(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113168389 B
(45)授权公告日 2023.03.31

(21)申请号 201980079575.8
(22)申请日 2019.04.24
(65)同一申请的已公布的文献号
申请公布号 CN 113168389 A
(43)申请公布日 2021.07.23
(30)优先权数据
62/785,778 2018.12.28 US
(85)PCT国际申请进入国家阶段日
2021.06.02
(86)PCT国际申请的申请数据
PCT/CN2019/084111 2019.04.24
(87)PCT国际申请的公布数据
WO2020/133841 EN 2020.07.02

(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 唐洪亮 万力 陈莉莉 汤志豪
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 陈聪

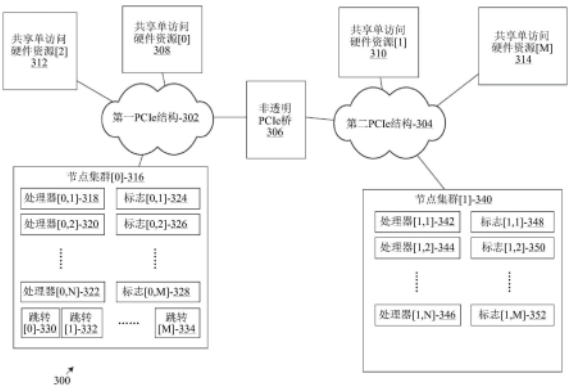
(51)Int.Cl.
G06F 13/364(2006.01)
(56)对比文件
CN 107015928 A,2017.08.04
CN 107066405 A,2017.08.18
CN 106462394 A,2017.02.22
CN 104715001 A,2015.06.17
审查员 王邦吉
权利要求书3页 说明书10页 附图6页

(54)发明名称

用于锁定具有非透明桥接的PCIe网络的装置和方法

(57)摘要

一种互连计算机系统,包括快捷外围部件互连标准(Peripheral Component Interconnect Express,PCIe)结构、以通信方式耦合到所述PCIe结构的第一计算机系统、以通信方式耦合到所述PCIe结构的第二计算机系统,以及耦合到所述PCIe结构的共享单访问硬件资源。所述第一计算机系统包括第一处理器和耦合到所述第一处理器的第一存储器,其中,所述第一存储器用于存储:第一标志(flag),用于指示希望所述第一计算机系统访问所述共享单访问硬件资源;和跳转(turn)变量,用于指示所述第一计算机系统和所述第二计算机系统中的哪个有权访问所述共享单访问硬件资源。所述第二计算机系统包括第二处理器和耦合到所述第二处理器的第二存储器,其中,所述第二存储器用于存储第二标志,所述第二标志指示希望所述第二计算机系统访问所述共享单访问硬件资源。



CN 113168389 B

1.一种互连计算机系统,其特征在于,包括:

PCIe(快捷外围部件互连标准)结构;

第一计算机系统,以通信方式耦合到所述PCIe结构;

第二计算机系统,以通信方式耦合到所述PCIe结构;

共享单访问硬件资源,耦合到所述PCIe结构;

所述第一计算机系统包括:

第一处理器;

第一本地存储器,耦合到所述第一处理器,用于存储:

第一标志,用于指示希望所述第一计算机系统访问所述共享单访问硬件资源,所述第一标志对应的变量存储在本地且通过本地写入和本地读取来执行访问;以及

跳转变量,用于指示所述第一计算机系统和所述第二计算机系统中的哪个有权访问所述共享单访问硬件资源;

所述第二计算机系统包括:

第二处理器;

第二本地存储器,耦合到所述第二处理器,用于存储第二标志,所述第二标志指示希望所述第二计算机系统访问所述共享单访问硬件资源,所述第二标志对应的变量存储在本地且通过本地写入和本地读取来执行访问。

2.根据权利要求1所述的互连计算机系统,其特征在于:

所述第一计算机系统位于所述PCIe结构的非透明桥的第一侧;

所述第二计算机系统位于所述PCIe结构的所述非透明桥的第二侧。

3.根据权利要求2所述的互连计算机系统,其特征在于,所述PCIe结构的所述非透明桥对经过其中的数据写入和数据读取进行地址转换。

4.根据权利要求3所述的互连计算机系统,其特征在于:

所述第一计算机系统包括多个第一子计算机系统,所述多个第一子计算机系统位于所述PCIe结构的非透明桥的第一侧;

所述第二计算机系统包括多个第二子计算机系统,所述多个第二子计算机系统位于所述PCIe结构的非透明桥的第二侧。

5.根据权利要求4所述的互连计算机系统,其特征在于:

所述第一本地存储器存储多个标志,所述多个标志指示对应地希望所述多个第一子计算机系统访问所述共享单访问硬件资源;

所述第二本地存储器存储多个标志,所述多个标志指示对应地希望所述多个第二子计算机系统访问所述共享单访问硬件资源。

6.根据权利要求3所述的互连计算机系统,其特征在于,所述共享单访问硬件资源是共享存储器。

7.根据权利要求3所述的互连计算机系统,其特征在于:

所述第一计算机系统是第一网络集群;

所述第二计算机系统是共享网络集群。

8.一种互连计算机系统,其特征在于,包括:

第一PCIe(快捷外围部件互连标准)结构;

第二PCIe结构；

PCIe非透明桥，将所述第一PCIe结构与所述第二PCIe结构互连；

第一节点集群，以通信方式耦合到所述第一PCIe结构；

第二节点集群，以通信方式耦合到所述第二PCIe结构；

共享单访问硬件资源，耦合到所述第一PCIe结构和所述第二PCIe结构之一；

所述第一节点集群包括第一本地存储器，所述第一本地存储器用于存储：

至少一个第一标志，用于指示希望所述第一节点集群访问所述共享单访问硬件资源，所述第一标志对应的变量存储在本地且通过本地写入和本地读取来执行访问；

跳转变量，用于指示所述第一节点集群和所述第二节点集群中的哪个有权访问所述共享单访问硬件资源；

所述第二节点集群包括第二本地存储器，所述第二本地存储器用于存储至少一个第二标志，所述至少一个第二标志指示希望所述第二节点集群访问所述共享单访问硬件资源，所述第二标志对应的变量存储在本地且通过本地写入和本地读取来执行访问。

9.根据权利要求8所述的互连计算机系统，其特征在于：

所述第一节点集群包括多个第一子计算机系统，所述多个第一子计算机系统位于所述PCIe结构的非透明桥的第一侧；

所述第二节点集群包括多个第二子计算机系统，所述多个第二子计算机系统位于所述PCIe结构的非透明桥的第二侧。

10.根据权利要求8或9所述的互连计算机系统，其特征在于，所述PCIe结构的所述非透明桥对经过其中的数据写入和数据读取进行地址转换。

11.根据权利要求10所述的互连计算机系统，其特征在于，所述共享单访问硬件资源是共享存储器。

12.一种操作互连计算机系统的方法，其中，所述互连计算机系统具有PCIe(快捷外围部件互连标准)结构、以通信方式耦合到所述PCIe结构的第一计算机系统、以通信方式耦合到所述PCIe结构的第二计算机系统，以及耦合到所述PCIe结构的共享单访问硬件资源，其特征在于，所述方法包括：

所述第一计算机系统在第一本地存储器中存储：

第一标志，用于指示希望所述第一计算机系统访问所述共享单访问硬件资源，所述第一标志对应的变量存储在本地且通过本地写入和本地读取来执行访问；

跳转变量，用于指示所述第一计算机系统和所述第二计算机系统中的哪个有权访问所述共享单访问硬件资源；

所述第二计算机系统在第二本地存储器中存储第二标志，所述第二标志指示希望所述第二计算机系统访问所述共享单访问硬件资源，所述第二标志对应的变量存储在本地且通过本地写入和本地读取来执行访问；

所述第一计算机系统根据所述第一标志和所述跳转变量的状态访问所述共享单访问硬件资源，其中，所述第一标志指示希望所述第一计算机系统访问所述共享单访问硬件资源；

所述第二计算机系统根据所述第二标志和所述跳转变量的状态访问所述共享单访问硬件资源，其中，所述第二标志指示希望所述第二计算机系统访问所述共享单访问硬件资

源。

13.根据权利要求12所述的方法,其特征在于:

所述第一计算机系统位于所述PCIe结构的非透明桥的第一侧;

所述第二计算机系统位于所述PCIe结构的所述非透明桥的第二侧。

14.根据权利要求13所述的方法,其特征在于,还包括所述PCIe结构的所述非透明桥对经过其中的数据写入和数据读取进行地址转换。

15.根据权利要求14所述的方法,其特征在于:

所述第一计算机系统包括多个第一子计算机系统,所述多个第一子计算机系统位于所述PCIe结构的非透明桥的第一侧;

所述第二计算机系统包括多个第二子计算机系统,所述多个第二子计算机系统位于所述PCIe结构的非透明桥的第二侧。

16.根据权利要求15所述的方法,其特征在于,还包括:

所述第一本地存储器存储多个标志,所述多个标志指示对应地希望所述多个第一子计算机系统访问所述共享单访问硬件资源;

所述第二本地存储器存储多个标志,所述多个标志指示对应地希望所述多个第二子计算机系统访问所述共享单访问硬件资源。

17.根据权利要求14所述的方法,其特征在于,所述共享单访问硬件资源是共享存储器。

18.根据权利要求14所述的方法,其特征在于:

所述第一计算机系统是第一网络集群;

所述第二计算机系统是共享网络集群。

用于锁定具有非透明桥接的PCIe网络的装置和方法

[0001] 相关申请案交叉申请

[0002] 本发明要求2018年12月28日递交的发明名称为“用于锁定具有非透明桥接的PCIe网络的装置和方法(Apparatus and Method for Locking PCIe Network Having Non-transparent Bridging)”的第62/785,778号美国临时申请的在先申请优先权,所述在先申请的全部内容以引入的方式并入本文本中。

技术领域

[0003] 本发明涉及通信技术;更具体地,涉及在互连计算机系统内共享单访问硬件资源。

背景技术

[0004] 互连计算机系统通常包括通过通信网络互连的多个单独计算机系统。互连计算机系统的实例包括网络存储中心、数据处理中心、万维网服务器中心以及包括多个计算机系统的其它类型的系统。服务于多个单独计算机系统的通信网络可以是光网络、局域网(Local Area Network,LAN)或其它类型的网络,例如PCI网络。

[0005] 互连计算机系统通常共享单访问硬件资源,例如存储器、共享输入/输出(Input/Output,I/O)端口、专用处理资源和大容量存储体,以及其它共享资源,所述资源在任何给定时间仅能由单个进程/处理器访问。很多时候,多个单独计算机系统希望同时访问共享单访问资源,这是不允许的。因此,已经建立了各种算法来促进对共享单访问资源的访问。

[0006] 一种此类算法被称为“Peterson算法”,由Gary L.Peterson在1981年提出。Peterson算法是实现互斥的并发程序算法,允许两个或更多个线程共享单访问资源而不发生访问冲突,仅使用共享存储器进行通信。虽然Peterson最初的公式只适用于两个线程,但所述算法后来泛化用于多于两个线程。通常,Peterson算法使用两个变量,即标志变量和跳转变量。标志[n]值为真(true)指示线程n希望进入临界区(要访问共享单访问资源)。如果线程[1]不希望进入其临界区(标志[1]＝假(false))或如果线程[1]已通过将跳转设置为0而给予线程[0]优先权,则授权线程[0]进入临界区。

[0007] 仅在标志[n]和跳转变量的写入可靠时Peterson算法才有用。在个人计算机互连(Personal Computer Interconnect,PCI)网络等一些网络中,Peterson算法效率低下是因为标志变量的远程写入不可靠,因此在远程写入标志变量后需要读取标志变量以确认标志变量已正确写入。当非透明桥形成PCI网络的一部分并分离线程时,访问共享单访问资源的效率受到严重影响。

发明内容

[0008] 为了克服现有系统及其操作的缺点,本文中公开了各种实施例。根据第一实施例,一种互连计算机系统包括快捷外围部件互连标准(Peripheral Component Interconnect Express,PCIe)结构、以通信方式耦合到所述PCIe结构的第一计算机系统、以通信方式耦合到所述PCIe结构的第二计算机系统,以及耦合到所述PCIe结构的共享单访问硬件资源。第

一计算机系统包括第一处理器，耦合到所述第一处理器的第一存储器，所述第一存储器用于存储：第一标志，用于指示希望第一计算机系统访问共享单访问硬件资源；和跳转变量，用于指示第一计算机系统和第二计算机系统中的哪个有权访问共享单访问硬件资源。第二计算机系统包括第二处理器和耦合到所述第二处理器的第二存储器，所述第二存储器用于存储第二标志，所述第二标志指示希望第二计算机系统访问共享单访问硬件资源。

[0009] 与现有系统相比，第一实施例的互连计算机系统(以及下文进一步描述的第二实施例和第三实施例)提供重要的性能增益。互连计算机系统可以实现对Peterson锁算法的改进，与现有实施方案相比，所述改进提供了重要的操作益处。传统上，Peterson锁算法包括了远程写入标志变量，即存储在线程[1]上的标志[0]和存储在线程[0]上的标志[1]。由于不可靠的远程写入，这种实施方案在PCIe网络中效率低下，因此所述远程写入之后需要进行远程读取以确认所述远程写入成功。另外，当PCIe网络包括PCIe非透明桥时，地址转换操作引入了额外的操作难度。

[0010] 因此，为了克服先前算法的问题，将标志变量存储在本地以使与远程写入(remote write，RW)和远程读取(remote read，RR)相比，所述标志变量可通过本地写入(local write，LW)和本地读取(local read，LR)访问。对于现有系统，要使线程P0访问单访问硬件资源，需要进行以下操作：RW+RR(设置标志[0])、RW+RR(设置跳转)、LR+LR(忙锁)和RW+RR(重新设置标志[0])，总共6次远程操作和2次本地操作。对于第一实施例，要使线程P0访问单访问硬件资源，需要进行以下操作：LW(设置标志[0])、RW(设置跳转)、RR(忙锁+确认RW)、LW(重新设置标志[0])，总共2次远程操作和2次本地操作。

[0011] 对于现有系统，要使线程P1访问单访问硬件资源，需要进行以下操作：RW+RR(设置标志[1])、LW(设置跳转)、LR+LR(忙锁)和RW+RR(重新设置标志[1])，总共4次远程操作和3次本地操作。对于图1的实施例，要使线程P1访问单访问硬件资源，需要进行以下操作：LW(设置标志[1]和跳转)、RR(忙锁)、LW(重新设置标志[1])，总共1次远程操作和3次本地操作。因此，对于先前系统，总共需要10次远程操作和5次本地操作。对于本文中所述的实施例，总共需要3次远程操作和5次本地操作，省去了7次远程操作。

[0012] 第一实施例包括许多可选方面，可以单独地或以任何组合应用。根据第一可选方面，第一计算机系统位于PCIe结构的非透明桥的第一侧，第二计算机系统位于PCIe结构的非透明桥的第二侧。对于此方面，PCIe结构的非透明桥可对经过其中的数据写入和数据读取进行地址转换。

[0013] 根据第二可选方面，第一计算机系统包括多个第一子计算机系统，所述多个第一子计算机系统位于PCIe结构的非透明桥的第一侧，第二计算机系统包括多个第二子计算机系统，所述多个第二子计算机系统位于PCIe结构的非透明桥的第二侧。另外，对于此方面，PCIe结构的非透明桥可对经过其中的数据写入和数据读取进行地址转换。另外，对于此可选方面，第一存储器可以存储多个标志，所述多个标志指示对应地希望多个第一子计算机系统访问共享单访问硬件资源，第二存储器可以存储多个标志，所述多个标志指示对应地希望多个第二子计算机系统访问共享单访问硬件资源。

[0014] 根据第一实施例的第三可选方面，共享单访问硬件资源是共享存储器。根据第一实施例的第四可选方面，第一计算机系统是第一网络集群，第二计算机系统是共享网络集群。

[0015] 根据本公开的第二实施例，一种互连计算机系统包括第一PCIe结构、第二PCIe结构、将第一PCIe结构与第二PCIe结构互连的PCIe非透明桥、以通信方式耦合到第一PCIe结构的第一节点集群、以通信方式耦合到第二PCIe结构的第二节点集群，以及耦合到第一PCIe结构和第二PCIe结构之一的共享单访问硬件资源。第一节点集群包括第一存储器，所述第一存储器用于存储：至少一个第一标志，用于指示希望第一节点集群访问共享单访问硬件资源；和跳转变量，用于指示第一节点集群和第二节点集群中的哪个有权访问共享单访问硬件资源。另外，第二节点集群包括第二存储器，所述第二存储器用于存储至少一个第二标志，所述至少一个第二标志指示希望第二节点集群访问共享单访问硬件资源。

