模型推理服务部署方法、装置、设备及存储介质

本申请要求在2025年02月20日提交中国专利局、申请号为202510188154.6的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算资源分配技术领域，例如涉及一种模型推理服务部署方法、装置、设备及存储介质。

背景技术

随着机器学习和人工智能技术的发展，汽车金融领域中多类业务结合自身的业务发展需要，逐渐利用机器学习和人工智能技术开发、训练多类业务垂直领域的人工智能模型并通过部署模型推理服务的方式应用到正常的业务领域中。由于汽车金融领域中的多类垂直业务之间隔离性较强，每个垂直业务所使用的模型差异性较强，业务之间基本不存在模型服用的场景，这也导致了随着越来越多的人工智能模型加入到多种垂直业务领域中，用于提供模型推理了的图形处理器（Graphics Processing Unit，GPU）资源需求日益增多，GPU资源压力逐渐增大。受限于GPU资源的稀缺性，扩充GPU资源的成本较高，亟需提供一种方法提升存量GPU资源的利用率。

发明内容

本申请提供了一种模型推理服务部署方法、装置、设备及存储介质，以根据需要动态调整调度策略，确保模型推理服务能够满足预定的性能要求，同时提升存量GPU资源的利用率。

本申请实施例提供了一种模型推理服务部署方法。该方法包括：

在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数；

针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式；

基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

本申请实施例提供了一种模型推理服务部署装置。该装置包括：

推理指标参数获取模块，设置为在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数；

目标调度模式确定模块，设置为针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式；

服务部署方式确定模块，设置为基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

本申请实施例提供了一种电子设备，所述电子设备包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行本申请任一实施例所述的模型推理服务部署方法。

本申请实施例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指令，所述计算机指令用于使处理器执行时实现本申请任一实施例所述的模型推理服务部署方法。

附图说明

图 1 是根据本申请实施例一提供的模型推理服务部署方法的流程图；

图 2 是根据本申请实施例二提供的模型推理服务部署方法的流程图；

图 3 是根据本申请实施例三提供的模型推理服务部署装置的结构图；

图 4 是实现本申请实施例的模型推理服务部署方法的电子设备的结构示意图。

具体实施方式

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，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等

是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

实施例一

图 1 为本申请实施例一提供的一种模型推理服务部署方法的流程图，本实施例可适用于对模型推理服务所需的计算资源进行优化分配的情况，该方法可以由模型推理服务部署装置来执行，该模型推理服务部署装置可以采用硬件和/或软件的形式实现，该模型推理服务部署装置可配置于电子设备中。如图 1 所示，该方法包括：

S101、在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数。

在机器学习和人工智能领域，模型推理服务是指利用已经训练好的人工智能模型对新的输入数据进行预测或分析的过程。这种服务通常部署在计算资源上，如 GPU，以提供高效的推理能力。在汽车金融等领域，模型推理服务被广泛应用于多类垂直业务中，以支持业务的智能化决策和运营。

计算资源集群可以是指一组相互连接的计算资源（如服务器、GPU 等），它们共同协作以提供模型推理服务。这些计算资源被组织在一起，以便能够高效地处理大量的模型推理请求。

推理指标参数可以是指模型推理服务在模型推理过程中的推理参数，如响应时间、吞吐量、资源利用率（如 GPU 显存占用率）和错误率等。

通过推理服务调度策略控制器选择相应的监控指标和时间段，筛选出与模型推理服务相关的监控项，导出或查看所需的推理指标参数。

S102、针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式。

所述目标调度模式包括服务质量优先模式和资源利用率优先模式。模式确定阈值可以根据实际情况进行设定，本申请对模式确定阈值的数值的确定不进行限定。

所述模式确定阈值的属性至少包括：以时间周期为窗口的平均请求数、以百分位为窗口的平均响应时间、以时间周期为窗口的最长响应时间、在指定时间内是否总是使用默认调度方式，以及在指定时间内总是不使用默认调度方式。示例性地，以时间周期为窗口的平均请求数可以包括每秒请求数、每分钟请求数和每小时请求数等。以百分位为窗口的平均响应时间可以包括 P99 响应时间、P95 响应时间和 P90 响应时间等。以时间周期为窗口的最长响应时间包括每分钟最长响应时间和每小时最长响应时间等。

