(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118819868 B
(45)授权公告日 2025.01.21

(21)申请号 202411303411.8

(22)申请日 2024.09.19

(65)同一申请的已公布的文献号
申请公布号 CN 118819868 A

(43)申请公布日 2024.10.22

(73)专利权人 山东浪潮科学研究院有限公司
地址 250101 山东省济南市高新区浪潮路1036号S02号楼

(72)发明人 许桂龙 赵鑫鑫 魏朝飞 姜凯

(74)专利代理机构 北京君慧知识产权代理事务所(普通合伙) 11716
专利代理师 董延丽

(51)Int.Cl.
G06F 9/50(2006.01)
G06T 1/20(2006.01)

(56)对比文件
CN 107577524 A,2018.01.12
CN 110968345 A,2020.04.07

审查员 王婕

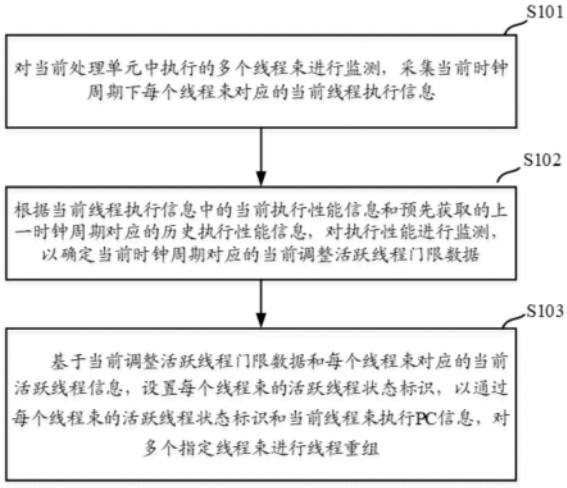
权利要求书3页 说明书9页 附图5页

(54)发明名称

一种基于GPGPU的线程重组方法、设备及介质

(57)摘要

本说明书实施例公开了一种基于GPGPU的线程重组方法、设备及介质,涉及线程分支管理技术领域,方法包括:对当前处理单元中执行的多个线程束进行监测,采集当前时钟周期下每个线程束对应的当前线程执行信息;根据当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定当前时钟周期对应的当前调整活跃线程门限数据;基于当前调整活跃线程门限数据和每个线程束对应的当前活跃线程信息,设置每个线程束的活跃线程状态标识,以通过每个线程束的活跃线程状态标识和当前线程束执行PC信息,对多个指定线程束进行线程重组,针对拥有不同活跃线程数量的线程束采取不同的重组方式,降低重组开销。



CN 118819868 B

1.一种基于GPGPU的线程重组方法,其特征在于,所述方法包括:

对当前处理单元中执行的多个线程束进行监测,采集当前时钟周期下每个所述线程束对应的当前线程执行信息,其中,所述当前线程执行信息包括当前执行性能信息、当前线程束执行PC信息和当前活跃线程信息;

根据所述当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定所述当前时钟周期对应的当前调整活跃线程门限数据,其中,所述执行性能信息包括执行过程访存次数和执行通道冲突次数;

基于所述当前调整活跃线程门限数据和每个所述线程束对应的当前活跃线程信息,设置每个所述线程束的活跃线程状态标识,以通过每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,对多个指定线程束进行线程重组;

通过每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,对多个指定线程束进行线程重组,具体包括:

通过每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,在多个所述线程束中进行筛选,以确定符合预设要求的多个第一线程束,其中,所述第一线程束的活跃线程状态标识为低活跃状态,且所述多个第一线程束对应的当前线程束执行PC信息相同;

确定所述多个第一线程束的待重组线程束数量,并获取预先生成的线程束重组数量阈值;

当所述待重组线程束数量大于所述线程束重组数量阈值时,根据每个所述第一线程束的指定执行性能评估要素,在多个所述第一线程束中确定符合预设要求的多个指定线程束,以进行线程重组,其中,所述多个指定线程束的数量为所述线程束重组数量阈值。

2.根据权利要求1所述的一种基于GPGPU的线程重组方法,其特征在于,根据所述当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定所述当前时钟周期对应的当前调整活跃线程门限数据,具体包括:

根据每个线程束对应的所述当前执行性能信息和所述历史执行性能信息,分别确定所述当前时钟周期对应的当前执行性能评估指标和所述上一时钟周期对应的历史执行性能评估指标;

通过所述当前执行性能评估指标和所述历史执行性能评估指标,确定所述当前时钟周期相对于所述上一时钟周期的执行性能变化类型,其中,所述执行性能变化类型包括性能提高和性能降低中的任意一项;

获取所述上一时钟周期对应的历史活跃线程门限数据,基于所述执行性能变化类型,对所述历史活跃线程门限数据进行调整,确定所述当前时钟周期对应的当前调整活跃线程门限数据。

3.根据权利要求2所述的一种基于GPGPU的线程重组方法,其特征在于,根据每个线程束对应的所述当前执行性能信息和所述历史执行性能信息,分别确定所述当前时钟周期对应的当前执行性能评估指标和所述上一时钟周期对应的历史执行性能评估指标,具体包括:

将每个线程束对应的所述当前执行性能信息中的当前执行过程访存次数和当前执行通道冲突次数进行累加,确定每个所述线程束的第一执行性能评估指标;

对每个所述线程束对应的第一执行性能评估指标取均值,确定所述当前时钟周期对应的当前执行性能评估指标;

基于每个所述历史执行性能信息中的历史执行过程访存次数和历史执行通道冲突次数进行累加,确定多个第二历史执行性能评估指标,以进行均值处理,确定所述上一时钟周期对应的历史执行性能评估指标。

4.根据权利要求2所述的一种基于GPGPU的线程重组方法,其特征在于,基于所述执行性能变化类型,对所述历史活跃线程门限数据进行调整,确定所述当前时钟周期对应的当前调整活跃线程门限数据,具体包括:

通过所述当前执行性能评估指标和所述历史执行性能评估指标,确定性能变化指标;

根据所述性能变化指标和所述历史执行性能评估指标,确定所述当前时钟周期的性能变化率,基于所述性能变化率和所述历史活跃线程门限数据,确定门限调整量;

通过所述执行性能变化类型,确定所述门限调整量对应的取整方式和门限调整方式,其中,所述取整方式包括向上取整和向下取整中的任意一项,所述门限调整方式包括门限上调和门限下调中的任意一项;

按照所述取整方式对所述门限调整量进行取整,确定门限调整步长;

按照所述门限调整方式,在所述历史活跃线程门限数据的基础上,根据所述门限调整步长进行调整,确定所述当前时钟周期对应的当前调整活跃线程门限数据。

5.根据权利要求1所述的一种基于GPGPU的线程重组方法,其特征在于,获取预先生成的线程束重组数量阈值,具体包括:

获取上一时间周期对应的历史线程束重组数量阈值;

确定当前时间周期内每个所述线程束对应的当前执行过程访存次数,以基于多个所述当前执行过程访问次数,确定所述当前时间周期内的当前平均访存次数;

