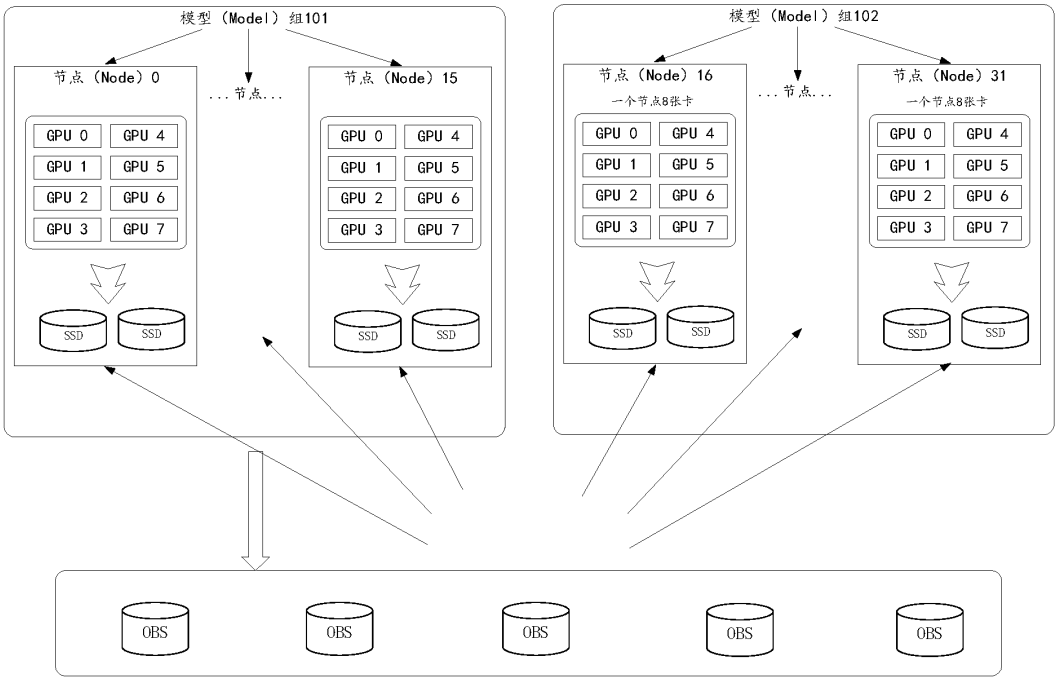

模型（Model）组101
节点（Node）0
...节点...
节点（Node）15
GPU 0
GPU 4
GPU 1
GPU 5
GPU 2
GPU 6
GPU 3
GPU 7
SSD
模型（Model）组102
节点（Node）16
一个节点8张卡
节点（Node）31
OBS

