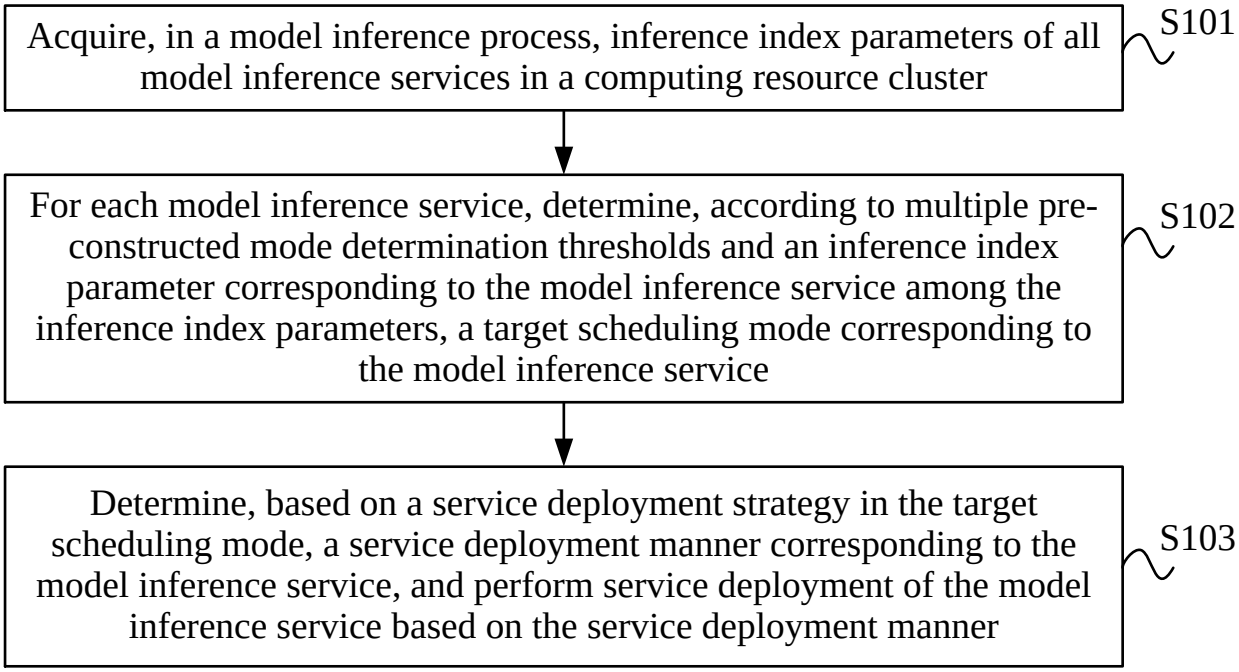
Acquire, in a model inference process, inference index parameters of all model inference services in a computing resource cluster
S101
For each model inference service, determine, according to multiple pre-constructed mode determination thresholds and an inference index parameter corresponding to the model inference service among the inference index parameters, a target scheduling mode corresponding to the model inference service
S102
Determine, based on a service deployment strategy in the target scheduling mode, a service deployment manner corresponding to the model inference service, and perform service deployment of the model inference service based on the service deployment manner
S103

