

Cloud Container Engine

FAQs

Issue 01
Date 2024-07-05



Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2024. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Cloud Computing Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Contents

1 Common Questions.....	1
2 Billing.....	2
2.1 How Is CCE Billed/Charged?.....	2
2.2 How Do I Change the Billing Mode of a CCE Cluster from Pay-per-Use to Yearly/Monthly?.....	2
2.3 Can I Change the Billing Mode of CCE Nodes from Pay-per-Use to Yearly/Monthly?.....	3
2.4 Which Invoice Modes Are Supported by Huawei Cloud?.....	5
2.5 Will I Be Notified When My Balance Is Insufficient?.....	5
2.6 Will I Be Notified When My Account Balance Changes?.....	5
2.7 Can I Delete a Yearly/Monthly-Billed CCE Cluster Directly When It Expires?.....	5
2.8 How Do I Unsubscribe From CCE?.....	5
3 Cluster.....	8
3.1 Cluster Creation.....	8
3.1.1 Why Cannot I Create a CCE Cluster?.....	8
3.1.2 Is Management Scale of a Cluster Related to the Number of Master Nodes?.....	9
3.1.3 How Do I Update the Root Certificate When Creating a CCE Cluster?.....	9
3.1.4 Which Resource Quotas Should I Pay Attention To When Using CCE?.....	10
3.2 Cluster Running.....	11
3.2.1 How Do I Locate the Fault When a Cluster Is Unavailable?.....	11
3.2.2 How Do I Reset or Reinstall a CCE Cluster?.....	12
3.2.3 How Do I Check Whether a Cluster Is in Multi-Master Mode?.....	12
3.2.4 Can I Directly Connect to the Master Node of a CCE Cluster?.....	13
3.2.5 How Do I Retrieve Data After a CCE Cluster Is Deleted?.....	13
3.2.6 Why Does CCE Display Node Disk Usage Inconsistently with Cloud Eye?.....	13
3.2.7 How Do I Change the Name of a CCE Cluster?.....	13
3.3 Cluster Deletion.....	14
3.3.1 What Can I Do If a Cluster Deletion Fails Due to Residual Resources in the Security Group?.....	14
3.3.2 How Do I Clear Residual Resources After Deleting a Non-Running Cluster?.....	16
3.4 Cluster Upgrade.....	18
3.4.1 What Do I Do If a Cluster Add-On Fails to be Upgraded During the CCE Cluster Upgrade?.....	18
4 Node.....	21
4.1 Node Creation.....	21
4.1.1 How Do I Troubleshoot Problems Occurred When Adding Nodes to a CCE Cluster?.....	21

4.1.2 How Do I Troubleshoot Problems Occurred When Accepting Nodes into a CCE Cluster?.....	25
4.1.3 What Should I Do If a Node Fails to Be Accepted Because It Fails to Be Installed?.....	26
4.2 Node Running.....	27
4.2.1 What Should I Do If a Cluster Is Available But Some Nodes Are Unavailable?.....	27
4.2.2 How Do I Troubleshoot the Failure to Remotely Log In to a Node in a CCE Cluster?.....	33
4.2.3 How Do I Log In to a Node Using a Password and Reset the Password?.....	33
4.2.4 How Do I Collect Logs of Nodes in a CCE Cluster?.....	34
4.2.5 What Can I Do If the Container Network Becomes Unavailable After yum update Is Used to Upgrade the OS?.....	35
4.2.6 What Should I Do If the vdb Disk of a Node Is Damaged and the Node Cannot Be Recovered After Reset?.....	35
4.2.7 Which Ports Are Used to Install kubelet on CCE Cluster Nodes?.....	37
4.2.8 How Do I Configure a Pod to Use the Acceleration Capability of a GPU Node?.....	37
4.2.9 What Should I Do If I/O Suspension Occasionally Occurs When SCSI EVS Disks Are Used?.....	38
4.2.10 What Should I Do If Excessive Docker Audit Logs Affect the Disk I/O?.....	39
4.2.11 How Do I Fix an Abnormal Container or Node Due to No Thin Pool Disk Space?.....	40
4.2.12 Where Can I Get the Listening Ports of CCE Worker Nodes?.....	43
4.2.13 How Do I Rectify Failures When the NVIDIA Driver Is Used to Start Containers on GPU Nodes?....	45
4.2.14 What Can I Do If the Time of CCE Nodes Is Not Synchronized with the NTP Server?.....	46
4.2.15 What Should I Do If the Data Disk Usage Is High Because a Large Volume of Data Is Written Into the Log File?.....	47
4.2.16 Why Does My Node Memory Usage Obtained by Running the kubelet top node Command Exceeds 100%?.....	47
4.3 Specification Change.....	49
4.3.1 How Do I Change the Node Specifications in a CCE Cluster?.....	49
4.3.2 What Should I Do If I Fail to Restart or Create Workloads on a Node After Modifying the Node Specifications?.....	50
4.3.3 Can I Change the IP Address of a Node in a CCE Cluster?.....	50
4.4 OSs.....	51
4.4.1 What Can I Do If cgroup kmem Leakage Occasionally Occurs When an Application Is Repeatedly Created or Deleted on a Node Running CentOS with an Earlier Kernel Version?.....	51
4.4.2 What Should I Do If There Is a Service Access Failure After a Backend Service Upgrade or a 1-Second Latency When a Service Accesses a CCE Cluster?.....	52
4.4.3 Why Are Pods Evicted by kubelet Due to Abnormal cgroup Statistics?.....	55
4.4.4 When Container OOM Occurs on the CentOS Node with an Earlier Kernel Version, the Ext4 File System Is Occasionally Suspended.....	56
4.4.5 What Should I Do If a DNS Resolution Failure Occurs Due to a Defect in IPVS?.....	57
4.4.6 What Should I Do If the Number of ARP Entries Exceeds the Upper Limit?.....	57
5 Node Pool.....	60
5.1 What Should I Do If a Node Pool Is Abnormal?.....	60
5.2 What Should I Do If No Node Creation Record Is Displayed When the Node Pool Is Being Expanding?.....	61
5.3 What Should I Do If a Node Pool Scale-Out Fails?.....	62
5.4 What Should I Do If Some Kubernetes Events Fail to Display After Nodes Were Added to or Deleted from a Node Pool in Batches?.....	63

5.5 How Do I Modify ECS Configurations When an ECS Cannot Be Managed by a Node Pool?.....	64
6 Workload.....	69
6.1 Workload Abnormalities.....	69
6.1.1 How Do I Use Events to Fix Abnormal Workloads?.....	69
6.1.2 What Should I Do If Pod Scheduling Fails?.....	71
6.1.3 What Should I Do If a Pod Fails to Pull the Image?.....	80
6.1.4 What Should I Do If Container Startup Fails?.....	88
6.1.5 What Should I Do If a Pod Fails to Be Evicted?.....	97
6.1.6 What Should I Do If a Storage Volume Cannot Be Mounted or the Mounting Times Out?.....	101
6.1.7 What Should I Do If a Workload Remains in the Creating State?.....	103
6.1.8 What Should I Do If Pods in the Terminating State Cannot Be Deleted?.....	105
6.1.9 What Should I Do If a Workload Is Stopped Caused by Pod Deletion?.....	105
6.1.10 What Should I Do If an Error Occurs When Deploying a Service on the GPU Node?.....	106
6.1.11 Why Does Pod Fail to Write Data?.....	107
6.1.12 Why Is Pod Creation or Deletion Suspended on a Node Where File Storage Is Mounted?.....	108
6.1.13 How to Locate Faults Using an Exit Code?.....	109
6.2 Container Configuration.....	112
6.2.1 When Is Pre-stop Processing Used?.....	113
6.2.2 How Do I Set an FQDN for Accessing a Specified Container in the Same Namespace?.....	113
6.2.3 What Should I Do If Health Check Probes Occasionally Fail?.....	113
6.2.4 How Do I Set the umask Value for a Container?.....	114
6.2.5 What Is the Retry Mechanism When CCE Fails to Start a Pod?.....	114
6.3 Alarm Monitoring.....	115
6.3.1 How Long Are the Events of a Workload Stored?.....	115
6.4 Scheduling Policies.....	115
6.4.1 How Do I Evenly Distribute Multiple Pods to Each Node?.....	115
6.4.2 How Do I Prevent a Container on a Node from Being Evicted?.....	116
6.4.3 Why Are Pods Not Evenly Distributed to Nodes?.....	117
6.4.4 How Do I Evict All Pods on a Node?.....	117
6.4.5 How Do I Check Whether a Pod Is Bound with CPU Cores?.....	118
6.4.6 What Should I Do If Pods cannot Be Rescheduled After the Node Is Stopped?.....	119
6.4.7 How Do I Prevent a Non-GPU or NPU Workload from Being Scheduled to a GPU or NPU Node?..	120
6.4.8 Why Cannot a Pod Be Scheduled to a Node?.....	122
6.5 Others.....	122
6.5.1 What Should I Do If a Scheduled Task Cannot Be Restarted After Being Stopped for a Period of Time?.....	122
6.5.2 What Is a Headless Service When I Create a StatefulSet?.....	123
6.5.3 What Should I Do If Error Message "Auth is empty" Is Displayed When a Private Image Is Pulled?.....	124
6.5.4 What Is the Image Pull Policy for Containers in a CCE Cluster?.....	124
6.5.5 Why Is the Mount Point of a Docker Container in the Kunpeng Cluster Uninstalled?.....	125
6.5.6 What Can I Do If a Layer Is Missing During Image Pull?.....	125
6.5.7 Why the File Permission and User in the Container Are Question Marks?.....	126

7 Networking	129
7.1 Network Planning	129
7.1.1 What Is the Relationship Between Clusters, VPCs, and Subnets?	129
7.1.2 How Do I View the VPC CIDR Block?	130
7.1.3 How Do I Set the VPC CIDR Block and Subnet CIDR Block for a CCE Cluster?	130
7.1.4 How Do I Set a Container CIDR Block for a CCE Cluster?	131
7.1.5 When Should I Use Cloud Native Network 2.0?	132
7.1.6 What Is an ENI?	133
7.1.7 How Can I Configure a Security Group Rule in a Cluster?	134
7.1.8 How Do I Configure the IPv6 Service CIDR Block When Creating a CCE Turbo Cluster?	145
7.1.9 Can Multiple NICs Be Bound to a Node in a CCE Cluster?	147
7.2 Network Fault	147
7.2.1 How Do I Locate a Workload Networking Fault?	147
7.2.2 Why the ELB Address Cannot Be used to Access Workloads in a Cluster?	150
7.2.3 Why the Ingress Cannot Be Accessed Outside the Cluster?	157
7.2.4 Why Does the Browser Return Error Code 404 When I Access a Deployed Application?	163
7.2.5 What Should I Do If a Container Fails to Access the Internet?	164
7.2.6 What Can I Do If a VPC Subnet Cannot Be Deleted?	165
7.2.7 How Do I Restore a Faulty Container NIC?	165
7.2.8 What Should I Do If a Node Fails to Connect to the Internet (Public Network)?	166
7.2.9 How Do I Resolve a Conflict Between the VPC CIDR Block and the Container CIDR Block?	166
7.2.10 What Should I Do If the Java Error "Connection reset by peer" Is Reported During Layer-4 ELB Health Check?	167
7.2.11 How Do I Locate the Service Event Indicating That No Node Is Available for Binding?	168
7.2.12 Why Does "Dead loop on virtual device gw_11cbf51a, fix it urgently" Intermittently Occur When I Log In to a VM using VNC?	168
7.2.13 Why Does a Panic Occasionally Occur When I Use Network Policies on a Cluster Node?	169
7.2.14 Why Are Lots of source ip_type Logs Generated on the VNC?	171
7.2.15 What Should I Do If Status Code 308 Is Displayed When the Nginx Ingress Controller Is Accessed Using the Internet Explorer?	172
7.2.16 What Should I Do If an Nginx Ingress Access in the Cluster Is Abnormal After the Add-on Is Upgraded?	173
7.3 Security Hardening	174
7.3.1 How Do I Prevent Cluster Nodes from Being Exposed to Public Networks?	174
7.3.2 How Do I Configure an Access Policy for a Cluster?	175
7.3.3 How Do I Obtain a TLS Key Certificate?	176
7.3.4 How Do I Change the Security Group of Nodes in a Cluster in Batches?	178
7.4 Network Configuration	178
7.4.1 How Does CCE Communicate with Other Huawei Cloud Services over an Intranet?	179
7.4.2 How Do I Set the Port When Configuring the Workload Access Mode on CCE?	179
7.4.3 How Can I Achieve Compatibility Between Ingress's property and Kubernetes client-go?	183
7.4.4 How Do I Obtain the Actual Source IP Address of a Client After a Service Is Added into Istio?	185
7.4.5 Why Cannot an Ingress Be Created After the Namespace Is Changed?	187

7.4.6 Why Is the Backend Server Group of an ELB Automatically Deleted After a Service Is Published to the ELB?.....	187
7.4.7 How Can Container IP Addresses Survive a Container Restart?.....	187
8 Storage.....	189
8.1 How Do I Expand the Storage Capacity of a Container?.....	189
8.2 What Are the Differences Among CCE Storage Classes in Terms of Persistent Storage and Multi-node Mounting?.....	190
8.3 Can I Create a CCE Node Without Adding a Data Disk to the Node?.....	191
8.4 Can EVS Volumes in a CCE Cluster Be Restored After They Are Deleted or Expired?.....	192
8.5 What Should I Do If the Host Cannot Be Found When Files Need to Be Uploaded to OBS During the Access to the CCE Service from a Public Network?.....	192
8.6 How Can I Achieve Compatibility Between ExtendPathMode and Kubernetes client-go?.....	193
8.7 What Can I Do If a Storage Volume Fails to Be Created?.....	195
8.8 Can CCE PVCs Detect Underlying Storage Faults?.....	196
8.9 An Error Is Reported When the Owner Group and Permissions of the Mount Point of the SFS 3.0 File System in the OS Are Modified.....	196
8.10 Why Cannot I Delete a PV or PVC Using the kubectl delete Command?.....	196
8.11 What Should I Do If a Yearly/Monthly EVS Disk Cannot Be Automatically Created?.....	197
9 Namespace.....	199
9.1 What Should I Do If a Namespace Fails to Be Deleted Due to an APIService Object Access Failure?.....	199
9.2 How Do I Delete a Namespace in the Terminating State?.....	200
10 Chart and Add-on.....	203
10.1 What Should I Do If the nginx-ingress Add-on Fails to Be Installed on a Cluster and Remains in the Creating State?.....	203
10.2 What Should I Do If Residual Process Resources Exist Due to an Earlier npd Add-on Version?.....	204
10.3 What Should I Do If a Chart Release Cannot Be Deleted Because the Chart Format Is Incorrect?.....	205
10.4 Does CCE Support nginx-ingress?.....	207
10.5 What Should I Do If Installation of an Add-on Fails and "The release name is already exist" Is Displayed?.....	207
10.6 What Should I Do If a Release Creation or Upgrade Fails and "rendered manifests contain a resource that already exists" Is Displayed?.....	209
10.7 What Can I Do If the kube-prometheus-stack Add-on Instance Fails to Be Scheduled?.....	210
10.8 What Can I Do If a Chart Fails to Be Uploaded?.....	212
10.9 How Do I Configure the Add-on Resource Quotas Based on Cluster Scale?.....	214
10.10 How Can I Clean Up Residual Resources After the NGINX Ingress Controller Add-on in the Unknown State Is Deleted?.....	217
10.11 Why TLS v1.0 and v1.1 Cannot Be Used After the NGINX Ingress Controller Add-on Is Upgraded?.....	218
11 API & kubectl FAQs.....	220
11.1 How Can I Access a Cluster API Server?.....	220
11.2 Can the Resources Created Using APIs or kubectl Be Displayed on the CCE Console?.....	220
11.3 How Do I Download kubeconfig for Connecting to a Cluster Using kubectl?.....	221
11.4 How Do I Rectify the Error Reported When Running the kubectl top node Command?.....	221

11.5 Why Is "Error from server (Forbidden)" Displayed When I Use kubectl?.....	222
12 DNS FAQs.....	224
12.1 What Should I Do If Domain Name Resolution Fails in a CCE Cluster?.....	224
12.2 Why Does a Container in a CCE Cluster Fail to Perform DNS Resolution?.....	227
12.3 Why Cannot the Domain Name of the Tenant Zone Be Resolved After the Subnet DNS Configuration Is Modified?.....	228
12.4 How Do I Optimize the Configuration If the External Domain Name Resolution Is Slow or Times Out?.....	228
12.5 How Do I Configure a DNS Policy for a Container?.....	229
13 Image Repository FAQs.....	231
13.1 How Do I Create a Docker Image and Solve the Problem of Slow Image Pull?.....	231
13.2 How Do I Upload My Images to CCE?.....	231
14 Permissions.....	232
14.1 Can I Configure Only Namespace Permissions Without Cluster Management Permissions?.....	232
14.2 Can I Use CCE APIs If the Cluster Management Permissions Are Not Configured?.....	232
14.3 Can I Use kubectl If the Cluster Management Permissions Are Not Configured?.....	233
15 Related Services.....	234
15.1 What Are the Differences Between CCE and CCI?.....	235
15.2 What Are the Differences Between CCE and ServiceStage?.....	238

1 Common Questions

Cluster Management

- [Why Cannot I Create a CCE Cluster?](#)
- [Is Management Scale of a Cluster Related to the Number of Master Nodes?](#)
- [How Do I Locate the Fault When a Cluster Is Unavailable?](#)

Node/Node Pool Management

- [What Should I Do If a Cluster Is Available But Some Nodes Are Unavailable?](#)
- [What Should I Do If a Node Fails to Be Accepted Because It Fails to Be Installed?](#)
- [What Should I Do If I/O Suspension Occasionally Occurs When SCSI EVS Disks Are Used?](#)

Workload Management

- [What Should I Do If Pod Scheduling Fails?](#)
- [What Should I Do If a Pod Fails to Pull the Image?](#)
- [What Should I Do If Container Startup Fails?](#)
- [What Should I Do If Pods in the Terminating State Cannot Be Deleted?](#)
- [What Is the Image Pull Policy for Containers in a CCE Cluster?](#)

Networking

[Why Does the Browser Return Error Code 404 When I Access a Deployed Application?](#)

[What Should I Do If a Node Fails to Connect to the Internet \(Public Network\)?](#)

[How Do I Optimize the Configuration If the External Domain Name Resolution Is Slow or Times Out?](#)

2 Billing

2.1 How Is CCE Billed/Charged?

Billing Modes

There are yearly/monthly and pay-per-use billing modes to meet your requirements. For details, see [Billing Modes](#).

- Yearly/Monthly is a prepaid billing mode. You pay in advance for a subscription term. Before purchasing yearly/monthly resources, ensure that your account has sufficient balance.
- Pay-per-use is a postpaid billing mode. You pay as you go and just pay for what you use.

After purchasing CCE clusters or cluster resources, you can change their billing modes if the current billing mode cannot meet your service requirements. For details, see [Billing Mode Changes](#).

Billing Items

You will be billed for clusters, nodes, and other cloud service resources. For details about the billing factors and formulas for each billed item, see [Billing Items](#).

For more information about the billing samples and the billing for each item, see [Billing Examples](#).

2.2 How Do I Change the Billing Mode of a CCE Cluster from Pay-per-Use to Yearly/Monthly?

Currently, clusters support **pay-per-use** and **yearly/monthly** billing modes. A pay-per-use cluster can be converted to a yearly/monthly-billed cluster.

Constraints and Limitations

- Only nodes in the default node pool **DefaultPool** can be changed to the yearly/monthly billing mode.

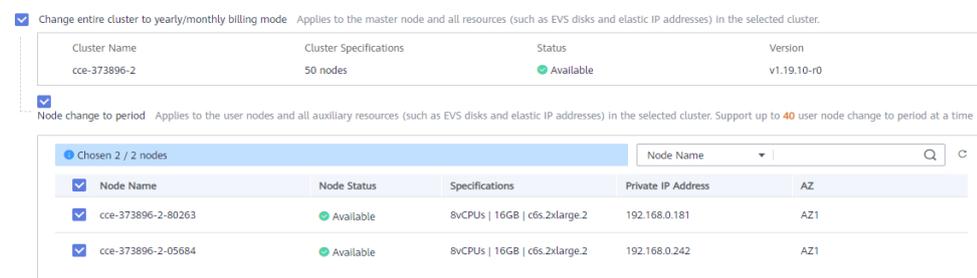
- Nodes whose billing mode is changed to yearly/monthly do not support auto scaling.

Changing to Yearly/Monthly Billing Mode

To change the billing mode of the clusters you have purchased from pay-per-use to yearly/monthly, perform the following steps:

- Step 1** Log in to the CCE console. In the navigation pane, choose **Clusters**.
- Step 2** Locate the target cluster, and click **Change Billing Mode** in the upper right corner.
- Step 3** On the **Change Billing Mode** page, select the target cluster. You can also select the nodes whose billing mode you want to change.

Figure 2-1 Changing the billing mode of a cluster to yearly/monthly



- Step 4** Click **OK**. Wait until the order is processed and the payment is complete.

----End

2.3 Can I Change the Billing Mode of CCE Nodes from Pay-per-Use to Yearly/Monthly?

Currently, nodes support **pay-per-use** and **yearly/monthly** billing modes.

Constraints and Limitations

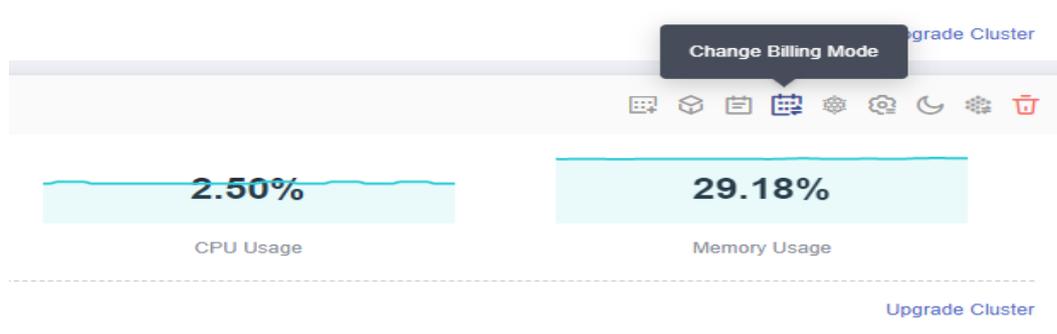
- You cannot change the nodes from pay-per-use to yearly/monthly on the ECS console.
- Only nodes in the default node pool **DefaultPool** can be changed to the yearly/monthly billing mode.
- Nodes whose billing mode is changed to yearly/monthly do not support auto scaling.

Procedure

To change the billing mode of the nodes you have purchased from pay-per-use to yearly/monthly, perform the following steps:

- Step 1** Log in to the CCE console. In the navigation pane, choose **Clusters**.
- Step 2** Click  next to the target cluster.

Figure 2-2 Changing to the yearly/monthly billing mode

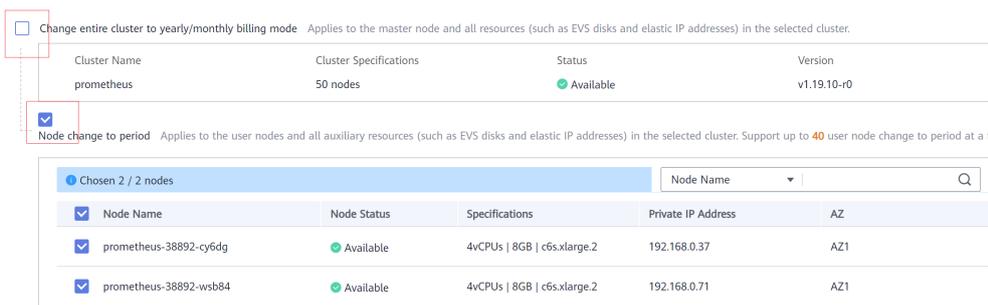


Step 3 On the **Change Billing Mode** page, choose the nodes that will be changed to the yearly/monthly billing mode.

NOTE

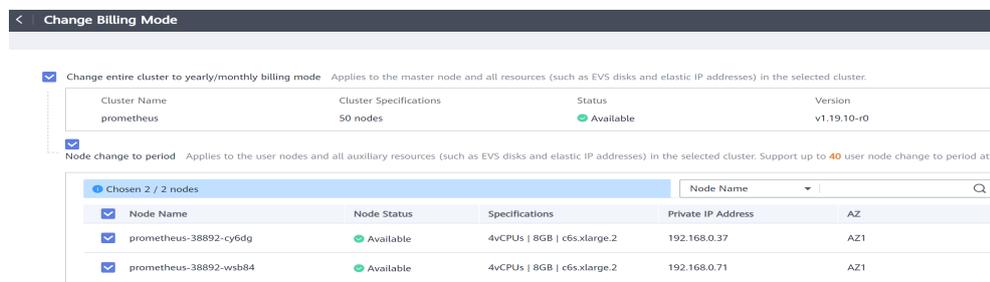
By default, the **Change entire cluster to yearly/monthly billing mode** is selected. If you want to change certain nodes of the cluster to the yearly/monthly billing mode, deselect this option.

Figure 2-3 Changing nodes to the yearly/monthly billing mode



If you want to change the whole cluster to the yearly/monthly billing mode, select this option and the nodes that need to be changed to the yearly/monthly billing mode.

Figure 2-4 Changing the cluster and nodes of the cluster to the yearly/monthly billing mode



Step 4 Click **OK**. Wait until the order is processed and the payment is complete.

----End

2.4 Which Invoice Modes Are Supported by Huawei Cloud?

Huawei Cloud supports two issuing invoices by billing cycle and by order.

You can issue invoices on the [Invoices](#) page in **Billing Center**.

2.5 Will I Be Notified When My Balance Is Insufficient?

In Billing Center, on the [Overview](#) page, click **Settings** in the **Available Credit** area. Then, you can enable or disable **Balance Alert**. Click **Modify** and configure a desired threshold.

- After the function is enabled, if your total balance (including cash balance, credit balance, common vouchers, and flexi-purchase coupons) is lower than the alert threshold, the system sends you SMS and email notifications every day for a maximum of three consecutive days.
- In Message Center, choose **Message Receiving Management > SMS & Email Settings** in the navigation pane, select **Account balance** under **Finance** to change contacts that receive the balance alerts.

2.6 Will I Be Notified When My Account Balance Changes?

The system will notify you via email or SMS message of your account balance changes, including whether your online topping up is successful.

2.7 Can I Delete a Yearly/Monthly-Billed CCE Cluster Directly When It Expires?

After a yearly/monthly-billed cluster expires, you can delete the cluster after all data is backed up.

If you do not renew or delete the cluster after it expires, the system will delete the cluster based on the resource expiration time. You are advised to renew the cluster and back up data in a timely manner.

2.8 How Do I Unsubscribe From CCE?

Yearly/monthly-billed CCE resources can be unsubscribed from, including the renewed part and currently used part. You cannot use these resources after unsubscription. A handling fee will be charged for unsubscribing from a resource.

Note

- Unsubscribing from CCE involves the renewed resources and the resources that are being used. After the unsubscription, these resources become unavailable.
- Solution product portfolios can only be unsubscribed from as a whole.
- If an order contains resources in a primary-secondary relationship, you need to unsubscribe from the resources separately.
- For details about unsubscribing from resources, see [Unsubscription Rules](#).

Procedure

CAUTION

- Before requesting an unsubscription, ensure that you have migrated or backed up any data saved on CCE that will be unsubscribed from. After the unsubscription is complete, CCE and any data it contains will be permanently deleted.
- The middle of the unsubscription page displays a message showing the number of unsubscriptions you have performed and the remaining allowed number.

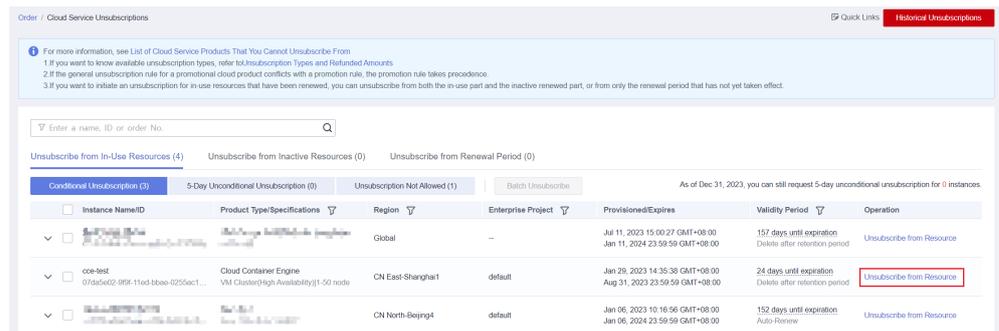
Step 1 Enter the [Unsubscriptions](#) page.

Step 2 Click the **Unsubscribe from In-Use Resources** tab.

Step 3 Unsubscribe from a single resource or from resources in a batch.

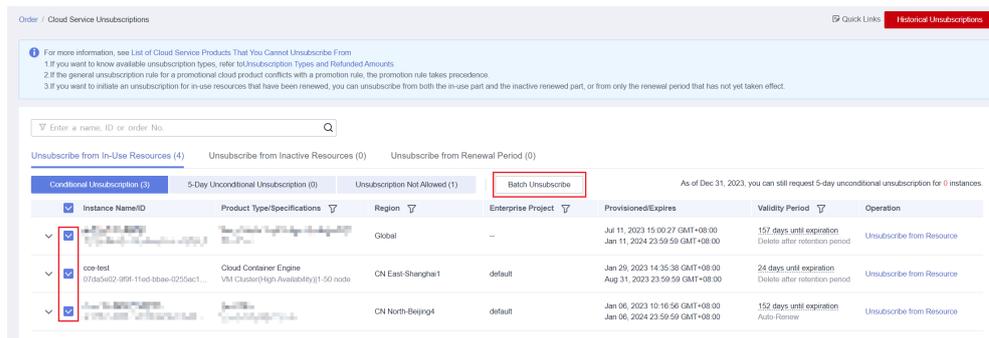
- To unsubscribe from a single resource, click **Unsubscribe from Resource** at the row of the target resource.

Figure 2-5 Unsubscribing from a single resource



- To unsubscribe from resources in a batch, select the target resources from the list and click **Batch Unsubscribe** above the list.

Figure 2-6 Unsubscribing from resources in a batch



Step 4 On the **Unsubscribe from In-Use Resources** page, confirm the information, select a reason for the unsubscription, and click **Confirm**.

----End

3 Cluster

3.1 Cluster Creation

3.1.1 Why Cannot I Create a CCE Cluster?

Overview

This section describes how to locate and rectify the fault if you fail to create a CCE cluster.

Details

Possible causes:

1. The Network Time Protocol daemon (ntpd) is not installed or fails to be installed, Kubernetes components fail to pass the pre-verification, or the disk partition is incorrect. The current solution is to create a cluster again. For details about how to locate the fault, see [Locating the Failure Cause](#).
2. Check whether your account is in arrears. If so, you cannot purchase resources, including using cash coupons. For details, see and [Topping Up an Account \(Prepaid Direct Customers\)](#).

Locating the Failure Cause

View the cluster logs to identify the cause and rectify the fault.

Step 1 Log in to the CCE console. In the navigation pane, click **Operation Records** above the cluster list to view operation records.

Step 2 Click the record of the **Failed** status to view error information.

Figure 3-1 Viewing the operation details

Operation Records

All Actions Failed

Cluster Name	Operation Type	Status	Time
r30027646-new	Create Cluster	Failed	May 07, 2022 11:39:40 GMT+08:00

Project	Start Time	End Time	Status
Create	May 07, 2022 11:39:41 GMT+08:00	May 07, 2022 11:39:41 GMT+08:00	Completed
Security	May 07, 2022 11:39:41 GMT+08:00	May 07, 2022 11:39:41 GMT+08:00	Completed
Group	May 07, 2022 11:39:41 GMT+08:00	May 07, 2022 11:39:46 GMT+08:00	Completed
Create	May 07, 2022 11:39:41 GMT+08:00	May 07, 2022 11:39:41 GMT+08:00	Completed
Network	May 07, 2022 11:39:41 GMT+08:00	May 07, 2022 11:39:44 GMT+08:00	Completed
Create Cluster	Create master node (5 minutes)[1/3]	May 07, 2022 11:39:44 GMT+08:00	Failed

Expected HTTP response code [200 201 202 203 204] when accessing [POST https://ecs-internal.cn-north-7.myhuaweicloud.com/v1/0524ea9c1a00d57e2fddc0190fc7dd97/cloudservers], but got 400 instead {"error": {"message": "The volume type[SSD] cannot be used with the specified flavor in the AZ [cn-north-7b].", "code": "Ecs.0044"}}

Step 3 Rectify the fault based on the error information and create a cluster again.

----End

3.1.2 Is Management Scale of a Cluster Related to the Number of Master Nodes?

Management scale indicates the maximum number of nodes that can be managed by a cluster. If you select **50 nodes**, the cluster can manage a maximum of 50 nodes.

The number of master nodes varies according to the cluster specification, but is not affected by the management scale.

After the multi-master node mode is enabled, three master nodes will be created. If one of them is faulty, the cluster can still run properly. The services will not be affected.

3.1.3 How Do I Update the Root Certificate When Creating a CCE Cluster?

The root certificate of CCE clusters is the basic certificate for Kubernetes authentication. Both the Kubernetes cluster control plane and the certificate are hosted on Huawei Cloud CCE. CCE will periodically update the certificate. This certificate is not open to users but will not expire.

The X.509 certificate is enabled on Kubernetes clusters by default. CCE will automatically maintain and update the X.509 certificate.

Obtaining a Cluster Certificate

You can obtain a cluster certificate on the CCE console to access Kubernetes. For details, see [Obtaining a Cluster Certificate](#).

3.1.4 Which Resource Quotas Should I Pay Attention To When Using CCE?

CCE restricts **only the number of clusters**. However, when using CCE, you may also be using other cloud services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Virtual Private Cloud (VPC), Elastic Load Balance (ELB), and SoftWare Repository for Containers (SWR).

What Is Quota?

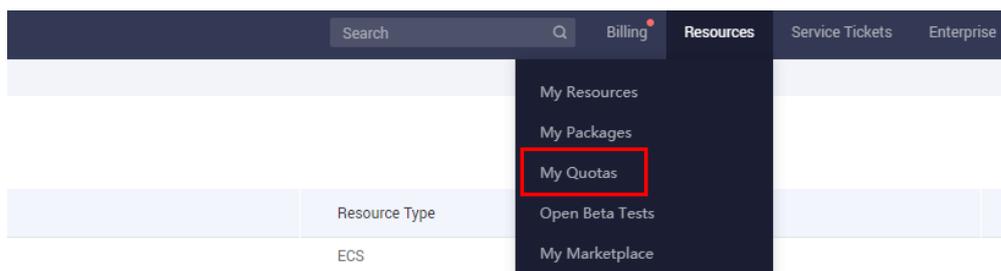
Quotas can limit the number or amount of resources available to users, such as the maximum number of ECSs or EVS disks that can be created.

If the existing resource quota cannot meet your service requirements, you can apply for a higher quota.

How Do I View My Quota?

1. Log in to the management console.
2. Click  in the upper left corner to select a region and a project.
3. In the upper right corner of the page, choose **Resources > My Quotas**.
The **Service Quota** page is displayed.

Figure 3-2 My Quotas

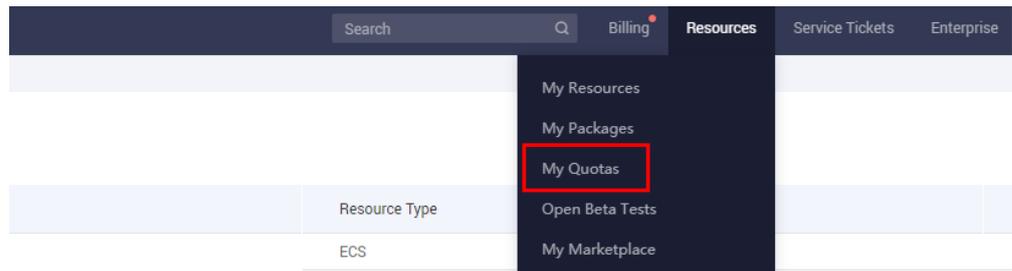


4. On this page, you can view the total quota and used quota of resources.
If a quota cannot meet your business requirements, click **Increase Quota**.

How Do I Increase My Quota?

1. Log in to the management console.
2. In the upper right corner of the page, choose **Resources > My Quotas**.
The **Service Quota** page is displayed.

Figure 3-3 My Quotas



3. Click **Increase Quota**.
4. On the **Create Service Ticket** page, configure parameters as required and submit a service ticket.
In the **Problem Description** area, enter the required quota and reason for the adjustment.
5. Select **I have read and agree to the Tenant Authorization Letter** and click **Submit**.

3.2 Cluster Running

3.2.1 How Do I Locate the Fault When a Cluster Is Unavailable?

If a cluster is **Unavailable**, perform the following operations to locate the fault.

Troubleshooting Process

The issues here are described in order of how likely they are to occur.

Check these causes one by one until you find the cause of the fault.

- **Check Item 1: Whether the Security Group Is Modified**
- **Check Item 2: Whether There Are Residual Listeners and Backend Server Groups on the Load Balancer**

If the fault persists, **submit a service ticket** and contact the customer service to help you locate the fault.

Figure 3-4 Fault locating



Check Item 1: Whether the Security Group Is Modified

- Step 1** Log in to the management console, and choose **Service List > Networking > Virtual Private Cloud**. In the navigation pane on the left, choose **Access Control > Security Groups** to find the security group of the master node in the cluster.

The name of this security group is in the format of *Cluster name-cce-control-ID*.

Step 2 Click the security group. On the details page displayed, ensure that the security group rules of the master node are correct.

For details, see [How Can I Configure a Security Group Rule in a Cluster?](#).

----End

Check Item 2: Whether There Are Residual Listeners and Backend Server Groups on the Load Balancer

Reproducing the Problem

A cluster exception occurs when a LoadBalancer Service is being created or deleted. After the fault is rectified, the Service is deleted successfully, but there are residual listeners and backend server group.

Step 1 Pre-create a CCE cluster. In the cluster, use the official Nginx image to create workloads, preset load balancers, Services, and ingresses.

Step 2 Ensure that the cluster is running properly and the Nginx workload is stable.

Step 3 Create and delete 10 LoadBalancer Services every 20 seconds.

Step 4 An injection exception occurs in the cluster. For example, the etcd pod is unavailable or the cluster is hibernated.

----End

Possible Causes

There are residual listeners and backend server groups on the load balancer.

Solution

Manually clear residual listeners and backend server groups.

Step 1 Log in to the management console and choose **Network > Elastic Load Balance** from the service list.

Step 2 In the load balancer list, click the name of the target load balancer to go to the details page. On the **Listeners** tab page, locate the target listener and delete it.

Step 3 On the **Backend Server Groups** tab page, locate the target backend server group and delete it.

----End

3.2.2 How Do I Reset or Reinstall a CCE Cluster?

CCE clusters cannot be reset or reinstalled. If a cluster becomes unavailable, [submit a service ticket](#) or delete the cluster and purchase a new one.

CCE supports resetting nodes. For details, see [Resetting a Node](#).

3.2.3 How Do I Check Whether a Cluster Is in Multi-Master Mode?

Log in to the CCE console and click the cluster. On the right of the cluster details page, view the number of master nodes.

- 3: The cluster is in multi-master mode.
- 1: The cluster is in single-master mode.

NOTICE

The number of master nodes cannot be changed after the cluster is created. If you want to adjust the number, you need to create a new cluster.

3.2.4 Can I Directly Connect to the Master Node of a CCE Cluster?

CCE allows you to use kubectl to connect a cluster. For details, see [Connecting to a Cluster Using kubectl](#).

However, you are not allowed to log in to the master node to perform related operations.

3.2.5 How Do I Retrieve Data After a CCE Cluster Is Deleted?

After a cluster is deleted, the workload on the cluster will also be deleted and cannot be restored. Therefore, exercise caution when deleting a cluster.

3.2.6 Why Does CCE Display Node Disk Usage Inconsistently with Cloud Eye?

Symptom

The disk usage of a node on the CCE cluster details page is higher than 80%, but the disk usage displayed on the Cloud Eye console is lower than 40%.

After fault locating on the node, it is found that the usage of a PVC disk reaches 92%. After the disk is cleared, the disk usage on CCE is the same as that on Cloud Eye.

Does CCE display only the highest disk usage?

Answer

In the CCE cluster monitoring information, the disk with the highest disk usage on the node is monitored.

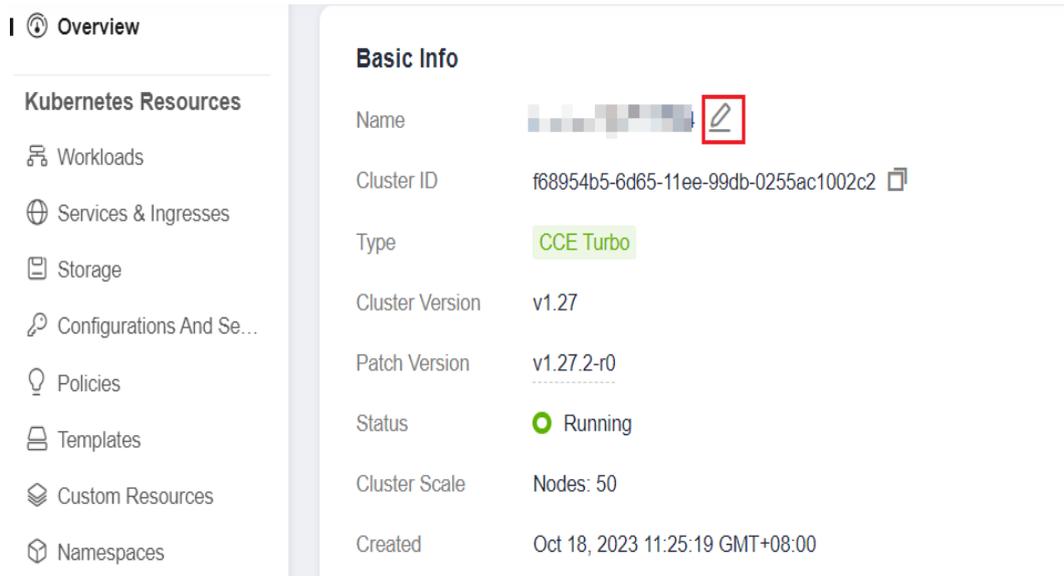
3.2.7 How Do I Change the Name of a CCE Cluster?

After a cluster is created, you can change its name.

Step 1 Log in to the CCE console and click the cluster name to access the cluster console.

Step 2 On the **Overview** page, click  next to **Name** in the **Basic Info** area.

Figure 3-5 Changing the cluster name



Step 3 Enter a new name and click **Save**.

----**End**

NOTICE

- The new name cannot be the same as its original name or the name of another cluster.
- If the related service logs use the original cluster name to name instances or configuration items, the name of the instances or configuration items will not be changed simultaneously. For example, the original cluster name will still be used for cluster log collection.

3.3 Cluster Deletion

3.3.1 What Can I Do If a Cluster Deletion Fails Due to Residual Resources in the Security Group?

When deleting a cluster, CCE obtains the cluster's resources, such as the elastic network interfaces (ENIs) or sub ENIs bound to a CCE Turbo cluster through kube-apiserver of the cluster. If the cluster is unavailable, frozen, or hibernated, the resources may fail to be obtained, and the cluster may not be deleted.

Symptom

The cluster cannot be deleted, and the following error information is displayed:

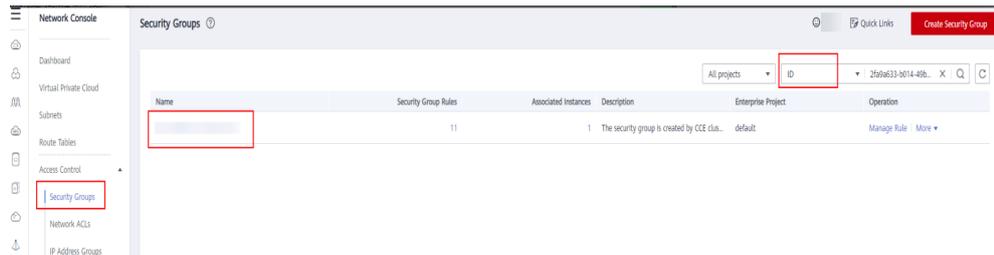
```
Expected HTTP response code [200 202 204 404] when accessing [DELETE https://vpc.xxx.com/v2.0/security-groups/46311976-7743-4c7c-8249-ccd293bcae91], but got 409 instead  
{  
  "code": "VPC.0602",  
  "message": "{\n    \"NeutronError\": {\n      \"message\": \"Security Group  
46311976-7743-4c7c-8249-ccd293bcae91 in use.\",  
      \"type\": \"SecurityGroupInUse\",  
      \"detail\": \"\"  
    }  
  }  
}
```

Possible Causes

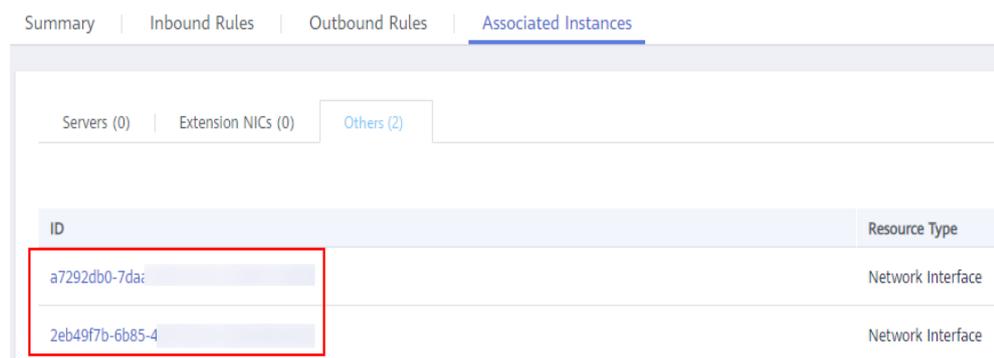
The cluster's security group has undeleted resources, preventing its deletion and causing the creation of the cluster to fail.

Procedure

- Step 1** Copy the resource ID in the error information, go to the **Security Groups** page of the VPC console, and obtain security groups by ID.



- Step 2** Click the security group to view its details, and click the **Associated Instances** tab.

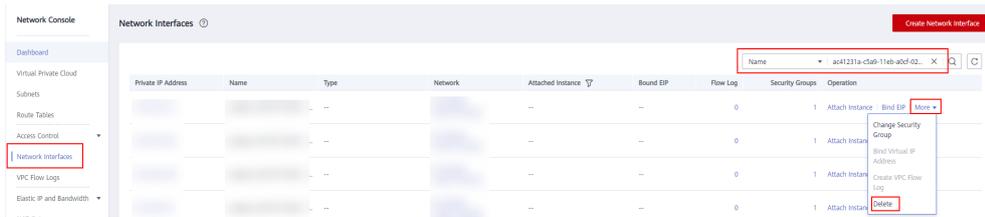


Obtain other resources associated with the security group, such as servers, ENIs, and sub-ENIs. You can delete residual resources. The sub ENIs will be automatically deleted.



- Step 3** For a residual ENI, go to the **Network Interfaces** page and delete the ENI obtained in the previous step.

You can search for the ENIs to be deleted by IDs or names.



Step 4 Go to the **Security Groups** page to confirm that the security group is not associated with any instance. Then, go to the CCE console to delete the cluster.

----End

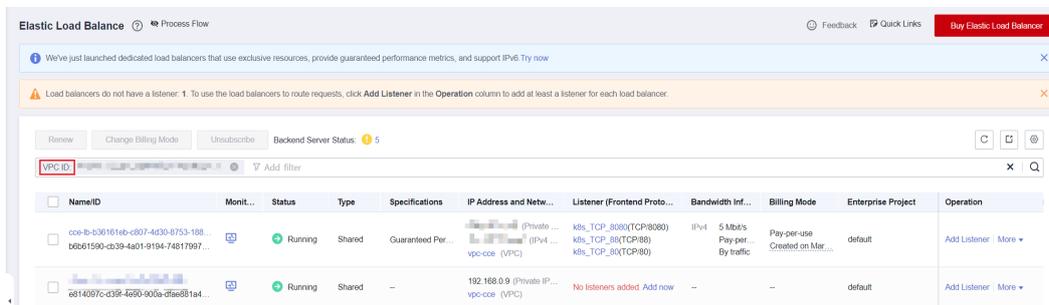
3.3.2 How Do I Clear Residual Resources After Deleting a Non-Running Cluster?

If a cluster is not in the running state (for example, frozen or unavailable), its resources such as PVCs, Services, and Ingresses cannot be obtained. After the cluster is deleted, residual network and storage resources may exist. In this case, manually delete these resources on their respective service console.

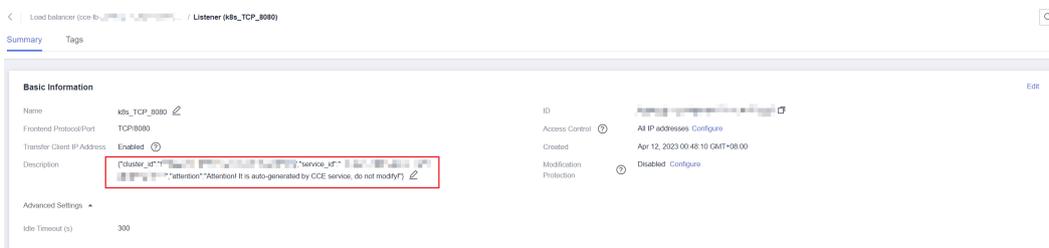
Deleting Residual ELB Resources

Step 1 Log in to the ELB console.

Step 2 Search for load balancers in the VPC by VPC ID used in the cluster.



Step 3 View the listener details of a load balancer. If the description contains the cluster ID and Service ID, the listener is created in the cluster.



Step 4 Delete the residual load balancer-related resources from the cluster based on the preceding information.

----End

Deleting Residual EVS Resources

An EVS disk dynamically created using a PVC is named in the format of **pvc-*{UID}***. The **metadata** field in the API contains the cluster ID. You can use this cluster ID to obtain these EVS disks automatically created in the cluster and delete them as required.

Step 1 Go to the EVS console.

Step 2 Search for EVS disks by **pvc-*{UID}*** to get all automatically created EVS disks in the cluster.

Disk Name	Status	Disk S...	Function	Server No...	Disk S...	Device...	Enrypt...	AZ	Billing...	Enterpris...	Operation
pvc-735f140b-1144-4765-b6bc-b1f61-af955fca-0129-45a4-a682-795244ab11	Available	Extreme S...	Data disk	--	Disabled	SCSI	No	AZ3	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More
pvc-aa8a8b1b-1023-4022-ba8a-d0a7f-0a293a42-eb60-4e25-bc8b-de88f2dd63c5	Available	High I/O S...	Data disk	--	Disabled	SCSI	No	AZ1	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More
pvc-02769831-887e-4057-83a5-b0d1-1262630a-080e-490a-5a7b-cc7f10d304c5	In-use	High I/O S...	Data disk	coe-1e4d-4...	Disabled	SCSI	No	AZ1	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More
pvc-18d8503b-8174-4307-a637-6961-9e70330a-1605-4e43-080a-4460767911d2	In-use	High I/O S...	Data disk	coe-1e4d-4...	Disabled	SCSI	No	AZ1	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More
pvc-18d8503b-8174-4307-a637-6961-2929735f-eb75-4650-b115-382340a37359	In-use	High I/O S...	Data disk	coe-1e4d-4...	Disabled	SCSI	No	AZ1	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More
pvc-385a0599-080a-48a7-8c3b-0a33-ef4e0d41-e569-4007-9631-ce507a06e081	In-use	High I/O S...	Data disk	coe-1e4d-4...	Disabled	SCSI	No	AZ1	Pay-per-use Created at...	default	Attach Expand Capacity Create Backup More

Step 3 Press **F12** to open the developer tools. Check whether the **metadata** field in the **detail** API contains the cluster ID. If yes, the EVS disks are automatically created in this cluster.

```

{count: 15, volumes: [{"id": "0e523ed2-eb60-4e25-bc8b-de88f2dd63c5", links: [{"rel": "self", "href": "/volumes/0e523ed2-eb60-4e25-bc8b-de88f2dd63c5"}], ...}]}
  
```

```

{
  "id": "0e523ed2-eb60-4e25-bc8b-de88f2dd63c5",
  "links": [{"rel": "self", "href": "/volumes/0e523ed2-eb60-4e25-bc8b-de88f2dd63c5"}],
  "attachments": [],
  "availability_zone": "cn-east-3a",
  "bootable": "false",
  "created_at": "2023-08-03T07:36:13.842134",
  "encrypted": false,
  "enterprise_project_id": "0",
  "id": "0e523ed2-eb60-4e25-bc8b-de88f2dd63c5",
  "links": [{"rel": "self", "href": "/volumes/0e523ed2-eb60-4e25-bc8b-de88f2dd63c5"}],
  "metadata": {
    "namespace": "monitoring",
    "readonly": "False",
    "hw:passthrough": "true",
    "cluster_id": "07da5e02-9f9f-11ed-bbae-0255ac1002c5"
  },
  "hw:passthrough": "true",
  "namespace": "monitoring",
  "readonly": "False",
  "multiattach": false
}
  
```

Step 4 Delete the residual EVS disk-related resources from the cluster based on the preceding information.

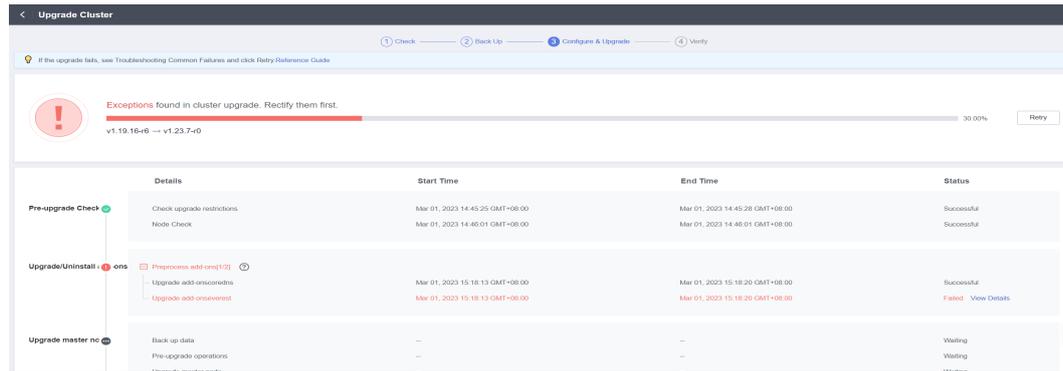
NOTE

Deleted data cannot be restored. Exercise caution when performing this operation.

----End

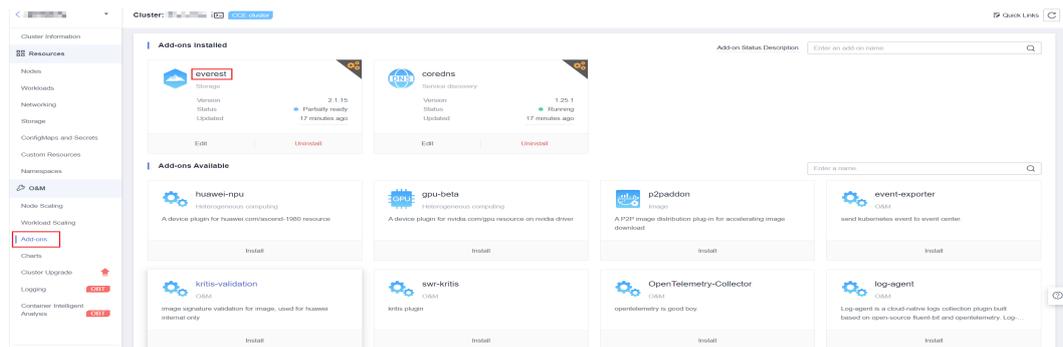
Deleting Residual SFS Resources

An SFS file system dynamically created using a PVC is named in the format of **pvc-*{UID}***. The **metadata** field in the API contains the cluster ID. You can use this cluster ID to obtain these SFS file systems automatically created in the cluster and delete them as required.



Procedure

- Step 1** If the add-on fails to be upgraded, try again first. If the retry fails, perform the following steps to rectify the fault.
- Step 2** If a failure message is displayed on the upgrade page, go to the **Add-ons** page to view the add-on status. For an abnormal add-on, click the add-on name to view details.



- Step 3** On the pod details page, click **View Events** in the **Operation** column of the abnormal pod.

Pods

Pod Name	Status	Names...	Pod IP	Node	Rest...	CPU Request/Limit/Usa	Memory Request/Limit/Usa	Created	Operation
everest-csi-driver- Host network	Running	kube-system	192.168.0.100	192.168.0.100	0	0.1 Cores 0.5 Cores 1.00%	300 MiB 300 MiB 26.25%	3 hours ago	Monitor View Events More
everest-csi-driver- Host network	Running	kube-system	192.168.0.230	192.168.0.230	0	0.1 Cores 0.5 Cores 0.80%	300 MiB 300 MiB 26.03%	3 hours ago	Monitor View Events More
everest-csi-control	Running	kube-system	10.0.0.3	192.168.0.230	0	0.25 Cores 0.25 Cores 0.80%	0.59 GiB 1.46 GiB 5.13%	3 hours ago	Monitor View Events More
everest-csi-control	Abnormal	kube-system	10.0.0.131	192.168.0.100	10	0.25 Cores 0.25 Cores 0.40%	0.59 GiB 1.46 GiB 3.92%	3 hours ago	Monitor View Events More

- Step 4** Rectify the fault based on the exception information. For example, delete the pod that is not started or restart it.

Events

X

💡 Event data is stored only for one hour and then automatically cleared.

