(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112035238 B
(45)授权公告日 2024.07.19

(21)申请号 202010957856.3

(22)申请日 2020.09.11

(65)同一申请的已公布的文献号
申请公布号 CN 112035238 A

(43)申请公布日 2020.12.04

(73)专利权人 曙光信息产业(北京)有限公司
地址 100089 北京市海淀区东北旺西路8号院36号楼
专利权人 中科曙光信息产业成都有限公司

(72)发明人 原帅 郝文静 张涛 王家尧 吕灼恒 李斌 沙超群 历军

(74)专利代理机构 北京超凡宏宇知识产权代理有限公司 11463
专利代理师 衡滔

(51)Int.Cl.
G06F 9/48(2006.01)

(56)对比文件
CN 111414234 A,2020.07.14
CN 110389826 A,2019.10.29

审查员 马贺

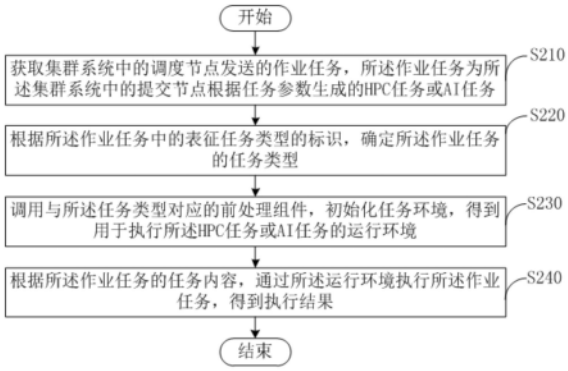
权利要求书3页 说明书11页 附图3页

(54)发明名称

任务调度处理方法、装置、集群系统及可读存储介质

(57)摘要

本申请提供一种任务调度处理方法、装置、集群系统及可读存储介质,涉及集群任务处理技术领域。方法包括:获取集群系统中的调度节点发送的作业任务,作业任务为集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;根据作业任务中的表征任务类型的标识,确定作业任务的任务类型;调用与任务类型对应的前处理组件,初始化任务环境,得到用于执行HPC任务或AI任务的运行环境;根据作业任务的任务内容,通过运行环境执行作业任务,得到执行结果,能够改善计算节点执行的任务类型单一,硬件资源利用率低的问题。



CN 112035238 B

1.一种任务调度处理方法,其特征在于,应用于集群系统中的计算节点,所述方法包括:

获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;

根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果;

调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境,包括:

当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;

当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境;

所述前处理组件包括通用处理组件、AI框架处理组件,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境,包括:

调用所述通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源,其中,每个子任务对应的目标硬件资源根据对应子任务所需的运算量从所有的目标计算节点的硬件资源中选择获得,所述目标硬件资源包括目标计算节点的身份标识、CPU的身份标识、内核的身份标识以及GPU的身份标识;

调用所述AI框架处理组件,选择与所述AI任务对应的处理框架、加速器,其中,所述AI任务重携带有表征执行该AI任务所需的处理框架信息和加速器信息;

根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括:

清除与所述作业任务对应的目标硬件资源的关联关系、所述容器。

3.根据权利要求1所述的方法,其特征在于,获取所述集群系统中的调度节点发送的作业任务,包括:

获取所述集群系统中通过所述调度节点的HPC调度器发送的作业任务。

4.一种任务调度处理方法,其特征在于,应用于集群系统,所述集群系统包括提交节点、调度节点及多个计算节点,所述方法包括:

所述提交节点,根据任务参数生成作业任务,所述作业任务包括HPC任务或AI任务;

所述调度节点,从所述提交节点获取所述作业任务;

所述调度节点,从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点;

所述目标计算节点,根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

所述目标计算节点,调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

所述目标计算节点,根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果;

所述目标计算节点,当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境;其中,所述调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境,包括:调用通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源,其中,每个子任务对应的目标硬件资源根据对应子任务所需的运算量从所有的目标计算节点的硬件资源中选择获得,所述目标硬件资源包括目标计算节点的身份标识、CPU的身份标识、内核的身份标识以及GPU的身份标识;调用AI框架处理组件,选择与所述AI任务对应的处理框架、加速器;根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境,其中,所述AI任务重携带有表征执行该AI任务所需的处理框架信息和加速器信息。

5.一种任务调度处理装置,其特征在于,应用于集群系统中的计算节点,所述装置包括:

获取单元,获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;

确定单元,用于根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

前处理单元,用于调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

执行单元,用于根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果;

所述前处理单元,具体用于当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境;其中,所述调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境,包括:调用通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源,其中,每个子任务对应的目标硬件资源根据对应子任务所需的运算量从所有的目标计算节点的硬件资源中选择获得,所述目标硬件资源包括目标计算节点的身份标识、CPU的身份标识、内核的身份标识以及GPU的身份标识;调用AI框架处理组件,选择与所述AI任务对应的处理框架、加速器,其中,所述AI任务重携带有表征执行该AI任务所需的处理框架信息和加速器信息;根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

6.一种服务器,其特征在于,所述服务器包括相互耦合的存储器、处理器,所述存储器内存储计算机程序,当所述计算机程序被所述处理器执行时,使得所述服务器执行如权利要求1-5中任一项所述的方法。

7.一种集群系统,其特征在于,所述集群系统包括提交节点、调度节点及多个计算节点,其中:

所述提交节点,用于根据任务参数生成作业任务,所述作业任务包括HPC任务或AI任务;

所述调度节点,用于从所述提交节点获取所述作业任务;

所述调度节点,还用于从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点;

所述目标计算节点,用于根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

所述目标计算节点,还用于调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

所述目标计算节点,还用于根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果;

所述目标计算节点,具体用于当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境;其中,所述调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境,包括:调用通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源;调用AI框架处理组件,选择与所述AI任务对应的处理框架、加速器;根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

8.一种计算机可读存储介质,其特征在于,所述可读存储介质中存储有计算机程序,当所述计算机程序在计算机上运行时,使得所述计算机执行如权利要求1-3中任意一项所述的方法。

