(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114489963 A
(43)申请公布日 2022.05.13

(21)申请号 202110172700.9

(22)申请日 2021.02.08

(66)本国优先权数据
202011262475.X 2020.11.12 CN

(71)申请人 华为云计算技术有限公司
地址 550025 贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心

(72)发明人 陈普

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 颜晶

(51)Int.Cl.
G06F 9/48(2006.01)
G06F 9/50(2006.01)
G06N 20/00(2019.01)

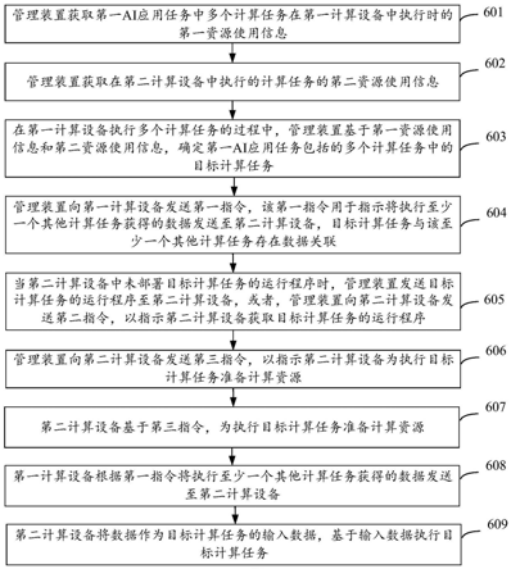
权利要求书4页 说明书22页 附图7页

(54)发明名称

人工智能应用任务的管理方法、系统、设备及存储介质

(57)摘要

本申请公开了一种人工智能应用任务的管理方法、系统、设备及存储介质，属于AI技术领域。AI应用任务涉及多个计算任务，该方法包括：在第一计算设备执行多个计算任务的过程中，管理装置在多个计算任务中，确定待由第二计算设备执行的目标计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联；管理装置向第一计算设备发送第一指令；第一计算设备根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备；第二计算设备将数据作为目标计算任务的输入数据，基于输入数据执行目标计算任务。本申请有利于提高执行AI应用任务的计算设备的资源利用率和/或计算任务的处理效率。



CN 114489963 A

1.一种人工智能AI应用任务的管理方法,其特征在于,AI应用任务涉及多个计算任务,所述方法包括:

在第一计算设备执行所述多个计算任务的过程中,管理装置确定所述多个计算任务中的目标计算任务,其中,所述目标计算任务为待由第二计算设备执行的计算任务,所述目标计算任务与所述AI应用任务涉及的至少一个其他计算任务存在数据关联;

所述管理装置向所述第一计算设备发送第一指令;

所述第一计算设备根据所述第一指令将执行所述至少一个其他计算任务获得的数据发送至所述第二计算设备;

所述第二计算设备将所述数据作为所述目标计算任务的输入数据,基于所述输入数据执行所述目标计算任务。

2.根据权利要求1所述的方法,其特征在于,在所述管理装置确定所述多个计算任务中的目标计算任务之前,所述方法还包括:

所述管理装置获取所述多个计算任务在所述第一计算设备中执行时的第一资源使用信息;

所述管理装置确定所述多个计算任务中的目标计算任务,包括:

所述管理装置基于所述第一资源使用信息确定所述多个计算任务中的目标计算任务。

3.根据权利要求1或2所述的方法,其特征在于,当所述第二计算设备中未部署所述目标计算任务的运行程序时,所述方法还包括:

所述管理装置发送所述目标计算任务的运行程序至所述第二计算设备,或者,所述管理装置向所述第二计算设备发送第二指令,以指示所述第二计算设备获取所述目标计算任务的运行程序。

4.根据权利要求1-3任一项所述的方法,其特征在于,所述方法还包括:

所述管理装置向所述第二计算设备发送第三指令,以指示所述第二计算设备为执行所述目标计算任务准备计算资源。

5.根据权利要求1-4任一项所述的方法,其特征在于,所述管理装置确定所述多个计算任务中的目标计算任务,包括:

当所述第一计算设备中的所述多个计算任务的运行效率不满足预设第一条件,和/或,所述第一计算设备中用于运行所述多个计算任务的资源利用情况不满足预设第二条件时,所述管理装置确定所述多个计算任务中的目标计算任务。

6.根据权利要求2所述的方法,其特征在于,所述方法还包括:所述管理装置获取在所述第二计算设备中执行的计算任务的第二资源使用信息;

所述管理装置确定所述多个计算任务中的目标计算任务,包括:

所述管理装置基于所述第一资源使用信息和所述第二资源使用信息确定所述多个计算任务中的目标计算任务。

7.根据权利要求2或6所述的方法,其特征在于,所述第一资源使用信息包括:所述多个计算任务中至少一个计算任务的运行信息和所述第一计算设备的资源信息。

8.根据权利要求7所述的方法,其特征在于,所述至少一个计算任务的运行信息由以下一个或多个运行参数获得:

所述计算任务在单位时长内的调用次数,所述计算任务的输入数据量,所述计算任务

的输出数据量，所述第一计算设备调用所述计算任务的运行时长，调用所述计算任务的处理器的消耗量，调用所述计算任务的内存的消耗量；

所述第一计算设备的资源信息由以下一个或多个参数获得：所述第一计算设备的内存的额定值和总消耗量，所述第一计算设备处理器内存向所述第一计算设备的AI计算内存传输数据的带宽和带宽的额定值，所述第一计算设备的AI计算内存向所述第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

9.根据权利要求1至8任一项所述的方法，其特征在于，所述管理装置部署在所述第一计算设备、所述第二计算设备或者第三计算设备中，其中，所述第三计算设备与所述第一计算设备和所述第二计算设备通过通信通路连接。

10.根据权利要求9所述的方法，其特征在于，

所述第一计算设备为显卡、AI计算芯片或服务器；

所述第二计算设备为显卡、AI计算芯片或服务器；

所述第三计算设备为显卡、AI计算芯片或服务器。

11.一种人工智能AI系统，其特征在于，所述AI系统包括第一计算设备、第二计算设备和管理装置，

所述管理装置，用于确定所述多个计算任务中的目标计算任务，其中，所述目标计算任务为待由第二计算设备执行的计算任务，所述目标计算任务与所述AI应用任务涉及的至少一个其他计算任务存在数据关联；

所述管理装置，还用于向所述第一计算设备发送第一指令；

所述第一计算设备，用于根据所述第一指令将执行所述至少一个其他计算任务获得的数据发送至所述第二计算设备；

所述第二计算设备，用于将所述数据作为所述目标计算任务的输入数据，基于所述输入数据执行所述目标计算任务。

12.根据权利要求11所述的系统，其特征在于，所述管理装置还用于：

获取所述多个计算任务在所述第一计算设备中执行时的第一资源使用信息；

基于所述第一资源使用信息确定所述多个计算任务中的目标计算任务。

13.根据权利要求11或12所述的系统，其特征在于，当所述第二计算设备中未部署所述目标计算任务的运行程序时，所述管理装置还用于：

发送所述目标计算任务的运行程序至所述第二计算设备，或者，向所述第二计算设备发送第二指令，以指示所述第二计算设备获取所述目标计算任务的运行程序。

14.根据权利要求11-13任一项所述的系统，其特征在于，所述管理装置还用于：

向所述第二计算设备发送第三指令，以指示所述第二计算设备为执行所述目标计算任务准备计算资源。

15.根据权利要求11-14任一项所述的系统，其特征在于，所述管理装置具体用于：

当所述第一计算设备中的所述多个计算任务的运行效率不满足预设第一条件，和/或，所述第一计算设备中用于运行所述多个计算任务的资源利用情况不满足预设第二条件时，确定所述多个计算任务中的目标计算任务。

16.根据权利要求12所述的系统，其特征在于，所述管理装置还用于：

获取在所述第二计算设备中执行的计算任务的第二资源使用信息；

所述管理装置具体用于：

基于所述第一资源使用信息和所述第二资源使用信息确定所述多个计算任务中的目标计算任务。

17.根据权利要求12或16所述的系统，其特征在于，所述第一资源使用信息包括：所述多个计算任务中至少一个计算任务的运行信息和所述第一计算设备的资源信息。

18.根据权利要求17所述的系统，其特征在于，所述至少一个计算任务的运行信息由以下一个或多个运行参数获得：

所述计算任务在单位时长内的调用次数，所述计算任务的输入数据量，所述计算任务的输出数据量，所述第一计算设备调用所述计算任务的运行时长，调用所述计算任务的处理器的消耗量，调用所述计算任务的内存的消耗量；

所述第一计算设备的资源信息由以下一个或多个参数获得：所述第一计算设备的内存的额定值和总消耗量，所述第一计算设备处理器内存向所述第一计算设备的AI计算内存传输数据的带宽和带宽的额定值，所述第一计算设备的AI计算内存向所述第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

19.根据权利要求11至18任一项所述的系统，其特征在于，所述管理装置部署在所述第一计算设备、所述第二计算设备或者第三计算设备中，其中，所述第三计算设备与所述第一计算设备和所述第二计算设备通过通信通路连接。

20.根据权利要求19所述的系统，其特征在于，

所述第一计算设备为显卡、AI计算芯片或服务器；

所述第二计算设备为显卡、AI计算芯片或服务器；

所述第三计算设备为显卡、AI计算芯片或服务器。

21.一种管理装置，其特征在于，所述管理装置包括控制模块和调度模块，

所述控制模块，用于在第一计算设备执行所述多个计算任务的过程中，确定所述多个计算任务中的目标计算任务，其中，所述目标计算任务为待由第二计算设备执行的计算任务，所述目标计算任务与所述AI应用任务涉及的至少一个其他计算任务存在数据关联；

所述调度模块，用于发送第一指令至所述第一计算设备，所述第一指令用于指示所述第一计算设备将执行所述至少一个其他计算任务获得的数据发送至所述第二计算设备中的所述目标计算任务。

22.根据权利要求21所述的装置，其特征在于，所述控制模块还用于：

获取所述多个计算任务在所述第一计算设备中执行时的第一资源使用信息；基于所述第一资源使用信息确定所述多个计算任务中的目标计算任务。

23.根据权利要求21或22所述的装置，其特征在于，当所述第二计算设备中未部署所述目标计算任务的运行程序时，所述调度模块，还用于发送所述目标计算任务的运行程序至所述第二计算设备，或者，向所述第二计算设备发送第二指令，以指示所述第二计算设备获取所述目标计算任务的运行程序。

24.根据权利要求21-23任一项所述的装置，其特征在于，所述调度模块，还用于向所述第二计算设备发送第三指令，以指示所述第二计算设备为执行所述目标计算任务准备计算资源。

25.根据权利要求21-24任一项所述的装置，其特征在于，所述控制模块，具体用于：

当所述第一计算设备中的所述多个计算任务的运行效率不满足预设第一条件，和/或，所述第一计算设备中用于运行所述多个计算任务的资源利用情况不满足预设第二条件时，确定所述多个计算任务中的目标计算任务。

26.根据权利要求22所述的装置，其特征在于，所述控制模块，还用于获取在所述第二计算设备中执行的计算任务的第二资源使用信息；

所述控制模块，具体用于基于所述第一资源使用信息和所述第二资源使用信息确定所述多个计算任务中的目标计算任务。

27.一种电子设备，其特征在于，所述电子设备包括存储器和处理器，所述处理器执行所述存储器中存储的计算机指令时，所述电子设备实现前述权利要求21-26任一项所述装置的功能。

28.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储程序指令，当所述程序指令被电子设备运行时，所述电子设备实现前述权利要求21-26任一项所述装置的功能。

29.一种计算机程序产品，其特征在于，所述计算机程序产品包括程序指令，当所述计算机程序产品中的程序指令被电子设备运行时，所述电子设备实现前述权利要求21-26任一项所述装置的功能。

人工智能应用任务的管理方法、系统、设备及存储介质

[0001] 本公开要求于2020年11月12日提交的申请号为202011262475.X、发明名称为“一种AI系统资源动态调度的方法和设备”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

[0002] 本申请涉及人工智能(artificial intelligence,AI)技术领域，特别涉及一种AI应用任务的管理方法、系统、设备、及存储介质。

背景技术

[0003] 随着人工智能技术的快速发展，越来越多的应用场景中使用人工智能技术解决问题。一个应用场景通常需要解决多个问题，例如，在对交通路口视频的分析识别场景中，需要解决车辆检测、跟踪、车型识别、颜色识别、红绿灯检测、红绿灯状态识别、人检测及跟踪、非机动车检测及跟踪等多个问题。

[0004] 相关技术中，当应用场景所涉及的解决方案包括多个待解决的问题时，一个应用场景下的AI应用任务可以包括多个计算任务，每个计算任务分别解决该应用场景下的部分问题，多个计算任务之间可以互相交互，共同形成该应用场景的一个解决方案。该多个计算任务可以集中部署在单个计算设备上，或者分布式地部署在多个计算设备上。

[0005] 当前，计算设备执行AI应用任务涉及的多个计算任务时，只能按照预先的部署方式进行计算任务的执行。在实际应用场景中，计算任务按照预先部署方式在计算设备中被执行的过程中，可能存在计算设备中的资源利用率较低或者计算任务处理效率较低的情况。

发明内容

[0006] 本申请提供了一种AI应用任务的管理方法、系统、设备、及存储介质，有利于提高执行AI应用任务的计算设备的资源利用率和/或计算任务的处理效率，提高了用于实现AI应用任务的整体系统的运行性能。本申请提供的技术方案如下：