Start Date – End Date Enter a Kubernetes event name

Kubernetes...	Event ...	Occurr...	Event Name	Kubernetes Event	First Occurred	Last Occurred
kubelet	Alarm	121	Failed to restar...	the failed container exited with ExitCod...	Mar 01, 2023 15:08:2...	Mar 01, 2023 15:33:1...
kubelet	Alarm	74	Failed to restar...	Back-off restarting failed container	Mar 01, 2023 15:08:2...	Mar 01, 2023 15:23:1...
kubelet	Alarm	4	PodsStart failed	Error: failed to start container "everest-...	Mar 01, 2023 15:08:0...	Mar 01, 2023 15:08:5...
kubelet	Normal	5	Image pulled	Container image "100.79.1.215:20202/...	Mar 01, 2023 11:53:5...	Mar 01, 2023 15:08:5...
kubelet	Normal	5	PodsCreated	Created container everest-csi-controller	Mar 01, 2023 11:53:5...	Mar 01, 2023 15:08:5...
kubelet	Normal	4	PodsVolume ...	Successfully mounted volumes for pod ...	Mar 01, 2023 11:53:5...	Mar 01, 2023 15:08:2...

Step 5 After the processing is successful, the add-on status changes to **Running**. Ensure that all add-ons are in the **Running** status.

Add-ons Installed



coredns
Service discovery

Version: 1.25.1
Status: ● **Running**
Updated: 8 days ago

[Edit](#) | [Uninstall](#)



everest
Storage

Version: 2.1.15
Status: ● **Running**
Updated: 8 days ago

[Edit](#) | [Uninstall](#)

Step 6 Go to the cluster upgrade page and click **Retry**.

① Check — ② Back Up — ③ **Configure & Upgrade** — ④ Verify

⚠️ If the upgrade fails, see [Troubleshooting Common Failures](#) and click [Retry Reference Guide](#).

Exceptions found in cluster upgrade. Rectify them first.

v1.19.16-k8s → v1.23.7-k8s 30.00%

	Details	Start Time	End Time	Status
Pre-upgrade Check ●	Check upgrade restrictions	Mar 01, 2023 14:45:25 GMT+08:00	Mar 01, 2023 14:45:28 GMT+08:00	Successful
	Node Check	Mar 01, 2023 14:46:01 GMT+08:00	Mar 01, 2023 14:46:01 GMT+08:00	Successful
Upgrade/Uninstall ● ons	❌ Preprocess add-ons [1/2] ⓘ			
	Upgrade add-ons/coredns	Mar 01, 2023 15:18:13 GMT+08:00	Mar 01, 2023 15:18:20 GMT+08:00	Successful
	Upgrade add-ons/everest	Mar 01, 2023 15:18:13 GMT+08:00	Mar 01, 2023 15:18:20 GMT+08:00	Failed View Details
Upgrade master node ●	Back up data	—	—	Waiting
	Pre-upgrade operations	—	—	Waiting

----End

4 Node

4.1 Node Creation

4.1.1 How Do I Troubleshoot Problems Occurred When Adding Nodes to a CCE Cluster?

Note

- The node images in the same cluster must be the same. Pay attention to this when creating, adding, or accepting nodes in a cluster.
- If you need to allocate user space from the data disk when creating a node, do not set the data storage path to any key directory. For example, to store data in the **/home** directory, set the directory to **/home/test** instead of **/home**.

NOTE

Do not set **Path inside a node** to the root directory **/**. Otherwise, the mounting fails. Set **Path inside a node** to any of the following:

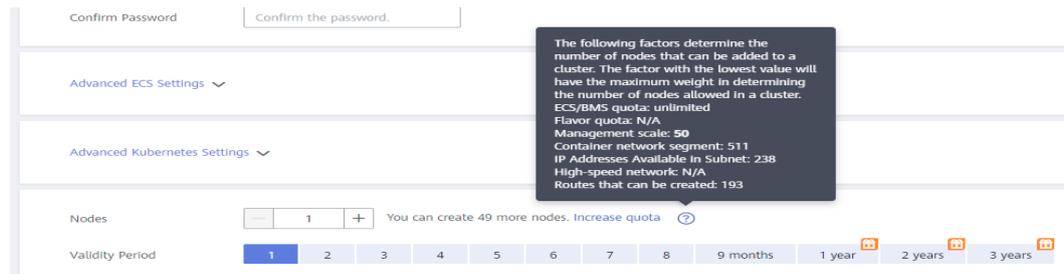
- **/opt/xxxx** (excluding **/opt/cloud**)
- **/mnt/xxxx** (excluding **/mnt/paas**)
- **/tmp/xxx**
- **/var/xxx** (excluding key directories such as **/var/lib**, **/var/script**, and **/var/paas**)
- **/xxxx** (It cannot conflict with the system directory, such as **bin**, **lib**, **home**, **root**, **boot**, **dev**, **etc**, **lost+found**, **mnt**, **proc**, **sbin**, **srv**, **tmp**, **var**, **media**, **opt**, **selinux**, **sys**, and **usr**.)

Do not set it to **/home/paas**, **/var/paas**, **/var/lib**, **/var/script**, **/mnt/paas**, or **/opt/cloud**. Otherwise, the system or node installation will fail.

Check Item 1: Subnet Quota

Symptom

New nodes cannot be added to a CCE cluster, and a message is displayed indicating that the subnet quota is insufficient.



Cause Analysis

Example:

VPC CIDR block: 192.168.66.0/24

Subnet CIDR block: 192.168.66.0/24

In 192.168.66.0/24, all 251 private IP addresses have been used.

Solution

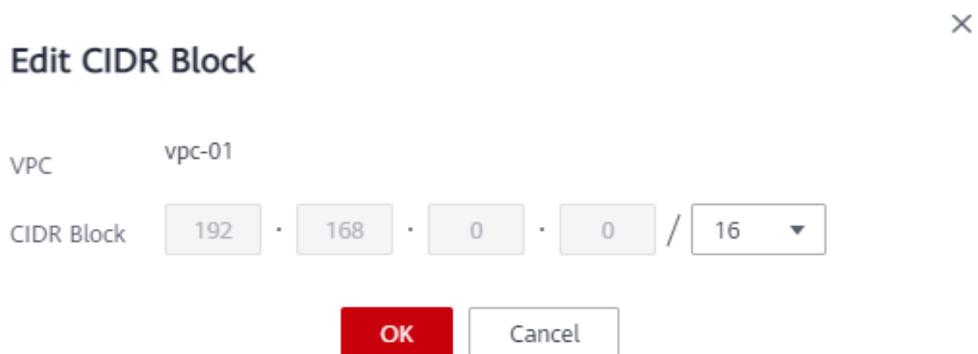
Step 1 Expand the VPC.

Log in to the console and choose **Virtual Private Cloud** from the service list. On the page displayed, locate the row containing the target VPC and click **Edit CIDR Block** in the **Operation** column.

See the following figure:



Step 2 Change the subnet mask to 16 and click OK.



Step 3 Click the VPC name. On the **Summary** tab page, click the number next to **Subnets** on the right and click **Create Subnet** to create a subnet.

✕

Create Subnet

* VPC 
IPv4 CIDR block: 192.168.0.0/16
 The VPC already contains 1 subnets.

* AZ 

* Name 
0-255

* IPv4 CIDR Block · · · / 
Available IP Addresses: 251
 The CIDR block cannot be modified after the subnet has been created.

IPv6 CIDR Block Enable 

Associated Route Table 

Advanced Settings  Gateway | DNS Server Address | Tag

Step 4 Return to the page for adding a node on the CCE console, and select the newly created subnet.

 **NOTE**

1. Adding subnets to the VPC does not affect the use of the existing 192.168.66.0/24 CIDR block.
 You can select a new subnet when creating a CCE node. The new subnet has a maximum of 251 private IP addresses. If the number of private IP addresses cannot meet service requirements, you can add more subnets.
2. Subnets in the same VPC can communicate with each other.

----End

Check Item 2: EIP Quota

Symptom

When a node is added, **EIP** is set to **Auto create**. The node cannot be created, and a message indicating that EIPs are insufficient is displayed.

Solution

Two methods are available to solve the problem.

- **Method 1:** Unbind the VMs bound with EIPs and add a node again.
 - a. Log in to the management console.

- b. Choose **Service List > Compute > Elastic Cloud Server**.
- c. In the ECS list, locate the target ECS and click its name.
- d. On the page displayed, click the **EIPs** tab. In the EIP list, locate the row containing the target EIP, click **Unbind**, and click **Yes**.

Figure 4-1 Unbinding an EIP



- e. Return to the page for adding a node on the CCE console, select **Use existing** for **EIP**, and add the node again.
- **Method 2:** Increase the EIP quota.

Check Item 3: Security Group

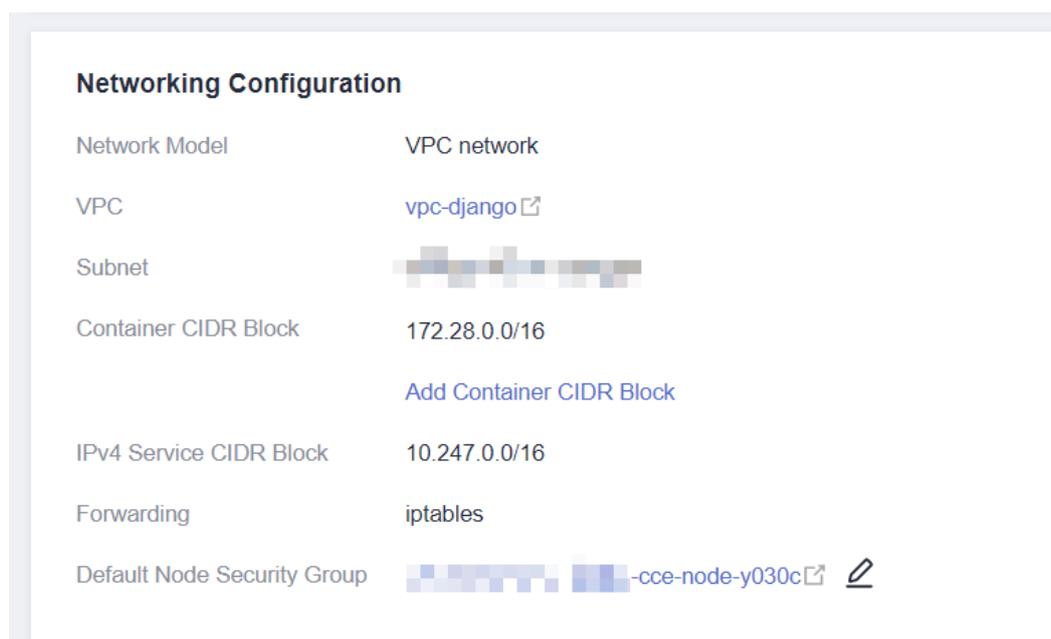
Symptom

A node cannot be added to a CCE cluster.

Solution

You can click the cluster name to view the cluster details. In the **Networking Configuration** area, click the icon next to **Default security group of the node** to check whether the default security group is deleted and whether the security group rules comply with [How Can I Configure a Security Group Rule in a Cluster?](#)

If your account has multiple clusters and you need to manage network security policies of nodes in a unified manner, you can specify custom security groups. For details, see [Changing the Default Security Group of a Node](#).



4.1.2 How Do I Troubleshoot Problems Occurred When Accepting Nodes into a CCE Cluster?

Overview

This section describes how to troubleshoot the problems occurred when you accept or add existing ECSs to a CCE cluster.

NOTICE

- While an ECS is being accepted into a cluster, the operating system of the ECS will be reset to the standard OS image provided by CCE to ensure node stability. The CCE console prompts you to select the operating system and the login mode during the reset.
- The ECS system and data disks will be formatted while the ECS is being accepted into a cluster. Ensure that data in the disks has been backed up.
- While an ECS is being accepted into a cluster, do not perform any operation on the ECS through the ECS console.

Notes and Constraints

- The cluster version must be 1.15 or later.
- Kunpeng nodes can be managed by clusters of v1.19 or later.
- You can manage ECS, BMS, and DeH nodes, but not HECS nodes.
- If **IPv6** is enabled for a cluster, only nodes in a subnet with IPv6 enabled can be accepted and managed. If **IPv6** is not enabled for the cluster, only nodes in a subnet without IPv6 enabled can be accepted.
- If a password or key has been set when the original VM node was created, reset the password or key during management. The original password or key will become invalid.
- Nodes in a CCE Turbo cluster must support sub-ENIs or be bound to at least 16 ENIs. For details about the node flavors, see the node flavors that can be selected on the console when you create a node.
- Data disks that have been partitioned will be ignored during node management. Ensure that there is at least one unpartitioned data disk meeting the specifications is attached to the node.

Prerequisites

A cloud server that meets the following conditions can be accepted:

- The node to be accepted must be in the **Running** state and not used by other clusters. In addition, the node to be accepted does not carry the CCE-Dynamic-Provisioning-Node tag.
- The node to be accepted and the cluster must be in the same VPC. (If the cluster version is earlier than v1.13.10, the node to be accepted and the CCE cluster must be in the same subnet.)

- Data disks must be attached to the nodes to be managed. A local disk (disk-intensive disk) or a data disk of at least 20 GiB can be attached to the node, and any data disks already attached cannot be smaller than 10 GiB. For details about how to attach a data disk, see [Adding a Disk to an ECS](#).
- The node to be accepted has 2-core or higher CPU, 4 GiB or larger memory, and only one NIC.
- If an enterprise project is used, the node to be accepted and the cluster must be in the same enterprise project. Otherwise, resources cannot be identified during management. As a result, the node cannot be accepted.
- Only cloud servers with the same data disk configurations can be added in batches.

Procedure

View the cluster log information to locate the failure cause and rectify the fault.

- Step 1** Log in to the CCE console. In the navigation pane, click **Operation Records** above the cluster list to view operation records.
- Step 2** Click the record of the **Failed** status to view error information.
- Step 3** Rectify the fault based on the error information and accept the node into a cluster again.

----End

Common Issues

If a node fails to be managed, a message will be displayed, indicating that the disk partitioning does not work:

```
Install config-prepare failed: exit status 1, output: [ Mon Jul 17 14:26:10 CST 2023 ] start install config-prepare\nNAME MAJ:MIN RM SIZE RO TYPE MOUNTPOINT\nsda 8:0 0 40G 0 disk \n└─sda1 8:1 0 40G 0 part\nnsdb 8:16 0 100G 0 disk \n└─sdb1 8:17 0 100G 0 part\ndisk /dev/sda has been partition, will skip this device\nRaw disk /dev/sdb has been partition, will skip this device\nwarning: selector can not match any evs volume
```

To resolve this issue, attach an unpartitioned data disk of 20 GiB or higher to the node. After the node is managed, the unpartitioned data disk is used to store the container engine and kubelet. You can perform operations on the partitioned data disk that does not work as required.

4.1.3 What Should I Do If a Node Fails to Be Accepted Because It Fails to Be Installed?

Symptom

A node fails to be accepted into a cluster.

Possible Causes

Log in to the node and check the `/var/paas/sys/log/baseagent/baseagent.log` installation log. The following error information is displayed:

```
net.core.somaxconn=32768
net.ipv4.tcp_max_syn_backlog=8096
DEERDNS=eno
failed because of no tenant.conf

10330 10:17:41.075997 6872 baseagent.go:330] install failed
E0330 10:17:41.076179 6872 install.go:181] Install Failed: Install Version(v1.13.7-rc) failed: Exec component plugins/config-prepare Install failed: exit status 1
output: [ Tue Mar 10 10:17:25 CST 2020 ] start install plugins/config-prepare
net.ipv4.ip_forward = 1
net.ipv4.neigh.default.gc_thresh1 = 2048
net.ipv4.neigh.default.gc_thresh2 = 4096
net.ipv4.neigh.default.gc_thresh3 = 8192
net.ipv4.ip_forward=1
```

Check the LVM settings of the node. It is found that the LVM logical volume is not created in `/dev/vdb`.

Solution

Run the following command to manually create a logical volume:

```
pvcreate /dev/vdb
vgcreate vgpaas /dev/vdb
```

After the node is reset on the GUI, the node becomes normal.

4.2 Node Running

4.2.1 What Should I Do If a Cluster Is Available But Some Nodes Are Unavailable?

If the cluster status is available but some nodes in the cluster are unavailable, perform the following operations to rectify the fault:

Mechanism for Detecting Node Unavailability

Kubernetes provides the heartbeat mechanism to help you determine node availability. For details about the mechanism and interval, see [Heartbeats](#).

Troubleshooting Process

The issues here are described in order of how likely they are to occur.

Check these causes one by one until you find the cause of the fault.

- [Check Item 1: Whether the Node Is Overloaded](#)
- [Check Item 2: Whether the ECS Is Deleted or Faulty](#)
- [Check Item 3: Whether You Can Log In to the ECS](#)
- [Check Item 4: Whether the Security Group Is Modified](#)
- [Check Item 5: Whether the Security Group Rules Contain the Security Group Policy for the Communication Between the Master Node and the Worker Node](#)
- [Check Item 6: Whether the Disk Is Abnormal](#)
- [Check Item 7: Whether Internal Components Are Normal](#)
- [Check Item 8: Whether the DNS Address Is Correct](#)
- [Check Item 9: Whether the vdb Disk on the Node Is Deleted](#)
- [Check Item 10: Whether the Docker Service Is Normal](#)

- [Check Item 11: Whether a Yearly/Monthly Node Is Being Unsubscribed](#)

Check Item 1: Whether the Node Is Overloaded

Symptom

The node connection in the cluster is abnormal. Multiple nodes report write errors, but services are not affected.

Fault Locating

Step 1 Log in to the CCE console and click the cluster name to access the cluster console. In the navigation pane, choose **Nodes** and click the **Nodes** tab. Locate the row that contains the unavailable node and click **Monitor**.

Step 2 On the top of the displayed page, click **View More** to go to the AOM console and view historical monitoring records.

A too high CPU or memory usage of the node will result in a high network latency or trigger system OOM. Therefore, the node is displayed as unavailable.

----End

Solution

1. You are advised to migrate services to reduce the workloads on the node and set the resource upper limit for the workloads.
2. Clear data on the CCE nodes in the cluster.
3. Limit the CPU and memory quotas of each container.
4. Add more nodes to the cluster.
5. You can also restart the node on the ECS console.
6. Add nodes to deploy memory-intensive containers separately.
7. Reset the node. For details, see [Resetting a Node](#).

After the node becomes available, the workload is restored.

Check Item 2: Whether the ECS Is Deleted or Faulty

Step 1 Check whether the cluster is available.

Log in to the CCE console and check whether the cluster is available.

- If the cluster is unavailable, for example, an error occurs, perform operations described in [How Do I Locate the Fault When a Cluster Is Unavailable?](#)
- If the cluster is running but some nodes in the cluster are unavailable, go to [Step 2](#).

Step 2 Log in to the ECS console and view the ECS status.

- If the ECS status is **Deleted**, go back to the CCE console, delete the corresponding node from the node list of the cluster, and then create another one.
- If the ECS status is **Stopped** or **Frozen**, restore the ECS first. It takes about 3 minutes to restore the ECS.
- If the ECS is **Faulty**, restart the ECS to rectify the fault.

- If the ECS status is **Running**, log in to the ECS to locate the fault according to [Check Item 7: Whether Internal Components Are Normal](#).

----End

Check Item 3: Whether You Can Log In to the ECS

Step 1 Log in to the ECS console.

Step 2 Check whether the node name displayed on the page is the same as that on the VM and whether the password or key can be used to log in to the node.

Figure 4-2 Checking the node name displayed on the page

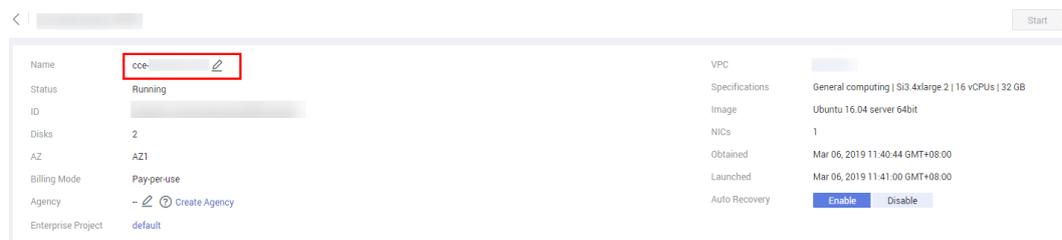
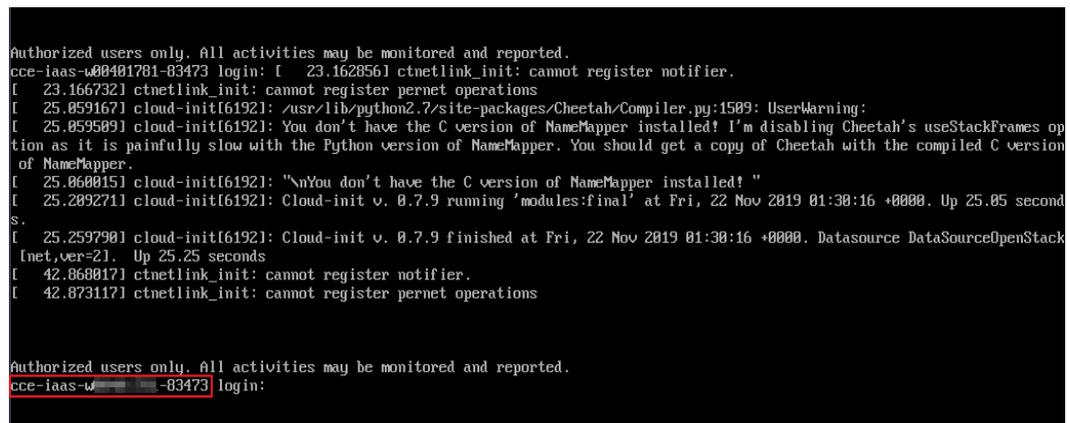


Figure 4-3 Checking the node name on the VM and whether the node can be logged in to



If the node names are inconsistent and the password and key cannot be used to log in to the node, Cloud-Init problems occurred when an ECS was created. In this case, restart the node and submit a service ticket to the ECS personnel to locate the root cause.

----End

Check Item 4: Whether the Security Group Is Modified

Log in to the VPC console. In the navigation pane, choose **Access Control > Security Groups** and locate the security group of the cluster master node.

The name of this security group is in the format of *Cluster name-cce-control-ID*. You can search for the security group by cluster name and **-cce-control-**.

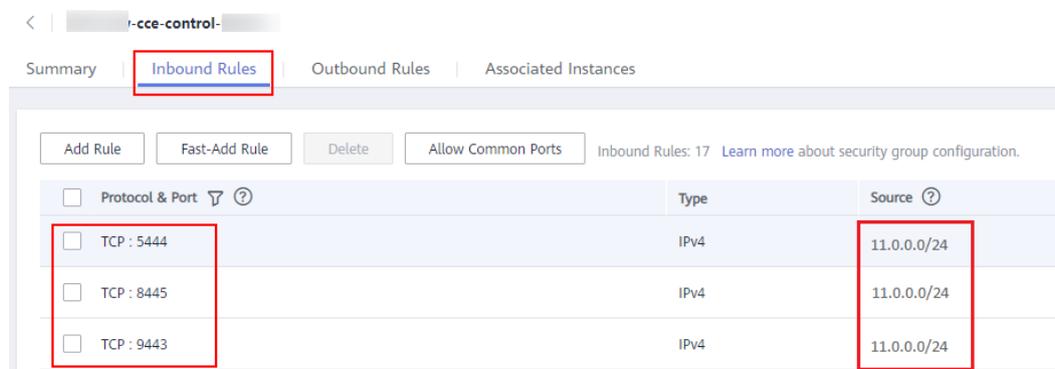
Check whether the security group rules have been modified. For details about security groups, see [How Can I Configure a Security Group Rule in a Cluster?](#)

Check Item 5: Whether the Security Group Rules Contain the Security Group Policy for the Communication Between the Master Node and the Worker Node

Check whether such a security group policy exists.

When a node is added to an existing cluster, if an extended CIDR block is added to the VPC corresponding to the subnet and the subnet is an extended CIDR block, you need to add the following three security group rules to the master node security group (the group name is in the format of **Cluster name-cce-control-Random number**). These rules ensure that the nodes added to the cluster are available. (This step is not required if an extended CIDR block has been added to the VPC during cluster creation.)

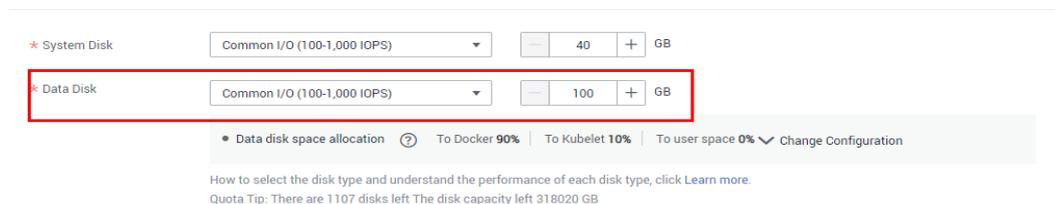
For details about security groups, see [How Can I Configure a Security Group Rule in a Cluster?](#)



Check Item 6: Whether the Disk Is Abnormal

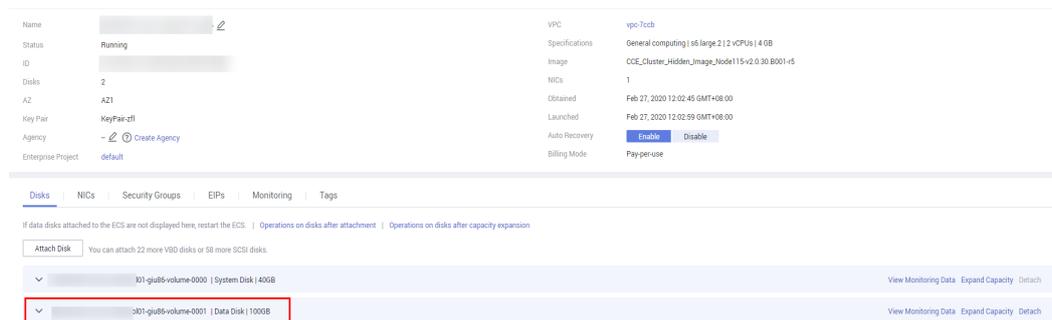
A 100 GiB data disk dedicated for Docker is attached to the new node. If the data disk is uninstalled or damaged, the Docker service becomes abnormal and the node becomes unavailable.

Figure 4-4 Data disk allocated when a node is created



Click the node name to check whether the data disk mounted to the node is uninstalled. If the disk is uninstalled, mount a data disk to the node again and restart the node. Then the node can be recovered.

Figure 4-5 Checking the disk

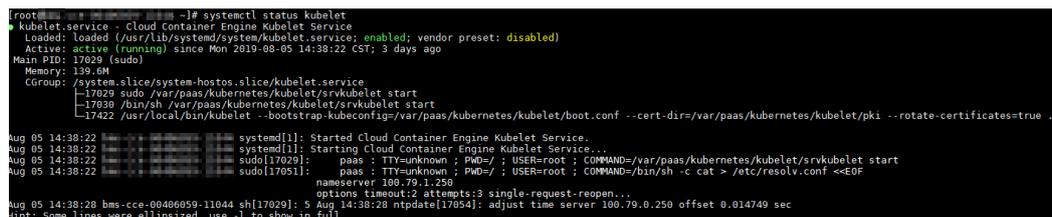


Check Item 7: Whether Internal Components Are Normal

Step 1 Log in to the ECS where the unavailable node is located.

Step 2 Run the following command to check whether the PaaS components are normal:
`systemctl status kubelet`

If the command is successfully executed, the status of each component is displayed as **active**, as shown in the following figure.



If the component status is not **active**, run the following commands (using the faulty component **canal** as an example):

Run `systemctl restart canal` to restart the component.

After restarting the component, run `systemctl status canal` to check the status.

Step 3 If the restart command fails to be run, run the following command to check the running status of the monitrc process:

`ps -ef | grep monitrc`

If the monitrc process exists, run the following command to kill this process. The monitrc process will be automatically restarted after it is killed.

`kill -s 9 `ps -ef | grep monitrc | grep -v grep | awk '{print $2}'``

----End

Check Item 8: Whether the DNS Address Is Correct

Step 1 After logging in to the node, check whether any domain name resolution failure is recorded in the `/var/log/cloud-init-output.log` file.

`cat /var/log/cloud-init-output.log | grep resolv`

If the command output contains the following information, the domain name cannot be resolved:

```
Could not resolve host: test.obs.ap-southeast-1.myhuaweicloud.com; Unknown error
```

Step 2 On the node, ping the domain name that cannot be resolved in the previous step to check whether the domain name can be resolved on the node.

ping test.obs.ap-southeast-1.myhuaweicloud.com

- If not, the DNS cannot resolve the IP address. Check whether the DNS address in the `/etc/resolv.conf` file is the same as that configured on the VPC subnet. In most cases, the DNS address in the file is incorrectly configured. As a result, the domain name cannot be resolved. Correct the DNS configuration of the VPC subnet and reset the node.
- If yes, the DNS address configuration is correct. Check whether there are other faults.

----End

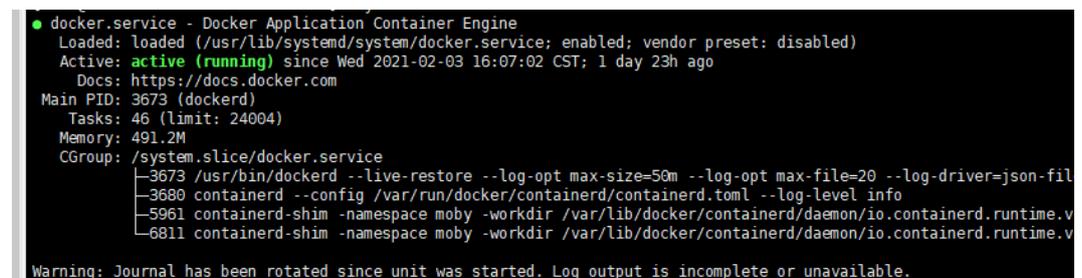
Check Item 9: Whether the vdb Disk on the Node Is Deleted

If the vdb disk on a node is deleted, you can refer to [this topic](#) to restore the node.

Check Item 10: Whether the Docker Service Is Normal

Step 1 Run the following command to check whether the Docker service is running:

```
systemctl status docker
```



```
● docker.service - Docker Application Container Engine
   Loaded: loaded (/usr/lib/systemd/system/docker.service; enabled; vendor preset: disabled)
   Active: active (running) since Wed 2021-02-03 16:07:02 CST; 1 day 23h ago
     Docs: https://docs.docker.com
   Main PID: 3673 (dockerd)
    Tasks: 46 (limit: 24004)
   Memory: 491.2M
   CGroup: /system.slice/docker.service
           └─3673 /usr/bin/dockerd --live-restore --log-opt max-size=50m --log-opt max-file=20 --log-driver=json-fil
             └─3680 containerd --config /var/run/docker/containerd/containerd.toml --log-level info
               └─5961 containerd-shim -namespace moby -workdir /var/lib/docker/containerd/daemon/io.containerd.runtime.v
                 └─6811 containerd-shim -namespace moby -workdir /var/lib/docker/containerd/daemon/io.containerd.runtime.v
Warning: Journal has been rotated since unit was started. Log output is incomplete or unavailable.
```

If the command fails or the Docker service status is not active, locate the cause or contact technical support if necessary.

Step 2 Run the following command to check the number of containers on the node:

```
docker ps -a | wc -l
```

If the command is suspended, the command execution takes a long time, or there are more than 1000 abnormal containers, check whether workloads are repeatedly created and deleted. If a large number of containers are frequently created and deleted, a large number of abnormal containers may occur and cannot be cleared in a timely manner.

In this case, stop repeated creation and deletion of the workload or use more nodes to share the workload. Generally, the nodes will be restored after a period

of time. If necessary, run the **docker rm** *{container_id}* command to manually clear abnormal containers.

----End

Check Item 11: Whether a Yearly/Monthly Node Is Being Unsubscribed

Once a node is unsubscribed, it will take some time to process the order, rendering the node unavailable during this period. Typically, the node is expected to be automatically cleared within 5 to 10 minutes.

4.2.2 How Do I Troubleshoot the Failure to Remotely Log In to a Node in a CCE Cluster?

After creating a node on CCE, you cannot remotely log in to the node using SSH. A message is displayed indicating that the selected key has not been registered on the remote host. In this case, the root user cannot directly log in to the node.

The cause is that the cloud-init is installed on the node on CCE. For the cloud-init, a default Linux user already exists and the user's key is also used for the device running Linux.

Solution

Log in to the device as the **Linux** user, and run the **sudo su** command to switch to the **root** user.

4.2.3 How Do I Log In to a Node Using a Password and Reset the Password?

Context

When creating a node on CCE, you selected a key pair or specified a password for login. If you forget your key pair or password, you can log in to the ECS console to reset the password of the node. After the password is reset, you can log in to the node using the password.

Procedure

- Step 1** Log in to the ECS console.
- Step 2** In the ECS list, select the cloud server type of the node. In the same row as the node, choose **More > Stop**.
- Step 3** After the node is stopped, choose **More > Reset Password**, and follow on-screen prompts to reset the password.
- Step 4** After the password is reset, choose **More > Start**, and click **Remote Login** to log in to the node using the password.

----End

4.2.4 How Do I Collect Logs of Nodes in a CCE Cluster?

The following tables list log files of CCE nodes.

Table 4-1 Node logs

Name	Path
kubelet log	<ul style="list-style-type: none"> For clusters of v1.21 or later: <code>/var/log/cce/kubernetes/kubelet.log</code> For clusters of v1.19 or earlier: <code>/var/paas/sys/log/kubernetes/kubelet.log</code>
kube-proxy log	<ul style="list-style-type: none"> For clusters of v1.21 or later: <code>/var/log/cce/kubernetes/kube-proxy.log</code> For clusters of v1.19 or earlier: <code>/var/paas/sys/log/kubernetes/kube-proxy.log</code>
yangtse log (networking)	<ul style="list-style-type: none"> For clusters of v1.21 or later: <code>/var/log/cce/yangtse</code> For clusters of v1.19 or earlier: <code>/var/paas/sys/log/yangtse</code>
canal log	<ul style="list-style-type: none"> For clusters of v1.21 or later: <code>/var/log/cce/canal</code> For clusters of v1.19 or earlier: <code>/var/paas/sys/log/canal</code>
System logs	<code>/var/log/messages</code>
Container engine Logs	<ul style="list-style-type: none"> For Docker nodes: <code>/var/lib/docker</code> For containerd nodes: <code>/var/log/cce/containerd</code>

Table 4-2 Add-on logs

Name	Path
everest log	<ul style="list-style-type: none"> For v2.1.41 or later: <ul style="list-style-type: none"> everest-csi-driver: <code>/var/log/cce/kubernetes</code> everest-csi-controller: <code>/var/paas/sys/log/kubernetes</code> For version earlier than v2.1.41: <ul style="list-style-type: none"> everest-csi-driver: <code>/var/log/cce/everest-csi-driver</code> everest-csi-controller: <code>/var/paas/sys/log/everest-csi-controller</code>
npd log	<ul style="list-style-type: none"> For v1.18.16 or later: <code>/var/paas/sys/log/kubernetes</code> For versions earlier than v1.18.16: <code>/var/paas/sys/log/cceaddon-npd</code>
cce-hpa-controller log	<ul style="list-style-type: none"> For v1.3.12 or later: <code>/var/paas/sys/log/kubernetes</code> For versions earlier than v1.3.12: <code>/var/paas/sys/log/ccehpa-controller</code>

4.2.5 What Can I Do If the Container Network Becomes Unavailable After yum update Is Used to Upgrade the OS?

The CCE console does not support OS upgrades on a node. You are advised not to upgrade the OS using the **yum update** command.

If you upgrade the OS using **yum update**, the container networking will be unavailable.

Perform the following operations to restore the container network:

NOTICE

This restoration method is valid only for EulerOS 2.2.

Step 1 Run the following script as user **root**:

```
#!/bin/bash
function upgrade_kmod()
{
    openvswitch_mod_path=$(rpm -qal openvswitch-kmod)
    rpm_version=$(rpm -qal openvswitch-kmod|grep -w openvswitch|head -1|awk -F "/" '{print $4}')
    sys_version=`cat /boot/grub2/grub.cfg | grep EulerOS|awk 'NR==1{print $3}' | sed 's/[()]/g`

    if [[ "${rpm_version}" != "${sys_version}" ]];then
        mkdir -p /lib/modules/"${sys_version}"/extra/openvswitch
        for path in ${openvswitch_mod_path[@]};do
            name=$(echo "$path" | awk -F "/" '{print $NF}')
            rm -f /lib/modules/"${sys_version}"/updates/"${name}"
            rm -f /lib/modules/"${sys_version}"/extra/openvswitch/"${name}"
            ln -s "${path}" /lib/modules/"${sys_version}"/extra/openvswitch/"${name}"
        done
    fi
    depmod "${sys_version}"
}
upgrade_kmod
```

Step 2 Restart the VM.

----End

Helpful Links

- [High-Risk Operations on Cluster Nodes](#)

4.2.6 What Should I Do If the vdb Disk of a Node Is Damaged and the Node Cannot Be Recovered After Reset?

Symptom

The vdb disk of a node is damaged and the node cannot be recovered after reset.

Error Scenarios

- On a normal node, delete the LV and VG. The node is unavailable.
- Reset an abnormal node, and a syntax error is reported. The node is unavailable.

The following figure shows the details.

```
vgcreate VG_new PV ...
create volume group error
, skip pause's work in case of failed dependency docker, skip fuxi's work in case of failed dependency docker, sk
work in case of failed dependency kubelet, skip kube-proxy's work in case of failed dependency config-prepare, sk
ork in case of failed dependency config-prepare, skip canal-agent's work in case of failed dependency fuxi, skip c
work in case of failed dependency config-prepare, skip docker's work in case of failed dependency config-prepare,
s work in case of failed dependency config-prepare]
18525 17:22:55.835685 7116 install.go:361 install failed
Install Failed: [Install config-prepare failed: exit status 1, output: [ Mon May 25 17:22:53 CST 2020 ] start inst
pare
success download the file
Checking device: /dev/vda
Raw disk /dev/vda has been partition, will skip this device
Checking device: /dev/vdb
Detected paas disk: /dev/vdb
Use to config lv(eg. docker(direct-lvm),kubelet,user)
No command with matching syntax recognised. Run 'vgcreate --help' for more information.
Correct command syntax is:
vgcreate VG_new PV ...

create volume group error
, skip pause's work in case of failed dependency docker, skip fuxi's work in case of failed dependency docker, sk
work in case of failed dependency kubelet, skip kube-proxy's work in case of failed dependency config-prepare, sk
ork in case of failed dependency config-prepare, skip canal-agent's work in case of failed dependency fuxi, skip c
work in case of failed dependency config-prepare, skip docker's work in case of failed dependency config-prepare,
s work in case of failed dependency config-prepare]
```

Fault Locating

If the volume group (VG) on the node is deleted or damaged and cannot be identified, you need to manually restore the VG first to prevent your data disks from being formatted by mistake during the reset.

Solution

Step 1 Log in to the node.

Step 2 Create a PV and a VG again. In this example, the following error message is displayed:

```
root@host1:~# pvcreate /dev/vdb
Device /dev/vdb excluded by a filter
```

This is because the added disk is created on another VM and has a partition table. The current VM cannot identify the partition table of the disk. You need to run the **parted** commands for three times to re-create the partition table.

```
root@host1:~# parted /dev/vdb
GNU Parted 3.2
Using /dev/vdb
Welcome to GNU Parted! Type 'help' to view a list of commands.
(parted) mklabel msdos
Warning: The existing disk label on /dev/vdb will be destroyed and all data on this disk will be lost. Do you
want to continue?
Yes/No? yes
(parted) quit
Information: You may need to update /etc/fstab.
```

Run **pvcreate** again. When the system asks you whether to erase the DOS signature, enter **y**. The disk is created as a PV.

```
root@host1:~# pvcreate /dev/vdb
WARNING: dos signature detected on /dev/vdb at offset 510. Wipe it? [y/n]: y
Wiping dos signature on /dev/vdb.
Physical volume "/dev/vdb" successfully created
```

Step 3 Create a VG.

Check the Docker disks of the node. If the disks are **/dev/vdb** and **/dev/vdc**, run the following command:

```
root@host1:~# vgcreate vgpaas /dev/vdb /dev/vdc
```

If there is only the **/dev/vdb** disk, run the following command:

```
root@host1:~# vgcreate vgpaas /dev/vdb
```

After the creation is complete, reset the node.

----End

4.2.7 Which Ports Are Used to Install kubelet on CCE Cluster Nodes?

The following ports are used:

- **10250 –port**: used to listen on kubelet. The related API can be used to check whether kubelet is running.
- **10248 –healthz-port**: used for health checks.
- **10255 –read-only-port**: read-only port, which can be accessed directly without authentication and authorization.
- **4194 –advisor-port**: used by cAdvisor.

4.2.8 How Do I Configure a Pod to Use the Acceleration Capability of a GPU Node?

Problem Description

I have purchased a GPU node, but the operating speed is still slow. How do I configure the pod to use the acceleration capability of the GPU node?

Solution

Solution 1:

You are advised to remove the unschedulable taints from the GPU nodes in the cluster, so that the GPU plug-in driver can be properly installed. In addition, you need to install the GPU driver of a later version.

If a container is not deployed on a GPU node in your cluster, you can configure affinity and anti-affinity policies to prevent the container from being scheduled to the GPU node.

Solution 2:

You are advised to install the GPU driver of a later version and use `kubectl` to update the GPU plug-in configuration. Add the following configuration:

```
tolerations:  
- operator: "Exists"
```

After the configuration is added, the GPU plug-in driver can be properly installed on the GPU node with a taint.

4.2.9 What Should I Do If I/O Suspension Occasionally Occurs When SCSI EVS Disks Are Used?

Symptom

When SCSI EVS disks are used and containers are created and deleted on a CentOS node, the disks are frequently mounted and unmounted. The read/write rate of the system disk may instantaneously surge. As a result, the system is suspended, affecting the normal node running.

When this problem occurs, the following information is displayed in the dmesg log:

```
Attached SCSI disk
task jdb2/xxx blocked for more than 120 seconds.
```

Example:

```
1128163.173120] sd 2:0:0:0: [sda] write Protect is 011
1128163.173457] sd 2:0:0:0: [sda] Mode Sense: 69 00 00 08
1128163.173573] sd 2:0:0:0: [sda] Write cache: disabled, read cache: enabled, doesn't support DPO or FUA
1128163.176426] sd 2:0:0:0: [sda] Attached SCSI disk
1128350.437941] INFO: task jbd2/dm-1-8:1604 blocked for more than 120 seconds.
1128350.438267] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
1128350.438564] jbd2/dm-1-8 D ffff9ede7f8420e0 0 1604 2 0x00000000
1128350.438829] Call Trace:
1128350.439120] [<ffffffffffaab5a585>] ? blk_mq_dispatch_rq_list+0x325/0x620
1128350.439394] [<ffffffffffaaf7f229>] schedule+0x29/0x70
```

Possible Causes

After a PCI device is hot added to BUS 0, the Linux OS kernel will traverse all the PCI bridges mounted to BUS 0 for multiple times, and these PCI bridges cannot work properly during this period. During this period, if the PCI bridge used by the device is updated, due to a kernel defect, the device considers that the PCI bridge is abnormal, and the device enters a fault mode and cannot work normally. If the front end is writing data into the PCI configuration space for the back end to process disk I/Os, the write operation may be deleted. As a result, the back end cannot receive notifications to process new requests on the I/O ring. Finally, the front-end I/O suspension occurs.

Impact

CentOS Linux kernels of versions earlier than 3.10.0-1127.el7 are affected.

Solution

Upgrade the kernel to a later version **by resetting the node**. For details, see [Resetting a Node](#).

Verification Method

Run the following command to check whether the fault is rectified:

```
auditctl -l | grep "/var/lib/docker -p rwx -k docker"
```

If no command output is displayed, the problem has been resolved.

4.2.11 How Do I Fix an Abnormal Container or Node Due to No Thin Pool Disk Space?

Problem Description

When the disk space of a thin pool on a node is about to be used up, the following exceptions occasionally occur:

Files or directories fail to be created in the container, the file system in the container is read-only, the node is tainted disk-pressure, or the node is unavailable.

You can run the **docker info** command on the node to view the used and remaining thin pool space to locate the fault. The following figure is an example.

```
Storage Driver: devicemapper
Pool Name: vgpaas-thinpool
Pool Blocksize: 524.3kB
Base Device Size: 10.74GB
Backing Filesystem: ext4
Udev Sync Supported: true
Data Space Used: 7.794GB
Data Space Total: 71.94GB
Data Space Available: 64.15GB
Metadata Space Used: 3.076MB
Metadata Space Total: 3.221GB
Metadata Space Available: 3.218GB
Thin Pool Minimum Free Space: 7.194GB
Deferred Removal Enabled: true
Deferred Deletion Enabled: true
Deferred Deleted Device Count: 0
Library Version: 1.02.146-RHEL7 (2018-01-22)
```

Possible Cause

When Docker device mapper is used, although you can configure the **basesize** parameter to limit the size of the **/home** directory of a single container (to 10 GB by default), all containers on the node still share the thin pool of the node for storage. They are not completely isolated. When the sum of the thin pool space used by certain containers reaches the upper limit, other containers cannot run properly.

In addition, after a file is deleted in the **/home** directory of the container, the thin pool space occupied by the file is not released immediately. Therefore, even if **basesize** is set to 10 GB, the thin pool space occupied by files keeps increasing until 10 GB when files are created in the container. The space released after file deletion will be reused only after a while. If **the number of service containers on the node multiplied by basesize** is greater than the thin pool space size of the node, there is a possibility that the thin pool space has been used up.

Solution

When the thin pool space of a node is used up, some services can be migrated to other nodes to quickly recover services. But you are advised to use the following solutions to resolve the root cause:

Solution 1:

Properly plan the service distribution and data plane disk space to avoid the scenario where **the number of service containers multiplied by basesize** is greater than the thin pool size of the node. To expand the thin pool size, perform the following steps:

- Step 1** Expand the capacity of a data disk on the EVS console. For details, see [Expanding EVS Disk Capacity](#).

Only the storage capacity of the EVS disk is expanded. You also need to perform the following steps to expand the capacity of the logical volume and file system.

- Step 2** Log in to the CCE console and click the cluster. In the navigation pane, choose **Nodes**. Click **More > Sync Server Data** in the row containing the target node.

- Step 3** Log in to the target node.

- Step 4** Run the **lsblk** command to check the block device information of the node.

A data disk is divided depending on the container storage **Rootfs**:

Overlayfs: No independent thin pool is allocated. Image data is stored in **dockersys**.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME        MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda         8:0   0  50G  0 disk
├─vda1      8:1   0  50G  0 part /
└─vdb       8:16  0 200G  0 disk # Data disk has been expanded but not allocated
   └─vgpaas-dockersys 253:0  0  90G  0 lvm  /var/lib/containerd # Space used by the
      container engine
         └─vgpaas-kubernetes 253:1  0  10G  0 lvm  /mnt/paas/kubernetes/kubelet # Space used by
            Kubernetes
```

2. Expand the disk capacity.

Add the new disk capacity to the **dockersys** logical volume used by the container engine.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/dockersys` specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <90.00 GiB (23039 extents) to <190.00 GiB (48639 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. `/dev/vgpaas/dockersys` specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/containerd; on-line resizing required
old_desc_blocks = 12, new_desc_blocks = 24
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

Devicemapper: A thin pool is allocated to store image data.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME                                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda                                  8:0    0  50G  0 disk
├─vda1                               8:1    0  50G  0 part /
└─vgdb                                8:16   0 200G  0 disk
   └─vgpaas-dockersys                 253:0   0  18G  0 lvm  /var/lib/docker
      └─vgpaas-thinpool_tmeta          253:1   0   3G  0 lvm
         └─vgpaas-thinpool             253:3   0  67G  0 lvm          # Space used by thinpool
            ...
      └─vgpaas-thinpool_tdata          253:2   0  67G  0 lvm
         └─vgpaas-thinpool             253:3   0  67G  0 lvm
            ...
   └─vgpaas-kubernetes                 253:4   0  10G  0 lvm  /mnt/paas/kubernetes/kubelet
```

2. Expand the disk capacity.

Option 1: Add the new disk capacity to the thin pool disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where thinpool is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/thinpool` specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/thinpool
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/thinpool changed from <67.00 GiB (23039 extents) to <167.00 GiB (48639 extents).
Logical volume vgpaas/thinpool successfully resized.
```

- c. Do not need to adjust the size of the file system, because the thin pool is not mounted to any devices.

Option 2: Add the new disk capacity to the **dockersys** disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/dockersys` specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <18.00 GiB (7679 extents) to <118.00 GiB (33279 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. `/dev/vgpaas/dockersys` specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/docker; on-line resizing required
old_desc_blocks = 4, new_desc_blocks = 16
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

----End

Solution 2:

Create and delete files in service containers in the local storage (such as `emptyDir` and `hostPath`) or cloud storage directory mounted to the container. Such files do not occupy the thin pool space.

Solution 3:

If the OS uses OverlayFS, services can be deployed on such nodes to prevent the problem that the disk space occupied by files created or deleted in the container is not released immediately.

4.2.12 Where Can I Get the Listening Ports of CCE Worker Nodes?

Table 4-3 Listening ports of a worker node

Destination Port	Protocol	Description
10248	TCP	Health check port for kubelet
10250	TCP	Service port of kubelet to provide monitoring information about workloads on nodes and access channels for containers
10255	TCP	Read-only port of kubelet to provide monitoring information about workloads on the node
Dynamic port (related to the range of the host machine, for example, the kernel parameter <code>net.ipv4.ip_local_port_range</code>)	TCP	Random port listened by kubelet, which is used to communicate with CRI Shim to obtain the EXEC URL.
10249	TCP	kube-proxy metric port to provide kube-proxy monitoring information
10256	TCP	Health check port for kube-proxy

Destination Port	Protocol	Description
Dynamic port (32768-65535)	TCP	WebSocket listening port for functions such as docker exec
Dynamic port (32768-65535)	TCP	WebSocket listening port for functions such as containerd exec
28001	TCP	Local listening port of ICAgent to receive syslog logs of the node
28002	TCP	Health check port for ICAgent
20101	TCP	Health check port of yangtse-agent/canal-agent (involved when the container tunnel network model is used)
20104	TCP	Metric port of yangtse-agent/canal-agent to provide component monitoring information (involved when the container tunnel network model is used)
3125	TCP	Health check listening port of everest-csi-driver
3126	TCP	everest-csi-driver pprof port
19900	TCP	Server port for the health check of node-problem-detector
19901	TCP	Port for connecting node-problem-detector to Prometheus to collect monitoring data
4789	UDP	OVS listening port, which is used to transmit VXLAN packets in container networking (involved when the container tunnel network model is used)
4789	UDIPv6	OVS listening port, which is used to transmit VXLAN packets in container networking (involved when the container tunnel network model is used)
Dynamic port (30000-32767)	TCP	Listening port of kube-proxy for layer-4 load balancing. Kubernetes allocates a random port to NodePort and Loadbalancer Services. The default port number ranges from 30000 to 32767.

Destination Port	Protocol	Description
Dynamic port (30000-32767)	UDP	Listening port of kube-proxy for layer-4 load balancing. Kubernetes allocates a random port to NodePort and Loadbalancer Services. The default port number ranges from 30000 to 32767.
123	UDP	Listening port of ntpd used for time synchronization
20202	TCP	Listening port of PodLB for layer-7 load balancing, which forwards container image pull requests.

4.2.13 How Do I Rectify Failures When the NVIDIA Driver Is Used to Start Containers on GPU Nodes?

Did a Resource Scheduling Failure Event Occur on a Cluster Node?

Symptom

A node is running properly and has GPU resources. However, the following error information is displayed:

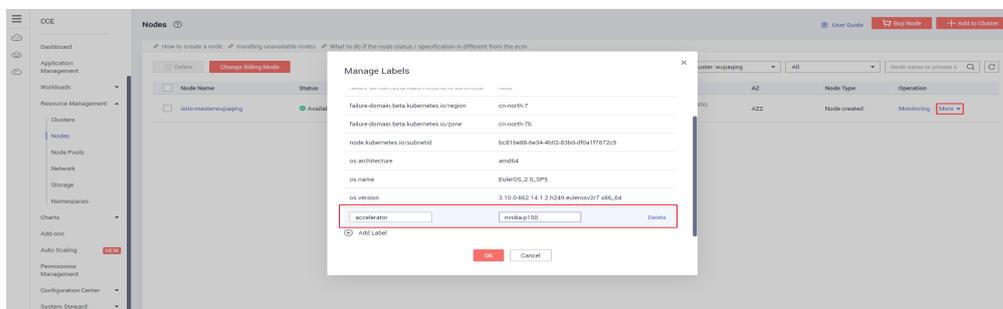
0/9 nodes are available: 9 insufficient nvidia.com/gpu

Analysis

1. Check whether the node is attached with NVIDIA label.

