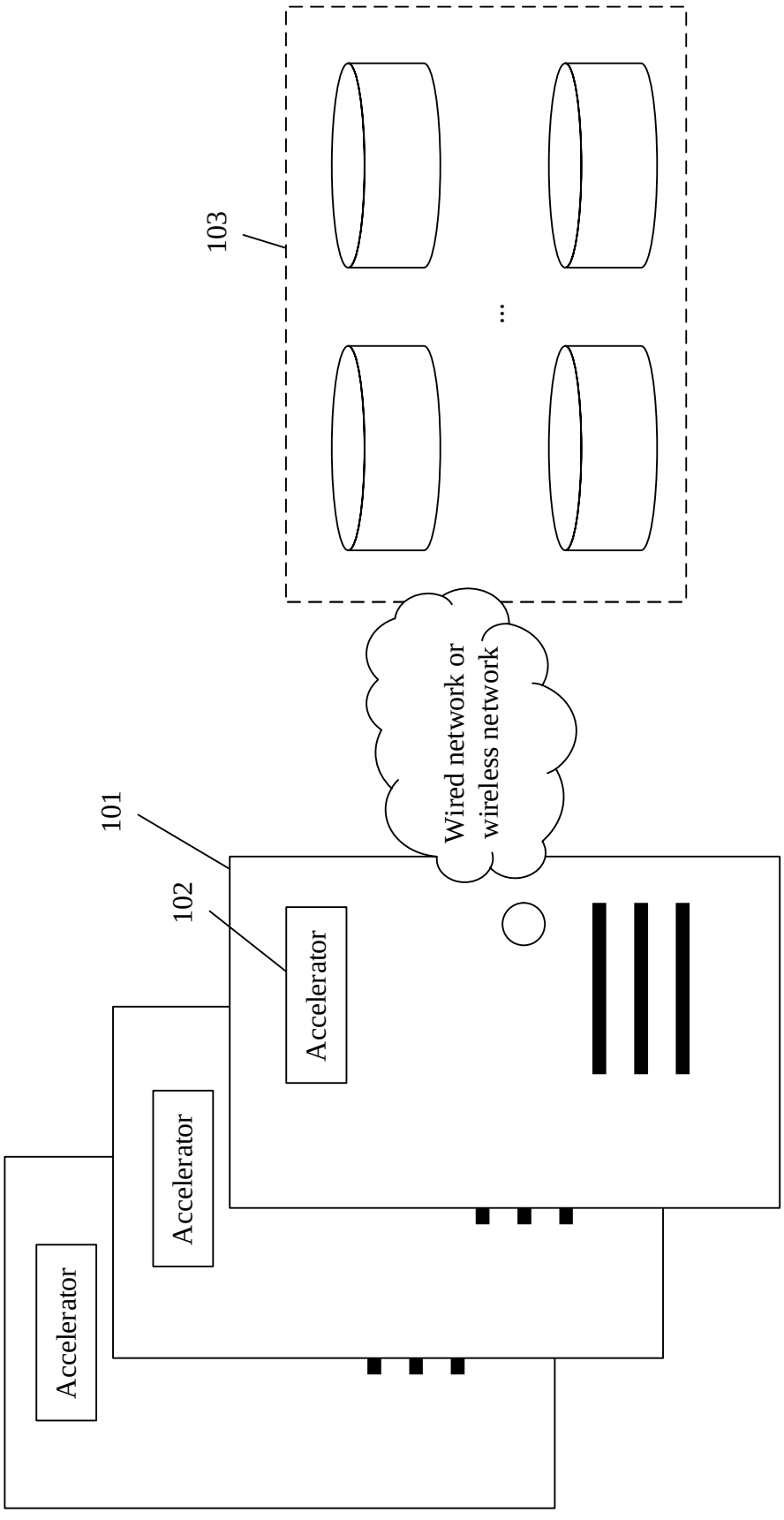

103
...
Wired network or wireless network
101
102
Accelerator
Accelerator
Accelerator
...

