



(12) 发明专利申请

(10) 申请公布号 CN 117377943 A

(43) 申请公布日 2024. 01. 09

(21) 申请号 202280025112.5

(22) 申请日 2022.03.29

(30) 优先权数据

17/217,792 2021.03.30 US

(85) PCT国际申请进入国家阶段日

2023.09.26

(86) PCT国际申请的申请数据

PCT/US2022/022357 2022.03.29

(87) PCT国际申请的公布数据

W02022/212383 EN 2022.10.06

(71) 申请人 超威半导体公司

地址 美国加利福尼亚州

(72) 发明人 穆罕默德·安布尔·哈桑

迈克尔·L·朱 阿斯温·阿吉

(74) 专利代理机构 上海胜康律师事务所 31263

专利代理师 樊英如 童礼翎

(51) Int.Cl.

G06F 9/30 (2006.01)

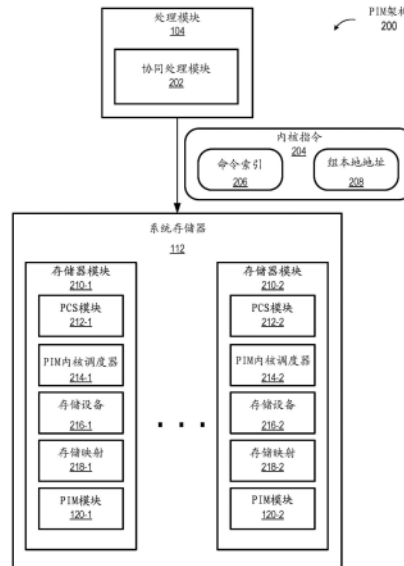
权利要求书2页 说明书8页 附图5页

(54) 发明名称

存算一体化并行处理系统和方法

(57) 摘要

本发明公开了一种处理系统(100),其包括处理模块(104,106)和存储器设备(112)。该存储器设备包括存算一体化(PIM)模块(120),该PIM模块代表该处理模块执行处理操作。该PIM模块的指令集架构(ISA)具有比该处理模块的ISA少的指令。从该处理模块接收的指令被翻译,使得该PIM模块的处理资源被虚拟化。因此,该PIM模块同时执行该处理模块的多个线程或应用的处理操作。



1. 一种存储器设备,包括:

多个存储设备,所述多个存储设备被配置为代表在处理单元处执行的程序来存储数据,其中所述处理单元在所述存储器设备外部;和

存算一体化 (PIM) 模块,所述PIM模块被配置为响应于与来自所述处理单元的所述程序相对应的内核指令而处理PIM命令,其中包括所述PIM命令且由所述PIM模块实施的指令集架构 (ISA) 具有比由所述处理单元实施的ISA少的指令,并且其中所述PIM模块的处理资源被配置为被虚拟化使得所述PIM模块同时处理来自多个PIM内核的PIM命令。

2. 根据权利要求1所述的存储器设备,还包括:

PIM命令存储 (PCS) 模块,所述PCS模块被配置为存储所述PIM命令并且向所述PIM模块提供所述PIM命令,其中所述PIM模块的所述处理资源包括所述PCS模块。

3. 根据权利要求2所述的存储器设备,还包括:

PIM内核调度器,所述PIM内核调度器被配置为组合与至少两个PIM内核相对应的至少两个内核指令的部分,并且在所述PCS模块处存储与所述至少两个内核指令相对应的单个PIM命令。

4. 根据权利要求3所述的存储器设备,其中:

所述PIM内核调度器被进一步配置为基于组合度量来组合所述至少两个内核指令。

5. 根据权利要求4所述的存储器设备,其中:

所述组合度量基于以下中的至少一者:在所述PIM模块处等待执行的PIM内核之间的各个内核指令的共性;基于对对应PIM内核的资源使用的预期的内核指令的加权组合;或者从所述处理单元接收的所述PIM内核的依赖图。

6. 根据权利要求3所述的存储器设备,其中:

所述多个存储设备还包括被配置为寄存器组的寄存器文件;并且

所述PIM内核调度器被进一步配置为将由所述内核指令指示的虚拟寄存器地址翻译成所述寄存器文件的物理寄存器地址。

7. 根据权利要求3所述的存储器设备,其中:

所述多个存储设备还包括被配置为寄存器组的寄存器文件;并且

所述PIM模块的所述处理资源包括所述寄存器文件,并且其中所述寄存器文件的至少一些寄存器被配置为被分配给所述至少两个PIM内核中的第一PIM内核,而所述寄存器文件的所述寄存器中的至少一些被分配给所述至少两个PIM内核中的第二PIM内核。

8. 根据权利要求3所述的存储器设备,其中:

所述PIM内核调度器被进一步配置为基于来自所述处理单元的所述内核指令来调度所述PIM命令以便在所述PIM模块处执行。

9. 一种方法,包括:

在存算一体化 (PIM) 算术逻辑单元 (ALU) 处从处理单元接收用于将由所述PIM ALU同时处理的多个内核的内核指令,其中由所述PIM ALU实施的指令集架构 (ISA) 具有比由所述处理单元实施的ISA少的指令;

从PIM命令存储 (PCS) 模块中检索与所述内核指令相对应的PIM ALU命令,其中至少一个PIM指令对应于至少两个内核;以及

执行所述PIM ALU命令包括将所述内核中的至少两个的数据存储在包括所述PIM ALU的存储器的寄存器组中。

10. 根据权利要求9所述的方法, 其中:

所述内核指令包括指示相应PIM ALU命令的相应命令索引和至少一个相应组本地地址。

11. 根据权利要求9所述的方法, 还包括:

在从所述PCS模块检索所述PIM ALU命令之前, 响应于检测到由所述内核指令指示的所述PIM ALU命令中的至少两个请求相同的操作, 将对所述至少两个PIM ALU命令的指示存储在所述PCS模块的单个条目处。