```

unknown flag: --show-labels
[root@chengyindu-test-98835 ~]# kubectl get node --show-labels
NAME              STATUS    ROLES    AGE    VERSION    LABELS
172.16.0.180      Ready    <none>   6h26m v1.13.10-r1-CCE2.0.28.B001  accelerator=nvidia-p100, beta.kubernetes.io/arch=amd64, beta.kubernetes.io/os=linux, failure-domain.beta.kubernetes.io/region=cn-east-2, failure-domain.beta.kubernetes.io/zone=cn-east-2b, kubernetes.io/availablezone=cn-east-2b, kubernetes.io/eni-quota=12, kubernetes.io/hostname=172.16.0.180, node.kubernetes.io/subnet-id=4883a3c2-f09f-412d-bd3a-5a2892c5033a, os.architecture=amd64, os.name=EulerOS_2.0_SP5, os.version=3.10.0-862.14.1.2.h249.el8rosv2r7.x86_64
[root@chengyindu-test-98835 ~]#
    
```



2. Check whether the NVIDIA driver is running properly.

Log in to the node where the add-on is running and view the driver installation log in the following path:

`/opt/cloud/cce/nvidia/nvidia_installer.log`

View standard output logs of the NVIDIA container.

Filter the container ID by running the following command:

```
docker ps -a | grep nvidia
```

View logs by running the following command:

```
docker logs Container ID
```

What Should I Do If the NVIDIA Version Reported by a Service and the CUDA Version Do Not Match?

Run the following command to check the CUDA version in the container:

```
cat /usr/local/cuda/version.txt
```

Check whether the CUDA version supported by the NVIDIA driver version of the node where the container is located contains the CUDA version of the container.

Helpful Links

[What Should I Do If an Error Occurs When Deploying a Service on the GPU Node?](#)

4.2.14 What Can I Do If the Time of CCE Nodes Is Not Synchronized with the NTP Server?

Symptom

In special scenarios, if the ntpd on a node cannot access the NTP server for a long time, the time offset may be too large and cannot be automatically restored.

Problem Detection

You can detect the problem by installing the CCE Node Problem Detector (npd) add-on that has the node time synchronization check item. For details, see [CCE Node Problem Detector](#).

Possible Causes

This is a known issue on nodes running EulerOS or CentOS. Nodes running other types of OSs are not affected by this issue.

NOTE

This issue has been resolved in clusters of v1.19.16-r7, v1.21.9-r10, and v1.23.7-r10.

Solution

- If your cluster version is v1.19.16-r7, v1.21.9-r10, v1.23.7-r10, or later, the node of this version has been switched to chronyd for time synchronization. Reset the node OS to the latest version to rectify the fault.
- If your cluster version does not meet the requirements, you are advised to upgrade the cluster to v1.19.16-r7, v1.21.9-r10, v1.23.7-r10, or later, and then reset the node OS to the latest version.

4.2.15 What Should I Do If the Data Disk Usage Is High Because a Large Volume of Data Is Written Into the Log File?

Symptom

Service containers on nodes that use containerd as the container runtime continuously write a large volume of data into the log file, resulting in full space of the `/var/lib/containerd` directory and slowing down the creation and deletion of containers on the node. This may evict pods and cause problems like high disk usage and abnormal nodes.

Possible Causes

For such service containers, if their logs are generated to the STDOUT, the kubelet will dump the logs. The kubelet also maintains the lifecycle of all containers on the node.

The kubelet will be overloaded if there are too many service containers on a node and they write a large volume of data into the log files. If the load exceeds a certain threshold, kubelet will dump the logs to the disk, which further results in a high disk usage. Operations such as container creation and deletion on the node will be affected.

Solution

Typically, for a node with 8 vCPUs, 16 GiB memory, and a 100 GiB data disk, the standard log output rate of a single container should be less than or equal to 512 KB/s and the overall standard log output rate of all containers on the node should be less than or equal to 5 MB/s. If a large number of logs are generated, resolve this issue in either of the following ways:

- Do not schedule containers which generate too many logs on the same node. For example, configure anti-affinity policies for pods running such containers or reduce the maximum number of pods on a single node.
- Attach an additional data disk separately. For example, you can attach an extra data disk or mount a dynamically provisioned storage volume when creating a node so that logs can be written to files in it.

4.2.16 Why Does My Node Memory Usage Obtained by Running the `kubelet top node` Command Exceeds 100%?

Symptom

The memory usage of the node obtained on the CCE console is not high, but that obtained by running the `kubelet top node` command exceeded 100%.

NAME	CPU(cores)	CPU%	MEMORY(bytes)	MEMORY%
192.168.0.243	79m	4%	2357Mi	109%

Possible Causes

kubectl top node calls the kubelet metrics API to obtain data, and the displayed information indicates the total number of used resources on the node divided by all allocatable resources.

For details, see <https://github.com/kubernetes/kubernetes/issues/86499>.

Example Scenarios

To obtain the parameters of a node, run **kubectl describe node**. The following is an example:

```
...
Capacity:
  cpu:                2
  ephemeral-storage: 51286496Ki
  hugepages-1Gi:     0
  hugepages-2Mi:     0
  localssd:          0
  localvolume:       0
  memory:            3494556Ki
  pods:              40
Allocatable:
  cpu:                1960m
  ephemeral-storage: 47265634636
  hugepages-1Gi:     0
  hugepages-2Mi:     0
  localssd:          0
  localvolume:       0
  memory:            2213604Ki
  pods:              40
...
```

- The **Capacity.memory** field whose value is **4030180Ki** indicates the total memory of the node.
- The **Allocatable.memory** field whose value is **2213604Ki** indicates the allocatable memory of the node.
- The value of the node's used memory in this example is **2413824Ki**. To obtain the value, run the following command:

```
kubectl get --raw /apis/metrics.k8s.io/v1beta1/nodes/
```

Information similar to the following will be displayed:

```
{
  "kind": "NodeMetricsList",
  "apiVersion": "metrics.k8s.io/v1beta1",
  "metadata": {},
  "items": [
    {
      ...
      "timestamp": "2023-08-15T14:09:38Z",
      "window": "1m0.177s",
      "usage": {
        "cpu": "78528126n",
        "memory": "2413824Ki"
      }
    }
  ]
}
```

To check the memory usage of the node, run **kubelet top node**.

Memory usage of a node = Used memory of the node/Allocatable memory of the node = 2413824Ki/2213604Ki = 109%

The actual memory usage of the node is calculated as follows:

Actual memory usage of the node = Used memory of the node/Total memory of the node = 2413824Ki/4030180Ki = 59.9%

4.3 Specification Change

4.3.1 How Do I Change the Node Specifications in a CCE Cluster?

Constraints

- CCE Turbo cluster nodes of certain specifications can be created only on the CCE console and cannot be modified on the ECS console. You can call the ECS API to modify the specifications. For details, see [Modifying the Specifications of an ECS](#).

Solution

CAUTION

If the node whose specifications need to be changed is accepted into the cluster for management, remove the node from the cluster and then change the node specifications to avoid affecting services.

- Step 1** Log in to the CCE console and click the cluster. In the navigation pane, choose **Nodes**. Click the name of the node to display the ECS details page.
- Step 2** In the upper right corner of the ECS details page, click **Stop**. After the ECS is stopped, choose **More > Modify Specifications**.
- Step 3** On the **Modify ECS Specifications** page, select a flavor name and click **Submit** to finish the specification modification. Return to ECS list page and choose **More > Start** to start the ECS.
- Step 4** Log in to the CCE console and click the cluster name to access the cluster console. In the navigation pane, choose **Nodes**. Locate the target node in the node list, and click **Sync Server Data** in the **Operation** column. After the synchronization is complete, you can view that the node specifications are the same as the modified specifications of the ECS.

----End

Common Issues

After the specifications of a node configured with CPU management policies are changed, the node may fail to be rebooted or workloads may fail to be created. In this case, see [What Should I Do If I Fail to Restart or Create Workloads on a Node After Modifying the Node Specifications?](#) to rectify the fault.

4.3.2 What Should I Do If I Fail to Restart or Create Workloads on a Node After Modifying the Node Specifications?

Context

The kubelet option **cpu-manager-policy** defaults to **static**, allowing pods with certain resource characteristics to be granted increased CPU affinity and exclusivity on the node. If you modify CCE node specifications on the ECS console, the original CPU information does not match the new CPU information. As a result, workloads on the node cannot be restarted or created.

For more information, see [Control CPU Management Policies on the Node](#).

Impact

The clusters that have enabled a CPU management policy will be affected.

Solution

Step 1 Log in to the CCE node (ECS) and delete the **cpu_manager_state** file.

Example command for the file deletion:

```
rm -rf /mnt/paas/kubernetes/kubelet/cpu_manager_state
```

Step 2 Restart the node or kubelet. The following is the kubelet restart command:

```
systemctl restart kubelet
```

Step 3 Verify that workloads on the node can be successfully restarted or created.

----End

4.3.3 Can I Change the IP Address of a Node in a CCE Cluster?

- A private IP address cannot be changed. CCE clusters use the private IP address of a node as the Kubernetes node name, which cannot be changed. Changing the name will make the node unavailable.
- The public IP address of a node can be changed on the ECS console.

How Do I Restore a Node After Its Private IP Is Changed?

After the private IP of a node is changed, the node becomes unavailable. You need to change it back to the original IP address.

Step 1 On the CCE console, view the node details and find the IP address and subnet of the node.

Figure 4-6 Private IP address and subnet of the node

Node Name	Status	Node Pool	Configure	IP Address
<input type="checkbox"/>	Running Schedulable	DefaultPool	AZ3 ac7.large.2 2vCPUs 4GiB	45 (Private)

Step 2 Log in to the ECS console, locate and stop the node, go to the node details page, and change the private IP address on the **Network Interfaces** tab page. Note that you need to select the corresponding subnet.

Figure 4-7 Changing the private IP address

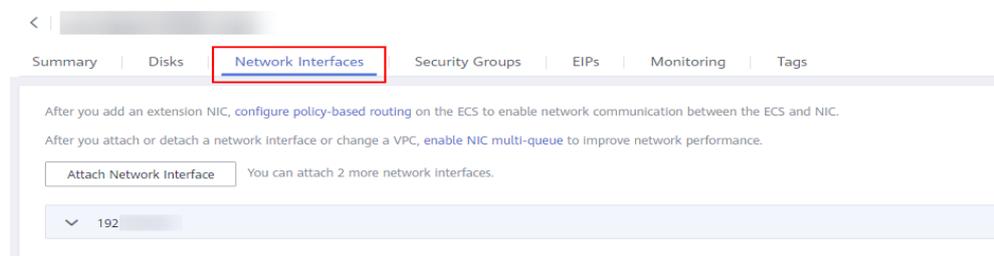
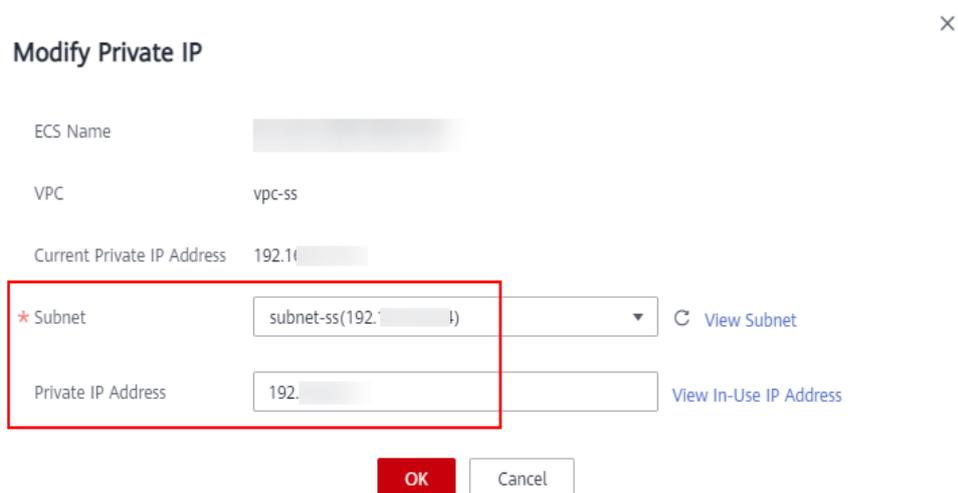


Figure 4-8 Changing the private IP address



Step 3 After the modification is complete, restart the node.

----End

4.4 OSs

4.4.1 What Can I Do If cgroup kmem Leakage Occasionally Occurs When an Application Is Repeatedly Created or Deleted on a Node Running CentOS with an Earlier Kernel Version?

Symptom

When an application is repeatedly created on a node running CentOS 7.6 with a kernel version earlier than 3.10.0-1062.12.1.el7.x86_64, (Such nodes mainly run in clusters 1.17.9.) cgroup kmem leakage occurs. As a result, although there is available memory on the node, new pods still cannot be added to it, and the error message "Cannot allocate memory" displays.

Possible Causes

A temporary memory cgroup is created along with the creation of the application. When the application is deleted, the cgroup (the corresponding **cgroup** directory in **/sys/fs/cgroup/memory**) has already been deleted from the kernel. But in the kernel, **cssid** is not released, which results in the number of cgroups considered by the kernel is different from the actual number. When the number of residual cgroups exhausts the limit on the node, pods cannot be added to the node.

Solution

- Use the **cgroup.memory=nokmem** parameter globally at the kernel to disable **kmem** to prevent leakage.
- Clusters of v1.17 are no longer maintained. To resolve this problem, upgrade the cluster to v1.19 or later and reset the OS of the node to the latest version. Ensure that the kernel version is later than 3.10.0-1062.12.1.el7.x86_64.

4.4.2 What Should I Do If There Is a Service Access Failure After a Backend Service Upgrade or a 1-Second Latency When a Service Accesses a CCE Cluster?

Symptom

If the kernel version of a node is earlier than 5.9 and a CCE cluster runs in IPVS forwarding mode, there may be a service access failure after a backend service upgrade or a 1-second latency when a service accesses the CCE cluster. This is caused by a bug in reusing Kubernetes IPVS connections.

IPVS Connection Reuse Parameters

The port reuse policy of IPVS is determined by the kernel parameter **net.ipv4.vs.conn_reuse_mode**.

1. If **net.ipv4.vs.conn_reuse_mode** is set to **0**, IPVS does not reschedule a new connection, but forwards the new connection to the original RS (IPVS backend).
2. If **net.ipv4.vs.conn_reuse_mode** is set to **1**, IPVS reschedules a new connection.

Problems Caused by IPVS Connection Reuse

- **Problem 1**
If **net.ipv4.vs.conn_reuse_mode** is set to **0**, IPVS does not proactively schedule new connections with port reuse or trigger any connection termination or drop operations. Data packets of the new connections will be directly forwarded to the previously used backend pod. If the backend pod has been deleted or recreated, an exception occurs. However, according to the current implementation logic, in a high-concurrency service access scenario, connection requests for port reuse are continuously forwarded, while kube-proxy did not delete the old ones, resulting in a service access failure.
- **Problem 2**

If **net.ipv4.vs.conn_reuse_mode** is set to **1** and the source port is the same as that of a previous connection in a high-concurrency scenario, the connection is not reused but rescheduled. According to the processing logic of `ip_vs_in()`, if **net.ipv4.vs.conntrack** is enabled, the first SYN packet is dropped. As a result, the SYN packet will be retransmitted, leading to a 1-second latency, and the performance deteriorates.

Community Settings and Impact on CCE Clusters

The default value of **net.ipv4.vs.conn_reuse_mode** on a node is **1**. However, the Kubernetes kube-proxy resets this parameter.

Cluster Version	kube-proxy Action	Impact on CCE Cluster
1.17 or earlier	By default, kube-proxy sets net.ipv4.vs.conn_reuse_mode to 0 . For details, see Fix IPVS low throughput issue .	If CCE clusters of 1.17 or earlier versions use the IPVS service forwarding mode, kube-proxy will set the net.ipv4.vs.conn_reuse_mode value of all nodes to 0 by default. This causes Problem 1 : The RS cannot be removed when the port is reused.

Cluster Version	kube-proxy Action	Impact on CCE Cluster
1.19 or later	<p>kube-proxy sets the value of net.ipv4.vs.conn_reuse_mode based on the kernel version. For details, see ipvs: only attempt setting of sysctlconnreuse on supported kernels.</p> <ul style="list-style-type: none"> If the kernel version is later than 4.1, kube-proxy will set net.ipv4.vs.conn_reuse_mode to 0. In other cases, the default value 1 will be retained. <p>NOTE This issue has been resolved in Linux kernel 5.9. Since Kubernetes 1.22, kube-proxy does not modify the net.ipv4.vs.conn_reuse_mode parameter of nodes that use the kernel 5.9 or later. For details, see Don't set sysctl net.ipv4.vs.conn_reuse_mode for kernels >=5.9.</p>	<p>If the IPVS service forwarding mode is used in CCE clusters of 1.19.16-r0 or later, the value of net.ipv4.vs.conn_reuse_mode varies with the kernel versions of node OSs.</p> <ul style="list-style-type: none"> For a node running EulerOS 2.5 or CentOS 7.6, if the kernel version is earlier than 4.1, kube-proxy will keep net.ipv4.vs.conn_reuse_mode at 1. This results in Problem 2, which is, there is a 1-second latency in the high-concurrency scenarios. For a node running Ubuntu 18.04, if the kernel version is later than 4.1, kube-proxy will set net.ipv4.vs.conn_reuse_mode to 0. This causes Problem 1: The RS cannot be removed when the port is reused. For a node running EulerOS 2.9, if the kernel version is too early, kube-proxy will set net.ipv4.vs.conn_reuse_mode to 0. This results in Problem 1. To resolve this problem, upgrade the kernel version. For details, see Rectification Plan. For a node running Huawei Cloud EulerOS 2.0 or Ubuntu 22.04, if the kernel version is later than 5.9, the problem has been resolved.

Suggestions

Evaluate the impact of these problems. If they affect your services, take the following measures:

1. Use an OS that is not affected by the preceding issues, for example, Huawei Cloud EulerOS 2.0 or Ubuntu 22.04. The newly created nodes which run EulerOS 2.9 are not affected by the preceding issues. Upgrade the earlier kernel versions used by existing nodes to the fixed version. For details, see [Rectification Plan](#).
2. Use a cluster whose forwarding mode is iptables.

Rectification Plan

If you use a node running EulerOS 2.9, check whether the kernel version meets the requirements. If the kernel version of the node is too early, reset the node or create a new one.

The following kernel versions are recommended:

- x86: 4.18.0-147.5.1.6.h686.eulerosv2r9.x86_64
- Arm: 4.19.90-vhulk2103.1.0.h584.eulerosv2r9.aarch64

Kubernetes community issue: <https://github.com/kubernetes/kubernetes/issues/81775>

4.4.3 Why Are Pods Evicted by kubelet Due to Abnormal cgroup Statistics?

Symptom

On an Arm node, pods are evicted by kubelet due to the abnormal cgroup statistics. As a result, the node runs abnormally.

kubelet keeps evicting pods. After all containers are killed, kubelet still considers that the memory is insufficient.

```

0000 [3] #41745002.1730/5151 (duration:0.000000000000000000) Error: UnregisterPlugin: dial failed at socket /mnt/paas/kubernetes/kubelet
dial socket /mnt/paas/kubernetes/kubelet/plugins_registry/disk.csi.everest.io-reg.sock, err: context deadline exceeded"
[0621 14:33:26.820449] 5176 setters.go:74] Using node IP: "192.168.160.181"
[0621 14:33:27.866390] 5176 eviction_manager.go:395] eviction manager: attempting to reclaim memory
[0621 14:33:27.866453] 5176 eviction_manager.go:406] eviction manager: must evict pod(s) to reclaim memory
[0621 14:33:27.866466] 5176 eviction_manager.go:417] eviction manager: eviction thresholds have been met, but no pods are active to evict
[0621 14:33:36.826267] 5176 setters.go:74] Using node IP: "192.168.160.181"
[0621 14:33:37.953876] 5176 eviction_manager.go:395] eviction manager: attempting to reclaim memory
[0621 14:33:37.953941] 5176 eviction_manager.go:406] eviction manager: must evict pod(s) to reclaim memory
[0621 14:33:37.953954] 5176 eviction_manager.go:417] eviction manager: eviction thresholds have been met, but no pods are active to evict
[0621 14:33:46.830638] 5176 setters.go:74] Using node IP: "192.168.160.181"
[0621 14:33:48.041573] 5176 eviction_manager.go:395] eviction manager: attempting to reclaim memory
[0621 14:33:48.041639] 5176 eviction_manager.go:406] eviction manager: must evict pod(s) to reclaim memory
[0621 14:33:48.041654] 5176 eviction_manager.go:417] eviction manager: eviction thresholds have been met, but no pods are active to evict
[0621 14:33:56.842191] 5176 setters.go:74] Using node IP: "192.168.160.181"
[0621 14:33:58.129728] 5176 eviction_manager.go:395] eviction manager: attempting to reclaim memory
[0621 14:33:58.129794] 5176 eviction_manager.go:406] eviction manager: must evict pod(s) to reclaim memory
[0621 14:33:58.129809] 5176 eviction_manager.go:417] eviction manager: eviction thresholds have been met, but no pods are active to evict
[0621 14:34:06.846538] 5176 setters.go:74] Using node IP: "192.168.160.181"
[0621 14:34:08.217755] 5176 eviction_manager.go:395] eviction manager: attempting to reclaim memory
[0621 14:34:08.217832] 5176 eviction_manager.go:406] eviction manager: must evict pod(s) to reclaim memory
[0621 14:34:08.217845] 5176 eviction_manager.go:417] eviction manager: eviction thresholds have been met, but no pods are active to evict

```

The resource usage is normal.

```

top - 14:34:46 up 135 days, 22:42,  2 users,  load average: 0.09, 0.17, 0.17
Tasks: 222 total,  1 running, 221 sleeping,  0 stopped,  0 zombie
%Cpu(s):  0.3 us,  1.4 sy,  0.0 ni, 98.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
MiB Mem : 22.9/15509.0 [|||||||||||||||||||||]
MiB Swap:  0.0/0.0  [

```

The value of `usage_in_bytes` of cgroup in the `/sys/fs/cgroup/memory` directory is abnormal.

```

# cd /sys/fs/cgroup/memory
# cat memory.usage_in_bytes
17618837504

```

Possible Causes

On an Arm node, the kernel of EulerOS 2.8 and 2.9 has a bug, which causes kubelet to evict pods and results in service unavailability.

NOTE

This issue has been resolved in the following versions:

- EulerOS 2.8: kernel-4.19.36-vhulk1907.1.0.h1252.eulerosv2r8.aarch64
- EulerOS 2.9: kernel-4.19.90-vhulk2103.1.0.h819.eulerosv2r9.aarch64

Solution

- If your cluster version is 1.19.16-r0, 1.21.7-r0, 1.23.5-r0, 1.25.1-r0, or later, reset the OS of the node to the latest version.
- If your cluster version does not meet the requirements, upgrade the cluster to the specified version and then reset the node OS to the latest version.

4.4.4 When Container OOM Occurs on the CentOS Node with an Earlier Kernel Version, the Ext4 File System Is Occasionally Suspended

Symptom

If the kernel version of a CentOS 7.6 node is earlier than 3.10.0-1160.66.1.el7.x86_64 and OOM occurs on containers on the node, all containers on the node may fail to be accessed, and processes such as Docker and jdb are in the D state. The fault is rectified after the node is restarted.

```
[<ffffffff99986832>] ? mutex_lock+0x12/0x2f
[<ffffffff9944d243>] do_sync_write+0x93/0xe8
[<ffffffff9944dd38>] vfs_write+0xc8/0x1f8
[<ffffffff9944eb8f>] Sys_write+0x7f/0xf8
[<ffffffff99994f92>] system call fastpath+0x25/0x2a
INFO: task dockerd:4393 blocked for more than 120 seconds.
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this
dockerd      D ffff8bc      0  4393      1 0x00000000
Call Trace:
[<ffffffff99987dd9>] schedule+0x29/0x78
[<ffffffffffc0358885>] wait_transaction_locked+0x85/0xd8 [jbd2]
[<ffffffffff992c6f00>] ? wake_up_atomic_t+0x38/0x38
[<ffffffffffc0358378>] add_transaction_credits+0x278/0x318 [jbd2]
```

Possible Cause

When the memory usage of a service container exceeds its memory limit, cgroup OOM is triggered and the container is terminated by the system kernel. Container cgroup OOM occasionally triggers ext4 file system suspension on CentOS 7, and ext4/jbd2 is permanently suspended due to deadlock. All tasks that perform I/O operation on the file system are affected.

Solution

- Temporary solution: Restart the node to temporarily rectify the fault.
- Long-term solution:
 - If your cluster version is 1.19.16-r0, 1.21.7-r0, 1.23.5-r0, 1.25.1-r0, or later, reset the OS of the node to the latest version.

- If your cluster version does not meet the requirements, upgrade the cluster to the specified version and then reset the node OS to the latest version.

4.4.5 What Should I Do If a DNS Resolution Failure Occurs Due to a Defect in IPVS?

Symptom

In IPVS forwarding mode used in a CCE cluster, packet loss may occur after CoreDNS is upgraded on the node. This results in a Domain Name System (DNS) resolution failure.

Possible Causes

This problem is caused by a defect in IPVS. The community has fixed it in IPVS v5.9-rc1. For details, see [ipvs: queue delayed work to expire no destination connections if expire_nodest_conn=1](#)

Nodes running Ubuntu 22.04 or Huawei Cloud EulerOS 2.0 are not affected by this problem. Nodes running CentOS, Ubuntu18.04, EulerOS 2.5, EulerOS 2.9 (with earlier kernel version), or Huawei Cloud EulerOS 1.1 are affected by this problem.

Solution

- The impact of the IPVS packet loss can be reduced by using NodeLocal DNSCache. For details, see .
- Use unaffected OSs, such as Huawei Cloud EulerOS 2.0 and Ubuntu 22.04.
- If the OS of your node is EulerOS 2.9, check whether the kernel version of the node meets the following requirements (If the kernel version of the node is too early, reset the node to rectify the fault. If the kernel version of the node meets the requirements, the node is not affected by this issue and no further action is required):
 - x86 node: The kernel version is 4.18.0-147.5.1.6.h998.eulerosv2r9.x86_64 or later.
 - Arm node: The kernel version is 4.19.90-vhulk2103.1.0.h990.eulerosv2r9.aarch64 or later.

4.4.6 What Should I Do If the Number of ARP Entries Exceeds the Upper Limit?

Symptom

The ARP cache exceeds the upper limit, resulting in the abnormal inter-container access, for example, the coredns DNS resolution failure.

Possible Causes

The number of ARP entries cached in the containers on the node exceeds the upper limit.

Fault Locating

- If the OS kernel of a node is later than 4.3, **neighbor table overflow** will display in the dmsg log. For details, see [GitHub](#).

```
# dmesg -T
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:55 2023] neighbour: arp_cache: neighbor table overflow!
[Tue May 30 18:35:58 2023] print_fib4_table_status: 7 callbacks suppressed
[Tue May 30 18:35:59 2023] print_fib4_table_status: 23 callbacks suppressed
[Tue May 30 18:36:00 2023] print_fib4_table_status: 16 callbacks suppressed
[Tue May 30 18:36:03 2023] print_fib4_table_status: 7 callbacks suppressed
[Tue May 30 18:36:04 2023] print_fib4_table_status: 17 callbacks suppressed
[Tue May 30 18:37:38 2023] net_ratelimit: 7966 callbacks suppressed
[Tue May 30 18:37:38 2023] neighbour: arp_cache: neighbor table overflow!
```

- If the kernel version of the node OS is earlier than 4.3, **neighbor table overflow** will not display. If **callbacks suppressed** is displayed, the number of ARP entries may exceed the upper limit.

```
[Wed Jun 14 21:08:58 2023] net_ratelimit: 198 callbacks suppressed
[Wed Jun 14 21:09:05 2023] net_ratelimit: 11 callbacks suppressed
[Wed Jun 14 21:12:35 2023] nr_pdlush_threads exported in /proc is scheduled for removal
[Wed Jun 14 21:39:03 2023] net_ratelimit: 337 callbacks suppressed
[Wed Jun 14 21:39:10 2023] net_ratelimit: 236 callbacks suppressed
[Wed Jun 14 22:09:18 2023] net_ratelimit: 53 callbacks suppressed
[Wed Jun 14 22:14:04 2023] net_ratelimit: 266 callbacks suppressed
[Wed Jun 14 22:14:10 2023] net_ratelimit: 350 callbacks suppressed
[Wed Jun 14 22:15:28 2023] net_ratelimit: 81 callbacks suppressed
[Wed Jun 14 22:34:12 2023] net_ratelimit: 178 callbacks suppressed
[Wed Jun 14 22:34:19 2023] net_ratelimit: 18 callbacks suppressed
[Wed Jun 14 22:39:17 2023] net_ratelimit: 95 callbacks suppressed
[Wed Jun 14 22:44:24 2023] net_ratelimit: 135 callbacks suppressed
[Wed Jun 14 22:51:43 2023] ip_finish_output2: No header cache and no neighbour!
```

Solution

The maximum number of non-permanent entries allowed by a node is determined by the **net.ipv4.neigh.default.gc_thresh3** parameter of the kernel. This parameter is not isolated by namespace. The node and containers running on the node share the ARP table size. In containers, set this parameter to **163790**.

How to calculate the kernel parameter

- In CCE Turbo clusters and clusters using the container tunnel networks:
net.ipv4.neigh.default.gc_thresh3 = Number of containers on a single node x Number of available IP addresses on the container subnet (If there are multiple container subnets in a CCE Turbo cluster, use the maximum number of available IP addresses on a container subnets and the maximum number of containers that can be deployed on a single node.)

For example, if a container subnet is **192.168.0.1/20**, there will be 4,096 IP addresses available and there can be at most 35 containers deployed on a single node, so you can set **net.ipv4.neigh.default.gc_thresh3** to **143360** (4096 x 35).

- Clusters using the VPC networks
net.ipv4.neigh.default.gc_thresh3 = Number of containers on a single node squared

For example, if the subnet mask of a node is 25, there will be 128 container IP addresses available, so you can set **net.ipv4.neigh.default.gc_thresh3** to **16384** (128 x 128).

NOTE

The preceding formulas are used in extreme scenarios.

1. All containers on a node proactively access all IP addresses in the container CIDR block. For example, a gateway container needs to access all other containers in the same cluster.
2. All available IP addresses in a container CIDR block are used up.

Step 1 In **88-k8s.conf**, change the value of **net.ipv4.neigh.default.gc_thresh3** to **163790**.

```
vi /etc/sysctl.d/88-k8s.conf
```

NOTE

The **net.ipv4.neigh.default.gc_thresh1** and **net.ipv4.neigh.default.gc_thresh2** parameters cannot be modified.

Step 2 Run the following command to reload the configuration file:

```
sysctl -p /etc/sysctl.d/88-k8s.conf
```

Step 3 Check whether the configuration takes effect.

```
sysctl -a | grep gc_thresh3
```

```
[root@xxxxxxxxx-turbo-readinessgate-08342-r05vt ~]# sysctl -a | grep gc_thresh3
net.ipv4.neigh.default.gc_thresh3 = 163790
net.ipv6.neigh.default.gc_thresh3 = 1024
```

----End

5 Node Pool

5.1 What Should I Do If a Node Pool Is Abnormal?

Fault Locating

Locate the fault based on the status of the abnormal node pool, as shown in [Table 5-1](#).

Table 5-1 Node pool exceptions

Abnormal Node Pool Status	Description	Solution
Error	The node pool cannot be deleted.	Delete the node pool again. If the node pool still cannot be deleted, submit a service ticket and delete the node pool.
QuotaInsufficient	The node pool cannot be scaled out due to insufficient quota.	Submit a service ticket and increase the quota.
SoldOut	The underlying resources are insufficient.	Update the node pool configuration and select other available resources.

Abnormal Node Pool Status	Description	Solution
Configuration invalid	<p>The ECS group does not exist (ServerGroupNotExists).</p> <p>The ECS group to which the node pool belongs is not present. This may be because you manually deleted the ECS group.</p>	<ol style="list-style-type: none"> 1. Log in to the CCE console. In the navigation pane, choose Nodes, click the Node Pools tab, and click the name of the target node pool. Click the Overview tab, click Expand, and check the ECS group to which the node pool belongs. 2. Log in to the ECS console. In the navigation pane, choose Elastic Cloud Server > ECS Group and see if the target ECS group is present. 3. If the ECS group is not present, log in to the CCE console. In the navigation pane, choose Nodes, click the Node Pools tab, locate the row containing the target node pool, and click Update. In the Advanced Settings area, unbind or change the ECS group.

5.2 What Should I Do If No Node Creation Record Is Displayed When the Node Pool Is Being Expanding?

Symptom

The node pool keeps being in the expanding state, but no node creation record is displayed in the operation record.

Troubleshooting

Check and rectify the following faults:

- Whether a tenant is in arrears.
- Whether the specifications configured for the node pool are insufficient.
- Whether the ECS or memory quota of the tenant is insufficient.
- The ECS capacity verification of the tenant may fail if too many nodes are created at a time.

Solution

- If the tenant is in arrears, renew the account as soon as possible.
- If the resources of the ECS flavor cannot meet service requirements, use ECSs of another flavor.
- If the ECS or memory quota is insufficient, increase the quota.

- If the ECS capacity verification fails, perform the verification again.

5.3 What Should I Do If a Node Pool Scale-Out Fails?

Fault Locating

Locate the fault based on the event of the failure to scale out a node pool, as shown in [Table 5-2](#).

Table 5-2 Node pool scale-out failure

Event	Possible Cause	Reference
...call fsp to query keypair fail, error code : Ecs.0314, reason is : the keypair *** does not match the user_id ***...	<p>The possible causes are as follows:</p> <ul style="list-style-type: none"> • The key pair selected for logging in to the node pool has been deleted. • The key pair selected for logging in to the node pool is a private one which cannot be used by the current user to log in to the node pool and create nodes in the node pool. 	Failed to Obtain the Key Pair Used for Logging In to a Node Pool
{"error": {"message": "encrypted key id [***] is invalid.", "code": "Ecs.0912"}}	<p>The possible causes are as follows:</p> <ul style="list-style-type: none"> • The KMS key ID entered during node pool creation does not exist. • The KMS key ID entered during node pool creation is the key of another user, but the user has not authorized you. 	Invalid KMS Key ID

Failed to Obtain the Key Pair Used for Logging In to a Node Pool

If a node pool scale-out fails, the event contains **Ecs.0314**. This error code indicates that the key pair used for logging in to the node pool cannot be obtained, which results in the creation failure of a new ECS.

```
...call fsp to query keypair fail, error code : Ecs.0314, reason is : the keypair *** does not match the user_id ***
...
```

The possible causes are as follows:

- The key pair selected for logging in to the node pool has been deleted.

- The key pair selected for logging in to the node pool is a private one which cannot be used by the current user to log in to the node pool and create nodes in the node pool.

Solution:

- If the scale-out fails due to the first cause, you can create a key pair and then create a node pool which can be logged in to using this key pair.
- If the scale-out fails due to the second cause, only the user who created the private key pair can scale out the node pool. You can use another key pair when creating a new node pool.

Invalid KMS Key ID

When a node pool fails to be expanded, the reported event contains **Ecs.0912**.

```
{"error":{"message":"encrypted key id [***] is invalid.,"code":"Ecs.0912"}}
```

The possible causes are as follows:

- The KMS key ID entered during node pool creation does not exist.
- The KMS key ID entered during node pool creation is the key of another user, but the user has not authorized you.

Solution:

- If the scale-out fails due to the first cause, ensure that the entered key ID exists.
- If the scale-out fails due to second cause, use the ID of the shared key that has been authorized to you.

5.4 What Should I Do If Some Kubernetes Events Fail to Display After Nodes Were Added to or Deleted from a Node Pool in Batches?

Symptom

After nodes were scaled in or out in a node pool in batches, some Kubernetes events failed to display.

For instance, if 10 nodes were deleted from a cluster in batches, 10 **Delete node** events are printed by CCE, while only 4 **StartScaleDownEmpty** Kubernetes events were printed.

View Events ×

Dec 20, 2023 11:33:44 GMT+0...	● Delete node 192.168.10.198 successfully[id:6c3234...	--
Dec 20, 2023 11:33:44 GMT+0...	● Delete node 192.168.10.125 successfully[id:6c322e...	--
Dec 20, 2023 11:33:44 GMT+0...	● Delete node 192.168.10.133 successfully[id:6c3237...	--
Dec 20, 2023 11:33:44 GMT+0...	● Delete node 192.168.10.39 successfully[id:6c322f...	--
Dec 20, 2023 11:33:44 GMT+0...	● Delete node 192.168.10.201 successfully[id:6c3232...	--
Dec 20, 2023 11:33:30 GMT+0...	● Delete node 192.168.10.220 successfully[id:6c3229...	--
Dec 20, 2023 11:33:30 GMT+0...	● Delete node 192.168.10.60 successfully[id:6c32310...	--
Dec 20, 2023 11:33:30 GMT+0...	● Delete node 192.168.10.228 successfully[id:6c3236...	--
Dec 20, 2023 11:33:30 GMT+0...	● Delete node 192.168.10.76 successfully[id:6c322d0...	--
Dec 20, 2023 11:33:30 GMT+0...	● Delete node 192.168.10.17 successfully[id:6c32338...	--
Dec 20, 2023 11:32:06 GMT+0...	● StartScaleDownEmpty	--
Dec 20, 2023 11:32:06 GMT+0...	● StartScaleDownEmpty	--
Dec 20, 2023 11:32:06 GMT+0...	● StartScaleDownEmpty	--
Dec 20, 2023 11:32:06 GMT+0...	● StartScaleDownEmpty	--
Dec 20, 2023 11:28:31 GMT+0...	● Creation result: 10 succeeded, 0 failed	--

Possible Causes

Kubernetes limits, aggregates, and counts events before printing to ensure the availability of etcd. Therefore, Kubernetes events are not fully printed, particularly when there are numerous identical events being printed.

This is achieved through the [EventCorrelate](#) method in Kubernetes source code. For details, see [design proposal](#) in GitHub.

You do not need to pay attention to this problem, as it is caused by the Kubernetes design mechanism.

5.5 How Do I Modify ECS Configurations When an ECS Cannot Be Managed by a Node Pool?

If an ECS cannot be managed by a node pool due to the reasons listed in this section, you can modify the configuration to manage the ECS.

Cause	Solution	Reference
Inconsistent flavors	Change the ECS flavor to that contained in the node pool.	Modifying the Flavor of an ECS
Inconsistent VPC and subnet	Change the VPC and subnet where the ECS resides to the same VPC and subnet as the node pool.	Changing the VPC and Subnet of an ECS

Cause	Solution	Reference
Different billing modes	Change the billing mode of the ECS to the same as that of the node pool.	Changing the Billing Mode of an ECS
Different data disk configuration	Change the data disk configuration of the ECS to be the same as that of the node pool.	Changing Data Disk Configuration of an ECS
Different enterprise projects	Change the enterprise project of the ECS to be the same as that of the node pool.	Changing the Enterprise Project of an ECS
Different ECS groups	Change the ECS group of the ECS to be the same as that of the node pool.	Change the ECS Group of an ECS

Modifying the Flavor of an ECS

NOTE

The flavor of the ECS to be managed must be changed to that contained in the target node pool.

For more operation guides, see [General Operations](#).

- Step 1** Log in to the ECS console.
- Step 2** Click the name of the target ECS. On the page displayed, click **Stop** in the upper right corner. After the ECS is stopped, choose **More > Modify Specifications** in the **Operation** column.
- Step 3** On the **Modify ECS Specifications** page, select the needed flavor and submit the application.
- Step 4** Go back to the ECS list page and start the ECS.

----End

Changing the VPC and Subnet of an ECS

NOTE

The VPC and subnet to which the ECS to be managed belongs must be changed to the same VPC and subnet as the target node pool.

For details, see [Changing a VPC](#).

- Step 1** Log in to the ECS console.
- Step 2** Locate the row containing the target ECS and choose **More > Manage Network > Change VPC** in the **Operation** column.
- Step 3** Configure the parameters for changing the VPC.

- **VPC:** Select the target VPC.
- **Subnet:** Select the target subnet.
- **Private IP Address:** Select **Assign new** or **Use existing** as required.

Step 4 Click **OK**.

----End

Changing the Billing Mode of an ECS

NOTE

The billing mode of the ECS to be managed must be the same as that of the target node pool.

From Pay-per-Use to Yearly/Monthly

For details, see [Pay-per-Use to Yearly/Monthly](#).

Step 1 Log in to the ECS console.

Step 2 Locate the row containing the target ECS and choose **More > Change Billing Mode** in the **Operation** column.

Step 3 Click **OK**. Then you are switched to Billing Center.

Step 4 Select the usage duration, determine whether to enable auto-renewal, confirm the expected expiration date and price, and click **Pay**.

Step 5 Select a payment method and make your payment. Once the order is paid, yearly/monthly billing is applied.

----End

From Yearly/Monthly to Pay-per-Use

For details, see [Yearly/Monthly to Pay-per-Use](#).

Step 1 Log in to the ECS console.

Step 2 Locate the row containing the target ECS and choose **More > Change to Pay-per-Use > Change to Pay-per-Use Immediately** in the **Operation** column.

Step 3 Click **OK**. Then you are switched to Billing Center.

Step 4 Select the resources to be changed to pay-per-use resources following instructions.

Step 5 Confirm the refund information and click **Change to Pay-Per-Use**.

Step 6 Confirm the resources to be changed to pay-per-use resources again and click **OK**.

----End

Changing Data Disk Configuration of an ECS

NOTE

The number, space, and type of data disks of the ECS to be managed must be the same as those of data disks in the node pool.

Data Disk Number

For more operation guides, see [Adding a Disk to an ECS](#) or [Detaching an EVS Disk from a Running ECS](#).

Step 1 Log in to the ECS console.

Step 2 Click the name of the target ECS to access the ECS details page.

Step 3 Click the **Disks** tab.

- If there are fewer data disks on the node to be managed than the number of data disks configured for the target node pool, you need to add more disks. Click **Add Disk** and configure parameters for the new disk. For details about how to configure EVS disks, see [Step 2: Purchase an EVS Disk](#).

NOTICE

The specifications and space of the new disk must be the same as those configured for the target node pool. You need also select **SCSI** for **Advanced Settings**.

- If there are more data disks on the node to be managed than the number of data disks configured for the target node pool, you need to remove some disks. Click **Detach** on the right of the EVS disk to be removed.

----End

Data Disk Space

For more operation guides, see [Expanding the Capacity of an EVS Disk](#).

Step 1 Log in to the ECS console.

Step 2 Click the name of the target ECS to access the ECS details page.

Step 3 Click the **Disks** tab and click **Expand Capacity** on the right of the EVS disk to be expanded.

Step 4 Configure **New Capacity** following instructions.

Step 5 Click **Next** and submit the order following instructions.

----End

Data Disk Type

For more operation guides, see [Changing the EVS Disk Type \(OBT\)](#).

Step 1 Log in to the ECS console.

Step 2 Click the name of the target ECS to access the ECS details page.

Step 3 Click the **Disks** tab and click **Modify Specifications** on the right of the right of the EVS disk to be expanded.

Step 4 Configure **Disk Type** following instructions.

Step 5 Click **Submit**.

----End

Changing the Enterprise Project of an ECS

NOTE

The enterprise project of the ECS to be managed must be the same as that of the target node pool.

For more operation guides, see [Removing Resources from an Enterprise Project](#).

- Step 1** Log in to the Huawei Cloud management console.
- Step 2** Choose **Enterprise > Project Management** in the upper right corner of the page.
- Step 3** On the page displayed, select an enterprise project and click **View Resource** in the **Operation** column.
- Step 4** Select the resources to be removed and click **Remove**.
- Step 5** Select **ECSs and ECS associated resources**. Resources associated with the ECS will be automatically removed simultaneously.
- Step 6** Select the destination enterprise project and click **OK**.

----End

Change the ECS Group of an ECS

NOTE

The ECS group of the ECS to be managed must be the same as that of the target node pool.

For more operation guides, see [Managing ECS Groups](#).

- Step 1** Log in to the ECS console.
- Step 2** In the navigation pane, choose **Elastic Cloud Server > ECS Group**.
- Step 3** Locate the row containing the target ECS group and click **Add ECS** in the **Operation** column.
- Step 4** In the dialog box displayed, select the ECS to be added.
- Step 5** Click **OK** to add the ECS to the ECS group.

----End

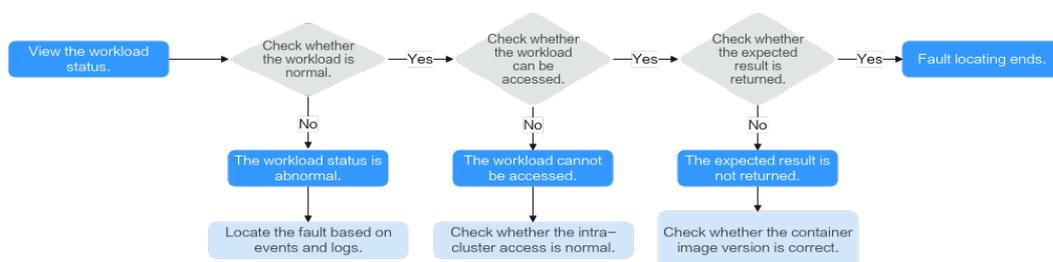
6 Workload

6.1 Workload Abnormalities

6.1.1 How Do I Use Events to Fix Abnormal Workloads?

If a workload is abnormal, you can check the pod events first to locate the fault and then rectify the fault.

Fault Locating



To check whether there is an abnormal pod in the workload, perform the following steps:

- Step 1** Log in to the CCE console.
- Step 2** Click the cluster name to access the cluster console. In the navigation pane, choose **Workloads**.
- Step 3** In the upper left corner of the page, select a namespace, locate the target workload, and view its status.
 - If the workload is not ready, view pod events, and determine the cause. For details, see [Viewing Pod Events](#).
 - If the workload is processing, wait patiently.
 - If the workload is running, no action is required. If the workload status is normal but it cannot be accessed, check whether intra-cluster access is normal.

Log in to the CCE console or use `kubectl` to obtain the pod IP address. Then, log in to the node where this pod locates and run `curl` or use other methods to manually call the APIs. Check whether the expected result is returned.

If `{Container IP address}:{Port}` cannot be accessed, log in to the service container and access `127.0.0.1:{Port}` to locate the fault.

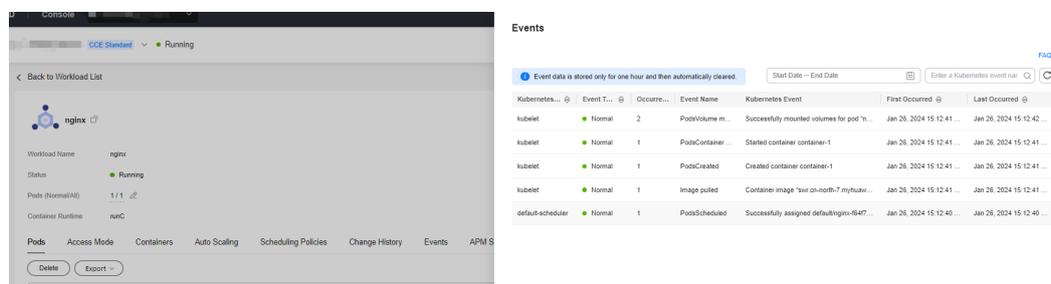
----End

Viewing Pod Events

Method 1

On the CCE console, click the workload name to go to the workload details page, locate the row containing the abnormal pod, and click **View Events** in the **Operation** column.

Figure 6-1 Viewing pod events



Method 2

Run `kubectl describe pod {Pod name}` to view pod events. The following shows an example:

```
$ kubectl describe pod prepare-58bd7bdf9-fthrp
...
Events:
  Type          Reason          Age   From          Message
  ----          -
  Warning       FailedScheduling 49s   default-scheduler  0/2 nodes are available: 2 Insufficient cpu.
  Warning       FailedScheduling 49s   default-scheduler  0/2 nodes are available: 2 Insufficient cpu.
```

Table 6-1 Troubleshooting methods

Event	Pod Status	Solution
PodsScheduling failed	Pending	For details, see What Should I Do If Pod Scheduling Fails?
PodsFailed to pull image Failed to re-pull image	FailedPullImage ImagePullBackOff	For details, see What Should I Do If a Pod Fails to Pull the Image?

Event	Pod Status	Solution
PodsCreation failed Failed to restart container	CreateContainerError CrashLoopBackOff	For details, see What Should I Do If Container Startup Fails?
The pod status is Evicted , and the pod keeps being evicted.	Evicted	For details, see What Should I Do If a Pod Fails to Be Evicted?
The storage volume fails to be mounted to the pod.	Pending	For details, see What Should I Do If a Storage Volume Cannot Be Mounted or the Mounting Times Out?
The pod stays Creating .	Creating	For details, see What Should I Do If a Workload Remains in the Creating State?
The pod stays Terminating .	Terminating	For details, see What Should I Do If Pods in the Terminating State Cannot Be Deleted?
The pod status is Stopped .	Stopped	For details, see What Should I Do If a Workload Is Stopped Caused by Pod Deletion?

6.1.2 What Should I Do If Pod Scheduling Fails?

Fault Locating

If the pod is in the **Pending** state and the event contains pod scheduling failure information, locate the cause based on the event information. For details about how to view events, see [How Do I Use Events to Fix Abnormal Workloads?](#)

Troubleshooting Process

Determine the cause based on the event information, as listed in [Table 6-2](#).

Table 6-2 Pod scheduling failure

Event Information	Cause and Solution
no nodes available to schedule pods.	No node is available in the cluster. Check Item 1: Whether a Node Is Available in the Cluster

Event Information	Cause and Solution
<p>0/2 nodes are available: 2 Insufficient cpu. 0/2 nodes are available: 2 Insufficient memory.</p>	<p>Node resources (CPU and memory) are insufficient. Check Item 2: Whether Node Resources (CPU and Memory) Are Sufficient</p>
<p>0/2 nodes are available: 1 node(s) didn't match node selector, 1 node(s) didn't match pod affinity rules, 1 node(s) didn't match pod affinity/anti-affinity.</p>	<p>The node and pod affinity configurations are mutually exclusive. No node meets the pod requirements. Check Item 3: Affinity and Anti-Affinity Configuration of the Workload</p>
<p>0/2 nodes are available: 2 node(s) had volume node affinity conflict.</p>	<p>The EVS volume mounted to the pod and the node are not in the same AZ. Check Item 4: Whether the Workload's Volume and Node Reside in the Same AZ</p>
<p>0/1 nodes are available: 1 node(s) had taints that the pod didn't tolerate.</p>	<p>Taints exist on the node, but the pod cannot tolerate these taints. Check Item 5: Taint Toleration of Pods</p>
<p>0/7 nodes are available: 7 Insufficient ephemeral-storage.</p>	<p>The ephemeral storage space of the node is insufficient. Check Item 6: Ephemeral Volume Usage</p>
<p>0/1 nodes are available: 1 everest driver not found at node</p>	<p>The everest-csi-driver on the node is not in the running state. Check Item 7: Whether everest Works Properly</p>
<p>Failed to create pod sandbox: ... Create more free space in thin pool or use dm.min_free_space option to change behavior</p>	<p>The node thin pool space is insufficient. Check Item 8: Thin Pool Space</p>
<p>0/1 nodes are available: 1 Too many pods.</p>	<p>The number of pods scheduled to the node exceeded the maximum number allowed by the node. Check Item 9: Number of Pods Scheduled onto the Node</p>

Check Item 1: Whether a Node Is Available in the Cluster

Log in to the CCE console and check whether the node status is **Available**. Alternatively, run the following command to check whether the node status is **Ready**:

```
$ kubectl get node
NAME          STATUS    ROLES    AGE   VERSION
192.168.0.37  Ready    <none>   21d  v1.19.10-r1.0.0-source-121-gb9675686c54267
192.168.0.71  Ready    <none>   21d  v1.19.10-r1.0.0-source-121-gb9675686c54267
```

If the status of all nodes is **Not Ready**, no node is available in the cluster.

Solution

- Add a node. If an affinity policy is not configured for the workload, the pod will be automatically migrated to the new node to ensure that services are running properly.
- Locate the unavailable node and rectify the fault. For details, see [What Should I Do if a Cluster Is Available But Some Nodes Are Unavailable?](#)
- Reset the unavailable node. For details, see [Resetting a Node](#).

Check Item 2: Whether Node Resources (CPU and Memory) Are Sufficient

0/2 nodes are available: 2 Insufficient cpu. This means insufficient CPUs.

0/2 nodes are available: 2 Insufficient memory. This means insufficient memory.

If the resources requested by the pod exceed the allocatable resources of the node where the pod runs, the node cannot provide the resources required to run new pods and pod scheduling onto the node will definitely fail.

Node Name	Status	Node...	Configure	IP Address...	Pods (Allocat...)	CPU Request...	Memory Request...	Runtime Versio...
example2-75620...	Run... Scheduled	DefaultP...	AZ3	10.1.0.55...	8 / 110	52.3% 88.01%	57.36% 93.37%	docker://18.9.0 EulerOS 2.0 (S...

If the number of resources that can be allocated to a node is less than the number of resources that a pod requests, the node does not meet the resource requirements of the pod. As a result, the scheduling fails.

Solution

Add nodes to the cluster. Scale-out is the common solution to insufficient resources.

Check Item 3: Affinity and Anti-Affinity Configuration of the Workload

Inappropriate affinity policies will cause pod scheduling to fail.

Example:

An anti-affinity relationship is established between workload 1 and workload 2. Workload 1 is deployed on node 1 while workload 2 is deployed on node 2.

When you try to deploy workload 3 on node 1 and establish an affinity relationship with workload 2, a conflict occurs, resulting in a workload deployment failure.

0/2 nodes are available: 1 node(s) didn't match **node selector**, 1 node(s) didn't match **pod affinity rules**, 1 node(s) didn't match **pod affinity/anti-affinity**.

- **node selector** indicates that the node affinity is not met.
- **pod affinity rules** indicate that the pod affinity is not met.
- **pod affinity/anti-affinity** indicates that the pod affinity/anti-affinity is not met.

Solution

- When adding workload-workload affinity and workload-node affinity policies, ensure that the two types of policies do not conflict each other. Otherwise, workload deployment will fail.
- If the workload has a node affinity policy, make sure that **supportContainer** in the label of the affinity node is set to **true**. Otherwise, pods cannot be scheduled onto the affinity node and the following event is generated:
No nodes are available that match all of the following predicates: MatchNode Selector, NodeNotSupportsContainer

If the value is **false**, the scheduling fails.

Check Item 4: Whether the Workload's Volume and Node Reside in the Same AZ

0/2 nodes are available: 2 node(s) had volume node affinity conflict. An affinity conflict occurs between volumes and nodes. As a result, the scheduling fails.

This is because EVS disks cannot be attached to nodes across AZs. For example, if the EVS volume is located in AZ 1 and the node is located in AZ 2, scheduling fails.

The EVS volume created on CCE has affinity settings by default, as shown below.

```
kind: PersistentVolume
apiVersion: v1
metadata:
  name: pvc-c29bfac7-efa3-40e6-b8d6-229d8a5372ac
spec:
  ...
  nodeAffinity:
    required:
      nodeSelectorTerms:
        - matchExpressions:
            - key: failure-domain.beta.kubernetes.io/zone
              operator: In
              values:
                - ap-southeast-1a
```

Solution

In the AZ where the workload's node resides, create a volume. Alternatively, create an identical workload and select an automatically assigned cloud storage volume.

Check Item 5: Taint Toleration of Pods

0/1 nodes are available: 1 node(s) had taints that the pod didn't tolerate. This means the node is tainted and the pod cannot be scheduled to the node.

Check the taints on the node. If the following information is displayed, taints exist on the node:

```
$ kubectl describe node 192.168.0.37
Name:          192.168.0.37
...
Taints:        key1=value1:NoSchedule
...
```

In some cases, the system automatically adds a taint to a node. The current built-in taints include:

- `node.kubernetes.io/not-ready`: The node is not ready.
- `node.kubernetes.io/unreachable`: The node controller cannot access the node.
- `node.kubernetes.io/memory-pressure`: The node has memory pressure.
- `node.kubernetes.io/disk-pressure`: The node has disk pressure. Follow the instructions described in [Check Item 4: Whether the Node Disk Space Is Insufficient](#) to handle it.
- `node.kubernetes.io/pid-pressure`: The node is under PID pressure. Follow the instructions in [Changing Process ID Limits \(kernel.pid_max\)](#) to handle it.
- `node.kubernetes.io/network-unavailable`: The node network is unavailable.
- `node.kubernetes.io/unschedulable`: The node cannot be scheduled.
- `node.cloudprovider.kubernetes.io/uninitialized`: If an external cloud platform driver is specified when kubelet is started, kubelet adds a taint to the current node and marks it as unavailable. After **cloud-controller-manager** initializes the node, kubelet deletes the taint.

Solution

To schedule the pod to the node, use either of the following methods:

- If the taint is added by a user, you can delete the taint on the node. If the taint is **automatically added by the system**, the taint will be automatically deleted after the fault is rectified.
- Specify a toleration for the pod containing the taint. For details, see [Taints and Tolerations](#).

```
apiVersion: v1
kind: Pod
metadata:
  name: nginx
spec:
  containers:
  - name: nginx
    image: nginx:alpine
  tolerations:
  - key: "key1"
    operator: "Equal"
    value: "value1"
    effect: "NoSchedule"
```

Check Item 6: Ephemeral Volume Usage

0/7 nodes are available: 7 Insufficient ephemeral-storage. This means insufficient ephemeral storage of the node.

Check whether the size of the ephemeral volume in the pod is limited. If the size of the ephemeral volume required by the application exceeds the existing capacity of the node, the application cannot be scheduled. To solve this problem, change the size of the ephemeral volume or expand the disk capacity of the node.

```
apiVersion: v1
kind: Pod
metadata:
  name: frontend
spec:
  containers:
  - name: app
    image: images.my-company.example/app:v4
    resources:
      requests:
        ephemeral-storage: "2Gi"
      limits:
        ephemeral-storage: "4Gi"
    volumeMounts:
    - name: ephemeral
      mountPath: "/tmp"
  volumes:
  - name: ephemeral
    emptyDir: {}
```

To obtain the total capacity (**Capacity**) and available capacity (**Allocatable**) of the temporary volume mounted to the node, run the **kubectl describe node** command, and view the application value and limit value of the temporary volume mounted to the node.

The following is an example of the output:

```
...
Capacity:
  cpu:          4
  ephemeral-storage: 61607776Ki
  hugepages-1Gi: 0
  hugepages-2Mi: 0
  localssd:     0
  localvolume:  0
  memory:       7614352Ki
  pods:         40
Allocatable:
  cpu:          3920m
  ephemeral-storage: 56777726268
  hugepages-1Gi: 0
  hugepages-2Mi: 0
  localssd:     0
  localvolume:  0
  memory:       6180752Ki
  pods:         40
...
Allocated resources:
(Total limits may be over 100 percent, i.e., overcommitted.)
Resource           Requests          Limits
-----
cpu                 1605m (40%)      6530m (166%)
memory              2625Mi (43%)     5612Mi (92%)
ephemeral-storage  0 (0%)           0 (0%)
hugepages-1Gi      0 (0%)           0 (0%)
hugepages-2Mi      0 (0%)           0 (0%)
localssd            0                0
```

```
localvolume    0    0
Events:        <none>
```

Check Item 7: Whether everest Works Properly

0/1 nodes are available: 1 everest driver not found at node. This means the everest-csi-driver of everest is not started properly on the node.

Check the daemon named **everest-csi-driver** in the kube-system namespace and check whether the pod is started properly. If not, delete the pod. The daemon will restart the pod.

Check Item 8: Thin Pool Space

A data disk dedicated for kubelet and the container engine will be attached to a new node. For details, see [Data Disk Space Allocation](#). If the data disk space is insufficient, the pod cannot be created.

Solution 1: Clearing images

Perform the following operations to clear unused images:

- Nodes that use containerd
 - a. Obtain local images on the node.
`crictl images -v`
 - b. Delete the images that are not required by image ID.
`crictl rmi Image ID`
- Nodes that use Docker
 - a. Obtain local images on the node.
`docker images`
 - b. Delete the images that are not required by image ID.
`docker rmi Image ID`

NOTE

Do not delete system images such as the cce-pause image. Otherwise, pods may fail to be created.

Solution 2: Expanding the disk capacity

To expand a disk capacity, perform the following steps:

Step 1 Expand the capacity of a data disk on the EVS console. For details, see [Expanding EVS Disk Capacity](#).

Only the storage capacity of the EVS disk is expanded. You also need to perform the following steps to expand the capacity of the logical volume and file system.

Step 2 Log in to the CCE console and click the cluster. In the navigation pane, choose **Nodes**. Click **More > Sync Server Data** in the row containing the target node.

Step 3 Log in to the target node.

Step 4 Run the **lsblk** command to check the block device information of the node.

A data disk is divided depending on the container storage **Rootfs**:

Overlayfs: No independent thin pool is allocated. Image data is stored in **dockersys**.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda                  8:0    0  50G  0 disk
└─vda1               8:1    0  50G  0 part /
vdb                  8:16    0 200G  0 disk # Data disk has been expanded but not allocated
├─vgpaas-dockersys 253:0    0  90G  0 lvm  /var/lib/containerd # Space used by the
container engine
└─vgpaas-kubernetes 253:1    0  10G  0 lvm  /mnt/paas/kubernetes/kubelet # Space used by
Kubernetes
```

2. Expand the disk capacity.

Add the new disk capacity to the **dockersys** logical volume used by the container engine.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/dockersys* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <90.00 GiB (23039 extents) to <190.00
GiB (48639 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. */dev/vgpaas/dockersys* specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/containerd; on-line resizing required
old_desc_blocks = 12, new_desc_blocks = 24
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

Devicemapper: A thin pool is allocated to store image data.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda                  8:0    0  50G  0 disk
└─vda1               8:1    0  50G  0 part /
vdb                  8:16    0 200G  0 disk
├─vgpaas-dockersys 253:0    0  18G  0 lvm  /var/lib/docker
├─vgpaas-thinpool_tmeta 253:1    0   3G  0 lvm
├─vgpaas-thinpool    253:3    0  67G  0 lvm # Space used by thinpool
├─...
├─vgpaas-thinpool_tdata 253:2    0  67G  0 lvm
├─vgpaas-thinpool    253:3    0  67G  0 lvm
├─...
└─vgpaas-kubernetes 253:4    0  10G  0 lvm  /mnt/paas/kubernetes/kubelet
```

2. Expand the disk capacity.

Option 1: Add the new disk capacity to the thin pool disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where thinpool is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/thinpool* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/thinpool
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/thinpool changed from <67.00 GiB (23039 extents) to <167.00 GiB (48639 extents).
Logical volume vgpaas/thinpool successfully resized.
```

- c. Do not need to adjust the size of the file system, because the thin pool is not mounted to any devices.

