(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111597135 A
(43)申请公布日 2020.08.28

(21)申请号 202010561517.3

(22)申请日 2020.06.18

(71)申请人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区砚瓦池正街47号

(72)发明人 刘威 龚锐 石伟 周宏伟 张剑锋 任巨 杨乾明 张见 王永文

(74)专利代理机构 湖南兆弘专利事务所(普通合伙) 43008
代理人 谭武艺

(51)Int.Cl.
G06F 13/42(2006.01)
H04L 12/935(2013.01)

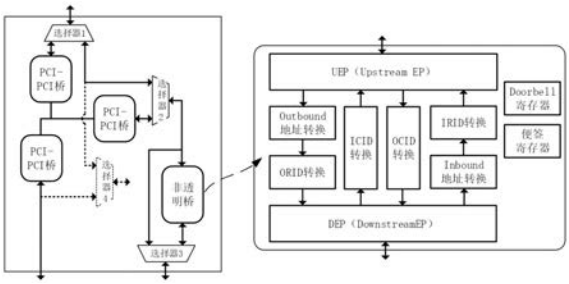
权利要求书3页 说明书10页 附图2页

(54)发明名称

一种透明桥和非透明桥功能可选的PCIE交换器及多主机系统

(57)摘要

本发明公开了一种透明桥和非透明桥功能可选的PCIE交换器及多主机系统，PCIE交换器包括透明桥、第一选择器、第二选择器、第三选择器以及非透明桥，第一选择器的固定端口作为上游端口，第一选择器的选择端口分别与透明桥的上游端口、第二选择器的一个选择端口相连，第二选择器的选择端口与透明桥的下游端口相连，第二选择器的固定端口分为两路且其中一路串接非透明桥后与第三选择器的一个选择端口相连、另一路作为非透明桥的旁路通道与第三选择器的另一个选择端口相连，第三选择器的固定端口作为PCIE交换器的一个下游端口。本发明能够实现透明桥和非透明桥功能可选，实现地址路由和ID路由兼容。



CN 111597135 A

1.一种透明桥和非透明桥功能可选的PCIE交换器,包括透明桥,其特征在于,还包括第一选择器、第二选择器、第三选择器以及非透明桥,所述第一选择器的固定端口作为PCIE交换器的上游端口,所述第一选择器的选择端口分别与透明桥的上游PCI-PCI桥的上游端口、第二选择器的一个选择端口相连,所述第二选择器的选择端口与透明桥的一个下游PCI-PCI桥的下游端口相连,所述第二选择器的固定端口分为两路且其中一路串接非透明桥后与第三选择器的一个选择端口相连、另一路作为非透明桥的旁路通道与第三选择器的另一个选择端口相连,第三选择器的固定端口作为PCIE交换器的一个下游端口。

2.根据权利要求1所述的透明桥和非透明桥功能可选的PCIE交换器,其特征在于,所述非透明桥包括上游端点设备模块UEP、下游端点设备模块DEP、门铃寄存器模块以及便签寄存器模块,所述上游端点设备模块UEP和下游端点设备模块DEP之间并联有四条通路:第一条通路包括出站地址转换模块、出站请求报文ID转换模块;第二条通路包括响应出站请求完成报文ID转换模块;第三条通路包括响应入站请求完成报文ID转换模块;第四条通路包括入站地址转换模块、入站请求报文ID转换模块;所述上游端点设备模块UEP上游方向与第二选择器的固定端口相连、下游方向分别与出站地址转换模块、响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块、入站请求报文ID转换模块、门铃寄存器模块、便签寄存器模块相连,所述上游端点设备模块UEP、下游端点设备模块DEP为标准PCIE端点设备。

3.根据权利要求2所述的透明桥和非透明桥功能可选的PCIE交换器,其特征在于,所述上游端点设备模块UEP中带有多个基地址寄存器,用于分配连接在PCIE交换器的上游端口的第一系统中的配置、IO地址和内存空间;所述下游端点设备模块DEP中带有多个基地址寄存器,用于分配连接在PCIE交换器的上游端口的第二系统中的配置、IO地址和内存空间;所述出站地址转换模块、入站地址转换模块用于实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换,所述出站请求报文ID转换模块、入站请求报文ID转换模块用于实现第一系统、第二系统之间的请求报文ID转换,所述响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块用于实现第一系统、第二系统之间的响应出站请求完成报文或响应入站请求完成报文的ID转换。

4.根据权利要求3所述的透明桥和非透明桥功能可选的PCIE交换器,其特征在于,出站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括:出站地址转换模块对收到的来自第一系统请求报文的地址域进行基地址寄存器中的基地址匹配,如果匹配上了,就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第二系统地址域里的地址,所述基地址转换寄存器里保存的是第二系统域里的基地址,如果没有匹配上,就直接丢弃或指示上游端点设备模块UEP返回不支持该请求;所述入站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括:入站地址转换模块对收到的来自第二系统请求报文的地址域进行基地址寄存器中的基地址匹配,如果匹配上了,就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第一系统地址域里的地址,所述基地址转换寄存器里保存的是第一系统域里的基地址,如果没有匹配上,就直接丢弃或指示下游端点设备模块DEP返回不支持该请求。

5.根据权利要求3所述的透明桥和非透明桥功能可选的PCIE交换器,其特征在于,所述出站请求报文ID转换模块有一个全局表,全局表需要在第一系统枚举配置完成之后,且在

第一系统发出请求之前被初始化完毕，该表需要设计8项，每项有4个字段：INDEX字段、Bus字段、Dev字段和Func字段，其中Func字段为第一系统域的功能号，INDEX字段为第一系统域的功能号对应的第二系统域的功能号，Bus字段用于记录总线号，Dev字段用于记录设备号，长度分别为3位、8位、5位和3位；出站请求报文ID转换模块收到第一系统域请求报文ID={FBus，FDev，FFunc}后，其中FBus表示总线号，FDev表示设备号，FFunc表示功能号；首先将第一系统域请求报文的请求ID采用内容匹配查找方式查找全局表，FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段，如果没有命中全局表中的任意一项，则返回不支持该请求；则将请求ID的FFunc字段用命中表项的INDEX字段替代，使得命中表项的INDEX字段为转换后的第二系统请求报文的请求ID的新FFunc字段，同时将请求ID的FDev字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPBus替代，从而得到转换到第二系统域的请求报文ID={DEPBus，DEPDev，INDEX}。

6.根据权利要求5所述的透明桥和非透明桥功能可选的PCIE交换器，其特征在于，响应入站请求完成报文ID转换模块实现第一系统、第二系统之间的响应入站请求完成报文的ID转换时，接收来自第二系统域的响应入站请求完成报文ID={DEPBus，DEPDev，INDEX}，利用其中的INDEX字段访问出站请求报文ID转换模块的全局表，读出命中表项的Bus字段、Dev字段和Func字段，并将其填入转换后的进入第一系统的响应入站请求完成报文ID={Bus，Dev，Func}。