图 1

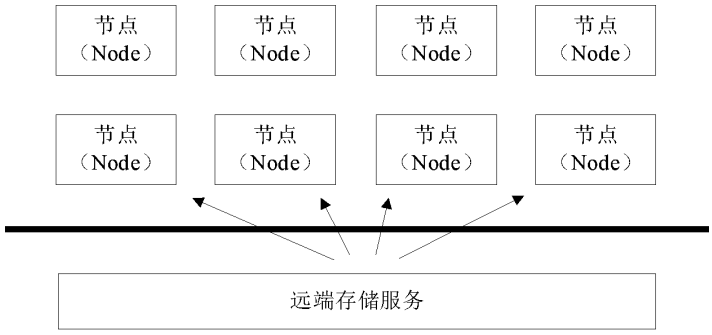

节点
（Node）
远端存储服务

图 2

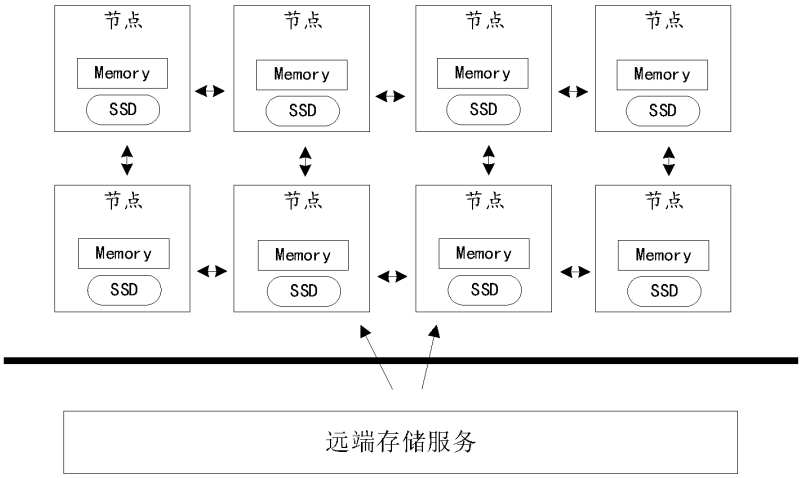


图 3

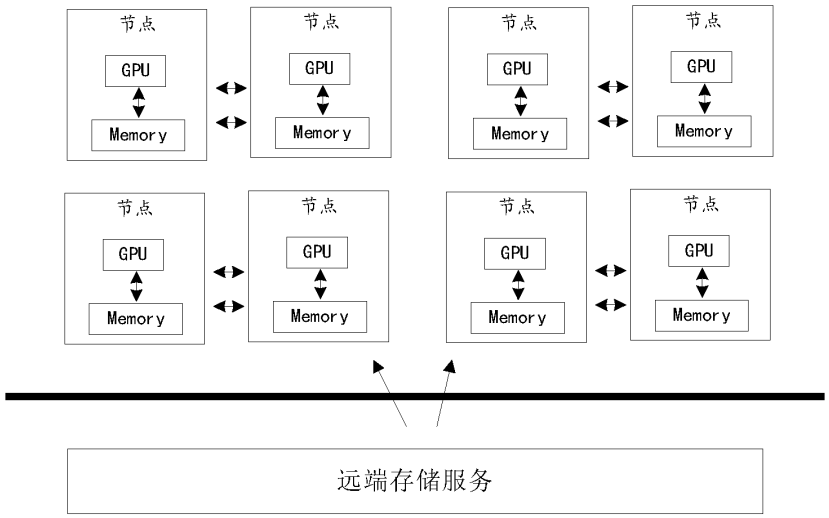


图 4

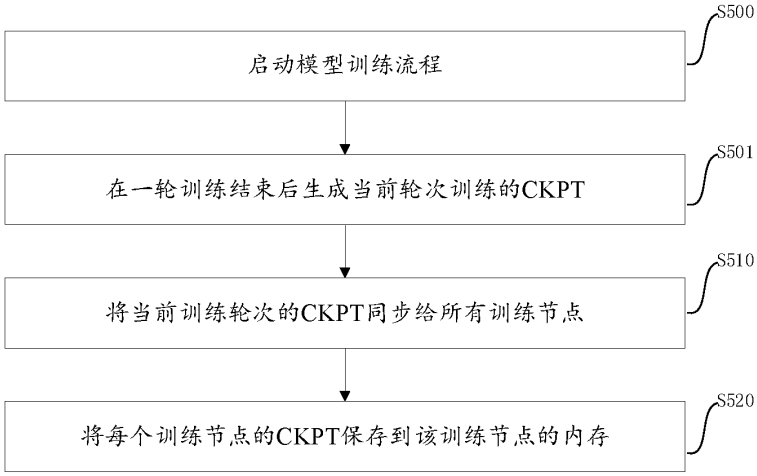

S500
启动模型训练流程
S501
在一轮训练结束后生成当前轮次训练的CKPT
S510
将当前训练轮次的CKPT同步给所有训练节点
S520
将每个训练节点的CKPT保存到该训练节点的内存

图 5

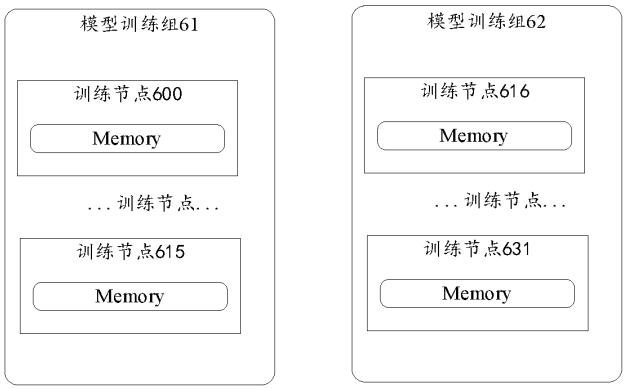

模型训练组61
训练节点600
Memory
. . . 训练节点. . .
训练节点615
Memory
模型训练组62
训练节点616
Memory
. . . 训练节点. . .
训练节点631
Memory

