(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 107430584 B
(45)授权公告日 2020.11.17

(21)申请号 201580078592.1
(22)申请日 2015.11.24
(65)同一申请的已公布的文献号
申请公布号 CN 107430584 A
(43)申请公布日 2017.12.01
(30)优先权数据
14/673,103 2015.03.30 US
(85)PCT国际申请进入国家阶段日
2017.09.30
(86)PCT国际申请的申请数据
PCT/US2015/062543 2015.11.24
(87)PCT国际申请的公布数据
WO2016/160070 EN 2016.10.06

(73)专利权人 伊姆西公司
地址 美国马萨诸塞州
(72)发明人 J·B·戴维斯
(74)专利代理机构 北京润平知识产权代理有限公司 11283
代理人 金旭鹏 肖冰滨

(51)Int.Cl.
G06F 15/173(2006.01)
H04L 29/08(2006.01)
审查员 于萍

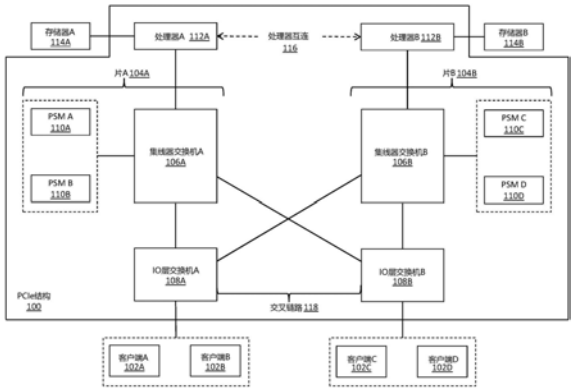
权利要求书2页 说明书8页 附图3页

(54)发明名称

经由具有完全连接网格拓扑的PCI EXPRESS结构从存储读取数据

(57)摘要

描述了用于从通信结构中的持久存储模块(“PSM”)读取数据的方法。读取请求可以经由处理器被提交给PSM。作为响应,所请求的数据可以被写入到客户端。读取完成可以通过通信结构沿着与数据相同的路径。



CN 107430584 B

1.一种用于从通信结构中的持久存储模块PSM读取数据的方法,该方法包括:

在处理器处接收来自客户端的持久存储模块PSM读取请求,其中所述PSM读取请求包括针对数据的逻辑地址;

在所述处理器处识别与所述逻辑地址相关联的PSM;

将所述PSM读取请求传送到与所述逻辑地址相关联的所述PSM;

将来自所述PSM的所述数据写入到所述客户端,其特征在于,写入所述数据包括:

通过与所述客户端相关联的下游IO层交换机传送所述数据,其中所述下游IO层交换机是网格通信结构的多层交换机之一,其中每一层交换机耦合到一个或多个客户端,其中所述处理器能够通过所述网格通信结构上的任意层交换机的与所述一个或多个客户端中的任意一者进行通信,其中所述多层交换机包括集线器层交换机和IO层交换机;

从所述PSM向所述处理器传送读取完成,其中所述读取完成在所述数据之后通过所述下游IO层交换机被传送;以及

从所述处理器向所述客户端传送所述读取完成。

2.根据权利要求1所述的方法,其中所述通信结构是包括完全连接的网格拓扑的PCIe结构。

3.根据权利要求1所述的方法,其中将来自所述PSM的所述数据写入到所述客户端包括直接存储器访问("DMA")写入。

4.根据权利要求1所述的方法,其中将所述PSM读取请求传送到所述PSM包括:

从所述处理器向所述PSM传送门铃;以及

在所述PSM处通过直接访问存储器("DMA")读取所述读取请求,所述读取请求来自与所述处理器相关联的存储器。

5.一种用于从通信结构中的持久存储模块PSM读取数据的设备,该设备包括:

用于在处理器处接收来自客户端的PSM读取请求的装置,其中所述PSM读取请求包括针对数据的逻辑地址;

用于在所述处理器处识别与所述逻辑地址相关联的PSM的装置;

用于与将所述PSM读取请求传送到与所述逻辑地址相关联的所述PSM的装置;

用于将来自所述PSM的所述数据写入到所述客户端的装置,其特征在于,写入所述数据包括通过与所述客户端相关联的下游IO层交换机传送所述数据,其中所述下游IO层交换机是网格通信结构的多层交换机之一,其中每一层交换机耦合到一个或多个客户端,其中所述处理器能够通过所述网格通信结构上的任意层交换机与所述一个或多个客户端中的任意一者进行通信,其中所述多层交换机包括集线器层交换机和IO层交换机;

用于从所述PSM向所述处理器传送读取完成的装置,其中所述读取完成在所述数据之后通过所述下游IO层交换机被传送;以及

用于从所述处理器向所述客户端传送所述读取完成的装置。

6.根据权利要求5所述的设备,其中所述通信结构是包括完全连接的网格拓扑的PCIe结构。

7.根据权利要求5所述的设备,其中将来自所述PSM的所述数据写入到所述客户端包括直接存储器访问("DMA")写入。

8.根据权利要求5所述的设备,其中将所述PSM读取请求传送到所述PSM包括:

用于从所述处理器向所述PSM传送门铃的装置;以及

用于在所述PSM处通过直接访问存储器(“DMA”)读取所述读取请求的装置,所述读取请求来自与所述处理器相关联的存储器。

9.一种用于从通信结构中的持久存储模块PSM读取数据的系统,该系统包括处理器,该处理器用于:

在所述处理器处接收来自客户端的PSM读取请求,其中所述PSM读取请求包括针对数据的逻辑地址;

在所述处理器处识别与所述逻辑地址相关联的PSM;

将所述PSM读取请求传送到与所述逻辑地址相关联的所述PSM;

