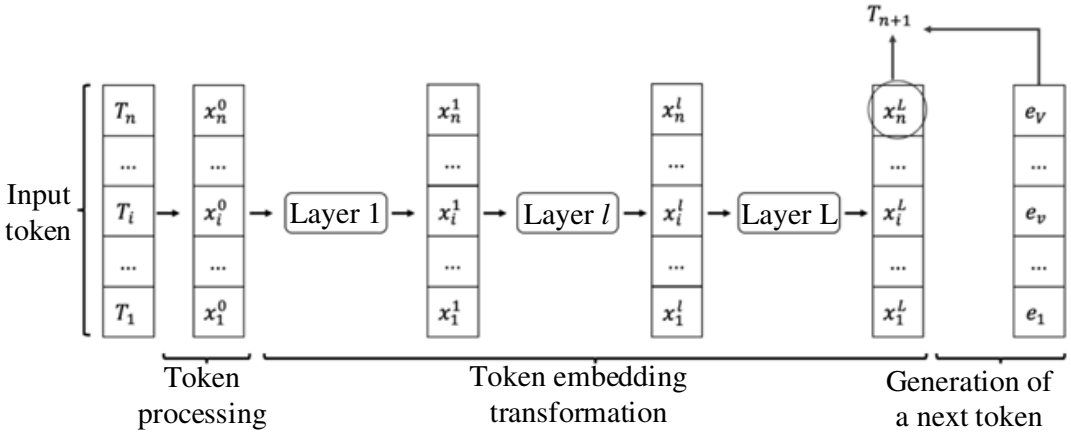
Input token
Layer 1
Layer l
Layer L
Token processing
Token embedding transformation
Generation of a next token

FIG. 1

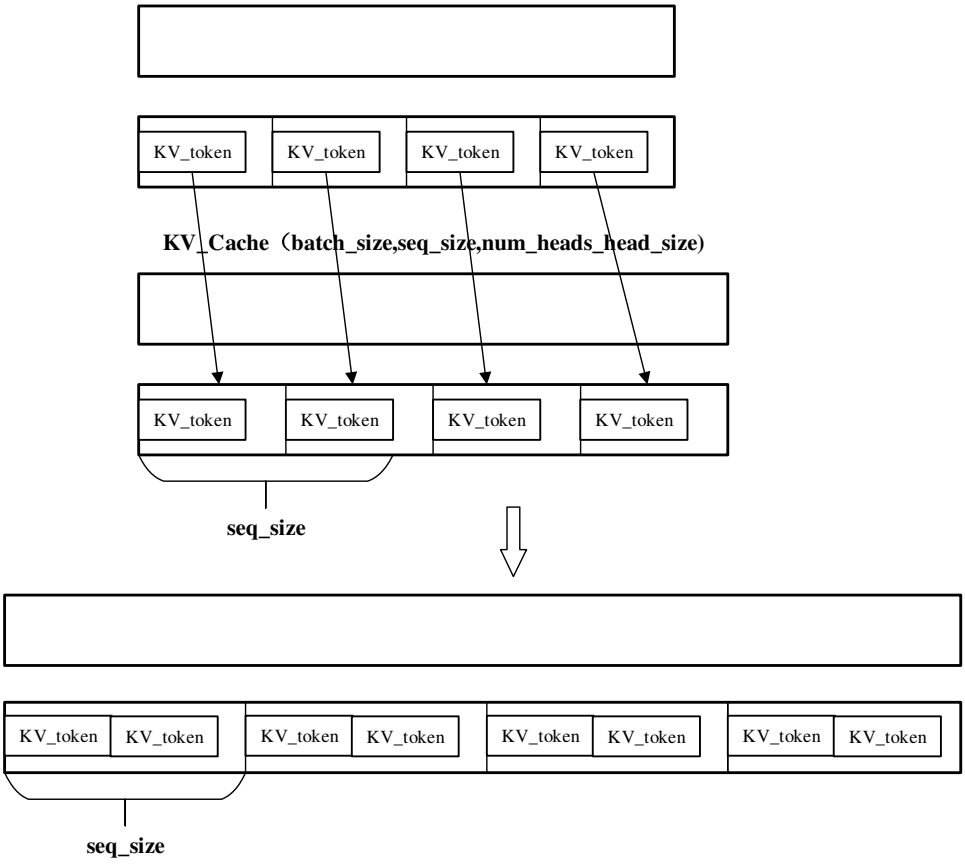
KV_token
KV_Cache（batch_size,seq_size,num_heads_head_size)
seq_size
seq_size

