(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 112988364 B
(45)授权公告日 2021.09.24

(21)申请号 202110550092.0
(22)申请日 2021.05.20
(65)同一申请的已公布的文献号
申请公布号 CN 112988364 A
(43)申请公布日 2021.06.18
(73)专利权人 西安芯瞳半导体技术有限公司
地址 710065 陕西省西安市高新区丈八街办丈八一路3号旺都1幢2单元11层21101号
(72)发明人 张竞丹 陈成 孙建康 樊良辉
(74)专利代理机构 西安维英格知识产权代理事务所(普通合伙) 61253
代理人 李斌栋 沈寒酉
(51)Int.Cl.
G06F 9/48(2006.01)
G06F 9/50(2006.01)

(56)对比文件
CN 103207814 A,2013.07.17
CN 102508704 A,2012.06.20
US 10937119 B2,2021.03.02
CN 103699445 A,2014.04.02

审查员 周静奇

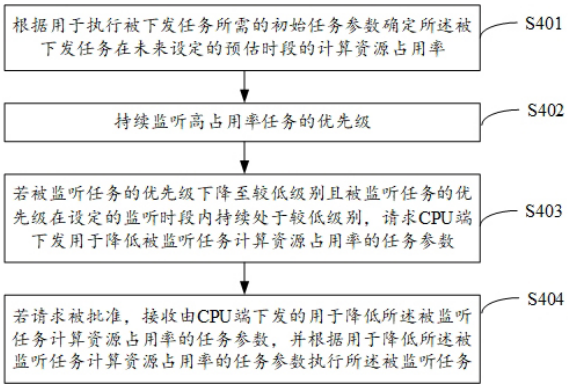
权利要求书2页 说明书10页 附图4页

(54)发明名称

一种动态的任务调度方法、装置及存储介质

(57)摘要

本发明实施例公开了一种动态的任务调度方法、装置及存储介质;该方法可以包括:根据用于执行被下发任务所需的当前任务参数确定所述被下发任务在未来设定的预估时段的计算资源占用率;持续监听高占用率任务的优先级;若被监听任务的优先级下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别,请求CPU端下发用于降低所述被监听任务计算资源占用率的任务参数;若所述请求被批准,接收由CPU端下发的用于降低所述被监听任务计算资源占用率的任务参数,并根据所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。



CN 112988364 B

1.一种动态的任务调度装置,其特征在于,所述装置包括:GPU驱动程序以及处于GPU结构内的命令处理器、处理器集群和资源监控器;其中,

所述命令处理器,经配置为将由所述GPU驱动程序下发的待执行任务分配至所述处理器集群中的执行核后,根据所述待执行任务的优先级以及当前任务参数更新所述资源监控器中的资源列表;

所述资源监控器,经配置为根据所述资源列表中的待执行任务的当前任务参数估算在未来设定的预估时段内所述待执行任务的计算资源占用率;并持续监听所述待执行任务中高占用率任务的优先级;以及,若被监听任务的优先级基于用户操作对象的改变下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别,向所述命令处理器发出第一通知指示;

所述命令处理器,还经配置为基于所述第一通知指示向所述GPU驱动程序请求下发用于降低所述被监听任务计算资源占用率的任务参数;以及,接收由所述GPU驱动程序基于所述请求被批准而下发的用于降低所述被监听任务计算资源占用率的任务参数,并将所述用于降低所述被监听任务计算资源占用率的任务参数下发至所述被监听任务的执行核以按照所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

2.根据权利要求1所述的任务调度装置,其特征在于,所述资源监控器,经配置为:相应于所述待执行任务的计算资源占用率高于设定的阈值,确定所述待执行任务为高占用率任务,并持续监听高占用率任务的优先级。

3.根据权利要求1所述的任务调度装置,其特征在于,所述GPU驱动程序,经配置为对接收到的请求进行仲裁;若仲裁结果为所述请求被批准,则向所述命令处理器下发用于降低所述被监听任务计算资源占用率的任务参数;否则,向所述命令处理器反馈拒绝指令,以拒绝下发用于降低所述被监听任务计算资源占用率的任务参数。

4.根据权利要求1所述的任务调度装置,其特征在于,所述资源监控器,还经配置为:若所述被监听任务的优先级由较低级别转至高级别且所述被监听任务的优先级在设定的监听时段内持续处于高级别,向所述命令处理器发出第二通知指示;

所述命令处理器,还经配置为基于所述第二通知指示向所述GPU驱动程序请求下发所述被监听任务的当前任务参数;以及,接收由所述GPU驱动程序基于所述请求被批准而下发的所述被监听任务的当前任务参数,并将所述被监听任务的当前任务参数下发至所述被监听任务的执行核以重新按照所述被监听任务的当前任务参数执行所述被监听任务。

5.根据权利要求1所述的任务调度装置,其特征在于,所述资源监控器,还经配置为:若所述被监听任务根据所述用于降低所述被监听任务计算资源占用率的任务参数仍被确定为高占用率任务,继续向所述命令处理器发出第一通知指示;

所述命令处理器,还经配置为基于所述第一通知指示继续向所述GPU驱动程序请求下发用于继续降低所述被监听任务计算资源占用率的任务参数;以及,接收由所述GPU驱动程序基于所述请求被批准而下发的用于继续降低所述被监听任务计算资源占用率的任务参数,并将所述用于继续降低所述被监听任务计算资源占用率的任务参数下发至所述被监听任务的执行核以按照所述用于继续降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

6.一种动态的任务调度方法,其特征在于,所述方法包括:

根据用于执行被下发任务所需的当前任务参数确定所述被下发任务在未来设定的预估时段的计算资源占用率；

持续监听高占用率任务的优先级；

若被监听任务的优先级基于用户操作对象的改变下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别，请求CPU端下发用于降低所述被监听任务计算资源占用率的任务参数；

若所述请求被批准，接收由CPU端下发的用于降低所述被监听任务计算资源占用率的任务参数，并根据所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

7.根据权利要求6所述的任务调度方法，其特征在于，所述持续监听高占用率任务的优先级，包括：

判断所述待执行任务的计算资源占用率是否大于设定的阈值：

若大于，则确定所述待执行任务为高占用率任务；

否则确定所述待执行任务为低占用率任务。

8.根据权利要求6所述的任务调度方法，其特征在于，所述方法还包括：

通过所述CPU端的GPU驱动程序对接收到的请求进行仲裁；

若仲裁结果为所述请求被批准，则下发用于降低所述被监听任务计算资源占用率的任务参数；否则，反馈拒绝指令，以拒绝下发用于降低所述被监听任务计算资源占用率的任务参数。

9.根据权利要求6所述的任务调度方法，其特征在于，所述方法还包括：

若所述被监听任务的优先级由较低级别转至高级别且所述被监听任务的优先级在设定的监听时段内持续处于高级别，根据所述被监听任务的当前任务参数执行所述被监听任务。

10.根据权利要求6所述的任务调度方法，其特征在于，所述方法还包括：

若所述被监听任务根据所述用于降低所述被监听任务计算资源占用率的任务参数仍被确定为高占用率任务，继续请求CPU端下发用于继续降低所述被监听任务计算资源占用率的任务参数，直至所述被监听任务的计算资源占用率下降至设定的范围。