[0016] 第二实施例还包括多个可选方面，可以单一地或以任何各种组合应用。根据第一可选方面，第一节点集群包括多个第一子计算机系统，所述多个第一子计算机系统位于PCIe结构的非透明桥的第一侧，第二节点集群包括多个第二子计算机系统，所述多个第二子计算机系统位于PCIe结构的非透明桥的第二侧。对于此方面，PCIe结构的非透明桥可对经过其中的数据写入和数据读取进行地址转换。根据第二可选方面，共享单访问硬件资源是共享存储器。

[0017] 根据本公开的第三实施例，公开了一种操作互连计算机系统的方法，所述互连计算机系统具有PCIe结构、以通信方式耦合到PCIe结构的第一计算机系统、以通信方式耦合到PCIe结构的第二计算机系统，以及耦合到PCIe结构的共享单访问硬件资源。根据第三实施例，操作包括第一计算机系统在第一本地存储器中存储：第一标志，用于指示希望第一计算机系统访问共享单访问硬件资源；和跳转变量，用于指示第一计算机系统和第二计算机系统中的哪个有权访问共享单访问硬件资源。第三实施例的操作还包括第二计算机系统在第二本地存储器中存储第二标志，所述第二标志指示希望第二计算机系统访问共享单访问硬件资源。这些操作还包括：第一计算机系统根据第一标志和跳转变量的状态访问共享单访问硬件资源，所述第一标志指示希望第一计算机系统访问共享单访问硬件资源。此外，这些操作可以还包括：第二计算机系统根据第二标志和跳转变量的状态访问共享单访问硬件资源，所述第二标志指示希望第二计算机系统访问共享单访问硬件资源。

[0018] 第三实施例还包括可选方面，可以单一地或以任何各种组合实施。根据第一可选方面，第一计算机系统位于PCIe结构的非透明桥的第一侧，第二计算机系统位于PCIe结构的非透明桥的第二侧。对于此可选方面，PCIe结构的非透明桥可以对经过其中的数据写入和数据读取进行地址转换。

[0019] 根据第三实施例的第二可选方面，第一计算机系统包括多个第一子计算机系统，所述多个第一子计算机系统位于PCIe结构的非透明桥的第一侧，第二计算机系统包括多个第二子计算机系统，所述多个第二子计算机系统位于PCIe结构的非透明桥的第二侧。对于此可选方面，PCIe结构的非透明桥可以对经过其中的数据写入和数据读取进行地址转换。

[0020] 根据第三实施例的第三可选方面，第一本地存储器存储多个标志，所述多个标志指示对应地希望多个第一子计算机系统访问共享单访问硬件资源，第二本地存储器存储多个标志，所述多个标志指示对应地希望多个第二子计算机系统访问共享单访问硬件资源。

[0021] 根据第三实施例的第四可选方面，共享单访问硬件资源是共享存储器。根据第三实施例的第五可选方面，第一计算机系统是第一网络集群，第二计算机系统是共享网络集群。

[0022] 根据下文结合附图以及权利要求书进行的详细描述,将更清楚地理解这些以及其它特征。

附图说明

[0023] 为了更完整地理解本公开,现在参考以下结合附图和具体实施方式进行的简要描述,其中相同参考标号表示相同部分。

[0024] 图1是示出根据第一所描述实施例的互连计算机系统的系统图。

[0025] 图2是示出根据第二所描述实施例的互连计算机系统的系统图。

[0026] 图3是示出根据第三所描述实施例的互连计算机系统的系统图。

[0027] 图4是示出根据第四所描述实施例的操作的流程图。

[0028] 图5是示出根据第五所描述实施例的操作的流程图。

[0029] 图6是示出根据第六所描述实施例的计算机系统的图。

具体实施方式

[0030] 图1是示出根据第一所描述实施例的互连计算机系统的系统图。互连计算机系统100包括快捷外围部件互连标准(Peripheral Component Interconnect Express,PCIe)结构102、以通信方式耦合到PCIe结构102的第一计算机系统106、以通信方式耦合到PCIe结构102的第二计算机系统108,以及耦合到PCIe结构的共享单访问硬件资源104。共享单访问硬件资源104是可以仅由单个线程/计算机系统一次性访问的部件,也可以是静态存储器、动态存储器磁存储器、通信接口或可以由单个线程/计算机系统访问而不会妨碍连贯一致的资源操作的另一类型的硬件资源中的一个或多个。

[0031] 第一计算机系统106包括至少第一处理器和耦合到第一处理器的第一存储器。图6中示出第一计算机系统106或第二计算机系统108的结构的实例并进行进一步描述。第一计算机系统106的存储器用于存储第一标志(标志[0])110,所述第一标志指示希望第一计算机系统106访问共享单访问硬件资源104。第二计算机系统包括至少第二处理器和耦合到第二处理器的第二存储器。第二存储器用于存储第二标志(标志[1])112,所述第二标志指示希望第二计算机系统访问共享单访问硬件资源104。第二计算机系统108的存储器还用于存储跳转变量114,所述跳转变量指示第一计算机系统106和第二计算机系统108中的哪个有权访问共享单访问硬件资源104。将参考图4进一步描述图1的互连计算机系统100的操作。

[0032] 图1的互连计算机系统100可以包括各种可选方面。根据第一可选方面,第一计算机系统106位于PCIe结构102的非透明桥的第一侧,第二计算机系统108位于PCIe结构102的非透明桥的第二侧。根据所述可选方面的变化形式,PCIe结构的非透明桥对经过其中的数据写入和数据读取进行地址转换。

[0033] 根据图1的互连计算机系统100的其它可选方面,第一计算机系统106包括多个第一子计算机系统,所述多个第一子计算机系统位于PCIe结构的非透明桥的第一侧,第二计算机系统108包括多个第二子计算机系统,所述多个第二子计算机系统位于PCIe结构的非透明桥的第二侧。根据所述可选方面的变化形式,PCIe结构的非透明桥对经过其中的数据写入和数据读取进行地址转换。

[0034] 与现有系统相比,图1的互连计算机系统100提供了重要的性能增益。互连计算机

系统100可以实现对Peterson锁算法的改进，与现有实施方案相比，所述改进提供了重要的操作益处。传统上，Peterson锁算法包括了远程写入标志变量，即存储在线程[1]上的标志[0]和存储在线程[0]上的标志[1]。由于不可靠的远程写入，这种实施方案在PCIe网络中效率低下，因此所述远程写入之后需要进行远程读取以确认所述远程写入成功。另外，当PCIe网络包括PCIe非透明桥时，地址转换操作引入了额外的操作难度。

[0035] 因此，为了克服先前算法的问题，将标志变量存储在本地以使与远程写入(remote write，RW)和远程读取(remote read，RR)相比，所述标志变量可通过本地写入(local write，LW)和本地读取(local read，LR)访问。对于现有系统，要使线程P0访问单访问硬件资源，需要进行以下操作：RW+RR(设置标志[0])、RW+RR(设置跳转)、LR+LR(忙锁)和RW+RR(重新设置标志[0])，总共6次远程操作和2次本地操作。对于图1的实施例，要使线程P0访问单访问硬件资源，需要进行以下操作：LW(设置标志[0])、RW(设置跳转)、RR(忙锁+确认RW)、LW(重新设置标志[0])，总共2次远程操作和2次本地操作。

[0036] 对于现有系统，要使线程P1访问单访问硬件资源，需要进行以下操作：RW+RR(设置标志[1])、LW(设置跳转)、LR+LR(忙锁)和RW+RR(重新设置标志[1])，总共4次远程操作和3次本地操作。对于图1的实施例，要使线程P1访问单访问硬件资源，需要进行以下操作：LW(设置标志[1]和跳转)、RR(忙锁)、LW(重新设置标志[1])，总共1次远程操作和3次本地操作。因此，对于先前系统，总共需要10次远程操作和5次本地操作。对于图1的实施例，总共需要3次远程操作和5次本地操作，省去了7次远程操作。

[0037] 图1的实施例100将PCIe写后读取与算法工作流相结合。对于非冲突请求，远程访问大大减少。因为大部分远程访问发生在忙等待循环中，所以图1的实施例允许使用循环中的简单sleep轻松控制。假设远程操作是工作流的主要成本，且假设两个线程的权重相等，则平均效率增益将使两个线程在1个占空比内将10次远程访问降低到3次。因此，实现时延减少70％或空载吞吐量增加超过200％。

[0038] 图2是示出根据第二所描述实施例的互连计算机系统的系统图。互连计算机系统200包括第一计算机系统208、第二计算机系统210、第一PCIe结构202、第二PCIe结构204、非透明PCIe桥206、第一共享单访问硬件资源212和第二共享单访问硬件资源214。