根据所述上一时间周期对应的多个历史执行过程访存次数,确定所述上一时间周期内的历史平均访存次数;

在所述当前平均访存次数大于所述历史平均访存次数的情况下,在所述历史线程束重组数量阈值的基础上,进行单位数量缩减,确定所述线程束重组数量阈值。

6.根据权利要求1所述的一种基于GPGPU的线程重组方法,其特征在于,根据每个所述第一线程束的指定执行性能评估要素,在多个所述第一线程束中确定符合预设要求的多个指定线程束,以进行线程重组,具体包括:

将每个第一线程束对应的所述当前执行性能信息中的当前执行过程访存次数和当前执行通道冲突次数进行累加,确定每个所述第一线程束的指定执行性能评估要素;

按照由大到小的顺序,根据每个所述第一线程束的指定执行性能评估要素,对所述多个第一线程束进行排序,并按照所述线程束重组数量阈值,依次确定对应数量个指定线程束,以进行线程重组。

7.根据权利要求1所述的一种基于GPGPU的线程重组方法,其特征在于,通过每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,对多个指定线程束进行线程重组,具体包括:

获取所述当前处理单元对应的PC-Warp查找表,其中,所述PC-Warp查找表包括线程束标识表项、线程束执行PC表项和活跃线程状态标识表项;

每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,在所述PC-Warp查找表中进行表项更新,确定更新PC-Warp查找表,以根据所述更新PC-Warp查找表,确定多个指定线程束,以对所述多个指定线程束进行线程重组。

8.一种基于GPGPU的线程重组设备,其特征在于,所述设备包括:

至少一个处理器;以及,

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行如权利要求1-7任一所述的方法。

9.一种非易失性计算机存储介质,存储有计算机可执行指令,其特征在于,所述计算机可执行指令设置为:执行如权利要求1-7任一所述的方法。

一种基于GPGPU的线程重组方法、设备及介质

技术领域

[0001] 本说明书涉及线程分支管理技术领域,尤其涉及一种基于GPGPU的线程重组方法、设备及介质。

背景技术

[0002] 图形处理器通用计算(General-Purpose computing on Graphics Processing Units,即GPGPU)在执行时会产生线程分支,线程分支会导致计算资源的浪费,进而影响程序的性能。目前硬件上主要存在着多路交替执行和重组两种方法来对GPU中的非一致控制流进行优化,其中多路交替执行的方法对线程级并行的提升有限,在本质上是对基于栈的重汇聚机制的改进,而且在重汇聚时机的选择上仍然没有很好的解决。线程束(Warp)的重组策略虽然能够很好的提升线程级并行,但是并不是所有的重组都是有效的,而且重组在硬件上的开销也是不容忽视。

[0003] 对于活跃线程数较少的Warp重组,可以显著提高并行效率并减少Warp数量。但对于活跃线程数较多的Warp,重组可能会破坏线程间数据的连续性,特别是当这些线程原本访存连续时,重组会导致访存请求数量的增加,从而抵消了GPGPU访存合并机制带来的性能优势,增加了额外的重组开销。因此,目前传统的线程重组策略未考虑不同活跃线程数之间的线程重组差异性,对拥有不同活跃线程数量的线程束采取相同的重组方式,存在破坏线程间数据的连续性的风险,并且增加了额外的重组开销。

发明内容

[0004] 本说明书一个或多个实施例提供了一种基于GPGPU的线程重组方法、设备及介质,用于解决如下技术问题:传统的线程重组策略未考虑不同活跃线程数之间的线程重组差异性,对拥有不同活跃线程数量的线程束采取相同的重组方式,存在破坏线程间数据的连续性的风险,并且增加了额外的重组开销。

[0005] 本说明书一个或多个实施例采用下述技术方案:

[0006] 本说明书一个或多个实施例提供一种基于GPGPU的线程重组方法,所述方法包括:对当前处理单元中执行的多个线程束进行监测,采集当前时钟周期下每个所述线程束对应的当前线程执行信息,其中,所述当前线程执行信息包括当前执行性能信息、当前线程束执行PC信息和当前活跃线程信息;根据所述当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定所述当前时钟周期对应的当前调整活跃线程门限数据,其中,所述执行性能信息包括执行过程访存次数和执行通道冲突次数;基于所述当前调整活跃线程门限数据和每个所述线程束对应的当前活跃线程信息,设置每个所述线程束的活跃线程状态标识,以通过每个所述线程束的活跃线程状态标识和所述当前线程束执行PC信息,对多个指定线程束进行线程重组。

[0007] 本说明书一个或多个实施例提供一种基于GPGPU的线程重组设备,包括:

[0008] 至少一个处理器;以及,

[0009] 与所述至少一个处理器通信连接的存储器;其中,

[0010] 所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行上述方法。

[0011] 本说明书一个或多个实施例提供的一种非易失性计算机存储介质,存储有计算机可执行指令,所述计算机可执行指令设置为:执行上述方法。

[0012] 本说明书实施例采用的上述至少一个技术方案能够达到以下有益效果:通过上述技术方案,通过动态监测和调整线程束的活跃线程状态以及进行有针对性的线程重组,有效解决了传统线程重组策略中未考虑不同活跃线程数之间差异性的问题;通过动态调整活跃线程门限数据,并根据每个线程束的当前活跃线程信息来设置活跃线程状态标识,能够更精确地控制哪些线程束需要进行重组,有针对性的重组策略减少了不必要的重组操作,从而降低了额外的重组开销;能够根据当前时钟周期下的执行性能信息动态调整活跃线程门限数据和进行线程重组,这使得能够更灵活地应对不同的工作负载和场景,无论是处理高负载任务还是低负载任务,系统都能够通过优化线程束的活跃线程状态和重组策略来提高性能;通过综合考虑执行性能信息和活跃线程信息,能够避免因为盲目重组而导致的性能下降或系统崩溃,提高了并行效率、保护了数据连续性、减少了重组开销、提高了系统灵活性和稳定性。

附图说明

[0013] 为了更清楚地说明本说明书实施例或现有技术中的技术方案,下面将对实施例或现有技术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说明书中记载的一些实施例,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。在附图中:

[0014] 图1为本说明书实施例提供的一种基于GPGPU的线程重组方法的流程示意图;

[0015] 图2为本说明书实施例提供的一种线程自调流水框架图;

[0016] 图3为本说明书实施例提供的一种性能监测模块的功能示意图;

[0017] 图4为本说明书实施例提供的一种自调线程重组单元的功能示意图;

[0018] 图5为本说明书实施例提供的一种基于GPGPU的线程重组设备的结构示意图。

具体实施方式

[0019] 为了使本技术领域的人员更好地理解本说明书中的技术方案,下面将结合本说明书实施例中的附图,对本说明书实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本说明书一部分实施例,而不是全部的实施例。基于本说明书实施例,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,都应当属于本说明书保护的范围。