11.一种计算机存储介质，其特征在于，所述计算机存储介质存储有动态的任务调度程序，所述动态的任务调度程序被至少一个处理器执行时实现权利要求6至10任一项所述的动态的任务调度方法的步骤。

一种动态的任务调度方法、装置及存储介质

技术领域

[0001] 本发明实施例涉及图像处理技术领域，尤其涉及一种动态的任务调度方法、装置及存储介质。

背景技术

[0002] 目前，图形处理器(GPU，Graphics Processing Unit)在执行CPU所下发的任务过程中，通常会根据任务的优先级的顺序执行；此外，各任务之间的复杂程度各不相同，高复杂度的任务将会占用GPU内较多的计算资源。如果当前正在处理的优先级较高的高复杂度任务基于事件或操作导致优先级降低，此时该任务所占用的GPU资源仍然持续被占用以继续处理该任务，即使GPU剩余的计算资源任然能够完成其他任务的处理，但是会导致GPU的功耗提升。

发明内容

[0003] 有鉴于此，本发明实施例期望提供一种动态的任务调度方法、装置及存储介质；能够降低GPU的在计算资源占用率，节省GPU的功耗。

[0004] 本发明实施例的技术方案是这样实现的：

[0005] 第一方面，本发明实施例提供了一种动态的任务调度装置，所述装置包括：GPU驱动程序以及处于GPU结构内命令处理器、处理器集群和资源监控器；其中，

[0006] 所述命令处理器，经配置为将由所述GPU驱动程序下发的待执行任务分配至所述处理器集群中的执行核后，根据所述待执行任务的优先级以及当前任务参数更新所述资源监控器中的资源列表；

[0007] 所述资源监控器，经配置为根据所述资源列表中的待执行任务的当前任务参数估算在未来设定的预估时段内所述待执行任务的计算资源占用率；并持续监听所述待执行任务中高占用率任务的优先级；以及，若被监听任务的优先级下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别，向所述命令处理器发出第一通知指示；

[0008] 所述命令处理器，还经配置为基于所述第一通知指示向所述GPU驱动程序请求下发用于降低所述被监听任务计算资源占用率的任务参数；以及，接收由所述GPU驱动程序基于所述请求被批准而下发的用于降低所述被监听任务计算资源占用率的任务参数，并将所述用于降低所述被监听任务计算资源占用率的任务参数下发至所述被监听任务的执行核以按照所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

[0009] 第二方面，本发明实施例提供了一种动态的任务调度方法，所述方法包括：

[0010] 根据用于执行被下发任务所需的当前任务参数确定所述被下发任务在未来设定的预估时段的计算资源占用率；

[0011] 持续监听高占用率任务的优先级；

[0012] 若被监听任务的优先级下降至较低级别且所述被监听任务的优先级在设定的监

听时段内持续处于较低级别，请求CPU端下发用于降低所述被监听任务计算资源占用率的任务参数；

[0013] 若所述请求被批准，接收由CPU端下发的用于降低所述被监听任务计算资源占用率的任务参数，并根据所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

[0014] 第三方面，本发明实施例提供了一种计算机存储介质，所述计算机存储介质存储有动态的任务调度程序，所述动态的任务调度程序被至少一个处理器执行时实现第二方面所述的动态的任务调度方法的步骤。

[0015] 本发明实施例提供了一种动态的任务调度方法、装置及存储介质；当计算资源占用率较高的任务所对应的优先级下降并持续一段时间之后，可以通过调整该任务的任务参数以降低执行该任务所需消耗的计算资源，从而降低GPU的计算资源占用率，节省GPU的功耗。

附图说明

[0016] 图1为可实施本发明实施例一个或多个方面的计算装置的框图。

[0017] 图2为说明图1中处理器、GPU和系统存储器的实例实施方案的框图。

[0018] 图3为可实施本发明实施例一个或多个方面的GPU和结构的实施方案框图。

[0019] 图4为本发明实施例提供的一种动态的任务调度方法流程示意图。

[0020] 图5为本发明实施例提供的一种判断高占用率任务示意图。

具体实施方式

[0021] 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。

[0022] 参见图1，其示出了能够实现本发明实施例技术方案的计算装置2，该计算装置2的实例包括但不限于：无线装置、移动或蜂窝电话(包含所谓的智能电话)、个人数字助理(PDA)、视频游戏控制台(包含视频显示器、移动视频游戏装置、移动视频会议单元)、膝上型计算机、桌上型计算机、电视机顶盒、平板计算装置、电子书阅读器、固定或移动媒体播放器，等。在图1的实例中，该计算装置2可以包括：处理器6、系统存储器10和GPU 12。计算装置2还可包含显示处理器14、收发器模块3、用户接口4和显示器8。收发器模块3和显示处理器14两者可为与处理器6和/或GPU 12相同的集成电路(IC)的部分，两者可在包含处理器6和/或GPU 12的一或多个IC的外部，或可形成于在包含处理器6和/或GPU 12的IC外部的IC中。

[0023] 为清楚起见，计算装置2可包含图1中未图示的额外模块或单元。举例来说，计算装置2可在其中计算装置2为移动无线电话或的实例中包含扬声器和麦克风(两者均未在图1中示出)来实现电话通信，或在计算装置2为媒体播放器的情况下包含扬声器。计算装置2还可包含摄像机。此外，计算装置2中所示的各种模块和单元可能不是在计算装置2的每个实例中都是必需的。举例来说，在计算装置2为桌上型计算机或经装备以与外部用户接口或显示器介接的其它装置的实例中，用户接口4和显示器8可在计算装置2外部。

[0024] 用户接口4的实例包含(但不限于)轨迹球、鼠标、键盘和其它类型的输入装置。用户接口4还可为触摸屏，并且可作为显示器8的部分并入。收发器模块3可包含电路以允许计

