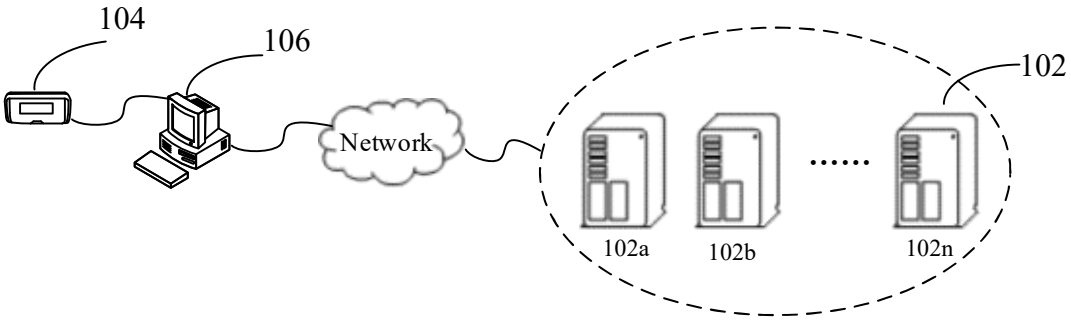
104
106
102
Network
102a
102b
102n

FIG. 1

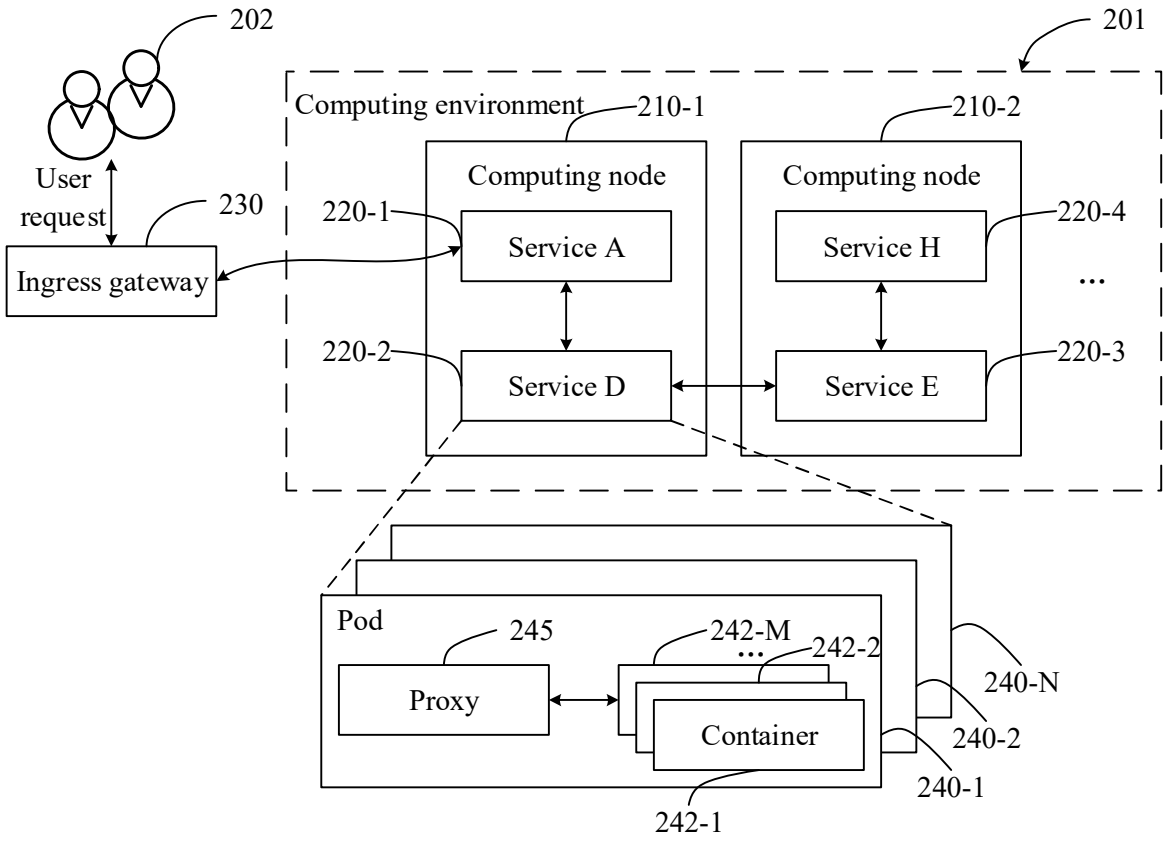
202
201
User request
230
Ingress gateway
Computing environment
210-1
210-2
Computing node
Computing node
220-1
Service A
Service H
220-4
220-2
Service D
Service E
220-3
Pod
245
242-M
242-2
Proxy
Container
240-N
240-2
240-1
242-1

