MULTI-CLUSTER SERVICE SYSTEM, SERVICE ACCESS METHOD, INFORMATION CONFIGURATION METHOD, DEVICE, AND MEDIUM

[0001] The present application claims priority to Chinese Patent Application No. 202210346226.1, filed with the China Patent Office on March 31, 2022 and titled "MULTI-CLUSTER SERVICE SYSTEM, SERVICE ACCESS METHOD, INFORMATION CONFIGURATION METHOD, DEVICE, AND MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present application relates to the field of cloud computing technology, and, in particular, to a multi-cluster service system, a service access method and an information configuration method, a device, and a medium.

BACKGROUND

[0003] With the development of cloud computing technology, a container orchestration system based on kubernetes (K8s) has become the de facto standard in the industry, and K8s multi-cluster deployment has become a mainstream model. The deployment of multiple K8s clusters can be based on factors such as application layering and deployment location, and there are a large number of access demands for services between the multiple K8s clusters.

[0004] In existing technologies, discovery and synchronous deployment of service information between different clusters are implemented by extending a CRD (Custom Resource Definition). However, such an approach is mainly applicable to multi-cluster services located in a same virtual private cloud (VPC). If two clusters are distributed in different VPCs, addresses of the VPCs are required not to overlap. Otherwise, access between multi-cluster services across the VPCs cannot be implemented through a private network.

A1-VPC2/pod3/eni-c1/port number P1; VIP:eni-b1/port number P1→RS2:service A1-VPC2/pod4/eni-c1/port number P1pod4. Assuming that Port D1 in service cluster 11 in VPC1 that serves as a client VPC serves as a client of service A1, a mapped port number of service A1 in VPC1 is P2, ENI information in VPC1 is eni-d1, and this eni-d1 carries endpoint endpoint1 corresponding to service A1, and then a second information mapping relationship includes VIP:eni-b1/port number P1→VPC1:endpoint1/eni-d1/port number P2.

[0075] The multi-cluster load balancing node also can create a listener for the target service, which is used for listening to the port number of the target service. Certainly, the port number can also be empty by default, or all port numbers can be legal by default.

[0076] In this embodiment, based on the above first information mapping relationship and second information mapping relationship, as shown in FIG. 1a, in a case where service D1 needs to access service A1, service instance pod1 sends an access request by accessing eni-d1 in VPC1, and provides this access request to lb1 node through a privatelink between endpoint1 and lb1 node, and a listener of lb1 node listens to an access request from endpoint1, this access request containing an IP address of eni-d1 or endpoint1; according to the second information mapping relationship, eni-b1 that has a mapping relationship with eni-d1 or endpoint1 is determined, then the first information mapping relationship is queried according to eni-b1 to determine service instances pod3 and pod4 which can provide service A1, and eni-c1 in VPC2, and one of pod3 and pod4 is finally selected as a target service instance through load balancing; taking pod3 as selected as an example, then a source IP address of the access request is converted into an IP address of eni-c1, a destination IP address is converted into an IP address of pod3, and the access request after address conversion is provided for pod3 through eni-c1 in VPC2, so as to enable pod3 to provide service A1 for pod1, achieving the purpose of access of pod1 to service A1. In FIG. 1a, a process of accessing service A1 by service D1 is indicated by a dashed line.

[0077] In an embodiment of the present application, structures of implementation of cluster management and control node 105 and multi-cluster service management and control node 104 are not limited. In an optional embodiment, cluster management and control node 105 includes: a cluster service management plane and a cluster interface server. As shown in FIG. 2b, taking a K8s cluster as an example, the cluster service

selected as the target service instance. It should be noted that the weight of the service VPC can be flexibly adjusted according to an application demand. For example, in a case where a local VPC needs to be upgraded, a remote service VPC can be set to have a higher load weight, to facilitate preferable selection of the remote service VPC to perform a service.

[0082] Optionally, load balancing weights corresponding to the plurality of service VPCs are determined according to a location relationship between regions where the plurality of service VPCs are located and a region where the multi-cluster load balancing node is located. For example, a load balancing weight of a service VPC located in a same region as the multi-cluster load balancing node is greater than a load balancing weight of a service VPC located in a different region from the multi-cluster load balancing node. Further, for service VPCs located in a same region as the multi-cluster load balancing node, load balancing weights of these service VPCs can be determined according to a distance between the region where the service VPCs are located and a region where a client VPC is located; the closer the distance between the region where the service VPCs are located and the region where the client VPC is located is, the greater the load balancing weights of the service VPCs are. In this way, an effect of providing a service nearby can be achieved, which is beneficial to reducing a service delay.

