(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118295936 B
(45)授权公告日 2024.08.02

(21)申请号 202410726081.7
(22)申请日 2024.06.06
(65)同一申请的已公布的文献号
申请公布号 CN 118295936 A
(43)申请公布日 2024.07.05
(73)专利权人 北京开源芯片研究院
地址 100084 北京市海淀区海淀大街31号3层312
(72)发明人 刘宇航 周嘉鹏 陈泓佚
(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
专利代理师 莎日娜
(51)Int.Cl.
G06F 12/122(2016.01)
G06F 12/123(2016.01)

(56)对比文件
CN 113157605 A,2021.07.23
CN 113297098 A,2021.08.24

审查员 杨继爽

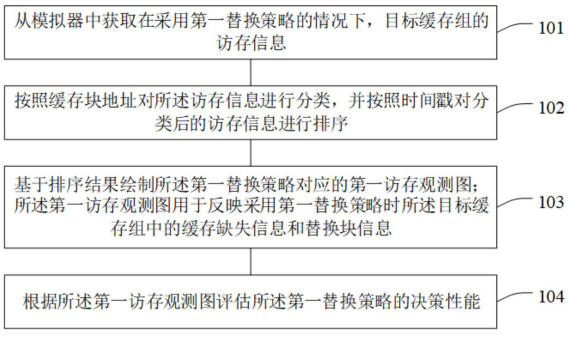
权利要求书3页 说明书13页 附图4页

(54)发明名称

高速缓存替换策略的管理方法、装置及电子设备

(57)摘要

本发明实施例提供一种高速缓存替换策略的管理方法、装置及电子设备，涉及计算机技术领域，该方法包括：从模拟器中获取在采用第一替换策略的情况下，目标缓存组的访存信息；所述模拟器用于对处理器系统的运行过程进行模拟；按照缓存块地址对所述访存信息进行分类，并按照时间戳对分类后的访存信息进行排序；基于排序结果绘制所述第一替换策略对应的第一访存观测图；根据所述第一访存观测图评估所述第一替换策略的决策性能。本发明实施例能够为替换策略的设计提供反馈，实现了对高速缓存运行过程的透视，增加了策略的可解释性，有利于辅助调试和设计。



CN 118295936 B

1.一种高速缓存替换策略的管理方法,其特征在于,所述方法包括:

从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息和替换块信息;

根据所述第一访存观测图评估所述第一替换策略的决策性能;

其中,所述排序结果包括所述目标缓存组对应的访问点图;所述基于排序结果绘制所述第一替换策略对应的第一访存观测图,包括:

在所述访问点图上按照事件类型对不同的事件进行标记,得到所述目标缓存组对应的第一访存观测图;

所述根据所述第一访存观测图评估所述第一替换策略的决策性能,包括:

基于所述第一访存观测图统计出所述目标缓存组内的缓存缺失次数,对所述第一替换策略进行评估。

2.根据权利要求1所述的方法,其特征在于,所述访存信息包括缓存块访问事件的相关信息;所述基于排序结果绘制所述第一替换策略对应的第一访存观测图,包括:

根据所述缓存块访问事件对应的缓存块地址确定访问点的纵坐标,根据所述缓存块访问事件对应的排序结果确定访问点的横坐标;

根据所述缓存块访问事件的访问类型,确定所述访问点的标识符;

根据所述缓存块访问事件的命中情况,确定所述标识符的颜色;

按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,绘制所述第一替换策略对应的第一访存观测图。

3.根据权利要求2所述的方法,其特征在于,所述访存信息还包括缓存块回填事件的回填信息;所述方法还包括:

在所述目标缓存组中存在缓存回填事件的情况下,根据所述回填信息,确定回填块地址和替换块地址;

根据所述回填块地址和所述替换块地址在所述第一访存观测图中添加替换标记;

其中,所述替换标记的横坐标与回填块地址对应的访问点的横坐标相同,所述替换标记的纵坐标为所述替换块地址对应的纵坐标值。

4.根据权利要求1所述的方法,其特征在于,所述根据所述第一访存观测图评估所述第一替换策略的决策性能,包括:

根据所述第一访存观测图确定所述第一替换策略对应的第一缓存缺失信息;

根据所述第一访存观测图中反映的访存踪迹信息,确定在采用目标替换策略的情况下,所述目标缓存组中的每个回填块对应的目标替换块;

根据所述访存踪迹信息和所述目标替换块,绘制所述目标替换策略对应的目标访存观测图;

根据所述目标访存观测图确定所述目标替换策略对应的目标缓存组缺失信息;

基于所述第一缓存缺失信息和所述目标缓存组缺失信息,评估所述第一替换策略的决

策性能。

5.根据权利要求1所述的方法,其特征在于,所述从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息,包括:

获取待观测的缓存组标识、第一替换策略和模拟指令数;

根据所述缓存组标识、第一替换策略和模拟指令数配置所述模拟器,以使所述模拟器按照所述第一替换策略和所述模拟指令数执行测试程序,并输出所述缓存组标识对应的目标缓存组的访存信息。

6.根据权利要求1所述的方法,其特征在于,所述方法还包括:

绘制第二替换策略对应的第二访存观测图;

根据所述第一访存观测图和所述第二访存观测图分别确定所述第一替换策略和所述第二替换策略的性能评分;

在所述第一替换策略的性能评分低于所述第二替换策略的性能评分的情况下,将所述目标缓存组的缓存替换策略更新为所述第二替换策略。

7.一种高速缓存替换策略的管理装置,其特征在于,所述装置包括:

获取模块,用于从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

预处理模块,用于按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

第一绘制模块,用于基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息和替换块信息;

评估模块,用于根据所述第一访存观测图评估所述第一替换策略的决策性能;

其中,所述排序结果包括所述目标缓存组对应的访问点图;所述第一绘制模块具体用于:

在所述访问点图上按照事件类型对不同的事件进行标记,得到所述目标缓存组对应的第一访存观测图;

所述评估模块具体用于:

基于所述第一访存观测图统计出所述目标缓存组内的缓存缺失次数,对所述第一替换策略进行评估。

8.根据权利要求7所述的装置,其特征在于,所述访存信息包括缓存块访问事件的相关信息;所述第一绘制模块,包括:

坐标确定子模块,用于根据所述缓存块访问事件对应的缓存块地址确定访问点的纵坐标,根据所述缓存块访问事件对应的排序结果确定访问点的横坐标;

标识符确定子模块,用于根据所述缓存块访问事件的访问类型,确定所述访问点的标识符;

颜色确定子模块, 用于根据所述缓存块访问事件的命中情况,确定所述标识符的颜色;

绘制子模块,用于按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,绘制所述第一替换策略对应的第一访存观测图。

9.一种电子设备,其特征在于,所述电子设备包括处理器、存储器、通信接口和通信总线,所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信;所述存储器用于存放可执行指令,所述可执行指令使所述处理器执行如权利要求1至6中任一项所述的高速缓存替换策略的管理方法。

10.一种可读存储介质,其特征在于,当所述可读存储介质中的指令由电子设备的处理器执行时,使得所述处理器能够执行如权利要求1至6中任一项所述的高速缓存替换策略的管理方法。