针对每个模型推理服务，可以根据推理指标参数和预先构建的多个模式确定阈值之间的大小关系，进而确定所述模型推理服务对应的目标调度模式为服务质量优先模式，还是资源利用率优先模式。

示例性地，所述根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，包括：

将所述推理指标参数和所述模式确定阈值进行对比处理，根据所述推理指标参数与所述模式确定阈值的匹配结果，确定所述模型推理服务对应的目标调度模式。

也就是说，将推理指标参数与所述模式确定阈值进行对比匹配处理，即将推理指标参数分别从以时间周期为窗口的平均请求数的维度确定，从以百分位为窗口的平均响应时间的维度确定，从以时间周期为窗口的最长响应时间的维度确定，从在指定时间内总是使用默认调度方式的维度确定，以及从是否在指定时间内总是不使用默认调度方式的维度确定，模型推理服务对应的目标调度模式为服务质量优先模式，还是资源利用率优先模式。

针对多个模式确定阈值的数值，本申请不进行定，可以根据实际使用情况或历史经验进行设定。

S103、基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

在本申请技术方案中，在所述目标调度模式为服务质量优先模式的情况下，所述服务部署方式包括独立部署方式和联合部署方式，在所述目标调度模式为资源利用率优先模式的情况下，所述服务部署方式包括联合部署方式和卡共享部署方式。

所述独立部署方式为单个模型推理服务独自占据所述计算资源集群中的一个计算资源。所述联合部署方式为使用同一个推理服务框架的相同推理类型的多个模型推理服务，以共享上下文方式占据所述计算资源集群中的同一个计算资源。卡共享部署方式为多个模型推理服务以时间片轮转方式轮转占据所述计

算资源集群中的同一个计算资源。

独立部署方式为模型推理服务独占整个 GPU 卡资源，模型推理服务对 GPU 的使用具有排他性，模型推理服务对 GPU 的占用一直持续到模型推理服务的生命周期结束（模型推理服务正式下线）后才释放。联合部署方式为多个模型推理服务使用同一个推理服务框架（例如 Nvidia Triton 框架）以共享上下文方式（Nvidia MPS）联合使用同一个计算节点的同一个 GPU，这种部署的方式避免了多个工作进程使用 GPU 时的上下文切换带来的延迟。共享卡部署方式为多个模型推理服务直接配置使用同一个计算节点的同一个 GPU，这种部署方式在不同进程访问 GPU 时，会进行显式的 GPU 上下文切换；多个进程间通过时间片轮转的方式获取 GPU 的使用权。

在本申请中，不同的目标调度模式具有不同的服务部署策略，针对每种目标调度模式，采用相应的服务部署策略，以确定所述模型推理服务对应的服务部署方式。在确定服务部署方式之后，通知模型部署服务按新的服务部署方式重新部署模型推理服务。

本申请实施例的技术方案，通过在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数。针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式。基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署，解决了服务资源利用不足或不平衡的问题，实现了根据需要动态调整调度策略，确保模型推理服务能够满足预定的性能要求，优化资源分配，提高整体资源利用率，降低了运营成本。

实施例二

图 2 为本申请实施例二提供的一种模型推理服务部署方法的流程图，本实施例在上述多个实施例的基础上，对基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式进行说明。如图 2 所示，该方法包括：

S201、在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数。

S202、针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模

式。

S203、在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式。

部署方式阈值可以是指确定部署方式的判断阈值，用于在所述目标调度模式为服务质量优先模式的情况下，判断服务部署方式为独立部署方式或联合部署方式。示例性地，部署方式阈值可以设定为单个模型推理使用超过单卡 75%显存。值得注意的是，在服务质量优先模式下，不使用卡共享部署的方式。

示例性地，所述基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式，包括：在所述推理指标参数大于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为独立部署方式；在所述推理指标参数小于或等于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为联合部署方式。

也就是说，在服务质量优先模式下，需要尽可能满足模型推理服务的服务质量调度策略。该策略会尽可能将高资源占用（即推理指标参数大于部署方式阈值）的模型推理服务调度为单独部署方式。将中、低资源占用（即推理指标参数小于或等于部署方式阈值）的模型推理服务调度为联合部署方式。