算装置2与另一装置或网络之间的无线或有线通信。收发器模块3可包含调制器、解调器、放大器和用于有线或无线通信的其它此类电路。

[0025] 处理器6可为微处理器，例如中央处理单元(CPU)，其经配置以处理供执行的计算机程序的指令。处理器6可包括控制计算装置2的运算的通用或专用处理器。用户可将输入提供到计算装置2，以致使处理器6执行一或多个软件应用程序。在处理器6上执行的软件应用程序可包含(例如)操作系统、文字处理器应用程序、电子邮件应用程序、电子表格应用程序、媒体播放器应用程序、视频游戏应用程序、图形用户接口应用程序或另一程序。另外，处理器6可执行用于控制GPU 12的运算的GPU驱动程序22。用户可经由一或多个输入装置(未图示)(例如，键盘、鼠标、麦克风、触摸垫或经由用户输入接口4耦合到计算装置2的另一输入装置)将输入提供到计算装置2。

[0026] 在处理器6上执行的软件应用程序可包含一或多个图形渲染指令，其指令处理器6来致使将图形数据渲染到显示器8。在一些实例中，所述软件指令可符合图形应用程序编程接口(API)，例如开放式图形库API、开放式图形库嵌入系统(OpenGL ES)API、Direct3D API、X3D API、RenderMan API、WebGL API、开放式计算语言(OpenCLT M)、RenderScript或任何其它异构计算API，或任何其它公用或专有标准图形或计算API。所述软件指令还可为针对无渲染算法(例如计算摄影、卷积神经网络、视频处理、科学应用程序等)的指令。为了处理图形渲染指令，处理器6可向GPU 12发出一或多个图形渲染命令发(例如，通过GPU驱动程序22)，以致使GPU 12执行图形数据的渲染中的一些或全部。在一些实例中，待渲染的图形数据可包含例如点、线、三角形、四边形、三角形带等图形图元的列表。

[0027] GPU 12可经配置以执行图形运算，从而将一或多个图形图元渲染到显示器8。因此，当在处理器6上执行的软件应用中的一者需要图形处理时，处理器6可将图形命令和图形数据提供到GPU 12以用于渲染到显示器8。图形数据可包含(例如)绘制命令、状态信息、图元信息、纹理信息等。在一些情况下，GPU 12可内置有高度并行结构，其提供比处理器6高效的对复杂图形相关运算的处理。举例来说，GPU 12可包含经配置来以并行方式对多个顶点或像素进行运算的多个处理元件，例如着色器单元。在一些情况下，GPU 12的高度并行性质允许GPU 12比使用处理器6直接将场景绘制到显示器8更快速地将图形图像(例如，GUI和二维(2D)和/或三维(3D)图形场景)绘制到显示器8上。

[0028] 在一些情况下，可将GPU 12集成到计算装置2的母板中。在其它情况下，GPU 12可存在于图形卡上，所述图形卡安装在计算装置2的母板中的端口中，或可以其它方式并入在经配置以与计算装置2互操作的外围装置内。GPU 12可包含一或多个处理器，例如一或多个微处理器、专用集成电路(ASIC)、现场可编程门阵列(FPGA)、数字信号处理器(DSP)或其它等效的集成或离散逻辑电路。GPU 12还可包含一或多个处理器核心，使得GPU 12可被称作多核处理器。

[0029] 图形存储器40可为GPU 12的一部分。因此，GPU 12可在不使用总线的情况下从图形存储器40读取数据且将数据写入到图形存储器40。换句话说，GPU 12可使用本地存储装置而不是芯片外存储器在本地处理数据。此类图形存储器40可被称作芯片上存储器。这允许GPU 12通过消除GPU 12经由总线读取和写入数据的需要来以更高效的方式操作，其中经由总线操作可经历繁重的总线业务。然而，在一些情况下，GPU 12可不包含单独的存储器，而是经由总线利用系统存储器10。图形存储器40可包含一或多个易失性或非易失性存储器

或存储装置，例如，随机存取存储器(RAM)、静态RAM(SRAM)、动态RAM(DRAM)、可擦除可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)、快闪存储器、磁性数据媒体或光学存储媒体。

[0030] 在一些实例中，GPU 12可将完全形成的图像存储在系统存储器10中。显示处理器14可从系统存储器10检索图像，且输出致使显示器8的像素照亮以显示所述图像的值。显示器8可为计算装置2的显示器，其显示由GPU 12产生的图像内容。显示器8可为液晶显示器(LCD)、有机发光二极管显示器(OLED)、阴极射线管(CRT)显示器、等离子显示器或另一类型的显示装置。

[0031] 图2是进一步详细说明图1中处理器6、GPU 12和系统存储器10的实例实施方案的框图。如图2所示，处理器6可执行至少一个软件应用程序18、图形API 20和GPU 驱动程序22，其中的每一者可为一或多个软件应用程序或服务。在一些实例中，图形API 20和GPU驱动程序22可实施为CPU 6的硬件单元。

[0032] 可供处理器6和GPU 12使用的存储器可包含系统存储器10和输出缓冲器16。输出缓冲器16可为系统存储器10的部分或可与系统存储器10分离。输出缓冲器16可存储经渲染图像数据，例如像素数据，以及任何其它数据。输出缓冲器16还可被称为帧缓冲器或显存。

[0033] 图形存储器40可包含片上存储装置或存储器，其物理上集成到GPU12的集成电路芯片中。如果图形存储器40是在芯片上，那么与经由系统总线从系统存储器10读取值或将值写入到系统存储器10相比，GPU12能够更加快速地从图形存储器40读取值或将值写入到图形存储器40。

[0034] 输出缓冲器16存储GPU 12的目的地像素。每个目的地像素可与唯一屏幕像素位置相关联。在一些实例中，输出缓冲器16可存储每个目的地像素的色彩分量和目的地α值。举例来说，输出缓冲器16可存储每个像素的红色、绿色、蓝色、α(RGBA)分量，其中“RGB”分量对应于色彩值，并且“A”分量对应于目的地α值(例如，用于图像合成的不透明度值)。尽管将输出缓冲器16和系统存储器10说明为单独的存储器单元，但在其它实例中，输出缓冲器16可以是系统存储器10的一部分。此外，输出缓冲器16还可能够存储除像素之外的任何合适的数据。

[0035] 软件应用程序18可为利用GPU 12的功能性的任何应用程序。举例来说，软件应用程序18可为GUI应用程序、操作系统、便携式制图应用程序、用于工程或艺术应用的计算机辅助设计程序、视频游戏应用程序或使用2D或3D图形的另一类型的软件应用程序。

[0036] 软件应用程序18可包含指令GPU 12渲染图形用户接口(GUI)和/或图形场景的一或多个绘制指令。举例来说，绘制指令可包含界定将由GPU 12渲染的一组一或多个图形图元的指令。在一些实例中，绘制指令可共同地界定用于GUI中的多个开窗表面的全部或部分。在额外实例中，所述绘制指令可共同地定义图形场景的全部或部分，所述图形场景包含在由应用程序定义的模型空间或世界空间内的一或多个图形对象。

