

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 12 月 27 日 (27.12.2018)



(10) 国际公布号
WO 2018/233299 A1

- (51) 国际专利分类号:
G06F 9/50 (2006.01)
- (21) 国际申请号: PCT/CN2018/076009
- (22) 国际申请日: 2018 年 2 月 9 日 (09.02.2018)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201710481998.5 2017年6月22日 (22.06.2017) CN
- (71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区八卦岭八卦三路平安大厦6楼, Guangdong 518000 (CN)。
- (72) 发明人: 王健宗(WANG, Jianzong); 中国广东省深圳市福田区八卦岭八卦三路平安大厦6楼, Guangdong 518000 (CN)。张懿(ZHANG, Yi); 中

国广东省深圳市福田区八卦岭八卦三路平安大厦6楼, Guangdong 518000 (CN)。郭卉(GUO, Hui); 中国广东省深圳市福田区八卦岭八卦三路平安大厦6楼, Guangdong 518000 (CN)。肖京(XIAO, Jing); 中国广东省深圳市福田区八卦岭八卦三路平安大厦6楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳中一专利商标事务所(SHENZHEN ZHONGYI PATENT AND TRADEMARK OFFICE); 中国广东省深圳市福田区深南中路1014号老特区报社四楼(5号信箱), Guangdong 518028 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: METHOD, APPARATUS AND DEVICE FOR SCHEDULING PROCESSOR, AND MEDIUM

(54) 发明名称: 调度处理器的方法、装置、设备及介质

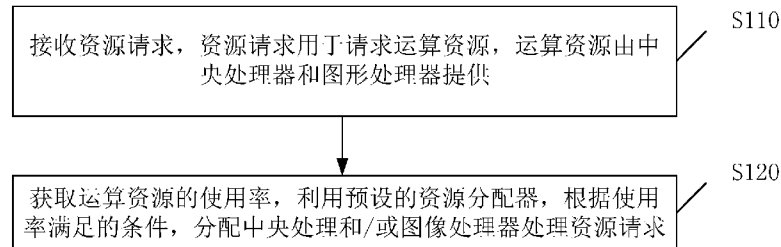


图 1

- S110 Receive resource requests, wherein the resource requests are used to request operation resources, and the operation resources are provided by a central processing unit and a graphics processing unit
- S120 Acquire the usage rate of the operation resources, and allocate, by using a pre-set resource allocator and according to the condition that the usage rate is satisfied, the resource requests processed by the central processing unit and/or the graphics processing unit

(57) Abstract: A method and apparatus for scheduling a processor, a device for scheduling a processor, and a medium, wherein same are applicable to the field of software. The method comprises: receiving resource requests, wherein the resource requests are used to request operation resources, and the operation resources are provided by a central processing unit and a graphics processing unit (S110); and acquiring the usage rate of the operation resources, and allocating, by using a pre-set resource allocator and according to the condition that the usage rate is satisfied, the resource requests processed by the central processing unit and/or the graphics processing unit (S120). The method and apparatus allocate resource requests by means of the usage rate of a CPU and a GPU. Compared with the fixed division of labor in the prior art, the existing operation resources can be more fully utilized to improve the processing speed, and furthermore, the GPU shares the operational burden of the CPU, such that the stable running of a system is ensured.

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 一种调度处理器的方法、装置、调度处理器的设备及介质, 适用于软件领域, 该方法包括: 接收资源请求, 资源请求用于请求运算资源, 运算资源由中央处理器和图形处理器提供 (S110); 获取运算资源的使用率, 利用预设的资源分配器, 根据使用率满足的条件, 分配中央处理和/或图像处理器处理资源请求 (S120)。该方法及装置通过CPU和GPU的使用率进行资源请求的分配, 相比于现有技术中各自的固定分工, 可以更充分地利用已有的运算资源, 提高处理速度, 进一步地, 通过GPU分担CPU的运算压力, 保证了系统的稳定运转。

调度处理器的方法、装置、设备及介质

[0001] 本申请要求于2017年06月22日提交中国专利局、申请号为201710481998.5、发明名称为“调度处理器的方法及设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请属于软件领域，尤其涉及一种调度处理器的方法、装置、设备及介质。

背景技术

[0003] 显卡的处理器称为图形处理器（Graphics Processing Unit，简称为GPU），其与中央处理器（Central Processing Unit，简称为CPU）类似，只不过GPU最初是专为执行复杂的数学和几何计算而设计的。

[0004] 随着计算机技术的发展，图形处理器得到了越来越充分的利用，通过GPU完成部分CPU的工作，可以减少计算机对CPU的依赖。例如，在以前的很多应用中，CPU是负责所有运算的，而GPU则只是负责最后的显示工作，因此一旦出现处理复杂图形数据的时候，很多使用性能较弱的CPU的电脑系统就开始缓慢无比，而使用性能较强的CPU的电脑系统也会出现CPU资源被大量占用的情况。而在GPU开始参与运算之后，原本会消耗CPU大量宝贵资源的图形数据处理部分就全部交给GPU进行处理，从而降低了CPU的负担，并且利用自身的特长，使得图形数据处理的效率更快，从而提升系统性能。

[0005] 现有技术对CPU和GPU进行了越来越精细的分工，即依照预设规则，将特定的任务分配给GPU处理，以分担CPU的压力。但是上述预设规则往往滞后于实际使用场景，因此当出现新的使用场景时，GPU往往无法及时分担CPU的压力。

[0006] 针对现有技术中GPU无法灵活地分担CPU压力的问题，目前业界没有理想的解决方式。

发明概述

技术问题

[0007] 有鉴于此，本申请实施例提供了一种调度处理器的方法、装置、设备及介质，

以解决现有技术中GPU无法灵活地分担CPU压力的问题。

问题的解决方案

技术解决方案

- [0008] 本申请实施例的第一方面提供了一种调度处理器的方法，包括：
- [0009] 接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供；
- [0010] 获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。
- [0011] 本申请实施例的第二方面提供了一种调度处理器的装置，包括：
- [0012] 接收模块，用于接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供。
- [0013] 获取模块，用于获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。
- [0014] 本申请实施例的第三方面提供了一种调度处理器的设备，包括存储器以及处理器，所述存储器中存储有可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现如下步骤：
- [0015] 接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供；
- [0016] 获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。
- [0017] 本申请实施例的第四方面提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如下步骤：
- [0018] 接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供；
- [0019] 获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。