12. 根据权利要求9所述的方法, 其中:

检索所述PIM ALU命令包括基于对应PIM内核的依赖性来延迟所述PIM ALU命令中的至少一个的执行。

13. 一种系统, 包括:

处理单元, 所述处理单元被配置为执行多个程序; 和

存储器设备, 所述存储器设备被配置为从所述处理单元接收内核指令, 其中所述内核指令致使所述存储器设备的存算一体化 (PIM) 模块同时处理用于多个PIM内核的PIM命令, 其中同时处理PIM命令包括将用于所述PIM内核中的至少一些的数据存储在所述存储器设备的被配置为充当虚拟寄存器组的多个存储设备中, 并且其中包括所述PIM命令且由所述PIM模块实施的指令集架构 (ISA) 具有比由所述处理单元实施的ISA少的指令。

14. 根据权利要求13所述的系统, 其中:

所述存储器设备包括PIM命令存储 (PCS) 模块, 所述PCS模块被配置为存储所述PIM命令并且向所述PIM模块提供所述PIM命令; 并且

所述处理单元包括PIM内核调度器的一部分, 所述PIM内核调度器的所述部分将应用地址翻译成PCS地址并将所述PCS地址作为所述内核指令的一部分发送。

15. 根据权利要求13所述的系统, 其中:

所述存储器设备包括第二PIM模块, 所述第二PIM模块被配置为同时处理用于第二多个PIM内核的第二PIM命令。

16. 根据权利要求13所述的系统, 其中:

所述处理单元包括被配置为向所述存储器设备发送所述内核指令的协处理器。

存算一体化并行处理系统和方法

背景技术

[0001] 处理系统通常包括可由处理器(例如,中央处理单元(CPU)或图形处理单元(GPU))相对快速地存取的本地存储器资源以及花费相对较长时间量来存取的外部存储器资源。在一些情况下,外部存储器资源位于与处理器不同的芯片上。然而,在一些情况下,由于较长的存取时间,涉及存储于外部存储器资源处的数据的指令花费不期望的长时间量来处理。另外,在一些情况下,外部存储器资源与处理器之间的数据通信消耗不期望量的互连电容。在一些情况下,为了减少对外部存储器资源的存取次数,使用存算一体化(PIM)电路来处理外部存储器资源处的数据,而非将数据带到处理器。PIM电路对存储在外部存储器资源处的数据执行有限的处理应用,而无需较通用的处理器的功率消耗或面积使用。因此,在一些情况下,减少了处理器与外部存储器资源之间的通信次数且缩短了涉及存储在外部存储器资源处的数据的一些指令的处理时间。然而,由于此有限架构,利用硬件并行性来同时处理多个内核是困难的。例如,由于虚拟化的缺乏,PIM电路一次仅加载一个内核的命令,在一些情况下由于数据依赖性而产生串行化瓶颈。

发明内容

[0002] 在一个总体方面,一种存储器设备包括:多个存储设备,该多个存储设备被配置为代表在处理单元处执行的程序来存储数据,其中该处理单元在该存储器设备外部;和存算一体化(PIM)模块,该PIM模块被配置为响应于与来自该处理单元的该程序相对应的内核指令而处理PIM命令,其中包括该PIM命令且由该PIM模块实施的指令集架构(ISA)具有比由该处理单元实施的ISA少的指令,并且其中该PIM模块的处理资源被配置为被虚拟化使得该PIM模块同时处理来自多个PIM内核的PIM命令。

[0003] 根据另一总体方面,一种方法包括在存算一体化(PIM)算术逻辑单元(ALU)处从处理单元接收用于将由该PIM ALU同时处理的多个内核的内核指令,其中由该PIM ALU实施的指令集架构(ISA)具有比由该处理单元实施的ISA少的指令。该方法还包括从PIM命令存储(PCS)模块中检索与该内核指令相对应的PIM ALU命令,其中至少一个PIM指令对应于至少两个内核。该方法还包括执行该PIM ALU命令包括将该内核中的至少两个的数据存储在包括该PIM ALU的存储器的寄存器组中。

[0004] 根据又一总体方面,一种系统包括被配置为执行多个程序的处理单元。该系统还包括存储器设备,该存储器设备被配置为从该处理单元接收内核指令,其中该内核指令致使该存储器设备的存算一体化(PIM)模块同时处理用于多个PIM内核的PIM命令,其中同时处理PIM命令包括将用于该PIM内核中的至少一些的数据存储在该存储器设备的被配置为充当虚拟寄存器组的多个存储设备中,并且其中包括该PIM命令且由该PIM模块实施的指令集架构(ISA)具有比由该处理单元实施的ISA少的指令。

附图说明

[0005] 通过参考附图,本公开被更好地理解,并且其许多特征和优点对于本领域技术人

员是显而易见的。在不同附图中使用相同的附图标记表示类似或相同的项目。

[0006] 图1是根据一些实施方案的PIM并行处理系统的框图。

[0007] 图2是根据一些实施方案的PIM架构的框图。

[0008] 图3是根据一些实施方案的示例性虚拟到物理内核寄存器映射的框图。

[0009] 图4是根据一些实施方案的示例性内核指令合并和映射的框图。

[0010] 图5是根据一些实施方案的使用PIM并行处理系统来同时处理用于多个内核的内核指令的方法的流程图。

具体实施方式

[0011] 一种处理系统包括处理单元和在处理单元外部的存储器设备。存储器设备包括代表处理系统处理数据的存算一体化 (PIM) 模块。PIM模块的指令集架构 (ISA) 小于处理单元的ISA,使得在一些情况下,PIM命令是处理单元的ISA的指令的子集。例如,在一些情况下,PIM模块是算术逻辑单元 (ALU)。因此,与诸如CPU的处理器将指令发送到诸如GPU的另一处理器相比,将附加指令发送到PIM模块。在一些实施方案中,附加指令作为指示要执行的PIM命令的内核指令来发送。响应于内核指令,PIM模块使用存储在存储器设备处的数据来处理PIM命令。