任务调度处理方法、装置、集群系统及可读存储介质

技术领域

[0001] 本发明涉及集群任务处理技术领域,具体而言,涉及一种任务调度处理方法、装置、集群系统及可读存储介质。

背景技术

[0002] 随着计算机集群处理技术的发展,超级计算机性能越来越高。集群系统通常需要支持高性能计算(High Performance Computing,HPC)任务的计算,还要支持人工智能(Artificial Intelligence,AI)任务的计算。目前,通常是将集群系统的硬件资源划分成面向不同领域的小集群或计算节点。每个小集群或计算节点执行的任务类型单一。例如,用于执行HPC任务的小集群便无法执行AI任务,从而使得集群的硬件资源的利用率低。

发明内容

[0003] 本申请提供一种任务调度处理方法、装置、集群系统及可读存储介质,能够改善集群中计算节点执行的任务类型单一,硬件资源利用率低的问题。

[0004] 为了实现上述目的,本申请实施例所提供的技术方案如下所示:

[0005] 第一方面,本申请实施例提供一种任务调度处理方法,应用于集群系统中的计算节点,所述方法包括:

[0006] 获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;

[0007] 根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0008] 调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0009] 根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0010] 在上述的实施方式中,计算节点可以根据任务类型,对任务环境进行前处理,以得到用于执行HPC任务或AI任务的运行环境,然后便可以基于得到的运行环境执行HPC任务或AI任务,从而改善计算节点执行的任务类型单一,硬件资源利用率低的问题。

[0011] 结合第一方面,在一些可选的实施方式中,调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境,包括:

[0012] 当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;

[0013] 当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境。

[0014] 在上述的实施方式中,针对HPC任务、AI任务,通过分别对任务环境进行前处理,得到相应的运行环境,使得计算节点能够执行不同任务类型的作业任务。

[0015] 结合第一方面,在一些可选的实施方式中,所述前处理组件包括通用处理组件、AI

框架处理组件,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境,包括:

[0016] 调用所述通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源;

[0017] 调用所述AI框架处理组件,选择与所述AI任务对应的处理框架、加速器;

[0018] 根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

[0019] 在上述的实施方式中,通过创建用于执行AI任务的容器与运行环境,使得计算节点能够执行AI任务。

[0020] 结合第一方面,在一些可选的实施方式中,所述处理框架包括DL框架。

[0021] 结合第一方面,在一些可选的实施方式中,所述方法还包括:

[0022] 清除与所述作业任务对应的目标硬件资源的关联关系、所述容器。

[0023] 在上述的实施方式中,在得到执行结果后,通过删除关联关系、容器等,有利于计算节点对新任务的执行,避免当前的作业任务的运行环境影响新任务的执行。

[0024] 结合第一方面,在一些可选的实施方式中,获取所述集群系统中的调度节点发送的作业任务,包括:

[0025] 获取所述集群系统中通过所述调度节点的HPC调度器发送的作业任务。

[0026] 在上述的实施方式中,HPC调度器可以对AI任务及HPC任务进行调度,改善HPC调度器仅能对HPC任务调度的问题。

[0027] 第二方面,本申请实施例还提供一种任务调度处理方法,应用于集群系统,所述集群系统包括提交节点、调度节点及多个计算节点,所述方法包括:

[0028] 所述提交节点,根据任务参数生成作业任务,所述作业任务包括HPC任务或AI任务;

[0029] 所述调度节点,从所述提交节点获取所述作业任务;

[0030] 所述调度节点,从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点;

[0031] 所述目标计算节点,根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0032] 所述目标计算节点,调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0033] 所述目标计算节点,根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0034] 第三方面,本申请实施例还提供一种任务调度处理装置,应用于集群系统中的计算节点,所述装置包括:

[0035] 获取单元,获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;

[0036] 确定单元,用于根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0037] 前处理单元,用于调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0038] 执行单元,用于根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0039] 第四方面,本申请实施例还提供一种服务器,所述服务器包括相互耦合的存储器、处理器,所述存储器内存储计算机程序,当所述计算机程序被所述处理器执行时,使得所述服务器执行上述的方法。

[0040] 第五方面,本申请实施例还提供一种集群系统,所述集群系统包括提交节点、调度节点及多个计算节点,其中:

[0041] 所述提交节点,用于根据任务参数生成作业任务,所述作业任务包括HPC任务或AI任务;

[0042] 所述调度节点,用于从所述提交节点获取所述作业任务;

[0043] 所述调度节点,还用于从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点;

[0044] 所述目标计算节点,用于根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0045] 所述目标计算节点,还用于调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0046] 所述目标计算节点,还用于根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0047] 第六方面,本申请实施例还提供一种计算机可读存储介质,所述可读存储介质中存储有计算机程序,当所述计算机程序在计算机上运行时,使得所述计算机执行上述的方法。

附图说明

[0048] 为了更清楚地说明本申请实施例的技术方案,下面将对实施例中所需要使用的附图作简单地介绍。应当理解,以下附图仅示出了本申请的某些实施例,因此不应被看作是对范围的限定,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他相关的附图。

[0049] 图1为本申请实施例提供的集群系统的通信连接示意图。

[0050] 图2为本申请实施例提供的计算节点的硬件资源的方框示意图。

[0051] 图3为本申请实施例提供的任务调度处理方法的流程示意图之一。

[0052] 图4为本申请实施例提供的任务调度处理方法的流程示意图之二。

[0053] 图5为本申请实施例提供的任务调度处理装置的功能框图。

[0054] 图标:10-集群系统;20-计算节点;30-调度节点;40-提交节点;300-任务调度处理装置;310-获取单元;320-确定单元;330-前处理单元;340-执行单元。

具体实施方式

[0055] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行描述。需要说明的是,术语“第一”、“第二”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0056] 申请人发现,目前集群系统的硬件资源通常需要被划分成面向不同领域的小集群。一个小集群通常包括一个或多个计算节点。通常而言,不同领域的任务所需要的运行环境不同,因此,每个小集群通常只能执行所划分领域的一个任务,而无法执行其他领域的任务。例如,用于执行HPC任务的小集群无法执行AI任务。因此,目前的集群系统中,计算节点所执行的任务类型单一,存在利用率的问题。