S204、在所述目标调度模式为资源利用率优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式。

资源利用率优先模式即为以最大化资源利用率的模式，需要以最大化资源利用率的调度策略，确定模型推理服务对应的服务部署方式。

示例性地，基于贪心算法，参照最大化资源利用率的调度策略，分别确定所述联合部署方式和所述卡共享部署方式在所述服务部署方式中的部署占比，或者，分别计算采用所述联合部署方式时对应的第一资源利用率，以及采用所述卡共享部署方式时对应的第二资源利用率，在所述第一资源利用率大于所述第二资源利用率的情况下采用所述联合部署方式，在所述第一资源利用率小于或等于所述第二资源利用率的情况下采用所述卡共享部署方式。

也就是说，在所述目标调度模式为资源利用率优先模式的情况下，可以采用多种方式确定服务部署方式，至少包括基于贪心算法，确定所述联合部署方式和所述卡共享部署方式在所述服务部署方式中的部署占比；或者根据采用联合部署方式和卡共享部署方式的资源利用率，确定模型推理服务对应的服务部署方式。

在所述目标调度模式为资源利用率优先模式的情况下，需要尽可能满足最

大化资源利用率的调度策略，该策略默认通过贪心算法以联合部署和卡共享部署的方式尽可能将多种资源占用的模型推理服务部署到同一个计算资源上。另一方面，在所述第一资源利用率大于所述第二资源利用率的情况下采用所述联合部署方式，在所述第一资源利用率小于或等于所述第二资源利用率的情况下采用所述卡共享部署方式。

示例性地，在所述目标调度模式为资源利用率优先模式的情况下，基于贪心算法，参照最大化资源利用率的调度策略的详细调度过程包括：

向资源管理服务提交查询请求，获取当前集群中所有标记正在用于资源利用优先策略的资源节点列表；

遍历资源节点列表，获取资源节点列表中剩余可用显存最大的资源节点和对应的 GPU 卡编号；

查询正在部署推理服务的部署亲和性，获取推理服务的部署方式使用联合部署还是共享卡部署，若推理服务未指定部署方式，则默认使用剩余资源最多的部署方式；

绑定推理服务和对应的资源节点、GPU 卡编号，生成调度方案，通知模型部署服务部署模型推理服务。

若资源管理服务没有返回正在用于资源利用优先策略的节点列表，或返回的资源节点列表中所有节点均不满足推理服务所需的 GPU 资源时：

向资源管理服务提交查询请求，获取当前集群中空闲的资源节点列表；

遍历资源节点列表，获取资源节点列表中总显存最大的资源节点和对应的 GPU 卡编号；

向资源管理服务提交资源分配请求，将指定节点标记为正在用于资源利用优先调度策略节点；

查询正在部署推理服务的部署亲和性，获取推理服务的部署方式使用联合部署还是共享卡部署，若推理服务未指定部署方式，则默认使用联合部署方式进行部署；

绑定推理服务和对应的资源节点、GPU 卡编号，生成调度方案，通知模型部署服务部署模型推理服务。

S205、基于所述服务部署方式，将所述模型推理服务进行服务部署。

本申请实施例技术方案，通过在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式。在所述目标调度模式为资源利用率

优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式。根据目标调度模式的不同，采用不同的服务部署方式，可以合理规划任务在硬件资源上的分配，避免多个任务同时争用同一资源导致性能下降或系统崩溃的情况，促进了模型推理服务的高效、稳定和可靠运行。

实施例三

图 3 为本申请实施例三提供的一种模型推理服务部署装置的结构示意图。如图 3 所示，该装置包括：

推理指标参数获取模块 301，设置为在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数；

目标调度模式确定模块 302，设置为针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式；

服务部署方式确定模块 303，设置为基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

本申请实施例的技术方案，通过在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数。针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式。基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署，解决了服务资源利用不足或不平衡的问题，实现了根据需要动态调整调度策略，确保模型推理服务能够满足预定的性能要求，优化资源分配，提高整体资源利用率，降低了运营成本。

可选的，所述模式确定阈值的属性至少包括：以时间周期为窗口的平均请求数、以百分位为窗口的平均响应时间、以时间周期为窗口的最长响应时间、在指定时间内总是使用默认调度方式，以及在指定时间内总是不使用默认调度方式。