将来自所述PSM的所述数据写入到所述客户端,其特征在于,写入所述数据包括通过与所述客户端相关联的下游IO层交换机传送所述数据,其中所述下游IO层交换机是网格通信结构的多层交换机之一,其中每一层交换机耦合到一个或多个客户端,其中所述处理器能够通过所述网格通信结构上的任意层交换机与所述一个或多个客户端中的任意一者进行通信,其中所述多层交换机包括集线器层交换机和IO层交换机;

从所述PSM向所述处理器传送读取完成,其中所述读取完成在所述数据之后通过所述下游IO层交换机被传送;以及

从所述处理器向所述客户端传送所述读取完成。

10.根据权利要求9所述的系统,其中所述通信结构是包括完全连接的网格拓扑的PCIe结构。

11.根据权利要求9所述的系统,其中将来自所述PSM的所述数据写入到所述客户端包括直接存储器访问(“DMA”)写入。

12.根据权利要求9所述的系统,其中将所述PSM读取请求传送到所述PSM包括:

从所述处理器向所述PSM传送门铃;以及

在所述PSM处通过直接访问存储器(“DMA”)读取所述读取请求,所述读取请求来自与所述处理器相关联的存储器。

经由具有完全连接网格拓扑的PCI EXPRESS结构从存储读取数据

[0001] 相关申请的交叉引用

[0002] 本申请涉及同时申请的共同未决的题为PCI EXPRESS FABRIC ROUTING FOR A FULLY-CONNECTED MESH TOPOLOGY的美国专利申请No.14/224,846和题为WRITING DATA TO STORAGE VIA A PCI EXPRESS FABRIC HAVING A FULLY-CONNECTED MESH TOPOLOGY的美国专利申请No.14/673,073,其通过引用的方式结合于此用于各种目的。

技术领域

[0003] 本发明一般涉及PCIe结构路由,且更特别地涉及用于在具有完全连接的网格拓扑的PCIe结构中从端点读取数据的系统和方法。

背景技术

[0004] 为了两个系统能够使得两个独立的系统通信,每一个系统需要包括足够的硬件和/或软件来使得这两个系统对接。

[0005] 从以下描述以及权利要求书中可以明白本发明的其他方面。

附图说明

[0006] 图1示出了根据本发明的一个或多个实施方式的包括PCIe结构的系统;

[0007] 图2示出了与本公开的实施方式保持一致的读取数据的方法;

[0008] 图3示出了用于与本公开的实施方式保持一致的写入数据的方法。

具体实施方式

[0009] 以下结合示出本发明原理的附图提供本发明的一个或多个实施方式的详细描述。虽然结合这些实施方式描述本发明,但是应当理解本发明不限于任意一个实施方式。相反,本发明的范围仅由权利要求书来限定且本发明包括许多替换、修改以及等同。处于示例的目的,在以下描述中提出了许多具体的细节以提供对本发明的全面理解。是处于示例的目的提供这些细节,且可以根据权利要求书在没有这些具体细节的一些或全部的情况下实施本发明。处于清楚的目的,与本发明相关的技术领域中公知的技术材料没有详细描述以避免不必要地使本发明晦涩。

[0010] 应当理解可以以许多方式来实施本发明,包括实施为过程、装置、系统、设备、方法或计算机可读介质,例如计算机可读存储介质或计算机网络,其中计算机程序指令通过光或电通信链路被发送。应用可以采用在通用计算机上执行的软件的形式或在硬件中硬线连接或硬编码。在本说明书中,这些实施或本发明可以采用的任意其他形式可以称为技术。一般来说,公开的过程的步骤的顺序可以在本发明的范围内改变。

[0011] 将参考以被配置成存储文件的存储系统的形式的数据存储系统来描述本发明的实施方式,但是应当理解本发明的原理不限于这种配置。而是,其可应用于能够以模拟、数

字或其他形式存储并处理各种类型的对象的任意系统。虽然诸如文档、文件、对象等的术语可以以示例的方式被使用，但是本发明的原理不限于任意特定形式的代表并存储数据或其他信息；而是，它们同等地可应用能够代表信息的任意对象。

[0012] 一般来说，本发明的实施方式涉及PCIe结构，其包括至少两层交换机，被连接以形成至少两层之间的完全连接的网格拓扑。此外，本发明的实施方式涉及PCIe结构，其使得连接到PCIe结构中的一个片(slice)的客户端执行对在PCIe结构的不同片中的存储器、网络端点设备和/或持久存储的操作例如(读和写操作)。

[0013] 在典型的通信结构(例如以太网或无限带宽)中，通过将唯一端点地址与每个端点设备相关联，并将端点地址指定为请求的部分来在通信结构中的交换机间路由请求。在PCIe结构中，基于被访问的存储器地址而不是端点地址在结构交换机之间路由读和写请求操作。结果，用于完全连接的网格的基于地址的路由的典型布置不允许所有客户端访问所有端点设备。本发明的一个或多个实施方式提供能够克服这一局限的机制。具体地，基于地址的路由可以用于实现所有客户端访问所有端点的层之间的完全连接的网格(下文所述)。

[0014] 在本发明的一个或多个实施方式中，PCIe结构中的组件通信和/或实施外部部件互联Express(PCIe)标准。本发明的实施方式不限于PCIe标准的任何过去、当前或将来的版本。此外，本发明的实施方式可以用与用于实施本发明各种实施方式的PCIe标准的特征类似的特征的其他标准来实施。

[0015] 本公开还提出用于通过PCIe结构执行I/O操作的方法和过程。由于该结构存在多条路径，因此当客户端尝试读取或写入数据时可以发生竞争条件。例如，客户端可以在数据被完全传输到客户端之前接收“读取完成”指示。类似地，可以存在写竞争条件，其中PSM和/或客户端在数据被完全写入到PSM或其他端点之前接收“写入完成”指示。这种问题可能在数据正被写入或从多个PSM读取的时候加剧。如本文所述，这些竞争条件可以通过使得某些系统传输沿着与数据相同的PCIe结构的通信路径来避免。