[0039] 如图所示，非透明PCIe桥206桥接第一PCIe结构202与第二PCIe结构204之间的PCIe通信。PCIe结构和非透明PCIe桥206的结构和操作通常是已知的。对于图2的互连计算机系统200，第一PCIe结构202具有第一地址域，第二PCIe结构204具有不同于第一地址域的第二地址域。非透明PCIe桥206解析经过其中的事务的地址，所述事务例如读取、写入等。因此，非透明PCIe桥206对标志变量和跳转变量的读取和写入进行地址转换，从而增加了这些事务的时延。另外，因为第一PCIe结构202和第二PCIe结构204在不同地址域中，所以确保变量的远程写入正确需要对应的远程读取。

[0040] 第一共享单访问硬件资源212耦合到第一PCIe结构202，而第二单访问硬件资源耦合到第二PCIe结构204。这些共享单访问硬件资源212和214中的每一个是可以仅由单个线程/计算机系统一次性访问的部件，也可以是静态存储器、动态存储器磁存储器、通信接口，或可以由单个线程/计算机系统访问而不会妨碍连贯一致的资源操作的另一类型的硬件资源中的一个或多个。

[0041] 第一计算机系统208包括至少第一处理器和耦合到第一处理器的第一存储器。图6

中示出第一计算机系统208的结构的实例并进行进一步描述。第一计算机系统208的存储器用于存储:标志(标志[0,0])216,所述标志指示希望第一计算机系统208访问第一共享单访问硬件资源212;和标志(标志[0,1])218,所述标志指示希望第一计算机系统208访问第二共享单访问硬件资源214。第一计算机系统208的存储器还用于存储跳转变量(跳转[0])220,所述跳转变量指示第一计算机系统212和第二计算机系统214中的哪个有权访问第一共享单访问硬件资源212。

[0042] 同样地,第二计算机系统210包括至少第二处理器和耦合到第二处理器的第二存储器。图6中示出第二计算机系统210的结构的实例并进行进一步描述。第二计算机系统210的存储器用于存储:标志(标志[1,0])222,所述标志指示希望第二计算机系统210访问第一共享单访问硬件资源212;和标志(标志[1,1])224,所述标志指示希望第二计算机系统210访问第二共享单访问硬件资源214。第二计算机系统210的存储器还用于存储跳转变量(跳转[1])226,所述跳转变量指示第一计算机系统208和第二计算机系统210中的哪个有权访问第二共享单访问硬件资源214。将参考图4和5进一步描述图1的互连计算机系统200的操作。

[0043] 根据图2的互连计算机系统200的其它可选方面,第一计算机系统208包括多个第一子计算机系统,第二计算机系统210包括多个第二子计算机系统。根据此可选方面,第一存储器存储多个标志,所述多个标志指示对应地希望多个第一子计算机系统访问共享单访问硬件资源,第二存储器存储多个标志,所述多个标志指示对应地希望多个第二子计算机系统访问共享单访问硬件资源。

[0044] 图3是示出根据第三所描述实施例的互连计算机系统的系统图。互连计算机系统300包括第一PCIe结构302、第二PCIe结构304、将第一PCIe结构302与第二PCIe结构304互连的非透明PCIe桥304、以通信方式耦合到第一PCIe结构302的第一节点集群316,以及以通信方式耦合到第二PCIe结构304的第二节点集群340。互连计算机系统还包括多[M]个共享单访问硬件资源312、308、310、……、314,每一共享单访问硬件资源耦合到第一PCIe结构302和第二PCIe结构304之一。

[0045] 第一节点集群316包括N个处理器,即处理器[0,1]318、处理器[0,2]320、……、处理器[0,N]322,以及服务于所述N个处理器的至少一个第一存储器。至少一个第一存储器用于存储M个标志,所述M个标志针对N个处理器中的每一个与M个共享单访问硬件资源312、308、310、……、314相对应,所述M个标志即标志[0,1]324、标志[0,2]326、……、标志[0,M]328,总共N×M个标志变量。标志[N,M]指示希望第N线程/处理器访问第M共享单次使用资源314。至少一个第一存储器进一步用于存储跳转变量330、332、……、334,所述跳转变量指示第一节点集群316和第二节点集群340中的哪个(或哪一线程/处理器)有权访问对应的共享单访问硬件资源312、308、310、……、314。替代地,跳转变量中的一些可以存储在第二节点集群340中。在一个具体实例中,第一节点集群316存储与耦合到第一PCIe结构302的共享单访问硬件资源312和308相对应的跳转变量,而第二节点集群340存储与耦合到第二PCIe结构304的共享单访问硬件资源310和314相对应的跳转变量。

[0046] 第二节点集群340包括N个处理器,即处理器[0,1]342、处理器[0,2]344、……、处理器[0,N]346,以及服务于所述N个处理器的至少一个第二存储器。至少一个第二存储器用于存储M个标志,所述M个标志针对N个处理器中的每一个与M个共享单访问硬件资源相对

应,所述M个标志即标志[1,1]348、标志[1,2]350、……、标志[1,M]352,总共N×M个标志变量。标志[N,M]指示希望第二节点集群340的第N线程/处理器访问第M共享单访问资源314。至少一个第二存储器可用于存储跳转变量330、332、……、334,所述跳转变量指示第二节点集群316和第二节点集群340中的哪个(或哪一线程/处理器)有权访问对应的共享单访问硬件资源312、308、310、……、314。

[0047] 与图2的互连计算机系统200的情况一样,非透明PCIe桥306对经过其中的数据写入和数据读取进行地址转换。因此,如果第一节点集群316的处理器访问共享单访问硬件资源308,则不需要进行地址转换,因为仅通过第一PCIe结构302进行访问。然而,如果第一节点集群316的处理器访问共享单访问资源310,则非透明PCIe桥306对事务进行地址转换,从而导致访问较慢且容易出现读取/写入错误。

[0048] 图4是示出根据第四所描述实施例的操作的流程图。图4的操作400可以与图1的互连计算机系统一致,所述互连计算机系统具有PCIe结构102、以通信方式耦合到PCIe结构102的第一计算机系统106、以通信方式耦合到PCIe结构102的第二计算机系统108,以及耦合到PCIe结构102的共享单访问硬件资源104。所述方法在第一计算机系统106和第二计算机系统108的空闲状态下开始(步骤402)。对于图4的操作400,第一计算机系统在第一本地存储器中存储第一标志,用于指示希望第一计算机系统访问共享单访问硬件资源。同样地,第二计算机系统存储:第二标志,用于指示希望第二计算机系统访问共享单访问硬件资源;和跳转变量,用于指示第一计算机系统和第二计算机系统中的哪个有权访问共享单访问硬件资源。

[0049] 从步骤402,第一线程可能希望访问共享单访问硬件(步骤404)。第一线程然后在LW操作中将标志[0]设置为真,在RW操作中将跳转设置为1(步骤406)。然后,第一线程在RR操作中读取标志[1],在RR操作中读取跳转(步骤408)。只要标志[1]=真且跳转=1,第一线程就保持在等待状态(步骤410)。一旦标志[1]=假(如由第二线程设置的)或跳转不等于1(如由第二线程或另一线程写入的),第一线程就进入临界区,所述第一线程在所述临界区中访问共享单访问硬件资源(步骤412)。然后,第一线程在本地写入操作中将标志[0]设置为假(步骤414)。第一线程还可以在步骤414将跳转设置为1。从步骤414,操作返回到步骤402。

[0050] 从步骤402,第二线程可能希望访问共享单访问硬件资源(步骤416)。然后,第二线程在LW操作中将标志[1]设置为真,在LW操作中将跳转设置为0(步骤418)。然后,第二线程在RR操作中读取标志[0],在LR操作中读取跳转(步骤420)。只要标志[0]=真且跳转=0,第二线程就保持在等待状态(步骤422)。一旦标志[0]=假(如由第一线程设置的)或跳转不等于0(如由第一线程或另一线程写入的),第二线程就进入临界区,所述第二线程在所述临界区中访问共享单访问硬件资源(步骤424)。然后,第二线程在本地写入操作中将标志[1]设置为假(步骤426)。第二线程还可以在步骤426将跳转设置为0。从步骤426,操作返回到步骤402。

[0051] 因此,通常且与步骤410一致,第一计算机系统根据第一标志和跳转变量的状态访问共享单访问硬件资源,所述第一标志指示希望第一计算机系统访问共享单访问硬件资源。同样地,第二计算机系统根据第二标志和跳转变量的状态访问共享单访问硬件资源,所述第二标志指示希望第二计算机系统访问共享单访问硬件资源。

[0052] 在各种方面中，第一计算机系统位于PCIe结构的非透明桥的第一侧，第二计算机系统位于PCIe结构的非透明桥的第二侧。因此，在各种方面中，操作400包括PCIe结构的非透明桥对经过其中的数据写入和数据读取进行地址转换。

[0053] 可根据图4执行的算法的一个实例如下：

[0054] 从P0本地存储器分配标志[0]

[0055] 在一个结构中，从P1本地存储器连续分配标志[1]和跳转。

[0056] 线程[0]执行以下操作：

[0057] P0:标志[0]＝真;//本地写入

[0058] 跳转＝1;//远程写入

[0059] 同时(标志[1]＝＝真&&跳转＝＝1)

[0060] //单次远程读取+确认

[0061] {

[0062] //忙等待

[0063] }

[0064] //临界区

[0065] ...