[0057] 鉴于上述问题,本申请申请人经过长期研究探索,提出以下实施例以解决上述问题。下面结合附图,对本申请实施例作详细说明。在不冲突的情况下,下述的实施例及实施例中的特征可以相互组合。

[0058] 第一实施例

[0059] 请参照图1,本申请实施例提供一种集群系统10,可以用于执行下述的任务调度处理方法中的各步骤,能够改善计算节点20执行的任务类型单一,使得硬件资源无法得到充分利用的问题。

[0060] 在本实施例中,集群系统10可以包括提交节点40、调度节点30及多个计算节点20。其中,集群系统10中的一个节点(例如,提交节点40、调度节点30、计算节点20等)为一个服务器。一个节点可以以提交节点40、调度节点30与计算节点20中至少一个身份运行。例如,一个提交节点40可以以提交节点40的身份运行,另外,该提交节点40还可以以调度节点30、计算节点20的身份运行。通常而言,提交节点40、调度节点30与计算节点20为不同的节点。

[0061] 在本实施例中,提交节点40可以通过网络与用户终端建立通信连接,以进行数据交互。提交节点40可以通过网络与调度节点30建立通信连接,以进行数据交互。调度节点30可以通过网络与一个或多个计算节点20建立通信连接,以进行数据交互。

[0062] 例如,用户终端可以将需要执行的作业任务的相关信息发送至提交节。提交节点40可以基于作业任务的相关信息,生成作业任务的脚本文件。该脚本文件即为计算机“可理解”的作业任务。另外,提交节点40可以将作业任务的脚本文件发送至调度节点30。调度节点30可以将脚本文件发送至相应的目标计算节点20。然后由目标计算节点20执行脚本文件对应的作业任务。其中,目标计算节点20可以为一个或多个计算节点20。

[0063] 用户终端可以是,但不限于,智能手机、个人电脑(Personal Computer,PC)、平板电脑、个人数字助理(Personal Digital Assistant,PDA)、移动上网设备(Mobile Internet Device,MID)等。网络可以是,但不限于,有线网络或无线网络。

[0064] 请参照图2,在本实施例中,计算节点20所包括的硬件资源包括但不限于中央处理器(Central Processing Unit,CPU)、图形处理器(Graphics Processing Unit,GPU)、内存。可理解地,一个CPU可以设置有一个或多个内核,处理器所包括的内核数量可以根据实际情况进行设置。例如,CPU可以为单个内核的处理器,或者为双内核的处理器。

[0065] 在一个计算节点20中,内核及图形处理器的数量均可以根据实际情况进行设置。作为一个示例,计算节点20可以包括如图2所示的N个中央处理器、M个图形处理器。N、M均为大于2的整数,可以相同或不同,可以根据实际情况进行设置。不同的计算节点20的硬件资源可以相同或不同,可以根据实际情况进行设置。例如,不同计算节点20的内核数量、图形处理器的数量、内核的运行参数、图形处理器的运行参数可以均不相同。

[0066] 请参照图3,本申请实施例还提供一种任务调度处理方法,可以应用于上述的集群系统10中,由集群系统10中的相应节点相互配合以执行方法中的各步骤。方法可以包括步

骤,如下:

[0067] 步骤S110,提交节点,根据任务参数生成作业任务,所述作业任务包括HPC任务或AI任务;

[0068] 步骤S120,调度节点,从所述提交节点获取所述作业任务;

[0069] 步骤S130,所述调度节点,从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点;

[0070] 步骤S140,目标计算节点,根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0071] 步骤S150,所述目标计算节点,调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0072] 步骤S160,所述目标计算节点,根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0073] 在本实施例中,计算节点可以根据任务类型,对任务环境进行前处理,以得到用于执行HPC任务或AI任务的运行环境,然后便可以基于得到的运行环境执行HPC任务或AI任务,从而改善计算节点执行的任务类型单一,硬件资源利用率低的问题。

[0074] 下面将对方法中的各步骤进行详细阐述,如下:

[0075] 在步骤S110中,提交节点在获取到任务参数后,可以自动根据任务参数,生成作业脚本。该作业脚本即为计算机“可理解”的作业任务。若任务参数包括表征AI任务的第一标识,提交节点便可以根据该任务参数,生成AI任务。若任务参数包括表征HPC任务的第二标识,提交节点便可以根据该任务参数,生成HPC任务。第一标识与第二标识不相同,可以为数字或字符,用于区分AI任务与HPC任务,可以根据实际情况进行设置。另外,提交节点生成的作业任务中,包括表征该作业任务的任务类型的标识,以便于计算节点根据不同类型的作业任务来执行。例如,AI任务中可以包括表征该任务为AI任务的第一标识,HPC任务中可以包括表征该任务为HPC任务的任务标识。

[0076] 在本实施例中,提交节点可以从用户终端获取到任务参数。用户终端提交的任务参数的格式可以为指定格式,例如,指定格式为JSON格式,以便于提交节点读取任务参数中的各项子参数。任务参数通常为用户根据实际需求,通过用户终端上传至提交节点的参数,可以根据实际情况进行设置。任务参数可以包括但不限于表征任务类型的标识、执行该任务所需的硬件要求(例如,执行任务所需的内核数量、内核/CPU运行时的额定时钟频率、GPU数量、GPU运行时的额定时钟频率)、用户信息、任务内容、环境变量等。例如,若作业任务为AI任务,该AI任务的任务参数包括但不限于用户信息、处理框架、镜像文件、执行该任务所需的硬件要求、DL(Deep Learning,深度学习)参数等。