FIG. 1

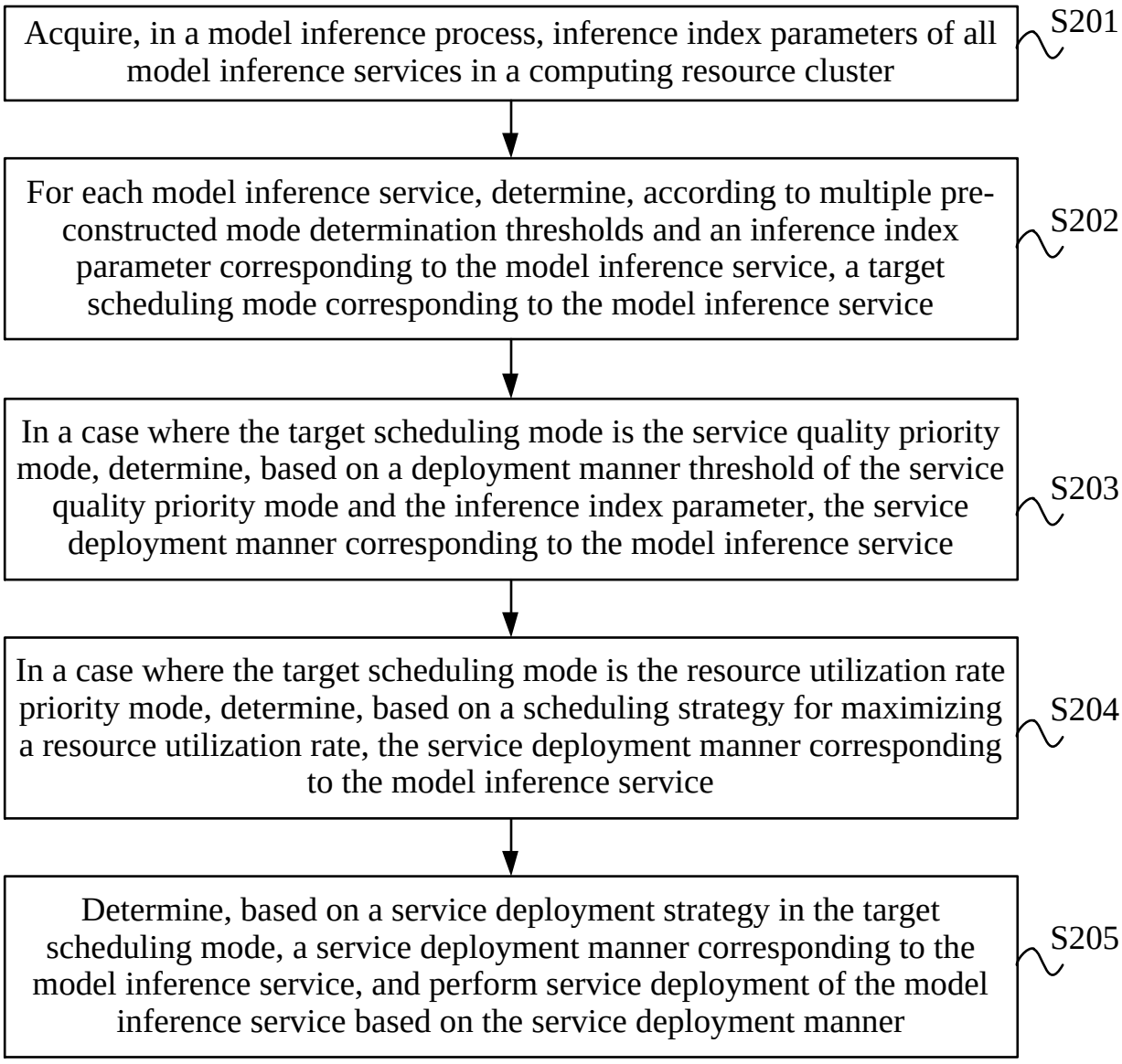
Acquire, in a model inference process, inference index parameters of all model inference services in a computing resource cluster
S201
For each model inference service, determine, according to multiple pre-constructed mode determination thresholds and an inference index parameter corresponding to the model inference service, a target scheduling mode corresponding to the model inference service
S202
In a case where the target scheduling mode is the service quality priority mode, determine, based on a deployment manner threshold of the service quality priority mode and the inference index parameter, the service deployment manner corresponding to the model inference service
S203
In a case where the target scheduling mode is the resource utilization rate priority mode, determine, based on a scheduling strategy for maximizing a resource utilization rate, the service deployment manner corresponding to the model inference service
S204
Determine, based on a service deployment strategy in the target scheduling mode, a service deployment manner corresponding to the model inference service, and perform service deployment of the model inference service based on the service deployment manner
S205

