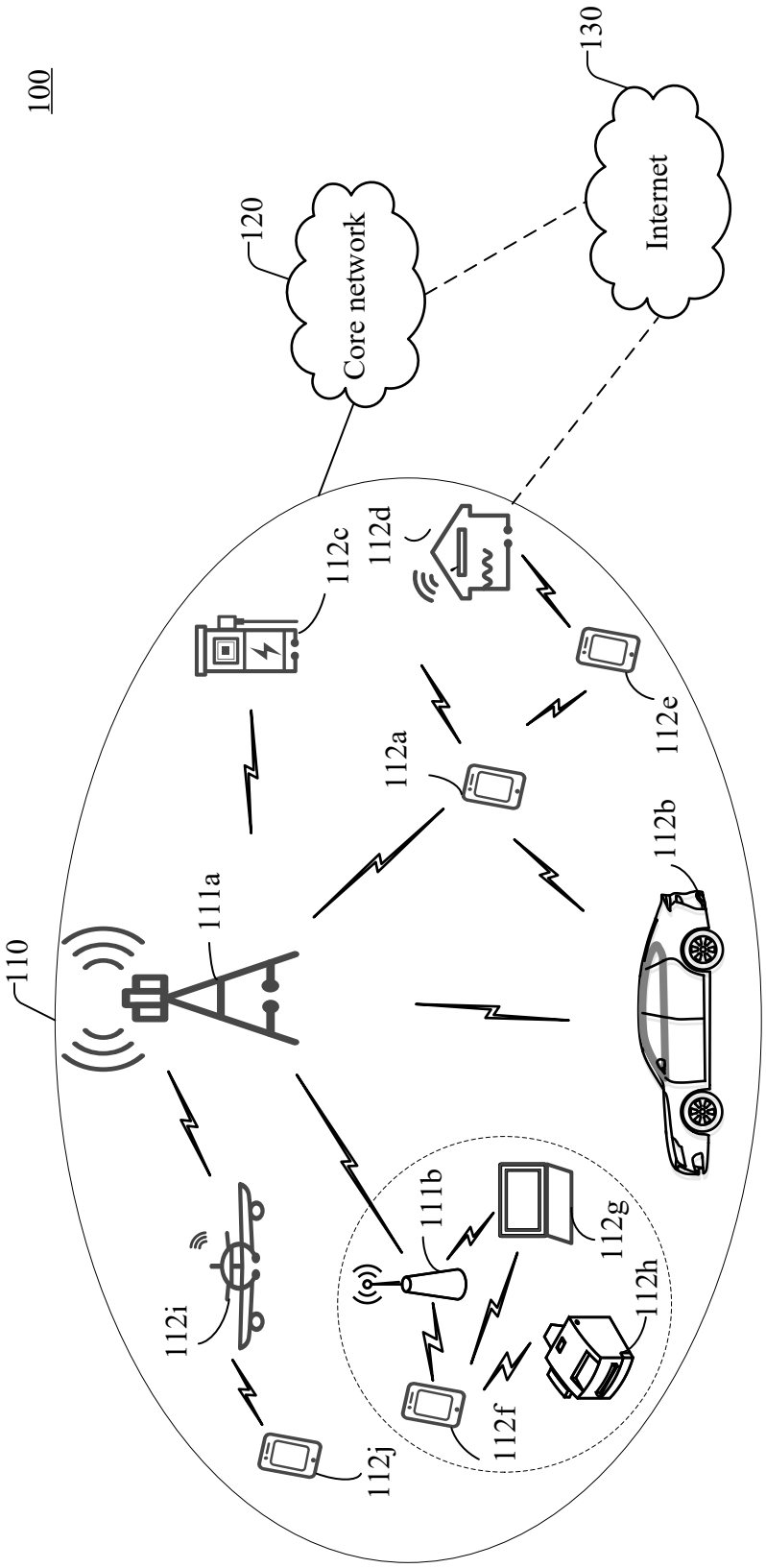

100
130
Internet
120
Core network
110
111a
112c
112d
112e
112a
112b
111b
112g
112h
112i
112f
112j