[0020] 本说明书实施例提供一种基于GPGPU的线程重组方法,需要说明的是,本说明书实施例中的执行主体可以是服务器,也可以是任意一种具备数据处理能力的设备。图1为本说明书实施例提供的一种基于GPGPU的线程重组方法的流程示意图,如图1所示,主要包括如下步骤:

[0021] 步骤S101,对当前处理单元中执行的多个线程束进行监测,采集当前时钟周期下

每个线程束对应的当前线程执行信息。

[0022] 在本说明书的一个实施例中,对当前处理单元中执行的多个线程束进行监测,采集当前时钟周期下每个线程束对应的当前线程执行信息,其中,该当前线程执行信息包括当前执行性能信息、当前线程束执行PC信息和当前活跃线程信息。当前处理单元是指流多处理器(Streaming Multiprocessor,SM),即SM处理单元。需要说明的是,此处的当前执行性能信息包括执行过程访存次数和执行通道冲突次数,访存次数指的是线程束在执行过程中访问存储器的次数,由于存储器的访问速度远低于处理器的执行速度,因此过多的访存操作会导致处理器等待存储器响应,从而降低整体性能;执行通道冲突次数指的是线程束在执行过程中由于资源竞争(如寄存器文件、执行单元等)而导致的冲突次数,这些冲突会导致线程束的执行被暂停或延迟;可以通过GPGPU通常配备的硬件性能计数器实时记录线程束的访存次数,通过专门的性能分析工具检测执行通道冲突次数。需要说明的是,线程束执行PC中PC对应的英文全称为Program Counter,即程序计数器,用于存执行指令的地址。

[0023] 步骤S102,根据当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定当前时钟周期对应的当前调整活跃线程门限数据。

[0024] 其中,该执行性能信息包括执行过程访存次数和执行通道冲突次数;

[0025] 在本说明书的一个实施例中,在每个时钟周期内对执行的多个线程束进行监测时,对每个时钟周期内每个线程束对应的线程执行信息进行存储,以获取位于当前时钟周期的上一时钟周期对应的历史执行性能信息。通过当前线程执行信息中的当前执行性能信息和上一时钟周期对应的历史执行性能信息,对执行性能进行监测,通过对线程束执行性能的变化,确定当前时钟周期对应的当前调整活跃线程门限数据。

[0026] 根据该当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息,对执行性能进行监测,以确定该当前时钟周期对应的当前调整活跃线程门限数据,具体包括:根据每个线程束对应的该当前执行性能信息和该历史执行性能信息,分别确定该当前时钟周期对应的当前执行性能评估指标和该上一时钟周期对应的历史执行性能评估指标;通过该当前执行性能评估指标和该历史执行性能评估指标,确定该当前时钟周期相对于该上一时钟周期的执行性能变化类型,其中,该执行性能变化类型包括性能提高和性能降低中的任意一项;获取该上一时钟周期对应的历史活跃线程门限数据,基于该执行性能变化类型,对该历史活跃线程门限数据进行调整,确定该当前时钟周期对应的当前调整活跃线程门限数据。

[0027] 在本说明书的一个实施例中,根据每个线程束对应的该当前执行性能信息和该历史执行性能信息,分别确定该当前时钟周期对应的当前执行性能评估指标和该上一时钟周期对应的历史执行性能评估指标。当该当前平均执行性能评估指标大于该历史平均执行性能评估指标时,确定该执行性能变化类型为性能降低;当该当前平均执行性能评估指标小于该历史平均执行性能评估指标时,确定该执行性能变化类型为性能提高。获取上一时钟周期对应的历史活跃线程门限数据,需要说明的是,若上一时钟周期不是第一个时钟周期,则上一时钟周期对应的历史活跃线程门限数据是根据上一时钟周期的前一时钟周期得到的,针对第一轮时钟周期的执行过程,其对应的初始活跃线程门限数据可以根据实际需求设置,例如,可设置为线程束中可以容纳最大的线程的数量的一半。在每一轮的执行过程

中,基于该执行性能变化类型,对上一轮的历史活跃线程门限数据进行调整,确定当前时钟周期对应的当前调整活跃线程门限数据。

[0028] 根据每个线程束对应的该当前执行性能信息和该历史执行性能信息,分别确定该当前时钟周期对应的当前执行性能评估指标和该上一时钟周期对应的历史执行性能评估指标,具体包括:将每个线程束对应的该当前执行性能信息中的当前执行过程访存次数和当前执行通道冲突次数进行累加,确定每个该线程束的第一执行性能评估指标;对每个该线程束对应的第一执行性能评估指标取均值,确定该当前时钟周期对应的当前执行性能评估指标;基于每个该历史执行性能信息中的历史执行过程访存次数和历史执行通道冲突次数进行累加,确定多个第二历史执行性能评估指标,以进行均值处理,确定该上一时钟周期对应的历史执行性能评估指标。

[0029] 在本说明书的一个实施例中,执行性能信息包括执行过程访存次数和执行通道冲突次数,对于每个线程束,统计在当前时钟周期内所有线程访问内存的总次数,统计在当前时钟周期内,由于线程束内线程尝试同时访问同一资源(如寄存器文件、共享内存或全局内存)而导致的冲突次数。对每个线程束,将上述两个指标相加,得到该线程束的第一执行性能评估指标。对所有线程束的第一执行性能评估指标取均值,得到当前时钟周期的当前执行性能评估指标,用于表示当前时钟周期内所有线程束的平均执行效率。

[0030] 按照同样的步骤对历史执行性能信息进行处理,得到上一个时钟周期内的历史执行性能评估指标。对每个线程束,将历史执行过程访存次数和历史执行通道冲突次数相加,得到线程束的第二历史执行性能评估指标,计算上一时钟周期的第二历史执行性能评估指标取均值,得到上一时钟周期的历史执行性能评估指标,用于表示上一时钟周期内所有线程束的平均执行效率。

[0031] 基于该执行性能变化类型,对该历史活跃线程门限数据进行调整,确定该当前时钟周期对应的当前调整活跃线程门限数据,具体包括:通过该当前执行性能评估指标和该历史执行性能评估指标,确定性能变化指标;根据该性能变化指标和该历史执行性能评估指标,确定该当前时钟周期的性能变化率,基于该性能变化率和该历史活跃线程门限数据,确定门限调整量;通过该执行性能变化类型,确定该门限调整量对应的取整方式和门限调整方式,其中,该取整方式包括向上取整和向下取整中的任意一项,该门限调整方式包括门限上调和门限下调中的任意一项;按照该取整方式对该门限调整量进行取整,确定门限调整步长;按照该门限调整方式,在该历史活跃线程门限数据的基础上,根据该门限调整步长进行调整,确定该当前时钟周期对应的当前调整活跃线程门限数据。

