(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 102016808 B
(45)授权公告日 2016.08.10

(21)申请号 200880128994.8

(22)申请日 2008.05.01

(85)PCT国际申请进入国家阶段日
2010.10.29

(86)PCT国际申请的申请数据
PCT/US2008/062154 2008.05.01

(87)PCT国际申请的公布数据
WO2009/134264 EN 2009.11.05

(73)专利权人 惠普发展公司，有限责任合伙企业
地址 美国得克萨斯州

(72)发明人 N·茹匹 A·戴维斯 N·阿加瓦尔 R·考夫曼

(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 宋颖娉 罗正云

(51)Int.Cl.
G06F 12/02(2006.01)
G06F 12/00(2006.01)

(56)对比文件
US 6795966 B1,2004.09.21,
US 6336161 B1,2002.01.01,
US 2007/0180217 A1,2007.08.02,

审查员 武晓冬

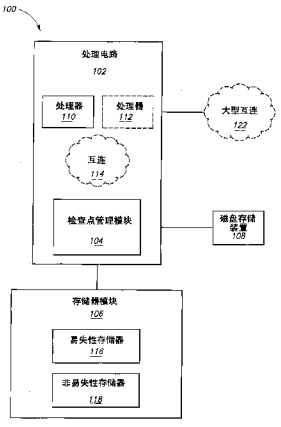
权利要求书3页 说明书8页 附图4页

(54)发明名称

将检查点数据存储于非易失性存储器中

(57)摘要

本发明描述了将检查点数据存储于非易失性存储器的方法和系统。根据一个实施例，数据存储方法包括使用处理电路执行应用程序，以及在执行期间将由应用程序的执行所生成的数据写入易失性存储器。写入数据之后提供检查点的指示。在已提供指示之后，该方法包括将数据从易失性存储器复制到非易失性存储器，以及复制之后继续执行应用程序。该方法可包括将应用程序的执行挂起。根据另一实施例，数据存储方法包括接收与一个或多个应用程序的执行相关联的检查点的指示，且响应于该接收，启动将由一个或多个应用程序的执行所生成的数据从易失性存储器复制到非易失性存储器。在一些实施例中，非易失性存储器可以是固态非易失性存储器。



CN 102016808 B

1.一种数据存储方法,包括:

使用处理电路执行应用程序;

在所述执行期间,将由所述应用程序的所述执行生成的数据写入易失性存储器;

在所述写入之后,提供检查点指示,其中,所述检查点指示由所述应用程序中的检查点指令提供,该检查点指令位于执行所述应用程序中一计算或者一功能的指令片断的末尾;

在所述提供之后,将所述数据从所述易失性存储器复制到非易失性存储器;

在所述复制期间将所述应用程序的所述执行挂起;以及

在所述复制之后,继续所述应用程序的所述执行,

其中所述非易失性存储器是非易失性固态存储器,且所述非易失性固态存储器和所述易失性存储器两者均为单个双列直插存储器模块DIMM的一部分。

2.如权利要求1所述的方法,其特征在于,还包括:

在继续所述执行之后,检测所述应用程序的所述执行中的错误;

响应于所述检测,将所述数据从所述非易失性固态存储器复制到所述易失性存储器;以及

在将所述数据从所述非易失性固态存储器复制到所述易失性存储器之后,使用存储于所述易失性存储器中的所述复制数据从所述检查点开始执行所述应用程序。

3.如权利要求1所述的方法,其特征在于,所述非易失性固态存储器和所述易失性存储器被组织成多个双列直插存储器模块DIMM,其中,每个DIMM包括所述易失性存储器的不同部分以及所述非易失性固态存储器的不同部分;所述非易失性固态存储器包括多个集成电路,所述易失性存储器包括多个集成电路,其中,每个DIMM包括构成所述易失性存储器一部分的多个集成电路以及构成所述非易失性固态存储器一部分的多个集成电路;

所述数据的所述复制包括:并发地,将所述数据的第一部分从所述易失性存储器的第一部分复制到所述非易失性固态存储器的第一部分,以及将所述数据的第二部分从所述易失性存储器的第二部分复制到所述非易失性固态存储器的第二部分;

其中,所述易失性存储器的第一部分和所述非易失性固态存储器的第一部分位于同一DIMM,所述易失性存储器的第二部分和所述非易失性固态存储器的第二部分位于同一DIMM。

4.如权利要求1所述的方法,其特征在于,提供所述检查点的所述指示包括响应于所述处理电路完成所述应用程序的一部分的执行而提供所述指示。

5.如权利要求1所述的方法,其特征在于,所述提供包括使用由所述处理电路执行的操作系统来提供所述指示。

6.一种数据存储方法,包括:

接收与一个或多个应用程序的执行相关联的检查点的指示,其中,所述检查点指示由所述一个或多个应用程序的检查点指令提供,该检查点指令位于执行所述一个或多个应用程序中一计算或者一功能的指令片断的末尾;以及

响应于所述接收,启动将所述一个或多个应用程序的执行所生成的数据从易失性存储器复制到非易失性存储器,

其中在将所述数据从所述易失性存储器复制到所述非易失性存储器期间将所述一个或多个应用程序的执行挂起,并且

其中所述非易失性存储器是非易失性固态存储器，且所述非易失性固态存储器和所述易失性存储器两者均为单个双列直插存储器模块DIMM的一部分。

7.如权利要求6所述的方法，其特征在于，所述接收包括从处理电路接收，且所述方法进一步包括确定所述数据已被复制到所述非易失性固态存储器，以及通知所述处理电路所述数据已被复制到所述非易失性固态存储器。

8.如权利要求6所述的方法，其特征在于，所述指示描述所述数据在所述易失性存储器内存储的位置。

9.如权利要求6所述的方法，其特征在于，所述非易失性固态存储器和所述易失性存储器被组织成多个DIMM，其中，每个DIMM包括所述易失性存储器的不同部分以及所述非易失性固态存储器的不同部分；第一DIMM包括所述非易失性固态存储器的第一部分和所述易失性存储器的第一部分，且第二DIMM包括所述非易失性固态存储器的第二部分和所述易失性存储器的第二部分；

所述启动所述复制包括：从所述易失性存储器的所述第一部分到所述非易失性固态存储器的所述第一部分的在所述第一DIMM上的第一启动复制，以及并发地，从所述易失性存储器的所述第二部分到所述非易失性固态存储器的所述第二部分的在所述第二DIMM上的第二启动复制。

