

# **INSTANCE DEPLOYMENT METHOD, INSTANCE MANAGEMENT NODE, COMPUTING NODE, AND COMPUTING DEVICE**

## **TECHNICAL FIELD**

**[0001]** This application relates to the database field, and in particular, to an instance  
5 deployment method, an instance management node, a computing node, a computing device, a  
computer-readable storage medium, and a computer program product.

## **BACKGROUND**

**[0002]** As shown in FIG. 1, a cache instance 11 is manually created on a computing node 11  
in advance. Subsequently, a tenant may specify to create a service container 12 on the computing  
10 node 11 by using a container management node.

## **SUMMARY**

**[0003]** In view of this, this application provides an instance deployment method, an instance  
management node, a computing node, a computing device, a computer-readable storage medium,  
and a computer program product, to automatically deploy a cache instance and a service instance  
15 through matching.

**[0004]** According to a first aspect, an instance deployment method is provided. In the instance  
deployment method, an instance management node receives a request for creating a service  
instance. The instance management node obtains a cache configuration corresponding to the  
service instance. For example, the cache configuration corresponding to the service instance may  
20 be a cache configuration that matches the service instance. The instance management node creates  
the service instance on a computing node, and creates a cache instance on the computing node  
based on the cache configuration. In this way, the service instance may provide a service by using  
the matched cache instance. This improves efficiency of providing a service by the service instance,  
and improves service performance.

**[0005]** In a possible design of the first aspect, in the method, the cache configuration includes at least one of the following configurations:

a cache size;

a cache medium;

5 a cache write policy; and

a cache algorithm.

**[0006]** In a possible design of the first aspect, in the method, the cache configuration is carried in the request for creating the service instance. For example, a tenant or a user of a tenant may specify, by using the request, the cache configuration that matches the service instance.

10 **[0007]** In this way, the instance management node obtains the cache configuration from the request, and may create, based on the cache configuration, the cache instance that is manually customized for the service instance.

**[0008]** In a possible design of the first aspect, in the method, the instance management node obtains the stored cache configuration based on the service instance. In this way, when creating  
15 the service instance, the instance management node may automatically obtain the cache configuration that matches the service instance, to create the cache instance that matches the service instance. This improves efficiency of providing a service by the service instance, and improves service performance.

**[0009]** In a possible design of the first aspect, in the method, the instance management node  
20 selects one computing node that meets the service instance and the cache instance from a plurality of computing nodes.

**[0010]** In this way, the instance management node selects a computing node that has sufficient resources and whose configuration meets the cache configuration from the plurality of computing nodes. This can avoid a creation failure caused by directly creating the service instance and the  
25 cache instance on a computing node that has insufficient resources or whose configuration does not meet the cache configuration.

**[0011]** In a possible design of the first aspect, in the method, the instance management node receives a request for creating another service instance. The instance management node obtains a cache configuration corresponding to the another service instance. For example, the cache  
30 configuration corresponding to the another service instance may be a cache configuration that matches the another service instance. The instance management node creates the another service

instance on the computing node, and creates a cache instance on the computing node based on the cache configuration corresponding to the another service instance. In this way, the another service instance may provide a service by using the matched cache instance. Therefore, two types of cache instances having different cache configurations may be provided on a same computing node, and different services are provided by running different service instances based on the cache instances having different cache configurations. This increases computing node resource utilization, and also provides a manner of deploying heterogeneous service instances.

**[0012]** In a possible design of the first aspect, in the method, cache instances having different cache configurations are dedicated to different service instances, that is, a cache instance is customized for a service instance.

**[0013]** In this way, the service instance may provide a service by using the customized cache instance.

**[0014]** In a possible design of the first aspect, in the method, the computing node receives a request sent by the instance management node. The request carries the cache configuration.

**[0015]** The computing node creates the service instance specified in the request, and creates, based on the cache configuration, the cache instance specified in the request. The cache instance is used by the service instance.

**[0016]** According to a second aspect, this application provides an instance management node. The instance management node includes one or more function modules, configured to implement the steps implemented by the instance management node in the method according to any one of the first aspect and the possible designs of the first aspect.

**[0017]** This application provides a computing node. The computing node includes one or more function modules, configured to implement the steps implemented by the computing node in the method according to any one of the first aspect and the possible designs of the first aspect.

