



(12) 发明专利

(10) 授权公告号 CN 111475112 B

(45) 授权公告日 2023. 03. 14

(21) 申请号 202010250410.7

(22) 申请日 2020.04.01

(65) 同一申请的已公布的文献号

申请公布号 CN 111475112 A

(43) 申请公布日 2020.07.31

(73) 专利权人 苏州浪潮智能科技有限公司

地址 215100 江苏省苏州市吴中区吴中经

济开发区郭巷街道官浦路1号9幢

(72) 发明人 张一可 段利宁 张远斌

(74) 专利代理机构 济南舜源专利事务有限公

司 37205

专利代理师 刘雪萍

(51) Int.Cl.

G06F 3/06 (2006.01)

G06F 16/21 (2019.01)

(56) 对比文件

CN 110502188 A, 2019.11.26

CN 110007870 A, 2019.07.12

审查员 吴昊

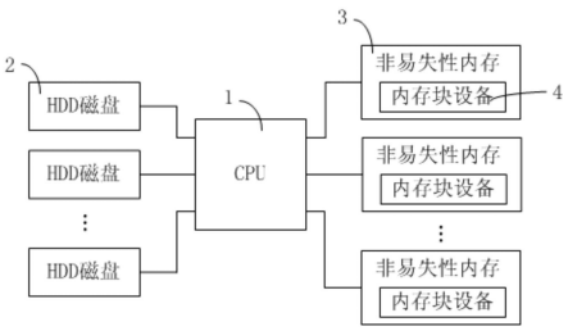
权利要求书1页 说明书4页 附图3页

(54) 发明名称

一种提升Oracle数据库性能的装置及数据
读写方法

(57) 摘要

本发明公开一种提升Oracle数据库性能的装置及数据读写方法,Oracle数据库配置于服务器,服务器的CPU上连接至少一个HDD磁盘和至少一个非易失性内存;HDD磁盘与非易失性内存一一对应,数量相同;一个HDD磁盘和一个非易失性内存配置成为一个块设备;非易失性内存上创建有内存块设备,该内存块设备被配置为对应HDD磁盘的缓存设备。写数据时,Oracle数据首先写到非易失性内存上,再以writeback方式回写到HDD磁盘上;读数据时,首先检测数据是否非易失性内存上,若在,则直接从非易失性内存上读取数据,否则继续从HDD磁盘上读取数据。本发明可有效提升数据库的读写能力,且降低服务器成本。



1. 一种提升Oracle数据库性能的装置, Oracle数据库配置于服务器, 其特征在于, 服务器的CPU上连接至少一个HDD磁盘和至少一个非易失性内存; HDD磁盘与非易失性内存一一对应, 数量相同;

一个HDD磁盘和一个非易失性内存配置成为一个块设备;

非易失性内存上创建有内存块设备, 该内存块设备被配置为对应HDD磁盘的缓存设备;

服务器上配置有Oracle ASM管理软件, 将每颗CPU下的所有块设备组成一组磁盘组, 进行数据条带化存储;

服务器上设置至少一对CPU, 每对中的两颗CPU所连接HDD磁盘数量相同; 每对CPU下的两个磁盘组互为镜像关系, Oracle ASM管理软件将每对CPU下的两个磁盘组设置为故障转移组;

CPU通过SAS卡连接HDD磁盘, 其中SAS卡连接到CPU的PCIe插槽上。

2. 根据权利要求1所述的提升Oracle数据库性能的装置, 其特征在于, 服务器上配置有块设备配置模块, 将非易失性内存上的内存块设备配置为缓存设备, 将HDD磁盘设置为后端设备, 从而将各个HDD磁盘及其对应非易失性内存配置成为各个块设备。

3. 根据权利要求2所述的提升Oracle数据库性能的装置, 其特征在于, 块设备配置模块为bcache缓存管理模块。

4. 根据权利要求3所述的提升Oracle数据库性能的装置, 其特征在于, 服务器上配置有内存块设备创建模块, 将非易失性内存设置为region模式, 将非易失性内存的region创建为内存块设备。