10.一种计算机系统，包括：

被配置成处理应用程序的指令的处理电路；

存储器模块，包括：

易失性存储器，其被配置成存储在所述应用程序的指令处理期间由所述处理电路生成的数据；以及

非易失性存储器，其被配置成从所述易失性存储器接收所述数据并存储所述数据；以及

其中所述处理电路被配置成响应于正被指示的检查点启动将所述数据从所述易失性存储器复制到所述非易失性存储器，并且在将所述数据从所述易失性存储器复制到所述非易失性存储器期间将所述应用程序的执行挂起，其中，所述检查点的指示由所述应用程序中的检查点指令提供，该检查点指令位于执行所述应用程序中一计算或者一功能的指令片断的末尾，并且

其中所述非易失性存储器是非易失性固态存储器，且所述非易失性固态存储器和所述易失性存储器两者均为单个双列直插存储器模块DIMM的一部分。

11.如权利要求10所述的系统，其特征在于，基于处理所述应用程序的所述指令的所述处理电路指示所述检查点。

12.如权利要求10所述的系统，其特征在于，所述存储器模块被配置成并行地将数据的不同部分并发复制到所述非易失性固态存储器。

13.如权利要求10所述的系统，其特征在于，所述处理电路进一步被配置成响应于在所述应用程序的执行期间正被检测的错误而启动将所述数据从所述非易失性固态存储器复制到所述易失性存储器。

14.如权利要求10所述的系统，其特征在于，所述处理电路被配置成经由大型互连与其他处理电路通信，所述其他处理电路也被配置成执行所述应用程序的所述指令。

15.如权利要求10所述的系统，其特征在于：

所述易失性存储器包括多个集成电路芯片，所述多个集成电路芯片中的每一个存储所述数据的不同部分；以及

所述处理电路被配置成并发地启动将所述数据的多个部分从所述多个集成电路芯片复制到所述非易失性固态存储器。

16.如权利要求10所述的系统，其特征在于：

所述存储器模块包括多个DIMM，每个DIMM包括所述易失性存储器的不同部分以及所述非易失性存储器的不同部分；以及

所述多个DIMM中的各个DIMM被配置成独立于所述多个DIMM中的其他DIMM将存储于所述各个DIMM的易失性存储器部分中的数据复制到所述各个DIMM的所述非易失性固态存储器部分，其中，所述处理电路被配置成，并发地，启动将第一DIMM中所述易失性存储器的第一部分保存的数据复制到所述第一DIMM中所述非易失性固态存储器的第一部分，并启动将第二DIMM中所述易失性存储器的第二部分保存的数据复制到所述第二DIMM中所述非易失性固态存储器的第二部分。

将检查点数据存储于非易失性存储器中

技术领域

[0001] 本发明的方面涉及将检查点数据存储于非易失性存储器中。

背景技术

[0002] 随着半导体制造技术持续缩小至越来越小的特征尺寸，硬件的故障率预期增加。至少两种类型的故障有可能发生：瞬态错误，其可以是临时的，但持续少时间量；硬错误，其可以是永久的。瞬态错误可具有诸多起因。示例瞬态错误包括，由电源波动、热效应、α粒子撞击引起的晶体管故障，以及由串扰、环境噪声和/或信号完整性问题所致的线路故障。硬错误的起因包括，例如，由工艺差异和过热的组合引起的晶体管故障，以及由工艺瑕疵或由超过线路材料临界电流所引起的金属迁移为原因的线路故障。

[0003] 硬错误和临时错误两者均可使用细或者大粒度级的冗余机制在内部加以纠正。细粒度机制包括存储器组件中的纠错码、分组传输信道上的循环冗余码、以及磁盘系统中的擦除编码机制。大粒度机制包括配置多个处理器以执行相同指令，然后比较来自多个处理器的执行结果以确定正确结果。在这种情况下，为了检测错误，执行相同指令的处理器的数量要为两个或两个以上。若处理器的数量为两个，错误可被检测。若处理器的数量为三个或三个以上，错误既可被检测又可被纠正。然而使用这种冗余机制对于大型并行系统可能价格过高。

[0004] 大型并行系统可包括处理器集群，其执行单个长期应用程序。在一些情况下，大型并行系统可包括数百万个集成电路，其在数日或数周里执行该单个长期应用程序。这些大型并行系统可通过在一个或多个磁盘上存储应用程序的中间状态来周期性地对该应用程序建立检查点。当发生故障时，计算可回退并且从最近记录的检查点重新开始，而非从计算的起点重新开始，其可能节省数小时或者数日的计算时间。

[0005] 因此，至少在一些计算设备(例如，大型并行系统)中使用检查点建立随着半导体制造技术的特征尺寸的减小和故障率的上升会变得越来越重要。已知系统将检查点数据写入磁盘。然而，磁盘带宽和磁盘存取时间可能没有发展得足够快以跟上计算系统的需求。此外，使用如磁盘等机械介质建立检查点数据时所消耗的功率量是重要缺点。

发明内容

[0006] 根据本发明的一些方面，描述了用于在非易失性存储器中存储检查点数据的方法和系统。

[0007] 根据一个方面，数据存储方法包括使用处理电路执行应用程序，并在执行期间将由应用程序的执行所生成的数据写入易失性存储器。该方法亦包括将数据写入易失性存储器之后提供检查点的指示(例如，检查点完成的指示)。在已提供检查点指示之后，该方法包括将数据从易失性存储器复制到非易失性存储器，并在复制之后继续执行该应用程序。在一些实施例中，非易失性存储器可以是固态存储器和/或随机存取存储器。

[0008] 在继续执行之后，在一些实施例中，该方法可包括检测应用程序执行中的错误。响

应于该检测，将数据从非易失性存储器复制到易失性存储器。接下来，可使用存储于易失性存储器中的所复制的数据从检查点开始执行应用程序。

[0009] 根据另一方面，数据存储方法包括接收与一个或多个应用程序的执行相关联的检查点的指示，并响应于该接收启动将一个或多个应用程序的执行所生成的数据从易失性存储器复制到非易失性存储器。在一些实施例中，该指示可描述易失性存储器内存储该数据的位置。

[0010] 根据另一个方面，计算机系统包括处理电路和存储器模块。处理电路配置成处理应用程序的指令。存储器模块可包括配置成存储应用程序指令的处理期间由处理电路生成的数据的易失性存储器。存储器模块亦可包括配置成接收来自易失性存储器的数据并存储该数据的非易失性存储器。在一个实施例中，处理电路配置成响应于被指示的检查点启动将数据从易失性存储器复制到非易失性存储器。