**[0018]** According to a third aspect, this application provides a computing device. The computing device includes a processor and a memory. The memory stores computer instructions. The processor executes 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 the first aspect and the possible designs of the first aspect, or the computing device implements the instance management node according to the second aspect. Alternatively, the processor executes 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 the first aspect and the possible designs of the first aspect, or the computing device implements the computing node according to the second aspect.

**[0019]** According to a fourth aspect, this application provides a computer-readable storage medium. The computer-readable storage medium stores computer instructions. When a processor in a computing device executes the computer instructions, the computing device performs the steps implemented by the instance management node in the method according to any one of the first aspect and the possible designs of the first aspect, or the computing device implements the instance management node according to the second aspect. Alternatively, when a processor in a computing device executes the computer instructions, the computing device performs the steps implemented by the computing node in the method according to any one of the first aspect and the possible designs of the first aspect, or the computing device implements the computing node according to the second aspect.

**[0020]** According to a fifth aspect, this application provides a computer program product. The computer program product includes computer instructions, and the computer instructions are stored in a computer-readable storage medium. A processor in a computing device may read the computer instructions from the computer-readable storage medium. The processor executes the computer instructions, so that the computing device performs the steps implemented by the instance management node in the method according to any one of the first aspect and the possible designs of the first aspect, or the computing device implements the instance management node according to the second aspect. Alternatively, the processor executes the computer instructions, so that the computing device performs the steps implemented by the computing node in the method according to any one of the first aspect and the possible designs of the first aspect, or the computing device implements the computing node according to the second aspect.

## **BRIEF DESCRIPTION OF DRAWINGS**

**[0021]** FIG. 1 is a schematic diagram of cache instance deployment;

**[0022]** FIG. 2 is a schematic diagram of an application scenario to which this application is applicable;

**[0023]** FIG. 3 is a schematic diagram of an application scenario to which this application is

applicable;

**[0024]** FIG. 4 is a schematic flowchart of an instance deployment method according to this application;

**[0025]** FIG. 5 is a schematic diagram of a logical structure of an instance management node  
5 20 according to this application;

**[0026]** FIG. 6 is a schematic diagram of a logical structure of a computing node according to this application; and

**[0027]** FIG. 7 is a schematic diagram of a structure of a computing device 700 according to this application.

10

## DESCRIPTION OF EMBODIMENTS

**[0028]** The following describes technical solutions provided in this application with reference to the accompanying drawings in this application.

**[0029]** In this application, a cache instance is created based on a cache configuration, so that the cache configuration can be adjusted to customize the cache instance for a service instance.

15

**[0030]** The cache configuration includes at least one of the following configurations:

a cache size (cache size);

a cache medium;

a cache write policy; and

a cache algorithm.

20

**[0031]** For example, the cache size is used to specify a size of cache space that can be provided by the cache instance for the service instance; or the cache size is used to specify a size of cache space of the cache instance.

25

**[0032]** For example, the cache medium may be a solid-state drive (solid-state drive, SSD), a static random access memory (static random access memory, SRAM), a flash, or another storage medium; or the cache medium may be a storage area network (storage area network, SAN), a network attached storage (network attached storage, NAS), or another storage manner.

**[0033]** For example, the cache write policy may be write through (write through, WT), or may be write back (write back, WB), or may be another write manner for writing data to a cache.

**[0034]** For example, the cache algorithm may be a least frequently used (Least Frequently

Used, LFU) eviction algorithm, a least recently used (Least Recently Used, LRU) eviction algorithm, an adaptive replacement cache (Adaptive Replacement Cache, ARC) algorithm, or another cache eviction algorithm; or the cache algorithm may be a no replacement cache (no replacement cache, NO REP) policy.

5   **[0035]**     Therefore, if two cache configurations have different cache sizes, the two cache configurations are different; if two cache configurations have different cache media, the two cache configurations are different; if two cache configurations have different cache write policies, the two cache configurations are different; if two cache configurations have different cache algorithms, the two cache configurations are different.

10   **[0036]**     FIG. 2 is a diagram of an architecture of instance deployment according to this application.

**[0037]**     As shown in FIG. 2, an instance management node 20 may manage service instances and cache instances on a plurality of computing nodes (for example, a computing node 21 and a computing node 22).

