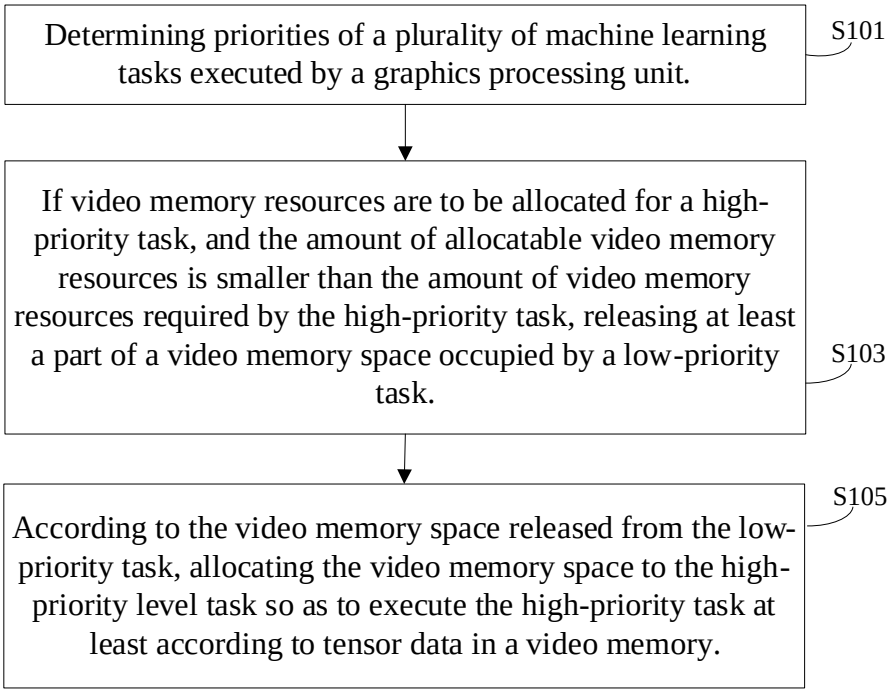

Determining priorities of a plurality of machine learning tasks executed by a graphics processing unit.
S101
If video memory resources are to be allocated for a high-priority task, and the amount of allocatable video memory resources is smaller than the amount of video memory resources required by the high-priority task, releasing at least a part of a video memory space occupied by a low-priority task.
S103
According to the video memory space released from the low-priority task, allocating the video memory space to the high-priority level task so as to execute the high-priority task at least according to tensor data in a video memory.
S105

