[0040] In this application, a cache instance is used to cache data stored in a backend storage; and a service instance uses the data cached by the cache instance, that is, uses the data stored in the backend storage. For example, as shown in FIG. 2, data required by the service instance 211 is stored in a backend storage 23, the cache instance 212 obtains the data required by the service instance 211 from the backend storage 23 and caches the data, and the service instance 211 uses the data cached by the cache instance 212. This can increase a speed of obtaining data by the service instance 211.

[0041] Optionally, as shown in FIG. 2, the backend storage 23, a backend storage 24, a backend storage 25, and a backend storage 26 may be different or may be the same. At least two of the backend storage 23, the backend storage 24, the backend storage 25, and the backend storage 26 may be the same.

[0042] Optionally, as shown in FIG. 2, the service instance 211 and the service instance 213 are different service instances running a same service. Therefore, the cache instance 212 and the cache instance 214 may have a same cache configuration.

[0043] Optionally, the service instance 211 and the service instance 213 are different service instances running a same service. Therefore, the cache instance 212 and the cache instance 214 may be the same cache instance 212, as shown in FIG. 3.

[0044] The following describes an instance deployment method according to the present invention with reference to FIG. 2, FIG. 3, and FIG. 4. In the method, when a service instance running a specific service is deployed, a matched cache instance is automatically deployed for the service instance. A cache configuration of the cache instance is customized for the service or the service instance. This improves efficiency of providing a service by the service instance, and improves service performance.

[0045] Step S41: The instance management node 20 receives a request for creating the service instance 211.

[0046] The service instance ~~11~~211 is a service instance running a specific service.

[0047] A scenario of triggering the request of the service instance 211~~11~~ is not limited in this application. For example, when a tenant needs to deploy the service, the tenant or a user of the tenant sends the request for creating the service instance 211 to the instance management node 20.

[0048] Optionally, the request carries a cache configuration of a cache instance used by the service instance 211.

CLAIMS

~~What is claimed is:~~

1. An instance deployment method, wherein the method comprises:

receiving, by an instance management node, a request for creating a service instance;

obtaining, by the instance management node, a cache configuration corresponding to the service instance; and

creating, by the instance management node, the service instance on a first computing node, wherein the service instance uses a cache instance which has the cache configuration.

2. The method according to claim 1, wherein the cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

3. The method according to claim 1 or 2, wherein the method comprises:

creating, by the instance management node, the cache instance on the first computing node.

4. The method according to any one of claims 1 to 3, wherein the method comprises:

creating, by the instance management node according to an affinity principle, the service instance and the cache instance on the first computing node.

5. The method according to claim 1 or 2, wherein the method comprises:

creating, by the instance management node, the cache instance on a second computing node.

6. The method according to any one of claims 1 to 5, wherein the method comprises:

selecting, by the instance management node, a computing node that meets the cache configuration from a plurality of computing nodes, wherein the selected computing node is used to create the cache instance.

7. The method according to any one of claims 1 to 6, wherein the service instance is a container.

8. An system, wherein the system comprises:

a plurality of computing nodes; and a instance management node, configured to receive a

request for creating a service instance, obtain a cache configuration corresponding to the service instance, and create the service instance on a first computing node of the plurality of computing nodes, wherein the service instance uses a cache instance which has the cache configuration.

9. The system according to claim 8, wherein the cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

10. The system according to claim 8 or 9, wherein the instance management node is configured to create the cache instance on the first computing node.

11. The method according to any one of claims 8 to 10, wherein the instance management node is configured to create the service instance and the cache instance on the first computing node according to an affinity principle.

12. The system according to claim 8 or 9, wherein the instance management node is configured to create the cache instance on a second computing node of the plurality of computing nodes.

13. The system according to any one of claims 8 to 12, wherein the instance management node is configured to select a computing node that meets the cache configuration from the plurality of computing nodes, wherein the selected computing node is used to create the cache instance.

14. The system according to any one of claims 8 to 13, wherein the service instance is a container.

15. A computing device, comprising a processor and a memory, wherein

the memory is configured to store computer instructions; and

the processor is configured to execute the computer instructions stored in the memory, so that the computing device is implemented as a instance management node which is configured to:

receive a request for creating a service instance, obtain a cache configuration corresponding to the service instance, and create the service instance on a first computing node, wherein the service instance uses a cache instance which has the cache configuration.

16. The computing device according to claim 15, wherein the cache configuration comprises

at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

17. The computing device according to claim 15 or 16, wherein the instance management node is configured to create the cache instance on the first computing node.

18. The computing device according to any one of claims 15 to 17, wherein the instance management node is configured to create the service instance and the cache instance on the first computing node according to an affinity principle.

19. The computing device according to claim 15 or 16, wherein the instance management node is configured to create the cache instance on a second computing node.

20. The computing device according to any one of claims 15 to 19, wherein the instance management node is configured to select a computing node that meets the cache configuration from a plurality of computing nodes, wherein the selected computing node is used to create the cache instance.