FIG. 2

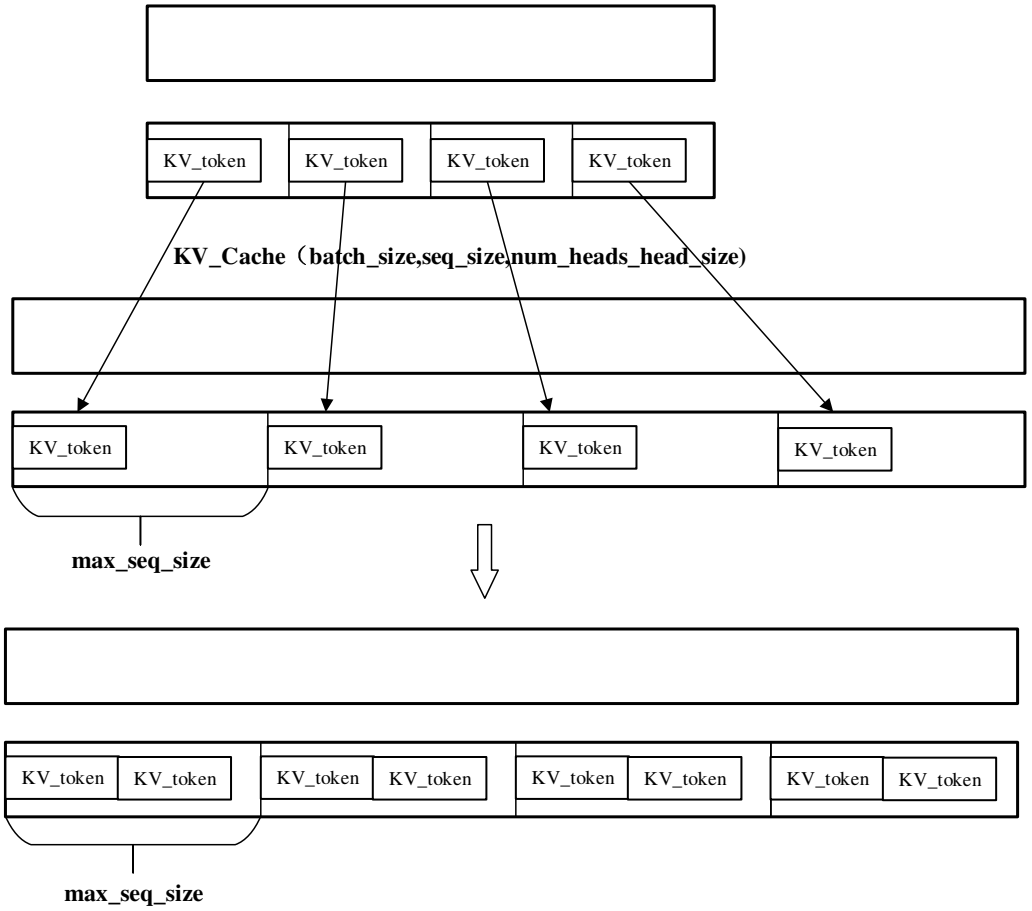
KV_token
KV_token
KV_token
KV_token
KV_Cache（batch_size,seq_size,num_heads_head_size)
KV_token
KV_token
KV_token
KV_token
max_seq_size
KV_token
KV_token
KV_token
KV_token
KV_token
KV_token
KV_token
KV_token
max_seq_size

FIG. 3

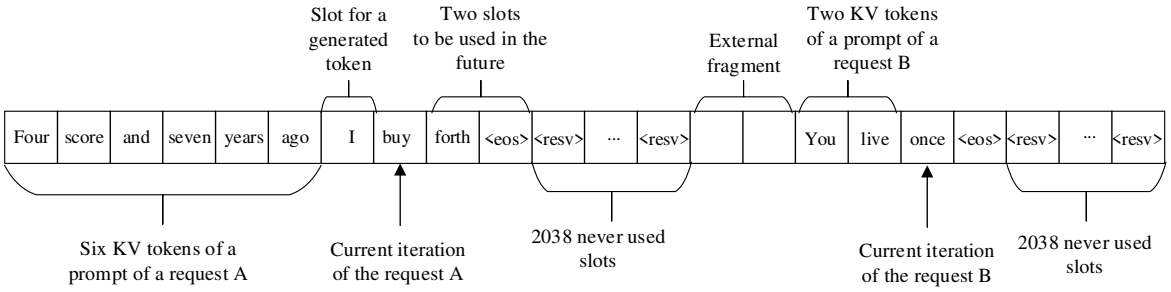
Slot for a generated token
Two slots to be used in the future
External fragment
Two KV tokens of a prompt of a request B
Four
score
and
seven
years
ago
I
buy
forth
<eos>
<resv>
···
<resv>
You
live
once
<eos>
<resv>
···
<resv>
Six KV tokens of a prompt of a request A
Current iteration of the request A
2038 never used slots
Current iteration of the request B
2038 never used slots