FIG. 1

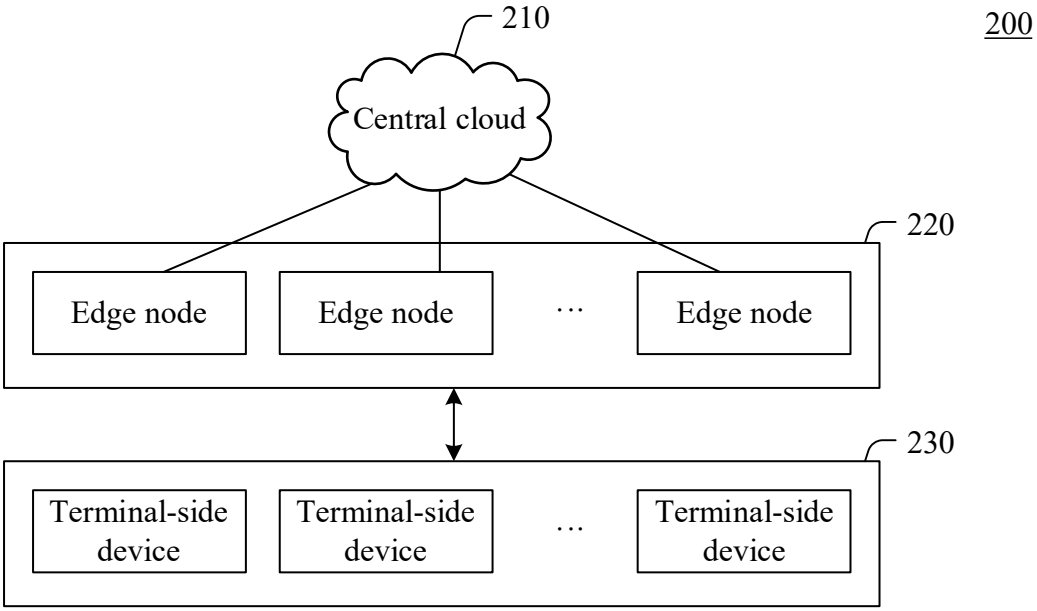

200
210
Central cloud
220
Edge node
Edge node
…
Edge node
230
Terminal-side device
Terminal-side device
…
Terminal-side device

FIG. 2

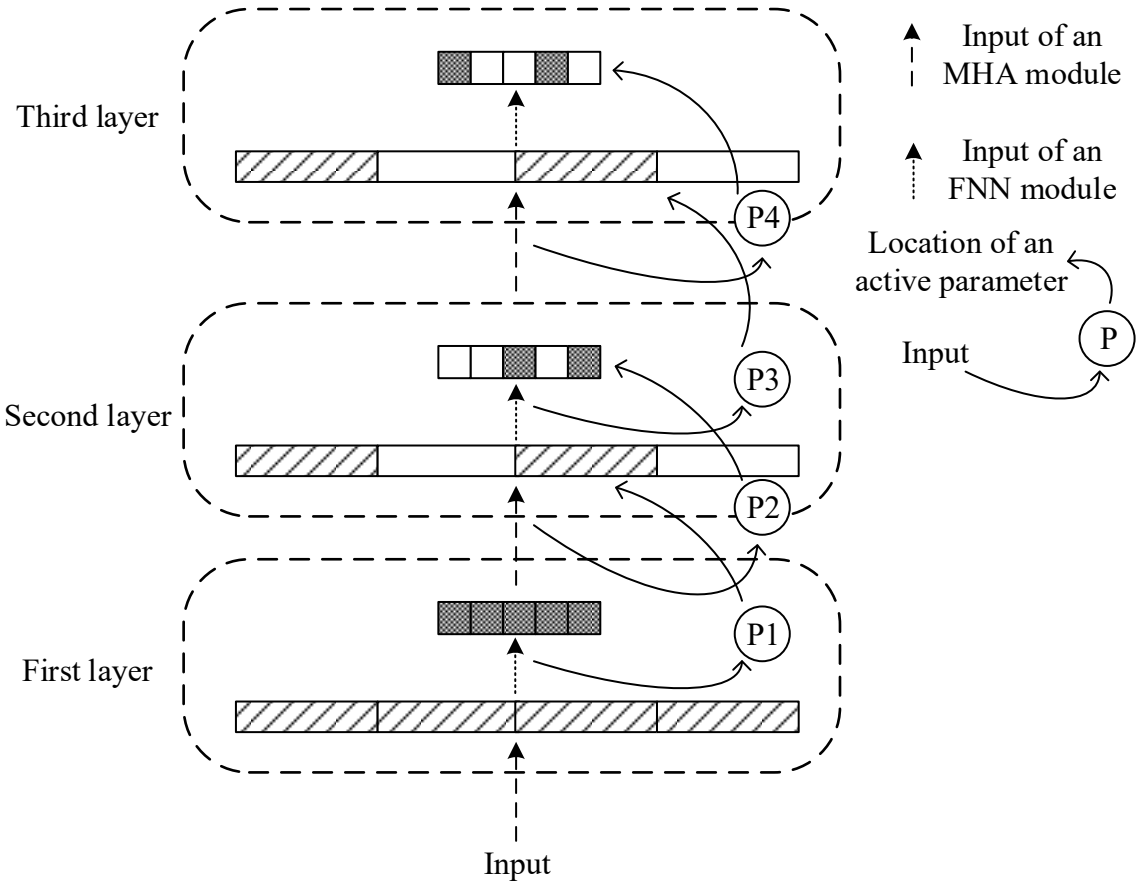

Third layer
Second layer
First layer
Input
P4
P3
P2
P1
Input of an MHA module
Input of an FNN module
Location of an active parameter
P
Input

