(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 104272249 B
(45)授权公告日 2019.01.08

(21)申请号 201280072796.0

(22)申请日 2012.06.08

(65)同一申请的已公布的文献号
申请公布号 CN 104272249 A

(43)申请公布日 2015.01.07

(85)PCT国际申请进入国家阶段日
2014.10.30

(86)PCT国际申请的申请数据
PCT/US2012/041546 2012.06.08

(87)PCT国际申请的公布数据
WO2013/184125 EN 2013.12.12

(73)专利权人 慧与发展有限责任合伙企业
地址 美国德克萨斯州

(72)发明人 凯文·T·林 阿尔温·奥杨

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 严芬 康泉

(51)Int.Cl.
G06F 9/06(2006.01)
G06F 13/00(2006.01)

(56)对比文件
CN 102016808 A,2011.04.13,
CN 101903866 A,2010.12.01,
CN 101256526 A,2008.09.03,
US 20110252181 A1,2011.10.13,
US 5175837 A,1992.12.29,
US 20090070391 A1,2009.03.12,

审查员 董泽华

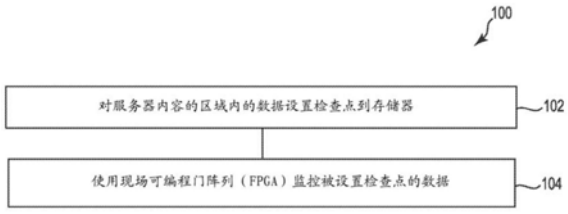
权利要求书2页 说明书6页 附图2页

(54)发明名称

使用FPGA设置检查点

(57)摘要

方法、系统以及计算机可读和可执行指令被提供用于使用现场可编程门阵列(FPGA)设置检查点。使用FPGA设置检查点可包括对服务器内容的区域内的数据设置检查点到存储器,并且使用FPGA监控被设置检查点的数据。



CN 104272249 B

1.一种计算机实现的方法,用于使用现场可编程门阵列FPGA设置检查点,该方法包括:

使用所述现场可编程门阵列FPGA临时地锁定分布式缓存散列表以免存取;

在锁定所述分布式缓存散列表的同时,对分布式服务器内容的第一区域设置检查点到非易失性存储器;

解除所述锁定达一时间段;并且

使用所述FPGA监控被设置检查点的数据,其中所述FPGA相干联接至所述存储器,使得利用所述FPGA捕获所述第一区域内的存取来自动执行所述监控,并且所述FPGA无效与设置检查点的所述第一区域相关的缓存行。

2.根据权利要求1所述的方法,其中到所述存储器的存取被广播到所述FPGA。

3.根据权利要求1所述的方法,其中对数据设置检查点到存储器进一步包括对数据设置检查点到非易失性存储器。

4.根据权利要求1所述的方法,进一步包括:响应于所述服务器崩溃利用所述被设置检查点的数据来恢复所述服务器。

5.根据权利要求1所述的方法,进一步包括:压缩所述被设置检查点的数据。

6.根据权利要求1所述的方法,其中监控所述数据进一步包括:从所述区域排除过期的数据条目。

7.一种非暂时性计算机可读介质,存储可由处理资源执行的一组指令,以便:

使用现场可编程门阵列FPGA临时地锁定分布式缓存散列表以免存取;

在锁定所述分布式缓存散列表的同时,对分布式缓存服务器内容的第一区域设置检查点到非易失性存储器;

解除所述锁定达一时间段;

重新锁定所述分布式缓存散列表;

在重新锁定所述分布式缓存散列表的同时,对所述分布式缓存服务器内容的第二区域设置检查点到所述非易失性存储器;以及

使用所述FPGA监控在所述第一区域和所述第二区域内被设置检查点的数据,其中所述FPGA相干附接至所述非易失性存储器,使得利用所述FPGA捕获所述第一区域和所述第二区域内的存取来自动执行所述监控,并且所述FPGA无效与设置检查点的所述第一区域和所述第二区域相关的缓存行。