5. 根据权利要求4所述的提升Oracle数据库性能的装置, 其特征在于, 内存块设备创建模块包括ipmctl工具和ndctl工具;

ipmctl工具将非易失性内存设置为region模式; ndctl工具将非易失性内存的region创建为内存块设备。

6. 一种基于权利要求 1-5任一项所述装置的数据读写方法, 其特征在于, 包括以下步骤:

写数据时, Oracle数据首先写到非易失性内存上, 再以writeback方式回写到HDD磁盘上;

将Oracle热点数据持久缓存在非易失性内存上;

读数据时, 首先检测数据是否非易失性内存上, 若在, 则直接从非易失性内存上读取数据, 否则继续从HDD磁盘上读取数据。

7. 根据权利要求6所述的数据读写方法, 其特征在于, 还包括步骤:

设置读请求时间阈值和写请求时间阈值;

写数据时, 当IO的时间超过写请求时间阈值时, 绕过非易失性内存, 直接写入HDD磁盘;

读数据时, 当IO的时间超过读请求时间阈值时, 绕过非易失性内存, 直接读取HDD磁盘。

一种提升Oracle数据库性能的装置及数据读写方法

技术领域

[0001] 本发明涉及Oracle数据库领域,具体涉及一种提升Oracle数据库性能的装置及数据库数据读写方法。

背景技术

[0002] Oracle数据库(又名OracleRDBMS,是甲骨文公司的一款关系数据库管理系统)系统因其可移植性好、使用方便、功能强,适用于各类大、中、小、微机环境,成为目前流行的关系数据库管理系统。对于OLTP类型业务系统,一个系统的吞吐量(承压能力)与request对CPU的消耗、外部接口、IO等等紧密关联。单个request对CPU消耗越高,外部系统接口、IO响应速度越慢,系统吞吐能力越低,反之越高。随着CPU性能提升,Oracle数据库最大的性能瓶颈在于IO性能。仅使用HDD磁盘会严重的降低数据库的性能,为了解决该问题,业界普遍采用全闪磁盘(SSD磁盘阵列)阵列提升Oracle数据库存储IO性能,随着业务的数据量越来越大,被迫选择更大容量的SASSSD或NVMeSSD,随之又产生高昂的成本问题及性能过剩问题。

发明内容

[0003] 为解决上述问题,本发明提供一种提升Oracle数据库性能的装置及数据读写方法,基于非易失性内存,构建非易失性内存和HDD磁盘的多个块设备,提升Oracle数据库IO性能。

[0004] 本发明的技术方案是:一种提升Oracle数据库性能的装置,Oracle数据库配置于服务器,

[0005] 服务器的CPU上连接至少一个HDD磁盘和至少一个非易失性内存;HDD磁盘与非易失性内存一一对应,数量相同;

[0006] 一个HDD磁盘和一个非易失性内存配置成为一个块设备;

[0007] 非易失性内存上创建有内存块设备,该内存块设备被配置为对应HDD磁盘的缓存设备。

[0008] 进一步地,服务器上配置有块设备配置模块,将非易失性内存上的内存块设备配置为缓存设备,将HDD磁盘设置为后端设备,从而将各个HDD磁盘及其对应非易失性内存配置成为各个块设备。

[0009] 进一步地,块设备配置模块为bcache缓存管理模块。

[0010] 进一步地,服务器上配置有内存块设备创建模块,将非易失性内存设置为region模式,将非易失性内存的region创建为内存块设备。

[0011] 进一步地,内存块设备创建模块包括ipmctl工具和ndctl工具;

[0012] ipmctl工具将非易失性内存设置为region模式;ndctl工具将非易失性内存的region创建为内存块设备。

[0013] 进一步地,服务器上配置有Oracle ASM管理软件,将所有块设备组成一组磁盘组,进行数据条带化存储。

[0014] 进一步地,服务器上设置至少一对CPU,每对中的两颗CPU所连接HDD磁盘数量相同;Oracle ASM管理软件将每对CPU下的两个磁盘组设置为故障转移组。