[0011] 在一个实施例中，非易失性存储器和易失性存储器可被组织成一个或多个双列直插内存模块(DIMM)，使得单个DIMM包括非易失性存储器的全部或部分和易失性存储器的全部或部分。在一个实施例中，非易失性存储器可包括多个集成电路芯片，且数据的复制可包括并发地将数据的第一子集复制到多个集成电路芯片中的第一个以及将数据的第二子集复制到多个集成电路芯片中的第二个。

[0012] 如根据下文讨论显而易见地，描述了其他实施例和诸方面。

[0013] 附图描述

[0014] 图1是根据一个实施例的处理系统的框图。

[0015] 图2是根据一个实施例的计算机系统的框图。

[0016] 图3是根据一个实施例的存储器模块的框图。

[0017] 图4是根据一个实施例的处理系统的框图。

[0018] 具体描述

[0019] 本发明涉及诸如处理系统、计算机、处理器和计算机系统的装置以及包括将检查点数据存储于非易失性存储器中的方法的方法。根据本发明的一些方面，使用处理电路执行应用程序。在一个实施例中，当应用程序的执行到达检查点时，应用程序的进一步执行可被挂起。存储于易失性存储器中的涉及应用程序的数据可被复制到非易失性存储器中。在一些实施例中，非易失性存储器可以是诸如NAND闪存或者相变存储器等固态非易失性存储器。非易失性存储器附加地或者替代地可以是随机存取存储器。

[0020] 在一些实施例中，一旦已复制了数据，应用程序的执行可继续。若应用程序执行期间出现错误，则可将存储于非易失性存储器中的数据复制回易失性存储器中。一旦数据已恢复到易失性存储器，应用程序可从检查点重启动。下文中讨论其他或者替代实施例。

[0021] 参考图1，示出根据一个实施例的处理系统100。系统100包括处理电路102、存储器模块106以及磁盘存储装置108。提供图1的实施例以示出一个可能的实施例，且包括较少、较多或者替代组件的其他实施例是可能的。附加地，图1的一些组件可组合。

[0022] 在一个实施例中，系统100可以是单个计算机。在这个实施例中，处理电路102可包括一个处理器110，但可能不包括互连114且可能不与大型互连122通信，两者均用幻影示出且在下文中进一步讨论。在这个实施例中，处理器110可以是单核处理器或者多核处理器。

[0023] 在另一实施例中，系统100可以是处理器集群。在这个实施例中，处理电路102可包

括多个处理器。虽然图1中只示出两个处理器——处理器110和处理器112，但处理电路102可包括两个以上的处理器。在一些情况下，处理电路102的处理器可并发执行单个应用程序。因此，可并行执行应用程序。在这个实施例中，处理电路102可包括互连114，其允许处理器110和112之间的通信以及允许应用程序执行的协作。此外，在各个实施例中，如下文中根据图2将要讨论，处理电路102可经由大型互连122与其它处理器集群(其亦可执行该应用程序)通信。

[0024] 在一个实施例中，存储器模块106包括易失性存储器116和非易失性存储器118。易失性存储器116可存储由处理电路102生成的数据以及从磁盘存储装置108取回的数据。在本文中这些数据称作应用程序数据。易失性存储器116可按照使用电子、磁、光、电磁或者用于存储信息的其他技术的诸多不同方式来具体化。一些具体示例包括但不限于DRAM以及SRAM。在一个实施例中，易失性存储器116可存储由处理电路102实现的程序。

[0025] 非易失性存储器118存储从易失性存储器116接收的检查点数据。检查点数据可以与应用程序数据相同，或者检查点数据可以是应用程序数据的子集。在一些实施例中，即使不向非易失性存储器118供电，非易失性存储器118也可永久地存储检查点数据。如上所述，在一个实施例中，应用程序数据和检查点数据存储于存储器中。存储于存储器中包括将数据存储于集成电路存储介质中。在一个实施例中，非易失性存储器118可以是固态和/或随机存取非易失性存储器(例如，NAND闪存、FeRAM(铁磁RAM)、MRAM(磁阻RAM)、PCRAM(相变RAM)、RRAM(阻抗RAM)、探针存储装置、以及NRAM(纳米管RAM))。在一个实施例中，从非易失性存储器118读取检查点数据不使用移动部件。在另一实施例中，非易失性存储器118可按照随机顺序存取。此外，无关于数据在非易失性存储器118内的物理位置、无关于数据与之前存取的数据是否相关，非易失性存储器118可在基本恒定时间里返回数据。

[0026] 在一个实施例中，处理电路102包括检查点管理模块104。在一个实施例中，检查点管理模块104被配置成控制和实现检查点操作。例如，检查点管理模块104可控制将检查点数据从易失性存储器116复制到非易失性存储器118，以及控制将检查点数据从非易失性存储器118复制到易失性存储器116。在一个实施例中，检查点管理模块104可包括诸如处理器的处理电路。在其他实施例中，检查点管理模块104可具体化于处理器110和/或处理器112中(例如作为微码或者软件)。

[0027] 作为示例，处理电路102可执行由磁盘存储装置108(例如，一个或多个硬盘)存储的应用程序。该应用程序可包括多个指令。一些或者全部指令可从磁盘存储装置108被复制到易失性存储器116中。一些或者全部指令然后可从易失性存储器116转移到处理电路102，以使处理电路102可处理指令。处理指令的结果是，处理电路102可从易失性存储器116或者磁盘存储装置108取回应用程序数据和/或将应用程序数据写入易失性存储器116或者磁盘存储装置108。因此，随着处理电路102处理应用程序指令，易失性存储器116和/或磁盘存储装置108的内容可改变。

[0028] 易失性存储器116在特定时间点的一些或者全部内容可保留作为检查点数据。例如，在处理电路102处理应用程序的一个或多个初始指令之后，可将存储于易失性存储器116中的检查点数据(其可以是应用程序数据的全部或者子集)复制到不同于易失性存储器116的位置。一旦已复制了检查点数据，处理电路102可继续处理应用程序的一个或多个后续指令。随后，可确定处理初始指令之后在执行应用程序时发生了错误。为了从错误中恢

