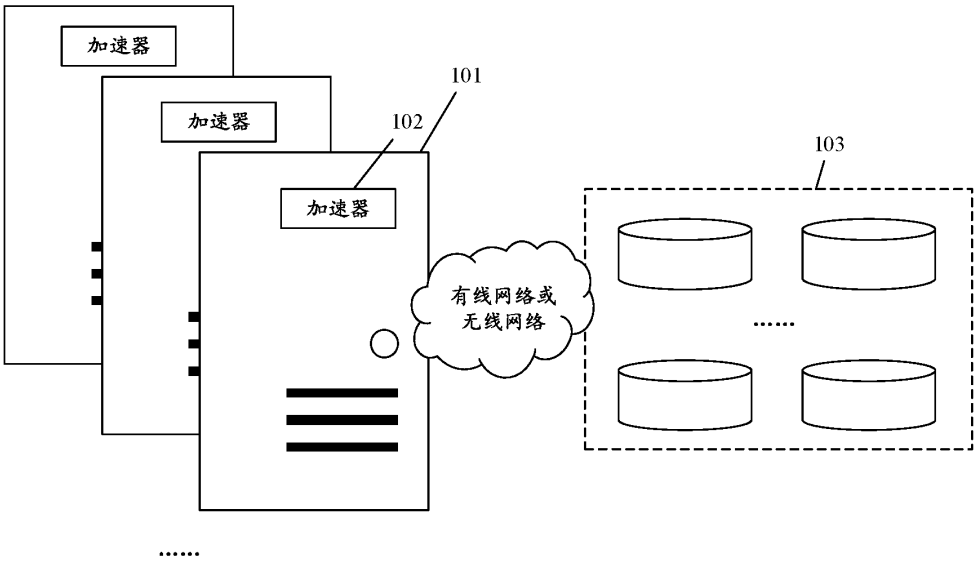
加速器
加速器
加速器
101
102
103
有线网络或
无线网络
.....
.....

图1

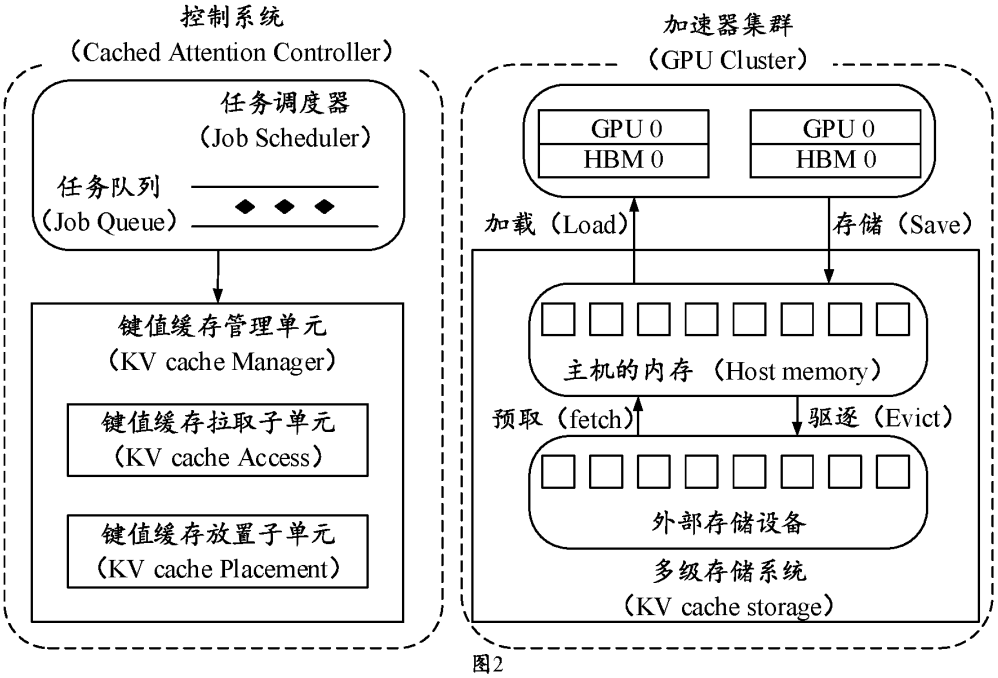
控制系统
（Cached Attention Controller）
任务调度器
（Job Scheduler）
任务队列
（Job Queue）
键值缓存管理单元
（KV cache Manager）
键值缓存拉取子单元
（KV cache Access）
键值缓存放置子单元
（KV cache Placement）
加速器集群
（GPU Cluster）
GPU 0
HBM 0
GPU 0
HBM 0
加载（Load）
存储（Save）
主机的内存（Host memory）
预取（fetch）
驱逐（Evict）
外部存储设备
多级存储系统
（KV cache storage）