发明的有益效果

有益效果

[0020] 本申请实施例灵活地通过CPU和GPU的使用率进行资源请求的分配，相比于现有技术中各自的固定分工，可以更充分地利用已有的运算资源，提高处理速度，进一步地，通过GPU分担CPU的运算压力，保证了系统的稳定运转。

对附图的简要说明

附图说明

[0021] 为了更清楚地说明本申请实施例中的技术方案，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申请的一些实施例，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。

[0022] 图1是本申请实施例提供的调度处理器的方法的流程图；

[0023] 图2是本申请实施例提供的分配资源请求方法的流程图；

[0024] 图3是本申请优选实施例提供的一种分配资源请求方法的流程图；

[0025] 图4是图3所示优选实施例的示意图；

[0026] 图5是本申请优选实施例提供的另一种分配资源请求方法的流程图；

[0027] 图6是图5所示优选实施例的示意图；

[0028] 图7是本申请实施例提供的调度处理器的装置的结构框图；

[0029] 图8是本申请实施例提供的调度处理器的设备的示意图。

发明实施例

本发明的实施方式

[0030] 以下描述中，为了说明而不是为了限定，提出了诸如特定系统结构、技术之类的具体细节，以便透彻理解本申请实施例。然而，本领域的技术人员应当清楚，在没有这些具体细节的其它实施例中也可以实现本申请。在其它情况中，省略对众所周知的系统、装置、电路以及方法的详细说明，以免不必要的细节妨碍本申请的描述。

[0031] 为了说明本申请所述的技术方案，下面通过具体实施例来进行说明。

[0032] 图1是本申请实施例提供的调度处理器的方法的流程图，如图1所示，该方法包括步骤S110和步骤S120。

- [0033] 步骤S110, 接收资源请求, 资源请求用于请求运算资源, 运算资源由中央处理器和图形处理器提供。
- [0034] 资源请求可以是HTTP请求(超文本传输协议HyperText Transfer Protocol简称为HTTP), 即是从客户端到服务器端的请求消息, 本申请实施例中HTTP请求可以指客户端需要运算资源时向服务器发出的请求。
- [0035] 运算资源是具有运算能力的硬件资源, 在本实施例中, 运算资源可以是算术逻辑单元(Arithmetic Logic Unit, 简称为ALU)。CPU和GPU通过ALU完成各种运算操作。
- [0036] 步骤S120, 获取运算资源的使用率, 利用预设的资源分配器, 根据使用率满足的条件, 分配中央处理和/或图像处理器处理资源请求。
- [0037] 运算资源的使用率也就是运算资源的使用比例, 一般用百分数表示, 例如当前运算资源的使用率为50%。
- [0038] 一般而言, 在一台设备中, 由于硬件的空间和成本的限制, 所以CPU和GPU中的ALU数量也是有限的, 因此在资源请求数量过多时, CPU和GPU将承受较大的运算压力。
- [0039] 具体地, CPU数量是远小于GPU数量的, CPU一般结构较复杂, 可以用来进行复杂运算, 因此既可以处理简单运算, 也可以负责维持整个系统的运转; GPU结构简单, 一般只能用于进行简单运算, 多个GPU可以并行使用。
- [0040] CPU和GPU之所以大不相同, 是由于其设计目标的不同, 它们分别针对了两种不同的应用场景。CPU需要很强的通用性来处理各种不同的数据类型, 同时又要逻辑判断, 会引入大量的分支跳转和中断的处理。这些都使得CPU的内部结构异常复杂。而GPU面对的则是类型高度统一的、相互无依赖的大规模数据和不需要被打断的纯净的计算环境。
- [0041] 资源请求一般请求的是简单运算, 如果资源请求数量过多, 且没有对资源请求进行合理分配, 导致占用较大CPU的运算资源时, 整个系统的运转都会相应地受到影响。
- [0042] GPU采用了数量众多的计算单元和超长的流水线, 但只有非常简单的控制逻辑并省去了Cache。与CPU擅长逻辑控制和通用类型数据运算不同, GPU擅长的是

大规模并发计算，因此对于例如声纹特征提取的需求，GPU的计算速度远优于CPU。

[0043] 在一种优选实现方式中，为了避免CPU压力过大，同时为了更充分地利用GPU的运算资源，当GPU使用率不高时，可以优先使用GPU处理资源请求。图2是本申请实施例提供的分配资源请求方法的流程图，该方法是步骤S110的一种优选实现方式，如图2所示，该方法包括步骤S210和步骤S220。

[0044] 步骤S210，获取图形处理器使用率。

[0045] 通过java、C语言和Linux等均有对应的获取CPU和GPU使用率的方式，例如调用对应的应用程序编程接口（Application Programming Interface, 简称为API）等。

[0046] 步骤S220，若图形处理器使用率低于预设值，则优先分配图形处理器处理资源请求。

[0047] 由于GPU的数量更多，且CPU在系统整体运转中起到的作用比GPU更大，因此，为了分担CPU的压力，让CPU更多地投入到系统整体运转中，可以在分配资源请求时优先考虑分配给GPU处理。例如，当GPU的使用率低于预设值的时候，将资源请求分配给GPU，预设值可以根据不同的情况来进行相应设置，例如设为70%-90%。

[0048] 分配资源请求这一操作可以通过负载均衡策略完成，作为一种优选的实现方式，可以利用Nginx负载均衡服务器分配中央处理器和/或图像处理器处理资源请求。

[0049] Nginx既可以在内部直接支持Rails框架和PHP程序对外进行服务，也可以支持作为HTTP代理服务器对外进行服务。Nginx可以采用C语言进行编写，系统资源开销比Perlbal要好很多。

[0050] 本申请实施例根据GPU的使用率进行资源请求的分配，相比于现有技术中各自的固定分工，可以更充分地利用已有的运算资源，提高处理速度，进一步地，通过GPU分担CPU的运算压力，保证了系统的稳定运转。