7.根据权利要求3所述的透明桥和非透明桥功能可选的PCIE交换器，其特征在于，所述入站请求报文ID转换模块有一个全局表，全局表需要在第一系统枚举配置完成之后，且在第二系统发出请求之前被初始化完毕，该表需要设计8项，每项有4个字段：INDEX字段、Bus字段、Dev字段和Func字段，其中Func字段为第二系统域的功能号，INDEX字段为第二系统域的功能号对应的第一系统域的功能号，Bus字段用于记录总线号，Dev字段用于记录设备号，长度分别为3位、8位、5位和3位；出站请求报文ID转换模块收到第二系统域请求报文ID={FBus，FDev，FFunc}后，其中FBus表示总线号，FDev表示设备号，FFunc表示功能号；首先将第二系统域请求报文的请求ID采用内容匹配查找方式查找全局表，FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段，如果没有命中全局表中的任意一项，则返回不支持该请求；则将请求ID的FFunc字段用命中表项的INDEX字段替代，使得命中表项的INDEX字段为转换后的第一系统请求报文的请求ID的新FFunc字段，同时将请求ID的FDev字段使用上游端点设备模块UEP捕获的第一系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第一系统的设备号DEPBus替代，从而得到转换到第一系统域的请求报文ID={DEPBus，DEPDev，INDEX}。

8.根据权利要求7所述的透明桥和非透明桥功能可选的PCIE交换器，其特征在于，响应出站请求完成报文ID转换模块实现第一系统、第二系统之间的响应出站请求完成报文的ID转换时，接收来自第一系统域的响应入站请求完成报文ID={DEPBus，DEPDev，INDEX}，利用其中的INDEX字段访问入站请求报文ID转换模块的全局表，读出命中表项的Bus字段、Dev字段和Func字段，并将其填入转换后的进入第二系统的响应入站请求完成报文ID={Bus，Dev，Func}。

9.根据权利要求2所述的透明桥和非透明桥功能可选的PCIE交换器，其特征在于，所述

门铃寄存器模块包含一组寄存器，用来第一系统和第二系统之间传递中断，所述门铃寄存器模块的寄存器包括：第一系统中断状态寄存器、第一系统中断请求寄存器、第一系统中断掩饰置位寄存器、第一系统中断掩饰清零寄存器、第二系统中断状态寄存器、第二系统中断请求寄存器、第二系统中断掩饰置位寄存器、第二系统中断掩饰清零寄存器，上述每个寄存器可被第一系统和第二系统通过内存地址或IO地址空间访问它们，只要请求寄存器有位被置起，且没有被掩饰，那么就输出有效中断；如果请求位被清掉了，或是被掩饰了，那么就将中断无效；便签寄存器模块也包含一组寄存器，第一系统和第二系统都可通过内存地址或IO地址空间访问它们，可读可写，这组寄存器用于传递控制、状态信息，或单纯作为读写状态寄存器。

10.一种多主机系统，至少包括第一系统、第二系统，所述第一系统、第二系统之间通过PCIE交换器相连，其特征在于，所述PCIE交换器为权利要求1～9中任意一项所述的透明桥和非透明桥功能可选的PCIE交换器。

一种透明桥和非透明桥功能可选的PCIE交换器及多主机系统

技术领域

[0001] 本发明涉及集成电路设计领域，具体涉及一种透明桥和非透明桥功能可选的PCIE交换器及多主机系统。

背景技术

[0002] 高速外围组件互联PCI-Express(Peripheral Component Interconnect Express，简称PCIE)是一种高速串行计算机扩展总线标准，PCIE设备可分为三种类型：根复合体、SWITCH和端点设备(Endpoint)。典型的PCIE树形拓扑如图1所示，在树形拓扑中只有一个根复合体，它负责发现整个拓扑的结构，包括其中的总线和各节点设备，并分配对应的总线号和地址空间。PCIE SWITCH内部包含多个PCI-PCI桥，这些PCI-PCI桥通常被称为透明桥。

[0003] 近年，分布式系统得到广泛发展，多主机系统可以提供高带宽的同时，还能提供更好的稳定行。但是，PCIE协议中树形拓扑中只有一个根复合体的规定，是得传统PCIE SWITCH不能友好地构建多主机系统，从而需要在SWITCH中实现非透明桥，用来隔离多个主机系统地地址空间。非透明桥通过假装成Endpoint，向两个主机端暴露的是Type0型地配置空间，那么两边的枚举软件就会把非透明桥都当作是拓扑中的叶节点，即两个主机系统都看不到对方拓扑中的设备了，从而达到被隔离的目的。同时非透明桥还利用Type0配置头中的BAR寄存器对双向的请求进行地址和ID转换，将这些请求的发起者从主机系统转换为非透明桥，然后在另一个主机系统中进行路由，这样就实现两个主机系统之间地址空间共享。

[0004] PCIE协议定义了三种路由方式：地址路由；ID路由和隐式路由。其中ID的含义是节点编号，有该节点的BUS号、DEV号和FUNC号组成，通常也记为BDF。PCIE协议还定义四种事务：存储读写；IO读写；配置读写；消息。MEM读写和IO读写是地址路由，配置读写和完成报文是ID路由，消息通常是隐式路由。非透明桥要实现两个主机系统之间的交互，就必须支持上述四种报文的转换处理。在SWITCH非透明桥的系统架构中，主机系统可以连接在非透明桥的端口上。这种架构的应用场景可以是两个主机系统交换内存数据，也可以是非透明端口的主机系统与另一个地址域的设备进行数据交换。这两种应用场景对系统架构的需求是不一样的，与不在本拓扑树上的设备交换数据，那么必须要有访问这些设备的路径，也就是转换后的请求也要能上SWITCH内部的虚拟总线。这些访问也是延迟敏感的，当然，这种敏感在内存数据交换应用中更突出。

[0005] 综述所述，PCIE SWITCH中非透明桥的结构需要面向上述两种应用场景，完成各类请求的转换和地址空间共享。

发明内容

[0006] 本发明要解决的技术问题：针对现有技术的上述问题，提供一种透明桥和非透明桥功能可选的PCIE交换器及多主机系统，本发明能够实现透明桥和非透明桥功能可选，实现地址路由和ID路由兼容。

[0007] 为了解决上述技术问题,本发明采用的技术方案为:

一种透明桥和非透明桥功能可选的PCIE交换器,包括透明桥,还包括第一选择器、第二选择器、第三选择器以及非透明桥,所述第一选择器的固定端口作为PCIE交换器的上游端口,所述第一选择器的选择端口分别与透明桥的上游PCI-PCI桥的上游端口、第二选择器的一个选择端口相连,所述第二选择器的选择端口与透明桥的一个下游PCI-PCI桥的下游端口相连,所述第二选择器的固定端口分为两路且其中一路串接非透明桥后与第三选择器的一个选择端口相连、另一路作为非透明桥的旁路通道与第三选择器的另一个选择端口相连,第三选择器的固定端口作为PCIE交换器的一个下游端口。

[0008] 可选地,所述非透明桥包括上游端点设备模块UEP、下游端点设备模块DEP、门铃寄存器模块以及便签寄存器模块,所述上游端点设备模块UEP和下游端点设备模块DEP之间并联有四条通路:第一条通路包括出站地址转换模块、出站请求报文ID转换模块;第二条通路包括响应出站请求完成报文ID转换模块;第三条通路包括响应入站请求完成报文ID转换模块;第四条通路包括入站地址转换模块、入站请求报文ID转换模块;所述上游端点设备模块UEP上游方向与第二选择器的固定端口相连、下游方向分别与出站地址转换模块、响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块、入站请求报文ID转换模块、门铃寄存器模块、便签寄存器模块相连,所述上游端点设备模块UEP、下游端点设备模块DEP为标准PCIE端点设备。

[0009] 可选地,所述上游端点设备模块UEP中带有多个基地址寄存器,用于分配连接在PCIE交换器的上游端口的第一系统中的配置、IO地址和内存空间;所述下游端点设备模块DEP中带有多个基地址寄存器,用于分配连接在PCIE交换器的上游端口的第二系统中的配置、IO地址和内存空间;所述出站地址转换模块、入站地址转换模块用于实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换,所述出站请求报文ID转换模块、入站请求报文ID转换模块用于实现第一系统、第二系统之间的请求报文ID转换,所述响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块用于实现第一系统、第二系统之间的响应出站请求完成报文或响应入站请求完成报文的ID转换。

[0010] 可选地,出站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括:出站地址转换模块对收到的来自第一系统请求报文的地址域进行基地址寄存器中的基地址匹配,如果匹配上了,就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第二系统地址域里的地址,所述基地址转换寄存器里保存的是第二系统域里的基地址,如果没有匹配上,就直接丢弃或指示上游端点设备模块UEP返回不支持该请求;所述入站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括:入站地址转换模块对收到的来自第二系统请求报文的地址域进行基地址寄存器中的基地址匹配,如果匹配上了,就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第一系统地址域里的地址,所述基地址转换寄存器里保存的是第一系统域里的基地址,如果没有匹配上,就直接丢弃或指示下游端点设备模块DEP返回不支持该请求。

[0011] 可选地,所述出站请求报文ID转换模块有一个全局表,全局表需要在第一系统枚举配置完成之后,且在第一系统发出请求之前被初始化完毕,该表需要设计8项,每项有4个字段:INDEX字段、Bus字段、Dev字段和Func字段,其中Func字段为第一系统域的功能号,

INDEX字段为第一系统域的功能号对应的第二系统域的功能号,Bus字段用于记录总线号,Dev字段用于记录设备号,长度分别为3位、8位、5位和3位;出站请求报文ID转换模块收到第一系统域请求报文ID={FBus,FDev,FFunc}后,其中FBus表示总线号,FDev表示设备号,FFunc表示功能号;首先将第一系统域请求报文的请求ID采用内容匹配查找方式查找全局表,FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段,如果没有命中全局表中的任意一项,则返回不支持该请求;则将请求ID的FFunc字段用命中表项的INDEX字段替代,使得命中表项的INDEX字段为转换后的第二系统请求报文的请求ID的新FFunc字段,同时将请求ID的FDev字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPBus替代,从而得到转换到第二系统域的请求报文ID={DEPBus,DEPDev,INDEX}。

[0012] 可选地,响应入站请求完成报文ID转换模块实现第一系统、第二系统之间的响应入站请求完成报文的ID转换时,接收来自第二系统域的响应入站请求完成报文ID={DEPBus,DEPDev,INDEX},利用其中的INDEX字段访问出站请求报文ID转换模块的全局表,读出命中表项的Bus字段、Dev字段和Func字段,并将其填入转换后的进入第一系统的响应入站请求完成报文ID={Bus,Dev,Func}。

[0013] 可选地,所述入站请求报文ID转换模块有一个全局表,全局表需要在第一系统枚举配置完成之后,且在第二系统发出请求之前被初始化完毕,该表需要设计8项,每项有4个字段:INDEX字段、Bus字段、Dev字段和Func字段,其中Func字段为第二系统域的功能号,INDEX字段为第二系统域的功能号对应的第一系统域的功能号,Bus字段用于记录总线号,Dev字段用于记录设备号,长度分别为3位、8位、5位和3位;出站请求报文ID转换模块收到第二系统域请求报文ID={FBus,FDev,FFunc}后,其中FBus表示总线号,FDev表示设备号,FFunc表示功能号;首先将第二系统域请求报文的请求ID采用内容匹配查找方式查找全局表,FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段,如果没有命中全局表中的任意一项,则返回不支持该请求;则将请求ID的FFunc字段用命中表项的INDEX字段替代,使得命中表项的INDEX字段为转换后的第一系统请求报文的请求ID的新FFunc字段,同时将请求ID的FDev字段使用上游端点设备模块UEP捕获的第一系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第一系统的设备号DEPBus替代,从而得到转换到第一系统域的请求报文ID={DEPBus,DEPDev,INDEX}。

[0014] 可选地,响应出站请求完成报文ID转换模块实现第一系统、第二系统之间的响应出站请求完成报文的ID转换时,接收来自第一系统域的响应入站请求完成报文ID={DEPBus,DEPDev,INDEX},利用其中的INDEX字段访问入站请求报文ID转换模块的全局表,读出命中表项的Bus字段、Dev字段和Func字段,并将其填入转换后的进入第二系统的响应入站请求完成报文ID={Bus,Dev,Func}。

[0015] 可选地,所述门铃寄存器模块包含一组寄存器,用来第一系统和第二系统之间传递中断,所述门铃寄存器模块的寄存器包括:第一系统中断状态寄存器、第一系统中断请求寄存器、第一系统中断掩饰置位寄存器、第一系统中断掩饰清零寄存器、第二系统中断状态寄存器、第二系统中断请求寄存器、第二系统中断掩饰置位寄存器、第二系统中断掩饰清零寄存器,上述每个寄存器可被第一系统和第二系统通过内存地址或IO地址空间访问它们,只要请求寄存器有位被置起,且没有被掩饰(mask),那么就输出有效中断;如果请求位被清