21. The computing device according to any one of claims 15 to 20, wherein the service instance is a container.

~~receiving, by an instance management node, a request for creating a first service instance;~~

~~obtaining, by the instance management node, a first cache configuration corresponding to the first service instance; and~~

~~creating, by the instance management node, the first service instance on a computing node, and creating a first cache instance on the computing node based on the first cache configuration, wherein the first cache instance is used by the first service instance.~~

~~2. The method according to claim 1, wherein the method comprises:~~

~~receiving, by the instance management node, a request for creating a second service instance;~~

~~obtaining, by the instance management node, a second cache configuration corresponding to the second service instance; and~~

~~creating, by the instance management node, the second service instance on the computing node, and creating a second cache instance on the computing node based on the second cache configuration, wherein the second cache instance is used by the second service instance.~~

3. The method according to claim 1 or 2, wherein the first cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

4. The method according to any one of claims 1 to 3, wherein the first cache configuration is carried in the request for creating the first service instance.

5. The method according to any one of claims 1 to 3, wherein the obtaining, by the instance management node, a first cache configuration corresponding to the first service instance comprises:

obtaining, by the instance management node, the stored first cache configuration based on the first service instance.

6. The method according to any one of claims 1 to 5, wherein the method comprises:

selecting, by the instance management node, one computing node that meets the first service instance and the first cache instance from a plurality of computing nodes.

7. The method according to any one of claims 1 to 6, wherein the first cache instance having the first cache configuration is dedicated to the first service instance.

8. An instance management node, wherein the instance management node comprises:

a receiving unit, configured to receive a request for creating a first service instance;

an obtaining unit, configured to obtain a first cache configuration corresponding to the first service instance; and

a creation unit, configured to: create the first service instance on a computing node, and create a first cache instance on the computing node based on the first cache configuration, wherein the first cache instance is used by the first service instance.

9. The instance management node according to claim 8, wherein

the receiving unit is configured to receive a request for creating a second service instance;

the obtaining unit is configured to obtain a second cache configuration corresponding to the second service instance; and

the creation unit is configured to: create the second service instance on the computing node, and create a second cache instance on the computing node based on the second cache configuration, wherein the second cache instance is used by the second service instance.

10. The instance management node according to claim 8 or 9, wherein the first cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

11. The instance management node according to any one of claims 8 to 10, wherein the first cache configuration is carried in the request for creating the first service instance.

12. The instance management node according to any one of claims 8 to 10, wherein

the obtaining unit is configured to obtain the stored first cache configuration based on the first service instance.

13. The instance management node according to any one of claims 8 to 12, wherein the instance management node comprises:

a selection unit, configured to select one computing node that meets the first service instance and the first cache instance from a plurality of computing nodes.

14. The instance management node according to any one of claims 8 to 13, wherein the first cache instance having the first cache configuration is dedicated to the first service instance.

15. A computing device, comprising a processor and a memory, wherein

the memory is configured to store computer instructions; and

the processor is configured to execute the computer instructions stored in the memory, so that the computing device performs the steps implemented by the instance management node in the method according to any one of claims 1 to 7.

16. A computer program product, wherein the computer program product comprises computer instructions, and the computer instructions instruct a computing device to perform the steps implemented by the instance management node in the method according to any one of claims 1 to 7

17. A computer-readable storage medium, wherein the computer-readable storage medium comprises computer instructions, and the computer instructions instruct a computing device to perform the steps implemented by the instance management node in the method according to any one of claims 1 to 7

18. An instance deployment method, wherein the method comprises:

receiving, by a computing node, a request sent by an instance management node, wherein the request carries a cache configuration; and

creating, by the computing node, a service instance specified in the request, and creating, based on the cache configuration, a cache instance specified in the request, wherein the cache instance is used by the service instance.

19. The method according to claim 18, wherein the cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

20. The method according to claim 18 or 19, wherein the cache instance having the cache configuration is dedicated to the service instance.

21. A computing node, wherein the computing node comprises:

a receiving unit, configured to receive a request sent by an instance management node, wherein the request carries a cache configuration; and

a creation unit, configured to: create a service instance specified in the request, and create, based on the cache configuration, a cache instance specified in the request, wherein the cache instance is used by the service instance.

22. The computing node according to claim 21, wherein the cache configuration comprises at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

a cache algorithm.

23. The computing node according to claim 21 or 22, wherein the cache instance having the cache configuration is dedicated to the service instance.

24. A computing device, comprising a processor and a memory, wherein

the memory is configured to store computer instructions; and

the processor is configured to execute the computer instructions stored in the memory, so that the computing device performs the steps implemented by the computing node in the method

~~according to any one of claims 18 to 20.~~

~~25. A computer program product, wherein the computer program product comprises computer instructions, and the computer instructions instruct a computing device to perform the steps implemented by the computing node in the method according to any one of claims 18 to 20~~

~~26. A computer-readable storage medium, wherein the computer-readable storage medium comprises computer instructions, and the computer instructions instruct a computing device to perform the steps implemented by the computing node in the method according to any one of claims 18 to 20~~

~~27. A system, wherein the system comprises an instance management node and a computing node, wherein~~

~~the instance management node is configured to: receive a request for creating a service instance, obtain a cache configuration corresponding to the service instance, and send a request carrying the cache configuration to the computing node; and~~

~~the computing node is configured to: receive the request sent by the instance management node, create the service instance specified in the request, and create, based on the cache configuration, a cache instance specified in the request, wherein the cache instance is used by the service instance.~~