FIG. 2

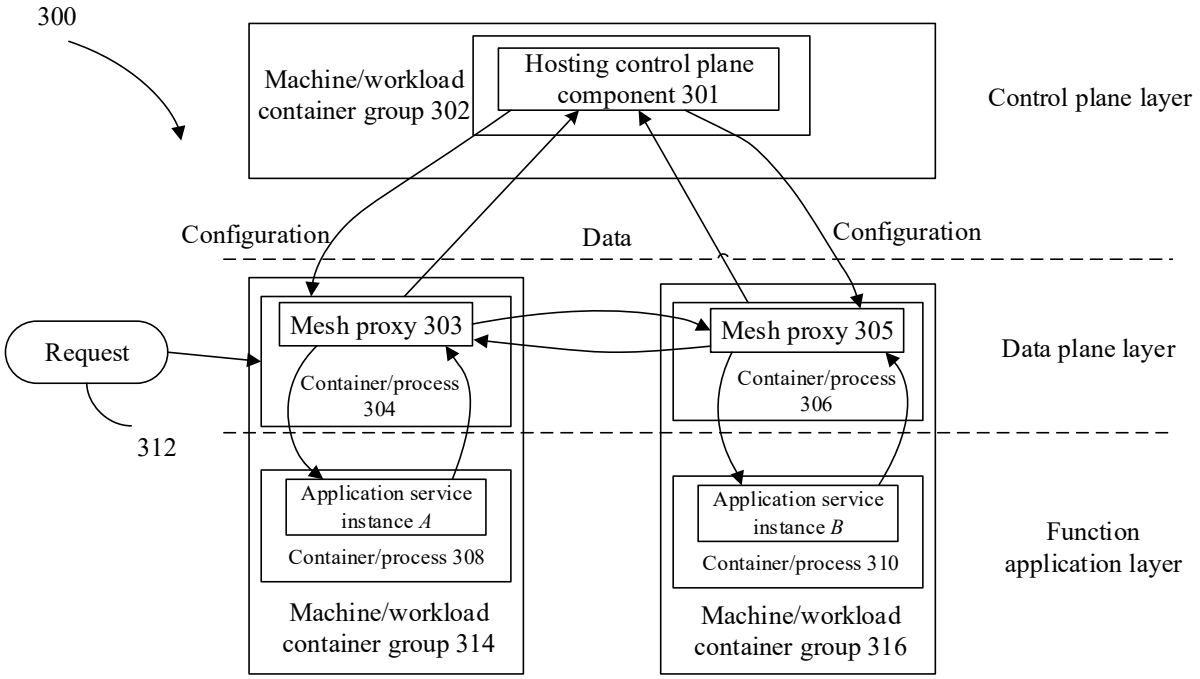
300
Machine/workload container group 302
Hosting control plane component 301
Control plane layer
Configuration
Data
Configuration
Mesh proxy 303
Container/process 304
Mesh proxy 305
Container/process 306
Data plane layer
Request
312
Application service instance A
Container/process 308
Application service instance B
Container/process 310
Function application layer
Machine/workload container group 314
Machine/workload container group 316

FIG. 3

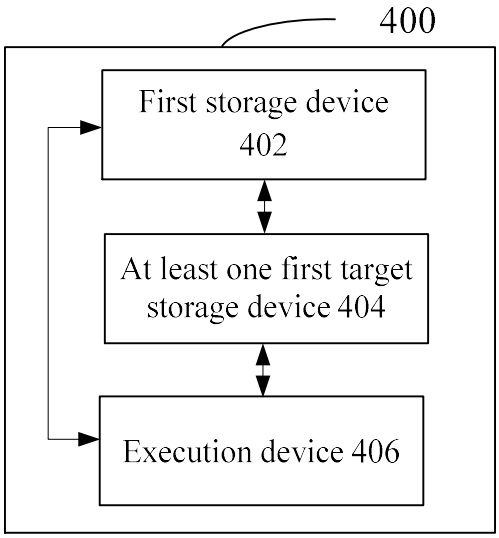
400
First storage device 402
At least one first target storage device 404
Execution device 406