8.根据权利要求7所述的非暂时性计算机可读介质,其中所述指令进一步是可执行的,以通过追踪所述第一区域和所述第二区域中的每个区域的更新来监控所述数据。

9.根据权利要求7所述的非暂时性计算机可读介质,其中所述指令进一步是可执行的,以拆分分布式缓存散列表锁并将附加的分布式缓存散列表锁增加到所述第一区域和所述第二区域中的每个区域。

10.根据权利要求7所述的非暂时性计算机可读介质,其中所述指令进一步是可执行的,以在设置检查点期间临时地减少区域更新。

11.根据权利要求7所述的非暂时性计算机可读介质,其中所述指令进一步是可执行的,以接收关于所述第一区域和所述第二区域中每个区域的散列表以及所述第一区域和所述第二区域中每个区域内的数据的信息到所述FPGA。

12.一种系统,用于使用现场可编程门阵列FPGA设置检查点,所述系统包括:

存储器资源;以及

联接至所述存储器资源的处理资源,用以:

接收分布式缓存服务器内的多个地址区域到所述FPGA以设置检查点;

使用所述现场可编程门阵列FPGA临时地锁定分布式缓存散列表以免存取;

在锁定所述分布式缓存散列表的同时,对所述多个地址区域中的第一区域设置检查点到非易失性存储器;

解除所述锁定达一时间段;并且

使用所述FPGA自动地监控所述多个地址区域中的每个内的数据,其中所述FPGA相干联接至所述非易失性存储器,使得利用所述FPGA捕获所述第一区域内的存取来自动执行所述监控,并且所述FPGA无效与设置检查点的所述第一区域相关的缓存行。

13.根据权利要求12所述的系统,其中所述处理资源进一步联接至所述存储器资源,以允许所述FPGA访问所述多个区域中的每个内的分布式缓存散列表锁。

14.根据权利要求12所述的系统,其中所述处理资源进一步联接至所述存储器资源,以向所述非易失性存储器保存所述多个地址区域中之一开始的指针、地址区域检查点的大小、和分布式缓存散列表开始的地址中的至少一个。

15.根据权利要求12所述的系统,其中所述FPGA位于独立的网络套接字中而不是位于所述非易失性存储器中。

使用FPGA设置检查点

背景技术

[0001] 分布式缓存(Memcached)是提供对象缓存的存储器中键值缓存,并且可用在网络服务器层中。分布式缓存服务器为来自数据库的内容提供缓存,并且可被放置在前端网络服务器和后端数据库之间。

附图说明

[0002] 图1是图示根据本公开的用于设置检查点的方法示例的框图。

[0003] 图2图示根据本公开的用于设置检查点的示例系统的框图。

[0004] 图3是图示根据本公开的处理资源、存储器资源和计算机可读介质的框图。

具体实施方式

[0005] 分布式缓存是一种存储器中键值缓存,其相比于其它缓存方法提供了较高的吞吐量和/或较低延迟的对象缓存。分布式缓存服务器可在网络服务器层中发挥作用,被放置在前端网络服务器和后端数据库(例如,结构化查询语言数据库)之间。这些服务器缓存来自数据库的内容,这减轻了对访问后端数据库的需求。该缓存降低了前端服务器的对象检索存取的延迟,并且也降低了后端服务器上的负荷。

[0006] 然而,与其它缓存方法相比,缓存(例如,有效地缓存)网络层级处的内容会需要增加量的存储器(例如动态随机存取存储器(DRAM))容量。在实际的部署中(例如,社交网络),与其它缓存方法相比,分布式缓存服务器可具有每个服务器较大的存储器容量(例如,每个服务器超过64GB的存储器),并且在整个服务器集群中具有成千上万的分布式缓存服务器。在这些部署中,通过增加服务器的数量和/或增加单个服务器的容量,可满足服务器池内总存储器(例如,DRAM)容量的增加。