复，可将所存储的检查点数据恢复到易失性存储器116，且处理电路102可从后续指令开始重启动应用程序的执行。

[0029] 在一个实施例中，检查点管理模块104可管理检查点数据的存储。在一个实施例中，检查点管理模块104可接收来自处理电路102的与一个或多个应用程序的执行相关联的检查点指示。执行检查点操作的指示可由不同源提供和/或针对不同启动标准，如下文的说明性示例中所讨论。在处理电路102已将处理电路102的一个或多个高速缓存存储器(未示出)的内容转储清除到易失性存储器116之后，处理电路102可向检查点管理模块104提供指示。处理电路102中的各种实体中的一个或多个实体可提供指示。例如，操作系统、虚拟机、管理程序或者应用程序可生成检查点的指示。用于生成指示的其他标准源是可能的且在下文中讨论。

[0030] 响应于接收指示，检查点管理模块104可启动将由易失性存储器116存储的所有或者部分应用程序数据复制到非易失性存储器118。在一个实施例中，在向检查点管理模块104提供指示之前或提供指示之后，处理电路102可使要建立检查点的应用程序的执行挂起，以使要建立检查点的应用程序的应用程序数据在将检查点数据从易失性存储器116复制到非易失性存储器118的同时不改变。

[0031] 在一些实施例中，处理电路102可向易失性存储器116和非易失性存储器118写入应用程序数据。在其他实施例中，处理电路102可向易失性存储器116写入应用程序数据，但是可能不能够向非易失性存储器118写入应用程序数据。不过，可将检查点数据从易失性存储器116复制到非易失性存储器118。因此，为了向非易失性存储器118写入检查点数据，可能先要向易失性存储器116写入检查点数据。

[0032] 易失性存储器116和非易失性存储器118的相对容量可按照任何适当的配置进行配置。例如，因为错误可能正好在检查点操作完成之前发生，一个实施例的非易失性存储器118可具有至少为易失性存储器116容量两倍的容量，以使非易失性存储器118可存储两组检查点数据。此外，至少在一个实施例中，可将对应于不同检查点的诸多不同检查点数据并发地存储于非易失性存储器118中。

[0033] 检查点指示可指定由易失性存储器116存储的应用程序数据的哪些部分为检查点数据。例如，该指示可指出由易失性存储器116存储的基本上所有应用程序数据为检查点数据，仅相关于特定应用程序的应用程序数据为检查点数据，和/或易失性存储器116的特定位置内的应用程序数据为检查点数据。在一个实施例中，该指示可包括描述检查点数据的保存向量(save vector)。

[0034] 在一个实施例中，处理电路102可通过控制易失性存储器116和非易失性存储器118来实现将检查点数据从易失性存储器116复制到非易失性存储器118。例如，处理电路102可向易失性存储器116和非易失性存储器118提供控制信号和指令。在另一实施例中，检查点管理模块104可通过控制存储器116和118来实现检查点数据的复制。一旦检查点数据已被成功地复制到非易失性存储器118，检查点管理模块104可通知处理电路102。

[0035] 在另一实施例中，存储器模块106可包括单独的处理电路(未示出)，且处理电路102或检查点管理模块104可向该处理电路提供描述检查点数据的信息(例如，易失性存储器116内存储检查点数据的位置)，并指示该处理电路将检查点数据复制到非易失性存储器118。一旦检查点数据已被成功地复制到非易失性存储器118中，存储器模块106的处理电路

可通知检查点管理模块104和/或处理电路102。

[0036] 在确定检查点数据已被成功地复制到非易失性存储器118之后，检查点控制模块104可通知处理电路102检查点数据已被复制到非易失性存储器118。作为响应，处理电路102可继续执行处理电路102之前在将检查点数据复制到非易失性存储器118时已挂起的应用程序。在应用程序执行期间，系统100可多次重复上述将检查点数据存储到非易失性存储器118中的方法。

[0037] 如上所述，可使用多种方法来确定何时应当生成检查点。根据一个方法，可周期性地存储检查点，并且可为由处理电路102正在执行的多个应用程序存储检查点。在这个实施例中，如上所述，处理电路102(例如，经由由处理电路102执行的操作系统、虚拟机、管理程序等)可周期性地向检查点管理模块104指示检查点。在一些示例中，检查点操作的周期可通过定时器中断来控制或者由周期性操作系统干预来控制。在一个实施例中，可将由易失性存储器116存储的基本上所有的应用程序数据复制到非易失性存储器118。替代地，可将相关于正在由处理电路102执行的仅一个应用程序的应用程序数据复制到非易失性存储器118。这个方法可称作自动建立检查点。

[0038] 根据另一个方法，正由处理电路102执行的应用程序可确定何时应当生成检查点数据。在一个实施例中，应用程序可指定哪些应用程序数据应被存储为检查点数据以及何时存储该检查点数据。在一个实施例中，应用程序可包括检查点指令。检查点指令可位于整个应用程序之中，以使应用程序被分割成由检查点指令界定的指令片断。在一个实施例中，检查点指令可位于执行特定计算或者功能的指令片断的末尾。例如，若应用程序为更新账户余额的银行应用程序，则该应用程序可包括紧跟在更新账户余额的指令之后的检查点指令。在另一实施例中，应用程序可响应于要遇上的情形请求生成检查点数据。这个方法可称作应用程序检查点建立。

[0039] 在存储检查点数据和继续执行应用程序之后，处理电路102和/或检查点管理模块104可检测应用程序执行中的错误(例如，经冗余计算检查)。在一个实施例中，根据错误的检测，处理电路102可将应用程序的进一步执行挂起。

[0040] 为了从错误中恢复，从与存储于非易失性存储器118中的检查点数据相关联的检查点处开始重新执行该应用程序。响应于错误的检测，检查点管理模块104可将检查点数据从非易失性存储器118复制到易失性存储器116。一旦已将检查点数据复制到易失性存储器116，检查点管理模块104可通知处理电路102。处理电路102然后可使用该检查点数据从检查点开始重新执行该应用程序，其中处理电路102现在能够使用易失性存储器116中的该检查点数据。

[0041] 在一个实施例中，该检查点数据可以是多个应用程序的检查点数据，且所检测的错误会影响所有的多个应用程序。在这个实施例中，一旦已恢复检查点数据，则可从该检查点处开始重新执行多个应用程序中的每一个。

[0042] 参考图2，示出大型计算机系统200。系统200包括关于图1在上文中所描述的多个处理系统100。在一个实施例中，系统100可用于并行地执行单个应用程序或者用于执行不同应用程序。相比于在一个处理器或者一个处理器集群上执行单个应用程序，并行地执行单个应用程序可提供显著的速度优势。系统200可包括附加的处理系统，为简单起见未示出。