[0051] 图3是本申请优选实施例提供的一种分配资源请求方法的流程图，如图3所示，该方法包括步骤S310至步骤S330。

- [0052] 步骤S310, 部署第一服务器和第二服务器, 第一服务器用于接收分配给图形处理器的资源请求, 第二服务器用于接收分配给中央处理器的资源请求。
- [0053] 本申请实施例中, 可以在硬件设备上部署多个服务器, 在一个服务器的资源被用完之后, 再运行下一个服务器, 这样可以更加合理地使用运算资源。本步骤中部署的第一服务器和第二服务器可以是2个虚拟服务器 (server)。
- [0054] 虚拟服务器是节省服务器硬件成本的一种技术, 可以将一台服务器硬件的某项或者全部服务内容逻辑划分为多个服务单位, 对外表现为多个服务器, 从而可以充分利用服务器硬件资源。
- [0055] 当单个服务器无法满足使用需求, 或者单个服务器能够达到这个要求的成本太高时, 可以使用虚拟服务器来达到运用需求。
- [0056] 多个虚拟服务器, 可以通过软件架设在同一个服务器硬件上, 而每个虚拟服务器又可以提供服务器的所有功能, 因此降低了运行成本。
- [0057] 具体地, 虚拟服务器可以是tomcat、jboss或weblogic等服务器程序。第一服务器对应GPU, 第二服务器对应CPU, 用Nginx作为负载均衡服务器为每个server按配置分配请求。
- [0058] 步骤S320, 设置第一服务器的权重高于第二服务器。
- [0059] 即GPU版本的server设置较高的权重, 从而可以优先给该server分配资源请求。
- [0060] 步骤S330, 在接收到资源请求之后, 根据权重和使用率进行资源请求分配。
- [0061] 本申请实施例可以用各服务器的权重乘以对应处理器的剩余资源的占比, 将资源请求分配给数值较大的服务器。例如第一服务器的权重是0.7, GPU的使用率30%, 第二服务器的权重是0.3, CPU的使用率是25%, 那么第一服务器中权重与剩余资源的乘积为 $0.7 * (1 - 30\%) = 0.49$, 第二服务器中权重与剩余资源的乘积为 $0.3 * (1 - 25\%) = 0.225$, 由于第一服务器中权重与剩余资源的乘积更大, 因此此时将资源请求分配给第一服务器, 即分配给GPU。
- [0062] 进一步地, 在分配资源请求时, 除了考虑处理器使用率之外, 还可以考虑线程池使用率, 通过上述这两个使用率最终决定调度方案。
- [0063] 在分配资源请求之后, 如果分配给CPU, 则采用切片轮询的方式进行任务执行方式, 使CPU的利用率达到最大; 如果分配给GPU, 则采用了单颗独占的会话保

持方式，即，在任务结束之前，GPU会被当前用户独占。这是因为GPU核心线程需要由CPU线程进行调度，若不采用独占方式，就会造成任务冲突，降低系统效率。

[0064] 图4是图3所示优选实施例的示意图，如图4所示，在收到HTTP请求以后，Nginx进行事件分发，将HTTP请求分配给GPU对应的server或CPU对应的server，server通过init.py进行初始化模块，其中CPU对应的server可以直接初始化统一业务模型（UBM模型）；而GPU对应的server需要先启动spring容器，加载指定GPU之后，再初始化UBM模型。完成初始化后，GPU对应的server通过get/set使用GPU的UBM池（线程池）中的资源；CPU通过get使用CPU的UBM池中的资源。每个server同一时间只能处理一定数量maxprocess(最大连接请求数)的请求，分别等于GPU和CPU的个数，资源请求如果超出maxprocess，则会放入Nginx请求队列重新分配。

[0065] 本优选实施例把分配权交给Nginx，核心负载较小。

[0066] 图5是本申请优选实施例提供的另一种分配资源请求的流程图，如图5所示，该方法包括步骤S510和步骤S520。

[0067] 步骤S510，部署多个服务器，服务器对应预设数量的图形处理器和中央处理器，若图形处理器使用率低于预设值，则优先分配图形处理器处理资源请求，设置每个服务器的最大运算资源值，最大运算资源值等于该服务器中中央处理器的运算资源与图形处理器的运算资源之和。

[0068] 本优选实施例中，不是对GPU和CPU分开管理，而是每个server都包括一定的CPU和GPU运算资源，在一个server达到一定负荷之后，再使用下一个server。

[0069] 步骤S520，按预设顺序使用服务器，在给当前服务器分配的资源请求大于最大运算资源值时，将超出最大运算资源值的资源请求分配至下一个服务器。

[0070] 预设顺序可以是对多个server进行编号，例如server1，server2，server3...当资源请求大于最大运算资源值，即maxprocess时，将资源请求分配给下一个server。

[0071] 图6是图5所示优选实施例的示意图，如图5所示，在收到HTTP请求以后，nginx进行事件分发，将HTTP请求分配server，server通过init.py进行初始化模块，具体地，先启动spring容器，加载指定GPU之后，再初始化UBM模型，然后初始化

CPU的UBM，并放入spring容器中。在每个server中，通过get/set使用GPU的UBM池中的资源，判断是否获取到GPU的运算资源，如果获取到GPU的运算资源，则返回该资源；如果没有获取到GPU的运算资源，在获取CPU运算资源。如果资源请求大于maxprocess，则将资源请求放入nginx的请求队列，在下一个server中进行分配。

[0072] 本优选实施例将GPU/CPU的选择权交给了Server，GPU利用充分，并且代码维护简单，部署时不同的server只需要简单的配置GPU参数即可。

[0073] 本申请实施例还提供了一种调度处理器的装置，图7是本申请实施例提供的调度处理器的装置的结构框图，如图7所示，该装置包括接收模块710和获取模块720。