图2

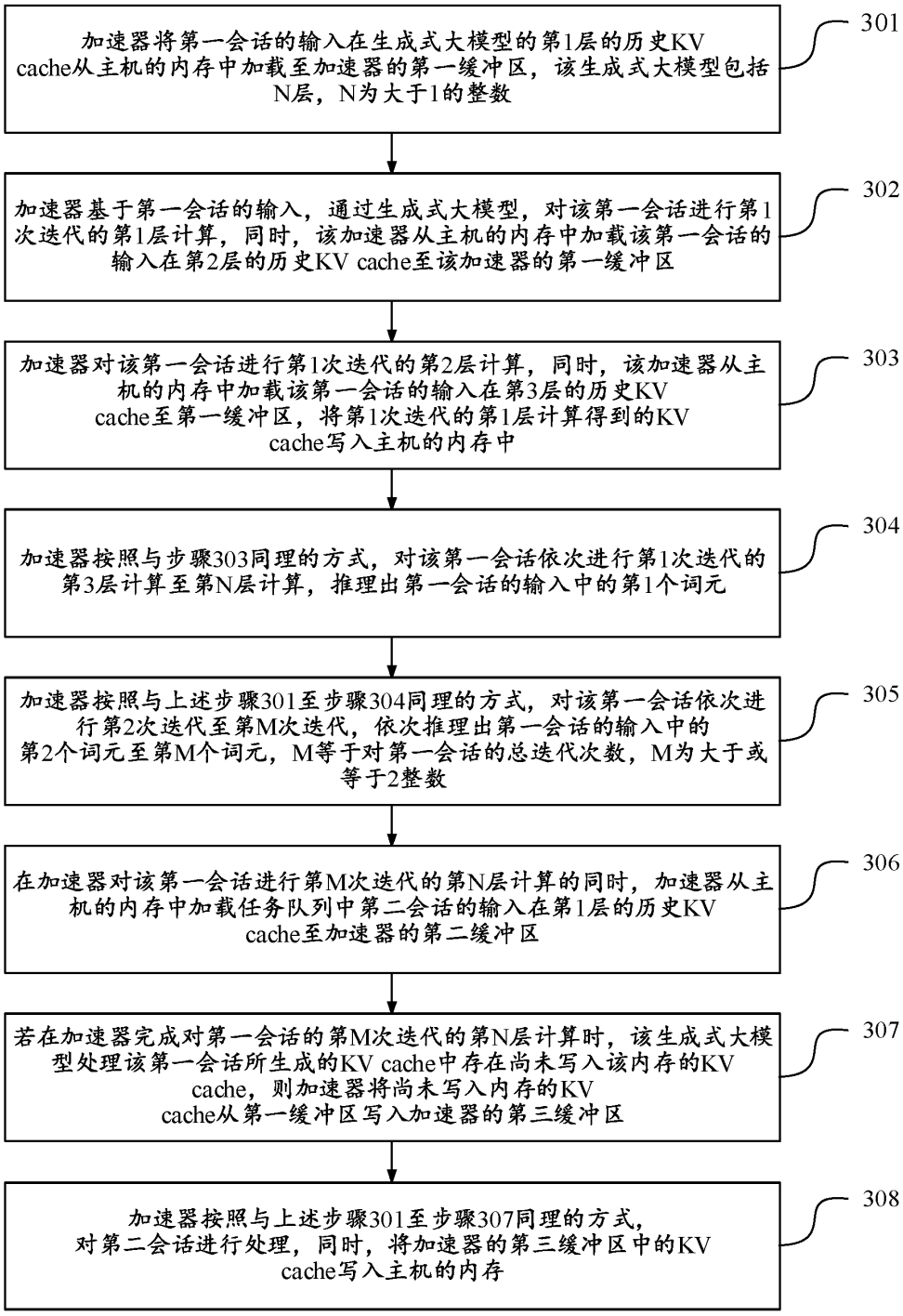
加速器将第一会话的输入在生成式大模型的第1层的历史KV cache从主机的内存中加载至加速器的第一缓冲区，该生成式大模型包括N层，N为大于1的整数
301
加速器基于第一会话的输入，通过生成式大模型，对该第一会话进行第1次迭代的第1层计算，同时，该加速器从主机的内存中加载该第一会话的输入在第2层的历史KV cache至该加速器的第一缓冲区
302
加速器对该第一会话进行第1次迭代的第2层计算，同时，该加速器从主机的内存中加载该第一会话的输入在第3层的历史KV cache至第一缓冲区，将第1次迭代的第1层计算得到的KV cache写入主机的内存中
303
加速器按照与步骤303同理的方式，对该第一会话依次进行第1次迭代的第3层计算至第N层计算，推理出第一会话的输入中的第1个词元
304
加速器按照与上述步骤301至步骤304同理的方式，对该第一会话依次进行第2次迭代至第M次迭代，依次推理出第一会话的输入中的第2个词元至第M个词元，M等于对第一会话的总迭代次数，M为大于或等于2整数
305
在加速器对该第一会话进行第M次迭代的第N层计算的同时，加速器从主机的内存中加载任务队列中第二会话的输入在第1层的历史KV cache至加速器的第二缓冲区
306
若在加速器完成对第一会话的第M次迭代的第N层计算时，该生成式大模型处理该第一会话所生成的KV cache中存在尚未写入该内存的KV cache，则加速器将尚未写入内存的KV cache从第一缓冲区写入加速器的第三缓冲区
307
加速器按照与上述步骤301至步骤307同理的方式，对第二会话进行处理，同时，将加速器的第三缓冲区中的KV cache写入主机的内存
308