[0066] //临界区的末端

[0067] 标志[0]＝假;//本地写入

[0068] 线程[1]执行以下操作：

[0069] P1:标志[1]＝真;

[0070] 跳转＝0;//单次本地写入

[0071] 同时(标志[0]＝＝真&&跳转＝＝0)

[0072] //远程读取+本地读取

[0073] {

[0074] //忙等待

[0075] }

[0076] //临界区

[0077] ...

[0078] //临界区的末端

[0079] 标志[1]＝假;//本地写入

[0080] 图5是示出根据第五所描述实施例的操作的流程图。图5的操作500可与图3的互连节点集群一致。所述方法在第一节点集群316和第二节点集群340的空闲状态下开始(步骤502)。对于图5的操作500，第一节点集群316在第一本地存储器中存储标志，所述标志指示希望第一节点集群316的N个线程/处理器中的每一个访问M个共享单访问硬件资源中的每一个。同样地，第二节点集群340将标志以及跳转变量存储在第二本地存储器中，所述标志指示希望第二节点集群340的N个线程/处理器中的每一个访问M个共享单访问硬件资源中的每一个，所述跳转变量还指示第一节点集群316和第二节点集群340中的哪个(或其线程/处理器)有权访问M个共享单访问硬件资源。

[0081] 从步骤502，第一节点集群316或其线程/处理器可能希望访问第M共享单访问硬件

(步骤504)。然后,第一节点集群在LW操作中将标志[0,M]设置为真(步骤506),在LW操作中将跳转[M]设置为1(步骤507)。然后,第一节点集群在RR操作中读取标志[1,N,M],在RR(或LR)操作中读取跳转[M](步骤508)。只要对于第二节点集群340的至少一个线程/处理器的标志[1,N,M]=真且跳转[M]=1,第一节点集群316就保持在等待状态(步骤510)。一旦标志[1,N,M]=假(如由第二节点集群设置的)或跳转[M]不等于1(如由第二节点集群340或另一线程写入的),第一节点集群316就进入临界区,所述第一节点集群在所述临界区中访问第M共享单访问硬件资源(步骤512)。然后,第一节点集群在本地写入操作中将标志[0,M]设置为假(步骤514)。第一节点集群316还可以在步骤514将跳转[M]设置为1。从步骤514,操作返回到步骤502。

[0082] 从步骤502,第二(第N)节点集群340可能希望访问第M共享单访问硬件资源(步骤516)。然后,第二节点集群在LW操作中将标志[N,M]设置为真(步骤518),在RW操作中将跳转[M]设置为0(步骤519)。然后,第二节点集群在RR操作中读取标志[0],在LR操作中读取跳转(步骤520)。只要标志[0]=真且跳转[M]=0,第二(第N)节点集群就保持在等待状态(步骤522)。一旦标志[0]=假(如由第一节点集群设置的)或跳转[M]不等于0(如由第一节点集群或另一节点集群写入的),第二(第N)节点集群就进入临界区,所述第二节点集群在所述临界区中访问第M共享单访问硬件资源(步骤524)。然后,第二(第N)节点集群在本地写入操作中将标志[N,M]设置为假(步骤526)。第二(第N)节点集群还可以在步骤526根据节点集群中的哪个希望访问第M资源将跳转[M]设置为0(或设置为另一状态。从步骤526,操作返回到步骤502。

[0083] 在各种方面中,第一节点集群位于PCIe结构的非透明桥的第一侧,第二节点集群位于PCIe结构的非透明桥的第二侧。因此,在各种方面中,操作400包括PCIe结构的非透明桥对经过其中的数据写入和数据读取进行地址转换。

[0084] 图6是示出根据第六所描述实施例的计算机系统的图。计算机系统600包括处理电路604、存储器606、一个或多个用户接口608、射频(Radio Frequency,RF)接口610、近场通信(Near Field Communication,NFC)接口612、有线/光接口614和电池616。计算机系统600可以是图1所示的计算机系统106或108之一,图2所示的计算机系统208或210之一,或图3所示的节点集群316或340之一。

[0085] 处理电路604可以是微处理器、数字信号处理器、专用处理电路,和/或能够根据预编程指令或软件指令的执行而执行逻辑运算的其它电路中的一个或多个。存储器606可以是动态RAM、静态RAM、闪存RAM、ROM、EEPROM、可编程ROM、磁存储装置、光存储装置,或能够存储指令和数据的其它存储装置。存储的数据可以是NFC天线调谐数据、音频数据、视频数据、用户数据、软件指令、配置数据或其它数据。用户接口608支持视频监视器、键盘、音频接口或其它用户接口设备中的一个或多个。

[0086] RF接口610支持蜂窝式通信、WLAN通信、WPAN通信、WWAN通信、60GHz(MMW)通信、NFC通信和/或其它无线通信中的一个或多个。这些无线通信在大多数实施例中是标准化的且在其它实施例中是专用的。NFC接口612耦合到NFC天线618并支持NFC通信,如本文中将进一步描述的。有线/光接口614支持有线通信和/或支持光通信,所述有线通信可以是LAN通信、WAN通信、电缆网络通信、直接数据链路通信或其它有线通信,所述光通信在一些实施例中是标准化的且在其它实施例中是专用的。

[0087] 计算机系统600的部件604、606、608、610、612和614中的多个部件可以在单个集成电路管芯上构造。相当常见的是在单个集成电路上形成所有通信部件，例如RF接口610、NFC接口612和有线/光接口614。在一些情况下，甚至支持RF接口610的天线也可以在单个集成电路上形成。在其它情况下，计算机系统600的部件中的一些或全部可以在印刷电路板(Printed Circuit Board,PCB)上形成。

[0088] 根据本公开的实施例，存储器606存储标志324、326、……、328以及跳转变量330、332、……、334。存储器606还存储软件指令618和其它数据620。这些指令由处理电路604执行，且在所述处理电路上操作数据以实施图1、2和3的结构以及图4和5的操作。

[0089] 首先应理解，尽管下文提供一个或多个实施例的说明性实施方案，但所公开的系统和/或方法可以使用任何数目的技术来实施，无论所述技术是当前已知还是现有的。本公开决不应限于下文所示出的说明性实施方案、图式和技术，包括本文所示出并描述的示例性设计和实施方案，而是可以在所附权利要求书的范围以及其等效物的整个范围内修改。

[0090] 尽管本公开提供多个实施例，但应理解，所公开的系统和方法也可以在不脱离本公开的精神或范围的情况下以许多其它具体形式体现。本公开的实例应被视为说明性而非限制性的，且本公开并不限于本文本所给出的细节。例如，各种元件或部件可以在另一系统中组合或整合，或某些特征可以省去或不予以实施。

[0091] 另外，在不脱离本公开的范围的情况下，在各种实施例中描述和示出为独立或单独的技术、系统、子系统和方法可以与其它系统、模块、技术或方法组合或整合。示出或论述为彼此耦合或直接耦合或通信的其它项也可以采用电方式、机械方式或其它方式通过某一接口、设备或中间部件间接地耦合或通信。改变、替代及更改的其它实例可以由所属领域的技术人员确认，且可以在不脱离本文中所公开的精神和范围的情况下作出。

