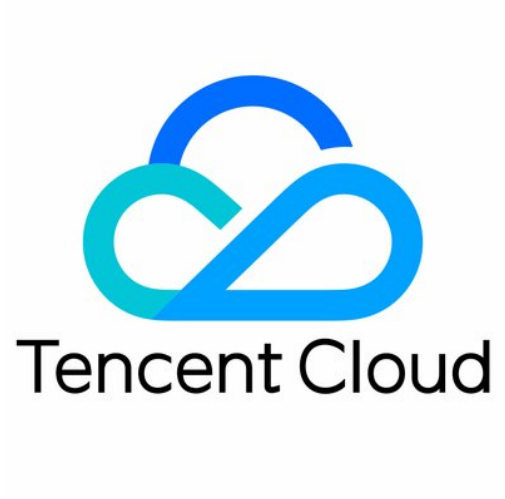
Tencent Kubernetes Engine

Purchase Guide

Product Documentation



Contents

Purchase Guide
Purchasing Clusters
TKE Billing Overview

Last updated：2022-07-11 15:05:26

Billable Items

The service fees for TKE consists of two parts, **cluster management fees** and **Tencent Cloud service resources fees**.

- **Cluster management fees**

> Note：
> Tencent Cloud starts charging for managed clusters from 10:00, March 21, 2022 (UTC +8). See Starting Charging for Managed Clusters.

Managed clusters incur the cluster management fees based on their cluster models. For more information, see Cluster Management Fees.

- **Tencent Cloud service resources fees**
  Other Tencent Cloud services resources (such as CVM, CBS and CLB) created during the usage of TKE will be charged based on the billing mode for each resource. For more information, see Tencent Cloud Services Resources Fees.

Cluster Management Fees

Billing mode

The billing mode of pay-as-you-go is usually adopted for TKE.

Billing Item	Billing Mode	Payment Method	Billing Unit
Number of clusters	Pay-as-you-go	Freeze the fees at the time of purchase, and the service is billed at an hourly basis	USD/hour

Recommendations for small clusters

- If your cluster has only a few nodes (less than 20), we highly recommend you use Elastic Kubernetes Service (EKS). With EKS, you can deploy workloads and pay for actual container usage, with no need to purchase nodes and pay cluster management fees.
- You can choose to migrate your existing TKE clusters as needed in the following ways:
  - Conduct smooth business migration through supernodes to reduce the number of nodes in the TKE cluster and thereby lower the cluster management fees (such fees are not charged for supernodes; for more information, see Pricing below).
  - Completely migrate the TKE cluster to the EKS cluster through the migration tool. For more information, see Guide on Migrating Resources in a TKE Managed Cluster to an EKS Cluster . You can submit a ticket for assistance.

Pricing

Note：

- The unit prices are varied depending on the region. Please refer to the prices displayed in the console.

- Please read the Purchase Instructions carefully before you choose the specification.

Cluster Specification	Price (USD/hour)
L5	0.02040816
L20	0.06279435
L50	0.11459969
L100	0.19152276
L200	0.40031397
L500	0.8021978
L1000	1.47252747
L3000	2.44897959
L5000	4.40188383

Tencent Cloud Service Resources Fees

Other Tencent Cloud service resources (such as CVM, CBS and CLB) created during the usage of TKE will be charged based on each billing mode. For more information, see billing description for each resource.

Tencent Cloud Service	Documentation
CVM	CVM Billing Mode
CBS	CBS Billing Overview
CLB	CLB Billing Description

Note：
TKE is a declarative service based on Kubernetes. When you do not need CLB, CBS or other IaaS service resources created by TKE, you must delete them in TKE console, otherwise, TKE will re-create them and continue to charge fees. For example, if you delete a CLB instance in CLB console instead of in TKE console, TKE will re-create a CLB instance based on declarative APIs.

Purchase Instructions

Last updated：2022-06-27 10:59:28

Purchase Notes

Note：
If the service is unavailable due to the failure of you to abide by the following suggestions, the corresponding service downtime shall not be counted towards the service unavailability period. For more information, see TKE Service Level Agreement.

The availability of TKE clusters are relevant to the number of resources (such as Pod, ConfigMap, CRD and Event) in the cluster, as well as QPS of Get/List read operations and Patch/Delete/Create/Update write operations of the resources. To improve the cluster availability, do not initiate List-like operations for clusters with large amount of resources, and do not write too many ConfigMap/CRDs/EndPoints into the cluster.

The common **List-like operations** are as follows (take Pod resources as an example):

- Query with a label

```
kubectl get pod -l app=nginx
```

- Query for a specified namespace

```
kubectl get pod -n default
```

- Query the Pods across the cluster

```
kubectl get pod --all-namespaces
```

- Initiate a List request through client-go

```
k8sClient.CoreV1().Pods("").List(metav1.ListOptions{})
```

If you want to **query all resources in the cluster**, it is recommended that you use the **informer mechanism** to query through local cache. In some simple scenarios, you can add the ResourceVersion parameter in List to query in kube-apiserver cache. For example, `k8sClient.CoreV1().Pods("").List(metav1.ListOptions{ResourceVersion: "0"})`. Note: When you query in kube-apiserver cache, if you initiate List requests frequently to many resources, it still causes high pressure on kube-apiserver memory. It is recommended that you use this query method for low-frequency requests.