FIG. 1

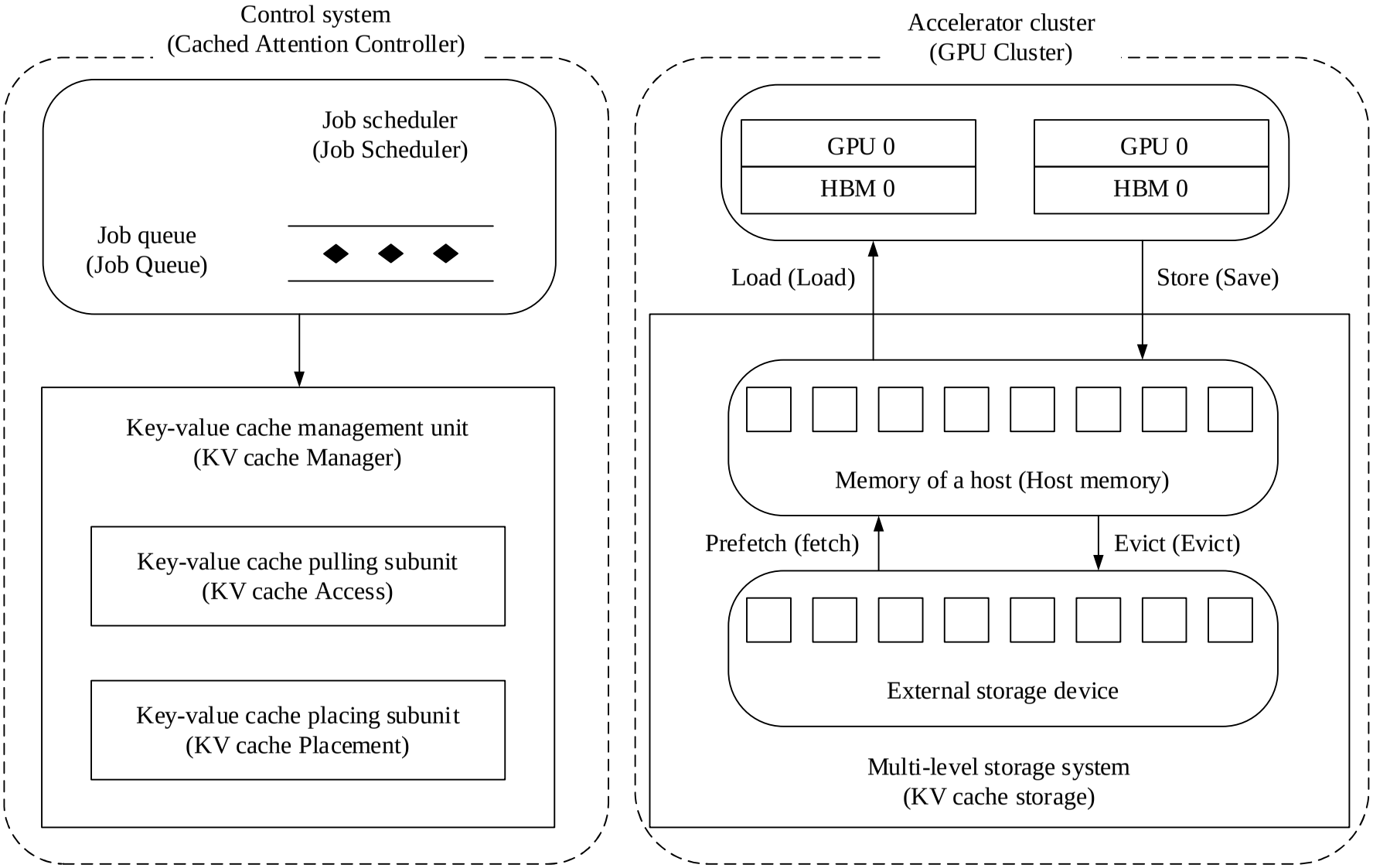

Control system
(Cached Attention Controller)
Job scheduler
(Job Scheduler)
Job queue
(Job Queue)
Key-value cache management unit
(KV cache Manager)
Key-value cache pulling subunit
(KV cache Access)
Key-value cache placing subunit
(KV cache Placement)
Accelerator cluster
(GPU Cluster)
GPU 0
HBM 0
GPU 0
HBM 0
Load (Load)
Store (Save)
Memory of a host (Host memory)
Prefetch (fetch)
Evict (Evict)
External storage device
Multi-level storage system
(KV cache storage)