FIG. 3

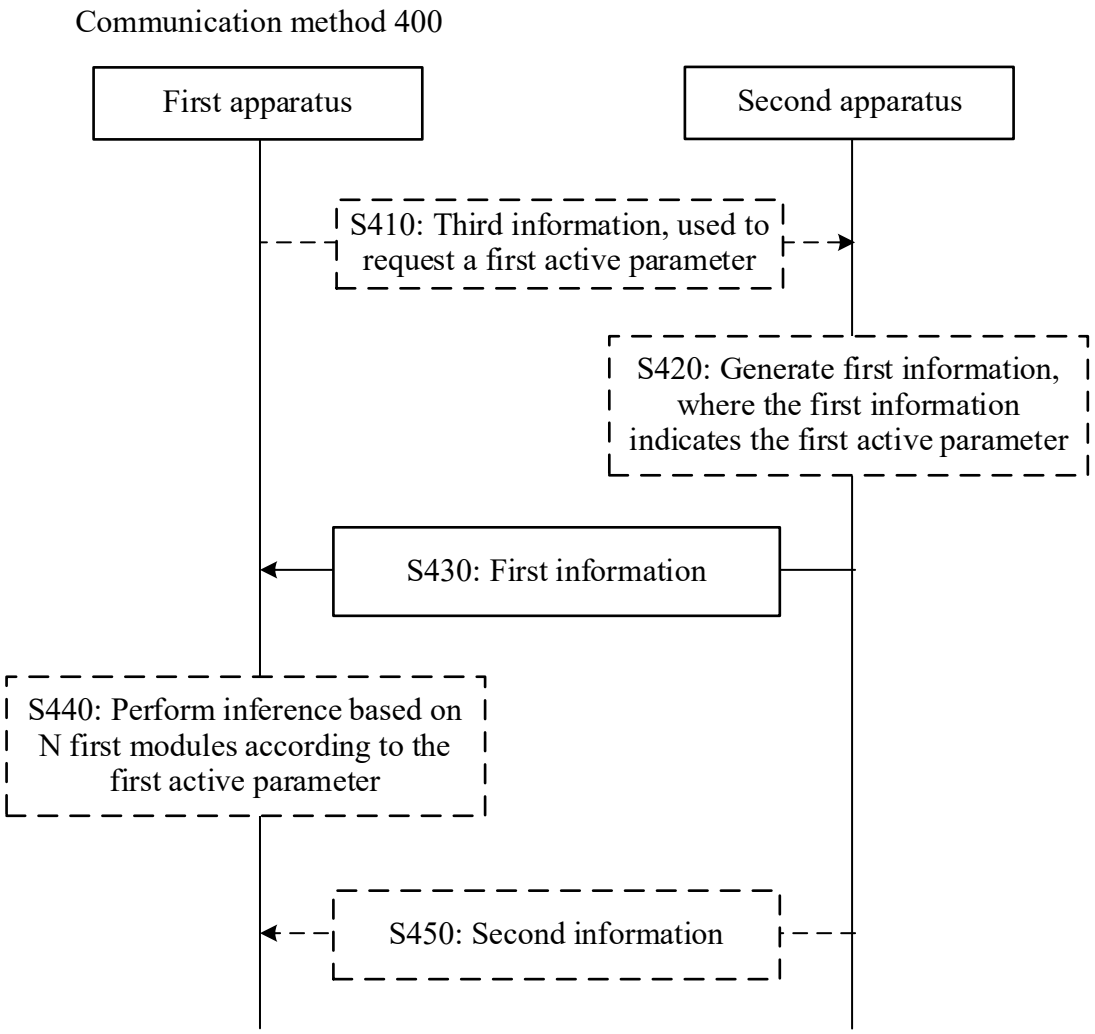

Communication method 400
First apparatus
Second apparatus
S410: Third information, used to request a first active parameter
S420: Generate first information, where the first information indicates the first active parameter
S430: First information
S440: Perform inference based on N first modules according to the first active parameter
S450: Second information