[0077] 其中,镜像文件可理解为任务参数中除去镜像文件后的数据形成的镜像文件,可以作为任务参数的备份文件。处理框架可以包括DL框架或其他框架。处理框架、DL参数为本领域技术人员所熟知。例如,处理框架可以是但不限于TensorFlow、PyTorch、MXNet、Caffe、Keras等为本领域技术人员所熟知的框架。DL参数包括但不限于学习率、阈值等参数。

[0078] 在步骤S120中,调度节点可以自动从提交节点获取提交节点生成的作业任务,比如,调度节点可以每隔预设时长,从提交节点获取该预设时长内所生成的作业任务,预设时长可以根据实际情况进行设置,例如,预设时长可以为1分钟、10分钟、1小时等时长。或者,

提交节点可以自动将生成的作业任务发送至调度节点,以使调度节点获取到作业任务。可理解地,调度节点获取的作业任务的方式可以根据实际情况进行设置,这里不做具体限定。

[0079] 在步骤S130中,调度节点可以根据集群系统中各个计算节点当前的运行情况,结合作业任务携带的执行该任务所需的硬件要求信息,从多个计算节点中选取能够满足执行当前的作业任务的硬件要求的一个或多个计算节点,作为目标计算节点,然后将作业任务发送至目标计算节点。

[0080] 可理解地,所选取的目标计算节点的硬件性能能够满足执行该作业任务的需求。即,目标计算节点的各项硬件资源的参数均大于或等于执行该作业任务所需的硬件要求所表征的各项硬件资源的参数。

[0081] 在本实施例中,调度节点可以实时获取到集群系统中各个计算节点的运行参数,或者,在接收到作业任务时,获取集群系统中各个计算节点的运行参数。运行参数包括每个节点的总硬件资源信息、闲置硬件资源信息。其中,总硬件资源信息包括但不限于节点所包括的CPU数量、每个CPU的内核数量、每个CPU运行时的额定时钟频率、GPU数量、每个GPU运行时的额定时钟频率、内存的总容量、内核的身份标识、GPU的身份标识等。闲置硬件资源信息包括但不限于未执行作业任务的CPU的身份标识、未执行作业任务的CPU的内核的身份标识、内存的剩余容量等。其中,额定时钟频率越大的CPU或GPU的运算能力越强。

[0082] 请再次参照图2,假设,一个集群系统包括计算节点A和计算节点B,计算节点A包括8个CPU,每个CPU包括8个内核,每个CPU的额定工作频率(主频)为主频4.0GHz,4个GPU,每个GPU的显存为8GB,额定工作频率为1500MHz。计算节点B包括8个CPU,每个CPU包括4个内核,每个CPU的额定工作频率(主频)为主频4.0GHz,2个GPU,每个GPU的显存为4GB,额定工作频率为1000MHz。若当前的工作任务为AI任务,执行该AI任务的所需硬件要求包括:内核个数至少16个,CPU/内核的主频不小于4.0GHz,GPU数量至少4个,GPU的显存不小于8GB,工作频率不小于1000MHz。由于计算节点A满足执行任务所需去硬件要求,计算节点B不满足该硬件要求,此时,调度节点可以基于执行任务所需的硬件要求,选择计算节点A作为目标计算节点,然后向目标计算节点发送作业任务。

[0083] 在步骤S140中,目标计算节点可以根据作业任务中携带的标识,确定该作业任务的任务类型。例如,若作业任务的标识为表征AI任务的第一标识,则确定该作业任务为AI任务,任务类型为AI类。若作业任务的标识为表征HPC任务的第二标识,则确定该作业任务为HPC任务,任务类型为HPC类。

[0084] 在步骤S150中,目标计算节点可以预先存储前处理组件与任务类型的关联关系。即,AI任务的前处理组件与AI类的标识关联,HPC任务的前处理组件与HPC类的标识关联。在确定作业任务的任务类型后,目标计算节点可以自动根据任务类型的标识,选择与任务类型对应的前处理组件。然后运行前处理组件,初始化任务环境,便可以得到用于执行当前的作业任务的运行环境。

[0085] 在步骤S160中,目标计算节点在得到用于执行当前作业任务的运行环境后,便可以通过该运行环境,执行该作业任务,从而得到执行结果。其中,计算节点执行作业任务的过程为本领域技术人员所熟知,这里不再赘述。执行结果与执行任务对应,可以根据实际情况进行确定。例如,HPC任务的目的为创建天气预报模型,得到的执行结果便为一个天气预报模型。AI任务的目的为创建人脸识别模型,得到的执行结果便为一个人脸识别模型。

[0086] 若目标计算节点为多个计算节点,各个目标计算节点可以相互协商,将作业任务细分成多个子任务,然后由各个目标计算节点执行相应的子任务。其中,作业任务的细分与协商处理为本领域技术人员所熟知,这里不再赘述。

[0087] 作为一种可选的实施方式,步骤S110还可以包括:获取所述集群系统中通过所述调度节点的HPC调度器发送的作业任务。

[0088] 可理解地,在本实施例中,HPC调度器可以具有调度HPC任务和AI任务的功能。由提交节点根据作业参数生成作业任务后,调度节点可以结合作业任务的任务类型及执行任务所需的硬件要求,选择相应的目标计算节点,实现任务的调度,改善HPC调度器无法调度AI任务的问题。

[0089] HPC调度器可以是但不限于LSF(Load Sharing Facility,加载共享设施)、Slurm等调度器。Slurm工具是面向Linux和Unix类似内核的开源工作调度程序,可以供计算机集群使用。

[0090] 作为一种可选的实施方式,步骤S150可以包括:

[0091] 当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;

[0092] 当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所述AI任务的运行环境。

[0093] 可理解地,目标计算节点可以根据作业任务的具体任务类型,选择相应的前处理组件。若作业任务为HPC任务,目标计算节点便调用与HPC任务对应的前处理组件,通过该前处理组件初始化任务环境,得到用于执行HPC任务的运行环境。若作业任务为AI任务,目标计算节点便调用与AI任务对应的前处理组件,初始化任务环境,得到用于执行AI任务的运行环境。基于此,目标计算节点可以根据作业任务的任务类型,搭建与任务类型对应的运行环境,从而可以执行AI任务、HPC任务,改善计算节点只能执行单一类型的任务的问题。