Recommended Configuration

Refer to the recommended configuration below and choose the model that best suits your requirements, so as to prevent cluster unavailability caused by heavy load of the control plane.
Assume that you want to deploy 50 nodes in a cluster and need 2,000 Pods. You need to select a model with the maximum number of nodes of 100, but not 50.

Note：

- The nodes indicate Kubnernetss nodes, including CVM nodes, BM nodes and external nodes. **Supernodes are excluded**.
- The number of Pods includes Pods in all namespaces and in any status, and excludes the Pods related to system components such as cni-agent.
- ConfigMap does not include the Pods related to system components such as cni-agent.
- **Maximum other resources** refers to the number of resources in the manged cluster excluding the Pods, nodes and ConfigMap. For example, for a L100 cluster, the number of the resources, such as ClusterRole, Services and Endpoints, cannot exceed 2,500 respectively.
- It is recommended that the size of all objects under each type of resource does not exceed 800 MiB, and the size of each object does not exceed 100 KB.

Cluster specification	Max nodes	Max Pods (recommended)	Max ConfigMap (recommended)	Maximum CRDs/Maximum other resources (recommended)
L5	5	150	128	150
L20	20	600	256	600
L50	50	1,500	512	1,250
L100	100	3,000	1,024	2,500
L200	200	6,000	2,048	5,000
L500	500	15,000	4,096	10,000

Cluster specification	Max nodes	Max Pods (recommended)	Max ConfigMap (recommended)	Maximum CRDs/Maximum other resources (recommended)
L1000	1,000	30,000	6,144	20,000
L3000	3,000	90,000	8,192	50,000
L5000	5,000	150,000	10,240	100,000

Payment Overdue

Last updated：2022-06-22 11:32:31

Overdue Payment

Your managed clusters will be processed as instructed below since **your account balance falls below 0**.

- Within 24 hours: Your TKE managed clusters can be used and are billed.
- After 24 hours: All managed clusters under the account are in **isolated** status and are not billed. You cannot access the API Server. Applications deployed on the nodes are not affected.

> Note：
> If your account balance is below 0 **before 10:00, April 1, 2022 (UTC +8)**, the **TKE managed clusters created before 10:00, March 21, 2022 (UTC +8)** are isolated after 10:00, April 1, 2022 (UTC +8).

Processing for Overdue Payments

When the managed clusters are in **isolated** status, Tencent Cloud will take the following actions:

> Note：
> The following description of overdue payment is only for managed clusters. For Overdue Payment of CVM Instances, see the corresponding descriptions.

Time Since Isolation	Description
≤ 15 days	If your account is topped up to a positive balance, billing will resume and the clusters will automatically resume the running status.
	If your account is not topped up to a positive balance, the clusters will remain isolated.
＞ 15 days	If overdue payment of your account persists, the clusters will be deleted and cannot be recovered. All nodes remaining in the clusters will be removed. We will notify the creator and all collaborators of the Tencent Cloud account through email and SMS when the clusters under the account are deleted.

Regions and Availability Zones

Last updated：2021-11-15 15:01:05

Regions

Overview

A region is the physical location of an IDC. In Tencent Cloud, regions are fully isolated from each other, ensuring cross-region stability and fault tolerance. We recommend that you choose the region closest to your end users to minimize access latency and improve access speed.

You can view the following table or use the DescribeRegions API to get a complete region list.

Characteristics

- The networks of different regions are fully isolated. Tencent Cloud services in different regions **cannot communicate via a private network by default**.
- Tencent Cloud services across regions can communicate with each other through Public IPs over the Internet, while those in different VPCs can communicate with each other through Cloud Connect Network that is faster and steadier.
- Cloud Load Balancer currently supports intra-region traffic forwarding and being bound to a CVM in the same region by default. If you enable the cross-region binding feature, a CLB instance can be bound to CVM instances in another region.

Availability Zones

Overview

An availability zone (AZ) is a physical IDC of Tencent Cloud with independent power supply and network in the same region. It can ensure business stability, as failures (except for major disasters or power failures) in one AZ are isolated without affecting other AZs in the same region. By starting an instance in an independent availability zone, users can protect their applications from being affected by a single point of failure.
You can view the following table or use the DescribeZones API to get a complete availability zone list.

Characteristics

Tencent Cloud services in the same VPC are interconnected via the private network, which means they can communicate using private IPs, even if they are in different availability zones of the same region.

> Note：
> Private network interconnection refers to the interconnection of resources under the same account. Resources under different accounts are completely isolated on the private network.

China

Region	Availability Zone
South China (Guangzhou) ap-guangzhou	Guangzhou Zone 3 ap-guangzhou-3
	Guangzhou Zone 4 ap-guangzhou-4
	Guangzhou Zone 6 ap-guangzhou-6
East China (Shanghai) ap-shanghai	Shanghai Zone 2 ap-shanghai-2
	Shanghai Zone 3 ap-shanghai-3
	Shanghai Zone 4 ap-shanghai-4
	Shanghai Zone 5 ap-shanghai-5
East China (Nanjing) ap-nanjing	Nanjing Zone 1 ap-nanjing-1
	Nanjing Zone 2 ap-nanjing-2
North China (Beijing) ap-beijing	Beijing Zone 3 ap-beijing-3
	Beijing Zone 4 ap-beijing-4