[0012] 在一些情况下,PIM模块同时处理多个应用或同一应用的不同线程的PIM命令。地址虚拟化通常用于防止由于多个应用尝试使用单个地址而导致的错误,并且减少由于应用由于物理地址的使用而接收关于其他应用的信息而导致的潜在安全问题。然而,与处理单元相比,由于PIM模块相对简单,因此在一些PIM模块中难以进行地址虚拟化。因此,那些PIM模块连续地给予对资源(例如,PIM模块和诸如寄存器文件之类的相关联的存储设备)的访问。在一些实施方案中,即时存储器设备还包括PIM内核调度器,该PIM内核调度器代表PIM模块执行虚拟到物理寄存器映射,从而实现存储器设备处的寄存器地址虚拟化,且使得PIM模块能够同时处理多个内核。

[0013] PIM命令被存储在PIM命令存储 (PCS) 模块处。然而,在一些实施方案中,PCS模块中的存储空间是有限的。因此,在一些情况下,PIM内核调度器进一步分析传入内核指令以识别多组指令是否请求相同的操作。响应于检测到请求相同的操作,PIM内核调度器组合内核指令,使得PCS模块仅存储单个PIM命令。因此,增加了PCS模块的资源利用。此外,在一些实施方案中,PCS地址被虚拟化,使得多个内核的指令同时被存储在PCS模块中。因此,在一些情况下,响应于检测到多个内核请求相同的操作,PIM内核调度器组合来自该多个内核的内核指令。

[0014] 如本文中所使用,“同时处理”是指应用或内核的至少一部分在开始之后但在另一应用或内核的一部分完成之前的执行。例如,如果以交错方式执行两个内核的部分使得第二内核的一部分在第一内核的第一部分之后但在第一内核的第二部分之前执行,那么认为该第一内核和该第二内核“被同时处理”。作为另一示例,其中第一内核的一部分在流水线的第二部分中且第二内核的一部分在流水线的第二部分中的流水线处理器被认为“同时处理”该两个内核。这两个应用或内核不需要由完全相同的硬件设备在完全相同的时间处理才被认为是“同时处理的”。

[0015] 在不同的实施方案中,本文所述的技术使用多种并行处理器(例如,矢量处理器、

图形处理单元 (GPU)、通用GPU (GPGPU)、非标量处理器、高度并行处理器、人工智能 (AI) 处理器、推断引擎、机器学习处理器、其他多线程处理单元等) 中的任一者得以采用。为了便于说明, 本文参考其中采用处理模块的示例性系统和方法。然而, 应当理解, 除非另外指出, 否则本文中所描述的系统和技术同样适用于其他类型的并行处理器的使用。

[0016] 图1示出了根据至少一些实施方案的在PIM模块处同时处理应用的处理系统100。处理系统100包括用于互连处理系统100的各个部件 (包括多个处理模块诸如处理模块104-106、一个或多个存储器控制器108以及一个或多个I/O集线器110) 的数据架构102。每个存储器控制器108又耦接到系统存储器112, 并且每个I/O集线器110又耦接到一个或多个I/O设备, 诸如I/O设备114、115。至少一个系统存储器112包括PIM模块120。

[0017] 处理模块104-106包括一个或多个处理器核心和本地高速缓存分级结构。处理器核心包括例如CPU核心、GPU核心、DSP核心、并行处理器核心或它们的组合。处理客户端的本地高速缓存分级结构包括一级或多级高速缓存。在一些实施方案中, 处理模块104-106中的至少一者不同于处理模块104-106中的至少另一者 (即, 处理模块104-106是异构的)。在其他实施方案中, 处理模块104-106是同构的。在一些情况下, 如以下参考图2进一步讨论的, 处理模块104-106中的一个或多个包括代表处理模块104-106执行任务的一个或多个协处理器。

[0018] 存储器控制器108作为对应系统存储器112与处理系统100的其他部件之间的接口来操作。因此, 要被缓存在处理客户端的本地高速缓存分级结构中的数据通常作为被称为“高速缓存行”的数据块来操纵, 并且利用系统存储器112的物理地址在存储器分级结构中对其进行寻址或以其他方式定位。存储器控制器108响应于来自处理客户端的存取请求而从系统存储器112存取高速缓存行, 并且高速缓存行被安置或缓存在处理客户端的一个或多个高速缓存中。同样地, 当包含经修改数据的高速缓存行被从处理客户端的本地高速缓存分级结构中逐出并且因此需要在系统存储器112中更新时, 存储器控制器108管理这个回写过程。此外, 如以下参考图2至图5所讨论的, 在一些情况下, PIM模块120代表处理模块104-106中的一个或多个来处理数据。在一些实施方案中, 处理系统100包括多个系统存储器112。另外, 在一些实施方案中, 系统存储器112包括多个PIM模块120。

[0019] I/O设备114、115操作以利用DMA存取操作将数据传送到处理系统100中以及传送出该处理系统。例如, I/O设备114、115中的一者可包括用于将该节点连接到网络以用于接收和传输数据的网络接口卡 (NIC)、或者用于非易失性存储相对大量的数据以供处理模块104-106使用的硬盘驱动器 (HDD) 或其他大容量存储设备, 等等。在至少一个实施方案中, I/O集线器110管理I/O设备114、115, 并且充当数据架构102与I/O设备114、115之间的接口。为了说明, 在一些实施方案中, I/O集线器110包括外围部件快速互连 (PCIe) 根联合体, 以作为I/O设备114、115与数据架构102之间的PCIe互连来操作。