掉了,或是被掩饰了,那么就将中断无效;便签寄存器模块也包含一组寄存器,第一系统和第二系统都可通过内存地址或IO地址空间访问它们,可读可写,这组寄存器用于传递控制、状态信息,或单纯作为读写状态寄存器。

[0016] 此外,本发明还提供一种多主机系统,至少包括第一系统、第二系统,所述第一系统、第二系统之间通过PCIE交换器相连,所述PCIE交换器为前述的透明桥和非透明桥功能可选的PCIE交换器。

[0017] 和现有技术相比,本发明具有下述优点:本发明包括透明桥、第一选择器、第二选择器、第三选择器以及非透明桥,第一选择器的固定端口作为PCIE交换器的上游端口,第一选择器的选择端口分别与透明桥的上游PCI-PCI桥的上游端口、第二选择器的一个选择端口相连,第二选择器的选择端口与透明桥的一个下游PCI-PCI桥的下游端口相连,第二选择器的固定端口分为两路且其中一路串接非透明桥后与第三选择器的一个选择端口相连、另一路作为非透明桥的旁路通道与第三选择器的另一个选择端口相连,第三选择器的固定端口作为PCIE交换器的一个下游端口,本发明能够实现透明桥和非透明桥功能可选,实现地址路由和ID路由兼容,能够适应两个主机系统交换内存数据、非透明端口的主机系统与另一个地址域的设备进行数据交换这两种典型的应用场景完成各类请求的转换和地址空间共享。

附图说明

[0018] 图1为现有技术的PCIE透明桥SWITCH系统结构示意图。

[0019] 图2为本发明实施例中PCIE交换器的结构示意图。

[0020] 图3为本发明实施例中地址转换的实现方式示意图。

[0021] 图4为本发明实施例中ID转换的实现方式示意图。

具体实施方式

[0022] 如图2所示,本实施例透明桥和非透明桥功能可选的PCIE交换器包括透明桥,还包括第一选择器(图中为选择器1)、第二选择器(图中为选择器2)、第三选择器(图中为选择器3)以及非透明桥,所述第一选择器的固定端口作为PCIE交换器的上游端口,所述第一选择器的选择端口分别与透明桥的上游PCI-PCI桥的上游端口、第二选择器的一个选择端口相连,所述第二选择器的选择端口与透明桥的一个下游PCI-PCI桥的下游端口相连,所述第二选择器的固定端口分为两路且其中一路串接非透明桥后与第三选择器的一个选择端口相连、另一路作为非透明桥的旁路通道与第三选择器的另一个选择端口相连,第三选择器的固定端口作为PCIE交换器的一个下游端口。参见图2可知,本实施例中在传统PCIE交换器的上游(Upstream)端口和下流(Downstream)端口间增加选择器1、选择器2和一条旁路通道(bypass通路),通过不同的配置实现透明桥和旁路通道之间的选择。选择器2和选择器3中间也增加一条旁路通道,以及非透明桥模块。本实施例透明桥和非透明桥功能可选的PCIE交换器可以通过不同的配置实现透明桥和非透明桥功能之间的选择。选择器1与上游方向的根复合体或是其他PCIE交换器相连。本实施例中,记选择器1上游方向的主机系统为第一系统。选择器3与下游方向的根复合体或其他PCIE交换器相连。即选择器3下游方向的主机系统为第二系统。需要说明的是:下述的上游、下游、出站(Outbound)和入站(Inbound)等表

示请求方向的定义,都是针对第一系统而言。例如出站(Outbound)方向是第一系统发出的请求,入站(Inbound)是第二系统发出的请求,其他类推。此外,还可以根据需要在其他下游PCI-PCI桥的下游端口和透明桥的上游PCI-PCI桥的上游端口之间添加更多的选择器来实现更多通路的选择, 例如图中的选择器4。

[0023] 第一系统主机的根复合体连接在上游端口,即图2中的选择器1。该系统架构可以工作在4种应用场景。(1)第二系统主机的根复合体连接在非透明桥端口,即图2中的选择器3。如果使用该SWITCH的应用场景是第一系统和第二系统角色对等,即它们主要是交换内存数据,那么配置选择器1和选择器2,将上游和下游的PCI-PCIE透明桥旁路,配置选择器3的选择信号,选择输出非透明桥到下游端口,从而实现第一系统和第二系统之间低延迟的通信。(2)如果第一系统是主,第二系统是从,即第二系统扮演的是备份的角色,那么配置选择器1和选择器2的选择信号,不旁路上游和下游的PCI-PCIE透明桥,配置选择器3的选择信号,选择输出非透明桥到下游端口,那么第二系统的主机可以访问到SWITCH内部的虚拟总线,即可与虚拟总线上的其他下游端点设备。(3)还可以通过选择器4的实现与否以及相应的选择信号控制,将非透明桥模块与其他下游端口连接,实现非透明端口的可配置。(4)本发明提出的SWITCH非透明桥的架构完全兼容图1种传统的SWITCH架构。即可以配置选择器1、选择器2和选择器3,将非透明桥旁路掉,实现兼容性的扩展。

[0024] 如图2所示,本实施例中的非透明桥包括上游端点设备模块UEP(Upstream UEP)、下游端点设备模块DEP(DownstreamEP)、门铃(Doorbell)寄存器模块以及便签寄存器模块,上游端点设备模块UEP和下游端点设备模块DEP之间并联有四条通路:

第一条通路包括出站地址转换模块(图2中简称为Outbound地址转换)、出站请求报文ID转换模块(图2中简称为ORID转换);

第二条通路包括响应出站请求完成报文ID转换模块(图2中简称为OCID转换);

第三条通路包括响应入站请求完成报文ID转换模块(图2中简称为ICID转换);

第四条通路包括入站地址转换模块(图2中简称为Inbound地址转换)、入站请求报文ID转换模块(图2中简称为IRID转换);

上游端点设备模块UEP上游方向与第二选择器的固定端口相连、下游方向分别与出站地址转换模块、响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块、入站请求报文ID转换模块、门铃寄存器模块、便签寄存器模块相连,上游端点设备模块UEP、下游端点设备模块DEP为标准PCIE端点设备。

[0025] 参见图2,上游端点设备模块UEP中带有多个基地址寄存器(简称BAR),用于分配连接在PCIE交换器的上游端口的第一系统中的配置、IO地址和内存空间;所述下游端点设备模块DEP中带有多个基地址寄存器,用于分配连接在PCIE交换器的上游端口的第二系统中的配置、IO地址和内存空间;所述出站地址转换模块、入站地址转换模块用于实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换,所述出站请求报文ID转换模块、入站请求报文ID转换模块用于实现第一系统、第二系统之间的请求报文ID转换,所述响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块用于实现第一系统、第二系统之间的响应出站请求完成报文或响应入站请求完成报文的ID转换。