可选地，目标调度模式确定模块 302，是设置为：

将所述推理指标参数和所述模式确定阈值进行对比处理，根据所述推理指标参数与所述模式确定阈值的匹配结果，确定所述模型推理服务对应的目标调

度模式。

可选地，在所述目标调度模式为服务质量优先模式的情况下，所述服务部署方式包括独立部署方式和联合部署方式，其中，所述独立部署方式为单个模型推理服务独自占据所述计算资源集群中的一个计算资源，所述联合部署方式为使用同一个推理服务框架的相同推理类型的多个模型推理服务，以共享上下文方式占据所述计算资源集群中的同一个计算资源；

在所述目标调度模式为资源利用率优先模式的情况下，所述服务部署方式包括联合部署方式和卡共享部署方式，其中，所述卡共享部署方式为多个模型推理服务以时间片轮转方式轮转占据所述计算资源集群中的同一个计算资源。

可选地，服务部署方式确定模块303，包括：

第一部署方式确定单元，设置为在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式；

第二部署方式确定单元，设置为在所述目标调度模式为资源利用率优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式。

可选地，第一部署方式确定单元，是设置为：

在所述推理指标参数大于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为独立部署方式；

在所述推理指标参数小于或等于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为联合部署方式。

可选地，第二部署方式确定单元，是设置为：

基于贪心算法，参照最大化资源利用率的调度策略，分别确定所述联合部署方式和所述卡共享部署方式在所述服务部署方式中的部署占比，

或者，

分别计算采用所述联合部署方式时对应的第一资源利用率，以及采用所述卡共享部署方式时对应的第二资源利用率，在所述第一资源利用率大于所述第二资源利用率的情况下采用所述联合部署方式，在所述第一资源利用率小于或等于所述第二资源利用率的情况下采用所述卡共享部署方式。

本申请实施例所提供的模型推理服务部署装置可执行本申请任意实施例所提供的模型推理服务部署方法，具备执行方法相应的功能模块。

实施例四

图 4 示出了可以用来实施本申请的实施例的电子设备 10 的结构示意图。电子设备旨在表示多种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示多种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备（如头盔、眼镜、手表等）和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

如图 4 所示，电子设备 10 包括至少一个处理器 11，以及与至少一个处理器 11 通信连接的存储器，如只读存储器（Read-Only Memory，ROM）12、随机访问存储器（Read-Only Memory，RAM）13 等，其中，存储器存储有可被至少一个处理器执行的计算机程序，处理器 11 可以根据存储在只读存储器（ROM）12 中的计算机程序或者从存储单元 18 加载到随机访问存储器（RAM）13 中的计算机程序，来执行多种适当的动作和处理。在 RAM 13 中，还可存储电子设备 10 操作所需的多种程序和数据。处理器 11、ROM 12 以及 RAM 13 通过总线 14 彼此相连。输入/输出（Input/Output，I/O）接口 15 也连接至总线 14。

电子设备 10 中的多个部件连接至 I/O 接口 15，包括：输入单元 16，例如键盘、鼠标等；输出单元 17，例如多种类型的显示器、扬声器等；存储单元 18，例如磁盘、光盘等；以及通信单元 19，例如网卡、调制解调器、无线通信收发机等。通信单元 19 允许电子设备 10 通过诸如因特网的计算机网络和/或多种电信网络与其他设备交换信息/数据。

处理器 11 可以是多种具有处理和计算能力的通用和/或专用处理组件。处理器 11 的一些示例包括但不限于中央处理单元（Central Processing Unit，CPU）、图形处理单元（Graphics Processing Unit，GPU）、多种专用的人工智能（Artificial Intelligence，AI）计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器（Digital Signal Processor，DSP）、以及任何适当的处理器、控制器、微控制器等。处理器 11 执行上文所描述的多个方法和处理，例如模型推理服务部署方法。

在一些实施例中，模型推理服务部署方法可被实现为计算机程序，其被有形地包含于计算机可读存储介质，例如存储单元 18。在一些实施例中，计算机程序的部分或者全部可以经由 ROM 12 和/或通信单元 19 而被载入和/或安装到电子设备 10 上。当计算机程序加载到 RAM 13 并由处理器 11 执行时，可以执行上文描述的模型推理服务部署方法的一个或多个步骤。备选地，在其他实施例中，处理器 11 可以通过其他任何适当的方式（例如，借助于固件）而被配置