FIG. 2

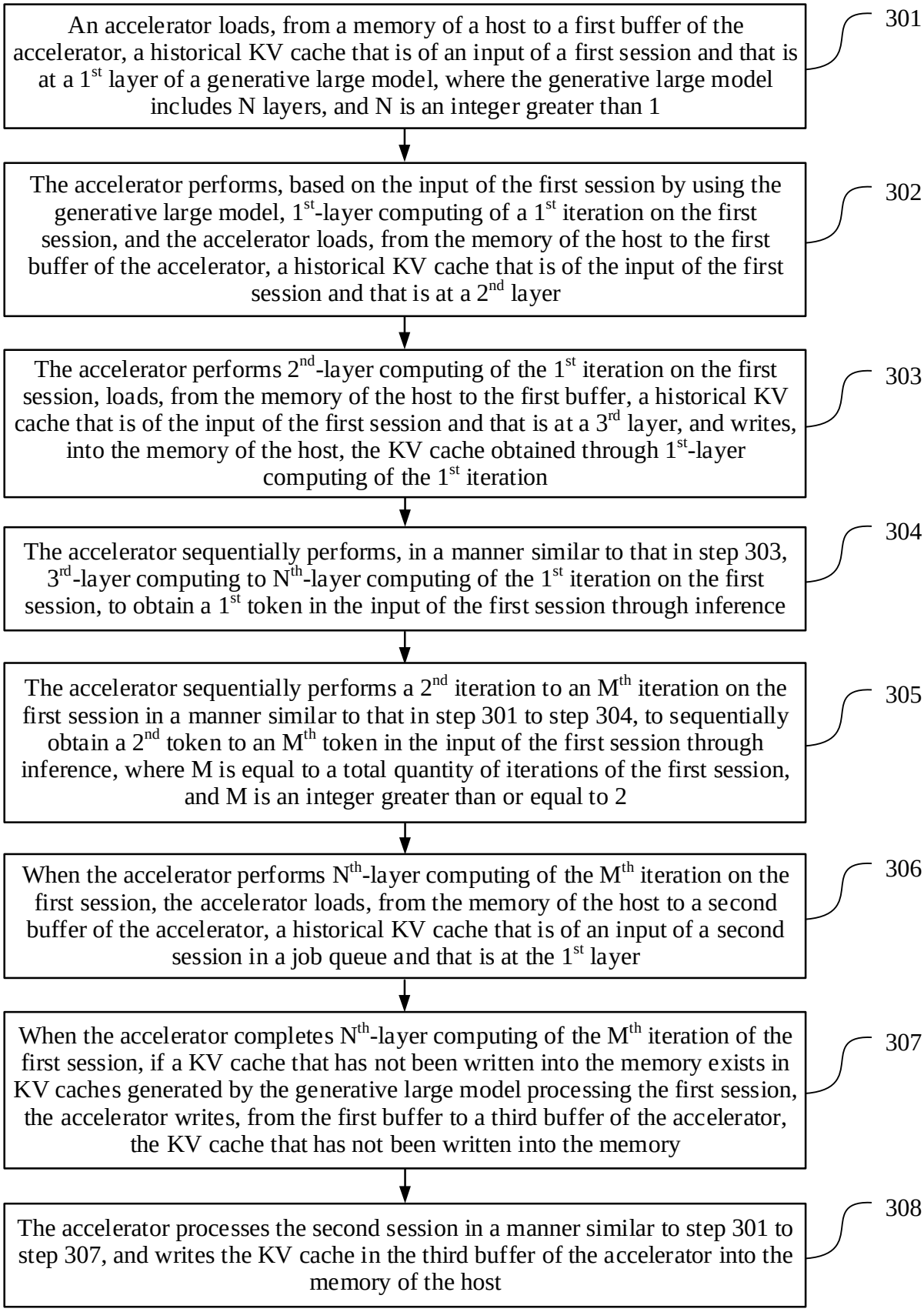

An accelerator loads, from a memory of a host to a first buffer of the accelerator, a historical KV cache that is of an input of a first session and that is at a 1st layer of a generative large model, where the generative large model includes N layers, and N is an integer greater than 1
301
The accelerator performs, based on the input of the first session by using the generative large model, 1st-layer computing of a 1st iteration on the first session, and the accelerator loads, from the memory of the host to the first buffer of the accelerator, a historical KV cache that is of the input of the first session and that is at a 2nd layer
302
The accelerator performs 2nd-layer computing of the 1st iteration on the first session, loads, from the memory of the host to the first buffer, a historical KV cache that is of the input of the first session and that is at a 3rd layer, and writes, into the memory of the host, the KV cache obtained through 1st-layer computing of the 1st iteration
303
The accelerator sequentially performs, in a manner similar to that in step 303, 3rd-layer computing to Nth-layer computing of the 1st iteration on the first session, to obtain a 1st token in the input of the first session through inference
304
The accelerator sequentially performs a 2nd iteration to an Mth iteration on the first session in a manner similar to that in step 301 to step 304, to sequentially obtain a 2nd token to an Mth token in the input of the first session through inference, where M is equal to a total quantity of iterations of the first session, and M is an integer greater than or equal to 2
305
When the accelerator performs Nth-layer computing of the Mth iteration on the first session, the accelerator loads, from the memory of the host to a second buffer of the accelerator, a historical KV cache that is of an input of a second session in a job queue and that is at the 1st layer
306
When the accelerator completes Nth-layer computing of the Mth iteration of the first session, if a KV cache that has not been written into the memory exists in KV caches generated by the generative large model processing the first session, the accelerator writes, from the first buffer to a third buffer of the accelerator, the KV cache that has not been written into the memory
307
The accelerator processes the second session in a manner similar to step 301 to step 307, and writes the KV cache in the third buffer of the accelerator into the memory of the host
308