[0043] 在一个实施例中，系统200还包括管理节点204、大型互连122、I/O节点206、网络208以及存储电路210。在一个实施例中，管理节点204可确定单个应用程序的哪一部分要由处理系统执行。管理节点204可经由大型互连122与处理系统100通信。

[0044] 在应用程序的执行期间，处理系统100和/或处理系统202可将数据存储于存储电路210中。为了实现该目的，处理系统可经大型互连122和I/O节点206向存储电路210发送数据。类似地，处理系统可经大型互连122和I/O节点206从存储电路210取回数据。例如，处理系统100可将数据从磁盘存储装置108移动到存储电路210，存储电路210相比于磁盘存储装置108具有较大容量。在一些实施例中，处理系统100和202可经由I/O节点206和网络208与其他计算机系统通信。在一个实施例中，网络208可以是因特网。

[0045] 在一个实施例中，存储电路210可包括非易失性存储器，并且管理节点204可启动将检查点数据从处理系统100经由大型互连122复制到存储电路210的非易失性存储器。

[0046] 现在回到图1，存储器模块106可配置成并行地将存储于易失性存储器116的检查点数据的不同部分并发复制到非易失性存储器118中，而非串行地复制检查点数据。这样做可显著减少用于将检查点数据从易失性存储器116复制到非易失性存储器118的时间量。

[0047] 参考图3，示出存储器模块106的一个实施例。所公开的实施例仅仅是示意性的且其他实施例是可能的。所描绘的实施例中，存储器模块106包括三个双列直插存储器模块(DIMM)302、304和306。当然，存储器模块106可包括少于或者多于三个DIMM，为了简单起见示出三个DIMM。替代地或者附加地，存储器模块106可包括不同于DIMM的其他形式的存储器。

[0048] DIMM 302、304和306的每一个可包括易失性存储器116的一部分和非易失性存储器118的一部分。如图3所示，DIMM 302包括易失性存储器(VM)308和非易失性存储器(NVM)310，DIMM 304包括易失性存储器(VM)312和非易失性存储器(NVM)314，以及DIMM 306包括易失性存储器(VM)316和非易失性存储器(NVM)318。易失性存储器308、312和316可各为图1的易失性存储器116的不同部分。类似地，非易失性存储器310、314和318可各为图1的非易失性存储器118的不同部分。

[0049] 在一个实施例中，DIMM 302、304和306可各为不同电路板。此外，易失性存储器308、312和316可各包括多于一个集成电路，且非易失性存储器310、314和318可各包括多于一个集成电路。相应地，例如，DIMM可包括构成易失性存储器308的多个易失性存储器集成电路，以及构成非易失性存储器310的多个非易失性存储器集成电路。

[0050] DIMM 302、304和306可各存储不同的应用程序数据。因此，当遇到检查点时，检查点管理模块104可启动将检查点数据从易失性存储器308复制到非易失性存储器310，从易失性存储器312复制到非易失性存储器314，以及从易失性存储器316复制到非易失性存储器318。在一个实施例中，检查点管理模块104可使用完全缓冲的DIMM控制协议与DIMM 302、304和306通信。

[0051] 在一个实施例中，检查点管理模块104和/或处理电路102可个别地与每个DIMM 302、304和306通信，以启动将检查点数据从易失性存储器116复制到非易失性存储器118。DIMM 302可独立于DIMM 304和306在易失性存储器308和非易失性存储器310之间复制数据。实际上，可将检查点数据的第一部分从易失性存储器308复制到非易失性存储器310，同时将检查点数据的第二部分从易失性存储器312复制到非易失性存储器314，且同时将检查

点数据的第三部分从易失性存储器316复制到非易失性存储器318。这样做可能显著地快于等待直到第一部分已复制才复制检查点数据的第二部分，以及等待直到第二部分已复制才复制检查点数据的第三部分。

[0052] 当将检查点数据从非易失性存储器118恢复到易失性存储器116时，可使用类似方法。根据这个方法，检查点管理模块104和/或处理电路102可个别地与每个DIMM 302、304和306通信，以启动将检查点数据从非易失性存储器118复制到易失性存储器116。并发地，可将检查点数据的第一部分从非易失性存储器310复制到易失性存储器308，并将检查点数据的第二部分从非易失性存储器314复制到易失性存储器312，以及可将检查点数据的第三部分从非易失性存储器318复制到易失性存储器316。

[0053] 参考图4，示出处理系统100的替代实施例作为系统100a。在这个实施例中，处理电路102包括处理器110、112以及互连114，如同图1所示的处理电路102的实施例。此外，处理电路102包括北桥402和南桥404，其可各自包括相应的处理器。

[0054] 北桥402可经由互连114从处理器110和112接收控制和/或数据事务。对于每个事务，北桥402可确定该事务是否去往存储器模块106、磁盘存储装置108或者大型互连122。若事务去往存储器模块106，北桥402可将该事务转发到存储器模块106。若事务去往磁盘存储装置108或大型互连122，北桥402可将该事务转发到南桥404，然后南桥404可向磁盘存储装置108或者大型互连122两者转发该事务。南桥404可将该要求转变成适合于磁盘存储装置108或者大型互连122两者的协议。

[0055] 在一个实施例中，北桥402包括检查点管理模块104。在这个实施例中，检查点管理模块104可存储指令，该指令被转移到处理器110和/或处理器112以供执行。替代地或者附加地，北桥401可包括控制逻辑，该逻辑实现检查点管理模块104的全部或部分。替代地，在另一实施例中，检查点管理模块104可被实现为由处理器110和/或处理器112处理的指令(例如，隐藏管理程序或者固件)。

[0056] 与上文描述所公开的系统和方法对比，在发生错误时，不包括非易失性存储器的其他计算机系统可将检查点数据从易失性存储器复制到磁盘存储装置，以及可将检查点数据从磁盘存储装置取回到易失性存储器。与这些其他计算机系统相比，将检查点数据存储于非易失性存储器中而非存储于磁盘存储装置中可提供一些优势。

[0057] 在一个实施例中，由于非易失性存储器比磁盘存储装置快很多，因此向非易失性存储器存储检查点数据可能比向磁盘存储装置存储检查点数据快一个以上数量级。此外，可在易失性存储器和非易失性存储器之间并行地复制检查点数据。