高速缓存替换策略的管理方法、装置及电子设备

技术领域

[0001] 本发明涉及计算机技术领域,尤其涉及一种高速缓存替换策略的管理方法、装置及电子设备。

背景技术

[0002] 在现代处理器中,高速缓存(Cache)往往占用了芯片较大比例(有的达到80%以上)的面积。高速缓存能利用应用程序存储访问的时间局部性和空间局部性,将频繁被访问的块留在高速存储介质中,从而提高程序性能。但由于上层(L1和L2)高速缓存过滤等因素,L3高速缓存的数据访问的局部性较差,利用率低,死块(Dead Block)较多。死块在被替换前一次也没有被再次访问,造成了芯片面积和功耗的浪费。但同时,仍有一些存在复用机会的数据块由于容量原因,在复用前被替换出高速缓存。高速缓存替换策略是管理高速缓存的重要途径之一,通过合理规划各缓存块的留存时间,可以起到增加高速缓存等效容量的作用。

[0003] 为了评估处理器的性能和功耗,以及相应的优化方法,通常采用时钟精确的全系统模拟器,对处理器核及高速缓存各个部件的模拟。模拟器能够输出较为精确的计数器值,用于性能和功耗评估。处理器性能通常用每周期指令数(Instructions Per Cycle,IPC)来评估。在高速缓存层次结构(Cache hierarchy)子系统中,常用的评价指标有缺失率(Miss Rate)、缺失延迟(Miss Latency)等。但是,这些指标仅仅是对替换策略结果的评价,不能反映替换策略的决策过程与应用的访存特征,从而不能给替换策略设计提供足够的反馈。

发明内容

[0004] 本发明实施例提供一种高速缓存替换策略的管理方法、装置及电子设备,可以解决相关技术中不能反映替换策略的决策过程与应用的访存特征,无法为缓存替换策略提供足够的反馈的问题。

[0005] 为了解决上述问题,本发明实施例公开了一种高速缓存替换策略的管理方法,所述方法包括:

[0006] 从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

[0007] 按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

[0008] 基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息和替换块信息;

[0009] 根据所述第一访存观测图评估所述第一替换策略的决策性能。

[0010] 可选地,所述访存信息包括缓存块访问事件的相关信息;所述基于排序结果绘制所述第一替换策略对应的第一访存观测图,包括:

[0011] 根据所述缓存块访问事件对应的缓存块地址确定访问点的纵坐标,根据所述缓存块访问事件对应的排序结果确定访问点的横坐标;

[0012] 根据所述缓存块访问事件的访问类型,确定所述访问点的标识符;

[0013] 根据所述缓存块访问事件的命中情况,确定所述标识符的颜色;

[0014] 按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,绘制所述第一替换策略对应的第一访存观测图。

[0015] 可选地,所述访存信息还包括缓存块回填事件的回填信息;所述方法还包括:

[0016] 在所述目标缓存组中存在缓存回填事件的情况下,根据所述回填信息,确定回填块地址和替换块地址;

[0017] 根据所述回填块地址和所述替换块地址在所述第一访存观测图中添加替换标记;

[0018] 其中,所述替换标记的横坐标与回填块地址对应的访问点的横坐标相同,所述替换标记的纵坐标为所述替换块地址对应的纵坐标值。

[0019] 可选地,所述根据所述第一访存观测图评估所述第一替换策略的决策性能,包括:

[0020] 根据所述第一访存观测图确定所述第一替换策略对应的第一缓存缺失信息;

[0021] 根据所述第一访存观测图中反映的访存踪迹信息,确定在采用目标替换策略的情况下,所述目标缓存组中的每个回填块对应的目标替换块;

[0022] 根据所述访存踪迹信息和所述目标替换块,绘制所述目标替换策略对应的目标访存观测图;

[0023] 根据所述目标访存观测图确定所述目标替换策略对应的目标缓存组缺失信息;

[0024] 基于所述第一缓存缺失信息和所述目标缓存组缺失信息,评估所述第一替换策略的决策性能。

[0025] 可选地,所述从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息,包括:

[0026] 获取待观测的缓存组标识、第一替换策略和模拟指令数;

[0027] 根据所述缓存组标识、第一替换策略和模拟指令数配置所述模拟器,以使所述模拟器按照所述第一替换策略和所述模拟指令数执行测试程序,并输出所述缓存组标识对应的目标缓存组的访存信息。

[0028] 可选地,所述方法还包括:

[0029] 绘制第二替换策略对应的第二访存观测图;

[0030] 根据所述第一访存观测图和所述第二访存观测图分别确定所述第一替换策略和所述第二替换策略的性能评分;

[0031] 在所述第一替换策略的性能评分低于所述第二替换策略的性能评分的情况下,将所述目标缓存组的缓存替换策略更新为所述第二替换策略。

[0032] 另一方面,本发明实施例公开了一种高速缓存替换策略的管理装置,所述装置包括:

[0033] 获取模块,用于从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

[0034] 预处理模块,用于按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

[0035] 第一绘制模块,用于基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息

和替换块信息；

[0036] 评估模块，用于根据所述第一访存观测图评估所述第一替换策略的决策性能。

[0037] 再一方面，本发明实施例还公开了一种电子设备，所述电子设备包括处理器、存储器、通信接口和通信总线，所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信；所述存储器用于存放可执行指令，所述可执行指令使所述处理器执行前述的访存方法。

[0038] 本发明实施例还公开了一种可读存储介质，当所述可读存储介质中的指令由电子设备的处理器执行时，使得电子设备能够执行前述的访存方法。

[0039] 本发明实施例包括以下优点：

[0040] 本发明实施例提供了一种高速缓存替换策略的管理方法，可以通过模拟器提供的目标缓存组的访存信息，以缓存块地址为视角绘制访存观测图，展示了一段时间内目标缓存组发生的所有访问与高速缓存块回填事件，从而清晰地展示每个缓存块的回填与替换时机，更直观地展示替换策略的决策与影响，为替换策略的设计提供反馈。本发明实施例能够实现对高速缓存运行过程的透视，增加了策略的可解释性，有利于辅助调试和设计。

附图说明

[0041] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1是本发明的一种高速缓存替换策略的管理方法实施例的步骤流程图；

[0043] 图2是本发明的一种访存观测图的示例；

[0044] 图3是本发明的另一种访存观测图的示例；

[0045] 图4是本发明的两种替换策略的访存观测图的示例；

[0046] 图5是本发明的一种访存观测图的示例；

[0047] 图6是本发明的另一种访存观测图的示例；

[0048] 图7是本发明的又一种访存观测图的示例；

[0049] 图8是本发明的一种高速缓存替换策略的管理装置的结构框图；

[0050] 图9是本发明示例提供的一种电子设备的结构框图。

具体实施方式

[0051] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0052] 本发明的说明书和权利要求书中的术语“第一”、“第二”等是用于区别类似的对象，而不用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施，且“第一”、“第二”等所区分的对象通常为一类，并不限定对象的个数，例如第一对象可以是一个，也可

以是多个。此外,说明书以及权利要求中的术语“和/或”用于描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。本发明实施例中术语“多个”是指两个或两个以上,其它量词与之类似。

[0053] 方法实施例

[0054] 参照图1,示出了本发明的一种高速缓存替换策略的管理方法实施例的步骤流程图,所述方法具体可以包括如下步骤:

[0055] 步骤101、从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

[0056] 步骤102、按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

[0057] 步骤103、基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息和替换块信息;