[0074] 接收模块710用于接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供。

[0075] 获取模块720用于获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。

[0076] 优选地，获取模块包括：获取子模块，用于获取图形处理器使用率；分配子模块，用于在图形处理器使用率低于预设值时，优先分配图形处理器处理资源请求。

[0077] 优选地，获取模块还用于：利用Nginx负载均衡服务器分配中央处理器和/或图像处理器处理资源请求。

[0078] 优选地，获取模块包括：第一部署子模块，用于部署第一服务器和第二服务器，第一服务器用于接收分配给图形处理器的资源请求，第二服务器用于接收分配给中央处理器的资源请求；第一设置子模块，用于设置第一服务器的权重高于第二服务器；发送子模块，用于在接收到资源请求之后，根据权重和使用率进行资源请求分配。

[0079] 优选地，获取模块包括：第二部署子模块，用于部署多个服务器，服务器对应预设数量的图形处理器和中央处理器，若图形处理器使用率低于预设值，则优先分配图形处理器处理资源请求，第二设置子模块，用于设置每个服务器的最大运算资源值，最大运算资源值等于该服务器中中央处理器的运算资源与图形

处理器的运算资源之和；使用子模块，用于按预设顺序使用多个服务器，在给当前服务器分配的资源请求大于该服务器的最大运算资源值时，将超出最大运算资源值的资源请求分配至下一个服务器。

[0080] 本申请实施例灵活地通过CPU和GPU的使用率进行资源请求的分配，相比于现有技术中各自的固定分工，可以更充分地利用已有的运算资源，提高处理速度，进一步地，通过GPU分担CPU的运算压力，保证了系统的稳定运转。

[0081] 图8是本申请一实施例提供的调度处理器的设备的示意图。如图8所示，该实施例的调度处理器的设备8包括：处理器80以及存储器81，所述存储器81中存储有可在所述处理器80上运行的计算机可读指令82，例如调度处理器的程序。所述处理器80执行所述计算机可读指令82时实现上述各个调度处理器的方法实施例中的步骤，例如图1所示的步骤101至102。或者，所述处理器80执行所述计算机可读指令82时实现上述各装置实施例中各模块/单元的功能，例如图5所示模块710至720的功能。

[0082] 示例性的，所述计算机可读指令82可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器81中，并由所述处理器80执行，以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令82在所述调度处理器的设备8中的执行过程。

[0083] 所述调度处理器的设备8可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述调度处理器的设备可包括，但不仅限于，处理器80、存储器81。本领域技术人员可以理解，图8仅仅是调度处理器的设备8的示例，并不构成对调度处理器的设备8的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如所述调度处理器的设备还可以包括输入输出设备、网络接入设备、总线等。

[0084] 所称处理器80可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其他可编程逻辑器件、分立门或者

晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0085] 所述存储器81可以是所述调度处理器的设备8的内部存储单元，例如调度处理器的设备8的硬盘或内存。所述存储器81也可以是所述调度处理器的设备8的外部存储设备，例如所述调度处理器的设备8上配备的插接式硬盘，智能存储卡（Smart Media Card，SMC），安全数字（Secure Digital，SD）卡，闪存卡（Flash Card）等。进一步地，所述存储器81还可以既包括所述调度处理器的设备8的内部存储单元也包括外部存储设备。所述存储器81用于存储所述计算机可读指令以及所述调度处理器的设备8所需的其他程序和数据。所述存储器81还可以用于暂时地存储已经输出或者将要输出的数据。

[0086] 所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。实施例中的各功能单元、模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中，上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。另外，各功能单元、模块的具体名称也只是为了便于相互区分，并不用于限制本申请的保护范围。上述系统中单元、模块的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0087] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述或记载的部分，可以参见其它实施例的相关描述。

[0088] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0089] 在本申请所提供的实施例中，应该理解到，所揭露的装置/终端设备和方法，

可以通过其它的方式实现。例如，以上所描述的装置/终端设备实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。

[0090] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0091] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0092] 所述集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，也可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一计算机可读存储介质中，该计算机可读指令在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机可读指令包括计算机可读指令代码，所述计算机可读指令代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读存储介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（Read-Only Memory，ROM）、随机存取存储器（Random Access Memory，RAM）、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读存储介质包含的内容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读存储介质不包括电载波信号和电信信号。

[0093] 以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权利要求书

- [权利要求 1] 一种调度处理器的方法，其特征在于，包括：
接收资源请求，所述资源请求用于请求运算资源，所述运算资源由中央处理器和图形处理器提供；
获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求。
- [权利要求 2] 如权利要求1所述的调度处理器的方法，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：
获取图形处理器使用率；
若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求。
- [权利要求 3] 如权利要求1所述的调度处理器的方法，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：
利用Nginx负载均衡服务器分配所述中央处理器和/或所述图像处理器处理所述资源请求。
- [权利要求 4] 如权利要求1所述的调度处理器的方法，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：
部署第一服务器和第二服务器，所述第一服务器用于接收分配给所述图形处理器的所述资源请求，所述第二服务器用于接收分配给所述中央处理器的所述资源请求；
设置所述第一服务器的权重高于所述第二服务器；
在接收到所述资源请求之后，根据所述权重和所述使用率进行资源请求分配。

- [权利要求 5] 如权利要求1所述的调度处理器的方法，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：
- 部署多个服务器，所述服务器对应预设数量的所述图形处理器和所述中央处理器，若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求；
- 设置每个所述服务器的最大运算资源值，所述最大运算资源值等于该服务器中所述中央处理器的运算资源与所述图形处理器的运算资源之和；
- 按预设顺序使用所述多个服务器，在给当前服务器分配的资源请求大于该服务器的所述最大运算资源值时，将超出所述最大运算资源值的所述资源请求分配至下一个服务器。
- [权利要求 6] 一种调度处理器的装置，其特征在于，包括：
- 接收模块，用于接收资源请求，资源请求用于请求运算资源，运算资源由中央处理器和图形处理器提供。
- 获取模块，用于获取运算资源的使用率，利用预设的资源分配器，根据使用率满足的条件，分配中央处理和/或图像处理器处理资源请求。
- [权利要求 7] 根据权利要求6所述的调度处理器的装置，其特征在于，所述获取模块包括：
- 获取子模块，用于获取图形处理器使用率；
- 分配子模块，用于在图形处理器使用率低于预设值时，优先分配图形处理器处理资源请求。
- [权利要求 8] 根据权利要求6所述的调度处理器的装置，其特征在于，所述获取模块还用于：利用Nginx负载均衡服务器分配中央处理器和/或图像处理器处理资源请求。
- [权利要求 9] 根据权利要求6所述的调度处理器的装置，其特征在于，所述获取模