FIG. 4

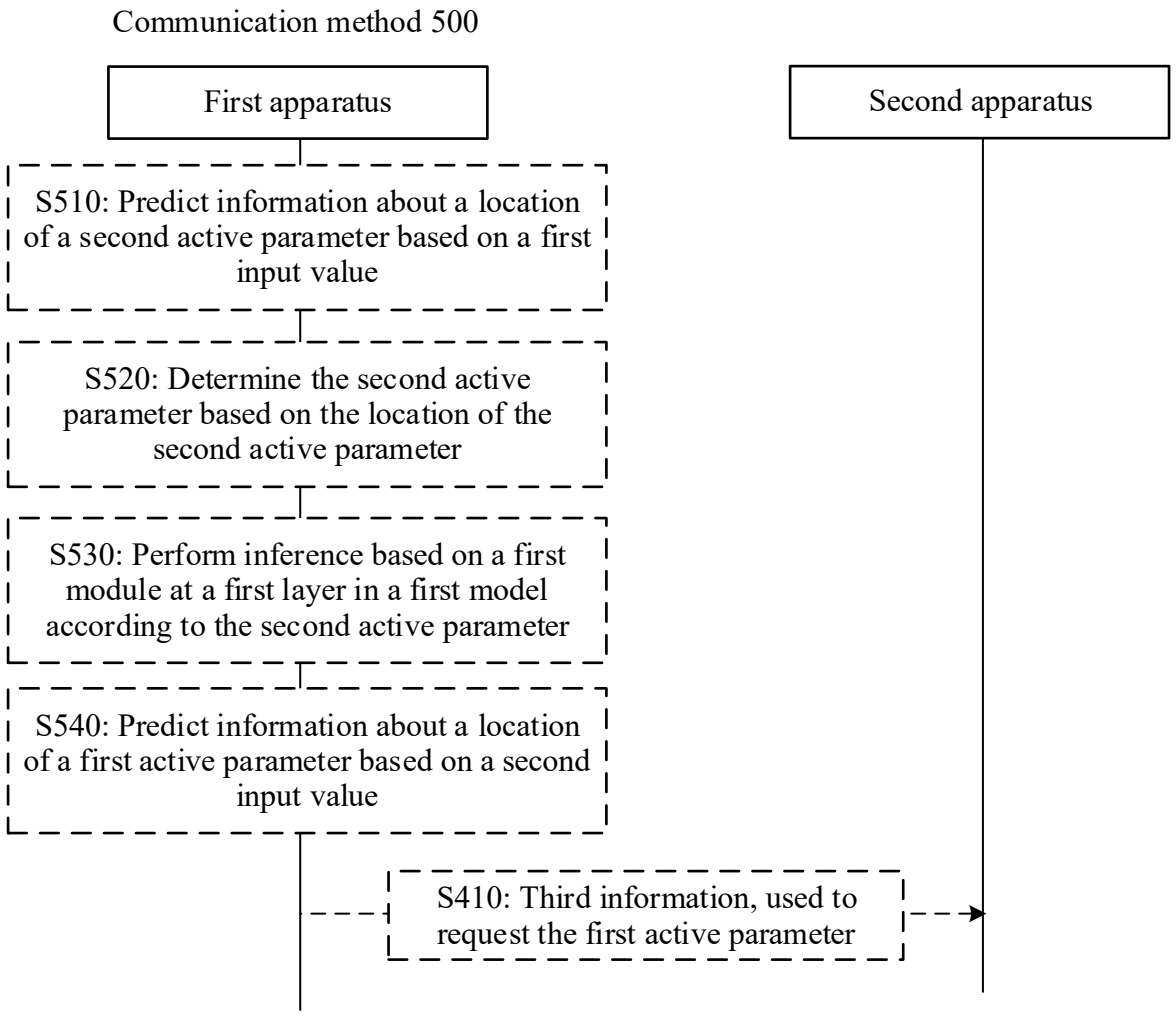

Communication method 500
First apparatus
Second apparatus
S510: Predict information about a location of a second active parameter based on a first input value
S520: Determine the second active parameter based on the location of the second active parameter
S530: Perform inference based on a first module at a first layer in a first model according to the second active parameter
S540: Predict information about a location of a first active parameter based on a second input value
S410: Third information, used to request the first active parameter