[0007] 第一方面，本申请提供了一种AI应用任务的管理方法。一个AI应用任务可以包括多个计算任务，每个计算任务用于实现解决方案的部分功能。该方法包括：在第一计算设备执行多个计算任务的过程中，管理装置确定多个计算任务中的目标计算任务，其中，目标计算任务为待由第二计算设备执行的计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联；管理装置向第一计算设备发送第一指令；第一计算设备根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备；第二计算设备将数据作为目标计算任务的输入数据，基于输入数据执行目标计算任务。其中，其他计算任务为该AI应用任务中除目标计算任务外的计算任务。计算任务之间存在数据关联是指：一个计算任务的输入数据为已被调用的另一个或多个计算任务的输出数据。

[0008] 通过在第一计算设备执行AI应用任务包括的多个计算任务的过程中，管理装置在

该多个计算任务中确定待在第二计算设备中执行的目标计算任务，然后第一计算设备将执行目标计算任务所需的数据发送至第二计算设备，第二计算设备将第一计算设备发送的数据作为目标计算任务的输入数据，并基于该输入数据执行目标计算任务，能够在第一计算设备执行AI应用任务的过程中，将目标计算任务调整至由第二计算设备执行，能够在执行AI应用任务的过程中对计算任务进行灵活的调度，有利于提高该第一计算设备和第二计算设备的资源利用率和/或计算任务处理效率，提高了用于实现AI应用任务的整体系统的运行性能。

[0009] 在一种可实现方式中，在管理装置确定多个计算任务中的目标计算任务之前，该方法还包括：管理装置获取多个计算任务在第一计算设备中执行时的第一资源使用信息。相应的，管理装置确定多个计算任务中的目标计算任务，包括：管理装置基于第一资源使用信息确定多个计算任务中的目标计算任务。

[0010] 第一资源使用信息用于反映第一AI应用任务中计算任务使用的资源情况和第一计算设备的资源情况。通过获取多个计算任务在第一计算设备中执行时的第一资源使用信息，并基于该第一资源使用信息确定多个计算任务中的目标计算任务，能够根据第一AI应用任务中计算任务和第一计算设备使用资源的实际情况在多个计算任务中筛选出目标计算任务，并对该目标计算任务进行调度，有利于提高该第一计算设备的资源利用率和/或计算任务处理效率。

[0011] 可选地，当第二计算设备中未部署目标计算任务的运行程序时，则在使用第二计算设备执行目标计算任务之前，需要在该第二计算设备中部署该目标计算任务的运行程序。则根据该第二计算设备获取该目标计算任务的运行程序的不同实现方式，该方法还包括：管理装置发送目标计算任务的运行程序至第二计算设备，或者，管理装置向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序。

[0012] 在一种可实现方式中，为了保证第二计算设备执行目标计算任务的运行性能，该方法还包括：管理装置向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0013] 在管理装置确定多个计算任务中的目标计算任务的一种可实现方式中，可以在第一计算设备中的多个计算任务的运行效率不满足预设第一条件，和/或，第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时，管理装置确定多个计算任务中的目标计算任务。

[0014] 通过在第一AI应用任务中计算任务的运行效率和/或第一AI应用任务中计算任务的资源利用情况不满足对应条件时，在第一AI应用任务中的多个计算任务中，确定待由第二计算设备执行的目标计算任务，能够有效提高第一AI应用任务的运行效率和/或资源利用率。

[0015] 管理装置可以根据该管理装置管理的多个计算设备的资源使用情况，对计算设备中执行的计算任务进行管理，以改善包括该多个计算设备的整体系统的运行性能。则该方法还包括：管理装置获取在第二计算设备中执行的计算任务的第二资源使用信息。相应的，管理装置确定多个计算任务中的目标计算任务，包括：管理装置基于第一资源使用信息和第二资源使用信息确定多个计算任务中的目标计算任务。

[0016] 通过根据第一资源使用信息和第二资源使用信息确定由第二计算设备执行目标

计算任务,能够从全局考虑管理装置管理的多个计算设备中运行的AI应用任务的运行效率和资源利用率,能够在第一AI应用任务的整体运行效率较差和/或资源利用情况较差时,对第一AI应用任务中的AI计算任务进行管理,能够以执行调度操作为代价减小整体系统的短木板效应,从而提升管理装置管理的多个计算设备的整体系统的运行性能。

[0017] 在一种可实现方式中,第一资源使用信息包括:多个计算任务中至少一个计算任务的运行信息和第一计算设备的资源信息。

[0018] 可选地,至少一个计算任务的运行信息由以下一个或多个运行参数获得:计算任务在单位时长内的调用次数,计算任务的输入数据量,计算任务的输出数据量,第一计算设备调用计算任务的运行时长,调用计算任务的处理器的消耗量,调用计算任务的内存的消耗量。

[0019] 第一计算设备的资源信息由以下一个或多个参数获得:第一计算设备的内存的额定值和总消耗量,第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽和带宽的额定值,第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

[0020] 在一种可实现方式中,管理装置部署在第一计算设备、第二计算设备或者第三计算设备中,其中,第三计算设备与第一计算设备和第二计算设备通过通信通路连接。

[0021] 其中,第一计算设备为显卡、AI计算芯片或服务器;第二计算设备为显卡、AI计算芯片或服务器;第三计算设备为显卡、AI计算芯片或服务器。

[0022] 第二方面,本申请提供了一种AI系统,该AI系统包括第一计算设备、第二计算设备和管理装置,管理装置,用于确定多个计算任务中的目标计算任务,其中,目标计算任务为待由第二计算设备执行的计算任务,目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联;管理装置,还用于向第一计算设备发送第一指令;第一计算设备,用于根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备;第二计算设备,用于将数据作为目标计算任务的输入数据,基于输入数据执行目标计算任务。

[0023] 可选地,管理装置还用于:获取多个计算任务在第一计算设备中执行时的第一资源使用信息;基于第一资源使用信息确定多个计算任务中的目标计算任务。

[0024] 可选地,当第二计算设备中未部署目标计算任务的运行程序时,管理装置还用于:发送目标计算任务的运行程序至第二计算设备,或者,向第二计算设备发送第二指令,以指示第二计算设备获取目标计算任务的运行程序。

[0025] 可选地,管理装置还用于:向第二计算设备发送第三指令,以指示第二计算设备为执行目标计算任务准备计算资源。

[0026] 可选地,管理装置具体用于:当第一计算设备中的多个计算任务的运行效率不满足预设第一条件,和/或,第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时,确定多个计算任务中的目标计算任务。

[0027] 可选地,管理装置还用于:获取在第二计算设备中执行的计算任务的第二资源使用信息;管理装置具体用于:基于第一资源使用信息和第二资源使用信息确定多个计算任务中的目标计算任务。

[0028] 可选地,第一资源使用信息包括:多个计算任务中至少一个计算任务的运行信息和第一计算设备的资源信息。

[0029] 可选地，至少一个计算任务的运行信息由以下一个或多个运行参数获得：计算任务在单位时长内的调用次数，计算任务的输入数据量，计算任务的输出数据量，第一计算设备调用计算任务的运行时长，调用计算任务的处理器的消耗量，调用计算任务的内存的消耗量。

[0030] 第一计算设备的资源信息由以下一个或多个参数获得：第一计算设备的内存的额定值和总消耗量，第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽和带宽的额定值，第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

[0031] 可选地，管理装置部署在第一计算设备、第二计算设备或者第三计算设备中，其中，第三计算设备与第一计算设备和第二计算设备通过通信通路连接。

[0032] 可选地，第一计算设备为显卡、AI计算芯片或服务器；第二计算设备为显卡、AI计算芯片或服务器；第三计算设备为显卡、AI计算芯片或服务器。

[0033] 第三方面，本申请提供了一种管理装置，管理装置包括控制模块和调度模块，控制模块，用于在第一计算设备执行多个计算任务的过程中，确定多个计算任务中的目标计算任务，其中，目标计算任务为待由第二计算设备执行的计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联；调度模块，用于发送第一指令至第一计算设备，第一指令用于指示第一计算设备将执行至少一个其他计算任务获得的数据发送至第二计算设备中的目标计算任务。

[0034] 可选地，控制模块还用于：获取多个计算任务在第一计算设备中执行时的第一资源使用信息；基于第一资源使用信息确定多个计算任务中的目标计算任务。

[0035] 可选地，当第二计算设备中未部署目标计算任务的运行程序时，调度模块，还用于发送目标计算任务的运行程序至第二计算设备，或者，向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序。

[0036] 可选地，调度模块，还用于向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0037] 可选地，控制模块，具体用于：当第一计算设备中的多个计算任务的运行效率不满足预设第一条件，和/或，第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时，确定多个计算任务中的目标计算任务。

[0038] 可选地，控制模块，还用于获取在第二计算设备中执行的计算任务的第二资源使用信息；控制模块，具体用于基于第一资源使用信息和第二资源使用信息确定多个计算任务中的目标计算任务。

[0039] 可选地，第一资源使用信息包括：多个计算任务中至少一个计算任务的运行信息和第一计算设备的资源信息。

[0040] 可选地，至少一个计算任务的运行信息由以下一个或多个运行参数获得：计算任务在单位时长内的调用次数，计算任务的输入数据量，计算任务的输出数据量，第一计算设备调用计算任务的运行时长，调用计算任务的处理器的消耗量，调用计算任务的内存的消耗量。

[0041] 第一计算设备的资源信息由以下一个或多个参数获得：第一计算设备的内存的额定值和总消耗量，第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽

和带宽的额定值，第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

[0042] 可选地，管理装置部署在第一计算设备、第二计算设备或者第三计算设备中，其中，第三计算设备与第一计算设备和第二计算设备通过通信通路连接。

[0043] 可选地，第一计算设备为显卡、AI计算芯片或服务器；第二计算设备为显卡、AI计算芯片或服务器；第三计算设备为显卡、AI计算芯片或服务器。

[0044] 第四方面，本申请提供了一种电子设备，电子设备包括存储器和处理器，处理器执行存储器中存储的计算机指令时，电子设备实现前述第三方面提供的装置的功能。

[0045] 第五方面，本申请提供了一种计算机可读存储介质，该计算机可读存储介质为非易失性计算机可读存储介质，计算机可读存储介质中存储程序指令，当程序指令被电子设备运行时，电子设备实现前述第三方面提供的装置的功能。

[0046] 第六方面，本申请提供了一种包含指令的计算机程序产品，当计算机程序产品在计算机上运行时，使得计算机实现前述第三方面提供的装置的功能。

附图说明

[0047] 图1是本申请实施例提供的一种AI应用任务的执行逻辑的示意图；

[0048] 图2是本申请实施例提供的一种AI应用任务的管理方法涉及的AI系统的示意图；

[0049] 图3是本申请实施例提供的另一种AI应用任务的管理方法涉及的AI系统的示意图；

[0050] 图4是本申请实施例提供的一种管理装置的功能能够由云服务提供商在云平台抽象成一种管理云服务的示意图；

[0051] 图5是本申请实施例提供的又一种AI应用任务的管理方法涉及的AI系统的示意图；

[0052] 图6是本申请实施例提供的一种AI应用任务的管理方法的流程图；

[0053] 图7是本申请实施例提供的一种AI应用任务的管理方法采用功能模块实现的流程图；

[0054] 图8是本申请实施例提供的一种AI应用任务的执行逻辑的示意图；

[0055] 图9是本申请实施例提供的一种在将第一AI应用任务中的目标计算任务调度至第二计算设备中执行后，第一AI应用任务和第二AI应用任务的执行逻辑的示意图；

[0056] 图10是本申请实施例提供的另一种在将第一AI应用任务中的目标计算任务调度至第二计算设备中执行后，第一AI应用任务和第二AI应用任务的执行逻辑的示意图；

[0057] 图11是本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

[0058] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0059] 随着AI技术的发展，越来越多的应用场景中使用AI技术解决问题。可以通过AI应用任务来实现针对某一应用场景的特定目标的解决方案，例如：针对交通场景，可以提供一套实时视频分析解决方案，以获得该交通场景下的车辆轨迹、车辆属性和该交通场景下的

实时红绿灯状态。一个AI应用任务表示针对一个应用场景的特定目标提供的解决方案，AI应用任务可以由计算设备执行，以实现特定的功能。通常一个AI应用任务涉及多个计算任务，每个计算任务用于实现解决方案的部分功能，因此，也可以认为一个AI应用任务包括多个计算任务。AI应用任务中一些计算任务之间可能存在数据关联，例如：一个计算任务的输入数据为已被调用的另一个或多个计算任务的输出数据。因此，通常一个AI应用任务中的一些计算任务的执行时间存在先后顺序。

[0060] 依然以对交通场景的视频流分析为例，如图1所示，采用AI应用任务能够对交通视频流进行分析。该AI应用任务包括多个计算任务，该多个计算任务分别用于实现视频解码、车辆目标检测、车辆目标跟踪、车辆属性检测、红绿灯检测、红绿灯状态检测和数据输出。每个计算任务采用图1中一个圆圈表示，该图1中的箭头表示计算任务之间的数据流向。在执行AI应用任务时，计算设备针对同一视频流根据箭头指示的路径依次执行计算任务，以获得针对该视频流程的分析结果。

[0061] 在一种实现方式中，多个计算任务采用集中式的方式部署，也即是，采用同一个计算设备执行AI应用任务包括的多个计算任务。

[0062] 在另一种实现方式中，多个计算任务采用分布式的方式部署。例如，多个计算任务分别部署在多个计算设备上，每个计算设备通过执行其上部署的计算任务实现对应的功能，多个计算设备协同实现包括多个计算任务的AI应用任务的功能。

[0063] 应理解，本申请中的计算设备可以是AI芯片(例如：包括中央处理器(central processing unit，CPU)、图形处理器(graphics processing unit，GPU)、现场可编程门阵列(field programmable gate array，FPGA)或专用集成电路(application-specific integrated circuit，ASIC)的芯片)、显卡、服务器或者虚拟机等。

