(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115225497 A
(43)申请公布日 2022.10.21

(21)申请号 202210827417.X

(22)申请日 2022.07.13

(71)申请人 上海壁仞智能科技有限公司
地址 201100 上海市闵行区陈行公路2388号16幢13层1302室

(72)发明人 不公告发明人

(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 万里晴

(51)Int.Cl.
H04L 41/0823(2022.01)
H04L 41/0893(2022.01)

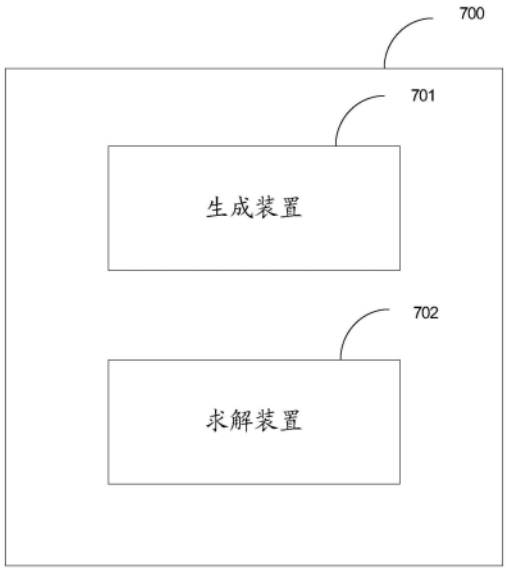
权利要求书3页 说明书18页 附图10页

(54)发明名称

优化策略的服务器、客户端、系统、方法、设备和介质

(57)摘要

提供用于优化策略的服务器、客户端、系统、方法、电子设备和非暂时存储介质。策略服务器，包括：生成装置，被配置为对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；求解装置，被配置为基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略。



CN 115225497 A

1.一种用于求解人工智能模型的优化策略的策略服务器,包括:

生成装置,被配置为对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略,并分配给至少一个计算设备来运行;

求解装置,被配置为基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据,求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略。

2.根据权利要求1所述的策略服务器,还包括:接收装置,被配置为接收来自客户端的查询最优模型优化策略的请求以及所述客户端的属性。

3.根据权利要求2所述的策略服务器,其中,在策略服务器已经求解出适合于所述客户端的当前最优模型优化策略的情况下,所述策略服务器的求解装置被配置为向所述至少一个计算设备发送适合于所述客户端的当前最优模型优化策略。

4.根据权利要求2所述的策略服务器,其中,在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略的情况下:所述生成装置被配置为基于所述请求,向所述客户端分配所述一个或多个待验证模型优化策略中的一个或多个待验证策略以便所述客户端运行通过所述一个或多个待验证策略所优化的一个或多个待验证优化模型,所述客户端被配置为向所述策略服务器发送运行所述一个或多个待验证优化模型得到的性能数据。

5.根据权利要求4所述的策略服务器,其中,所述求解装置被配置为基于所述客户端发送的客户端的属性和得到的性能数据,确定所述一个或多个待验证策略中是否存在适合于所述客户端的当前最优模型优化策略。

6.根据权利要求4所述的策略服务器,其中,所述求解装置被配置为在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解,且在存储器中缓存计算设备的当前最优模型优化策略、与之相关联的当前时间以及所述部分性能数据。

7.根据权利要求1所述的策略服务器,其中,所述生成装置被配置为向不同的客户端分配不同的待验证模型优化策略以便所述求解装置确定客户端的属性是否适合于运行被分配的待验证模型优化策略中的算子改变。

8.根据权利要求7所述的策略服务器,其中,如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变,所述生成装置被配置为向具有该属性的客户端分配具有该算子改变的待验证模型优化策略。

9.根据权利要求1所述的策略服务器,其中,所述性能数据包括运行时间、时钟数、内存用量、数据传输量、出错数量、温度中的至少一个,所述计算设备的属性包括:计算设备的芯片的型号、芯片的类型、计算设备的芯片内存量、支持带宽、芯片提供的卷积运算量的能力、计算设备的计算核的组成中的至少一个,其中,所述策略服务器被配置为将使得至少一个计算设备的性能数据最优的待验证模型优化策略确定为适合于具有属性的至少一个计算设备的当前最优模型优化策略。

10.根据权利要求1所述的策略服务器,其中,所述生成装置被配置为:对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成所述一个或多个待验证模型优化策略。

11.根据权利要求1所述的策略服务器,其中,所述生成装置被配置为基于所述至少一

个计算设备的属性，从生成的所述一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。

12.根据权利要求1所述的策略服务器，其中，所述求解装置被配置为：定时进行所述求解步骤；或响应于接收到新的性能数据，进行所述求解步骤；或以上两者的结合。

13.一种用于求解人工智能模型的优化策略的方法，包括：

对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；

基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略。

14.一种用于应用人工智能模型的优化策略的客户端，包括：

发送装置，被配置为向策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性；

应用装置，被配置为从所述策略服务器接收策略服务器根据所述客户端的属性而向所述客户端发送的适合于所述客户端的模型优化策略，并应用所述模型优化策略。

15.根据权利要求14所述的客户端，其中，

所述客户端还包括接收装置，被配置为接收从所述策略服务器分配的一个或多个待验证策略，且所述客户端的应用装置运行通过所述一个或多个待验证策略所优化的一个或多个待验证优化模型，

所述发送装置被配置为向所述策略服务器发送运行所述一个或多个待验证优化模型得到的性能数据，

所述策略服务器被配置为基于所述客户端发送的所述客户端的属性和得到的性能数据，确定所述一个或多个待验证策略中是否存在适合于所述客户端的当前最优模型优化策略，

或者所述客户端被配置为在被设置为保密的情况下不向所述策略服务器发送运行所述一个或多个待验证策略得到的性能数据。

16.根据权利要求14所述的客户端，其中，

所述客户端还包括确定装置，被配置为在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化策略或不为所述客户端分配待验证模型优化策略的情况下：确定是否缓存了先前为所述客户端分配的时间上最近的模型优化策略，且在确定缓存的情况下，所述应用装置被配置为应用缓存的所述最近的模型优化策略，且在确定未缓存的情况下，所述发送装置被配置为向所述策略服务器发送先前缓存的性能数据。

17.一种用于应用人工智能模型的优化策略的方法，包括：

向策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性；

从所述策略服务器接收策略服务器根据所述客户端的属性而向所述客户端发送的适合于所述客户端的模型优化策略，并应用所述模型优化策略。

18.一种用于求解和应用人工智能模型的优化策略的系统，包括：

策略服务器，被配置为：

对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；

基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略；

客户端，被配置为向所述策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性，其中，所述策略服务器被配置为根据所述客户端的属性而向所述客户端发送适合于所述客户端的模型优化策略。

19.一种用于求解和应用人工智能模型的优化策略的方法，包括：

由策略服务器：

对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；

基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略；

由客户端：

向所述策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性，其中，所述策略服务器被配置为根据所述客户端的属性而向所述客户端发送适合于所述客户端的模型优化策略。

20.一种电子设备，包括：

存储器，用于存储指令；

处理器，用于读取所述存储器中的指令，并执行如权利要求13、17、19中任一项所述的方法。

21.一种非暂时存储介质，其上存储有指令，

其中，所述指令在被处理器读取时，使得所述处理器执行如权利要求13、17、19中任一项所述的方法。

优化策略的服务器、客户端、系统、方法、设备和介质

技术领域

[0001] 本申请涉及人工智能领域，且更具体地，涉及用于求解和应用人工智能模型的优化策略的服务器、客户端、系统、方法、电子设备和非暂时存储介质。

背景技术

[0002] 人工智能算法模型通常是由多种算子组成。通常利用例如图形处理单元(graphics processing unit,GPU)来运行人工智能算法模型。为了加快GPU运行人工智能算法模型的速度，需要对人工智能算法模型进行优化。当前，这种优化包括静态策略优化和动态优化以提高人工智能算法模型的运行效率和性能。静态策略优化是在人工智能算法模型加载到GPU处理器上运行之前进行的对人工智能算法模型的优化。而动态优化是在人工智能算法模型加载后的运行时进行的。

[0003] 仍需要对人工智能算法模型进行优化以便提高人工智能算法模型在特定硬件、例如特定GPU上的运行效率和性能。

发明内容

[0004] 根据本申请的一个方面，提供一种用于求解人工智能模型的优化策略的策略服务器，包括：生成装置，被配置为对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；求解装置，被配置为基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略。

[0005] 在一个实施例中，策略服务器还包括：接收装置，被配置为接收来自客户端的查询最优模型优化策略的请求以及所述客户端的属性。

[0006] 在一个实施例中，在策略服务器已经求解出适合于所述客户端的当前最优模型优化策略的情况下，所述策略服务器的求解装置被配置为向所述至少一个计算设备发送适合于所述客户端的当前最优模型优化策略。

[0007] 在一个实施例中，在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略的情况下：所述生成装置被配置为基于所述请求，向所述客户端分配所述一个或多个待验证模型优化策略中的一个或多个待验证策略以便所述客户端运行通过所述一个或多个待验证策略所优化的一个或多个待验证优化模型，所述客户端被配置为向所述策略服务器发送运行所述一个或多个待验证优化模型得到的性能数据。

[0008] 在一个实施例中，所述求解装置被配置为基于所述客户端发送的客户端的属性和得到的性能数据，确定所述一个或多个待验证策略中是否存在适合于所述客户端的当前最优模型优化策略。

[0009] 在一个实施例中，所述求解装置被配置为在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解，且在存储器中缓存计算设备的当前最优模型优化策

略、与之相关联的当前时间以及所述部分性能数据。

[0010] 在一个实施例中，所述生成装置被配置为向不同的客户端分配不同的待验证模型优化策略以便所述求解装置确定客户端的属性是否适合于运行被分配的待验证模型优化策略中的算子改变。

[0011] 在一个实施例中，如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变，所述生成装置被配置为向具有该属性的客户端分配具有该算子改变的待验证模型优化策略。

[0012] 在一个实施例中，所述性能数据包括运行时间、时钟数、内存用量、数据传输量、出错数量、温度中的至少一个，所述计算设备的属性包括：计算设备的芯片的型号、芯片的类型、计算设备的芯片内存量、支持带宽、芯片提供的卷积运算量的能力、计算设备的计算核kernel的组成中的至少一个，其中，所述策略服务器被配置为将使得至少一个计算设备的性能数据最优的待验证模型优化策略确定为适合于具有属性的至少一个计算设备的当前最优模型优化策略。

[0013] 在一个实施例中，所述生成装置被配置为：对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成所述一个或多个待验证模型优化策略。

[0014] 在一个实施例中，所述生成装置被配置为基于所述至少一个计算设备的属性，从生成的所述一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。

[0015] 在一个实施例中，所述求解装置被配置为：定时进行所述求解步骤；或响应于接收到新的性能数据，进行所述求解步骤；或以上两者的结合。

[0016] 根据本申请的另一方面，提供一种用于求解人工智能模型的优化策略的方法，包括：对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略。

[0017] 根据本申请的另一方面，提供一种用于应用人工智能模型的优化策略的客户端，包括：发送装置，被配置为向策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性；应用装置，被配置为从所述策略服务器接收策略服务器根据所述客户端的属性而向所述客户端发送的适合于所述客户端的模型优化策略，并应用所述模型优化策略。

[0018] 在一个实施例中，所述客户端还包括接收装置，被配置为接收从所述策略服务器分配的一个或多个待验证策略，且所述客户端的应用装置运行通过所述一个或多个待验证策略所优化的一个或多个待验证优化模型，所述发送装置被配置为向所述策略服务器发送运行所述一个或多个待验证优化模型得到的性能数据，所述策略服务器被配置为基于所述客户端发送的所述客户端的属性和得到的性能数据，确定所述一个或多个待验证策略中是否存在适合于所述客户端的当前最优模型优化策略，或者所述客户端被配置为在被设置为保密的情况下不向所述策略服务器发送运行所述一个或多个待验证策略得到的性能数据。

[0019] 在一个实施例中，所述客户端还包括确定装置，被配置为在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化