[0016] 图1示出了根据本发明的一个或多个实施方式的包括PCIe结构的系统。PCIe结构(100)由两个或更多个片(104A，104B)组成，其中每个片直接连接到一个处理器(112A，112B)和一个或多个客户端(102A-102D)。下面描述之前提到的组件中的每一个。

[0017] 在本发明的一个实施方式中，每一个客户端(102A-102D)是物理设备，其包括处理器(或另一类型的处理组件)、存储器以及用于使其连接到PCIe结构(100)的物理接口。此外，每个客户端包括实施本发明的一个或多个实施方式所需的用于实施PCIe标准(或其部分)的功能。客户端还包括用于发送和/或接收事务层分组(TLP)的功能。TLP对应于根据PCIe标准被定义的一种类型的分组。在本发明的一个实施方式中，TLP使得客户端能够从PCIe结构读取数据并向PCIe结构写入数据。例如另一种方式，TLP使得客户端能够将数据传输到PCIe结构中的位置并从PCIe结构中的位置传输数据。在本发明的一个实施方式中，客户端的一个或多个操作为PCIe端点，即，发起事务的设备和/或是事务的目标的设备。每一个客户端可以经由链路(即，客户端与PCIe结构之间的物理连接)连接到PCIe结构。

[0018] 继续图1的描述，每个片(104A，104B)包括输入/输出(IO)层交换机(ITS)(108A，108B)、集线器层交换机(HTS)(106A、106B)以及一个或多个持久存储模块(PSM)。下面描述这些组件的每一个。

[0019] 关于ITS，每个ITS是连接到一个或多个客户端(102A-102D)的物理PCIe交换机。每

个ITS还连接到ITS所位于的相同片中的HTS。此外，每个ITS可以连接到与ITS所位于的片不同的片中的一个或多个HTS。在本发明的一个实施方式中，每个ITS连接到PCI结构中的每一个HTS，由此产生PCIe结构中层之间的完全连接的网格。在不偏离本发明的情况下可以不需要层间完全连接的网格来实施本发明的实施方式。

[0020] 在本发明的一个实施方式中，每个ITS被配置成：(i)从其连接的客户端接收TLP并使用地址路由(例如存储器地址路由)将TLP路由到ITS上的合适的出口端口(上游端口或下游端口中的一个下游端口)，以及(ii)从ITS连接的一个或多个HTS接收TLP并使用地址路由将TLP路由到ITS上的合适的出口端口(典型地下游端口)。例如，在图1中，ITS B(108B)可以从客户端C(102A)、客户端D(102D)、HTS A(106A)以及HTS B(106B)接收TLP。

[0021] 关于HTS，每个HTS是连接到一个或多个ITS(108A-108B)以及一个或多个持久存储模块(PSM(110A-110D))的物理PCIe交换机。每个HTS连接到HTS所位于的相同片中的ITS。此外，每个HTS可以连接到与HTS所位于的片不同的片中的零个或多个ITS。在本发明的一个实施方式中，每个HTS连接到PCIe结构中的每一个其他ITS由此产生PCIe结构中层间的完全连接的网格。每个HTS还可以经由其根端口(未示出)连接到处理器。在不偏离本发明的情况下可以不需要层间完全连接的网格来实施本发明的实施方式。

[0022] 在本发明的一个实施方式中，每个HTS被配置成：(i)从其连接的持久存储模块(PSM)接收TLP并使用地址路由将TLP路由到HTS上的合适的出口端口(典型地下游端口)以及(ii)从HTS连接的一个或多个ITS接收TLP并使用地址路由将TLP路由到HTS上的合适的出口端口(上游端口和/或下游端口中的一个或多个下游端口)。例如，在图1中，HTS B(106B)可以从PSM C(110C)、PSM D(110D)、ITS A(108A)和ITS B(108B)接收TLP。下面参考图3提供关于HTS的另外的细节。

[0023] 在本发明的一个实施方式中，持久存储模块(100A-110D)的每一个包括持久存储(未示出)以及可选地包括易失性存储器(未示出)(例如动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM以及DDR SDRAM)。持久存储可以包括但不限于NAND闪存、NOR闪存、磁性RAM存储器(M-RAM)、转矩磁性RAM存储器(ST-MRAM)、相变存储器(PCM)、记忆存储器、被定义为非易失性存储级存储器(SCM)的任意其他存储器、磁盘以及光盘。本领域技术人员可以理解本发明的实施方式不限于存储级存储器。在本发明的一个实施方式中，PSM的每一个是仅一个片的部分。

[0024] 继续PCIe结构的讨论，如上所述，PCIe结构中的每一个片直接连接到至少一个处理器(112A、112B)。每个处理器是具有被配置成执行指令的单核或被配置成执行指令的多核的电路组。可以使用复杂指令集(CISC)架构或精简指令集(RISC)架构来实施处理器。在本发明的一个或多个实施方式中，处理器包括根复合体(如由PCIe标准定义的)(未示出)。根复合体将处理器连接到至少一个片以及存储器(114A、114B)(例如，动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM以及DDR SDRAM)，其可经由PCIe结构被访问但不是PCIe架构中的任意片的部分。

[0025] 在本发明的一个实施方式中，PCIe结构内的处理器(112A、112B)能够使用例如处理器互连(116)(例如英特尔QuickPath互连、英特尔前端总线或AMD HyperTransport)直接通信。本领域技术人员可以理解在不偏离本发明的情况下其他端到端通信机制可以用于允许处理器(112A、112B)之间的直接通信。

[0026] 本发明不限于图1中示出的系统。

[0027] 虽然图1示出了客户端连接到ITS和PSM连接到HTS，本发明的实施方式可以被实施为客户端连接到HTS以及PSM连接到ITS。

[0028] 在本发明的另一实施方式中，PCIe结构可以被实施由此其不包括任何PSM；而是ITS和HTS都连接到分开的客户端集合，其中PCIe结构促进客户端之间的通信。