[0064] 但是，在以上两种实现方式中，相关技术在完成AI应用任务的部署后，仅能够按照固定的部署方式完成对应的计算任务，导致部署有AI应用任务的整体系统的资源利用率较差。例如，在将AI应用任务部署在第一计算设备上时，第一计算设备执行AI应用任务中的各计算任务的过程中，可能存在一些计算任务在执行的过程中所需的算力较小，但是所需的计算内存的占用较大，此时，由于能够提供给其他计算任务的内存不足，会导致在执行这类型计算任务的过程中该计算设备中的其他算力无法用于执行其他计算任务，使得该计算设备的资源利用率不足。再例如，第一计算设备执行AI应用任务中的各计算任务的过程中，可能存在一些计算任务在执行的过程中所需的算力较大，但是所需的计算内存的占用较小，此时，由于能提供给其他计算任务的算力不足，会导致在执行这类型计算任务的过程中该计算设备中的计算内存存在浪费，也使得该计算设备的资源利用率不足。

[0065] 其中，计算设备的资源指执行计算任务所需的计算资源，计算设备中包括算力资源和内存资源等。例如，通用处理器(如CPU)、AI处理器(如GPU)提供的算力资源、处理器内存(如CPU内存)和AI计算内存(如GPU内存)提供的内存资源。其中，处理器内存为向通用处理器分配的内存，AI计算内存为向AI处理器分配的内存，当计算任务由AI处理器执行时，需要占用AI计算内存。

[0066] 应理解，本申请中计算任务在执行过程中所需的算力表示执行该计算任务时单位时间内需使用计算设备中的计算资源的数量或数量占比，或者，计算任务在单位时间内使用计算资源的时间或时间占比。例如：当计算设备为包括GPU和内存的显卡时，计算任务执

行过程中所需的算力可以用计算任务每秒钟所占用GPU计算资源的实际耗时表示，如：每秒钟(1秒＝1000毫秒)占用GPU的实际耗时为10毫秒，且每秒调用1次计算任务时，则该计算任务所需的算力可以表示为(10毫秒/1000毫秒)×1×100％＝1％。

[0067] 还应理解，本申请中计算任务执行过程中所需的计算内存表示执行计算任务时所占用的计算设备内的内存量或者内存比例。仍以计算设备为包括GPU和内存的显卡为例，计算任务执行过程中所需的计算内存可以用计算任务运行所需占用的显卡中的内存(也称为显存)与该显卡的内存额定量的比值来表示。

[0068] 本申请实施例提供了一种AI应用任务的管理方法。该方法包括：在第一计算设备执行一个AI应用任务涉及的多个计算任务的过程中，管理装置在该多个计算任务中确定待在第二计算设备中执行的目标计算任务，然后第一计算设备将执行目标计算任务所需的数据发送至第二计算设备，第二计算设备将第一计算设备发送的数据作为目标计算任务的输入数据，并基于该输入数据执行目标计算任务。其中，执行该目标计算任务所需的数据包括：目标计算任务与该AI应用任务涉及的至少一个其他计算任务存在数据关联的数据。该其他计算任务为该AI应用任务中除目标计算任务外的计算任务。计算任务之间存在数据关联是指：一个计算任务的输入数据为已被调用的另一个或多个计算任务的输出数据。

[0069] 通过该AI应用任务的管理方法，能够在第一计算设备执行AI应用任务的过程中，将目标计算任务调整至由第二计算设备执行，也即，该方法能够在执行AI应用任务的过程中对计算任务进行灵活的调度，有利于提高该第一计算设备和第二计算设备的资源利用率和/或计算任务处理效率，提高了用于实现AI应用任务的整体系统的运行性能。

[0070] 图2是本申请实施例提供的一种AI应用任务的管理方法涉及的AI系统的示意图。如图2所示，该AI系统包括：第一计算设备10、第二计算设备20和第三计算设备30，第三计算设备30与第一计算设备10和第二计算设备20通过通信通路连接。可选地，该AI系统还可以包括更多计算设备。下面以图2所示的AI系统为例，对本申请实施例提供的一种AI应用任务的管理方法涉及的AI系统的工作原理进行说明。

[0071] 如图2所示，第一计算设备10中部署有第一AI应用任务101。第二计算设备20中部署有第二AI应用任务201。第三计算设备30中部署有管理装置301。管理装置301用于在第一AI应用任务的执行过程中，在第一AI应用任务101包括的多个计算任务中，确定待由第二计算设备20执行的目标计算任务，并向第一计算设备10发送第一指令。相应的，第一计算设备10还用于根据该第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备20，该至少一个其他计算任务与目标计算任务存在数据关联。并且，第二计算设备20还用于基于执行至少一个其他计算任务获得的数据执行目标计算任务。根据该图2可知，在该图2所示的AI系统中，管理装置可以单独部署在第三计算设备30上。

[0072] 可选地，第三计算设备30、第一计算设备10和第二计算设备20均可以为显卡、AI计算芯片、物理机、裸金属服务器或云服务器。例如：第三计算设备30、第一计算设备10和第二计算设备20均可以为AI计算芯片。当第一计算设备10和第二计算设备20为显卡或AI计算芯片时，第一计算设备10和第二计算设备20可以分别部署在不同的主机上，或者，部署在同一主机上。当第三计算设备30、第一计算设备10和第二计算设备20为显卡或AI计算芯片时，该第三计算设备30可以相对于第一计算设备10和第二计算设备20部署在单独的主机上，或者，该第三计算设备30可以与第一计算设备10和第二计算设备20中的部分或全部部署在同

一主机上。

[0073] 图3是本申请实施例提供的另一种AI应用任务的管理方法涉及的AI系统的示意图。如图3所示，该AI系统包括第一计算设备10和第二计算设备20，第一计算设备10和第二计算设备20通过通信通路连接。可选地，该AI系统还可以包括更多计算设备。下面以图3所示的AI系统为例，对本申请实施例提供的另一种AI应用任务的管理方法涉及的AI系统的工作原理进行说明。

[0074] 如图3所示，第一计算设备10中部署有第一AI应用任务101。第二计算设备20中部署有第二AI应用任务201。并且，第一计算设备10中还部署有管理装置103。或者，管理装置103也可以部署在第二计算设备20中。第一计算设备10、第二计算设备20和管理装置103的功能可相应参考图2所示的AI系统中的相关说明。根据该图3可知，在该图3所示的AI系统中，管理装置103部署在用于执行AI应用任务的计算设备中。当管理装置部署在用于执行AI应用任务的计算设备中时，能够提升管理装置与该计算设备中用于执行AI应用任务的装置之间的通信效率，减小网络等外部因素的影响。

[0075] 可选地，第一计算设备10和第二计算设备20可以为显卡、AI计算芯片、物理机、裸金属服务器或云服务器。第一计算设备10和第二计算设备20为显卡或AI计算芯片时，第一计算设备10和第二计算设备20可以分别部署在不同的主机上，或者，第一计算设备10和第二计算设备20可以部署在同一主机上。

[0076] 在一种可实现方式中，AI系统可以为云上系统，其可以利用云上的计算资源为用户提供云服务。相应的，第一计算设备10、第二计算设备20、第三计算设备30可以是云平台中的计算设备。例如：云平台中的服务器的显卡、AI计算芯片或者主机(例如：云服务器)。第一计算设备10和第二计算设备20执行的AI应用任务可以是拥有云平台资源的云服务提供商部署在云平台上的计算设备中并提供给用户使用的，也可以是AI算法提供商在云平台中的计算设备进行部署并提供给用户使用的。

[0077] 此时，如图4所示，管理装置301的功能能够由云服务提供商在云平台1抽象成一种管理云服务提供给用户。该管理云服务能够通过在第一计算设备执行AI应用任务张多个计算任务的过程中，在该多个计算任务中确定待由第二计算设备执行的目标计算任务，并向第一计算设备发送第一指令，以实现对目标计算任务的管理。其中，图4为管理装置301部署在第三计算设备30中的示意图。

[0078] 可选地，如图4所示，第一AI应用任务101和/或第二AI应用任务201的功能也能够由云服务商在云平台1中抽象成一种AI业务云服务，该AI业务云服务能够通过执行AI应用任务实现用户的AI业务。该AI业务云服务与管理云服务能够配合使用。其中，图4为第一计算设备10用于执行第一AI应用任务，第二计算设备20用于执行第二AI应用任务的示意图。

[0079] 例如，用户购买AI业务云服务后，云平台可以为用户购买的AI业务云服务自动提供管理云服务。即云平台在向用户提供AI业务云服务的过程中，监测该云服务的服务质量，并在服务质量较差时，通过运行本申请实施例提供的AI应用任务的管理方法，对AI应用任务中的计算任务进行调度，向用户购买的AI业务云服务提供管理云服务，以保证用户购买的AI业务云服务的服务质量。

[0080] 又例如，用户在购买该AI业务云服务时可以选择是否购买管理云服务。当用户购买AI业务云服务，且购买管理云服务时，管理云服务用于监测AI业务云服务的服务质量，并

在服务质量较差时，通过运行本申请实施例提供的AI应用任务的管理方法，对AI应用任务中的计算任务进行调度，向用户购买的AI业务云服务提供管理云服务，以保证用户购买的AI业务云服务的服务质量。

[0081] 在另一种可能实现中，管理云服务可以为云平台提供的云服务中的一项独立的云服务。即用户可以在云平台独立地购买该管理云服务。用户利用其它平台提供的资源运行AI应用任务时，可以仅在云平台购买管理云服务，以通过该管理云服务对其它资源中的AI应用任务中的计算任务进行调度。

[0082] 需要说明的是，在本申请实施例中，云平台1可以是中心云的云平台、边缘云的云平台或包括中心云和边缘云的云平台，本申请实施例对其不做具体限定。并且，当部署有管理装置的计算设备和用于执行AI应用任务的计算设备均部署在云平台中时，部署有管理装置的计算设备和用于执行AI应用任务的计算设备可以部署在同一朵云上，或者部署在不同的云上。例如，用于执行AI应用任务的计算设备部署在中心云上，部署有管理装置的计算设备部署在边缘云上。

[0083] 在一种可实现方式中，本申请实施例提供的AI应用任务的管理方法可以通过多个功能模块协同实现。下面以图2所示的AI系统为例，对通过多个功能模块实现该方法过程进行说明。

[0084] 如图5所示，在本申请实施例提供的AI应用任务的管理方法中，管理装置301的功能可以通过调度模块3011和控制模块3012实现，第一计算设备10的功能通过第一采集模块102、第一任务调度执行模块103和第一资源调度执行模块104实现，第二计算设备20的功能通过第二采集模块202、第二任务调度执行模块203和第二资源调度执行模块204实现。其中，该图5是AI应用任务包括5个计算任务的示意图，图5中黑色圆点表示计算任务。各个功能模块的作用如下：

[0085] 第一采集模块102和第二采集模块202用于在执行AI应用任务的过程中，采集该AI应用任务的运行参数，并向调度模块3011发送运行参数，或者，对运行参数进行处理得到资源使用信息，并向调度模块3011发送资源使用信息。其中，运行参数为用于反映AI应用任务中各个计算任务对资源的使用情况和用于执行AI应用任务的计算设备对资源的使用情况的基本参数。资源使用信息为对运行参数进行处理后得到的信息，该资源使用信息用于反映AI应用任务中计算任务使用的资源情况和用于执行AI应用任务的计算设备对资源的使用情况。例如，运行参数为AI应用任务中各个计算任务在执行过程中调用CPU的时间信息和第一计算设备调用CPU的时间信息，资源使用信息为根据运行参数得到的各个计算任务在执行过程中对CPU的调用频率、调用时长和消耗量等信息，及第一计算设备对CPU的调用频率、调用时长和消耗量等信息。

[0086] 调度模块3011用于基于接收到的运行参数和/或资源使用信息，向控制模块3012提供资源使用信息。

[0087] 控制模块3012用于基于资源使用信息，在AI应用任务包括的多个计算任务中确定待由第二计算设备执行的目标计算任务，并向调度模块3011发送指示由第二计算任务执行目标计算任务的通知，向第一任务调度执行模块103发送第一指令。

[0088] 相应的，调度模块3011还用于基于指示由第二计算任务执行目标计算任务的通知，生成资源调度策略，向第二资源调度执行模块204提供该资源调度策略。该资源调度策

略用于指示第二资源调度执行模块204为执行目标计算任务准备计算资源。可选地，调度模块3011还用于向第一资源调度模块1023发送资源调度策略。该资源调度策略用于指示第一资源调度模块1023对执行目标计算任务所需的资源进行资源回收。

[0089] 第一任务调度执行模块103用于基于接收到的第一指令，将执行至少一个其他计算任务获得的数据发送至第二计算设备，以便于第二计算设备根据该数据执行目标计算任务。第二任务调度执行模块203的功能请相应参考第一任务调度执行模块103的功能。

[0090] 第一资源调度执行模块104和第二资源调度执行模块204用于基于资源调度策略，对资源进行调度，以便于计算设备利用调度的资源执行计算任务。

[0091] 下面对本申请实施例提供的一种AI应用任务的管理方法的实现过程进行说明。如图6所示，该AI应用任务的管理方法的实现过程包括以下步骤：

[0092] 步骤601、管理装置获取第一AI应用任务中多个计算任务在第一计算设备中执行时的第一资源使用信息。

[0093] 第一计算设备执行第一AI应用任务(为便于区分，下文将第一计算设备执行的AI应用任务称为第一AI应用任务)的过程中，该第一计算设备可以获取运行参数，并向管理装置提供该运行参数，使得管理装置根据该运行参数得到该第一资源使用信息。或者，第一计算设备获取运行参数后，根据运行参数进行处理，得到第一资源使用信息，并向管理装置提供该第一资源使用信息。其中，第一AI应用任务包括一个或多个计算任务，每个计算任务用于实现解决方案的部分功能，且每个计算任务实现的功能可以通过执行一种或多种算法实现。