[0020] 数据架构102通常在处理系统100的其他部件之间以及在处理系统100与其他节点141之间传输命令、数据、请求、状态通信和其他信令。这些传输操作的一个这样的子集是由I/O设备114、115在系统存储器112处提供的用于供处理模块104-106中的一者或多者使用的数据的存储。如关于图2至图5进一步讨论的, 在一些情况下, 响应于经由数据架构102从处理模块104-16中的一者或多者发送的内核指令, 使用PIM模块120来处理存储在系统存储器112处的数据。I/O代理140作为用于I/O集线器110和I/O设备114、115的相干代理来操作。

此外,在一些实施方案中,传输层130经由一个或多个桥部件或相干代理(未示出)耦接到一个或多个其他节点141的对应传输层或耦接到处理模块104-16。在各种实施方案中,数据架构102与一个或多个标准化互连规范兼容,诸如超传输(TM)规范或无限架构(TM)规范。

[0021] 图2是示出了根据一些实施方案的处理系统100的PIM架构200的框图的框图。在其他实施方案中,PIM架构200对应于包括比处理系统100更多或更少部件的不同处理系统。在所示实施方案中,PIM架构至少包括处理模块104和系统存储器112。处理模块104包括协同处理模块202。系统存储器112包括存储器模块210。存储器模块210包括PCS模块212、PIM内核调度器214、存储设备216、存储映射218和PIM模块120。尽管例示的实施方案示出了部件的特定配置,但是在各种实施方案中,设想了部件的其他组合。例如,在一些实施方案中,系统存储器112仅包括单个存储器模块210。作为另一示例,在一些实施方案中,处理模块104不包括协同处理模块202。此外,在一些实施方案中,设想了附加部件(诸如缓冲器或数据架构102),但为了便于讨论,并未示出附加部件。

[0022] 处理模块104执行各种程序。那些程序中的至少一些的数据被存储在系统存储器112处(例如,在存储设备216中)。在一些情况下,确定在系统存储器112处执行至少一些处理功能(例如,使用PIM模块120中的一个或多个),而不是在处理模块104与系统存储器112之间重复地发送数据请求和响应。因此,处理模块104或者在一些情况下协同处理模块202将内核指令204发送到系统存储器112。在所示出的实施方案中,内核指令204对应于多个应用、单个应用内的不同线程或两者,且作为系统存储器112中的单独PIM内核来执行。在一些实施方案中,处理模块104另外执行一些内核指令(例如,不是PIM模块120的ISA的一部分的内核指令或PIM模块120不具有要执行的带宽的内核指令)。

[0023] 在所示的实施方案中,内核指令204包括一个或多个命令索引206和一个或多个组本地地址208。命令索引206指示将由PIM模块(例如,PIM模块120-1)执行的PIM命令。在一些情况下,命令索引206指示将经由相应指令识别符执行的相应PIM命令。此外,如下文参考图3所描述,在一些情况下,命令索引206中的至少一些包括指示系统存储器112内(例如,存储设备216-2内)将用作执行PIM命令的一部分的寄存器的虚拟寄存器地址。如果PIM命令执行存储器操作(例如,加载操作),则组本地地址208指示系统存储器112的存储设备(例如,存储设备216-1)内要访问的一个或多个存储器地址。在一些实施方案中,组本地地址208是全存储器地址。在其他实施方案中,组本地地址208由系统存储器112修改或生成(例如,通过将存储器阵列的基址添加到数据元素的索引)。因为与例如处理模块104-106相比,PIM模块120具有较小的ISA和较少的硬件资源,所以在一些实施方案中,针对每个PIM命令发送单独的内核指令204(例如,发送一个内核指令以执行加载操作,并且发送第二内核指令以执行添加操作)。在一些实施方案中,用于一个应用或多个应用的多组内核指令204(例如,五个内核指令或五十个内核指令)被一起发送,在一些情况下,使得多个PIM命令被存储在PCS模块212的相应条目中,以供一个或多个PIM模块120处理。然而,在一些实施方案中,处理资源(例如,PCS模块212内的存储空间)是有限的,并且因此发送更小组的内核指令204。另外,如上所述,在一些情况下,在处理模块104处执行一些内核指令。此外,在一些情况下,至少一些PIM命令不同于将在处理模块104处执行以执行类似动作的对应指令。因此,在一些情况下,命令索引206在处理模块104处被翻译成对应的PIM命令。在其他实施方案中,PIM内核调度器214将内核指令204所指示的命令翻译成对应的PIM命令。

[0024] 系统存储器112包括一个或多个存储器模块210。在各种实施方案中,存储器模块210针对相同应用或针对不同应用执行操作。在一些实施方案中,存储器模块210是动态随机存取存储器(DRAM)设备。然而,在其他实施方案中,存储器模块210是其他类型的存储器设备(例如,静态随机存取存储器(SRAM)设备、同步DRAM(SDRAM)存储器设备、快闪存储器设备等)。在一些实施方案中,存储器模块210是相同类型的存储器设备。在其他实施方案中,存储器模块210是不同的存储器设备。在一些实施方案中,存储器模块210中的一个或多个包括更少或附加的部件。例如,在一些实施方案中,PIM内核调度器214-1对应于存储器模块210-1和210-2两者,并且存储器模块210-2不包括PIM内核调度器214-2。在一些实施方案中,存储器模块210包括在存储设备216中组织为寄存器组(例如,物理组织或虚拟组织)的相应寄存器。在其他实施方案中,寄存器是与一个或多个存储器模块210相关联的一个或多个单独设备的一部分。在各种实施方案中,PIM模块120

[0025] PCS模块212存储PIM命令并将PIM命令发送到一个或多个PIM模块120以供执行。如在图4的示例中进一步描述的,在一些实施方案中,PCS模块212中的一个或多个被虚拟化。例如,在一些情况下,PCS模块212-2存储将由PIM模块120-2执行的PIM命令序列,其中该PIM命令是对应于多个相应应用的多个PIM内核的部分。

