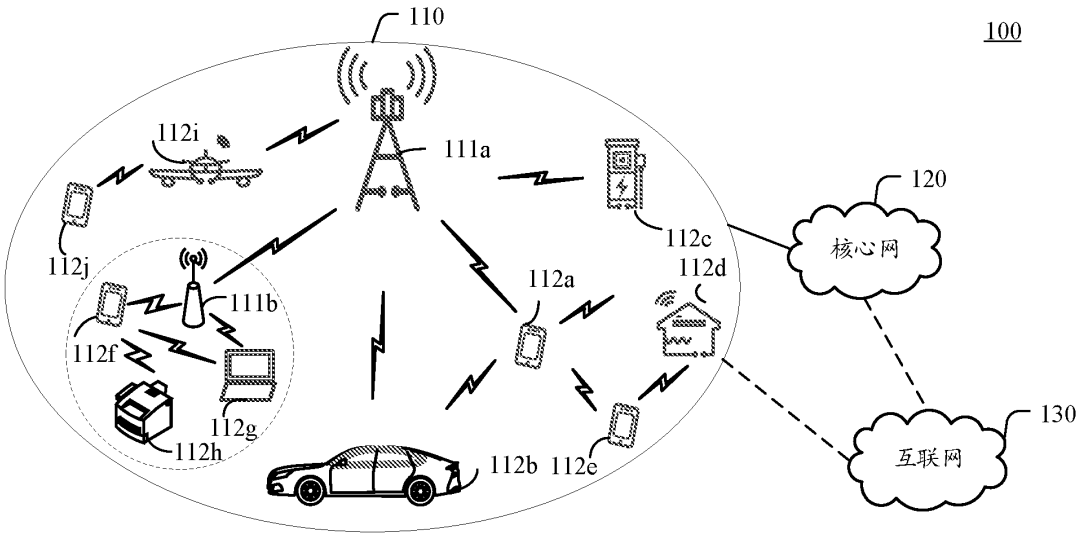

110
100
112i
111a
120
112j
112c
核心网
112d
111b
112a
112f
112g
112h
112b
112e
130
互联网