[0094] 在一种可实现方式中，当本申请实施例提供的AI应用任务的管理方法通过图5所示的AI系统实现时，第一计算设备中设置有第一采集模块102，第三计算设备中设置有调度模块3011。如图7所示，在第一计算设备执行第一AI应用任务的过程中，该第一采集模块102能够获取运行参数，并向调度模块3011提供该运行参数，使得该调度模块3011基于该运行参数获取该第一资源使用信息。其中，运行参数可以为用于反映第一AI应用任务中各个计算任务使用的资源情况和第一计算设备对资源的使用情况的基本参数。第一资源使用信息为对运行参数进行处理后得到的信息，该第一资源使用信息用于反映第一AI应用任务中计算任务使用的资源情况和第一计算设备对资源的使用情况。例如，运行参数为AI应用任务中各个计算任务在执行过程中，调用CPU的时间信息和第一计算设备调用CPU的时间信息，第一资源使用信息为根据运行参数得到的各个计算任务在执行过程中，对CPU的调用频率、调用时长和消耗量等信息，及第一计算设备对CPU的调用频率、调用时长和消耗量等信息。

[0095] 可选地，第一资源使用信息包括：第一AI应用任务中的至少一个计算任务的运行信息和第一计算设备的资源信息。

[0096] 其中，至少一个计算任务的运行信息可以由以下一个或多个运行参数获得：计算任务在单位时长内的调用次数；和，计算任务的输入数据量；和，计算任务的输出数据量；和，第一计算设备调用计算任务的运行时长；和，第一计算设备调用计算任务的处理器的消耗量；和，第一计算设备调用计算任务的内存的消耗量。其中，第一计算设备调用计算任务的内存的消耗量包括：第一计算设备调用计算任务的处理器内存和/或AI计算内存的消耗量。

[0097] 第一计算设备的资源信息可以由以下一个或多个参数获得：第一计算设备的内存

的额定值和总消耗量;和,第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽和带宽的额定值;和,第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。第一计算设备的内存的额定值和总消耗量包括:第一计算设备的处理器内存的额定值和总消耗量,和/或,第一计算设备的AI计算内存的额定值和总消耗量。

[0098] 其中,调用计算任务的处理器的消耗量通过执行调用使用的处理器的核数和时长表示。例如,通过执行调用时对单核的占用时长表示,如用于执行调用的处理器的消耗量为一个核10毫秒的消耗。内存的消耗量通过调用计算任务时,计算任务使用的内存的范围表示。例如,某次调用计算任务时,AI计算内存的消耗量为10千兆字节(gigabyte,GB)。相应的,第一资源使用信息可以为内存占用率和/或算力占比等。管理装置可以采用预置的算法根据运行参数计算得到资源使用信息。

[0099] 例如,假设第一计算设备执行第一AI应用任务的过程中时,第一采集模块102获取的运行参数包括:对一路数据流执行某一项计算任务时,每秒调用该计算任务的平均次数为0.5次;每次调用该计算任务时,计算任务的输入数据量为100千字节(kilobyte,KB);每次调用该计算任务时,计算任务的输出数据量为1KB;每次调用该计算任务时,第一计算设备调用计算任务的运行时长为10毫秒;每次调用该计算任务时,AI计算内存的消耗量为500至600兆字节(megabyte,MB);计算设备中所有通用处理器使用的处理器内存的总消耗量为7.5GB;计算设备中所有计算任务使用的AI计算内存的总消耗量为7.5GB;计算设备的处理器内存向计算设备的AI计算内存传输数据的带宽为1千兆字节每秒(GB/S),计算设备的AI计算内存向计算设备的处理器内存传输数据的带宽为1GB/S。

[0100] 第一采集模块102将上述运行参数发送至调度模块3011后,调度模块3011可以根据上述运行参数采用预置的算法计算得到第一资源使用信息。其中,当计算任务执行过程中所需的算力用计算任务每秒钟所占用GPU计算资源的实际耗时表示时,由于每秒调用该计算任务的平均次数为0.5次,每次调用该计算任务时,第一计算设备调用计算任务的运行时长为10毫秒,且1秒等于1000毫秒,则该计算任务执行过程中所需的算力可以表示为(10毫秒/1000毫秒)×0.5×100%=0.5%。当计算任务执行过程中所需的AI计算内存用计算任务运行所需占用的AI计算内存与第一计算设备中所有计算任务使用的AI计算内存的总消耗量的比值来表示时,由于每次调用该计算任务时,AI计算内存的消耗量为500至600MB,计算设备中所有计算任务使用的AI计算内存的消耗量为7.5GB,且1GB=1024MB,则按照AI计算内存的消耗量为600MB计算,可得计算任务执行过程中所需的AI计算内存可以表示为(600MB/(7.5GB×1024))×100%=7.8%。即第一资源使用信息包括:该计算任务在1秒内占用的AI算力占比为0.5%,占用的AI计算内存占比为7.8%。

[0101] 可选地,第一计算设备获取运行参数的实现方式包括:基于AI计算开发使用的框架进行收集,或者,通过特定应用程序接口(application program interface,API)查询或统计得到,或者,也可以由用于实现第一AI应用任务的应用程序自身采集并上报。并且,对于一些消耗比较固定的功能模块,其运行参数可以是固定值,每次在需要获取该运行参数时,可以将预先配置的固定值用作该运行参数,而无需通过采集方式获取。例如,当视频流分析的对象为分辨率为指定值的视频流时,视频解码流所需的消耗是固定的,则可以将视频解码流对应的运行参数的取值统一设置为固定消耗值,在需要获取运行参数时,将该固

定消耗值确定为视频解码流对应的运行参数的取值。其中，该固定消耗值可以通过统计方式确定。

[0102] 需要说明的是，获取第一资源使用信息的操作可以由管理装置自动地执行，或管理装置在人为触发下执行。例如，管理装置可以根据预设时间周期，在第一AI应用任务的执行过程中，周期性地自动获取第一AI应用任务在第一计算设备中执行的第一资源使用信息。又例如，管理装置的系统维护人员或者用户可以在有调度需求时，向管理装置发送触发指令，使得管理装置在该触发指令的指示下，获取第一AI应用任务在第一计算设备中执行的第一资源使用信息。

[0103] 并且，当本申请实施例提供的AI应用任务的管理方法由云服务抽象成一种管理云服务提供给用户时，用户在购买管理云服务时，可以根据需求配置获取第一资源使用信息的操作是自动地执行还是在人为触发下执行。

[0104] 步骤602、管理装置获取在第二计算设备中执行的计算任务的第二资源使用信息。

[0105] 管理装置可以根据该管理装置管理的多个计算设备的资源使用情况，对计算设备中执行的计算任务进行管理，以改善包括该多个计算设备的整体系统的运行性能。并且，该多个计算设备中除第一计算设备外的其他计算设备中也可以执行AI应用任务。则本申请实施例提供的AI应用任务的管理方法还包括：管理装置获取在第二计算设备中执行的计算任务的第二资源使用信息。该第二计算设备为管理装置管理的多个计算设备中除第一计算设备外的任一计算设备。且该步骤602的实现过程请相应参考步骤601的实现过程，此处不再赘述。

[0106] 需要说明的是，该步骤602为可选步骤。在执行AI应用任务的管理方法时，可以根据应用需求确定是否执行该步骤602。当执行该步骤602时，能够考虑该管理装置管理的多个计算设备的资源使用情况，对计算任务进行管理，能够有效保证整体系统的运行性能。

[0107] 应理解，在一些实现方式中，第二计算设备执行的第二AI应用任务可以与第一计算设备执行的第一AI应用任务相同，也即第一AI应用任务和第二AI应用任务中包括的各计算任务均相同。例如：图1为第一AI应用任务和第二AI应用任务的逻辑原理图，第一AI应用任务和第二AI应用任务均用于对交通场景的视频流进行分析，第一AI应用任务和第二AI应用任务均包括多个计算任务，且该多个计算任务分别用于对视频流进行视频解码、车辆目标检测、车辆目标跟踪、车辆属性检测、红绿灯检测、红绿灯状态检测和数据输出。

[0108] 在另一些实现方式中，第二计算设备执行的第二AI应用任务也可以与第一计算设备执行的第一AI应用任务不同。其中，该不同可以指第二AI应用任务包括的多个计算任务部分不同或全部不同。例如，第一计算设备执行的第一AI应用任务包括5个计算任务，该5个计算任务分别为视频解码、车辆目标检测、车辆目标跟踪、车辆属性检测和数据输出，第二计算设备执行的第二AI应用任务包括3个计算任务，该3个计算任务分别为红绿灯检测、红绿灯状态检测和数据输出，可知该第一AI应用任务中的计算任务与第二AI应用任务中的AI计算任务部分相同。又例如，第一计算设备执行的第一AI应用任务包括4个计算任务，该4个计算任务分别为视频解码、车辆目标检测、车辆目标跟踪和车辆属性检测，第二计算设备执行的第二AI应用任务包括3个计算任务，该3个计算任务分别为红绿灯检测、红绿灯状态检测和数据输出，可知第一AI应用任务中的计算任务与第二AI应用任务中的AI计算任务全部相同。

[0109] 步骤603、在第一计算设备执行多个计算任务的过程中，管理装置基于第一资源使用信息和第二资源使用信息，确定第一AI应用任务包括的多个计算任务中的目标计算任务。

[0110] 管理装置获取第一资源使用信息和第二资源使用信息后，可以根据该第一资源使用信息和第二资源使用信息，对AI应用任务中的计算任务进行管理决策，在AI应用任务包括的多个计算任务中确定目标计算任务。其中，目标计算任务为待由第二计算设备执行的计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联。可选地，确定目标计算任务的输出结果可以采用图结构表示。应理解，在另一些实现方式中，当不执行步骤602时，该步骤603的实现过程为：管理装置基于第一资源使用信息确定第一AI应用任务包括的多个计算任务中的目标计算任务。

[0111] 在根据第一资源使用信息和第二资源使用信息确定目标计算任务的一种可实现方式中，管理装置可以根据第一资源使用信息，确定第一AI应用任务中的多个计算任务的运行效率和/或第一计算设备中用于运行该多个计算任务的资源利用情况，并在第一计算设备中的多个计算任务的运行效率不满足预设第一条件，和/或，第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时，管理装置确定多个计算任务中的目标计算任务。

[0112] 其中，第一条件和第二条件可以根据应用需求进行设置。如第一条件可以为第一AI应用任务中计算任务的运行效率达到参考效率阈值，第二条件可以为第一AI应用任务中计算任务对第一资源的资源利用率与第一AI应用任务中计算任务对第二资源的资源利用率的差值小于参考差值阈值。例如，第二条件可以为第一AI应用任务中计算任务的AI算力占比与AI计算内存占比差值小于参考差值阈值。并且，根据分析结果确定第一AI应用任务中待调度的目标计算任务时，也可以根据第一AI应用任务中每个计算任务的运行效率和/或资源利用情况确定。例如，当第一AI应用任务中某个计算任务的AI算力占比与AI计算内存占比的差值大于指定差值阈值时，将该计算任务确定为目标计算任务。另外，确定目标AI计算任务时，还可以参考其他因素。例如，当第一AI应用任务中某个计算任务的AI算力占比与AI计算内存占比的差值大于指定差值阈值，且该第一计算设备的内存拷贝带宽和网络带宽等资源能够支持将该计算任务调度至由第二计算设备执行时，将该计算任务确定为目标计算任务。

[0113] 通过在第一AI应用任务中计算任务的运行效率和/或第一AI应用任务中计算任务的资源利用情况不满足对应条件时，在第一AI应用任务中的多个计算任务中，确定待由第二计算设备执行的目标计算任务，能够有效提高第一AI应用任务的运行效率和/或资源利用率。

[0114] 可选地，管理装置确定由第二计算设备执行目标计算任务的过程，也可以结合第一计算设备和第二计算设备的执行计算任务的运行效率和资源利用情况执行。例如，在第二计算设备中的多个计算任务的运行效率也不满足第一条件，和/或，第二计算设备中用于运行多个计算任务的资源利用情况也不满足第二条件时，管理装置确定由第二计算设备执行目标计算任务。又例如，管理装置可以在目标计算任务对第一资源的资源利用率小于目标计算任务对第二资源的资源利用率，第二AI应用任务中计算任务对第二资源的资源利用率小于第二AI应用任务中计算任务对第一资源的资源利用率时，确定由第二计算设备执行

目标计算任务。

[0115] 通过根据第一资源使用信息和第二资源使用信息确定由第二计算设备执行目标计算任务，能够从全局考虑管理装置管理的多个计算设备中运行的AI应用任务的运行效率和资源利用率，能够在第一AI应用任务的整体运行效率较差和/或资源利用情况较差时，对第一AI应用任务中的AI计算任务进行管理，能够以执行调度操作为代价减小整体系统的短木板效应，从而提升管理装置管理的多个计算设备的整体系统的运行性能。

[0116] 在一种可实现方式中，当本申请实施例提供的AI应用任务的管理方法通过图5所示的AI系统实现时，第三计算设备中设置有调度模块3011和控制模块3012，如图7所示，调度模块3011获取第一资源使用信息和第二资源使用信息后，将该第一资源使用信息和第二资源使用信息发送至控制模块3012，该控制模块3012根据该第一资源使用信息和第二资源使用信息，在第一AI应用任务中的多个计算任务中，确定待由第二计算设备执行的目标计算任务。