[0015] 进一步地,CPU通过SAS卡连接HDD磁盘,其中SAS卡连接到CPU的PCIe插槽上。

[0016] 本发明的技术方案还包括一种基于上述装置的数据读写方法,包括以下步骤:

[0017] 写数据时,Oracle数据首先写到非易失性内存上,再以writeback方式回写到HDD磁盘上;

[0018] 将Oracle热点数据持久缓存在非易失性内存上;

[0019] 读数据时,首先检测数据是否在非易失性内存上,若在,则直接从非易失性内存上读取数据,否则继续从HDD磁盘上读取数据。

[0020] 进一步地,还包括步骤:

[0021] 设置读请求时间阈值和写请求时间阈值;

[0022] 写数据时,当IO的时间超过写请求时间阈值时,绕过非易失性内存,直接写入HDD磁盘;

[0023] 读数据时,当IO的时间超过读请求时间阈值时,绕过非易失性内存,直接读取HDD磁盘。

[0024] 本发明提供的提升Oracle数据库性能的装置及数据读写方法,在处理器下配置数量相同的非易失性内存和HDD磁盘,构建成为非易失性内存和HDD磁盘的多个块设备,写数据时Oracle数据首先写到非易失性内存上,再以writeback方式回写到HDD磁盘上,读数据时首先检测数据是否在非易失性内存上,若在,则直接从非易失性内存上读取数据,否则继续从HDD磁盘上读取数据,有效提升数据库的读写能力。另外,本发明HDD不需要通过RAID磁盘阵列卡组成RAID阵列,只需要单块盘使用,减少了RAID磁盘阵列校验带来的性能损耗;非易失性内存条和HDD磁盘故障时不影响系统正常运行;非易失性内存和HDD磁盘同处于一颗处理器下,数据的处理无需跨CPU;对于任何Oracle数据库一体机或者Oracle数据库均可使用该装置;且相比采用大容量的全闪磁盘阵列大大降低了服务器成本。

附图说明

[0025] 图1是本发明实施例一结构示意图。

[0026] 图2是本发明实施例一具体实现方式原理架构图。

[0027] 图3是本发明实施例二方法流程示意图。

具体实施方式

[0028] 下面结合附图并通过具体实施例对本发明进行详细阐述,以下实施例是对本发明的解释,而本发明并不局限于以下实施方式。

[0029] 如图1所示,本实施例提供一种提升Oracle数据库性能的装置,Oracle数据库配置于服务器上,为实现该装置,服务器的CPU1上连接至少一个HDD磁盘2和至少一个非易失性内存3;HDD磁盘2与非易失性内存3一一对应,数量相同。一个HDD磁盘2和一个非易失性内存3配置成为一个块设备5,其中非易失性内存3上创建有内存块设备4,该内存块设备4被配置为对应HDD磁盘2的缓存设备。需要说明的是,CPU1通过SAS卡连接HDD磁盘2,其中SAS卡连接到CPU1的PCIe插槽上。

[0030] 本实施例中,服务器上配置有块设备5配置模块,将非易失性内存3上的内存块设备4配置为缓存设备,将HDD磁盘设置为后端设备,从而将各个HDD磁盘及其对应非易失性内存配置成为各个块设备5。具体的,块设备5配置模块为bcache(系统内核)缓存管理模块。

[0031] 服务器上配置有内存块设备4创建模块,将非易失性内存3设置为region(区域)模式,将非易失性内存3的region创建为内存块设备4。具体地,内存块设备4创建模块包括ipmctl工具(Intel持久内存控制工具)和ndctl工具(non-volatile memory devices非易失性内存控制工具);ipmctl工具将非易失性内存3设置为region模式;ndctl工具将非易失性内存3的region创建为内存块设备4。