图3

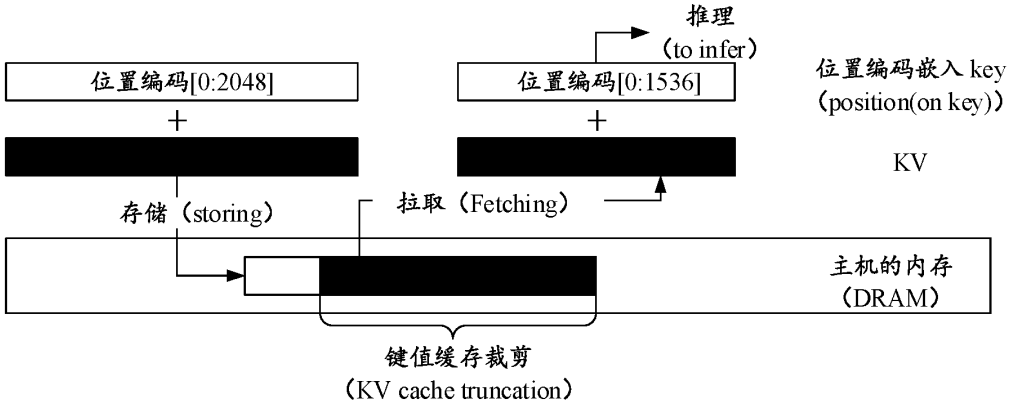
推理
(to infer)
位置编码[0:2048]
位置编码[0:1536]
位置编码嵌入 key
(position(on key))
KV
存储（storing）
拉取（Fetching）
主机的内存
(DRAM)
键值缓存裁剪
(KV cache truncation)