[0117] 例如，假设第一计算设备执行第一AI应用任务，该第一AI应用任务的执行逻辑图请参考图8，该第一AI应用任务包括图8所示的1至5个计算任务，且该第一AI应用任务运行在第一计算设备的过程中的第一资源使用信息包括：第一AI应用任务中第5个计算任务在1秒内占用的AI算力占比为0.5％，占用的AI计算内存占比为7.8％。第二计算设备执行第二AI应用任务，该第二AI应用任务的执行逻辑图请继续参考图8，该第二AI应用任务包括图8所示的1至5个计算任务，且该第二AI应用任务运行在第二计算设备的过程中的第二资源使用信息包括：第二AI应用任务中5个计算任务在1秒内占用的AI算力占比约为8％，占用的AI计算内存占比约为0.5％。

[0118] 控制模块3012根据该第一资源使用信息和第二资源使用信息，可知该第一计算设备执行的第5个计算任务的AI计算内存占比大于AI算力占比，且两者的差值大于参考差值阈值，第二计算设备执行的第二AI应用任务的AI算力占比大于AI计算内存占比，且两者的差值也大于参考差值阈值，第一计算设备和第二计算设备均无法较好地利用资源。控制模块3012考虑到第5个计算任务完成一次输出的数据量相对于拷贝总带宽较有限，认为可以将第5个计算任务的计算带宽分布到第二计算设备上，则控制模块3012根据该分析结果确定将该第一AI应用任务中的第5个计算任务调度至第二计算设备中执行。

[0119] 步骤604、管理装置向第一计算设备发送第一指令，该第一指令用于指示将执行至少一个其他计算任务获得的数据发送至第二计算设备，目标计算任务与该至少一个其他计算任务存在数据关联。

[0120] 管理装置确定目标计算任务后，可向第一计算设备发送第一指令，以使得该第一计算设备将执行目标计算任务所需的数据至第二计算设备，以便于第二计算设备根据该数据执行目标计算任务。其中，指定目标计算任务所需的数据包括执行该至少一个其他计算任务获得的数据。

[0121] 其中，管理装置向第一计算设备发送第一指令指管理装置将第一指令发送至该第一计算设备中用于向第二计算设备发送执行目标计算任务所需的数据的功能模块。当管理装置部署在第一计算设备中时，管理装置可以为第一计算设备中的具备前述功能的虚拟装置，则该发送动作为第一计算设备内部不同功能模块之间的发送动作。当管理装置部署在第三计算设备中时，该发送动作为不同计算设备间的发送动作。

[0122] 在一种可实现方式中，当本申请实施例提供的AI应用任务的管理方法通过图5所示的AI系统实现时，第三计算设备中设置有控制模块3012，第一计算设备中设置有第一任务调度执行模块103。如图7所示，控制模块3012确定目标计算任务后，可以向第一任务调度执行模块103发送第一指令，以便于第一任务调度执行模块103根据该第一指令将执行目标计算任务所需的数据发送至第二计算设备。此时，管理装置向第一计算设备发送第一指令是不同设备间的发送动作。

[0123] 步骤605、当第二计算设备中未部署目标计算任务的运行程序时，管理装置发送目标计算任务的运行程序至第二计算设备，或者，管理装置向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序。

[0124] 第二计算设备中有可能部署有目标计算任务的运行程序，也可能未部署有目标计算任务的运行程序。当第二计算设备中未部署有目标计算任务的运行程序时，在第二计算设备执行目标计算任务之前，还需要先在该第二计算设备中部署该目标计算任务的运行程序。

[0125] 可选地，管理装置可以从存储有该目标计算任务的运行程序的计算设备中。获取该目标计算任务的运行程序，并将该目标计算任务的运行程序发送至该第二计算设备。例如，当不同计算设备之间的部分交互事务需要通过管理装置实现时，假设第一计算设备中存储有目标计算任务的运行程序，管理装置可以先从第一计算设备中获取该目标计算任务的运行程序，再由管理装置将该目标计算任务的运行程序发送至第二计算设备。并且，管理装置从第一计算设备中获取该目标计算任务的运行程序可以包括：管理装置主动从第一计算设备中读取该目标计算任务的运行程序，或，管理装置接收第一计算设备发送的该目标计算任务的运行程序。

[0126] 或者，管理装置可以向第二计算设备发送第二指令，以指示第二计算设备从存储有该目标计算任务的运行程序的计算设备中获取该目标计算任务的运行程序。相应的，第二计算设备接收到该第二指令后，即可直接根据该第二指令从存储有该目标计算任务的运行程序的计算设备中，获取该目标计算任务的运行程序。并且，第二计算设备从存储有该目标计算任务的运行程序的计算设备中，获取该目标计算任务的运行程序的过程可以包括：第二计算设备向存储有该目标计算任务的运行程序的计算设备发送获取请求，以请求该计算设备向该第二计算设备发送该目标计算任务的运行程序，或，第二计算设备存储有该目标计算任务的运行程序的计算设备中读取该目标计算任务的运行程序。

[0127] 需要说明的是，该步骤605为可选步骤，当第二计算设备中未部署目标计算任务的运行程序时，需要执行该步骤605，当第二计算设备中部署有目标计算任务的运行程序时，无需执行该步骤605，则在完成步骤604后，可以直接执行步骤606。

[0128] 步骤606、管理装置向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0129] 为了保证第二计算设备执行目标计算任务的运行性能，管理装置确定目标计算任务后，还需要向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0130] 在一种可实现方式中，管理装置可以获取目标计算任务的计算需求，基于该计算需求确定运行该目标计算任务所需的计算资源，并结合第二计算设备中的资源，生成针对

该目标计算任务的资源调度策略，并向第二计算设备发送携带有该资源调度策略的第三指令，以指示第二计算设备按照该资源调度为执行目标计算任务准备计算资源。可选地，该资源调度策略还指示对第一计算设备中的资源进行调度，如指示第一计算设备回收用于执行目标计算任务的资源。应理解，当资源调度策略还指示对第一计算设备中的资源进行调度时，管理装置还可将该资源调度策略发送至第一计算设备，以便于第一计算设备根据该资源调度策略对第一计算设备中的资源进行调度。其中，目标计算任务的计算需求可以根据该目标计算任务的输入数据量、输出数据量和该目标计算任务实现的算法确定。

[0131] 例如，假设第三指令携带的资源调度策略指示第二计算设备中为执行目标计算任务分配计算资源，且向目标计算任务分配的资源包括：提供两个核的AI算力、7.5GB的内存、计算设备的处理器内存向计算设备的AI计算内存传输数据的带宽调整为1GB/S，计算设备的AI计算内存向计算设备的处理器内存传输数据的带宽调整为1GB/S。

[0132] 管理装置可以按照预置的资源调度规则进行调度决策，生成资源调度策略。通过按照该资源调度规则进行调度决策，能够以执行调度操作为代价减小整体系统的短木板效应，提升整体系统的运行性能。

[0133] 在一种可实现方式中，当本申请实施例提供的AI应用任务的管理方法通过图5所示的AI系统实现时，第三计算设备中设置有调度模块3011和控制模块3012，第一计算设备中设置有第一资源调度执行模块104，第二计算设备中设置有第二资源调度执行模块204。如图7所示，控制模块3012确定目标计算任务后，向调度模块3011发送了指示由第二计算设备执行目标计算任务的通知，调度模块3011可以根据该通知生成对应的资源调度策略，并向第二资源调度执行模块204和第一资源调度执行模块104发送该资源调度策略，以便于第一资源调度执行模块104和第二资源调度执行模块204根据该资源调度策略进行资源调度。

[0134] 需要说明的是，生成资源调度策略的过程也可以由第二计算设备执行。例如，管理装置向第二计算设备发送第三指令后，第二计算设备可以根据该第三指令获取目标计算任务的计算需求，基于该计算需求确定执行该目标计算任务所需的计算资源，并结合第二计算设备中的资源，生成针对该目标计算任务的资源调度策略。

[0135] 步骤607、第二计算设备基于第三指令，为执行目标计算任务准备计算资源。

[0136] 第二计算设备接收资源调度策略后，可以按照该资源调度策略指示的方式，对第二计算设备的资源进行调度，为执行目标计算任务准备计算资源。如图7所示，该步骤607可以由第二资源调度执行模块204执行。应理解，当资源调度策略还指示对第一计算设备中的资源进行调度时，本申请实施例提供的AI应用任务的管理方法还包括：第一计算设备根据该资源调度策略对第一计算设备中的资源进行调度。例如，对用于执行目标计算任务的资源进行资源回收。并且，在回收资源前需要确定使用该资源执行的计算任务均已被卸载。其中，第一计算设备根据该资源调度策略对第一计算设备中的资源进行调度可以由第一资源调度执行模块104执行。

[0137] 步骤608、第一计算设备根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备。

[0138] 第一计算设备确定由第二计算设备执行目标计算任务后，需要将执行目标计算任务所需的数据发送至第二计算设备。该执行目标计算任务所需的数据包括执行至少一个其他计算任务获得的数据。该至少一个其他计算任务与目标计算任务存在数据关联。

[0139] 在一种可实现方式中，如图7所示，将执行目标计算任务所需的数据发送至第二计算设备的操作可以由第一任务调度执行模块103执行。并且，由于发送执行目标计算任务所需的数据的过程可能存在时延，该第一任务调度执行模块103还负责对执行目标计算任务所需的数据进行缓存管理，如暂时缓存执行目标计算任务所需的数据。

[0140] 步骤609、第二计算设备将数据作为目标计算任务的输入数据，基于输入数据执行目标计算任务。

[0141] 在完成为目标计算任务准备计算资源后，即可使用为该目标计算任务准备的计算资源运行目标计算任务的运行程序，并将至少一个其他计算任务获得的数据作为目标计算任务的输入数据，以执行目标计算任务。

[0142] 例如，假设目标计算任务为第一AI应用任务中的第5个计算任务，管理装置指示回收第一计算设备中用于运行该目标AI计算任务的资源，且在将目标计算任务调度至第二计算设备中之前，第一计算设备用于执行第一AI应用任务，第二计算设备用于执行第二AI应用任务，且第一AI应用任务和第二AI应用任务均包括目标计算任务，则在将该目标计算任务调度至第二计算设备中执行后，第一AI应用任务和第二AI应用任务的执行逻辑图如图9所示。即该第一计算设备中不再执行第5个计算任务，该第一计算设备中原来部署的第5个计算任务执行的操作由第二计算设备运行的第5个计算任务执行。

[0143] 又例如，假设目标计算任务为第一AI应用任务中的第5个计算任务，管理装置指示回收第一计算设备中用于执行该目标计算任务的资源，且在将目标计算任务调度至第二计算设备中之前，第二计算设备中未部署任何计算任务的运行程序，则在将该目标计算任务调度至第二计算设备中执行后，第一AI应用任务和第二AI应用任务的执行逻辑图如图10所示。即该第一计算设备中不再执行第5个计算任务，换做第二计算设备执行该第5个计算任务。

[0144] 又例如，假设目标计算任务为第一AI应用任务中的第5个计算任务，管理装置指示回收第一计算设备中用于执行该目标计算任务的资源，且在将目标计算任务调度至第二计算设备中之前，第二计算设备中未部署第一AI应用任务中的第5个计算任务的运行程序，则在将该目标计算任务调度至第二计算设备中执行后，第一AI应用任务和第二AI应用任务的执行逻辑图如图10所示。即该第一计算设备中不再执行第5个计算任务，由该第二计算设备执行该第5个计算任务。其中，为便于查看，该图10中未示出调度之前第二计算设备中已部署的计算任务。

[0145] 综上所述，在本申请实施例提供的AI应用任务的管理方法中，通过在第一计算设备执行AI应用任务包括的多个计算任务的过程中，管理装置在该多个计算任务中确定待在第二计算设备中执行的目标计算任务，然后第一计算设备将执行目标计算任务所需的数据发送至第二计算设备，第二计算设备将第一计算设备发送的数据作为目标计算任务的输入数据，并基于该输入数据执行目标计算任务，能够在第一计算设备执行AI应用任务的过程中，将目标计算任务调度至第二计算设备中执行，能够在执行AI应用任务的过程中对计算任务进行灵活的调度，有利于提高该第一计算设备和第二计算设备的资源利用率和/或计算任务处理效率，提高了用于实现AI应用任务的整体系统的运行性能。

[0146] 需要说明的是，本申请实施例提供的AI应用任务的管理方法的步骤先后顺序可以进行适当调整，步骤也可以根据情况进行相应增减，例如可以根据应用需求确定是否执行

步骤602，或者，可以根据应用需求确定是否执行步骤605。任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化的方法，都应涵盖在本申请的保护范围之内，因此不再赘述。

[0147] 本申请实施例还提供了一种AI系统。AI系统包括第一计算设备、第二计算设备和管理装置。该AI系统的架构请参考图2、图3或图5。其中，第一计算设备、第二计算设备和管理装置的作用如下：

[0148] 管理装置，用于确定多个计算任务中的目标计算任务，其中，目标计算任务为待由第二计算设备执行的计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联。

[0149] 管理装置，还用于向第一计算设备发送第一指令。

[0150] 第一计算设备，用于根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备。

[0151] 第二计算设备，用于将数据作为目标计算任务的输入数据，基于输入数据执行目标计算任务。

[0152] 可选地，管理装置还用于：获取多个计算任务在第一计算设备中执行时的第一资源使用信息，基于第一资源使用信息确定多个计算任务中的目标计算任务。

[0153] 可选地，当第二计算设备中未部署目标计算任务的运行程序时，管理装置还用于：发送目标计算任务的运行程序至第二计算设备，或者，向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序。

[0154] 可选地，管理装置还用于：向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0155] 可选地，管理装置具体用于：当第一计算设备中的多个计算任务的运行效率不满足预设第一条件，和/或，第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时，确定多个计算任务中的目标计算任务。