[0032] 在本说明书的一个实施例中,计算当前执行性能评估指标和历史执行性能评估指标之间的差值,这个差值的绝对值即为性能变化指标。根据性能变化指标和历史执行性能评估指标的比值,得到性能变化率。基于性能变化率和历史活跃线程门限数据,可以计算出门限调整量。需要说明的是,此处得到的门限调整量可能为非整数,在非整数的情况下,需要根据执行性能变化类型,对得到的门限调整量进行对应方向上的取整操作,例如向上取整、向下取整。例如当执行性能变化类型为性能下降时,为了保证最终得到的调整后的值与之前比时减少的,此处可以向上取整,保证在历史活跃线程门限数据减少门限调整步长后得到的为相对较小的值。当执行性能变化类型为性能提高时,此处可以向下取整,提供安全裕量。按照选定的取整方式,对门限调整量进行取整,得到门限调整步长,此处的门限调整

步长是指调整的门限量。此外,同样基于执行性能变化类型,选择门限上调或门限下调,此处的门限上调或门限下调是指调整的方向,即增加或减少数量。根据选定的门限调整方式和计算出的门限调整步长,在历史活跃线程门限数据的基础上进行调整,得到当前时钟周期对应的当前调整活跃线程门限数据。通过历史活跃线程门限数据增加或减少门限调整步长,得到当前调整活跃线程门限数据。

[0033] 步骤S103,基于当前调整活跃线程门限数据和每个线程束对应的当前活跃线程信息,设置每个线程束的活跃线程状态标识,以通过每个线程束的活跃线程状态标识和当前线程束执行PC信息,对多个指定线程束进行线程重组。

[0034] 在本说明书的一个实施例中,将每个线程束对应的当前活跃线程信息分别与当前调整活跃线程门限数据进行对比,根据对比结果,设置每个线程束的活跃线程状态标识,对低于当前调整活跃线程门限数据的线程束设置为低活跃状态,将不低于当前调整活跃线程门限数据的线程束设置为高活跃状态。通过每个线程束的活跃线程状态标识和当前线程束执行PC信息,对多个指定线程束进行线程重组。

[0035] 通过每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,对多个指定线程束进行线程重组,具体包括:通过每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,在多个该线程束中进行筛选,以确定符合预设要求的多个第一线程束,其中,该第一线程束的活跃线程状态标识为低活跃状态,且该多个第一线程束对应的当前线程束执行PC信息相同;确定该多个第一线程束的待重组线程束数量,并获取预先生成的线程束重组数量阈值;当该待重组线程束数量大于该线程束重组数量阈值时,根据每个该第一线程束的指定执行性能评估要素,在多个该第一线程束中确定符合预设要求的多个指定线程束,以进行线程重组,其中,该多个指定线程束的数量为该线程束重组数量阈值。

[0036] 在本说明书的一个实施例中,通过每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,遍历所有线程束,确定活跃线程状态标识为低活跃状态,并且当前执行PC信息相同的线程束,按照上述方式得到多个第一线程束。统计得到的第一线程束的数量,获取预先生成的线程束重组数量阈值,需要说明的是,此处的线程束重组数量阈值用于表示每次重组的线程束的数量阈值,通常为不小于2的正整数。每次重组时线程束的数量可能是较多数量的线程束,线程束重组数量阈值限制了一个门限,让参与此次重组的线程束不会高于这个门限值。当该待重组线程束数量不大于该线程束重组数量阈值时,则将得到的多个第一线程束全部进行重组。若该待重组线程束数量大于该线程束重组数量阈值,则需要在多个第一线程束中进行筛选,可以根据每个该第一线程束的指定执行性能评估要素,在多个该第一线程束中确定符合预设要求的多个指定线程束,对多个指定线程束进行线程重组,需要说明的是多个指定线程束的数量为该线程束重组数量阈值。

[0037] 获取预先生成的线程束重组数量阈值,具体包括:获取上一时间周期对应的历史线程束重组数量阈值;确定当前时间周期内每个该线程束对应的当前执行过程访存次数,以基于多个该当前执行过程访问次数,确定该当前时间周期内的当前平均访存次数;根据该上一时间周期对应的多个历史执行过程访存次数,确定该上一时间周期内的历史平均访存次数;在该当前平均访存次数大于该历史平均访存次数的情况下,在该历史线程束重组数量阈值的基础上,进行单位数量缩减,确定该线程束重组数量阈值。

[0038] 在本说明书的一个实施例中,获取上一时间周期对应的历史线程束重组数量阈

值,并确定当前时间周期内每个该线程束对应的当前执行过程访存次数,基于多个该当前执行过程访问次数,确定当前时间周期内的当前平均访存次数。按照同样的方式,计算上一时钟周期内的历史平均访存次数,即根据该上一时间周期对应的多个历史执行过程访存次数,确定该上一时间周期内的历史平均访存次数。若当前平均访存次数大于该历史平均访存次数,则在该历史线程束重组数量阈值的基础上减一,确定线程束重组数量阈值,但线程束重组数量阈值最小为2。需要说明的是,初始的线程束重组数量阈值可以设置为一个SM可容纳的最大线程束数量的一半。若当前平均访存次数不大于该历史平均访存次数,则采用历史线程束重组数量阈值,作为线程束重组数量阈值。

[0039] 根据每个该第一线程束的指定执行性能评估要素,在多个该第一线程束中确定符合预设要求的多个指定线程束,以进行线程重组,具体包括:将每个第一线程束对应的该当前执行性能信息中的当前执行过程访存次数和当前执行通道冲突次数进行累加,确定每个该第一线程束的指定执行性能评估要素;按照由大到小的顺序,根据每个该第一线程束的指定执行性能评估要素,对该多个第一线程束进行排序,并按照该线程束重组数量阈值,依次确定对应数量个指定线程束,以进行线程重组。

[0040] 在本说明书的一个实施例中,将每个第一线程束对应的该当前执行性能信息中的当前执行过程访存次数和当前执行通道冲突次数进行累加,确定每个该第一线程束的指定执行性能评估要素,此处可以通过上述步骤中各个线程束的第一执行性能评估指标确定。根据累加后的指定执行性能评估要素,对所有的第一线程束进行排序。排序的顺序应该是由大到小,因为访存次数和冲突次数较高的线程束更有可能成为性能瓶颈,因此应该优先考虑进行重组。按照排序后的顺序,依次选择指定数量的线程束作为指定线程束,指定数量等于之前确定的线程束重组数量阈值,得到的指定线程束将进行后续的线程重组操作。

[0041] 通过每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,对多个指定线程束进行线程重组,具体包括:获取该当前处理单元对应的PC-Warp查找表,其中,该PC-Warp查找表包括线程束标识表项、线程束执行PC表项和活跃线程状态标识表项;每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,在该PC-Warp查找表中进行表项更新,确定更新PC-Warp查找表,以根据该更新PC-Warp查找表,确定多个指定线程束,以对该多个指定线程束进行线程重组。