FIG. 4

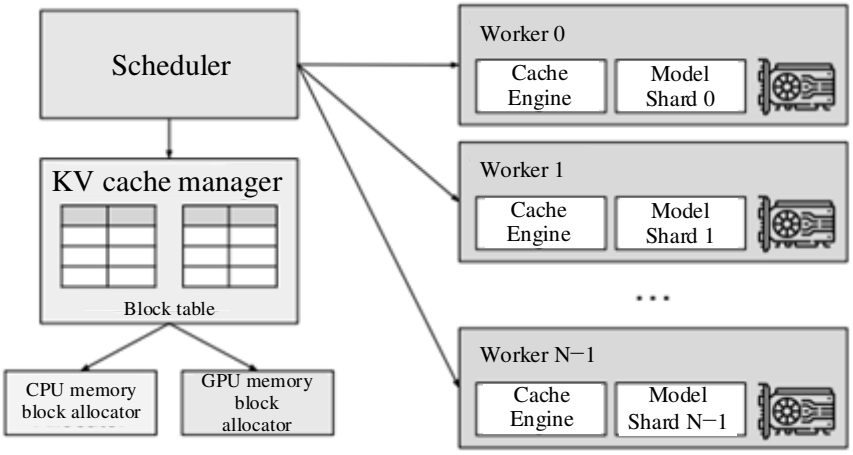

FIG. 5

Physical block number	#Filled slot

FIG. 6

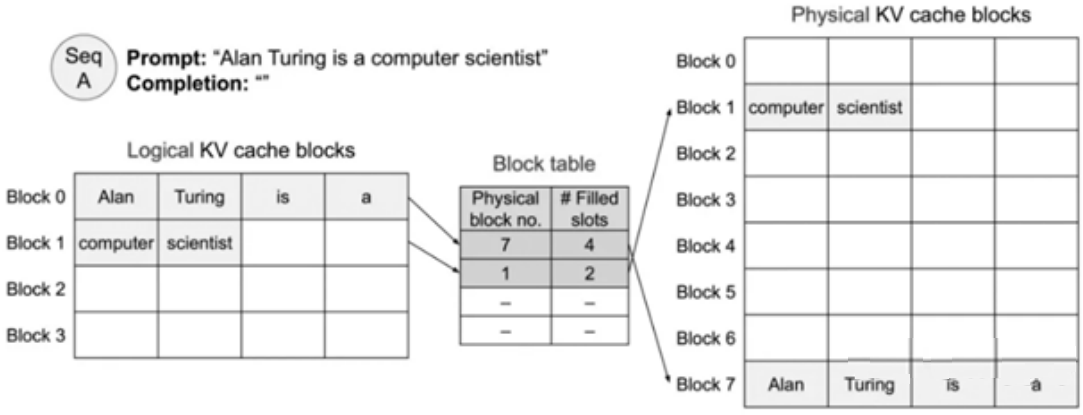

FIG. 7A

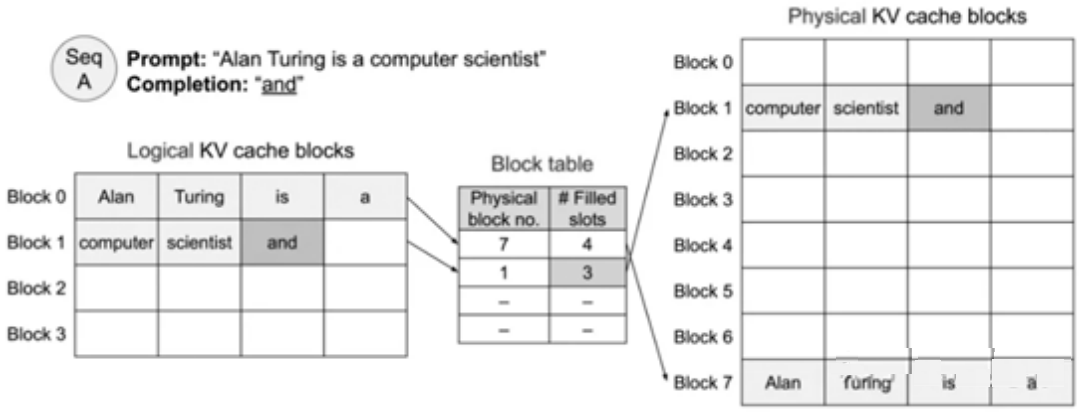
Seq A
Prompt: "Alan Turing is a computer scientist"
Completion: "and"
Logical KV cache blocks
Block 0
Alan
Turing
is
a
Block 1
computer
scientist
and
Block 2
Block 3
Block table
Physical block no.
Filled slots
7
4
1
3
–
–
–
–
Physical KV cache blocks
Block 0
Block 1
computer
scientist
and
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Alan
Turing
is
a

FIG. 7B

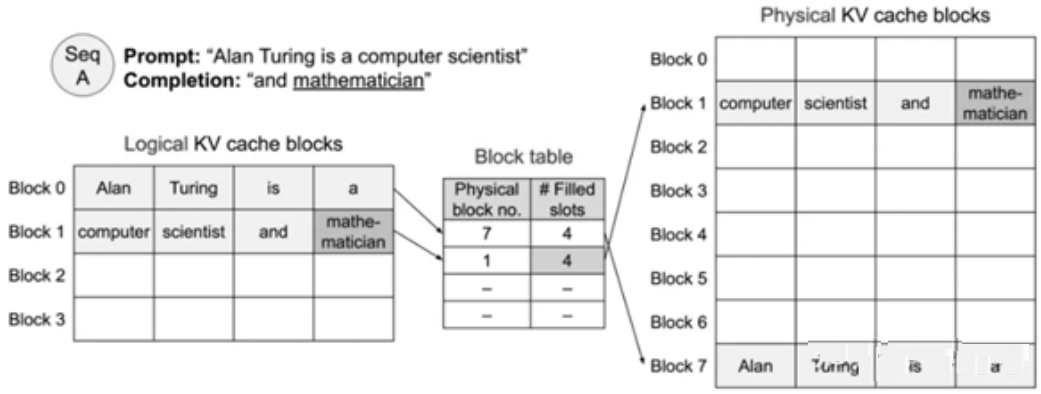
Seq A
Prompt: "Alan Turing is a computer scientist"
Completion: "and mathematician"
Logical KV cache blocks
Block 0
Alan
Turing
is
a
Block 1
computer
scientist
and
mathe-matician
Block 2
Block 3
Block table
Physical block no.
Filled slots
7
4
1
4
–
–
–
–
Physical KV cache blocks
Block 0
Block 1
computer
scientist
and
mathe-matician
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Alan
Turing
is
a