15   **[0038]**     For example, the instance management node 20 may deploy, on the computing node 21, a service instance 211 and a cache instance 212 used by the service instance 211, where a cache configuration of the cache instance 212 is set based on a service run by the service instance 211; and the instance management node 20 may deploy, on the computing node 21, a service instance 213 and a cache instance 214 used by the service instance 213, where a cache configuration of the  
20   cache instance 214 is set based on a service run by the service instance 213. If the service instance 211 and the service instance 213 run different services, the cache instance 212 and the cache instance 214 may have different cache configurations. In this way, cache instances having different cache configurations may be provided to cooperate with service instances running different services. This improves efficiency of providing a service or improves service performance.

25   **[0039]**     For example, if a service has a priority, a service instance running a high-priority service may be deployed on a high-performance computing node. For example, performance of the computing node 21 is higher than performance of the computing node 22, both the service instance 211 and the service instance 213 provide high-priority services, and the service instance 211 and the service instance 213 are deployed on the computing node 21; both a service instance  
30   221 and a service instance 223 provide low-priority services, and the service instance 221 and the service instance 223 are deployed on the computing node 22.

**[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  
5 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  
10 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  
15 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  
20 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  
25 instance 211.

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

**[0047]** A scenario of triggering the request of the service instance 211 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  
30 service instance 211.

**[0049]** Optionally, the request does not carry a cache configuration of a cache instance used by the service instance 211.

**[0050]** Step S42: The instance management node 20 obtains the cache configuration corresponding to the service instance 211.

5 **[0051]** In this application, the service instance 211 uses the matched cache instance to improve efficiency of providing a service or improve service performance. Specifically, the service or the service instance 211 has a customized cache configuration, and a cache instance created by using the cache configuration is used to provide a cache service for the service instance 211.

**[0052]** Optionally, if the request received in step S41 carries the cache configuration, the  
10 instance management node 20 obtains the cache configuration from the request, and may create, based on the cache configuration, the cache instance customized for the service instance 211.

**[0053]** Optionally, if the cache configuration of the cache instance that matches the service instance 211 is pre-stored, the instance management node 20 obtains the stored cache configuration, and may create, based on the cache configuration, the cache instance customized for the service  
15 instance 211.

**[0054]** Step S43: The instance management node 20 creates the service instance 211 on the computing node 21, and creates the cache instance 212 on the computing node 21 based on the cache configuration obtained in step S42.

**[0055]** The created cache instance 212 is used by the service instance 211.

20 **[0056]** Optionally, the instance management node 20 creates the service instance 211 and the cache instance 212 according to an affinity principle. This can ensure that the service instance 211 and the cache instance 212 are created on a same computing node 21.

**[0057]** Optionally, the backend storage 23 stores data for use by the service instance 211, and the cache instance 212 may be created based on the backend storage 23. In this way, the data in  
25 the backend storage 23 may be read into the cache instance 212, and subsequently the service instance 211 directly uses the data provided by the cache instance 212, that is, the service instance 211 uses the data in the backend storage 23 by using the cache instance 212.

**[0058]** For example, a data volume is created based on the backend storage 23, and a path of the data volume is stored in the cache instance 212. In this way, the cache instance 212 may obtain  
30 data of the backend storage 23 by using the data volume.

**[0059]** Optionally, the instance management node 20 selects one computing node that meets

the service instance 211 and the cache instance 212 from a plurality of computing nodes, to create the service instance 211 and the cache instance 212.

**[0060]** For example, the computing node 21 has an SSD, and the computing node 22 does not have an SSD. The instance management node 20 obtains the cache configuration corresponding to the service instance 211 in step S42, and a cache medium specified in the cache configuration is an SSD. Therefore, the instance management node 20 selects the computing node 21 to create the service instance 211 and the cache instance 212, creates, based on the cache configuration obtained in step S42, the cache instance 212 by using the SSD of the computing node 21, and creates the service instance 211 on the computing node 21.

**[0061]** Optionally, the instance management node 20 generates a request in step S43. The request carries the cache configuration corresponding to the service instance 211, and the request is used to indicate the computing node 21 to create the service instance 211 and the cache instance.