FIG. 4

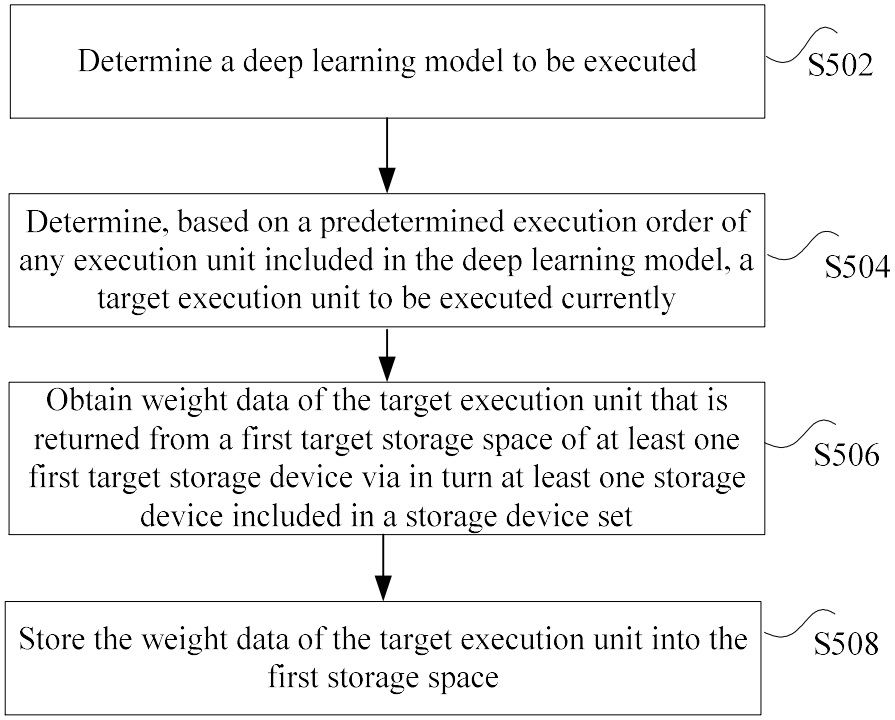
Determine a deep learning model to be executed
S502
Determine, based on a predetermined execution order of any execution unit included in the deep learning model, a target execution unit to be executed currently
S504
Obtain weight data of the target execution unit that is returned from a first target storage space of at least one first target storage device via in turn at least one storage device included in a storage device set
S506
Store the weight data of the target execution unit into the first storage space
S508

FIG. 5

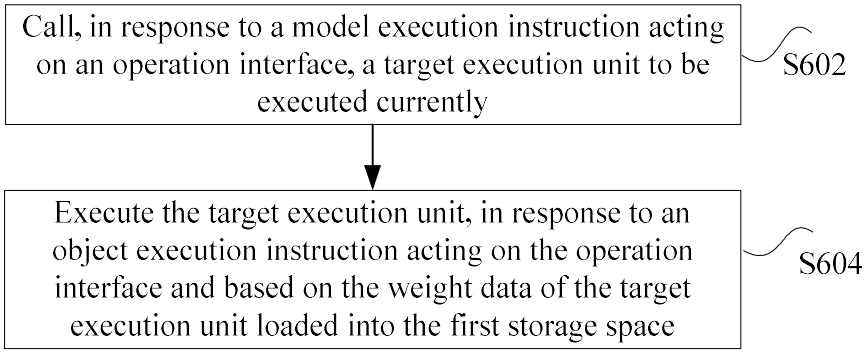
Call, in response to a model execution instruction acting on an operation interface, a target execution unit to be executed currently
S602
Execute the target execution unit, in response to an object execution instruction acting on the operation interface and based on the weight data of the target execution unit loaded into the first storage space
S604