策略或不为所述客户端分配待验证模型优化策略的情况下:确定是否缓存了先前为所述客户端分配的时间上最近的模型优化策略,且在确定缓存的情况下,所述应用装置被配置为应用缓存的所述最近的模型优化策略,且在确定未缓存的情况下,所述发送装置被配置为向所述策略服务器发送先前缓存的性能数据。

[0020] 根据本申请的另一方面,提供一种用于应用人工智能模型的优化策略的方法,包括:向策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性;从所述策略服务器接收策略服务器根据所述客户端的属性而向所述客户端发送的适合于所述客户端的模型优化策略,并应用所述模型优化策略。

[0021] 根据本申请的另一方面,提供一种用于求解和应用人工智能模型的优化策略的系统,包括:策略服务器,被配置为:对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略,并分配给至少一个计算设备来运行;基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据,求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略;客户端,被配置为向所述策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性,其中,所述策略服务器被配置为根据所述客户端的属性而向所述客户端发送适合于所述客户端的模型优化策略。

[0022] 在一个实施例中,在所述策略服务器已经求解出适合于所述客户端的当前最优模型优化策略的情况下,所述策略服务器被配置为向所述客户端发送所述适合于所述客户端的当前最优模型优化策略。

[0023] 在一个实施例中,在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略的情况下:所述策略服务器被配置为基于所述请求,向所述客户端分配所述一个或多个待验证模型优化策略中的一个或多个待验证策略以便所述客户端运行通过所述一个或多个待验证策略所优化的一个或多个待验证优化模型,所述客户端被配置为向所述策略服务器发送运行所述一个或多个待验证优化模型得到的性能数据,所述策略服务器被配置为基于所述客户端发送的所述客户端的属性和得到的性能数据,确定所述一个或多个待验证策略中是否存在适合于所述客户端的当前最优模型优化策略。

[0024] 在一个实施例中,所述客户端被配置为在所述策略服务器还未求解出适合于所述客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化策略或不为所述客户端分配待验证模型优化策略的情况下:确定是否缓存了先前为所述客户端分配的时间上最近的模型优化策略,且在确定缓存的情况下,应用缓存的所述最近的模型优化策略,且在确定未缓存的情况下,向所述策略服务器发送先前缓存的性能数据。

[0025] 在一个实施例中,所述客户端被配置为在被设置为保密的情况下不向所述策略服务器发送运行所述一个或多个待验证策略得到的性能数据。

[0026] 在一个实施例中,所述策略服务器被配置为在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解,且在存储器中缓存所述计算设备的当前最优模型优化策略、与之相关联的当前时间以及所述部分性能数据。

[0027] 在一个实施例中,所述策略服务器被配置为向不同的客户端分配不同的待验证模型优化策略以便确定所述客户端的属性是否适合于运行被分配的所述待验证模型优化策

略中的算子改变。

[0028] 在一个实施例中，如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变，所述策略服务器被配置为向具有所述属性的所述客户端分配具有所述算子改变的待验证模型优化策略。

[0029] 在一个实施例中，所述性能数据包括运行时间、时钟数、内存用量、数据传输量、出错数量、温度中的至少一个，计算设备的属性包括：计算设备的芯片的型号、芯片的类型、计算设备的芯片内存量、支持带宽、芯片提供的卷积运算量的能力、计算设备的计算核kernel的组成中的至少一个，其中，所述策略服务器被配置为将使得所述至少一个计算设备的所述性能数据最优的待验证模型优化策略确定为适合于具有所述属性的所述至少一个计算设备的当前最优模型优化策略。

[0030] 在一个实施例中，所述策略服务器被配置为：对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成一个或多个待验证模型优化策略。

[0031] 在一个实施例中，所述策略服务器被配置为：基于所述至少一个计算设备的属性，从生成的一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。

[0032] 在一个实施例中，所述策略服务器被配置为：定时进行所述求解步骤；或响应于接收到新的性能数据，进行所述求解步骤；或以上两者的结合。

[0033] 根据本申请的另一方面，提供一种用于求解和应用人工智能模型的优化策略的方法，包括：由策略服务器：对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略；由客户端：向所述策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性，其中，所述策略服务器被配置为根据所述客户端的属性而向所述客户端发送适合于所述客户端的模型优化策略。

[0034] 根据本申请的一个方面，提供一种电子设备，包括：存储器，用于存储指令；处理器，用于读取所述存储器中的指令，并执行根据本申请的各个实施方式所述的方法。

[0035] 根据本申请的一个方面，提供一种非暂时存储介质，其上存储有指令，其中，所述指令在被处理器读取时，使得所述处理器执行根据本申请的各个实施方式所述的方法。

附图说明

[0036] 为了更清楚地说明本公开实施例或现有技术中的技术方案，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公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0037] 图1示出了根据本申请实施方式的用于求解和应用人工智能模型的优化策略的系统的示意方框图。

[0038] 图2A示出了根据本申请实施方式的策略服务器和客户端分别负责的功能。

[0039] 图2B示出了根据本申请实施方式的用于求解和应用人工智能模型的优化策略的

系统的功能步骤的示意流程图。

[0040] 图3示出了根据本申请实施方式的客户端处请求和获得优化策略的步骤的示意流程图。

[0041] 图4示出了根据本申请实施方式的求解和应用人工智能模型的优化策略的系统的应用场景示意图。

[0042] 图5示出了根据本申请实施方式的策略服务器处主要进行的离线优化策略求解服务和与客户端交互的优化策略请求服务的步骤的示意流程图。

[0043] 图6示出根据本申请实施方式的用于求解和应用人工智能模型的优化策略的方法的示意流程图。

[0044] 图7示出了根据本申请实施方式的一种用于求解人工智能模型的优化策略的策略服务器的方框图。

[0045] 图8示出了根据本申请实施方式的用于求解人工智能模型的优化策略的方法的示意流程图。

[0046] 图9示出了根据本申请实施方式的一种用于应用人工智能模型的优化策略的客户端的方框图。

[0047] 图10示出了根据本申请实施方式的用于应用人工智能模型的优化策略的方法的示意流程图。

[0048] 图11示出了适于用来实现本申请实施方式的示例性电子设备的框图。

[0049] 图12示出了根据本公开的实施例的非暂时性计算机可读存储介质的示意图。

具体实施方式

[0050] 现在将详细参照本申请的具体实施例，在附图中例示了本申请的例子。尽管将结合具体实施例描述本申请，但将理解，不是想要将本申请限于描述的实施例。相反，想要覆盖由所附权利要求限定的在本申请的精神和范围内包括的变更、修改和等价物。应注意，这里描述的方法步骤都可以由任何功能块或功能布置来实现，且任何功能块或功能布置可被实现为物理实体或逻辑实体、或者两者的组合。

[0051] 人工智能算法模型通常是神经网络模型，例如图像推理模型、语音推理模型等。输入一张带有动物的图像，经过神经网络模型，直接输出一个标签，表示这张图中的物体是什么，例如是狗或者猫。神经网络模型通常包括卷积神经网络模型。

[0052] 神经网络模型可以由算子(operator)组成。其中，算子指的是神经网络模型中各层所做的各种运算，例如神经网络模型的卷积层对神经网络模型的输入数据所做的卷积运算即为卷积算子。神经网络模型可以包括众多种类的算子，例如卷积算子、全连接算子、池化算子、转置算子、Sobel算子、reshape算子等等。

[0053] 在实际的神经网络模型的应用中，可以采用计算硬件来实现(计算)神经网络模型中的各个算子。比如对于上述卷积算子，可以采用图形处理单元(GPU)、中央处理器(central processing unit，CPU)、机器学习单元(machine learning unit)或者现场可编程阵列(field programmable gate array，FPGA)等硬件实现神经网络模型中的算子。

[0054] 各个算子的计算量有的大有的小，但总的来说，人工智能算法模型的计算量是庞大且耗时的，需要对人工智能算法模型的计算进行优化，以便节省时间、计算量和计算成

本。

[0055] 而且不同的计算硬件和不同的硬件属性的处理能力、可支持带宽等性能以及存储空间等都不同，同一种神经网络模型在不同的计算硬件上运行时的性能也可能不同。

[0056] 先前提到的人工智能算法模型的静态策略优化和动态优化的步骤都是在单机上执行的，导致优化的动作必须根据单机(例如计算硬件)的任务执行安排来执行，因此，整个优化的过程可能很费时且低效。而且这种优化都是客户端根据经验的固定的优化程序及策略，而这种优化策略不一定是对运行该人工智能算法模型的计算硬件的硬件性能等来说最优的，因此现有技术的优化也缺少灵活性和适应性。在要对优化策略进行升级时，需要专门的升级软件来对各个计算硬件的优化策略进行分别的升级。

[0057] 图1示出了根据本申请实施方式的用于求解和应用人工智能模型的优化策略的系统100的示意方框图。

[0058] 如图1所示，用于求解和应用人工智能模型的优化策略的系统100包括：策略服务器101，被配置为：对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；基于运行一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解一个或多个待验证模型优化策略中适合于至少一个计算设备的当前最优模型优化策略；客户端102，被配置为向策略服务器发送查询最优模型优化策略的请求以及客户端的属性，其中，策略服务器被配置为根据客户端的属性而向客户端发送适合于客户端的模型优化策略。

[0059] 如此，将优化策略的求解放到策略服务器101上离线进行，可以将模型编译过程中的优化策略求解和客户端处的优化后的模型编译和运行在一定程度上剥离解耦，简化两部分的实现，在策略服务器处集中计算，利用大算力离线求解最优化策略，减小客户端单独运行时求解对性能要求的约束，减少客户端处的模型优化负担。传统的模型动态优化的运行时计算可以被转换为客户端发起的查询和服务器处的静态优化，而查询和静态优化是能量友好的和效率高效的，因此本技术可以实现较低的能量消耗和高效率的策略优化。而且，策略服务器101可以离线地逐步从客户端102接收和完善性能数据及求解最优策略，在不影响客户端102的工作负担的情况下有更多的时间和概率求解到最优策略。另外，策略服务器101可以结合计算设备的特定属性和该计算设备运行模型得到的性能数据来求解适合于该计算设备的当前最优模型优化策略，从而以后为发出请求的客户端发送最适合该客户端的属性的最优模型优化策略。而且，由于策略服务器自行求解和升级当前最优模型优化策略，因此不需要为每个客户端进行模型优化策略升级，减少了大量工作量。策略服务器还可以统筹不同客户端的属性与不同模型优化策略之间的关系，了解哪些属性适合于运行哪种优化策略(算子改变)，从而在待验证优化策略向不同客户端的分配以及生成哪种待验证优化策略上做出协调。

[0060] 图2A示出了根据本申请实施方式的策略服务器和客户端分别负责的功能。

[0061] 策略服务器负责更新包括策略和性能数据的数据、离线求解最优策略，以及存储求解的最优解。而客户端不进行求解策略等动作，而进行从策略服务器获取优化策略以及向策略服务器上传性能数据，具体地，客户端获取策略服务器的优化结果并应用(运算)优化后的模型，在本地缓存获取到的优化策略，然后获取应用(运算)优化后的模型后的性能

数据并上传给策略服务器。

[0062] 可见,将优化策略的求解放到策略服务器101上离线进行,可以将模型编译过程中的优化策略求解和客户端处的优化后的模型编译和运行在一定程度上剥离解耦,简化两部分的实现,在策略服务器处集中计算,利用大算力离线求解最优化策略,减小客户端单独运行时求解对性能要求的约束,减少客户端处的模型优化负担。