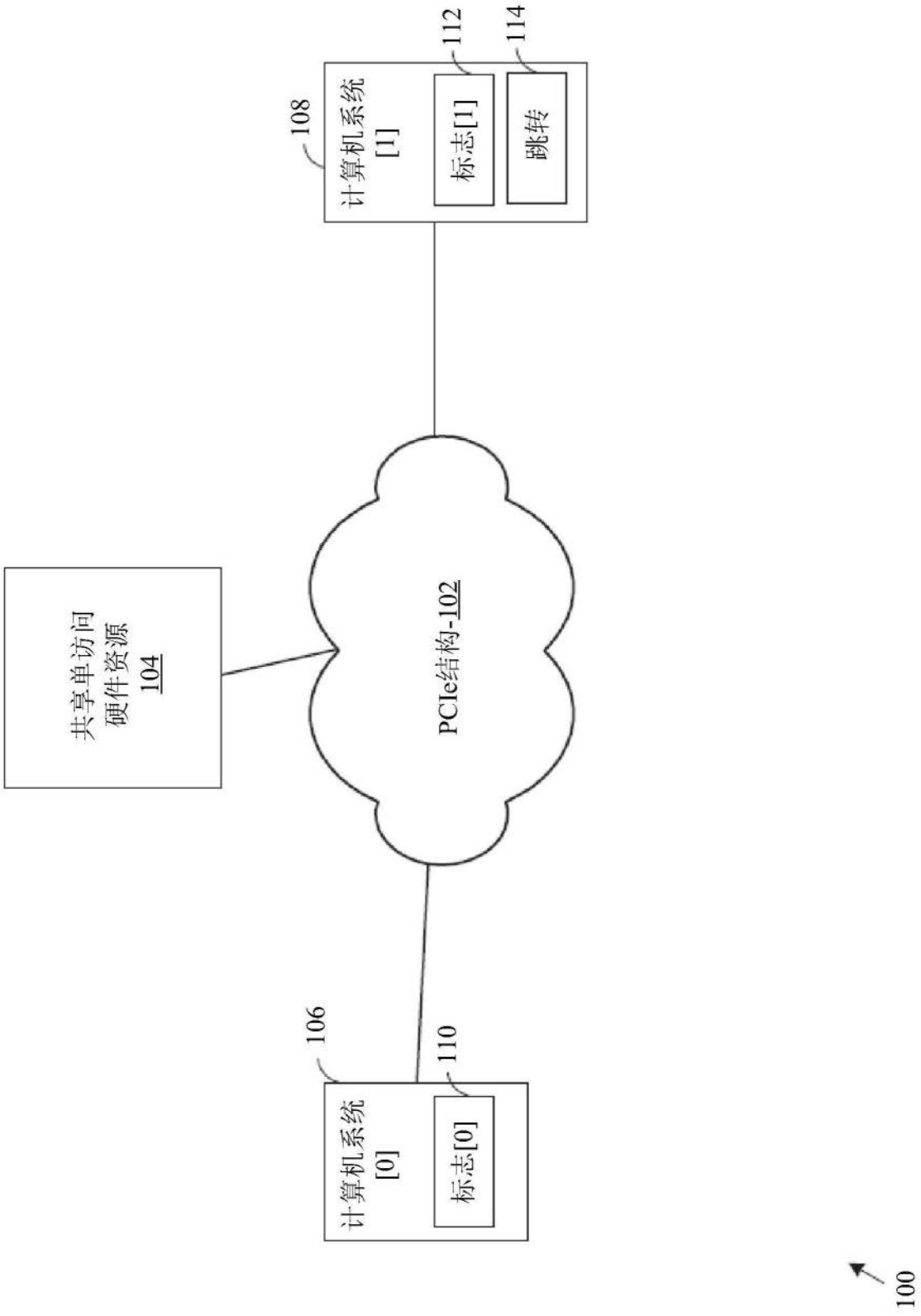

108
计算机系统
[1]
112
标志[1]
114
跳转
共享单访问
硬件资源
104
PCIe结构-102
106
计算机系统
[0]
110
标志[0]
100

图1

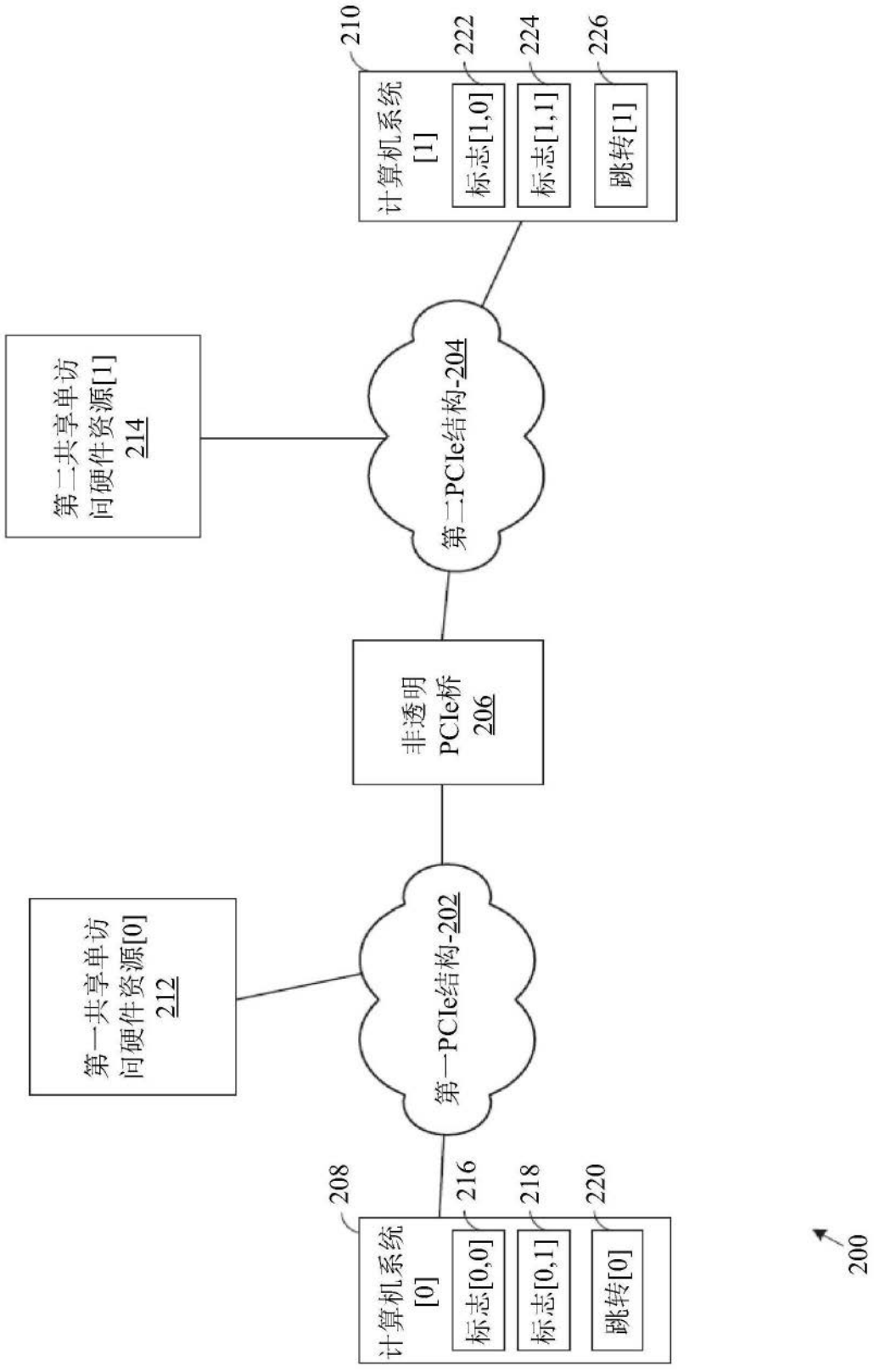

210
计算机系统
[1]
222
标志[1,0]
224
标志[1,1]
226
跳转[1]
第二共享单访
问硬件资源[1]
214
第二PCIe结构-204
非透明
PCIe桥
206
第一共享单访
问硬件资源[0]
212
第一PCIe结构-202
208
计算机系统
[0]
216
标志[0,0]
218
标志[0,1]
220
跳转[0]
200