FIG. 6

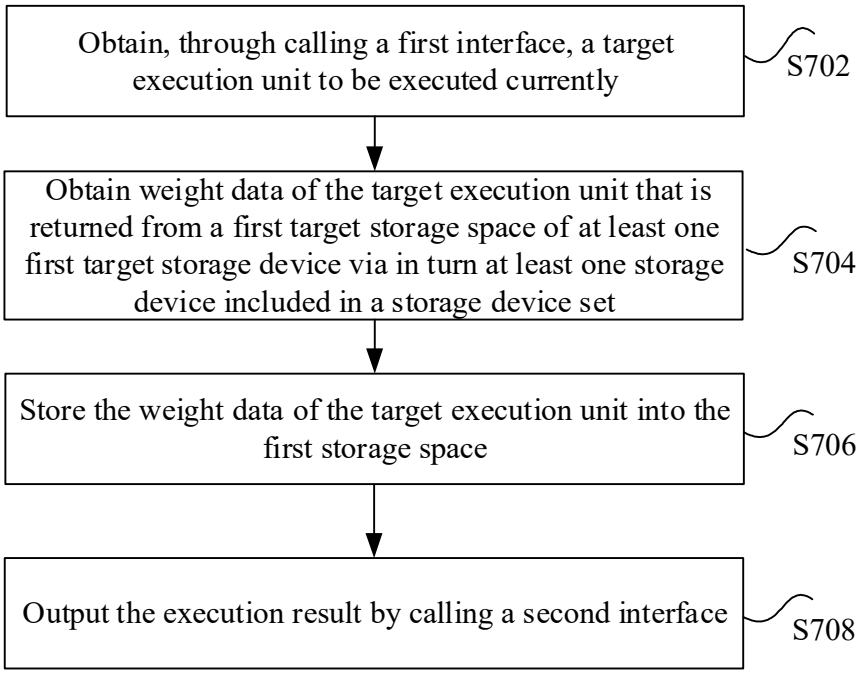
Obtain, through calling a first interface, a target execution unit to be executed currently
S702
Obtain weight data of the target execution unit that is returned from a first target storage space of at least one first target storage device via in turn at least one storage device included in a storage device set
S704
Store the weight data of the target execution unit into the first storage space
S706
Output the execution result by calling a second interface
S708

FIG. 7

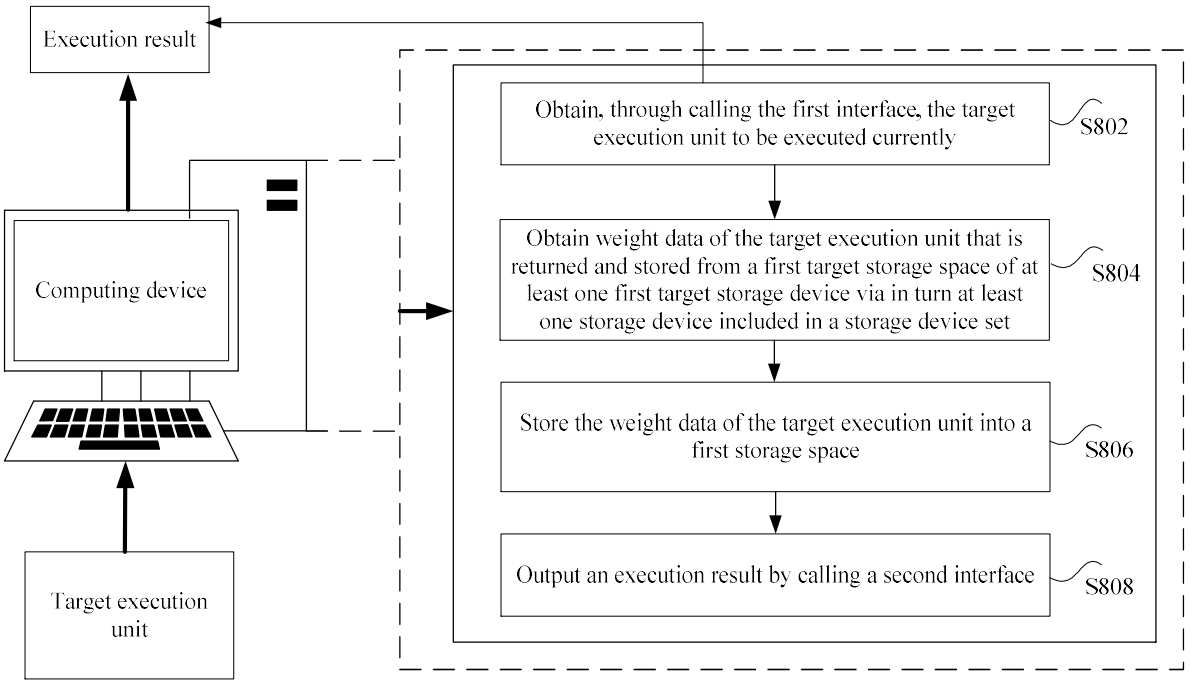
Execution result
Computing device
Target execution unit
Obtain, through calling the first interface, the target execution unit to be executed currently
S802
Obtain weight data of the target execution unit that is returned and stored from a first target storage space of at least one first target storage device via in turn at least one storage device included in a storage device set
S804
Store the weight data of the target execution unit into a first storage space
S806
Output an execution result by calling a second interface
S808

