

GPU COMPUTATIONAL RESOURCE SCHEDULING METHODS AND APPARATUSES

ABSTRACT

5 A GPU computational resource scheduling method includes: in response to a target
computing task created in a computing cluster, determining a task type of the target computing
task; if the target computing task is the first-type computing task, scheduling, for running, the
target computing task to a first GPU hardware that has remaining computational resources
satisfying a computational demand of the target computing task in the computing cluster; and in
response to a first indication indicating that is reported by a first computing node integrated with
10 the first GPU hardware and that indicates that the first-type computing task exclusively occupies
computational resources of the first GPU hardware, rescheduling, for running to a second GPU
hardware that has remaining computational resources satisfying a computational demand of the
second-type computing task in the computing cluster. A service level of the first-type computing
task is higher than that of the second-type computing task.

15

FIG. 1