FIG. 7C

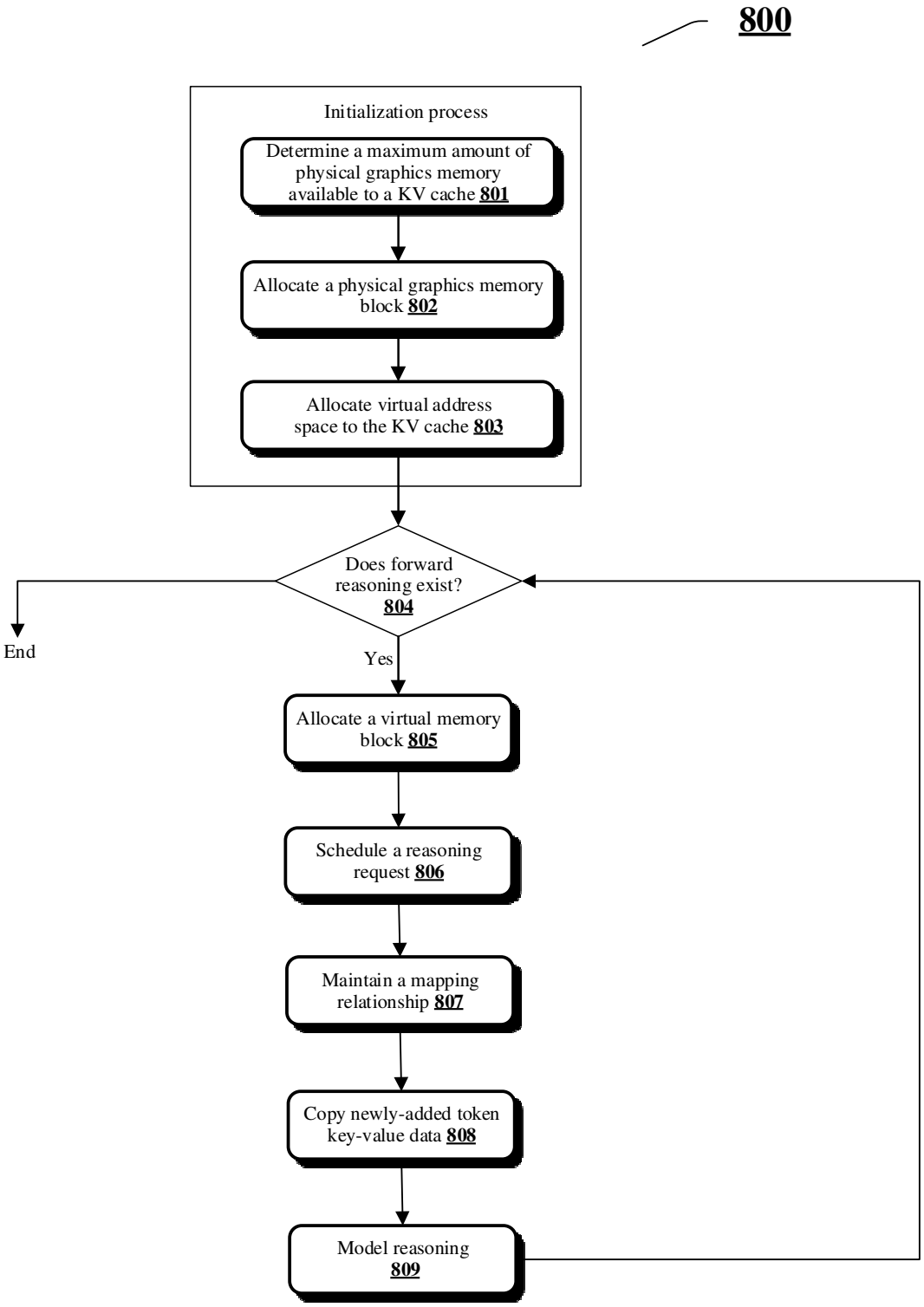
800
Initialization process
Determine a maximum amount of physical graphics memory available to a KV cache 801
Allocate a physical graphics memory block 802
Allocate virtual address space to the KV cache 803
Does forward reasoning exist? 804
End
Yes
Allocate a virtual memory block 805
Schedule a reasoning request 806
Maintain a mapping relationship 807
Copy newly-added token key-value data 808
Model reasoning 809

FIG. 8

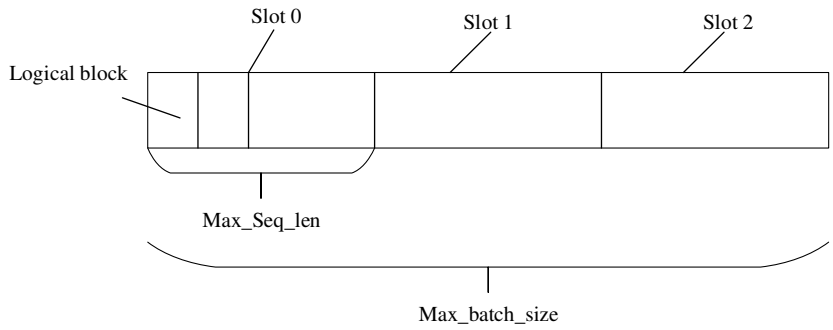
Slot 0
Slot 1
Slot 2
Logical block
Max_Seq_len
Max_batch_size