[0042] 在本说明书的一个实施例中,可以通过PC-Warp查找表的形式,通过每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,对多个指定线程束进行线程重组,接下来以PC-Warp查找表为例进行说明,首先,获取该当前处理单元对应的PC-Warp查找表,在PC-Warp查找表中包括线程束标识表项、线程束执行PC表项和活跃线程状态标识表项;线程束标识表项用于存储线程束ID,线程束执行PC表项用于存储线程束的执行PC,活跃线程状态标识表项用于存储线程束的活跃线程状态,除了上述表项之外,还可以额外添加表项用于存储每个线程束的执行性能评估指标等相关参数。根据每个该线程束的活跃线程状态标识和该当前线程束执行PC信息,在该PC-Warp查找表中进行表项更新,确定更新PC-Warp查找表,以根据该更新PC-Warp查找表,确定多个指定线程束,以对该多个指定线程束进行线程重组。在更新PC-Warp查找表针对表项进行检查,当线程束中活跃的线程数低于门限值,即对应低活跃状态时,便与PC-Warp中具有相同PC的,且同样为低活跃状态的线程束进行重组并送入Warp池当中等待调度。当高于门限值,即活跃线程状态标识为高活跃状态时,就判

定为不需要进行重组,就直接送入Warp池中等待调度。

[0043] 通过上述技术方案,通过动态监测和调整线程束的活跃线程状态以及进行有针对性的线程重组,有效解决了传统线程重组策略中未考虑不同活跃线程数之间差异性的问题;通过动态调整活跃线程门限数据,并根据每个线程束的当前活跃线程信息来设置活跃线程状态标识,能够更精确地控制哪些线程束需要进行重组,有针对性的重组策略减少了不必要的重组操作,从而降低了额外的重组开销;能够根据当前时钟周期下的执行性能信息动态调整活跃线程门限数据和进行线程重组,这使得能够更灵活地应对不同的工作负载和场景,无论是处理高负载任务还是低负载任务,系统都能够通过优化线程束的活跃线程状态和重组策略来提高性能;通过综合考虑执行性能信息和活跃线程信息,能够避免因为盲目重组而导致的性能下降或系统崩溃,提高了并行效率、保护了数据连续性、减少了重组开销、提高了系统灵活性和稳定性。

[0044] 在本说明书的一个实施例中,为了提高程序遇到分支时执行的性能,降低线程束重组时的开销,提高线程重组的效率,因此针对拥有不同活跃线程数量的线程束采取不同的重组方式。设置性能监测单元,监测当前执行的线程重组方式的性能,并根据执行性能动态调节线程重组条件,并设置自调线程重组单元对符合重组条件的线程束进行重组。

[0045] 图2为本说明书实施例提供的一种线程自调流水框架图,如图2所示,自调架构包含,线程调度器、流水单元、寄存器文件单元、并行执行单元、性能监测模块、自调线程重组单元、线程调度器、缓存存储单元。线程调度器根据得到的线程束信息,采用特定的调度策略进行线程束的调度执行;流水单元,包含取指、译码、记分牌、指令发射等功能,实现指令的流水执行,寄存器文件单元用于存储每个线程执行所需要的操作数信息。在图2中,并行执行单元以算术逻辑单元(Arithmetic Logic Unit,ALU)进行示例性展示,用于根据指令执行的信息进行操作数的特定运算。性能监测模块用于收集线程执行过程中的信息,并将整合的信息发送给自调线程重组单元。自调线程重组单元根据性能监测模块反馈的信息,自动调节线程束重组的方式,达到提升程序执行性能的结果。缓存存储单元用于存储指令执行的数据,可通过加载/存储单元(Load/Store Unit, LSU),即通过LSU模块加载。

[0046] 首先,线程调度器根据自调线程重组单元反馈的线程束进行调度取指,将对应线程束的指令从指令单元中取出,送入到流水单元中执行,主要进行指令译码,记分牌检查冲突,并将符合要求且已经从寄存器文件中获得操作数的线程束发送给后方的执行单元。执行线程束指令之后,性能监测模块会对执行过程中的访存次数以及执行通道冲突次数进行收集,以用来动态调整门限值。调整后的门限值会发送给自调线程重组单元,自调线程重组单元会根据门限值来筛选需要进行重组的线程束,并将重组后的和不符合重组条件的线程束都放入到Warp池中,以供调度器选择。

[0047] 图3为本说明书实施例提供的一种性能监测模块的功能示意图,如图3所示,性能监测模块功能会收集执行过程中的访存次数以及执行通道的冲突次数并进行累加,然后将活跃线程门限值上调(调大),此后在下一轮流水的线程束重组后,通过访存次数和执行通道冲突次数之和来判断性能是否提高,当总和次数变少就是性能提高,当总和次数变多就是性能降低。当性能提高了就继续上调活跃线程门限值,当性能降低了就下调活跃线程门限值,以此循环动态调整,该活跃线程门限值每一次调整后均会被送入自调线程重组单元中。除了对活跃线程门限值的调整之外,收集执行过程中的访存次数,将访存次数与历史访

存次数进行比较,判断访存次数是否增加,若是,则降低线程束数量门限值,若否,保持上一次的线程束数量门限值不变,将得到的线程束数量门限值送入自调线程重组单元。

[0048] 图4为本说明书实施例提供的一种自调线程重组单元的功能示意图,如图4所示,该模块在每一个时钟周期,生成一个PC-Warp的查找表,该查找表包含Warp的ID,以及它所执行的PC和活跃线程的数量。此后就针对查找表的表项进行检查,进行活跃的线程数与活跃线程束门限值的条件1判断过程,当Warp中活跃的线程数不低于门限值时,判定为不需要进行重组,直接进入Warp池等待调度。当Warp中活跃的线程数低于门限值时,确定活跃的线程数低于门限值的线程束数量,判断此数量是否大于线程束数量门限值,若否,则将PC-Warp中符合条件上述条件的所有同PC的线程束进行重组;若是,则优先选取PC-Warp中符合条件1的活跃线程数量较低的同PC的线程束重组,也可以在符合条件1且同PC的线程束中,选择与线程束数量门限值相同数量的访存次数和冲突次数较高的线程束进行重组。在重组后送入Warp池当中等待调度。

[0049] 通过上述技术方案,每次指令执行后,门限值都可以根据程序执行的性能进行自适应动态调整,以用来满足不同功能的程序块,降低了线程重组的开销,且具备更高的灵活性以及更好的分支执行性能。

[0050] 本说明书实施例还提供一种基于GPGPU的线程重组设备,如图5所示,设备包括:至少一个处理器;以及,与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行上述方法。

[0051] 本说明书实施例还提供一种非易失性计算机存储介质,存储有计算机可执行指令,计算机可执行指令设置为:执行上述方法。

[0052] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于装置、设备、非易失性计算机存储介质实施例而言,由于其基本相似于方法实施例,所以描述的比较简单,相关之处参见方法实施例的部分说明即可。

[0053] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0054] 本说明书实施例提供的设备和介质与方法是一一对应的,因此,设备和介质也具有与其对应的方法类似的有益技术效果,由于上面已经对方法的有益技术效果进行了详细说明,因此,这里不再赘述设备和介质的有益技术效果。