[0026] 其中,上游端点设备模块UEP暴露在第一系统中的地址空间是通过UEP中的6各BAR实现的。第一系统的软件通过配置UEP里的6个BAR寄存器(基地址寄存器)的类型和大小,分

配第一系统中的配置、IO和MEM空间。上游端点设备模块UEP是第一系统中PCIE拓扑结构的终点，它在第二系统中不可见。

[0027] 其中，下游端点设备模块DEP在下游方向与选择器3相连，上游方向与入站地址转换模块、响应出站请求完成报文ID转换模块、响应入站请求完成报文ID转换模块、出站请求报文ID转换模块、门铃寄存器模块、便签寄存器模块相连。下游端点设备模块DEP是标准的PCIE端点设备，暴露在第二系统中的地址空间是通过下游端点设备模块DEP中的6各BAR实现的。第二系统的软件通过配置下游端点设备模块DEP里的6个BAR寄存器的类型和大小，分配第二系统中的配置、IO和MEM空间。下游端点设备模块DEP是第二系统中PCIE拓扑结构的终点，它在第一系统中不可见。

[0028] 出站地址转换模块与上游端点设备模块UEP和出站请求报文ID转换模块相连。MEM读写和IO读写是地址路由，在出站地址转换模块将上游端点设备模块UEP发出的这类请求中的地址域进行转换，将地址域的信息从第一系统的地址空间转换到第二系统的地址空间中。

[0029] 本实施例中，出站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括：出站地址转换模块对收到的来自第一系统请求报文的地址域进行基地址寄存器中的基地址匹配，如果匹配上了，就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第二系统地址域里的地址，所述基地址转换寄存器里保存的是第二系统域里的基地址，如果没有匹配上，就直接丢弃或指示上游端点设备模块UEP返回不支持该请求；所述入站地址转换模块实现第一系统、第二系统之间的IO地址和内存空间的地址空间转换的详细步骤包括：入站地址转换模块对收到的来自第二系统请求报文的地址域进行基地址寄存器中的基地址匹配，如果匹配上了，就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第一系统地址域里的地址，所述基地址转换寄存器里保存的是第一系统域里的基地址，如果没有匹配上，就直接丢弃或指示下游端点设备模块DEP返回不支持该请求。

[0030] 地址转换的实现原理如图3所示，PCIE端点设备Type0配置头里的6个BAR，用于系统软件给该设备需要的各类空间进行地址分配。通过这6个BAR，设备可以申请6个32位空间或是3个64位地址空间。假设图3中第一系统地址域中的白色空白域是上游端点设备模块UEP的某个BAR声明的空间大小，第二系统地址域中的白色空白域是下游端点设备模块DEP的某个BAR声明的空间大小。当第一系统发出的请求命中了上游端点设备模块UEP的BAR，如图3左中斜线阴影部分，请求报文的地址对应于BAR基址还带有地址偏移。那么出站地址转换模块的功能是将该请求的地址域转换为第二系统中对应的地址。具体地址转换实现如如图3右所示，出站地址转换模块对收到的请求报文的地址域进行BAR基地址匹配，如果匹配上了，就将报文地址域的偏移和基地址转换寄存器的基地址重新组合成第二系统地址域里的地址；如果没有匹配上，就直接丢弃或指示上游端点设备模块UEP返回UR请求。其中基地址转换寄存器里保存的是第二系统域里的基地址。入站地址转换模块的功能与是出站地址转换模块功能一样，将第二系统里的请求的地址域转换为第一系统中对应的地址，不同之处在于Inbound地址转换时基地址转换寄存器里保存的是第一系统域里的基地址。

[0031] 出站请求报文ID转换模块与出站地址转换模块和下游端点设备模块DEP相连。出站请求报文ID转换模块与出站地址转换模块和下游端点设备模块DEP相连。根据PCIE协议

定义,MEM读写和IO读写请求报文头里,除了地址域外,还有请求(Requester)的ID域信息,该ID就是第一系统给其PCIE拓扑中各个节点分配的总线号(Bus)、设备号(Dev)和功能号(Func)。因此,对MEM读写和IO读写请求除了需要地址转换外,还需要将其Requester ID转换到第二系统中的请求ID。转换后的请求,在第二系统中被认为是下游端点设备模块DEP发起的请求。因此,转换后的请求ID为下游端点设备模块DEP在第二系统中的拓扑节点号。

[0032] 本实施例中,出站请求报文ID转换模块有一个全局表,全局表需要在第一系统枚举配置完成之后,且在第一系统发出请求之前被初始化完毕,该表需要设计8项,每项有4个字段:INDEX字段、Bus字段、Dev字段和Func字段,其中Func字段为第一系统域的功能号,INDEX字段为第一系统域的功能号对应的第二系统域的功能号,Bus字段用于记录总线号,Dev字段用于记录设备号,长度分别为3位、8位、5位和3位;出站请求报文ID转换模块收到第一系统域请求报文ID={FBus,FDev,FFunc}后,其中FBus表示总线号,FDev表示设备号,FFunc表示功能号;首先将第一系统域请求报文的请求ID采用内容匹配查找方式查找全局表,FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段,如果没有命中全局表中的任意一项,则返回不支持该请求;则将请求ID的FFunc字段用命中表项的INDEX字段替代,使得命中表项的INDEX字段为转换后的第二系统请求报文的请求ID的新FFunc字段,同时将请求ID的FDev字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第二系统的设备号DEPBus替代,从而得到转换到第二系统域的请求报文ID={DEPBus,DEPDev,INDEX}。

[0033] ID转换的实现原理如图4所示,出站请求报文ID转换模块的功能是将Outbound请求报文里的Requester ID域从第一系统域转换为第二系统域的ID值。记发起该请求的节点号为{FBus,FDev,FFunc},即出站请求报文ID转换模块收到的第一系统域请求报文的Requester ID={FBus,FDev,FFunc}。在出站请求报文ID转换模块里要有一个全局表项,全局表需要在第一系统枚举配置完成之后,且在第一系统发出请求之前被初始化完毕,该表需要设计8项,每项有4个字段:INDEX字段、Bus字段、Dev字段和Func字段,分别需要3位、8位、5位和3位。该表设计8项的原因是出站请求报文ID转换模块在转换Requester ID的时候,只利用第一系统域请求报文的Requester ID的FFunc字段,而PCIE协议定义的功能号字段只有3位,故该全局表只需要设计8项。转换过程首先将第一系统域请求报文的Requester ID采用CAM方式查找全局表,CAM方式即是内容匹配查找方式,如果{FBus,FDev,FFunc}没有命中了表中任何一项{Bus,Dev,Func},那么就返回不支持该请求(UR请求);如果命中某一项,那么就将第一系统域请求报文的Requester ID的FFunc字段用命中表项的INDEX字段替代,即INDEX字段为转换后的第二系统请求报文的Requester ID的新FFunc字段。请求报文经过转换后,在第二系统主机软件看来,该请求就是下游端点设备模块DEP发起的。所以,新的Requester ID的FBus字段和FDev字段,需要利用下游端点设备模块DEP先前捕获的下游端点设备模块DEPBus值和下游端点设备模块DEPDev值替换,即第二系统请求报文的Requester ID={下游端点设备模块DEPBus,下游端点设备模块DEPDev,INDEX}。那么这意味着第一系统和第二系统之间发送请求之前,各自的系统主机软件已经完成了各自系统中PCIE拓扑结构的枚举和配置,因为出站请求报文ID转换模块需要使用下游端点设备模块DEP捕获的下游端点设备模块DEPBus值和下游端点设备模块DEPDev值,而各拓扑节点的捕获时机就是在各自主机软件对其进行枚举后配置的过程中,当各节点的配置请求到达时,