[0058] 步骤104、根据所述第一访存观测图评估所述第一替换策略的决策性能。

[0059] 本发明实施例提供的高速缓存替换策略的管理方法,可以对高速缓存替换策略进行性能评估。

[0060] 高速缓存(Cache)是计算机处理器的重要组成部分之一,是存储速度介于寄存器和内存之间的一种存储器。在计算机系统中,可以把各种存储部件(包括寄存器、高速缓存、内存、硬盘等)根据工作速度和单位成本划分成不同层次。越靠近处理器端,存储部件的工作速度越快、容量越小、单位容量的成本越高;越靠近内存端,存储部件的容量越大、工作速度越慢、单位容量的成本越低。Cache的运行速度通常慢于处理器但是快于内存。利用程序执行的局部性原理,把最近可能反复访问的数据拷贝到工作速度更快的Cache中,当处理器需要数据时可以以很小的时间延迟把数据提交给处理器,能有效减少访存消耗的时间,从而在一定程度上掩盖处理器和内存之间的工作速度差距,提升处理器性能。

[0061] 由于Cache的容量较小,需要对Cache进行有效的管理,尽可能将处理器需要的数据放入高速缓存中,从而减少系统访存失效概率,降低访存代价,提升系统的整体性能。

[0062] 目前,主流的高速缓存替换策略主要包括随机替换策略(Random)、先入先出替换策略(first in first out,FIFO)、先进后出替换策略(last in first out,LIFO)、最近最少使用替换策略(Least Recently Used,LRU)、最近最不常用替换策略(Least Frequently Used,LFU)、自适应替换策略(Adaptive Replacement Cache,ARC)、双峰性复用间隔预测替换策略(Bimodal Re-Reference Interval Prediction,BRRIP) 等。

[0063] 在采用理想LRU替换策略的高速缓存中,每个缓存行会维护一个时间戳计数器,该时间戳计数器用来记录本缓存行上一次被访问时的时钟计。每次访问失效时,同一缓存组内,时间戳计数最小的缓存行中的数据就会被替换出高速缓存,该缓存行会被用来存放从内存中新读取到的数据。同时,每一次访问,包括失效后新添加数据的访问,都会更新相应数据的缓存行中的时间戳,从而保证每次被替换出的数据都是来自最近最少使用的缓存行,即同一组内时间戳最小的缓存行。

[0064] FIFO策略是按照数据进入缓存的先后顺序进行替换。最早进入缓存的数据将最早

被替换掉,而最新进入缓存的数据则被保留。

[0065] LFU策略是根据每个数据项被访问的频率来进行替换。较少被访问的数据将被优先替换掉,以便为更常被访问的数据腾出空间。

[0066] 随机替换策略是一种简单直接的方法,随机选择一个数据进行替换。由于随机性的存在,该策略无法保证缓存中是存放着最有用的数据。

[0067] 此外,针对局部性较差的应用程序,高速缓存可以采用不更新新加入数据缓存行时间戳的策略,即将新加入的数据放入最先被替换的位置,通常称为LRU位置插入策略(the LRU Position Insertion Policy,LIP)。

[0068] 在实际应用中,选择适当的缓存替换策略对计算机系统的系统性能具有重要影响。不同的应用场景可能对缓存替换策略有不同的要求。缓存替换策略常用的评估指标有:命中率(Hit Ratio)、替换开销(Replacement Overhead)、公平性(Fairness)、缺失率(Miss Rate)、缺失延迟(Miss Latency)等。其中,命中率是指在缓存访问中,缓存中已存在数据的比例,命中率高意味着缓存中的数据能够满足大部分访问需求,系统性能较好。替换开销是指进行缓存替换操作时所需要的时间和计算资源,较低的替换开销可以提高系统的响应速度。公平性是指缓存替换策略在对待不同的数据项时是否具有公正性,如果某些数据频繁被替换,而其他数据则很少被替换,可能导致系统性能不均衡。缺失率是指在缓存访问中,未找到所需数据的次数与总请求次数的比率,缺失率高意味着缓存中的数据不能满足大部分访问需求,系统性能较差。缺失延迟指的是因为缓存缺失导致的访问延迟,缺失延迟越高,系统性能越差。

[0069] 可以理解的是,上述这些评价指标仅仅是对缓存替换结果的评价,不能反映替换策略的决策过程与应用的访存特征,无法给替换策略设计提供足够的反馈。

[0070] 本发明实施例提供的高速缓存替换策略的管理方法,可以通过模拟器提供的目标缓存组的访存信息,以缓存块地址为视角,绘制Set内访存序列微观观测图,也即本发明中的第一访存观测图,展示了一段时间内目标缓存组发生的所有访问(Access)与高速缓存块回填(Cache Fill)事件,从而清晰地展示每个缓存块的回填与替换时机,更直观地展示替换策略的决策(选择的替换块)与影响(是否造成替换块后续访问缺失),为替换策略的设计提供反馈。

[0071] 需要说明的是,本发明实施例中的模拟器用于对处理器系统的运行过程进行模拟。处理器系统包括处理器核、高速缓存和内存,高速缓存通常包含多级高速缓存。模拟器可以对处理器核及高速缓存等各个部件进行模拟,提供目标缓存组的访存信息。其中,目标缓存组的访存信息,包括来自上级高速缓存的访存信息和本级高速缓存的预取信息。

[0072] 可选地,所述访存信息包括以下至少一项:被访问的缓存块地址、访问类型、访问命中情况、访问时间戳、替换块地址。

[0073] 作为一种示例,在利用模拟器对处理器系统进行全系统模拟前,用户可以在某一级高速缓存中指定需要观察的目标缓存组(Set)的编号,模拟器在目标缓存组内发现来自上级高速缓存的访问与本级预取时记录相关信息,并在模拟过程中输出到指定文件进行保存。其中,来自上级高速缓存的访问包括上级高速缓存发送的预取和来自CPU的读写请求,本级预取包括本级高速缓存发送的预取请求。发现这些请求后,需要记录的信息包括:被访问的缓存块地址(Address)、访问类型(Type)、访问命中情况、访问时间戳(Timestamp)和替

换块地址。其中,访存命中情况包括访问命中(Hit)或缺失(Miss)。替换块地址指的是发生缓存块回填时被替换的高速缓存块的地址。

[0074] 可选地,所述访问类型包括处理器核读写、上一级高速缓存的预取、本级高速缓存的预取。

[0075] 在本发明实施例中,从模拟器中获取到目标缓存组的访存信息之后,按照缓存块地址对这些访存信息进行分类,并按照时间顺序排列,形成目标缓存组在一段时间内的访问点图。在访问点图上按照事件类型,如缓存块回填事件、缓存块缺失事件、缓存块命中事件等,对不同的事件进行标记,就可以得到目标缓存组对应的第一访存观测图。

[0076] 根据第一访存观测图,就可以直观地反映采用第一替换策略时目标缓存组内的缓存缺失信息和替换块信息,例如,根据第一访存观测图可以看出目标缓存组中被替换的缓存块,也即替换块,以及当某个替换块被替换出高速缓存后是否会造成该替换块后续的访问缺失,还可以基于第一访存观测图统计出目标缓存组内的缓存缺失次数等信息。基于第一访存观测图可以从目标缓存组内微观的访存特征(如:访存是否存在复用,以及复用的特征等),对第一替换策略进行评估,并为替换策略设计提供反馈。