FIG. 5

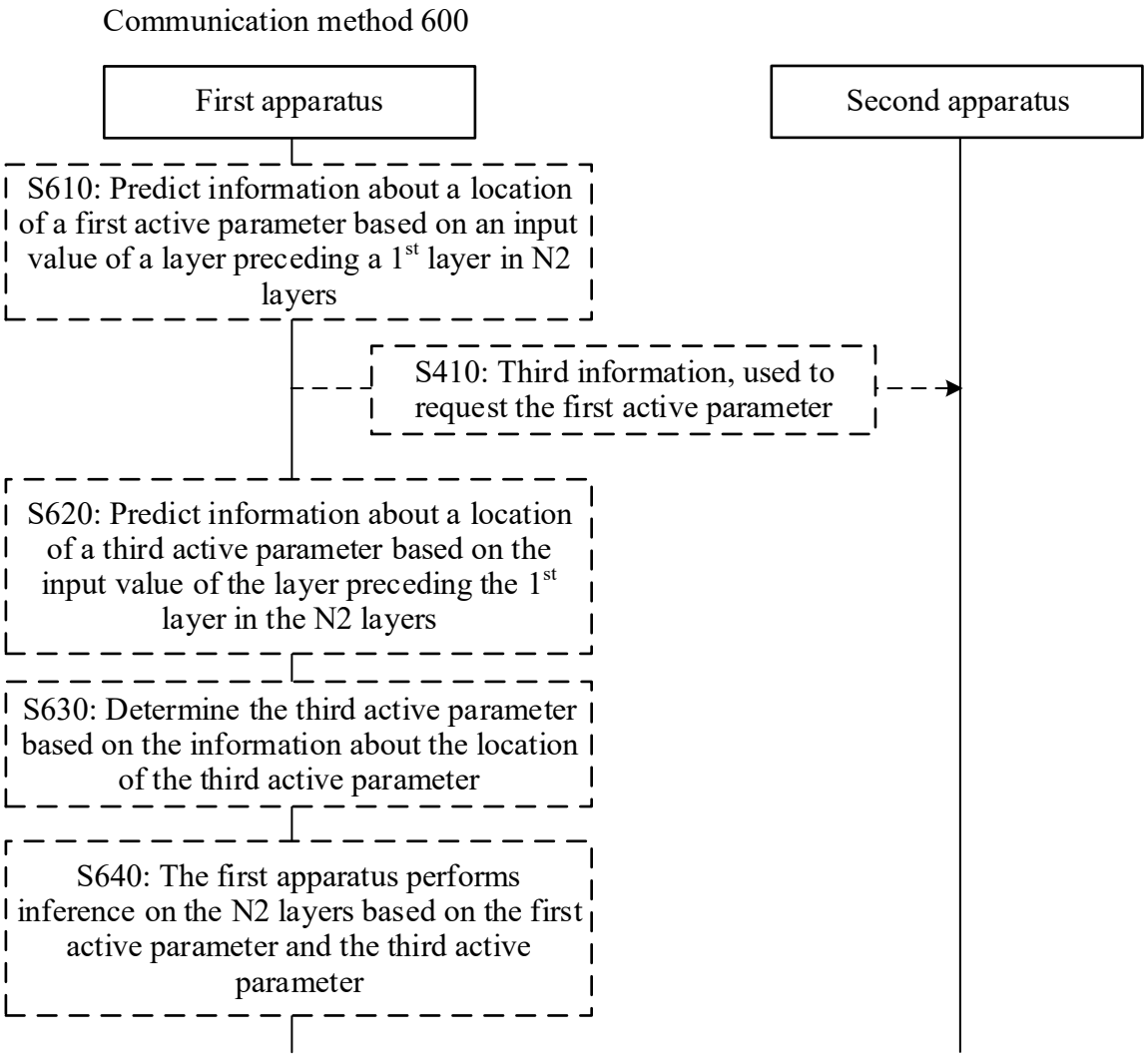

Communication method 600
First apparatus
Second apparatus
S610: Predict information about a location of a first active parameter based on an input value of a layer preceding a 1st layer in N2 layers
S410: Third information, used to request the first active parameter
S620: Predict information about a location of a third active parameter based on the input value of the layer preceding the 1st layer in the N2 layers
S630: Determine the third active parameter based on the information about the location of the third active parameter
S640: The first apparatus performs inference on the N2 layers based on the first active parameter and the third active parameter

FIG. 6

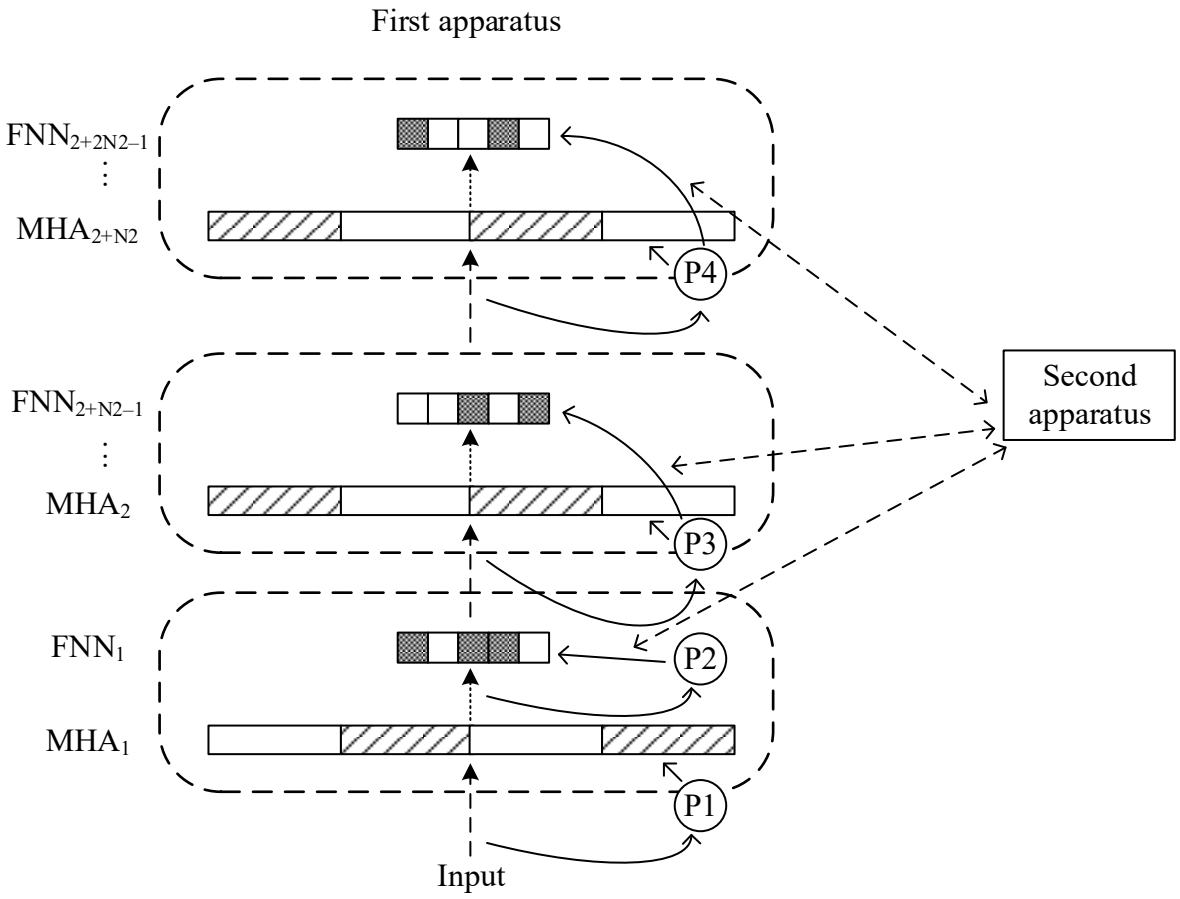

First apparatus
FNN2+2N2–1
MHA2+N2
P4
FNN2+N2–1
MHA2
P3
FNN1
MHA1
P2
P1
Input
Second apparatus