节点会将配置请求报文里的总线字段和设备字段记录下来，完成捕获动作。通过ORID转换后，该请求进入第二系统，第二系统主机软件认为该请求的发起节点为下游端点设备模块DEP。当该请求在第二系统中到达目的节点，目的节点会返回完成报文。

[0034] 响应入站请求完成报文ID转换模块分别与下游端点设备模块DEP、上游端点设备模块UEP相连。对应出站请求报文ID转换模块发起到第二系统中的请求报文，ICID模块接收其完成报文。PCIE协议定义的完成报文中既有请求 ID域和完成 ID域，前者对应其请求报文中的请求ID域，后者对应该完成报文的拓扑节点号。因为该请求是在第二系统中完成的，那么ICID需要同时将请求ID域和完成ID域转换为第一系统的拓扑节点，即将请求ID域从下游端点设备模块DEP的节点号转换为第一系统中真正的请求发起端点的节点号，将完成ID域由第二系统中的完成节点号转换为上游端点设备模块UEP在第一系统中的节点号，让第一系统软件感知到的是该请求是上游端点设备模块UEP完成的。

[0035] 本实施例中，响应入站请求完成报文ID转换模块实现第一系统、第二系统之间的响应入站请求完成报文的ID转换时，接收来自第二系统域的响应入站请求完成报文ID={DEPBus，DEPDev，INDEX}，利用其中的INDEX字段访问出站请求报文ID转换模块的全局表，读出命中表项的Bus字段、Dev字段和Func字段，并将其填入转换后的进入第一系统的响应入站请求完成报文ID={Bus，Dev，Func}。完成报文是ID路由，没有地址字段，因此完成报文穿过下游端点设备模块DEP后到达响应入站请求完成报文ID转换模块时，不需要对其进行地址转换，但是，完成报文里除了包含请求ID字段外，还有Completer ID字段，该字段表示的是第二系统里的目标节点的ID，即ICID模块需要完成两个ID的转换。首先是将第二系统的请求ID转换为该请求的发起节点在第一系统中的原始ID。在完成报文里，请求 ID={下游端点设备模块DEPBus，下游端点设备模块DEPDev，INDEX}，那么需要利用查找表方式对全局表进行索引，即利用INDEX为地址项访问全局表，读出命中表项的总线号、设备号和功能号字段，填入转换后的进入第一系统的完成报文，即此时Requester ID={Bus，Dev，Func}。因为在ORID模块进行ID转换时采用的CAM查找方式命中的全局表，这意味着{FBus，FDev，FFunc}={Bus，Dev，Func}，所以响应入站请求完成报文ID转换模块对完成报文里的Requester ID字段是正确的。完成报文里的完成 ID字段不需要查找全局表，只需要将各字段利用上游端点设备模块UEP捕获的字段替换即可。这样处理是合理的，是因为在第一系统的主机软件的视图里，第二系统地址域不可见，所有发往第二系统地址域的请求，在第一系统主机软件认为都是发往上游端点设备模块UEP的，所以所有完成报文的完成ID字段里应该填入上游端点设备模块UEP在第一系统PCIE拓扑里分配的节点ID。

[0036] 入站地址转换模块与下游端点设备模块DEP和响应入站请求完成报文ID转换模块相连。该模块将下游端点设备模块DEP发出的MEM读写和IO读写请求中的地址域进行转换，将地址域的信息从第二系统的地址空间转换到第一系统的地址空间中。同样地，来至第二系统中的这类Inbound请求除了需要地址转换外，还需要ID转换。ID转换在响应入站请求完成报文ID转换模块中实现。响应入站请求完成报文ID转换模块与入站地址转换模块和上游端点设备模块UEP相连。响应入站请求完成报文ID转换模块将入站请求的请求ID转换到第一系统中的请求 ID。转换后的请求，在第一系统中被认为是上游端点设备模块UEP发起的请求。因此，转换后的请求 ID为上游端点设备模块UEP在第一系统中的拓扑节点号。

[0037] 本实施例中，入站请求报文ID转换模块有一个全局表，全局表需要在第一系统枚

举配置完成之后，且在第二系统发出请求之前被初始化完毕，该表需要设计8项，每项有4个字段：INDEX字段、Bus字段、Dev字段和Func字段，其中Func字段为第二系统域的功能号，INDEX字段为第二系统域的功能号对应的第一系统域的功能号，Bus字段用于记录总线号，Dev字段用于记录设备号，长度分别为3位、8位、5位和3位；出站请求报文ID转换模块收到第二系统域请求报文ID={FBus，FDev，FFunc}后，其中FBus表示总线号，FDev表示设备号，FFunc表示功能号；首先将第二系统域请求报文的请求ID采用内容匹配查找方式查找全局表，FBus用于匹配Bus字段、FDev用于匹配Dev字段、FFunc用于匹配Func字段，如果没有命中全局表中的任意一项，则返回不支持该请求；则将请求ID的FFunc字段用命中表项的INDEX字段替代，使得命中表项的INDEX字段为转换后的第一系统请求报文的请求ID的新FFunc字段，同时将请求ID的FDev字段使用上游端点设备模块UEP捕获的第一系统的设备号DEPDev替代、请求ID的FBus字段使用下游端点设备模块DEP捕获的第一系统的设备号DEPBus替代，从而得到转换到第一系统域的请求报文ID={DEPBus，DEPDev，INDEX}。

[0038] 响应出站请求完成报文ID转换模块分别与下游端点设备模块DEP、上游端点设备模块UEP相连。对应响应入站请求完成报文ID转换模块发起到第一系统中的请求报文，响应出站请求完成报文ID转换模块接收其完成报文。因为该请求是在第一系统中完成的，那么响应出站请求完成报文ID转换模块需要同时将请求 ID域和完成 ID域转换为第二系统的拓扑节点，即将请求 ID域从上游端点设备模块UEP的节点号转换为第二系统中真正的请求发起端点的节点号，将完成 ID域由第一系统中的完成节点号转换为下游端点设备模块DEP在第二系统中的节点号，让第二系统软件感知到的是该请求是下游端点设备模块DEP完成的。