[0077] 可选地,步骤101所述从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息,包括:

[0078] 步骤S11、获取待观测的缓存组标识、第一替换策略和模拟指令数;

[0079] 步骤S12、根据所述缓存组标识、第一替换策略和模拟指令数配置所述模拟器,以使所述模拟器按照所述第一替换策略和所述模拟指令数执行测试程序,并输出所述缓存组标识对应的目标缓存组的访存信息。

[0080] 本发明实施例中的模拟器可以包括但不限于全系统模拟器、高速缓存模拟器等,模拟方式可以是功能模拟或性能模拟。模拟器运行前,需要指定观测的缓存组标识、第一替换策略和模拟指令数。其中,缓存组标识可以包括高速缓存层级、目标缓存组编号等。模拟器按照第一替换策略和模拟指令数执行测试程序,并目标缓存组的访存信息,例如,每次访问的访问地址、访问请求类型、访问是否命中、访问时间戳和发生缓存块回填时被替换的高速缓存块的地址,等等。

[0081] 在本发明的一种可选实施例中,所述访存信息包括缓存块访问事件的相关信息;步骤103所述基于排序结果绘制所述第一替换策略对应的第一访存观测图,包括:

[0082] 步骤S21、根据所述缓存块访问事件对应的缓存块地址确定访问点的纵坐标,根据所述缓存块访问事件对应的排序结果确定访问点的横坐标;

[0083] 步骤S22、根据所述缓存块访问事件的访问类型,确定所述访问点的标识符;

[0084] 步骤S23、根据所述缓存块访问事件的命中情况,确定所述标识符的颜色;

[0085] 步骤S24、按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,绘制所述第一替换策略对应的第一访存观测图。

[0086] 绘制第一访存观测图时,可以根据缓存访问事件对应的缓存块地址,确定访问点的纵坐标,根据缓存块访问事件对应的排序结果确定访问点的横坐标。示例性地,可以根据访问的缓存块地址确定访问地址标签,访问地址标签可以是对缓存块地址进行哈希运算得到的,也可以将缓存块地址的高位(例如高4位)作为访问地址标签。然后,根据访问地址标签确定纵坐标,若该标签第一次出现,则纵坐标为当前最大纵坐标值+1,否则等于原标签对

应的纵坐标值。按照访问时间戳对访存信息进行排序后,访问点的横坐标就可以等于上一个访问点的横坐标值+1。

[0087] 接下来,根据访问类型确定访问点的标识符。访问类型可以包括处理器核读写、上一级高速缓存的预取、本级高速缓存的预取。分别用不同的标识符表示不同的访问类型。例如,标记“□”、“○”和“△”分别用于表示CPU读写、上级预取和本级预取等不同的访问类型。对于每一次缓存块访问,使用不同的颜色来表示访问是否造成缺失。

[0088] 按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,就可以绘制出第一替换策略对应的第一访存观测图。

[0089] 可选地,所述访存信息还包括缓存块回填事件的回填信息;所述方法还包括:

[0090] 步骤S31、在所述目标缓存组中存在缓存回填事件的情况下,根据所述回填信息,确定回填块地址和替换块地址;

[0091] 步骤S32、根据所述回填块地址和所述替换块地址在所述第一访存观测图中添加替换标记。

[0092] 其中,所述替换标记的横坐标与回填块地址对应的访问点的横坐标相同,所述替换标记的纵坐标为所述替换块地址对应的纵坐标值。

[0093] 在本发明实施例中,如果目标缓存组内发生缓存块回填事件,可以在第一访存观测图中添加替换标记,用来表示回填块和替换块。其中,回填块指的是从下一级高速缓存或内存中回填至本级高速缓存中的数据块,替换块指的是高速缓存中被回填块替换掉的缓存块。缓存块回填事件可以是由缓存块访问缺失或本级预取访问触发的。

[0094] 示例性地,如果由于缓存块访问缺失或本级预取访问造成缓存块回填,在图中用竖虚线和替换标记“×”来表示替换旧缓存块。替换标记“×”的横坐标与访问点相同,纵坐标等于替换块地址标签对应的纵坐标值,代表左侧地址标签对应的缓存块被换出。

[0095] 参照图2,示出了本发明实施例提供的一种访存观测图。如图2所示,分别以高速缓存组和缓存块地址标签为视角,展示了一个2路(way)组相联的高速缓存的Set内访存观测图。示例展示了该组内对地址标签为A、B、C的三个缓存块的10次访问,对应的缓存块访问序列为“A-B-C-C-A-A-D-E-A-A”。左侧的以高速缓存组为视角的访存观测图是目前常见的观测手段,它展示了每一次访问前后,高速缓存组内存放的缓存块,但访问是否缺失和替换块信息只能与访问顺序相关联,非常零散,而且不能反映缓存块的复用情况。而在右侧的以缓存块地址标签为视角的访存观测图中,访问的缺失信息与替换块信息可以与缓存块地址相关联,清晰地展示了每个缓存块的访问、回填、替换与复用情况。此外,以缓存块地址标签为视角的访存观测图还可以方便地观察这段时间内的程序工作集大小。所有地址标签与组内所有访问围成的矩形空间可以看作为高速缓存组的访问时空图,“X”标记的数量表明该段时间内的高速缓存缺失数量,代表着相应替换策略的性能。

[0096] 在以高速缓存块地址标签(Tag)为视角的访存观测图中,虚线圈表示访存缺失,实线圈表示访存命中。从图中可以看出,缓存块A在第5和6次访问时存在复用,缓存块C在第4次访问存在复用。在第3次访问,缓存块C在回填时替换了缓存块A,从而造成了缓存块A在第5次访问复用时的访存缺失。

[0097] 参照图3,示出了本发明实施例提供的另一种访存观测图。需要说明的是,图3展示了带有预取的访存观测图示例,相比于图2展示的前6次高速缓存组访问,在第3次访问前加

入了对缓存块C的本级预取访问,在上级读写访问到来前将缓存块从下级高速缓存(或内存)中取出,从而减少一次上级读写访问缺失。

[0098] 可选地,步骤104所述根据所述第一访存观测图评估所述第一替换策略的决策性能,包括:

[0099] 步骤S41、根据所述第一访存观测图确定所述第一替换策略对应的第一缓存缺失信息;

[0100] 步骤S42、根据所述第一访存观测图中反映的访存踪迹信息,确定在采用目标替换策略的情况下,所述目标缓存组中的每个回填块对应的目标替换块;

[0101] 步骤S43、根据所述访存踪迹信息和所述目标替换块,绘制所述目标替换策略对应的目标访存观测图;

[0102] 步骤S44、根据所述目标访存观测图确定所述目标替换策略对应的目标缓存组缺失信息;

[0103] 步骤S45、基于所述第一缓存缺失信息和所述目标缓存组缺失信息,评估所述第一替换策略的决策性能。

[0104] 在本发明实施例中,可以根据第一替换策略对应的第一访存观测图中反映的访存踪迹信息,确定在采用目标替换策略的情况下,目标缓存组中的每个回填块对应的目标替换块。其中,目标替换策略指的是理论最优策略。示例性地,目标替换策略可以是在每次缓存块回填导致需要替换时,选择在更远的将来被用到的缓存块用于替换。即在已有的踪迹信息中从该次访问向后寻找,在高速缓存中优先替换掉未在踪迹中被访问的缓存块,然后是在踪迹中最晚被访问的缓存块。