FIG. 2

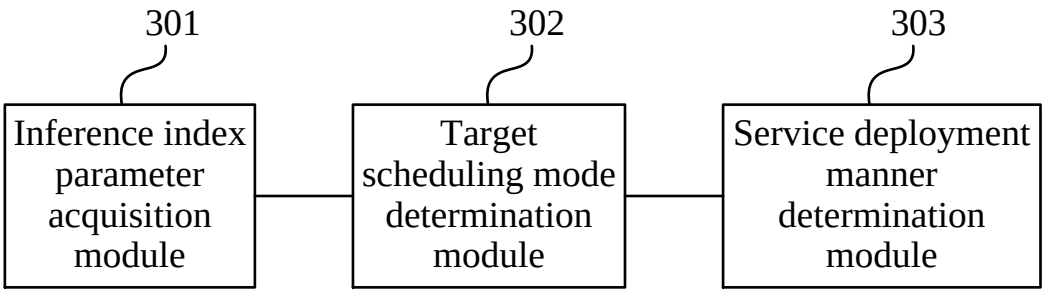
301
302
303
Inference index parameter acquisition module
Target scheduling mode determination module
Service deployment manner determination module

FIG. 3

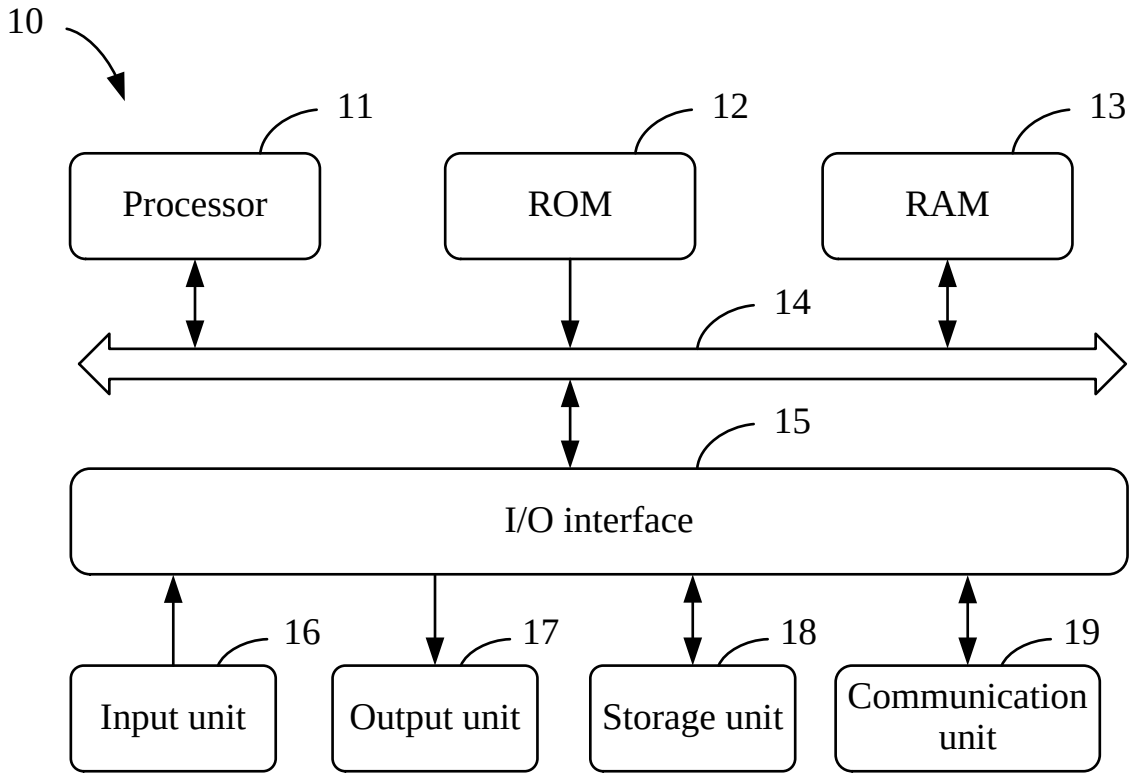
10
11
12
13
Processor
ROM
RAM
14
15
I/O interface
16
17
18
19
Input unit
Output unit
Storage unit
Communication unit

FIG. 4