[0007] 为了减轻管理负担并减少管理多个实体服务器的所有权的总成本,部署可增加单个服务器的容量(具体是存储器密度增加)来代替添加额外的服务器。然而,重新填充大缓存所需的时间量的问题会成为部署具有这些大存储器容量的分布式缓存服务器的障碍。问题可包括源自分布式缓存服务器崩溃的临时停机时间和/或严重的性能下降(例如,由于缓存丢失和在后端数据库中所引起的负荷);增加的存储器容量会增加分布式缓存服务器重新填充其缓存和恢复其正常运行机制所需的时间。

[0008] 如这里进一步讨论的,对数据设置检查点(例如,使用现场可编程门阵列(FPGA)对数据设置检查点)可使得这些分布式缓存服务器随着存储器密度的趋势而增加存储器容量。例如,与其它方法相比,分布式缓存服务器能使用非易失性存储器和相干附接(coherently-attached)的FPGA的组合来提供更快的恢复,而不会在正常的服务器运行时过度地影响性能。

[0009] 本公开的示例可包括方法、系统以及计算机可读和可执行指令和/或逻辑。使用FPGA设置检查点的示例方法可包括:对在服务器内容区域内的数据设置检查点到存储器,并且使用FPGA监控被设置检查点的数据。

[0010] 在本公开的下列详细描述中，参考形成本发明一部分的附图，并且在附图中以图示的方式示出本公开的示例可如何实施。这些实施例足够详细地被描述，以使得本领域普通技术人员能够实施本公开的示例，并且应当理解，还可利用其它示例，且可在不脱离本公开的范围的情况下做出过程、电子和/或结构的改变。

[0011] 本文中的图遵循以下编号惯例：其中第一位数字对应于图号，而其余数字标识图中的元件或组件。不同图之间的相似元件或组件通过使用相似的数字来标识。在本文的各个示例中示出的元件可以被增加、替换和/或去除，以便提供本公开的多个附加的实施例。

[0012] 此外，图中提供的元件的比例和相对大小旨在说明本公开的示例，而不应构成限制意义。如这里所用的，具体关于图中的附图标记，标志符“N”、P”、“R”和“S”指示这样标记的多个具体的特征可以包括在本公开的多个示例中。而且，如这里所用的，“多个”元件和/或特征可指代一个或多个这样的元件和/或特征。

[0013] 根据服务器的性能要求和从故障到恢复之间平均时间的预期，可以以特定的间隔(例如，由分布式缓存服务器管理员限定的间隔)来设置检查点。通过暂停分布式缓存服务器的运行、复制其全部数据到非易失性存储器(例如非易失性随机存取存储器(NVRAM))以及允许分布式缓存服务器重新恢复，来执行检查点。然而，考虑到动态随机存取存储器(DRAM)的容量(例如，64GB和/或更大)以及非易失性存储器的写入带宽(例如，1GB/s的写入带宽)，以这种方式完成完全检查点会花费比预期时间更长的时间(例如，数分钟而不是数秒钟)。

[0014] 图1是图示根据本公开的设置检查点的方法100的示例的框图。在102，对服务器内容区域中的数据设置检查点到存储器。在多个实施例中，设置检查点数据到非易失性存储器，包括到闪存等。相对于其它的检查点设置方法，对存储器区域设置检查点可提供具有减少的中断时间的检查点设置。在一些示例中，可按照区域对存储器区域设置检查点，这迭代通过存储器(例如，全部存储器)直到设置了检查点(例如，直到设置了全部检查点)。

[0015] FPGA可用于临时锁定分布式缓存散列表来避免存取、对散列表的特定区域(例如，2GB、4GB等)设置检查点(例如，读取和复制值至存储器)、解除锁定、以及允许整个系统恢复。例如，当对分布式缓存状态的特定区域(例如，DRAM)设置检查点时，采取防范措施以防客户端请求改变该区域的内容。例如，在复制区域时，“阻挡”和/或“锁定”客户端请求直到该区域完成复制。以这种方式来设置检查点(例如，降低的时间量)能防止正常客户端请求的中断。