[0055] 本领域内的技术人员应明白,本说明书的实施例可提供为方法、系统、或计算机程序产品。因此,本说明书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,本说明书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。

[0056] 本说明书是参照根据本说明书实施例的方法、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。

[0057] 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设备以特定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。

[0058] 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设备上,使得在计算机或其他可编程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。

[0059] 在一个典型的配置中,计算设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

[0060] 内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器 (RAM) 和/或非易失性内存等形式,如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

[0061] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体 (transitory media),如调制的数据信号和载波。

[0062] 还需要说明的是,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、商品或者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、商品或者设备所固有的要素。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、商品或者设备中还存在另外的相同要素。

[0063] 以上所述仅为本说明书的一个或多个实施例而已,并不用于限制本说明书。对于本领域技术人员来说,本说明书的一个或多个实施例可以有各种更改和变化。凡在本说明书的一个或多个实施例的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本说明书的权利要求范围之内。

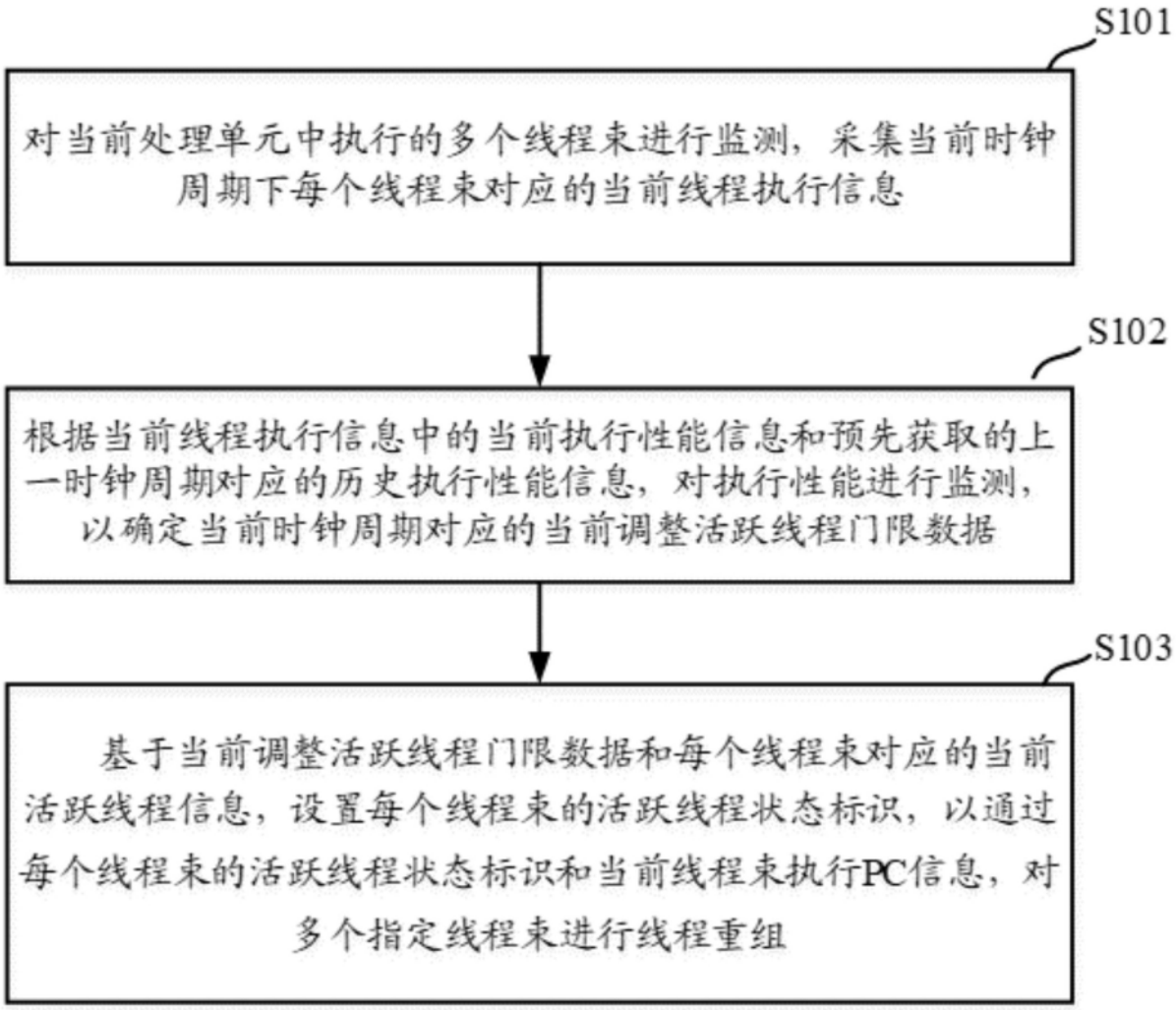

S101
对当前处理单元中执行的多个线程束进行监测，采集当前时钟周期下每个线程束对应的当前线程执行信息
S102
根据当前线程执行信息中的当前执行性能信息和预先获取的上一时钟周期对应的历史执行性能信息，对执行性能进行监测，以确定当前时钟周期对应的当前调整活跃线程门限数据
S103
基于当前调整活跃线程门限数据和每个线程束对应的当前活跃线程信息，设置每个线程束的活跃线程状态标识，以通过每个线程束的活跃线程状态标识和当前线程束执行PC信息，对多个指定线程束进行线程重组