[0105] 根据目标替换策略决策的目标替换块和访存踪迹信息,绘制出目标替换策略对应的目标访存观测图。具体的绘制过程可以参考前述第一访存观测图的绘制过程,本发明实施例在此不做进一步赘述。

[0106] 目标访存观测图可以用于确定替换策略的性能上限,对不同的替换策略进行性能评估。示例性地,对于待评估的第一替换策略,可以基于其对应的第一访存观测图确定出第一替换策略对应的第一缓存缺失信息,基于目标访存观测图确定出目标替换策略对应的目标缓存缺失信息,然后将两组缓存缺失信息进行比较,第一缓存缺失信息中的各项数据越接近目标缓存缺失信息,第一替换策略的性能越好。其中,缓存缺失信息可以包括上级读取缺失次数、上级写回缺失次数,等等。

[0107] 此外,也可以直接基于第一替换策略对应的第一访存观测图,统计采用第一替换策略时目标缓存组内各类访问的缺失数量,如CPU读写缺失、上级预取缺失,等等,基于统计结果评估第一替换策略的性能。

[0108] 可选地,所述方法还包括:

[0109] 步骤S51、绘制第二替换策略对应的第二访存观测图;

[0110] 步骤S52、根据所述第一访存观测图和所述第二访存观测图分别确定所述第一替换策略和所述第二替换策略的性能评分;

[0111] 步骤S53、在所述第一替换策略的性能评分低于所述第二替换策略的性能评分的情况下,将所述目标缓存组的缓存替换策略更新为所述第二替换策略。

[0112] 在本发明实施例中,还可以通过分别绘制不同替换策略对应的访存观测图,对不

同替换策略进行性能比较,并从中选择出性能较好的替换策略作为目标缓存组中实际采用的缓存替换策略,从而提升目标缓存组的缓存替换性能,提升整个计算机系统的访存性能。

[0113] 其中,第一替换策略、第二替换策略可以是任意一种替换策略,只要两者不同即可。

[0114] 示例性地,参照图4,示出了本发明实施例提供的两种替换策略的访存观测图。其中,策略X的替换决策与图2相同,在第3次访问时缓存块C回填时替换了缓存块A,从而在第5次访问块A复用时造成访问缺失。而对比策略Y则在第3次访问选择替换缓存块B,在块A复用时命中。从图4中可以看出,策略Y比策略X少了一次上级读写访问缺失,同时也是该访存场景下的最优替换策略。

[0115] 参照图5至图7,分别示出了本发明实施例提供的一种访存观测图。具体地,5至图7展示了采用本发明实施例提供的高速缓存替换策略的管理方法,对LRU和BRRIP两个策略在SPEC 2006测试集的libquantum应用中的一个实例。其中,图5为LRU替换策略对应的访存观测图,图6为BRRIP替换策略对应的访存观测图,图7为最优替换策略对应的访存观测图。在该实例配置中,L3高速缓存的关联度为16,共运行四千万条(40M)指令,绘制这两个替换策略在L3高速缓存的访存观测图。从图中可以看出,这段时间内共有32个不同的缓存块被依次循环访问。受限于L3组相联的特点,一个高速缓存组(Set)只能容纳16个不同的缓存块,BRRIP保护了前面访问的16个缓存块,避免了这部分访存再次访问时造成的缓存缺失。而LRU一直替换最早访问的缓存块,但由于所有缓存块的复用距离(同一缓存块两次访问间的不同的缓存块数量)都为32,大于组容量16,所以再次访问时,该缓存块已被替换出组,造成缓存缺失。在生成的最优策略中,也采用了保留前16个缓存块的方式。

[0116] 在实例中,有上级读取请求和上级写回请求,由于写回请求不在关键路径上,其缺失对性能的影响不大。而由上级高速缓存发出的读取请求与CPU的数据供应关系密切,其缺失将较大程度上影响性能。从缓存缺失数统计来看,LRU策略造成了68次上级读取访问缺失,而BRRIP策略只造成了50次上级读取访问缺失,接近于最优策略造成的48次缺失,性能更好。

[0117] 综上,本发明实施例提供了一种高速缓存替换策略的管理方法,可以通过模拟器提供的目标缓存组的访存信息,以缓存块地址为视角绘制访存观测图,展示了一段时间内目标缓存组发生的所有访问(Access)与高速缓存块回填事件,从而清晰地展示每个缓存块的回填与替换时机,更直观地展示替换策略的决策与影响,为替换策略的设计提供反馈。本发明实施例能够实现对高速缓存运行过程的透视,将以往的黑盒过程白盒化,增加了策略的可解释性,有利于辅助调试和设计。

[0118] 需要说明的是,对于方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本发明实施例并不受所描述的动作顺序的限制,因为依据本发明实施例,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作并不一定是本发明实施例所必须的。

[0119] 装置实施例

[0120] 参照图8,示出了本发明的一种高速缓存替换策略的管理装置的结构框图,所述装置具体可以包括:

[0121] 获取模块801,用于从模拟器中获取在采用第一替换策略的情况下,目标缓存组的访存信息;所述模拟器用于对处理器系统的运行过程进行模拟;

[0122] 预处理模块802,用于按照缓存块地址对所述访存信息进行分类,并按照时间戳对分类后的访存信息进行排序;

[0123] 第一绘制模块803,用于基于排序结果绘制所述第一替换策略对应的第一访存观测图;所述第一访存观测图用于反映采用第一替换策略时所述目标缓存组中的缓存缺失信息和替换块信息;

[0124] 评估模块804,用于根据所述第一访存观测图评估所述第一替换策略的决策性能。

[0125] 可选地,所述访存信息包括缓存块访问事件的相关信息;所述第一绘制模块,包括:

[0126] 坐标确定子模块,用于根据所述缓存块访问事件对应的缓存块地址确定访问点的纵坐标,根据所述缓存块访问事件对应的排序结果确定访问点的横坐标;

[0127] 标识符确定子模块,用于根据所述缓存块访问事件的访问类型,确定所述访问点的标识符;

[0128] 颜色确定子模块, 用于根据所述缓存块访问事件的命中情况,确定所述标识符的颜色;

[0129] 第一绘制子模块,用于按照各个访问点的纵坐标、横坐标、标识符和所述标识符的颜色,绘制所述第一替换策略对应的第一访存观测图。

[0130] 可选地,所述访存信息还包括缓存块回填事件的回填信息;所述装置还包括:

[0131] 回填确定模块,用于在所述目标缓存组中存在缓存回填事件的情况下,根据所述回填信息,确定回填块地址和替换块地址;

[0132] 标记添加模块,用于根据所述回填块地址和所述替换块地址在所述第一访存观测图中添加替换标记;

[0133] 其中,所述替换标记的横坐标与回填块地址对应的访问点的横坐标相同,所述替换标记的纵坐标为所述替换块地址对应的纵坐标值。

[0134] 可选地,所述评估模块,包括:

[0135] 第一确定子模块,用于根据所述第一访存观测图确定所述第一替换策略对应的第一缓存缺失信息;

[0136] 第二确定子模块,用于根据所述第一访存观测图中反映的访存踪迹信息,确定在采用目标替换策略的情况下,所述目标缓存组中的每个回填块对应的目标替换块;