[0016] 在特定时间段过去之后，FPGA能重新锁定散列表并且对下一个区域设置检查点，继续直到对整个存储器设置了检查点。基于本申请，可以改变检查点区域之间的特定时间段(例如，纳秒、毫秒或微秒等)。还可以改变全部检查点之间的特定时间段(例如，秒、分、小时等)。例如，与具有较轻负荷的服务器的分布式缓存部署相比，具有较重负荷的服务器的分布式缓存部署可能需要更长的时间。在一些示例中，与其它方法相比，FPGA可以在较短的时间段内锁定散列表，并且与单个通过(single pass)检查点方法相比，FPAG可以导致减少的中断时间，这例如使得在设置检查点的同时保持服务在线。

[0017] 在104，使用FPGA监控被设置检查点的数据。FPGA可相干附接和/或联接到存储器，使得主处理器广播到存储器的任何存取都被广播到FPGA，这允许FPGA自动追踪(例如，监控)和捕获每个区域内的存取。在一些示例中，FPGA可捕获每个区域内的更新。通过监控每

个区域,FPGA可维护每个检查点以及全部检查点的一致性视图。在多个示例中,FPGA位于单独的网络套接字中(例如,横贯计算机网络的进程间通信流的不同端点)而不是位于存储器中。

[0018] 通过无效与设置检查点的区域相关的缓存行,相干附接的FPGA能观测至该区域的通信量。FPGA能监听那些缓存行。例如,如果非易失性存储器不是字节可寻址的(例如,闪存),则可以缓冲所捕获的更新并且写入日志。该日志能回放已完成的检查点,这允许维护一致的状态。

[0019] 如果足够的更新(例如,特定数量)发生在单个块,则可将整个块重写到非易失性存储器。在一些实施例中,块(例如,单元)可包括FPGA将信息写到其自身本地存储器的粒度。例如,如果足够的更新(例如,日志更新)填满块,则FPGA可将该块写到非易失性存储器(例如,NVRAM)。在一些示例中,足够的更新可包括足够的写入,整个块可被编程为记录每个操作的块。一旦完成检查点设置,FPGA可停止监听与检查点相关的地址区域。

[0020] 如果需要恢复检查点,则执行可使得分布式缓存进程请求将数据从FPGA存储器转移到主存储器的系统调用。可以保存(例如,至存储器)和提供簿记状态(bookkeeping state),包括指向检查点区域开始的指针、检查点的大小和/或分布式缓存散列表开始的地址等等。在一些示例中,响应于服务器崩溃,利用被设置检查点的数据来恢复服务器。在一些示例中,附接到FPGA的存储器仅用于设置检查点,并且FPGA用于加速设置检查点。

[0021] 在一些实施例中,监控被设置检查点的数据可包括从检查点设置区域排除过期的数据条目。分布式缓存这样的系统能使用过期条目的惰性驱逐(lazy eviction)。例如,不会立即驱逐早于其过期时间的条目,而改为在下次存取时进行驱逐。在本公开中,为了降低写入到存储器(例如,非易失性存储器)的信息量,FPGA会对从该区域所读取的数据执行检查,并且排除早于其过期时间的任何条目(例如,使用整数比较)。这可以降低由于写入造成的有效期消耗,以及可以降低所消耗的带宽。

[0022] 为了减少和/或进一步减少写入的数据,可对正被设置检查点的数据应用压缩(例如,使用比原始表示更少比特的编码信息)。FPGA能读取和缓冲特定量的分布式缓存状态,在写入存储器之前可压缩该特定量的分布式缓存状态。例如,不是让FPGA将分布式缓存的DRAM状态的内容复制到非易失性存储器(例如,NVRAM)中,而是该FPGA可将内容复制到其自身的内部缓冲中、压缩内容并且将压缩的内容复制到非易失性存储器中。在一些示例中,通过将其自身存储器上的内容分阶段,FPGA能降低锁定分布式缓存DRAM内容(例如来自客户端请求)的时间量。当恢复检查点时,数据可以被解压缩。压缩的选择(例如,压缩模型的选择)可以向服务器管理员提供选项以平衡存储器检查点容量和速度。例如,与最终检查点保存相反,在检查点恢复时可回放已记录的更新。