[0058] 由于易失性存储器和非易失性存储器之间的物理距离比易失性存储器和磁盘存储装置之间的物理距离小，因此相比于在磁盘存储装置中存储检查点数据，在非易失性存储器中存储检查点数据可能消耗较少能量。这一较短物理距离亦可减少等待时间。此外，与在磁盘存储装置中存储检查点数据相比，在非易失性存储器中存储检查点数据可能消耗较少能量，因为与磁盘存储装置相比，非易失性存储器可能不包括移动部件。

[0059] 因为从非易失性存储器恢复检查点所用时间量可显著少于从磁盘存储装置恢复检查点所用时间量，所以作为向磁盘存储装置写入检查点数据的替代方案，向非易失性存储器写入检查点数据的结果是，处理器系统或处理器集群的可用度可增加。此外，因为磁盘存储装置存在机械错误模式(由于移动部件的使用)，而非易失性存储器不存在该问题，所

以与在磁盘存储装置中存储检查点数据相比，在非易失性存储器中存储检查点数据可导致较少错误。

[0060] 在一个实施例中，处理器系统的可用度计算可涉及处理器系统的未计划的停机时间量。在错误检测之后将检查点数据恢复到易失性存储器所花费的时间可看作未计划的停机时间。因为将检查点数据从非易失性存储器恢复到易失性存储器比将检查点数据从磁盘存储装置恢复到易失性存储器快，所以当向非易失性存储器建立检查点时的未计划停机时间量可能小于向磁盘存储装置建立检查点时的未计划停机时间量。

[0061] 处理器系统的一个示例可用度等式可以是：可用度＝1/(1+错误率×未计划停机时间)。作为示例，若每年发生1000个错误，且当从磁盘存储装置恢复检查点数据时的每个错误的停机时间为3秒，则处理器系统的可用度可大于99.99％而小于99.999％，且因此可称为具有“4个9”可靠性。相反，使用非易失性存储器时，若当从非易失性存储器恢复检查点数据时的每个错误的停机时间为300毫秒，则系统的可用度可大于99.999％而小于99.9999％，且因此可称为具有“5个9”可靠性。

[0062] 除了减少处理器系统的未计划停机时间，向非易失性存储器而非向磁盘存储装置写入检查点数据亦可减少处理器系统的未计划停机时间量。如上文所讨论，在将检查点数据写入非易失性存储器的同时，可将处理器系统对应用程序的执行挂起。应用程序挂起的时间量可看作处理器系统的计划停机时间。相比于向磁盘存储装置写入检查点数据，向非易失性存储器写入检查点数据可显著减少处理器系统的计划停机时间量，因为向非易失性存储器写入检查点数据要求较少时间。

[0063] 所寻求的保护不限于仅作为示例给出的公开的实施例，相反，所寻求的保护仅由所附权利要求的范围限定。

[0064] 此外，本文的诸方面为了本发明公开的说明性实施例的构造和/或操作中的指导而呈现。本文的申请人认为这些描述的说明性实施例亦包括、公开以及描述了除那些明确揭示的内容以外的进一步发明性方面。例如，与所描述的那些说明性实施例相比，附加发明性方面可包括更少、更多和/或替代特征。在更具体示例中，申请人认为该公开包括、揭示和描述了相比于那些明确揭示的方法包括较少、较多和/或替代步骤的方法，以及相比于明确揭示的结构包括更少、更多和/或替代结构的装置。

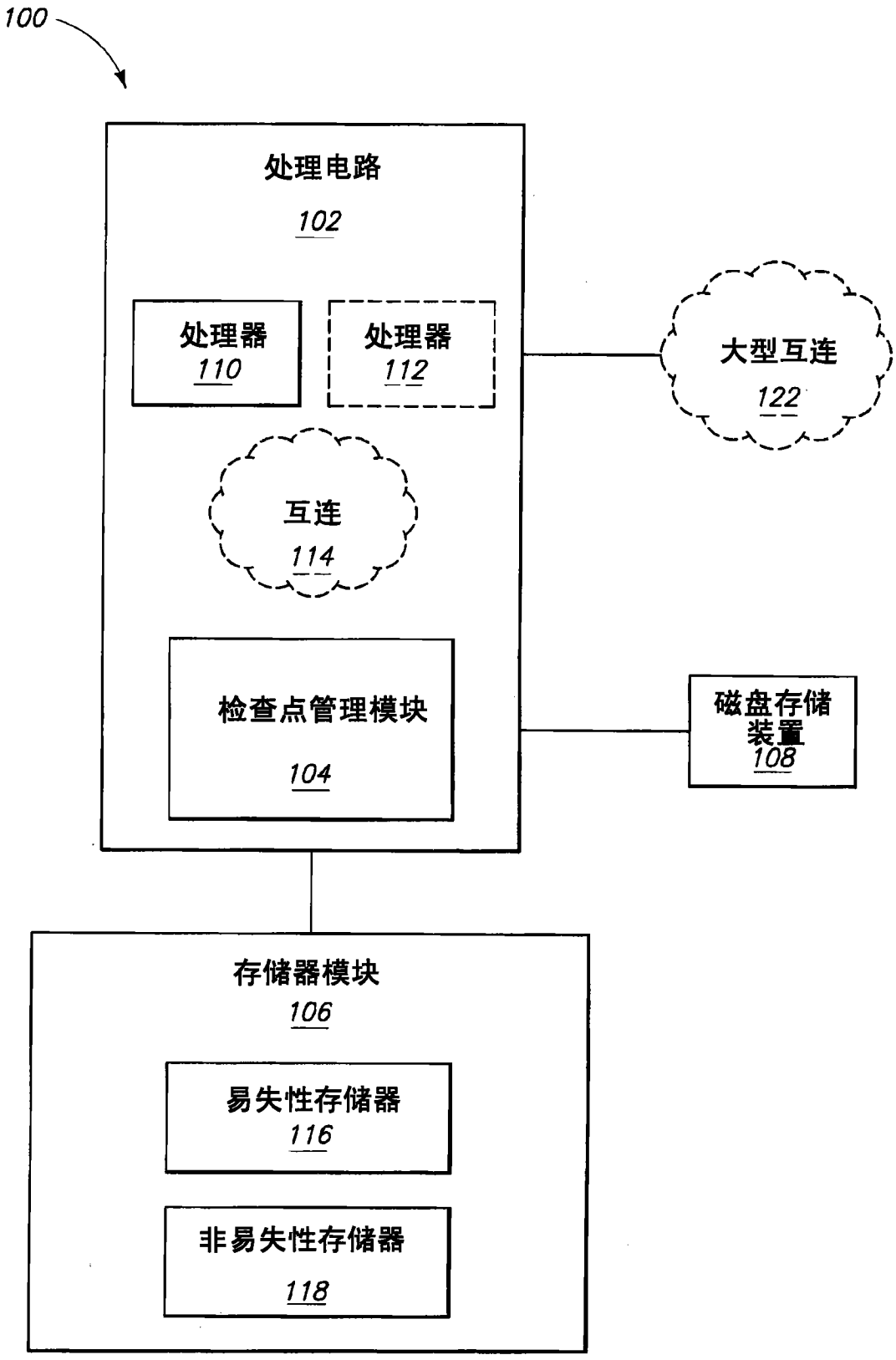

100
处理电路
102
处理器
110
处理器
112
互连
114
检查点管理模块
104
大型互连
122
磁盘存储装置
108
存储器模块
106
易失性存储器
116
非易失性存储器
118