图1

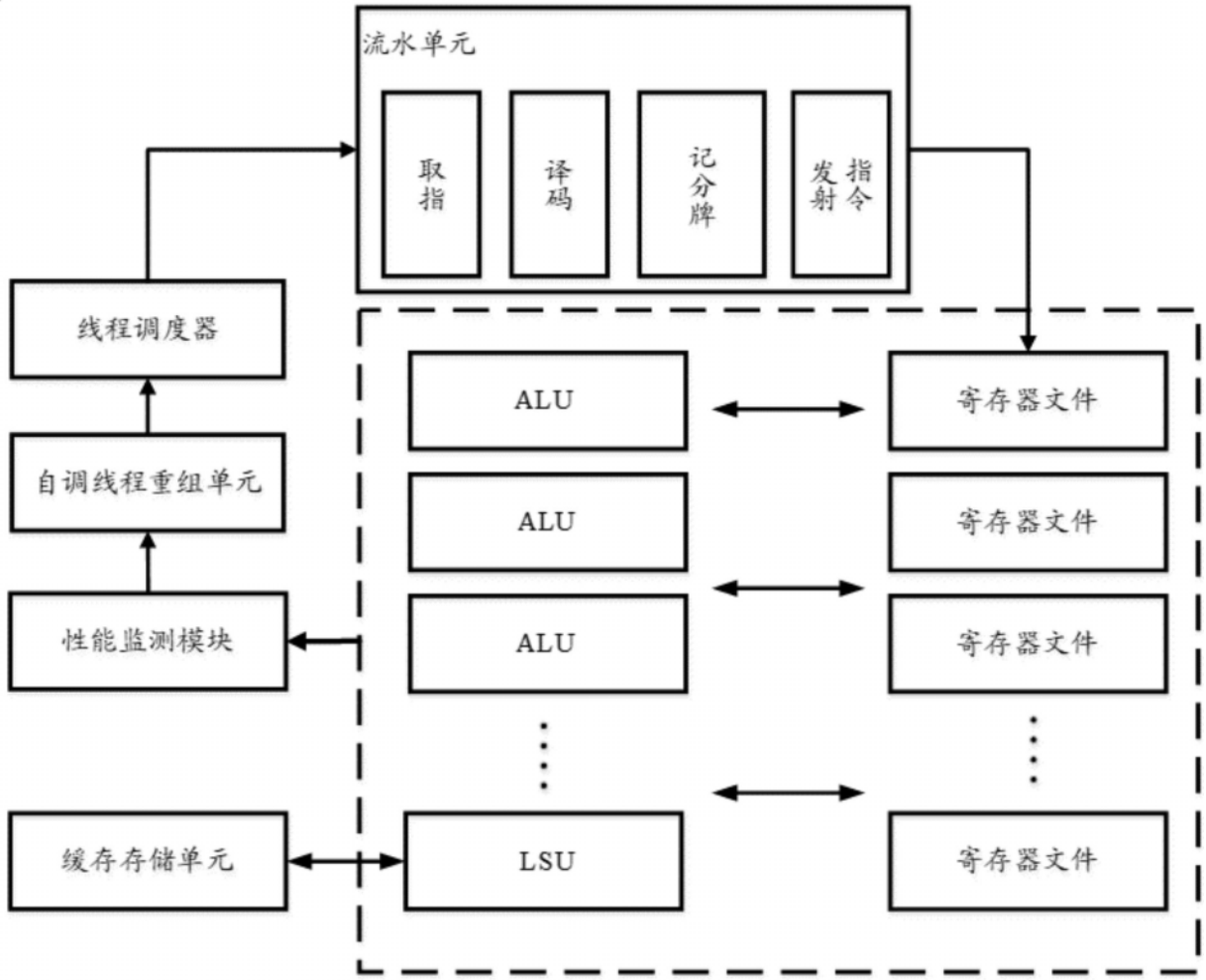

流水单元
取指
译码
记分牌
指令发射
线程调度器
自调线程重组单元
性能监测模块
缓存存储单元
ALU
ALU
ALU
LSU
寄存器文件
寄存器文件
寄存器文件
寄存器文件

图2

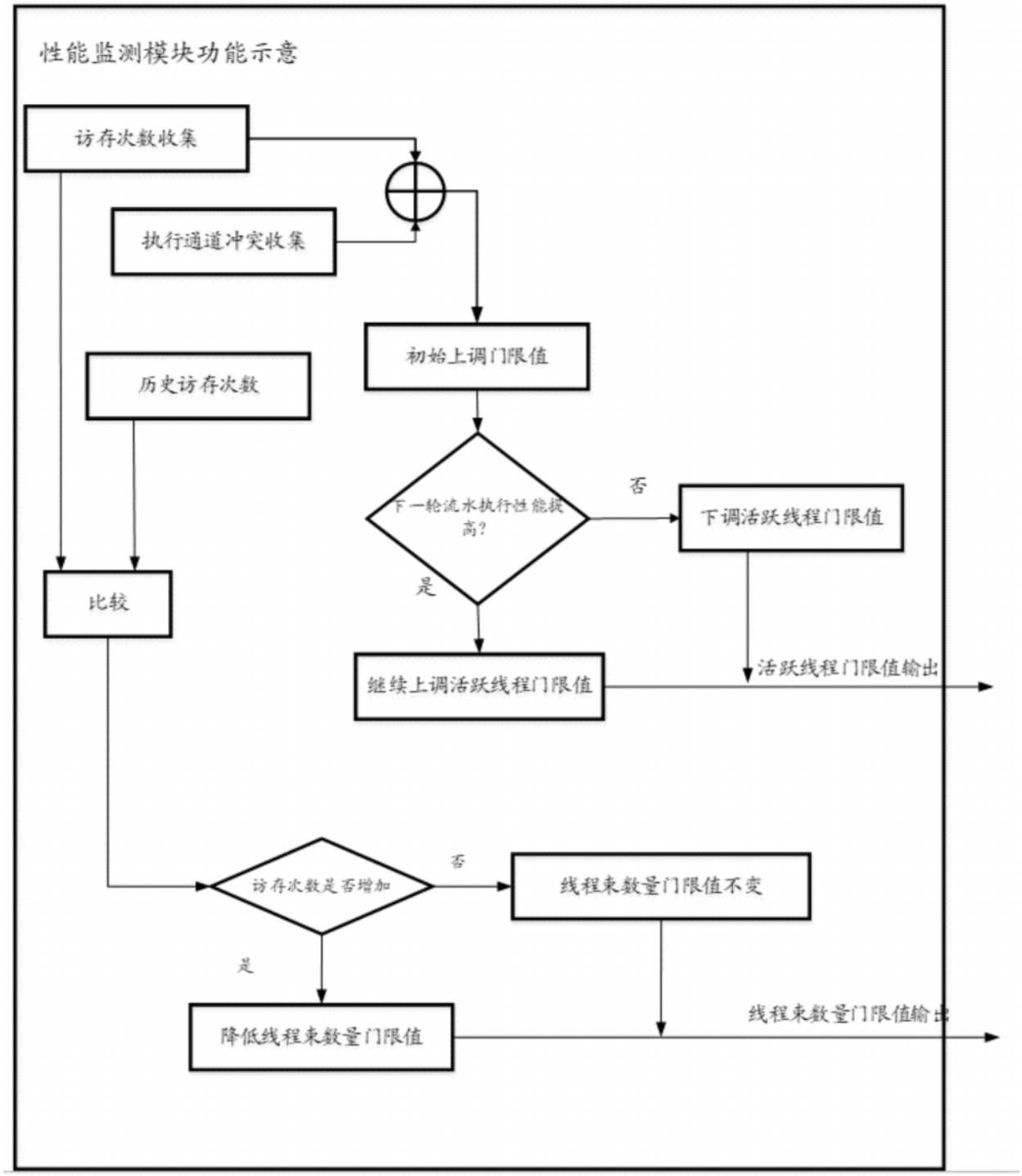

性能监测模块功能示意
访存次数收集
执行通道冲突收集
初始上调门限值
历史访存次数
下一轮流水执行性能提高?
否
下调活跃线程门限值
是
比较
继续上调活跃线程门限值
活跃线程门限值输出
访存次数是否增加
否
线程束数量门限值不变
是
降低线程束数量门限值
线程束数量门限值输出