块包括：

第一部署子模块，用于部署第一服务器和第二服务器，第一服务器用于接收分配给图形处理器的资源请求，第二服务器用于接收分配给中央处理器的资源请求；

第一设置子模块，用于设置第一服务器的权重高于第二服务器；发送子模块，用于在接收到资源请求之后，根据权重和使用率进行资源请求分配。

[权利要求 10] 根据权利要求6所述的调度处理器的装置，其特征在于，所述获取模块包括：

第二部署子模块，用于部署多个服务器，服务器对应预设数量的图形处理器和中央处理器，若图形处理器使用率低于预设值，则优先分配图形处理器处理资源请求；

第二设置子模块，用于设置每个服务器的最大运算资源值，最大运算资源值等于该服务器中中央处理器的运算资源与图形处理器的运算资源之和；

使用子模块，用于按预设顺序使用多个服务器，在给当前服务器分配的资源请求大于该服务器的最大运算资源值时，将超出最大运算资源值的资源请求分配至下一个服务器。

[权利要求 11] 一种调度处理器的设备，其特征在于，包括存储器以及处理器，所述存储器中存储有可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现如下步骤：

接收资源请求，所述资源请求用于请求运算资源，所述运算资源由中央处理器和图形处理器提供；

获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求。

[权利要求 12] 根据权利要求11所述的调度处理器的设备，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和

/或所述图像处理器处理所述资源请求，包括：

获取图形处理器使用率；

若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求。

[权利要求 13] 根据权利要求11所述的调度处理器的设备，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

利用Nginx负载均衡服务器分配所述中央处理器和/或所述图像处理器处理所述资源请求。

[权利要求 14] 根据权利要求11所述的调度处理器的设备，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

部署第一服务器和第二服务器，所述第一服务器用于接收分配给所述图形处理器的所述资源请求，所述第二服务器用于接收分配给所述中央处理器的所述资源请求；

设置所述第一服务器的权重高于所述第二服务器；

在接收到所述资源请求之后，根据所述权重和所述使用率进行资源请求分配。

[权利要求 15] 根据权利要求11所述的调度处理器的设备，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

部署多个服务器，所述服务器对应预设数量的所述图形处理器和所述中央处理器，若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求；

设置每个所述服务器的最大运算资源值，所述最大运算资源值等于该服务器中所述中央处理器的运算资源与所述图形处理器的运算资源之

和；

按预设顺序使用所述多个服务器，在给当前服务器分配的资源请求大于该服务器的所述最大运算资源值时，将超出所述最大运算资源值的所述资源请求分配至下一个服务器。

[权利要求 16] 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，其特征在于，所述计算机可读指令被至少一个处理器执行时实现如下步骤：

接收资源请求，所述资源请求用于请求运算资源，所述运算资源由中央处理器和图形处理器提供；

获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求。

[权利要求 17] 根据权利要求16所述的计算机可读存储介质，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

获取图形处理器使用率；

若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求。

[权利要求 18] 根据权利要求16所述的计算机可读存储介质，其特征在于，所述利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

利用Nginx负载均衡服务器分配所述中央处理器和/或所述图像处理器处理所述资源请求。

[权利要求 19] 根据权利要求16所述的计算机可读存储介质，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

部署第一服务器和第二服务器，所述第一服务器用于接收分配给所述

图形处理器的所述资源请求，所述第二服务器用于接收分配给所述中央处理器的所述资源请求；

设置所述第一服务器的权重高于所述第二服务器；

在接收到所述资源请求之后，根据所述权重和所述使用率进行资源请求分配。

[权利要求 20] 根据权利要求19所述的计算机可读存储介质，其特征在于，所述获取所述运算资源的使用率，利用预设的资源分配器，根据所述使用率满足的条件，分配所述中央处理和/或所述图像处理器处理所述资源请求，包括：