图 1

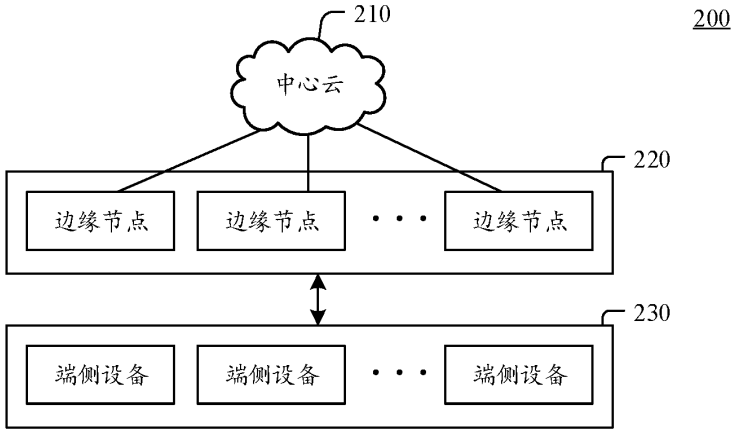

210
200
中心云
220
边缘节点
边缘节点
边缘节点
230
端侧设备
端侧设备
端侧设备

图 2

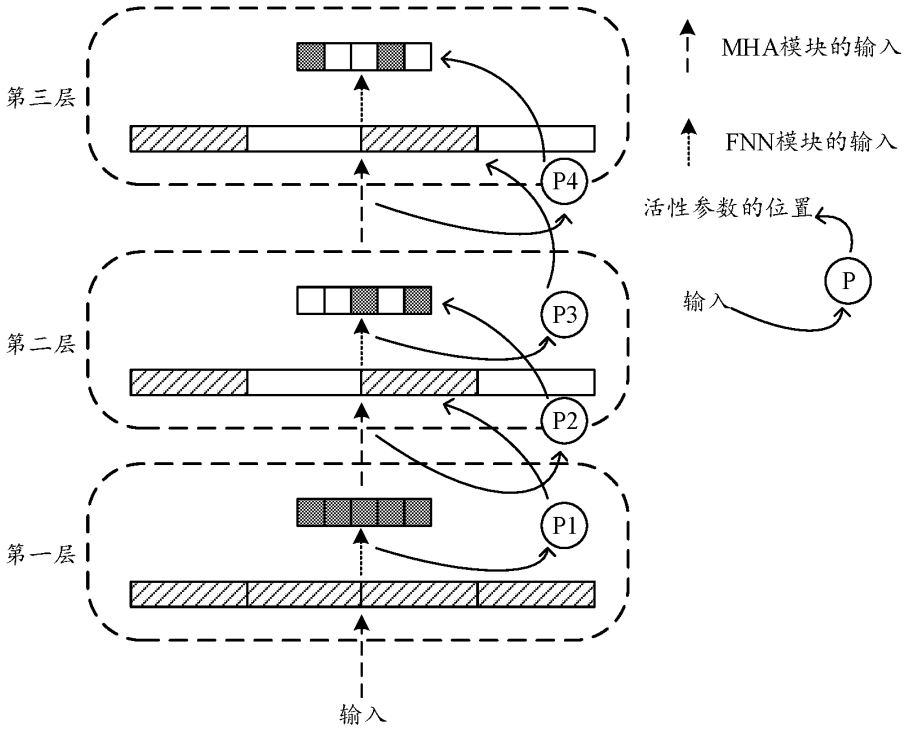


图 3

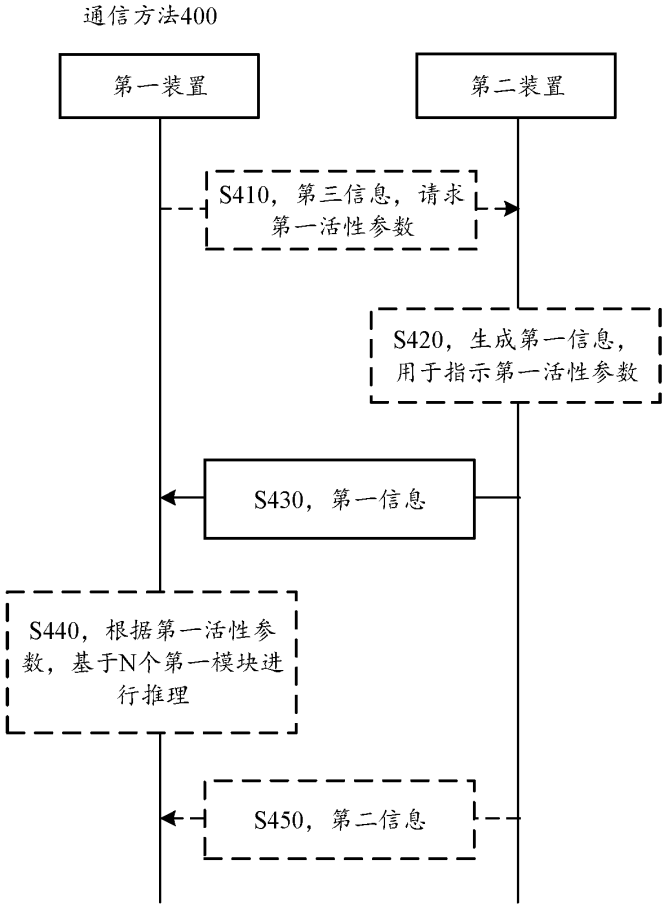

通信方法400
第一装置
第二装置
S410，第三信息，请求第一活性参数
S420，生成第一信息，用于指示第一活性参数
S430，第一信息
S440，根据第一活性参数，基于N个第一模块进行推理
S450，第二信息