图 6

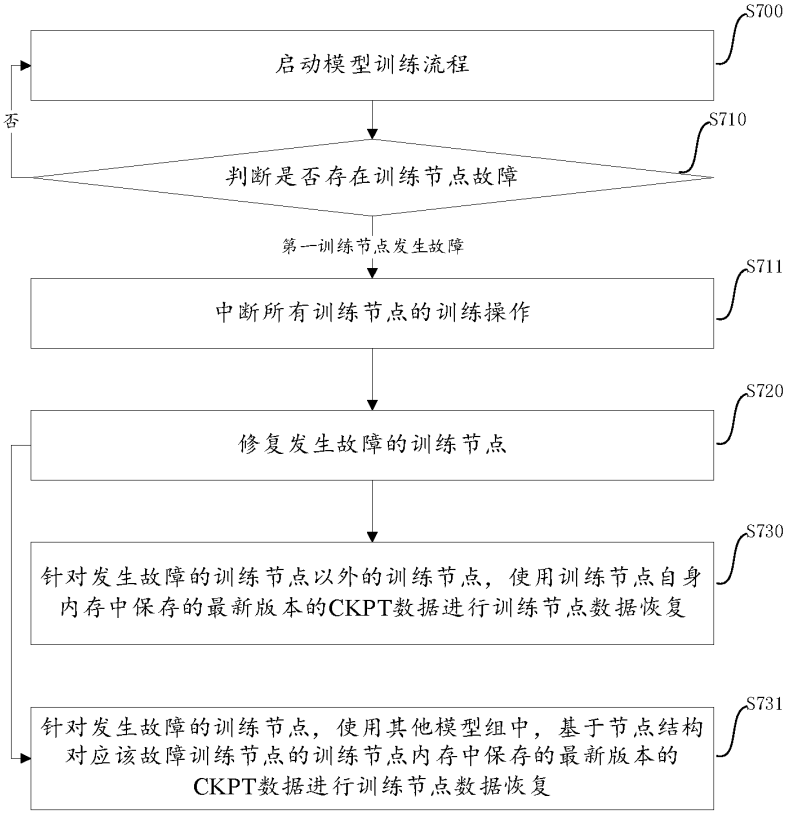

S700
启动模型训练流程
否
S710
判断是否存在训练节点故障
第一训练节点发生故障
S711
中断所有训练节点的训练操作
S720
修复发生故障的训练节点
S730
针对发生故障的训练节点以外的训练节点，使用训练节点自身内存中保存的最新版本的CKPT数据进行训练节点数据恢复
S731
针对发生故障的训练节点，使用其他模型组中，基于节点结构对应该故障训练节点的训练节点内存中保存的最新版本的CKPT数据进行训练节点数据恢复

图 7

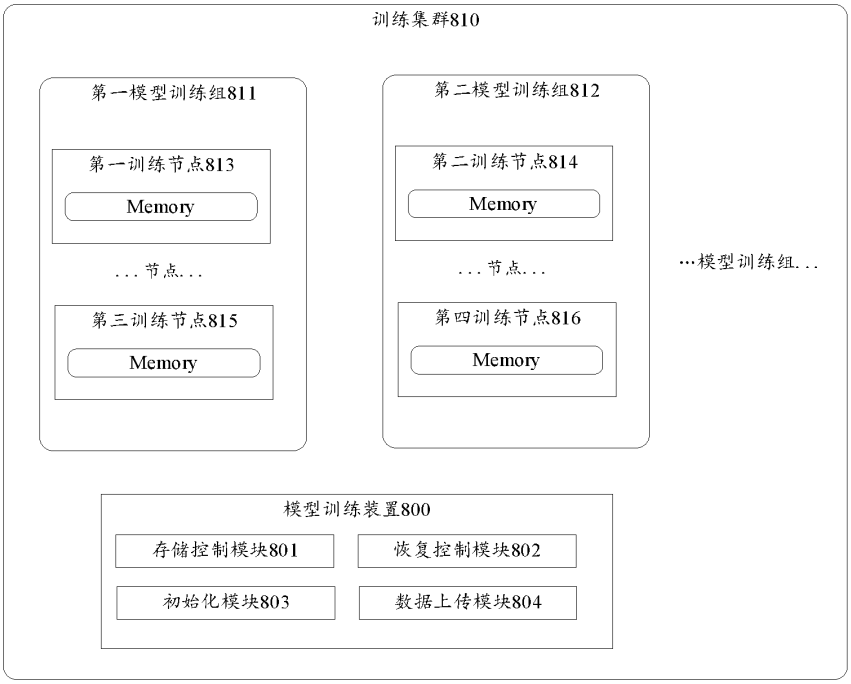

训练集群810
第一模型训练组811
第一训练节点813
Memory
...节点...
第三训练节点815
Memory
第二模型训练组812
第二训练节点814
Memory
...节点...
第四训练节点816
Memory
…模型训练组...
模型训练装置800
存储控制模块801
恢复控制模块802
初始化模块803
数据上传模块804

