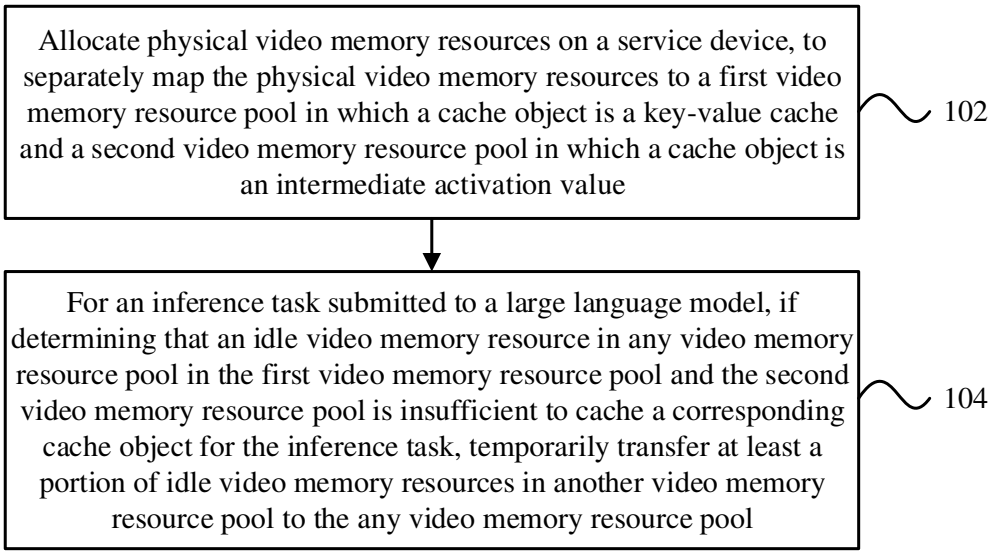

Allocate physical video memory resources on a service device, to separately map the physical video memory resources to a first video memory resource pool in which a cache object is a key-value cache and a second video memory resource pool in which a cache object is an intermediate activation value
102
For an inference task submitted to a large language model, if determining that an idle video memory resource in any video memory resource pool in the first video memory resource pool and the second video memory resource pool is insufficient to cache a corresponding cache object for the inference task, temporarily transfer at least a portion of idle video memory resources in another video memory resource pool to the any video memory resource pool
104