[0083] In another optional embodiment, load balancing weights corresponding to the plurality of service VPCs can be determined according to the number of service instances included in the plurality of service VPCs. For example, the more the number of service instances included in the service VPCs is, the greater the load balancing weights are; the less the number of service instances included in the service VPCs is, the smaller the load balancing weights are; and the target service instance is, in turn, determined from the service instances in the service VPCs according to the load balancing weights corresponding to the plurality of service VPCs.

[0084] In a further optional embodiment, load balancing weights corresponding to the plurality of service VPCs can be determined according to regions where the plurality of service VPCs are located and the number of service instances included. For example, initial load balancing weights corresponding to the plurality of service VPCs are determined according to a distance between the regions where the service VPCs are

located and a region where a client VPC is located; and then according to the number of service instances included in the plurality of service VPCs, the initial load balancing

where a first private network IP address of the target service is provided by a first virtual network interface card device, a first virtual network interface card device corresponding to the target service can also be deployed in each service VPC through the VPC network management and control node.

[0102] It should be noted that, an execution subject of respective steps of the methods provided in the above embodiments can be a same device, or different devices are used as an execution subject of the methods. For example, an execution subject of step 301a to step 302a can be device A; for another example, an execution subject of step 301a can be device A, and an execution subject of step 302a can be device B; and the like.

[0103] Additionally, in some flows described in the above embodiments and drawings, a plurality of operations that appear in a particular sequence are contained, but it should be clearly understood that these operations may not be executed in the sequence in which they appear herein, or executed in parallel. Sequence numbers of the operations, such as 301a and 302a, are only used to distinguish different operations, and the sequence numbers themselves do not represent any execution sequence. Additionally, these flows can include more or fewer operations, and these operations can be executed in sequence or in parallel. It should be noted that descriptions such as "first" and "second" herein are used to distinguish different messages, devices, modules, and the like, which do not represent a sequence, or limit "first" and "second" to be different types.

[0104] FIG. 4a is a schematic structural diagram of a load balancing apparatus provided in an exemplary embodiment of the present application, and this apparatus can be implemented as a multi-cluster load balancing node in a multi-cluster service system including at least one service VPC responsible for providing a target service and at least one client VPC which can access the target service, as shown in FIG. 4a, this apparatus including: a generation module 41a and a load balancing module 42a.

[0105] Generation module 41a is used for generating in advance an information mapping relationship among second service information which a target service has in each client VPC, first service information which the target service has in each service VPC, and information of a service instance in each service VPC which can provide the target service; and

the target service, and information of each client VPC which needs to access the target service, and second service information which the target service has in each client VPC.

[0134] In an optional embodiment, the processor is further used for: deploying, according to second service information which the target service has in each client VPC, by a VPC network management and control node, in each client VPC, an endpoint corresponding to the target service, that is, a second virtual network interface card device. Further, in a case where a first private network IP address of the target service is provided by a first virtual network interface card device, a first virtual network interface card device corresponding to the target service can also be deployed in each service VPC through the VPC network management and control node.

[0135] Further, as shown in FIG. 5, this node device further includes: other components such as a communication component 56, a display 57, a power supply component 58, and an audio component 59. Only some components are schematically shown in FIG. 5, which does not mean that the node device only includes the components shown in FIG 5. It should be noted that, the components indicated by the dashed lines in FIG. 5 are optional components, rather than mandatory components, and may be determined specifically depending on a product form of the node device.

[0136] Correspondingly, an embodiment of the present application further provides a computer-readable storage medium stored with a computer program, wherein the computer program, when executed by a processor, causes the processor to be capable of implementing steps in the methods shown in FIG. 3a and FIG. 3b.

[0137] Correspondingly, an embodiment of the present application further provides a computer program product including a computer program/instruction, wherein the computer program/instruction, when executed by a processor, causes the processor to implement steps in the methods shown in FIG. 3a and FIG. 3b.

[0138] The communication component in the above FIG. 5 is configured to facilitate wired or wireless communication between the device where the communication component is located and other devices. The device where the communication component is located can access a wireless network based on a communication standard, such as Wi-Fi, a mobile communication network such as 2G, 3G, 4G/LTE, or 5G, or a combination thereof. In one exemplary embodiment, the communication component receives, via a broadcast channel, a broadcast signal or

CLAIMS

What is claimed is:

1. A multi-cluster service system, comprising:

a plurality of service clusters for

providing services externally and distributed in a plurality of VPCs;

wherein the plurality of VPCs comprise at least one service VPC responsible for providing a target service and at least one client VPC which can access the target service, and the target service having first service information in each service VPC and second service information in each client VPC;

a multi-cluster load balancing node for

generating in advance an information mapping relationship among second service information which the target service has in each client VPC, first service information which the target service has in each service VPC, and information of a service instance in each service VPC which can provide the target service; and

load-balancing, based on the information mapping relationship, an access request for the target service from any client VPC onto a target service instance in the at least one service VPC, to enable the target service instance to provide the target service for a service instance corresponding to the access request.