图 8


训练集群
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练节点
GPU
内存
SDK
训练框架服务
远端存储服务

图 9

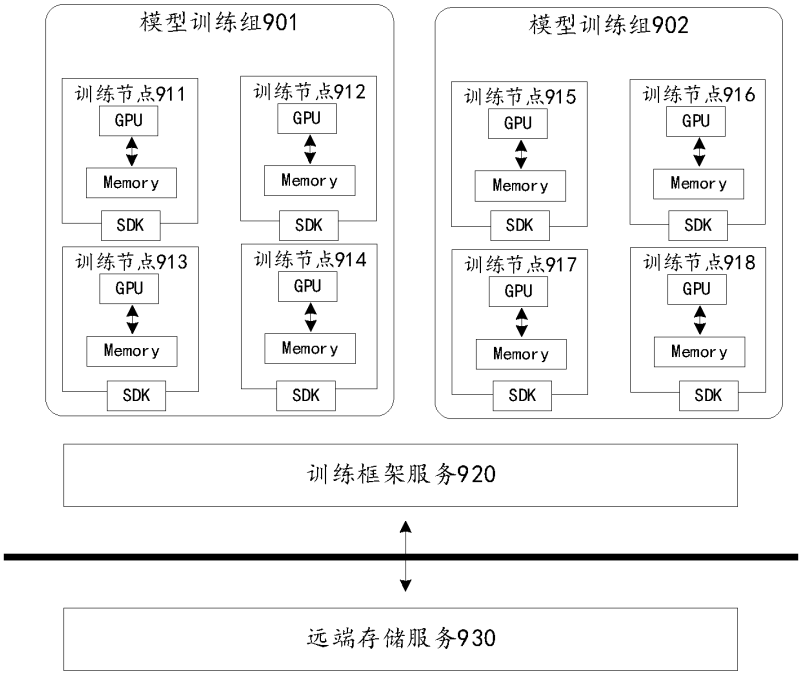

模型训练组901
训练节点911
GPU
Memory
SDK
训练节点912
GPU
Memory
SDK
训练节点913
GPU
Memory
SDK
训练节点914
GPU
Memory
SDK
模型训练组902
训练节点915
GPU
Memory
SDK
训练节点916
GPU
Memory
SDK
训练节点917
GPU
Memory
SDK
训练节点918
GPU
Memory
SDK
训练框架服务920
远端存储服务930