[0023] 服务中断可通过拆分散列表锁和向该区域中增加额外的锁而降低。因此,能使得发生在此时未设置检查点的区域的存取继续进行,这降低了整体服务中断。例如,FPGA会对多个区域中之一内的分布式缓存散列表锁进行存取。

[0024] 在一些示例中,FPGA能接收关于多个区域中每个区域的散列表和多个区域中每个区域内的数据的信息。例如,FPGA会对多个区域中之一内的分布式缓存散列表锁进行存取。

[0025] 在一些实施例中,FPGA追踪对已经设置了检查点的区域的更新。为了降低更新的量,能临时地降低在设置检查点期间的状态改变。例如,正当设置检查点时,服务器可避免

更新缓存策略簿记(例如,至少最近使用的列表)。这样做不会影响准确性,并且能降低对存储器的写入量和FPGA追踪的更新量。

[0026] 图2图示出根据本公开的用于设置检查点的示例系统220的框图。系统220可包括FPGA 254和存储器(例如,非易失性存储器)256。FPGA 254可相干附接和/或联接到存储器256,这意味着例如主处理器广播到存储器256的存取能广播到FPGA 254,使得FPGA 254自动追踪(例如,监控)和捕获每个设置了检查点的区域内的存取。

[0027] 系统220可包括具有存储器和处理资源的计算设备222,指令(例如,计算机可读指令(CRI)244)存储在存储器中并且由处理资源执行以对数据设置检查点。正如此处所述,计算设备222可以是硬件和/或配置为设置检查点的程序指令(例如,CRI)的任意组合。例如,硬件可包括一个或多个处理资源250-1、250-2…250-N、计算机可读介质(CRM)246等。程序指令可包括存储在CRM 246上的指令,该指令可由一个或多个处理资源执行以实现一个或多个不同功能或此处所述的特定动作(例如,检查点数据)。

[0028] 计算设备222可包括与处理资源250-1、250-2…250-N通信的CRM 246。CRM 246可与具有多于或少于250-1、250-2…250-N的处理资源的计算设备248(例如,Java®应用服务器等)通信。如此处所述,计算设备248可与有形非暂时性CRM 246通信,该CRM 246存储可由一个或多个处理资源250-1、250-2…250-N执行的一组计算机可读指令(CRI)244。CRI 244还可存储在由服务器管理的远程存储器中,并且呈现为可下载、安装和执行的安装包。计算设备248可包括存储器资源252,并且处理资源250-1、250-2…250-N可联接至存储器资源252。

[0029] 处理资源250-1、250-2…250-N能运行可存储在内部或外部非暂时性CRM 246上的CRI 244。处理资源250-1、250-2…250-N可运行CRI 244以执行包括方法100中描述的功能的各种功能。例如,处理资源250-1、250-2…250-N能运行CRI 244以对数据设置检查点。如此处所用,非暂时性CRM(例如,CRM 246)可包括易失性和/或非易失性存储器。易失性存储器可包括依靠电源来存储信息的存储器,如不同类型的动态随机存取存储器(DRAM)等。非易失性存储器可包括不依靠电源来存储信息的存储器。非易失性存储器的示例可包括如闪存、电可擦除可编程只读存储器(EEPROM)、相变随机存取存储器(PCRAM)的固态介质,如硬盘、磁带驱动器、软盘和/或磁带存储器之类的磁存储器,光盘,数字化通用盘(DVD),蓝光盘(BD),压缩盘(CD)和/或固态驱动器(SSD)等,以及其它类型的计算机可读介质。

[0030] 非暂时性CRM 246可以是集成的或者以有线和/或无线的方式可通信地联接到计算设备248。例如,非暂时性CRM 246可以是内部存储器、便携式存储器,便携式盘或与其它计算资源相关联的存储器。