FIG. 9

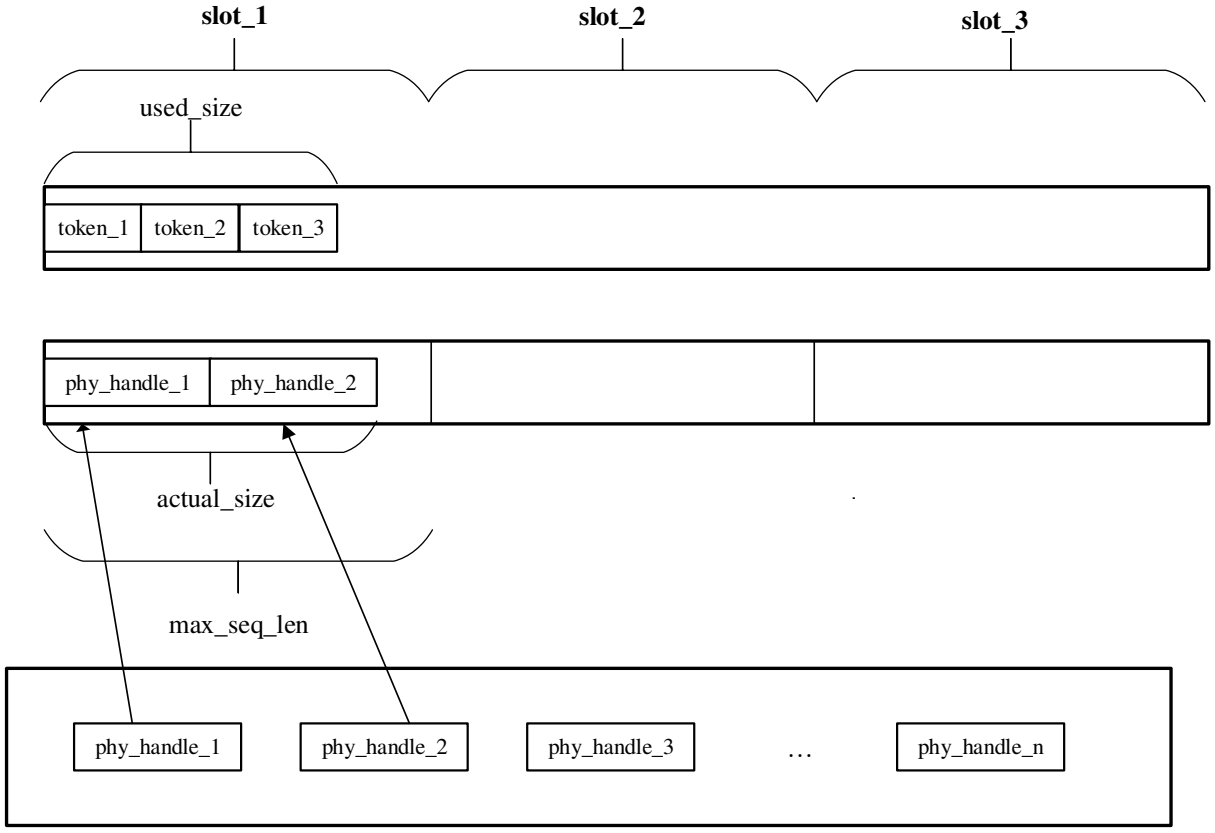
slot_1
slot_2
slot_3
used_size
token_1
token_2
token_3
phy_handle_1
phy_handle_2
actual_size
max_seq_len
phy_handle_1
phy_handle_2
phy_handle_3
…
phy_handle_n
Physical graphics memory pool

FIG. 10

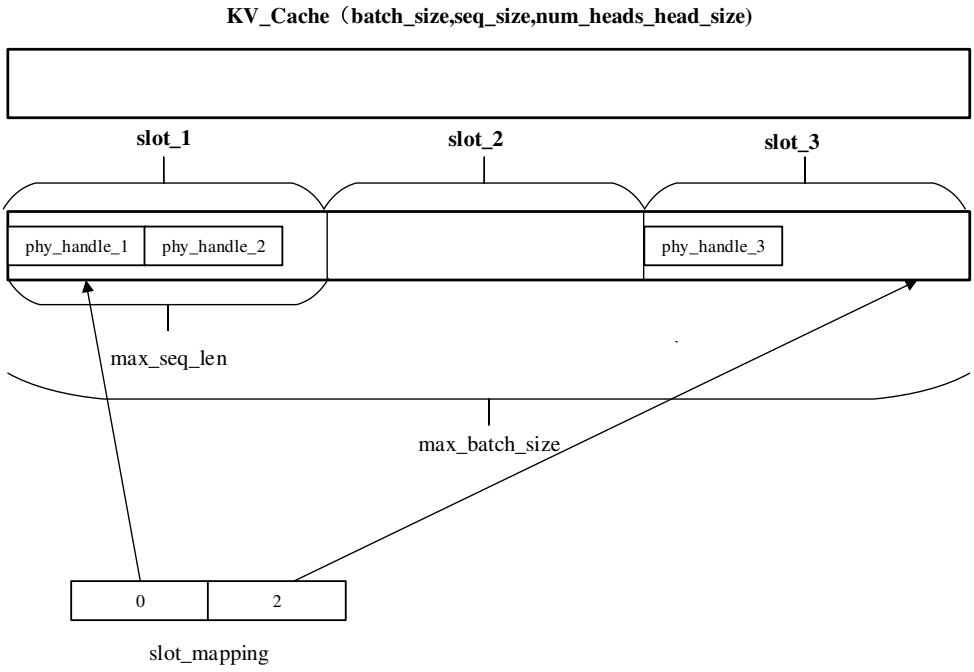
KV_Cache（batch_size,seq_size,num_heads_head_size)
slot_1
slot_2
slot_3
phy_handle_1
phy_handle_2
phy_handle_3
max_seq_len
max_batch_size
0
2
slot_mapping