	Beijing Zone 5 ap-beijing-5
	Beijing Zone 6 ap-beijing-6
	Beijing Zone 7 ap-beijing-7
Southwest China (Chengdu) ap-chengdu	Chengdu Zone 1 ap-chengdu-1
	Chengdu Zone 2 ap-chengdu-2
Hong Kong, Macao and Taiwan, China (Hong Kong) ap-hongkong	Hong Kong Zone 2 (Nodes in Hong Kong, China can cover Hong Kong/Macao/Taiwan regions) ap-hongkong-2

Other Countries and Regions

Region	Availability Zone
Southeast Asia (Singapore) ap-singapore	Singapore Zone 1 (Nodes in Singapore can cover Southeast Asia) ap-singapore-1
	Singapore Zone 2 (Nodes in Singapore can cover Southeast Asia) ap-singapore-2
Southeast Asia (Jakarta) ap-jakarta	Jakarta Zone 1 (Nodes in Jakarta can cover Southeast Asia) ap-jakarta-1
Northeast Asia (Seoul) ap-seoul	Seoul Zone 1 (Nodes in Seoul can cover Northeast Asia) ap-seoul-1
	Seoul Zone 2 (Nodes in Seoul can cover Northeast Asia) ap-seoul-2
Northeast Asia (Tokyo) ap-tokyo	Tokyo Zone 2 (Tokyo nodes can cover services in Northeast Asia) ap-tokyo-2
South Asia (Mumbai) ap-mumbai	Mumbai Zone 1 (Nodes in Mumbai can cover South Asia) ap-mumbai-1
	Mumbai Zone 2 (Nodes in Mumbai can cover South Asia)

	ap-mumbai-2
Southeast Asia (Bangkok) ap-bangkok	Bangkok Zone 1 (Nodes in Bangkok can cover Southeast Asia) ap-bangkok-1
North America (Toronto) na-toronto	Toronto Zone 1 (Nodes in Toronto can cover North America) na-toronto-1
Eastern US (Virginia) na-ashburn	Virginia Zone 1 (Nodes in Virginia can cover Eastern US) na-ashburn-1
	Virginia Zone 2 (Nodes in Virginia can cover Eastern US) na-ashburn-2
Europe (Frankfurt) eu-frankfurt	Frankfurt Zone 1 (Nodes in Frankfurt can cover Europe) eu-frankfurt-1
Europe (Moscow) eu-moscow	Moscow Zone 1 (Nodes in Moscow can cover Europe) eu-moscow-1

How to Select Regions and Availability Zones

When selecting a region and availability zone, take the following into consideration:

- Your location, the location of your users, and the region of the CVM instances.
  We recommend that you choose the region closest to your end users when purchasing CVM instances to minimize access latency and improve access speed.
- Other Tencent Cloud services you use.
  When you select other Tencent Cloud services, we recommend you try to locate them all in the same region and availability zone to allow them to communicate with each other through the private network, reducing access latency and increasing access speed.
- High availability and disaster recovery.
  Even if you have just one VPC, we still recommend that you deploy your businesses in different availability zones to prevent a single point of failure and enable cross-AZ disaster recovery.
- There may be network latency among different availability zones. We recommend that you assess your business requirements and find the optimal balance between high availability and low latency.
- If you need access to CVM instances in other countries or regions, we recommend you select a CVM in those other countries or regions. If you use a CVM instance in China to access servers in other countries and regions, you may encounter much higher network latency.

Resource Availability

The following table describes which Tencent Cloud resources are global, which are regional, and which are specific to availability zones.

Resource	Resource ID Format -8-Digit String of Numbers and Letters	Type	Description
User Account	No limit	Globally unique	Users can use the same account to access Tencent Cloud resources around the world.
SSH Keys	skey-xxxxxxxx	Global	Users can use an SSH key to bind a CVM in any region under the account.
CVM Instances	ins-xxxxxxxx	CVM instances are specific to an availability zone.	A CVM instance created in an availability zone is not available to other availability zones.
Custom Images	img-xxxxxxxx	Regional	Custom images created for the instance are available to all availability zones of the same region. Use **Copy Image** to copy a custom image if you need to use it in other regions.
EIPs	eip-xxxxxxxx	Regional	EIPs can only be associated with instances in the same region.
Security Groups	sg-xxxxxxxx	Regional	Security group can only be associated with instances in the same region. Tencent Cloud automatically creates three default security groups for users.
Cloud Block Storage	disk-xxxxxxxx	CVM instances are specific to an availability zone.	Users can only create a Cloud Block Storage disk in a specific AZ and mount it to instances in the same availability zone.
Snapshots	snap-xxxxxxxx	Regional	A snapshot created from a cloud disk can be used for other purposes (such as creating cloud disks) in this region.
Cloud Load Balancer	clb-xxxxxxxx	Regional	Cloud Load Balancer can be bound with CVMs in different availability zones of a single region for traffic forwarding.

VPC	vpc-xxxxxxxx	Regional	A VPC in one region can have resources created in different availability zones of the region.
Subnets	subnet-xxxxxxxx	CVM instances are specific to an availability zone.	Users cannot create subnets across availability zones.
Route Tables	rtb-xxxxxxxx	Regional	When creating a route table, users need to specify a VPC. Therefore, route tables are regional as well.

Related Operations

Migrating an instance to another availability zone

Once launched, an instance cannot be migrated. However, you can create a custom image of your CVM instance and use the image to launch or update an instance in a different availability zone.