[0031] CRM 246可经由通信路径242与处理资源250-1、250-2…250-N通信。通信路径242可相对于与处理资源250-1、250-2…250-N相关联的机器(例如,计算设备248)是本地的或远程的。本地通信路径242的示例可包括机器(例如,计算机)内部的电子总线,其中CRM 246是经由电子总线与处理资源250-1、250-2…250-N通信的易失性、非易失性、固定和/或可移除的存储介质之一。这种电子总线的示例可包括工业标准体系结构(ISA)、外围组件互连(PCI)、高级技术附件(ATA)、小型计算机系统接口(SCSI)、通用串行总线(USB)、其它类型的电子总线及其变体。

[0032] 通信路径242可以是这样的,使得CRM 246相对于处理资源250-1、250-2…250-N是

远程的，如在CRM 246和处理资源(例如，250-1、250-2…250-N)之间的网络连接中。也就是说，通信路径242可以是网络连接。这种网络连接的示例可包括局域网(LAN)、广域网(WAN)、个人局域网(PAN)和因特网等。在这些示例中，CRM 246可与第一计算设备相关联，并且处理资源250-1、250-2…250-N可与第二计算设备(例如，计算设备248)相关联。例如，处理资源250-1、250-2…250-N可与CRM 246通信，其中CRM 246包括一组指令并且其中处理资源250-1、250-2…250-N被设计为执行该组指令以对数据设置检查点。

[0033] 联接至存储器252的处理资源250-1、250-2…250-N可运行程序指令以对数据设置检查点。联接至存储器252的处理资源250-1、250-2…250-N可运行程序指令以接收分布式缓存服务器内的多个地址区域到FPGA来设置检查点。在本公开的各种示例中，联接至存储器252的处理资源250-1、250-2…250-N可运行程序指令以周期性地对多个地址区域中的每个区域设置检查点到非易失性存储器，并且在本公开的一些示例中，联接至存储器252的处理资源250-1、250-2…250-N可运行程序指令以使用FPGA自动地监控多个地址区域中每个区域内的数据。

[0034] 如此处所使用的，“逻辑”是运行这里所描述的动作和/或功能等的替代物或附加的处理资源，其包括硬件(例如，各种形式的晶体管逻辑电路、专用集成电路(ASIC)等)，与在存储器中存储的且通过处理可运行的计算机可执行指令(例如，软件、固件等)相反。

[0035] 图3图示出根据本公开的在云系统中平衡管理责任(management duties)的示例计算系统322的图。计算系统322可包括处理资源350。处理资源350例如可包括图2中所示的处理资源250-1、250-2…250-N。

[0036] 处理资源350可经由通信路径342可通信地联接至CRM 346。CRM 346可与图2中所示的CRM 246相似。CRM 346可包括多个模块378、380、382、384、386和388。该多个模块可包括例如可由处理资源350运行以执行多个功能的CRI。

[0037] 锁定模块378可例如包括多个CRI，这些CRI可由处理资源350运行以执行或实现特定动作，或者实施临时地锁定分布式缓存散列表以防使用FPGA存取的动作。

[0038] 第一检查点模块380可包括可由处理资源350运行的多个指令。例如，第一检查点模块380可对分布式缓存服务器内容的第一区域设置检查点到非易失性存储器，与此同时，分布式缓存散列表被锁定。

[0039] 解除模块382可包括可由处理资源350运行的多个指令。例如，解除模块382可解除锁定达一时间段，并且重锁模块384(例如，包括可由处理资源350运行的多个指令)可重新锁定分布式缓存散列表。在多个实施例中，锁定模块378可重新锁定分布式缓存散列表。锁定模块378和重锁模块384可包括例如相同的模块。例如，在一些实施例中，计算系统322不包括重锁模块。

[0040] 第二检查点模块386可包括可由处理资源350运行的多个指令。例如，第二检查点模块386可对分布式缓存服务器内容的第二区域设置检查点到非易失性存储器，与此同时，分布式缓存散列表被重新锁定。在多个实施例中，第一检查点模块380和第二检查点模块386包括相同的模块。

