COMPUTING SYSTEM, MULTI-TURN SESSION INFERENCE METHOD AND APPARATUS, AND COMPUTE DEVICE CLUSTER

ABSTRACT

This application provides a computing system, a multi-turn session inference method and apparatus, and a compute device cluster, and pertains to the field of cloud computing technologies. In the system, an external storage device stores a historical key-value cache of a completed session. Because capacity of the external storage device is far greater than capacity of an HBM in an accelerator, a hit rate of historical key-value caches can be improved, thereby avoiding key-value cache recomputation. In addition, when the accelerator processes a session, a host can preload a historical key-value cache of a to-be-processed session in a job queue from the external storage device to a memory of the host. When performing i^{th}-layer computing on the session, the accelerator can preload, from the memory of the host, a historical key-value cache required by $(i+1)^{th}$-layer computing. Because a computing process and a data loading process are synchronously performed, there is no need for the computing process to wait for completion of data loading. In this way, time overheads of accessing the external storage device by the accelerator can be hidden, and multi-turn session inference efficiency is improved.

(FIG.3)