1. Create a custom image from the current instance. For more information, see Creating Custom Images.
2. If the instance is on a VPC network environment and you want to retain its current private IP address after the migration, first delete the subnet in the current availability zone and then create a subnet in the new availability zone with the same IP address range. Note that a subnet can be deleted only when it contains no available instances. Therefore, all the instances in the current subnet should be migrated to the new subnet.
3. Create a new instance in the new availability zone by using the custom image you have just created. You can choose the same type and configuration as the original instance, or choose new settings. For more information, see Creating Instances via CVM Purchase Page.
4. If an elastic IP is associated with the original instance, dissociate it from the old instance and associate it with the new instance. For more information, see Elastic IP (EIP).
5. (Optional) If the original instance is pay-as-you-go, you can choose to terminate it. For more information, see Terminating Instances.

Copying images to other regions

Operations such as launching and viewing instances are region-specific. If the image of the instance that you need to launch does not exist in the region, copy the image to the desired region. For more information, see Copying Images.

Quotas and Limits

Last updated：2022-06-10 19:32:51

While using TKE services, you need to consider the service quota applied to TKE, CVM, and managed clusters.

TKE Quota Limit

The default TKE quota for each user is as follows. If you want to increase the quota, submit a ticket for application.

> Note：
> From October 21, 2019, the maximum node quota for a cluster has been adjusted to at least 5,000.

Item	Default Value	Where to Check	Quota Increase Allowed or Not
Clusters in a region	5	Bottom-right section of the TKE overview page	Yes
Nodes in a cluster	5,000		
Image namespaces in a region	10		
Image repositories in a region	500		
Tags of an image	100		

CVM Quota Limit

CVM instances generated by TKE are subject to purchase limits. For more information, see Purchase Limits. If you need more quotas than the default, submit a ticket for application.

Item	Default Value	Where to Check	Quota Increase Allowed or Not
Pay-as-you-go CVM instances in an AZ	30 or 60	CVM Instances page - Resources in each region	Yes

Cluster Configuration Limit

Note：
Cluster configuration limits the cluster size and cannot be modified currently.

Item	IP Address Range	Affected Scope	Where to Check	Modification Allowed or Not
VPC network - Subnet	Custom	Number of nodes that can be added to the subnet	VPC subnet list page for the cluster - Number of available IP addresses	• No • You can use a new subnet
Container CIDR block	Custom	• Maximum number of nodes per cluster • Maximum number of services per cluster • Maximum number of Pods per node	Basic information page for the cluster - Container CIDR block	No

K8s Resource Quota Description

Note：
The following quotas are automatically applied from April 30, 2022 (UTC +8) and cannot be adjusted. **You can increase the resource quota by upgrading the cluster model**.
To adjust your quota, submit a ticket for application.

Run the following command to check the quota:

```
kubectl get resourcequota tke-default-quota -o yaml
```

To check the `tke-default-quota` object of a specified namespace, add `--namespace` to specify the namespace.

Note：

- Other K8s resource limit means that the number of all K8s resources in the cluster except Pod, Node, and ConfigMap **cannot** exceed this value. For example, for an L100 cluster, the number of ClusterRole, Service, Endpoint, and other K8s resources **cannot** exceed 10,000.

- **CRD** refers to **the sum of all CRDs** in the cluster. If the number of some CRDs increases, the number of other CRDs will decrease.

Cluster Model	Pods	ConfigMap	CRDs/Other K8s Resources
L5	600	256	1,250
L20	1,500	512	2,500
L50	3,000	1,024	5,000
L100	6,000	2,048	10,000
L200	15,000	4,096	20,000
L500	30,000	6,144	50,000
L1000	90,000	8,192	100,000
L3000	150,000	10,240	150,000
L5000	200,000	20,480	200,000

Namespace quota

By default, **each namespace has the same margin (margin = quota for the current cluster level - amount already used by the entire cluster). If you create resources in a namespace, the margin will decrease, and the amount available in other namespaces will decrease accordingly after a certain period of time.**

If you want to customize the allocation ratio, you can create a `tke-quota-config` ConfigMap under `kube-system` to specify the **margin** allocation ratio for each namespace.

The following example sets the **margin** allocation ratio to `50%` for the `default` namespace, `40%` for the `kube-system` namespace, and `10%` for the rest of the namespaces. **If the sum of the set percentages exceeds `100%` , TKE considers the ratio invalid and will use the default allocation policy**.

```
apiVersion: v1
data:
default: "50"
kube-system: "40"
```

```
kind: ConfigMap
metadata:
name: tke-quota-config
namespace: kube-system
```

Container Node Disk Settings

Last updated：2019-09-02 16:35:15

Description

When creating or scaling a TKE cluster, you can set the type and size of the system disks and data disks of the container nodes to meet your actual business needs.

Suggestions

1. The directory of the container is stored in the system disk. We recommend creating a system disk with a capacity of 50 GB.
2. If you have specific requirements for the system disk, you can move the Docker's directory to the data disk when initializing the cluster.

Notes on the Public IP of a TKE Node

Last updated：2022-04-25 12:23:18

If you don't want to avoid exposing your company's IP while accessing the public network, you can use Tencent Cloud NAT Gateway. This document describes how to access the public network via an NAT gateway.

Public IP

When a cluster is created, public IPs are assigned to the nodes in the cluster by default. With these public IPs, you can:

- Log in to the nodes in the cluster.
- Access services on the public network.

Public Network Bandwidth

When a service is created on the public network, the public network CLB uses the bandwidth and traffic of the nodes. If the public network service is required, the nodes need to have public network bandwidth. You can choose not to purchase public network bandwidth if it is not needed.