[0029] 此外，虽然图1中示出的PCIe结构仅包括两个片、两个处理器以及四个PSM，但是在不偏离本发明的情况下PCIe结构可以被实施具有更少或更多的上述组件中的每一者。此外，虽然图1中的PCIe结构连接到四个客户端和两个存储器，但是在不偏离本发明的情况下本发明的实施方式可以被实施以使得PCIe结构能够连接更少或更多数量的客户端和/或存储器。

[0030] 此外，虽然关于包括存储(例如PSM(110A-110D))的PCIe结构描述了本发明的实施方式，但是本发明的实施方式可以被实施为使得任意两个设备能够使用PCIe结构进行通信。例如，在本发明的一个实施方式中，图1中示出的客户端可以是刀片式服务器，其中刀片式服务器不包括任何物理NIC卡且PSM可以用网络端点设备来替换。

[0031] 在该示例中，网络端点设备是被配置成使用互联协议与网络(即，有线网络、无线网络或其组合)对接以及经由PCIe与PCIe结构对接的设备。网络端点设备的示例是PICe NIC卡。网络端点设备的每一个可以包括持久存储(如以上关于PSM所述的)和存储端点设备存储器(例如动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM、DDR SDRAM或任意其他类型的易失性存储器)的组合。

[0032] 继续该示例，PCIe结构可以使得刀片式服务器能够与一个或多个网络端点设备通信。该实施方式可以允许刀片式服务器有效率地共享一个或多个网络端点设备。本发明不限于该示例。

[0033] 在另一示例中，在本发明的一个或多个实施方式中，PSM可以用存储端点设备(即，包括从客户端存储数据和服务读取以及写入请求的功能的设备)来替换。存储端点设备的每一个可以包括持久存储(如以上关于PSM所述的)和存储端点设备存储器(例如动态随机存取存储器(DRAM)、同步DRAM、SDR SDRAM、DDR SDRAM或任意其他类型的易失性存储器)的组合。存储端点设备的示例是存储设备。本发明不限于该示例。

[0034] 此外，本发明的实施方式可以被扩展到包括经由PCIe结构的两个或更多个设备通信。在一般情况中，PSM(图1中示出的)可以被一般化为目标设备，其中目标设备可以包括PSM、网络端点设备、存储端点设备或能够使用PCIe通信的任意其他设备。

[0035] 虽然图1中的PCIe结构已经被示出包括PSM(或更一般地目标设备)，但是PCIe结构可以被实施为由此其不包括目标设备；而是，PCIe结构仅包括用于连接到目标设备的必要的物理组件。

[0036] 现在参考图2，描述了用于从类似图1的系统中的PSM读取数据的过程。可以在处理器(例如处理器A或处理器B(112A、112B)处从客户端(例如客户端(102A-D))接收该读取请求。该请求可以针对位于系统中的一个或多个PSM(例如PSM(110A-D))的数据。在一些实施方式中，可以使用组播通过系统传送请求和所有其他传输，如在交叉引用的专利文献中记载的。在一些实施方式中，可以使用如上所述的TLP通过系统路由传输。

[0037] 在一个实施方式中，可以从一个或多个PSM读取数据并将其传回客户端。例如，响

应于来自客户端C(102C)的读取请求,可以从PSM A(110A)读取数据并将其传送到该客户端。一旦从PSM读取数据,PSM A可以发送完成到处理器(例如处理器A(112A)和/或处理器B(112B)),而其可以依次将读取完成传送到客户端(在该情况中是客户端C)。该读取完成可以向客户端指示数据已经从PSM被传送且不期望更多的数据。但是,在数据仍然在传输途中时(即,在数据仍然正通过PCIe结构从PSM传送时)存在客户端会接收到读取完成的可能性。例如,数据在到达客户端C(102C)之前可能流过集线器交换机A(106A)和IO层交换机B(108B)。但是来自PSM的读取完成可能在下传通过集线器交换机B(106B)和IO层交换机B(108B)之前流到处理器A(112A)和/或处理器B(112B)。如果线路上存在等待时间,例如在交叉链路(118)处存在等待时间,则读取完成可能在任意或所有数据之前到达客户端C(102C)。因此客户端可能认为其具有所有的读取数据,而实际上其一些仍然正流过该结构。图2中示出的方法解决这一竞争条件。

[0038] 在框200,可以在处理器处接收来自客户端的PSM读取请求。例如,客户端C(102C)可以通过IO层交换机B(108B)和集线器交换机B(106B)发送读取请求到处理器B(112B)。在一个实施方式中,PSM读取请求可以包括针对客户端希望读取的数据的逻辑地址。

[0039] 在框202,处理器(例如处理器B(112B))可以识别与客户端提供的逻辑地址相关联的PSM。在一些实施方式中,处理器可以查询存储器内数据结构以将逻辑地址解析成物理地址。存储器内数据结构可以是位于处理器存储器(例如存储器A(114A)或存储器B(114B))中的逻辑至物理地址映射。该识别的物理地址可以对应于系统中的数据的一个或多个位置。例如,物理地址可以识别包含所有或部分的数据的一个或多个PSM中的一个或多个区。

[0040] 在框204,可以从处理器向一个或多个物理位置传送门铃。例如,可以向PSM A(110A)传送门铃。在一些实施方式中,可以向多个PSM传送门铃,多个PSM可以在相同片上或不同片上。例如,可以向片A(104A)上的PSM A(110A)和片B(104B)上的PSM D(110D)传送门铃。这例如在所请求的数据的部分位于不同PSM的情况下是有利的。

[0041] 响应于门铃,一个或多个PSM可以从处理器存储器读取物理位置数据(在206)。在一些实施方式中,读取物理位置数据包括DMA读取请求。该读取请求可以包括PSM在PSM上物理定位数据所需的地址信息。