[0039] 本实施例中，响应出站请求完成报文ID转换模块实现第一系统、第二系统之间的响应出站请求完成报文的ID转换时，接收来自第一系统域的响应入站请求完成报文ID={DEPBus，DEPDev，INDEX}，利用其中的INDEX字段访问入站请求报文ID转换模块的全局表，读出命中表项的Bus字段、Dev字段和Func字段，并将其填入转换后的进入第二系统的响应入站请求完成报文ID={Bus，Dev，Func}。

[0040] 参见前文可知，出站请求和入站请求的报文地转换过程虽然方向相反，但是功能是一致地。入站完成报文和出站完成报文的报文地转换过程虽然方向相反，但是功能是一致地。

[0041] 门铃寄存器模块和便签寄存器模块都分别与下游端点设备模块DEP、上游端点设备模块UEP相连，其中，门铃寄存器用来传递中断；便签寄存器用来传递状态之类的信息。

[0042] 本实施例中，门铃寄存器模块包含一组寄存器(本实施例中每个寄存器16位)，用来第一系统和第二系统之间传递中断，所述门铃寄存器模块的寄存器包括：第一系统中断状态寄存器、第一系统中断请求寄存器、第一系统中断掩饰置位寄存器、第一系统中断掩饰清零寄存器、第二系统中断状态寄存器、第二系统中断请求寄存器、第二系统中断掩饰置位寄存器、第二系统中断掩饰清零寄存器，上述每个寄存器可被第一系统和第二系统通过内存地址或IO地址空间访问它们，只要请求寄存器有位被置起，且没有被掩饰，那么就输出有效中断；如果请求位被清掉了，或是被掩饰了，那么就将中断无效(deassert)；本实施例中在切换状态中，中断有效(assert)和中断无效(deassert)之间的跳转后，还需要向上传递相应的报文，例如中断报文INTx、消息告知中断报文MSI等。

[0043] 本实施例中，便签寄存器模块也包含一组寄存器(通常为8个)，第一系统和第二系统都可通过内存地址或IO地址空间访问它们，可读可写，这组寄存器用于传递控制、状态信息，或单纯作为读写状态寄存器。读写这组寄存器，不会捕出中断。

[0044] PCIE协议定义的其他类型请求，例如配置读写请求和消息。其中，配置读写请求不会穿过非透明桥，它到达非透明桥就结束了，因为第一系统的配置请求只命中上游端点设备模块UEP，第二系统的配置请求只命中下游端点设备模块DEP。消息类请求也不会穿过非透明桥，消息到达非透明桥后直接被丢弃，然后通过门铃寄存器模块和便签寄存器模块完成第一系统和第二系统之间的通信。

[0045] 综上所述，本发明透明桥和非透明桥功能可选的PCIE交换器能够实现透明桥和非透明桥功能可选，实现地址路由和ID路由兼容，能够适应两个主机系统交换内存数据、非透明端口的主机系统与另一个地址域的设备进行数据交换这两种典型的应用场景完成各类请求的转换和地址空间共享。

[0046] 此外，本发明还提供一种多主机系统，至少包括第一系统、第二系统，所述第一系统、第二系统之间通过PCIE交换器相连，所述PCIE交换器为前述的透明桥和非透明桥功能可选的PCIE交换器。

[0047] 以上所述仅是本发明的优选实施方式，本发明的保护范围并不仅局限于上述实施例，凡属于本发明思路下的技术方案均属于本发明的保护范围。应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理前提下的若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