图4

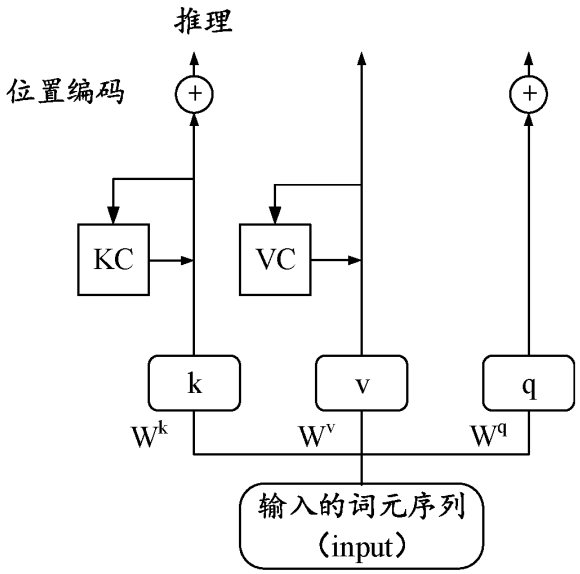
推理
位置编码
KC
VC
k
v
q
W^k
W^v
W^q
输入的词元序列
(input)

图5

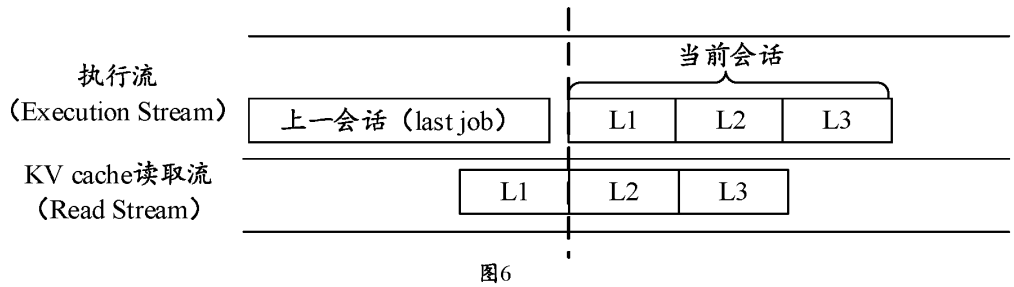
执行流
(Execution Stream)
当前会话
上一会话（last job）
L1
L2
L3
KV cache读取流
(Read Stream)
L1
L2
L3

图6

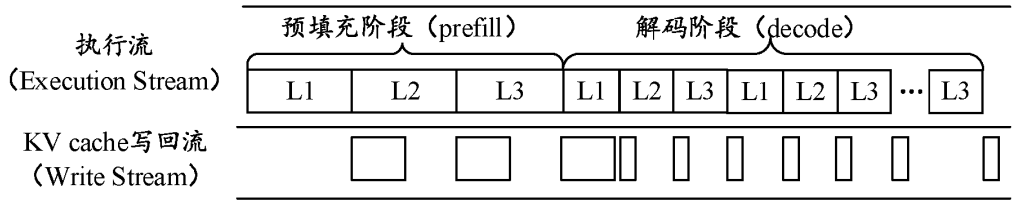
执行流
（Execution Stream）
预填充阶段（prefill）
解码阶段（decode）
L1
L2
L3
L1
L2
L3
L1
L2
L3
…
L3
KV cache写回流
（Write Stream）

图7

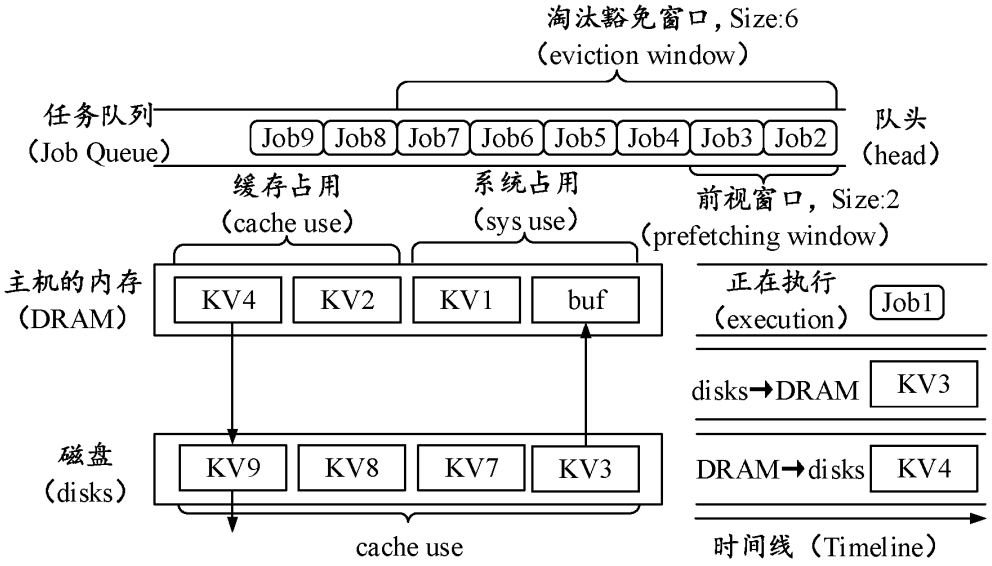
淘汰豁免窗口，Size:6
（eviction window）
任务队列
（Job Queue）
Job9
Job8
Job7
Job6
Job5
Job4
Job3
Job2
队头
（head）
缓存占用
（cache use）
系统占用
（sys use）
前视窗口，Size:2
（prefetching window）
主机的内存
（DRAM）
KV4
KV2
KV1
buf
正在执行
（execution）
Job1
disks→DRAM
KV3
磁盘
（disks）
KV9
KV8
KV7
KV3
DRAM→disks
KV4
cache use
时间线（Timeline）