**[0062]** Correspondingly, the computing node 21 receives the request sent by the instance management node 20. The computing node 21 creates the service instance 211 specified in the request, and creates, based on the cache configuration corresponding to the service instance 211, the cache instance 212 specified in the request. In this way, the cache instance 212 may be used by the service instance 211. The service instance 211 may provide a service by using the cache instance 212. This improves efficiency of providing a service by the service instance, and improves service performance.

**[0063]** In the instance deployment method provided in the present invention, a new specific application scenario is provided. To be specific, cache instances having different cache configurations are created on a same computing node to be used by service instances running different services. Correspondingly, the method further includes step S44, step S45, and step S46.

**[0064]** Step S44: The instance management node 20 receives a request for creating the service instance 213.

**[0065]** The service instance 213 and the service instance 211 are service instances that provide different services. The service instance 213 and the service instance 211 need to use cache instances having different cache configurations to provide different services.

**[0066]** An implementation of step S44 is similar to an implementation of step S41. Details are not described herein again.

**[0067]** Step S45: The instance management node 20 obtains a cache configuration

corresponding to the service instance 213.

**[0068]** The cache configuration corresponding to the service instance 213 and the cache configuration corresponding to the service instance 211 are different. For example, a cache size (for example, 50 megabytes) required by the cache configuration corresponding to the service instance 213 is different from a cache size (for example, 60 megabytes) required by the cache configuration corresponding to the service instance 211. For example, a cache algorithm (LRU) required by the cache configuration corresponding to the service instance 213 is different from a cache algorithm (NO REP) required by the cache configuration corresponding to the service instance 211. For example, a cache write policy (WT) required by the cache configuration corresponding to the service instance 213 is different from a cache write policy (WB) required by the cache configuration corresponding to the service instance 211.

**[0069]** An implementation of step S45 is similar to an implementation of step S42. Details are not described herein again.

**[0070]** Step S46: The instance management node 20 creates the service instance 213 on the computing node 21, and creates the cache instance 214 on the computing node 21 based on the cache configuration obtained in step S45.

**[0071]** The cache instance 214 is used by the service instance 213.

**[0072]** An implementation of step S46 is similar to an implementation of step S43. Details are not described herein again.

**[0073]** Optionally, when the computing node 21 meets both a resource requirement of the service instance 213 and a resource requirement of the cache instance 214, the instance management node 20 creates the service instance 213 and the cache instance 214 on the computing node 21. Optionally, the instance management node 20 generates a request in step S46. The request carries the cache configuration corresponding to the service instance 213, and the request is used to indicate the computing node 21 to create the service instance 213 and the cache instance.

**[0074]** Correspondingly, the computing node 21 receives the request sent by the instance management node 20. The computing node 21 creates the service instance 213 specified in the request, and creates, based on the cache configuration corresponding to the service instance 213, the cache instance 214 specified in the request. In this way, the cache instance 214 may be used by the service instance 213. The service instance 213 may provide a service by using the cache instance 214. This improves efficiency of providing a service by the service instance, and improves

service performance.

**[0075]** This application further provides an instance management node 20. Specifically, a function of the instance management node 20 may be implemented on a computing device (for example, a server or another device having a computing capability), so that the computing device serves as the instance management node 20.

**[0076]** The instance management node 20 includes function units configured to implement the instance deployment method in this application. A manner of dividing the instance management node 20 into function units is not limited in this application. The following provides an example of function unit division, as shown in FIG. 5.

**[0077]** The instance management node 20 for instance deployment in FIG. 5 includes:  
a receiving unit 51, configured to receive a request for creating a first service instance;  
an obtaining unit 52, configured to obtain a first cache configuration corresponding to the first service instance; and  
a creation unit 53, 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, where the first cache instance is used by the first service instance.

**[0078]** Optionally, the receiving unit 51 is configured to receive a request for creating a second service instance.

**[0079]** The obtaining unit 52 is configured to obtain a second cache configuration corresponding to the second service instance.

**[0080]** The creation unit 53 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, where the second cache instance is used by the second service instance.

**[0081]** Optionally, the first cache configuration includes at least one of the following configurations:

- a cache size;
- a cache medium;
- a cache write policy; and
- a cache algorithm.

**[0082]** Optionally, the first cache configuration is carried in the request for creating the first service instance.

**[0083]** Optionally, the obtaining unit 52 is configured to obtain the stored first cache configuration based on the first service instance.