FIG. 8

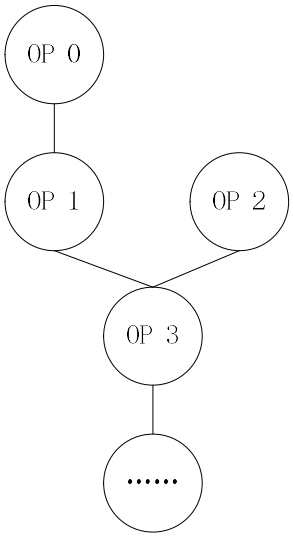
OP 0
OP 1
OP 2
OP 3
.....

FIG. 9

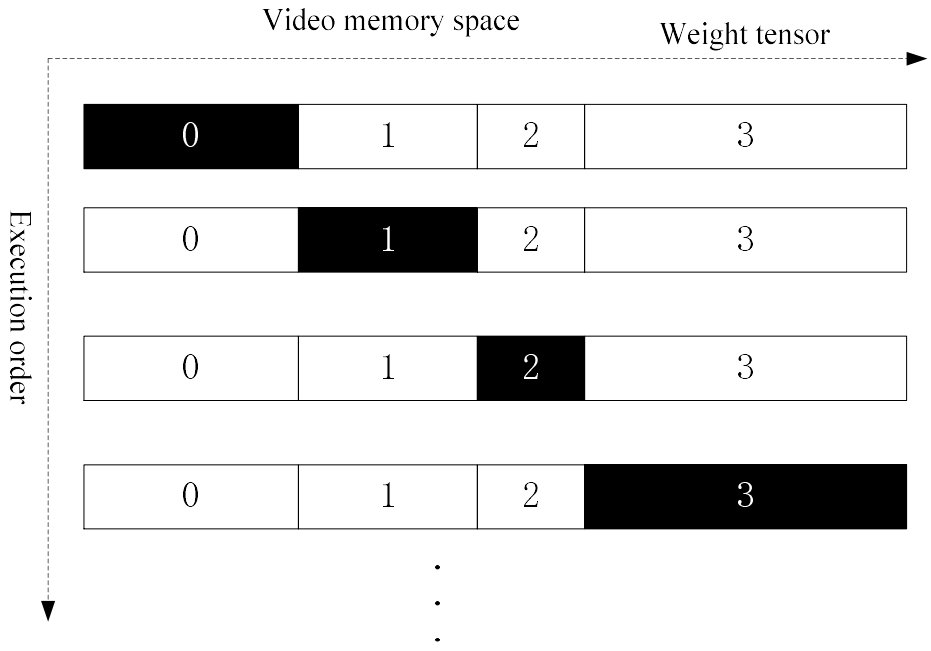
Video memory space
Weight tensor
Execution order
0
1
2
3
0
1
2
3
0
1
2
3
0
1
2
3