[0042] 可以存在针对框204和206的另外和/或可替换过程。例如,处理器可以直接将读取请求写入到PSM存储器。在一些实施方式中,读取请求可以被传送到PSM,带有标签(例如设置比特、标志和/或指示符,指示其是新请求)。此外或可替换地,可以实施硬件和/或软件FIFO队列,其中PSM知道队列中的一切是新请求。

[0043] 在一些实施方式中,所请求的数据可以位于PSM的两个位置中的至少一个。第一,数据可以在非持久PSM存储器中,例如拱形存储器(vaulted memory)。拱形存储器可以是例如属于PSM的DRAM。第二,数据可以在PSM的存储模块中,例如闪存模块,用于持久存储。在一个实施方式中,从拱形存储器读取数据可以提供更好的系统性能。

[0044] 一旦数据位于PSM中,其可以通过与客户端相关联的下游IO层交换机被写入到客户端(在框208)。例如,如果客户端C(102C)请求了数据,则其可以通过与客户端通信的IO层交换机B(108B)被写入到客户端。如果从PSM A(110A)读取数据,则该数据必须还经过集线器交换机A(106A)和将集线器交换机106A与IO层交换机B(108B)连接的交叉链路。

[0045] 在一些实施方式中,将数据写入到客户端包括DMA写入。例如,可以从PSM的拱形存

储器和/或存储模块将数据写入到客户端的存储器。在一些实施方式中，这可以通过位于客户端、PSM和/或处理器处的DMA引擎来促进。

[0046] 在框210，一旦数据从该结构被写出，则读取完成指示可以从一个或多个PSM被传送到处理器。如果PSM和客户端在同一个片中，则读取完成指示可以直接被传送到与该片相关联的处理器。但是，如果PSM和客户端在不同的片中，则读取完成指示可以通过与数据相同的下游IO层交换机传送。在本示例中，该下游IO层交换机可以是IO层交换机B(108B)。通过与客户端相关联的下游IO层交换机传送数据确保读取完成指示不会由于交叉链路上的等待时间而在数据之前到达客户端和/或处理器。这是因为读取完成跟随数据沿着相同的路径，且在被路由到处理器之前几乎到同一个端点。

[0047] 例如，在从IO层交换机B(108B)到集线器交换机B(106B)并最终到处理器B(112B)的上游传送之前，读取完成指示在数据之后通过集线器交换机A(106A)和IO层交换机B(108B)被传送。这可以与直接向处理器(例如通过集线器交换机A(106A)向处理器A(112A)并到处理器B(112B))传送读取完成指示形成对比。此外或可替换地，读取完成指示符可以被传送到处理器A(112A)，处理器A然后可以在数据之后将该指示发回交换机A(106A)和IO层交换机B(108B)。

[0048] 最后，在框212，读取完成指示可以从处理器被传送到客户端。例如，一旦处理器B(112B)已经从PSM接收到读取完成指示，则其可以向下游将该指示(或类似的通知)传回客户端。如果从多个PSM读取数据，则处理器可以等待传送通知直到其已经从所有PSM接收到完成。例如，如果响应于读取请求从PSM A(110A)和PSM D(110D)两者读取数据，则处理器可以等待传送读取完成指示直到其已经从PSM A(110A)和PSM D(110D)两者接收到类似的指示。这允许处理器在告知客户端读取完成之前确保所有数据确实已经被传送到客户端。

[0049] 现在参考图3，描述了用于避免包括完全连接的网格拓扑的PCIe结构中的写入竞争条件的方法。在一些实施方式中，写入竞争条件类似于上述的读取竞争条件。客户端可能希望将数据写入到连接到PCIe结构的一个或多个PSM。客户端可以传送写入请求到处理器，例如处理器A(112A)或处理器B(112B)。处理器可以用PSM位置来响应，以及客户端可以开始数据传输。一旦客户端已经将其数据写入到线路，其可以将写入完成发回处理器。处理器然后可以在PSM处设置指示写入完成的门铃。PSM然后可以对从客户端接收的数据进行非持久存储。但是，在一些实施方式中，可能在数据仍然在从客户端的传输过程中接收到门铃。例如，门铃可以沿着不同于数据的路径，且因此可以在数据之前到达PSM。作为响应，PSM可以在数据的其余部分到达之前对任意量的接收的数据进行非持久存储。图3中示出的过程可以防止这种竞争条件。

[0050] 在框300，处理器可以接收来自客户端的指示客户具有其希望写入到PSM的数据的写入请求。例如，处理器B(112B)可以从客户端C(102C)接收写入请求。在一些实施方式中，该写入请求可以经由位于客户端的DMA引擎被接收。DMA引擎可以使得客户端能够直接向存储器处理器(例如存储器B(114B))写入或从处理器存储器读取。此外或可替换地，客户端处理器将写入请求传送到处理器。这对于不期望有使用DMA引擎造成的开销的较小分组来说是有利的。

[0051] 响应于写入请求，处理器可以识别与一个或多个PSM相关联的一个或多个写入位置。例如，处理器可以识别应写入数据的物理地址。这些物理地址可以是在同一个PSM内或

可以在多个PSM中。在一些实施方式中，将数据写入到多个PSM可以在PSM故障的情况下提供冗余。在本示例中，处理器B(112B)可以将PSM A(110A)识别为用于数据的位置。

[0052] 在框304，处理器可以将指示符(例如比特、标志或指示符)写入到客户端DMA引擎。该比特可以指示在处理器存储器中有一个或多个数据写入位置可用。该数据写入位置可以是在框302中被识别的物理位置。例如，该比特可以向客户端通知在处理器B(112B)的存储器(114B)中有一个PSM位置可用。

[0053] 在框306，客户端可以从处理器存储器读取写入位置。该读取可以经由位于客户端的DMA引擎发生。一旦客户端已经从处理器存储器接收到PSM位置，则其可以开始经由PCIe结构将数据写入到PSM。