FIG. 7

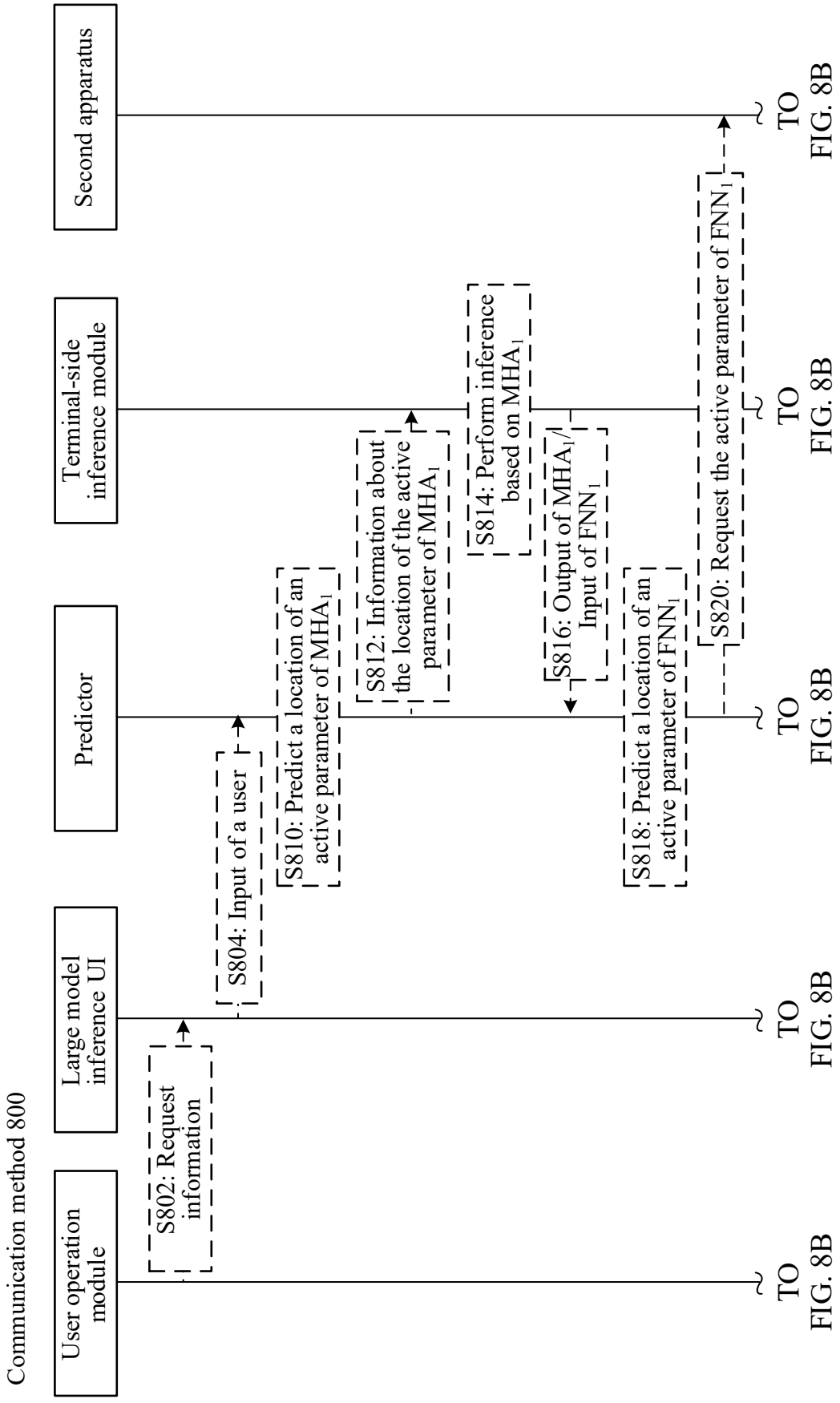

Communication method 800
User operation module
Large model inference UI
Predictor
Terminal-side inference module
Second apparatus
S802: Request information
S804: Input of a user
S810: Predict a location of an active parameter of MHA1
S812: Information about the location of the active parameter of MHA1
S814: Perform inference based on MHA1
S816: Output of MHA1/ Input of FNN1
S818: Predict a location of an active parameter of FNN1
S820: Request the active parameter of FNN1
TO FIG. 8B
TO FIG. 8B
TO FIG. 8B
TO FIG. 8B
TO FIG. 8B