[0137] 第二绘制子模块,用于根据所述访存踪迹信息和所述目标替换块,绘制所述目标替换策略对应的目标访存观测图;

[0138] 第三确定子模块,用于根据所述目标访存观测图确定所述目标替换策略对应的目标缓存组缺失信息;

[0139] 评估子模块,用于基于所述第一缓存缺失信息和所述目标缓存组缺失信息,评估所述第一替换策略的决策性能。

[0140] 可选地,所述获取模块,包括:

[0141] 获取子模块,用于获取待观测的缓存组标识、第一替换策略和模拟指令数;

[0142] 配置子模块,用于根据所述缓存组标识、第一替换策略和模拟指令数配置所述模

拟器,以使所述模拟器按照所述第一替换策略和所述模拟指令数执行测试程序,并输出所述缓存组标识对应的目标缓存组的访存信息。

[0143] 可选地,所述装置还包括:

[0144] 第二绘制模块,用于绘制第二替换策略对应的第二访存观测图;

[0145] 评分确定模块,用于根据所述第一访存观测图和所述第二访存观测图分别确定所述第一替换策略和所述第二替换策略的性能评分;

[0146] 策略更新模块,用于在所述第一替换策略的性能评分低于所述第二替换策略的性能评分的情况下,将所述目标缓存组的缓存替换策略更新为所述第二替换策略。

[0147] 综上,本发明实施例提供了一种高速缓存替换策略的管理装置,可以通过模拟器提供的目标缓存组的访存信息,以缓存块地址为视角绘制访存观测图,展示了一段时间内目标缓存组发生的所有访问与高速缓存块回填事件,从而清晰地展示每个缓存块的回填与替换时机,更直观地展示替换策略的决策与影响,为替换策略的设计提供反馈。本发明实施例能够实现对高速缓存运行过程的透视,将以往的黑盒过程白盒化,增加了策略的可解释性,有利于辅助调试和设计。

[0148] 对于装置实施例而言,由于其与方法实施例基本相似,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。

[0149] 本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。

[0150] 关于上述实施例中的处理器,其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述,此处将不做详细阐述说明。

[0151] 参照图9,是本发明实施例提供的一种用于高速缓存替换策略管理的电子设备的结构框图。如图9所示,所述电子设备包括:处理器、存储器、通信接口和通信总线,所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信;所述存储器用于存放可执行指令,所述可执行指令使所述处理器执行前述实施例的高速缓存替换策略的管理方法。

[0152] 所述处理器可以是CPU(Central Processing Unit,中央处理器),通用处理器、DSP(Digital Signal Processor,数字信号处理器),ASIC(Application Specific Integrated Circuit,专用集成电路),FPGA(Field Programmable Gate Array,现场可编程门阵列)或者其他可编辑器件、晶体管逻辑器件、硬件部件或者其任意组合。所述处理器也可以是实现计算功能的组合,例如包含一个或多个微处理器组合,DSP和微处理器的组合等。

[0153] 所述通信总线可包括一通路,在存储器和通信接口之间传送信息。通信总线可以是PCI(Peripheral Component Interconnect,外设部件互连标准)总线或EISA(Extended Industry Standard Architecture,扩展工业标准结构)总线等。所述通信总线可以分为地址总线、数据总线、控制总线等。为便于表示,图9中仅用一条线表示,但并不表示仅有一根总线或一种类型的总线。

[0154] 所述存储器可以是ROM(Read Only Memory,只读存储器)或可存储静态信息和指令的其他类型的静态存储设备、RAM(Random Access Memory,随机存取存储器)或者可存储信息和指令的其他类型的动态存储设备,也可以是EEPROM(Electrically Erasable

Programmable Read Only Memory,电可擦可编程只读存储器)、CD-ROM(Compact Disc Read Only Memory,只读光盘)、磁带、软盘和光数据存储设备等。

[0155] 本发明实施例还提供了一种非临时性计算机可读存储介质,当所述存储介质中的指令由电子设备(服务器或者终端)的处理器执行时,使得处理器能够执行图1所示的高速缓存替换策略的管理方法。

[0156] 本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。

[0157] 本领域内的技术人员应明白,本发明实施例的实施例可提供为方法、装置或计算机程序产品。因此,本发明实施例可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且,本发明实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0158] 本发明实施例是参照根据本发明实施例的方法、终端设备(系统)和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框,以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0159] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以预测方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0160] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上,使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0161] 尽管已描述了本发明实施例的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例做出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本发明实施例范围的所有变更和修改。

[0162] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0163] 以上对本发明所提供的一种高速缓存替换策略的管理方法、装置及电子设备,进行了详细介绍,本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施

例的说明只是用于帮助理解本发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

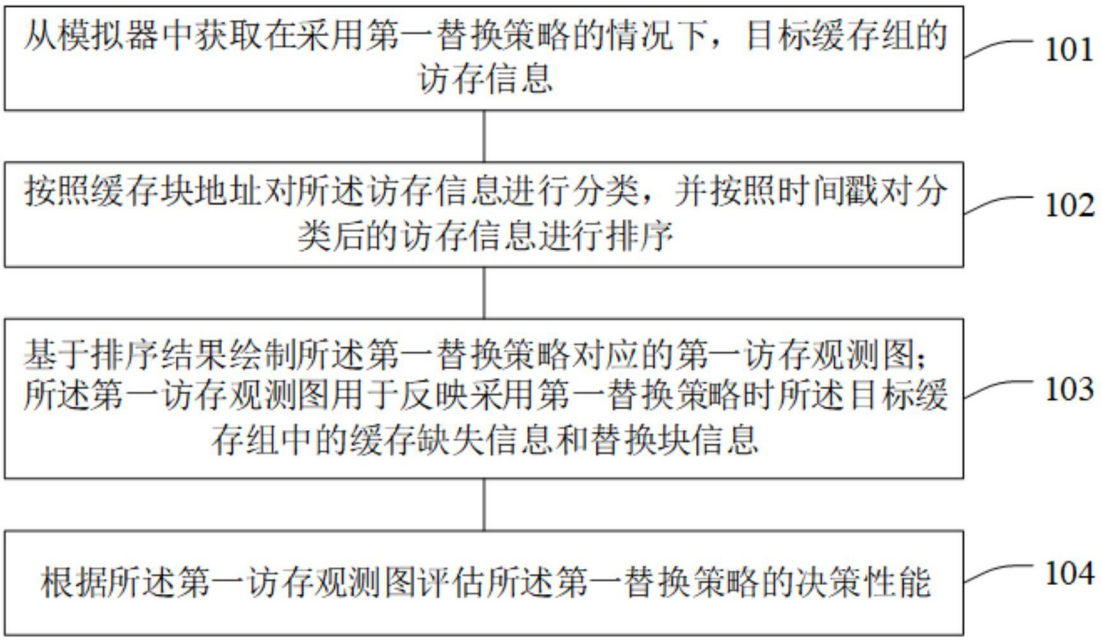
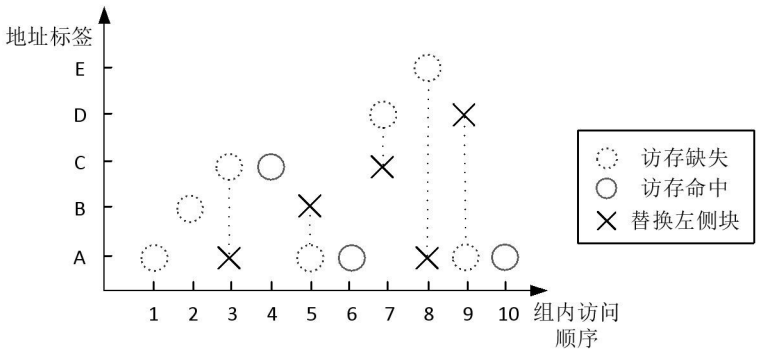