[0041] 监控模块388可例如包括可由处理资源350运行的多个指令。例如，监控模块388可使用FPGA监控在第一和第二区域内设置了检查点的数据，其中FPGA相干附接至非易失性存储器。

[0042] 说明书示例提供对本公开的系统和方法的应用和使用的描述。由于在不脱离本公开的系统和方法的精神和范围的情况下可进行多个示例,因此此说明书列出了众多可行示例配置和应用中的一些。


100
对服务器内容的区域内的数据设置检查点到存储器
102
使用现场可编程门阵列（FPGA）监控被设置检查点的数据
104

图1

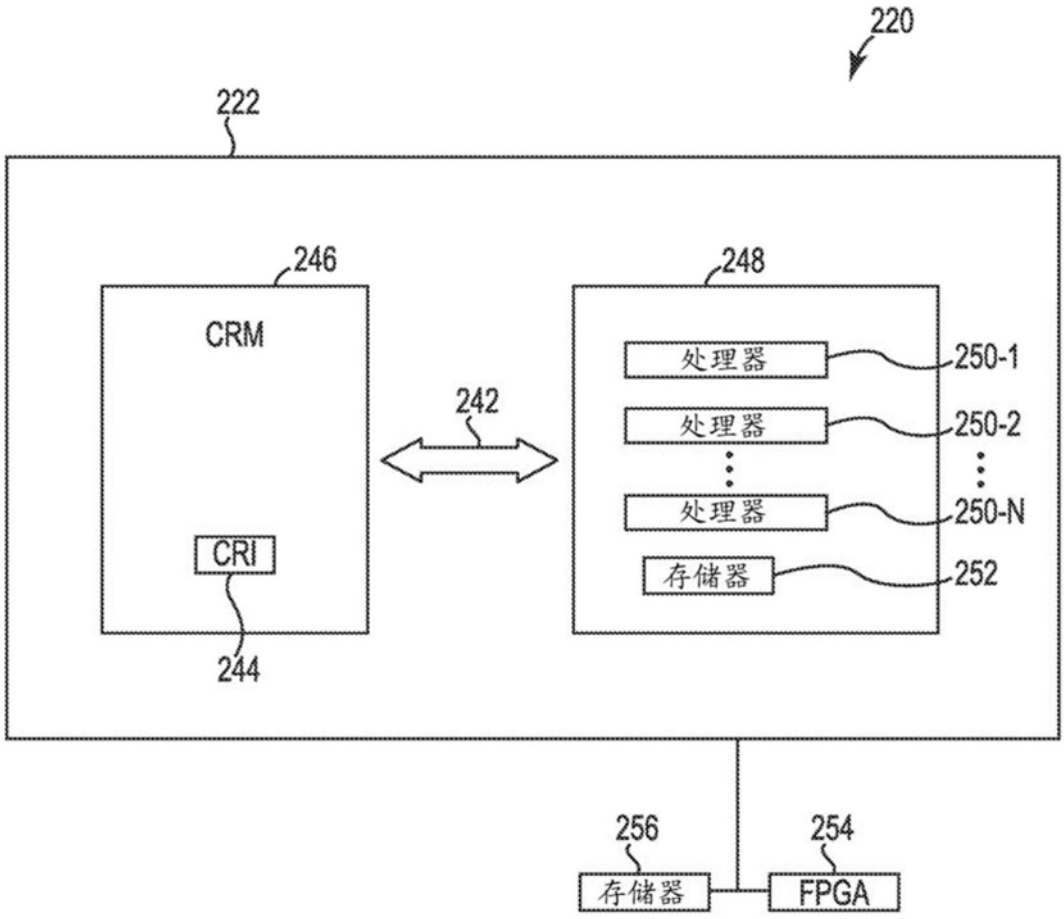

220
222
246
CRM
CRI
244
242
248
处理器
250-1
处理器
250-2
处理器
250-N
存储器
252
256
存储器
254
FPGA