[0063] 图2B示出了根据本申请实施方式的用于求解和应用人工智能模型的优化策略的系统100的功能步骤的示意流程图。

[0064] 如图2B所示,在服务端的策略服务器在201中对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略,并在202中分配给至少一个计算设备来运行。客户端此时可以在205中获得被分配的优化策略,并应用(运行)利用优化策略优化的人工智能模型。客户端可以在206中收集运行该优化的人工智能模型的性能数据。然后客户端在207中向策略服务器回传收集的该优化的人工智能模型的性能数据。策略服务器在接收到性能数据之后可以依据客户端的硬件信息(属性)与模型信息来将与该硬件信息(属性)和该模型信息(指示具有什么样的算子改变的哪个模型)相关的当前的性能数据更新为最新接收到的性能数据。策略服务器可以在203中基于运行一个或多个待验证模型优化策略的客户端的属性和性能数据,求解一个或多个待验证模型优化策略中适合于该客户端的当前最优模型优化策略,并能够将最优模型优化策略更新为当前最优模型优化策略,并可以更新待验证策略的集合(例如,表示其中一个待验证策略已经完成验证,从集合中删除该已完成验证的策略等等)。

[0065] 然后,如果客户端向策略服务器发送查询最优模型优化策略的请求以及客户端的属性,其中,策略服务器被配置为根据客户端的属性而向客户端发送适合于客户端的刚才更新的当前最优模型优化策略。

[0066] 在一个实施例中,为了生成一个或多个待验证模型优化策略,策略服务器可以被配置为:对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成一个或多个待验证模型优化策略。这里可以穷举所有的算子改变,即进行无差别的分支尝试,尝试各个算子的所有拆分和合并的可能性,以便生成所有可能的待验证模型优化策略。当然,在已知一些算子的改变(例如某两个算子的合并或融合)能够得到优化的结果时,可以先生成和验证具有这种改变的待验证模型优化策略。

[0067] 在一个实施例中,策略服务器被配置为:基于至少一个计算设备的属性,从生成的一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。例如,可以结合硬件特性(比如有无Tcore以及Tcore的大小等)对整个待验证模型优化策略的集合进行剔除,删除所有不适合的待验证模型优化策略,对剩余的可选择待验证模型优化策略进行随机选择来验证。如此,可以尽量地减少要验证的待验证模型优化策略的数量。

[0068] 图3示出了根据本申请实施方式的客户端处请求和获得优化策略的步骤的示意流程图。

[0069] 在301处,客户端可以向策略服务器发送查询和获取最优模型优化策略的请求以及客户端的属性。在策略服务器已经求解出适合于客户端的当前最优模型优化策略的情况下,策略服务器被配置为向客户端发送适合于客户端的当前最优模型优化策略。在302处,

客户端可以确认有适合于客户端的当前最优模型优化策略。在305处，客户端则应用该求解出的当前最优模型优化策略。

[0070] 在策略服务器还未求解出适合于客户端的当前最优模型优化策略的情况下：策略服务器被配置为基于请求，向客户端分配一个或多个待验证模型优化策略中的一个或多个待验证策略，以便客户端在303处，确认有待验证策略，在304处决定使用待验证策略，并在305处运行(或称应用)通过一个或多个待验证策略所优化的一个或多个待验证优化模型。

[0071] 在一个实施例中，客户可配置客户端的保密性。在308处判断客户端并未被设置为保密且能够回传性能数据的权利下，客户端被配置为在309处，向策略服务器发送运行一个或多个待验证优化模型得到的性能数据。如果客户端被配置为在被设置为保密的情况下不向策略服务器发送运行一个或多个待验证策略得到的性能数据，则在308处判断不回传性能数据，则客户端就不会回传性能数据，从而增加客户端的保密性。

[0072] 然后，策略服务器被配置为基于客户端发送的客户端的属性和得到的性能数据，确定一个或多个待验证策略中是否存在适合于客户端的当前最优模型优化策略(图3中未示出)。

[0073] 在一个实施例中，客户端被配置为在306处在策略服务器还未求解出适合于客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化策略或不为客户端分配待验证模型优化策略的情况下：确定是否缓存了先前为客户端分配的时间上最近的模型优化策略，在确定缓存的情况下，在305处，应用缓存的最近的模型优化策略。如果在确定没有缓存的最近的模型优化策略的情况下，在307处获取先前缓存的性能数据，例如客户端之前运行其他策略而得到的性能数据。之后，客户端可以在308处判断是否回传该先前缓存的性能数据。在策略服务器获得了客户端发送的该先前缓存的性能数据，之后可以根据该先前缓存的性能数据来对优化策略求解，以尽快得到求解出的适合于该客户端的优化策略，以便发送给客户端。

[0074] 如此，可以在客户端处复用模型优化策略，节省客户端重复计算模型带来的计算损耗。

[0075] 如果客户端既没有接收到当前求解的当前最优模型优化策略，也没有待验证策略，也没有缓存的最近的模型优化策略，则客户端可以继续等待策略服务器进行求解和分配，也可以结束。

[0076] 如此，客户端不进行具体优化算法的求解，仅进行优化策略的解析及应用的部分，这也减少在客户端处的软件升级。

[0077] 在一个实施例中，策略服务器被配置为在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解，且在存储器中缓存计算设备的当前最优模型优化策略、与之相关联的当前时间(例如时间戳)以及部分性能数据。从而，策略服务器可以收集多种类计算设备及多种计算工作运行各种模型时的(部分)性能数据，从而依据收集的性能数据对运行于各种计算设备上的计算模型进行动态优化，且可以用于共享复用。

[0078] 策略服务器可以根据收集的性能数据建立黑盒成本模型来计算优化模型中的算子要耗费多少时间周期，并根据性能数据求解最优的优化策略。如此，还可以提供性能估计数据为混合计算环境(如大规模集群)的任务调度提供参数。策略服务器还可以在逐步接收更多的性能数据时逐步求解更优的模型优化策略，且在客户端请求时发送当前部分求解的

最优策略，然后离线地继续基于收集的更多性能数据来进一步求解。

[0079] 在这里，性能数据可以包括：运行一个或多个待验证优化模型得到的运行时间、时钟数、内存用量、数据传输量、出错(bug)数量、计算设备温度等中的至少一种。总之，性能数据可以包括用来评估一种模型是否是最适合于在该计算设备上运行的各种参数。在大多数情况下，通常可以用执行时间来评估，例如，计算设备运行某个待验证优化模型的执行时间最短，则可以认为该待验证优化模型在该计算设备上是最优的，或最适合该计算设备。

[0080] 计算设备的属性可以包括：例如计算设备的芯片的型号、芯片(例如GPU芯片)的类型、计算设备的芯片内存量、支持带宽、芯片提供的卷积运算量的能力(5ms、10ms等)、计算设备的计算核kernel的组成等。也就是说，计算设备的属性通常是与计算设备的计算性能有关，尤其是与人工智能模型中的各类算子的计算能力有关。例如，不同GPU的特性不同可能导致模型优化策略不同，而相同GPU的优化策略可能相同或相近。

[0081] 策略服务器被配置为将使得至少一个计算设备的性能数据最优的待验证模型优化策略确定为适合于具有属性的至少一个计算设备的当前最优模型优化策略。如此，可以了解某种客户端的计算设备的芯片的属性适合哪种待验证模型优化策略中的哪些算子的优化运算。

[0082] 例如，在某个计算设备具有特定芯片型号、特定芯片的类型、特定计算设备的芯片内存量等时，某个待验证模型优化策略在该计算设备上运行的执行时间最短、时钟数最少、内存用量最少、数据传输量最少、出错数量最少(当然，只采用其中一个参数作为评价标准也可以)等，则可以判断该待验证模型优化策略是最适合于运行在该计算设备上的最优模型优化策略。

[0083] 在一个实施例中，性能数据包括运行时间、时钟数、内存用量、数据传输量中的至少一个，其中，策略服务器被配置为将使得至少一个计算设备的性能数据最优的待验证模型优化策略确定为适合于至少一个计算设备的当前最优模型优化策略。

[0084] 在一个实施例中，策略服务器被配置为向不同的客户端分配不同的待验证模型优化策略以便确定客户端的属性是否适合于运行被分配的待验证模型优化策略中的算子改变。如此，可以了解不同客户端的计算设备的不同芯片在运行不同待验证模型优化策略中的不同的改变后算子的性能，从而可以分析和推导计算设备的不同芯片的属性与不同的改变后算子之间的性能关系。例如，第一待验证模型优化策略是合并第一算子和第二算子。第二待验证模型优化策略是合并第二算子和第三算子。在第一计算设备的具有第一属性的第一GPU芯片上运行了合并第一算子和第二算子后的优化后的待验证模型的运行时间比在第二计算设备的具有第二属性的第二GPU芯片上运行了合并第二算子和第三算子后的优化后的待验证模型的运行时间短，则可以确定第一算子和第二算子合并的这种优化策略适合于在具有第一属性的第一GPU芯片上运行。从而，在一个实施例中，如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变，策略服务器被配置为向具有该属性的客户端分配具有该算子改变的待验证模型优化策略。例如，策略服务器可以在接收到来自具有第一属性的第一GPU芯片的第一计算设备的查询优化策略请求时，从一个或多个待验证模型优化策略中为其分配具有第一算子和第二算子合并的待验证优化策略，从而可以剔除不适合的待验证优化策略，以更高效地确定适合于客户端的当前最优模型优化策略。

[0085] 在一个实施例中，策略服务器可以在没有客户端的查询请求的情况下自动地从待

验证模型优化策略中求解当前最优模型优化策略。策略服务器被配置为:定时进行求解步骤;或响应于接收到新的性能数据,进行求解步骤;或以上两者的结合。例如,策略服务器可以每一小时进行上述求解,或者在响应于接收到某个计算设备发送的性能数据再进行求解。总之,策略服务器可以与客户端离线地、异步地、自动地进行当前最优模型优化策略的求解任务,如此可以使得模型优化策略的求解与性能数据收集、客户端的优化策略的实际应用分离,获得高效的优化效果。

[0086] 图4示出了根据本申请实施方式的求解和应用人工智能模型的优化策略的系统的应用场景示意图。

[0087] 如图4所示,该应用场景包括离线的策略优化服务集群,包括多个策略服务器401。每个策略服务器401完成如下服务:调度服务、优化策略请求服务、离线优化策略求解服务、数据访问接口服务。

[0088] 调度服务负责总的调度策略服务器内的各种服务的进行。

[0089] 优化策略请求服务负责接收来自客户端的查询优化策略的请求,查询策略数据库,并决定向客户端发送哪种优化策略。该请求也可能是来自客户端的获取性能测试任务的请求,此时,优化策略服务需要收集来自客户端的客户端运行优化策略后的性能数据来决定哪种优化策略最适合。优化策略请求服务还可以接收来自客户端的性能数据并持久化该性能数据,例如在持久化装置中存储该性能数据。

[0090] 离线优化策略求解服务负责依据来自客户端的性能数据使用优化策略探索算法,例如启发式、深度学习、动态规划等来进行离线计算获得适合于硬件和框架的优化策略,将求解出来的优化策略存储到优化策略数据库中。在该优化策略数据库也可以存储与该优化策略相关联的时间戳,例如求解出该优化策略的时间戳。离线优化策略求解服务还可以获得待验证策略并持久化这些待验证策略。离线优化策略求解服务还可以获得竞争失败策略(求解为不适合的策略,可组成优化策略黑名单)并存储到优化策略数据库中,用于在以后从待验证策略中剔除这些竞争失败策略,以优化计算剪枝。

[0091] 数据访问接口主要负责与持久化装置、客户端所连接的服务网关进行发送和接收。

[0092] 在持久化装置402中持久化存储性能数据(包括历史性能数据)、优化策略(包括已经获得的模型优化策略、优化策略黑名单、待验证优化策略)。注意,持久化可以指的是将数据保存到持久化存储,例如本地磁盘或本地分布式存储、云端存储等,以供后续读取。

[0093] 服务网关404连接客户端406。而客户端406可以包括多种种类的节点,例如推理服务节点、训练节点、性能测试节点、其他性能数据提供节点等。

[0094] 推理服务节点包括本地优化时优化器、本地优化模型缓存、性能数据收集模块、推理服务引擎。训练节点与推理服务节点一样可包括本地优化时优化器、本地优化模型缓存、性能数据收集模块、推理服务引擎。性能测试节点可以只包括性能数据收集模块和测试任务模块。其他性能数据提供节点提供其他性能数据。