**[0084]** Optionally, the instance management node 20 includes:

5 a selection unit 54, configured to select one computing node that meets the first service instance and the first cache instance from a plurality of computing nodes.

**[0085]** Optionally, the first cache instance having the first cache configuration is dedicated to the first service instance.

**[0086]** Optionally, the first service instance may be the service instance 211, the first cache instance may be the service instance 212, the second service instance may be the service instance 10 213, and the second cache instance may be the service instance 214.

**[0087]** This application further provides a computing node. For example, the computing node is the computing node 21 or the computing node 22 in this application. Specifically, a function of the computing node may be implemented on a computing device (for example, a server or another device having a computing capability), so that the computing device serves as the computing node.

15 **[0088]** The computing node includes function units configured to implement the instance deployment method in this application. A manner of dividing the computing node into function units is not limited in this application. The following provides an example of function unit division, as shown in FIG. 6.

**[0089]** The computing node for instance deployment in FIG. 6 includes:

20 a receiving unit 61, configured to receive a request sent by an instance management node, where the request carries a first cache configuration; and

a creation unit 62, configured to: create a first service instance specified in the request, and create, based on the first cache configuration, a first cache instance specified in the request, where the first cache instance is used by the first service instance.

25 **[0090]** Optionally, the first cache configuration includes at least one of the following configurations:

a cache size;

a cache medium;

a cache write policy; and

30 a cache algorithm.

**[0091]** Optionally, the first cache instance having the first cache configuration is dedicated to

the first service instance.

**[0092]** Optionally, the first service instance may be the service instance 211, the first cache instance may be the service instance 212, the second service instance may be the service instance 213, and the second cache instance may be the service instance 214.

5 **[0093]** Optionally, FIG. 7 schematically provides a possible basic architecture of a computing device in this application.

**[0094]** As shown in FIG. 7, the computing device 700 includes a processor 701, a memory 702, a communications interface 703, and a bus 704.

**[0095]** The computing device 700 may include one or more processors 701. FIG. 7 shows only  
10 one processor 701. Optionally, the processor 701 may be a central processing unit (central processing unit, CPU). If the computing device 700 has a plurality of processors 701, the plurality of processors 701 may be of different types or may be of a same type. Optionally, the plurality of processors 701 in the computing device 700 may be further integrated into a multi-core processor.

**[0096]** The memory 702 stores computer instructions and data. The memory 702 may store  
15 computer instructions and data required for implementing the instance deployment method provided in this application. For example, the memory 702 stores computer instructions used to implement the steps implemented by the instance management node 20 in the instance deployment method. For example, the memory 702 stores computer instructions used to implement the steps implemented by the computing node in the instance deployment method. For another example, the  
20 memory 702 stores computer instructions used to implement the instance management node 20 or the computing node. The memory 702 may be any one or any combination of the following storage media: a nonvolatile memory (for example, a read-only memory (ROM), a solid-state drive (SSD), a hard disk drive (HDD), or an optical disc), and a volatile memory.

**[0097]** The communications interface 703 may be any one or any combination of components  
25 having a network access function such as a network interface (for example, an Ethernet interface) and a wireless network interface card.

**[0098]** The communications interface 703 is configured to perform data communication between the computing device 700 and another computing device or a terminal.

**[0099]** In FIG. 7, a thick line is used to represent the bus 704. The bus 704 may connect the  
30 processor 701 to the memory 702 and the communications interface 703. In this way, through the bus 704, the processor 701 may access the memory 702, and may further exchange data with

another computing device or a terminal through the communications interface 703.

**[00100]** In this application, the computing device 700 executes the computer instructions in the memory 702, so that the computing device 700 implements the instance deployment method provided in this application, for example, the computing device performs the steps implemented  
5 by the instance management node 20 in the instance deployment method, or the computing device performs the steps implemented by the computing node in the instance deployment method. Alternatively, the computing device 700 executes the computer instructions in the memory 702, so that the computing device 700 implements the instance management node 20 or the computing node in this application.

**[00101]** This application provides a computer-readable storage medium. The computer-readable storage medium stores computer instructions. When the processor 701 in the computing device 700 executes the computer instructions, the computing device 700 performs the steps implemented by the instance management node 20 in the instance deployment method provided in this application, or the computing device 700 implements functions of the instance management  
10 node 20 in this application. Alternatively, when the processor 701 in the computing device 700 executes the computer instructions, the computing device 700 performs the steps implemented by the computing node in the instance deployment method provided in this application, or the computing device implements the computing node provided in this application.