图1

2路组相联Cache			
访问顺序	Way 0	Way 1	命中?
0	-	-	
1	A	-	缺失
2	A	B	缺失
3	C	B	缺失
4	C	B	命中
5	C	A	缺失
6	C	A	命中
7	D	B	缺失
8	D	E	缺失
9	A	E	缺失
10	A	E	命中

以高速缓存组为视角



以缓存块地址标签为视角

图2

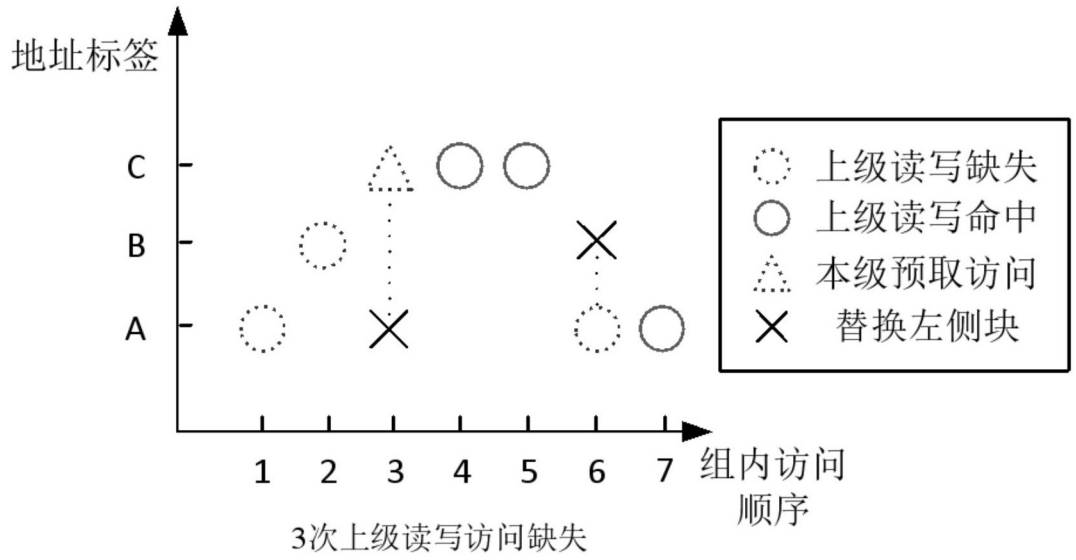

地址标签
C
B
A
1 2 3 4 5 6 7
组内访问顺序
上级读写缺失
上级读写命中
本级预取访问
替换左侧块
3次上级读写访问缺失

图3

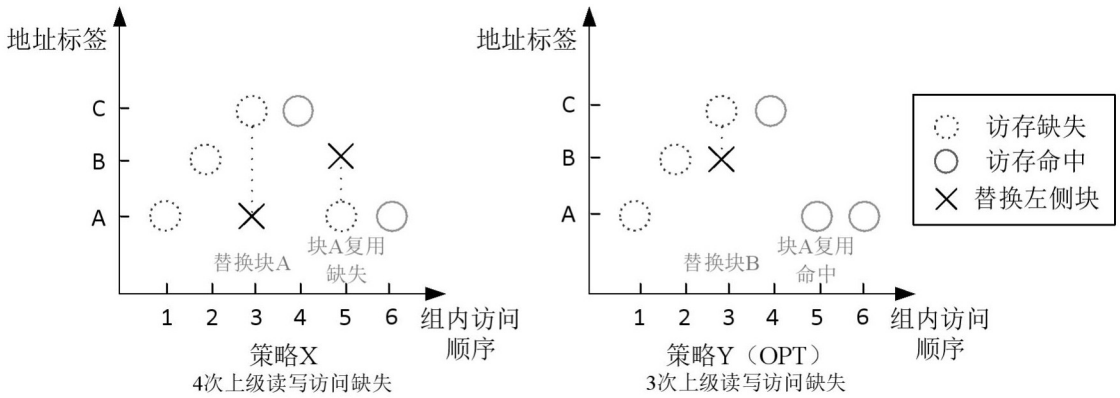

地址标签
C
B
A
替换块A
块A复用
缺失
1 2 3 4 5 6
组内访问顺序
策略X
4次上级读写访问缺失
地址标签
C
B
A
替换块B
块A复用
命中
1 2 3 4 5 6
组内访问顺序
策略Y（OPT）
3次上级读写访问缺失
访存缺失
访存命中
替换左侧块

图4

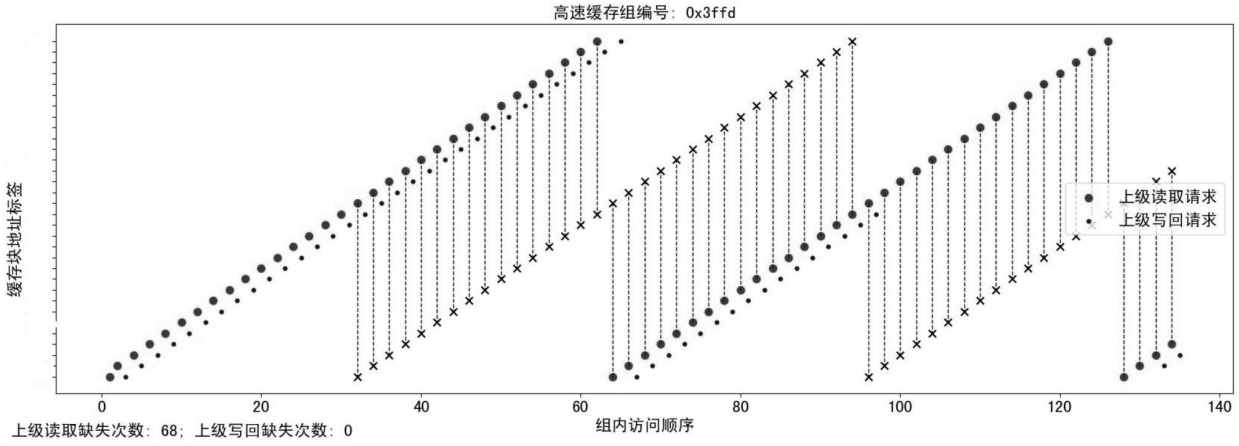

高速缓存组编号：0x3ffd
缓存块地址标签
0 20 40 60 80 100 120 140
组内访问顺序
上级读取请求
上级写回请求
上级读取缺失次数：68；上级写回缺失次数：0