[0037] 软件应用程序18可经由图形API 20调用GPU驱动程序22，以向GPU 12发出一或多个命令，以用于将一或多个图形图元渲染到可显示的图形图像中。举例来说，软件应用程序18可调用GPU驱动程序22，以向GPU 12提供图元定义。在一些情况下，图元定义可以例如三角形、矩形、三角形扇、三角形带等的绘制图元的列表的形式被提供到GPU 12。图元定义可包含指定与待呈现的图元相关联的一或多个顶点的顶点规格。所述顶点规格可包含每个顶点的位置坐标，且在一些情况下包含与顶点相关联的其它属性，例如色彩属性、法向量和纹

理坐标。图元定义还可包含图元类型信息(例如,三角形、矩形、三角形扇、三角形带等)、缩放信息、旋转信息及类似者。

[0038] 基于由软件应用程序18向GPU驱动程序22发出的指令,GPU驱动程序22可调配指定供GPU 12执行的一或多个运算以便渲染图元的一或多个命令。当GPU 12接收到来自CPU 6的命令时,GPU 12可使用处理器集群46执行图形处理管线,以便对命令进行解码,并对图形处理管线进行配置以执行命令中所制定的操作。所以,在一些示例中,处理器集群46也可被称之为GPU 12的计算资源。

[0039] 处理器集群46可包含一或多个可编程执行核24和/或一或多个固定功能执行核26。对于上述两种处理单元,可编程执行核24可包含例如被配置成执行从CPU 6下载到GPU 12上的一或多个着色器程序的可编程着色器单元。在一些实例中,可编程着色器单元可被称为“着色器处理器”或“统一着色器”,且可被配置为能够至少执行顶点和片元着色操作以呈现图形;可选地,可编程着色器单元还可以被配置执行几何或其它着色操作以呈现图形。因此,处理器集群46中的可编程着色器单元可至少包含顶点着色器单元、片元着色器单元,此外,还可以包含几何着色器单元、外壳着色器单元、域着色器单元、计算着色器单元和/或统一着色器单元。在具体实施过程中,可编程着色器单元可各自包含用于提取和解码操作的一或多个组件、用于进行算术计算的一或多个ALU、一或多个存储器、高速缓存和寄存器。

[0040] 并且,固定功能执行核26可包含经硬连线以执行某些功能的硬件。尽管固定功能硬件可经由例如一或多个控制信号而配置以执行不同功能,但所述固定功能硬件通常并不包含能够接收用户编译程序的程序存储器。在一些实例中,处理器集群46中的固定功能执行核26可包含例如执行图元装配的处理单元、执行光栅化操作的处理单元,像素后处理单元,所述像素后处理单元包括深度/模板测试、裁剪测试、α混合等。对于执行图元装配以及光栅化操作的处理单元来说,其能够将通过顶点着色器单元已完成着色的顶点按照原始连接关系还原出图形的网格结构,即图元,从而供后续片元着色器单元进行处理。

[0041] 一般来说,GPU 12从CPU 6所接收到的命令,其示例为,处理器6执行GPU驱动程序,以使得GPU驱动程序22可基于由软件应用程序18向GPU驱动程序22发出的指令产生定义用于由GPU 12执行的操作集合的命令流。所述命令流能够控制处理器集群46中可编程执行核24和固定功能执行核26(在不做区分的情况下,可统称之为“执行核”)的操作。

[0042] 如上所述,GPU 12可包含可从GPU驱动程序22接收命令流的命令处理器30。命令处理器30可以是被配置成接收并处理一或多个命令流的硬件与软件的任意组合。由此,命令处理器30可在本地控制GPU 资源而无需处理器6 的干预。举例来说,GPU 12的命令处理器30可从处理器6接收一个或一个以上“任务”。命令处理器30可独立地调度所述任务由GPU 12的计算资源(比如处理器集群46中的一或多个可编程执行核24和/或处理器集群46中的一或多个固定功能执行核26)执行。在一个实例中,命令处理器30可以是硬件处理器。在图2中所示出的实例中,命令处理器30可包含于GPU 12中。在其它实例中,命令处理器30可以是与CPU 6和GPU 12分离的单元。命令处理器30还可被称为流处理器、命令/流处理器及类似者,以指示其可以是被配置成接收命令和/或操作的流的任何处理器。

[0043] 命令处理器30可处理一或多个命令流,其包含调度操作,所述调度操作包含于由GPU 12执行的一或多个命令流中。具体地说,命令处理器30可处理一或多个命令流,且调度所述一或多个命令流中的操作,以由处理器集群46执行。在操作中,GPU驱动程序22可向命

令处理器30发送包括待由GPU 12执行的一系列操作的命令流。命令处理器30可接收包括命令流的操作流且可基于命令流中的操作次序依序地处理命令流的操作，且可调度命令流中的操作可以由处理器集群46中的一或多个执行核执行。

[0044] 详细来说，命令处理器30所接收到的任务通常会被划分优先级。在一些示例中，优先级可以由该任务所涉及的软件应用程序18是否为用户当前操作对象来确定，也就是说，作为当前用户操作对象的软件应用程序，其对应的任务优先级将会高于其他软件应用程序对应的任务，高优先级的任务也将会被命令处理器30优先下发至GPU 12的计算资源执行。但是，除了优先级方面的划分以外，任务还可以基于消耗计算资源的多寡按照复杂度进行区分，比如运行大型3D游戏或者播放高清视频均需要占用大量的计算资源；对于复杂度越高的任务，其所消耗的GPU 12的计算资源也就相应增加。而任务的优先级会随着用户的操作对象的改变而改变，但是任务的复杂度则几乎不会进行变化。因此，需要提供一种能够在任务优先级以及复杂度之间进行平衡的任务调度方案，以期能够降低GPU 12的计算资源占用率，节省GPU 12的功耗。

[0045] 结合图1、图2以上阐述，本发明实施例还在GPU 12内设置有如图3中所示的用于进行计算资源监控的资源监控器80，该资源监控器80可以连接命令处理器30以获知由命令处理器30下发且当前GPU 12的计算资源所执行的任务列表；此外，还能够连接处理器集群46以获知任务列表中各任务所占用的计算资源。在一些示例中，资源监控器80可以是硬件处理器。在图3中所示出的实例中，命令处理器30可作为与命令处理器30分离的单独个体包含于GPU 12中，也可以通过复用GPU 12中的各部件的资源形成逻辑个体。在其它实例中，资源监控器80也可以是与CPU 6和GPU 12分离的单元。

[0046] 基于上述图3所示的GPU结构，本发明实施例提供了一种动态的任务调度方案。在方案中，命令处理器30，经配置为将由GPU驱动程序22下发的待执行任务分配至处理器集群46中的执行核后，根据所述待执行任务的优先级以及当前任务参数更新资源监控器80中的资源列表；