[0156] 可选地，管理装置还用于：获取在第二计算设备中执行的计算任务的第二资源使用信息。

[0157] 相应的，管理装置具体用于：基于第一资源使用信息和第二资源使用信息确定多个计算任务中的目标计算任务。

[0158] 可选地，第一资源使用信息包括：多个计算任务中至少一个计算任务的运行信息和第一计算设备的资源信息。

[0159] 可选地，至少一个计算任务的运行信息由以下一个或多个运行参数获得：计算任务在单位时长内的调用次数，计算任务的输入数据量，计算任务的输出数据量，第一计算设备调用计算任务的运行时长，调用计算任务的处理器的消耗量，调用计算任务的内存的消耗量。

[0160] 可选地，第一计算设备的资源信息由以下一个或多个参数获得：第一计算设备的内存的额定值和总消耗量，第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽和带宽的额定值，第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

[0161] 可选地，管理装置部署在第一计算设备、第二计算设备或者第三计算设备中，其

中，第三计算设备与第一计算设备和第二计算设备通过通信通路连接。

[0162] 可选地，第一计算设备为显卡、AI计算芯片或服务器。第二计算设备为显卡、AI计算芯片或服务器。第三计算设备为显卡、AI计算芯片或服务器。

[0163] 综上所述，在本申请实施例提供的AI系统中，通过在第一计算设备执行AI应用任务包括的多个计算任务的过程中，管理装置在该多个计算任务中确定待在第二计算设备中执行的目标计算任务，然后第一计算设备将执行目标计算任务所需的数据发送至第二计算设备，第二计算设备将第一计算设备发送的数据作为目标计算任务的输入数据，并基于该输入数据执行目标计算任务，能够在第一计算设备执行AI应用任务的过程中，将目标计算任务调度至第二计算设备中执行，能够在执行AI应用任务的过程中对计算任务进行灵活的调度，有利于提高该第一计算设备和第二计算设备的资源利用率和/或计算任务处理效率，提高了用于实现AI应用任务的整体系统的运行性能。

[0164] 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，上述描述的管理装置、第一计算设备和第二计算设备的具体工作过程，可以参考前述方法实施例中的对应内容，在此不再赘述。

[0165] 本申请实施例还提供了一种管理装置。该管理装置的架构请参考图5，如图5所示，管理装置301包括控制模块3012和调度模块3011。

[0166] 控制模块3012，用于在第一计算设备执行多个计算任务的过程中，确定多个计算任务中的目标计算任务，其中，目标计算任务为待由第二计算设备执行的计算任务，目标计算任务与AI应用任务涉及的至少一个其他计算任务存在数据关联；

[0167] 调度模块3011，用于发送第一指令至第一计算设备，第一指令用于指示第一计算设备将执行至少一个其他计算任务获得的数据发送至第二计算设备中的目标计算任务。

[0168] 可选地，控制模块3012还用于：获取多个计算任务在第一计算设备中执行时的第一资源使用信息；基于第一资源使用信息确定多个计算任务中的目标计算任务。

[0169] 可选地，当第二计算设备中未部署目标计算任务的运行程序时，调度模块3011，还用于发送目标计算任务的运行程序至第二计算设备，或者，向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序。

[0170] 可选地，调度模块3011，还用于向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源。

[0171] 可选地，控制模块3012，具体用于：当第一计算设备中的多个计算任务的运行效率不满足预设第一条件，和/或，第一计算设备中用于运行多个计算任务的资源利用情况不满足预设第二条件时，确定多个计算任务中的目标计算任务。

[0172] 可选地，控制模块3012，还用于获取在第二计算设备中执行的计算任务的第二资源使用信息。相应的，控制模块3012，具体用于基于第一资源使用信息和第二资源使用信息确定多个计算任务中的目标计算任务。

[0173] 可选地，第一资源使用信息包括：多个计算任务中至少一个计算任务的运行信息和第一计算设备的资源信息。

[0174] 可选地，至少一个计算任务的运行信息由以下一个或多个运行参数获得：计算任务在单位时长内的调用次数，计算任务的输入数据量，计算任务的输出数据量，第一计算设备调用计算任务的运行时长，调用计算任务的处理器的消耗量，调用计算任务的内存的消

耗量。

[0175] 可选地，第一计算设备的资源信息由以下一个或多个参数获得：第一计算设备的内存的额定值和总消耗量，第一计算设备处理器内存向第一计算设备的AI计算内存传输数据的带宽和带宽的额定值，第一计算设备的AI计算内存向第一计算设备的处理器内存传输数据的带宽和带宽的额定值。

[0176] 可选地，管理装置部署在第一计算设备、第二计算设备或者第三计算设备中，其中，第三计算设备与第一计算设备和第二计算设备通过通信通路连接。

[0177] 可选地，第一计算设备为显卡、AI计算芯片或服务器；第二计算设备为显卡、AI计算芯片或服务器；第三计算设备为显卡、AI计算芯片或服务器。

[0178] 综上所述，在本申请实施例提供的管理装置中，通过在第一计算设备执行AI应用任务包括的多个计算任务的过程中，管理装置在该多个计算任务中确定待在第二计算设备中执行的目标计算任务，然后向第一计算设备发送第一指令，使得第一计算设备将执行目标计算任务所需的数据发送至第二计算设备，第二计算设备将第一计算设备发送的数据作为目标计算任务的输入数据，并基于该输入数据执行目标计算任务，能够在第一计算设备执行AI应用任务的过程中，将目标计算任务调度至第二计算设备中执行，能够在执行AI应用任务的过程中对计算任务进行灵活的调度，有利于提高该第一计算设备和第二计算设备的资源利用率和/或计算任务处理效率，提高了用于实现AI应用任务的整体系统的运行性能。

[0179] 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，上述描述的管理装置即模块的具体工作过程，可以参考前述方法实施例中的对应内容，在此不再赘述。

[0180] 本申请实施例还提供了一种电子设备，该电子设备用于通过运行指令以实现前述管理装置的功能。在如图2所示的场景中，该电子设备可以是第三计算设备30，或者可以是包括前述第三计算设备30的电子设备，例如：第三计算设备30为显卡，则该电子设备可以为包括该显卡的服务器。在如图3所述的实施例中，该电子设备可以是实现管理装置103的功能的第一计算设备10，或者可以是包括前述第一计算设备10的电子设备。如图11所示，该电子设备110包括总线1101、处理器1102、通信接口1103和存储器1104。处理器1102、存储器1104和通信接口1103之间通过总线1101通信。

[0181] 其中，处理器1102可以是一种集成电路芯片，具有信号的处理能力。在实现过程中，本申请实施例提供的AI应用任务的管理方法中管理装置的功能可以通过处理器502中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器502还可以是专用集成电路(application-specific integrated circuit,ASIC)，可编程逻辑器件(programmable logic device,PLD)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件及以上部分或全部的组合。上述PLD可以是复杂可编程逻辑器件(complex programmable logic device,CPLD)，现场可编程逻辑门阵列(field-programmable gate array,FPGA)，通用阵列逻辑(generic array logic,GAL)或其任意组合。

[0182] 处理器1102也可以是通用处理器，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。例如，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)或者，CPU、GPU和NP中部分或全部的组合。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译

码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器501，处理器502读取存储器501中的信息和计算机指令，结合其硬件实现本申请实施例的AI应用任务的管理方法中管理装置的功能。

[0183] 存储器1104存储计算机指令和数据。例如，存储器1104中存储有AI应用任务的管理方法中管理装置的功能的可执行代码，处理器1102读取存储器1104中的该可执行代码以执行本申请实施例提供的AI应用任务的管理方法。存储器1104可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者的结合。其中，非易失性存储器可以是只读存储器(read-only memory，ROM)、可编程只读存储器(programmable ROM，PROM)、可擦除可编程只读存储器(erasable PROM，EPROM)、电可擦除可编程只读存储器(electrically EPROM，EEPROM)、快闪存储器(flash memory)，硬盘(hard disk drive，HDD)或固态硬盘(solid-state drive，SSD)。易失性存储器可以是随机存取存储器(random access memory，RAM)，其用作外部高速缓存。通过示例性但不是限制性说明，许多形式的RAM可用，例如静态随机存取存储器(static RAM，SRAM)、动态随机存取存储器(DRAM)、同步动态随机存取存储器(synchronous DRAM，SDRAM)、双倍数据速率同步动态随机存取存储器(double data date SDRAM，DDR SDRAM)、增强型同步动态随机存取存储器(enhanced SDRAM，ESDRAM)、同步连接动态随机存取存储器(synchlink DRAM，SLDRAM)和直接内存总线随机存取存储器(direct rambus RAM，DR RAM)。并且，存储器1104中还可以包括操作系统等其他运行进程所需的软件模块。操作系统可以为LINUX™，UNIX™，WINDOWS™等。

[0184] 通信接口1103使用例如但不限于收发器一类的收发模块，来实现计算机110与其他设备或通信网络之间的通信。

[0185] 总线1101可包括在电子设备110各个部件(例如，处理器1102、通信接口1103和存储器1104)之间传送信息的通路。

[0186] 本申请还提供了一种计算机可读存储介质，该计算机可读存储介质可以为非瞬态的可读存储介质，当计算机可读存储介质中的程序指令被电子设备运行时，该电子设备实现本申请提供的AI应用任务的管理方法中管理装置的功能。该计算机可读存储介质包括但不限于易失性存储器，例如随机访问存储器，非易失性存储器，例如快闪存储器、硬盘(hard disk drive，HDD)、固态硬盘(solid state drive，SSD)。

[0187] 本申请实施例还提供了一种包含指令的计算机程序产品，当计算机程序产品在电子设备上运行时，使得电子设备实现本申请实施例提供的AI应用任务的管理方法中管理装置的功能。

[0188] 本申请实施例还提供了一种芯片，该芯片包括可编程逻辑电路和/或程序指令，当芯片运行时用于实现本申请实施例提供的AI应用任务的管理方法中管理装置的功能。

[0189] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0190] 在本申请实施例中，术语“第一”、“第二”和“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。术语“一个或多个”是指一个或多个，术语“多个”指两个或两个以

上,除非另有明确的限定。

[0191] 本申请中术语“和/或”,仅仅是一种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。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

[0192] 以上仅为本申请的可选实施例,并不用以限制本申请,凡在本申请的构思和原则之内,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。

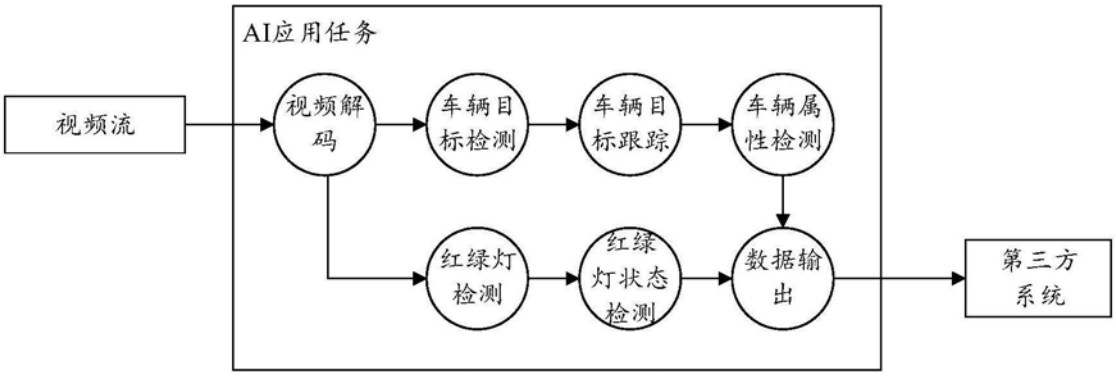

AI应用任务
视频流
视频解码
车辆目标检测
车辆目标跟踪
车辆属性检测
红绿灯检测
红绿灯状态检测
数据输出
第三方系统

图1

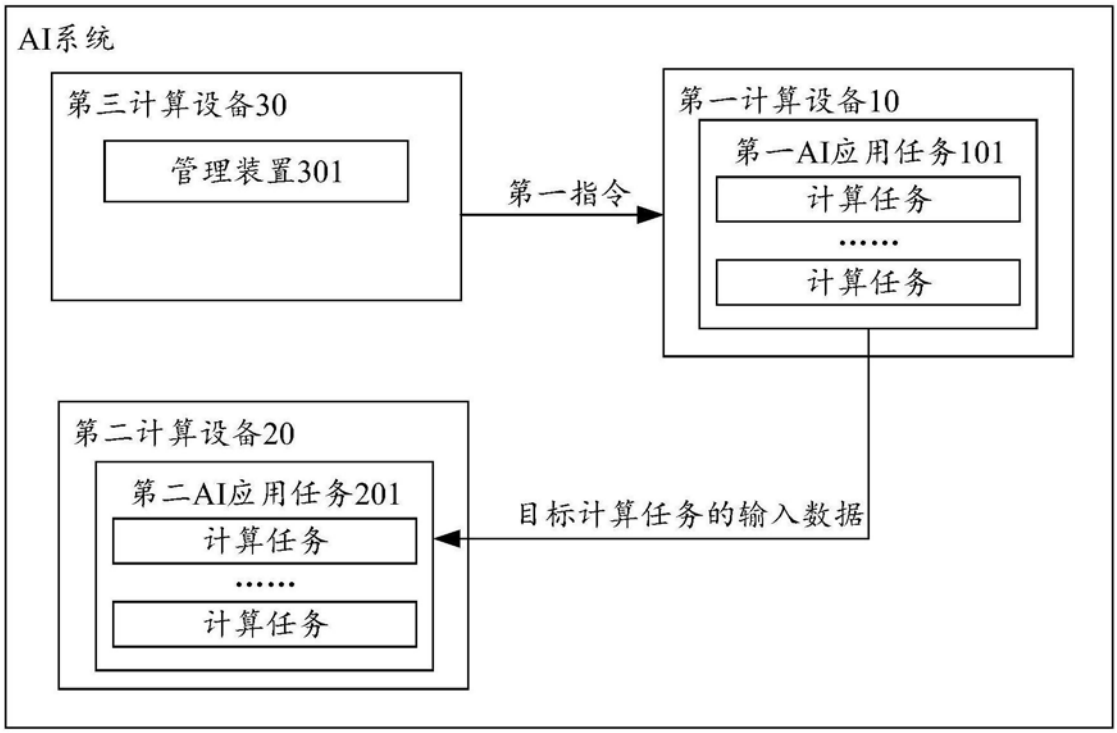

AI系统
第三计算设备30
管理装置301
第一指令
第一计算设备10
第一AI应用任务101
计算任务
……
计算任务
第二计算设备20
第二AI应用任务201
计算任务
……
计算任务
目标计算任务的输入数据