图5

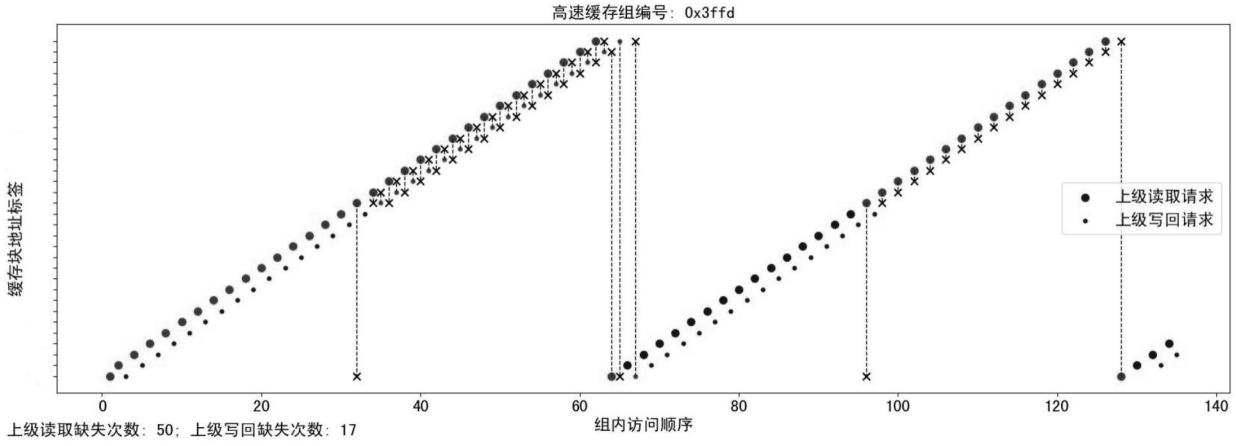

高速缓存组编号：0x3ffd
缓存块地址标签
上级读取请求
上级写回请求
0
20
40
60
80
100
120
140
组内访问顺序
上级读取缺失次数：50；上级写回缺失次数：17

图6

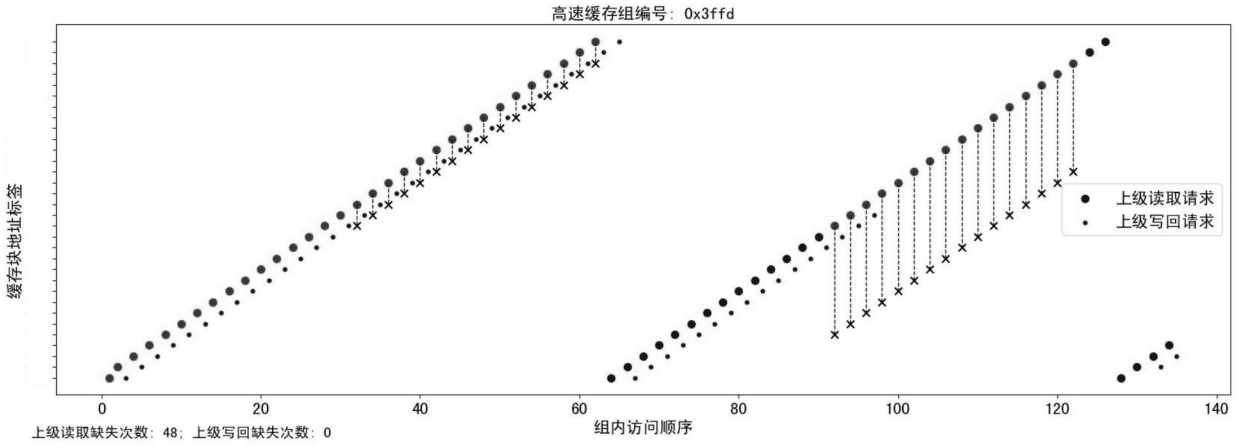

高速缓存组编号：0x3ffd
缓存块地址标签
上级读取请求
上级写回请求
0
20
40
60
80
100
120
140
组内访问顺序
上级读取缺失次数：48；上级写回缺失次数：0

图7

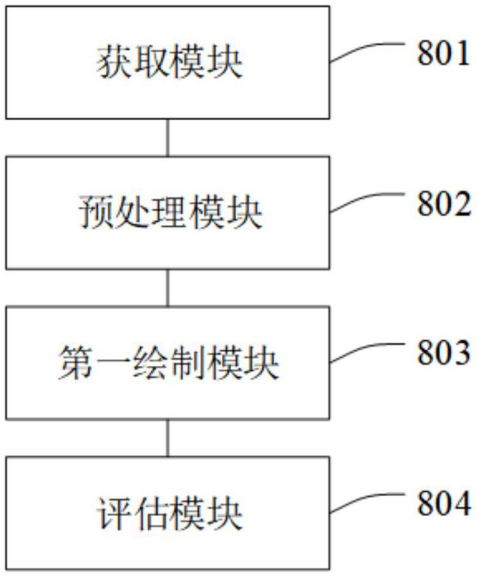

获取模块
801
预处理模块
802
第一绘制模块
803
评估模块
804

图8

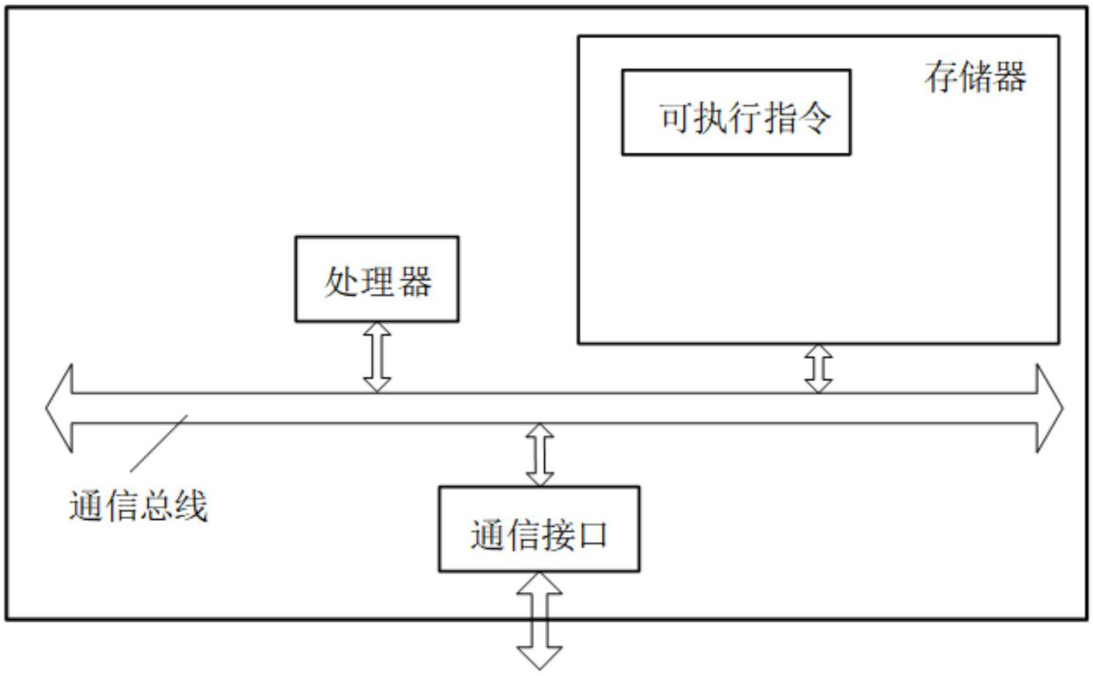

存储器
可执行指令
处理器
通信总线
通信接口

图9