FIG. 8A

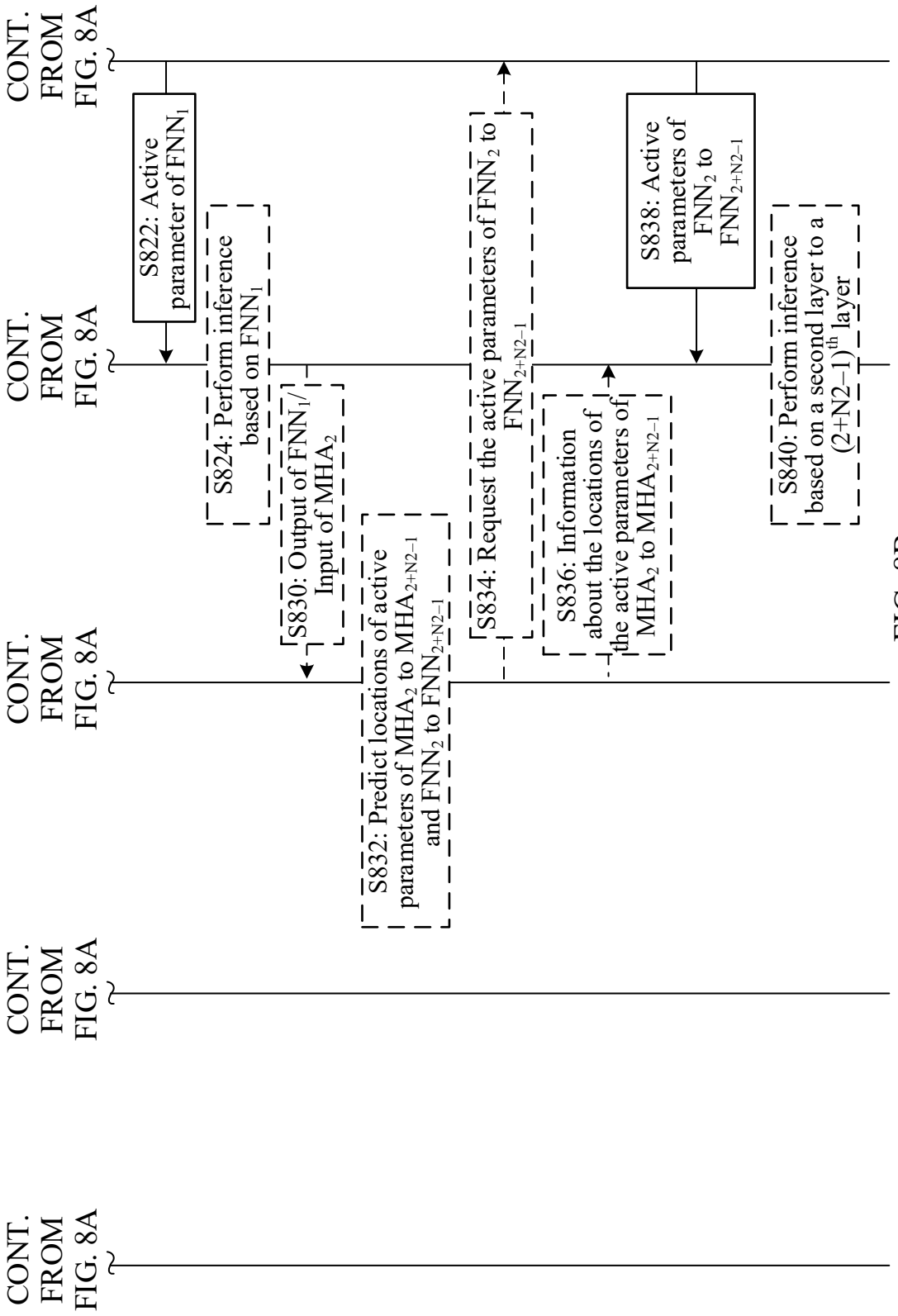

CONT. FROM FIG. 8A
S822: Active parameter of FNN_1
S824: Perform inference based on FNN_1
S830: Output of FNN_1/ Input of MHA_2
S832: Predict locations of active parameters of MHA_2 to MHA_{2+N2-1} and FNN_2 to FNN_{2+N2-1}
S834: Request the active parameters of FNN_2 to FNN_{2+N2-1}
S836: Information about the locations of the active parameters of MHA_2 to MHA_{2+N2-1}
S838: Active parameters of FNN_2 to FNN_{2+N2-1}
S840: Perform inference based on a second layer to a $(2+N2-1)^{th}$ layer