[0054] 在框308，PSM可以接收从客户端写入的数据。在一个实施方式中，该数据可以通过与客户端相关联的下游IO层交换机端口被接收。例如，数据可以从客户端C(102C)流过IO层交换机B(108)并经由集线器交换机A(106A)到PSM A(110A)。在一个实施方式中，在每个IO层交换机和/或集线器交换机处使用组播群组来通过PCIe结构路由数据。在其中在多个PSM处识别多个物理地址的实施方式中，数据可以通过与客户端相关联的下游IO层交换机流到两个或更多个PSM。

[0055] 在一些实施方式中，在PSM处接收的数据可以被写入到PSM的拱形存储器。该拱形存储器可以是例如与PSM相关联的DRAM或MRAM。该拱形存储器还可以是另一形式的非易失性/持久或非持久存储器。PSM可以将数据存储到拱形存储器中直到其从处理器接收到对数据进行长期持久存储的指示。

[0056] 一旦客户端已经完成将数据写入到PCIe结构，其可以向处理器传送写入完成通知。在框310，处理器可以从客户端接收该通知。例如，客户端C(102C)可以一旦其已经完成将数据写入到结构就向处理器B(112B)传送写入完成通知。该写入完成通知可以在数据仍然在往PSM的传输途中被传送和/或接收。在一些实施方式中，写入完成通知是客户端DMA写入，例如MWr(“存储器写入”)PCIe TLP。

[0057] 一旦从客户端接收写入完成，则处理器可以将分开的写入完成传送到接收来自客户端的数据的任意PSM。例如，如果PSM A(110A)接收到数据，则处理器可以将写入完成传送到该PSM。如果多个PSM接收到数据，则可以将写入完成发送到其中的每一个PSM。

[0058] 从处理器向PSM传送的写入完成可以流过与写入数据的客户端相关联的下游IO层交换机。例如，写入完成在经过交叉链路到集线器交换机A(106A)再到PSM A(110A)之前可以流到与客户端C(102C)相关联的IO层交换机B(108B)。使得写入完成经过与客户端相关联的IO层交换机确保其沿着与数据相同的路径到PSM。这可以避免上述的竞争条件，因为写入完成将在数据之后到达而不是提前到达。

[0059] 在框314，一旦接收到写入完成，则数据标签可以从处理器被传输到PSM。处理器可以为每一个物理位置创建一个数据标签，并将标签传送到每一个物理位置。在一个实施方式中，该数据标签包括对象标识符(例如逻辑单元标识符(“LUN”))和偏移。在一个实施方式中，数据标签和其构成成分可以识别数据的位置。标签可以被存储在存储器内数据结构中并与逻辑地址相关联，其在一个实施方式中是对象标识符。该逻辑地址可以用于在读取请求期间访问数据，如上面详细描述的。在一些实施方式中，将标签传送到PSM包括从PSM将DMA读取到处理器存储器。该读取可以响应于接收到写入完成指示而被执行。

[0060] 在一些实施方式中，一旦PSM从处理器接收到标签，其可以将标签传输完成通知传送回处理器。这可以向处理器指示在继续到框316之前在PSM处接收到标签。

[0061] 最后，在框316，处理器将系统写入完成传送到客户端。一旦处理器已经将标签传送到已经接收到数据的PSM的每一个，则该处理器可以传送该系统写入完成。该系统写入通知可以向客户端通知写入过程完成以及数据现在在PSM中可用。

[0062] 为了清楚起见，使用特定的流程示出了本申请的过程和方法，但是应当理解其他顺序是可能的且可以并行执行一些，而不偏离本发明的实质。此外，可以细分或组合步骤。如本申请所公开，根据本发明写的软件可以以某种计算机可读介质的形式存储，例如存储器或CD-ROM，或通过网络传送，以及由处理器执行。

[0063] 本申请中所有的引用意在通过引用的方式结合于此。虽然以上关于特定实施方式描述了本发明，但是可以理解对本发明的变形和修改对于本领域技术人员来说是明显的且可以在权利要求书的范围及其等同范围内被实施。可以使用多于一个计算机，例如通过并行使用多个计算机或在多个计算机间加载共享布置或分配任务，由此作为整体它们执行本文确定的组件的功能，即它们取代单个计算机。上述的各种功能可以通过在单个计算机或在一些计算机上分配的单个过程或一组过程来执行。过程可以调用其他过程来处理某些任务。单个存储设备可以被使用，或一些存储设备可以用于替代单个存储设备。公开的实施方式是示例性的而非限制性的，且本发明不限于本申请中给出的细节。有实施本发明的许多可替换方式。因此，本公开和权利要求书旨在被解释为覆盖落入本发明的实质和范围内的所有这样的变形和修改。