图2

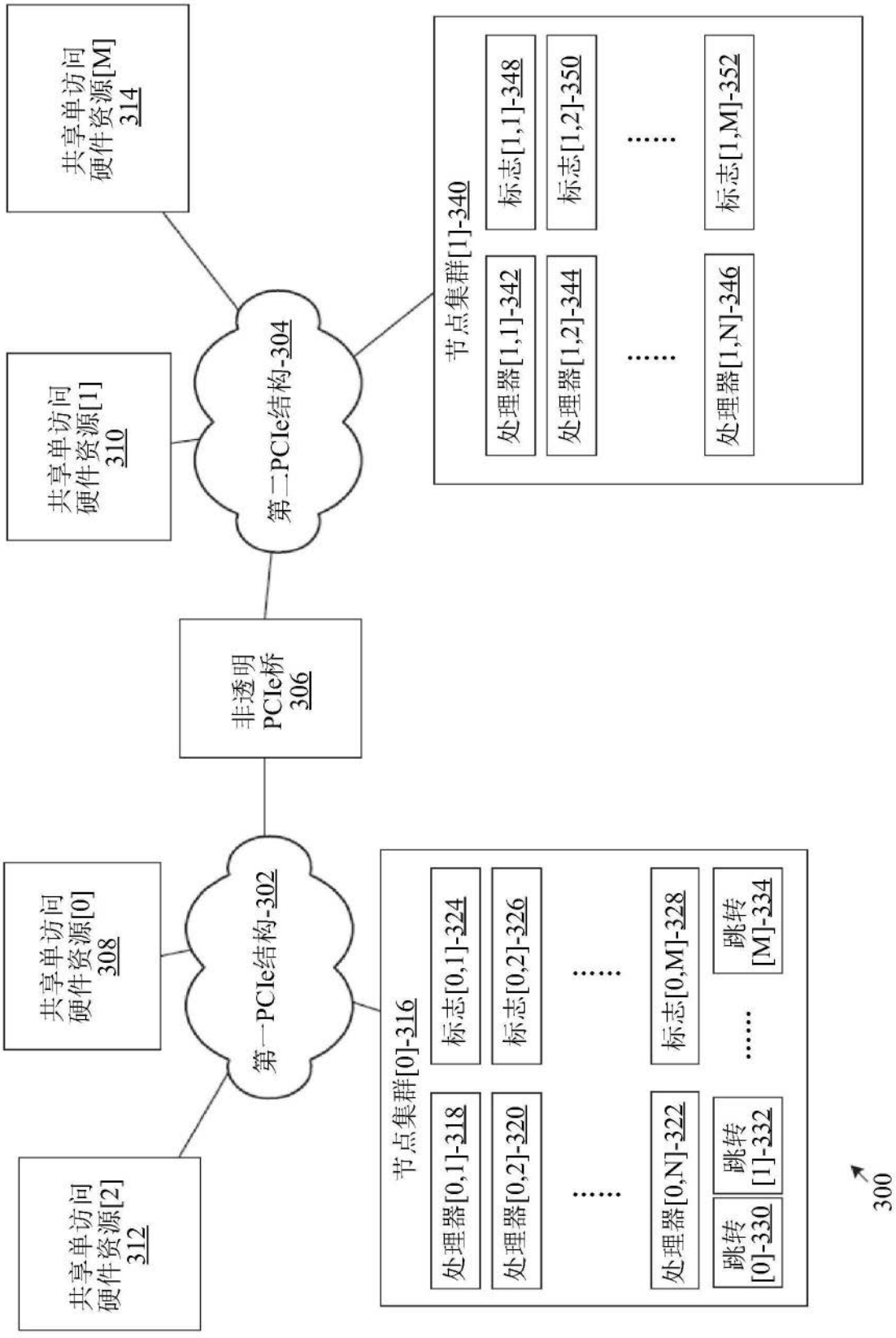

共享单访问硬件资源[M] 314
共享单访问硬件资源[1] 310
共享单访问硬件资源[2] 312
共享单访问硬件资源[0] 308
第二PCIe结构-304
第一PCIe结构-302
非透明PCIe桥 306
节点集群[1]-340
处理器[1,1]-342
处理器[1,2]-344
处理器[1,N]-346
标志[1,1]-348
标志[1,2]-350
标志[1,M]-352
节点集群[0]-316
处理器[0,1]-318
处理器[0,2]-320
处理器[0,N]-322
标志[0,1]-324
标志[0,2]-326
标志[0,M]-328
跳转[0]-330
跳转[1]-332
跳转[M]-334
300

图3

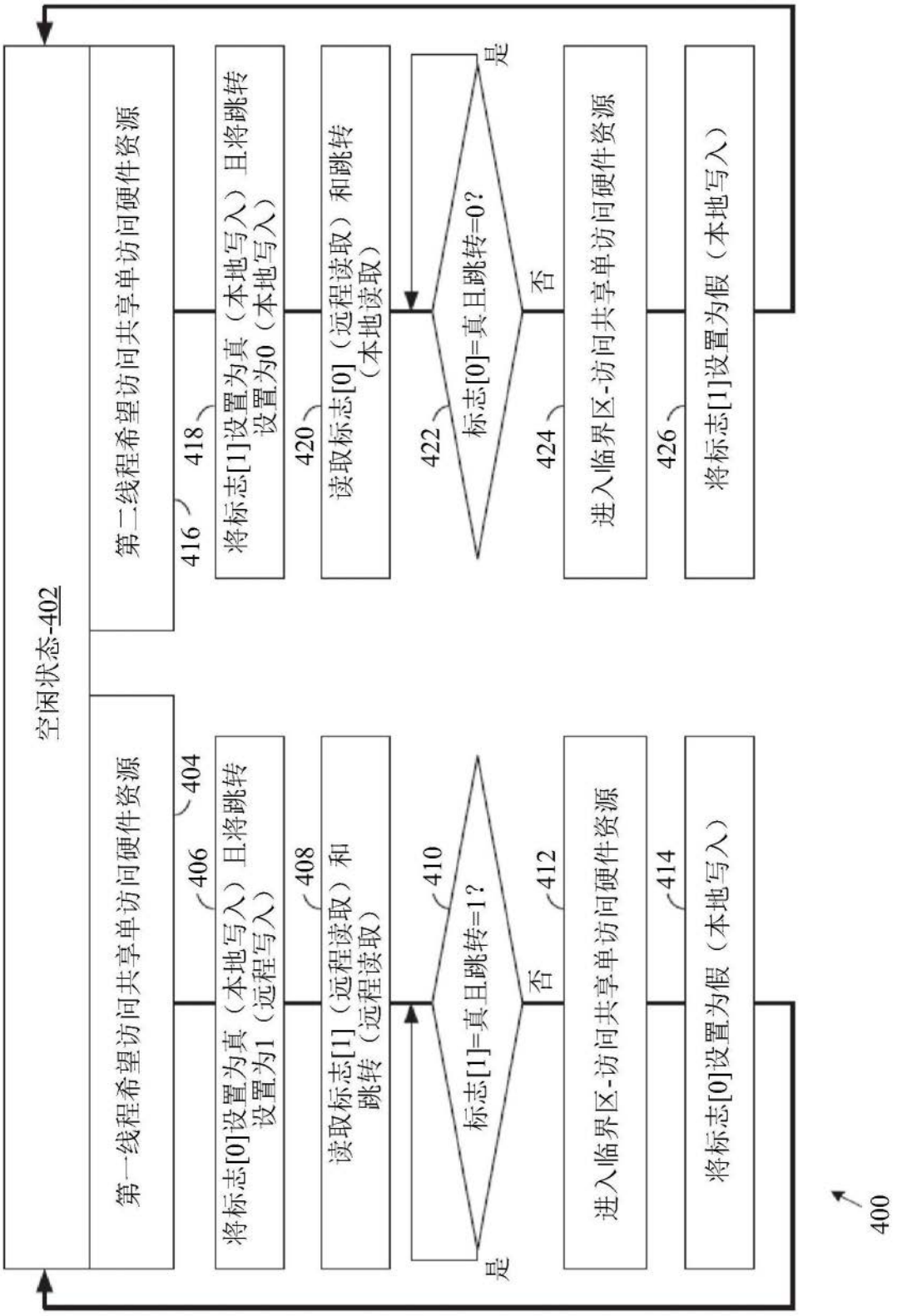

空闲状态-402
第一线程希望访问共享单访问硬件资源
404
将标志[0]设置为真（本地写入）且将跳转设置为1（远程写入）
406
读取标志[1]（远程读取）和跳转（远程读取）
408
标志[1]=真且跳转=1?
410
是
否
412
进入临界区-访问共享单访问硬件资源
414
将标志[0]设置为假（本地写入）
第二线程希望访问共享单访问硬件资源
416
将标志[1]设置为真（本地写入）且将跳转设置为0（本地写入）
418
读取标志[0]（远程读取）和跳转（本地读取）
420
标志[0]=真且跳转=0?
422
是
否
424
进入临界区-访问共享单访问硬件资源
426
将标志[1]设置为假（本地写入）
400