图 4

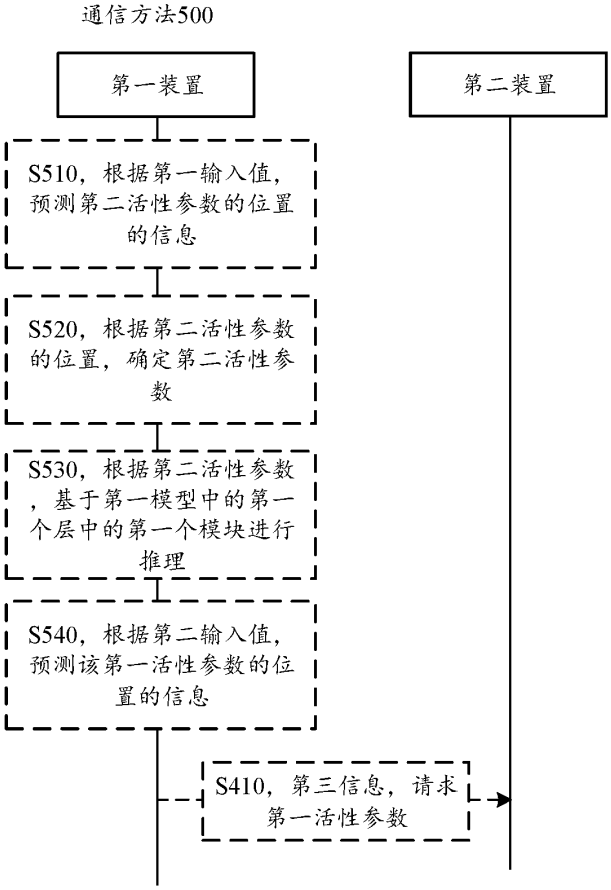

通信方法500
第一装置
第二装置
S510，根据第一输入值，预测第二活性参数的位置的信息
S520，根据第二活性参数的位置，确定第二活性参数
S530，根据第二活性参数，基于第一模型中的第一个层中的第一个模块进行推理
S540，根据第二输入值，预测该第一活性参数的位置的信息
S410，第三信息，请求第一活性参数

图 5

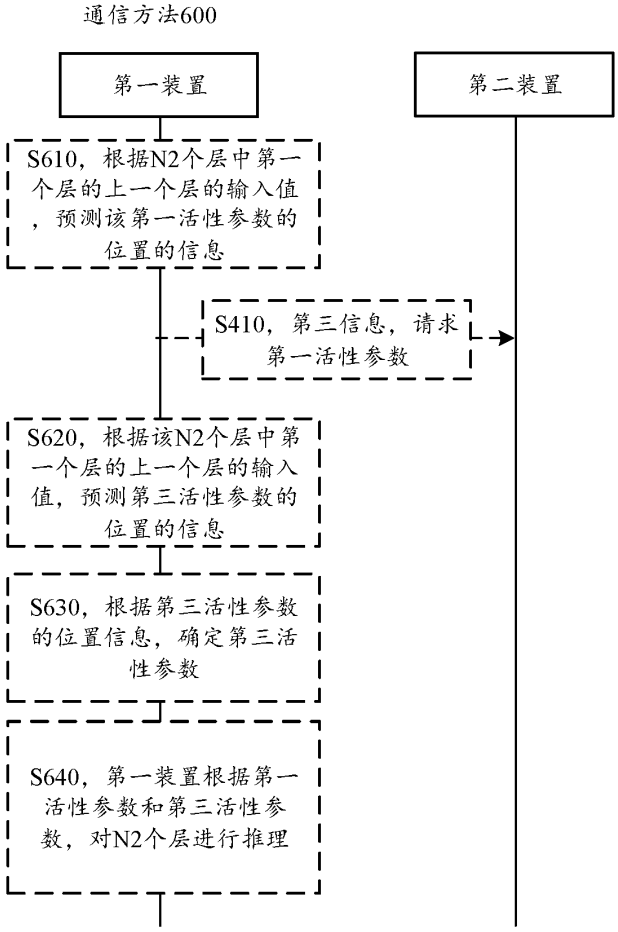

通信方法600
第一装置
第二装置
S610，根据N2个层中第一个层的上一个层的输入值，预测该第一活性参数的位置的信息
S410，第三信息，请求第一活性参数
S620，根据该N2个层中第一个层的上一个层的输入值，预测第三活性参数的位置的信息
S630，根据第三活性参数的位置信息，确定第三活性参数
S640，第一装置根据第一活性参数和第三活性参数，对N2个层进行推理