FIG. 11

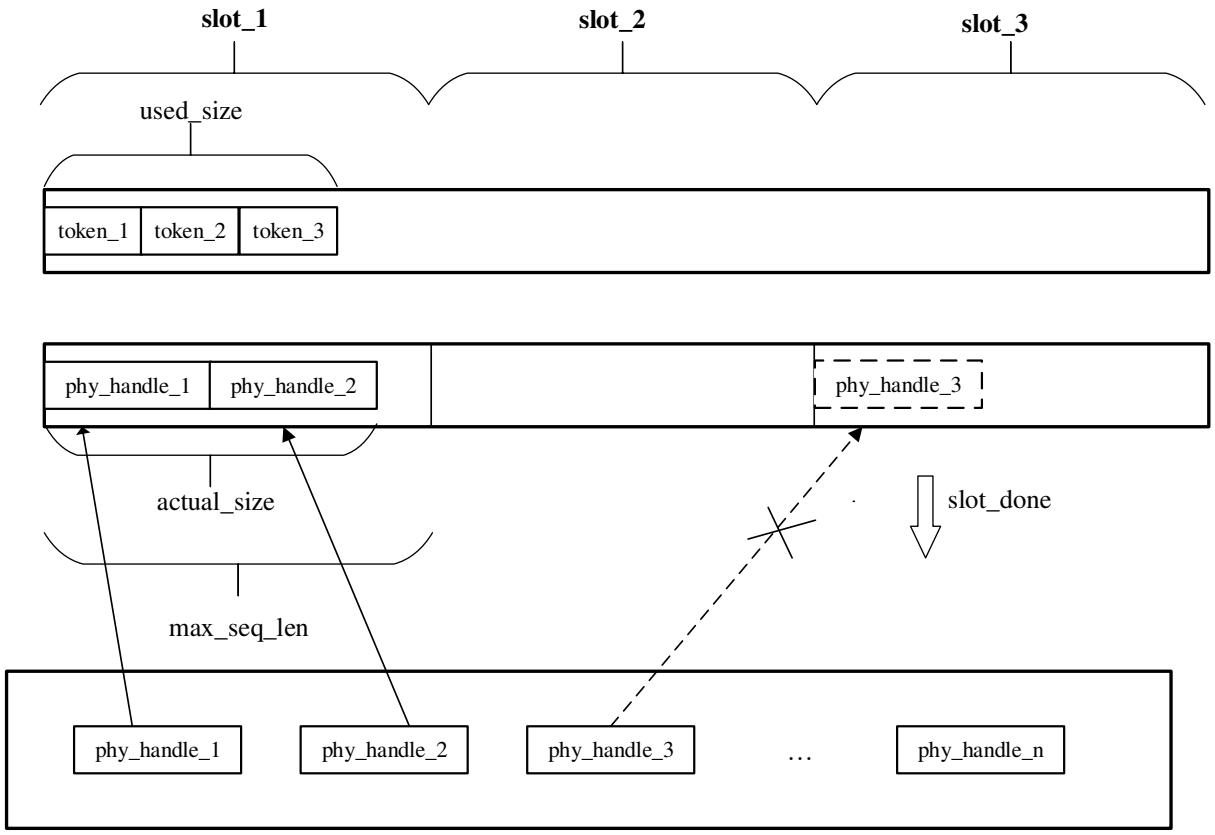
slot_1
slot_2
slot_3
used_size
token_1
token_2
token_3
phy_handle_1
phy_handle_2
phy_handle_3
actual_size
slot_done
max_seq_len
phy_handle_1
phy_handle_2
phy_handle_3
…
phy_handle_n
Physical graphics memory pool

FIG. 12

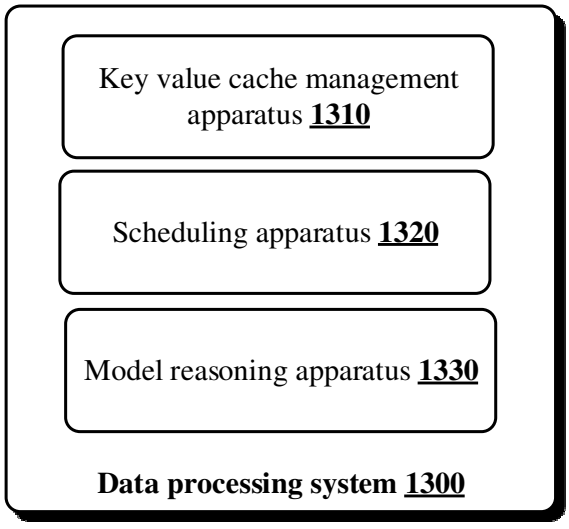
Key value cache management apparatus 1310
Scheduling apparatus 1320
Model reasoning apparatus 1330
Data processing system 1300

FIG. 13

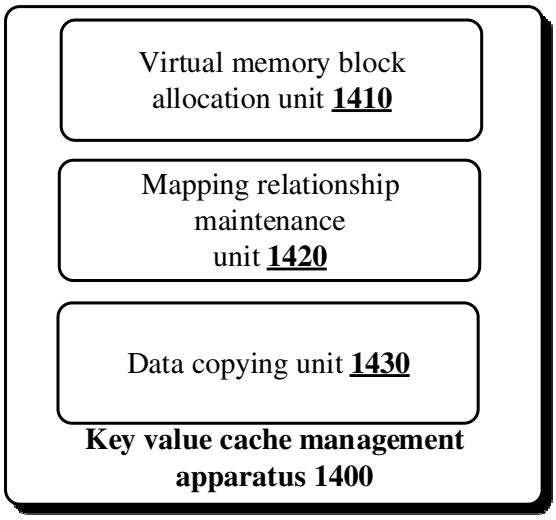
Virtual memory block allocation unit 1410
Mapping relationship maintenance unit 1420
Data copying unit 1430
Key value cache management apparatus 1400