图1

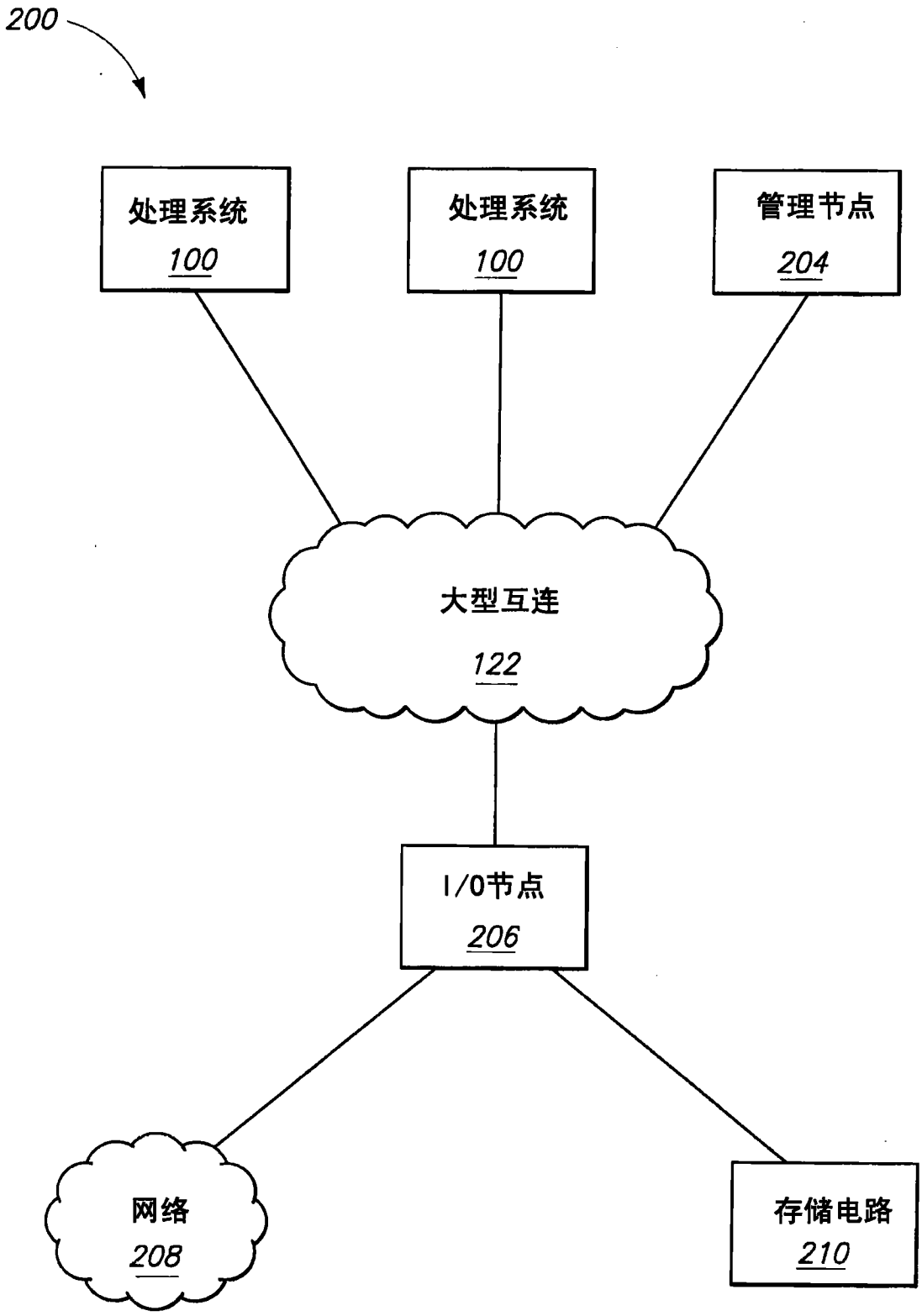

200
处理系统
100
处理系统
100
管理节点
204
大型互连
122
I/O节点
206
网络
208
存储电路
210

图2

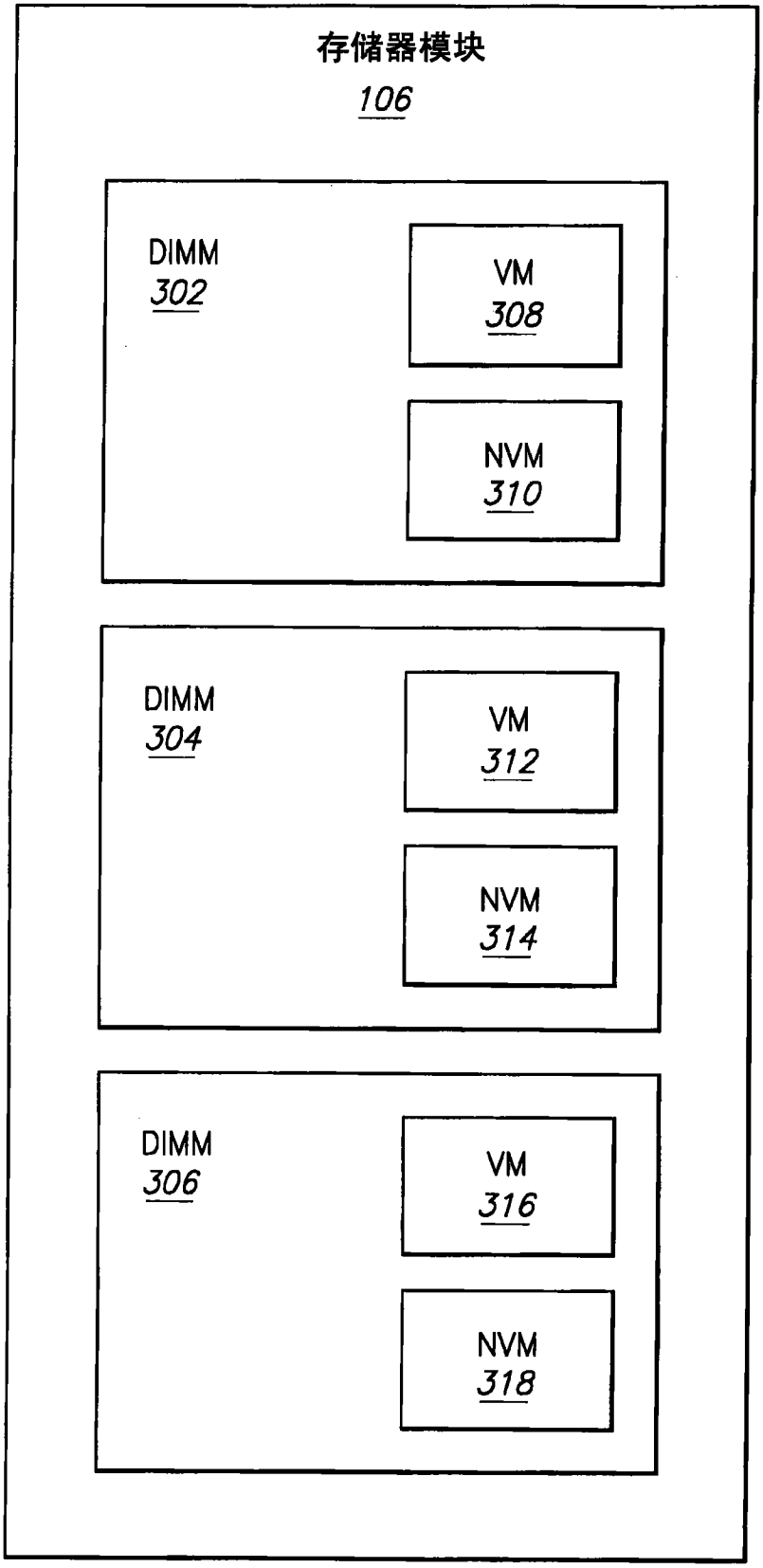

存储器模块
106
DIMM
302
VM
308
NVM
310