[0094] 在本实施例中,前处理组件通常可以包括多类组件,每类组件可以用于搭建相应的任务环境。在运行前处理组件时,各类组件可以相互配合,搭建得到用于执行当前作业任务的运行环境。

[0095] 作为一种可选的实施方式,当作业任务为AI任务时,前处理组件包括通用处理组件、AI框架处理组件。步骤S150还可以包括:

[0096] 调用所述通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源;

[0097] 调用所述AI框架处理组件,选择与所述AI任务对应的处理框架、加速器;

[0098] 根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

[0099] 可理解地,目标计算节点可以将作业任务划分为多个子任务,划分子任务的方式为本领域技术人员所熟知,这里不再赘述。当作业任务为AI任务时,计算节点可以通过调用通用处理组件,解析作业任务中的任务参数(如执行任务所需的硬件资源的参数)、环境变量、收集用户信息、用户组文件等,然后根据每个子任务所需的运算量,从所有的目标计算节点的硬件资源中,选择执行该子任务所需的硬件资源以作为该子任务的目标硬件资源。目标硬件资源包括但不限于目标计算节点的身份标识、CPU的身份标识、内核的身份标识、GPU的身份标识。其中,环境变量可以根据实际情况进行确定,例如,可以为计算节点的操作

系统运行环境的一些参数,如:临时文件夹位置和系统文件夹位置等。

[0100] 计算节点可以通过调用AI框架处理组件,与AI任务对应的处理框架、加速器。可理解地,AI任务中,可以携带有执行该AI任务所需的处理框架的信息、加速器的信息。例如,AI任务中携带有表征执行该AI任务所需的处理框架为TensorFlow,所需的加速器为Nvidia加速器。当然,加速器还可以为其他类型的加速器,比如AMD加速器,这里对加速器的类型不做具体限定。

[0101] 为了便于理解计算节点实现前处理的过程,下面将举例阐述计算节点通过进行前处理得到相应运行环境的实现过程:

[0102] 目标计算节点收到调度节点发送的作业任务时,便开始执行Prolog,然后检测作业任务的任务类型。其中,Prolog是一种面向演绎推理的逻辑型程序设计语言。Prolog可理解为是一段程序的前言,Epilog可理解为是一段程序的尾言。编译器会在每一个函数的开头塞入Prolog代码,在每一个函数的结尾塞入Epilog代码。

[0103] 当检测出作业任务为AI任务时,便执行AI任务的通用Prolog,调用AI任务的通用处理组件、AI框架处理组件。当检测出作业任务为HPC任务时,可以直接调用HPC任务的前处理组件。

[0104] 其中,调用AI任务的通用处理组件、AI框架处理组件的执行过程可以为:通过通用处理组件,获取作业任务的任务内容/任务参数、环境变量、用户信息、用户组文件等。然后为AI任务的子任务,选择相应的硬件资源以作为子任务的目标硬件资源,以及根据任务内容选择加速器的类型(Nvidia或AMD)、以及DL框架。然后,通过AI框架处理组件,基于AI任务的各个子任务,分配用于创建执行该子任务的容器所需的硬件资源。该创建容器所需的硬件资源即为该子任务的目标硬件资源。然后,根据所选取的目标硬件资源、处理框架、加速器创建用于执行子任务的容器,并记录容器的信息,例如,记录该容器与子任务、目标硬件资源的管理关系,此时,便可以创建得到用于执行AI任务的运行环境。

[0105] 当作业任务为HPC任务时,计算节点可以直接调用HPC任务的前处理组件,以使计算节点的当前任务环境变成为能够执行HPC任务的运行环境。

[0106] 在本实施例中,计算节点默认的任务环境可以为能够执行HPC任务的运行环境。当作业任务为HPC任务时,若任务环境不是用于执行HPC的运行环境,则调用HPC任务的前处理组件,以使任务环境恢复为默认的运行环境。

[0107] 作为一种可选的实施方式,方法还可以包括:清除与所述作业任务对应的目标硬件资源的关联关系、所述容器。

[0108] 可理解地,在步骤S140之后,计算节点还可以清除作业任务对应的目标硬件资源的关联关系、容器、环境变量、以及执行作业任务中产生的临时文件与临时数据等。通过清除该关联关系、容器等数据,有利于计算节点对新任务的执行,让任务环境恢复至执行作业任务前,避免当前作业任务的运行环境影响新任务的执行。

[0109] 在得到执行结果后,集群系统可以存储该执行结果,或者由计算节点将执行结果发送至用户终端,以供用户查看执行结果。或者由计算节点将执行结果通过调度节点发送至提交节点,然后由提交节点发送至用户终端。

[0110] 基于上述设计,集群系统的硬件资源共享,同一个计算节点,可同时承担高性能计算、人工智能等各种任务,提升硬件资源使用率。AI任务的硬件资源统一分配,避免AI分布

式任务占用部分硬件资源,又无法运行,而浪费硬件资源。另外,可以基于HPC调度器,通过计算节点的前处理和后处理,完成容器的创建和销毁,实现对容器的编排调度,实现HPC调度器对AI任务的支持。在保留容器灵活性、快捷方便的同时,可以实现对HPC任务和AI任务的融合调度。

[0111] 第二实施例

[0112] 请参照图4,本申请还提供另一种任务调度处理方法,可以应用于集群系统中的计算节点。方法可以包括如下步骤:

[0113] 步骤S210,获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;

[0114] 步骤S220,根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型;

[0115] 步骤S230,调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境;

[0116] 步骤S240,根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0117] 可理解地,与第一实施例中的任务调度处理方法相比,在第二实施例中,任务调度处理方法的实现过程及取得的技术效果与第一实施例提供的方法相类似,区别在于,第二实施例中的任务调度方法应用于计算节点,由计算节点执行方法中的各步骤。当然,第二实施例中的任务调度处理方法还可以包括其他步骤,例如还可以包括如第一实施例中计算节点所执行的其他步骤,这里不再赘述。其中,执行任务调度处理方法的计算节点即为调度节点所确定的目标计算节点。

