CLAIMS

1. A method for inference of large-scale language models, the method comprising:

receiving a first request of a user, wherein the first request is a request for a first service;

sending the first request to a first electronic device, wherein the first electronic device is configured to perform a first inference calculation in a first processing manner in a prefill phase;

receiving a second request from the first electronic device, wherein the second request comprises a first output information corresponding to the first service, and the first output information comprises a token output for the first request generated by the first electronic device through the first inference calculation;

in response to the first output information meeting a specified condition, ending the scheduling of the first service; and

in response to the first output information not meeting the specified condition, sending the second request to a second electronic device, and receiving a third request from the second electronic device, wherein the second electronic device is different from the first electronic device, the second electronic device is configured to perform a second inference calculation in a second processing manner in a decode phase, and the third request comprises a second output information corresponding to the first service, and the second output information comprises a token output for the second request generated by the second electronic device through the second inference calculation;

in response to the second output information meeting a specified condition, ending the scheduling of the first service; and

in response to the second output information not meeting the specified condition, sending the third request to the second electronic device.

2. The method according to claim 1, wherein the specified condition comprises at least one of:

a length of output information is greater than or equal to a first threshold; or output information comprises a terminator.

3. The method of claim 1, wherein the first inference calculation comprises processing M1 services in parallel, the M1 services includes the first service, M1 is a positive integer, and the method further comprises:

in response to the first inference calculation of the M1 services in a current round being completed, releasing the occupancy of computing service resources of the first electronic device by the M1 services.

4. The method of claim 3, wherein after releasing the occupancy of computing service resources of the first electronic device by the M1 services, the method further comprises:

inserting M2 services into idle computing service resources of the first electronic device to perform a next round of the first inference calculation, wherein M2 is a positive integer.

5. The method of claim 3, wherein M1 equals 1, the M1 services consists solely of the first service, and processing M1 services in parallel comprises processing only the first service.

6. The method of claim 1, wherein the second inference calculation comprises processing N1 services in parallel, the N1 services includes the first service, N1 is a positive integer, and in response to the first output information not meeting the specified condition, the method further comprises:

in response to output information corresponding to the first service in the calculation results of the second inference calculation for N1 services meeting the specified condition, releasing the occupancy of computing service resources of the second electronic device by the first service.

7. The method of claim 6, wherein after releasing the occupancy of computing service resources of the second electronic device by the first service, the method further comprises:

inserting a second service into idle computing service resources of the second electronic device to perform the second inference calculation for N2 services, wherein N2 is a positive integer, wherein the N1 services does not include the second service, and the N2 services includes the second service and does not include the first service.

8. The method of claim 7, wherein both N1 and N2 are greater than each of M1 and M2.

9. The method according to claim 7, wherein the intersection of the N1 services and the N2 services is non-empty.

10. The method according to claim 1, wherein

the first processing manner comprises:

after the inference calculation of all services in a current round has been completed, releasing the computing service resources occupied by the current round and inserting a next round of at least one service into the idle computing service resources to perform the first inference calculation;

the second processing manner comprises:

during the inference process of the current round, if there is a target request whose output information meets the specified condition, releasing the computing service resources of the target request and inserting a service of a next round into the idle computing service resources to perform the second inference calculation.

11. The method according to claim 1, further comprising:

transmitting first cache information generated through the first inference calculation to the second electronic device, wherein the first cache information is used by the second electronic device to perform the second inference calculation.

12. The method according to claim 1, further comprising:

storing second cache information generated through the second inference calculation in the second electronic device, wherein the second cache information is used by the second electronic device to perform the second inference calculation in a next round.

13. The method according to claim 1, wherein after sending the third request to the second

electronic device, the method further comprises:

receiving a fourth request from the second electronic device, wherein the fourth request comprises a third output information corresponding to the first service, the third output information comprises a token output for the third request, and the token output for the third request is generated by the second electronic device through the second inference calculation.

14. The method according to claim 1, wherein the first electronic device is a computing card, and the second electronic device is a computing card.

15. The method according to claim 13, wherein the computing card includes a GPU or an NPU.

16. A model inference system, comprising at least one first electronic device, at least one second electronic device, and a scheduling apparatus, wherein:

the first electronic device is configured to perform a first inference calculation in a prefill phase in a first processing manner;

the second electronic device is configured to perform a second inference calculation in a decode phase in a second processing manner;

the scheduling apparatus is configured to:

receive a first request of a user, wherein the first request is a request for a first service;

send the first request to the first electronic device;

receive a second request from the first electronic device, wherein the second request comprises a first output information corresponding to the first service, and the first output information comprises a token output for the first request generated by the first electronic device through the first inference calculation;

in response to first output information meeting a specified condition, end the scheduling of the first service; and

in response to the first output information not meeting the specified condition, send the second request to the second electronic device, and receive a third request from the second electronic device, wherein the third request comprises a second output information corresponding to the first service, and the second output information comprises a token output for the second request generated by the second electronic device through the second inference calculation;

in response to the second output information meeting a specified condition, end the scheduling of the first service; and

in response to the second output information not meeting the specified condition, send the third request to the second electronic device.