FIG. 3

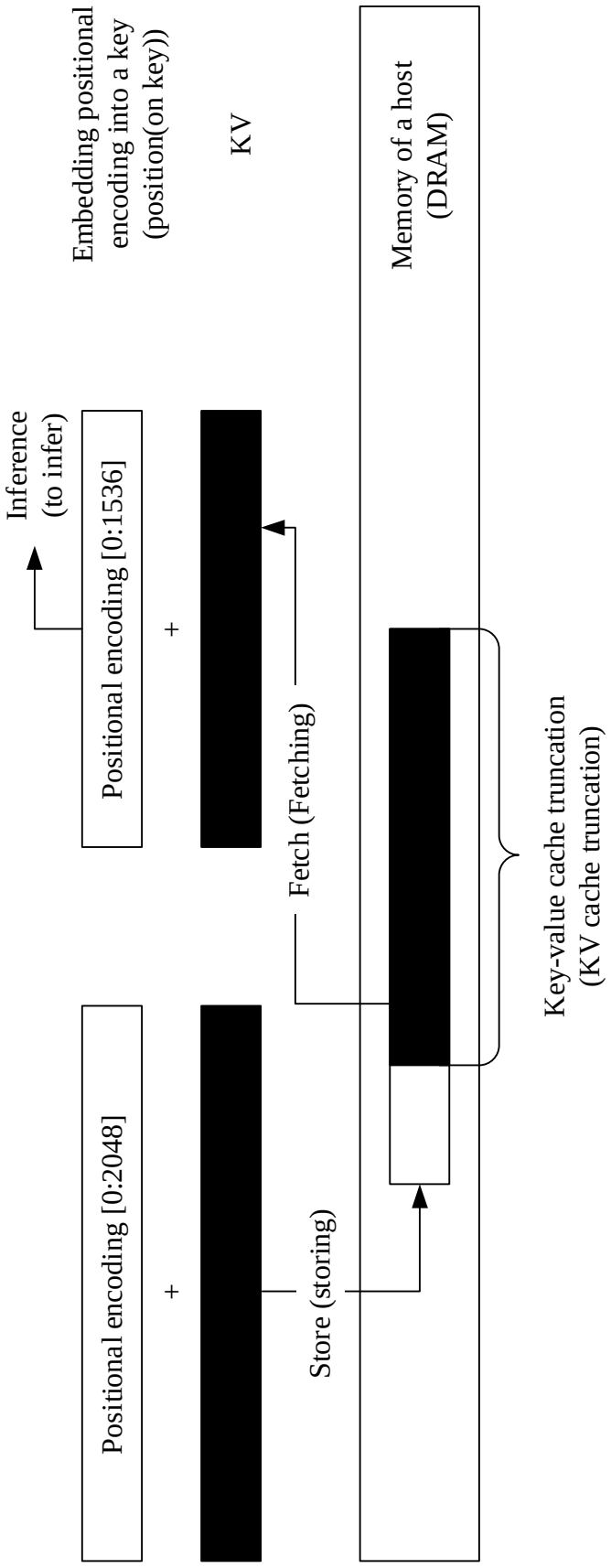

Inference
(to infer)
Positional encoding [0:1536]
+
Embedding positional
encoding into a key
(position(on key))
KV
Memory of a host
(DRAM)
Fetch (Fetching)
Positional encoding [0:2048]
+
Store (storing)
Key-value cache truncation
(KV cache truncation)

