID、设备识别码Device ID或硬盘序列号中的至少一个;或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

20.根据权利要求17所述的数据处理方法,其特征在于,所述获取请求的数据在所述存储设备中的第一存储地址包括:

根据所述第一数据请求中携带的PCIe存储设备的唯一标识,获取请求的数据的BAR中的基地址;

根据所述BAR中的基地址以及所述第一数据请求携带的LBA地址,获取请求的数据在所述PCIe存储设备中的第一存储地址,所述第一存储地址是MMIO地址。

21.根据权利要求13所述的数据处理方法,其特征在于,所述CPU将获取到的所述第一存储地址注册在所述网络适配器中。

22.根据权利要求13所述的数据处理方法,其特征在于,所述方法还包括:

所述数据处理系统将获取到的所述第一存储地址发送给所述第二数据处理系统。

23.根据权利要求13所述的数据处理方法,其特征在于,

所述PCIe控制器获取所述网络适配器发出的携带所述第一存储地址的数据请求,获取第二存储地址的数据并将获取到的数据返回给所述网络适配器,或将网络适配器发送的数据写入所述第二存储地址。

24.根据权利要求17所述的数据处理方法,其特征在于,所述方法还包括:

在所述数据处理系统接收到第二数据请求时,根据所述第二数据请求中携带的PCIe存储设备的唯一标识,获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址,所述第二数据请求用于向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据的请求;

根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的LBA地址,获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

25.根据权利要求24所述的数据处理方法,其特征在于,所述第二数据处理系统中的PCIe控制器,根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址,获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址,所述物理地址是能够直接读取数据的线性连续地址,所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

数据处理系统和数据处理的方法

技术领域

[0001] 本发明涉及信息技术领域，尤其涉及一种不同数据处理系统之间数据传输的设备、方法和系统。

背景技术

[0002] 在大数据的潮流下，通常会采用多副本的方式来保障数据的可靠性。而采用多副本方式，往往会导致节点间的数据迁移操作非常多。

[0003] RDMA(Remote Direct Memory Access，远程直接数据存取)技术是一种实现网络上两个节点间数据读取的技术。RDMA通过网络把数据直接传入计算机的内存，将数据从本地节点快速移动到远程节点内存中，而不对操作系统造成任何影响。

[0004] 在网络上传输的RDMA信息包含目标虚拟地址、内存钥匙和数据本身.请求完成既可以完全在用户空间中处理(通过轮询用户级完成排列)，或者在应用一直睡眠到请求完成时的情况下通过内核内存处理。RDMA操作使应用可以从一个远程应用的内存中读数据或向这个内存写数据。目标主机的网络适配器确认内存钥匙，直接将数据写人应用缓存中。

[0005] RDMA要求传输的数据必须要通过内存才能进行两个相互通信的服务器之间的数据传输。如果不通过内存的话，则无法进行数据传输。导致数据传输的时延和内存占用率较高。

发明内容

[0006] 本发明实施例提供一种系统和数据处理方法，以提高在两个数据处理系统之间传输数据时的效率和设备利用率。

[0007] 本发明实施例提供了一种数据处理系统，包括中央处理器CPU、内存、快捷外围部件互连标准PCIe控制器、网络适配器和至少一个PCIe存储设备，其特征在于，还包括：

[0008] 管理单元，用于在所述数据处理系统接收数据请求时，根据所述数据请求中携带的第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是内存映射输入输出MMIO地址；

[0009] 所述网络适配器，根据所述第一存储地址从所述PCIe存储设备直接读取数据传送给第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中，所述第二数据处理系统是与所述第一数据处理系统通过网络通信的数据处理系统；所述PCIe存储设备包括地址转换单元，用于根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址；所述第二存储地址是逻辑地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0010] 可选的，所述第二存储地址是物理地址，所述物理地址是能够直接读取数据的线性连续地址。

[0011] 可选的，所述地址转换单元还用于配置第一基地址寄存器BAR，所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系，所述第二存储地址是

线性连续的存储地址。

[0012] 可选的，所述地址转换单元还用于配置第二BAR地址寄存器，所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系，所述第二存储地址是非线性连续的存储地址，所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

[0013] 可选的，所述CPU为每个所述PCIe存储设备分配一个唯一标识，所述唯一标识用于标识每个所述PCIe存储设备。

[0014] 可选的，所述第一地址信息包括所述PCIe存储设备的唯一标识和逻辑区块地址LBA。

[0015] 可选的，所述管理单元还用于建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。

[0016] 所述唯一标识包括供应商识别码Vender ID、设备识别码Device ID或硬盘序列号中的至少一个；或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

[0017] 可选的，所述管理单元包括基地址获取单元和存储地址获取单元：

[0018] 所述基地址获取单元，用于在所述数据处理系统接收到与其通过网络通信的第二数据处理系统的第一数据请求时，根据所述第一数据请求中携带的PCIe存储设备的唯一标识，获取请求的数据的BAR中的基地址；

[0019] 存储地址获取单元，用于根据所述BAR中的基地址以及所述第一数据请求携带的LBA地址，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址。

[0020] 可选的，所述CPU将获取到的所述第一存储地址注册在所述网络适配器中。

[0021] 可选的，所述数据处理系统还包括发送单元，所述发送单元用于将所述管理单元获取的所述第一存储地址发送给所述第二数据处理系统。

[0022] 可选的，所述PCIe控制器获取所述网络适配器发出的携带所述第一存储地址的数据请求，所述地址转换单元获取第二存储地址的数据，并将获取到的数据返回给所述网络适配器，或将网络适配器发送的数据写入所述第二存储地址。

[0023] 可选的，所述管理单元还包括全局基地址获取单元和全局存储地址获取单元：

[0024] 所述全局基地址获取单元，用于在所述数据处理系统接收到第二数据请求时，根据所述第二数据请求中携带的PCIe存储设备的唯一标识，获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，所述第二数据请求是向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据的请求；

[0025] 所述全局存储地址获取单元，用于根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的所述LBA地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

[0026] 可选的，所述第二数据处理系统的PCIe控制器中的数据转换单元，根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址，所述物理地址是能够直接读取数据的线性连续地址，所述逻辑地址是对非线性连续的物理地址

进行线性排序后的地址。

[0027] 本发明实施例还提供了一种

[0028] 数据处理方法，所述方法应用于包括中央处理器CPU、内存、快捷外围部件互连标准PCIe控制器、网络适配器和至少一个PCIe存储设备的数据处理系统中，所述方法包括：

[0029] 在所述数据处理系统接收到数据请求时，获取所述数据请求携带的第一地址信息；

[0030] 根据所述第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址；

[0031] 所述网络适配器根据所述第一存储地址直接从所述PCIe存储设备读取数据传送给所述第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中，所述第二数据处理系统是与所述数据处理系统通过网络通信的数据处理系统；所述PCIe存储设备根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址；所述第二存储地址是逻辑地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0032] 可选的，所述第二存储地址是物理地址，所述物理地址是能够直接读取数据的线性连续地址。

[0033] 可选的，所述方法还包括：

[0034] 所述PCIe控制器配置第一BAR地址寄存器，所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系，所述第二存储地址是线性连续的存储地址。

[0035] 可选的，所述方法还包括：

[0036] 所述PCIe控制器配置第二BAR地址寄存器，所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系，所述第二存储地址是非线性连续的存储地址，所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