[0047] 资源监控器80，经配置为根据所述资源列表中的待执行任务的当前任务参数估算在未来设定的预估时段内所述待执行任务的计算资源占用率；并持续监听所述待执行任务中高占用率任务的优先级；以及，若被监听任务的优先级下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别，向所述命令处理器30发出第一通知指示；

[0048] 所述命令处理器30，还经配置为基于所述第一通知指示向GPU驱动程序22请求下发用于降低所述被监听任务计算资源占用率的任务参数；以及，接收由GPU驱动程序22基于所述请求被批准而下发的用于降低所述被监听任务计算资源占用率的任务参数，并将所述用于降低所述被监听任务计算资源占用率的任务参数下发至所述被监听任务的执行核以按照所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

[0049] 需要说明的是，对于上述技术方案中，在一些示例中，任务参数可以包括用于度量或衡量执行该任务过程中所需消耗计算资源的参数，比如，图像的帧率、片元内的图元数量、渲染所需的顶点数目等。基于上述示例对任务参数的定义，就可以通过任务参数估算计算装置2在执行任务时的计算资源占用率，并以此判断任务为高占用率任务或者低占用率任务，所以，可以通过调整任务参数来调节任务执行过程中资源消耗。通常来说，高清视频

播放程序在观看高清视频的过程中，其占用的计算资源将明显大于运行社交软件所占用的计算资源。除了度量或衡量所需消耗的计算资源以外，任务参数还关系到任务的执行效果。举例来说，关于高清视频播放程序的任务在执行过程中，图像的帧率高，那么所消耗的计算资源也就越多，而视频也就播放的越清晰；如果降低图像的帧率，随之所消耗的计算资源也就减少，视频的播放效果也会变更模糊一些。

[0050] 在一些示例中，所述资源列表的每一项均表示一种对应关系，该对应关系用于表示待执行任务、优先级以及执行所述待执行任务所需的当前任务参数之间的对应关系；因此，每当命令处理器30接收到由GPU驱动程序22下发的待执行任务之后，均可以对现有的资源列表进行更新操作，以将最新接收到的待执行任务、优先级以及当前任务参数增加至资源列表中。

[0051] 在一些示例中，优先级下降的触发可以是基于事件的，该事件可以基于用户当前操作的目标应用程序进行变更，比如，计算装置2同时运行着高清视频播放程序以及社交软件，在用户在观看高清视频的过程中，当前操作的目标应用程序则为高清视频播放程序，社交软件则在后台继续执行；当用户切换至社交软件进行聊天时，当前操作的目标应用程序则由高清视频播放程序转变为社交软件，高清视频播放程序则转为后台执行，此时高清视频播放程序的优先级将被降低；或者，当高清视频播放程序与社交软件程序基于特定的工作模式，例如分屏模式下同时出现在屏幕上时，高清视频播放程序将不再占用整个显示区域，此时GPU驱动程序22需要同时向GPU 12发送这两个程序的任务，此时，如果用户在社交软件上进行操作，则高清视频播放程序任务优先级同样也会被降低。进一步来说，在切换前优先级较高的高清视频播放程序对应的任务的优先级基于用户的切换操作而降低，在切换前优先级较低的社交软件对应的任务的优先级基于用户的切换操作而提升。

[0052] 对于上述技术方案，当计算资源占用率较高的任务所对应的优先级下降并持续一段时间之后，可以通过更新该任务的任务参数以降低执行该任务所需消耗的计算资源，从而降低GPU 12的计算资源占用率，节省GPU 12的功耗。

[0053] 对于前述技术方案，在一些可能的实现方式中，所述资源监控器80，经配置为：相应于所述待执行任务的计算资源占用率高于设定的阈值，确定所述待执行任务为高占用率任务，并持续监听高占用率任务的优先级。针对本实现方式，具体来说，资源监控器80还可以维持一张优先级监听列表，该列表中的每一项均表示高占用率任务及优先级之间的对应关系。在该列表中的任务均可以被认为是高占用率任务，持续的监听这些高占用率任务的优先级，从而能够及时对优先级下降的高占用率任务调整其对应的任务参数，从而降低该任务的计算资源占用率。

[0054] 对于前述技术方案，在一些实现方式中，所述GPU驱动程序22，经配置为对接收到的请求进行仲裁；若仲裁结果为所述请求被批准，则向所述命令处理器30下发用于降低所述被监听任务计算资源占用率的任务参数；否则，向所述命令处理器30反馈拒绝指令，以拒绝下发用于降低所述被监听任务计算资源占用率的任务参数。可以理解地，某些任务尽管优先级并不高，但在一些特定的条件下仍然无法降低其运行效果，对于这类任务，尽管GPU驱动程序22收到了命令处理器30发送的请求，但为了必须保证其运行效果，仍旧不能批准该请求。

[0055] 对于前述技术方案，在一些可能的实现方式中，所述资源监控器80，还经配置为：

若所述被监听任务的优先级由较低级别转至高级别且所述被监听任务的优先级在设定的监听时段内持续处于高级别，向所述命令处理器30发出第二通知指示；

[0056] 所述命令处理器30，还经配置为基于所述第二通知指示向GPU驱动程序22请求下发所述被监听任务的当前任务参数；以及，接收由GPU驱动程序22基于所述请求被批准而下发的所述被监听任务的当前任务参数，并将所述被监听任务的当前任务参数下发至所述被监听任务的执行核以重新按照所述被监听任务的当前任务参数执行所述被监听任务。

[0057] 在上述实现方式中，具体来说，尽管被监听的任务的计算资源占用率较高，但是，如果被监听的任务由低优先级转至高优先级，那么此时的关键应当由资源占用转为运行效果，因此，对于重新恢复至高优先级的任务，就需要恢复按照该任务的当前任务参数进行执行以保证执行效果。

[0058] 对于前述技术方案，在一些可能的实现方式中，所述资源监控器80，还经配置为：若所述被监听任务根据所述用于降低所述被监听任务计算资源占用率的任务参数仍被确定为高占用率任务，继续向所述命令处理器30发出第一通知指示；

[0059] 所述命令处理器30，还经配置为基于所述第一通知指示继续向GPU驱动程序22请求下发用于继续降低所述被监听任务计算资源占用率的任务参数；以及，接收由GPU驱动程序22基于所述请求被批准而下发的用于继续降低所述被监听任务计算资源占用率的任务参数，并将所述用于继续降低所述被监听任务计算资源占用率的任务参数下发至所述被监听任务的执行核以按照所述用于继续降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

[0060] 对于上述实现方式，需要说明的是，当对任务参数进行一次调整后仍然无法将其对应任务的计算资源占有率下降至正常的范围，那么可以继续通知命令处理器30向GPU驱动程序22发送请求，直至任务的计算资源占有率下降至正常的范围或者GPU驱动程序22拒绝请求。