DIMM
304
VM
312
NVM
314
DIMM
306
VM
316
NVM
318

图3

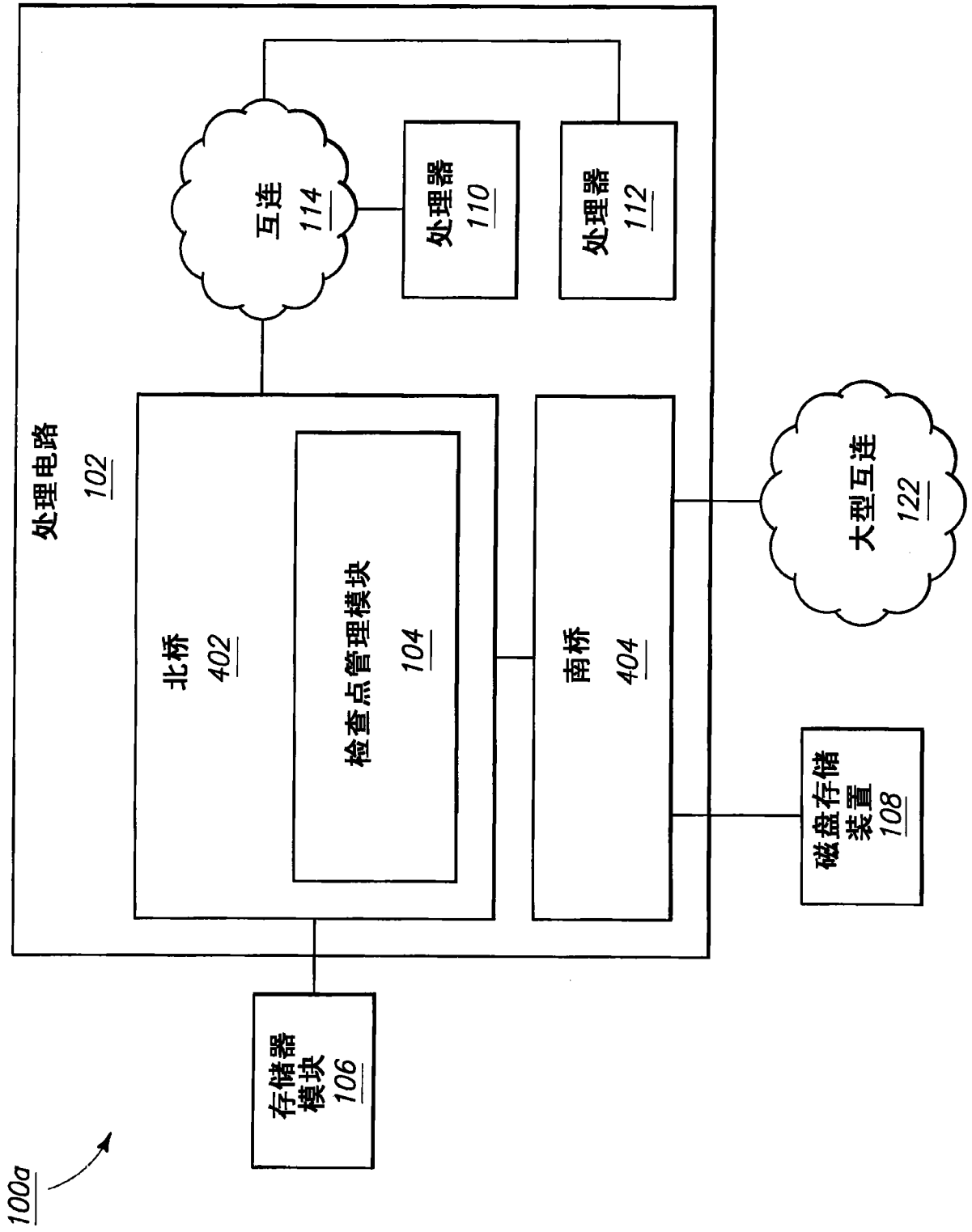

处理电路
102
互连
114
处理器
110
处理器
112
北桥
402
检查点管理模块
104
南桥
404
大型互连
122
磁盘存储装置
108
存储器模块
106
100a

图4