FIG. 4

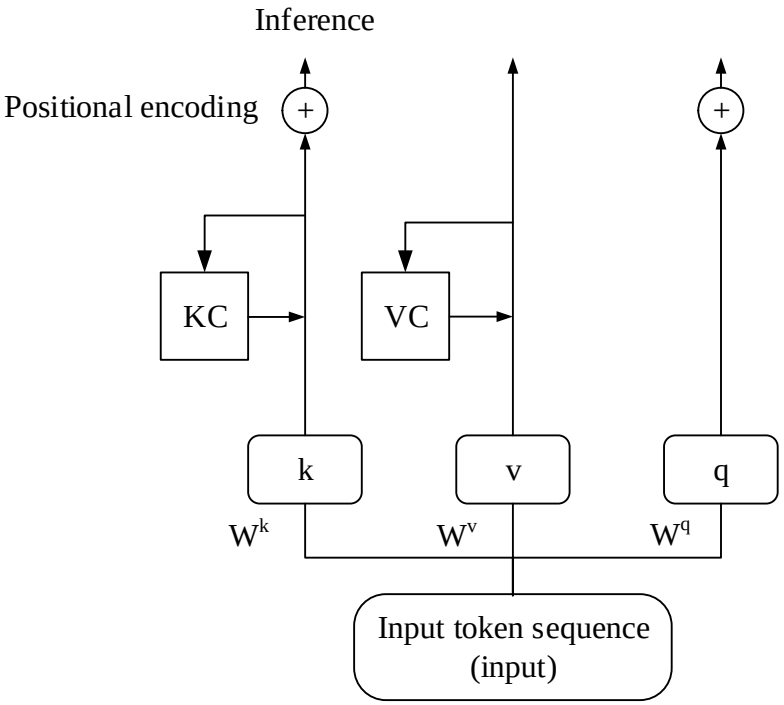


FIG. 5

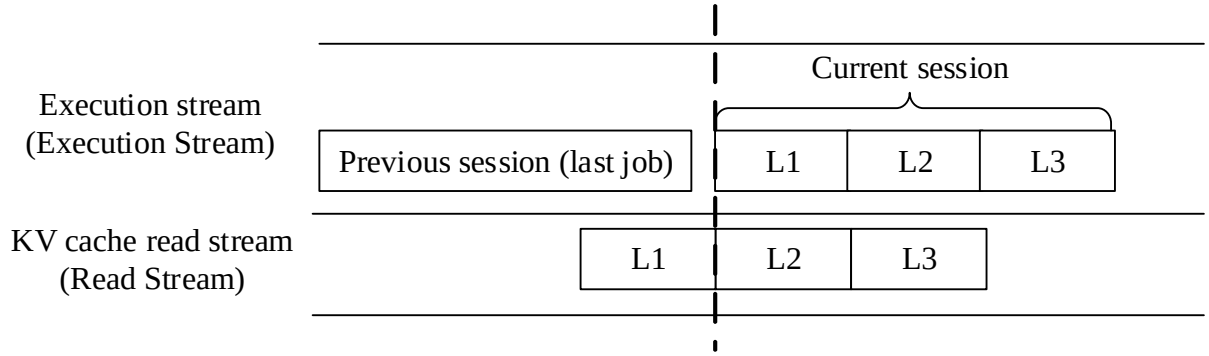


FIG. 6

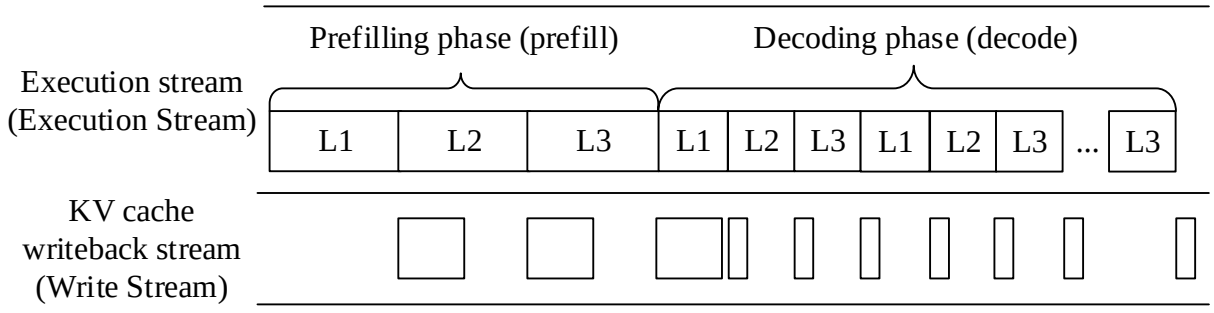


FIG. 7

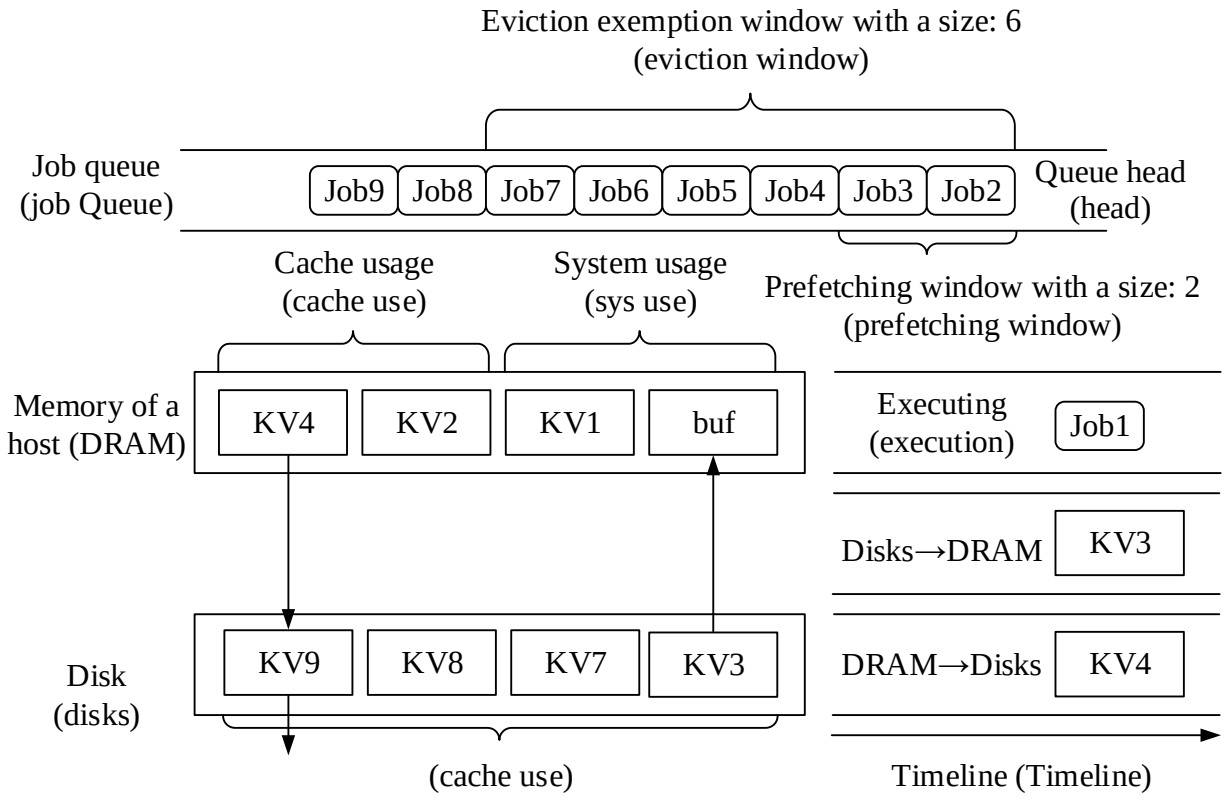

Eviction exemption window with a size: 6
(eviction window)
Job queue
(job Queue)
Job9
Job8
Job7
Job6
Job5
Job4
Job3
Job2
Queue head
(head)
Cache usage
(cache use)
System usage
(sys use)
Prefetching window with a size: 2
(prefetching window)
Memory of a
host (DRAM)
KV4
KV2
KV1
buf
Executing
(execution)
Job1
Disks→DRAM
KV3
DRAM→Disks
KV4
Disk
(disks)
KV9
KV8
KV7
KV3
(cache use)
Timeline (Timeline)