[0037] 可选的，所述CPU为每个所述PCIe存储设备分配一个唯一标识，所述唯一标识用于标识每个所述PCIe存储设备。

[0038] 可选的，所述第一地址信息包括所述PCIe存储设备的唯一标识和逻辑区块地址LBA。

[0039] 可选的，所述方法还包括：

[0040] 建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。

[0041] 可选的，所述唯一标识包括供应商识别码Vender ID、设备识别码Device ID或硬盘序列号中的至少一个；或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

[0042] 可选的，所述获取请求的数据在所述存储设备中的第一存储地址包括：

[0043] 根据所述第一数据请求中携带的PCIe存储设备的唯一标识，获取请求的数据的BAR中的基地址；

[0044] 根据所述BAR中的基地址以及所述第一数据请求中的LBA地址，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址。

[0045] 可选的，所述CPU将获取到的所述第一存储地址注册在所述网络适配器中。

[0046] 可选的，所述方法还包括：

[0047] 所述数据处理系统将获取到的所述第一存储地址发送给所述第二数据处理系统。

[0048] 可选的，所述PCIe控制器获取所述网络适配器发出的携带所述第一存储地址的数据请求，并获取第二存储地址的数据，将获取到的数据返回给所述网络适配器，或将网络适配器发送的数据写入所述第二存储地址。

[0049] 可选的，所述方法还包括：

[0050] 在所述数据处理系统接收到第二数据请求时，根据所述第二数据请求中携带的PCIe存储设备的唯一标识，获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，所述第二数据请求用于向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据的请求；

[0051] 根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的所述LBA地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

[0052] 可选的，所述第二数据处理系统中的PCIe控制器，根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址，所述物理地址是能够直接读取数据的线性连续地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0053] 本发明实施例提供的数据处理系统和数据处理方法，通过获取数据请求的MMIO地址，该MMIO地址能够直接获取PCIe存储设备中存储的数据，网络适配器根据该MMIO地址，能够直接从所述数据处理系统的PCIe存储设备中读取数据并传送给第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中。使得所述处理系统能够实现在网络通信的两个数据处理系统间传递数据时，数据的传输从PCIe存储设备直接到网络适配器之间传递，不需要经过内存。降低了在两个数据处理系统件传递数据时对内存、CPU等资源的占用率，并提高了数据传输的效率。

附图说明

[0054] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0055] 图1为现有技术中一种将远程节点存储设备中的数据搬到本地节点存储设备中的流程示意图；

[0056] 图2为本发明实施例一种数据处理系统的结构示意图；

[0057] 图3为本发明实施例数据处理系统的一种具体实现结构示意图；

[0058] 图4为本发明实施例一种数据处理方法的流程示意图；

[0059] 图5为本发明实施例数据处理系统一种实现方式的基本硬件结构示意图；

[0060] 图6为本发明实施例管理单元存储的PCIe存储设备唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系的示例图；

[0061] 图7为本发明实施例一中配置PCIe存储设备内部地址到CPU的MMIO地址之间映射的结构关系示意图；

[0062] 图8为本发明实施例两个数据处理系统之间传递数据的数据流向示意图。

具体实施方式

[0063] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明的一部分实施例，而不是全部实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都应属于本发明保护的范围。

[0064] 图1为现有技术中一种将远程节点存储设备中的数据搬到本地节点存储设备中的流程示意图。其中节点可以是服务器等实现计算或存储功能的设备。当节点Node 1需要通过网络从节点Node2读取存储在Node 2的存储设备中的数据时，其实现过程如下：

[0065] 步骤1，Node 1的CPU发起请求读取数据的远程连接；

[0066] 步骤2，Node 1中的网络适配器将该请求报文发送至指定节点(即Node 2)的网络适配器；

[0067] 步骤3，Node 2的网络适配器将该请求报文转发给Node 2的CPU；

[0068] 步骤4，Node 2的CPU解析请求报文后，向其PCIe控制器发起数据请求；

[0069] 步骤5，PCIe控制器将请求的数据通过DMA的方式读至内存；

[0070] 步骤6，Node 2的CPU将读到内存的请求数据发送到其网络适配器；

[0071] 步骤7，Node 2的网络适配器将请求的数据通过网络发送至Node 1的网络适配；

[0072] 步骤8，Node 1的CPU从其网络适配器读取请求的数据后将该数据缓冲至内存；

[0073] 步骤9，Node1的CPU将缓存在内存中的数据发送给PCIe控制器以请求将数据写入存储设备；

[0074] 步骤10，Node 1的PCIe控制器将接收到的数据写入存储设备中。

[0075] 从上述现有技术中数据远程读取的过程可以看出，参与读取的节点的CPU参与到了数据读取与写入，每个节点都需要申请一段内存空间来存储CPU读取或是待写入的数据。这样在整个数据搬移过程中，由于数据搬移的次数较多必然导致时延迟的增大，同时CPU与内存的占用率也居高不下。

[0076] 为解决现有技术中远程数据迁移时时延大，CPU和内存占用率高的问题，本发明实施例提供了一种数据处理系统，以解决现有技术中数据处理系统间远程读写数据时，因占用内存和CPU资源带来的资源消耗和时延的问题。

[0077] 为解决现有技术中在跨节点数据传输时占用CPU、内存资源多，传输延迟大的问题，本发明实施例提供了一种数据处理系统，结合网络技术与PCIe存储设备的MMIO地址映射技术，使用节点间数据远程直接访问技术来进行数据的直接拷贝，拷贝过程中无须CPU参与数据的搬移，只需要CPU进行控制，同时也不需要将数据预先搬移到内存处理，这样就降低了CPU与内存的使用率；同时减少了数据在CPU和内存之间的迁移过程，降低了数据处理的时延，提高了数据传输的效率。

[0078] 参考图2，图2为本发明实施例一种数据处理系统200的结构示意图。该数据处理系统200包括中央处理器CPU 202、内存206、快捷外围部件互连标准PCIe控制器203、网络适配

器205和至少一个PCIe存储设备204,其特征在于,还包括:

[0079] 管理单元201,用于在所述数据处理系统接收数据请求时,根据所述数据请求中携带的第一地址信息,获取请求的数据在所述存储设备中的第一存储地址,所述第一存储地址是MMIO(Memory mapping I/O,内存映射输入输出)地址;

[0080] 所述网络适配器205,根据所述第一存储地址从所述PCIe存储设备204直接读取数据传送给第二数据处理系统,或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备204中,所述第二数据处理系统是与所述第一数据处理系统通过网络通信的数据处理系统。

[0081] 通过上述数据处理系统中的管理单元201获取数据请求的MMIO地址,该MMIO地址能够直接获取PCIe存储设备中存储的数据,网络适配器205根据该MMIO地址,能够直接从所述数据处理系统200的PCIe存储设备204中读取数据并传送给第二数据处理系统,或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备204中;使得所述处理系统能够实现在网络通信的两个数据处理系统间传递数据时,数据的传输从PCIe存储设备直接到网络适配器之间传递,不需要经过内存。降低了在两个数据处理系统件传递数据时对内存、CPU等资源的占用率,并提高了数据传输的效率。

[0082] 参考图3,图3为本发明实施例数据处理系统200的一种具体实现结构示意图。如图3所示,所述PCIe存储设备203还包括地址转换单元2031,用于根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址。所述第二存储地址可以是物理地址或逻辑地址,所述物理地址是能够直接读取数据的线性连续地址,所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0083] 通过所述地址转换单元2031将第一存储地址,即MMIO地址转换为第二存储地址,所述第二存储地址是所述PCIe存储设备203的可访问介质的物理地址,能够使得所述PCIe存储设备在收到网络适配器205的数据请求时,根据数据请求中携带的MMIO地址,获取与该MMIO地址对应的可访问介质的物理地址,并通过该物理地址读取数据。在PCIe存储设备中可访问介质地址不是用MMIO地址表示时,能够使得网络适配器获取所请求数据的可访问介质地址,实现了数据的直接读取和写入。