FIG. 1

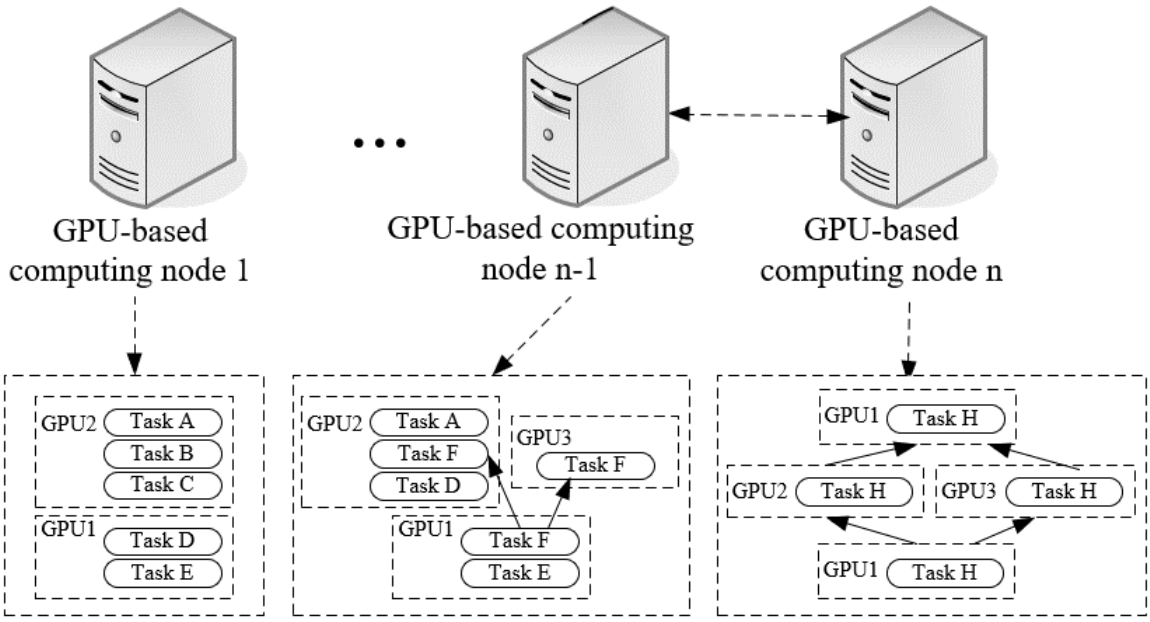

GPU-based computing node 1
GPU-based computing node n-1
GPU-based computing node n
GPU2
Task A
Task B
Task C
GPU1
Task D
Task E
GPU2
Task A
Task F
Task D
GPU3
Task F
GPU1
Task F
Task E
GPU1
Task H
GPU2
Task H
GPU3
Task H
GPU1
Task H

FIG. 2

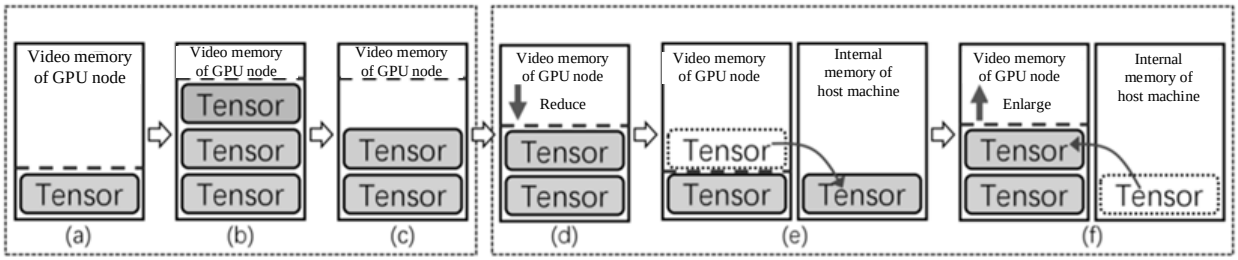

Video memory of GPU node
Tensor
(a)
Video memory of GPU node
Tensor
Tensor
Tensor
(b)
Video memory of GPU node
Tensor
Tensor
(c)
Video memory of GPU node
Reduce
Tensor
Tensor
(d)
Video memory of GPU node
Tensor
Tensor
Internal memory of host machine
Tensor
(e)
Video memory of GPU node
Enlarge
Tensor
Tensor
Internal memory of host machine
Tensor
(f)

FIG. 3

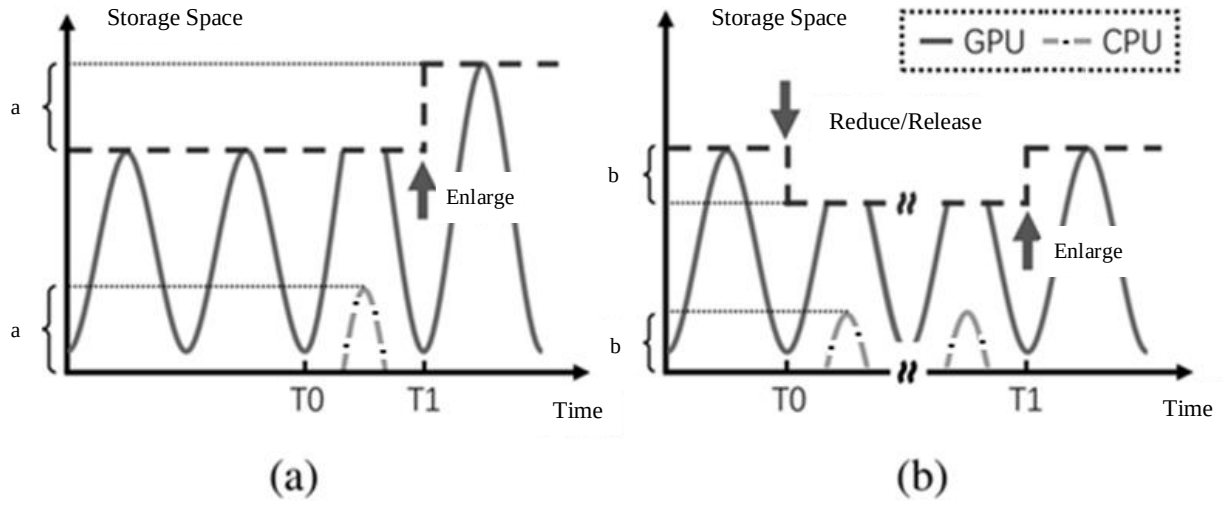

Storage Space
a
a
Enlarge
T0
T1
Time
(a)
Storage Space
GPU
CPU
Reduce/Release
b
b
Enlarge
T0
T1
Time
(b)