[0095] 这里的节点是服务的逻辑概念,及一台物理服务器可以同时为作为几种不同类型的节点的一个或多个客户端服务。

[0096] 本地运行时优化器负责根据从策略服务器获得的优化策略对模型进行优化。本地优化模型缓存负责缓存该优化策略。

[0097] 推理服务引擎加载待验证的优化后模型，进行静态优化，动态优化，内存使用规划，推理任务调度等。

[0098] 性能数据收集模块负责收集运行待验证的优化后模型的性能数据，包括但不限于运行时间、时钟数、内存用量、数据传输量、出错数量、温度等。

[0099] 图5示出了根据本申请实施方式的策略服务器处主要进行的离线优化策略求解服务和与客户端交互的优化策略请求服务的步骤的示意流程图。

[0100] 策略服务器进行求解的触发条件可以是：定时进行求解步骤；或响应于接收到新的性能数据，进行求解步骤；或以上两者的结合。图5示出了以接收到新的性能数据作为触发求解步骤的一个例子。

[0101] 如图5所示，为了进行离线优化策略求解服务，在501处，策略服务器确定是否接收到新的性能数据并更新了该性能数据。如果还没有，则在502处继续定时等待，定时到期之后继续确定是否接收到新的性能数据并更新了该性能数据。也就是说，仍然是以接收到新的性能数据作为触发求解步骤。

[0102] 如果在501处，策略服务器确定接收到新的性能数据并更新了该性能数据，则触发求解步骤。具体地，在503处，策略服务器获取更新的性能数据以及对应的模型的元数据(例如计算设备的硬件信息(属性)、模型信息、被优化的算子(例如改变了哪些算子)、优化方法等)。在504处，策略服务器获取相应的已有的优化策略、数据剪枝策略等，并根据各种信息，来调用优化策略求解算法进行策略求解和最优优化策略更新。

[0103] 如图5所示，为了进行与客户端交互的优化策略请求服务，在505中，策略服务器接收客户端的优化策略查询请求，在506中，策略服务器可以检索优化策略数据库中的优化策略，如果已有当前最优优化策略，则在506中确定当前最优优化策略，如果没有最优优化策略，则生成待验证策略，在507中为不同的客户端分配不同的待验证策略。在508中，策略服务器将确定的当前最优优化策略或待验证策略返回给请求的客户端。

[0104] 注意，本文中提到的服务器和客户端都可以作为应用程序存在，它们可以运行在于不同的物理机器上，也可以共存在同一台的物理机器上，在此并不做限制。

[0105] 图6示出根据本申请实施方式的用于求解和应用人工智能模型的优化策略的方法600的示意流程图。

[0106] 如图6所示，一种用于求解和应用人工智能模型的优化策略的方法，包括：由策略服务器：步骤601，对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；步骤602，基于运行一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解一个或多个待验证模型优化策略中适合于至少一个计算设备的当前最优模型优化策略；步骤603，由客户端：向策略服务器发送查询最优模型优化策略的请求以及客户端的属性，其中，策略服务器被配置为根据客户端的属性而向客户端发送适合于客户端的模型优化策略。

[0107] 在一个实施例中，方法600包括：在策略服务器已经求解出适合于客户端的当前最优模型优化策略的情况下，由策略服务器向客户端发送适合于客户端的当前最优模型优化策略。

[0108] 在一个实施例中，方法600包括：在策略服务器还未求解出适合于客户端的当前最

优模型优化策略的情况下:由策略服务器基于请求,向客户端分配一个或多个待验证模型优化策略中的一个或多个待验证策略以便客户端运行通过一个或多个待验证策略所优化的一个或多个待验证优化模型,由客户端向策略服务器发送运行一个或多个待验证优化模型得到的性能数据,由策略服务器基于客户端发送的客户端的属性和得到的性能数据,确定一个或多个待验证策略中是否存在适合于客户端的当前最优模型优化策略。

[0109] 在一个实施例中,方法600包括:由客户端在策略服务器还未求解出适合于客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化策略或不为客户端分配待验证模型优化策略的情况下:确定是否缓存了先前为客户端分配的时间上最近的模型优化策略,且在确定缓存的情况下,应用缓存的最近的模型优化策略,且在确定未缓存的情况下,向策略服务器发送先前缓存的性能数据。

[0110] 在一个实施例中,方法600包括:由客户端在被设置为保密的情况下不向策略服务器发送运行一个或多个待验证策略得到的性能数据。

[0111] 在一个实施例中,方法600包括:由策略服务器在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解,且在存储器中缓存计算设备的当前最优模型优化策略、与之相关联的当前时间以及部分性能数据。

[0112] 在一个实施例中,方法600包括:由策略服务器向不同的客户端分配不同的待验证模型优化策略以便确定客户端的属性是否适合于运行被分配的待验证模型优化策略中的算子改变。

[0113] 在一个实施例中,方法600包括:由策略服务器被配置为如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变,向具有属性的客户端分配具有算子改变的待验证模型优化策略。

[0114] 在一个实施例中,方法600包括:由策略服务器对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成一个或多个待验证模型优化策略。

[0115] 在一个实施例中,方法600包括:由策略服务器基于至少一个计算设备的属性,从生成的一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。

[0116] 在一个实施例中,方法600包括:由策略服务器:定时进行求解步骤;或响应于接收到新的性能数据,进行求解步骤;或以上两者的结合。

[0117] 图7示出了根据本申请实施方式的一种用于求解人工智能模型的优化策略的策略服务器700的方框图。

[0118] 该策略服务器700包括:生成装置701,被配置为对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略,并分配给至少一个计算设备来运行;求解装置702,被配置为基于运行一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据,求解一个或多个待验证模型优化策略中适合于至少一个计算设备的当前最优模型优化策略。

[0119] 如此,将优化策略的求解放到策略服务器上离线进行,计算设备仅设计运行优化后的模型以及反馈性能数据,简化两部分的实现。

[0120] 策略服务器700还可以包括接收装置,被配置为接收来自客户端的查询最优模型

优化策略的请求以及客户端的属性。

[0121] 在一个实施例中，在策略服务器700已经求解出适合于所述客户端的当前最优模型优化策略的情况下，策略服务器700的求解装置被配置为向至少一个计算设备发送适合于客户端的当前最优模型优化策略。

[0122] 在一个实施例中，在策略服务器700还未求解出适合于客户端的当前最优模型优化策略的情况下：策略服务器700的生成装置701被配置为基于请求，向客户端分配一个或多个待验证模型优化策略中的一个或多个待验证策略以便客户端运行通过一个或多个待验证策略所优化的一个或多个待验证优化模型，客户端被配置为向策略服务器发送运行一个或多个待验证优化模型得到的性能数据。

[0123] 在一个实施例中，策略服务器700的求解装置702被配置为基于客户端发送的客户端的属性和得到的性能数据，确定一个或多个待验证策略中是否存在适合于客户端的当前最优模型优化策略。

[0124] 在一个实施例中，策略服务器700的求解装置702被配置为在接收到部分性能数据的情况下部分地进行当前最优模型优化策略的求解，且在存储器中缓存计算设备的当前最优模型优化策略、与之相关联的当前时间以及部分性能数据。

[0125] 策略服务器700的生成装置701被配置为向不同的客户端分配不同的待验证模型优化策略以便求解装置702确定客户端的属性是否适合于运行被分配的待验证模型优化策略中的算子改变。

[0126] 如果确定客户端的属性适合于运行被分配的待验证模型优化策略中的算子改变，策略服务器700的生成装置701被配置为向具有该属性的客户端分配具有该算子改变的待验证模型优化策略。

[0127] 在一个实施例中，性能数据包括运行时间、时钟数、内存用量、数据传输量、出错数量、温度中的至少一个，计算设备的属性包括：计算设备的芯片的型号、芯片的类型、计算设备的芯片内存量、支持带宽、芯片提供的卷积运算量的能力、计算设备的计算核kernel的组成中的至少一个，其中，策略服务器被配置为将使得至少一个计算设备的性能数据最优的待验证模型优化策略确定为适合于具有属性的至少一个计算设备的当前最优模型优化策略。

[0128] 策略服务器700的生成装置701被配置为：对原始人工智能模型中的一个或多个算子进行算子类型转换、算子替换、算子拆分、算子合并中的一种或多种改变以生成一个或多个待验证模型优化策略。

[0129] 策略服务器700的生成装置701被配置为基于至少一个计算设备的属性，从生成的一个或多个待验证模型优化策略中剔除不适合的待验证模型优化策略。

[0130] 策略服务器700的求解装置702被配置为：定时进行求解步骤；或响应于接收到新的性能数据，进行求解步骤；或以上两者的结合。

[0131] 如此，将优化策略的求解放到策略服务器上离线进行，可以将模型编译过程中的优化策略求解和客户端处的优化后的模型编译和运行在一定程度上剥离解耦，简化两部分的实现，在策略服务器处集中计算，利用大算力离线求解最优化策略，减小客户端单独运行时求解对性能要求的约束，减少客户端处的模型优化负担。传统的模型动态优化的运行时计算可以被转换为客户端发起的查询和服务器处的静态优化，而查询和静态优化是能量友

好的和效率高效的，因此本技术可以实现较低的能量消耗和高效率的策略优化。而且，策略服务器可以离线地逐步从客户端接收和完善性能数据及求解最优策略，在不影响客户端的工作负担的情况下有更多的时间和概率求解到最优策略。另外，策略服务器可以结合计算设备的特定属性和该计算设备运行模型得到的性能数据来求解适合于该计算设备的当前最优模型优化策略，从而以后为发出请求的客户端发送最适合该客户端的属性的最优模型优化策略。而且，由于策略服务器自行求解和升级当前最优模型优化策略，因此不需要为每个客户端进行模型优化策略升级，减少了大量工作量。策略服务器还可以统筹不同客户端的属性与不同模型优化策略之间的关系，了解哪些属性适合于运行哪种优化策略(算子改变)，从而在待验证优化策略向不同客户端的分配以及生成哪种待验证优化策略上做出协调。

[0132] 图8示出了根据本申请实施方式的用于求解人工智能模型的优化策略的方法800的示意流程图。

[0133] 如图8所示，方法800包括：步骤801，对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行；步骤802，基于运行一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解一个或多个待验证模型优化策略中适合于至少一个计算设备的当前最优模型优化策略。

[0134] 如此，将优化策略的求解放到策略服务器上离线进行，计算设备仅设计运行优化后的模型以及反馈性能数据，简化两部分的实现。

[0135] 图9示出了根据本申请实施方式的一种用于应用人工智能模型的优化策略的客户端900的方框图。

[0136] 该客户端900包括：发送装置901，被配置为向策略服务器发送查询最优模型优化策略的请求以及客户端的属性；应用装置902，被配置为从策略服务器接收策略服务器根据客户端的属性而向客户端发送的适合于客户端的模型优化策略，并应用该模型优化策略。

[0137] 在一个实施例中，客户端还包括接收装置，被配置为接收从策略服务器分配的一个或多个待验证策略，且客户端的应用装置运行通过一个或多个待验证策略所优化的一个或多个待验证优化模型。

[0138] 在一个实施例中，客户端还包括发送装置，被配置为向策略服务器发送运行一个或多个待验证优化模型得到的性能数据。

[0139] 在一个实施例中，策略服务器被配置为基于客户端发送的客户端的属性和得到的性能数据，确定一个或多个待验证策略中是否存在适合于客户端的当前最优模型优化策略。

[0140] 或者，在一个实施例中，客户端被配置为在被设置为保密的情况下不向策略服务器发送运行一个或多个待验证策略得到的性能数据。

[0141] 在一个实施例中，客户端还包括确定装置，被配置为在策略服务器还未求解出适合于客户端的当前最优模型优化策略、且也没有一个或多个待验证模型优化策略或不为客户端分配待验证模型优化策略的情况下：确定是否缓存了先前为客户端分配的时间上最近的模型优化策略，且在确定缓存的情况下，应用装置被配置为应用缓存的最近的模型优化策略，且在确定未缓存的情况下，发送装置被配置为向策略服务器发送先前缓存的性能数