图2

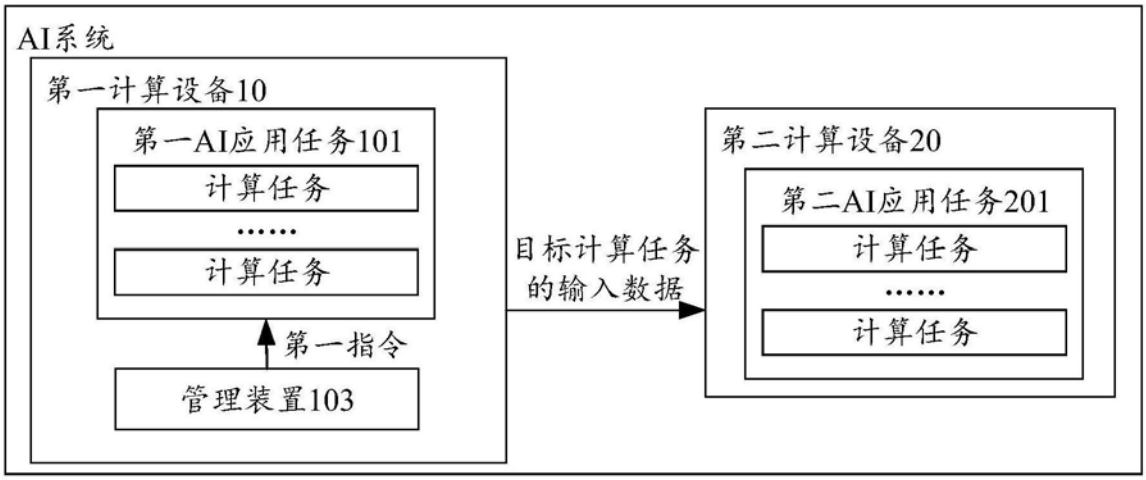

AI系统
第一计算设备10
第一AI应用任务101
计算任务
……
计算任务
第一指令
管理装置103
目标计算任务
的输入数据
第二计算设备20
第二AI应用任务201
计算任务
……
计算任务

图3

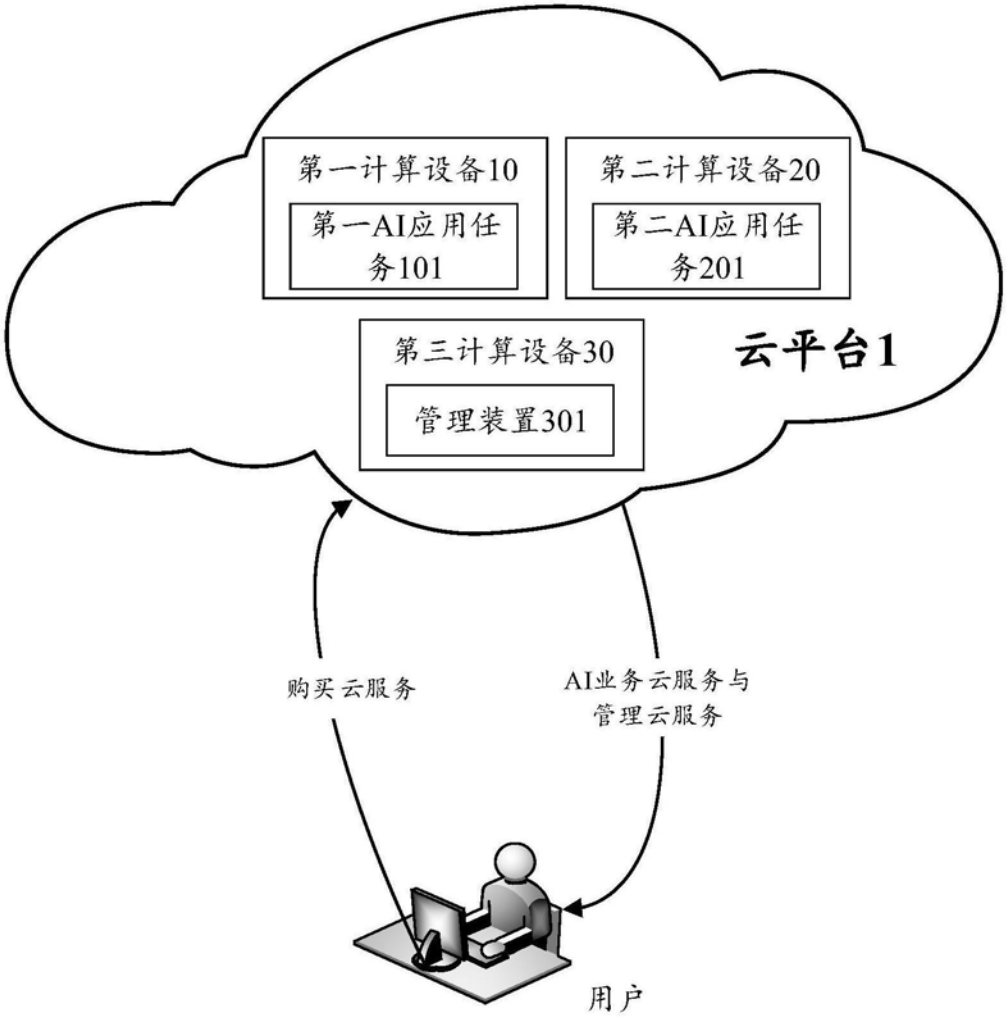

第一计算设备10
第一AI应用任务101
第二计算设备20
第二AI应用任务201
第三计算设备30
管理装置301
云平台1
购买云服务
AI业务云服务与
管理云服务
用户

图4

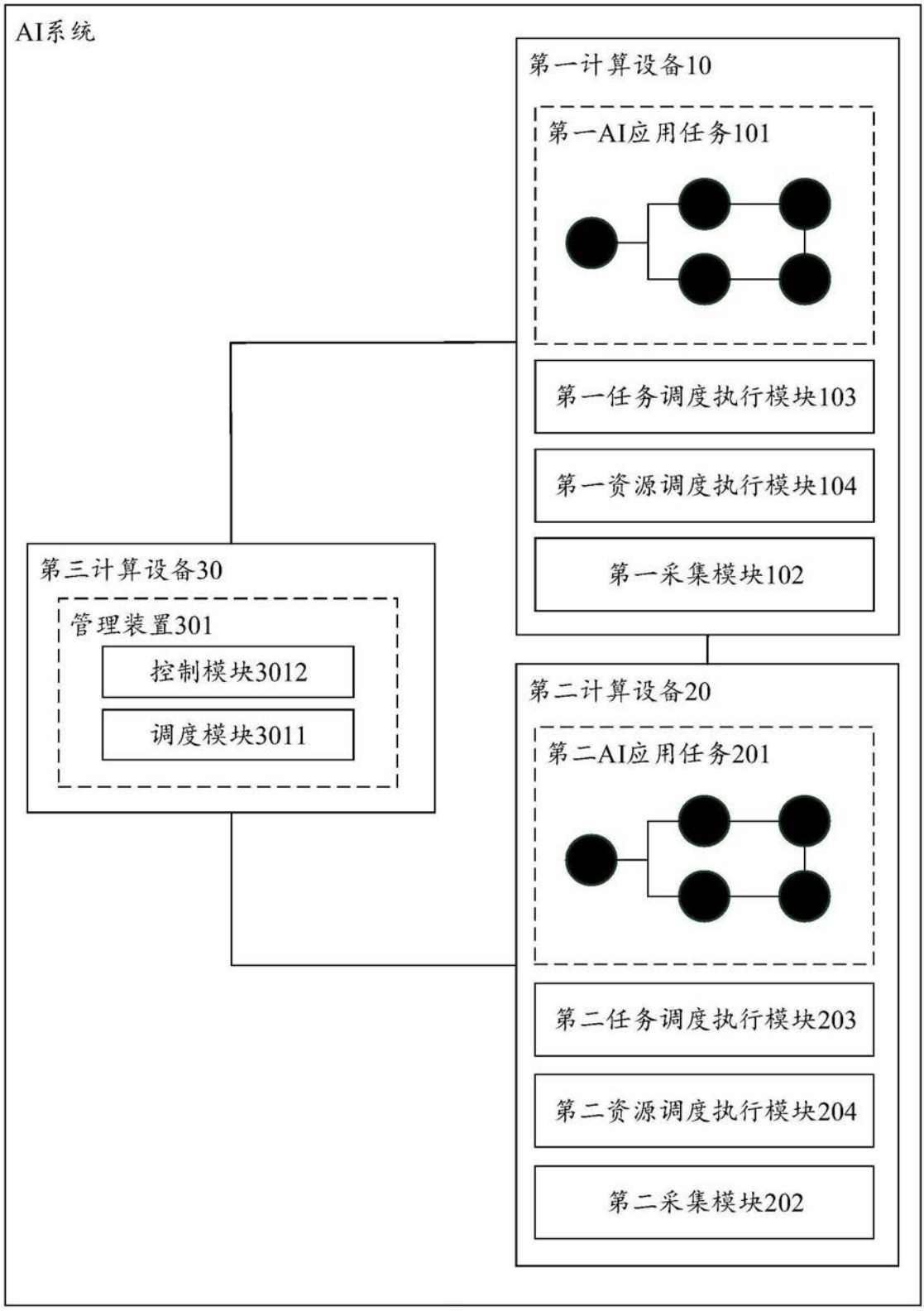

AI系统
第一计算设备10
第一AI应用任务101
第一任务调度执行模块103
第一资源调度执行模块104
第一采集模块102
第三计算设备30
管理装置301
控制模块3012
调度模块3011
第二计算设备20
第二AI应用任务201
第二任务调度执行模块203
第二资源调度执行模块204
第二采集模块202

图5

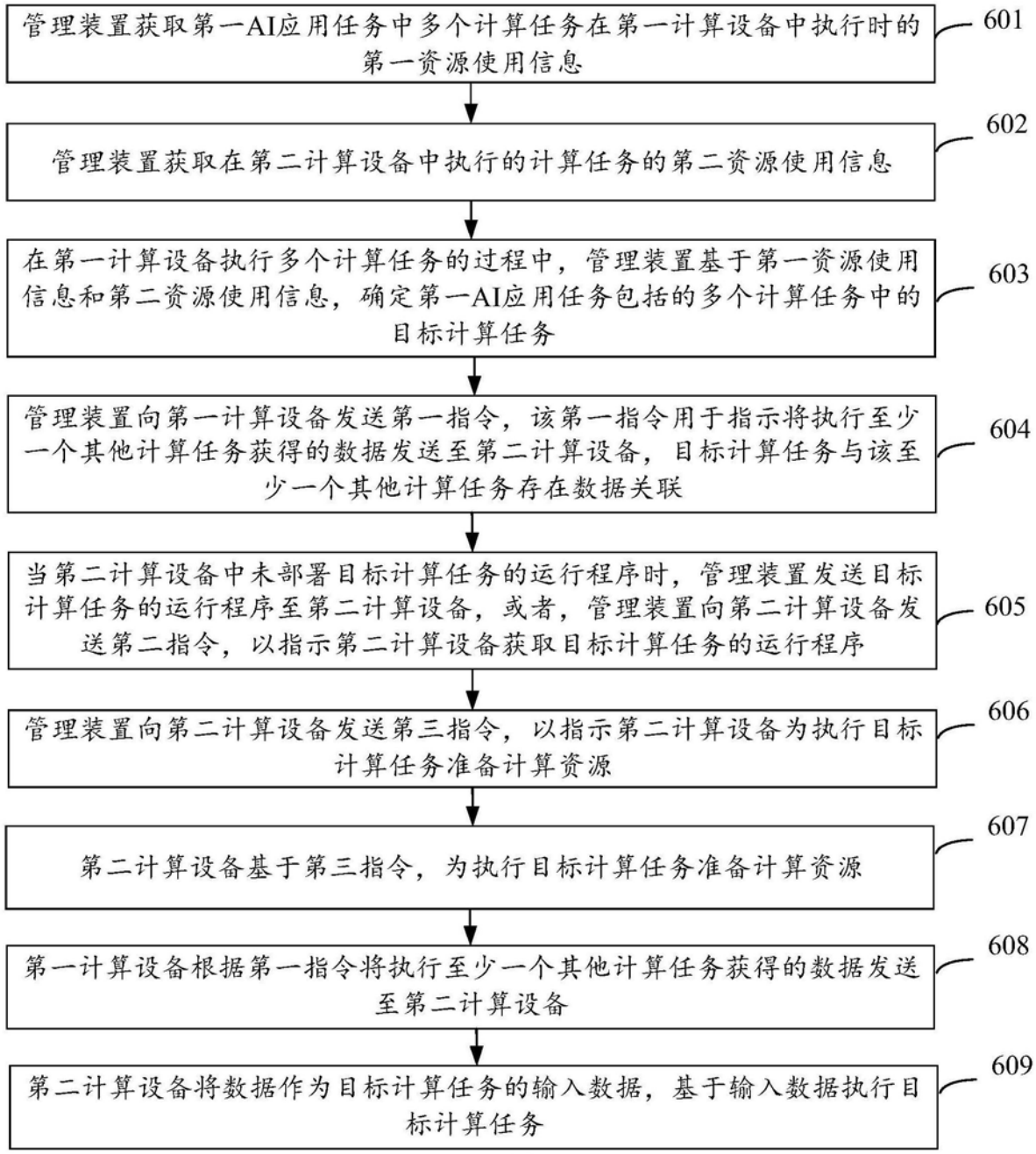

管理装置获取第一AI应用任务中多个计算任务在第一计算设备中执行时的第一资源使用信息
601
管理装置获取在第二计算设备中执行的计算任务的第二资源使用信息
602
在第一计算设备执行多个计算任务的过程中，管理装置基于第一资源使用信息和第二资源使用信息，确定第一AI应用任务包括的多个计算任务中的目标计算任务
603
管理装置向第一计算设备发送第一指令，该第一指令用于指示将执行至少一个其他计算任务获得的数据发送至第二计算设备，目标计算任务与该至少一个其他计算任务存在数据关联
604
当第二计算设备中未部署目标计算任务的运行程序时，管理装置发送目标计算任务的运行程序至第二计算设备，或者，管理装置向第二计算设备发送第二指令，以指示第二计算设备获取目标计算任务的运行程序
605
管理装置向第二计算设备发送第三指令，以指示第二计算设备为执行目标计算任务准备计算资源
606
第二计算设备基于第三指令，为执行目标计算任务准备计算资源
607
第一计算设备根据第一指令将执行至少一个其他计算任务获得的数据发送至第二计算设备
608
第二计算设备将数据作为目标计算任务的输入数据，基于输入数据执行目标计算任务
609