为执行模型推理服务部署方法。

本文中以上描述的系统和技术的多种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列（Field-Programmable Gate Array，FPGA）、专用集成电路（Application Specific Integrated Circuit，ASIC）、专用标准产品（Application Specific Standard Parts，ASSP）、芯片上系统的系统（System on Chip，SOC）、负载可编程逻辑设备（Complex Programmable Logic Device，CPLD）、计算机硬件、固件、软件、和/或它们的组合中实现。这些多种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

用于实施本申请的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器，使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

在本申请的上下文中，计算机可读存储介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。备选地，计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（Erasable Programmable Read-Only Memory，EPROM）或快闪存储器、光纤、便捷式紧凑盘只读存储器（Compact Disc Read-Only Memory，CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

为了提供与用户的交互，可以在电子设备上实施此处描述的系统和技术，该电子设备具有：用于向用户显示信息的显示装置（例如，阴极射线管（Cathode Ray Tube，CRT）或者液晶显示器（Liquid Crystal Display，LCD）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或

者、触觉输入）来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（Local Area Network，LAN）、广域网（Wide Area Network，WAN）、区块链网络和互联网。

计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器，又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决了传统物理主机与虚拟专用服务器（VirtualPrivateServer，VPS）服务中，存在的管理难度大，业务扩展性弱的缺陷。

应该理解，可以使用上面所示的多种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的多个步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请的技术方案所期望的结果，本文在此不进行限制。

1.一种模型推理服务部署方法，包括：

在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数；

针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式；

基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

2.根据权利要求 1 所述的方法，其中，所述模式确定阈值的属性至少包括：以时间周期为窗口的平均请求数、以百分位为窗口的平均响应时间、以时间周期为窗口的最长响应时间、在指定时间内总是使用默认调度方式，以及在指定时间内总是不使用默认调度方式。

3.根据权利要求 1 所述的方法，其中，所述根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，包括：

将所述推理指标参数和所述模式确定阈值进行对比处理，根据所述推理指标参数与所述模式确定阈值的匹配结果，确定所述模型推理服务对应的目标调度模式。

4.根据权利要求 1 所述的方法，还包括：

在所述目标调度模式为服务质量优先模式的情况下，所述服务部署方式包括独立部署方式和联合部署方式，其中，所述独立部署方式为单个模型推理服务独自占据所述计算资源集群中的一个计算资源，所述联合部署方式为使用同一个推理服务框架的相同推理类型的多个模型推理服务，以共享上下文方式占据所述计算资源集群中的同一个计算资源；

在所述目标调度模式为资源利用率优先模式的情况下，所述服务部署方式包括联合部署方式和卡共享部署方式，其中，所述卡共享部署方式为多个模型推理服务以时间片轮转方式轮转占据所述计算资源集群中的同一个计算资源。

5.根据权利要求 4 所述的方法，其中，所述基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，包括：

在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式；

在所述目标调度模式为资源利用率优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式。

6.根据权利要求5所述的方法，其中，所述基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式，包括：

在所述推理指标参数大于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为独立部署方式；

在所述推理指标参数小于或等于所述部署方式阈值的情况下，确定所述模型推理服务对应的服务部署方式为联合部署方式。

7.根据权利要求5所述的方法，其中，所述基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式，包括：

基于贪心算法，参照最大化资源利用率的调度策略，分别确定所述联合部署方式和所述卡共享部署方式在所述服务部署方式中的部署占比，

或者，

分别计算采用所述联合部署方式时对应的第一资源利用率，以及采用所述卡共享部署方式时对应的第二资源利用率，在所述第一资源利用率大于所述第二资源利用率的情况下采用所述联合部署方式，在所述第一资源利用率小于或等于所述第二资源利用率的情况下采用所述卡共享部署方式。