图8

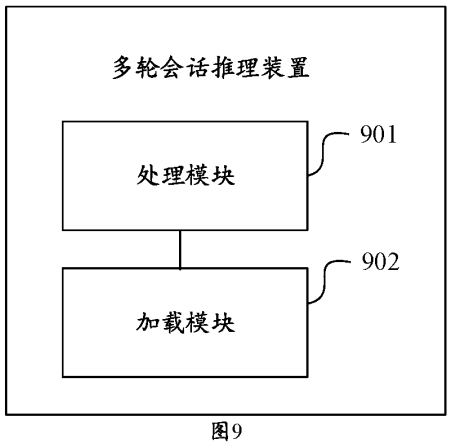
多轮会话推理装置
处理模块
901
加载模块
902

图9

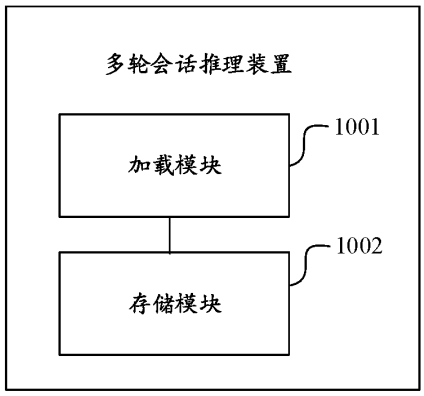
多轮会话推理装置
加载模块
1001
存储模块
1002

图10

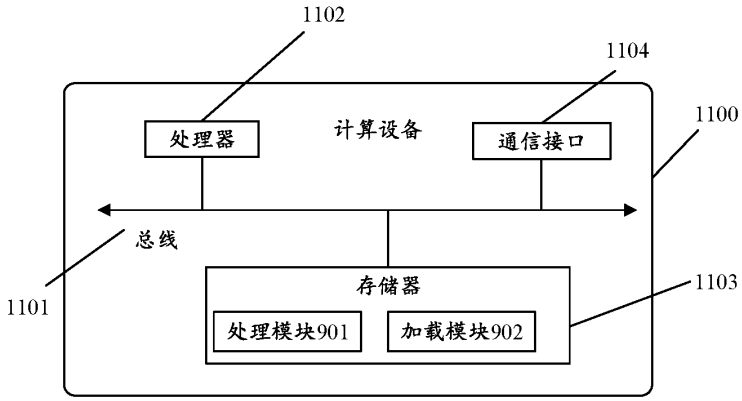
1102
1104
处理器
计算设备
通信接口
1100
总线
1101
存储器
1103
处理模块901
加载模块902