Option 2: Add the new disk capacity to the **dockersys** disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/dockersys* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <18.00 GiB (7679 extents) to <118.00 GiB (33279 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. */dev/vgpaas/dockersys* specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/docker; on-line resizing required
old_desc_blocks = 4, new_desc_blocks = 16
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

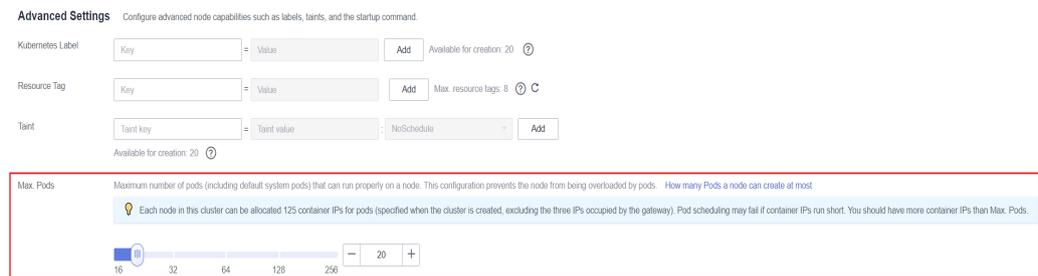
----End

Check Item 9: Number of Pods Scheduled onto the Node

0/1 nodes are available: 1 Too many pods. indicates excessive number of pods have been scheduled to the node.

When creating a node, configure **Max. Pods** in **Advanced Settings** to specify the maximum number of pods that can run properly on the node. The default value varies with the node flavor. You can change the value as needed.

Figure 6-2 Maximum number of pods



On the **Nodes** page, obtain the **Pods (Allocated/Total)** value of the node, and check whether the number of pods scheduled onto the node has reached the upper limit. If so, add nodes or change the maximum number of pods.

To change the maximum number of pods that can run on a node, do as follows:

- For nodes in the default node pool: Change the **Max. Pods** value when resetting the node.
- For nodes in a customized node pool: Change the value of the node pool parameter **max-pods**. For details, see [Configuring a Node Pool](#).

Figure 6-3 Checking the number of pods



6.1.3 What Should I Do If a Pod Fails to Pull the Image?

Fault Locating

When a workload enters the state of "Pod not ready: Back-off pulling image "xxxxx", a Kubernetes event of **PodsFailed to pull image** or **Failed to re-pull image** will be reported. For details about how to view Kubernetes events, see [Viewing Pod Events](#).

Troubleshooting Process

Determine the cause based on the event information, as listed in [Table 6-3](#).

Table 6-3 FailedPullImage

Event Information	Cause and Solution
Failed to pull image "xxx": rpc error: code = Unknown desc = Error response from daemon: Get xxx: denied: You may not login yet	You have not logged in to the image repository. Check Item 1: Whether imagePullSecret Is Specified When You Use kubectl to Create a Workload
Failed to pull image "nginx:v1.1": rpc error: code = Unknown desc = Error response from daemon: Get https://registry-1.docker.io/v2/: dial tcp: lookup registry-1.docker.io: no such host	The image address is incorrectly configured. Check Item 2: Whether the Image Address Is Correct When a Third-Party Image Is Used Check Item 3: Whether an Incorrect Secret Is Used When a Third-Party Image Is Used
Failed create pod sandbox: rpc error: code = Unknown desc = failed to create a sandbox for pod "nginx-6dc48bf8b6-l8xrw": Error response from daemon: mkdir xxxxx: no space left on device	The disk space is insufficient. Check Item 4: Whether the Node Disk Space Is Insufficient
Failed to pull image "xxx": rpc error: code = Unknown desc = error pulling image configuration: xxx x509: certificate signed by unknown authority	An unknown or insecure certificate is used by the third-party image repository from which the image is pulled. Check Item 5: Whether the Remote Image Repository Uses an Unknown or Insecure Certificate
Failed to pull image "xxx": rpc error: code = Unknown desc = context canceled	The image size is too large. Check Item 6: Whether the Image Size Is Too Large
Failed to pull image "docker.io/bitnami/nginx:1.22.0-debian-11-r3": rpc error: code = Unknown desc = Error response from daemon: Get https://registry-1.docker.io/v2/: net/http: request canceled while waiting for connection (Client.Timeout exceeded while awaiting headers)	Check Item 7: Connection to the Image Repository

Event Information	Cause and Solution
<p>ERROR: toomanyrequests: Too Many Requests.</p> <p>Or</p> <p>you have reached your pull rate limit, you may increase the limit by authenticating an upgrading</p>	<p>The rate is limited because the number of image pull times reaches the upper limit.</p> <p>Check Item 8: Whether the Number of Public Image Pull Times Reaches the Upper Limit</p>

Check Item 1: Whether imagePullSecret Is Specified When You Use kubectl to Create a Workload

If the workload status is abnormal and a Kubernetes event is displayed indicating that the pod fails to pull the image, check whether the **imagePullSecrets** field exists in the YAML file.

Items to Check

- If an image needs to be pulled from SWR, the **name** parameter must be set to **default-secret**.

```

apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx
spec:
  replicas: 1
  selector:
    matchLabels:
      app: nginx
  strategy:
    type: RollingUpdate
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
        - image: nginx
          imagePullPolicy: Always
          name: nginx
          imagePullSecrets:
            - name: default-secret

```

- If an image needs to be pulled from a third-party image repository, the **imagePullSecrets** parameter must be set to the created secret name.

When you use kubectl to create a workload from a third-party image, specify the **imagePullSecret** field, in which **name** indicates the name of the secret used to pull the image. For details about how to create a secret, see [Using kubectl](#).

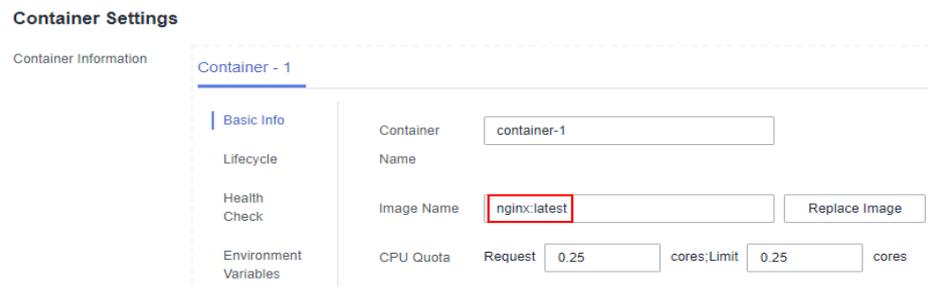
Check Item 2: Whether the Image Address Is Correct When a Third-Party Image Is Used

CCE allows you to create workloads using images pulled from third-party image repositories.

Enter the third-party image address according to requirements. The format must be **ip:port/path/name:version** or **name:version**. If no tag is specified, **latest** is used by default.

- For a private repository, enter an image address in the format of **ip:port/path/name:version**.
- For an open-source Docker repository, enter an image address in the format of **name:version**, for example, **nginx:latest**.

Figure 6-4 Using a third-party image



The following information is displayed when you fail to pull an image due to incorrect image address provided.

```
Failed to pull image "nginx:v1.1": rpc error: code = Unknown desc = Error response from daemon: Get https://registry-1.docker.io/v2/: dial tcp: lookup registry-1.docker.io: no such host
```

Solution

You can either edit your YAML file to change the image address or log in to the CCE console to replace the image on the **Upgrade** tab on the workload details page.

Check Item 3: Whether an Incorrect Secret Is Used When a Third-Party Image Is Used

Generally, a third-party image repository can be accessed only after authentication (using your account and password). CCE uses the secret authentication mode to pull images. Therefore, you need to create a secret for an image repository before pulling images from the repository.

Solution

If your secret is incorrect, images will fail to be pulled. In this case, create a new secret.

To create a secret, see [Using kubectl](#).

Check Item 4: Whether the Node Disk Space Is Insufficient

If the Kubernetes event contains information "no space left on device", there is no disk space left for storing the image. As a result, the image will fail to be pulled. In this case, clear the image or expand the disk space to resolve this issue.

```
Failed create pod sandbox: rpc error: code = Unknown desc = failed to create a sandbox for pod "nginx-6dc48bf8b6-l8xrw": Error response from daemon: mkdir xxxxx: no space left on device
```

Run the following command to obtain the disk space for storing images on a node:

```
lvs
```

```
[root@zhouxu-20650 ~]# lvs
LV          VG      Attr      LSize   Pool Origin Data%  Meta%  Move Log Cpy%Sync Convert
kubernetes  vgpaas  -wi-ao--- <10.00g
thinpool    vgpaas  twi-aot--- 84.00g  5.05  0.07
```

Solution 1: Clearing images

Perform the following operations to clear unused images:

- Nodes that use containerd
 - a. Obtain local images on the node.


```
crictl images -v
```
 - b. Delete the images that are not required by image ID.


```
crictl rmi Image ID
```
- Nodes that use Docker
 - a. Obtain local images on the node.


```
docker images
```
 - b. Delete the images that are not required by image ID.


```
docker rmi Image ID
```

NOTE

Do not delete system images such as the cce-pause image. Otherwise, pods may fail to be created.

Solution 2: Expanding the disk capacity

To expand a disk capacity, perform the following steps:

Step 1 Expand the capacity of a data disk on the EVS console. For details, see [Expanding EVS Disk Capacity](#).

Only the storage capacity of the EVS disk is expanded. You also need to perform the following steps to expand the capacity of the logical volume and file system.

Step 2 Log in to the CCE console and click the cluster. In the navigation pane, choose **Nodes**. Click **More > Sync Server Data** in the row containing the target node.

Step 3 Log in to the target node.

Step 4 Run the **lsblk** command to check the block device information of the node.

A data disk is divided depending on the container storage **Rootfs**:

Overlayfs: No independent thin pool is allocated. Image data is stored in **dockersys**.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME          MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda           8:0    0  50G  0 disk
├─vda1        8:1    0  50G  0 part /
└─vdb         8:16   0 200G  0 disk # Data disk has been expanded but not allocated
   └─vgpaas-dockersys 253:0  0  90G  0 lvm  /var/lib/containerd # Space used by the
      container engine
        └─vgpaas-kubernetes 253:1  0  10G  0 lvm  /mnt/paas/kubernetes/kubelet # Space used by
          Kubernetes
```

2. Expand the disk capacity.

Add the new disk capacity to the **dockersys** logical volume used by the container engine.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/dockersys* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <90.00 GiB (23039 extents) to <190.00 GiB (48639 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. */dev/vgpaas/dockersys* specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/containerd; on-line resizing required
old_desc_blocks = 12, new_desc_blocks = 24
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

Devicemapper: A thin pool is allocated to store image data.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME                                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda                                  8:0    0  50G  0 disk
├─vda1                               8:1    0  50G  0 part /
└─vdb                                 8:16   0 200G  0 disk
   └─vgpaas-dockersys                 253:0   0  18G  0 lvm  /var/lib/docker
      └─vgpaas-thinpool_tmeta         253:1   0   3G  0 lvm
         └─vgpaas-thinpool            253:3   0  67G  0 lvm          # Space used by thinpool
            ...
      └─vgpaas-thinpool_tdata         253:2   0  67G  0 lvm
         └─vgpaas-thinpool            253:3   0  67G  0 lvm
            ...
      └─vgpaas-kubernetes             253:4   0  10G  0 lvm  /mnt/paas/kubernetes/kubelet
```

2. Expand the disk capacity.

Option 1: Add the new disk capacity to the thin pool disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where thinpool is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/thinpool* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/thinpool
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/thinpool changed from <67.00 GiB (23039 extents) to <167.00 GiB (48639 extents).
Logical volume vgpaas/thinpool successfully resized.
```

- c. Do not need to adjust the size of the file system, because the thin pool is not mounted to any devices.

Option 2: Add the new disk capacity to the **dockersys** disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. */dev/vdb* specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed  
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. *vgpaas/dockersys* specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <18.00 GiB (7679 extents) to <118.00  
GiB (33279 extents).  
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. */dev/vgpaas/dockersys* specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/docker; on-line resizing required  
old_desc_blocks = 4, new_desc_blocks = 16  
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

----End

Check Item 5: Whether the Remote Image Repository Uses an Unknown or Insecure Certificate

When a pod pulls an image from a third-party image repository that uses an unknown or insecure certificate, the image fails to be pulled from the node. The pod event list contains the event "Failed to pull the image" with the cause "x509: certificate signed by unknown authority".

NOTE

The security of EulerOS 2.9 images is enhanced. Some insecure or expired certificates are removed from the system. It is normal that this error is reported in EulerOS 2.9 but not on some third-party images on other types of nodes. You can also perform the following operations to rectify the fault.

Solution

- Step 1** Check the IP address and port number of the third-party image server for which the error message "unknown authority" is displayed.

You can see the IP address and port number of the third-party image server for which the error is reported in the event information "Failed to pull image".

```
Failed to pull image "bitnami/redis-cluster:latest": rpc error: code = Unknown desc = error pulling image configuration: Get https://production.cloudflare.docker.com/registry-v2/docker/registry/v2/blobs/sha256/e8/e83853f03a2e792614e7c1e6de75d63e2d6d633b4e7c39b9d700792ee50f7b56/data?verify=1636972064-AQbl5RActnudzV%2F3EShZwnqOe8%3D: x509: certificate signed by unknown authority
```

The IP address of the third-party image server is *production.cloudflare.docker.com*, and the default HTTPS port number is *443*.

Step 2 Load the root certificate of the third-party image server to the node where the third-party image is to be downloaded.

Run the following commands on the EulerOS and CentOS nodes with `{server_url}`: `{server_port}` replaced with the IP address and port number obtained in Step 1, for example, **production.cloudflare.docker.com:443**:

If the container engine of the node is containerd, replace **systemctl restart docker** with **systemctl restart containerd**.

```
openssl s_client -showcerts -connect {server_url}:{server_port} < /dev/null | sed -ne '/-BEGIN
CERTIFICATE-/,/-END CERTIFICATE-/p' > /etc/pki/ca-trust/source/anchors/tmp_ca.crt
update-ca-trust
systemctl restart docker
```

Run the following command on Ubuntu nodes:

```
openssl s_client -showcerts -connect {server_url}:{server_port} < /dev/null | sed -ne '/-BEGIN
CERTIFICATE-/,/-END CERTIFICATE-/p' > /usr/local/share/ca-certificates/tmp_ca.crt
update-ca-trust
systemctl restart docker
```

----End

Check Item 6: Whether the Image Size Is Too Large

The pod event list contains the event "Failed to pull image". This may be caused by a large image size.

```
Failed to pull image "XXX": rpc error: code = Unknown desc = context canceled
```

However, the image can be manually pulled by running the **docker pull** command.

Possible Causes

In Kubernetes clusters, there is a default timeout period for pulling images. If the image pulling progress is not updated within a certain period of time, the download will be canceled. If the node performance is poor or the image size is too large, the image may fail to be pulled and the workload may fail to be started.

Solution

- Solution 1 (recommended):
 - a. Log in to the node and manually pull the image.
 - containerd nodes:

```
crictl pull <image-address>
```
 - Docker nodes:

```
docker pull <image-address>
```
 - b. When creating a workload, ensure that **imagePullPolicy** is set to **IfNotPresent** (the default configuration). In this case, the workload uses the image that has been pulled to the local host.
- Solution 2 (applies to clusters of v1.25 or later): Modify the configuration parameters of the node pools. The configuration parameters for nodes in the **DefaultPool** node pool cannot be modified.
 - a. Log in to the CCE console.
 - b. Click the cluster name to access the cluster console. Choose **Nodes** in the navigation pane and click the **Node Pools** tab.

- c. Locate the row that contains the target node pool and click **Manage**.
- d. In the window that slides out from the right, modify the **image-pull-progress-timeout** parameter under **Docker/containerd**. This parameter specifies the timeout interval for pulling an image.
- e. Click **OK**.

Check Item 7: Connection to the Image Repository

Symptom

The following error message is displayed during workload creation:

```
Failed to pull image "docker.io/bitnami/nginx:1.22.0-debian-11-r3": rpc error: code = Unknown desc = Error response from daemon: Get https://registry-1.docker.io/v2/: net/http: request canceled while waiting for connection (Client.Timeout exceeded while awaiting headers)
```

Possible Causes

Failed to connect to the image repository due to the disconnected network. SWR allows you to pull images only from the official Docker repository. For image pulls from other repositories, you need to access the Internet.

Solution

- Bind a public IP address to the node which needs to pull the images.
- Upload the image to SWR and then pull the image from SWR.

Check Item 8: Whether the Number of Public Image Pull Times Reaches the Upper Limit

Symptom

The following error message is displayed during workload creation:

```
ERROR: toomanyrequests: Too Many Requests.
```

Or

```
you have reached your pull rate limit, you may increase the limit by authenticating an upgrading: https://www.docker.com/increase-rate-limits.
```

Possible Causes

Docker Hub sets the maximum number of container image pull requests. For details, see [Understanding Your Docker Hub Rate Limit](#).

Solution

Push the frequently used image to SWR and then pull the image from SWR.

6.1.4 What Should I Do If Container Startup Fails?

Fault Locating

On the details page of a workload, if an event is displayed indicating that the container fails to be started, perform the following steps to locate the fault:

- Step 1** Log in to the node where the abnormal workload is located.

Step 2 Check the ID of the container where the workload pod exits abnormally.

```
docker ps -a | grep $podName
```

Step 3 View the logs of the corresponding container.

```
docker logs $containerID
```

Rectify the fault of the workload based on logs.

Step 4 Check the error logs.

```
cat /var/log/messages | grep $containerID | grep oom
```

Check whether the system OOM is triggered based on the logs.

----End

Troubleshooting Process

Determine the cause based on the event information, as listed in [Table 6-4](#).

Table 6-4 Container startup failure

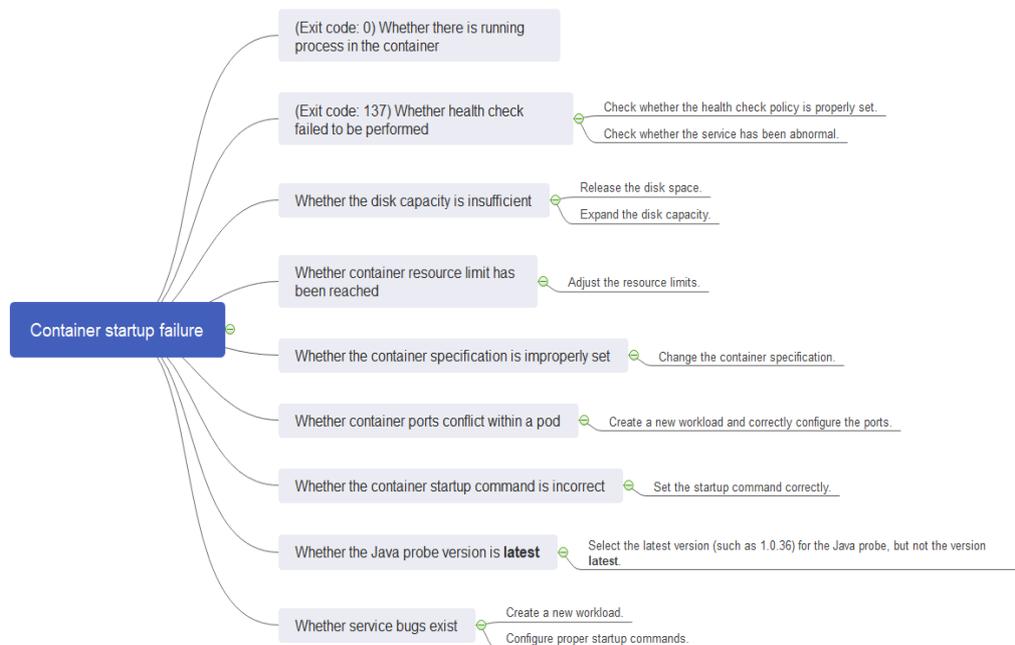
Log or Event	Cause and Solution
The log contains exit(0) .	No process exists in the container. Check whether the container is running properly. Check Item 1: Whether There Are Processes that Keep Running in the Container (Exit Code: 0)
Event information: Liveness probe failed: Get http... The log contains exit(137) .	Health check fails. Check Item 2: Whether Health Check Fails to Be Performed (Exit Code: 137)
Event information: Thin Pool has 15991 free data blocks which are less than minimum required 16383 free data blocks. Create more free space in thin pool or use dm.min_free_space option to change behavior	The disk space is insufficient. Clear the disk space. Check Item 3: Whether the Container Disk Space Is Insufficient
The keyword OOM exists in the log.	The memory is insufficient. Check Item 4: Whether the Upper Limit of Container Resources Has Been Reached Check Item 5: Whether the Resource Limits Are Improperly Configured for the Container

Log or Event	Cause and Solution
Address already in use	A conflict occurs between container ports in the pod. Check Item 6: Whether the Container Ports in the Same Pod Conflict with Each Other
Error: failed to start container "filebeat": Error response from daemon: OCI runtime create failed: container_linux.go:330: starting container process caused "process_linux.go:381: container init caused \"setenv: invalid argument\": unknown	A secret is mounted to the workload, and the value of the secret is not encrypted using Base64. Check Item 7: Whether the Value of the Secret Mounted to the Workload Meets Requirements

In addition to the preceding possible causes, there are some other possible causes:

- **Check Item 8: Whether the Container Startup Command Is Correctly Configured**
- **Check Item 9: Whether the Java Probe Version Is latest**
- **Check Item 10: Whether the User Service Has a Bug**
- Use the correct image when you create a workload on an Arm node.

Figure 6-5 Troubleshooting process of the container restart failure



Check Item 1: Whether There Are Processes that Keep Running in the Container (Exit Code: 0)

Step 1 Log in to the node where the abnormal workload is located.

Step 2 View the container status.

```
docker ps -a | grep $podName
```

Example:

```
Process ID | Name | State | Exit Code | Reason | Time Since Started | Time Since Last State Change
-----|-----|-----|-----|-----|-----|-----
1f99b744e77 | k8s_container-0_test-66b79c9db7-htc | Exited (0) | 10 seconds ago | /bin/bash | 10 seconds ago
```

If no running process exists in the container, the status code **Exited (0)** is displayed.

----End

Check Item 2: Whether Health Check Fails to Be Performed (Exit Code: 137)

The health check configured for a workload is performed on services periodically. If an exception occurs, the pod reports an event and the pod fails to be restarted.

If the liveness-type (workload liveness probe) health check is configured for the workload and the number of health check failures exceeds the threshold, the containers in the pod will be restarted. On the workload details page, if Kubernetes events contain **Liveness probe failed: Get http...**, the health check fails.

Solution

Click the workload name to go to the workload details page, click the **Containers** tab. Then select **Health Check** to check whether the policy is proper or whether services are running properly.

Check Item 3: Whether the Container Disk Space Is Insufficient

The following message refers to the thin pool disk that is allocated from the Docker disk selected during node creation. You can run the **lvs** command as user **root** to view the current disk usage.

Thin Pool has 15991 free data blocks which are less than minimum required 16383 free data blocks. Create more free space in thin pool or use dm.min_free_space option to change behavior

```
# lvs
LV          VG      Attr       LSize   Pool Origin Data%  Meta%  Move Log Cpy%Sync Convert
dockersys  vgpas  -wi-ao---- <18.00g
kubernetes vgpas  -wi-ao---- <18.00g
thinpool   vgpas  twi-aot--- 67.00g   98.04  1.32
```

Solution

Solution 1: Clearing images

Perform the following operations to clear unused images:

- Nodes that use containerd
 - a. Obtain local images on the node.


```
crictl images -v
```

- b. Delete the images that are not required by image ID.
`crictl rmi Image ID`
- Nodes that use Docker
 - a. Obtain local images on the node.
`docker images`
 - b. Delete the images that are not required by image ID.
`docker rmi Image ID`

 **NOTE**

Do not delete system images such as the cce-pause image. Otherwise, pods may fail to be created.

Solution 2: Expanding the disk capacity

To expand a disk capacity, perform the following steps:

- Step 1** Expand the capacity of a data disk on the EVS console. For details, see [Expanding EVS Disk Capacity](#).

Only the storage capacity of the EVS disk is expanded. You also need to perform the following steps to expand the capacity of the logical volume and file system.

- Step 2** Log in to the CCE console and click the cluster. In the navigation pane, choose **Nodes**. Click **More > Sync Server Data** in the row containing the target node.

- Step 3** Log in to the target node.

- Step 4** Run the **lsblk** command to check the block device information of the node.

A data disk is divided depending on the container storage **Rootfs**:

Overlayfs: No independent thin pool is allocated. Image data is stored in **dockersys**.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME        MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda         8:0   0  50G  0 disk
└─vda1      8:1   0  50G  0 part /
vdb         8:16   0 200G  0 disk  # Data disk has been expanded but not allocated
└─vgpaas-dockersys 253:0   0  90G  0 lvm  /var/lib/containerd # Space used by the
  container engine
└─vgpaas-kubernetes 253:1   0  10G  0 lvm  /mnt/paas/kubernetes/kubelet # Space used by
  Kubernetes
```

2. Expand the disk capacity.

Add the new disk capacity to the **dockersys** logical volume used by the container engine.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where dockersys is located.

```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/dockersys` specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <90.00 GiB (23039 extents) to <190.00 GiB (48639 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. `/dev/vgpaas/dockersys` specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/containerd; on-line resizing required
old_desc_blocks = 12, new_desc_blocks = 24
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

Devicemapper: A thin pool is allocated to store image data.

1. Check the disk and partition sizes of the device.

```
# lsblk
NAME                MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
vda                  8:0    0  50G  0 disk
├─vda1                8:1    0  50G  0 part /
└─vgdb                 8:16   0 200G  0 disk
   └─vgpaas-dockersys 253:0   0  18G  0 lvm  /var/lib/docker
      └─vgpaas-thinpool_tmeta 253:1   0   3G  0 lvm
         └─vgpaas-thinpool 253:3   0  67G  0 lvm          # Space used by thinpool
            ...
      └─vgpaas-thinpool_tdata 253:2   0  67G  0 lvm
         └─vgpaas-thinpool 253:3   0  67G  0 lvm
            ...
   └─vgpaas-kubernetes 253:4   0  10G  0 lvm  /mnt/paas/kubernetes/kubelet
```

2. Expand the disk capacity.

Option 1: Add the new disk capacity to the thin pool disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where thinpool is located.


```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/thinpool` specifies the logical volume used by the container engine.


```
lvextend -l+100%FREE -n vgpaas/thinpool
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/thinpool changed from <67.00 GiB (23039 extents) to <167.00 GiB (48639 extents).
Logical volume vgpaas/thinpool successfully resized.
```

- c. Do not need to adjust the size of the file system, because the thin pool is not mounted to any devices.

Option 2: Add the new disk capacity to the **dockersys** disk.

- a. Expand the PV capacity so that LVM can identify the new EVS capacity. `/dev/vdb` specifies the physical volume where dockersys is located.


```
pvresize /dev/vdb
```

Information similar to the following is displayed:

```
Physical volume "/dev/vdb" changed
1 physical volume(s) resized or updated / 0 physical volume(s) not resized
```

- b. Expand 100% of the free capacity to the logical volume. `vgpaas/dockersys` specifies the logical volume used by the container engine.

```
lvextend -l+100%FREE -n vgpaas/dockersys
```

Information similar to the following is displayed:

```
Size of logical volume vgpaas/dockersys changed from <18.00 GiB (7679 extents) to <118.00 GiB (33279 extents).
Logical volume vgpaas/dockersys successfully resized.
```

- c. Adjust the size of the file system. */dev/vgpaas/dockersys* specifies the file system path of the container engine.

```
resize2fs /dev/vgpaas/dockersys
```

Information similar to the following is displayed:

```
Filesystem at /dev/vgpaas/dockersys is mounted on /var/lib/docker; on-line resizing required
old_desc_blocks = 4, new_desc_blocks = 16
The filesystem on /dev/vgpaas/dockersys is now 49807360 (4k) blocks long.
```

----End

Check Item 4: Whether the Upper Limit of Container Resources Has Been Reached

If the upper limit of container resources has been reached, OOM will be displayed in the event details as well as in the log:

```
cat /var/log/messages | grep 96feb0a425d6 | grep oom
```

```
[root@xxx ~]#
[root@xxx ~]# cat /var/log/messages | grep 96feb0a425d6 | grep oom
2019-07-22T11:57:49.441756+08:00 xxx dockerd: time="2019-07-22T11:57:49.440755329+08:00" level=info msg=event OOMKilled=true containerID=96feb0a425d6669f8f062cf3a6096868617a10711334fd5bce4a6ee6eadc82d module=libcontainerd namespace=moby topic=/tasks/oom
2019-07-22T11:59:55.828162+08:00 xxx [/bin/bash]: [2019-07-22T11:57:49.441756+08:00 xxx dockerd: time="2019-07-22T11:57:49.440755329+08:00" level=info msg=event OOMKilled=true containerID=96feb0a425d6669f8f062cf3a6096868617a10711334fd5bce4a6ee6eadc82d module=libcontainerd namespace=moby topic=/tasks/oom] return code=[127], execute failed by [root(uid=0)] from [pts/0 (192.168.0.7)]
2019-07-22T12:01:47.621029+08:00 xxx [/bin/bash]: [cat /var/log/messages | grep 96feb0a425d6 | grep oom] return code=[0], execute success by [root(uid=0)] from [pts/0 (192.168.0.7)]
[root@xxx ~]#
```

When a workload is created, if the requested resources exceed the configured upper limit, the system OOM is triggered and the container exits unexpectedly.

Check Item 5: Whether the Resource Limits Are Improperly Configured for the Container

If the resource limits set for the container during workload creation are less than required, the container fails to be restarted.

Check Item 6: Whether the Container Ports in the Same Pod Conflict with Each Other

Step 1 Log in to the node where the abnormal workload is located.

Step 2 Check the ID of the container where the workload pod exits abnormally.

```
docker ps -a | grep $podName
```

Step 3 View the logs of the corresponding container.

```
docker logs $containerID
```

Rectify the fault of the workload based on logs. As shown in the following figure, container ports in the same pod conflict. As a result, the container fails to be started.

Figure 6-6 Container restart failure due to a container port conflict

```
[root@k8s-01 ~]# docker ps -a|grep test2
aebc17c4d66c          94818572c4ef          "nginx -g 'daemon ..." 8 se
conds ago           Exited (1) 5 seconds ago    k8s_container-1_test2-65dbb945d6-xh9n2_defau
lt_38892324-94b7-11e9-aa5f-fa163e07fc60_3
0c43d629292e          nginx                  "nginx -g 'daemon ..."  Abou
t a minute ago     Up About a minute          k8s_container-0_test2-65dbb945d6-xh9n2_defau
lt_38892324-94b7-11e9-aa5f-fa163e07fc60_0
3484b34393ce          cfe-pause:11.23.1     "/pause"                  Abou
t a minute ago     Up About a minute          k8s_POD_test2-65dbb945d6-xh9n2_default_38892
324-94b7-11e9-aa5f-fa163e07fc60_0
[root@k8s-01 ~]# docker logs aebc17c4d66c
2019/06/22 06:31:29 [emerg] 1#1: bind() to 0.0.0.0:80 failed (98: Address already in use)
nginx: [emerg] bind() to 0.0.0.0:80 failed (98: Address already in use)
2019/06/22 06:31:29 [emerg] 1#1: bind() to 0.0.0.0:80 failed (98: Address already in use)
nginx: [emerg] bind() to 0.0.0.0:80 failed (98: Address already in use)
2019/06/22 06:31:29 [emerg] 1#1: bind() to 0.0.0.0:80 failed (98: Address already in use)
nginx: [emerg] bind() to 0.0.0.0:80 failed (98: Address already in use)
2019/06/22 06:31:29 [emerg] 1#1: bind() to 0.0.0.0:80 failed (98: Address already in use)
nginx: [emerg] bind() to 0.0.0.0:80 failed (98: Address already in use)
2019/06/22 06:31:29 [emerg] 1#1: bind() to 0.0.0.0:80 failed (98: Address already in use)
nginx: [emerg] bind() to 0.0.0.0:80 failed (98: Address already in use)
2019/06/22 06:31:29 [emerg] 1#1: still could not bind()
nginx: [emerg] still could not bind()
```

----End

Solution

Re-create the workload and set a port number that is not used by any other pod.

Check Item 7: Whether the Value of the Secret Mounted to the Workload Meets Requirements

Information similar to the following is displayed in the event:

```
Error: failed to start container "filebeat": Error response from daemon: OCI runtime create failed:
container_linux.go:330: starting container process caused "process_linux.go:381: container init caused
\"setenv: invalid argument\\\": unknown
```

The root cause is that a secret is mounted to the workload, but the value of the secret is not encrypted using Base64.

Solution:

Create a secret on the console. The value of the secret is automatically encrypted using Base64.

If you use YAML to create a secret, you need to manually encrypt its value using Base64.

```
# echo -n "Content to be encoded" | base64
```

Check Item 8: Whether the Container Startup Command Is Correctly Configured

The error messages are as follows:

```
[root@k8s-POD_test1-abc59fc55-8gr9f_defau ~]# docker ps -a | grep test1
2ae258d570c2      94818572c4ef      "/bin/sh -c 'sleep..." 14 s
econds ago      Up 12 seconds      k8s_container-0_test1-abc59fc55-8gr9f_defau
1t_19f0d2a0-94ba-11e9-aa5f-fa163e07fc60_1
492b258c1e89      94818572c4ef      "/bin/sh -c 'sleep..." Abou
t a minute ago Exited (1) 14 seconds ago k8s_container-0_test1-abc59fc55-8gr9f_defau
1t_19f0d2a0-94ba-11e9-aa5f-fa163e07fc60_0
2fcd00990111      cfe-pause:11.23.1      "/pause" Abou
t a minute ago Up About a minute k8s_POD_test1-abc59fc55-8gr9f_defau1t_19f0d
2a0-94ba-11e9-aa5f-fa163e07fc60_0
[root@k8s-POD_test1-abc59fc55-8gr9f_defau ~]# docker logs 492b258c1e89
cat: /tmp/test: No such file or directory
```

Solution

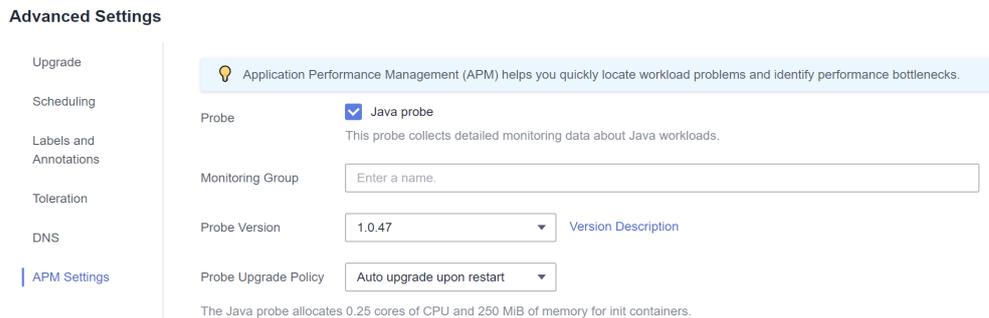
Click the workload name to go to the workload details page, click the **Containers** tab. Choose **Lifecycle**, click **Startup Command**, and ensure that the command is correct.

Check Item 9: Whether the Java Probe Version Is latest

Kubernetes event "Created container init-pinpoint" occurs.

Solution

1. When creating a workload, select the specific latest Java probe version (for example, **1.0.36**, not the **latest** option) on the **APM Settings** tab in the **Advanced Settings** area.



2. If you selected **latest** for the Java probe during workload creation, you can upgrade the workload and change it to the specific latest version (for example, **1.0.36**).

Check Item 10: Whether the User Service Has a Bug

Check whether the workload startup command is correctly executed or whether the workload has a bug.

Step 1 Log in to the node where the abnormal workload is located.

Step 2 Check the ID of the container where the workload pod exits abnormally.

```
docker ps -a | grep $podName
```

Step 3 View the logs of the corresponding container.

```
docker logs $containerID
```

Note: In the preceding command, *containerID* indicates the ID of the container that has exited.

Figure 6-7 Incorrect startup command of the container

```
[root@dcb-ha-11638 ~]# docker ps -a |grep nginx
cf0357f617f9          3f8a4339aadd        "/bin/bash /tmp/test." 2 minutes ago
      Exited (127) 2 minutes ago          k8s_container-0_nginx-267
0177225-kt929_test_d6402ef7-4e0f-11e8-b4f7-fa163e74044e_5
c2176ce394a1         cfe-pause:3.7.6     "/pause"                5 minutes ago
      Up 5 minutes                       k8s_POD_nginx-2670177225-
kt929_test_d6402ef7-4e0f-11e8-b4f7-fa163e74044e_0
[root@dcb-ha-11638 ~]# docker logs cf035
/bin/bash: /tmp/test.sh: No such file or directory
[root@dcb-ha-11638 ~]#
```

As shown in the figure above, the container fails to be started due to an incorrect startup command. For other errors, rectify the bugs based on the logs.

----End

Solution

Create a new workload and configure a correct startup command.

6.1.5 What Should I Do If a Pod Fails to Be Evicted?

Principle of Eviction

When a node is abnormal, Kubernetes will evict pods on the node to ensure workload availability.

In Kubernetes, both kube-controller-manager and kubelet can evict pods.

- **Eviction implemented by kube-controller-manager**

kube-controller-manager consists of multiple controllers, and eviction is implemented by node controller. node controller periodically checks the status of all nodes. If a node is in the **NotReady** state for a period of time, all pods on the node will be evicted.

kube-controller-manager supports the following startup parameters:

- **pod-eviction-timeout**: indicates an interval when a node is down, after which pods on that node are evicted. The default interval is 5 minutes.
- **node-eviction-rate**: indicates the number of nodes to be evicted per second. The default value is **0.1**, indicating that pods are evicted from one node every 10 seconds.
- **secondary-node-eviction-rate**: specifies a rate at which nodes are evicted in the second grade. If a large number of nodes are down in the cluster, the eviction rate will be reduced to **secondary-node-eviction-rate**. The default value is **0.01**.
- **unhealthy-zone-threshold**: specifies a threshold for an AZ to be considered unhealthy. The default value is **0.55**, meaning that if the percentage of faulty nodes in an AZ exceeds 55%, the AZ will be considered unhealthy.
- **large-cluster-size-threshold**: specifies a threshold for a cluster to be considered large. The parameter defaults to **50**. If there are more nodes than this threshold, the cluster is considered as a large one. If there are more than 55% faulty nodes in a cluster, the eviction rate is reduced to 0.01. If the cluster is a small one, the eviction rate is reduced to 0, which means, pods running on the nodes in the cluster will not be evicted.

- **Eviction implemented by kubelet**

If resources of a node are to be used up, kubelet executes the eviction policy based on the pod priority, resource usage, and resource request. If pods have the same priority, the pod that uses the most resources or requests for the most resources will be evicted first.

kube-controller-manager evicts all pods on a faulty node, while kubelet evicts some pods on a faulty node. kubelet periodically checks the memory and disk resources of nodes. If the resources are insufficient, it will evict some pods based on the priority. For details about the pod eviction priority, see [Pod selection for kubelet eviction](#).

There are soft eviction thresholds and hard eviction thresholds.

- **Soft eviction thresholds:** A grace period is configured for node resources. kubelet will reclaim node resources associated with these thresholds if that grace period elapses. If the node resource usage reaches these thresholds but falls below them before the grace period elapses, kubelet will not evict pods on the node.

You can configure soft eviction thresholds using the following parameters:

- **eviction-soft:** indicates a soft eviction threshold. If a node's [eviction signal](#) reaches a certain threshold, for example, **memory.available<1.5Gi**, kubelet will not immediately evict some pods on the node but wait for a grace period configured by **eviction-soft-grace-period**. If the threshold is reached after the grace period elapses, kubelet will evict some pods on the node.
- **eviction-soft-grace-period:** indicates an eviction grace period. If a pod reaches the soft eviction threshold, it will be terminated after the configured grace period elapses. This parameter indicates the time difference for a terminating pod to respond to the threshold being met. The default grace period is 90 seconds.
- **eviction-max-pod-grace-period:** indicates the maximum allowed grace period to use when terminating pods in response to a soft eviction threshold being met.
- **Hard eviction thresholds:** Pods are immediately evicted once these thresholds are reached.

You can configure hard eviction thresholds using the following parameters:

eviction-hard: indicates a hard eviction threshold. When the [eviction signal](#) of a node reaches a certain threshold, for example, **memory.available<1Gi**, which means, when the available memory of the node is less than 1 GiB, a pod eviction will be triggered immediately.

kubelet supports the following default hard eviction thresholds:

- **memory.available<100Mi**
- **nodefs.available<10%**
- **imagefs.available<15%**
- **nodefs.inodesFree<5%** (for Linux nodes)

kubelet also supports other parameters:

- **eviction-pressure-transition-period**: indicates a period for which the kubelet has to wait before transitioning out of an eviction pressure condition. The default value is 5 minutes. If the time exceeds the threshold, the node is set to **DiskPressure** or **MemoryPressure**. Then some pods running on the node will be evicted. This parameter can prevent mistaken eviction decisions when a node is oscillating above and below a soft eviction threshold in some cases.
- **eviction-minimum-reclaim**: indicates the minimum number of resources that must be reclaimed in each eviction. This parameter can prevent kubelet from repeatedly evicting pods because only a small number of resources are reclaimed during pod evictions in some cases.

Fault Locating

If the pods are not evicted when the node is faulty, perform the following steps to locate the fault:

After the following command is run, the command output shows that many pods are in the **Evicted** state.

```
kubectl get pods
```

Check results will be recorded in kubelet logs of the node. You can run the following command to search for the information:

```
cat /var/log/cce/kubernetes/kubelet.log | grep -i Evicted -C3
```

Troubleshooting Process

The issues here are described in order of how likely they are to occur.

Check these causes one by one until you find the cause of the fault.

- [Check Item 1: Whether the Node Is Under Resource Pressure](#)
- [Check Item 2: Whether Tolerations Have Been Configured for the Workload](#)
- [Check Item 3: Whether the Conditions for Stopping Pod Eviction Are Met](#)
- [Check Item 4: Whether the Allocated Resources of the Pod Are the Same as Those of the Node](#)
- [Check Item 5: Whether the Workload Pod Fails Continuously and Is Redeployed](#)

Check Item 1: Whether the Node Is Under Resource Pressure

If a node suffers resource pressure, kubelet will change the **node status** and add taints to the node. Perform the following steps to check whether the corresponding taint exists on the node:

```
$ kubectl describe node 192.168.0.37
Name:          192.168.0.37
...
Taints:        key1=value1:NoSchedule
...
```

Table 6-5 Statuses of nodes with resource pressure and solutions

Node Status	Taint	Eviction Signal	Description	Solution
MemoryPressure	node.kubernetes.io/memory-pressure	memory.available	The available memory on the node reaches the eviction thresholds.	You can scale out node specifications. For details, see How Do I Change the Node Specifications in a CCE Cluster?
DiskPressure	node.kubernetes.io/disk-pressure	nodefs.available, nodefs.inodesFree, imagefs.available or imagefs.inodesFree	The available disk space and inode on the root file system or image file system of the node reach the eviction thresholds.	You can expand the storage space of the node. For details, see Expanding the Storage Space .
PIDPressure	node.kubernetes.io/pid-pressure	pid.available	The available process identifier on the node is below the eviction thresholds.	You can modify the upper limit of PIDs on the node. For details, see Changing Process ID Limits (kernel.pid_max) .

Check Item 2: Whether Tolerations Have Been Configured for the Workload

Use kubectl or locate the row containing the target workload and choose **More > Edit YAML** in the **Operation** column to check whether tolerance is configured for the workload. For details, see [Taints and Tolerations](#).

Check Item 3: Whether the Conditions for Stopping Pod Eviction Are Met

In a cluster that runs less than 50 worker nodes, if the number of faulty nodes accounts for over 55% of the total nodes, the pod eviction will be suspended. In this case, Kubernetes will not attempt to evict the workload on the faulty node. For details, see [Rate limits on eviction](#).

Check Item 4: Whether the Allocated Resources of the Pod Are the Same as Those of the Node

An evicted pod will be frequently scheduled to the original node.

Possible Causes

Pods on a node are evicted based on the node resource usage. The evicted pods are scheduled based on the allocated node resources. Eviction and scheduling are based on different rules. Therefore, an evicted container may be scheduled to the original node again.

Solution

Properly allocate resources to each container.

Check Item 5: Whether the Workload Pod Fails Continuously and Is Redeployed

A workload pod fails and is being redeployed constantly.

Analysis

After a pod is evicted and scheduled to a new node, if pods in that node are also being evicted, the pod will be evicted again. Pods may be evicted repeatedly.

If a pod is evicted by kube-controller-manager, it would be in the **Terminating** state. This pod will be automatically deleted only after the node where the container is located is restored. If the node has been deleted or cannot be restored due to other reasons, you can forcibly delete the pod.

If a pod is evicted by kubelet, it would be in the **Evicted** state. This pod is only used for subsequent fault locating and can be directly deleted.

Solution

Run the following command to delete the evicted pods:

```
kubectl get pods <namespace> | grep Evicted | awk '{print $1}' | xargs kubectl delete pod <namespace>
```

In the preceding command, <namespace> indicates the namespace name. Configure it based on your requirements.

References

[Kubelet does not delete evicted pods](#)

Submitting a Service Ticket

If the problem persists, [submit a service ticket](#).

6.1.6 What Should I Do If a Storage Volume Cannot Be Mounted or the Mounting Times Out?

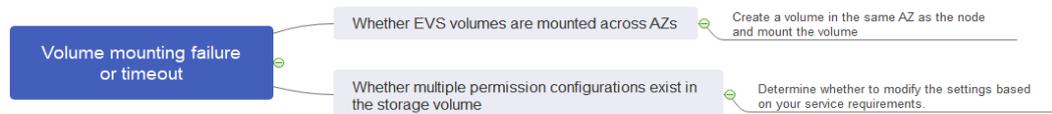
Troubleshooting Process

The issues here are described in order of how likely they are to occur.

Check these causes one by one until you find the cause of the fault.

- [Check Item 1: Whether EVS Volumes Are Mounted Across AZs](#)
- [Check Item 2: Whether Multiple Permission Configurations Exist in the Storage Volume](#)
- [Check Item 3: Whether There Is More Than One Replica for a Deployment with EVS Volumes](#)
- [Check Item 4: Whether the EVS Disk File System Is Damaged](#)

Figure 6-8 Troubleshooting for storage volume mounting failure or mounting timeout



Check Item 1: Whether EVS Volumes Are Mounted Across AZs

Symptom

Mounting an EVS volume to a StatefulSet times out.

Fault Locating

If your node is in **AZ 1** but the volume to be mounted is in **AZ 2**, the mounting times out and the volume cannot be mounted.

Solution

Create a volume in the same AZ as the node and mount the volume.

Check Item 2: Whether Multiple Permission Configurations Exist in the Storage Volume

If the volume to be mounted stores too many data and involves permission-related configurations, the file permissions need to be modified one by one, which results in mounting timeout.

Fault Locating

- Check whether the **securityContext** field contains **runAsuser** and **fsGroup**. **securityContext** is a Kubernetes field that defines the permission and access control settings of pods or containers.
- Check whether the startup commands contain commands used to obtain or modify file permissions, such as **ls**, **chmod**, and **chown**.

Solution

Determine whether to modify the settings based on your service requirements.

Check Item 3: Whether There Is More Than One Replica for a Deployment with EVS Volumes

Symptom

The pod fails to be created, and an event indicating that the storage fails to be added is reported.

```
Multi-Attach error for volume "pvc-62a7a7d9-9dc8-42a2-8366-0f5ef9db5b60" Volume is already used by pod(s) testttt-7b774658cb-lc98h
```

Fault Locating

Check whether the number of replicas of the Deployment is greater than 1.

If the Deployment uses an EVS volume, the number of replicas can only be 1. If you specify more than two pods for the Deployment on the backend, CCE does not restrict the creation of the Deployment. However, if these pods are scheduled to different nodes, some pods cannot be started because the EVS volumes used by the pods cannot be mounted to the nodes.

Solution

Set the number of replicas of the Deployment that uses an EVS volume to 1 or use other volume types.

Check Item 4: Whether the EVS Disk File System Is Damaged

Symptom

The pod fails to be created, and information similar to the following is displayed, indicating that the disk file system is damaged.

```
MountVolume.MountDevice failed for volume "pvc-08178474-c58c-4820-a828-14437d46ba6f" : rpc error: code = Internal desc = [09060def-afd0-11ec-9664-fa163eef47d0] /dev/sda has file system, but it is detected to be damaged
```

Solution

Back up the disk in EVS and run the following command to restore the file system:

```
fsck -y {Drive letter}
```

6.1.7 What Should I Do If a Workload Remains in the Creating State?

Symptom

The workload remains in the creating state.

Troubleshooting Process

The issues here are described in order of how likely they are to occur.

Check these causes one by one until you find the cause of the fault.

- [Check Item 1: Whether the cce-pause Image Is Deleted by Mistake](#)
- [Check Item 2: Modifying Node Specifications After the CPU Management Policy Is Enabled in the Cluster](#)

Check Item 1: Whether the cce-pause Image Is Deleted by Mistake

Symptom

When creating a workload, an error message indicating that the sandbox cannot be created is displayed. This is because the **cce-pause:3.1** image fails to be pulled.

```
Failed to create pod sandbox: rpc error: code = Unknown desc = failed to get sandbox image "cce-pause:3.1": failed to pull image "cce-pause:3.1": failed to pull and unpack image "docker.io/library/cce-pause:3.1": failed to resolve reference "docker.io/library/cce-pause:3.1": pulling from host **** failed with status code [manifests 3.1]: 400 Bad Request
```

Possible Causes

The image is a system image added during node creation. If the image is deleted by mistake, the workload cannot be created.

Solution

Step 1 Log in to the faulty node.

Step 2 Decompress the cce-pause image installation package.

```
tar -xvzf /opt/cloud/cce/package/node-package/pause-*.tgz
```

Step 3 Import the image.

- Docker nodes:

```
docker load -i ./pause/package/image/cce-pause-*.tar
```

- containerd nodes:

```
ctr -n k8s.io images import --all-platforms ./pause/package/image/cce-pause-*.tar
```

Step 4 Create a workload.

----End

Check Item 2: Modifying Node Specifications After the CPU Management Policy Is Enabled in the Cluster

The kubelet option **cpu-manager-policy** defaults to **static**. This allows granting enhanced CPU affinity and exclusivity to pods with certain resource characteristics on the node. If you modify CCE node specifications on the ECS console, the original CPU information does not match the new CPU information. As a result, workloads on the node cannot be restarted or created.

Step 1 Log in to the CCE node (ECS) and delete the **cpu_manager_state** file.

Example command for deleting the file:

```
rm -rf /mnt/paas/kubernetes/kubelet/cpu_manager_state
```

Step 2 Restart the node or kubelet. The following is the kubelet restart command:

```
systemctl restart kubelet
```

Verify that workloads on the node can be successfully restarted or created.

For details, see [What Should I Do If I Fail to Restart or Create Workloads on a Node After Modifying the Node Specifications?](#)

----End

6.1.8 What Should I Do If Pods in the Terminating State Cannot Be Deleted?

Symptom

When a node is in the Unavailable state, CCE migrates container pods on the node and sets the pods running on the node to the **Terminating** state.

After the node is restored, the pods in the **Terminating** state are automatically deleted.

However, some pods remain in the **Terminating** state.

```
#kubectl get pod -n aos
NAME                                READY   STATUS    RESTARTS   AGE
aos-apiserver-5f8f5b5585-s9l92     1/1    Terminating    0         3d1h
aos-cmdbserver-789bf5b497-6rwrq    1/1    Running         0         3d1h
aos-controller-545d78bs8d-vm6j9    1/1    Running         3         3d1h
```

Running **kubectl delete pods <podname> -n <namespace>** cannot delete the pods.

```
kubectl delete pods aos-apiserver-5f8f5b5585-s9l92 -n aos
```

Solution

You can run the following command to forcibly delete the pods created in any ways:

```
kubectl delete pods <pod> --grace-period=0 --force
```

Therefore, run the following command to delete the pod:

```
kubectl delete pods aos-apiserver-5f8f5b5585-s9l92 --grace-period=0 --force
```

6.1.9 What Should I Do If a Workload Is Stopped Caused by Pod Deletion?

Problem

A workload is in **Stopped** state.

Cause:

The **metadata.enable** field in the YAML file of the workload is **false**. As a result, the pod of the workload is deleted and the workload is in the stopped status.

```
kind: Deployment
apiVersion: apps/v1
metadata:
  name: test
  namespace: default
  selfLink: /apis/apps/v1/namespaces/default/deployments/test
  uid: b130db9f-9306-11e9-a2a9-fa163eaff9f7
  resourceVersion: '7314771'
  generation: 1
  creationTimestamp: '2019-06-20T02:54:16Z'
  labels:
    appgroup: ''
  annotations:
    deployment.kubernetes.io/revision: '1'
    description: ''
  enable: false
spec:
```

Solution

Delete the **enable** field or set it to **true**.

6.1.10 What Should I Do If an Error Occurs When Deploying a Service on the GPU Node?

Symptom

The following exceptions occur when services are deployed on the GPU nodes in a CCE cluster:

1. The GPU memory of containers cannot be queried.
2. Seven GPU services are deployed, but only two of them can be accessed properly. Errors are reported during the startup of the remaining five services.
 - The CUDA versions of the two services that can be accessed properly are 10.1 and 10.0, respectively.
 - The CUDA versions of the failing services are also 10.0 and 10.1.
3. Files named **core.*** are found in the GPU service containers. No such files existed in any of the previous deployments.

Fault Locating

1. The driver version of the gpu add-on is too old. After a new driver is downloaded and installed, the fault is rectified.
2. The workloads do not declare that GPU resources are required.

Suggested Solution

After you install `gpu-beta` (`gpu-device-plugin`) on a node, `nvidia-smi` will be automatically installed. If an error is reported during GPU deployment, this issue is

typically caused by an NVIDIA driver installation failure. Check whether the NVIDIA driver has been downloaded.

- GPU node:
If the add-on version is earlier than 2.0.0, run the following command:
`cd /opt/cloud/cce/nvidia/bin && ./nvidia-smi`

If the add-on version is 2.0.0 or later and the driver installation path is changed, run the following command:
`cd /usr/local/nvidia/bin && ./nvidia-smi`
- Container:
`cd /usr/local/nvidia/bin && ./nvidia-smi`

If GPU information is returned, the device is available and the add-on has been installed.

If the driver address is incorrect, uninstall the add-on, reinstall it, and configure the correct address.

NOTE

You are advised to store the NVIDIA driver in the OBS bucket and set the bucket policy to public read.

Helpful Links

- [How Do I Rectify Failures When the NVIDIA Driver Is Used to Start Containers on GPU Nodes?](#)
- [Installing the gpu add-on](#)

6.1.11 Why Does Pod Fail to Write Data?

Pod Events

The file system of the node where the pod is located is damaged. As a result, the newly created pod cannot write data to `/var/lib/kubelet/device-plugins/.xxxxx`. Events similar to the following may occur in the pod:

Message: Pod Update Plugin resources failed due to failed to write checkpoint file "kubelet_internal_checkpoint": open /var/lib/kubelet/device-plugins/.xxxxxx: read-only file system, which is unexpected.

```
[root@10.0.0-213 node]# kubectl describe pod trunlport-test1-d84dc649-zxfpk
Name:          trunlport-test1-d84dc649-zxfpk
Namespace:    default
Priority:      0
Node:         10.0.0.77/
Start Time:   Sat, 20 Feb 2021 16:43:35 +0800
Labels:       app=trunlport-test1
              pod-template-hash=d84dc649
              version=v1
Annotations:  kubernetes.io/psp: psp-global
              metrics.alpha.kubernetes.io/custom-endpoints: [{"api":"","path":"","port":"","names":""}]
Status:       Failed
Reason:       UnexpectedAdmissionError
Message:      Pod Update plugin resources failed due to failed to write checkpoint file "kubelet_internal_checkpoint": open /var/lib/kubelet/device-plugins/.762832416: read-only file system, which is unexpected.
IP:           <none>
IPs:          <none>
Controlled By: ReplicaSet/trunlport-test1-d84dc649
Containers:
  container-0:
    Image:      100.125.4.7:20202/cce-test/tomcat:latest
    Port:       <none>
    Host Port:  <none>
    Limits:
      cpu:      250m
      memory:   512Mi
    Requests:
      cpu:      250m
      memory:   512Mi
    Environment:
      ENV_NAME:  trunlport-test1
```

Such abnormal pods are recorded in error events but do not occupy system resources.

Procedure

There are many causes for file system exceptions, for example, the physical master node is powered on or off unexpectedly. If the file systems are not restored and a large number of pods becomes abnormal (which do not affect services), perform the following steps:

- Step 1** Run the **kubectl drain <node-name>** command to mark the node as unschedulable, and evict existing pods to other nodes.

```
kubectl drain <node-name>
```

- Step 2** Locate the cause of the file system exception and rectify the fault.

- Step 3** Run the following command to make the node schedulable:

```
kubectl uncordon <node-name>
```

----End

Clearing Abnormal Pods

- The garbage collection mechanism of kubelet is the same as that of the community. After the owner (for example, Deployment) of the pod is cleared, the abnormal pod is also cleared.
- You can run the kubelet command to delete the pod recorded as abnormal.

6.1.12 Why Is Pod Creation or Deletion Suspended on a Node Where File Storage Is Mounted?

Symptom

On the node to which SFS or SFS Turbo volumes are mounted, pod deletion tasks stay in the **Stopping** state, and pod creation tasks remain **Creating**.