据。

[0142] 如此,客户端只需要在需要优化策略时向策略服务器请求即可。而在策略服务器处集中计算,利用大算力离线求解最优化策略,减小客户端单独运行时求解对性能要求的约束,减少客户端处的模型优化负担。传统的模型动态优化的运行时计算可以被转换为客户端发起的查询和服务器处的静态优化,而查询和静态优化是能量友好的和效率高效的,因此本技术可以实现较低的能量消耗和高效率的策略优化。因此也不需要为每个客户端进行模型优化策略升级,减少了大量工作量。

[0143] 图10示出了根据本申请实施方式的用于应用人工智能模型的优化策略的方法1000的示意流程图。

[0144] 该方法1000包括:步骤1001,向策略服务器发送查询最优模型优化策略的请求以及客户端的属性;步骤1002,从策略服务器接收策略服务器根据客户端的属性而向客户端发送的适合于客户端的模型优化策略,并应用该模型优化策略。

[0145] 如此,在策略服务器处集中计算,利用大算力离线求解最优化策略,减小客户端单独运行时求解对性能要求的约束,减少客户端处的模型优化负担。传统的模型动态优化的运行时计算可以被转换为客户端发起的查询和服务器处的静态优化,而查询和静态优化是能量友好的和效率高效的,因此本技术可以实现较低的能量消耗和高效率的策略优化。因此也不需要为每个客户端进行模型优化策略升级,减少了大量工作量。

[0146] 图11示出了适于用来实现本申请实施方式的示例性电子设备的框图。

[0147] 电子设备可以包括处理器(H1);存储介质(H2),耦合于处理器(H1),且在其中存储计算机可执行指令,用于在由处理器执行时进行本申请的实施例的各个方法的步骤。

[0148] 处理器(H1)可以包括但不限于例如一个或者多个处理器或者或微处理器等。

[0149] 存储介质(H2)可以包括但不限于例如,随机存取存储器(RAM)、只读存储器(ROM)、快闪存储器、EPROM存储器、EEPROM存储器、寄存器、计算机存储介质(例如硬碟、软碟、固态硬盘、可移动碟、CD-ROM、DVD-ROM、蓝光盘等)。

[0150] 除此之外,该电子设备还可以包括数据总线(H3)、输入/输出(I/O)总线(H4),显示器(H5)以及输入/输出设备(H6)(例如,键盘、鼠标、扬声器等)等。

[0151] 处理器(H1)可以通过I/O总线(H4)经由有线或无线网络(未示出)与外部设备(H5、H6等)通信。

[0152] 存储介质(H2)还可以存储至少一个计算机可执行指令,用于在由处理器(H1)运行时执行本技术所描述的实施例中的各个功能和/或方法的步骤。

[0153] 在一个实施例中,该至少一个计算机可执行指令也可以被编译为或组成一种软件产品,其中一个或多个计算机可执行指令被处理器运行时执行本技术所描述的实施例中的各个功能和/或方法的步骤。

[0154] 图12示出了根据本公开的实施例的非暂时性计算机可读存储介质的示意图。

[0155] 如图12所示,计算机可读存储介质1220上存储有指令,指令例如是计算机可读指令1210。当计算机可读指令1210由处理器运行时,可以执行参照以上描述的各个方法。计算机可读存储介质包括但不限于例如易失性存储器和/或非易失性存储器。易失性存储器例如可以包括随机存取存储器(RAM)和/或高速缓冲存储器(cache)等。非易失性存储器例如可以包括只读存储器(ROM)、硬盘、闪存等。例如,计算机可读存储介质1220可以连接于诸如

计算机等的计算设备，接着，在计算设备运行计算机可读存储介质1220上存储的计算机可读指令1210的情况下，可以进行如上描述的各个方法。

[0156] 当然，上述的具体实施例仅是例子而非限制，且本领域技术人员可以根据本申请的构思从上述分开描述的各个实施例中合并和组合一些步骤和装置来实现本申请的效果，这种合并和组合而成的实施例也被包括在本申请中，在此不一一描述这种合并和组合。

[0157] 注意，在本公开中提及的优点、优势、效果等仅是示例而非限制，不能认为这些优点、优势、效果等是本申请的各个实施例必须具备的。另外，上述公开的具体细节仅是为了示例的作用和便于理解的作用，而非限制，上述细节并不限制本申请为必须采用上述具体的细节来实现。

[0158] 本公开中涉及的器件、装置、设备、系统的方框图仅作为例示性的例子并且不意图要求或暗示必须按照方框图示出的方式进行连接、布置、配置。如本领域技术人员将认识到的，可以按任意方式连接、布置、配置这些器件、装置、设备、系统。诸如“包括”、“包含”、“具有”等等的词语是开放性词汇，指“包括但不限于”，且可与其互换使用。这里所使用的词汇“或”和“和”指词汇“和/或”，且可与其互换使用，除非上下文明确指示不是如此。这里所使用的词汇“诸如”指词组“诸如但不限于”，且可与其互换使用。

[0159] 本公开中的步骤流程图以及以上方法描述仅作为例示性的例子并且不意图要求或暗示必须按照给出的顺序进行各个实施例的步骤。如本领域技术人员将认识到的，可以按任意顺序进行以上实施例中的步骤的顺序。诸如“其后”、“然后”、“接下来”等等的词语不意图限制步骤的顺序；这些词语仅用于引导读者通读这些方法的描述。此外，例如使用冠词“一个”、“一”或者“该”对于单数的要素的任何引用不被解释为将该要素限制为单数。

[0160] 另外，本文中的各个实施例中的步骤和装置并非仅限定于某个实施例中实行，事实上，可以根据本申请的概念来结合本文中的各个实施例中相关的部分步骤和部分装置以构思新的实施例，而这些新的实施例也包括在本申请的范围内。

[0161] 以上描述的方法的各个操作可以通过能够进行相应的功能的任何适当的手段而进行。该手段可以包括各种硬件和/或软件组件和/或模块，包括但不限于硬件的电路、专用集成电路(ASIC)或处理器。

[0162] 可以利用被设计用于进行在此描述的功能的通用处理器、数字信号处理器(DSP)、ASIC、场可编程门阵列信号(FPGA)或其他可编程逻辑器件(PLD)、离散门或晶体管逻辑、离散的硬件组件或者其任意组合而实现或进行描述的各个例示的逻辑块、模块和电路。通用处理器可以是微处理器，但是作为替换，该处理器可以是任何商业上可获得的处理器、控制器、微控制器或状态机。处理器还可以实现为计算设备的组合，例如DSP和微处理器的组合，多个微处理器、与DSP核协作的微处理器或任何其他这样的配置。

[0163] 结合本公开描述的方法或算法的步骤可以直接嵌入在硬件中、处理器执行的软件模块中或者这两种的组合中。软件模块可以存在于任何形式的有形存储介质中。可以使用的存储介质的一些例子包括随机存取存储器(RAM)、只读存储器(ROM)、快闪存储器、EPROM存储器、EEPROM存储器、寄存器、硬碟、可移动碟、CD-ROM等。存储介质可以耦接到处理器以便该处理器可以从该存储介质读取信息以及向该存储介质写信息。在替换方式中，存储介质可以与处理器是整体的。软件模块可以是单个指令或者许多指令，并且可以分布在几个不同的代码段上、不同的程序之间以及跨过多个存储介质。

[0164] 在此公开的方法包括用于实现描述的方法的动作。方法和/或动作可以彼此互换而不脱离权利要求的范围。换句话说，除非指定了动作的具体顺序，否则可以修改具体动作的顺序和/或使用而不脱离权利要求的范围。

[0165] 上述功能可以按硬件、软件、固件或其任意组合而实现。如果以软件实现，功能可以作为指令存储在切实的计算机可读介质上。存储介质可以是可以由计算机访问的任何可用的切实介质。通过例子而不是限制，这样的计算机可读介质可以包括RAM、ROM、EEPROM、CD-ROM或其他光碟存储、磁碟存储或其他磁存储器件或者可以用于携带或存储指令或数据结构形式的期望的程序代码并且可以由计算机访问的任何其他切实介质。如在此使用的，碟(disk)和盘(disc)包括紧凑盘(CD)、激光盘、光盘、数字通用盘(DVD)、软碟和蓝光盘，其中碟通常磁地再现数据，而盘利用激光光学地再现数据。

[0166] 因此，计算机程序产品可以进行在此给出的操作。例如，这样的计算机程序产品可以是具有有形存储(和/或编码)在其上的指令的计算机可读的有形介质，该指令可由处理器执行以进行在此描述的操作。计算机程序产品可以包括包装的材料。

[0167] 软件或指令也可以通过传输介质而传输。例如，可以使用诸如同轴电缆、光纤光缆、双绞线、数字订户线(DSL)或诸如红外、无线电或微波的无线技术的传输介质从网站、服务器或者其他远程源传输软件。

[0168] 此外，用于进行在此描述的方法和技术的模块和/或其他适当的手段可以在适当时由用户终端和/或基站下载和/或其他方式获得。例如，这样的设备可以耦接到服务器以促进用于进行在此描述的方法的手段的传送。或者，在此描述的各种方法可以经由存储部件(例如RAM、ROM、诸如CD或软碟等的物理存储介质)提供，以便用户终端和/或基站可以在耦接到该设备或者向该设备提供存储部件时获得各种方法。此外，可以利用用于将在此描述的方法和技术提供给设备的任何其他适当的技术。

[0169] 其他例子和实现方式在本公开和所附权利要求的范围和精神内。例如，由于软件的本质，以上描述的功能可以使用由处理器、硬件、固件、硬连线或这些的任意的组合执行的软件实现。实现功能的特征也可以物理地位于各个位置，包括被分发以便功能的部分在不同的物理位置处实现。而且，如在此使用的，包括在权利要求中使用的，在以“至少一个”开始的项的列举中使用的“或”指示分离的列举，以便例如“A、B或C的至少一个”的列举意味着A或B或C，或AB或AC或BC，或ABC(即A和B和C)。此外，措辞“示例的”不意味着描述的例子是优选的或者比其他例子更好。

[0170] 可以不脱离由所附权利要求定义的教导的技术而进行对在此描述的技术的各种改变、替换和更改。此外，本公开的权利要求的范围不限于以上描述的处理、机器、制造、事件的组成、手段、方法和动作的具体方面。可以利用与在此描述的相应方面进行基本相同的功能或者实现基本相同的结果的当前存在的或者稍后要开发的处理、机器、制造、事件的组成、手段、方法或动作。因而，所附权利要求包括在其范围内的这样的处理、机器、制造、事件的组成、手段、方法或动作。

[0171] 提供所公开的方面的以上描述以使本领域的任何技术人员能够做出或者使用本申请。对这些方面的各种修改对于本领域技术人员而言是非常显而易见的，并且在此定义的一般原理可以应用于其他方面而不脱离本申请的范围。因此，本申请不意图被限制到在此示出的方面，而是按照与在此公开的原理和新颖的特征一致的最宽范围。

[0172] 为了例示和描述的目的已经给出了以上描述。此外，此描述不意图将本申请的实施例限制到在此公开的形式。尽管以上已经讨论了多个示例方面和实施例，但是本领域技术人员将认识到其某些变型、修改、改变、添加和子组合。