NAT Gateway

The CVM instance is not bound to an EIP, and all the traffic accessing the internet is forwarded via an NAT gateway. In this way, the traffic accessing the internet of the instance is forwarded to the NAT gateway over the private network. This means that the traffic is not subject to the upper limit of public network bandwidth specified when you purchase the instance, and the traffic generated from the NAT gateway does not occupy the public network bandwidth egress of the instance. To access the internet via an NAT gateway, follow the steps below:

Step 1. Create an NAT gateway

1. Log in to the VPC Console and click **NAT Gateway** in the left sidebar.
2. On the NAT gateway management page, click **Create**.
3. In the **Create an NAT Gateway** window that pops up, enter the following parameters.

- Gateway Name: Custom.
- Network: Select the VPC of the NAT gateway service;

- Gateway Type: Select based on actual needs. The type of the gateway can be changed after it is created.
- Outbound Bandwidth Cap: Set based on actual needs.
- Elastic IP: Assign an EIP to the NAT gateway. You can choose an existing EIP or purchase a new one.

4. Click **Create** to complete the creation of the NAT gateway.

> The rental fee of 1 hour will be frozen during the creation of the NAT gateway.

Step 2. Configure the route table associated with the subnet

> After the NAT gateway is created, you need to configure the routing rules on the route table page in the VPC Console to redirect the subnet traffic to the NAT gateway.

1. Click **Route Table** in the left sidebar.
2. In the route table list, click the route table ID/name associated with the subnet that needs to access the internet.
3. In the "Routing Policy" section, click **+ New routing policies**.
4. In the **Add routing** page, enter the **Destination**, select **NAT gateway** for **Next Hop Type**, and select the ID of the created NAT gateway for **Next Hop**.
5. Click **OK**.
   Now, the traffic generated when the CVM instance in the subnet associated with the route table accesses the internet will be directed to the NAT gateway.

Other Solutions

Solution 1. Use an EIP

The CVM instance is only bound with an EIP but does not use an NAT gateway. With this solution, all the traffic of the instance accessing the internet goes out through the EIP and is subject to the upper limit of public network bandwidth specified when you purchase the instance. The fees for accessing the internet are charged based on the billing method of the instance's network.
For more information, see Elastic Public IP.

Solution 2. Use both an NAT gateway and an EIP

If both an NAT gateway and an EIP are used, all the traffic of the CVM instance accessing the internet is forwarded to the NAT gateway over the private network, and the response packets are returned to the instance through the NAT

gateway. This means that the traffic is not subject to the upper limit of public network bandwidth specified when you purchase the instance, and the traffic generated by the NAT gateway does not occupy the public network bandwidth egress of the instance. If the traffic from the internet proactively accesses the EIP of the instance, the response packets of the instance are all returned through the EIP. In this case, the resulting outbound public network traffic is subject to the upper limit of public network bandwidth specified when you purchase the instance. The fees for accessing the public network are charged based on the billing method of the instance's network.

> If the bandwidth package (BWP) feature is activated in your account, fees of the outbound traffic generated by the NAT gateway will be deducted from the BWP (which means the network traffic will not be repeatedly billed at 0.12 USD/GB). It is recommended that you limit the outbound bandwidth of the NAT gateway so as to avoid high BWP fees due to excessive outbound bandwidth.

TKE Security Group Settings

Last updated：2022-03-23 18:17:29

Security is a matter of utmost importance. Tencent Cloud considers security as a top priority in product design and requires all its products to be fully isolated and provides multiple layers of security protection with its basic network. TKE is a typical example. It adopts VPC as the underlying network of container services. This document describes the best practice of security group usage in TKE to help you select the most appropriate security group policy.