[0026] PIM内核调度器214管理一个或多个PCS模块212的虚拟化。例如,如上所述,命令索引206指示要执行的相应PIM命令。在一些实施方案中,PIM内核调度器214-2填充PCS模块212-2,并将寻址到PIM模块120-2的命令索引206翻译成PCS模块212-2内的位置,其中相应位置存储对应的PIM命令。另外,在一些实施方案中,PIM内核调度器214中的一个或多个维护并填充存储映射218-2、指示(例如,存储设备216-2的)其中存储PIM命令使用的数据的物理寄存器的虚拟到物理映射表。在一些实施方案中,存储映射218中的一个或多个存储在物理上靠近相应PIM模块120(例如,在同一芯片上)的存储设备中。在一些实施方案中,存储映射218中的一个或多个与存储器模块210分离,但是与一个或多个存储器模块210相关联。

[0027] 在一些实施方案中,为了在填充PCS模块212时减少PCS资源的使用,PIM内核调度器214组合对应于至少两个PIM内核的至少两个内核指令204的部分。例如,如以下参考图4进一步描述的,如果两个内核指令从PIM模块120-1请求“加载R0←[ADDR]”,则PIM内核调度器214-1仅用“加载R0←[ADDR]”的单个指示来填充PCS模块212-1,从而节省PCS模块212-1的存储空间并减少PCS模块212-1被编程的次数。另外,在一些实施方案中,PCS模块212、PIM内核调度器214或两者中的一者或多者在多个PIM模块120之间共享。因此,在一些实施方案中,PIM内核调度器检测到PIM模块120-1和120-2各自要为不同的PIM内核执行“加载R0←[ADDR]”,并且在PCS模块212-1中仅存储“加载R0←[ADDR]”的单个指示。

[0028] 在各种实施方案中,PIM内核调度器214另外确定PIM内核的组合以在PIM模块120处进行调度,在一些情况下,延迟至少一些PIM命令的执行。基于组合度量来确定这些组合,该组合度量指示各种因素,诸如PIM内核之间的相应内核指令的共性、PIM内核的相应优先级、PIM内核的预期资源使用(例如,PIM模块使用、存储器带宽使用、通信带宽等)、PIM内核的依赖性(例如,在从处理模块104接收的依赖图中所指示)。在一些实施方案中,组合由处理模块104指示(例如,来自编译器)。在其他实施方案中,组合由PIM内核调度器214在运行时确定,诸如通过接收PIM内核的依赖图。PIM命令的顺序也基于各种因素来确定,诸如PIM内核之间的相应内核指令的共性、PIM内核的相应优先级、PIM内核的预期资源使用、PIM内

核的依赖性(例如,在从处理模块104接收的依赖图中所指示),或接收到对应内核指令的时间。

[0029] PIM模块120从PCS模块212接收PIM命令,在一些情况下使用存储在存储设备216和存储设备216的寄存器处的数据来处理那些PIM命令。在一些实施方案中,一个或多个PIM模块120是算术逻辑单元(ALU)。另外,在一些实施方案中,存储设备216是被组织为用于PIM模块120的物理寄存器组或虚拟寄存器组的寄存器文件。在一些情况下,寄存器组的至少一些寄存器被分配给第一内核,而寄存器组的至少一些寄存器被分配给第二内核。响应于来自处理模块104的请求(例如,经由内核指令204),PIM模块120将数据返回到处理模块104。因此,存储器模块210的处理资源被虚拟化,使得在一些情况下,PIM模块120同时处理多个PIM内核。

[0030] 在一些实施方案中,一个或多个存储映射218和对应于该一个或多个存储映射218的PIM内核调度器214的一部分被存储在处理模块104上的暂存器中。因此,在那些实施方案中,在内核指令204被发送到系统存储器112之前,对组本地地址208进行翻译。

[0031] 图3示出了根据至少一些实施方案的虚拟到物理内核寄存器映射310的示例300。如上所述,在一些实施方案中,虚拟到物理内核寄存器映射310由PIM内核调度器214(诸如图2的PIM内核调度器214-1)执行。在所示示例300中,内核1虚拟寄存器映射302和内核2虚拟寄存器映射304在PIM内核调度器214处接收。内核1虚拟寄存器映射302的指令所指示的寄存器是R0、R1和R2。PIM内核调度器214将所接收的虚拟寄存器地址分别翻译成物理寄存器地址[0]、[2]和[7],其中该寄存器地址对应于与PIM内核调度器214相关联的存储设备(诸如存储设备216-1的寄存器)中的地址。类似地,内核2虚拟寄存器映射304的指令所指示的寄存器是R0、R1、R2和R3。PIM内核调度器将那些接收到的虚拟寄存器地址分别翻译成物理寄存器地址[3]、[4]、[8]和[9]。因此,PIM模块能够处理指示虚拟索引的内核指令。另外,在一些情况下,将至少一些寄存器分配给第一内核,而将至少一些寄存器分配给第二内核。

[0032] 图4示出了根据至少一些实施方案的内核指令合并和映射420的示例400。如上所述,在一些实施方案中,合并和映射420由诸如图2的PIM内核调度器214-2之类的PIM内核调度器执行。在所示出的示例400中,内核1指令402和内核2指令404各自对应于相应索引。在存储在诸如PCS模块212-2之类的PCS模块中之前,指令被合并以减少PCS模块内的存储利用。因此,内核1指令402和内核2指令404两者的“加载R0←[ADDR]”被合并到内核1指令索引406和内核2指令索引408中的物理索引0中。另外,PIM内核调度器确定指令的顺序,从而产生PCS合并内核内容410。因此,内核1指令402和内核2指令404由PIM模块(诸如PIM模块120-2)以所指示的顺序同时处理。因为各个PIM内核的组本地地址由主机指令生成且与相应PCS模块索引一起发送,所以在一些情况下,即使相同PCS模块条目由多个内核使用,也使用不同物理寄存器。因此,PIM模块能够同时处理多个内核的内核指令,并且减少了PCS模块的资源消耗。