Possible Causes

- The backend file storage is deleted. As a result, the mount point cannot be accessed.
- The network between the node and the file storage is abnormal. As a result, the mount point cannot be accessed.

Solution

- Step 1** Log in to the node to which the file storage is mounted and run the following command to find the mount path of the file storage:

```
findmnt
```

```
Example mount path: /mnt/paas/kubernetes/kubelet/pods/  
7b88feaf-71d6-4e6f-8965-f5f0766d9f35/volumes/kubernetes.io~csi/sfs-turbo-  
ls/mount
```

- Step 2** Run the following command to access the file storage folder:

```
cd /mnt/paas/kubernetes/kubelet/pods/7b88feaf-71d6-4e6f-8965-f5f0766d9f35/  
volumes/kubernetes.io~csi/sfs-turbo-ls/mount
```

If the access fails, the file storage is deleted or the network between the file storage and the node is abnormal.

Step 3 Run the **umount -l** command to unmount the file storage.

```
umount -l /mnt/paas/kubernetes/kubelet/pods/7b88feaf-71d6-4e6f-8965-  
f5f0766d9f35/volumes/kubernetes.io~csi/sfs-turbo-ls/mount
```

Step 4 Restart kubelet.

```
systemctl restart kubelet
```

----End

Root Cause

This problem usually occurs when the hard mounts are used for file storage. In this mode, all processes that access the mount point are hung until the access is successful. You can use soft mounts to avoid this issue. For details, see [Setting Mount Options](#).

6.1.13 How to Locate Faults Using an Exit Code?

When a container fails to be started or terminated, the exit code is recorded by Kubernetes events to report the cause. This section describes how to locate faults using an exit code.

Viewing an Exit Code

You can use `kubectl` to connect to the cluster and run the following command to check the pod:

```
kubectl describe pod {pod name}
```

In the command output, the **Exit Code** field indicates the status code of the last program exit. If the value is not **0**, the program exits abnormally. You can further analyze the cause through this code.

```
Containers:  
container-1:  
  Container ID: ...  
  Image: ...  
  Image ID: ...  
  Ports: ...  
  Host Ports: ...  
  Args: ...  
  State: Running  
    Started: Sat, 28 Jan 2023 09:06:53 +0000  
  Last State: Terminated  
    Reason: Error  
    Exit Code: 255  
    Started: Sat, 28 Jan 2023 09:01:33 +0000  
    Finished: Sat, 28 Jan 2023 09:05:11 +0000  
  Ready: True  
  Restart Count: 1
```

Description

The exit code ranges from 0 to 255.

- If the exit code is 0, the container exits normally.

- Generally, the abnormal exit is caused by the program and such exit code ranges from 1 to 128. In special scenarios, the exit code ranges from 129 to 255.
- When a program exits due to external interrupts, the exit code ranges from 129 to 255. When the operating system sends **an interrupt signal** to the program, the exist code is the interrupt signal value plus 128. For example, if the interrupt signal value of **SIGKILL** is 9, the exit status code is 137 (9 + 128).
- If the exist code is not in the range of 0 to 255, for example, exit(-1), the exit code is automatically converted to a value that is within the range of 0 to 255.

If the exist code is a positive number, the conversion formula is as follows:

`code % 256`

If the exit code is a negative number, the conversion formula is as follows:

`256 - (|code| % 256)`

For details, see [Exit Codes With Special Meanings](#).

Common Exit Codes

Table 6-6 Common exit codes

Exit Code	Name	Description
0	Normal exit	The container exits normally. This status code does not necessarily indicate that an exception occurs. When there is no process in the container, it may also be displayed.
1	Common program error	There are many causes for this exception, most of which are caused by the program. You need to further locate the cause through container logs. For example, this error occurs when an x86 image is running on an Arm node.
125	The container is not running.	The possible causes are as follows: <ul style="list-style-type: none"> • An undefined flag is used in the command, for example, docker run --abcd. • The user-defined command in the image has insufficient permission on the local host. • The container engine is incompatible with the host OS or hardware.
126	Command calling error	The command called in the image cannot be executed. For example, the file permission is insufficient or the file cannot be executed.
127	The file or directory cannot be found.	The file or directory specified in the image cannot be found.

Exit Code	Name	Description
128	Invalid exit parameter	The container exits but no valid exit code is provided. There are multiple possible causes. You need to further locate the cause. For example, an application running on the containerd node attempts to call the docker command.
137	Immediate termination (SIGKILL)	The program is terminated by the SIGKILL signal. The common causes are as follows: <ul style="list-style-type: none"> • The memory usage of the container in the pod reaches the resource limit. For example, out of memory (OOM) causes cgroup to forcibly stop the container. • If OOM occurs, the kernel of the node stops some processes to release the memory. As a result, the container may be terminated. • If the container health check fails, kubelet stops the container. • Other external processes, such as malicious scripts, forcibly stop the container.
139	Segmentation error (SIGSEGV)	The container receives the SIGSEGV signal from the OS because the container attempts to access an unauthorized memory location.
143	Graceful termination (SIGTERM)	The container is correctly closed as instructed by the host. Generally, this exit code 143 does not require troubleshooting.
255	The exit code is out of range.	The container exit code is out of range. For example, exit(-1) may be used for abnormal exit, and -1 is automatically converted to 255. Further troubleshooting is required.

Linux Standard Interrupt Signal

You can run the **kill -l** command to view the signals and corresponding values in the Linux OS.

Table 6-7 Common Linux standard interrupt signals

Signal	Value	Action	Commit
SIGHUP	1	Term	Sent when the user terminal connection (normal or abnormal) ends.
SIGINT	2	Term	Program termination signal, which is sent by the terminal by pressing Ctrl+C .

Signal	Value	Action	Commit
SIGQUIT	3	Core	Similar to SIGINT , the exit command is sent by the terminal. Generally, the exit command is controlled by pressing Ctrl+\ .
SIGILL	4	Core	Invalid instruction, usually because an error occurs in the executable file.
SIGABRT	6	Core	Signal generated when the abort function is invoked. The process ends abnormally.
SIGFPE	8	Core	A floating-point arithmetic error occurs. Other arithmetic errors such as divisor 0 also occur.
SIGKILL	9	Term	Any process is terminated.
SIGSEGV	11	Core	Attempt to access an unauthorized memory location.
SIGPIPE	13	Term	The pipe is disconnected.
SIGALRM	14	Term	Indicates clock timing.
SIGTERM	15	Term	Process end signal, which is usually the normal exit of the program.
SIGUSR1	10	Term	This is a user-defined signal in applications.
SIGUSR2	12	Term	This is a user-defined signal in applications.
SIGCHLD	17	Ign	This signal is generated when a subprocess ends or is interrupted.
SIGCONT	18	Cont	Resume a stopped process.
SIGSTOP	19	Stop	Suspend the execution of a process.
SIGTSTP	20	Stop	Stop a process.
SIGTTIN	21	Stop	The background process reads the input value from the terminal.
SIGTTOU	22	Stop	The background process reads the output value from the terminal.

6.2 Container Configuration

6.2.1 When Is Pre-stop Processing Used?

Service processing takes a long time. Pre-stop processing makes sure that during an upgrade, a pod is killed only when the service in the pod has been processed.

6.2.2 How Do I Set an FQDN for Accessing a Specified Container in the Same Namespace?

Context

When creating a workload, users can specify a container, pod, and namespace as an FQDN for accessing the container in the same namespace.

FQDN stands for Fully Qualified Domain Name, which contains both the host name and domain name. These two names are combined using a period (.).

For example, if the host name is **bigserver** and the domain name is **mycompany.com**, the FQDN is **bigserver.mycompany.com**.

Solution

Solution 1: Use the domain name for service discovery. The host name and namespace must be pre-configured. The domain name of the registered service is in the format of *service name.namespace name.svc.cluster.local*. The limitation of this solution is that the registration center must be deployed using containers.

Solution 2: Use the host network to deploy containers and then configure affinity between the containers and a node in the cluster. In this way, the service address (that is, the node address) of the containers can be determined. The registered address is the IP address of the node where the service is located. This solution allows you to deploy the registration center using VMs, whereas the disadvantage is that the host network is not as efficient as the container network.

6.2.3 What Should I Do If Health Check Probes Occasionally Fail?

When the liveness and readiness probes fail to perform the health check, locate the service fault first.

Common causes are as follows:

- The service processing takes a long time. As a result, the response times out.
- The Tomcat connection setup and waiting time are too long (for example, too many connections or threads). As a result, the response times out.
- The performance of the node where the container is located, such as the disk I/O, reaches the bottleneck. As a result, the service processing times out.

6.2.4 How Do I Set the umask Value for a Container?

Symptom

A container is started in **tailf /dev/null** mode and the directory permission is **700** after the startup script is manually executed. If the container is started by Kubernetes itself without **tailf**, the obtained directory permission is **751**.

Solution

The reason is that the umask values set in the preceding two startup modes are different. Therefore, the permissions on the created directories are different.

The umask value is used to set the default permission for a newly created file or directory. If the umask value is too small, group users or other users will have excessive permissions, posing security threats to the system. Therefore, the default umask value for all users is set to **0077**. That is, the default permission on directories created by users is **700**, and the default permission on files is **600**.

You can add the following content to the startup script to set the permission on the created directory to **700**:

1. Add **umask 0077** to the **/etc/bashrc** file and all files in **/etc/profile.d/**.
2. Run the following command:

```
echo "umask 0077" >> $FILE
```

NOTE

FILE indicates the file name, for example, **echo "umask 0077" >> /etc/bashrc**.

3. Set the owner and group of the **/etc/bashrc** file and all files in **/etc/profile.d/** to **root**.
4. Run the following command:

```
chown root.root $FILE
```

6.2.5 What Is the Retry Mechanism When CCE Fails to Start a Pod?

CCE is a fully managed Kubernetes service and is fully compatible with Kubernetes APIs and kubectl.

In Kubernetes, the spec of a pod contains a **restartPolicy** field. The value of **restartPolicy** can be **Always**, **OnFailure**, or **Never**. The default value is **Always**.

- **Always**: When a container fails, kubelet automatically restarts the container.
- **OnFailure**: When a container stops running and the exit code is not **0** (indicating normal exit), kubelet automatically restarts the container.
- **Never**: kubelet does not restart the container regardless of the container running status.

restartPolicy applies to all containers in a pod.

restartPolicy only refers to restarts of the containers by kubelet on the same node. When containers in a pod exit, kubelet restarts them with an exponential back-off delay (10s, 20s, 40s, ...), which is capped at five minutes. Once a

container has been running for 10 minutes without any problems, kubelet resets the restart backoff timer for the container.

The settings of **restartPolicy** vary depending on the controller:

- **Replication Controller (RC)** and **DaemonSet**: **restartPolicy** must be set to **Always** to ensure continuous running of the containers.
- **Job**: **restartPolicy** must be set to **OnFailure** or **Never** to ensure that containers are not restarted after being executed.

6.3 Alarm Monitoring

6.3.1 How Long Are the Events of a Workload Stored?

In a cluster of v1.7.3-r12, 1.9.2-r3, or a later version, the event information of a workload is stored for one hour, after which the data is automatically cleared.

In clusters earlier than 1.7.3-r12, events are stored for 24 hours.

6.4 Scheduling Policies

6.4.1 How Do I Evenly Distribute Multiple Pods to Each Node?

The kube-scheduler component in Kubernetes is responsible pod scheduling. For each newly created pod or other unscheduled pods, kube-scheduler selects an optimal node from them to run on. kube-scheduler selects a node for a pod in a 2-step operation: filtering and scoring. In the filtering step, all nodes where it is feasible to schedule the pod are filtered out. In the scoring step, kube-scheduler ranks the remaining nodes to choose the most suitable pod placement. Finally, kube-scheduler schedules the pod to the node with the highest score. If there is more than one node with the equal scores, kube-scheduler selects one of them at random.

BalancedResourceAllocation is only one of the scoring priorities. Other scoring items may also cause uneven distribution. For details about scheduling, see [Kubernetes Scheduler](#) and [Scheduling Policies](#).

You can configure pod anti-affinity policies to evenly distribute pods onto different nodes.

Example:

```
kind: Deployment
apiVersion: apps/v1
metadata:
  name: nginx
  namespace: default
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
```

```
  app: nginx
  spec:
    containers:
      - name: container-0
        image: nginx:alpine
        resources:
          limits:
            cpu: 250m
            memory: 512Mi
          requests:
            cpu: 250m
            memory: 512Mi
    affinity:
      podAntiAffinity:
        # Workload anti-affinity
        preferredDuringSchedulingIgnoredDuringExecution:
          # Ensure that the following conditions are met:
          - weight: 100 # Priority that can be configured when the best-effort policy is used. The value
            ranges from 1 to 100. A larger value indicates a higher priority.
            podAffinityTerm:
              labelSelector:
                # Select the label of the pod, which is anti-affinity with the
                workload.
              matchExpressions:
                - key: app
                  operator: In
                  values:
                    - nginx
              namespaces:
                - default
            topologyKey: kubernetes.io/hostname # It takes effect on the node.
    imagePullSecrets:
      - name: default-secret
```

6.4.2 How Do I Prevent a Container on a Node from Being Evicted?

Context

During workload scheduling, two containers on a node may compete for resources. As a result, kubelet evicts both containers. This section describes how to set a policy to retain one of the containers.

Solution

kubelet uses the following criteria to evict a pod:

- Quality of Service (QoS) class: **BestEffort**, **Burstable**, and **Guaranteed**
- Consumed resources based on the pod scheduling request

Pods of different QoS classes are evicted in the following sequence:

BestEffort -> Burstable -> Guaranteed

- BestEffort pods: These pods have the lowest priority. They will be the first to be killed if the system runs out of memory.
- Burstable pods: These pods will be killed if the system runs out of memory and no BestEffort pods exist.
- Guaranteed pods: These pods will be killed if the system runs out of memory and no Burstable or BestEffort pods exist.

 NOTE

- If a pod is killed because of excessive resource usage (while the node resources are still sufficient), the system tends to restart the pod on the same node.
- If resources are sufficient, you can assign the QoS class of Guaranteed to all pods. In this way, more compute resources are used to improve service performance and stability, reducing troubleshooting time and costs.
- To improve resource utilization, assign the QoS class of Guaranteed to service pods and Burstable or BestEffort to other pods (for example, filebeat).

6.4.3 Why Are Pods Not Evenly Distributed to Nodes?

The kube-scheduler component in Kubernetes is responsible pod scheduling. For each newly created pod or other unscheduled pods, kube-scheduler selects an optimal node from them to run on. kube-scheduler selects a node for a pod in a 2-step operation: filtering and scoring. In the filtering step, all nodes where it is feasible to schedule the pod are filtered out. In the scoring step, kube-scheduler ranks the remaining nodes to choose the most suitable pod placement. Finally, kube-scheduler schedules the pod to the node with the highest score. If there is more than one node with the equal scores, kube-scheduler selects one of them at random.

BalancedResourceAllocation is only one of the scoring priorities. Other scoring items may also cause uneven distribution. For details about scheduling, see [Kubernetes Scheduler](#) and [Scheduling Policies](#).

6.4.4 How Do I Evict All Pods on a Node?

You can run the **kubectl drain** command to safely evict all pods from a node.

 NOTE

By default, the **kubectl drain** command retains some system pods, for example, everest-csi-driver.

Step 1 Use kubectl to connect to the cluster.

Step 2 Check the nodes in the cluster.

```
kubectl get node
```

Step 3 Select a node and view all pods on the node.

```
kubectl get pod --all-namespaces -owide --field-selector spec.nodeName=192.168.0.160
```

The pods on the node before eviction are as follows:

NAMESPACE	NAME	READY	STATUS	RESTARTS	AGE	IP
NODE	NOMINATED NODE	READINESS GATES				
default	nginx-5bcc57c74b-lgcvh	1/1	Running	0	7m25s	10.0.0.140
192.168.0.160	<none>	<none>				
kube-system	coredns-6fcd88c4c-97p6s	1/1	Running	0	3h16m	10.0.0.138
192.168.0.160	<none>	<none>				
kube-system	everest-csi-controller-56796f47cc-99dtm	1/1	Running	0	3h16m	10.0.0.139
192.168.0.160	<none>	<none>				
kube-system	everest-csi-driver-dpfzl	2/2	Running	2	12d	192.168.0.160
192.168.0.160	<none>	<none>				
kube-system	icagent-tpfpv	1/1	Running	1	12d	192.168.0.160
192.168.0.160	<none>	<none>				

Step 4 Evict all pods on the node.

```
kubectl drain 192.168.0.160
```

If a pod mounted with local storage or controlled by a DaemonSet exists on the node, the message "error: unable to drain node "192.168.0.160" due to error: cannot delete DaemonSet-managed Pods..." will be displayed. The eviction command does not take effect. You can add the following parameters to the end of the preceding command to forcibly evict the pod:

- **--delete-emptydir-data**: forcibly evicts pods mounted with local storage, for example, coredns.
- **--ignore-daemonsets**: forcibly evicts the DaemonSet pods, for example, everest-csi-driver.

In the example, both types of pods exist on the node. Therefore, the eviction command is as follows:

```
kubectl drain 192.168.0.160 --delete-emptydir-data --ignore-daemonsets
```

Step 5 After the eviction, the node is automatically marked as unschedulable. That is, the node is tainted **node.kubernetes.io/unschedulable = : NoSchedule**.

After the eviction, only system pods are retained on the node.

NAMESPACE	NAME	READY	STATUS	RESTARTS	AGE	IP	NODE
NOMINATED	NODE	READINESS	GATES				
kube-system	everest-csi-driver-dpfzl	2/2	Running	2	12d	192.168.0.160	192.168.0.160
<none>	<none>						
kube-system	icagent-tpfpv	1/1	Running	1	12d	192.168.0.160	192.168.0.160
<none>	<none>						

----End

Related Operations

Drain, cordon, and uncordon operations of kubectl:

- **drain**: Safely evicts all pods from a node and marks the node as unschedulable.
- **cordon**: Marks the node as unschedulable. That is, the node is tainted **node.kubernetes.io/unschedulable = : NoSchedule**.
- **uncordon**: Marks the node as schedulable.

For more information, see the [kubectl documentation](#).

6.4.5 How Do I Check Whether a Pod Is Bound with CPU Cores?

Take a node with 4 vCPUs and 8 GiB memory as an example. Deploy a workload whose CPU request is 1 and limit is 2 in the cluster in advance.

Step 1 Log in to a node in the node pool and view the **/var/lib/kubelet/cpu_manager_state** output.

```
cat /var/lib/kubelet/cpu_manager_state
```

Information similar to the following will be displayed:

```
{"policyName":"static","defaultCpuSet":"0,2-3","entries":{"c1fcd22d-8a83-4aef-a27a-4c037e482b16":{"container-1":"1"},"checksum":1500530529}}
```

If the value of **policyName** is **static**, the policy has been configured.

Step 2 Check the cgroup setting of **cpuset.preferred_cpus** of the container. The output is the ID of the CPU that is preferentially used.

```
cat /sys/fs/cgroup/cpuset/kubepods/pod {pod uid} / {Container id} /cpuset.cpus
```

- *{pod uid}* indicates the pod UID. It can be obtained by running the following command on the host that has been connected to the cluster using **kubectl**:

```
kubectl get po {pod name} -n {namespace} -ojsonpath='{.metadata.uid}'
```

In the preceding command, *{pod name}* indicates the pod name and *{namespace}* indicates the namespace to which the pod belongs.

- *{Container id}* must be a complete container ID. To obtain the ID, run the following command on the node where the container is running:

Docker node pool: In the command, *{pod name}* indicates the pod name.

```
docker ps --no-trunc | grep {pod name} | grep -v cce-pause | awk '{print $1}'
```

containerd node pool: In the command, *{pod name}* indicates the pod name, *{pod id}* indicates the pod ID, and *{container name}* indicates the container name.

```
# Obtain the pod ID.
```

```
crictl pods | grep {pod name} | awk '{print $1}'
```

```
# Obtain the complete container ID.
```

```
crictl ps --no-trunc | grep {pod id} | grep {container name} | awk '{print $1}'
```

A complete example is as follows:

```
cat /sys/fs/cgroup/cpuset/kubepods/podc1fcd22d-8a83-4aef-  
a27a-4c037e482b16/5cb15f55f429e4496172bef05994477caa96e0ca468563208695c1ad5cc141e0/  
cpuset.cpus
```

The command output shows that CPU 1 is bound.

```
1
```

----End

6.4.6 What Should I Do If Pods cannot Be Rescheduled After the Node Is Stopped?

Symptom

After a node is stopped, pods on the node are still running. The latest pod event obtained by running **kubectl describe pod <pod-name>** is displayed as follows:

```
Warning NodeNotReady 17s node-controller Node is not ready
```

Possible Causes

After a node is stopped, the system automatically adds taints to the node.

- node.kubernetes.io/unreachable:NoExecute
- node.cloudprovider.kubernetes.io/shutdown:NoSchedule
- node.kubernetes.io/unreachable:NoSchedule
- node.kubernetes.io/not-ready:NoExecute

If a pod has tolerations for these taints, it will not be rescheduled. Therefore, check the tolerations of the pod.

Solution

Check the tolerations by viewing the YAML file of the pod or workload. The tolerations of a workload consist of the following fields:

```
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoSchedule"
```

Or:

```
tolerations:
- key: "key1"
  operator: "Exists"
  effect: "NoSchedule"
```

If the preceding tolerations are incorrectly configured, the scheduling may fail. For example:

```
tolerations:
- operator: "Exists"
```

In this example, the **operator** parameter is set to **Exists**. In this case, the **value** parameter cannot be configured.

- If the **operator** parameter of a toleration is set to **Exists** but the **key** parameter is empty, the toleration can match any key, value, and effect. It can tolerate any taint.
- If the **effect** parameter of a toleration is empty but the **key** parameter is configured, the toleration can match the effects of all keys.

For details, see [Taints and Tolerations](#).

Restore the default tolerations configuration by modifying the YAML file of the workload as follows:

```
tolerations:
- key: node.kubernetes.io/not-ready
  operator: Exists
  effect: NoExecute
  tolerationSeconds: 300
- key: node.kubernetes.io/unreachable
  operator: Exists
  effect: NoExecute
  tolerationSeconds: 300
```

This default toleration indicates that the pod can run on the node with the preceding taints for 300s and then be evicted.

6.4.7 How Do I Prevent a Non-GPU or NPU Workload from Being Scheduled to a GPU or NPU Node?

Symptom

If there are GPU/NPU nodes and other types of nodes running in your cluster, the non-NPU/GPU workloads may be scheduled to the GPU/NPU nodes. In this case, the GPU/NPU resources cannot be used properly.

Possible Causes

The non-GPU/NPU workloads use CPU and memory resources. The GPU/NPU nodes can provide these resources. The scheduler may schedule the non-GPU/NPU workloads to these nodes, even if the workloads do not claim to use the GPU/NPU nodes. This may result in the idle GPU/NPU resources.

Solution

Add taints to the GPU/NPU nodes and configure tolerations to prevent non-GPU/NPU workloads from being scheduled to these nodes.

- For the GPU/NPU workloads, add tolerations so that they can be scheduled to the GPU/NPU nodes.
- For the non-GPU/NPU workload, if tolerations are not configured, they cannot be scheduled to the GPU/NPU nodes.

The procedure is as follows:

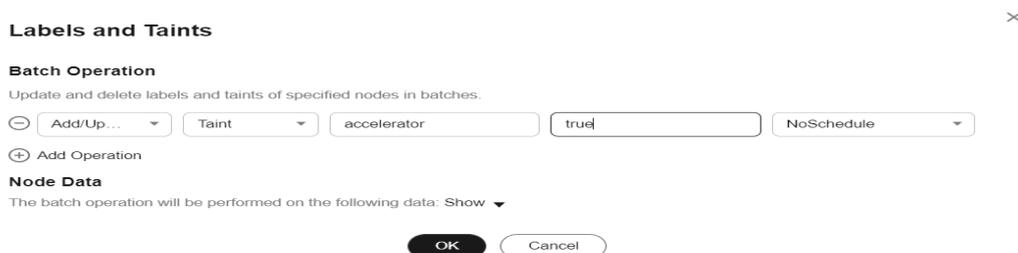
Step 1 Log in to the CCE console and click the cluster name to access the cluster console.

Step 2 In the navigation pane, choose **Nodes**. Click the **Nodes** tab, select a GPU/NPU node, and click **Labels and Taints** above the list.

Step 3 Click **Add Operation** under **Batch Operation** and add a taint to the node.

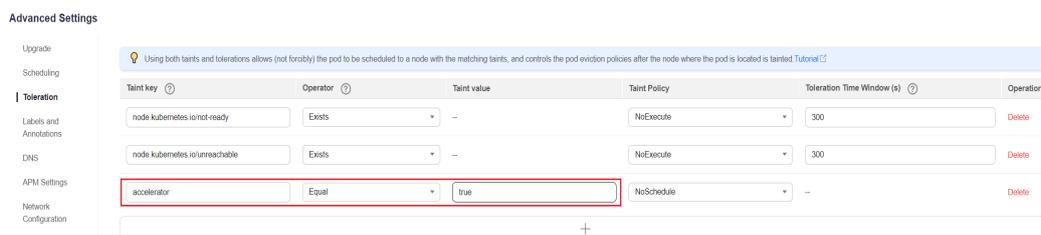
Select **Taint**. Enter the key and value and select the taint effect. The following example shows how to add the **accelerator=true:NoSchedule** taint to the GPU or NPU nodes.

Figure 6-9 Adding a taint



Step 4 When creating a GPU/NPU workload, manually add a toleration in the **Advanced Settings** area.

Figure 6-10 Adding a toleration



Step 5 When creating a non-GPU/NPU workload, add no tolerations, and this workload will not be scheduled to the GPU/NPU nodes.

----End

6.4.8 Why Cannot a Pod Be Scheduled to a Node?

Step 1 Check whether the node and Docker are normal. For details, see [Check Item 7: Whether Internal Components Are Normal](#).

Step 2 If the node and Docker are normal, check whether an affinity policy is configured for the pod. For details, see [Check Item 3: Affinity and Anti-Affinity Configuration of the Workload](#).

Step 3 Check whether the resources on the node are sufficient. If the resources are insufficient, expand the capacity or add nodes.

----End

6.5 Others

6.5.1 What Should I Do If a Scheduled Task Cannot Be Restarted After Being Stopped for a Period of Time?

If a scheduled task is stopped during running, before its restart, the system calculates the difference between the last time the task was successfully executed and the current time and compares the time difference with the scheduled task period multiplied by 100. If the time difference is greater than the period multiplied by 100, the scheduled task will not be triggered again. For details, see [CronJob Limitations](#).

For example, assume that a cron job is set to create a job every minute from 08:30:00 and the **startingDeadlineSeconds** field is not set. If the cron job controller stops running from 08:29:00 to 10:21:00, the job will not be started because the time difference between 08:29:00 and 10:21:00.00 exceeds 100 minutes, that is, the number of missed scheduling times exceeds 100 (in the example, a scheduling period is 1 minute).

If the **startingDeadlineSeconds** field is set, the controller calculates the number of missed jobs in the last x seconds (x indicates the value of **startingDeadlineSeconds**). For example, if **startingDeadlineSeconds** is set to **200**, the controller counts the number of jobs missed in the last 200 seconds. In this case, if the cron job controller stops running from 08:29:00 to 10:21:00, the job will start again at 10:22:00, because only three scheduling requests are missed in the last 200 seconds (in the example, one scheduling period is 1 minute).

Solution

Configure the **startingDeadlineSeconds** parameter in a cron job. This parameter can be created or modified only by using `kubectl` or APIs.

Example YAML:

```
apiVersion: batch/v1
kind: CronJob
```

```
metadata:
  name: hello
spec:
  startingDeadlineSeconds: 200
  schedule: "* * * * *"
  jobTemplate:
    spec:
      template:
        spec:
          containers:
            - name: hello
              image: busybox:1.28
              imagePullPolicy: IfNotPresent
              command:
                - /bin/sh
                - -c
                - date; echo Hello
              restartPolicy: OnFailure
```

If you create a cron job again, you can temporarily avoid this issue.

6.5.2 What Is a Headless Service When I Create a StatefulSet?

The inter-pod discovery service of CCE corresponds to the headless Service of Kubernetes. Headless Services specify **None** for the cluster IP (spec:clusterIP) in YAML, which means no cluster IP is allocated.

Differences Between Headless Services and Common Services

- **Common Services:**
One Service may be backed by multiple endpoints (pods). A client accesses the cluster IP address and the request is forwarded to the real server based on the iptables or IPVS rules to implement load balancing. For example, a Service has two endpoints, but only the Service address is returned during DNS query. The iptables or IPVS rules determine the real server that the client accesses. The client cannot access the specified endpoint.
- **Headless Services:**
When a headless Service is accessed, the actual endpoint (pod IP addresses) is returned. The headless Service points directly to each endpoint, that is, each pod has a DNS domain name. In this way, pods can access each other, achieving inter-pod discovery and access.

Headless Service Application Scenarios

If there is no difference between multiple pods of a workload, you can use a common Service and use the cluster kube-proxy to implement load balancing, for example, an Nginx Deployment.

However, in some application scenarios, pods of a workload have different roles. For example, in a Redis cluster, each Redis pod is different. They have a master/slave relationship and need to communicate with each other. In this case, a common Service cannot access a specified pod through the cluster IP address. Therefore, you need to allow the headless Service to directly access the real IP address of the pod to implement mutual access among pods.

Headless Services work with **StatefulSet** to deploy stateful applications, such as Redis and MySQL.

6.5.3 What Should I Do If Error Message "Auth is empty" Is Displayed When a Private Image Is Pulled?

Problem Description

When you replace the image of a container in a created workload and use an uploaded image on the CCE console, an error message "Auth is empty, only accept X-Auth-Token or Authorization" is displayed when the uploaded image is pulled.

```
Failed to pull image "IP address:Port number /magicdoom/tidb-operator:latest": rpc error: code = Unknown desc = Error response from daemon: Get https://IP address:Port number /v2/magicdoom/tidb-operator/manifests/latest: error parsing HTTP 400 response body: json: cannot unmarshal number into Go struct field Error.code of type errcode.ErrorCode: "{\n  \"errors\": [\n    {\n      \"code\": 400,\n      \"message\": \"Auth is empty, only accept X-Auth-Token or Authorization.\\\"}\n    ]\n  }\""
```

Solution

You can select a private image to create an application on the CCE console. In this case, CCE automatically carries the secret. This problem will not occur during the upgrade.

When you create a workload using an API, you can include the secret in Deployments to avoid this problem during the upgrade.

```
imagePullSecrets:  
- name: default-secret
```

6.5.4 What Is the Image Pull Policy for Containers in a CCE Cluster?

A container image is required to create a container. Images may be stored locally or in a remote image repository.

The **imagePullPolicy** field in the Kubernetes configuration file is used to describe the image pull policy. This field has the following value options:

- **Always:** Always force a pull.
imagePullPolicy: Always
- **IfNotPresent:** The image is pulled only if it is not already present locally.
imagePullPolicy: IfNotPresent
- **Never:** The image is assumed to exist locally. No attempt is made to pull the image.
imagePullPolicy: Never

Description

1. If this field is set to **Always**, the image is pulled from the remote repository each time a container is started or restarted.
If **imagePullPolicy** is left blank, the policy defaults to **Always**.
2. If the policy is set to **IfNotPresent**:
 - a. If the required image does not exist locally, it will be pulled from the remote repository.
 - b. If the content, except the tag, of the required image is the same as that of the local image, and the image with that tag exists only in the remote repository, Kubernetes will not pull the image from the remote repository.

6.5.5 Why Is the Mount Point of a Docker Container in the Kunpeng Cluster Uninstalled?

Symptom

The mount point of a Docker container in the Kunpeng cluster is uninstalled.

Possible Cause

If the Kunpeng cluster node runs EulerOS 2.8 and the **MountFlags=shared** field is configured in the Docker service file, the container mount point will be uninstalled due to the systemd feature.

Solution

Modify the Docker file, delete the **MountFlags=shared** field, and restart Docker.

Step 1 Log in to the node.

Step 2 Run the following command to delete the **MountFlags=shared** field from the configuration file and save the file:

```
vi /usr/lib/systemd/system/docker.service
```

```
[Service]
MountFlags=shared
Type=notify
EnvironmentFile=-/etc/sysconfig/docker
EnvironmentFile=-/etc/sysconfig/docker-storage
EnvironmentFile=-/etc/sysconfig/docker-network
Environment=GOTRACEBACK=crash
```

Step 3 Run the following command to restart Docker:

```
systemctl restart docker
```

----End

6.5.6 What Can I Do If a Layer Is Missing During Image Pull?

Symptom

When containerd is used as the container engine, there is a possibility that the image layer is missing when an image is pulled to a node. As a result, the workload container fails to be created.

```
Events:
Type      Reason      Age      From      Message
-----
Normal    Scheduled   54s     default-scheduler      Successfully assigned cattle-prometheus/prometheus-server-6c69469c-f4-nfs7f to 10.14.11.139
Normal    SuccessfulMountVolume  55s     kubelet      Successfully mounted volumes for pod "prometheus-server-6c69469c-f4-nfs7f_cattle-prometheus(48ac202a-649a-429c-91ca-573d8aabc872)"
Normal    SuccessfulUpdateSecurityGroup  52s     yangtse-controller      Successfully updated security group to "c8a07f89-6fd6-431a-b901-e8075805198c"
Normal    Pulled      8s (x6 over 51s)  kubelet      Container image "100.125.0.29:20202/registry.cn-hangzhou.aliyuncs.com/yangtse/prometheus-server:1.29.2" already present on machine
Warning   FailedCreate  7s (x6 over 50s)  kubelet      Error: failed to create container: error unpacking image: failed to extract layer: sha256:f9f9e4ef62f0689cd75239e14ade4880ec61f488a05af5ab2f9cca54c299d: failed to get reader from content store: content digest sha256:8c9a7d1afbc602695fc2c95445743ccc5ff32053ea399ea9b88773b7068185: not found
```

Possible Causes

Docker earlier than v1.10 supports the layer whose **mediaType** is **application/octet-stream**. However, containerd does not support **application/octet-stream**. As a result, the layer is not pulled.

Solution

You can use either of the following methods to solve this problem:

- Use Docker v1.11 or later to repackage the image.
- Manually pull the image.
 - a. Log in to the node.
 - b. Run the following command to pull the image:
ctr -n k8s.io images pull --user u:p images
 - c. Use the newly pulled image to create a workload.

6.5.7 Why the File Permission and User in the Container Are Question Marks?

Symptom

If the node OS is CentOS 7.6 or EulerOS 2.5 and the Debian GNU/Linux 11 (bullseye) kernel is used as the base image container, exceptions occur on file permissions and users.

```
[root@ ]# docker run -it debian:11 bash
root@a6b8fa7fcdea:/# ls -al
ls: cannot access 'dev': Operation not permitted
ls: cannot access 'root': Operation not permitted
ls: cannot access 'run': Operation not permitted
ls: cannot access 'lib': Operation not permitted
ls: cannot access 'mnt': Operation not permitted
ls: cannot access '.': Operation not permitted
ls: cannot access 'tmp': Operation not permitted
ls: cannot access 'proc': Operation not permitted
ls: cannot access 'bin': Operation not permitted
ls: cannot access 'srv': Operation not permitted
ls: cannot access 'sys': Operation not permitted
ls: cannot access 'var': Operation not permitted
ls: cannot access 'etc': Operation not permitted
ls: cannot access 'media': Operation not permitted
ls: cannot access 'usr': Operation not permitted
ls: cannot access 'sbin': Operation not permitted
ls: cannot access 'home': Operation not permitted
ls: cannot access 'boot': Operation not permitted
ls: cannot access 'lib64': Operation not permitted
ls: cannot access '..': Operation not permitted
ls: cannot access 'opt': Operation not permitted
ls: cannot access '.dockerenv': Operation not permitted
total 0
d????????? ? ? ? ?      ? .
d????????? ? ? ? ?      ? ..
-????????? ? ? ? ?      ? .dockerenv
d????????? ? ? ? ?      ? bin
d????????? ? ? ? ?      ? boot
d????????? ? ? ? ?      ? dev
d????????? ? ? ? ?      ? etc
d????????? ? ? ? ?      ? home
d????????? ? ? ? ?      ? lib
d????????? ? ? ? ?      ? lib64
d????????? ? ? ? ?      ? media
d????????? ? ? ? ?      ? mnt
d????????? ? ? ? ?      ? opt
d????????? ? ? ? ?      ? proc
d????????? ? ? ? ?      ? root
d????????? ? ? ? ?      ? run
d????????? ? ? ? ?      ? sbin
```

Impact

Exceptions occur on file permissions and users in a container.

Solution

CCE provides two solutions:

- Use Debian 9 or 10 as the base image of the service container.
- Use EulerOS 2.9 or Ubuntu 18.04 as the node OS.

7 Networking

7.1 Network Planning

7.1.1 What Is the Relationship Between Clusters, VPCs, and Subnets?

A Virtual Private Cloud (VPC) is similar to a private local area network (LAN) managed by a home gateway whose IP address is 192.168.0.0/16. A VPC is a private network built on the cloud and provides basic network environment for running elastic cloud servers (ECSs), elastic load balances (ELBs), and middleware. Networks of different scales can be configured based on service requirements. Generally, you can set the CIDR block to 10.0.0.0/8–24, 172.16.0.0/12–24, or 192.168.0.0/16–24. The largest CIDR block is 10.0.0.0/8, which corresponds to a class A network.

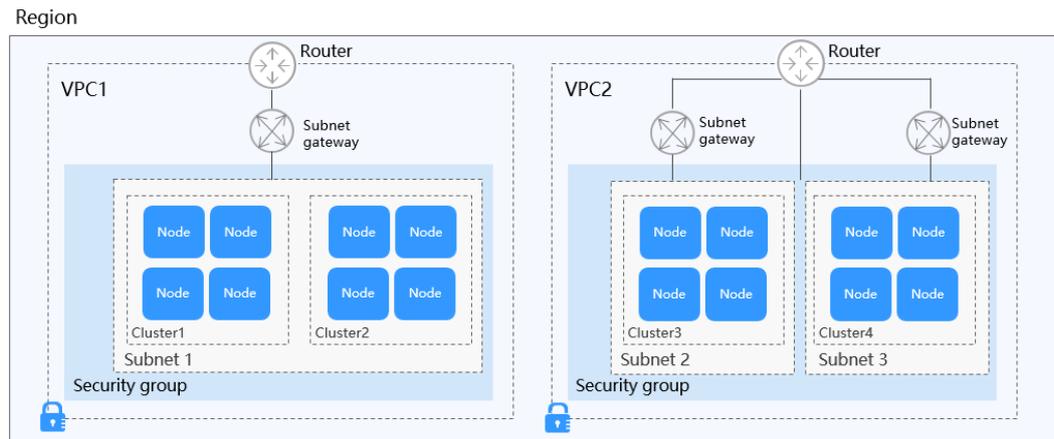
A VPC can be divided into multiple subnets. Security groups are configured to determine whether these subnets can communicate with each other. This ensures that subnets can be isolated from each other, so that you can deploy different services on different subnets.

A cluster is one or a group of cloud servers (also known as nodes) in the same VPC. It provides computing resource pools for running containers.

As shown in [Figure 7-1](#), a region may comprise of multiple VPCs. A VPC consists of one or more subnets. The subnets communicate with each other through a subnet gateway. A cluster is created in a subnet. There are three scenarios:

- Different clusters are created in different VPCs.
- Different clusters are created in the same subnet.
- Different clusters are created in different subnets.

Figure 7-1 Relationship between clusters, VPCs, and subnets



References

Planning CIDR Blocks for a Cluster

7.1.2 How Do I View the VPC CIDR Block?

On the home page of the VPC console, view the **Name/ID** and **CIDR Block** of VPCs. You can modify the CIDR block of a VPC or re-create a VPC.

Figure 7-2 Viewing the CIDR block of VPCs

You can create 45 more VPCs.

Name	IPv4 CIDR Block	Status	Subnet
vpc-demo	192.168.0.0/16	Normal	2
vpcdrs-database-fwx514861	192.168.0.0/16	Normal	1
vpc-test	192.168.0.0/16	Normal	1
CCE-AutoCreate-VPC-w5xvc	192.168.0.0/16	Normal	1
vpc-8d9c	192.168.0.0/16	Normal	1

7.1.3 How Do I Set the VPC CIDR Block and Subnet CIDR Block for a CCE Cluster?

The CIDR block of a VPC cannot be changed after the VPC is created. When creating a VPC, allocate sufficient IP addresses for the VPC and subnets.

The subnet CIDR block can be set on the VPC console by choosing **Create VPC > Subnet CIDR Block** on the **Create VPC** page. You can view the number of available IP addresses under the CIDR block setting.

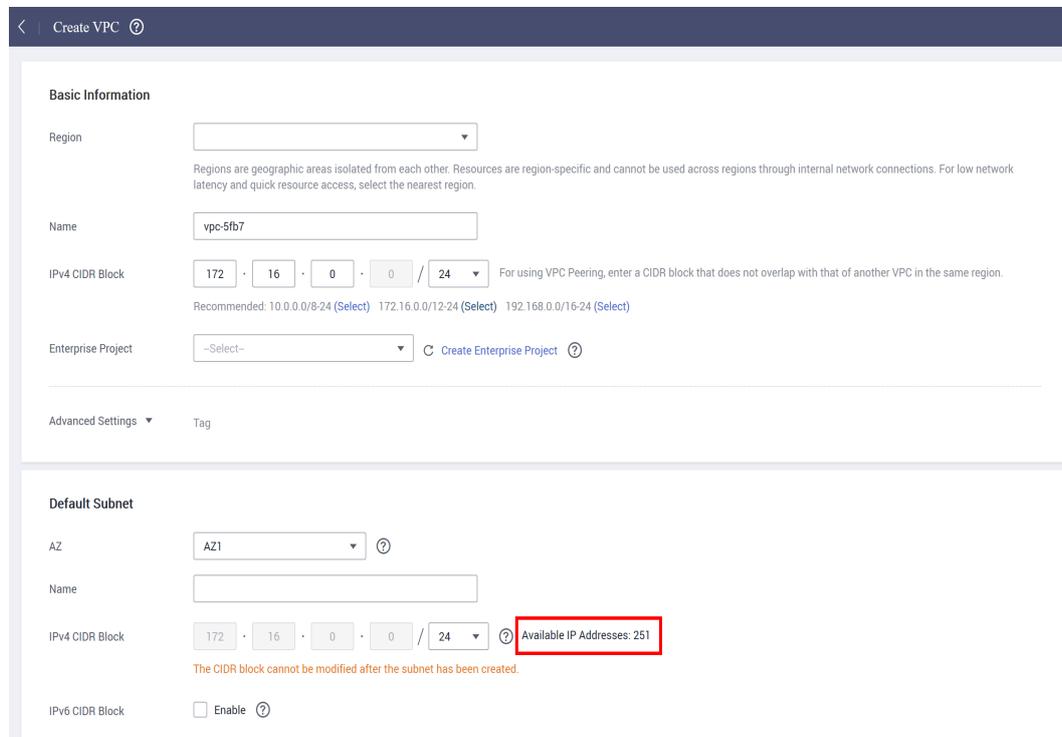
If the subnet mask is not set properly, the number of available nodes in the cluster may be insufficient.

Example:

- If the cluster has 1000 nodes, you can set the subnet CIDR block to 192.168.0.0/20, which supports 4090 nodes.

- If VPC CIDR block is set to 192.168.0.0/16 and the subnet CIDR block is set to 192.168.0.0/25, only 122 nodes are supported. If you create a cluster with 200 nodes using this VPC, only 122 nodes (including master nodes) can be added.

Figure 7-3 Viewing the number of available IP addresses



References

Planning CIDR Blocks for a Cluster

7.1.4 How Do I Set a Container CIDR Block for a CCE Cluster?

Log in to the CCE console and set **Container CIDR Block** when creating a cluster.

Available container CIDR blocks are 10.0.0.0/8-18, 172.16.0.0/16-18, and 192.168.0.0/16-18.

To add a container CIDR block after a cluster is created, go to the cluster information page and click **Add Container CIDR Block**.

NOTICE

- Currently, container CIDR blocks cannot be added to clusters that use the container tunnel network.
- The added container CIDR block cannot be deleted.
- The default service CIDR block is 10.247.0.0/16. Therefore, the container CIDR block cannot be 10.247.0.0/16.

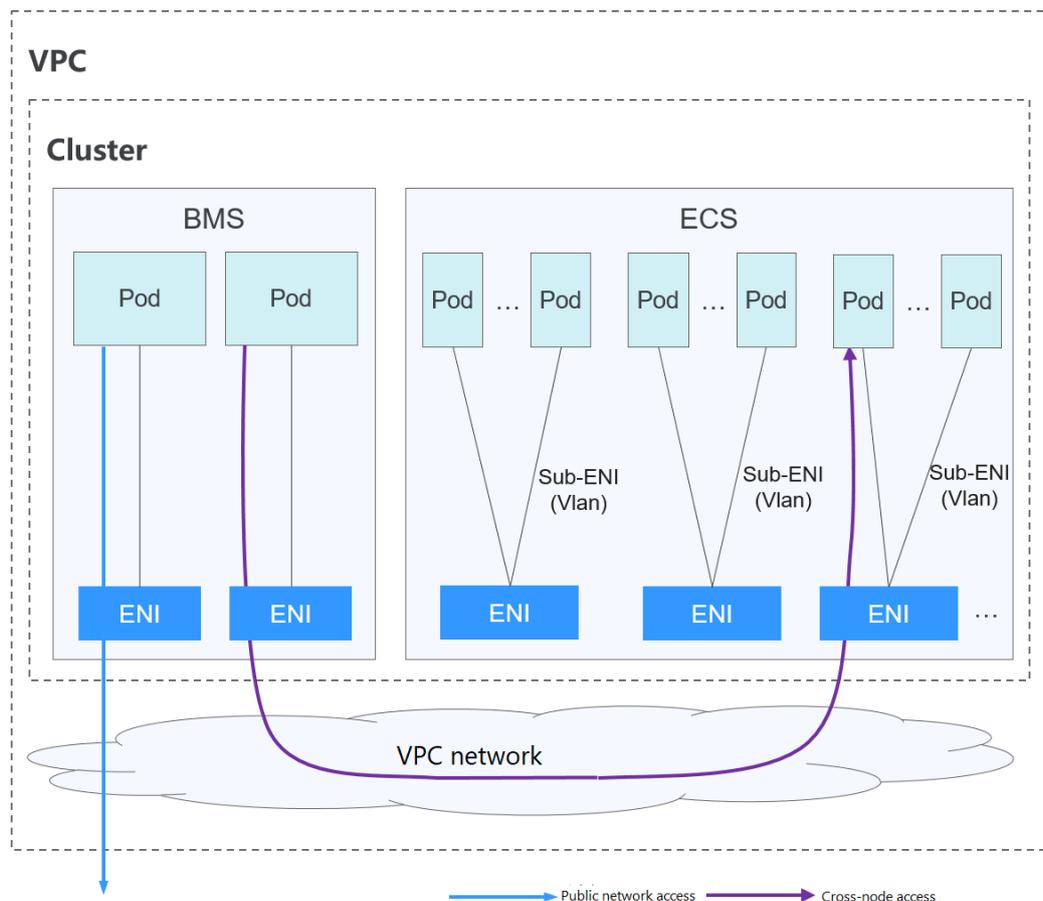
The screenshot displays the configuration page for a CCE cluster named 'cce-test'. The left sidebar shows navigation options like 'Cluster Information', 'Resources', and 'O&M'. The main content area is divided into two sections: 'Basic Info' and 'Networking Configuration'. The 'Basic Info' section lists various cluster attributes such as name, ID, type, version, patch, status, scale, creation time, and project. The 'Networking Configuration' section details network settings including the VPC network, VPC, subnet, CIDR blocks for containers and services, forwarding rules, and the default node security group. A red box highlights the 'Add Container CIDR Block' button next to the 'Container CIDR Block' field.

7.1.5 When Should I Use Cloud Native Network 2.0?

Cloud Native Network 2.0

Cloud Native Network 2.0 is a new container networking solution. This network model deeply integrates the native elastic network interfaces (ENIs) of VPC, uses the VPC CIDR block to allocate container addresses, and supports passthrough networking to containers through a load balancer.

Figure 7-4 Cloud Native Network 2.0



Notes and Constraints

This network model is available only to CCE Turbo clusters.

Application Scenarios

- High performance requirements and use of other VPC network capabilities: Cloud Native Network 2.0 directly uses VPC, which delivers almost the same performance as the VPC network. Therefore, it is applicable to scenarios that have high requirements on bandwidth and latency, such as online live broadcast and e-commerce seckill.
- Large-scale networking: Cloud Native Network 2.0 supports a maximum of 2000 ECS nodes and 100,000 containers.

7.1.6 What Is an ENI?

An elastic network interface (ENI) is a virtual network card. You can create and configure ENIs and attach them to your cloud server instances (ECSs and BMSs) to build flexible and highly available networks.

ENI Types

- A primary network interface is created together with an ECS instance by default, which cannot be detached from its ECS.

- An extension network interface can be created and attached to an ECS, and can be detached from the ECS. The number of extension network interfaces that you can attach to an ECS varies by ECS flavor.

Application Scenarios

- **Flexible migration**
You can detach an ENI from a cloud server instance and then attach it to another instance. The ENI retains its private IP address, EIP, and security group rules. In this way, service traffic on the faulty instance can be quickly migrated to the standby instance, implementing quick service recovery.
- **Independent traffic management**
You can attach multiple ENIs that belong to different subnets in a VPC to the same instance, and specify them to carry the private network traffic, public network traffic, and management network traffic of the instance, respectively. You can configure access control policies and routing policies for each subnet, and configure security group rules for each ENI to isolate networks and service traffic.

Restrictions

- An instance and its extension network interfaces must be in the same AZ, VPC, and subnet. However, they can belong to different security groups.
- A primary network interface cannot be detached from its ECS.
- The number of extension network interfaces that you can attach to an ECS varies by ECS flavor.

7.1.7 How Can I Configure a Security Group Rule in a Cluster?

CCE is a universal container platform. Its default security group rules apply to common scenarios. When a cluster is created, a security group is automatically created for the master node and worker node, separately. The security group name of the master node is *{Cluster name}-cce-control-{Random ID}*, and the security group name of the worker node is *{Cluster name}-cce-node-{Random ID}*. If a CCE Turbo cluster is used, an additional ENI security group named *{Cluster name}-cce-eni-{Random ID}* will be created.

You can log in to the management console, choose **Service List > Networking > Virtual Private Cloud**. On the page displayed, choose **Access Control > Security Groups** in the navigation pane, locate the security group of the cluster, and modify the security group rules as required.

If you need to specify a node security group when creating a cluster, allow specific ports based on the rules of the default security group automatically created in the cluster to ensure normal communication in your cluster.

The default security group rules of the clusters using different networks are as follows:

- [Security Group Rules of a Cluster Using a VPC Network](#)
- [Security Group Rules of a Cluster Using the Tunnel Network](#)
- [Security Group Rules of a CCE Turbo Cluster Using the Cloud Native 2.0 Network](#)

NOTICE

- Modifying or deleting security group rules may affect cluster running. Exercise caution when performing this operation. If you need to modify security group rules, do not modify the rules of the port on which CCE running depends.
- When adding a new security group rule to a cluster, ensure that the new rule does not conflict with the original rules. Otherwise, the original rules may become invalid, affecting the cluster running.

Security Group Rules of a Cluster Using a VPC Network

Security group of worker nodes

A security group named *{Cluster name}-cce-node-{Random ID}* is automatically created for worker nodes in a cluster. For details about the default ports, see [Table 7-1](#).

Table 7-1 Default ports in the security group for worker nodes that use a VPC network

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	All UDP ports	VPC CIDR block	Allow access between worker nodes and between worker nodes and the master nodes.	No	N/A
	All TCP ports				
	All ICMP ports	Security group of master nodes	Allow master nodes to access worker nodes.	No	N/A
	TCP port range: 30000 to 32767	All IP addresses : 0.0.0.0/0	Allow access from NodePort.	Yes	These ports must permit requests from VPC, container, and ELB CIDR blocks.
	UDP port range: 30000 to 32767				
All	Container CIDR block	Allow containers in a cluster to access nodes.	No	N/A	

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
	All	Security group of worker nodes	Access outside the security group of the worker nodes is restricted, but mutual access between instances in the security group of the worker nodes is not restricted.	No	N/A
	TCP port 22	All IP addresses : 0.0.0.0/0	Allow SSH access to ECSs.	Recommended	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default. You are advised to retain this setting.	Yes	If you want to harden security by allowing traffic only on specific ports, remember to allow such ports. For details, see Hardening Outbound Rules .

Security group of master nodes

A security group named *{Cluster name}-cce-control-{Random ID}* is automatically created for master nodes in a cluster. For details about the default ports, see [Table 7-2](#).

Table 7-2 Default ports in the security group for master nodes that use a VPC network

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	TCP port 5444	VPC CIDR block	Allow access from kube-apiserver, which provides lifecycle management for Kubernetes resources.	No	N/A
	TCP port 5444	Container CIDR block			

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
	TCP port 9443	VPC CIDR block	Allow the network add-on of the worker nodes to access master nodes.	No	N/A
	TCP port 5443	All IP addresses : 0.0.0.0/0	Allow kube-apiserver of the master nodes to listen to the worker nodes.	Recommended	The port must allow traffic from the CIDR blocks of the VPC, container, and the control plane of the hosted service mesh. NOTE To use CloudShell, you need to allow traffic from 198.19.0.0/16 on port 5443. Otherwise, you cannot access the cluster using CloudShell.
	TCP port 8445	VPC CIDR block	Allow the storage add-on of worker nodes to access master nodes.	No	N/A
	All	Security group of master nodes	Access outside the security group of the master nodes is restricted, but mutual access between instances in the security group of the master nodes is not restricted.	No	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default.	No	N/A

Security Group Rules of a Cluster Using the Tunnel Network

Security group of worker nodes

A security group named *{Cluster name}-cce-node-{Random ID}* is automatically created for worker nodes in a cluster. For details about the default ports, see [Table 7-3](#).

Table 7-3 Default ports in the security group for worker nodes that use a tunnel network

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	UDP port 4789	All IP addresses : 0.0.0.0/0	Allow access between containers.	No	N/A
	TCP port 10250	CIDR block of master nodes	Allow master nodes to access kubelet on worker nodes, for example, by running kubectl exec {pod} .	No	N/A
	TCP port range: 30000 to 32767	All IP addresses : 0.0.0.0/0	Allow access from NodePort.	Yes	These ports must permit requests from VPC, container, and ELB CIDR blocks.
	UDP port range: 30000 to 32767				
	TCP port 22	All IP addresses : 0.0.0.0/0	Allow SSH access to EC2s.	Recommended	N/A
	All	Security group of worker nodes	Access outside the security group of the worker nodes is restricted, but mutual access between instances in the security group of the worker nodes is not restricted.	No	N/A

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default. You are advised to retain this setting.	Yes	If you want to harden security by allowing traffic only on specific ports, remember to allow such ports. For details, see Hardening Outbound Rules .

Security group of master nodes

A security group named *{Cluster name}-cce-control-{Random ID}* is automatically created for master nodes in a cluster. For details about the default ports, see [Table 7-4](#).

Table 7-4 Default ports in the security group for master nodes that use a tunnel network

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	UDP port 4789	All IP addresses : 0.0.0.0/0	Allow access between containers.	No	N/A
	TCP port 5444	VPC CIDR block	Allow access from kube-apiserver, which provides lifecycle management for Kubernetes resources.	No	N/A
	TCP port 5444	Container CIDR block			
	TCP port 9443	VPC CIDR block	Allow the network add-on of the worker nodes to access master nodes.	No	N/A

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
	TCP port 5443	All IP addresses : 0.0.0.0/0	Allow kube-apiserver of the master nodes to listen to the worker nodes.	Recommended	The port must allow traffic from the CIDR blocks of the VPC, container, and the control plane of the hosted service mesh. NOTE To use CloudShell, you need to allow traffic from 198.19.0.0/16 on port 5443. Otherwise, you cannot access the cluster using CloudShell.
	TCP port 8445	VPC CIDR block	Allow the storage add-on of worker nodes to access master nodes.	No	N/A
	All	Security group of master nodes	Access outside the security group of the master nodes is restricted, but mutual access between instances in the security group of the master nodes is not restricted.	No	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default.	No	N/A

Security Group Rules of a CCE Turbo Cluster Using the Cloud Native 2.0 Network

Security group of worker nodes

A security group named *{Cluster name}-cce-node-{Random ID}* is automatically created for worker nodes in a cluster. For details about the default ports, see [Table 7-5](#).

Table 7-5 Default ports in the security group for worker nodes

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	TCP port 10250	CIDR block of master nodes	Allow master node to access kubelet on worker nodes, for example, by running kubectl exec {pod} .	No	N/A
	TCP port range: 30000 to 32767	All IP addresses : 0.0.0.0/0	Allow access from NodePort.	Yes	These ports must permit requests from VPC, container, and ELB CIDR blocks.
	UDP port range: 30000 to 32767				
	TCP port 22	All IP addresses : 0.0.0.0/0	Allow SSH access to EC2s.	Recommended	N/A
	All	Security group of worker nodes	Access outside the security group of the worker nodes is restricted, but mutual access between instances in the security group of the worker nodes is not restricted.	No	N/A
	All	Container subnet CIDR block	Allow containers in a cluster to access nodes.	No	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default. You are advised to retain this setting.	Yes	If you want to harden security by allowing traffic only on specific ports, remember to allow such ports. For details, see Hardening Outbound Rules .

Security group of master nodes

A security group named *{Cluster name}-cce-control-{Random ID}* is automatically created for master nodes in a cluster. For details about the default ports, see [Table 7-6](#).