[0084] 作为一种可选的实现方式,所述地址转换单元2031还用于配置第一基地址寄存器BAR(base address register,基地址寄存器),所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系,所述第二存储地址是线性连续的存储地址。如果所述第二存储地址是非线性连续的存储地址,所述地址转换单元用于配置第二BAR地址寄存器,所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系,所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

[0085] 通过上述地址转换单元2031配置BAR地址寄存器,将PCIe存储设备中的线性连续的物理地址与MMIO地址一一对应,将PCIe存储设备中的非线性连续的物理地址经过线性排序后的逻辑地址与MMIO地址一一对应,实现MMIO地址与PCIe存储设备可访问介质地址的映射,使得网络适配器根据MMIO地址,通过PCIe控制器映射到与MMIO地址一一对应的PCIe可访问存储介质地址,实现了数据的直接读取和写入。

[0086] 作为一种可选的实现方式,所述CPU 202为每个所述PCIe存储设备分配一个唯一

标识，所述唯一标识用于标识每个所述存储设备。相应的，所述管理单元201还用于建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。通过PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系，管理单元201能够根据数据处理系统200接收到的数据请求消息中包含的PCIe存储设备的唯一标识，获取与该唯一标识对应的PCIe存储设备的BAR中的基地址。由于所述数据处理系统接收的数据请求中的第一地址信息包括PCIe存储设备的唯一标识和LBA(Logical Block Address，逻辑区块地址)地址，管理单元201通过获取到的PCIe存储设备的BAR中的基地址与LBA地址，能够获取请求数据的MMIO地址。

[0087] 可选的，所述唯一标识包括Vender ID、Device ID或硬盘序列号中的至少一个；或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

[0088] 作为一种可选的实现方式，如图3所示，所述管理单元201包括基地址获取单元2011和存储地址获取单元2012：

[0089] 所述基地址获取单元2011，用于在所述数据处理系统200接收到与其通过网络通信的第二数据处理系统的第一数据请求时，根据所述第一数据请求中携带的PCIe存储设备204的唯一标识，获取请求的数据的BAR中的基地址；

[0090] 存储地址获取单元2012，用于根据所述BAR中的基地址以及所述第一数据请求中的LBA地址，获取请求的数据在所述存储设备中的第一存储地址，所述第一存储地址是MMIO地址。

[0091] 作为一种可选的实现方式，所述CPU 202将获取到的所述第一存储地址注册在所述网络适配器中。将所述第一存储地址注册在所述网络适配器中，网络适配器205在接收到第二数据处理系统携带有第一存储地址的数据请求时，能够将所述第一存储地址在PCIe总线上发布，PCIe控制器203接收到网络适配器205发送的携带第一存储地址的数据请求时，获取该请求，并将请求的数据发送给网络适配器205，或将网络适配器接收到的第二数据处理系统发送的数据写入PCIe存储设备中所述第一存储地址对应的位置。

[0092] 作为一种可选的实现方式，所述数据处理系统200还包括发送单元207，所述发送单元207用于将所述管理单元200获取的所述第一存储地址发送给所述第二数据处理系统。

[0093] 可选的，所述PCIe控制器203获取所述网络适配器205发出的携带所述第一存储地址的数据请求，所述地址转换单元2031获取第二存储地址的数据，并将获取到的数据返回给所述网络适配器，或将网络适配器发送的数据写入所述第二存储地址。

[0094] 本发明实施例中，所述数据处理系统200与其它的数据处理系统，例如第二数据处理系统，通过网络通信，实现数据在不同数据处理系统之间的传输。所述网络包括但不限于以太网、支持多并发链接的转换线缆技术IB网络或FC(fiber channel，光纤信道)网络等。上述第二数据处理系统可以是实现本发明实施例所述方案的系统，也可以现有技术中的数据处理系统。当所述第二数据处理系统为实现本发明实施例所述方案的系统时，第二数据处理系统也能够实现网络适配器到PCIe存储设备的直接数据读取或写入。

[0095] 作为一种可选的实现方案，所述管理单元203还包括全局基地址获取单元2033和全局存储地址获取单元2034：

[0096] 所述全局基地址获取单元2033，用于在所述数据处理系统接收到第二数据请求

时，根据所述第二数据请求中携带的PCIe存储设备的唯一标识，获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，所述第二数据请求是向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据；

[0097] 所述全局存储地址获取单元2034，用于根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的所述LBA地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

[0098] 相应的，所述第二数据处理系统的PCIe控制器中的数据转换单元，根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址，所述物理地址是能够直接读取数据的线性连续地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0099] 上述实施例中，管理单元203还保存有第二数据处理系统中PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系，在数据处理系统200接收到向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据的第二数据请求时，能够获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，并进一步获取第二数据请求所请求数据的MMIO地址，从而实现两个数据处理系统数据的直接传输，不需要CPU和内存的参与，节省了CPU和内存的资源，并挺高了数据传输的效率。例如图8所示的第一数据处理系统与第二数据处理系统之间实现数据的传输，其中黑色虚线部分即为两个数据处理系统之间数据传输的轨迹和流向。

[0100] 本发明实施例的数据处理系统200还可以与多个数据处理系统通过通信网络连接，并进行数据的传输。所述数据处理系统200可以获取其它多个数据处理系统发送的PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系，以实现数据处理系统之间的直接传输。也可以向其它多个数据处理系统请求PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系并保存，以实现数据处理系统之间的直接传输。当然，在所述数据处理系统200获取其它数据处理系统PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系后，当其它数据处理系统PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系发生变化时，发生变化的数据处理系统可以将变化后的对应关系发送给数据处理系统200。

[0101] 参考图4，图4为本发明实施例一种数据处理方法的流程示意图。本发明实施例的数据处理方法应用于包括中央处理器CPU、内存、快捷外围部件互连标准PCIe控制器、网络适配器和至少一个存储设备的数据处理系统中，其特征在于，所述方法包括：

[0102] 步骤400：在所述数据处理系统接收到数据请求时，获取所述数据请求携带的第一地址信息；

[0103] 步骤402：根据所述第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址；

[0104] 步骤404：所述网络适配器根据所述第一存储地址直接从所述PCIe存储设备读取数据传送给所述第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中，所述第二数据处理系统是与所述第一数据处理通过网络通信的数据处理系统。

[0105] 上述方法实施例的步骤400和步骤402,可以由所述数据处理系统中的管理单元来实现,该管理单元可以是所述CPU中的一个模块或逻辑单元,也可以是所述数据处理系统中单独的硬件实体,本发明实施例不限定管理单元的具体实现方式。

[0106] 通过上述实施例中步骤402获取数据请求的MMIO地址,该MMIO地址能够直接获取PCIe存储设备中存储的数据,网络适配器根据该MMIO地址,能够直接从所述数据处理系统的PCIe存储设备中读取数据并传送给第二数据处理系统,或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中;使得所述处理系统能够实现在网络通信的两个数据处理系统间传递数据时,数据的传输从PCIe存储设备直接到网络适配器之间传递,不需要经过内存。降低了在两个数据处理系统件传递数据时对内存、CPU等资源的占用率,并提高了数据传输的效率。