图6

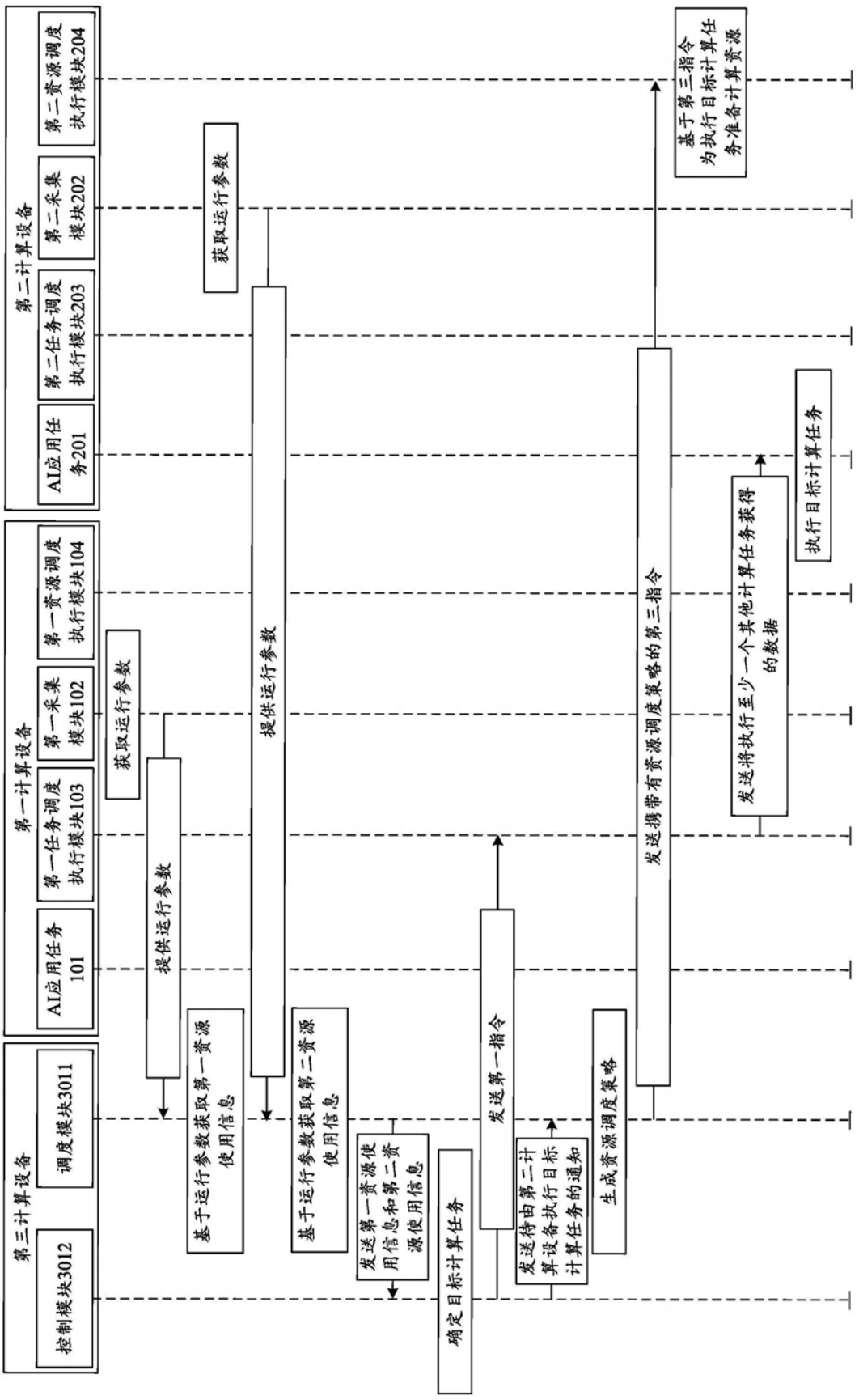

第三计算设备
控制模块3012
调度模块3011
第一计算设备
AI应用任务101
第一任务调度执行模块103
第一采集模块102
第一资源调度执行模块104
第二计算设备
AI应用任务201
第二任务调度执行模块203
第二采集模块202
第二资源调度执行模块204
提供运行参数
获取运行参数
基于运行参数获取第一资源使用信息
基于运行参数获取第二资源使用信息
发送第一资源使用信息和第二资源使用信息
确定目标计算任务
发送第一指令
发送待由第二计算设备执行目标计算任务的通知
生成资源调度策略
发送携带有资源调度策略的第三指令
基于第三指令为执行目标计算任务准备计算资源
发送将执行至少一个其他计算任务获得的数据
执行目标计算任务

图7

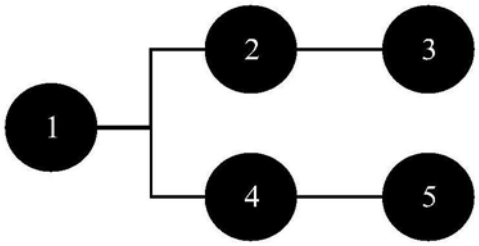

1
2
3
4
5

图8

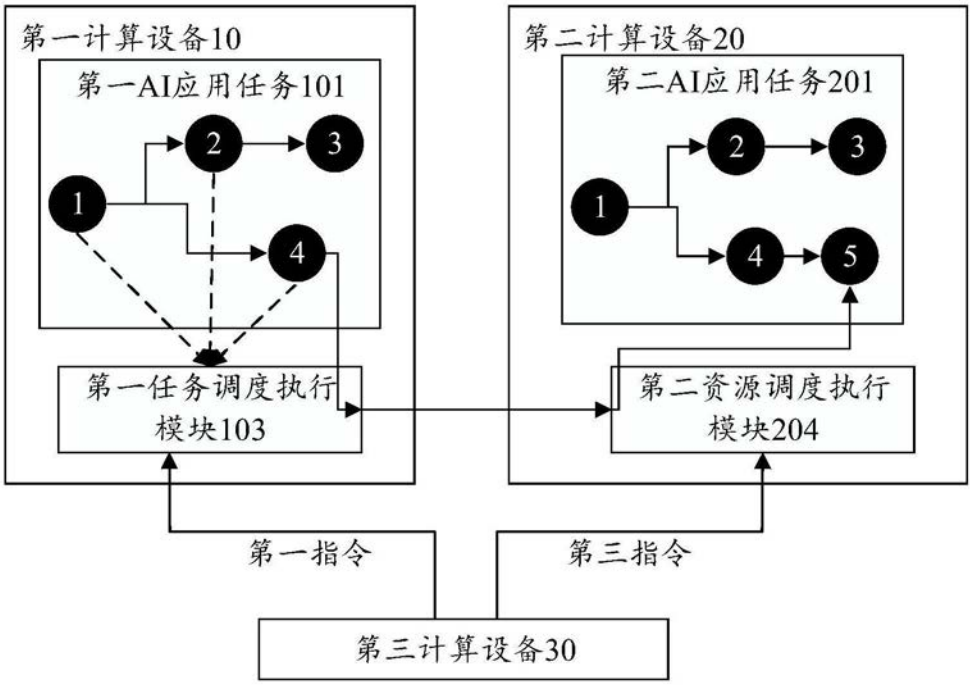

第一计算设备10
第一AI应用任务101
1
2
3
4
第一任务调度执行模块103
第二计算设备20
第二AI应用任务201
1
2
3
4
5
第二资源调度执行模块204
第一指令
第三指令
第三计算设备30

图9

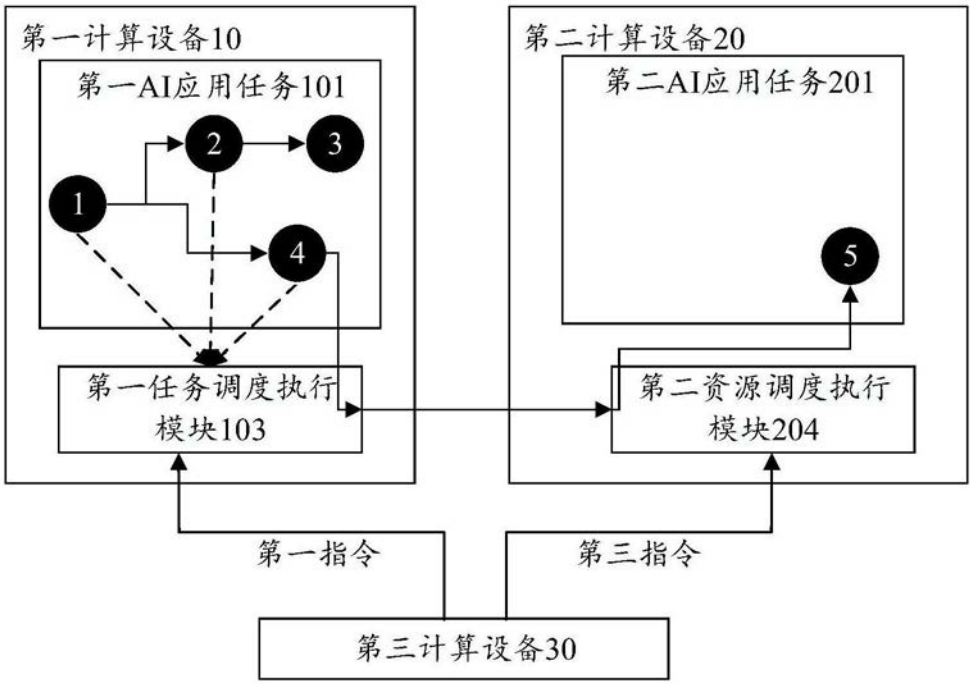

第一计算设备10
第一AI应用任务101
2
3
1
4
第一任务调度执行模块103
第二计算设备20
第二AI应用任务201
5
第二资源调度执行模块204
第一指令
第三指令
第三计算设备30

图10

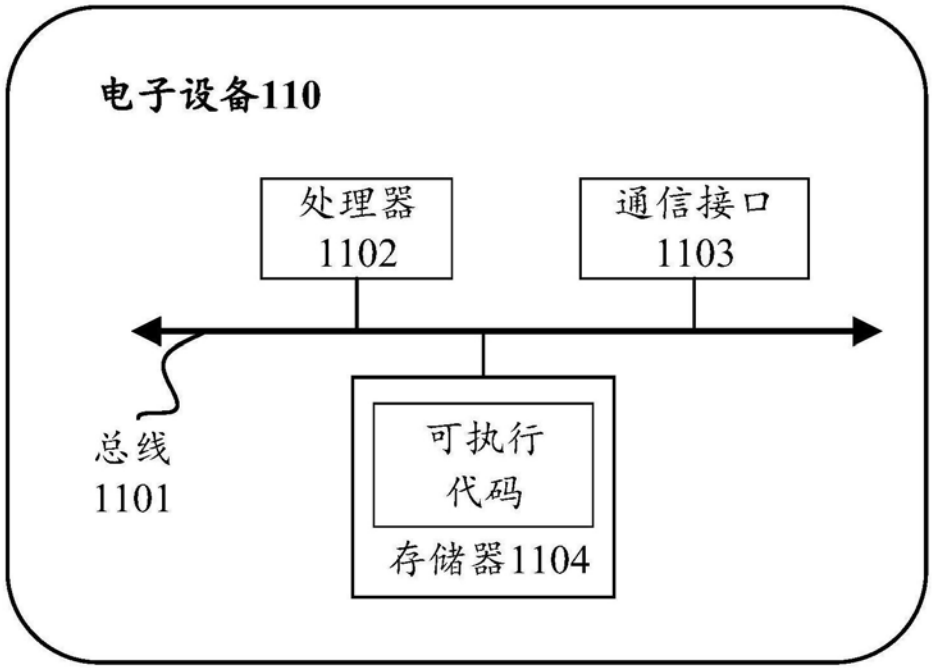

电子设备110
处理器1102
通信接口1103
总线1101
可执行代码
存储器1104

图11