**[00102]** This application provides a computer program product. The computer program product  
20 includes computer instructions, and the computer instructions are stored in a computer-readable storage medium. The processor 701 in the computing device 700 may read the computer instructions from the computer-readable storage medium, and the processor 701 executes the computer instructions, so that the computing device 700 performs the steps implemented by the instance management node 20 in the instance deployment method, or the computing device 700  
25 implements functions implemented by the instance management node 20 in this application. Alternatively, the processor 701 in the computing device 700 may read the computer instructions from the computer-readable storage medium, and the processor 701 executes the computer instructions, so that the computing device 700 performs the steps implemented by the computing node in the instance deployment method provided in this application, or the computing device 700  
30 implements functions implemented by the computing node in this application.

**[00103]** Embodiments are merely intended for describing the technical solutions of the present

invention, but not for limiting the present invention. Although the present invention is described in detail with reference to embodiments, persons of ordinary skill in the art should understand that they may still make modifications to the technical solutions described in embodiments, without departing from the protection scope of the claims.

## CLAIMS

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

receiving, by an instance management node, a request for creating a first service instance, wherein the first service instance is a first container;

5 obtaining, by the instance management node, a first cache configuration corresponding to the first service instance, wherein the first cache configuration comprises a cache size, a cache write policy, a cache algorithm, and a cache medium; and

creating, by the instance management node, the first service instance on a computing node, wherein the first service instance is configured to use a first cache instance which has the first  
10 cache configuration, is on a computing node that meets the first cache configuration, and is for caching data that are stored in a backend storage and that are for use by the first service instance; and

creating the first cache instance on the computing node for the first cache instance to run on the computing node based on the first cache configuration and an affinity principle between the  
15 first service instance and the first cache instance, 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  
20 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 is carried in  
25 the request for creating the first service instance.

4. The method according to claim 1 or 2, 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.

30 5. The method according to any one of claims 1 to 4, 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.

6. The method according to any one of claims 1 to 5, wherein the first cache instance having the first cache configuration is dedicated to the first service instance.

5 7. An instance management node, wherein the instance management node comprises:

a receiving unit, configured to receive a request for creating a first service instance, wherein the first service instance is a first container;

an obtaining unit, configured to obtain a first cache configuration corresponding to the first service instance, wherein the first cache configuration comprises a cache size, a cache write policy,  
10 a cache algorithm, and a cache medium; and

a creation unit, configured to: create the first service instance on a computing node, wherein the first service instance is configured to use a first cache instance which has the first cache configuration, is on a computing node that meets the first cache configuration, and is for caching data that are stored in a backend storage and that are for use by the first service instance; and

15 the creation unit, further configured to: create the first cache instance on the computing node for the first cache instance to run on the computing node based on the first cache configuration and an affinity principle between the first service instance and the first cache instance, wherein the first cache instance is used by the first service instance.

8. The instance management node according to claim 7, wherein

20 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,  
25 wherein the second cache instance is used by the second service instance.

9. The instance management node according to claim 8 or 9, wherein the first cache configuration is carried in the request for creating the first service instance.

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

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

11. The instance management node according to claim 8 or 9, 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.

12. The instance management node according to claim 8 or 9, wherein the first cache instance  
5 having the first cache configuration is dedicated to the first service instance.

13. 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  
10 method according to any one of claims 1 to 6.

14. 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  
6.

15. 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 6.

16. A system, wherein the system comprises an instance management node and a computing  
20 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, wherein the service instance is a container,  
and the cache configuration comprises a cache size, a cache write policy, a cache algorithm, and a  
25 cache medium; and

the computing node is configured to: receive the request sent by the instance management  
node, create the service instance specified in the request, wherein the service instance is configured  
to use a cache instance which has the cache configuration, is on a computing node that meets the  
cache configuration, and is for caching data that are stored in a backend storage and that are for  
30 use by the service instance; and create, based on the cache configuration and an affinity principle  
between the service instance and the cache instance, the cache instance specified in the request,

wherein the cache instance is used by the service instance.