[0107] 作为一种可选的实现方式,所述数据处理方法还包括,所述PCIe存储设备根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址。所述第二存储地址是物理地址或逻辑地址,所述物理地址是能够直接读取数据的线性连续地址,所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0108] 可选的,所述PCIe控制器配置第一BAR地址寄存器,所述第一BAR地址寄存器存储所述第一存储地址与所述第二存储地址之间的对应关系,所述第二存储地址是线性连续的存储地址。或者,所述PCIe控制器配置第二BAR地址寄存器,所述第二BAR地址寄存器存储所述第一存储地址与所述第二存储地址的虚拟地址之间的对应关系,所述第二存储地址是非线性连续的存储地址,所述第二存储地址的虚拟地址是所述第二存储地址经过线性排序后的地址。

[0109] 通过上述配置BAR地址寄存器,将PCIe存储设备中的线性连续的物理地址与MMIO地址一一对应,将PCIe存储设备中的非线性连续的物理地址经过线性排序后的逻辑地址与MMIO地址一一对应,实现MMIO地址与PCIe存储设备可访问介质地址的映射,使得网络适配器根据MMIO地址,通过PCIe控制器映射到与MMIO地址一一对应的PCIe可访问存储介质地址,实现了数据的直接读取和写入。

[0110] 作为一种可选的实现方式,所述数据处理方法还包括:所述CPU为每个所述PCIe存储设备分配一个唯一标识,所述唯一标识用于标识每个所述PCIe存储设备。根据所述唯一标识,建立所述唯一标识与所述PCIe存储设备的BAR中的基地址之间的对应关系。

[0111] 可选的,所述第一地址信息包括所述PCIe存储设备的唯一标识和逻辑区块地址LBA。

[0112] 可选的,所述唯一标识包括Vender ID、Device ID或硬盘序列号中的至少一个;或者所述唯一标识是对Vender ID、Device ID或硬盘序列号中的至少一个进行哈希处理后得到的标识。

[0113] 通过PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系,能够根据所述数据处理系统接收到的数据请求消息中包含的PCIe存储设备的唯一标识,获取与该唯一标识对应的PCIe存储设备的BAR中的基地址。由于所述数据处理系统接收的数据请求中的第一地址信息包括PCIe存储设备的唯一标识和LBA地址,通过获取到的PCIe存储设备的BAR中的基地址与LBA地址,能够获取请求数据的MMIO地址。

[0114] 作为一种可选的实现方式,所述获取请求的数据在所述存储设备中的第一存储地

址包括：

[0115] 根据所述第一数据请求中携带的PCIe存储设备的唯一标识，获取请求的数据的BAR中的基地址；

[0116] 根据所述BAR中的基地址以及所述第一数据请求中的LBA地址，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址。

[0117] 可选的，所述CPU将获取到的所述第一存储地址注册在所述网络适配器中。所述数据处理系统将获取到的所述第一存储地址发送给所述第二数据处理系统。在所述数据处理系统的网络适配器接收到所述第二数据处理系统发送的携带有所述第一存储地址的数据请求时，所述网络适配器在PCIe总线上发布接收到的数据请求，该请求能够被所述数据处理系统中的PCIe控制器接收到。所述PCIe控制器获取所述网络适配器发出的携带所述第一存储地址的数据请求后，根据所述第一存储地址获取对应的第二存储地址，并从所述第二存储地址获取数据，将获取到的数据返回给所述网络适配器，或将网络适配器发送的数据写入所述第二存储地址。

[0118] 作为一种可选的实现方式，所述方法还包括：

[0119] 在所述数据处理系统接收到第二数据请求时，根据所述第二数据请求中携带的PCIe存储设备的唯一标识，获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，所述第二数据请求用于向所述第二数据处理系统发送数据或从所述第二数据处理系统读取数据；

[0120] 根据所述第二数据处理系统的BAR中的基地址以及所述第二数据请求中携带的所述LBA地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址。

[0121] 相应的，所述第二数据处理系统中的PCIe控制器，根据所述数据处理系统发送的所述第二数据请求所请求的数据在所述第二数据处理系统的MMIO地址，获取所述第二数据请求所请求的数据在所述第二数据处理系统的物理地址或逻辑地址，所述物理地址是能够直接读取数据的线性连续地址，所述逻辑地址是对非线性连续的物理地址进行线性排序后的地址。

[0122] 通过上述获取第二数据处理系统中PCIe存储设备的唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系，能够获取所述第二数据请求所请求的数据的在所述第二数据处理系统的BAR中的基地址，并进一步获取第二数据请求所请求数据的MMIO地址，从而实现两个数据处理系统数据的直接传输，不需要CPU和内存的参与，节省了CPU和内存的资源，并挺高了数据传输的效率。

[0123] 图5示意性的示出了根据本发明实施例数据处理系统一种实现方式的基本硬件结构示意图。该数据处理系统包括CPU、内存、PCIe控制器、PCIe总线、PCIe存储设备和网络适配器等基本硬件组成。其中，该数据处理系统中的网络适配器是支持网络直接访问技术的基于PCIe总线技术的网络适配器，所述网络直接访问技术包括但不限于RDMA(远程直接数据存取，Remote Direct Memory Access)技术等；所述网络适配器包括但不限于以太网卡、IB HCA(Infiniband Host Channel Adapter，支持多并发链接的转换线缆技术主机通道适配器)、iWarp HCA(internet wide area RDMA protocol Host Channel Adapter，支持互联网广域RDMA协议主机通道适配器)、Rapid IO HCA(Rapid IO Host Channel Adapter，支持快速读写的主机通道适配器)等；所述基于PCIe总线技术是指此网络适配器的上行总线

接口为PCIe。该数据处理系统还包括至少一个基于PCIe总线的PCIe存储设备，该PCIe存储设备包括但不限于内存、硬盘、SSD(Solid State Disk，固态硬盘)、Flash、NVRAM(Non-Volatile Random Access Memory，非易失性随机访问存储器)等。该数据处理系统的实现方式，包括但不限于服务器(机架式、刀塔式、机框式等)、存储设备或小型机等设备。

[0124] 本发明实施例的数据处理系统，在图5所示的硬件基本架构的基础上，增加一个管理单元，该管理单元用于在所述数据处理系统接收数据请求时，根据所述数据请求中携带的第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址。该管理单元可以是在CPU中实现，也可以通过单独的硬件实现，本发明实施例不限定管理单元在所述数据处理系统中的实现方式。

[0125] 具体的，所述管理单元通过建立的所述PCIe存储设备的BAR中的基地址与所述PCIe存储设备的唯一标识之间的对应关系，获取所述数据处理系统接收到的数据请求的数据在所述PCIe存储设备中的地址。

[0126] 所述PCIe存储设备的唯一标识，例如GUID(Globally Unique Identifier，全球唯一标识)，是所述数据处理系统为每个PCIe存储设备分配的唯一确定该PCIe存储设备的标识。该唯一标识可以是所述CPU分配的唯一标识，也可以是所述管理单元分配的唯一标识。该唯一标识可以由PCIe存储设备的Vender ID(vender identity，供应商识别码)、Device ID(device identity，设备识别码)以及PCIe存储设备下挂的存储设备的唯一标记ID(如硬盘序列号)等组成一个唯一的一组字符串，或者对所述字符串进行哈希后得到的标识。本发明实施例不限定该唯一标识的组成，只要能够唯一的标记此节点内部的存储设备信息即可。