FIG. 8B

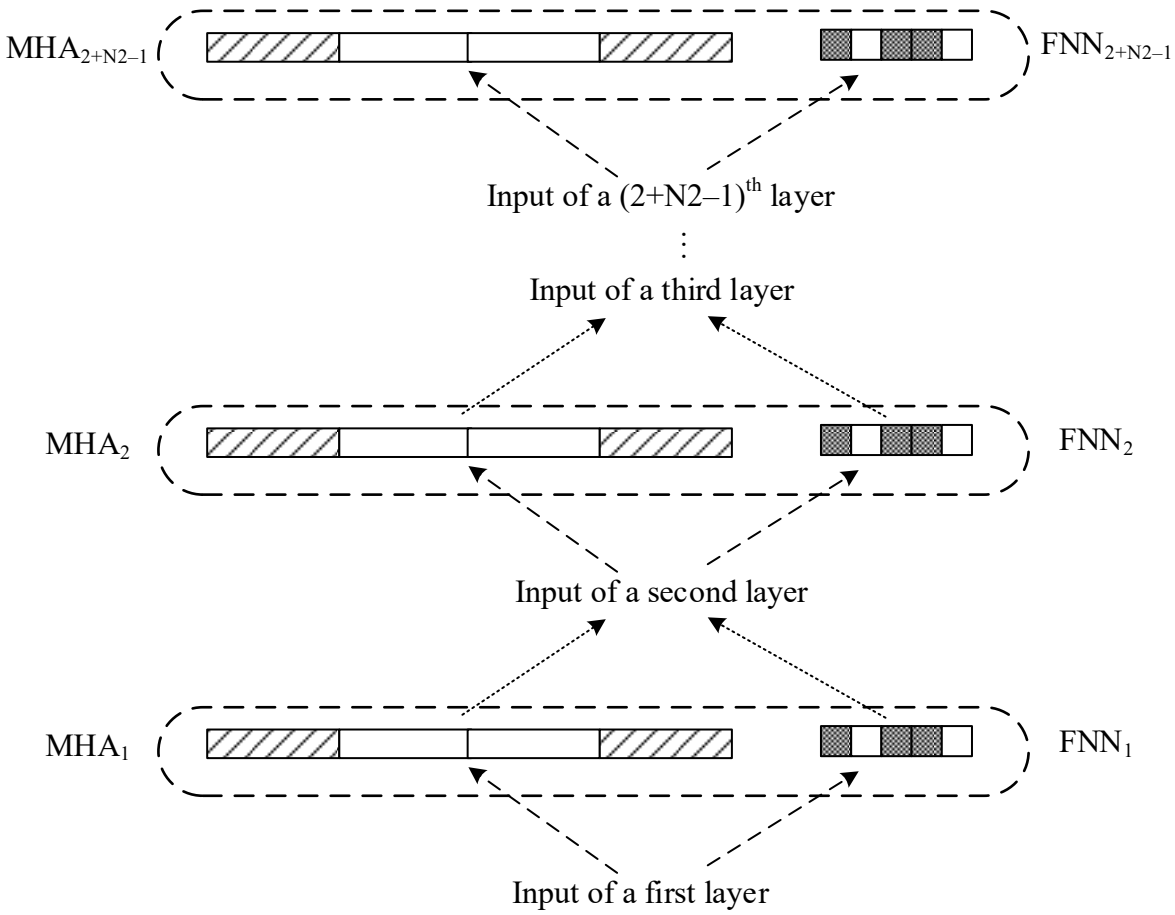

MHA2+N2–1
FNN2+N2–1
Input of a (2+N2–1)th layer
Input of a third layer
MHA2
FNN2
Input of a second layer
MHA1
FNN1
Input of a first layer

FIG. 9

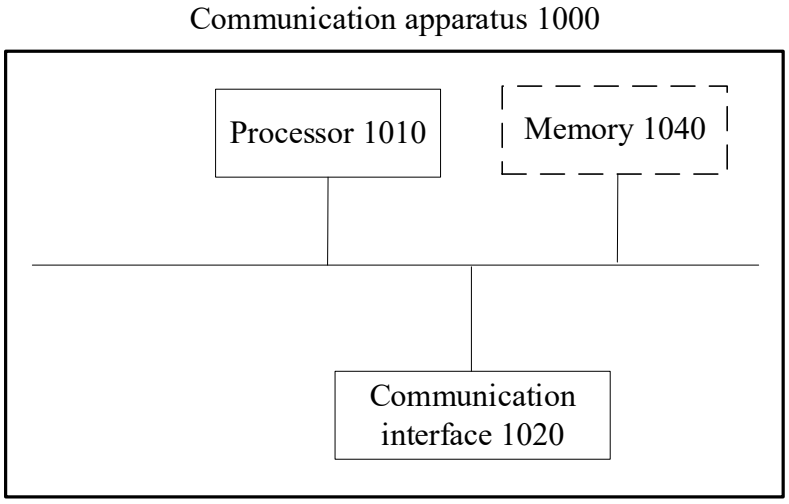

Communication apparatus 1000
Processor 1010
Memory 1040
Communication interface 1020

FIG. 10

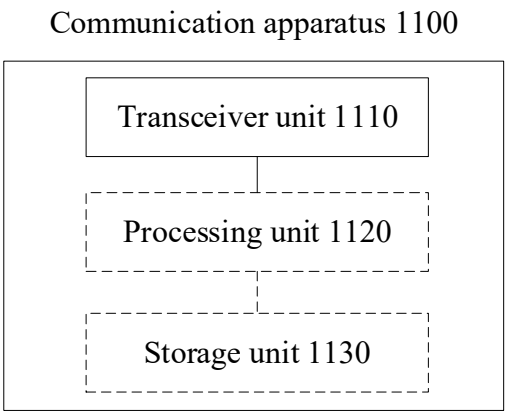

Communication apparatus 1100
Transceiver unit 1110
Processing unit 1120
Storage unit 1130

FIG. 11