FIG. 10

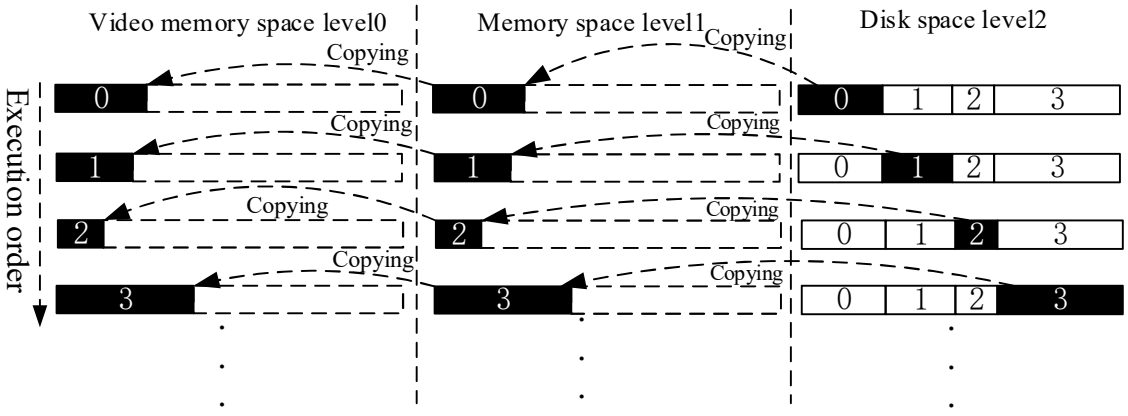

FIG. 11

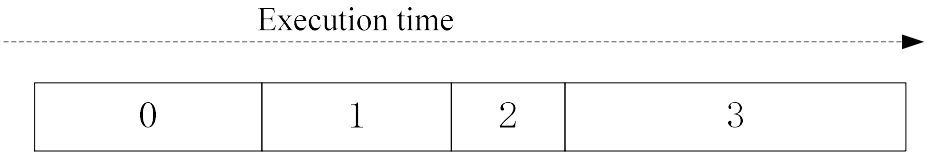

FIG. 12(a)

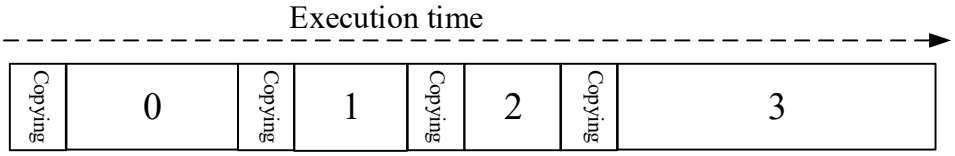

FIG. 12(b)

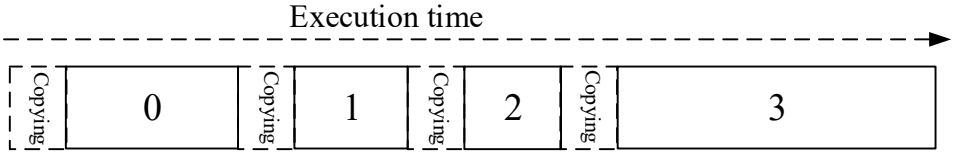

FIG. 12(c)