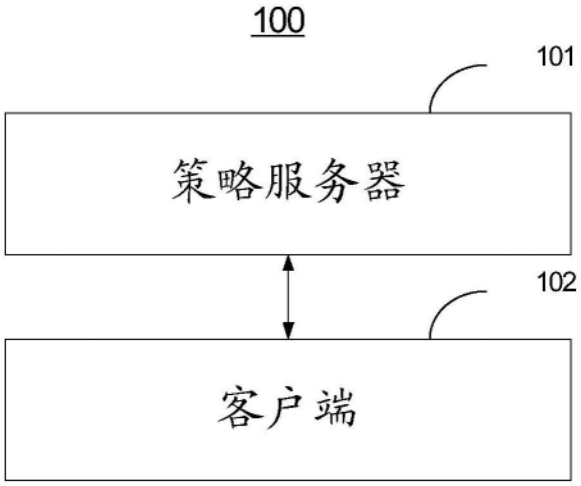

100
101
策略服务器
102
客户端

图1

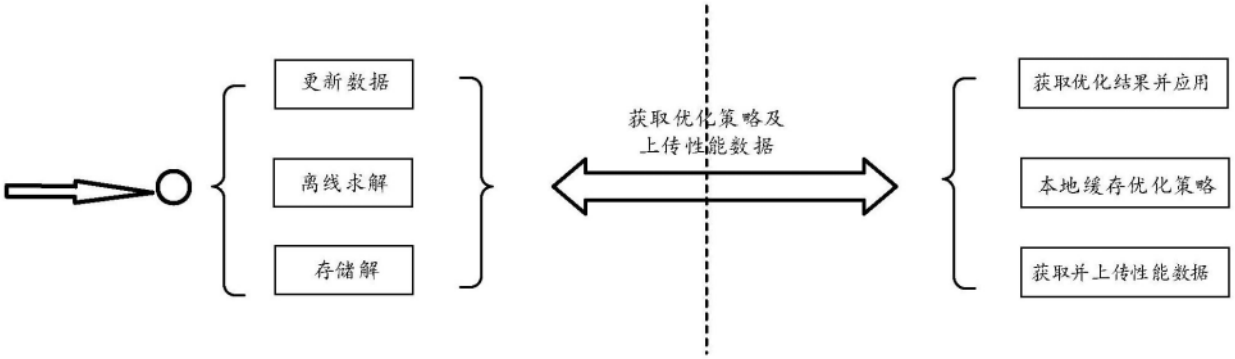

更新数据
离线求解
存储解
获取优化策略及
上传性能数据
获取优化结果并应用
本地缓存优化策略
获取并上传性能数据

图2A

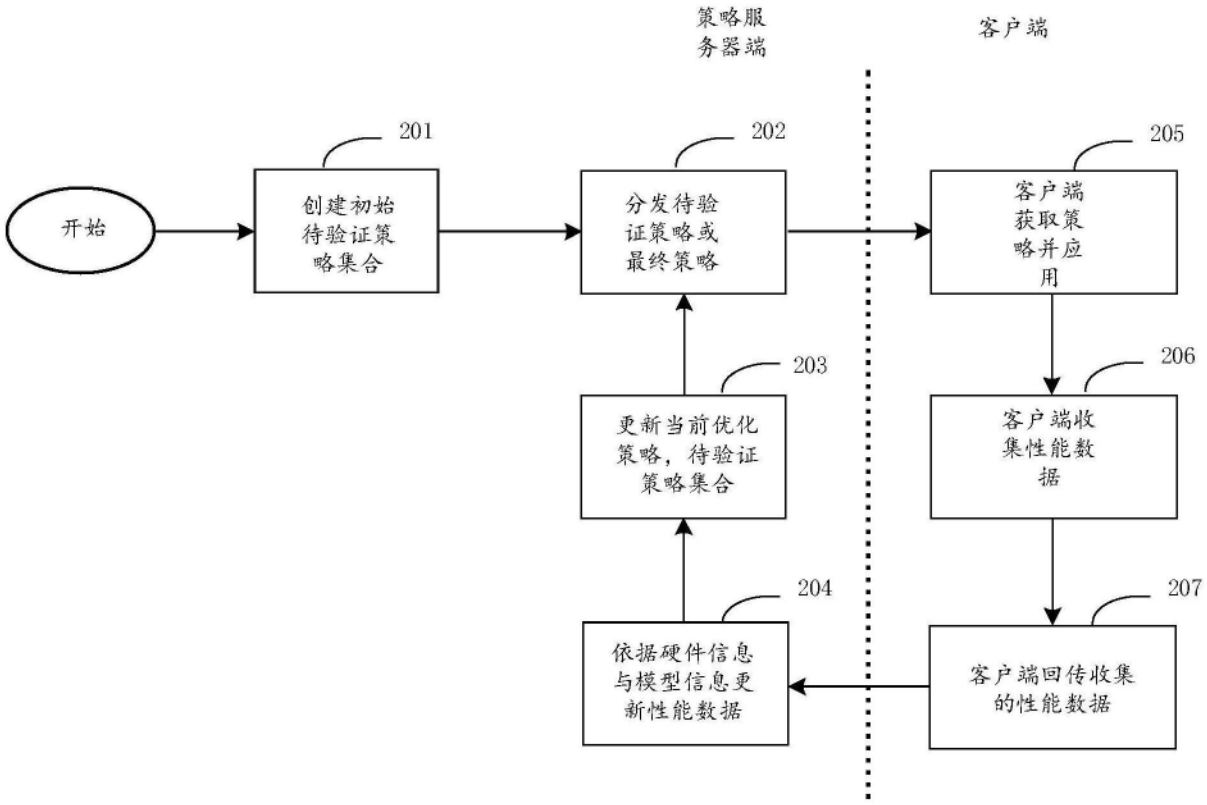

策略服务器端
客户端
开始
201
创建初始待验证策略集合
202
分发待验证策略或最终策略
205
客户端获取策略并应用
203
更新当前优化策略，待验证策略集合
206
客户端收集性能数据
204
依据硬件信息与模型信息更新性能数据
207
客户端回传收集的性能数据

图2B

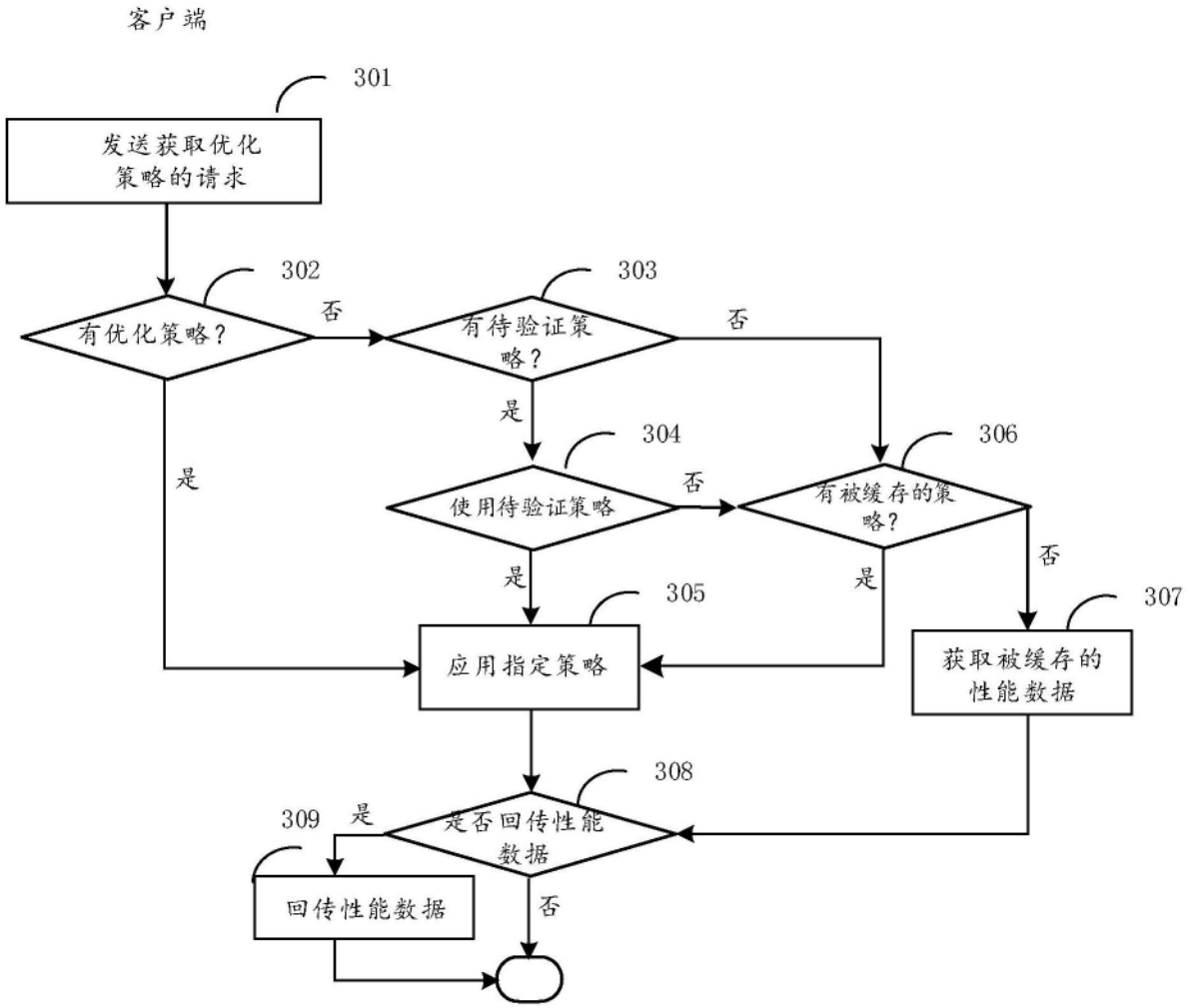

客户端
301
发送获取优化策略的请求
302
有优化策略？
否
303
有待验证策略？
否
是
是
304
使用待验证策略
否
306
有被缓存的策略？
是
否
305
应用指定策略
307
获取被缓存的性能数据
308
是否回传性能数据
309
是
回传性能数据
否

图3

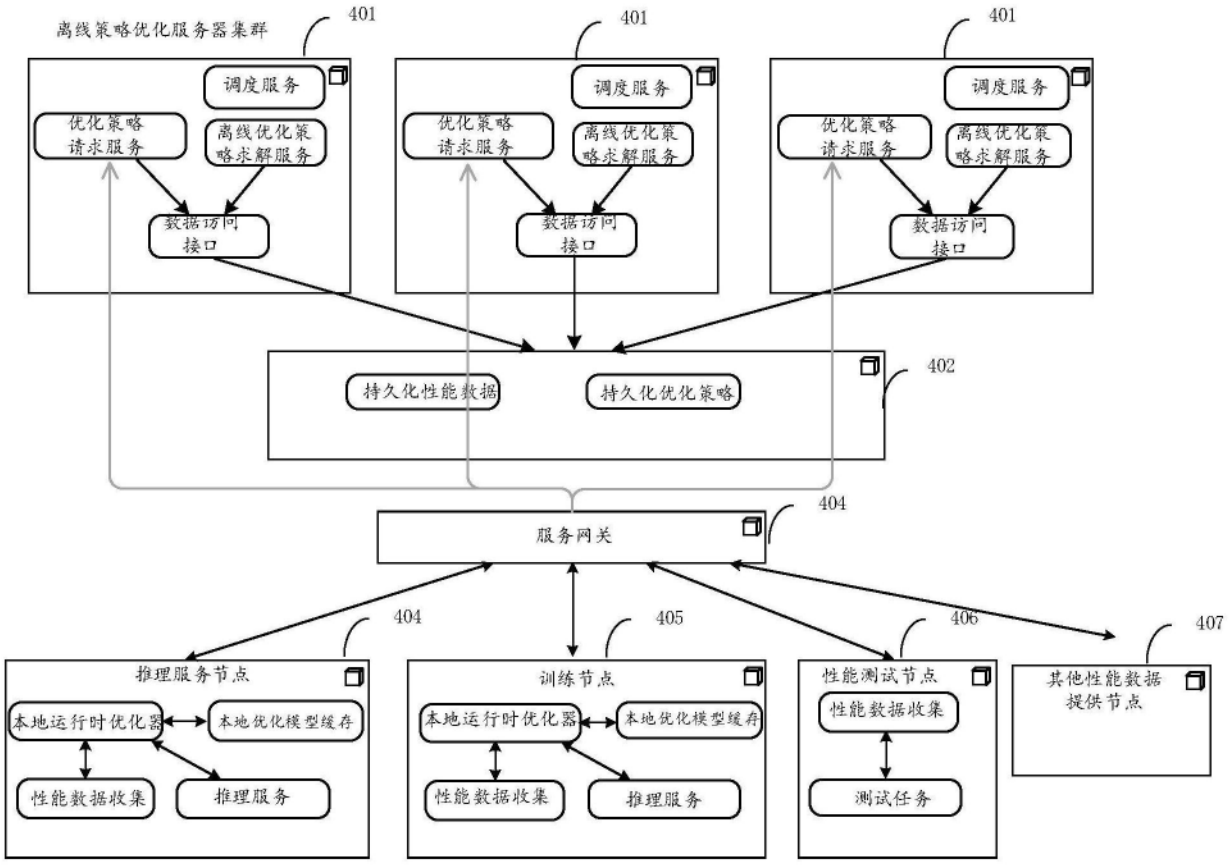

离线策略优化服务器集群
401
调度服务
优化策略请求服务
离线优化策略求解服务
数据访问接口
402
持久化性能数据
持久化优化策略
404
服务网关
404
推理服务节点
本地运行时优化器
本地优化模型缓存
性能数据收集
推理服务
405
训练节点
406
性能测试节点
测试任务
407
其他性能数据提供节点