Table 7-6 Default ports in the security group for master nodes

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	TCP port 5444	All IP addresses : 0.0.0.0/0	Allow access from kube-apiserver, which provides lifecycle management for Kubernetes resources.	No	N/A
	TCP port 5444	VPC CIDR block		No	N/A
	TCP port 9443	VPC CIDR block	Allow the network add-on of the worker nodes to access master nodes.	No	N/A
	TCP port 5443	All IP addresses : 0.0.0.0/0	Allow kube-apiserver of the master nodes to listen to the worker nodes.	Recommended	The port must allow traffic from the CIDR blocks of the VPC, container, and the control plane of the hosted service mesh. NOTE To use CloudShell, you need to allow traffic from 198.19.0.0/16 on port 5443. Otherwise, you cannot access the cluster using CloudShell.
	TCP port 8445	VPC CIDR block	Allow the storage add-on of worker nodes to access master nodes.	No	N/A

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
	All	Security group of master nodes	Access outside the security group of the master nodes is restricted, but mutual access between instances in the security group of the master nodes is not restricted.	No	N/A
	All	Container subnet CIDR block	Allow traffic from all source IP addresses in the container subnet CIDR block.	No	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default.	No	N/A

Security group of ENI

In a CCE Turbo cluster, an additional security group named *{Cluster name}-cce-eni-{Random ID}* is created. By default, containers in the cluster are bound to this security group. For details about the default ports, see [Table 7-7](#).

Table 7-7 Default ports of the ENI security group

Direction	Port	Default Source Address	Description	Modifiable	Modification Suggestion
Inbound rules	All	ENI security group	Allow containers in a cluster to access each other.	No	N/A
		VPC CIDR block	Allow instances in the cluster VPC to access containers.	No	N/A
Outbound rule	All	All IP addresses : 0.0.0.0/0	Allow traffic on all ports by default.	No	N/A

Hardening Outbound Rules

By default, all security groups created by CCE allow all the **outbound** traffic. You are advised to retain this configuration. To harden outbound rules, ensure that the ports listed in the following table are enabled.

Table 7-8 Minimum configurations of outbound security group rules for a worker node

Port	Allowed CIDR	Description
UDP port 53	DNS server of the subnet	Allow traffic on the port for domain name resolution.
UDP port 4789 (required only for clusters that use the tunnel networks)	All IP addresses	Allow access between containers.
TCP port 5443	CIDR block of master nodes	Allow kube-apiserver of the master nodes to listen to the worker nodes.
TCP port 5444	CIDR blocks of the VPC and container	Allow access from kube-apiserver, which provides lifecycle management for Kubernetes resources.
TCP port 6443	CIDR block of master nodes	None
TCP port 8445	VPC CIDR block	Allow the storage add-on of worker nodes to access master nodes.
TCP port 9443	VPC CIDR block	Allow the network add-on of the worker nodes to access master nodes.
All ports	198.19.128.0/17	Allow worker nodes to access the VPC Endpoint (VPCEP) service.
UDP port 123	100.125.0.0/16	Allow worker nodes to access the internal NTP server.
TCP port 443	100.125.0.0/16	Allow worker nodes to access OBS over internal networks to pull the installation package.
TCP port 6443	100.125.0.0/16	Allow worker nodes to report that the worker nodes are installed.

7.1.8 How Do I Configure the IPv6 Service CIDR Block When Creating a CCE Turbo Cluster?

Context

To create an IPv4/IPv6 dual-stack CCE Turbo cluster, you need to set an IPv6 Service CIDR block. The default CIDR block is **fc00::/112**, which contains 65536 IPv6 addresses. If you need to customize a Service CIDR block, you can refer to this section.

IPv6

IPv6 address

An IPv6 address is 128-bit binary string, four times the length of an IPv4 address. Therefore, the decimal format of IPv4 addresses is no longer applicable. IPv6 addresses are expressed in hexadecimal format. To convert a 128-bit binary string, it is transformed into a 32-bit hexadecimal string. These hexadecimal strings are grouped into sets of four (case insensitive) and separated by a colon (:). IPv6 addresses are divided into eight groups.

An IPv6 address can be omitted in the following ways:

- Omission of leading 0s: 0s can be omitted if the colon group starts with 0s. The following IPv6 addresses are the same.
 - ff01:0d28:03ee:0000:0000:0000:0000:0c23
 - ff01:d28:3ee:0000:0000:0000:0000:c23
 - ff01:d28:3ee:0:0:0:c23
- Omission of all-0s hextets: You can use a double colon (::) to represent a single contiguous string of all-0s segments. A double colon (::) can be used only once.

Example:

Before Omission	After Omission
ff01:d28:3ee:0:0:0:c23	ff01:d28:3ee::c23
0:0:0:0:0:0:1	::1
0:0:0:0:0:0:0	::

IPv6 address segment

An IPv6 address segment is usually expressed in Classless Inter-Domain Routing (CIDR) format. It is usually represented by a slash (/) followed by a number, that is, *IPv6 address/Prefix length*. The function of the prefix is similar to that of the mask of the IPv4 address segment. The number of binary bits occupied by the network part represents the binary bits occupied by the network part. An IPv6 address consists of the network part and host part. The prefix specifies the number of bits occupied by the network part, and the remaining bits are the host part.

For example, **fc00:d28::/32** indicates an IPv6 address segment with a 32-bit prefix. The first 32 bits (**fc00:d28** in binary mode) are the network part and the last 96 bits are available host part.

Constraints on IPv6 Service CIDR Blocks

When setting the cluster service CIDR block, note the following constraints:

- The IPv6 Service CIDR block must belong to the **fc00::/8** CIDR block. The address is a unique local address (ULA). The ULA has a fixed prefix **fc00::/7**, including **fc00::/8** and **fd00::/8**. The two ranges are similar to the dedicated IPv4 network addresses **10.0.0.0/8**, **172.16.0.0/12**, and **192.168.0.0/16**. They are equivalent to private CIDR blocks and can be used only on the local network.
- The prefix ranges from 112 to 120. You can adjust the number of addresses by adjusting the prefix value. The maximum number of addresses is 65536.

Example of an IPv6 Service CIDR Block

According to the constraints, this section provides an example of setting an IPv6 CIDR block that contains 8192 addresses for your reference.

- Step 1** Set the prefix length based on the number of addresses. The prefix length ranges from 112 to 120.

In this example, 8,192 IP addresses are required, which are represented by 13-bit binary strings. An IPv6 address has a total length of 128-bit binary strings. As a result, the prefix length of the IPv6 CIDR block is 115 (128-13). This means that the first 115 bits are used to distinguish the CIDR block, while the last 13 bits indicate 8,192 host IP addresses.

The following table shows how many IP addresses are there in an IPv6 CIDR block with the prefix ranging from 112 to 120.

Prefix Length	Number of IP Addresses
112	65,536
113	32,768
114	16,384
115	8,192
116	4,096
117	2,048
118	1,024
119	512
120	256

- Step 2** Set the IPv6 network address, which must belong to the **fc00::/8** CIDR block.

Check Item 1: Container and Container Port

Log in to the CCE console or use `kubectl` to query the IP address of the pod. Then, log in to the node or container in the cluster and run the `curl` command to manually call the API. Check whether the expected result is returned.

If `<container IP address>:<port>` cannot be accessed, you are advised to log in to the application container and access `<127.0.0.1>:<port>` to locate the fault.

Common issues:

1. The container port is incorrectly configured (the container does not listen to the access port).
2. The URL does not exist (no related path exists in the container).
3. A Service exception (a Service bug in the container) occurs.
4. Check whether the cluster network kernel component is abnormal (container tunnel network model: openswitch kernel component; VPC network model: ipvlan kernel component).

Check Item 2: Node IP Address and Node Port

Only NodePort or LoadBalancer Services can be accessed using the node IP address and node port.

- **NodePort Services:**

The access port of a node is the port exposed externally by the node.

- **LoadBalancer Service:**

You can view the node port of a LoadBalancer Service by editing the YAML file.

Example:

nodePort: 30637 indicates the exposed node port. **targetPort: 80** indicates the exposed pod port. **port: 123** is the exposed Service port. LoadBalancer Services also use this port to configure the ELB listener.

```
spec:
  ports:
    - name: cce-service-0
      protocol: TCP
      port: 123
      targetPort: 80
      nodePort: 30637
```

After finding the node port (nodePort), access `<IP address>:<port>` of the node where the container is located and check whether the expected result is returned.

Common issues:

1. The service port is not allowed in the inbound rules of the node.
2. A custom route is incorrectly configured for the node.
3. The label of the pod does not match that of the Service (created using `kubectl` or API).

Check Item 3: ELB IP Address and Port

There are several possible causes if <IP address>:<port> of the ELB cannot be accessed, but <IP address>:<port> of the node can be accessed.

Possible causes:

- The backend server group of the port or URL does not meet the expectation.
- The security group on the node has not exposed the related protocol or port to the ELB.
- The health check of the layer-4 load balancing is not enabled.
- The certificate used for Services of layer-7 load balancing has expired.

Common issues:

1. When exposing a layer-4 ELB load balancer, if you have not enabled health check on the console, the load balancer may route requests to abnormal nodes.
2. For UDP access, the ICMP port of the node has not been allowed in the inbound rules.
3. The label of the pod does not match that of the Service (created using kubectl or API).

Check Item 4: NAT Gateway + Port

Generally, no EIP is configured for the backend server of NAT. Otherwise, exceptions such as network packet loss may occur.

Check Item 5: Whether the Security Group of the Node Where the Container Is Located Allows Access

Log in to the management console, choose **Service List > Networking > Virtual Private Cloud**. On the Network console, choose **Access Control > Security Groups**, locate the security group rule of the CCE cluster, and modify and harden the security group rule.

- CCE cluster:
The security group name of the node is **{Cluster name}-cce-node-{Random characters}**.
- CCE Turbo cluster:
The security group name of the node is **{Cluster name}-cce-node-{Random characters}**.
The name of the security group associated with the containers is **{Cluster name}-cce-eni-{Random characters}**.

Check the following:

- IP address, port, and protocol of an external request to access the workloads in the cluster. They must be allowed in the inbound rule of the cluster security group.
- IP address, port, and protocol of a request by a workload to visit external applications outside the cluster. They must be allowed in the outbound rule of the cluster security group.

For details about security group configuration, see [How Can I Configure a Security Group Rule in a Cluster?](#)

7.2.2 Why the ELB Address Cannot Be used to Access Workloads in a Cluster?

Symptom

In a cluster (on a node or in a container), the ELB address cannot be used to access workloads.

Possible Cause

If the service affinity of a Service is set to the node level, that is, the value of **externalTrafficPolicy** is **Local**, the Service may fail to be accessed from within the cluster (specifically, nodes or containers). Information similar to the following is displayed:

```
upstream connect error or disconnect/reset before headers. reset reason: connection failure
Or
curl: (7) Failed to connect to 192.168.10.36 port 900: Connection refused
```

It is common that a load balancer in a cluster cannot be accessed. The reason is as follows: When Kubernetes creates a Service, kube-proxy adds the access address of the load balancer as an external IP address (External-IP, as shown in the following command output) to iptables or IPVS. If a client inside the cluster initiates a request to access the load balancer, the address is considered as the external IP address of the Service, and the request is directly forwarded by kube-proxy without passing through the load balancer outside the cluster.

When the value of **externalTrafficPolicy** is **Local**, the access failures in different container network models and service forwarding modes are as follows:

NOTE

- For a multi-pod workload, ensure that all pods are accessible. Otherwise, there is a possibility that the access to the workload fails.
- In a CCE Turbo cluster that utilizes a Cloud Native 2.0 network model, node-level affinity is supported only when the Service backend is connected to a HostNetwork pod.
- The table lists only the scenarios where the access may fail. Other scenarios that are not listed in the table indicate that the access is normal.

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
NodePort Service	Public/Private network	Same node as the service pod	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
		Different nodes from the service pod	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	The access is successful.	The access is successful.

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
		Other containers on the same node as the service pod	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	The access failed.	<p>Access the IP address and NodePort on the node where the server is located: The access is successful.</p> <p>Access the IP address and NodePort on a node other than the node where the server is located: The access failed.</p>	The access failed.

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
		Other containers on different nodes from the service pod	Access the IP address and NodePort on the node where the server is located: The access is successful. Access the IP address and NodePort on a node other than the node where the server is located: The access failed.	Access the IP address and NodePort on the node where the server is located: The access is successful. Access the IP address and NodePort on a node other than the node where the server is located: The access failed.	Access the IP address and NodePort on the node where the server is located: The access is successful. Access the IP address and NodePort on a node other than the node where the server is located: The access failed.	Access the IP address and NodePort on the node where the server is located: The access is successful. Access the IP address and NodePort on a node other than the node where the server is located: The access failed.
LoadBalancer Service using a dedicated load balancer	Private network	Same node as the service pod	The access failed.	The access failed.	The access failed.	The access failed.

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
		Other containers on the same node as the service pod	The access failed.	The access failed.	The access failed.	The access failed.
DNAT gateway Service	Public network	Same node as the service pod	The access failed.	The access failed.	The access failed.	The access failed.
		Different nodes from the service pod	The access failed.	The access failed.	The access failed.	The access failed.
		Other containers on the same node as the service pod	The access failed.	The access failed.	The access failed.	The access failed.
		Other containers on different nodes from the service pod	The access failed.	The access failed.	The access failed.	The access failed.

Service Type Released on the Server	Access Type	Request Initiation Location on the Client	Tunnel Network Cluster (IPVS)	VPC Network Cluster (IPVS)	Tunnel Network Cluster (iptables)	VPC Network Cluster (iptables)
nginx-ingress add-on connected with a dedicated load balancer (Local)	Private network	Same node as cceaddon-nginx-ingress-controller pod	The access failed.	The access failed.	The access failed.	The access failed.
		Other containers on the same node as the cceaddon-nginx-ingress-controller pod	The access failed.	The access failed.	The access failed.	The access failed.

Solution

The following methods can be used to solve this problem:

- **(Recommended)** In the cluster, use the ClusterIP Service or service domain name for access.
- Set **externalTrafficPolicy** of the Service to **Cluster**, which means cluster-level service affinity. Note that this affects source address persistence.

```

apiVersion: v1
kind: Service
metadata:
  annotations:
    kubernetes.io/elb.class: union
    kubernetes.io/elb.autocreate: '{"type":"public","bandwidth_name":"cce-
bandwidth","bandwidth_chargemode":"traffic","bandwidth_size":5,"bandwidth_sharetype":"PER","eip_t
ype":"5_bgp","name":"james"}'
  labels:
    app: nginx
    name: nginx
spec:
  externalTrafficPolicy: Cluster
  ports:
  - name: service0
    port: 80
    protocol: TCP
    targetPort: 80
  selector:
    app: nginx
  type: LoadBalancer
    
```

- Leveraging the pass-through feature of the Service, kube-proxy is bypassed when the ELB address is used for access. The ELB load balancer is accessed first, and then the workload.

NOTE

- In a CCE standard cluster, after passthrough networking is configured for a dedicated load balancer, the private IP address of the load balancer cannot be accessed from the node where the workload pod resides or other containers on the same node as the workload.
- Passthrough networking is not supported for clusters of v1.15 or earlier.
- In IPVS network mode, the passthrough settings of Services connected to the same load balancer must be the same.
- If node-level (local) service affinity is used, **kubernetes.io/elb.pass-through** is automatically set to **onlyLocal** to enable pass-through.

```
apiVersion: v1
kind: Service
metadata:
  annotations:
    kubernetes.io/elb.pass-through: "true"
    kubernetes.io/elb.class: union
    kubernetes.io/elb.autocreate: '{"type":"public","bandwidth_name":"cce-
bandwidth","bandwidth_chargemode":"traffic","bandwidth_size":5,"bandwidth_sharetype":"PER","eip_t
ype":"5_bgp","name":"james"}'
  labels:
    app: nginx
    name: nginx
spec:
  externalTrafficPolicy: Local
  ports:
  - name: service0
    port: 80
    protocol: TCP
    targetPort: 80
  selector:
    app: nginx
  type: LoadBalancer
```

7.2.3 Why the Ingress Cannot Be Accessed Outside the Cluster?

Ingresses forward requests based on layer-7 HTTP and HTTPS protocols. As an entry of cluster traffic, ingresses use domain names and paths to achieve finer granularities. After an ingress is added to a cluster, the cluster may fail to be accessed. This section describes how to locate the fault when an ingress fails to be added or cannot be accessed. Before rectifying ingress issues, read the following precautions and perform a self-check:

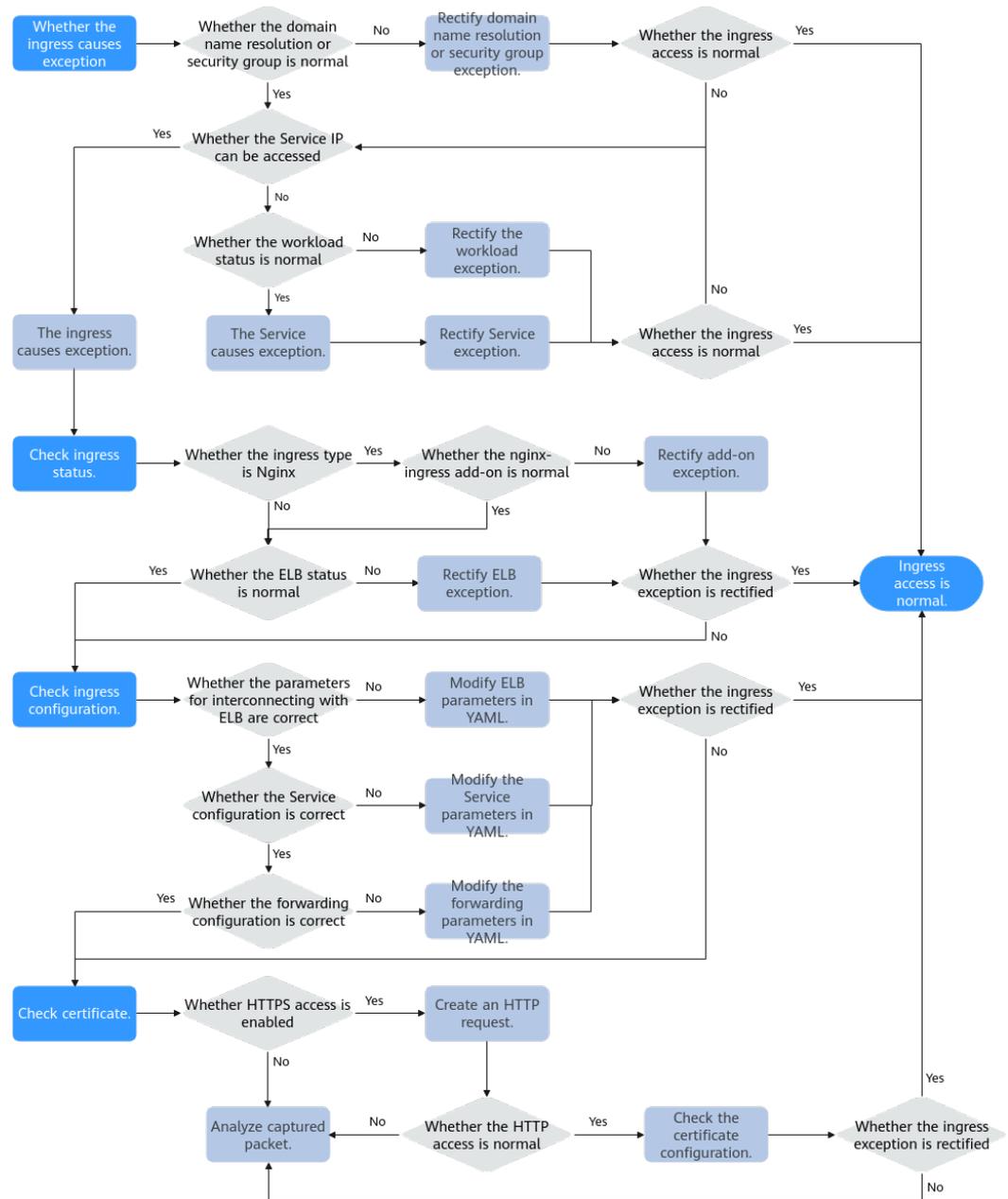
NOTICE

- If the host address is specified in the ingress, the IP address cannot be used for access.
 - Check the node security group of the cluster and ensure that the service ports in the range of 30000 to 32767 are accessible to all network segments for inbound traffic.
-

Troubleshooting Process

This section provides an overview of troubleshooting ingress external access exceptions, as shown in [Figure 7-5](#).

Figure 7-5 Overview of troubleshooting ingress external access exceptions



1. Checking Whether the Exception Is Caused by the Ingress.

Check whether the problem is caused by the ingress. Ensure that the external domain name resolution is normal, the security group rules are correct, and the service and workload corresponding to the ingress are working properly.

2. Checking the Ingress Status.

When the service and workload are normal, ensure that the load balancer on which the ingress depends is normal. If the ingress is of the Nginx type, ensure that the nginx-ingress add-on is normal.

3. **Checking Whether the Ingress Is Configured Correctly.**

If the preceding check results are normal, the ingress configuration may be incorrect.

- Check whether the parameters for interconnecting with the load balancer are correct.
- Check whether the Service configuration is correct.
- Check whether the forwarding configuration is correct.

4. **Checking Certificate.**

If HTTPS access is enabled on the ingress, you also need to check whether the fault is caused by incorrect certificate configuration. You can use the same load balancer to create an HTTP ingress. If the access is normal, the HTTPS certificate may be faulty.

5. If the fault persists, capture packets for analysis or submit a service ticket for help.

Checking Whether the Exception Is Caused by the Ingress

Check whether the access exception is caused by the ingress. If the domain name resolution exception, security group rule error, service exception, or workload exception occurs, the ingress access may fail.

The following check sequence complies with the rules from outside to inside:

Step 1 Check whether the domain name resolution or security group rules are normal.

1. Run the following command to check whether record sets of the domain name take effect on the authoritative DNS server:
`nslookup -qt= Type Domain name Authoritative DNS address`
2. Check the security group rules of the cluster nodes and ensure that the service ports in the 30000–32767 range are accessible to all network segments for inbound traffic. For details about how to harden the security group, see [How Can I Configure a Security Group Rule in a Cluster?](#)

Priority	Action	Protocol & Port	Type	Source	Description	Last Modified	Operation
1	Allow	TCP: All	IPv4	192.168.0.0/16	-	Nov 15, 2022 15:04:16 GMT+08...	Modify Replicate Delete
1	Allow	TCP: 30000-32767	IPv4	0.0.0.0	-	Nov 15, 2022 15:04:16 GMT+08...	Modify Replicate Delete
1	Allow	UDP: 30000-32767	IPv4	0.0.0.0	-	Nov 15, 2022 15:04:16 GMT+08...	Modify Replicate Delete

Step 2 Check whether the Service can access services in the container.

You can create a pod in the cluster and use the cluster IP address to access the Service. If the Service type is NodePort, you can also use **EIP.Port** to access the service over the Internet.

1. Use kubectl to connect to the cluster and query the Service in the cluster.

```
# kubectl get svc
NAME      TYPE      CLUSTER-IP   EXTERNAL-IP  PORT(S)    AGE
kubernetes ClusterIP  10.247.0.1   <none>       443/TCP    34m
nginx     ClusterIP  10.247.138.227 <none>       80/TCP     30m
```

2. Create a pod and log in to the container.

```
kubectl run -i --tty --image nginx:alpine test --rm /bin/sh
```

3. Run the **curl** command to access *ClusterIP address:Port* of the Service to check whether the Service in the cluster is accessible.

```
curl 10.247.138.227:80
```

If the Service can be accessed, the backend workload status is normal. It can be preliminarily determined that the exception is caused by the ingress. For details, see [Checking the Ingress Status](#).

If the Service access is abnormal, check the workload status to determine the cause.

Step 3 Check whether the workload status is normal.

If the workload is normal but the Service cannot be accessed, the exception may be caused by the Service. Check the Service configuration. For example, check whether the container port is correctly configured to an open service port of the container.

If the workload is normal but the access result is not as expected, check the service code running in the container.

----End

Checking the Ingress Status

CCE supports two types of ingresses. The Nginx Ingress Controller is provided by the open source community and needs to be maintained by installing the add-on in the cluster. The ELB Ingress Controller runs on the master node and is maintained by a dedicated Huawei Cloud team.

- ### Step 1
- If you use an Nginx ingress, you need to install the nginx-ingress add-on in the cluster. If you use an ELB ingress, skip this step.

Go to **Add-ons > Add-ons Installed** and check whether the nginx-ingress add-on is running. Ensure that node resources are sufficient in the cluster. If not, the add-on instance cannot be scheduled.

- ### Step 2
- Go to the ELB console to check the ELB status.

- ELB ingress

The access port can be customized. Check whether the listener and backend server group created on the ELB are not deleted or modified.

When creating an ELB ingress, you can choose **Auto Create** on the console to automatically create a load balancer. Do not modify the load balancer. Otherwise, ingress exceptions may be caused.

- Nginx ingress

The access ports are fixed to 80 and 443. Custom ports are not supported. Installing the nginx-ingress add-on occupies both ports 80 and 443. Do not delete them. Otherwise, you need to reinstall the add-on.

You can also determine whether the fault is caused by the load balancer based on the error code. If the following error code is displayed, there is a high probability that the fault is caused by the load balancer. In this case, you need to pay special attention to the load balancer.

404 Not Found

ELB

----End

Checking Whether the Ingress Is Configured Correctly

If the preceding check items are normal, check whether the exception is caused by parameter settings. When using `kubectl` to create an ingress, a large number of parameters need to be set, which is prone to errors. You are advised to use the console to create ingresses and set parameters as required to automatically filter out load balancers and Services that do not meet requirements. This effectively prevents incorrect formats or missing of key parameters.

Check the ingress configuration according to the following steps:

- **Check whether the parameters for interconnecting with the load balancer are correct.**

Load balancers are defined by parameters in the **annotations** field. Kubernetes does not verify the parameters in the **annotations** field when creating resources. If key parameters are incorrect or missing, an ingress can be created but cannot be accessed.

The following problems frequently occur:

- The interconnected ELB load balancer is not in the **same VPC** as the cluster.
 - Key fields **kubernetes.io/elb.id**, **kubernetes.io/elb.ip**, **kubernetes.io/ingress.class**, and **kubernetes.io/elb.port** in **annotations** are missing when an ELB ingress is added to connect to an existing ELB load balancer.
 - When you add an Nginx ingress, the nginx-ingress add-on is not installed. As a result, the ELB connection is unavailable.
 - When you add an Nginx ingress, key fields **kubernetes.io/ingress.class** and **kubernetes.io/elb.port** are missing in **annotations**.
 - When you add an Nginx ingress, the **kubernetes.io/elb.port** field does not support custom ports. If HTTP is used, the value is fixed to **80**. If HTTPS is used, the value is fixed to **443**.
- **Check whether Service is configured correctly.**
 - Check whether the Service type connected to the ingress is correct. For details about the Service types supported by the ingress, see the following table.

Table 7-9 Services supported by LoadBalancer ingresses

Cluster Type	ELB Type	ClusterIP	NodePort
CCE standard cluster	Shared load balancer	Not supported	Supported
	Dedicated load balancer	Not supported (Failed to access the dedicated load balancers because no ENI is bound to the associated pod of the ClusterIP Service.)	Supported
CCE Turbo cluster	Shared load balancer	Not supported	Supported
	Dedicated load balancer	Supported	Not supported (Failed to access the dedicated load balancers because an ENI has been bound to the associated pod of the NodePort Service.)

Table 7-10 Services supported by Nginx ingresses

Cluster Type	ELB Type	ClusterIP	NodePort
CCE standard cluster	Shared load balancer	Supported	Supported
	Dedicated load balancer	Supported	Supported
CCE Turbo cluster	Shared load balancer	Supported	Supported
	Dedicated load balancer	Supported	Supported

- Check whether the access port number of the Service is correct. The access port number (**port** field) of the Service must be different from the container port number (**targetPort** field).
- **Check whether the forwarding configuration is correct.**
 - The forwarding URL added must exist in the backend application. Otherwise, the forwarding fails.
For example, the default access URL of the Nginx application is **/usr/share/nginx/html**. When adding **/test** to the ingress forwarding policy, ensure that your Nginx application contains the same URL, that is, **/usr/share/nginx/html/test**, otherwise, 404 is returned.

 **NOTE**

When using the Nginx Ingress Controller, you can add the **rewrite** comment to the **annotations** field for redirection to rewrite the path that does not exist in the Service to avoid the error that the access path does not exist. For details, see [Rewrite](#).

- If the domain name (host) is specified when an ingress is created, the ingress cannot be accessed using an IP address.

Checking Certificate

The ingress secret certificate type of CCE is **IngressTLS** or **kubernetes.io/tls**. If the certificate type is incorrect, the ingress cannot create a listener on the load balancer. As a result, the ingress access becomes abnormal.

Step 1 Remove HTTPS parameters from YAML and create an HTTP ingress to check whether the ingress can be accessed.

If the HTTP access is normal, check whether the HTTPS secret certificate is correct.

Step 2 Check whether the secret type is correct. Check whether the secret type is **IngressTLS** or **kubernetes.io/tls**.

```
# kubectl get secret
NAME          TYPE          DATA  AGE
ingress       IngressTLS    2      36m
```

Step 3 Create test certificates to rectify the certificate fault.

```
# openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout tls.key -out tls.crt -subj "/"
CN={YOUR_HOST}/O={YOUR_HOST}'
```

Step 4 Use the test certificates **tls.key** and **tls.crt** to create a secret and check whether the secret can be accessed normally. The following example shows how to create an IngressTLS secret.

Specifically, the IngressTLS secret is created using kubectl:

```
kind: Secret
apiVersion: v1
type: IngressTLS
metadata:
  name: ingress
  namespace: default
data:
  tls.crt: LS0tLS1CRU*****FURS0tLS0t
  tls.key: LS0tLS1CRU*****VZLS0tLS0=
```

NOTE

In the preceding information, **tls.crt** and **tls.key** are only examples. Replace them with the actual files. The values of **tls.crt** and **tls.key** are the content encrypted using Base64.

----End

7.2.4 Why Does the Browser Return Error Code 404 When I Access a Deployed Application?

CCE does not return any error code when you fail to access your applications using a browser. Check your services first.

404 Not Found

If the error code shown in the following figure is returned, it indicates that the ELB cannot find the corresponding forwarding policy. Check the forwarding policies.

Figure 7-6 404:ALB

404 Not Found

ALB

If the error code shown in the following figure is returned, it indicates that errors occur on Nginx (your services). In this case, check your services.

Figure 7-7 404:nginx/1.**.*

404 Not Found

nginx/1.14.0

7.2.5 What Should I Do If a Container Fails to Access the Internet?

If a container cannot access the Internet, check whether the node where the container is located can access the Internet. Then check whether the network configuration of the container is correct. For example, check whether the DNS configuration can resolve the domain name.

Check Item 1: Whether the Node Can Access the Internet

Step 1 Log in to the ECS console.

Step 2 Check whether an EIP has been bound to the ECS (node) or whether the ECS has a NAT gateway configured.

Figure 7-8 shows that an EIP has been bound. If no EIP is displayed, bind an EIP to the ECS.

Figure 7-8 Node with an EIP bound

Name/ID	Moni...	AZ	S...	Specifications/Image	IP Address	Bill...	Enterpris...	Operation
...		AZ2	Ru...	2 vCPUs 8 GB s3 large-4 OCE_Cluster_Hidden_ima...	192.168.0.4 (Private IP)	Pay-per-u...	default	Remote Login More
...		AZ2	Ru...	2 vCPUs 8 GB s3 large-4 OCE_Cluster_Hidden_ima...	192.168.0.85 (EIP)	Pay-per-u...	default	Remote Login More
...		AZ2	Ru...	2 vCPUs 8 GB s3 large-4 OCE_Cluster_Hidden_ima...	119.2.145.57 (EIP) 5 Mbit/s 192.168.0.51 (Private IP)	Pay-per-u...	default	Remote Login More

----End

Check Item 2: Whether a Network ACL Has Been Configured for the Node

Step 1 Log in to the VPC console.

Step 2 In the navigation pane on the left, choose **Access Control > Network ACLs**.

Step 3 Check whether a network ACL has been configured for the subnet where the node is located and whether external access is restricted.

----End

Check Item 3: Whether the DNS Configuration of the Container Is Correct

Run `cat /etc/resolv.conf` in the container to check the DNS configuration. An example is as follows:

```
nameserver 10.247.x.x
search default.svc.cluster.local svc.cluster.local cluster.local
options ndots:5
```

If **nameserver** is set to **10.247.x.x**, DNS is connected to the CoreDNS of the cluster. Ensure that the CoreDNS of the cluster is running properly. If another IP address is displayed, an in-cloud or on-premises DNS server is used. Ensure that the domain name resolution is correctly configured.

7.2.6 What Can I Do If a VPC Subnet Cannot Be Deleted?

A VPC subnet may fail to be deleted if you have used the VPC subnet in the CCE cluster. Therefore, you need to delete the corresponding cluster on the CCE console before deleting the VPC subnet.

NOTICE

- If you delete a cluster, all nodes, applications, and services in the cluster will be deleted. Exercise caution when deleting a cluster.
- You are not advised to delete nodes in a CCE cluster on the ECS page.

7.2.7 How Do I Restore a Faulty Container NIC?

If a container NIC is faulty, the container restarts repeatedly and cannot provide services for external systems. To rectify the fault, perform the following steps:

Procedure

Step 1 Run the following command to delete the pod of the faulty container:

```
kubectl delete pod {podName} -n {podNamespace}
```

Where,

- **{podName}**: Enter the name of the pod of the faulty container.
- **{podNamespace}**: Enter the namespace where the pod is located.

Step 2 After the pod of the faulty container is deleted, the system automatically recreates a pod for the container. In this way, the container NIC is restored.

----End

7.2.8 What Should I Do If a Node Fails to Connect to the Internet (Public Network)?

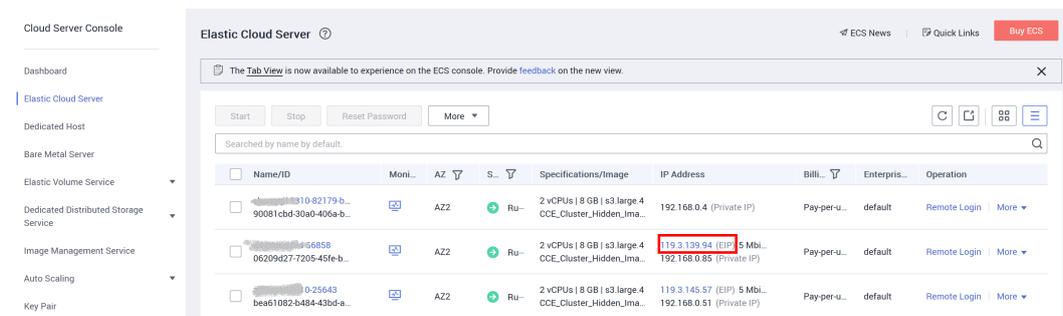
If a node fails to be connected to the Internet, perform the following operations:

Check Item 1: Whether an EIP Has Been Bound to the Node

Log in to the ECS console and check whether an EIP has been bound to the ECS corresponding to the node.

If there is an IP address in the EIP column, an EIP has been bound. If there is no IP address in that column, bind one.

Figure 7-9 Node with an EIP bound



Check Item 2: Whether a Network ACL Has Been Configured for the Node

Log in to the VPC console. In the navigation pane, choose **Access Control > Network ACLs**. Check whether a network ACL has been configured for the subnet where the node is located and whether external access is restricted.

7.2.9 How Do I Resolve a Conflict Between the VPC CIDR Block and the Container CIDR Block?

When you create a cluster, if the container CIDR block conflicts with the VPC CIDR block, an error message will be displayed. In this case, change the container CIDR block.

Figure 7-10 Conflict error message

Network Settings Select the VPC and CIDR blocks for creating nodes and containers in the cluster.

Network Model: **VPC network** | Tunnel network | [? Network Model Overview](#)
 Model used for container networking in a cluster. Not editable after creation

Number of container IP addresses reserved for each node (cannot be changed after creation): [Learn more](#)

VPC: [Create VPC](#)
 CIDR block used by master nodes and worker nodes in the cluster. Not editable after creation

Master Node Subnet: [Create Subnet](#) Available Subnet IP Addresses: **4,089**
 Subnet used by the master node in the cluster. At least 4 IP addresses are required. Not editable after creation

Container CIDR Block: **Manually set** | Auto select | [? How to plan CIDR blocks?](#)

. . . /

✘ The pod CIDR block conflicts with the subnet CIDR block. Select another one.

💡 Max. nodes supported by the current networking configuration: **131,069**

7.2.10 What Should I Do If the Java Error "Connection reset by peer" Is Reported During Layer-4 ELB Health Check

Complete Error Information

```
java.io.IOException: Connection reset by peer
at sun.nio.ch.FileDispatcherImpl.read0(Native Method)
at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:39)
at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223)
at sun.nio.ch.IOUtil.read(IOUtil.java:197)
at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:380)
at com.wanyu.smarthome.gateway.EquipmentSocketServer.handleReadEx(EquipmentSocketServer.java:245)
at com.wanyu.smarthome.gateway.EquipmentSocketServer.run(EquipmentSocketServer.java:115)
```

Analysis Results

A socket server is established using Java Non-blocking I/O (NIO). When the client is shut down unexpectedly rather than sending a specified notification to instruct the server to exit, an error is reported.

TCP Health Check Process

1. The ELB node that performs health checks sends a SYN packet to the backend server (private IP address+health check port) based on the health check configuration.
2. After receiving the packet, the backend server returns a SYN-ACK packet over its port.
3. If the ELB node does not receive the SYN-ACK packet within the timeout duration, the backend server is declared unhealthy. Then, the ELB node sends an RST packet to the backend server to terminate the TCP connection.

4. If the ELB node receives the SYN-ACK packet from the backend server within the timeout duration, it sends an ACK packet to the backend server and declares that the backend server is healthy. Then, the ELB node sends an RST packet to the backend server to terminate the TCP connection.

Note

After a normal TCP three-way handshake, there will be data transfer. However, an RST packet will be sent to terminate the TCP connection during the health check. The applications on the backend server may determine a connection error and reports a message, for example, "Connection reset by peer".

This error is justified and unavoidable. You can ignore it.

7.2.11 How Do I Locate the Service Event Indicating That No Node Is Available for Binding?

- Step 1** Log in to the CCE console, click the cluster, and choose **Networking** in the navigation pane.
- Step 2** Check whether the Service has an associated workload, or whether the pods of the associated workload are normal.

----End

7.2.12 Why Does "Dead loop on virtual device gw_11cbf51a, fix it urgently" Intermittently Occur When I Log In to a VM using VNC?

Symptom

In a cluster that uses the VPC network model, the message "Dead loop on virtual device gw_11cbf51a, fix it urgently" is displayed after login to the VM.

```
[7520230.908741] Dead loop on virtual device gw_11cbf51a, fix it urgently!  
[7764908.323899] Dead loop on virtual device gw_11cbf51a, fix it urgently!  
[7876345.412678] Dead loop on virtual device gw_11cbf51a, fix it urgently!  
[7886952.430199] Dead loop on virtual device gw_11cbf51a, fix it urgently!  
[8053806.787694] Dead loop on virtual device gw_11cbf51a, fix it urgently!
```

Cause

The VPC network model uses the open-source Linux IPvlan module for container networking. In IPvlan L2E mode, layer-2 forwarding is preferentially performed, and then layer-3 forwarding.

Scene reproduction

Assume that there is a service pod A, which provides services externally and is constantly accessed by the node you log in to via the container gateway port through the host Kubernetes Service. Another scenario can be that pods on this node directly access each other. When pod A exits due to upgrade, scale-in, or other reasons, and the corresponding network resources are reclaimed, if the node still attempts to send packets to the IP address of pod A, the IPvlan module in the

kernel first attempts to forward these packets at Layer 2 based on the destination IP address. However, as the NIC to which the pod A IP address belongs can no longer be found, the IPvlan module determines that the packet may be an external packet. Therefore, the module attempts to forward the packet at Layer 3 and matches the gateway port based on the routing rule. After the gateway port receives the packet again, it forwards the packet through the IPvlan module, and this process repeats. The `dev_queue_xmit` function in the kernel detects that the packet is repeatedly sent for 10 times. As a result, the packet is discarded and this log was generated.

After a packet is lost, the access initiator generally performs backoff retries for several times. Therefore, several logs are printed until the ARP in the container of the access initiator ages or the service terminates the access.

For communication between containers on different nodes, the destination and source IP addresses do not belong to the same node-level dedicated subnet (note that this subnet is different from the VPC subnet). Therefore, packets will not be repeatedly sent, and this problem will not occur.

Pods on different nodes in the same cluster can be accessed through a NodePort Service. However, the IP address of the NodePort Service is translated into the IP address of the gateway interface of the accessed container by SNAT, which may generate the logs you see above.

Impact

The normal running of the accessed container is not affected. When a container is destroyed, there is a slight impact that packets are repeatedly sent for 10 times and then discarded. This process is fast in the kernel and has little impact on the performance.

If the ARP ages, the service does not retry, or a new container is started, the container service packets are redirected to the new service through kube-proxy.

Handling in the Open-Source Community

Currently, this problem still exists in the community when the IPvlan L2E mode is used. The problem has been reported to the community for a better solution.

Solution

The dead loop problem does not need to be resolved.

However, it is recommended that the service pod gracefully exit. Before the service is terminated, set your pod to the deleting state. After the service processing is complete, the pod exits.

7.2.13 Why Does a Panic Occasionally Occur When I Use Network Policies on a Cluster Node?

Scenario

Cluster version: v1.15.6-r1

Cluster type: CCE cluster

Network model: Container tunnel network

Node operating system: CentOS 7.6

After a network policy is configured for the cluster, the canal-agent network component on the node is incompatible with the CentOS 7.6 kernel. As a result, a kernel panic may occur.

Conditions

If any of the following conditions is not met, this issue will not occur:

- The cluster version is v1.15.6-r1 and the container tunnel network model is used.
- The CentOS 7.6 node uses the canal-agent component whose version is 1.0.RC10.1230.B005 or earlier. (CentOS 7.6 nodes created on or before February 23, 2021 use such component.)
- You plan to use or have used network policies.

Fault Locating

Quick locating (for pay-per-use nodes)

Check whether your CentOS 7.6 node was created after February 24, 2021 on the CCE console.

Accurate locating (General)

If the cluster version is v1.15.6-r1, the network model is container tunnel network, the node OS is CentOS 7.6, and the canal-agent component version is 1.0.RC10.1230.B005.sp1 or later, the problem will not occur. If an earlier version is used (for example, 1.0.RC10.1230.B002), you are advised to reset or delete the node before configuring network policies.

Perform the following steps to query the version of the network component on the node:

Step 1 Prepare a node where kubectl can be used.

Step 2 Run the following command to query the CentOS node list:

```
for node_item in $(kubectl get nodes --no-headers | awk '{print $1}'); do kubectl get node ${node_item} -o yaml | grep CentOS >/dev/null; if [[ "$?" == "0" ]];then echo "${node_item} is CentOS node";fi;done
```

The command output is as follows:

```
10.0.50.187 is CentOS node
10.0.50.220 is CentOS node
10.0.50.43 is CentOS node
```

Step 3 Assume that the IP address of the target CentOS node is 10.0.50.187. Run the following command to check the canal-agent version:

```
kubectl get packageversions.version.cce.io 10.0.50.187 -o yaml | grep -A 1 canal-agent
```

The command output is as follows:

```
- name: canal-agent  
  version: 1.0.RC10.1230.B005.sp1
```

----End

Solution

If you still want to use the node, reset the CentOS 7.6 nodes in the cluster to upgrade the networking components to the latest version. For details, see [Resetting a Node](#).

If you want to delete the risky node and purchase a new one, see [Deleting a Node](#) and [Buying a Node](#).

7.2.14 Why Are Lots of source ip_type Logs Generated on the VNC?

Scenario

Cluster version: v1.15.6-r1

Cluster type: CCE cluster

Network model: VPC network

Node operating system: CentOS 7.6

When containers on the preceding nodes communicate with each other, the container networking component reports a large number of source ip_types or "not ipvlan but in own host logs" on the VNC. As a result, the VNC page on the node and the container networking performance in high-load scenarios are affected. Symptoms of this problem are as follows:

```
[ 3840.916433] =====source ip_type 2, ipv4 10.0.0.128, mac fa:16:3e:57:f2:8f  
[ 3840.916527] =====source ip_type 2, ipv4 10.0.0.129, mac fa:16:3e:57:f2:8f  
[ 3840.916736] =====source ip_type 2, ipv4 10.0.0.129, mac fa:16:3e:57:f2:8f  
  
[16739.000551] =====not ipvlan but in own host, mac_src=fa:16:3e:34:23:93 mac_dst=ff:ff:ff:ff:ff:ff  
[16740.000968] =====not ipvlan but in own host, mac_src=fa:16:3e:34:23:93 mac_dst=ff:ff:ff:ff:ff:ff
```

Fault Locating

1. Quick Check

This method applies to pay-per-use nodes. Check the node creation time on the CCE console. CentOS 7.6 nodes created on or after February 24, 2021 do not have this problem.

2. Accurate Check (General)

You can perform the following steps to check whether a node is affected:

Step 1 Log in to each CCE node as user **root**.

Step 2 Run the following command to check whether the node is risky:

```
ETH0_IP=$(ip addr show eth0 | grep "inet " | head -n 1 | awk '{print $2}' | awk -F '/' '{print $1}');arping -w  
0.2 -c 1 -I gw_11cbf51a 1.1.1.1 >/dev/null 2>&1 ; echo ;dmesg -T | grep -E "=="not ipvlan but in own host|
```

```
==source ip_type" 1>/dev/null 2>&1 ; if [[ "$?" == "0" ]];then echo "WARNING, node ${ETH0_IP} is affected"; else echo "node ${ETH0_IP} works well"; fi;
```

NOTE

In this command, *1.1.1.1* is an example IP address, which is used only to trigger ARP packet sending. You can use it or replace it with a valid IP address.

Step 3 If the following information is displayed, the node has potential risks. *10.2.0.35* is the IP address of the eth0 NIC on the node. The actual IP address will be displayed in your practice.

```
WARNING, node 10.2.0.35 is affected
```

If the following information is displayed, the node does not have this problem:

```
node 10.2.0.35 works well
```

----End

Procedure

If you still want to use the node, reset the CentOS 7.6 nodes in the cluster to upgrade the networking components to the latest version. For details, see [Resetting a Node](#).

If you want to delete the risky node and purchase a new one, see [Deleting a Node](#) and [Buying a Node](#).

7.2.15 What Should I Do If Status Code 308 Is Displayed When the Nginx Ingress Controller Is Accessed Using the Internet Explorer?

Symptom

After the Nginx Ingress Controller is upgraded, existing services cannot be accessed by Using the Internet Explorer, and the status code is 308.

Possible Causes

After the Nginx Ingress Controller is upgraded, the default permanent redirection status code changes from 301 to 308. However, Internet Explorer of some earlier versions does not support this code. As a result, the Nginx Ingress Controller cannot be accessed.

Nginx Ingress Controller community issue: <https://github.com/kubernetes/ingress-nginx/issues/1825>

Solution

When creating an Ingress, you can use the **nginx.ingress.kubernetes.io/permanent-redirect-code** annotation to specify that the permanent redirection status code is 301.

An example is as follows:

```
apiVersion: networking.k8s.io/v1
kind: Ingress
```

```

metadata:
  name: ingress-test
  namespace: default
  annotations:
    nginx.ingress.kubernetes.io/permanent-redirect-code: '301'
...

```

7.2.16 What Should I Do If an Nginx Ingress Access in the Cluster Is Abnormal After the Add-on Is Upgraded?

Symptom

An Nginx Ingress whose type is not specified (**kubernetes.io/ingress.class: nginx** is not added to annotations) exists in the cluster. After the nginx-ingress add-on is upgraded from 1.x to 2.x, services are interrupted.

Fault Locating

For an Nginx Ingress, check the YAML. If the Ingress type is not specified in the YAML file and the Ingress is managed by the Nginx Ingress Controller, the Ingress is at risk.

Step 1 Check the Ingress type.

Run the following command:

```
kubectl get ingress <ingress-name> -oyaml | grep -E ' kubernetes.io/ingress.class: | ingressClassName:'
```

- Fault scenario: If the command output is empty, the Ingress type is not specified.
- Normal scenario: The command output is not empty, indicating that the Ingress type has been specified by **annotations** or **ingressClassName**.

```

[root@192-168-0-31 paas]# kubectl get ingress test -oyaml | grep -E ' kubernetes.io/ingress.class: | ingressClassName:' -B 1
Warning: extensions/v1beta1 Ingress is deprecated in v1.14+, unavailable in v1.22+; use networking.k8s.io/v1 Ingress
annotations:
  kubernetes.io/ingress.class: nginx
spec:
  ingressClassName: nginx

```

Step 2 Ensure that the Ingress is managed by the Nginx Ingress Controller. The LoadBalancer Ingresses are not affected by this issue.

- For clusters of v1.19, confirm this issue using **managedFields**.

```
kubectl get ingress <ingress-name> -oyaml | grep 'manager: nginx-ingress-controller'
```

```

[root@192-168-0-31 paas]# kubectl get ingress test -oyaml | grep 'manager: nginx-ingress-controller'
Warning: extensions/v1beta1 Ingress is deprecated in v1.14+, unavailable in v1.22+; use networking.k8s.io/v1 Ingress
manager: nginx-ingress-controller

```

- For clusters of other versions, check the logs of the Nginx Ingress Controller pod.

```
kubectl logs -nkube-system cceaddon-nginx-ingress-controller-545db6b4f7-bv74t | grep 'updating Ingress status'
```

```

[root@192-168-0-31 paas]# kubectl logs -nkube-system cceaddon-nginx-ingress-controller-545db6b4f7-bv74t | grep 'updating Ingress status'
+-----+
8 status.go:281] "updating Ingress status" namespace="default" ingress="test" currentValue=[] newV
alue={{IP: +-----+ Hostname: Ports:[]}} {IP: +-----+ Hostname: Ports:[]}}

```

If the fault persists, contact technical support personnel.

----End

Solution

Add an annotation to the Nginx Ingress as follows:

```
kubectl annotate ingress <ingress-name> kubernetes.io/ingress.class=nginx
```

NOTICE

There is no need to add this annotation to LoadBalancer Ingresses. [Verify](#) that these Ingresses are managed by Nginx Ingress Controller.

Possible Causes

The nginx-ingress add-on is developed based on the Nginx Ingress Controller template and image of the open source community.

For the Nginx Ingress Controller of an earlier version (community version v0.49 or earlier, corresponding to CCE nginx-ingress version v1.x.x), the Ingress type is not specified as nginx during Ingress creation, which is, **kubernetes.io/ingress.class: nginx** is not added to annotations. This Ingress can also be managed by Nginx Ingress Controller. For details, see the [GitHub code](#).

For the Nginx Ingress Controller of a later version (community version v1.0.0 or later, corresponding to CCE nginx-ingress version 2.x.x), if the Ingress type is not specified as nginx during Ingress creation, this Ingress will be ignored by the Nginx Ingress Controller and the Ingress rules become invalid. The services will be interrupted. For details, see the [GitHub code](#).

Related link: <https://github.com/kubernetes/ingress-nginx/pull/7341>

You can specify the Ingress type in either of the following ways:

- Add the **kubernetes.io/ingress.class: nginx** annotation to the Ingress.
- Use spec. Set the **.spec.ingressClassName** field to **nginx**. IngressClass resources are required.

An example is as follows:

```
apiVersion: networking.k8s.io/v1beta1
kind: Ingress
metadata:
  name: test
  namespace: default
  annotations:
    kubernetes.io/ingress.class: nginx
spec:
  ingressClassName: nginx
  rules:
  ...
status:
  loadBalancer: {}
```

7.3 Security Hardening

7.3.1 How Do I Prevent Cluster Nodes from Being Exposed to Public Networks?

- If access to port 22 of a cluster node is not required, you can define a security group rule that disables access to port 22.

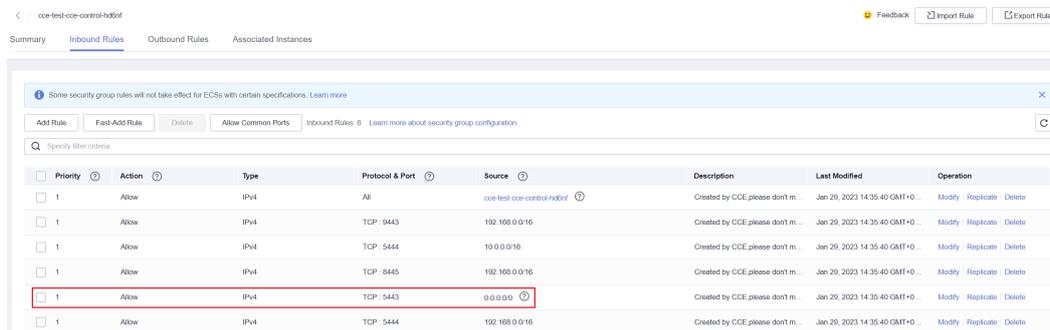
- Do not bind an EIP to a cluster node unless necessary.

If remote login to a cluster node is required, you are advised to use Huawei Cloud Bastion Host (CBH) as the transit node to connect to the cluster node.

7.3.2 How Do I Configure an Access Policy for a Cluster?

After the public API Server address is bound to the cluster, modify the security group rules of port 5443 on the master node to harden the access control policy of the cluster.

- Step 1** Log in to the CCE console and click the cluster name to access the cluster console. On the **Overview** page, copy the cluster ID in the **Basic Info** area.
- Step 2** Log in to the VPC console. In the navigation pane, choose **Access Control** > **Security Groups**.
- Step 3** Select **Description** as the filter criterion and paste the cluster ID to search for the target security groups.
- Step 4** Locate the row that contains the security group (starting with *{CCE cluster name}-cce-control*) of the master node and click **Manage Rules** in the **Operation** column.
- Step 5** On the page displayed, locate the row that contains port 5443 and click **Modify** in the **Operation** column to modify its inbound rules.

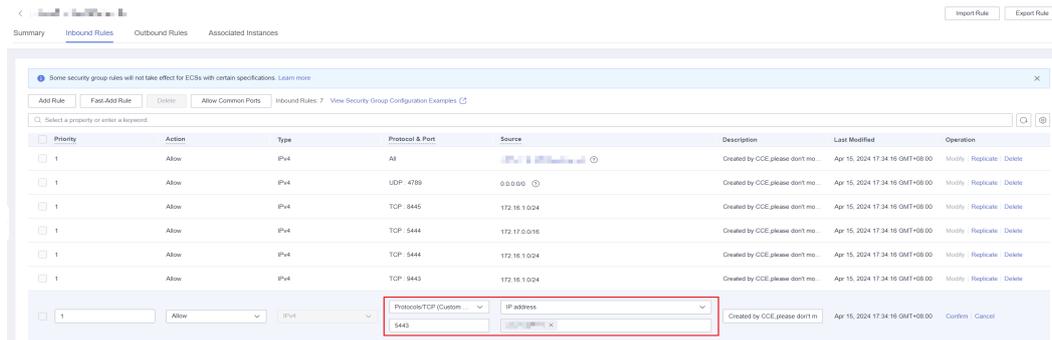


- Step 6** Change the source IP address that can be accessed as required. For example, if the IP address used by the client to access the API Server is **100.*.***, you can add an inbound rule for port 5443 and set the source IP address to **100.*.***.

NOTE

In addition to the client IP address, the port must allow traffic from the CIDR blocks of the VPC, container, and the control plane of the hosted service mesh to ensure that the API Server can be accessed from within the cluster.

To use CloudShell, you need to allow traffic from 198.19.0.0/16 on port 5443. Otherwise, you cannot access the cluster using CloudShell.



Step 7 Click **Confirm**.

----End

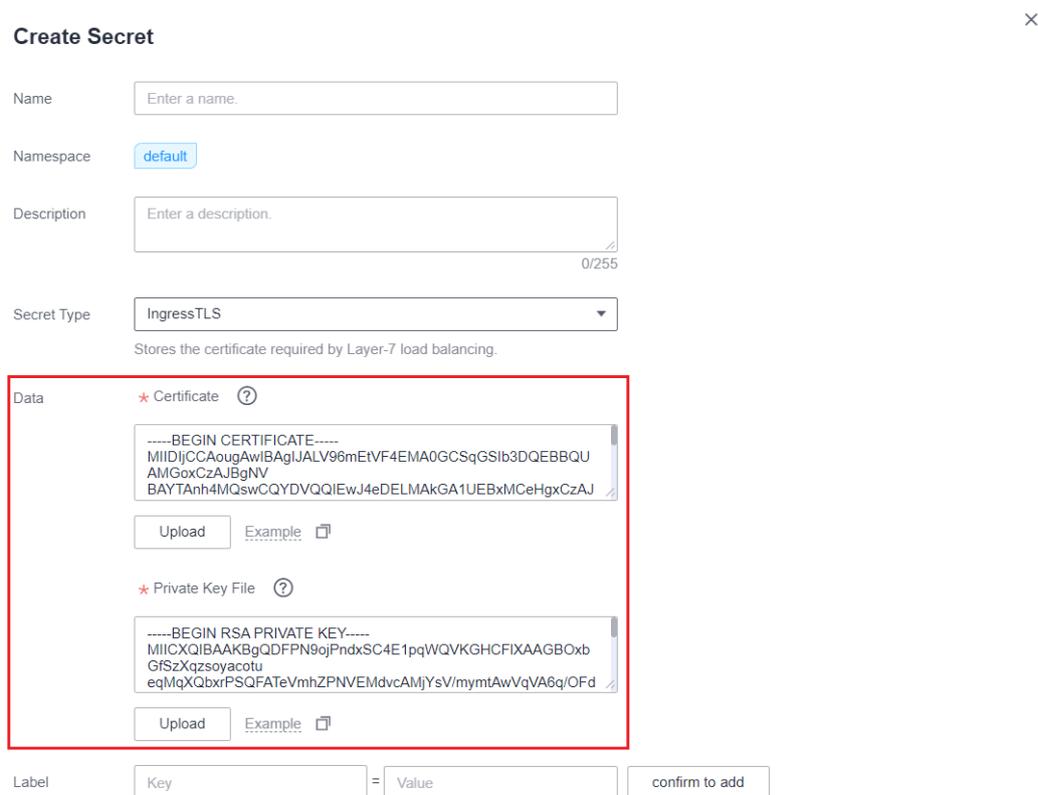
7.3.3 How Do I Obtain a TLS Key Certificate?