[0032] 如图2所示服务器为例,服务器有两颗CPU,在服务器的每颗CPU下连接一块SAS卡,每个SAS卡上连接1-6块HDD磁盘2,每颗CPU下配置与HDD磁盘2数量相等的非易失性内存3设备。

[0033] 在系统下安装ipmctl和ndctl工具将每个非易失性内存3设备配置为内存块设备4(pmem),如果非易失性内存设备数为12个,那么在系统下创建名为pmem0、pmem1...pmem11的内存块设备4。

[0034] 创建步骤:

[0035] 使用ipmctl将服务器的每个非易失性内存3的类型设置为每个非易失性内存3为一个region模式,如果有12个非易失性内存则对应会生成region0到region11:ipmctl create -goal PersistentMemoryType=AppDirectNotInterleaved

[0036] 使用ndctl工具将非易失内存的region创建为pmem设备:

[0037] ndctl create-namespace -region=region0

[0038] 在系统下安装激活内核bcache缓存管理模块,通过bcache模块的管理命令,将每个内存块设备4设置为前端缓存(caching)设备,每个HDD磁盘阵列设置为后端(backing)设备,最终生成一个bcache0、bcache1、bcache2、bcache3等块设备5。

[0039] 假设每个CPU1下配置6个HDD磁盘2和6个非易失性内存3,那么需要在CPU_0下创建名为bcache0~6的块(bcache)设备5,在CPU_1下创建名为bcache7~12的块(bcache)设备5。

[0040] 命令如下:

[0041] make-bcache --cache_replacement_policy=lru -C /dev/pmem0 -B /dev/sdb

[0042] 其中,--cache_replacement_policy为缓存算法,lru为最近最久未使用算法,即是在非易失性内存3上缓存热点数据。

[0043] -C参数指定非易失性内存3设备为caching设备,

[0044] -B参数指定HDD磁盘2为后端存储设备。

[0045] 本实施例中,服务器上配置有Oracle ASM(Automatic Storage Management自动存储管理)管理软件,将所有块设备5组成一组磁盘组,进行数据条带化存储。将Oracle数据均匀分布到生成的多个块设备5上。

[0046] 同时,本实施例还采用ASM多副本模式解决非易失性内存3和HDD磁盘2容错问题。具体地,服务器上设置至少一对CPU1,每对中的两颗CPU所连接HDD磁盘2数量相同。Oracle ASM管理软件将每对CPU下的两个磁盘组设置为故障转移组。

[0047] 以图2所示服务器连接两个CPU1为例,每颗CPU1下对应的一块HDD磁盘2和一个非易失性内存3组成1块设备5,通过OracleASM管理工具将每颗CPU1下的所有块设备5组成一

组磁盘组,冗余类型为双副本。假设磁盘组名为+DATA。磁盘组+DATA的双副本分别位于不同CPU一侧,即CPU_0和CPU_1一侧的HDD互为镜像组。也即是2个磁盘组互为镜像关系,CPU_1下的磁盘组设置为CPU_0下磁盘组的故障转移组,保证HDD磁盘2和非易失性内存3缓存层故障不影响数据层设备的数据。

[0048] 实施例二

[0049] 如图3所示,本实施例提供一种基于实施例一的数据读写方法,包括以下步骤:

[0050] 写数据时,Oracle数据首先写到非易失性内存3上,再以writeback方式回写到HDD磁盘2上;

[0051] 将Oracle热点数据持久缓存在非易失性内存3上;

[0052] 读数据时,首先检测数据是否在非易失性内存3上,若在,则直接从非易失性内存3上读取数据,否则继续从HDD磁盘2上读取数据。

[0053] 进一步,可设置读请求时间阈值和写请求时间阈值,写数据时,当IO的时间超过写请求时间阈值时,绕过非易失性内存3,直接写入HDD磁盘2;读数据时,当IO的时间超过读请求时间阈值时,绕过非易失性内存3,直接读取HDD磁盘2。保证所有的离散读和写都会经过非易失性内存3,提高数据块在非易失性内存3上的命中率。