2. The system according to claim 1, wherein a first virtual network interface card device is deployed in each service VPC as a traffic access point of a target service, the first virtual network interface card device has a first private network IP address, and first service information which the target service has in the service VPC comprises the first private network IP address and a port number of the target service.

3. The system according to claim 2, wherein a second virtual network interface card device bound to the target service is deployed in each client VPC as a traffic

routing point of the target service, the second virtual network interface card device has a second private network IP address, and second service information which the target service has in the client VPC comprises the second private network IP address and a port number or mapped port number of the target service.

4. The system according to claim 3, wherein the multi-cluster load balancing node is specifically used for:

receiving an access request sent by any service instance in any client VPC through access to a second virtual network interface card device in the client VPC, wherein the access request comprises identification information of the any client VPC and a second private network IP address;

determining, according to the identification information of the any client VPC and the second private network IP address, in conjunction with the information mapping relationship, the target service instance and a target private network IP address, wherein the target private network IP address is a first private network IP address which the target service has in a service VPC where the target service instance is located; and

sending, according to an IP address of the target service instance and the target private network IP address, the access request to the target service instance, to enable the target service instance to provide the target service for a service instance corresponding to the access request.

5. The system according to claim 1, wherein the multi-cluster load balancing node is located in a system VPC and is further used for:

deploying, in the system VPC, a third virtual network interface card device bound to the target service, wherein the third virtual network interface card has a third private network IP address, and third service information which the target service has in the system VPC comprises the third private network IP address and a port number of the target service; and

establishing, through the third service information which the target service has in the system VPC, an information mapping relationship among the second service information which the target service has in each client VPC, the first service

information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service.

6. The system according to any one of claims 1 to 5, further comprising:

a multi-cluster service management and control node for acquiring and delivering the second service information which the target service has in each client VPC, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service to the multiple-cluster load balancing node, to be used by the multiple-cluster load balancing node to locally generate the information mapping relationship.

7. The system according to claim 6, further comprising: a cluster management and control node, and a VPC network management and control node;

wherein the multi-cluster service management and control node is further used for: automatically discovering, in a case where the target service is a containerized service, based on a service exposure mechanism, a target service exposed externally by the at least one service VPC, and at least one client VPC which needs to access the target service, configuring the second service information which the target service has in each client VPC, and acquiring, by the cluster management and control node, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service; and

deploying, according to the second service information which the target service has in each client VPC, by the VPC network management and control node, in each client VPC, an endpoint corresponding to the target service.

8. The system according to claim 7, wherein the multi-cluster service management and control node is further used for:

receiving, in a case where the target service is a virtual machine-based service, service configuration information submitted by a user, wherein the service configuration information comprises information of at least one service VPC where the

target service is located, the first service information which the target service has in each service VPC, the information of the service instance in each service VPC which can provide the target service, and information of each client VPC which needs to access the target service, and the second service information which the target service has in each client VPC; and

deploying, according to the second service information which the target service has in each client VPC, by the VPC network management and control node, in each client VPC, an endpoint corresponding to the target service.

9. A service access method, applicable to a multi-cluster load balancing node in a multi-cluster service system comprising at least one service VPC responsible for providing a target service and at least one client VPC which can access the target service, the method comprising:

generating in advance an information mapping relationship among second service information which the target service has in each client VPC, first service information which the target service has in each service VPC, and information of a service instance in each service VPC which can provide the target service; and

load-balancing, based on the information mapping relationship, an access request for the target service from any client VPC onto a target service instance in the at least one service VPC, to enable the target service instance to provide the target service for a service instance corresponding to the access request.

10. The method according to claim 9, wherein load-balancing, based on the information mapping relationship, the access request for the target service from the any client VPC onto the target service instance in the at least one service VPC, to enable the target service instance to provide the target service for the service instance corresponding to the access request, comprises:

receiving an access request sent by any service instance in any client VPC through access to a second virtual network interface card device in the client VPC, wherein the access request comprises identification information of the any client VPC and a second private network IP address which the second virtual network interface

card device has;

determining, according to the identification information of the any client VPC and the second private network IP address, in conjunction with the information mapping relationship, the target service instance and a target private network IP address, wherein the target private network IP address is a first private network IP address which the target service has in a first virtual network interface card device in a service VPC where the target service instance is located; and

sending, according to an IP address of the target service instance and the target private network IP address, the access request to the target service instance, to enable the target service instance to provide the target service for a service instance corresponding to the access request.