Scenario

If your ingress needs to use HTTPS, you must configure a secret of the IngressTLS or kubernetes.io/tls type when creating an ingress.

Create an IngressTLS key certificate, as shown in [Figure 7-11](#).

Figure 7-11 Creating a secret



The certificate file to be uploaded must match the private key file. Otherwise, the certificate file becomes invalid.

Solution

Generally, you need to obtain a valid certificate from the certificate provider. If you want to use it in the test environment, you can create a certificate and private key by the performing the following steps.

NOTE

Self-created certificates apply only to test scenarios. Such certificates are invalid and will affect browser access. Manually upload a valid one to ensure secure connections.

1. Generate a `tls.key`.

```
openssl genrsa -out tls.key 2048
```

The command will generate a private `tls.key` in the directory where the command is executed.

2. Generate a certificate using the private `tls.key`.

```
openssl req -new -x509 -key tls.key -out tls.crt -subj /C=CN/ST=Beijing/O=Devops/CN=example.com -days 3650
```

The generated key must be in the following format:

```
-----BEGIN RSA PRIVATE KEY-----  
.....  
-----END RSA PRIVATE KEY-----
```

The generated certificate must be in the following format:

```
-----BEGIN CERTIFICATE-----  
.....  
-----END CERTIFICATE-----
```

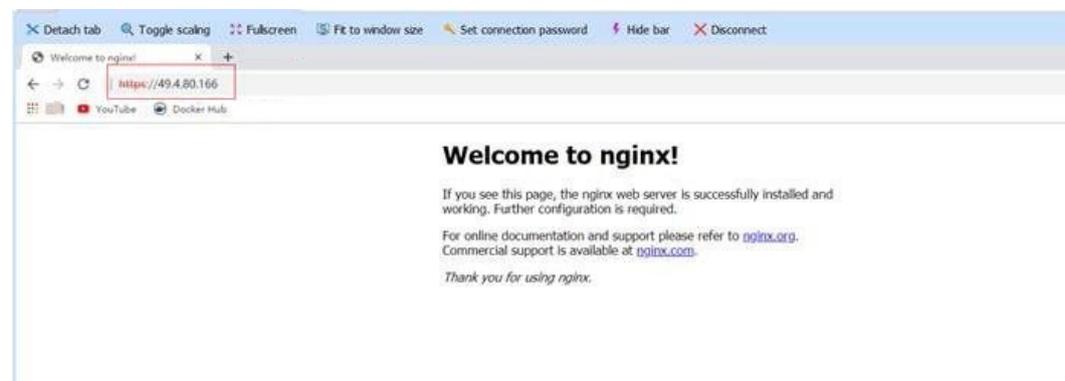
3. Import the certificate.

When creating a TLS secret, import the certificate and private key file to the corresponding location.

Verification

Using a browser to access the ingress is successful. However, the certificate and secret are not issued by CA and the address bar shows the connection to nginx is not secure.

Figure 7-12 Verification result



7.3.4 How Do I Change the Security Group of Nodes in a Cluster in Batches?

Notes and Constraints

Do not add more than 1000 instances to the same security group. Otherwise, the security group performance may deteriorate. For more restrictions on security groups, see [Security Group Constraints](#).

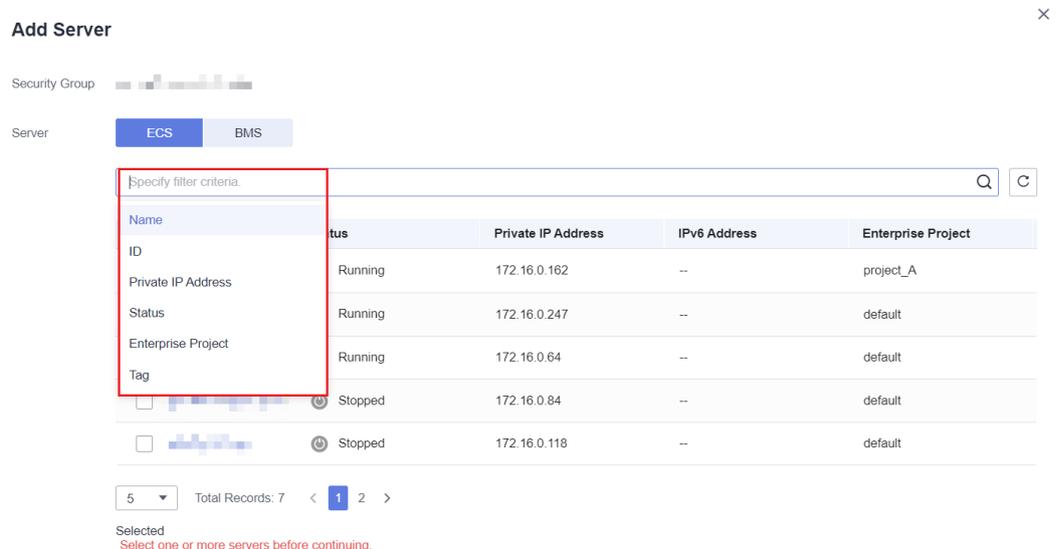
Procedure

- Step 1** Log in to the VPC console and select the desired region and project in the upper left corner.
- Step 2** In the navigation pane on the left, choose **Access Control > Security Groups**.
- Step 3** On the **Security Groups** page, click **Manage Instance** in the **Operation** column.
- Step 4** On the **Servers** tab page, click **Add**.
- Step 5** Select the servers to be added to the security group and click **OK**. You can also search for servers by name, ID, private IP address, status, enterprise project, or tag.

You can change the maximum number of servers displayed on a page in the lower left corner to add a maximum of 20 servers to a security group at a time.

NOTE

After the node is added to a new security group, the original security group is retained. To remove the instance, click **Manage Instance** of the original security group and select the node servers to be removed.



----End

7.4 Network Configuration

7.4.1 How Does CCE Communicate with Other Huawei Cloud Services over an Intranet?

Common HUAWEI CLOUD services that communicate with CCE over the intranet include RDS, DMS, Kafka, RabbitMQ, VPN, and ModelArts. The following two scenarios are involved:

- In the same VPC network, CCE nodes can communicate with all services. When CCE nodes communicate with other services, check whether the security group rule in the inbound direction of the container CIDR block is enabled on the peer end. (This restriction applies only to CCE clusters that use the VPC network model.)
- If CCE nodes and other services are in different VPCs, you can use a peering connection or VPN to connect two VPCs. Note that the two VPC CIDR blocks cannot overlap with the container CIDR block. In addition, you need to configure a return route for the peer VPC or private network. (This restriction applies only to CCE clusters that use the VPC network model.) For details, see [VPC Peering Connection](#).

NOTICE

- This logic works for all Huawei Cloud services.
- Clusters using the container tunnel network support internal communication of Services with no additional configuration required.
- Pay attention to the following points when configuring a cluster using the VPC network:
 1. The source IP address displayed on the peer end is the container IP address.
 2. Custom routing rules added on CCE enable containers to communicate with each other on nodes in a VPC.
 3. When a CCE container accesses other Services, **check whether the inbound security group rule or firewall of the container CIDR block is configured on the peer end (destination end)**. For details, see [Security Group Configuration Examples](#).
 4. If a VPN or VPC peering connection is used to enable communication between private networks, you need to configure a **VPC peering connection route that points to the container CIDR block** on the path and destination.
- Clusters using **Cloud Native 2.0 networks** need to allow traffic from the container security groups based on service requirements. The default container security group is named in the format of *{Cluster name}-cce-eni-{Random ID}*. For details, see [Security Group Rules of a CCE Turbo Cluster Using the Cloud Native 2.0 Network](#).

7.4.2 How Do I Set the Port When Configuring the Workload Access Mode on CCE?

Workloads in a CCE cluster can access each other and can be accessed from the Internet.

- Workloads can access each other through ClusterIP (a virtual cluster IP address) and NodePort (a node IP address).

Table 7-11 Internal access description

Access Type	Description	Guide
Cluster IP (a virtual cluster IP address)	<p>Used for mutual access between workloads in a cluster. For example, if a backend workload needs to communicate with a frontend workload, use this access type.</p> <p>When this access type is selected, a cluster IP address is automatically allocated.</p>	<ul style="list-style-type: none"> • Container port: the port on a container on which the workload listens. The container port varies with the Service. Typically, a container port is specified in the container image. • Service port: the port configured for the workload after the workload was deployed in a Service. Enter an integer from 1 to 65535. Workloads in a cluster can access each other through <i>{Cluster IP}:{Access port number}</i>.

Access Type	Description	Guide
NodePort (through a node IP addresses)	The workload can be accessed through <i>{Node IP address}:{Node port number}</i> . If an EIP is bound to the node, the workload can be accessed from the external networks.	<ul style="list-style-type: none"> • Container port: the port on a container on which the workload listens. The container port varies with the Service. Typically, a container port is specified in the container image. • Service port: the port configured for the workload after the workload was deployed in a Service. Enter an integer from 1 to 65535. • Node port: the port on the node to which the container is mapped. After the configuration is complete, an actual port is open on all nodes in the project to which the user belongs. The workload can be accessed through <i>{Node IP}:{Node port number}</i>. If there are no special requirements, select Automatically generated so that the system automatically assigns an access port. If you select Specified port, enter an integer ranging from 30000 to 32767 and ensure that the value is unique in the cluster.

- A workload can be accessed from the Internet through NodePort (using an EIP), LoadBalancer, or DNAT.

Table 7-12 External access description

Access Type	Description	Guide
NodePort (using an EIP)	<p>If the node where the workload runs is bound with an EIP, the workload can be accessed through <i>{Node EIP}:{Node port number}</i>. The workload can then be accessed from the Internet.</p>	<ul style="list-style-type: none"> • Container port: the port on a container on which the workload listens. The container port varies with the Service. Typically, a container port is specified in the container image. • Service port: the port configured for the workload after the workload was deployed in a Service. Enter an integer from 1 to 65535. • Node port: the port on the node to which the container is mapped. After the configuration is complete, an actual port is open on all nodes in the project to which the user belongs. The workload can be accessed through <i>{Node IP}:{Node port number}</i>. If there are no special requirements, select Automatically generated so that the system automatically assigns an access port. If you select Specified port, enter an integer ranging from 30000 to 32767 and ensure that the value is unique in the cluster.
LoadBalancer	<p>ELB automatically distributes access traffic to multiple nodes to balance their service load. It supports higher levels of fault tolerance for workloads and expands workload service capabilities.</p> <p>You need to create an ELB instance in advance and select ELB as the CCE access type.</p>	<ul style="list-style-type: none"> • Container port: the port on a container on which the workload listens. The container port varies with the Service. Typically, a container port is specified in the container image. • Service port: the port registered with a load balancer. Enter an integer ranging from 1 to 65535. External users can use <i>{Virtual IP address of the load balancer}:{Service port number}</i> to access the workload.

Access Type	Description	Guide
DNAT	<p>NAT gateways provide network address translation (NAT) for cloud servers so that multiple cloud servers can share an EIP.</p> <p>You need to buy a public NAT gateway in advance.</p>	<ul style="list-style-type: none"> • Container port: the port on a container on which the workload listens. The container port varies with the Service. Typically, a container port is specified in the container image. • Service port: the port registered on your NAT gateway. Enter an integer ranging from 1 to 65535. The system automatically creates DNAT rules. External users can access the workload through <i>{EIP of the NAT gateway}:{Service port number}</i>.

7.4.3 How Can I Achieve Compatibility Between Ingress's property and Kubernetes client-go?

Scenario

The Kubernetes ingress structure does not contain the **property** attribute. Therefore, the ingress created by client-go through API calling does not contain the **property** attribute. CCE provides a solution to ensure compatibility with the Kubernetes client-go.

Solution

When using client-go to create an ingress instance, make the following declaration in **annotation**:

```
kubernetes.io/ingress.property: '[{"host":"test.com","path":"/test","matchmode":"STARTS_WITH"}, {"host":"test.com","path":"/dw","matchmode":"EQUAL_TO"}]'
```

Matching rule: When a user calls the Kubernetes interface of CCE to create an ingress instance, CCE attempts to match the **host** and **path** fields in ingress rules. If the **host** and **path** fields in ingress rules are the same as those in annotation, CCE injects the **property** attribute to the path. The following is an example:

```
kind: Ingress
apiVersion: extensions/v1beta1
metadata:
  name: test
  namespace: default
  resourceVersion: '2904229'
  generation: 1
  labels:
    isExternal: 'true'
    zone: data
  annotations:
    kubernetes.io/ingress.class: cce
```

```
kubernetes.io/ingress.property: '[{"host":"test.com","path":"/test","matchmode":"STARTS_WITH"},
{"Path":"/dw","MatchMode":"EQUAL_TO}]'
spec:
  rules:
  - host: test.com
    http:
      paths:
      - path: /ss
        backend:
          serviceName: zlh-test
          servicePort: 80
      - path: /dw
        backend:
          serviceName: zlh-test
          servicePort: 80
```

The format after conversion is as follows:

```
kind: Ingress
apiVersion: extensions/v1beta1
metadata:
  name: test
  namespace: default
  resourceVersion: '2904229'
  generation: 1
  labels:
    isExternal: 'true'
    zone: data
  annotations:
    kubernetes.io/ingress.class: cce
    kubernetes.io/ingress.property: '[{"host":"test.com","path":"/ss","matchmode":"STARTS_WITH"},
{"host":"","path":"/dw","matchmode":"EQUAL_TO}]'
spec:
  rules:
  - host: test.com
    http:
      paths:
      - path: /ss
        backend:
          serviceName: zlh-test
          servicePort: 80
        property:
          ingress.beta.kubernetes.io/url-match-mode: STARTS_WITH
      - path: /dw
        backend:
          serviceName: zlh-test
          servicePort: 80
```

Table 7-13 Descriptions of key parameters

Parameter	Type	Description
host	String	Domain name configuration. If this parameter is not set, path is automatically matched.
path	String	Matching path.

Parameter	Type	Description
ingress.beta.kubernetes.io/url-match-mode	String	Route matching policy. The values are as follows: <ul style="list-style-type: none"> • REGEX: indicates regular expression match. • STARTS_WITH: indicates prefix match. • EQUAL_TO: indicates exact match.

Helpful Links

[Layer-7 Load Balancing \(Ingress\)](#)

7.4.4 How Do I Obtain the Actual Source IP Address of a Client After a Service Is Added into Istio?

Symptom

After Istio is enabled, the source IP address of the client cannot be obtained from access logs.

Solution

This section uses the Nginx application bound to an ELB Service as an example. The procedure is as follows:

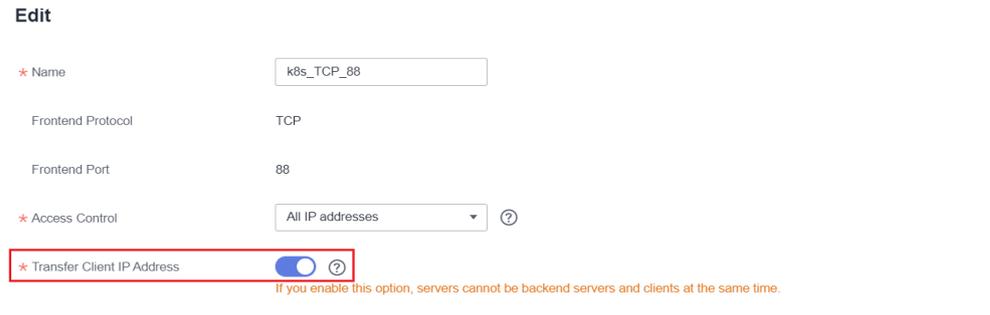
Step 1 Enabling the function of obtaining the client IP address on the load balancer

NOTE

Transparent transmission of source IP addresses is enabled for dedicated load balancers by default. You do not need to manually enable this function.

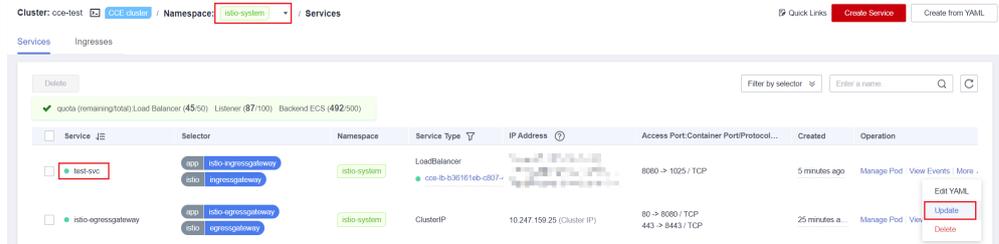
1. Log in to the ELB console.
2. Click  in the upper left corner of the management console and select a region and a project.
3. Click **Service List**. Under **Networking**, click **Elastic Load Balance**.
4. On the **Load Balancers** page, click the name of the load balancer.
5. Click the **Listeners** tab, locate the row containing the target listener, and click **Edit**. If modification protection exists, disable the protection on the basic information page of the listener and try again.
6. Enable **Transfer Client IP Address**.

Figure 7-13 Enabling the function

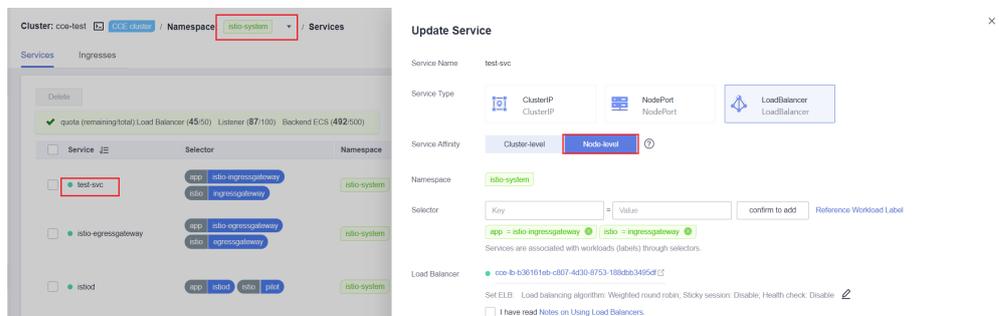


Step 2 Updating the gateway associated with a Service

1. Log in to the CCE console and click the cluster name to access the cluster console. In the navigation pane, choose **Networking**.
2. On the displayed page, switch to the **istio-system** namespace and update the gateway associated with the Service.



3. Change the level of the Service automatically generated in the **istio-system** namespace to the node level.



Step 3 Verifying the obtained source IP address

1. Use `kubectl` to connect to the cluster.
2. Query the Nginx application logs.
`kubectl logs <pod_name>`

In this example, the source IP address obtained by the Nginx application is as follows:

```
2023/04/11 16:56:18 [notice] 181: using the "opoll" event method
2023/04/11 16:56:18 [notice] 181: nginx/1.23.4
2023/04/11 16:56:18 [notice] 181: built by gcc 10.2.1 20210110 (Debian 10.2.1-6)
2023/04/11 16:56:18 [notice] 181: OS: Linux 4.18.0-147.5.1.el8.x86_64
2023/04/11 16:56:18 [notice] 181: getrlimit(RLIMIT_NOFILE): 1048576:1048576
2023/04/11 16:56:18 [notice] 181: start worker process
2023/04/11 16:56:18 [notice] 181: start worker process 30
2023/04/11 16:56:18 [notice] 181: start worker process 31
2023/04/11 16:56:18 [notice] 181: start worker process 32
2023/04/11 16:56:18 [notice] 181: start worker process 33
172.0.0.6 - [11/Apr/2023:16:56:18 +0800] "GET / HTTP/1.1" 304 0 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36 Edg/112.0.1722.34" "-"
172.0.0.6 - [11/Apr/2023:16:56:13 +0800] "GET / HTTP/1.1" 304 0 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36 Edg/112.0.1722.34" "-"
172.0.0.6 - [11/Apr/2023:16:56:08 +0800] "GET / HTTP/1.1" 304 0 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36 Edg/112.0.1722.34" "-"
172.0.0.6 - [11/Apr/2023:16:56:58 +0800] "GET / HTTP/1.1" 304 0 "-" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36 Edg/112.0.1722.34" "-"
```

----End

7.4.5 Why Cannot an Ingress Be Created After the Namespace Is Changed?

Symptom

An ingress can be created in the default namespace, but cannot be created in other namespaces.

Cause Analysis

After a load balancer is created, an HTTP listener is created in the default namespace for port 80. In CCE, only ingresses of the same port can be created in the same namespace (the actual forwarding policies can be distinguished based on domain names and Services). Therefore, ingresses of the same port cannot be created in other namespaces (a port conflict message is displayed).

Solution

You can use YAML files to create ingresses. Port conflicts occur when you create ingresses on the CCE console, but not when you do so in the backend.

7.4.6 Why Is the Backend Server Group of an ELB Automatically Deleted After a Service Is Published to the ELB?

Symptom

After a Service is published to ELB, the workload is normal, but the pod port of the Service is not published in time. As a result, the backend server group of the ELB is automatically deleted.

Answer

1. If the ELB monitoring check fails during ELB creation, the backend server group will be deleted and will not be added after the Service becomes normal. If an existing SVC is updated, the backend server group is not deleted.
2. When a node is added or deleted, the node access mode in the cluster may change due to the cluster status change. To ensure normal service running, the ELB performs a refresh operation. The process is similar to that of updating the ELB.

Suggestions

Optimize the application to speed up the startup.

7.4.7 How Can Container IP Addresses Survive a Container Restart?

If Containers Will Run in a Single-Node Cluster

Add **hostNetwork: true** to the **spec.spec** in the YAML file of the workload to which the containers will belong.

If Containers Will Run in a Multi-Node Cluster

Configure node affinity policies, in addition to perform the operations described in "If the Container Runs in a Single-Node Cluster". However, after the workload is created, the number of running pods cannot exceed the number of affinity nodes.

Expected Result

After the previous settings are complete and the workload is running, the IP addresses of the workload's pods are the same as the node IP addresses. After the workload is restarted, these IP addresses will keep unchanged.

8 Storage

8.1 How Do I Expand the Storage Capacity of a Container?

Application Scenarios

The default storage size of a container is 10 GB. If a large volume of data is generated in the container, expand the capacity using the method described in this topic.

Solution

- Step 1** Log in to the CCE console and click the cluster name to access the cluster console.
- Step 2** Choose **Nodes** from the navigation pane.
- Step 3** Click the Nodes tab, locate the row containing the target node, and choose **More** > **Reset Node** in the **Operation** column.

NOTICE

Resetting a node may make unavailable the node-specific resources (such as local storage and workloads scheduled to this node). Exercise caution when performing this operation to avoid impact on running services.

- Step 4** Click **Yes**.
- Step 5** Reconfigure node parameters.

If you need to adjust the container storage space, pay attention to the following configurations:

Storage Settings: Click **Expand** next to the data disk to set the following parameters:

Space Allocation for Pods: indicates the base size of a pod. It is the maximum size that a workload's pods (including the container images) can grow to in the

disk space. Proper settings can prevent pods from taking all the disk space available and avoid service exceptions. It is recommended that the value is less than or equal to 80% of the container engine space. This parameter is related to the node OS and container storage rootfs and is not supported in some scenarios.

For more information about container storage space allocation, see [Data Disk Space Allocation](#).

Step 6 After the node is reset, log in to the node and run the following command to access the container and check whether the container storage capacity has been expanded:

`docker exec -it container_id /bin/sh` or `kubectl exec -it container_id /bin/sh`

`df -h`

```
# df -h
Filesystem                Size      Used Avail Use% Mounted on
/dev/mapper/docker-253:1-787293-631c1bde2cbe82e39f32253b216ba914cb183b168b54780b3e5b9a54ee40a8d1 15G    229M    15G     2% /
tmpfs                      32G         0     32G     0% /dev
tmpfs                      32G         0     32G     0% /sys/fs/cgroup
/dev/mapper/vgpaas-kubernetes 9.8G    37M     9.2G     1% /etc/hosts
/dev/vda1                  48G     5.2G     33G    14% /etc/hostname
shm                        64M         0     64M     0% /dev/shm
tmpfs                      32G    16K     32G     1% /run/secrets/kubernetes.io/serviceaccount
tmpfs                      32G         0     32G     0% /proc/acpi
tmpfs                      32G         0     32G     0% /sys/firmware
tmpfs                      32G         0     32G     0% /proc/scsi
tmpfs                      32G         0     32G     0% /proc/kbox
tmpfs                      32G         0     32G     0% /proc/oom_extend
```

----End

8.2 What Are the Differences Among CCE Storage Classes in Terms of Persistent Storage and Multi-node Mounting?

Container storage provides storage for container workloads. It supports multiple storage classes. A pod can use any amount of storage.

Currently, CCE supports local, EVS, SFS, SFS Turbo, and OBS volumes.

The following table lists the differences among these storage classes.

Table 8-1 Differences among storage classes

Storage Class	Persistent Storage	Automatic Migration with Containers	Multi-node Mounting
Local disks	Supported	Not supported	Not supported
EVS	Supported	Supported	Not supported
OBS	Supported	Supported	Supported. This type of volumes can be shared among multiple nodes or workloads.
SFS	Supported	Supported	Supported. This type of volumes can be shared among multiple nodes or workloads.

Storage Class	Persistent Storage	Automatic Migration with Containers	Multi-node Mounting
SFS Turbo	Supported	Supported	Supported. This type of volumes can be shared among multiple nodes or workloads.

Selecting a Storage Class

You can use the following types of storage volumes when creating a workload. You are advised to store workload data on EVS volumes. If you store workload data on a local volume, the data cannot be restored when a fault occurs on the node.

- **Local volumes:** Mount the file directory of the host where a container is located to a specified container path (corresponding to `hostPath` in Kubernetes). Alternatively, you can leave the source path empty (corresponding to `emptyDir` in Kubernetes). If the source path is left empty, a temporary directory of the host will be mounted to the mount point of the container. A specified source path is used when data needs to be persistently stored on the host, while `emptyDir` is used when temporary storage is needed. A `ConfigMap` is a type of resource that stores configuration data required by a workload. Its contents are user-defined. A `Secret` is an object that contains sensitive data such as workload authentication information and keys. Information stored in a `Secret` is determined by users.
- **EVS volumes:** Mount an EVS volume to a container path. When the container is migrated, the mounted EVS volume is migrated together. This storage class is applicable when data needs to be stored permanently.
- **SFS volumes:** Create SFS volumes and mount them to a container path. The file system volumes created by the underlying SFS service can also be used. SFS volumes are applicable to persistent storage for frequent read/write in multiple workload scenarios, including media processing, content management, big data analysis, and workload analysis.
- **OBS volumes:** Create OBS volumes and mount them to a container path. OBS volumes are applicable to scenarios such as cloud workload, data analysis, content analysis, and hotspot objects.
- **SFS Turbo volumes:** Create SFS Turbo volumes and mount them to a container path. SFS Turbo volumes are fast, on-demand, and scalable, which makes them suitable for DevOps, containerized microservices, and enterprise office applications.

8.3 Can I Create a CCE Node Without Adding a Data Disk to the Node?

No. A data disk is mandatory.

A data disk dedicated for kubelet and the container engine will be attached to a new node. For details, see [Data Disk Space Allocation](#). By default, CCE uses

Logical Volume Manager (LVM) to manage data disks. With LVM, you can adjust the disk space ratio for different resources on a data disk. For details, see [LVM Overview](#).

If the data disk is uninstalled or damaged, the container engine will malfunction and the node becomes unavailable.

8.4 Can EVS Volumes in a CCE Cluster Be Restored After They Are Deleted or Expired?

You need to manually configure backup policies for EVS disks. If an EVS disk is deleted or released, you can use the VBS backup to restore data.

8.5 What Should I Do If the Host Cannot Be Found When Files Need to Be Uploaded to OBS During the Access to the CCE Service from a Public Network?

When a Service deployed on CCE attempts to upload files to OBS after receiving an access request from an offline machine, an error message is displayed, indicating that the host cannot be found. The following figure shows the error message.

Time	message
February 22nd 2020, 18:50:27.521	com.obs.services.exception.ObsException: OBS service Error Message. Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com
February 22nd 2020, 18:50:27.521	18:50:27.520 [XNIO-1 task-16] ERROR c.h.f.c.provider.ExceptionProvider - OBS service Error Message. Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com
February 22nd 2020, 18:50:27.298	18:50:27.298 [XNIO-1 task-9] ERROR c.h.f.c.provider.ExceptionProvider - OBS service Error Message. Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com
February 22nd 2020, 18:50:27.298	com.obs.services.exception.ObsException: OBS service Error Message. Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com
February 22nd 2020, 18:50:27.275	18:50:27.274 [XNIO-1 task-9] WARN c.o.s.internal.RestStorageService - com.obs.services.internal.ServiceException: Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com HEAD 'https://obs.oss-cn-szhuangzhihuang.com/obs-it-problem-management-media-test?apiversion' on Host 'obs.oss-cn-szhuangzhihuang.com'
February 22nd 2020, 18:50:27.275	com.obs.services.internal.ServiceException: Request Error : java.net.UnknownHostException: obs.oss-cn-szhuangzhihuang.com
February 22nd 2020, 18:50:27.275	2020-02-22 18:50:27 274 com.obs.services.internal.RestStorageService handleThrowable 205 com.obs.services.internal.ServiceException: Request Error : java.net.UnknownHostException:

Fault Locating

After receiving the HTTP request, the Service transfers files to OBS through the proxy.

If too many files are transferred, a large number of resources are consumed. Currently, the proxy is assigned 128 MiB memory. According to pressure test results, resource consumption is large, resulting in request failure.

The test results show that all traffic passes through the proxy. Therefore, if the service volume is large, more resources need to be allocated.

Solution

1. File transfer involves a large number of packet copies, which occupies a large amount of memory. In this case, increase the proxy memory based on the actual scenario and then try to access the Service and upload files again.
2. Additionally, remove the Service from the mesh because the proxy only forwards packets and does not perform any other operations. If requests pass through the ingress gateway, the grayscale release function of the Service is not affected.

8.6 How Can I Achieve Compatibility Between ExtendPathMode and Kubernetes client-go?

Application Scenarios

The Kubernetes pod structure does not contain **ExtendPathMode**. Therefore, when a user calls the API for creating a pod or deployment by using client-go, the created pod does not contain **ExtendPathMode**. CCE provides a solution to ensure compatibility with the Kubernetes client-go.

Solution

NOTICE

- When creating a pod, you need to add **kubernetes.io/extend-path-mode** to **annotation** of the pod.
- When creating a Deployment, you need to add **kubernetes.io/extend-path-mode** to **kubernetes.io/extend-path-mode** in the template.

The following is an example YAML of creating a pod. After the **kubernetes.io/extend-path-mode** keyword is added to **annotation**, the **containername**, **name**, and **mountpath** fields are matched, and the corresponding **extendpathmode** is added to **volumeMount**.

```
apiVersion: v1
kind: Pod
metadata:
  name: test-8b59d5884-96vdz
  generateName: test-8b59d5884-
  namespace: default
  selfLink: /api/v1/namespaces/default/pods/test-8b59d5884-96vdz
  labels:
    app: test
    pod-template-hash: 8b59d5884
  annotations:
    kubernetes.io/extend-path-mode:
'["containername":"container-0","name":"vol-156738843032165499","mountpath":"/
tmp","extendpathmode":"PodUID"]'
  metrics.alpha.kubernetes.io/custom-endpoints: '[{"api":"","path":"","port":"","names":""}]'
  ownerReferences:
```

```

- apiVersion: apps/v1
  kind: ReplicaSet
  name: test-8b59d5884
  uid: 2633020b-cd23-11e9-8f83-fa163e592534
  controller: true
  blockOwnerDeletion: true
spec:
  volumes:
  - name: vol-156738843032165499
    hostPath:
      path: /tmp
      type: ""
  - name: default-token-4s959
    secret:
      secretName: default-token-4s959
      defaultMode: 420
  containers:
  - name: container-0
    image: 'nginx:latest'
    env:
    - name: PAAS_APP_NAME
      value: test
    - name: PAAS_NAMESPACE
      value: default
    - name: PAAS_PROJECT_ID
      value: b6315dd3d0ff4be5b31a963256794989
  resources:
    limits:
      cpu: 250m
      memory: 512Mi
    requests:
      cpu: 250m
      memory: 512Mi
  volumeMounts:
  - name: vol-156738843032165499
    mountPath: /tmp
    extendPathMode: PodUID
  - name: default-token-4s959
    readOnly: true
    mountPath: /var/run/secrets/kubernetes.io/serviceaccount
  terminationMessagePath: /dev/termination-log
  terminationMessagePolicy: File
  imagePullPolicy: Always
  restartPolicy: Always
  terminationGracePeriodSeconds: 30
  dnsPolicy: ClusterFirst
  serviceAccountName: default
  serviceAccount: default
  nodeName: 192.168.0.24
  securityContext: {}
  imagePullSecrets:
  - name: default-secret
  - name: default-secret
  affinity: {}
  schedulerName: default-scheduler
  tolerations:
  - key: node.kubernetes.io/not-ready
    operator: Exists
    effect: NoExecute
    tolerationSeconds: 300
  - key: node.kubernetes.io/unreachable
    operator: Exists
    effect: NoExecute
    tolerationSeconds: 300
  priority: 0
  dnsConfig:
    options:
    - name: timeout
      value: ""

```

```
- name: ndots
  value: '5'
- name: single-request-reopen
  enableServiceLinks: true
```

Table 8-2 Descriptions of key parameters

Parameter	Type	Description
containername	String	Name of a container.
name	String	Name of a volume.
mountpath	String	Mount path.
extendpathmode	String	<p>A third-level directory is added to the created volume directory/subdirectory to facilitate the obtaining of a single pod output file.</p> <p>The following types are supported. For details, see Monitoring.</p> <ul style="list-style-type: none"> ● None: The extended path is not configured. ● PodUID: ID of a pod. ● PodName: Name of a pod. ● PodUID/ContainerName: ID of a pod or name of a container. ● PodName/ContainerName: Name of a pod or container.

8.7 What Can I Do If a Storage Volume Fails to Be Created?

Symptom

The PV or PVC fails to be created. The following information is displayed in the event:

```
{"message": "Your account is suspended and resources can not be used.", "code": 403}
```

Possible Causes

The event indicates that your account is suspended or permissions are not granted to the account. Check whether your account is normal.

If the account is normal, check whether you have the permissions to access the namespace. You must have one of the development, O&M, and administrator permissions of the namespace, or have the customized permission to read and write PVCs and PVs. For details, see [Configuring Namespace Permissions \(on the Console\)](#).

8.8 Can CCE PVCs Detect Underlying Storage Faults?

CCE PersistentVolumeClaims (PVCs) are implemented as they are in Kubernetes. A PVC is defined as a storage declaration and is decoupled from underlying storage. It is not responsible for detecting underlying storage details. Therefore, CCE PVCs cannot detect underlying storage faults.

Cloud Eye allows users to view cloud service metrics. These metrics are built-in based on cloud service attributes. After users enable a cloud service on the cloud platform, Cloud Eye automatically associates its built-in metrics. Users can track the cloud service status by monitoring these metrics.

It is recommended that users who have storage fault detection requirements use Cloud Eye to monitor underlying storage and send alarm notifications.

8.9 An Error Is Reported When the Owner Group and Permissions of the Mount Point of the SFS 3.0 File System in the OS Are Modified

Symptom

After the SFS 3.0 file system is mounted to a directory in the OS, the directory becomes the mount point of the SFS 3.0 file system. When you run the **chown** and **chmod** commands to modify the owner group or permissions of the mount point, the following error information is displayed:

```
chown: changing ownership of '***': Stale file handle
```

Or

```
chmod: changing permissions of '***': Stale file handle
```

Possible Cause

The owner group and permissions of the mount point of the SFS 3.0 file system in the OS cannot be modified.

8.10 Why Cannot I Delete a PV or PVC Using the `kubectl delete` Command?

Symptom

An existing PV or PVC cannot be deleted by running the **kubectl delete** command and it remains in the terminating state.

Possible Causes

To prevent data loss caused by mis-deletion of PVs or PVCs, Kubernetes provides a data protection mechanism. A PV or PVC cannot be directly deleted using the **kubectl delete** command.

Solution

Run the **kubectl patch** command first to remove the protection mechanism and then delete the PV or PVC.

If you have run **kubectl delete** to delete a PV or PVC, the PV or PVC remains in the terminating state. It will be directly deleted after you run the **kubectl patch** command.

- Run the following command to delete a PV:

```
kubectl patch pv <pv-name> -p '{"metadata":{"finalizers":null}}'
```

```
kubectl delete pv <pv-name>
```
- Run the following command to delete a PVC:

```
kubectl patch pvc <pvc-name> -p '{"metadata":{"finalizers":null}}'
```

```
kubectl delete pvc <pvc-name>
```

8.11 What Should I Do If a Yearly/Monthly EVS Disk Cannot Be Automatically Created?

Symptom

When creating a yearly/monthly EVS disk, payment permissions cannot be added to **cce_cluster_agency**.

Possible Causes

cce_cluster_agency is the system agency of CCE. It contains the cloud service resource operation permissions required by CCE components, but does not include the payment permissions. For details, see [System Entrustment Description](#). When creating yearly/monthly EVS disks, **cce_cluster_agency** must have the payment permissions, so you must manually add the **bss:order:pay** permission to **cce_cluster_agency**.

Solution

You can create a custom policy, add the **bss:order:pay** permission to it, and grant the policy to **cce_cluster_agency**.

Step 1 Create a custom policy.

1. Log in to the IAM console. In the navigation pane, choose **Permissions > Policies/Roles**. Then click **Create Custom Policy**.
2. Configure parameters for the policy.
 - **Policy Name:** Set it to **CCE Subscribe Operator**.
 - **Policy View:** Select **JSON**.
 - **Policy Content:** Configure it as follows:

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "bss:order:pay"
      ]
    }
  ]
}
```

3. Click **OK**.

Step 2 Grant the custom policy to **cce_cluster_agency**.

1. Log in to the IAM console. In the navigation pane, choose **Agencies**.
2. Locate the agency named **cce_cluster_agency** and click **Authorize**.
3. Search for the **CCE Subscribe Operator** custom policy, select it, and click **Next**.
4. Select an authorization scope as needed.
By default, **All resources** is selected.
5. Click **OK**.

Step 3 Go back to the CCE console, create a yearly/monthly EVS disk again, and verify that this problem has been resolved.

----End

9 Namespace

9.1 What Should I Do If a Namespace Fails to Be Deleted Due to an APIService Object Access Failure?

Symptom

The namespace remains in the Deleting state. The error message "DiscoveryFailed" is displayed in **status** in the YAML file.

```
76 status:
77   phase: Terminating
78   conditions:
79     - type: NamespaceDeletionDiscoveryFailure
80       status: 'True'
81       lastTransitionTime: '2022-07-04T13:44:55Z'
82       reason: DiscoveryFailed
83       message: 'Discovery failed for some groups, 1 failing: unable to retrieve the complete list of server
84 APIS: metrics.k8s.io/v1beta1: the server is currently unable to handle the request'
85     - type: NamespaceDeletionGroupVersionParsingFailure
86       status: 'False'
```

In the preceding figure, the full error message is "Discovery failed for some groups, 1 failing: unable to retrieve the complete list of server APIs: metrics.k8s.io/v1beta1: the server is currently unable to handle the request".

This indicates that the namespace deletion is blocked when kube-apiserver accesses the APIService resource object of the metrics.k8s.io/v1beta1 API.

Possible Causes

If an APIService object exists in the cluster, deleting the namespace will first access the APIService object. If the access fails, the namespace deletion will be blocked. In addition to the APIService objects created by users, add-ons like metrics-server and prometheus in the CCE cluster automatically create APIService objects.

NOTE

For details, see <https://kubernetes.io/docs/concepts/extend-kubernetes/api-extension/apiserver-aggregation/>.

Solution

Use either of the following methods:

- Rectify the APIService object in the error message. If the object is created by an add-on, ensure that the pod where the add-on locates is running properly.
- Delete the APIService object in the error message. If the object is created by an add-on, uninstall the add-on.

9.2 How Do I Delete a Namespace in the Terminating State?

A Kubernetes namespace is typically in the active or terminating state. If a namespace is deleted when there are still running resources, the namespace enters the terminating state. In this case, the namespace will be automatically deleted only after Kubernetes reclaims the resources in it.

However, in some cases, even if no resource is running in the namespace, the namespace in the terminating state still cannot be deleted.

To solve this problem, perform the following steps:

Step 1 View the namespace details.

```
$ kubectl get ns | grep rdb
rdbms          Terminating 6d21h

$ kubectl get ns rdbms -o yaml
apiVersion: v1
kind: Namespace
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"v1","kind":"Namespace","metadata":{"annotations":{},"name":"rdbms"}}
  creationTimestamp: "2020-05-07T15:19:43Z"
  deletionTimestamp: "2020-05-07T15:33:23Z"
  name: rdbms
  resourceVersion: "84553454"
  selfLink: /api/v1/namespaces/rdbms
  uid: 457788ddf-53d7-4hde-afa3-1fertg21ewe1
spec:
  finalizers:
  - kubernetes
status:
  phase: Terminating
```

Step 2 View resources in the namespace.

```
# View resources that can be isolated using namespaces in the cluster.
$ kubectl api-resources -o name --verbs=list --namespaced | xargs -n 1 kubectl get --show-kind --ignore-not-found -n rdbms
```

The command output shows that no resource is occupied in the **rdbms** namespace.

Step 3 Delete the namespace.

Directly delete the **rdbms** namespace.

```
$ kubectl delete ns rdbms
Error from server (Conflict): Operation cannot be fulfilled on namespaces "rdbms": The system is ensuring all content is removed from this namespace. Upon completion, this namespace will automatically be purged by the system.
```

The deletion fails and a message is displayed, indicating that the system will automatically delete the namespace after confirming that no resource is running in it.

Step 4 Forcibly delete the namespace.

```
$ kubectl delete ns rdbms --force --grace-period=0
warning: Immediate deletion does not wait for confirmation that the running resource has been
terminated. The resource may continue to run on the cluster indefinitely.
Error from server (Conflict): Operation cannot be fulfilled on namespaces "rdbms": The system is ensuring
all content is removed from this namespace. Upon completion, this namespace will automatically be
purged by the system.
```

After running this command, the namespace still cannot be deleted.

Step 5 Call the Kubernetes native APIs to delete resources in the namespace. In most cases, resources in a namespace cannot be forcibly deleted. Use the Kubernetes native APIs instead.

View the namespace details.

```
$ kubectl get ns rdbms -o json > rdbms.json
```

Check the JSON configuration defined by the namespace, edit the JSON file, and delete the **spec** part.

```
$ cat rdbms.json
{
  "apiVersion": "v1",
  "kind": "Namespace",
  "metadata": {
    "annotations": {
      "kubectl.kubernetes.io/last-applied-configuration": "{\"apiVersion\":\"v1\",\"kind\":\"Namespace\", \"name\":\"rdbms\"}"
    },
    "creationTimestamp": "2019-10-14T12:17:44Z",
    "deletionTimestamp": "2019-10-14T12:30:27Z",
    "name": "rdbms",
    "resourceVersion": "8844754",
    "selfLink": "/api/v1/namespaces/rdbms",
    "uid": "29067ddf-56d7-4cce-afa3-1fbdbb221ab1"
  },
  "spec": {
    "finalizers": [
      "kubernetes"
    ]
  },
  "status": {
    "phase": "Terminating"
  }
}
```

After the PUT request is executed, the namespace is automatically deleted.

```
$ curl --cacert /root/ca.crt --cert /root/client.crt --key /root/client.key -k -H "Content-Type:application/json" -X PUT --data-binary @rdbms.json https://x.x.x.x:5443/api/v1/namespaces/rdbms/finalize
{
  "kind": "Namespace",
  "apiVersion": "v1",
  "metadata": {
    "name": "rdbms",
    "selfLink": "/api/v1/namespaces/rdbms/finalize",
    "uid": "29067ddf-56d7-4cce-afa3-1fbdbb221ab1",
    "resourceVersion": "8844754",
    "creationTimestamp": "2019-10-14T12:17:44Z",
    "deletionTimestamp": "2019-10-14T12:30:27Z",
    "annotations": {
      "kubectl.kubernetes.io/last-applied-configuration": "{\"apiVersion\":\"v1\",\"kind\":\"Namespace\", \"name\":\"rdbms\"}"
    }
  }
}
```

```
}  
},  
"spec": {  
  
},  
"status": {  
  "phase": "Terminating"  
}
```

If the namespace still cannot be deleted, check whether the **finalizers** field exists in the metadata. If the field exists, run the following command to access the namespace and delete the field:

```
kubectl edit ns rdbms
```

NOTE

- For details about how to obtain the cluster certificate, see [Obtaining a Cluster Certificate](#).
- **https://x.x.x.x:5443** indicates the address for accessing the cluster. To obtain the private IP address, log in to the CCE console, access the cluster console, and view the connection information.

Step 6 Check whether the namespace has been deleted.

```
$ kubectl get ns | grep rdb
```

----End

10 Chart and Add-on

10.1 What Should I Do If the nginx-ingress Add-on Fails to Be Installed on a Cluster and Remains in the Creating State?

Context

You have purchased and set up a CCE cluster and want to access the deployed applications from public networks. Currently, the most efficient way is to register the Service paths of an application on the ingress to allow public network access.

However, after the nginx-ingress add-on is installed, the add-on is always in the **Creating** state, and the pod **nginx-ingress-controller** is always in the **Pending** state.

Solution

The memory resources for the nginx-ingress add-on are limited. As a result, the nginx-ingress add-on cannot be started. Cancel the resource limitation to ensure that nginx-ingress add-on can be started properly.

Scene Simulation

- Step 1** Create a cluster with three nodes, 2 vCPUs and 4 GB memory for each node.
- Step 2** Install the nginx-ingress add-on and select 2 vCPUs and 2 GB memory.
- Step 3** The nginx-ingress Deployment is successfully created, but the nginx-ingress-controller add-on fails to be installed.

Figure 10-1 nginx-ingress-controller add-on always in the Creating state

<input type="checkbox"/>	Name	Status
<input type="checkbox"/>	-7697b9f7...	Creating
<input type="checkbox"/>	-7697b9f7...	Running

Figure 10-2 nginx-ingress-controller add-on failing to be installed

```
[root@k8s-zwx767800-cluster-33393-1a7ex ~]# kubectl get po -n kube-system grep nginx
cceaddon-nginx-ingress-controller-577bc9c678-xz17d 0/1 Pending 0 27m
cceaddon-nginx-ingress-default-backend-77f6d77b6f-m5tth 1/1 Running 0 27m
```

Step 4 Check the error message. The following information indicates that resources are insufficient.

```
status:
  phase: Pending
  conditions:
    - type: PodScheduled
      status: 'False'
      lastProbeTime: '2020-02-13T01:20:57Z'
      lastTransitionTime: '2020-02-12T08:50:21Z'
      reason: Unschedulable
      message: '0/3 nodes are available: 3 Insufficient cpu, 3 Insufficient memory.'
  qosClass: Guaranteed
```

Step 5 Add a node with 4 vCPUs and 8 GB memory. After that, the nginx-ingress add-on is installed successfully.

----End

Possible Cause

Processes such as kubelet, kube-proxy, and Docker on each node are using system resources. As a result, the available resources of the node are less than required for the nginx-ingress add-on to be successfully installed.

Suggested Solution

Purchase a node with at least 4 vCPUs and 8 GB memory.

10.2 What Should I Do If Residual Process Resources Exist Due to an Earlier npd Add-on Version?

Problem Description

When the node load is heavy, residual npd process resources may exist.

Symptom

After successful login to the ECS node where the CCE cluster runs, it is found that there are a large number of npd processes exist.

```

paas 32763 16574 0 Oct23 ? 00:00:00 [sali] <defunct>
root 32764 16574 0 Oct23 ? 00:00:00 [sudo] <defunct>
root 32765 16574 0 Oct20 ? 00:00:00 [sudo] <defunct>
paas 32766 16574 0 Oct20 ? 00:00:00 [sali] <defunct>
paas 32767 16574 0 Oct13 ? 00:00:00 [grep] <defunct>
[Font:Emacs]--mesal--Inferencet4-gpu-z-89-44 --JE ^C
-bash: Fork: retry: No child processes
-bash: Fork: retry: No child processes
root@ecs-1:~# ps -ef | grep -E 'npd|nfd'
COMMAND PID USER FD TYPE DEVICE SIZE/OFF NODE NAME
node-prob 16574 paas cwd DIR 252,9 4896 2851 /
node-prob 16574 paas rtd DIR 252,9 4896 2851 /
node-prob 16574 paas txt REG 252,9 56274632 25180170 /var/paas/node-problem-detector/node-problem-detector
node-prob 16574 paas men REG 0,20 1806632386 1590580444 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000100a878-0005b24934459833_Journal1
node-prob 16574 paas men REG 0,20 134217278 1590591938 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000e06e-0005b27dd65f5ec3_Journal1
node-prob 16574 paas men REG 0,20 117448932 1359727668 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000c4e2f-0005b2721cc292123_Journal1
node-prob 16574 paas men REG 0,20 117448932 1343987821 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000c4e2f-0005b2721cc292123_Journal1
node-prob 16574 paas men REG 0,20 117448932 1311319384 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-00000000004051f-0005b1c473e8d8f9c_Journal1
node-prob 16574 paas men REG 0,20 189051904 1226454318 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000211ed-0005b2a206454584c_Journal1
node-prob 16574 paas men REG 0,20 117448932 1227802866 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000003576-0005b23f1c18d6edc_Journal1
node-prob 16574 paas men REG 0,20 117448932 1261644838 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000b7711-0005b133a7a45474c_Journal1
node-prob 16574 paas men REG 0,20 117448932 1244021972 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000a859-0005b218a8f5533_Journal1
node-prob 16574 paas men REG 0,20 117448932 1227480734 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000d6fd-0005b219c6d6f8db_Journal1
node-prob 16574 paas men REG 0,20 189051904 1211843384 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000025701-0005b28d5fd6d86c_Journal1
node-prob 16574 paas men REG 0,20 117448932 1195737894 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-00000000002687-0005b2018d099f6b_Journal1
node-prob 16574 paas men REG 0,20 189051904 1180118994 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000f676-0005b1f5dbed8c3a_Journal1
node-prob 16574 paas men REG 0,20 117448932 1163478567 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000ce18-0005b1e28b18345_Journal1
node-prob 16574 paas men REG 0,20 189051904 1147397412 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000311ed-0005b21c1f82b212_Journal1
node-prob 16574 paas men REG 0,20 117448932 1130671264 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000076a5-0005b1e0b5871c6a_Journal1
node-prob 16574 paas men REG 0,20 134217278 1114601808 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000e07-0005b13f0321c48e_Journal1
node-prob 16574 paas men REG 0,20 189051904 1109019008 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-00000000004051f-0005b1c473e8d8f9c_Journal1
node-prob 16574 paas men REG 0,20 134217278 1597373671 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000b70f-0005b1330b5d49776_Journal1
node-prob 16574 paas men REG 0,20 117448932 1456897748 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-0000000000268b-0005b2a206454584c_Journal1
node-prob 16574 paas men REG 0,20 189051904 1513834458 /run/log/journal/6d14d648c8d6f8a06d8b1b1412959c/system@73d6fe9d5143049e63f8508736d07-000000000073a07-0005b2f04934c375_Journal1
    
```

Solution

Upgrade the npd add-on to the latest version.

- Step 1** Log in to the CCE console and click the cluster. In the navigation pane, choose **Add-ons**. On the displayed page, click **Upgrade** under **npd**.

 **NOTE**

If the npd add-on version is 1.13.6 or later, you do not need to upgrade it.

- Step 2** On the **Specify Basic Information** page, select the cluster and the add-on version (for example, 1.13.6), and click **Next**.

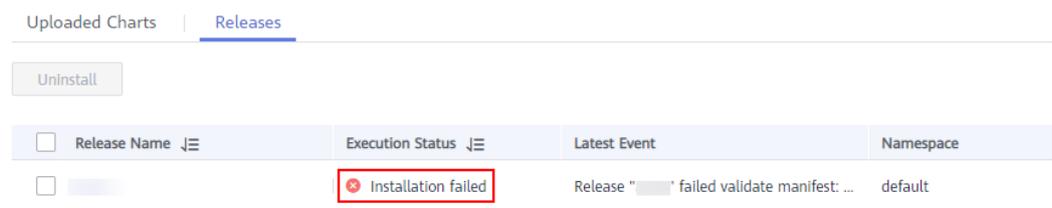
- Step 3** Click **Upgrade** to upgrade the npd add-on. Note that the npd add-on has no configurable parameters and can be directly upgraded.

----End

10.3 What Should I Do If a Chart Release Cannot Be Deleted Because the Chart Format Is Incorrect?

Symptom

If an uploaded chart contains incorrect or incompatible resources, the chart will fail to be installed.



In this case, the chart release cannot work properly. You may not be able to delete the release, the error message "deletion failed" is displayed, and the release is still on the GUI.

Uploaded Charts | Releases

Uninstall

<input type="checkbox"/>	Release Name	Execution Status	Latest Event	Namespace
<input type="checkbox"/>		Installation failed	deletion failed with 1 error(s): unable to d	default

Solution

In this case, you can run the kubectl commands to delete the release.

NOTE

This problem cannot be solved by deleting the residual chart release. To prevent this problem from occurring again, update the API version of the resources in the chart so that the API version of the resources matches the Kubernetes version.

During chart installation, some resources specified in the chart may have been successfully created. You need to manually delete these resources first. After the residual resources are deleted, you need to delete the chart instance.

For a Helm v2 chart release, query the ConfigMap corresponding to the chart release in the kube-system namespace. For example:

```
[paas@192-168-0-40 ~]$ kubectl -n kube-system get cm
NAME                                DATA  AGE
9a37566a.cce.io                     0      25d
aosredis.v1                          1      55s
cceaddon-coredns.v1                 1      25d
cceaddon-everest.v1                 1      17d
cceaddon-metrics-server.v1         1      25d
cceaddon-npd-custom-config          0      25d
cceaddon-npd.v1                     1      25d
cceaddon-prometheus.v1             1      25h
cluster-autoscaler-status           1      8d
cluster-versions                    1      25d
coredns                             1      25d
extension-apiserver-authentication  6      25d
ingress-controller-leader-nginx     0      22d
[paas@192-168-0-40 ~]$
```

After the ConfigMap is deleted, the chart release is deleted successfully.

```
[paas@192-168-0-40 ~]$ kubectl -n kube-system delete cm aosredis.v1
configmap "aosredis.v1" deleted
[paas@192-168-0-40 ~]$
```

For a Helm v3 chart release, query the Secret corresponding to the chart release in the namespace. For example:

```
[root@cce-1717-vpc-node2 ~]# kubectl -n default get secret
NAME                                TYPE                                DATA  AGE
default-secret                     kubernetes.io/dockerconfigjson     1      21h
default-token-978pv                kubernetes.io/service-account-token 3      21h
paas.elb                           cfe/secure-opaque                  3      21h
sh.helm.release.v1.test-nginx.v1   helm.sh/release.v1                  1      139m
[root@cce-1717-vpc-node2 ~]#
```

After the Secret is deleted, the chart release is deleted successfully.