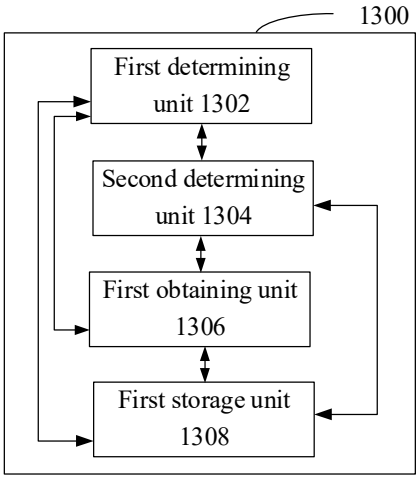

FIG. 13

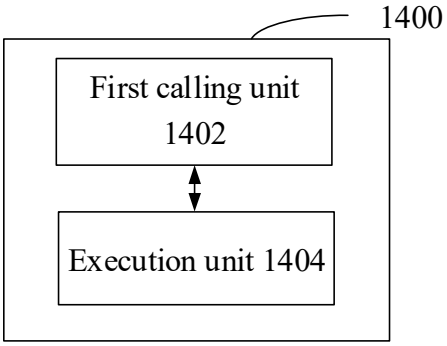

FIG. 14

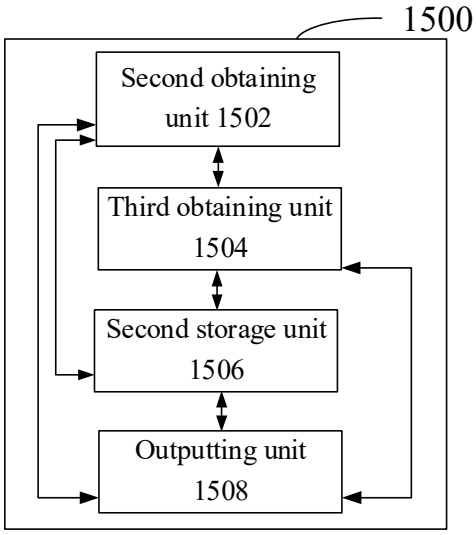

FIG. 15

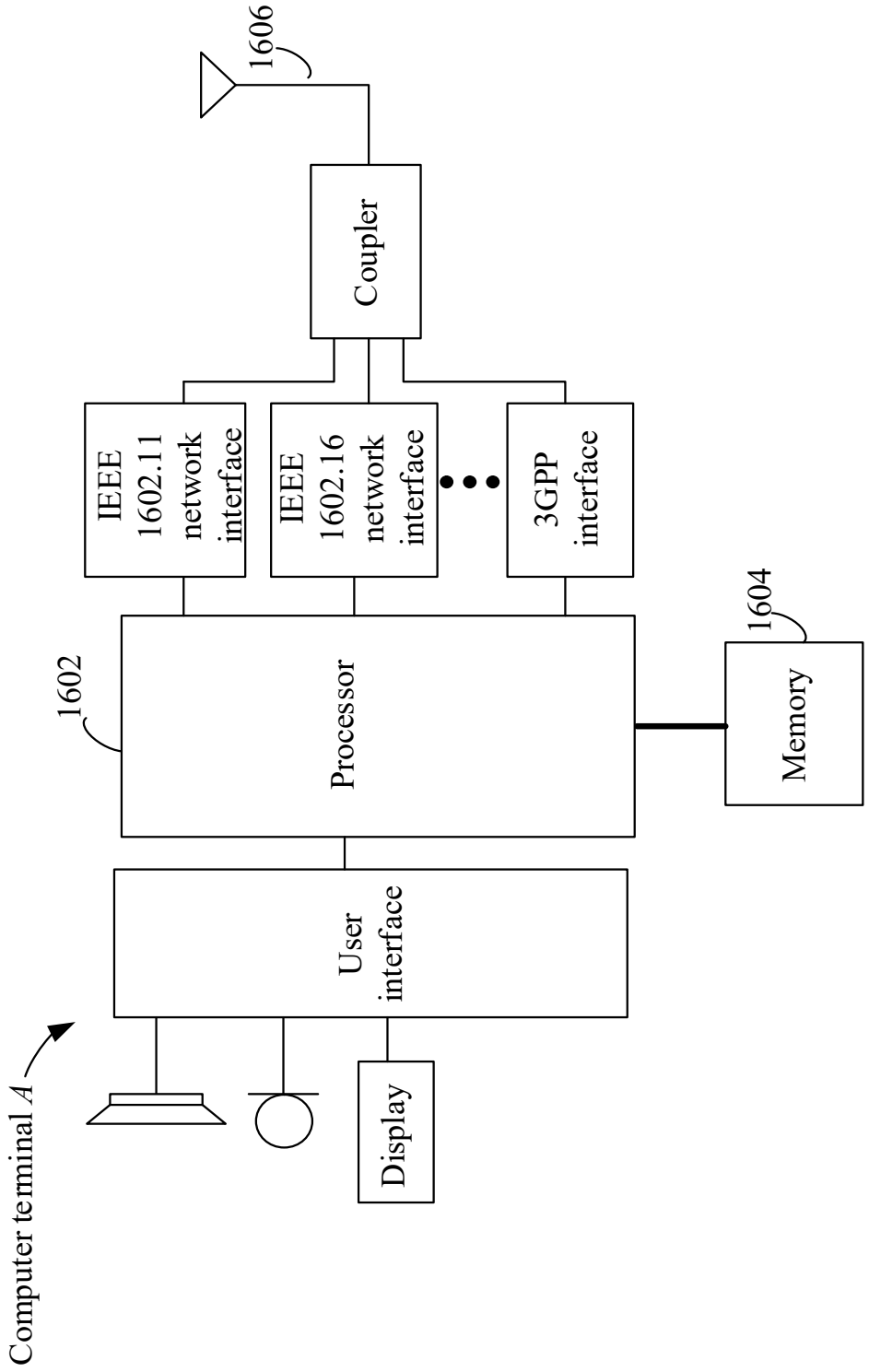
1606
Coupler
IEEE 1602.11 network interface
IEEE 1602.16 network interface
3GPP interface
1602
Processor
1604
Memory
User interface
Display
Computer terminal A

FIG. 16