FIG. 14

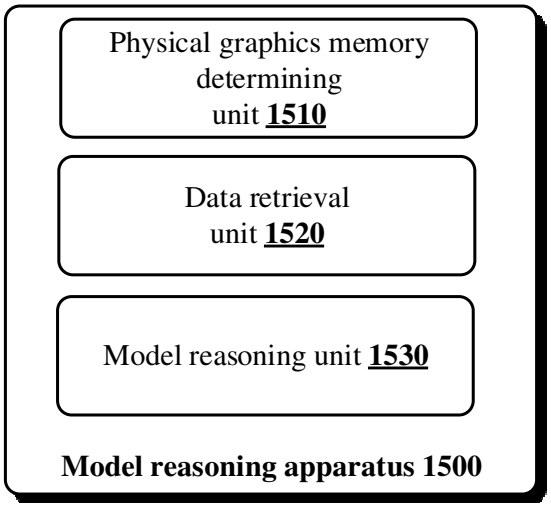
Physical graphics memory determining unit 1510
Data retrieval unit 1520
Model reasoning unit 1530
Model reasoning apparatus 1500

FIG. 15

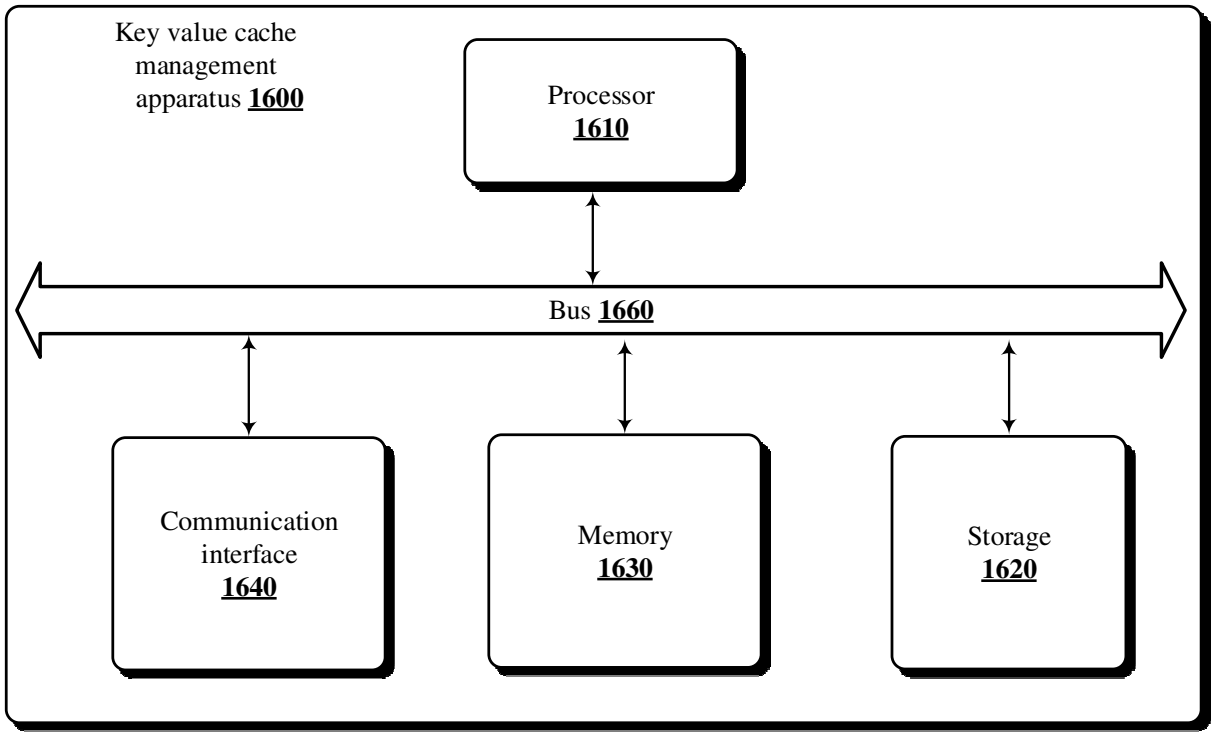
Key value cache management apparatus **1600**
Processor **1610**
Bus **1660**
Communication interface **1640**
Memory **1630**
Storage **1620**