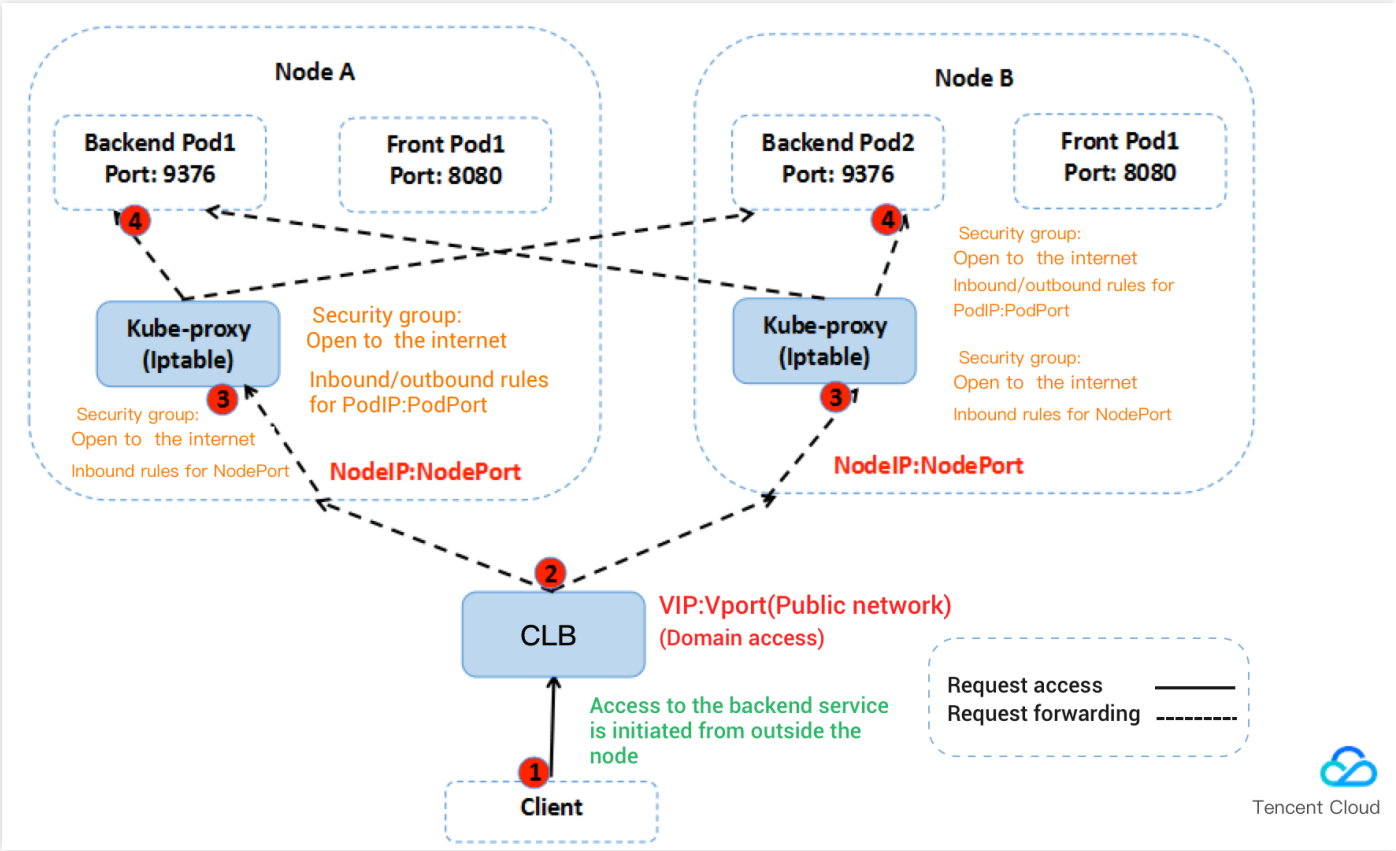
Security Groups

A security group is a virtual firewall capable of filtering stateful packets. As an important network security isolation means provided by Tencent Cloud, it can be used to configure network access control for one or more CVM instances. For more information, see Security Group.

How to Select a Security Group for TKE

- In a container cluster, service pods are distributed on different nodes. We recommend that you bind all CVM instances in one cluster to the same security group and do not add non-clustered CVMs to a security group for a cluster.
- A security group only grants the minimum permission externally.
- You must enable the following rules for using TKE:
- Open the container pod network and the cluster node network to the Internet.
  When a node receives a service access request, the node forwards the request to a service pod according to the iptables rule configured by the kube-proxy module. If the service pod is on another node, cross-node access occurs. For example, the destination IP addresses of the access request include the IP address of the service pod, IP addresses of other nodes in the cluster, and the IP address of the cluster's cbr0 bridge on the node. In this case, the container pod network and the cluster node network on the peer node must be open to the Internet.
- If clusters in the same VPC need to communicate with each other, you must open the container networks and node networks of the corresponding clusters to the Internet.
- Open port 22 to the Internet if SSH login is required.
- Open ports 30000 to 32768 on nodes to the Internet.
  In the access path, you must use a load balancer to forward data packets to NodeIP:NodePort of the container cluster. NodeIP is the CVM instance IP of any node in the cluster. NodePort is assigned by the container cluster by default when the service is created. NodePort ranges from 30000 to 32768.

The following figure uses service access from the public network as an example.



Default Security Group Rules for TKE

Default security group rules for node

Some ports must be opened to the Internet to ensure normal communication between cluster nodes. To avoid cluster creation failures due to binding to invalid security groups, TKE provides default security group rules, as described in the following table.

Note：
If the current default security group cannot meet your service requirements and you have created a cluster bound to this security group, you can view and modify the security group rules for the cluster. For more information, please see Managing Security Group Rules.

Inbound rules

Protocol	Port Number	Source IP Address	Rule	Description
All	All	CIDR of the container network	Allow	Enable the communication between pods in the container network.
All	All	CIDR of the cluster network	Allow	Enable the communication between nodes in the cluster network.
TCP	22	0.0.0.0/0	Allow	Open the SSH login port to the Internet.
tcp	30000 - 32768	0.0.0.0/0	Allow	Open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).
udp	30000 - 32768	0.0.0.0/0	Allow	Open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).
ICMP	-	0.0.0.0/0	Allow	Enable the support for Internet Control Message Protocol (ICMP) and ping operations.

Outbound rules

Protocol	Port Number	Source IP Address	Rule
All	All	0.0.0.0/0	Allow

Note：

- To customize outbound rules, you need to open the node IP range and container IP range.
- If you configure this rule for container nodes, the services in the cluster can be accessed using different access methods.
- For more information on how to access a service in a cluster, please see "Service Access" in Overview.

Default security group rules for master node in self-deployed cluster

When you create a self-deployed cluster, the default TKE security group will be bound to the master node by default to reduce the risks where the master node cannot communicate with other nodes normally or Services cannot be accessed normally. The configuration rules of default security group are as detailed below:

Note：
The security group creation permission is inherited from the TKE service role. For more information, see Description of Role Permissions Related to Service Authorization.