8.一种模型推理服务部署装置，包括：

推理指标参数获取模块，设置为在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数；

目标调度模式确定模块，设置为针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式；

服务部署方式确定模块，设置为基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署。

9.一种电子设备，其包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算

机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求 1-7 中任一项所述的模型推理服务部署方法。

10.一种计算机可读存储介质，存储有计算机指令，所述计算机指令用于使处理器执行时实现权利要求 1-7 中任一项所述的模型推理服务部署方法。

一种模型推理服务部署方法、装置、设备及存储介质。该方法包括，在模型推理过程中,获取计算资源集群中所有模型推理服务的推理指标参数(S101)。针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的各个模式确定阈值，确定所述模型推理服务对应的目标调度模式，其中，所述目标调度模式包括服务质量优先模式和资源利用率优先模式（S102）。基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署（S103）。

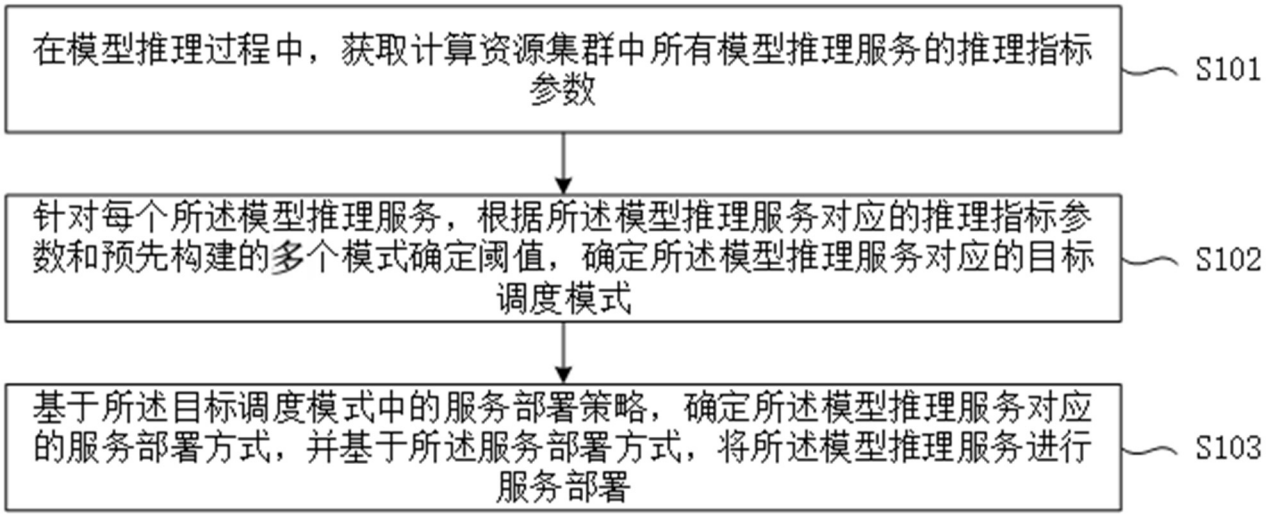

在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数
S101
针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式
S102
基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署
S103