[0061] 基于前述技术方案相同的发明构思，参见图4，其示出了本发明实施例提供的一种动态的任务调度方法，该方法可以适用于图3所示的GPU结构，该方法可以包括：

[0062] S401：根据用于执行被下发任务所需的当前任务参数确定所述被下发任务在未来设定的预估时段的计算资源占用率；

[0063] S402：持续监听高占用率任务的优先级；

[0064] S403：若被监听任务的优先级下降至较低级别且所述被监听任务的优先级在设定的监听时段内持续处于较低级别，请求CPU端下发用于降低所述被监听任务计算资源占用率的任务参数；

[0065] S404：若所述请求被批准，接收由CPU端下发的用于降低所述被监听任务计算资源占用率的任务参数，并根据所述用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务。

[0066] 对于图4所示的技术方案，在一些可能的实现方式中，如图5所示，所述持续监听高占用率任务的优先级，包括：

[0067] S51：判断所述待执行任务的计算资源占用率是否大于设定的阈值：

[0068] S52：若大于，则确定所述待执行任务为高占用率任务；

[0069] S53：否则确定所述待执行任务为低占用率任务。

[0070] 对于图4所示的技术方案，在一些可能的实现方式中，所述方法还包括：

[0071] 通过所述CPU端的GPU驱动程序对接收到的请求进行仲裁；

[0072] 若仲裁结果为所述请求被批准，则下发用于降低所述被监听任务计算资源占用率的任务参数；否则，反馈拒绝指令，以拒绝下发用于降低所述被监听任务计算资源占用率的任务参数。

[0073] 对于图4所示的技术方案，在一些可能的实现方式中，所述方法还包括：若所述被监听任务的优先级由较低级别转至高级别且所述被监听任务的优先级在设定的监听时段内持续处于高级别，根据所述被监听任务的当前任务参数执行所述被监听任务。

[0074] 对于图4所示的技术方案，在一些可能的实现方式中，所述方法还包括：

[0075] 若所述被监听任务根据所述用于降低所述被监听任务计算资源占用率的任务参数仍被确定为高占用率任务，继续请求CPU端下发用于继续降低所述被监听任务计算资源占用率的任务参数，直至所述被监听任务的计算资源占用率下降至设定的范围。

[0076] 可以理解地，上述动态的任务调度方法的示例性技术方案，与前述基于图3所示的GPU结构所提供的动态的任务调度方案属于同一构思，因此，上述动态的任务调度方法的示例性技术方案中执行各步骤的主体以及未详细描述的细节内容，均可以参见前述基于图3所示的GPU结构所提供的动态的任务调度方案的描述。本发明实施例对此不做赘述。

[0077] 在上述一或多个实例或示例中，所描述的功能可实施于，所描述功能可实施于硬件、软件、固件或其任何组合中。如果实施于软件中，那么可将功能作为一或多个指令或代码存储在计算机可读媒体上或经由计算机可读媒体传输。计算机可读媒体可包含计算机数据存储媒体或通信媒体，通信媒体包含促进将计算机程序从一处传递到另一处的任何媒体。数据存储媒体可为可由一或多个计算机或一或多个处理器存取以检索用于实施本发明中描述的技术的指令、代码和/或数据结构的任何可用媒体。举例来说且非限制，此类计算机可读媒体可包括U盘、移动硬盘、RAM、ROM、EEPROM、CD-ROM或其它光盘存储装置、磁盘存储装置或其它磁性存储装置，或可用于运载或存储呈指令或数据结构的形式的所要程序代码且可由计算机存取的任何其它媒体。并且，任何连接被恰当地称作计算机可读媒体。举例来说，如果使用同轴电缆、光纤电缆、双绞线、数字订户线(DSL)或例如红外线、无线电和微波等无线技术从网站、服务器或其它远程源传输软件，那么同轴电缆、光纤电缆、双绞线、DSL或例如红外线、无线电和微波等无线技术包含于媒体的定义中。如本文中所使用，磁盘和光盘包含压缩光盘(CD)、激光光盘、光学光盘、数字多功能光盘(DVD)、软性磁盘和蓝光光盘，其中磁盘通常以磁性方式再现数据，而光盘利用激光以光学方式再现数据。以上各项的组合也应包含在计算机可读媒体的范围内。

[0078] 代码可由一或多个处理器执行，所述一或多个处理器例如是一或多个数字信号处理器(DSP)、通用微处理器、专用集成电路(ASIC)、现场可编程逻辑阵列(FPGA)或其它等效的可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。因此，如本文中所使用的术语“处理器”和“处理单元”可指前述结构或适于实施本文中所描述的技术的任何其它结构中的任一者。另外，在一些方面中，本文中所描述的功能性可在经配置用于编码和解码的专用硬件和/或软件模块内提供，或者并入在组合式编解码器中。而且，所述技术可完全实施于一或多个电路或逻辑元件中。

[0079] 本发明实施例的技术可实施于各种各样的装置或设备中，所述装置或设备包含无

线手持机、集成电路(IC)或一组IC(即,芯片组)。本发明中描述各种组件、模块或单元是为了强调经配置以执行所公开的技术的装置的功能方面,但未必需要由不同硬件单元实现。实际上,如上文所描述,各种单元可结合合适的软件和/或固件组合在编码解码器硬件单元中,或者通过互操作硬件单元的集合来提供,所述硬件单元包含如上文所描述的一或多个处理器。

[0080] 已描述了本发明的各种方面。这些和其它实施例在所附权利要求书的范围内。需要说明的是:本发明实施例所记载的技术方案之间,在不冲突的情况下,可以任意组合。

[0081] 以上所述,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应以所述权利要求的保护范围为准。