FIG. 1

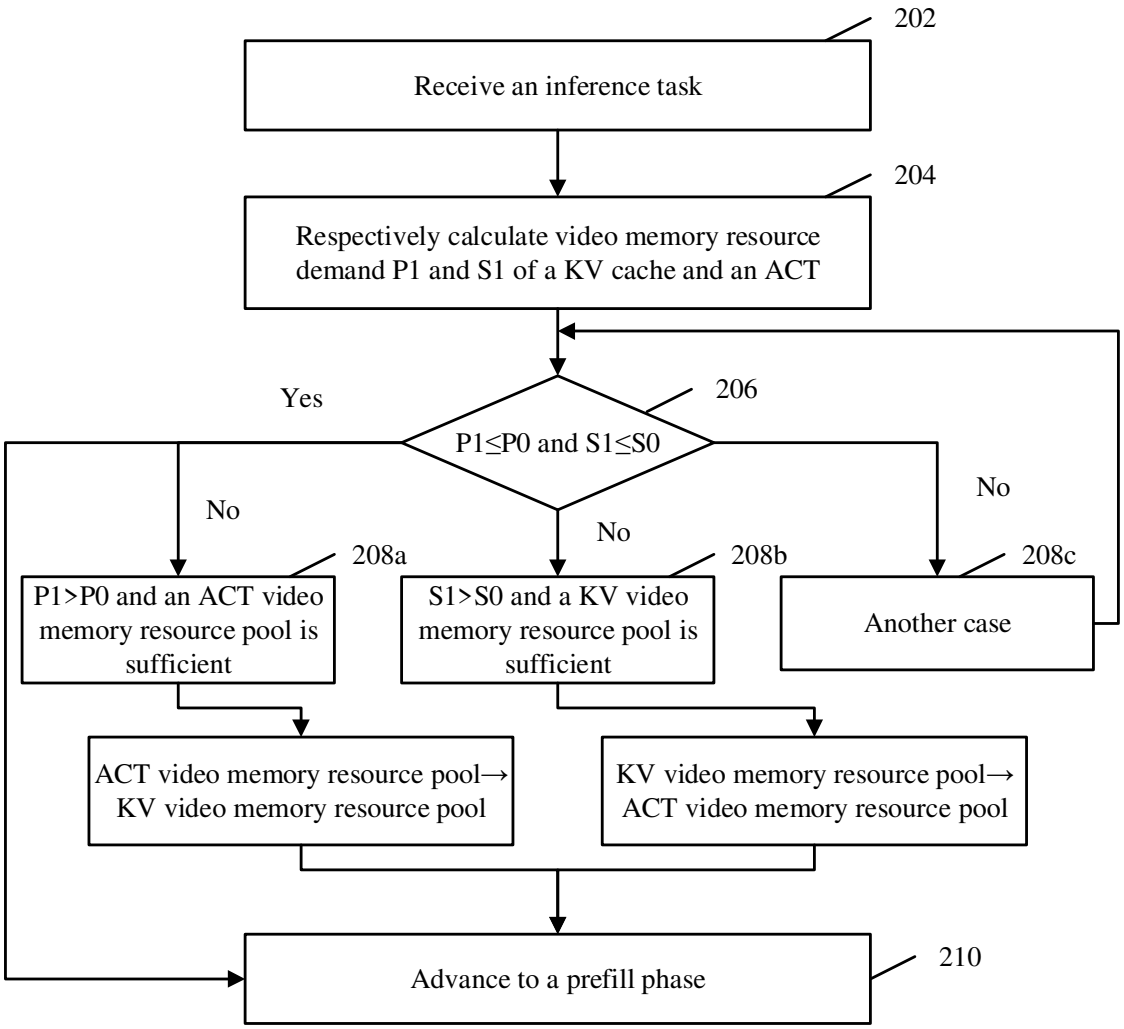

202
Receive an inference task
204
Respectively calculate video memory resource demand P1 and S1 of a KV cache and an ACT
206
P1≤P0 and S1≤S0
Yes
No
No
No
208a
P1>P0 and an ACT video memory resource pool is sufficient
208b
S1>S0 and a KV video memory resource pool is sufficient
208c
Another case
ACT video memory resource pool→ KV video memory resource pool
KV video memory resource pool→ ACT video memory resource pool
210
Advance to a prefill phase

FIG. 2

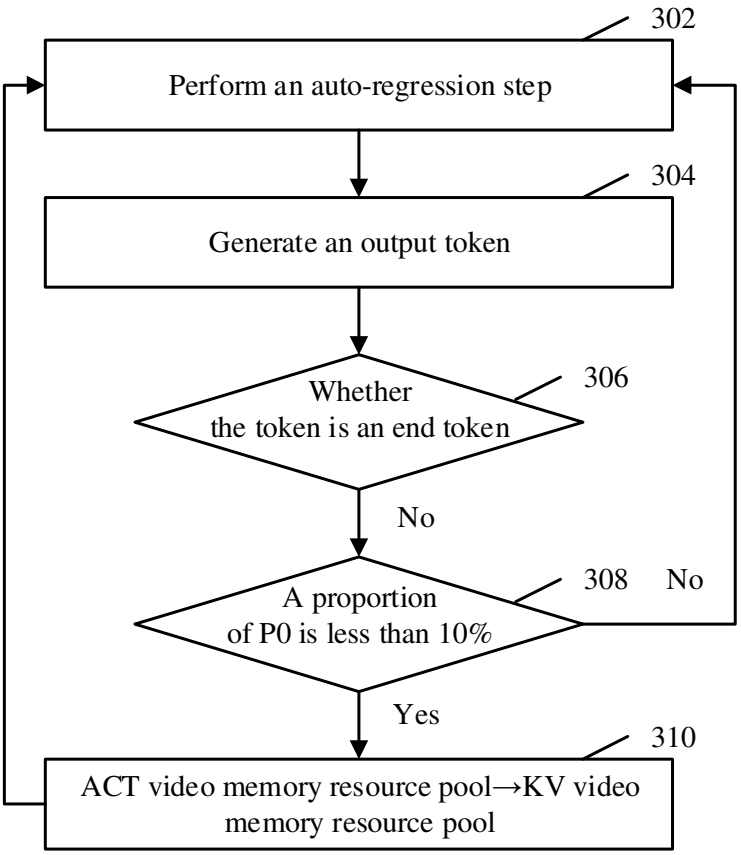

302
Perform an auto-regression step
304
Generate an output token
306
Whether
the token is an end token
No
308
No
A proportion
of P0 is less than 10%
Yes
310
ACT video memory resource pool→KV video
memory resource pool