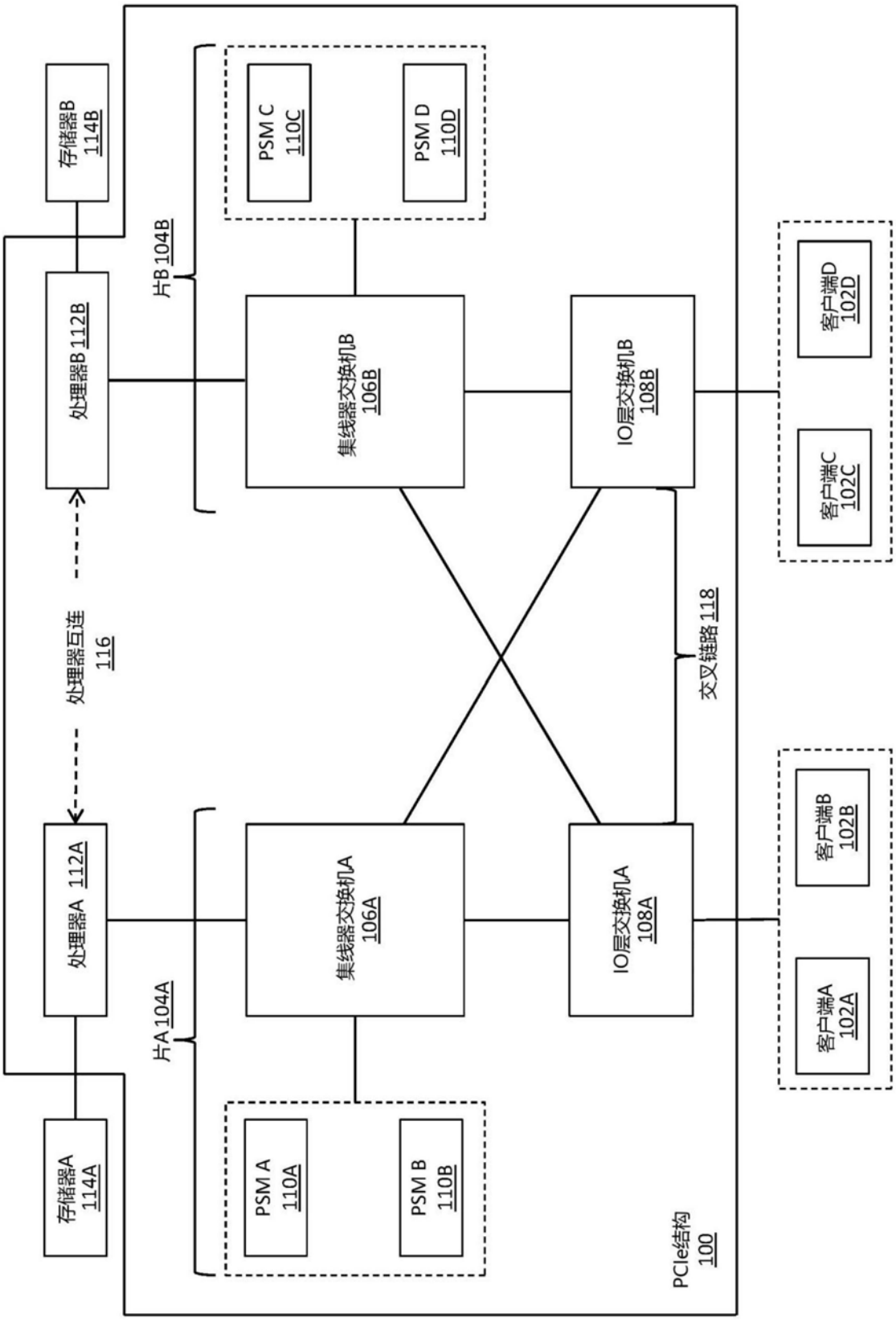

存储器B 114B
PSM C 110C
PSM D 110D
片B 104B
处理器B 112B
集线器交换机B 106B
IO层交换机B 108B
客户端D 102D
客户端C 102C
处理器互连 116
交叉链路 118
处理器A 112A
集线器交换机A 106A
IO层交换机A 108A
客户端B 102B
客户端A 102A
片A 104A
存储器A 114A
PSM A 110A
PSM B 110B
PCIe结构 100