图4

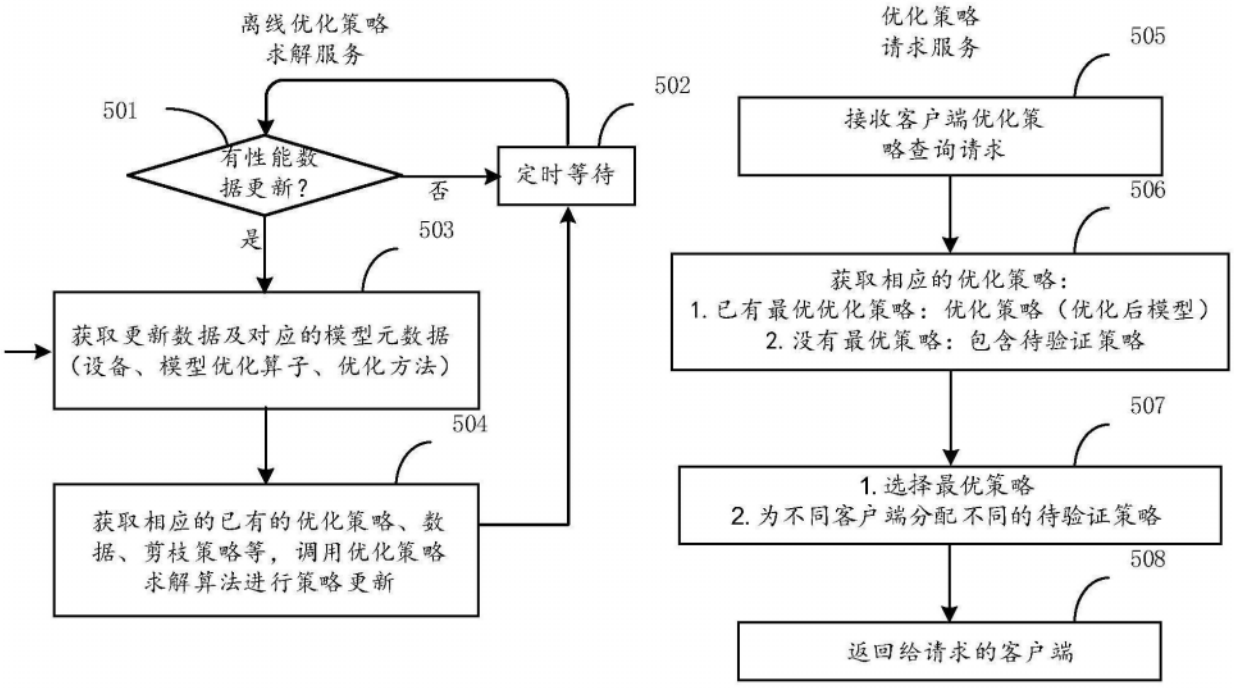

离线优化策略求解服务
501
有性能数据更新?
否
是
502
定时等待
503
获取更新数据及对应的模型元数据（设备、模型优化算子、优化方法）
504
获取相应的已有的优化策略、数据、剪枝策略等，调用优化策略求解算法进行策略更新
优化策略请求服务
505
接收客户端优化策略查询请求
506
获取相应的优化策略：
1. 已有最优优化策略：优化策略（优化后模型）
2. 没有最优策略：包含待验证策略
507
1. 选择最优策略
2. 为不同客户端分配不同的待验证策略
508
返回给请求的客户端

图5

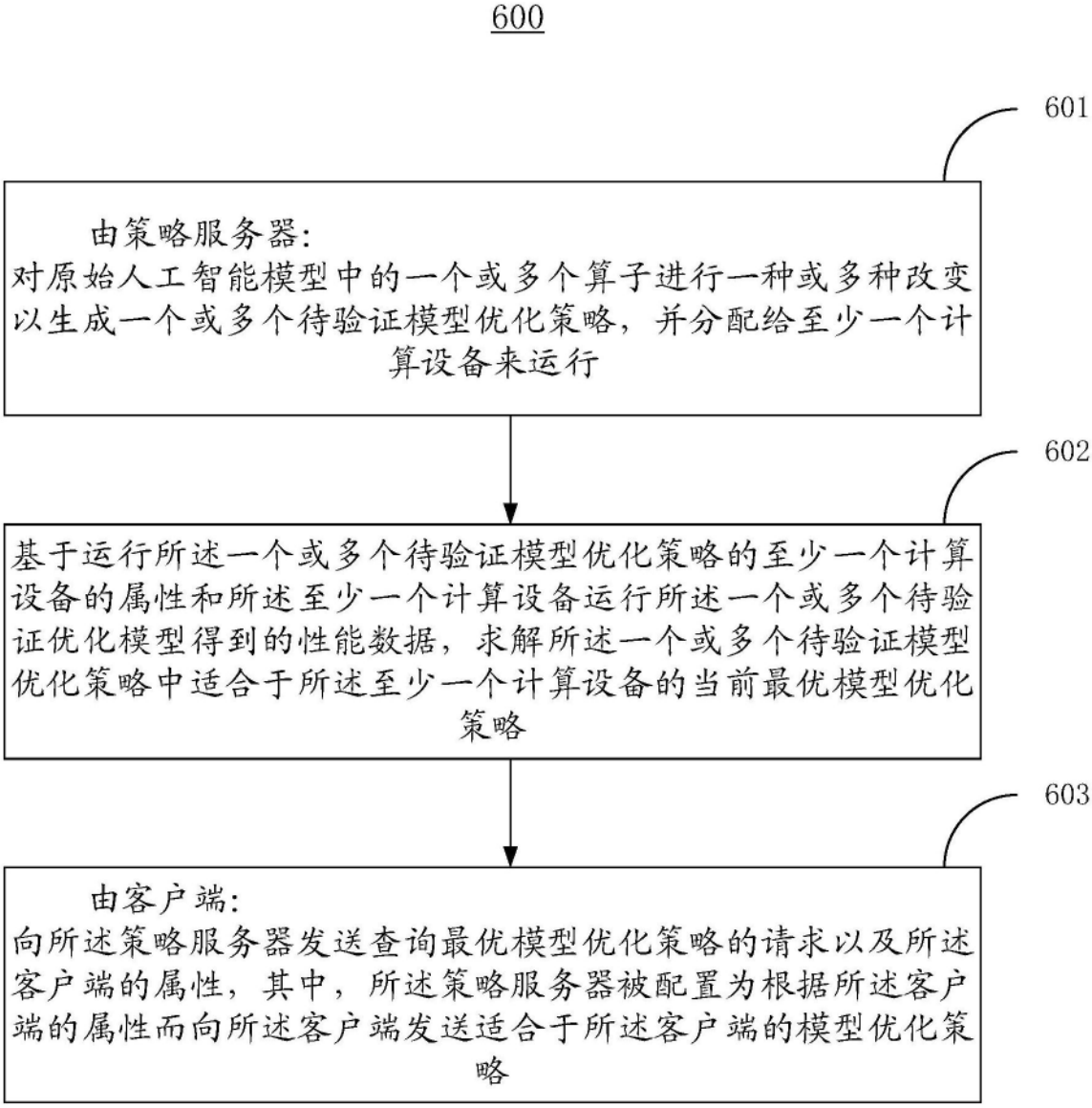

600
601
由策略服务器:
对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行
602
基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略
603
由客户端:
向所述策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性，其中，所述策略服务器被配置为根据所述客户端的属性而向所述客户端发送适合于所述客户端的模型优化策略