图3

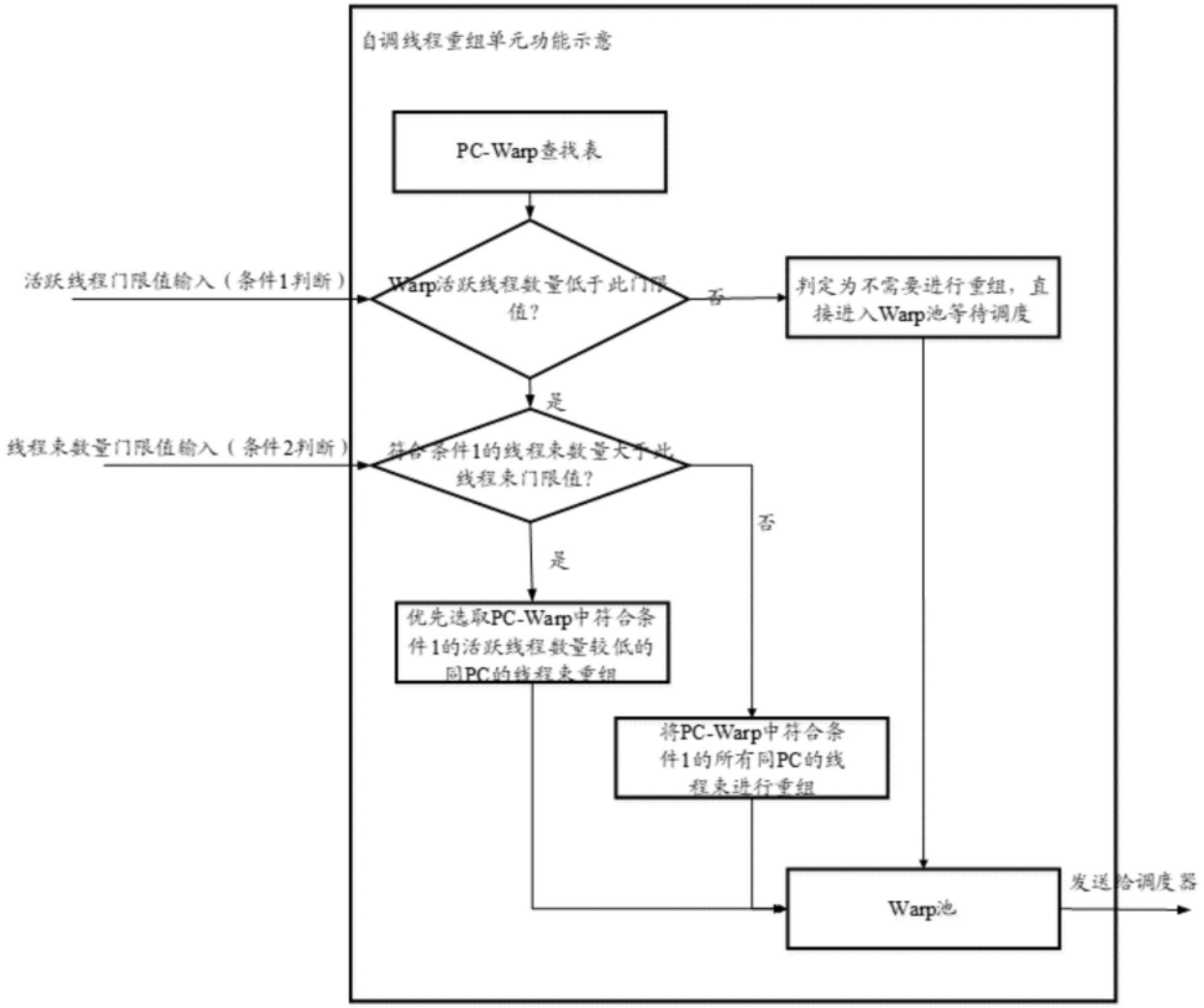

自调线程重组单元功能示意
PC-Warp查找表
活跃线程门限值输入（条件1判断）
Warp活跃线程数量低于此门限值?
否
判定为不需要进行重组，直接进入Warp池等待调度
是
线程束数量门限值输入（条件2判断）
符合条件1的线程束数量大于此线程束门限值?
否
是
优先选取PC-Warp中符合条件1的活跃线程数量较低的同PC的线程束重组
将PC-Warp中符合条件1的所有同PC的线程束进行重组
Warp池
发送给调度器

图4

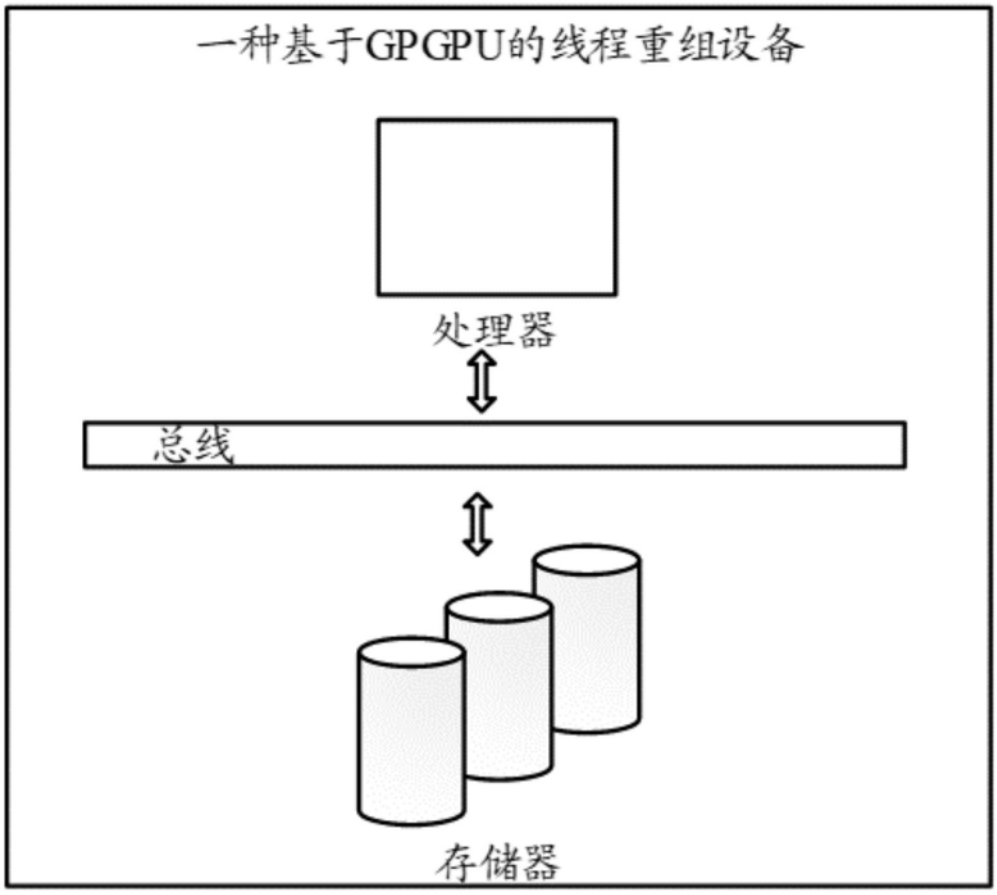


图5