图11

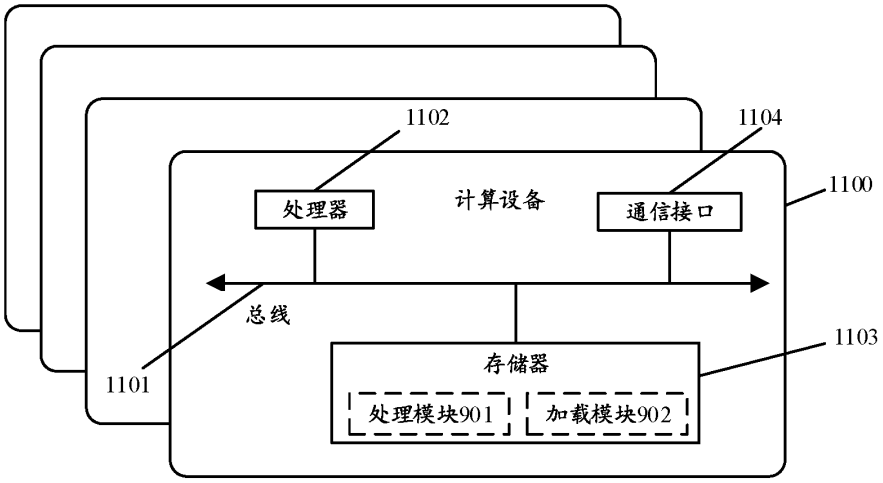
1102
1104
处理器
计算设备
通信接口
1100
总线
1101
存储器
1103
处理模块901
加载模块902

图12

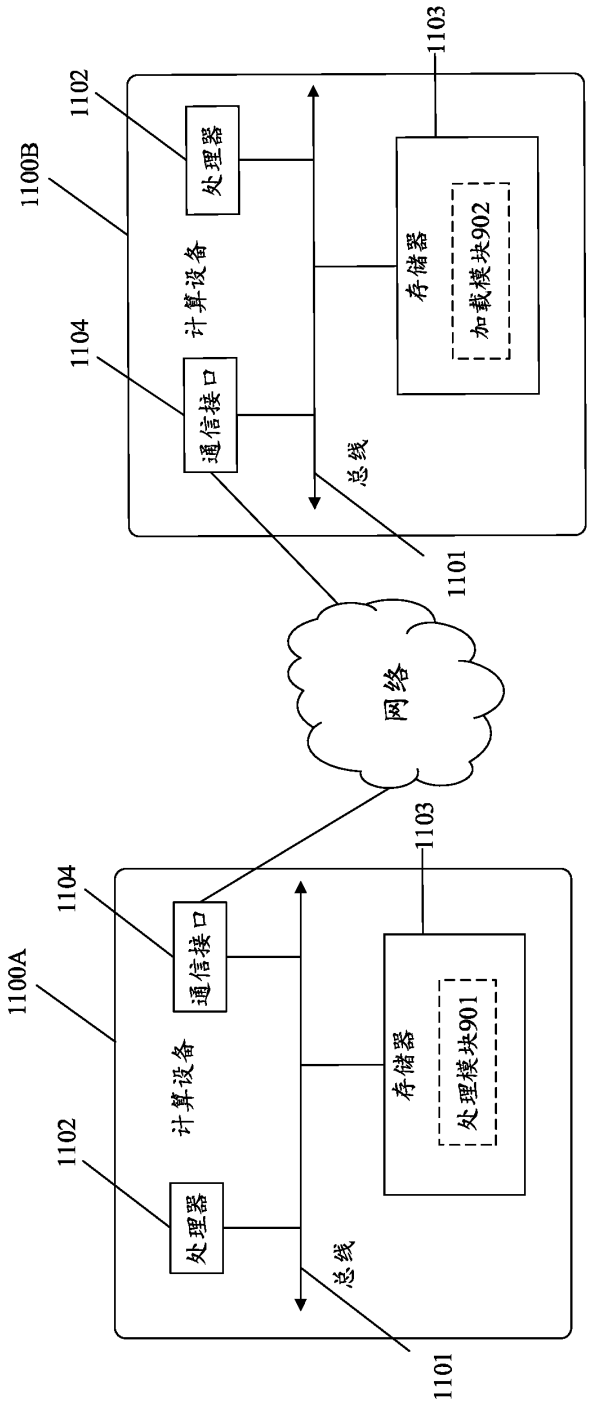
1100B
1102
处理器
1103
计算设备
存储器
加载模块902
1104
通信接口
总线
1101
网络
1103
1104
通信接口
1100A
计算设备
存储器
处理模块901
1102
处理器
总线
1101

图13