FIG. 4

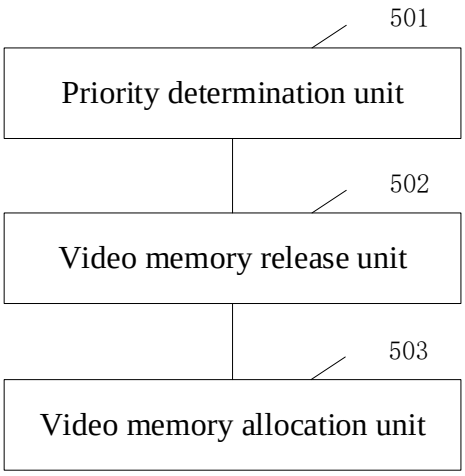

501
Priority determination unit
502
Video memory release unit
503
Video memory allocation unit

FIG. 5

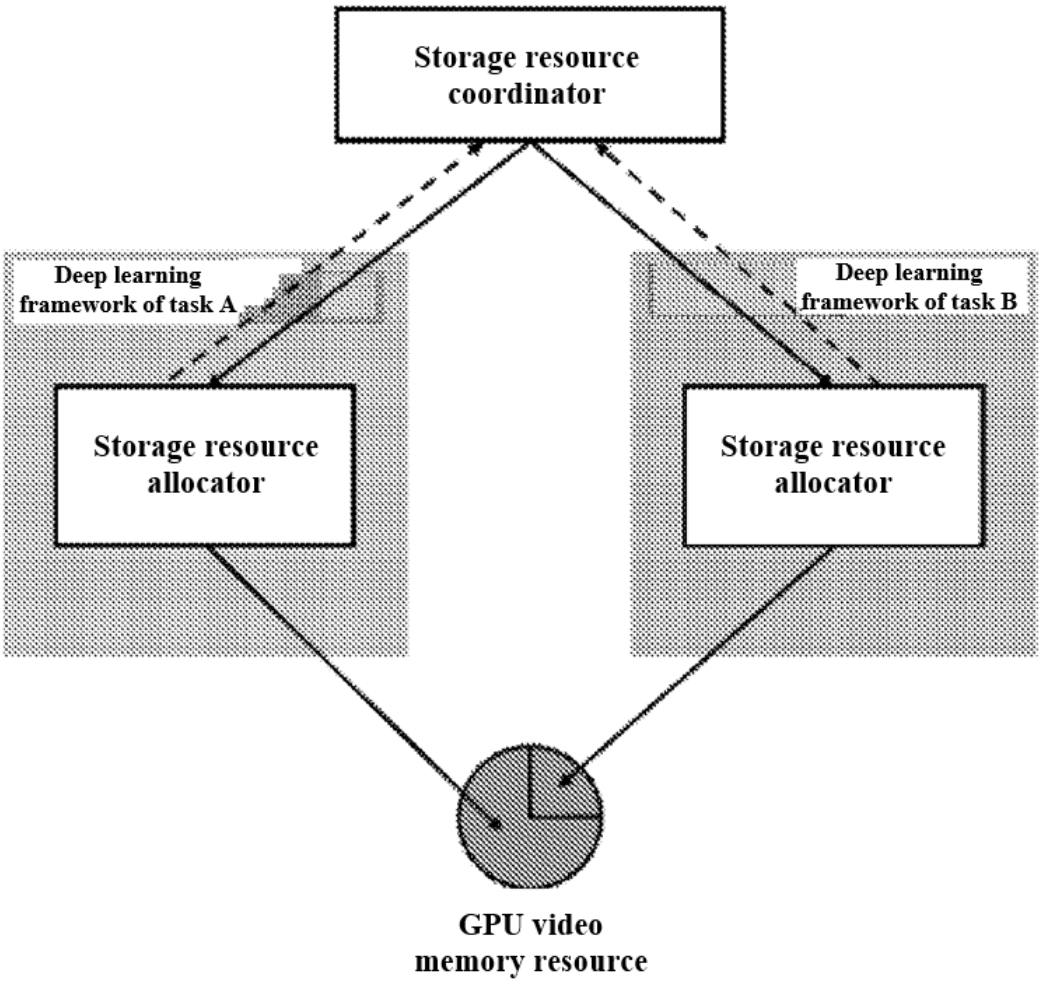

Storage resource coordinator
Deep learning framework of task A
Storage resource allocator
Deep learning framework of task B
Storage resource allocator
GPU video memory resource

FIG. 6