[0127] 所述PCIe存储设备的BAR中的基地址是在所述数据处理系统启动时分配的，所述数据处理系统启动完成后，所述管理单元获取每个PCIe存储设备的BAR中的基地址。所述管理单元可以通过扫描所述数据处理系统中所有的PCIe存储设备，获取每个PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识。

[0128] 所述管理单元根据获取到的每个PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识，并记录每个PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系。

[0129] 所述数据处理系统在接收到数据请求时，该数据请求会携带请求的数据所在的PCIe存储设备的唯一标识和LBA(Logical Block Address，逻辑区块地址)地址。所述管理单元根据所述数据请求中携带的唯一标识，以及建立的每个PCIe存储设备的BAR中的基地址与该PCIe存储设备的唯一标识之间的对应关系，获取所请求数据所在PCIe存储设备的BAR中的基地址；并结合LBA地址获取所请求的数据的MMIO地址。所述管理单元可以通过下述方式获取所请求的数据的MMIO地址的起始地址和结束地址：

[0130] 起始地址＝映射的BAR中的基地址+(LBA×块大小)

[0131] 结束地址＝映射的BAR中的基地址+((LBA+块数量)×块大小)-1。

[0132] 所述管理单元还用于维护在PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系，在PCIe存储设备的BAR中的基地址发生变化时，刷新PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系。例如，PCIe存储设备的BAR中的基地址可能会因为所述数据处理系统的重新启动而改变，即所述数据处理系

统为每个PCIe存储设备分配的PCIe存储设备的BAR中的基地址可能与上一次不同。所述管理单元需要根据每个PCIe存储设备的唯一标识,刷新PCIe存储设备唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系。

[0133] 所述管理单元还可以获取其它数据处理系统中PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系。通过该其它数据处理系统中PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系,能够实现在对其它数据处理系统进行数据读写时,获取该其它数据处理系统中PCIe存储设备的存储地址。获取其它数据处理系统中PCIe存储设备的BAR中的基地址以及该PCIe存储设备的唯一标识之间的对应关系,可以由本数据处理系统主动向其它数据处理系统发起请求以获取,也可以接收其它数据处理系统主动发送后获取。本发明实施例不限定具体获取的方式。

[0134] 参考图6,图6为本发明实施例管理单元存储的PCIe存储设备唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系的示例图。该示例图中,IP地址用于标识所述数据处理系统在其所在的网络内的唯一地址,GUID为PCIe存储设备的唯一标识,映射空间起始地址是PCIe存储设备地址在CPU寻址空间所映射区域的起始地址;设备逻辑地址是指由存储设备内部的逻辑起始地址;空间长度指系统所映射的这段区域的整个长度。

[0135] 本发明实施例中,PCIe存储设备唯一标识与PCIe存储设备的BAR中的基地址之间的对应关系,可以由PCIe存储设备的驱动建立,也可以由所述管理单元建立。具体可以通过添加脚本的方式在所述数据处理系统启动后自动加载,也可以通过手动方式加载。当由PCIe存储设备的驱动建立时,首先由PCIe存储设备的驱动配置PCIe BAR地址寄存器,管理单元读取配置后的映射关系。当由管理单元建立时,通过配置PCIe存储设备的寄存器来实现。

[0136] 本发明实施例的PCIe存储设备,还包括地址转换单元,用于根据所述第一存储地址获取所述数据请求所请求的数据在所述PCIe存储设备的第二存储地址。

[0137] 所述地址转换单元可以在PCIe设备控制器中实现,建立PCIe存储设备内的介质可访问地址与所述MMIO地址之间的对应关系,在接收到对所述MMIO地址进行的数据读写请求时,能够根据所述MMIO地址获取与该MMIO地址对应PCIe存储设备内的介质可访问地址,例如PCIe存储设备的存储地址,以进行数据的读写。

[0138] 以PCIe存储设备为PCIe NVRAM为例,根据其遵循的PCIe协议规范,将该PCIe NVRAM内部的可访问介质的地址空间直接映射到CPU的MMIO地址空间,使得所有对该PCIe NVRAM的读写请求等同于对MMIO地址的请求。MMIO地址空间对应NVRAM所有可访问的空间,即建立了PCIe存储设备内的介质可访问地址与所述MMIO地址之间的对应关系,通过该MMIO地址空间的访问,可以直接访问其对应NVRAM空间。将该PCIe NVRAM内部的可访问介质的地址空间直接映射到CPU的MMIO地址空间,可以通过配置BAR地址寄存器的方式实现。

[0139] 如图7所示,图7为本发明实施例一中配置PCIe存储设备内部地址到CPU的MMIO地址之间映射的结构关系示意图。图中PCIe控制器中的PCIe End Point作为地址转换单元,配置PCIe BAR地址寄存器,映射PCIe存储设备内的地址到CPU的MMIO地址,使得对CPU地址(可供DMA的地址)空间的访问,可以直接映射到PCIe存储设备可访问介质的地址。

[0140] 本发明实施例中建立PCIe存储设备内的介质可访问地址与所述MMIO地址之间的对应关系,即建立PCIe存储设备的地址空间与CPU中的MMIO地址空间之间的映射关系。该映

射关系的建立，依据PCIe存储设备的物理地址是否是线性连续的存储地址，实现方式有所不同。

[0141] 当PCIe存储设备的物理地址是线性连续的存储地址时，地址转换单元配置的BAR地址寄存器，映射一段与实际存储设备大小一致的CPU地址空间，由于PCIe存储设备地址是线性连续的，该PCIe存储设备地址与映射的CPU地址空间的地址一一对应。CPU对这段地址空间的操作能够被地址转换单元获取，地址转换单元将接收到的请求转换为实际的PCIe存储设备的物理地址。

[0142] 当PCIe存储设备的物理地址不是线性连续的存储地址时，地址转换单元将经过线性排序后的存储地址与CPU的地址建立映射。即配置BAR地址寄存器，映射一段与虚拟地址大小一致的CPU地址空间，该虚拟地址是该非线性连续的存储地址经过线性排序后的地址。例如，可以将扇区、块或其它最小单元单位连续的打上标签，形成一个“虚拟”的线性空间，再与CPU地址空间建立映射关系。

[0143] 通过上述数据处理系统中的管理单元建立的所述PCIe存储设备的BAR中的基地址与所述PCIe存储设备的唯一标识之间的对应关系，所述管理单元能够获取所述数据处理系统接收到的数据请求的请求的数据的MMIO地址，并依据该MMIO地址，向PCIe控制器发起数据请求；PCIe控制器中的地址转换单元依据建立的PCIe存储设备内的介质可访问地址与所述MMIO地址之间的对应关系，获取请求的数据在PCIe存储设备的介质可访问地址，即实际的物理地址，通过该介质可访问地址，能够直接读取数据或写入数据。

[0144] 相应的，所述数据处理系统中的网络适配器能够根据上述MMIO地址，直接从所述PCIe存储设备读取数据或向所述PCIe存储设备写入数据。避免了现有技术中需要将PCIe存储设备上相关的数据先读取到物理内存，然后再通过该物理内存传递到远程数据处理系统时对内存和CPU资源的消耗和带来的传输时延问题。

[0145] 下面以第一数据处理系统需要将PCIe SSD F的偏移地址0x1000～0x2000共4Kbytes的数据写入第二数据处理系统的PCIe SSD G的0x3000～0x4000的位置为例，对本发明实施例的数据处理系统和数据处理方法的实现方式做详细说明。其中，PCIe SSD即PCIe存储设备的一种具体实现方式。本实施例以第一数据处理系统不仅创建了自身的PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系，而且还获取了第二数据处理系统中PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系为例进行说明。