部署多个服务器，所述服务器对应预设数量的所述图形处理器和所述中央处理器，若所述图形处理器使用率低于预设值，则优先分配所述图形处理器处理所述资源请求；

设置每个所述服务器的最大运算资源值，所述最大运算资源值等于该服务器中所述中央处理器的运算资源与所述图形处理器的运算资源之和；

按预设顺序使用所述多个服务器，在给当前服务器分配的资源请求大于该服务器的所述最大运算资源值时，将超出所述最大运算资源值的所述资源请求分配至下一个服务器。

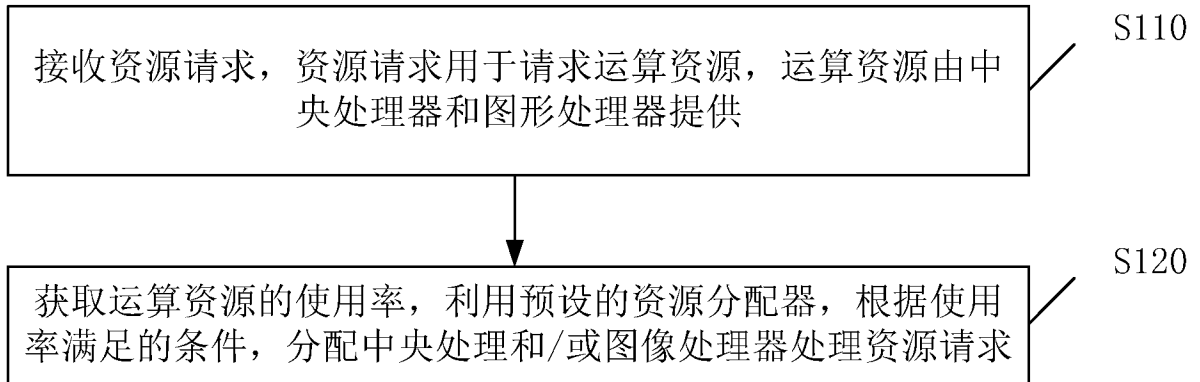


图 1

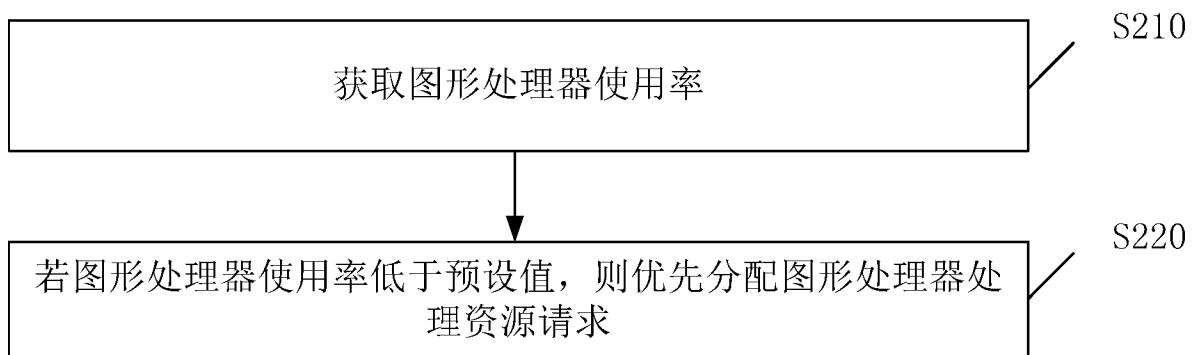


图 2

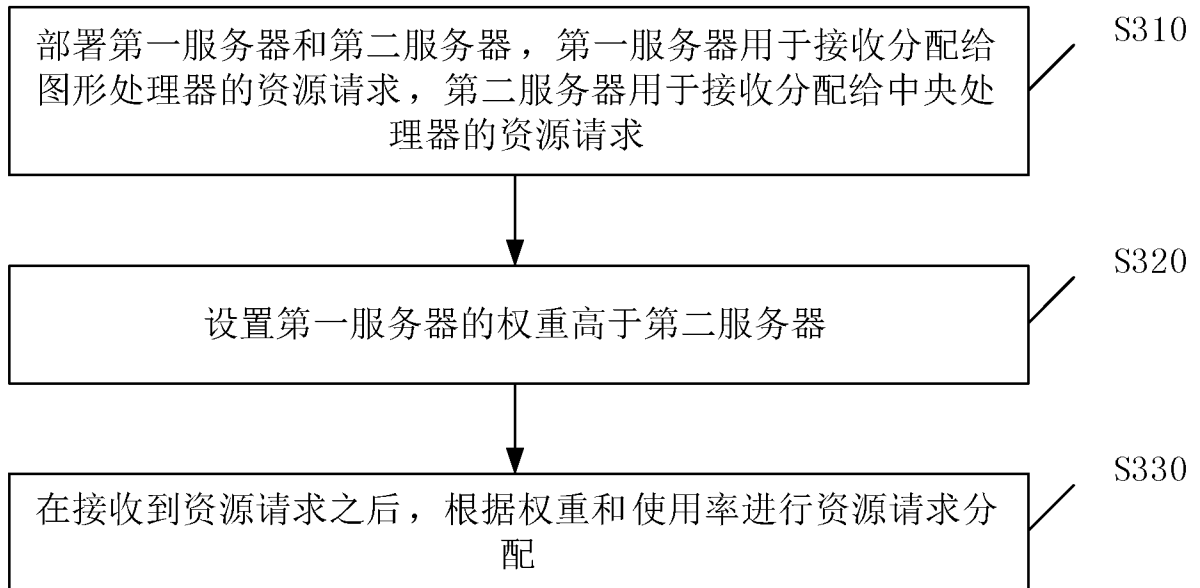


图 3

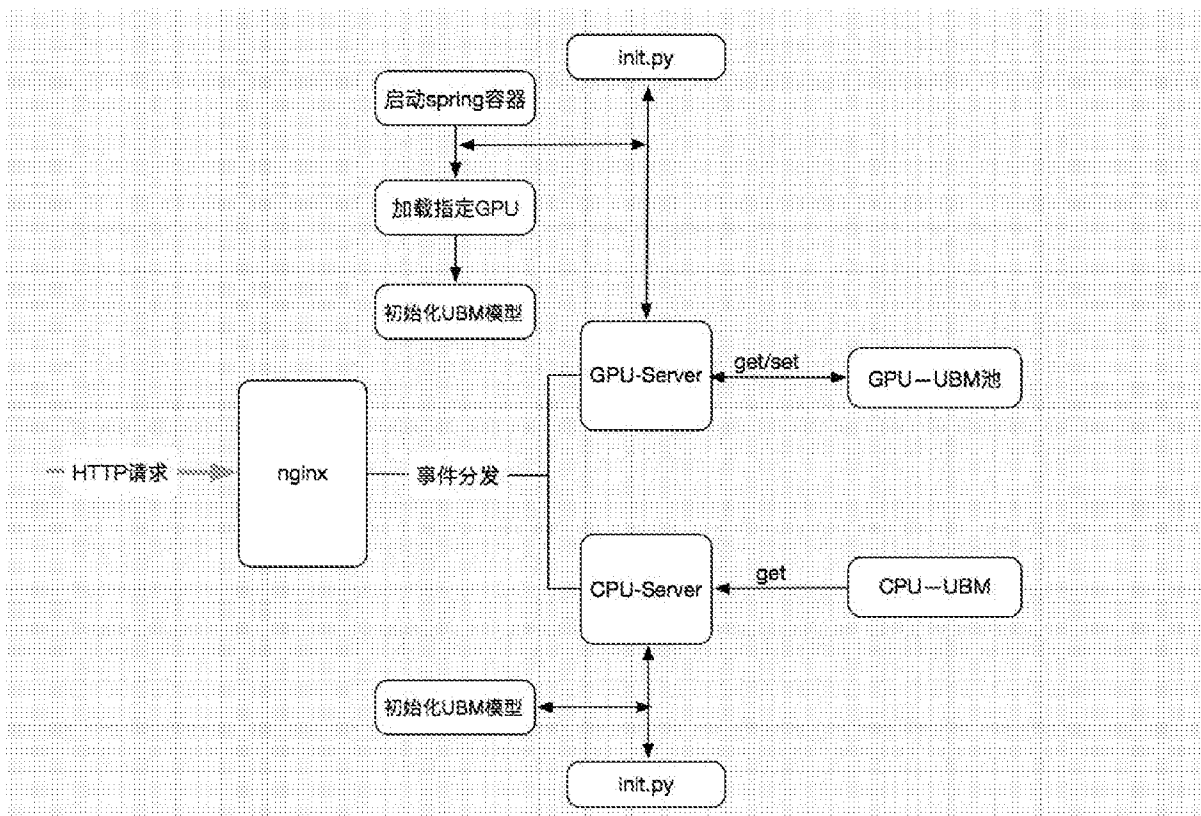


图 4

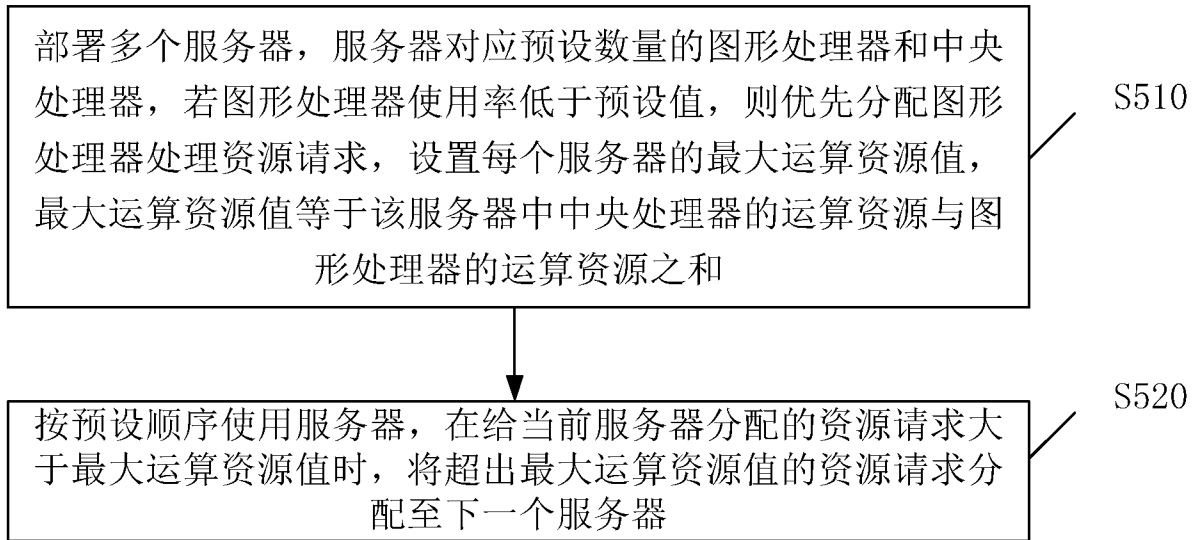


图 5

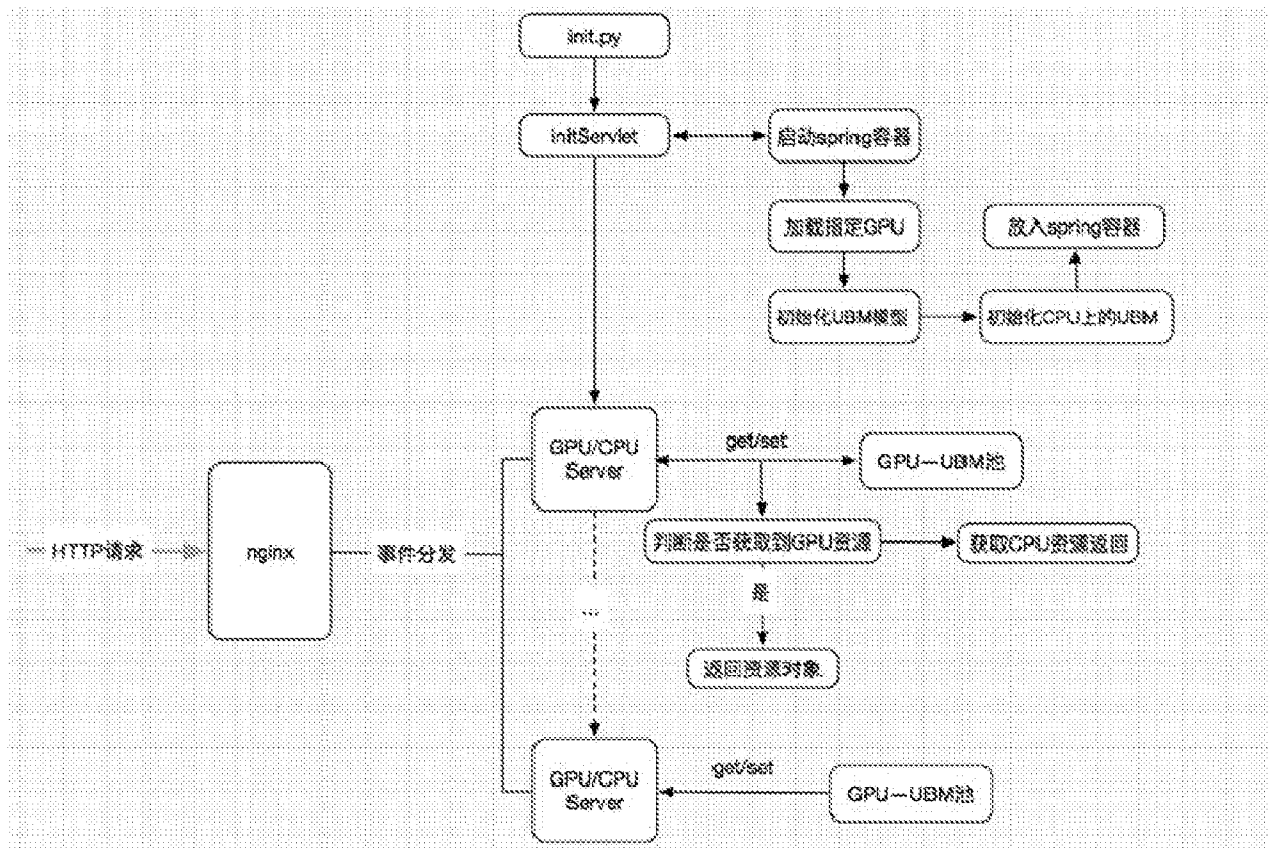


图 6

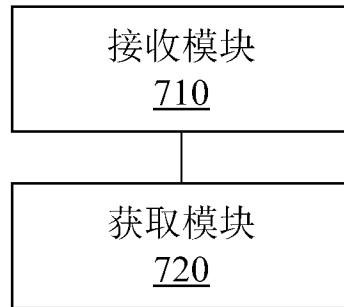


图 7

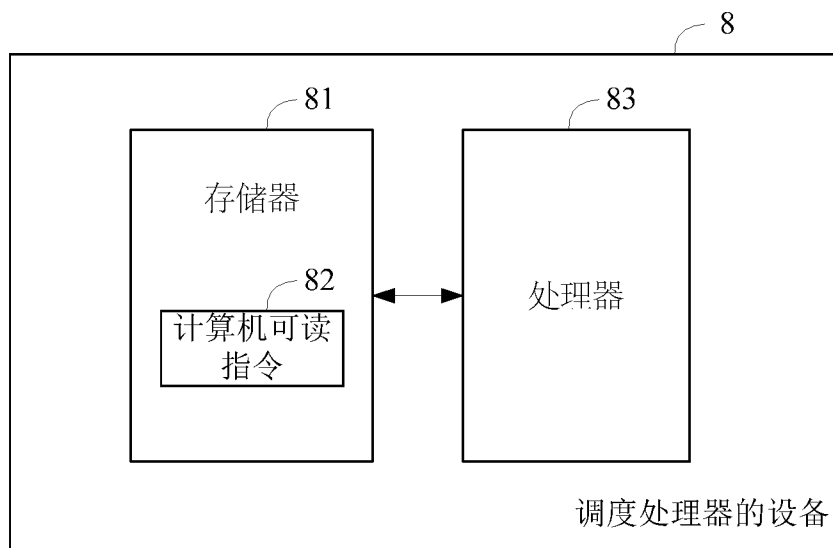


图 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/076009

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPODOC, WPI, CNPAT, CNKI, IEEE, GOOGLE: 调度, 资源, 图形处理器, 使用率, 请求, 中央处理器, 预设值, 阈值, 服务器, 权重, 分配, dispatch+, resource, GPU, usage rate, request, CPU, preset value, threshold, server, weight, assign

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106155811 A (ALIBABA GROUP HOLDING LIMITED), 23 November 2016 (23.11.2016), claims 8-10, description, paragraphs [0050]-[0068], and figures 1-3 and 5	1-3, 6-8, 11-13, 16-18
A	CN 102958166 A (HUAWEI TECHNOLOGIES CO., LTD.), 06 March 2013 (06.03.2013), entire document	1-20