[0118] 请参照图5,本申请实施例还提供一种任务调度处理装置300,可以应用于集群系统中的计算节点,用于执行计算节点所执行的各步骤。任务调度处理装置300包括至少一个可以软件或固件(Firmware)的形式存储于存储模块中或固化在服务器操作系统(Operating System,OS)中的软件功能模块。处理模块用于执行存储模块中存储的可执行模块,例如任务调度处理装置300所包括的软件功能模块及计算机程序等。

[0119] 任务调度处理装置300可以包括获取单元310、确定单元320、前处理单元330及执行单元340。

[0120] 获取单元310,获取所述集群系统中的调度节点发送的作业任务,所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务。

[0121] 确定单元320,用于根据所述作业任务中的表征任务类型的标识,确定所述作业任务的任务类型。

[0122] 前处理单元330,用于调用与所述任务类型对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务或AI任务的运行环境。

[0123] 执行单元340,用于根据所述作业任务的任务内容,通过所述运行环境执行所述作业任务,得到执行结果。

[0124] 可选地,前处理单元330用于:当所述作业任务为HPC任务时,调用与所述HPC任务对应的前处理组件,初始化任务环境,得到用于执行所述HPC任务的运行环境;当所述作业任务为AI任务时,调用与所述AI任务对应的前处理组件,初始化任务环境,得到用于执行所

述AI任务的运行环境。

[0125] 可选地,前处理组件包括通用处理组件、AI框架处理组件。前处理单元330还用于:调用所述通用处理组件,选择与所述AI任务中的子任务对应的目标硬件资源;调用所述AI框架处理组件,选择与所述AI任务对应的处理框架、加速器;根据所述目标硬件资源、所述处理框架、所述加速器,创建用于执行所述子任务的容器,得到用于执行所述AI任务的运行环境。

[0126] 可选地,任务调度处理装置300还可以包括清除单元,用于清除与所述作业任务对应的目标硬件资源的关联关系、所述容器。

[0127] 可选地,获取单元310用于:获取所述集群系统中通过所述调度节点的HPC调度器发送的作业任务。

[0128] 需要说明的是,所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的集群系统、任务调度处理装置300、计算节点的具体工作过程,可以参考前述方法中的各步骤对应过程,在此不再过多赘述。

[0129] 在本实施例中,集群系统中的服务器(例如,计算节点)可以包括处理模块、通信模块、存储模块以及任务调度处理装置300,处理模块、通信模块、存储模块以及任务调度处理装置300各个元件之间直接或间接地电性连接,以实现数据的传输或交互。例如,这些元件相互之间可通过一条或多条通讯总线或信号线实现电性连接。

[0130] 处理模块可以是一种集成电路芯片,具有信号的处理能力。上述处理模块可以是通用处理器。例如,该处理器可以是中央处理器(Central Processing Unit,CPU)、图形处理器(Graphics Processing Unit,GPU)、网络处理器(Network Processor,NP)等;还可以是数字信号处理器(Digital Signal Processing,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程门阵列(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,可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。

[0131] 存储模块可以是,但不限于,随机存取存储器,只读存储器,可编程只读存储器,可擦除可编程只读存储器,电可擦除可编程只读存储器等。在本实施例中,存储模块可以用于存储作业任务的相关信息。当然,存储模块还可以用于存储程序,处理模块在接收到执行指令后,执行该程序。

[0132] 通信模块用于通过网络建立节点自身与集群系统中的其他节点的通信连接,并通过网络收发数据。

[0133] 本申请实施例还提供一种计算机可读存储介质。可读存储介质中存储有计算机程序,当计算机程序在计算机上运行时,使得计算机执行如上述实施例中所述的任务调度处理方法。

[0134] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到本申请可以通过硬件实现,也可以借助软件加必要的通用硬件平台的方式来实现,基于这样的理解,本申请的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施场景所述的方法。

[0135] 综上所述,本申请提供一种任务调度处理方法、装置、集群系统及可读存储介质。

方法包括:获取集群系统中的调度节点发送的作业任务,作业任务为集群系统中的提交节点根据任务参数生成的HPC任务或AI任务;根据作业任务中的表征任务类型的标识,确定作业任务的任务类型;调用与任务类型对应的前处理组件,初始化任务环境,得到用于执行HPC任务或AI任务的运行环境;根据作业任务的任务内容,通过运行环境执行作业任务,得到执行结果。在本方案中,计算节点可以根据任务类型,对任务环境进行前处理,以得到用于执行HPC任务或AI任务的运行环境,然后便可以基于得到的运行环境执行HPC任务或AI任务,从而改善计算节点执行的任务类型单一,硬件资源利用率低的问题。

[0136] 在本申请所提供的实施例中,应该理解到,所揭露的装置、系统和方法,也可以通过其它的方式实现。以上所描述的装置、系统和方法实施例仅仅是示意性的,例如,附图中的流程图和框图显示了根据本申请的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或代码的一部分,所述模块、程序段或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也要注意的是,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。另外,在本申请各个实施例中的各功能模块可以集成在一起形成一个独立的部分,也可以是各个模块单独存在,也可以两个或两个以上模块集成形成一个独立的部分。

[0137] 以上所述仅为本申请的优选实施例而已,并不用于限制本申请,对于本领域的技术人员来说,本申请可以有各种更改和变化。凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