图 1

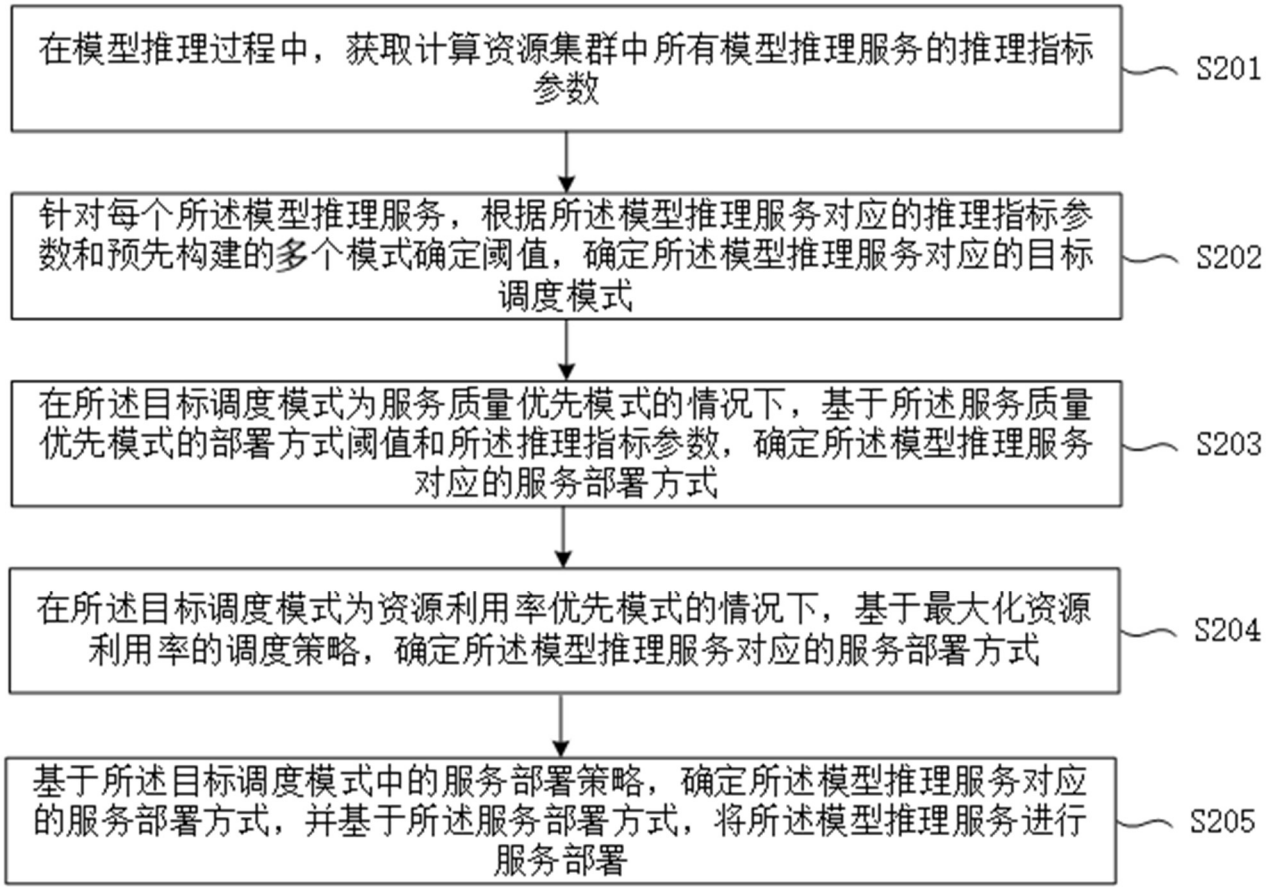

在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数
S201
针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式
S202
在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式
S203
在所述目标调度模式为资源利用率优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式
S204
基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署
S205

图 2


301
302
303
推理指标参数获取模块
目标调度模式确定模块
服务部署方式确定模块

图 3

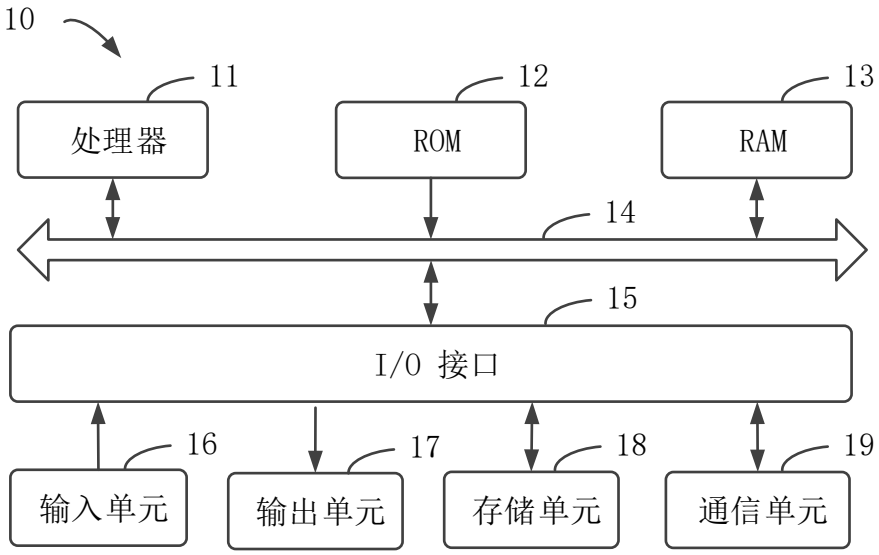

10
11
12
13
处理器
ROM
RAM
14
15
I/O 接口
16
17
18
19
输入单元
输出单元
存储单元
通信单元

图 4