图 10

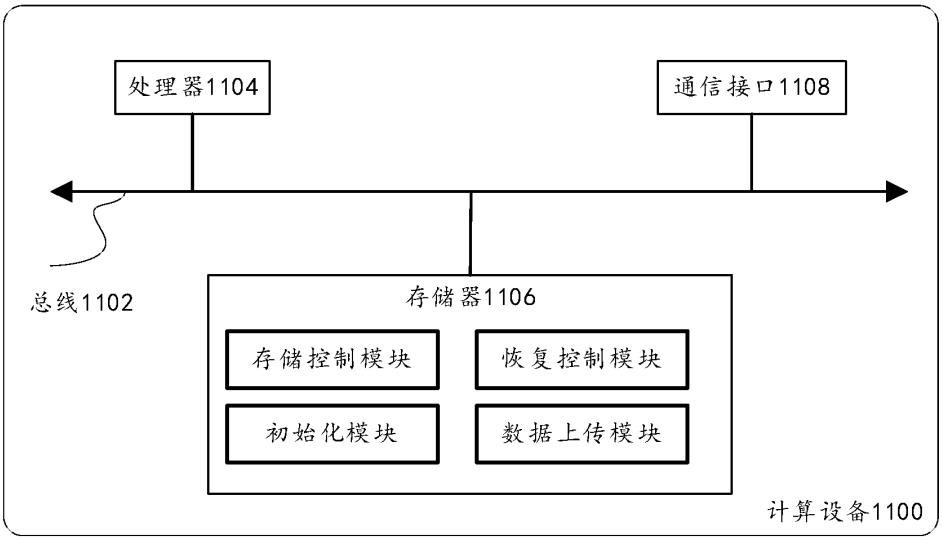

处理器1104
通信接口1108
总线1102
存储器1106
存储控制模块
恢复控制模块
初始化模块
数据上传模块
计算设备1100

图 11

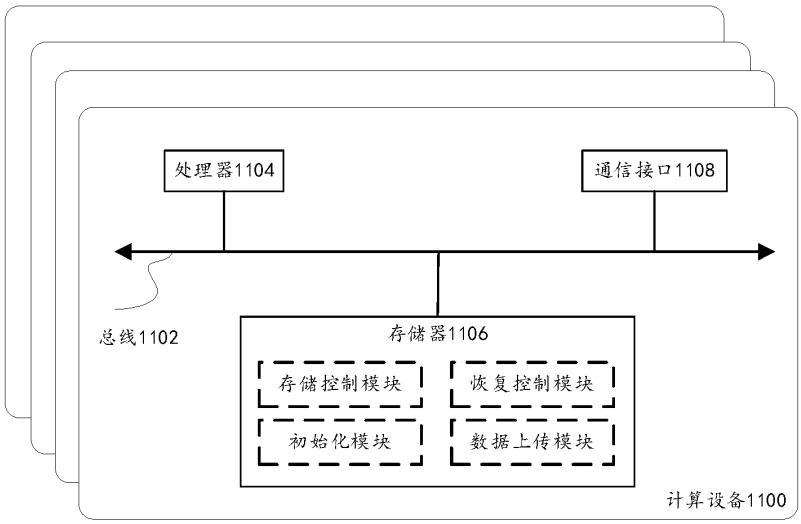

处理器1104
通信接口1108
总线1102
存储器1106
存储控制模块
恢复控制模块
初始化模块
数据上传模块
计算设备1100

图 12

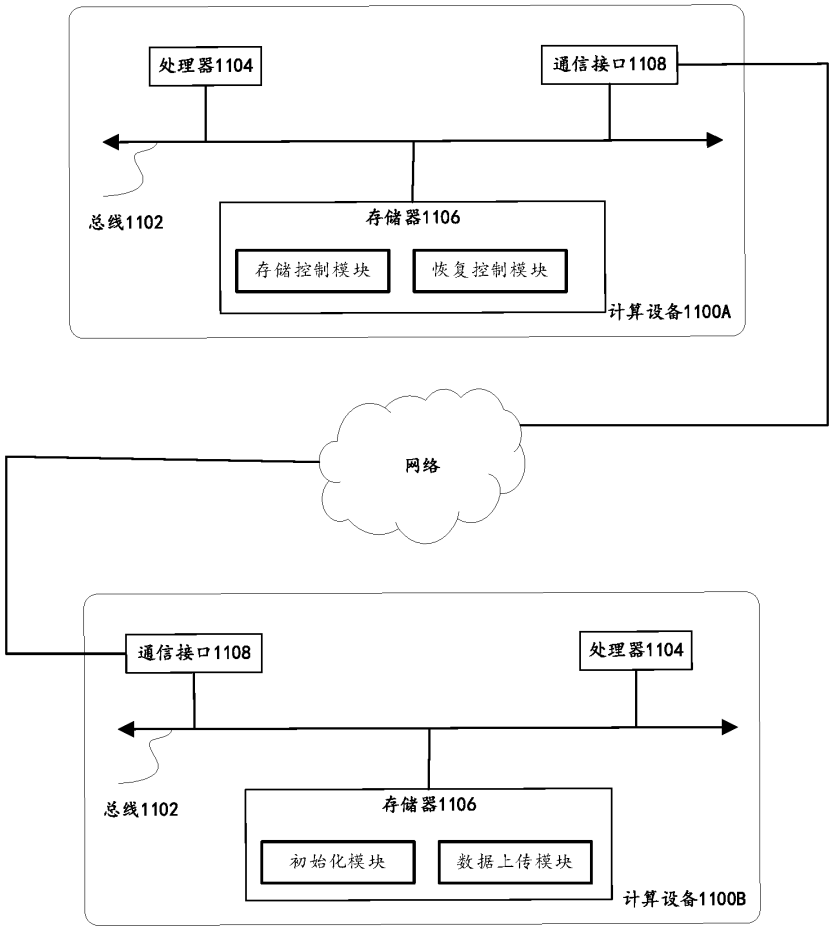

处理器1104
通信接口1108
总线1102
存储器1106
存储控制模块
恢复控制模块
计算设备1100A
网络
通信接口1108
处理器1104
总线1102
存储器1106
初始化模块
数据上传模块
计算设备1100B

图 13