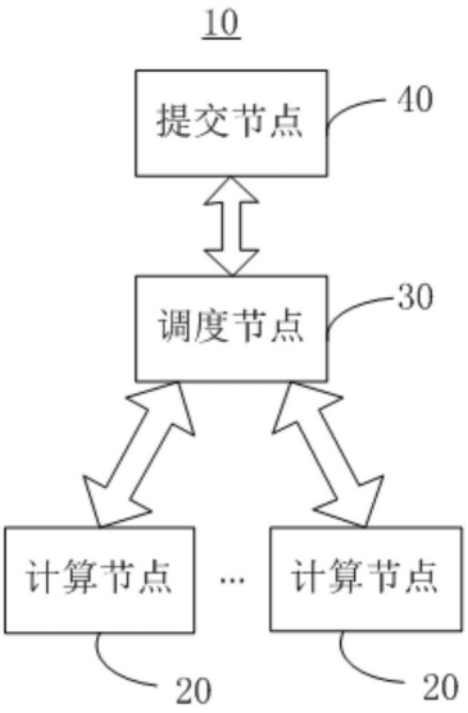

10
提交节点
40
调度节点
30
计算节点
...
计算节点
20
20

图1

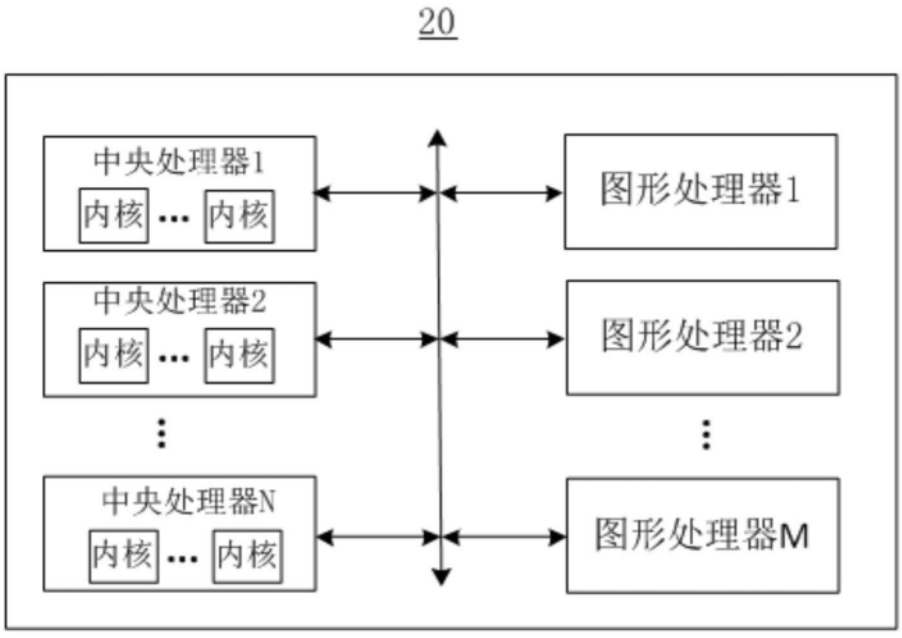

20
中央处理器1
内核
...
内核
中央处理器2
内核
...
内核
中央处理器N
内核
...
内核
图形处理器1
图形处理器2
图形处理器M

图2

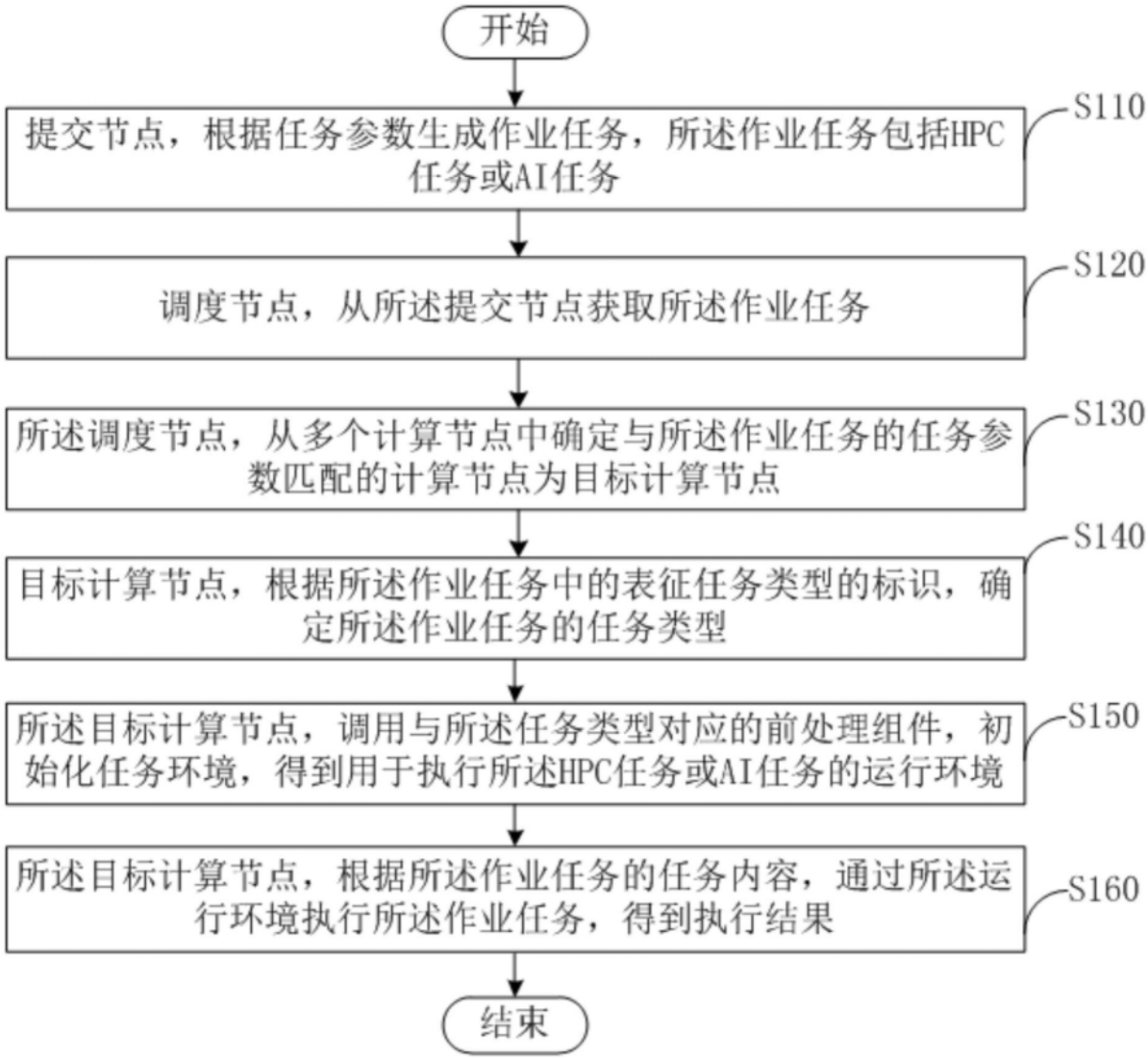

开始
提交节点，根据任务参数生成作业任务，所述作业任务包括HPC任务或AI任务
S110
调度节点，从所述提交节点获取所述作业任务
S120
所述调度节点，从多个计算节点中确定与所述作业任务的任务参数匹配的计算节点为目标计算节点
S130
目标计算节点，根据所述作业任务中的表征任务类型的标识，确定所述作业任务的任务类型
S140