FIG. 8

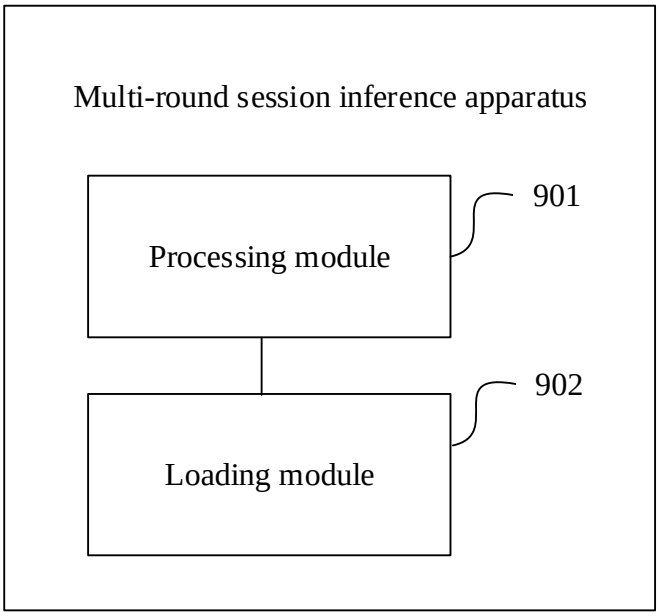

Multi-round session inference apparatus
Processing module
901
Loading module
902

FIG. 9

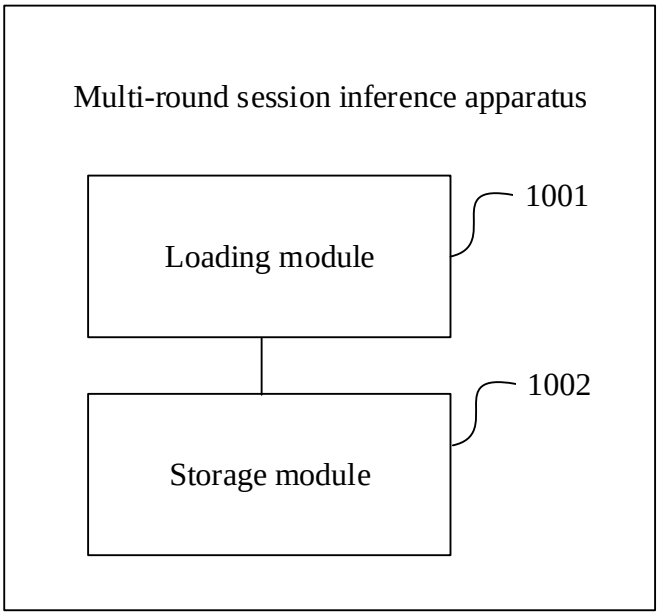

Multi-round session inference apparatus
Loading module
1001
Storage module
1002

FIG. 10

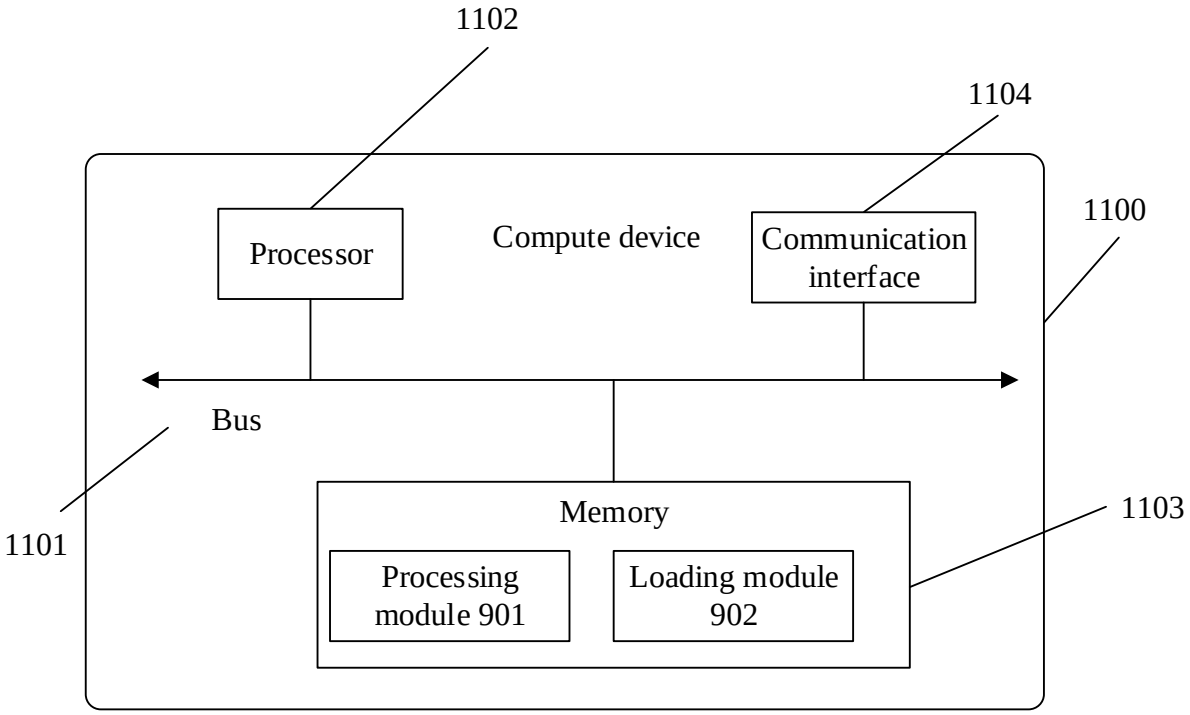