A	CN 105988874 A (ALIBABA GROUP HOLDING LIMITED), 05 October 2016 (05.10.2016), entire document	1-20
A	CN 101447939 A (CHINA MOBILE GROUP BEIJING CO., LTD.), 03 June 2009 (03.06.2009), entire document	1-20
A	US 9576332 B1 (VCE IP HOLDING COMPANY LLC), 21 February 2017 (21.02.2017), entire document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 02 April 2018	Date of mailing of the international search report 26 April 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer WU, Qing Telephone No. (86-10) 010-53961436

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/076009

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106155811 A	23 November 2016	WO 2016173450 A1	03 November 2016
CN 102958166 A	06 March 2013	WO 2013029487 A1	07 March 2013
		EP 2725862 A1	30 April 2014
		US 2014173620 A1	19 June 2014
CN 105988874 A	05 October 2016	WO 2016130487 A1	18 August 2016
		TW 201636838 A	16 October 2016
		US 2016232640 A1	11 August 2016
		HK 1229906 A0	24 November 2017
CN 101447939 A	03 June 2009	None	
US 9576332 B1	21 February 2017	None	

国际检索报告

国际申请号

PCT/CN2018/076009

A. 主题的分类

G06F 9/50(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

EPODOC, WPI, CNPAT, CNKI, IEEE, GOOGLE: 调度, 资源, 图形处理器, 使用率, 请求, 中央处理器, 预设值, 阈值, 服务器, 权重, 分配, dispatch+, resource, GPU, usage rate, request, CPU, preset value, threshold, server, weight, assign