FIG. 16

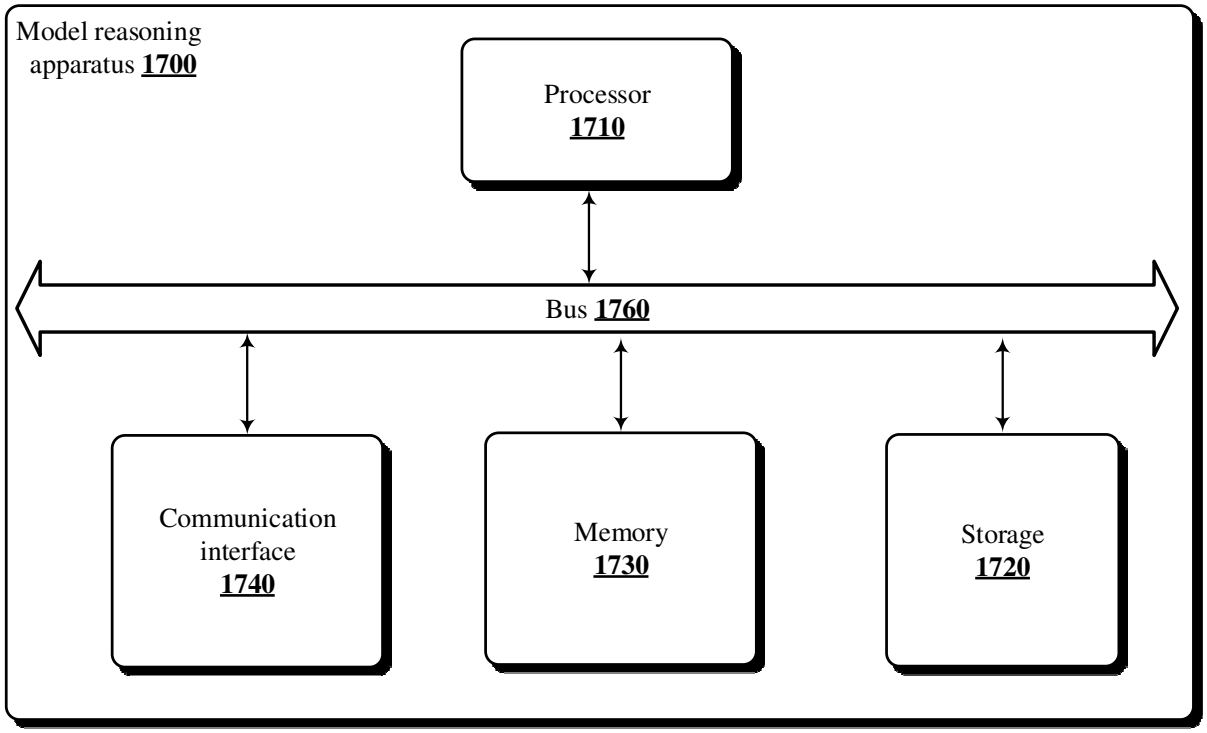
Model reasoning apparatus **1700**
Processor **1710**
Bus **1760**
Communication interface **1740**
Memory **1730**
Storage **1720**

FIG. 17

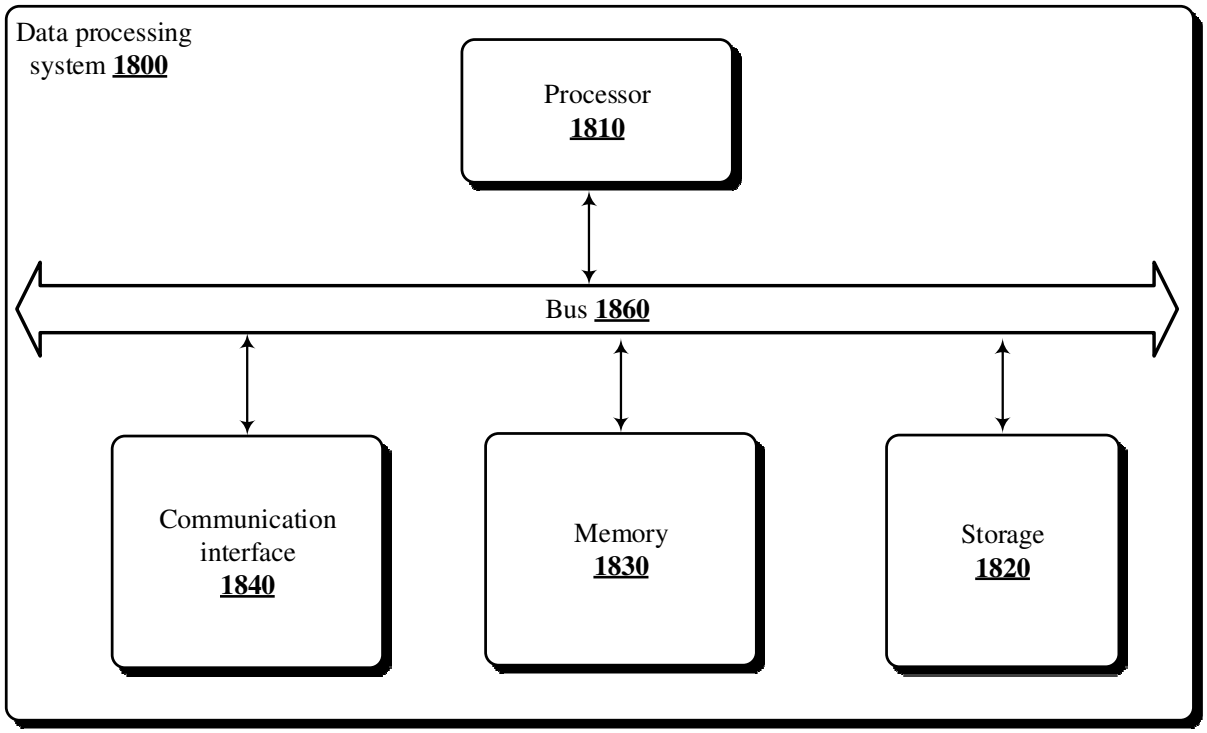
Data processing system **1800**
Processor **1810**
Bus **1860**
Communication interface **1840**
Memory **1830**
Storage **1820**

FIG. 18