1102
1104
Processor
Compute device
Communication interface
1100
Bus
1101
Memory
1103
Processing module 901
Loading module 902

FIG. 11

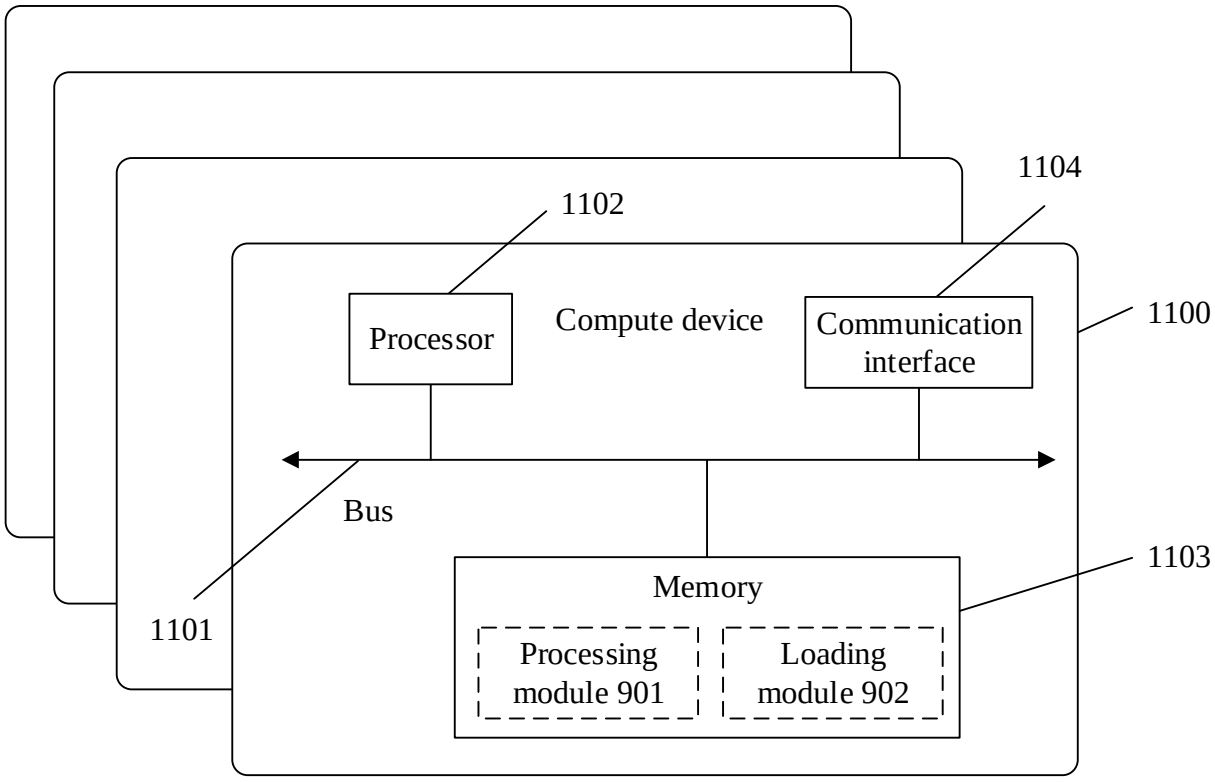

1104
1102
Processor
Compute device
Communication interface
1100
Bus
1101
Memory
1103
Processing module 901
Loading module 902