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	CN 106155811 A (阿里巴巴集团控股有限公司) 2016年 11月 23日 (2016 - 11 - 23) 权利要求8-10、说明书第[0050]-[0068]段、附图1-3、5	1-3, 6-8, 11-13, 16-18
A	CN 102958166 A (华为技术有限公司) 2013年 3月 6日 (2013 - 03 - 06) 全文	1-20
A	CN 105988874 A (阿里巴巴集团控股有限公司) 2016年 10月 5日 (2016 - 10 - 05) 全文	1-20
A	CN 101447939 A (中国移动通信集团北京有限公司) 2009年 6月 3日 (2009 - 06 - 03) 全文	1-20
A	US 9576332 B1 (VCE IP HOLDING COMPANY LLC) 2017年 2月 21日 (2017 - 02 - 21) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2018年 4月 2日

国际检索报告邮寄日期

2018年 4月 26日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

吴卿

电话号码 (86-10)010-53961436

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/076009

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	106155811	A	2016年 11月 23日	WO	2016173450	A1	2016年 11月 3日
CN	102958166	A	2013年 3月 6日	WO	2013029487	A1	2013年 3月 7日
				EP	2725862	A1	2014年 4月 30日
				US	2014173620	A1	2014年 6月 19日
CN	105988874	A	2016年 10月 5日	WO	2016130487	A1	2016年 8月 18日
				TW	201636838	A	2016年 10月 16日
				US	2016232640	A1	2016年 8月 11日
				HK	1229906	A0	2017年 11月 24日
CN	101447939	A	2009年 6月 3日	无			
US	9576332	B1	2017年 2月 21日	无			

表 PCT/ISA/210 (同族专利附件) (2009年7月)