Inbound rules

Protocol	Port	IP Range	Policy	Remarks
ICMP	All	0.0.0.0/0	Supported	Ping operations are supported.
TCP	30000-32768	Cluster network CIDR	Supported	It is used to open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).
UDP	30000-32768	Cluster network CIDR	Supported	It is used to open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).
TCP	60001, 60002, 10250, 2380, 2379, 53, 17443, 50055, 443, 61678	Cluster network CIDR	Supported	It is used to open API Server communication to the Internet.
TCP	60001, 60002, 10250, 2380, 2379, 53, 17443	Container network CIDR	Supported	It is used to open API Server communication to the internet.
TCP	30000-32768	Container network CIDR	Supported	It is used to open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).

Protocol	Port	IP Range	Policy	Remarks
UDP	30000-32768	Container network CIDR	Supported	It is used to open NodePort to the Internet (Services in LoadBalancer type need to be forwarded through NodePort).
UDP	53	Container network CIDR	Supported	It is used to open CoreDNS communication to the internet.
UDP	53	Cluster network CIDR	Supported	It is used to open CoreDNS communication to the internet.

Outbound rules

Protocol	Port Number	Source IP Address	Rule
All	All	0.0.0.0/0	Allow

Project of New Resources

Last updated：2022-08-26 11:18:00

Overview

To conduct financial accounting by projects, please take the following into the consideration:

1. Clusters are not project-specific, but CVMs, load balancers and other resources in a cluster are project-specific.
2. Project of New-added Resource: Only resources newly added to the cluster are allocated to the project.

Notes

1. We recommend you allocate all the resources in a cluster to the same project.
2. If you need to distribute the CVMs in a cluster to different projects, go to the CVM Console to migrate projects. For more information, see Adjusting Project Configuration.
3. If CVMs belong to different projects, they belong to different `security group instances`. Try to configure the same `security group rules` for the CVMs in the same cluster. For more information, see Changing Security Group.

Billing Description

Last updated：2022-04-06 10:29:26

TKE

Cluster management fees and cloud resource fees are charged based on the actual usage of managed clusters with different specifications. For the specific prices of TKE, see TKE Billing Overview.

EKS

EKS is billed by the CPU, GPU, and memory resources requested by workloads and the running time of workloads. It is suitable for different needs in different use cases. For the specific prices of EKS, see Billing Overview.

Purchasing EKS
Regions and Availability Zones

Last updated：2021-08-09 10:53:53

Region

Overview

A region is the physical location of an IDC. In Tencent Cloud, regions are fully isolated from each other, ensuring cross-region stability and fault tolerance. We recommend that you choose the region closest to your end users to minimize access latency and improve access speed.

The following table describes the regions, availability zones, and resource types that are currently supported by EKS.

Characteristics

- The networks of different regions are fully isolated. Tencent Cloud services in different regions **cannot communicate via a private network by default**.
- Tencent Cloud services across regions can communicate with each other through public IPs over the Internet, while those in different VPCs can communicate with each other through CCN, which is faster and more stable.
- Cloud Load Balancer (CLB) currently supports intra-region traffic forwarding by default. If you enable the cross-region binding feature, cross-region binding of CLB and CVM instances is supported.

Availability Zone

Overview

An availability zone (AZ) is a physical IDC of Tencent Cloud with independent power supply and network in the same region. Business stability can be ensured because failures (except for major disasters or power failures) in one AZ do not affect other AZs in the same region. By starting an instance in an independent availability zone, users can protect their applications from a single point of failure.
You can view the following table or use the DescribeZones API to get a complete AZ list.

Characteristics

Tencent Cloud services in the same VPC are interconnected via the private network, which means they can communicate using private IPs, even if they are in different AZs in the same region.

Note：
Private network interconnection refers to the interconnection of resources under the same account. Resources under different accounts are completely isolated on the private network.

China

Region	Availability Zone
South China (Guangzhou) ap-guangzhou	Guangzhou Zone 3 ap-guangzhou-3
	Guangzhou Zone 4 ap-guangzhou-4
	Guangzhou Zone 6 ap-guangzhou-6
East China (Shanghai) ap-shanghai	Shanghai Zone 2 ap-shanghai-2
	Shanghai Zone 3 ap-shanghai-3
	Shanghai Zone 4 ap-shanghai-4
	Shanghai Zone 5 ap-shanghai-5
East China (Nanjing) ap-nanjing	Nanjing Zone 1 ap-nanjing-1
	Nanjing Zone 2 ap-nanjing-2
North China (Beijing) ap-beijing	Beijing Zone 3 ap-beijing-3
	Beijing Zone 4 ap-beijing-4
	Beijing Zone 5 ap-beijing-5

	Beijing Zone 6 ap-beijing-6
	Beijing Zone 7 ap-beijing-7
Southwest China (Chengdu) ap-chengdu	Chengdu Zone 1 ap-chengdu-1
	Chengdu Zone 2 ap-chengdu-2
Hong Kong, Macao and Taiwan, China (Hong Kong) ap-hongkong	Hong Kong Zone 2 (Nodes in Hong Kong, China can cover services in Hong Kong/Macao/Taiwan regions) ap-hongkong-2