[0033] 图5是示出了根据一些实施方案的使用PIM并行处理系统来同时处理用于多个内核的内核指令的方法500的流程图。在一些实施方案中,由一个或多个处理器响应于由计算机可读存储介质所存储的一个或多个指令来发起方法500。

[0034] 在框502处,接收用于将由PIM ALU同时处理的多个内核的内核指令。例如,从处理模块104接收对应于将由PIM模块120-1处理的多个内核的内核指令204,其中由PIM模块

120-1实施的ISA具有比由处理模块104实施的ISA少的指令。在框504处,将内核指令所指示的命令的虚拟寄存器地址翻译成物理寄存器地址。例如,PIM内核调度器214-1将内核指令204所指示的虚拟寄存器地址翻译为存储设备216-1的寄存器的物理寄存器地址。

[0035] 在框506处,合并由内核指令指示的PIM命令中的至少一些。例如,PIM内核调度器214-1分析内核指令204所指示的PIM命令并基于组合度量来合并至少一些PIM命令。在框508处,将经翻译和合并的命令存储在PCS模块中。例如,经翻译和合并的命令被存储在PCS模块212-1中。在框510处,检索并执行至少一个合并的命令。例如,从PCS模块212-1检索对应于至少两个PIM内核的至少一个PIM命令且将其发送到PIM模块120-1以供执行。因此,描绘了同时处理用于多个内核的内核指令的方法。

[0036] 在一些实施方案中,计算机可读存储介质包括在使用期间可由计算机系统访问以向计算机系统提供指令和/或数据的任何非暂态存储介质或非暂态存储介质的组合。此类存储介质可以包括但不限于光学介质(例如,光盘(CD)、数字通用光盘(DVD)、蓝光光盘)、磁性介质(例如,软盘、磁带或磁性硬盘驱动器)、易失性存储器(例如,随机存取存储器(RAM)或高速缓冲存储器)、非易失性存储器(例如,只读存储器(ROM)或闪存存储器)、或基于微机电系统(MEMS)的存储介质。在一些实施方案中,计算机可读存储介质嵌入计算系统(例如,系统RAM或ROM)中,固定地附接到计算系统(例如,磁性硬盘驱动器),可移除地附接到计算系统(例如,光盘或基于通用串行总线(USB)的闪存存储器),或者经由有线或无线网络(例如,网络可访问存储装置(NAS))耦接到计算机系统。

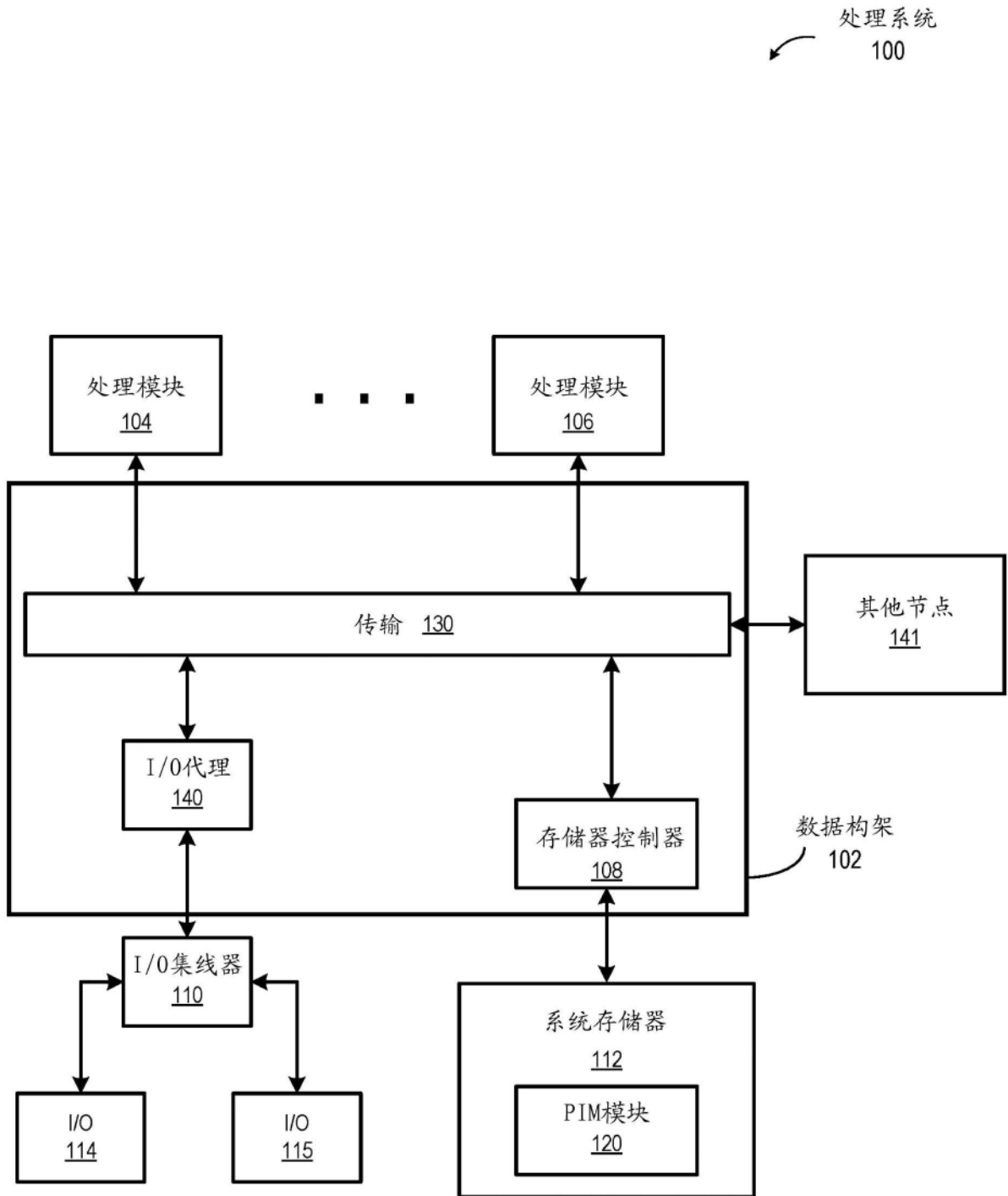
[0037] 在一些实施方案中,上述技术的某些方面由执行软件的处理系统的一个或多个处理器实现。软件包括可执行指令的一个或多个集合,该可执行指令存储在或以其他方式有形地体现在非暂态计算机可读存储介质上。软件可包括指令和某些数据,这些指令和数据在由一个或多个处理器执行时操纵一个或多个处理器以执行上述技术的一个或多个方面。非暂态计算机可读存储介质可包括例如磁盘或光盘存储设备、固态存储设备诸如闪存存储器、高速缓冲存储器、随机存取存储器(RAM)或其他一个或多个非易失性存储器设备等。在一些实施方案中,存储在非暂态计算机可读存储介质上的可执行指令是源代码、汇编语言代码、目标代码、或者被一个或多个处理器解释或以其他方式可执行的其他指令格式。