11. The method according to claim 10, wherein sending, according to the IP address of the target service instance and the target private network IP address, the access request to the target service instance, to enable the target service instance to provide the target service for the service instance corresponding to the access request, comprises:

converting a source IP address and a destination IP address of the access request into the target private network IP address and the IP address of the target service instance, respectively; and

sending, via the first virtual network interface card device in the service VPC where the target service instance is located, the access request after address conversion to the target service instance, to enable the target service instance to provide the target service for the service instance corresponding to the access request.

12. The method according to claim 10, wherein the at least one service VPC comprises a plurality of service VPCs distributed in different regions, and determining the target service instance according to the identification information of the any client VPC and the second private network IP address, in conjunction with the information mapping relationship, comprises:

acquiring, according to the identification information of the any client VPC and

the second private network IP address in conjunction with the information mapping relationship, the plurality of service VPCs and information of a service instance therein which can provide the target service;

determining, according to regions where the plurality of service VPCs are located and/or the number of service instances contained which can provide the target service, load balancing weights corresponding to the plurality of service VPCs; and

determining, according to the load balancing weights corresponding to the plurality of service VPCs, the target service instance from service instances in the plurality of service VPCs which can provide the target service.

13. The method according to claim 12, wherein determining, according to the regions where the plurality of service VPCs are located, the load balancing weights corresponding to the plurality of service VPCs, comprises:

determining, according to a location relationship between regions where the plurality of service VPCs are located and a region where the multi-cluster load balancing node is located, the load balancing weights corresponding to the plurality of service VPCs;

wherein a load balancing weight of a service VPC located in a same region as the multi-cluster load balancing node is greater than a load balancing weight of a service VPC located in a different region from the multi-cluster load balancing node.

14. The method according to any one of claims 9 to 13, wherein generating in advance the information mapping relationship among the second service information which the target service has in each client VPC, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service, comprises:

deploying, in a system VPC where the multi-cluster load balancing node is located, a third virtual network interface card device bound to the target service, wherein the third virtual network interface card has a third private network IP address, and third service information which the target service has in the system VPC comprises the third private network IP address and a port number of the target service; and

establishing, through the third service information which the target service has in the system VPC, an information mapping relationship among the second service information which the target service has in each client VPC, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service.

15. An information configuration method, applicable to a multi-cluster service management and control node in a multi-cluster service system comprising at least one service VPC responsible for providing a target service and at least one client VPC which can access the target service, the method comprising:

acquiring and delivering second service information which the target service has in each client VPC, first service information which the target service has in each service VPC, and information of a service instance in each service VPC which can provide the target service to a multi-cluster load balancing node, to be used by the multi-cluster load balancing node to locally generate an information mapping relationship;

wherein the information mapping relationship is used for load balancing an access request for the target service from any client VPC onto a target service instance in the at least one service VPC.

16. The method according to claim 15, wherein acquiring the second service information which the target service has in each client VPC, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service, comprises:

automatically discovering, in a case where the target service is a containerized service, based on a service exposure mechanism, a target service exposed externally by the at least one service VPC, and at least one client VPC which needs to access the target service, configuring the second service information which the target service has in each client VPC, and acquiring, by a cluster management and control node, the first service information which the target service has in each service VPC, and the information of the service instance in each service VPC which can provide the target service;

or

receiving, in a case where the target service is a virtual machine-based service, service configuration information submitted by a user, wherein the service configuration information comprises information of at least one service VPC where the target service is located, the first service information which the target service has in each service VPC, the information of the service instance in each service VPC which can provide the target service, and information of each client VPC which needs to access the target service, and the second service information which the target service has in each client VPC.

17. The method according to claim 16, further comprising:

deploying, according to the second service information which the target service has in each client VPC, by a VPC network management and control node, in each client VPC, an endpoint corresponding to the target service.

18. A node device for use in a multi-cluster service system, comprising: a memory for storing a computer program, and a processor coupled to the memory for executing the computer program to cause the processor to implement steps in the method according to any one of claims 9 to 17.

19. A computer-readable storage medium stored with a computer program, wherein the computer program, when executed by a processor, causes the processor to implement steps in the method according to any one of claims 9 to 17.