图6

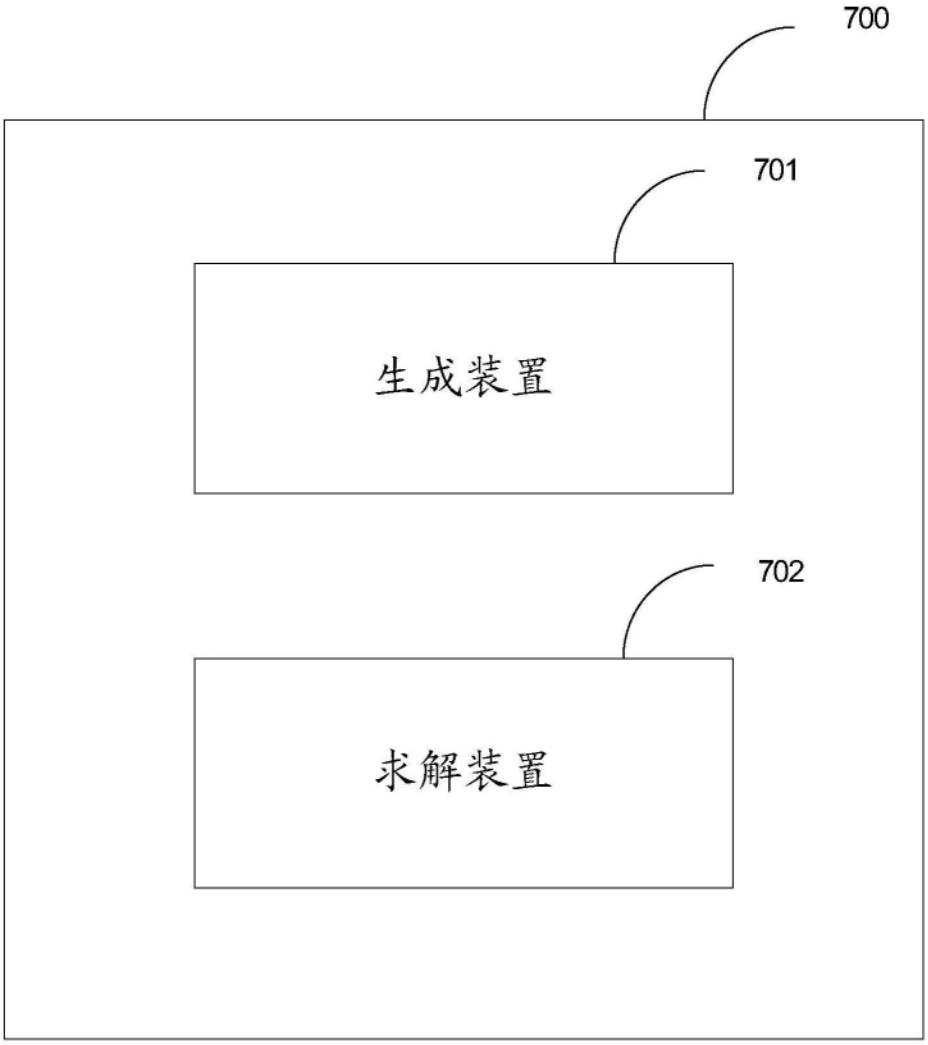


图7

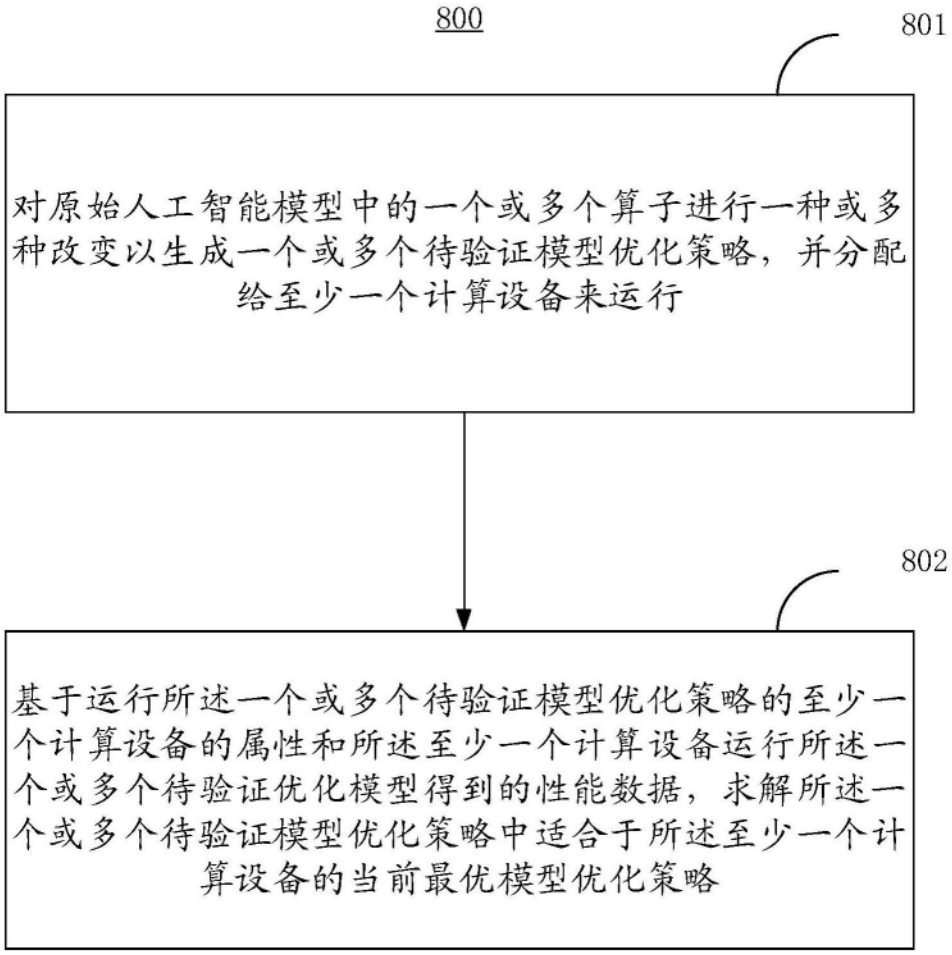

800
801
对原始人工智能模型中的一个或多个算子进行一种或多种改变以生成一个或多个待验证模型优化策略，并分配给至少一个计算设备来运行
802
基于运行所述一个或多个待验证模型优化策略的至少一个计算设备的属性和所述至少一个计算设备运行所述一个或多个待验证优化模型得到的性能数据，求解所述一个或多个待验证模型优化策略中适合于所述至少一个计算设备的当前最优模型优化策略

图8

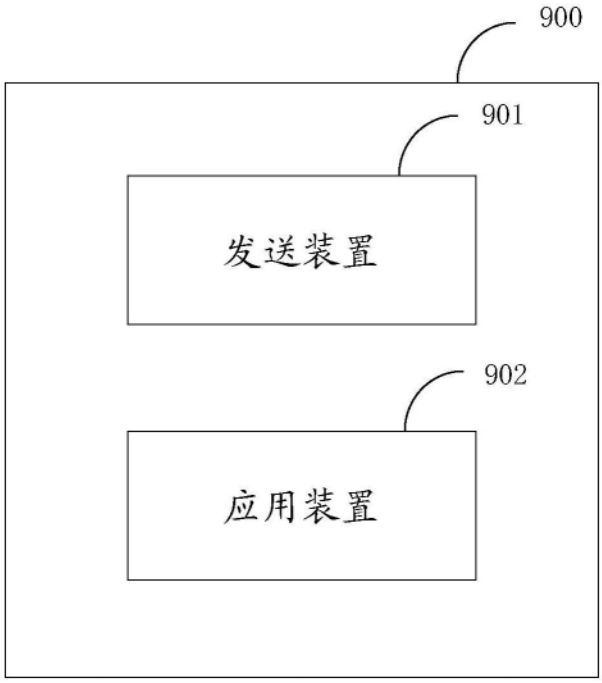


图9

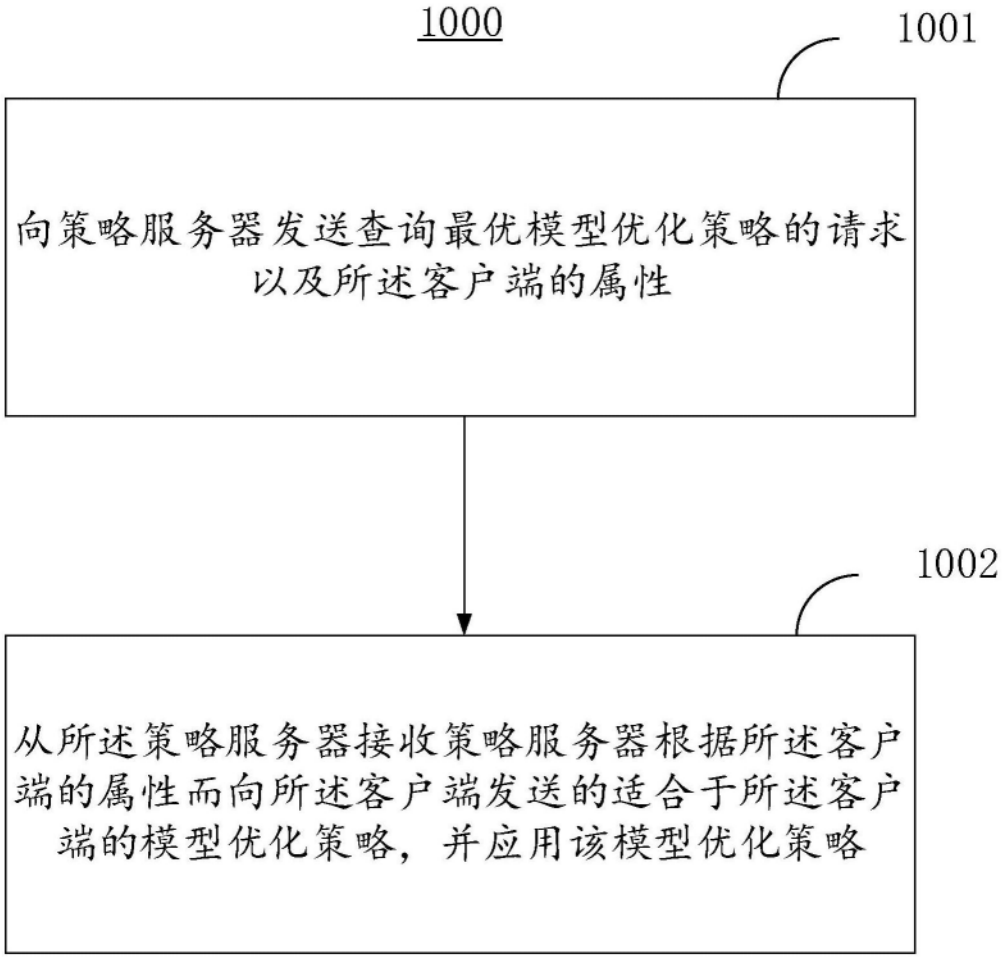

1000
1001
向策略服务器发送查询最优模型优化策略的请求以及所述客户端的属性
1002
从所述策略服务器接收策略服务器根据所述客户端的属性而向所述客户端发送的适合于所述客户端的模型优化策略，并应用该模型优化策略

图10

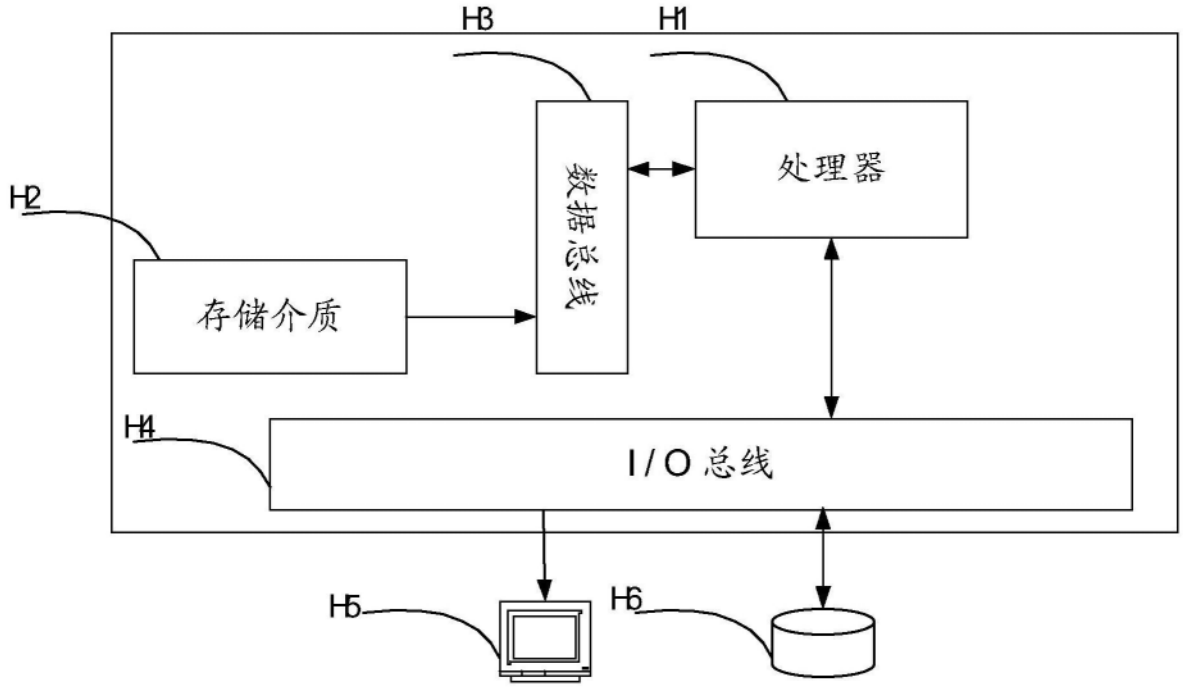


图11

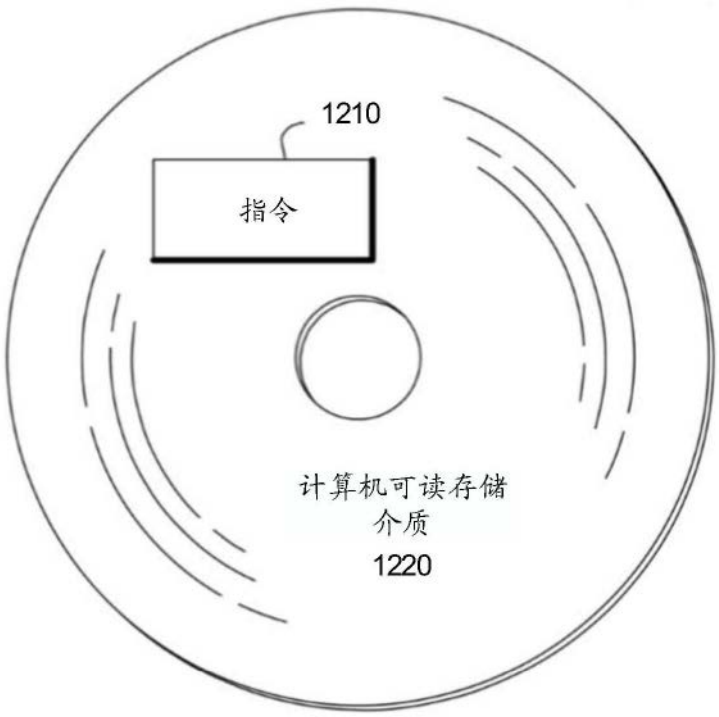


图12