```
[root@cce-1717-vpc-node2 ~]# kubectl -n default delete secret sh.helm.release.v1.test-nginx.v1
secret "sh.helm.release.v1.test-nginx.v1" deleted
[root@cce-1717-vpc-node2 ~]#
```

Note: If you perform operations on the console, CCE automatically bumps the original v2 chart release to v3 when you obtain or update the chart release. The release information is stored in the Secret. The release information in the original ConfigMap is not deleted. You are advised to query and delete the chart release in both the ConfigMap and Secret.

10.4 Does CCE Support nginx-ingress?

Introduction to nginx-ingress

nginx-ingress is a popular ingress-controller. It functions as a reverse proxy to import external traffic to a cluster and expose Services in Kubernetes clusters to external systems. In layer-7 load balancing (ingress), domain names are used to match Services. In this way, Services in a cluster can be accessed through domain names.

This chart is composed of ingress-controller and nginx.

- ingress-controller monitors Kubernetes ingresses and updates nginx configurations.

NOTE

For details about ingresses, see <https://kubernetes.io/docs/concepts/services-networking/ingress/>.

- nginx implements load balancing for requests and supports layer-7 request forwarding.

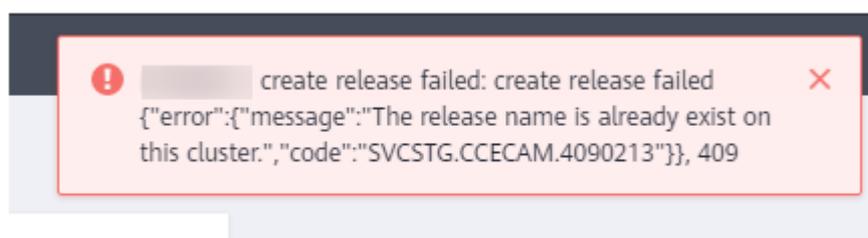
Installation Method

You can install the nginx-ingress add-on on the **Add-ons** page on the CCE console and configure its parameters.

10.5 What Should I Do If Installation of an Add-on Fails and "The release name is already exist" Is Displayed?

Symptom

When an add-on fails to be installed, the error message "The release name is already exist" is returned.



Possible Causes

The add-on release record remains in the Kubernetes cluster. Generally, it is because the cluster etcd has backed up and restored the add-on, or the add-on fails to be installed or deleted.

Solution

Use kubectl to connect to the cluster and manually clear the Secret and Configmap corresponding to add-on release. The following uses autoscaler add-on release as an example.

- Step 1** Connect to the cluster using kubectl, and run the following command to view the Secret list of add-on releases:

kubectl get secret -A |grep cceaddon

```
[root@cce-123-vpc-node2 ~]# kubectl get secret -nkube-system |grep cceaddon
sh.helm.release.v1.cceaddon-autoscaler.v1      helm.sh/release.v1      1      61s
sh.helm.release.v1.cceaddon-autoscaler.v2      helm.sh/release.v1      1      47s
sh.helm.release.v1.cceaddon-coredns.v1        helm.sh/release.v1      1      6h2m
sh.helm.release.v1.cceaddon-everest.v1        helm.sh/release.v1      1      6h2m
[root@cce-123-vpc-node2 ~]#
```

The Secret name of an add-on release is in the format of **sh.helm.release.v1.cceaddon-*{add-on name}*.v***. If there are multiple release versions, you can delete their Secrets at the same time.

- Step 2** Run the **release secret** command to delete the Secrets.

Example:

**kubectl delete secret sh.helm.release.v1.cceaddon-autoscaler.v1
sh.helm.release.v1.cceaddon-autoscaler.v2 -nkube-system**

```
[root@cce-123-vpc-node2 ~]# kubectl delete secret sh.helm.release.v1.cceaddon-autoscaler.v1 sh.helm.release.v1.cceaddon-autoscaler.v2 -nkube-system
secret "sh.helm.release.v1.cceaddon-autoscaler.v1" deleted
secret "sh.helm.release.v1.cceaddon-autoscaler.v2" deleted
[root@cce-123-vpc-node2 ~]#
```

- Step 3** If the add-on is created when Helm v2 is used, CCE automatically bumps the v2 release in Configmaps to v3 release in Secrets when viewing the add-ons and their details. The v2 release in the original Configmap is not deleted. Run the following command to view the ConfigMap list of add-on releases:

kubectl get configmap -A | grep cceaddon

```
cluster-autoscaler-th-config      1      7d10h
[paas@192-168-0-64 ~]$ kubectl get configmap -nkube-system | grep cceaddon
cceaddon-autoscaler.v1            1      7d10h
cceaddon-autoscaler.v2            1      52m
cceaddon-coredns.v1               1      14d
cceaddon-everest.v1               1      14d
[paas@192-168-0-64 ~]$
```

The ConfigMap name of an add-on release is in the format of **cceaddon-*{add-on name}*.v***. If there are multiple release versions, you can delete their ConfigMaps at the same time.

- Step 4** Run the **release configmap** command to delete the ConfigMaps.

Example:

kubectl delete configmap cceaddon-autoscaler.v1 cceaddon-autoscaler.v2 -nkube-system

```
[paas@192-168-0-64 ~]$ kubectl delete configmap cceaddon-autoscaler.v1 cceaddon-autoscaler.v2 -nkube-system
configmap "cceaddon-autoscaler.v1" deleted
configmap "cceaddon-autoscaler.v2" deleted
[paas@192-168-0-64 ~]$
```

 **CAUTION**

Deleting resources in kube-system is a high-risk operation. Ensure that the command is correct before running it to prevent resources from being deleted by mistake.

Step 5 On the CCE console, install add-on and then uninstall it. Ensure that the residual add-on resources are cleared. After the uninstall is complete, install the add-on again.

 **NOTE**

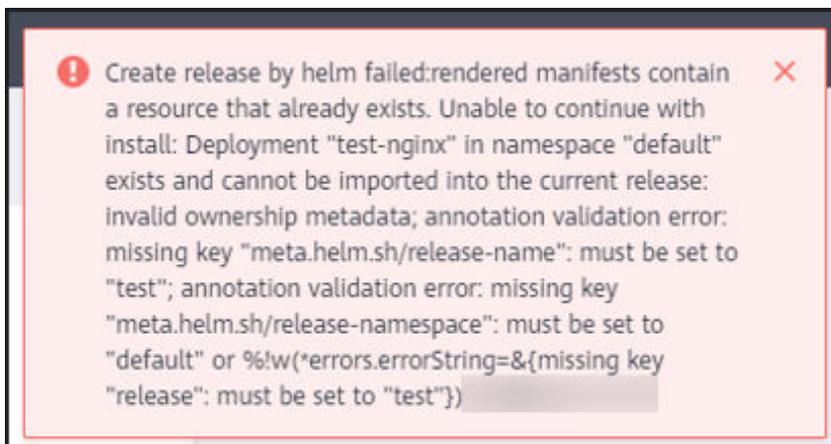
When installing the add-on for the first time, you may find it abnormal after the installation due to the residual resources of the previous add-on release, which is normal. In this case, you can uninstall the add-on on the console to ensure that the residual resources are cleared and the add-on can run properly after being installed again.

----End

10.6 What Should I Do If a Release Creation or Upgrade Fails and "rendered manifests contain a resource that already exists" Is Displayed?

Symptom

When a release cannot be created or upgraded, the error message "Create release by helm failed:rendered manifests contain a resource that already exists" is displayed. Unable to continue with install: ..., label validation error:missing key \"app.kubernetes.io/managed-by\":must be set to\"Helm\" ... Failed to create the release.



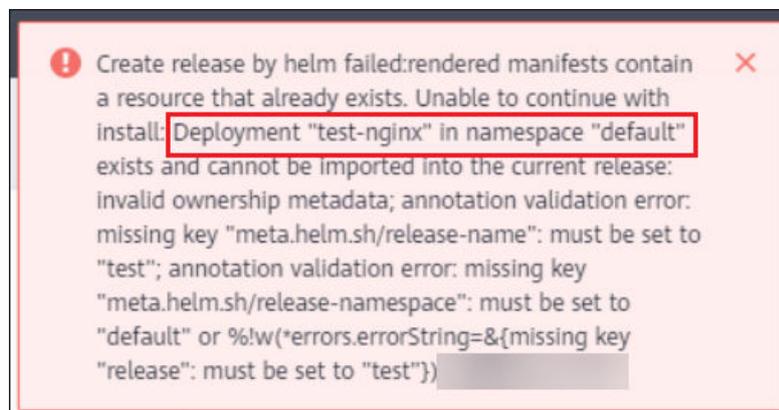
Possible Causes

If the preceding error information is displayed, the release is not created using Helm v3. If a release with the same name exists in the environment and does not have the home tag **app.kubernetes.io/managed-by: Helm** of Helm v3, a conflict message is displayed.

Solution

Delete the release and create it again using Helm.

- Step 1** Check the error message and locate the release that causes the conflict. Pay attention to the information following **Unable to continue with install:**. For example, the following error message indicates that a conflict occurs in the **test-nginx** Deployment in the **default** namespace.



- Step 2** Go to the cluster console or run the following kubectl command to delete the **test-nginx** Deployment. The preceding information is only an example. Perform operations according to the actual error information.

```
kubectl delete deploy test-nginx -n default
```

- Step 3** After the conflict is resolved, reinstall the chart.

----End

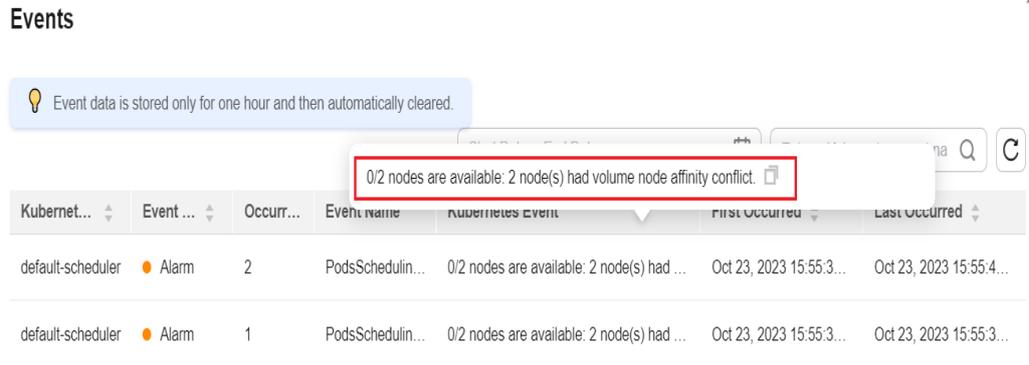
10.7 What Can I Do If the kube-prometheus-stack Add-on Instance Fails to Be Scheduled?

Symptom

During the installation of kube-prometheus-stack, the add-on retains in the partially ready state. The message "0/x nodes are available: x node(s) had volume node affinity conflict." is displayed in the event of the prometheus pod.

The same problem may occur during the installation of grafana.

Figure 10-3 Failed to schedule the prometheus pod



Possible Causes

The PV required by the prometheus pod already exists in the cluster, but the corresponding EVS disk is not in the same AZ as the node where the prometheus pod resides. As a result, the pod scheduling fails. This may be because kube-prometheus-stack is not installed for the first time in the cluster.

- If kube-prometheus-stack is installed for the first time, the attaching of an EVS disk (with the PVC named **pvc-prometheus-server-0**) to the prometheus pod will be delayed. When the EVS disk is created, it will automatically be in the same AZ as the node where the prometheus pod resides. For example, if the AZ of the node where the pod is running is AZ 1, the disk will be automatically created in AZ 1.
- When kube-prometheus-stack is uninstalled from the cluster, the PV mounted to the prometheus pod will not be deleted and the existing monitoring data will be retained. If the add-on is installed again, the nodes in the cluster may be newly created. If none of them is in AZ 1, the prometheus pod cannot run.

The causes may also result in the scheduling failure of a grafana pod.

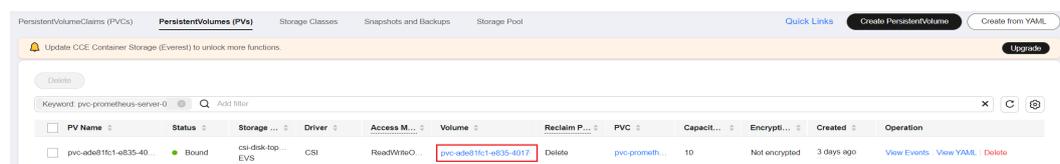
Solution

Check the AZ of the EVS disk corresponding to the existing PV mounted to the pod and create a node in the same AZ as this EVS disk.

Step 1 Log in to the CCE console and click the cluster name to access the cluster console.

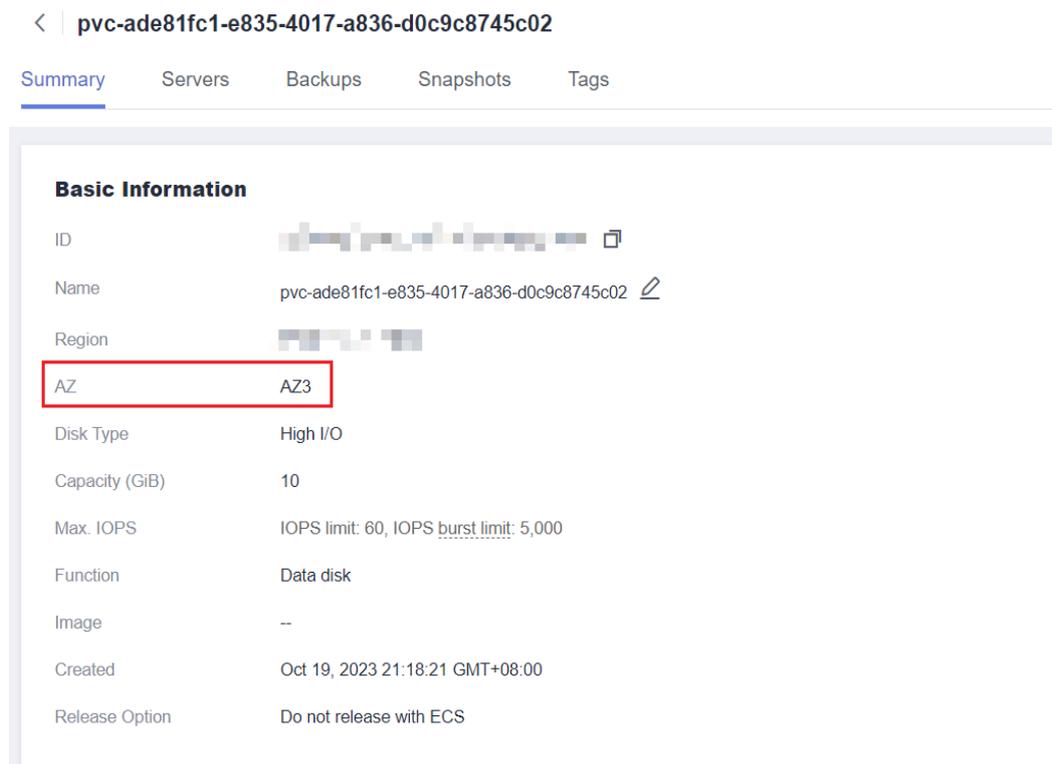
Step 2 In the navigation pane, choose **Storage**. Click the **PVs** tab, locate the row that contains the **pvc-prometheus-server-0** PVC in the **PVC** column, and click the volume name in the **Volume** column to go to the EVS disk details page.

Figure 10-4 Locating the target volume



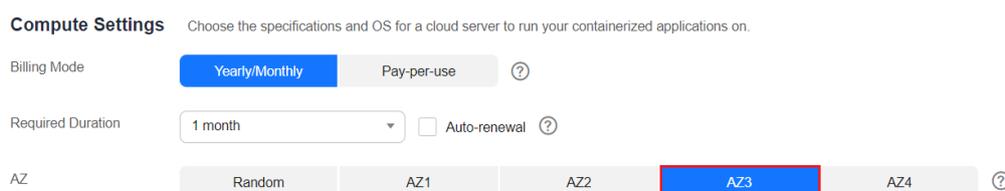
Step 3 In the **Basic Information** area, view the AZ of the EVS disk.

Figure 10-5 Viewing the details of the EVS disk



Step 4 On the CCE console, click the cluster name to access the cluster console. Choose **Nodes** in the navigation pane, click the **Nodes** tab, and click **Create Node** to create a node in the same AZ as the EVS disk.

Figure 10-6 Creating a node in a specified AZ



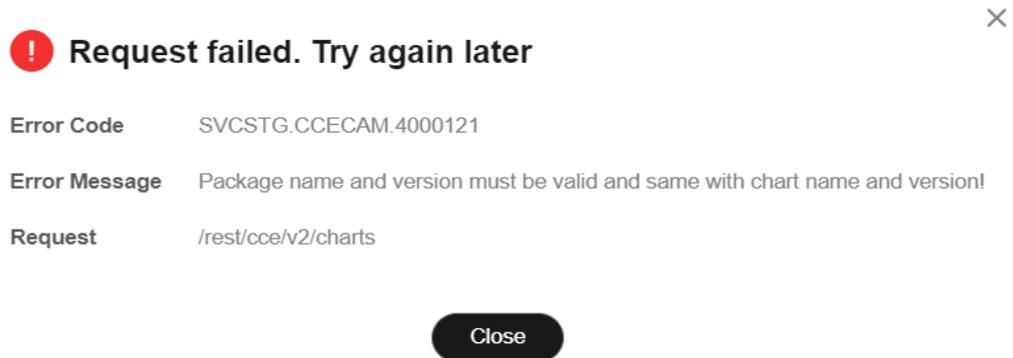
Step 5 Reschedule the node. This operation is automatically performed by the workload scheduler.

----End

10.8 What Can I Do If a Chart Fails to Be Uploaded?

Symptom

When a chart is uploaded, a request error "Request failed. Try again later" is displayed. The error code is **SVCSTG.CCECAM.4000121**, and the error message is "Package name and version must be valid and same with chart name and version!".

Figure 10-7 Chart upload failure

Possible Causes

If the preceding error information is displayed, the values of the **name** and **version** fields in the **Chart.yaml** file are inconsistent with those in the chart package.

NOTE

To customize the name and version of a chart package, modify the **name** and **version** fields in the **Chart.yaml** file.

Solution

Step 1 Check the **name** and **version** fields in the **Chart.yaml** file.

The following shows an example **Chart.yaml** file, where the chart name is **newer-nginx-ingress** and the version is **4.4.2**:

```
annotations:
  artifacthub.io/changes: |
    - Adding support for disabling liveness and readiness probes to the Helm chart
    - add:(admission-webhooks) ability to set securityContext
    - Updated Helm chart to use the fullname for the electionID if not specified
    - Rename controller-webhooks-networkpolicy.yaml
  artifacthub.io/prerelease: "false"
apiVersion: v2
appVersion: 1.5.1
description: Ingress controller for Kubernetes using NGINX as a reverse proxy and
  load balancer
home: https://github.com/kubernetes/ingress-nginx
icon: https://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Nginx_logo.svg/500px-
  Nginx_logo.svg.png
keywords:
  - ingress
  - nginx
kubeVersion: '>=1.20.0-0'
maintainers:
  - name: rikatz
  - name: strongjz
  - name: tao12345666333
name: newer-nginx-ingress
sources:
  - https://github.com/kubernetes/ingress-nginx
version: 4.4.2
```

Step 2 Change the name of the chart package based on the **Chart.yaml** file. The chart package is named in the format of **{Name}-{Version}.tgz**, for example, **newer-**

nginx-ingress-4.4.2.tgz. *{Version}* indicates the version number. It is named in the format of *{Major version number}.{Minor version number}.{Revision number}*.

Step 3 Upload the chart again.

----End

10.9 How Do I Configure the Add-on Resource Quotas Based on Cluster Scale?

After changing the cluster scale, adjust the add-on resource quotas based on the cluster scale to ensure that the add-on pods can run properly. For example, if you expand the cluster scale from 50 worker nodes to 200 worker nodes or more, increase the CPU and memory quotas of the add-on pods to avoid exceptions such as OOM caused by too many nodes required for scheduling the add-on pods.

Configuring Resource Quotas for coredns

Queries per Second (QPS) of the coredns add-on is positively correlated with the CPU consumption. If the number of nodes or containers in the cluster grows, the coredns pod will bear heavier workloads. Adjust the number of the coredns pods and their CPU and memory quotas based on the cluster scale.

Table 10-1 Recommended values for coredns

Node	Recommended Configuration	Pod	CPU Request	CPU Limit	Memory Request	Memory Limit
50	2500 QPS	2	500m	500m	512Mi	512Mi
200	5000 QPS	2	1000m	1000m	1024Mi	1024Mi
1000	10,000 QPS	2	2000m	2000m	2048Mi	2048Mi
2,000	20,000 QPS	4	2000m	2000m	2048Mi	2048Mi

Configuring Resource Quotas for everest

After the cluster scale is adjusted, the everest specifications need to be modified based on the cluster scale and the number of PVCs. The requested CPU and memory can be increased based on the number of nodes and PVCs. For details, see [Table 10-2](#).

In non-typical scenarios, the formulas for estimating the limit values are as follows:

- everest-csi-controller

- CPU limit: 250m for 200 or fewer nodes, 350m for 1000 nodes, and 500m for 2000 nodes
- Memory limit = (200 Mi + Number of nodes x 1 Mi + Number of PVCs x 0.2 Mi) x 1.2
- everest-csi-driver
 - CPU limit: 300m for 200 or fewer nodes, 500m for 1000 nodes, and 800m for 2000 nodes
 - Memory limit: 300 Mi for 200 or fewer nodes, 600 Mi for 1000 nodes, and 900 Mi for 2000 nodes

Table 10-2 Recommended configuration limits in typical scenarios

Configuration Scenario			everest-csi-controller		everest-csi-driver	
Nodes	PVs/PVCs	Add-on Instances	vCPUs (Limit = Requested)	Memory (Limit = Requested)	vCPUs (Limit = Requested)	Memory (Limit = Requested)
50	1000	2	250m	600 MiB	300m	300 MiB
200	1000	2	250m	1 GiB	300m	300 MiB
1000	1000	2	350m	2 GiB	500m	600 MiB
1000	5000	2	450m	3 GiB	500m	600 MiB
2000	5000	2	550m	4 GiB	800m	900 MiB
2000	10000	2	650m	5 GiB	800m	900 MiB

Configuring Resource Quotas for autoscaler

autoscaler automatically adjusts the number of nodes in a cluster based on workloads. Adjust the number of add-on pods and their CPU and memory quotas based on the cluster scale.

Table 10-3 Recommended values for autoscaler

Node	Pod	CPU Request	CPU Limit	Memory Request	Memory Limit
50	2	1000m	1000m	1000Mi	1000Mi
200	2	4000m	4000m	2000Mi	2000Mi
1,000	2	8000m	8000m	8000Mi	8000Mi
2,000	2	8000m	8000m	8000Mi	8000Mi

Configuring Resource Quotas for volcano

After the cluster scale is increased, the resource quotas required by volcano need to be modified based on the cluster scale.

- If the number of nodes is less than 100, retain the default configuration. The requested CPU is 500m, and the limit is 2000m. The requested memory is 500 MiB, and the limit is 2000 MiB.
- If the number of nodes is greater than 100, increase the requested CPU by 500m and the requested memory by 1000 MiB each time 100 nodes (10,000 pods) are added. Increase the CPU limit by 1500m and the memory limit by 1000 MiB.

NOTE

Formulas for calculating the requests:

- CPU request: Calculate the number of nodes multiplied by the number of pods, perform interpolation search using the product of the number of nodes in the cluster multiplied by the number of pods in [Table 10-4](#), and round up the request and limit that are closest to the specifications.

For example, for 2000 nodes (20,000 pods), the product of the number of nodes multiplied by the number of pods is 40 million, which is close to 700/70,000 in the specification (Number of nodes x Number of pods = 49 million). Set the CPU request to 4000m and the limit to 5500m.

- Memory request: Allocate 2.4 GiB of memory to every 1000 nodes and 1 GiB of memory to every 10,000 pods. The memory request is the sum of the two values. (The obtained value may be different from the recommended value in [Table 10-4](#). You can use either of them.)

Memory request = Number of nodes/1000 x 2.4 GiB + Number of pods/10000 x 1 GiB

For example, for 2000 nodes and 20,000 pods, the memory request value is 6.8 GiB (2000/1000 x 2.4 GiB + 20000/10000 x 1 GiB).

Table 10-4 Recommended values for volcano-controller and volcano-scheduler

Nodes/Pods in a Cluster	Requested vCPUs (m)	vCPU Limit (m)	Requested Memory (MiB)	Memory Limit (MiB)
50/5000	500	2000	500	2000
100/10,000	1000	2500	1500	2500
200/20,000	1500	3000	2500	3500
300/30,000	2000	3500	3500	4500
400/40,000	2500	4000	4500	5500
500/50,000	3000	4500	5500	6500
600/60,000	3500	5000	6500	7500
700/70,000	4000	5500	7500	8500

Configuring Resource Quotas for Other Add-ons

Resource quotas of other add-ons may also be insufficient due to cluster scale expansion. If, for example, the CPU or memory usage of the add-on pods increases and even OOM occurs, modify the resource quotas as required.

For example, the resources occupied by the kube-prometheus-stack add-ons are related to the number of pods in the cluster. If the cluster scale is expanded, the number of pods may also grow. In this case, increase the resource quotas of the prometheus pods.

10.10 How Can I Clean Up Residual Resources After the NGINX Ingress Controller Add-on in the Unknown State Is Deleted?

Symptom

The NGINX Ingress Controller add-on is in the unknown state, and after this add-on is uninstalled, residual components still remain.

Involved Kubernetes resources include:

- Namespace-level resources: secret, ConfigMap, Deployment, Service, Role, RoleBinding, lease, ServiceAccount, and job
- Cluster-level resources: ClusterRole, ClusterRoleBinding, IngressClass, and ValidatingWebhookConfiguration

Solution

Step 1 Use kubectl to access a cluster.

Step 2 Search for related resources.

```
className="nginx"
namespace="kube-system"
className=`if [[ ${className} == "nginx" ]]; then echo ""; else echo "-${className}";fi`
kubectl get -n ${namespace} secret sh.helm.release.v1.cceaddon-nginx-ingress${className}.v1 cceaddon-
nginx-ingress${className}-admission
kubectl get -n ${namespace} cm cceaddon-nginx-ingress${className}-controller
kubectl get -n ${namespace} deploy cceaddon-nginx-ingress${className}-controller cceaddon-nginx-ingress
${className}-default-backend
kubectl get -n ${namespace} svc cceaddon-nginx-ingress${className}-controller-admission cceaddon-nginx-
ingress${className}-default-backend cceaddon-nginx-ingress${className}-controller
kubectl get -n ${namespace} role cceaddon-nginx-ingress${className}
kubectl get -n ${namespace} rolebinding cceaddon-nginx-ingress${className}
kubectl get -n ${namespace} lease ingress-controller-leader${className}
kubectl get -n ${namespace} serviceAccount cceaddon-nginx-ingress${className}
kubectl get clusterRole cceaddon-nginx-ingress${className}
kubectl get clusterRoleBinding cceaddon-nginx-ingress${className}
kubectl get ingressClass ${className}
kubectl get ValidatingWebhookConfiguration cceaddon-nginx-ingress${className}-admission
```

className specifies the name of a controller. **namespace** specifies the namespace where NGINX Ingress Controller was installed.

Step 3 Manually delete the residual resources if the preceding resources are present.

----End

10.11 Why TLS v1.0 and v1.1 Cannot Be Used After the NGINX Ingress Controller Add-on Is Upgraded?

Symptom

After the NGINX Ingress Controller add-on is upgraded to 2.3.3 or later, if the TLS version of the client is earlier than v1.2, an error is reported during the negotiation between the client and NGINX Ingress Controller.

```
[root@~]# curl -I --tls-max 1.1 -kv https://192.168.0.141:443
* Trying 192.168.0.141:443...
* Connected to 192.168.0.141 (192.168.0.141) port 443 (#0)
* ALPN, offering h2
* ALPN, offering http/1.1
* successfully set certificate verify locations:
* CAfile: /etc/pki/tls/certs/ca-bundle.crt
* CApath: none
* TLSv1.1 (OUT), TLS handshake, Client hello (1):
* TLSv1.1 (IN), TLS alert, protocol version (582):
* error:1409442E:SSL routines:ssl3_read_bytes:tlsv1 alert protocol version
* Closing connection 0
curl: (35) error:1409442E:SSL routines:ssl3_read_bytes:tlsv1 alert protocol version
```

Solution

NGINX Ingress Controller 2.3.3 and later versions support only TLS v1.2 and v1.3 by default. If additional TLS versions are needed, you can add the `@SECLEVEL=0` field to the `ssl-ciphers` parameter configured for the NGINX Ingress Controller add-on. For details, see [TLS/HTTPS](#).

- Step 1** Log in to the CCE console and click the cluster name to access the cluster console. In the navigation pane, choose **Add-ons**. Choose **Installed** from the drop-down list, click **Manage** under **NGINX Ingress Controller**.
- Step 2** Click **Edit** of the corresponding instance.
- Step 3** Add the following configuration to the **Nginx Parameters**:

```
{
  "ssl-ciphers": "@SECLEVEL=0 ECDHE-ECDSA-AES128-GCM-SHA256:ECDHE-RSA-AES128-GCM-SHA256:ECDHE-ECDSA-AES256-GCM-SHA384:ECDHE-RSA-AES256-GCM-SHA384:ECDHE-ECDSA-CHACHA20-POLY1305:ECDHE-RSA-CHACHA20-POLY1305:DHE-RSA-AES128-GCM-SHA256:DHE-RSA-AES256-GCM-SHA384:DHE-RSA-CHACHA20-POLY1305:ECDHE-ECDSA-AES128-SHA256:ECDHE-RSA-AES128-SHA256:ECDHE-ECDSA-AES128-SHA:ECDHE-RSA-AES128-SHA:ECDHE-ECDSA-AES256-SHA384:ECDHE-RSA-AES256-SHA384:ECDHE-ECDSA-AES256-SHA:ECDHE-RSA-AES256-SHA:DHE-RSA-AES128-SHA256:DHE-RSA-AES256-SHA256:AES128-GCM-SHA256:AES256-GCM-SHA384:AES128-SHA256:AES256-SHA256:AES128-SHA:AES256-SHA:DES-CBC3-SHA",
  "ssl-protocols": "TLSv1 TLSv1.1 TLSv1.2 TLSv1.3"
}
```

- Step 4** Click **OK**.
- Step 5** Use TLS v1.1 for access again. The response is normal.

```
[root@... ]# curl -I --tls-max 1.1 -kv https://192.168.0.141:443
* Trying 192.168.0.141:443...
* Connected to 192.168.0.141 (192.168.0.141) port 443 (#0)
* ALPN, offering h2
* ALPN, offering http/1.1
* successfully set certificate verify locations:
* CAfile: /etc/pki/tls/certs/ca-bundle.crt
* CApath: none
* TLSv1.1 (OUT), TLS handshake, Client hello (1):
* TLSv1.1 (IN), TLS handshake, Server hello (2):
* TLSv1.1 (IN), TLS handshake, Certificate (11):
* TLSv1.1 (IN), TLS handshake, Server key exchange (12):
* TLSv1.1 (IN), TLS handshake, Server finished (14):
* TLSv1.1 (OUT), TLS handshake, Client key exchange (16):
* TLSv1.1 (OUT), TLS change cipher, Change cipher spec (1):
* TLSv1.1 (OUT), TLS handshake, Finished (20):
* TLSv1.1 (IN), TLS handshake, Finished (20):
```

----End

11 API & kubectl FAQs

11.1 How Can I Access a Cluster API Server?

You can use either of the following methods to access a cluster API server:

- (Recommended) Through the cluster API. This access mode uses certificate authentication. It is suitable for API calls on scale thanks to its direct connection to the API Server. This is a recommended option.
- API Gateway. This access mode uses token authentication. You need to obtain a token using your account. This access mode applies to small-scale API calls. API gateway flow control may be triggered when APIs are called on scale.

For details, see [Kubernetes APIs](#).

11.2 Can the Resources Created Using APIs or kubectl Be Displayed on the CCE Console?

The CCE console does not support the display of the following Kubernetes resources: DaemonSets, ReplicationControllers, ReplicaSets, and endpoints.

To query these resources, run the kubectl commands.

In addition, Deployments, StatefulSets, Services, and pods can be displayed on the console only when the following conditions are met:

- Deployments and StatefulSets: At least one label uses **app** as its key.
- Pods: Pods are displayed on the **Pods** tab page in the workload details only after a Deployment or StatefulSet has been created.
- Services: Services are displayed on the **Access Mode** tab page in the Deployment or StatefulSet details.

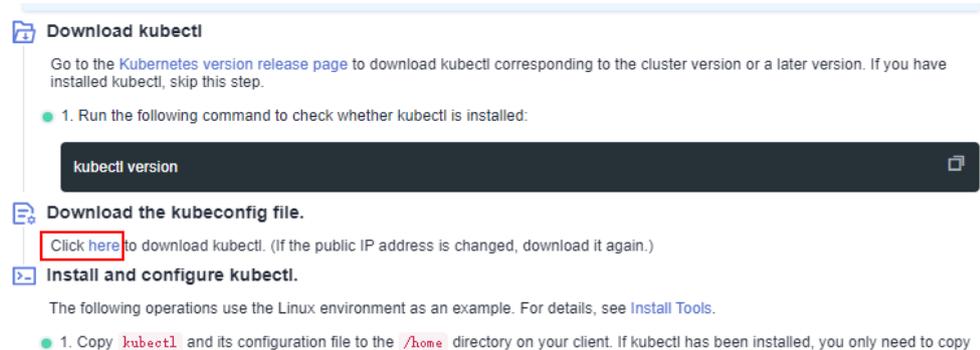
The Services displayed on this tab page are associated with the workload.

- a. At least one label of the workload uses **app** as its key.
- b. The label of a Service is the same as that of the workload.

11.3 How Do I Download kubeconfig for Connecting to a Cluster Using kubectl?

- Step 1** Log in to the CCE console. Click the target cluster to go to its details page.
- Step 2** In the **Connection Information** area, view the kubectl connection mode.
- Step 3** In the window that is displayed, download the kubectl configuration file (**kubeconfig.json**).

Figure 11-1 Downloading kubeconfig.json



----End

11.4 How Do I Rectify the Error Reported When Running the kubectl top node Command?

Symptom

The error message "Error from server (ServiceUnavailable): the server is currently unable to handle the request (get nodes.metrics.k8s.io)" is displayed after the **kubectl top node** command is executed.

Possible Causes

"Error from server (ServiceUnavailable)" indicates that the cluster is not connected. In this case, you need to check whether the network between kubectl and the master node in the cluster is normal.

Solution

- If the kubectl command is executed outside the cluster, check whether the cluster is bound to an EIP. If yes, download the **kubeconfig** file and run the kubectl command again.
- If the kubectl command is executed on a node in the cluster, check the security group of the node and check whether the TCP/UDP communication between the worker node and master node is allowed. For details about the security group, see [How Can I Configure a Security Group Rule in a Cluster?](#)

11.5 Why Is "Error from server (Forbidden)" Displayed When I Use kubectl?

Symptom

When you use kubectl to create or query Kubernetes resources, the following output is returned:

```
# kubectl get deploy Error from server (Forbidden): deployments.apps is forbidden:
User "0c97ac3cb280f4d91fa7c0096739e1f8" cannot list resource "deployments" in
API group "apps" in the namespace "default"
```

Possible Cause

This user has no permissions to operate Kubernetes resources.

Solution

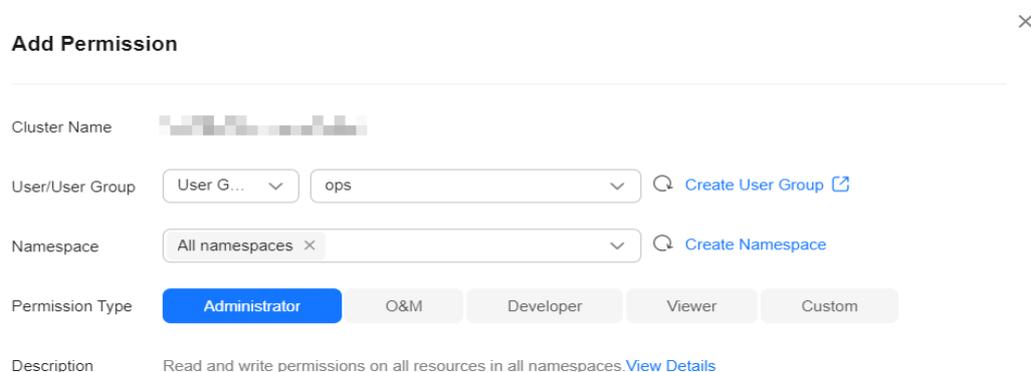
Assign permissions to the user.

- Step 1** Log in to the CCE console. In the navigation pane, choose **Permissions**.
- Step 2** Select a cluster for which you want to add permissions from the drop-down list on the right.
- Step 3** Click **Add Permissions** in the upper right corner.
- Step 4** Confirm the cluster name and select the namespace to assign permissions for. For example, select **All namespaces**, the target user or user group, and select the permissions.

NOTE

If you do not have IAM permissions, you cannot select users or user groups when configuring permissions for other users or user groups. In this case, you can enter a user ID or user group ID.

Figure 11-2 Configuring namespace permissions



Add Permission ×

Cluster Name

User/User Group 🔍 Create User Group

Namespace 🔍 Create Namespace

Permission Type Administrator O&M Developer Viewer Custom

Description Read and write permissions on all resources in all namespaces. [View Details](#)

Permissions can be customized as required. After selecting **Custom** for **Permission Type**, click **Add Custom Role** on the right of the **Custom** parameter. In the dialog

box displayed, enter a name and select a rule. After the custom rule is created, you can select a value from the **Custom** drop-down list box.

Custom permissions are classified into ClusterRole and Role. Each ClusterRole or Role contains a group of rules that represent related permissions. For details, see [Using RBAC Authorization](#).

- A ClusterRole is a cluster-level resource that can be used to configure cluster access permissions.
- A Role is used to configure access permissions in a namespace. When creating a Role, specify the namespace to which the Role belongs.

Figure 11-3 Custom permission

Add Custom Role ×

Name

Type ClusterRole Role

Rule ⓘ All operations: *
Read-only: get + list + watch
Read-write: get + list + watch + create + update + patch + delete

[+ Add](#)

Step 5 Click **OK**.

----End

12 DNS FAQs

12.1 What Should I Do If Domain Name Resolution Fails in a CCE Cluster?

Check Item 1: Whether the coredns Add-on Has Been Installed

- Step 1** Log in to the CCE console and click the cluster name to access the cluster console.
- Step 2** In the navigation pane, choose **Add-ons** and check whether the CoreDNS add-on has been installed.
- Step 3** If not, install the add-on. For details, see [Why Does a Container in a CCE Cluster Fail to Perform DNS Resolution?](#)

----End

Check Item 2: Whether the coredns Instance Reaches the Performance Limit

CoreDNS QPS is positively correlated with the CPU usage. If the QPS is high, adjust the coredns instance specifications based on the QPS. If a cluster has more than 100 nodes, you are advised to use NodeLocal DNSCache to improve DNS performance. For details, see [Using NodeLocal DNSCache to Improve DNS Performance](#).

- Step 1** Log in to the CCE console and click the cluster name to access the cluster console.
- Step 2** In the navigation tree, choose **Add-ons** and verify that CoreDNS is running.
- Step 3** Click the coredns add-on name to view the add-on list.
- Step 4** Click **Monitor** of the coredns add-on to view the CPU and memory usage.

If the add-on performance reaches the bottleneck, adjust the coredns add-on specifications.

1. In the **Add-ons Installed** area, locate the row that contains the coredns add-on and click **Edit**. The add-on details page is displayed.
2. In the **Specifications** area, configure the coredns add-on. You can use the CoreDNS QPS as required.

3. You can also select **Custom qps** and set the number of pods, CPU quota, and memory quota.

Specifications

Add-on Specifications: 2500 qps, 5000 qps, 10000 qps, 20000 qps, **Custom qps**

Pods:

Containers: coredns

CPU Quota: Request Limit Memory Quota: Request Limit

- The request value must be less than or equal to the limit value, otherwise it cannot be created successfully.
- Please ensure that the node resources under the cluster are sufficient, otherwise it cannot be created successfully.

4. Click **OK**.

----End

Check Item 3: Whether the External Domain Name Resolution Is Slow or Times Out

If the domain name resolution failure rate is lower than 1/10000, optimize parameters by referring to [How Do I Optimize the Configuration If the External Domain Name Resolution Is Slow or Times Out?](#) or add a retry policy in the service.

Check Item 4: Whether UnknownHostException Occurs

When service requests in the cluster are sent to an external DNS server, a domain name resolution error occurs due to occasional UnknownHostException. UnknownHostException is a common exception. When this exception occurs, check whether there is any domain name-related error or whether you have entered a correct domain name.

To locate the fault, perform the following steps:

- Step 1** Check the host name carefully (spelling and extra spaces).
- Step 2** Check the DNS settings. Before running the application, run the **ping hostname** command to ensure that the DNS server has been started and running. If the host name is new, you need to wait for a period of time before the DNS server is accessed.
- Step 3** Check the CPU and memory usage of the coredns add-on to determine whether the performance bottleneck has been reached. For details, see [Check Item 2: Whether the coredns Instance Reaches the Performance Limit](#).
- Step 4** Check whether traffic limiting is performed on the coredns add-on. If traffic limiting is triggered, the processing time of some requests may be prolonged. In this case, you need to adjust the coredns add-on specifications.

Log in to the node where the coredns add-on is installed and view the following content:

```
cat /sys/fs/cgroup/cpu/kubepods/pod<pod_uid>/<coredns container ID>/cpu.stat
```

- *<pod uid>* indicates the pod UID of the coredns add-on, which can be obtained by running the following command:

```
kubectl get po <pod name> -nkube-system -ojsonpath='{.metadata.uid}'
```

In the preceding command, *<pod name>* indicates the name of the coredns add-on running on the current node.

- *<coredns container ID>* must be a complete container ID, which can be obtained by running the following command:

Docker nodes:

```
docker ps --no-trunc | grep k8s_coredns | awk '{print $1}'
```

containerd nodes:

```
crictl ps --no-trunc | grep k8s_coredns | awk '{print $1}'
```

Example:

```
cat /sys/fs/cgroup/cpu/kubepods/  
pod27f58662-3979-448e-8f57-09b62bd24ea6/6aa98c323f43d689ac47190bc84cf4fadd23bd8dd25307f773df2  
5003ef0eef0/cpu.stat
```

Pay attention to the following metrics:

- **nr_throttled**: number of times that traffic is limited.
- **throttled_time**: total duration of traffic limiting, in nanoseconds.

----End

If the host name and DNS settings are correct, you can use the following optimization policies.

Optimization policies:

1. Change the coredns cache time.
2. Configure the stub domain.
3. Modify the value of **ndots**.

NOTE

- **Increasing the cache time of coredns** helps resolve the same domain name for the N time, reducing the number of cascading DNS requests.
- **Configuring the stub domain** can reduce the number of DNS request links.

How to modify:

1. Modifying the coredns cache time and configuring the stub domain:

[Configuring the Stub Domain for CoreDNS](#)

Restart the coredns add-on after you modify the configurations.

2. Modifying **ndots**:

[How Do I Optimize the Configuration If the External Domain Name Resolution Is Slow or Times Out?](#)

Example:

```
dnsConfig:  
  options:  
    - name: timeout  
      value: '2'  
    - name: ndots  
      value: '5'  
    - name: single-request-reopen
```

You are advised to change the value of **ndots** to **2**.

12.2 Why Does a Container in a CCE Cluster Fail to Perform DNS Resolution?

Symptom

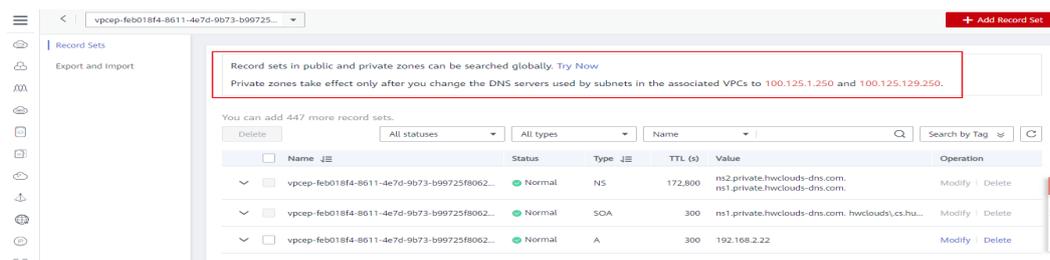
A customer bound its domain name to the private domain names in the DNS service and also to a specific VPC. It is found that the ECSs in the VPC can properly resolve the private domain name but the containers in the VPC cannot.

Application Scenario

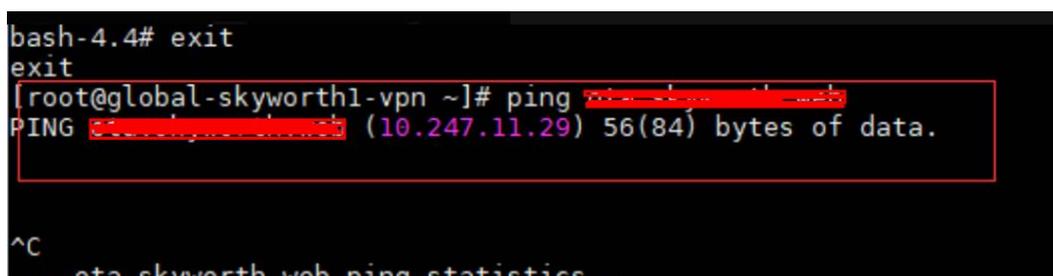
Containers in a VPC cannot resolve domain names.

Solution

According to the resolution rules of private domain names, the subnet DNS in the VPC must be set to the cloud DNS. You can find the details of the private network DNS service on its console.



The customer can perform domain name resolution on the ECSs in the VPC subnet, which indicates that the preceding configuration has been completed in the subnet.



However, when the domain name resolution is performed in a container, the message "bad address" is displayed, indicating that the domain name cannot be resolved.



Log in to the CCE console and check the add-ons installed in the cluster.

If you find that the `coredns` add-on does not exist in **Add-ons Installed**, the `coredns` add-on may have been incorrectly uninstalled.

Install it and add the corresponding domain name and DNS service address to resolve the domain name.

12.3 Why Cannot the Domain Name of the Tenant Zone Be Resolved After the Subnet DNS Configuration Is Modified?

Symptom

After a DNS server record, for example, 114.114.114.114, is added to the DNS configuration of the user cluster subnet, the domain name of the tenant zone cannot be resolved.

Cause Analysis

CCE configures the subnet DNS information of the user on the node, which is also used by the `coredns` add-on. As a result, the domain name fails to be resolved by the node container occasionally.

Solution

You are advised to modify the stub domain of the `coredns` add-on to update the DNS configuration of the user cluster subnet. For details, see [Configuring the Stub Domain for CoreDNS](#).

12.4 How Do I Optimize the Configuration If the External Domain Name Resolution Is Slow or Times Out?

The following is an example `resolv.conf` file for a container in a workload:

```
root@test-5dffddd95-vpt4m:/# cat /etc/resolv.conf
nameserver 10.247.3.10
search istio.svc.cluster.local svc.cluster.local cluster.local
options ndots:5 single-request-reopen timeout:2
```

In the preceding information:

- **nameserver**: IP address of the DNS. Set this parameter to the cluster IP address of CoreDNS.
- **search**: domain name search list, which is a common suffix of Kubernetes.
- **ndots**: If the number of dots (.) is less than the domain name, **search** is preferentially used for resolution.
- **timeout**: timeout interval.
- **single-request-reopen**: indicates that different source ports are used to send different types of requests.

By default, when you create a workload on the CCE console, the preceding parameters are configured as follows:

```
dnsConfig:
  options:
    - name: timeout
      value: '2'
    - name: ndots
      value: '5'
    - name: single-request-reopen
```

These parameters can be optimized or modified based on service requirements.

Scenario 1: Slow External Domain Name Resolution

Optimization Solution

1. If the workload does not need to access the Kubernetes Service in the cluster, see [How Do I Configure a DNS Policy for a Container?](#)
2. If the number of dots (.) in the domain name used by the working Service to access other Kubernetes Services is less than 2, set **ndots** to 2.

Scenario 2: External Domain Name Resolution Timeout

Optimization Solution

1. Generally, the timeout of a Service must be greater than the value of **timeout** multiplied by **attempts**.
2. If it takes more than 2s to resolve the domain name, you can set **timeout** to a larger value.

12.5 How Do I Configure a DNS Policy for a Container?

CCE uses **dnsPolicy** to identify different DNS policies for each pod. The value of **dnsPolicy** can be either of the following:

- **None:** No DNS policy is configured. In this mode, you can customize the DNS configuration, and **dnsPolicy** needs to be used together with **dnsConfig** to customize the DNS.
- **Default:** The pod inherits the name resolution configuration from the node where the pod is running. The container's DNS configuration file is the DNS configuration file that the kubelet's **--resolv-conf** flag points to. In this case, a cloud DNS is used for CCE clusters.
- **ClusterFirst:** In this mode, the DNS in the pod uses the DNS service configured in the cluster. That is, the kube-dns or CoreDNS service in the Kubernetes is used for domain name resolution. If the resolution fails, the DNS configuration of the host machine is used for resolution.

If the type of **dnsPolicy** is not specified, **ClusterFirst** is used by default.

- If the type of **dnsPolicy** is set to **Default**, the name resolution configuration is inherited from the worker node where the pod is running.
- If the type of **dnsPolicy** is set to **ClusterFirst**, DNS queries will be sent to the kube-dns service.

The kube-dns service responds to queries on the domains that use the configured cluster domain suffix as the root. All other queries (for example, `www.kubernetes.io`) are forwarded to the upstream name server inherited from the node. Before this feature was supported, stub domains were typically introduced by a custom resolver, instead of the upstream DNS. However, this causes the custom resolver itself to be the key path to DNS resolution, where scalability and availability issues can make the DNS functions unavailable to the cluster. This feature allows you to introduce custom resolvers without taking over the entire resolution path.

If a workload does not need to use CoreDNS in the cluster, you can use `kubectl` or call the APIs to set the `dnsPolicy` to `Default`.

13 Image Repository FAQs

13.1 How Do I Create a Docker Image and Solve the Problem of Slow Image Pull?

Creating a Docker Image

For details about how to use Dockerfile to customize a Docker image for a simple web application, see [Docker Basics](#) or [How Do I Create a Docker Image?](#).

Accelerating Image Pull

Public images may be pulled slowly due to carrier network. You can upload frequently used images to SWR to improve the image pull speed.

Introduction to SWR

SWR provides full-lifecycle container image management, which is easy-to-use, secure, and reliable. SWR enables users to quickly deploy containerized services. SWR can be used as an image repository to store and manage Docker images.

SWR FAQs

[General FAQs](#)

13.2 How Do I Upload My Images to CCE?

SoftWare Repository for Container (SWR) manages images for CCE. It provides the following ways to upload images:

- [Uploading an Image Through a Container Engine Client](#)
- [Uploading an Image Through SWR Console](#)

For details about how to smoothly migrate from Harbor to SWR, see [Synchronizing Images Across Clouds from Harbor to SWR](#).

14 Permissions

14.1 Can I Configure Only Namespace Permissions Without Cluster Management Permissions?

Namespace permissions and cluster management permissions are independent and complementary to each other.

- Namespace permissions: apply to clusters and are used to manage operations on cluster resources (such as creating workloads).
- Cluster management (IAM) permissions: apply to cloud services and used to manage CCE clusters and peripheral resources (such as VPC, ELB, and ECS).

Administrators of the IAM Admin user group can grant cluster management permissions (such as CCE Administrator and CCE FullAccess) to IAM users or grant namespace permissions on a cluster on the CCE console. However, the permissions you have on the CCE console are determined by the IAM system policy. If the cluster management permissions are not configured, you do not have the permissions for accessing the CCE console.

If you only run `kubectl` commands to work on cluster resources, you only need to obtain the `kubeconfig` file with the namespace permissions. For details, see [Can I Use kubectl If the Cluster Management Permissions Are Not Configured?](#)

Note that information leakage may occur when you use the `kubeconfig` file.

14.2 Can I Use CCE APIs If the Cluster Management Permissions Are Not Configured?

CCE has cloud service APIs and cluster APIs.

- Cloud service APIs: You can perform operations on the infrastructure (such as creating nodes) and cluster resources (such as creating workloads).

When using cloud service APIs, the cluster management (IAM) permissions must be configured.

- Cluster APIs: You can perform operations on cluster resources (such as creating workloads) through the Kubernetes native API server, but not on cloud infrastructure resources (such as creating nodes).

When using cluster APIs, you only need to add the cluster certificate. Only the users with the cluster management (IAM) permissions can **download** the cluster certificate. Note that information leakage may occur during certificate transmission.

14.3 Can I Use kubectl If the Cluster Management Permissions Are Not Configured?

IAM authentication is not required for running kubectl commands. Therefore, you can run kubectl commands without configuring cluster management (IAM) permissions. However, you need to obtain the kubectl configuration file (kubeconfig) with the namespace permissions. In the following scenarios, information leakage may occur during file transmission.

- Scenario 1

If an IAM user has been configured with the cluster management permissions and namespace permissions, downloads the kubeconfig authentication file and then deletes the cluster management permissions (reserving the namespace permissions), kubectl can still be used to perform operations on Kubernetes clusters. Therefore, if you want to permanently delete the permission of a user, you must also delete the cluster management permissions and namespace permissions of the user.

- Scenario 2

An IAM user has certain cluster management and namespace permissions and downloads the kubeconfig authentication file. In this case, CCE determines which Kubernetes resources can be accessed by kubectl based on the user information. That is, the authentication information of a user is recorded in kubeconfig. Anyone can use kubeconfig to access the cluster.

15 Related Services

15.1 What Are the Differences Between CCE and CCI?

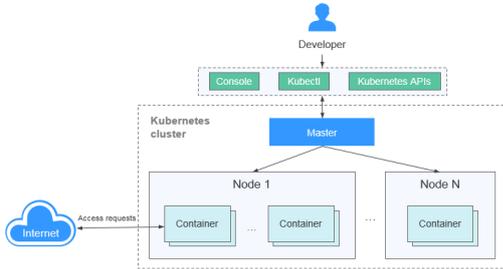
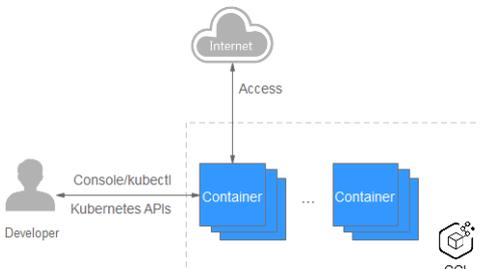
Description

Table 15-1 Introduction to CCE and CCI

Cloud Container Engine (CCE)	Cloud Container Instance (CCI)
<p>CCE provides highly scalable, high-performance, enterprise-class Kubernetes clusters and supports Docker containers. CCE is a one-stop container platform that provides full-stack container services from Kubernetes cluster management, lifecycle management of containerized applications, application service mesh, and Helm charts to add-on management, application scheduling, and monitoring and O&M. With CCE, you can easily deploy, manage, and scale containerized applications on Huawei Cloud.</p> <p>For details, see What Is Cloud Container Engine?</p>	<p>Cloud Container Instance (CCI) is a serverless container engine that allows you to run containers without creating or managing server clusters. With CCI, you only need to manage containerized services running on Kubernetes. You can quickly create and run container workloads on CCI without managing clusters and servers. Because of the serverless architecture, CCI frees you from containerized application O&M and allows you to focus on the services themselves.</p> <p>With the serverless architecture, you can focus on building and operating applications without having to create or manage servers, not to mention the issues caused by abnormal server running. All you have to do is to specify resource requirements (on CPU and memory, for example). This gives you a more focused approach to business needs and helps you reduce management and maintenance costs. Traditionally, to run containerized workloads using Kubernetes, you need to create a Kubernetes cluster first.</p>

Creation Mode

Table 15-2 Creation modes

Cloud Container Engine (CCE)	Cloud Container Instance (CCI)
<p>CCE is a hosted Kubernetes service for container management. It allows you to create native Kubernetes clusters with just a few clicks.</p> <p>You need to create clusters and nodes to use CCE. They are easy to create on an intuitive console and highly available. You do not need to manage master nodes.</p> 	<p>CCI provides a serverless container engine. When deploying containers on Huawei Cloud, you do not need to purchase and manage ECSs, eliminating the need for O&M and management.</p> <p>You do not need to create clusters, master nodes, or work nodes, but directly start applications.</p> 

Billing

Table 15-3 Different billing modes

Aspect	Cloud Container Engine (CCE)	Cloud Container Instance (CCI)
Pricing	Related resources (such as nodes and bandwidth) will be created when CCE is used. You need to pay for these resources.	CCI instance resources include CPUs, memory, and GPUs. You will be billed by the actual instance resource specifications.
Billing mode	Pay-per-use and yearly/monthly billing modes are supported.	Pay-per-use billing mode is supported.
Minimum pricing unit	By hour	Billed by second. The bill run period is hour.

Application Scenarios

Table 15-4 Different application scenarios

Cloud Container Engine (CCE)	Cloud Container Instance (CCI)
<p>Applicable to all scenarios. Generally, large-scale and long-term stable applications are running. For example:</p> <ul style="list-style-type: none"> • E-commerce • Service mid-end • IT system 	<p>Applicable to scenarios with obvious peak and off-peak hours. Resources can be flexibly requested to improve resource utilization. For example:</p> <ul style="list-style-type: none"> • Batch computing • High-performance computing • Scale-out upon traffic bursts • CI/CD test

Cluster Creation

Table 15-5 Creation modes

Cloud Container Engine (CCE)	Cloud Container Instance (CCI)
<p>Process of using CCE:</p> <ol style="list-style-type: none"> 1. Creating a cluster Configure basic information such as the name, region, and network. 2. Creating a node Specify the node specifications and data disk size. 3. Configuring the cluster Install cluster add-ons, such as networking, monitoring, and logs. 4. Creating a workload in the cluster 	<p>Process of using CCI:</p> <ol style="list-style-type: none"> 1. Creating a namespace Configure basic information such as the name, region, and network. 2. Creating a workload

Cooperation Between CCE and CCI

By installing the virtual-kubelet add-on, you can use CCI to deploy pods for your Deployments, StatefulSets, and jobs on CCE when service spikes occur, which can reduce consumption caused by cluster scaling.

Functions:

- Creates pods automatically in seconds. When CCE cluster resources are insufficient, you do not need to add nodes to the cluster. virtual-kubelet automatically creates pods on CCI, eliminating the overhead of resizing the CCE cluster.
- Seamlessly works with Huawei Cloud SWR for you to use public and private images.

- Supports event synchronization, monitoring, logging, remote command execution, and status query for CCI pods.
- Allows you to view the capacity information about virtual elastic nodes.
- Supports connectivity between CCE and CCI pods through Services.

For details, see [Elastic Scaling of CCE Pods to CCI](#).

15.2 What Are the Differences Between CCE and ServiceStage?

In terms of use, CCE focuses on pod deployment, and ServiceStage focuses on service usage.

In terms of technical implementation, ServiceStage encapsulates CCE capabilities.

Basic Concepts

Cloud Container Engine (CCE)

CCE provides highly scalable, high-performance, enterprise-class Kubernetes clusters and supports Docker containers. CCE is a one-stop container platform that provides full-stack container services from Kubernetes cluster management, lifecycle management of containerized applications, application service mesh, and Helm charts to add-on management, application scheduling, and monitoring and O&M. With CCE, you can easily deploy, manage, and scale containerized applications on HUAWEI CLOUD.

ServiceStage

ServiceStage is an application and microservice management platform that helps enterprises simplify application lifecycle management from deployment, monitoring, and O&M, to governance. ServiceStage provides a full-stack solution for enterprises to develop microservice, mobile, and web applications. This solution helps enterprises easily migrate various applications onto the cloud, allowing enterprises to focus on service innovation for digital transformation.