图4

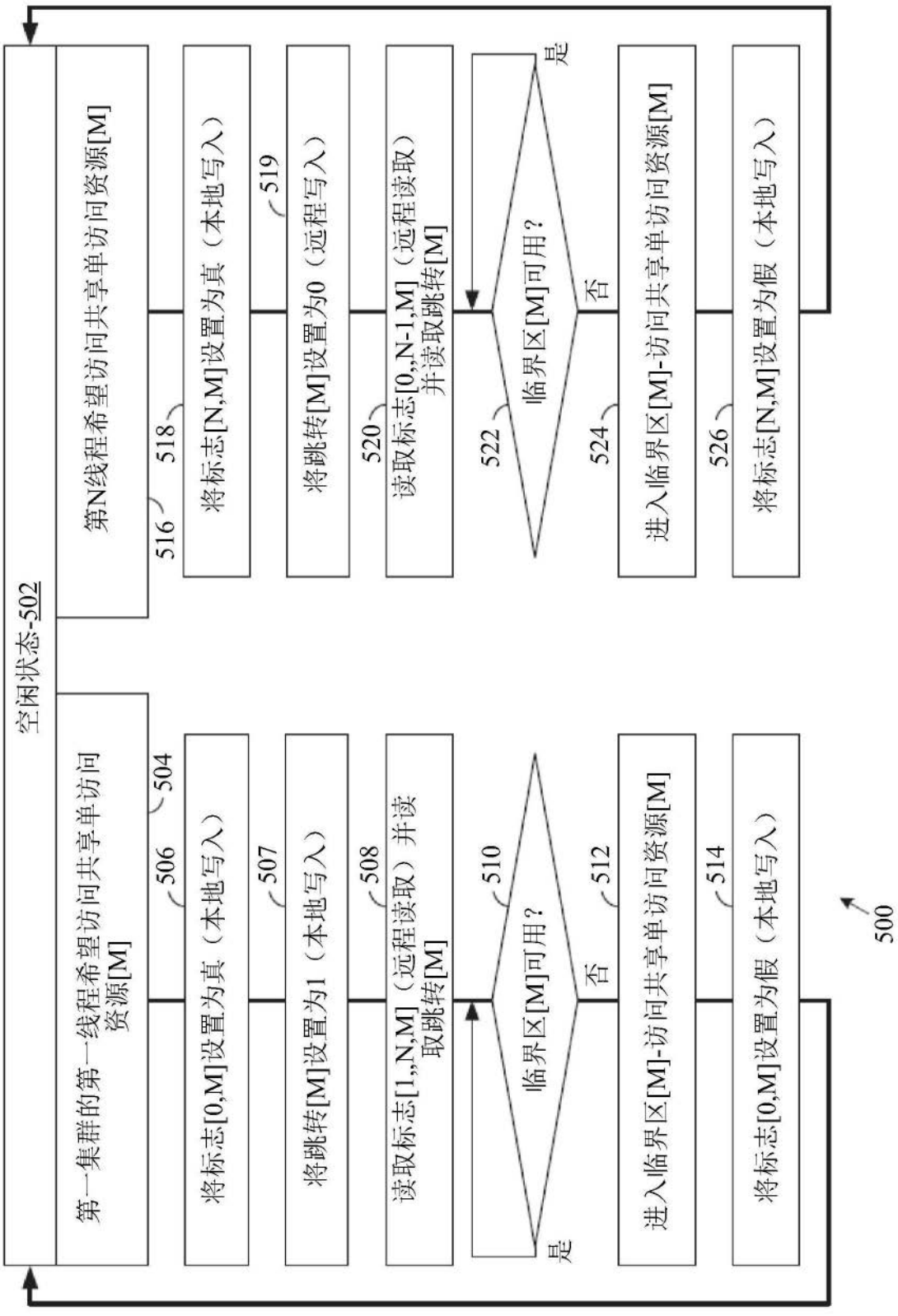

空闲状态-502
第一集群的第一线程希望访问共享单访问资源[M]
504
将标志[0,M]设置为真（本地写入）
506
将跳转[M]设置为1（本地写入）
507
读取标志[1,,N,M]（远程读取）并读取跳转[M]
508
临界区[M]可用？
510
是
否
512
进入临界区[M]-访问共享单访问资源[M]
514
将标志[0,M]设置为假（本地写入）
第N线程希望访问共享单访问资源[M]
516
将标志[N,M]设置为真（本地写入）
518
将跳转[M]设置为0（远程写入）
519
读取标志[0,,N-1,M]（远程读取）并读取跳转[M]
520
临界区[M]可用？
522
是
否
524
进入临界区[M]-访问共享单访问资源[M]
526
将标志[N,M]设置为假（本地写入）
500

图5

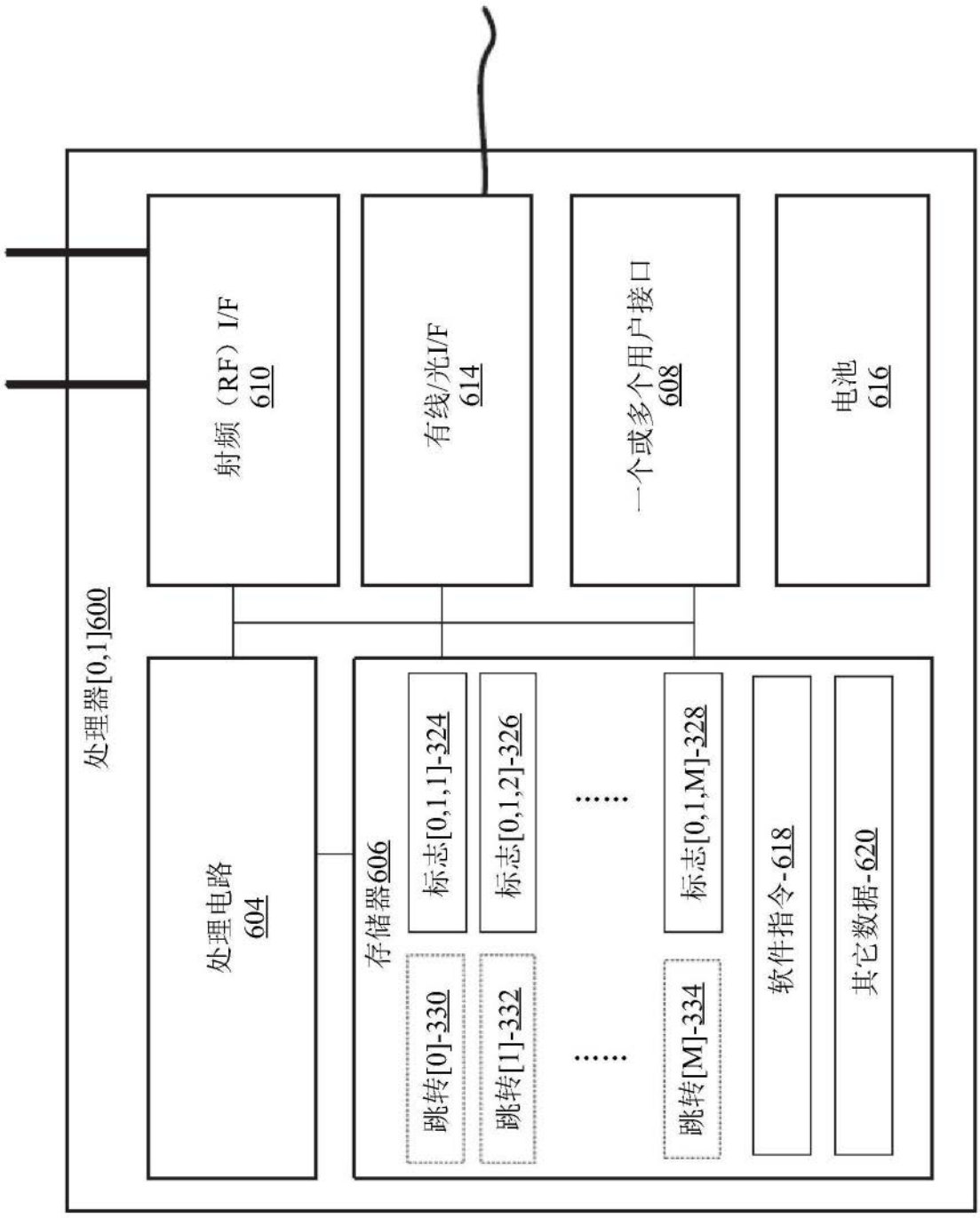

处理器[0,1]600
射频（RF）I/F
610
有线/光I/F
614
一个或多个用户接口
608
电池
616
处理电路
604
存储器606
标志[0,1,1]-324
标志[0,1,2]-326
标志[0,1,M]-328
跳转[0]-330
跳转[1]-332
跳转[M]-334
软件指令-618
其它数据-620

图6