[0146] 步骤500：第一数据处理系统获取PCIe SSD的BAR中的基地址；

[0147] 第一数据处理系统通过管理单元中PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系，获取PCIe SSD F的PCIe BAR中的基地址为0xff00 0000，得到偏移地址0x1000对应的MMIO地址，例如0xff00 1000；同时还获取PCIe SSD G的PCIe BAR中的基地址0xfe00 0000，得到偏移地址的0x3000～0x4000对应的MMIO地址，例如0xfe00 3000。

[0148] 步骤502：第一数据处理系统的网络适配器发起数据写请求；

[0149] 第一数据处理系统的网络适配器通过PCIe SSD F的新的MMIO地址0xff00 1000向PCIe SSD F设备发起数据请求，PCIe SSD F设备内部PCIe接口将MMIO地址转换为内部可访问的存储介质地址XXX，这个XXX的结果根据不同的映射单元实现会有所不同。PCIe SSD F读取存储介质地址XXX对应的数据，读取完成后将数据通过PCIe总线发送给网络适配器，这整个过程中CPU无需参与数据的传输过程，同时也不需要内存的占用。

[0150] 步骤504:第二数据处理系统的网络适配器接收所述第一数据处理系统的网络适配器发送的数据,并发送给PCIe SSD G;

[0151] 第二数据处理系统的网络适配器通过PCIe SSD G的新的MMIO地址0xfe003000向PCIe SSD G设备发起写数据请求,PCIe SSD G设备PCIe接口将MMIO地址转换为内部可访问的存储介质地址YYY,这个YYY的结果根据不同的映射单元实现会有所不同。PCIe SSD G将网络适配器传送来的数据写入存储介质地址YYY,这整个过程中CPU无需参与数据的传输过程,同时也不需要内存的占用。

[0152] 上述PCIe SSD F设备内部PCIe接口将MMIO地址转换为内部可访问的存储介质地址XXX,以及PCIe SSD G设备PCIe接口将MMIO地址转换为内部可访问的存储介质地址YYY,是针对第二存储地址是线连续的物理地址的情况。当第二存储地址是逻辑地址时,PCIe SSD F设备内部PCIe接口将MMIO地址转换为内部可访问的存储介质地址XXX,还需要一个从逻辑地址到非线性连续地址之间的转换,此处不再赘述。

[0153] 对于第一数据处理系统未获取第二数据处理系统中PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系的情况,第二数据处理系统的管理单元建立该第二数据处理系统中的PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系,在第二数据处理系统的网络适配器接收到第一数据处理系统发送的数据读写请求时,依据所述第二数据处理系统中的PCIe SSD的唯一标识与PCIe SSD的BAR中的基地址之间的对应关系,获取要写入数据的PCIe SSD的BAR中的基地址,直接写入对应的PCIe SSD中。

[0154] 本发明的上述实施例中,是以PCIe存储设备为例说明在不同数据处理系统之间传递数据时直接读取或写入数据的实现方式。对于通过内存接口的存储设备,例如SCM(Storage Class Memory)、RRAM(Resistive Random Access Memory)、NVDIMM(Non-Volatile DIMMs)等,也可以参照上述PCIe存储设备的实现方式,通过管理单元对SCM、RRAM和NVDIMM进行数据的直接读写处理。当系统启动时,存储设备(NVDIMM\RRAM\SCM)在初始化的时候会将其访问的空间在系统内注册,通过访问注册后的地址,即可以对存储设备进行访问。与上述实施例不同的是,无需进行PCIe存储设备内地址可访问介质地址与MMIO地址的映射,也不需要地址转换,但为了让网络上其它的数据处理系统获取所要访问的数据的地址,需要获取存储设备(NVDIMM\RRAM\SCM)在系统内的地址并同步其它的数据处理系统。

[0155] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本发明的范围。

[0156] 所属领域的技术人员可以清楚地了解到,为了描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0157] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另外,所显示或讨

论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口、装置或单元的间接耦合或通信连接，也可以是电的，机械的或其它的形式连接。

[0158] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本发明实施例方案的目的。

[0159] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以是两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0160] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分，或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0161] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