图1

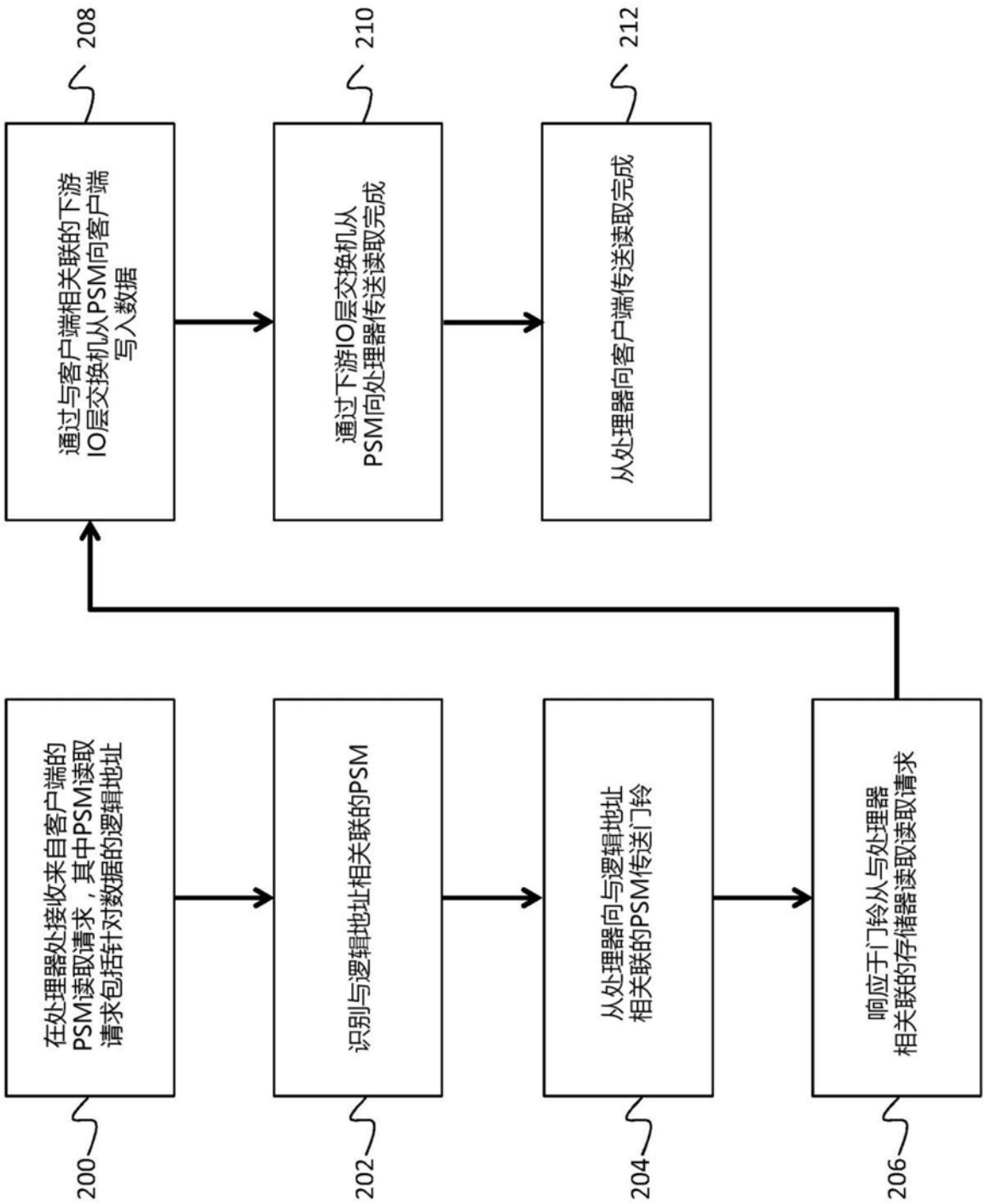

在处理器处接收来自客户端的PSM读取请求，其中PSM读取请求包括针对数据的逻辑地址
200
识别与逻辑地址相关联的PSM
202
从处理器向与逻辑地址相关联的PSM传送门铃
204
响应于门铃从与处理器相关联的存储器读取读取请求
206
通过与客户端相关联的下游IO层交换机从PSM向客户端写入数据
208
通过下游IO层交换机从PSM向处理器传送读取完成
210
从处理器向客户端传送读取完成
212

图2

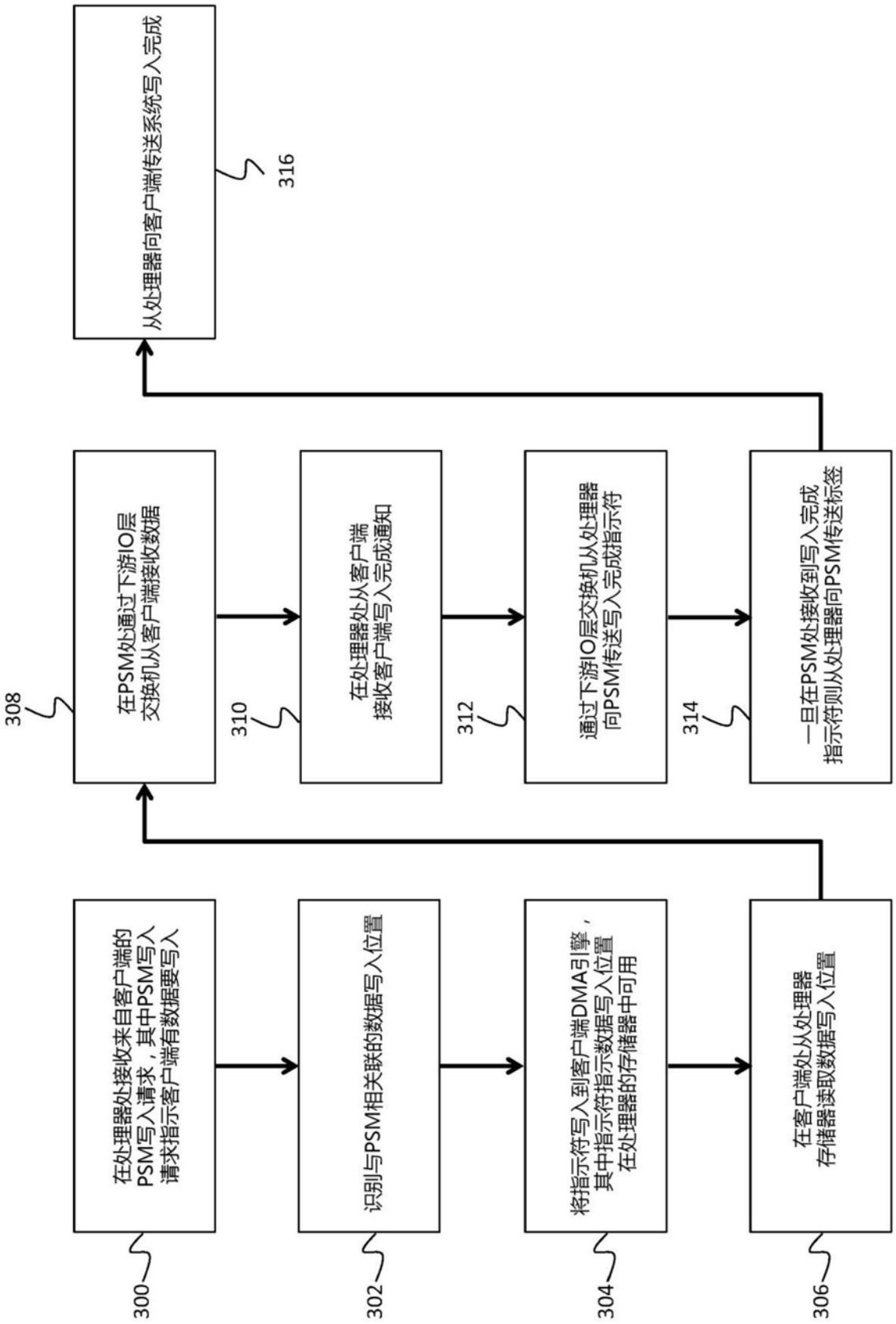

在处理器处接收来自客户端的PSM写入请求，其中PSM写入请求指示客户端有数据要写入
300
识别与PSM相关联的数据写入位置
302
将指示符写入到客户端DMA引擎，其中指示符指示数据写入位置在处理器的存储器中可用
304
在客户端处从处理器存储器读取数据写入位置
306
一旦在PSM处接收到写入完成指示符则从处理器向PSM传送标签
314
通过下游IO层交换机从处理器向PSM传送写入完成指示符
312
在处理器处从客户端接收客户端写入完成通知
310
在PSM处通过下游IO层交换机从客户端接收数据
308
从处理器向客户端传送系统写入完成
316

图3