FIG. 3

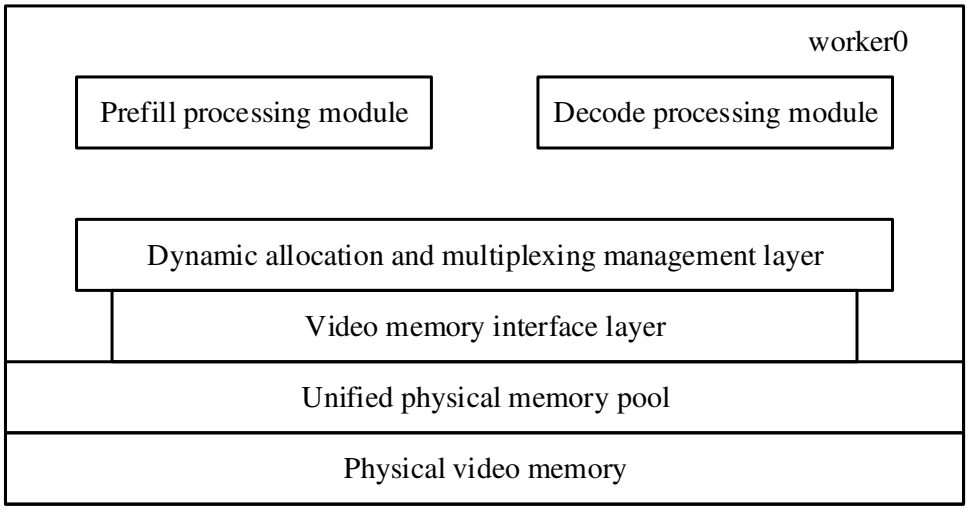

worker0
Prefill processing module
Decode processing module
Dynamic allocation and multiplexing management layer
Video memory interface layer
Unified physical memory pool
Physical video memory

FIG. 4

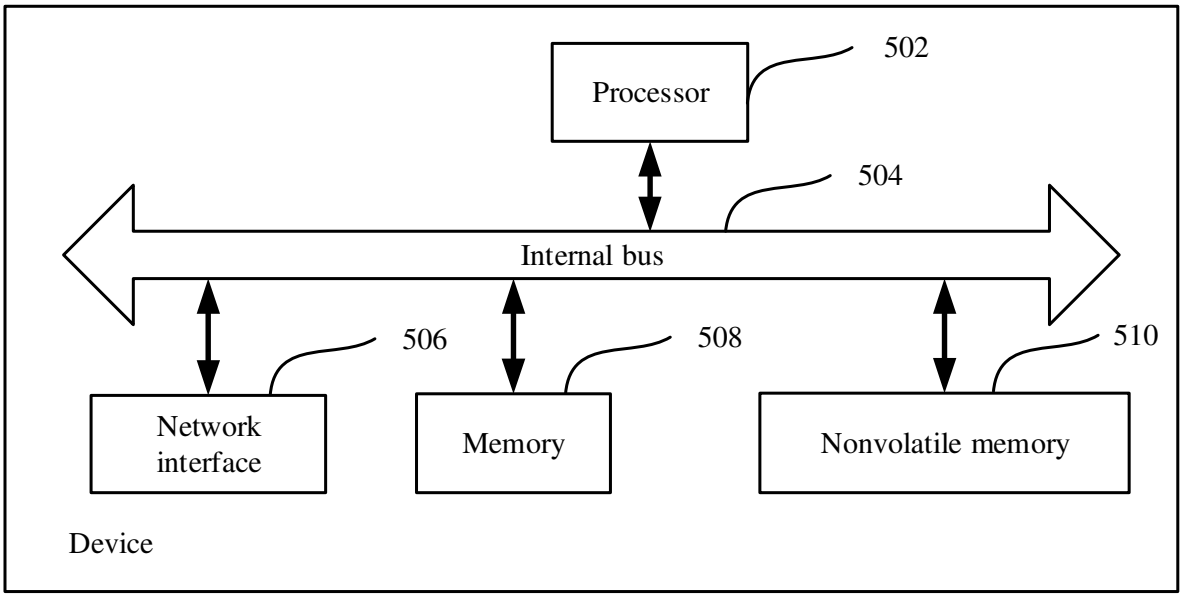

Processor
502
504
Internal bus
506
508
510
Network interface
Memory
Nonvolatile memory
Device

FIG. 5

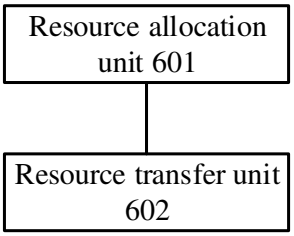

Resource allocation unit 601
Resource transfer unit 602

FIG. 6