FIG. 12

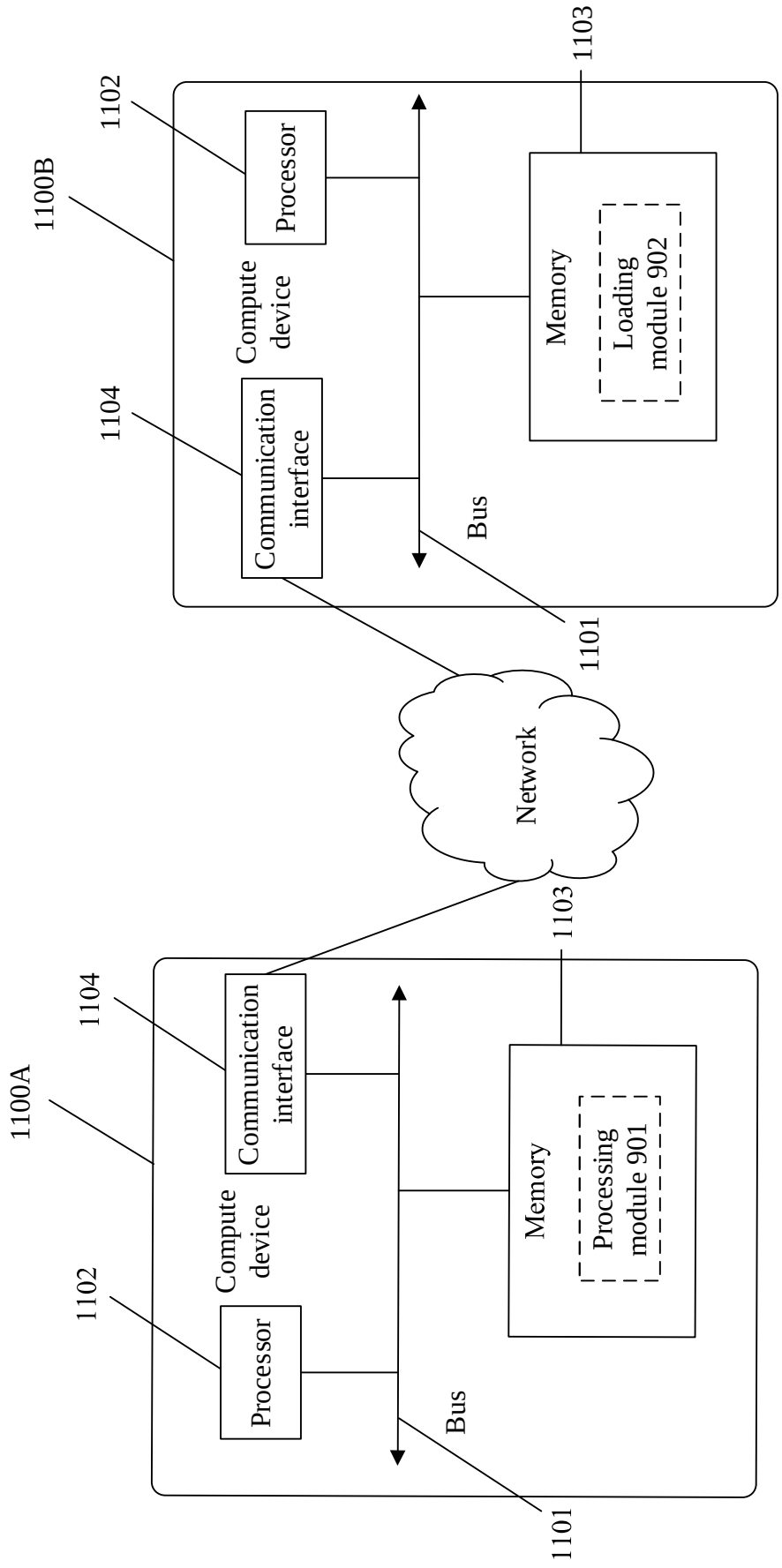

1100A
1104
Communication interface
Compute device
1102
Processor
Bus
1101
Memory
Processing module 901
1103
Network
1103
1100B
1102
Processor
Compute device
1104
Communication interface
Bus
1101
Memory
Loading module 902
1103

FIG. 13