图2

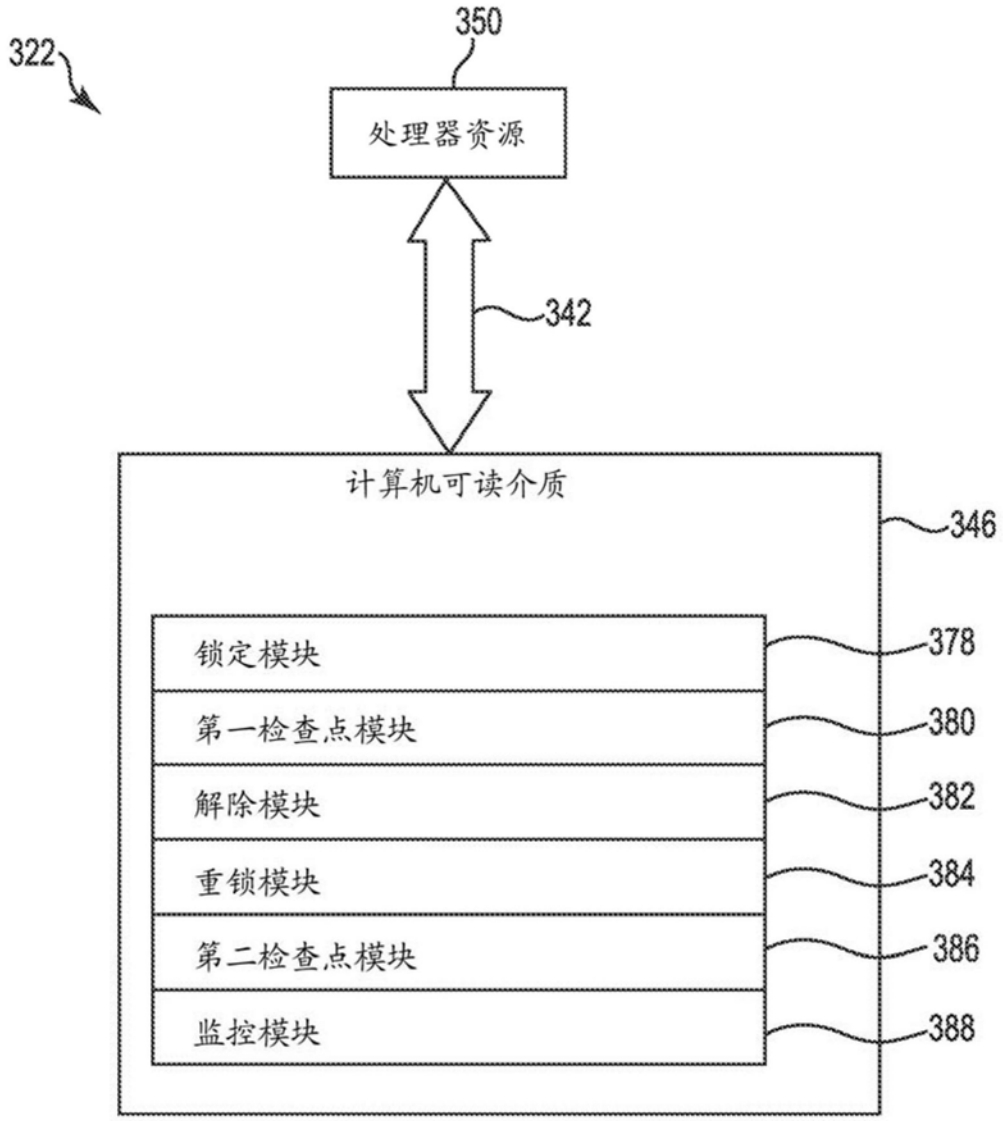

322
350
处理器资源
342
计算机可读介质
346
锁定模块
378
第一检查点模块
380
解除模块
382
重锁模块
384
第二检查点模块
386
监控模块
388

图3