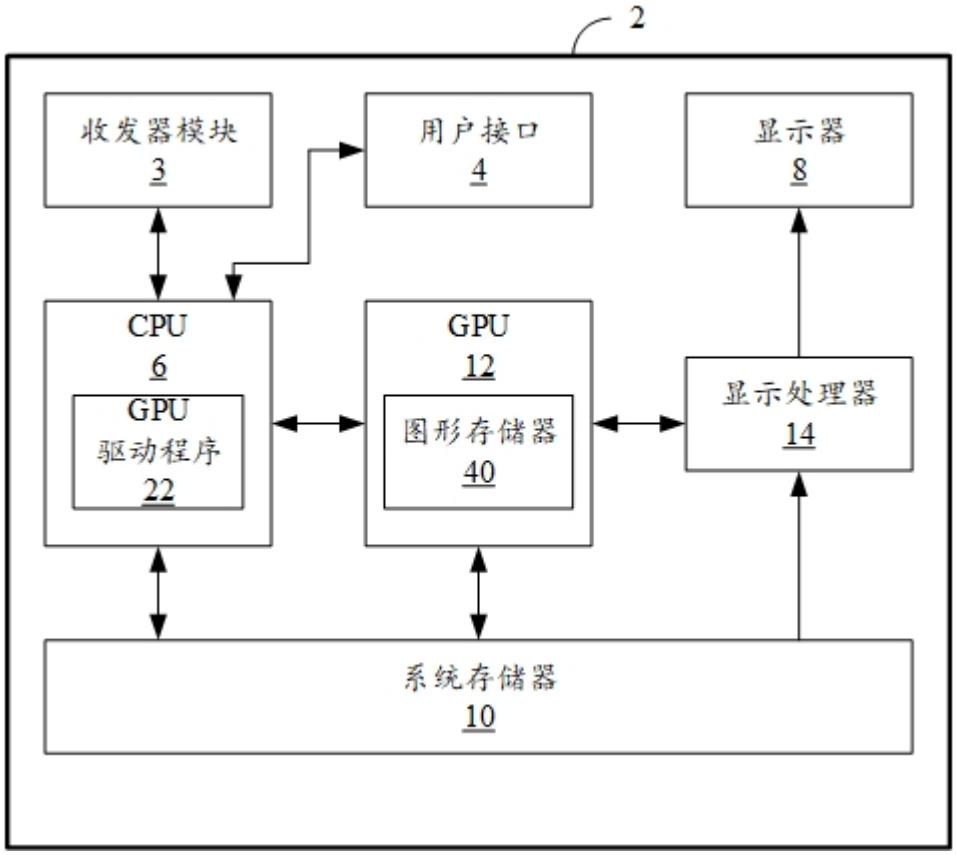

2
收发器模块
3
用户接口
4
显示器
8
CPU
6
GPU
驱动程序
22
GPU
12
图形存储器
40
显示处理器
14
系统存储器
10

图1

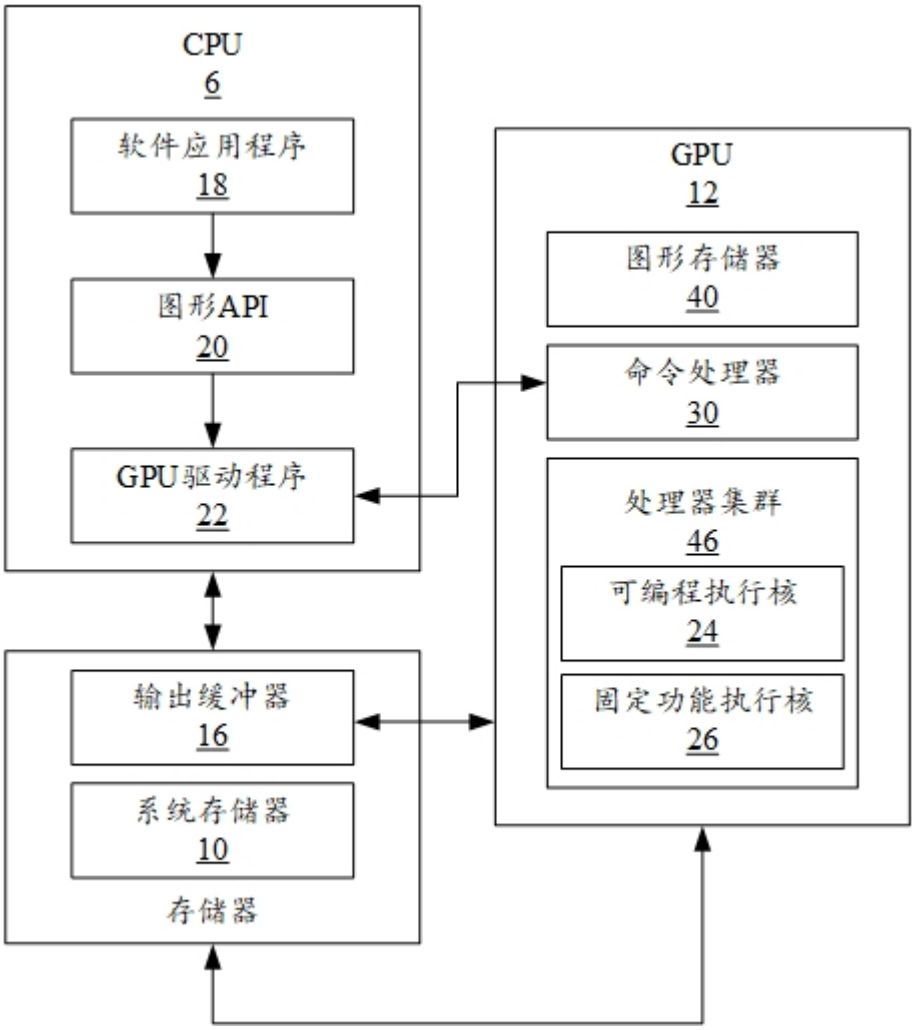

CPU
6
软件应用程序
18
图形API
20
GPU驱动程序
22
输出缓冲器
16
系统存储器
10
存储器
GPU
12
图形存储器
40
命令处理器
30
处理器集群
46
可编程执行核
24
固定功能执行核
26

