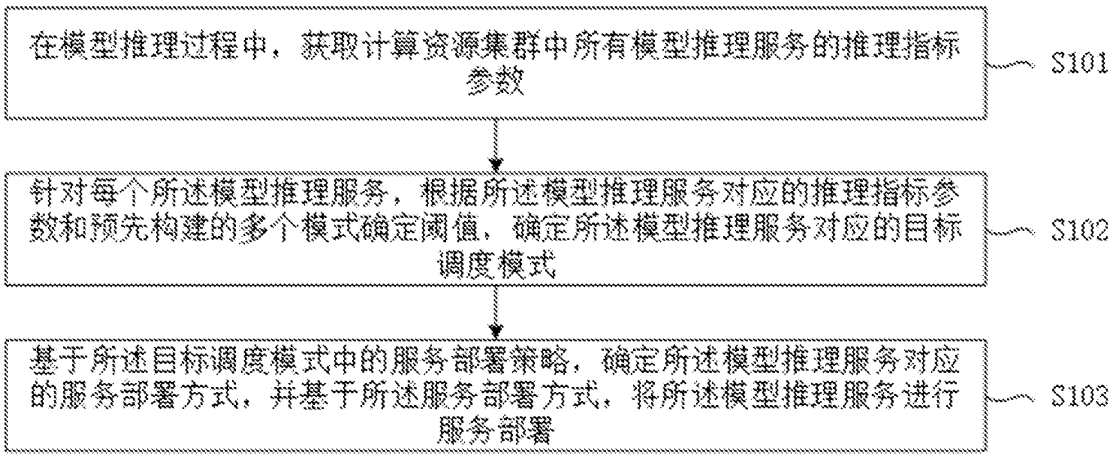
在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数
S101
针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式
S102
基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部署方式，将所述模型推理服务进行服务部署
S103

图 1

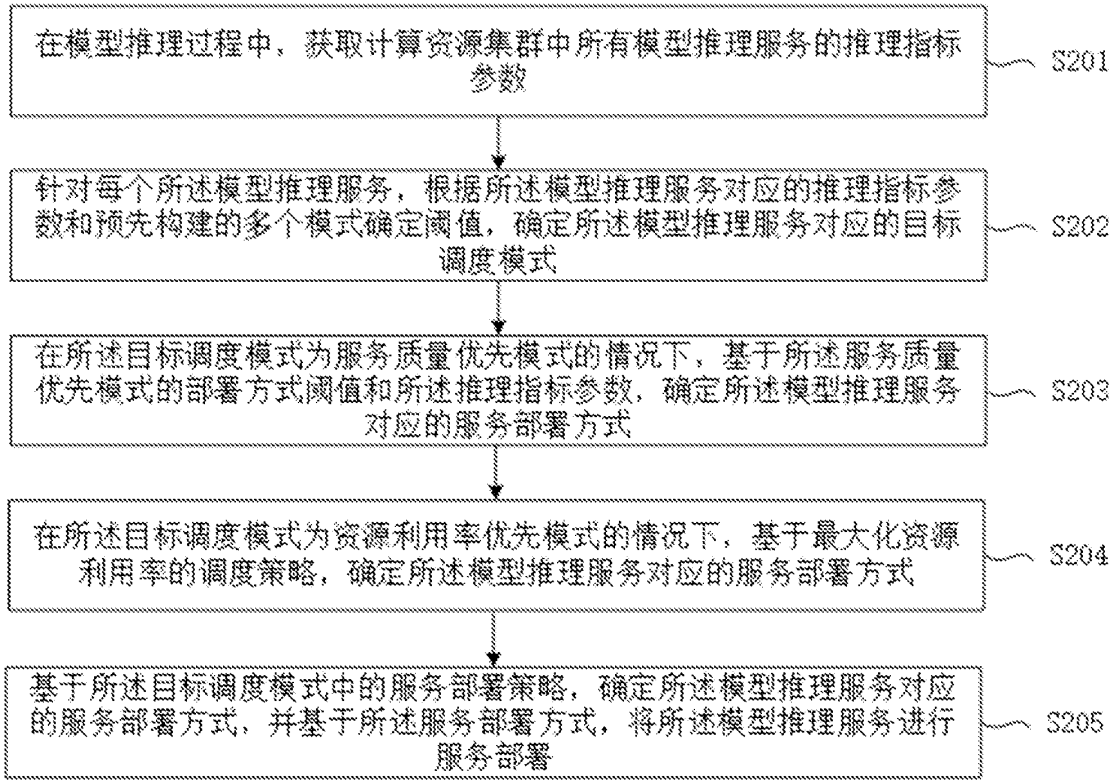
在模型推理过程中，获取计算资源集群中所有模型推理服务的推理指标参数
S201
针对每个所述模型推理服务，根据所述模型推理服务对应的推理指标参数和预先构建的多个模式确定阈值，确定所述模型推理服务对应的目标调度模式
S202
在所述目标调度模式为服务质量优先模式的情况下，基于所述服务质量优先模式的部署方式阈值和所述推理指标参数，确定所述模型推理服务对应的服务部署方式
S203
在所述目标调度模式为资源利用率优先模式的情况下，基于最大化资源利用率的调度策略，确定所述模型推理服务对应的服务部署方式
S204
基于所述目标调度模式中的服务部署策略，确定所述模型推理服务对应的服务部署方式，并基于所述服务部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

图 2

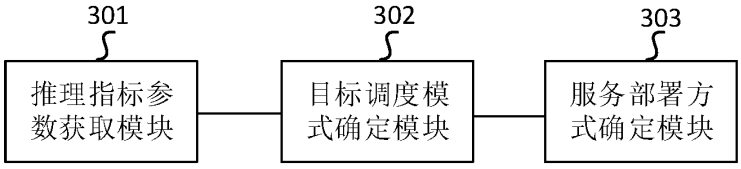
301
302
303
推理指标参数获取模块
目标调度模式确定模块
服务部署方式确定模块

图 3

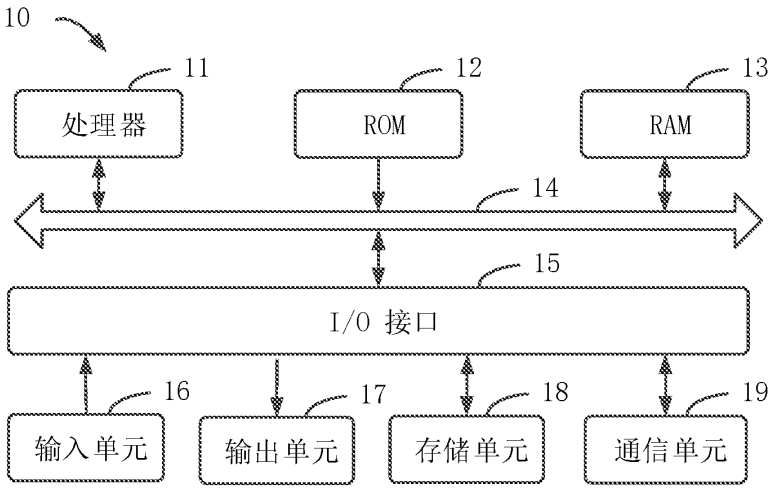
10
11
12
13
处理器
ROM
RAM
14
15
I/O 接口
16
17
18
19
输入单元
输出单元
存储单元
通信单元

图 4