[0038] 应当注意,并非以上在一般描述中描述的所有活动或元件都是必需的,特定活动或设备的一部分不是必需的,并且在一些情况下,除了所描述的那些之外,还执行一个或多个另外的活动,或者还包括一个或多个另外的元件。更进一步地,列出活动的顺序不一定是执行它们的顺序。另外,已经参考具体实施方案描述了这些概念。然而,本领域普通技术人员理解,在不脱离如以下权利要求中阐述的本公开的范围的情况下,可以进行各种修改和改变。因此,说明书和附图被认为是说明性的而非限制性的,并且所有此类修改旨在被包括在本公开的范围之内。

[0039] 上文已经关于具体实施方案描述了益处、其他优点和问题的解决方案。然而,益处、优点、问题的解决方案以及导致任何益处、优点或解决方案出现或变得更显著的任何特征不应被解释为任何或所有权利要求的关键的、必需的或基本的特征。此外,上文公开的特定实施方案仅是说明性的,因为所公开的主题可以以受益于本文中的教导内容的本领域的技术人员显而易见的不同但等效的方式来修改和实践。除了以下权利要求书中所描述的之外,不旨在对本文所示的构造或设计的细节进行限制。因此,显而易见的是,可以改变或修

改上文公开的特定实施方案,并且所有此类变化被认为是在所公开的主题的范围内。因此,本文寻求的保护如以下权利要求中所阐述。

[0040] 在本公开内,在一些情况下,不同实体(其被不同地称为“部件”、“单元”、“设备”等)被描述或宣称为“被配置”为执行一个或多个任务或操作。该阐述方式(即,被配置为[执行一个或多个任务]的[实体])在本文中用来指结构(即,某种物理结构,诸如电子电路)。更具体地,该阐述方式用于指示该结构被布置为在操作期间执行该一个或多个任务。结构可被称为“被配置为”执行某个任务,即使当前未操作该结构。“被配置为存储数据的存储器设备”旨在覆盖例如具有在操作期间存储数据的电路的集成电路,即使当前未使用所考虑的集成电路(例如,电源未连接至该集成电路)。因此,被描述或叙述为“被配置为”执行某个任务的实体是指某种物理实体,诸如设备、电路、存储程序指令(所述程序指令可执行以实现该任务)的存储器等。该短语在本文中不用来指某种无形实体。此外,术语“被配置为”并非旨在意指“可配置为”。例如,未编程的现场可编程门阵列将不会被视为“被配置为”执行某种特定功能,但其可以是“可配置为”在编程后执行该功能。另外,在所附权利要求书中叙述结构“被配置为”执行一个或多个任务明确地不旨在被解释为具有装置加功能元件。