所述目标计算节点，调用与所述任务类型对应的前处理组件，初始化任务环境，得到用于执行所述HPC任务或AI任务的运行环境
S150
所述目标计算节点，根据所述作业任务的任务内容，通过所述运行环境执行所述作业任务，得到执行结果
S160
结束

图3

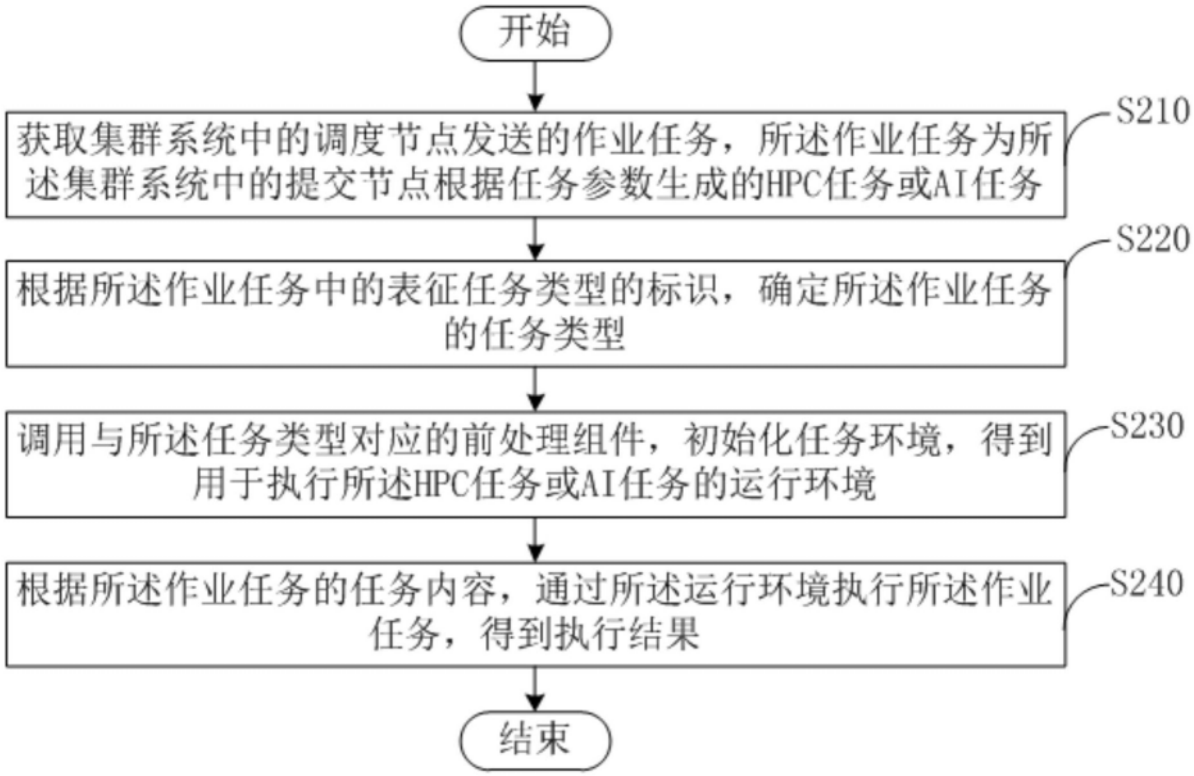

开始
获取集群系统中的调度节点发送的作业任务，所述作业任务为所述集群系统中的提交节点根据任务参数生成的HPC任务或AI任务
S210
根据所述作业任务中的表征任务类型的标识，确定所述作业任务的任务类型
S220
调用与所述任务类型对应的前处理组件，初始化任务环境，得到用于执行所述HPC任务或AI任务的运行环境
S230
根据所述作业任务的任务内容，通过所述运行环境执行所述作业任务，得到执行结果
S240
结束

图4

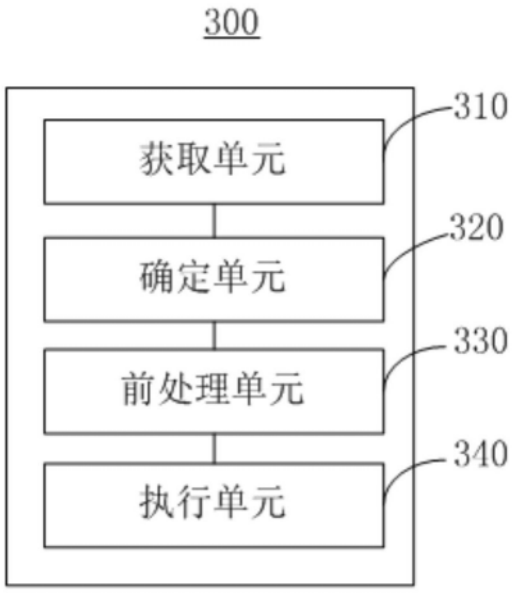

300
获取单元
310
确定单元
320
前处理单元
330
执行单元
340

图5