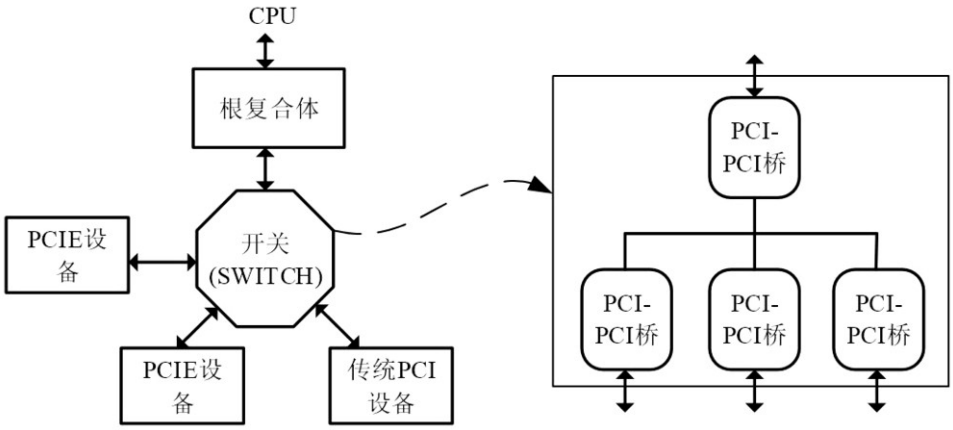

CPU
根复合体
PCIE设备
开关(SWITCH)
PCIE设备
传统PCI设备
PCI-PCI桥
PCI-PCI桥
PCI-PCI桥
PCI-PCI桥

图 1

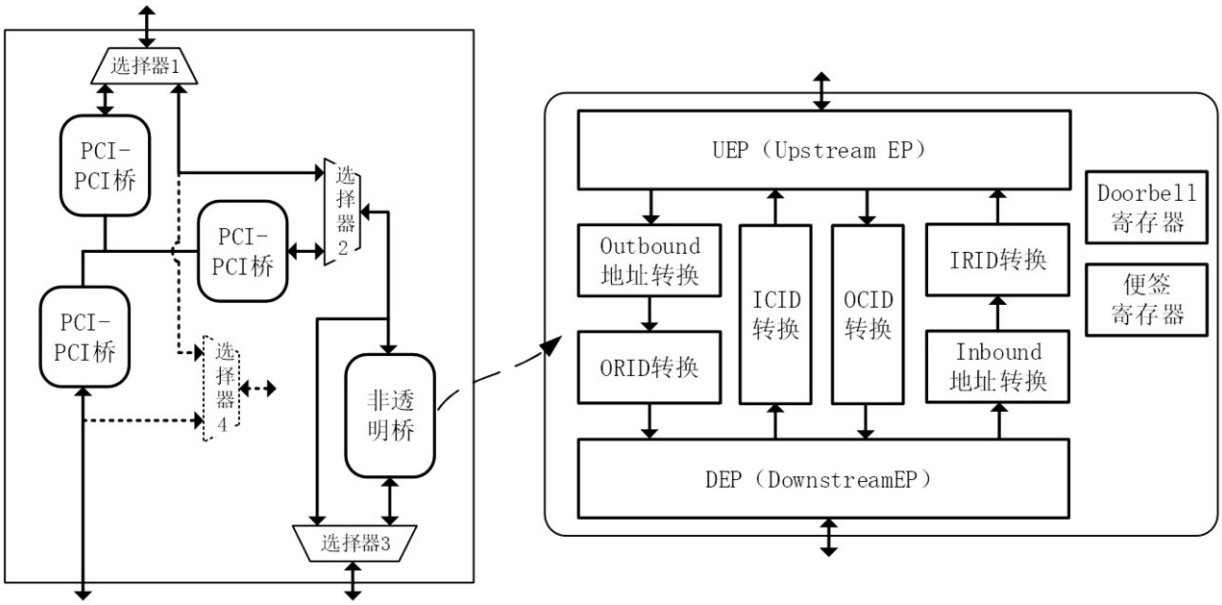

选择器1
PCI-PCI桥
PCI-PCI桥
PCI-PCI桥
选择器2
选择器4
非透明桥
选择器3
UEP（Upstream EP）
Outbound地址转换
ORID转换
ICID转换
OCID转换
IRID转换
Inbound地址转换
DEP（DownstreamEP）
Doorbell寄存器
便签寄存器

图 2

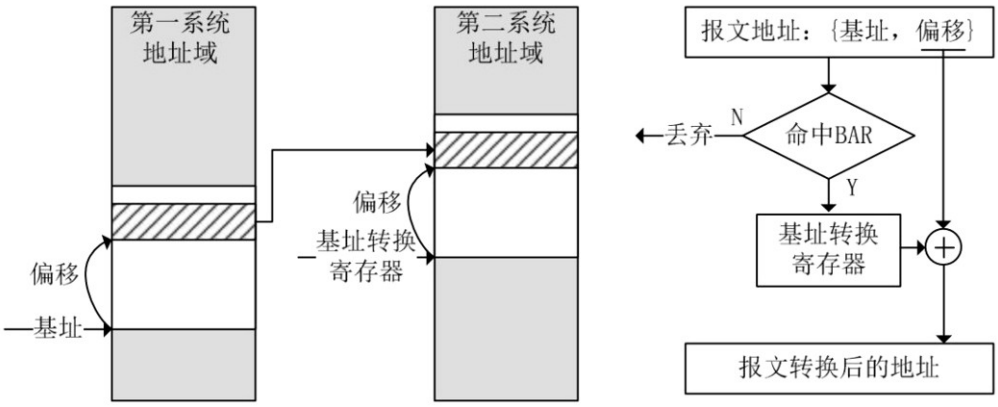

第一系统
地址域
第二系统
地址域
偏移
基址转换
寄存器
偏移
基址
报文地址：{基址，偏移}
丢弃
N
命中BAR
Y
基址转换
寄存器
报文转换后的地址

图 3

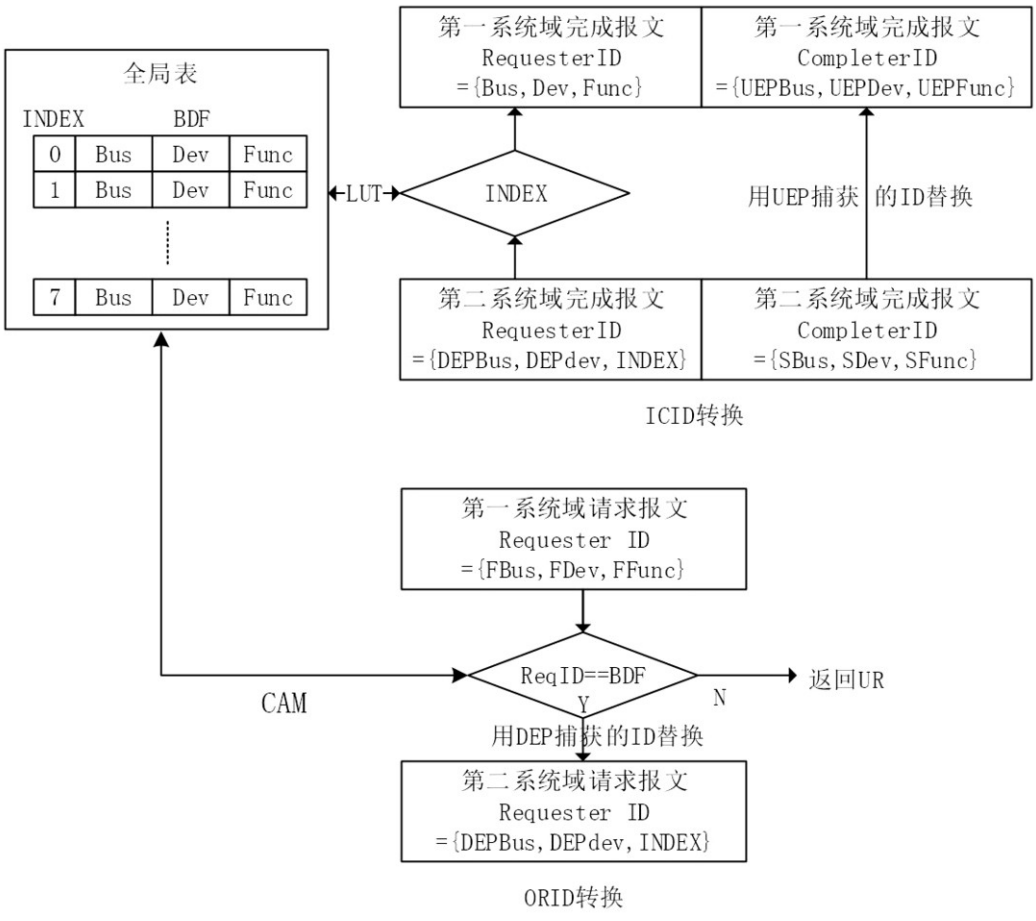

全局表
INDEX
BDF
0 Bus Dev Func
1 Bus Dev Func
7 Bus Dev Func
LUT
INDEX
第一系统域完成报文
RequesterID
={Bus,Dev,Func}
第一系统域完成报文
CompleterID
={UEPBus,UEPDev,UEPFunc}
用UEP捕获 的ID替换
第二系统域完成报文
RequesterID
={DEPBus,DEPdev,INDEX}
第二系统域完成报文
CompleterID
={SBus,SDev,SFunc}
ICID转换
第一系统域请求报文
Requester ID
={FBus,FDev,FFunc}
ReqID==BDF
Y
N
返回UR
CAM
用DEP捕获的ID替换
第二系统域请求报文
Requester ID
={DEPBus,DEPdev,INDEX}
ORID转换

图 4