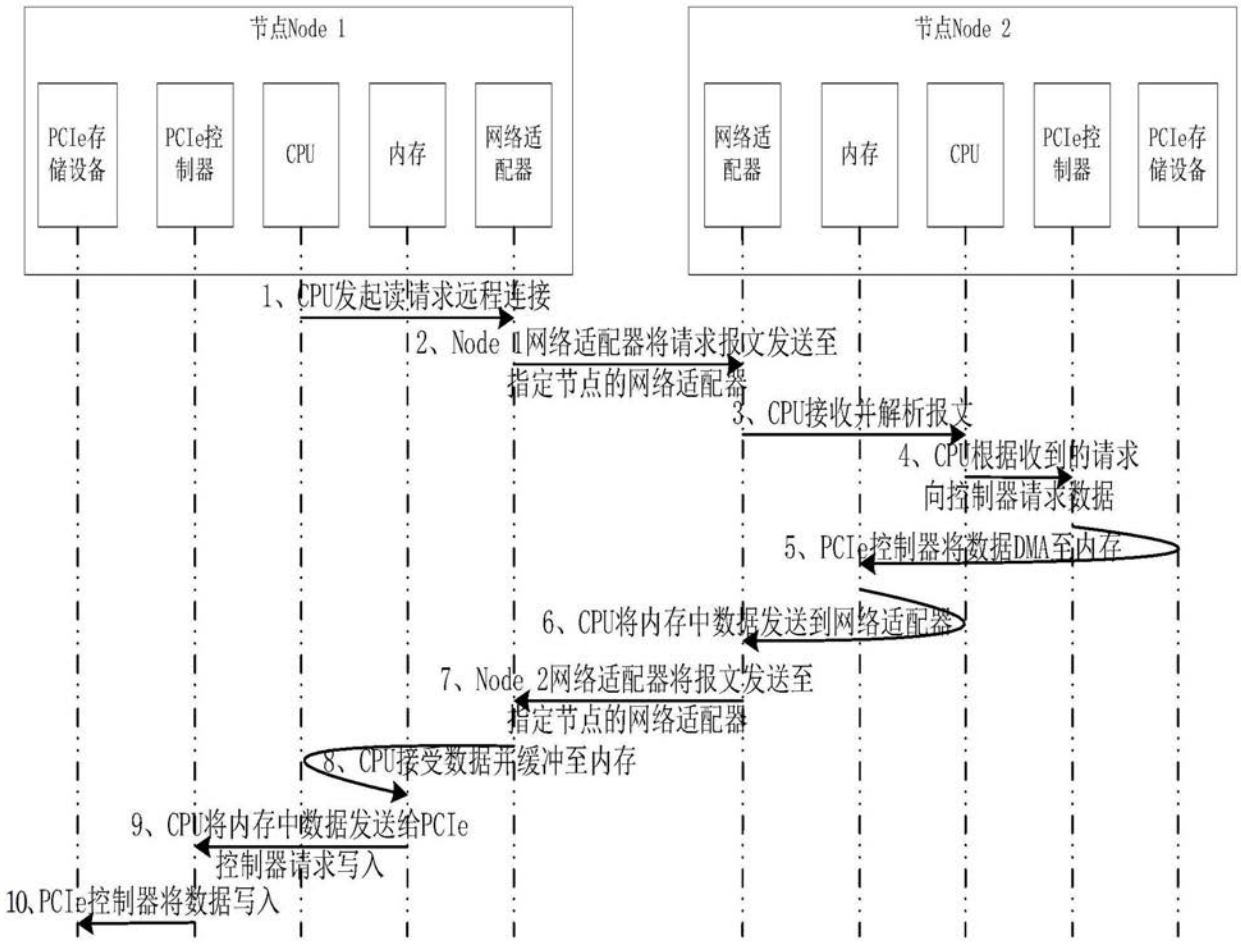

节点Node 1
节点Node 2
PCIe存储设备
PCIe控制器
CPU
内存
网络适配器
网络适配器
内存
CPU
PCIe控制器
PCIe存储设备
1、CPU发起读请求远程连接
2、Node 1网络适配器将请求报文发送至指定节点的网络适配器
3、CPU接收并解析报文
4、CPU根据收到的请求向控制器请求数据
5、PCIe控制器将数据DMA至内存
6、CPU将内存中数据发送到网络适配器
7、Node 2网络适配器将报文发送至指定节点的网络适配器
8、CPU接受数据并缓冲至内存
9、CPU将内存中数据发送给PCIe控制器请求写入
10、PCIe控制器将数据写入

图1

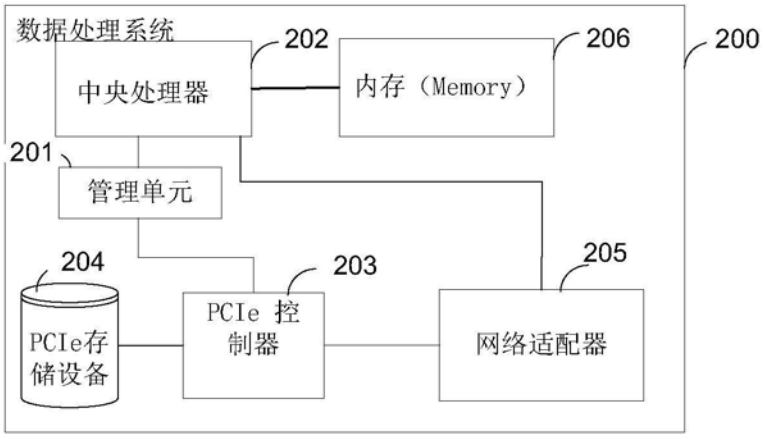

数据处理系统
200
202
中央处理器
206
内存（Memory）
201
管理单元
204
PCIe存储设备
203
PCIe 控制器
205
网络适配器

图2

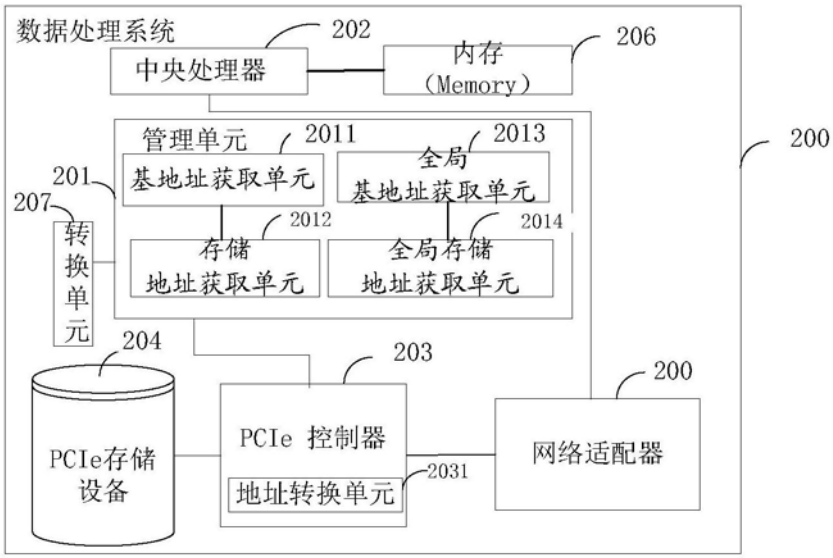

数据处理系统
202
中央处理器
内存
(Memory)
206
200
管理单元
2011
基地址获取单元
2013
全局
基地址获取单元
201
207
转换单元
2012
存储
地址获取单元
2014
全局存储
地址获取单元
204
203
200
PCIe存储设备
PCIe 控制器
地址转换单元
2031
网络适配器

图3

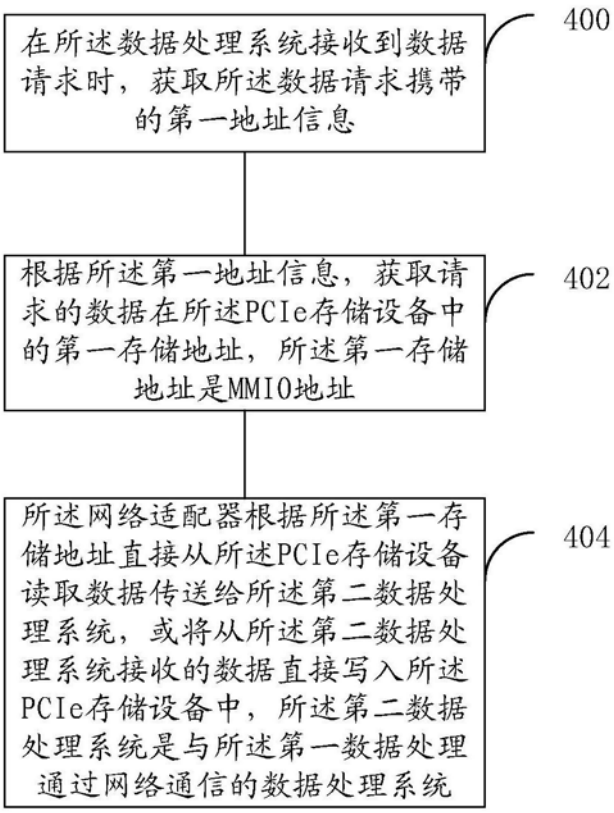

在所述数据处理系统接收到数据请求时，获取所述数据请求携带的第一地址信息
400
根据所述第一地址信息，获取请求的数据在所述PCIe存储设备中的第一存储地址，所述第一存储地址是MMIO地址
402
所述网络适配器根据所述第一存储地址直接从所述PCIe存储设备读取数据传送给所述第二数据处理系统，或将从所述第二数据处理系统接收的数据直接写入所述PCIe存储设备中，所述第二数据处理系统是与所述第一数据处理通过网络通信的数据处理系统
404