图 6

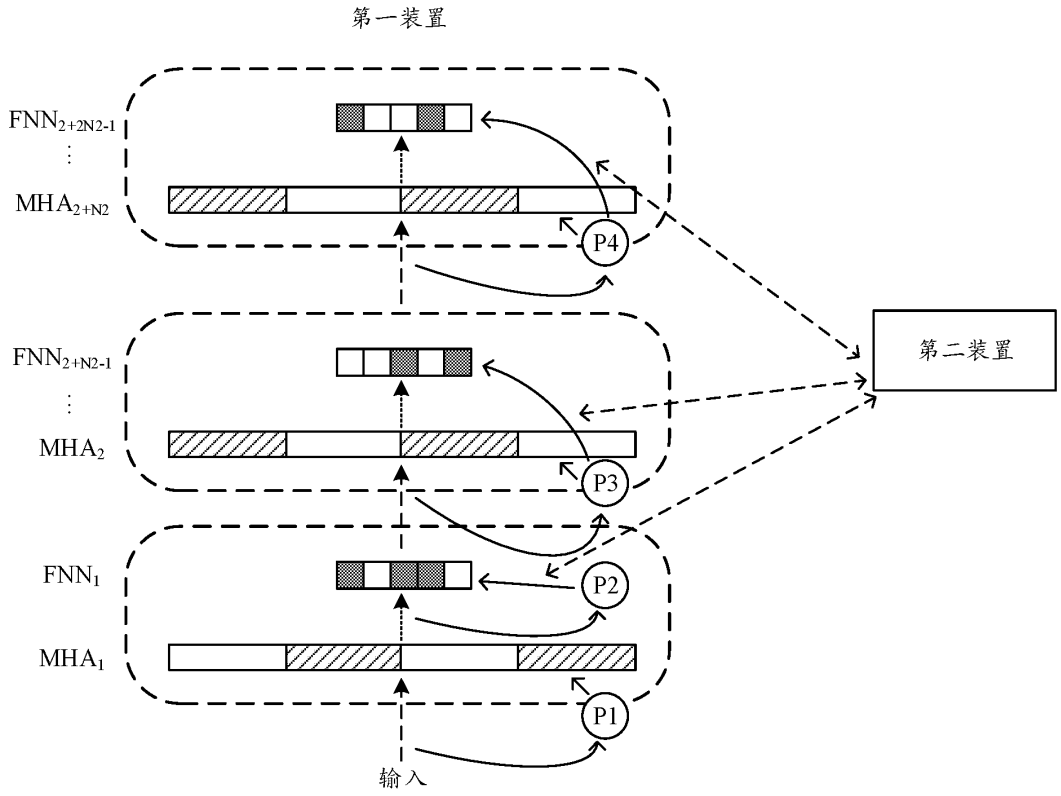

第一装置
FNN2+2N2-1
MHA2+N2
P4
FNN2+N2-1
MHA2
P3
FNN1
P2
MHA1
P1
输入
第二装置

图 7

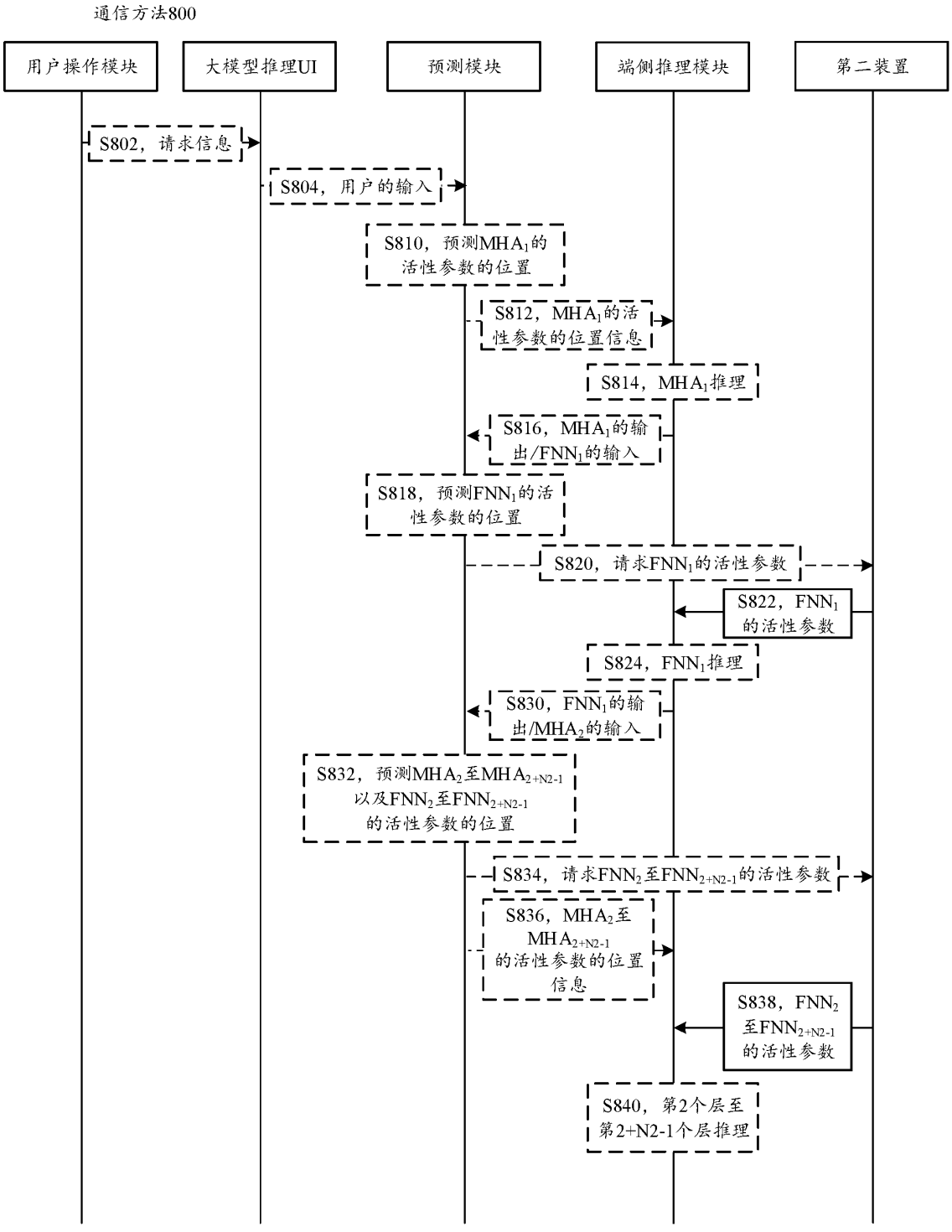

通信方法800
用户操作模块
大模型推理UI
预测模块
端侧推理模块
第二装置
S802，请求信息
S804，用户的输入
S810，预测MHA1的活性参数的位置
S812，MHA1的活性参数的位置信息
S814，MHA1推理
S816，MHA1的输出/FNN1的输入
S818，预测FNN1的活性参数的位置
S820，请求FNN1的活性参数
S822，FNN1的活性参数
S824，FNN1推理
S830，FNN1的输出/MHA2的输入
S832，预测MHA2至MHA2+N2-1以及FNN2至FNN2+N2-1的活性参数的位置
S834，请求FNN2至FNN2+N2-1的活性参数
S836，MHA2至MHA2+N2-1的活性参数的位置信息
S838，FNN2至FNN2+N2-1的活性参数
S840，第2个层至第2+N2-1个层推理