[0054] 具体实施例,该方法基于bcache内核模块参数实现,具体可调整以下参数:

[0055] 设置缓存策略为writeback回写策略,也就是数据流先经过非易失性内存3,非易失性内存3上的数据异步写入到HDD磁盘2;

[0056] 设置顺序IO缓存策略参数sequential_cutoff=0,意味着所有顺序数据块都要经过非易失性内存3;

[0057] 调整缓存写回比例参数 writeback_percent,设置在非易失性内存3上保留热点数据的比例,根据需要可设置最大在非易失性内存3上保留40%的热点数据;

[0058] 设置 congested_read_threshold_us和congested_write_threshold_us参数,默认情况下当读请求阈值2ms,写请求阈值为20ms,如果IO的时间超过阈值,则绕过非易失性内存3缓存设备,从而不会导致数据cachemissing问题,保证所有的离散读和写都会经过非易失性内存3,提高数据块在非易失性内存3上的命中率。

[0059] 以上公开的仅为本发明的优选实施方式,但本发明并非局限于此,任何本领域的技术人员能思之的没有创造性的变化,以及在不脱离本发明原理前提下所作的若干改进和润饰,都应落在本发明的保护范围内。

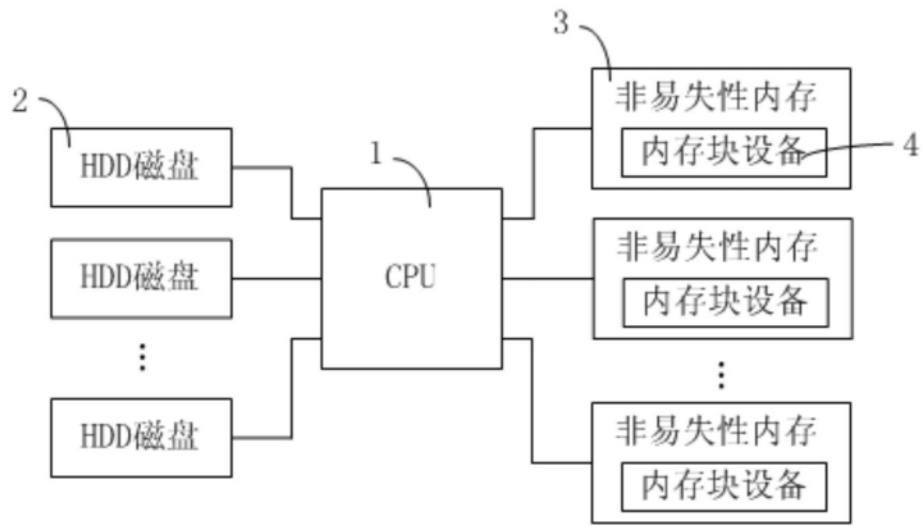


图1

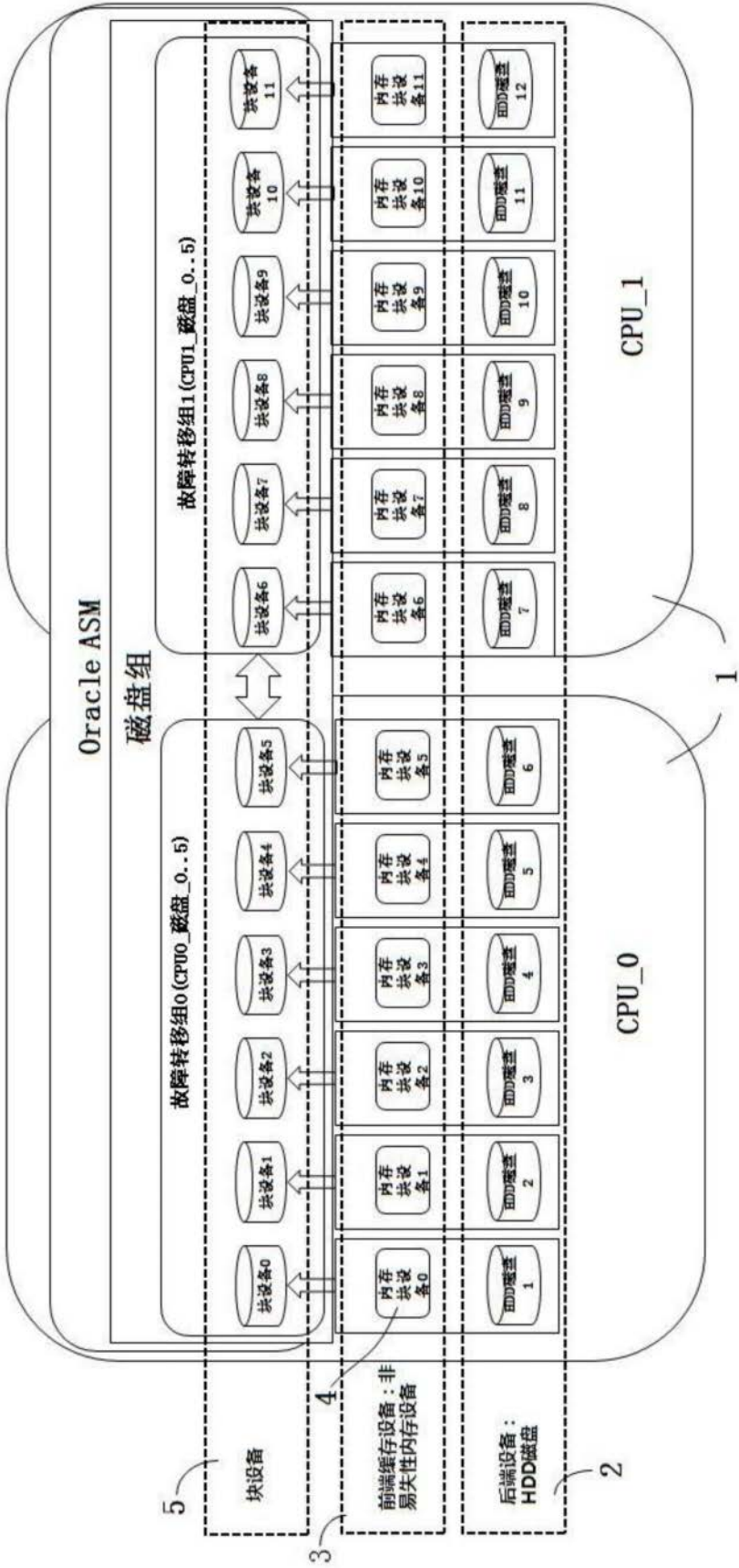


图2

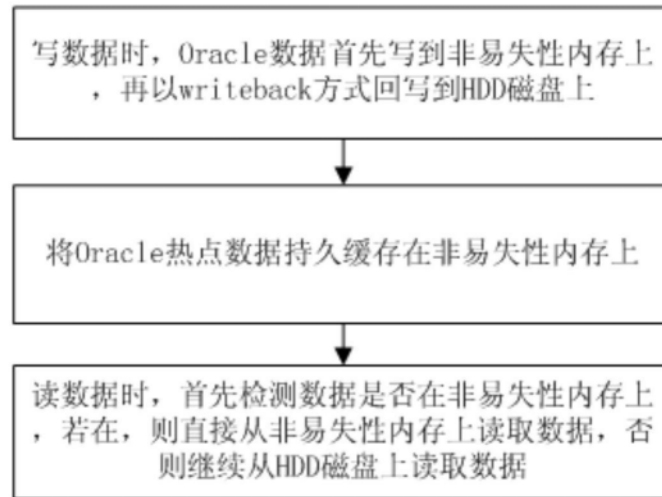


图3