图4

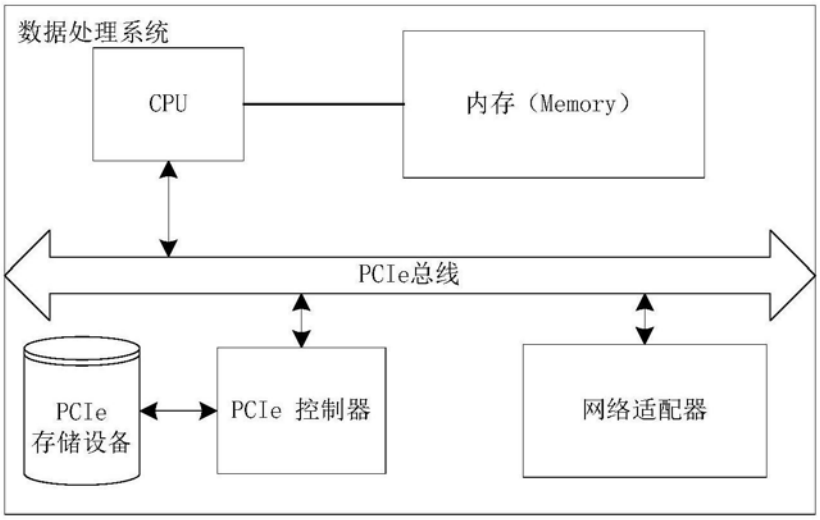

数据处理系统
CPU
内存（Memory）
PCIe总线
PCIe
存储设备
PCIe 控制器
网络适配器

图5

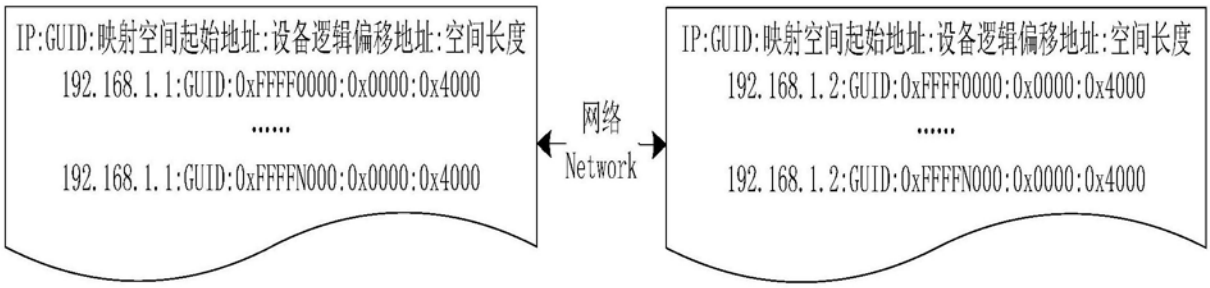

IP:GUID:映射空间起始地址:设备逻辑偏移地址:空间长度
192.168.1.1:GUID:0xFFFF0000:0x0000:0x4000
.....
192.168.1.1:GUID:0xFFFFN000:0x0000:0x4000
网络
Network
IP:GUID:映射空间起始地址:设备逻辑偏移地址:空间长度
192.168.1.2:GUID:0xFFFF0000:0x0000:0x4000
.....
192.168.1.2:GUID:0xFFFFN000:0x0000:0x4000

图6

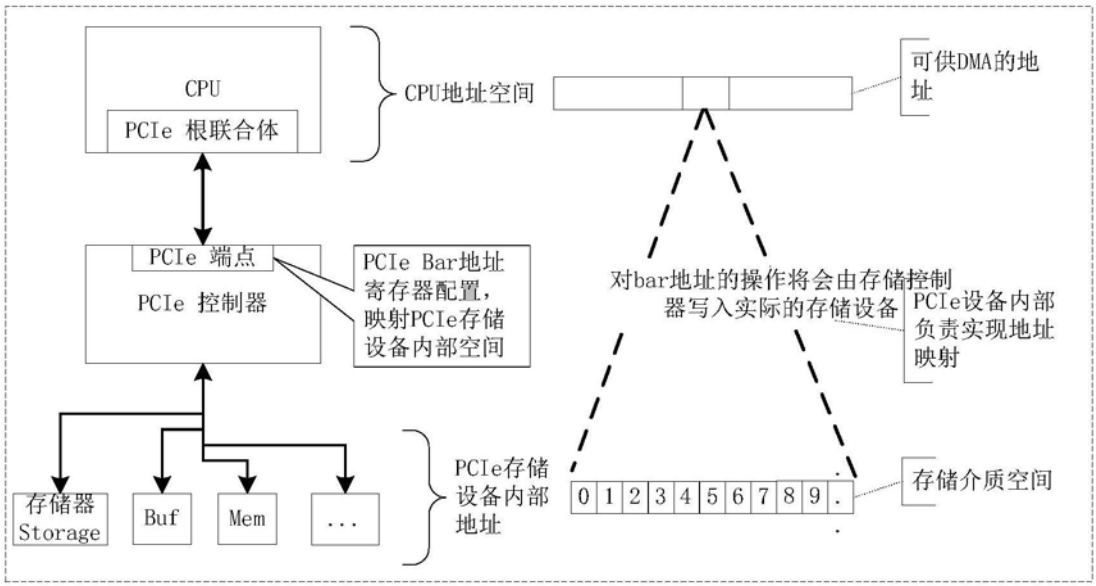

CPU
PCIe 根联合体
CPU地址空间
可供DMA的地址
PCIe 端点
PCIe 控制器
PCIe Bar地址寄存器配置，映射PCIe存储设备内部空间
对bar地址的操作将会由存储控制器写入实际的存储设备
PCIe设备内部负责实现地址映射
存储器
Storage
Buf
Mem
...
PCIe存储设备内部地址
0 1 2 3 4 5 6 7 8 9 .
存储介质空间

图7

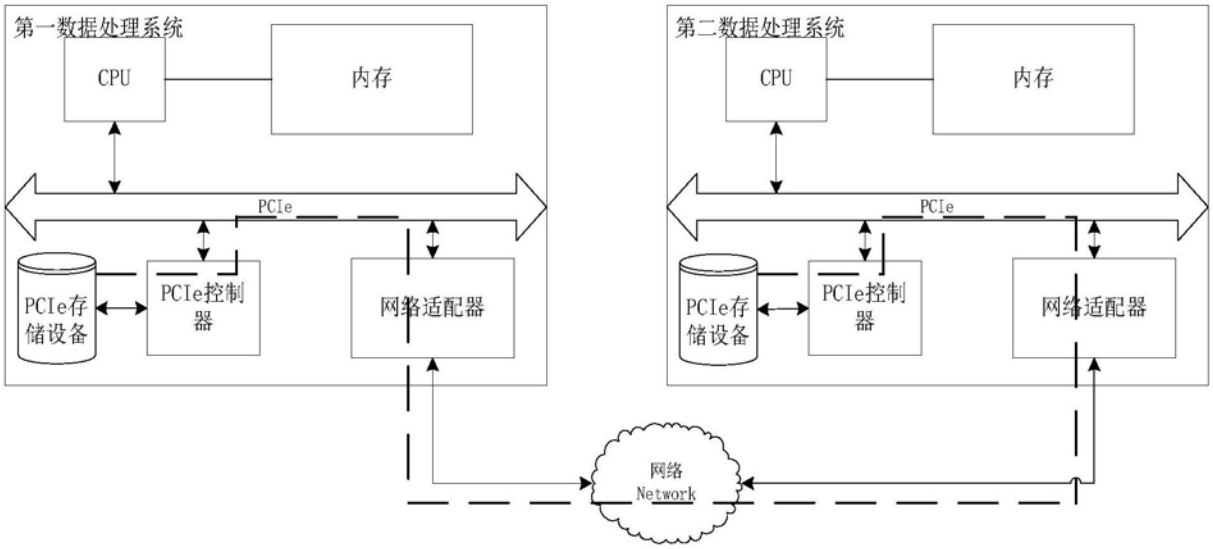

第一数据处理系统
CPU
内存
PCIe
PCIe存储设备
PCIe控制器
网络适配器
第二数据处理系统
CPU
内存
PCIe
PCIe存储设备
PCIe控制器
网络适配器
网络
Network

图8