图 8

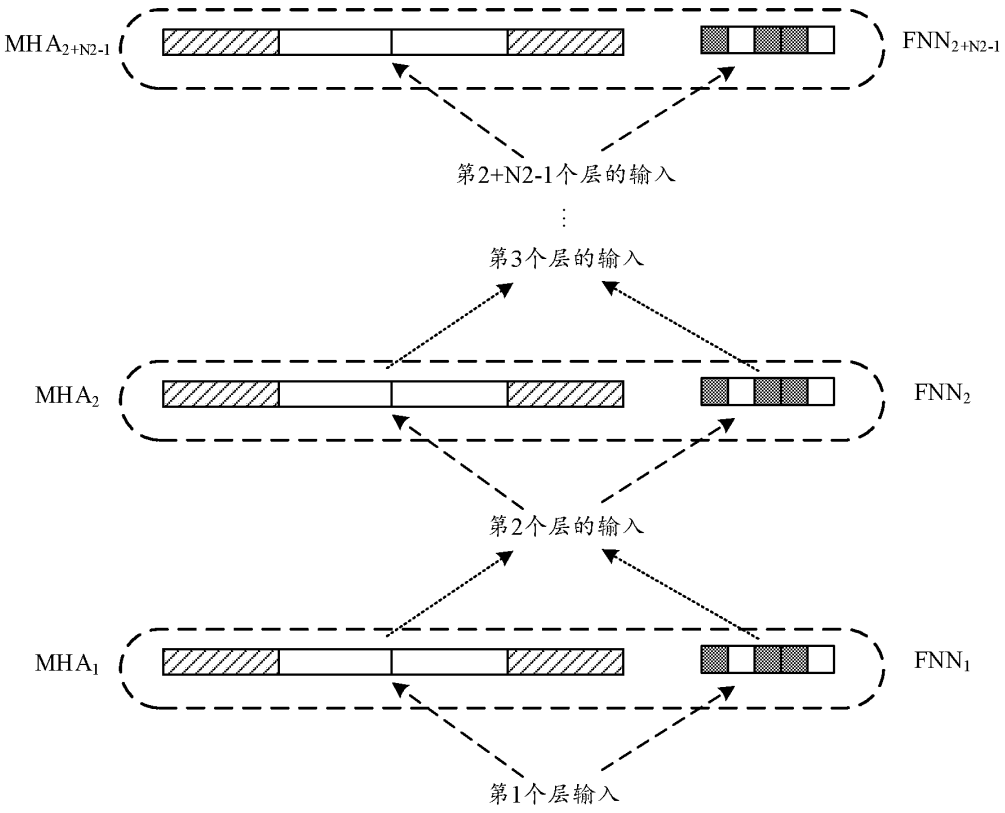

MHA2+N2-1
FNN2+N2-1
第2+N2-1个层的输入
第3个层的输入
MHA2
FNN2
第2个层的输入
MHA1
FNN1
第1个层输入

图 9

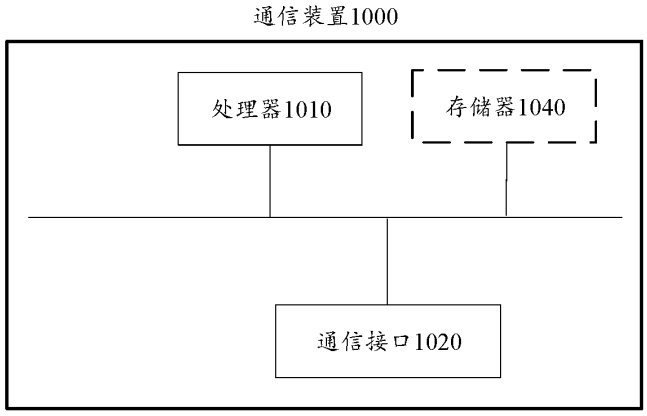

通信装置1000
处理器1010
存储器1040
通信接口1020

图 10

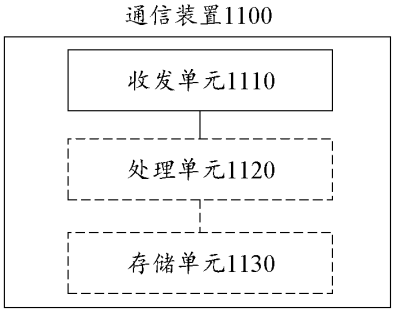


图 11