图2

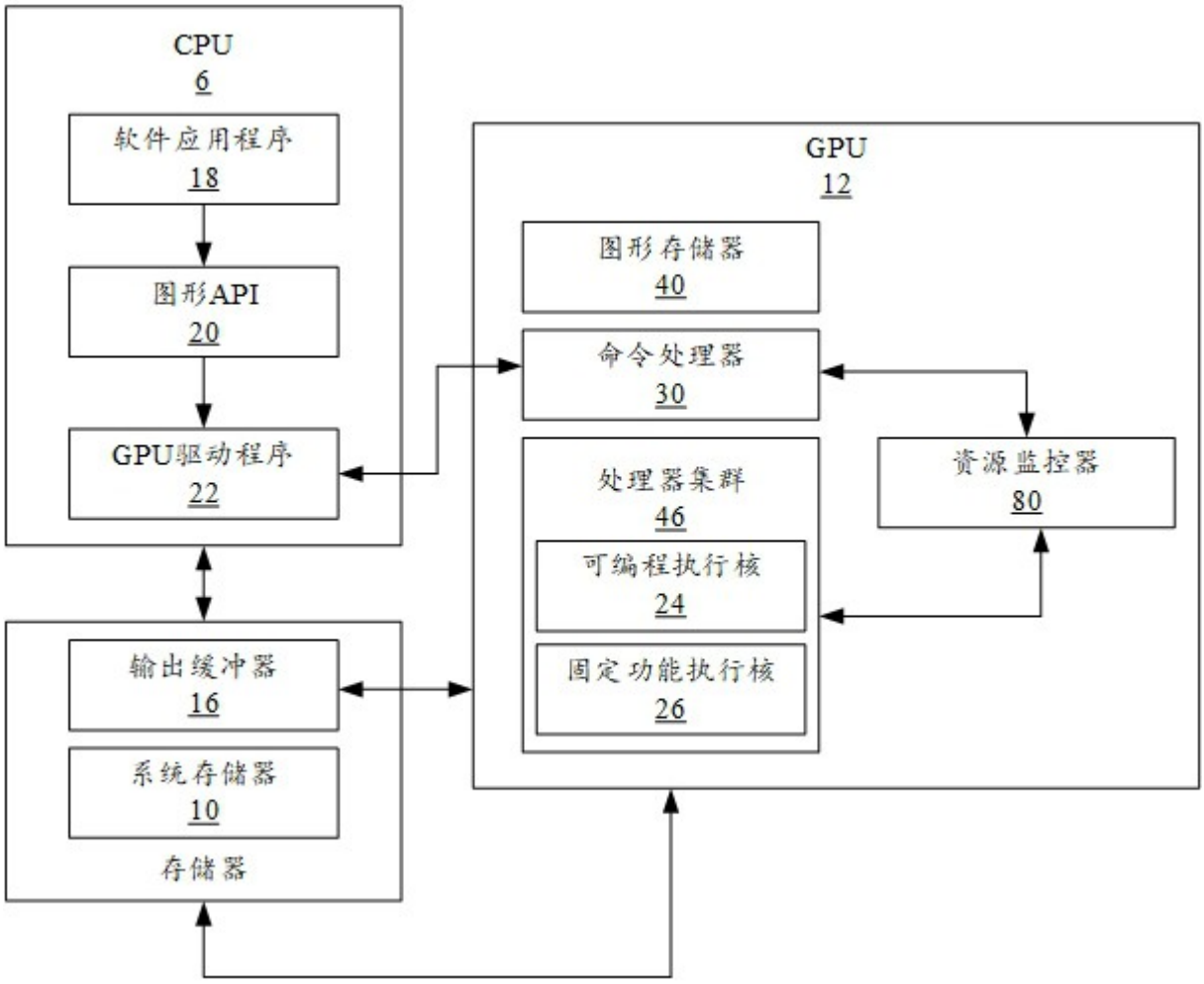

CPU
6
软件应用程序
18
图形API
20
GPU驱动程序
22
输出缓冲器
16
系统存储器
10
存储器
GPU
12
图形存储器
40
命令处理器
30
处理器集群
46
可编程执行核
24
固定功能执行核
26
资源监控器
80

图3

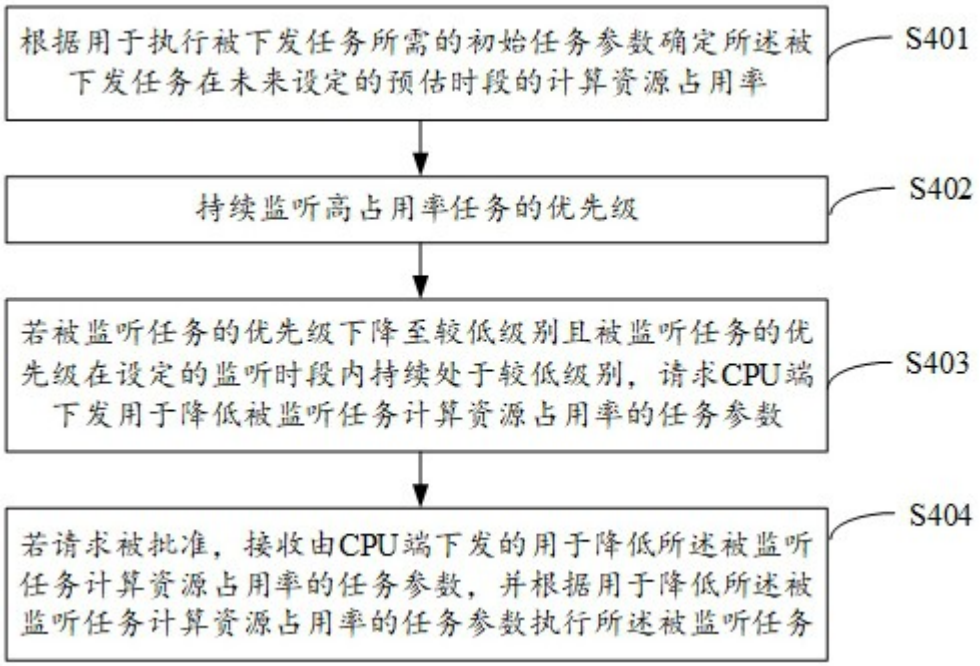

根据用于执行被下发任务所需的初始任务参数确定所述被下发任务在未来设定的预估时段的计算资源占用率
S401
持续监听高占用率任务的优先级
S402
若被监听任务的优先级下降至较低级别且被监听任务的优先级在设定的监听时段内持续处于较低级别，请求CPU端下发用于降低被监听任务计算资源占用率的任务参数
S403
若请求被批准，接收由CPU端下发的用于降低所述被监听任务计算资源占用率的任务参数，并根据用于降低所述被监听任务计算资源占用率的任务参数执行所述被监听任务
S404

图4

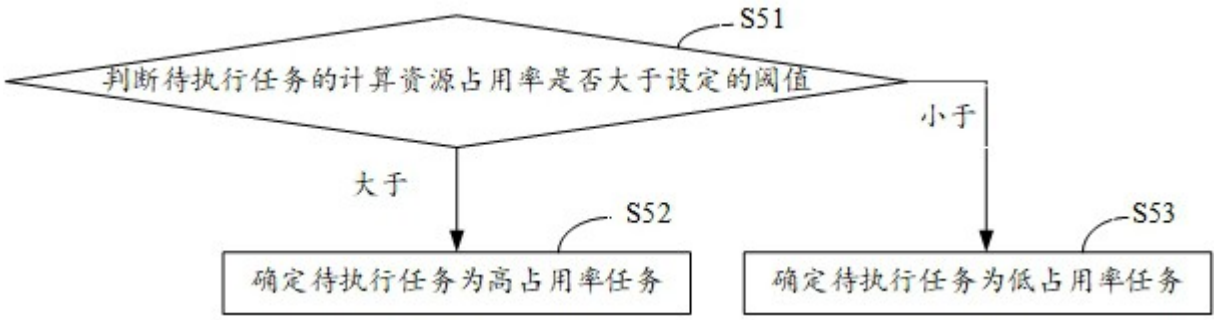

S51
判断待执行任务的计算资源占用率是否大于设定的阈值
大于
小于
S52
确定待执行任务为高占用率任务
S53
确定待执行任务为低占用率任务

图5