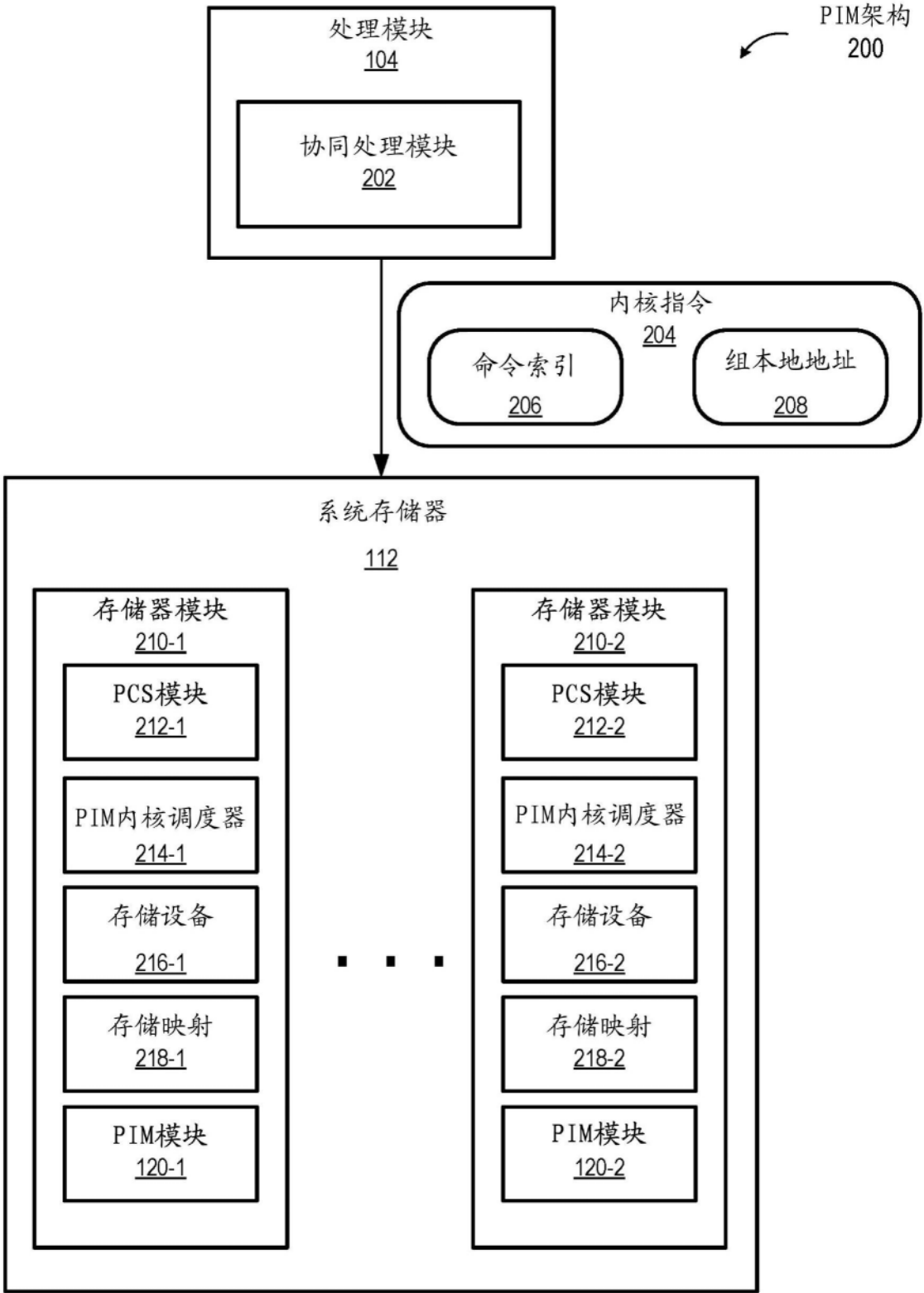


图2



图3



图4

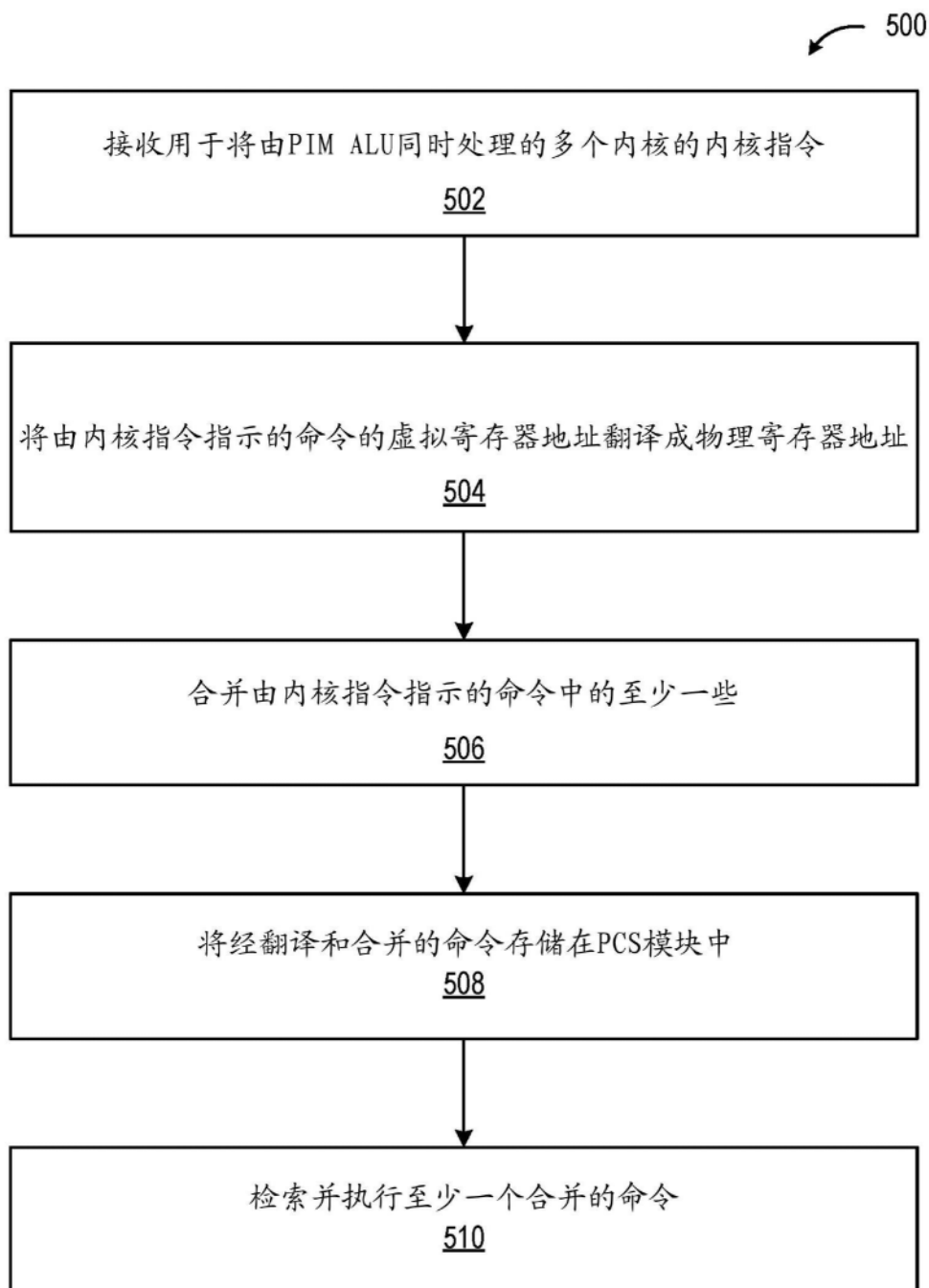


图5