17. The model inference system according to claim 16, wherein the at least one first electronic device and the at least one second electronic device are deployed in an independent separate deployment manner.

18. The model inference system according to claim 16, wherein the resource ratio between the at least one first electronic device and the at least one second electronic device is dynamically adjusted.

19. The model inference system according to claim 16, wherein

the first processing manner comprises:

after the inference calculation of all services in a current round has been completed, releasing the computing service resources occupied by the current round and inserting a next round of at least one service into the idle computing service resources to perform the first inference calculation;

the second processing manner comprises:

during the inference process of the current round, if there is a target request whose output information meets the specified condition, releasing the computing service resources of the target request and inserting a service of a next round into the idle computing service resources to perform the second inference calculation.

20. The model inference system according to claim 16, wherein the specified condition comprises at least one of:

a length of output information is greater than or equal to a first threshold; or output information comprises a terminator.

21. The model inference system according to claim 16, wherein

the second electronic device is configured to receive first cache information generated through the first inference calculation, wherein the first cache information is used by the second electronic device to perform the second inference calculation; and

the second electronic device is configured to store second cache information generated through the second inference calculation, wherein the second cache information is used by the second electronic device to perform the second inference calculation in a next round.

22. The model inference system according to claim 16, wherein the scheduling apparatus is configured to:

after sending the third request to the second electronic device, receive a fourth request from the second electronic device, wherein the fourth request comprises a third output information corresponding to the first service, the third output information comprises a token output for the third request, and the token output for the third request is generated by the second electronic device through the second inference calculation.

23. The model inference system according to claim 16, wherein the first electronic device is a computing card, and the second electronic device is a computing card.

24. The model inference system according to claim 23, wherein the computing card includes a GPU or an NPU.

25. A model inference system, comprising at least one first electronic device and at least one second electronic device, wherein:

the first electronic device is configured to perform a first inference calculation in a prefill phase in a first processing manner;

the second electronic device is configured to perform a second inference calculation in a

decode phase in a second processing manner;

the first processing manner comprises: after the inference calculation of all services in a current round has been completed the inference calculation, releasing the computing service resources occupied by the current round and inserting a next round of at least one service into the idle computing service resources to perform the first inference calculation;

the second processing manner comprises: during the inference process of the current round, if there is a target request whose output information meets the specified condition, releasing the computing service resources of the target request and inserting a service of a next round into the idle computing service resources to perform the second inference calculation.

26. The model inference system of claim 25, wherein in a single round of inference computation, the number of services executed by the first electronic device is less than the number of services executed by the second electronic device.

27. The model inference system according to claim 25, wherein the at least one first electronic device and the at least one second electronic device are deployed in an independent separate deployment manner.

28. The model inference system according to claim 25, wherein the resource ratio between the at least one first electronic device and the at least one second electronic device is dynamically configured.

29. The model inference system according to claim 25, wherein

the second electronic device is configured to receive first cache information generated through the first inference calculation, wherein the first cache information is used by the second electronic device to perform the second inference calculation; and

the second electronic device is configured to store second cache information generated through the second inference calculation, wherein the second cache information is used by the second electronic device to perform the second inference calculation in a next round.

30. The model inference system according to claim 25, wherein the first electronic device is a computing card, and the second electronic device is a computing card.

31. The model inference system according to claim 30, wherein the computing card includes a GPU or an NPU.

32. The model inference system according to claim 25, wherein the model inference system comprises a scheduling apparatus, and the scheduling apparatus is configured to:

receive a first request of a user, wherein the first request is a request for a first service;

send the first request to the first electronic device;

receive a second request from the first electronic device, wherein the second request comprises a first output information corresponding to the first service, and the first output information comprises a token output for the first request generated by the first electronic device through the first inference calculation;

in response to the first output information meeting a specified condition, end the scheduling of the first service;

in response to the first output information not meeting the specified condition, send the second request to the second electronic device, and receive a third request from the second electronic device, wherein the third request comprises a second output information corresponding to the first service, and the second output information comprises a token output for the second request generated by the second electronic device through the second inference calculation;

in response to the second output information meeting a specified condition, end the scheduling of the first service; and

in response to the second output information not meeting the specified condition, send the third request to the second electronic device.

33. The model inference system according to claim 32, wherein the specified condition comprises at least one of:

a length of output information is greater than or equal to a first threshold; or output information comprises a terminator.

34. The model inference system according to claim 32, wherein the scheduling apparatus is configured to:

after sending the third request to the second electronic device, receive a fourth request from the second electronic device, wherein the fourth request comprises a third output information corresponding to the first service, the third output information comprises a token output for the third request, and the token output for the third request is generated by the second electronic device through the second inference calculation.