Other Countries and Regions

Region	Availability Zone
Southeast Asia (Singapore) ap-singapore	Singapore Zone 1 (Singapore nodes can cover services in Southeast Asia) ap-singapore-1
	Singapore Zone 2 (Singapore nodes can cover services in Southeast Asia) ap-singapore-2
Southeast Asia (Jakarta) ap-jakarta	Jakarta Zone 1 (Jakarta nodes can cover services in Southeast Asia) ap-jakarta-1
Northeast Asia (Seoul) ap-seoul	Seoul Zone 1 (Seoul nodes can cover services in Northeast Asia) ap-seoul-1
	Seoul Zone 2 (Seoul nodes can cover services in Northeast Asia) ap-seoul-2
Northeast Asia (Tokyo) ap-tokyo	Tokyo Zone 2 (Tokyo nodes can cover services in Northeast Asia) ap-tokyo-2
South Asia (Mumbai) ap-mumbai	Mumbai Zone 1 (Mumbai nodes can cover services in South Asia) ap-mumbai-1
	Mumbai Zone 2 (Mumbai nodes can cover services in South Asia) ap-mumbai-2
Southeast Asia (Bangkok)	Bangkok Zone 1 (Bangkok nodes can cover services in Southeast Asia)

ap-bangkok	ap-bangkok-1
North America (Toronto) na-toronto	Toronto Zone 1 (Toronto nodes can cover services in North America) na-toronto-1
Eastern US (Virginia) na-ashburn	Virginia Zone 1 (Virginia nodes can cover services in Eastern US) na-ashburn-1
	Virginia Zone 2 (Virginia nodes can cover services in Eastern US) na-ashburn-2
Europe (Frankfurt) eu-frankfurt	Frankfurt Zone 1 (Frankfurt nodes can cover services in Europe) eu-frankfurt-1
Europe (Moscow) eu-moscow	Moscow Zone 1 (Moscow nodes can cover services in Europe) eu-moscow-1

Billing Overview

Last updated：2020-04-28 18:47:37

Billing Mode

Elastic Kubernetes Service (EKS) is a pay-as-you-go service. The fees are calculated based on the configured amount of resources and the actual period of using them.

Billing Method

EKS calculates fees based on the specifications of the CPU, GPU, and memory for a workload and the running time of the workload. For more information, see Product Pricing.

Other Fees

If you use EKS with other paid products such as CLB, CBS, and CFS, these products are billed according to their own billing rules. For more information, see the purchase guide for the specific product.

Product Pricing

Last updated：2021-01-12 19:51:14

Elastic Kubernetes Service (EKS) bills a pod by multiplying the resource specification of the pod by the **Unit price of the resource** and **Running time**. The following table shows the unit prices of the CPU and memory resources.

Billing Item	Price (per second)	Price (per hour)
CPU	0.000004976 USD per core per second	0.0179 USD per core per hour
Memory	0.000002073 USD per GiB per second	0.0075 USD per GiB per hour

Star Lake AMD

Based on Tencent Cloud's self-developed Star Lake servers, EKS provides reliable, secure, and stable high performance. For more information, see CVM Standard SA2 Introduction.

Billing Item	Price (per second)	Price (per hour)
CPU	0.00000219 USD per core per second	0.0079 USD per core per hour
Memory	0.00000127 USD per GiB per second	0.0046 USD per GiB per hour

Running Time

The running time is the time that elapses from when a pod fetches the first container image until the pod stops running. The pod is billed for the resources used during this period, which is measured in seconds.

Billing Examples

Sample 1

Assume that the specification of pods managed by a Deployment is 2 cores and 4 GB memory, and the number of replicas is fixed at 2. If the period from the time when the Deployment is launched to the time when the Deployment is terminated is 5 minutes, the Deployment is billed for the resources used during the running time of 300 seconds (5 minutes × 60 seconds).

In this case, the running fees of the Deployment = 2 × (2 × 0.000004976 + 4 × 0.000002073) × 300 = 0.019495 USD.

Sample 2

Assume that a CronJob needs to launch 10 pods with 4 cores and 8 GB memory each time and terminate the pods 10 minutes later. If the CronJob executes the job twice a day and this task is managed by EKS, the task is billed as follows:

Daily task fees = 2 × 10 × (4 × 0.000004976 + 8 × 0.000002073) × 600 = 0.437856 USD.

Purchase Limits

Last updated：2022-06-14 15:39:17

Use Restrictions

Before using EKS, you need to sign up for a Tencent Cloud account and complete identity verification.

Supported Regions

Please see Regions and Availability Zones for regions supported by EKS, and see Resource Specifications for information about resource specifications.

Resource Quotas

Resource	Limit	Description
Clusters in one region	5	Includes clusters that are being created and running.
Pods in one cluster	100	Includes all namespaces, workloads, and stateless and stateful pods.
Pod replicas for one workload	100	Includes all stateless and stateful pods in the workload.
The largest container instance size for the same region	500	Includes all stateless and stateful container instances.

If the number of required resources exceeds the quota limit shown in the preceding table, you can submit a ticket to apply for a higher quota. Tencent Cloud will assess your actual needs and increase your quota as appropriate.

Applying for a higher quota

1. Log in to the ticket system console. On the **Submit ticket** page, select **Other Problems** and then click **Create Now** to go to the page for creating a ticket.
2. In the **Problem description** field, enter a description such as "I want to apply for a higher quota for the elastic cluster." Then, enter the region where the elastic cluster is located and your desired quota. Finally, enter your mobile number and other information as instructed.
3. Click **ticket system console**.

Resource Specifications

Last updated：2021-07-14 16:10:04

Overview

Elastic Kubernetes Service (EKS) frees you from managing cluster nodes. However, to properly allocate resources and accurately calculate fees, you need to specify resource specifications for Pods when deploying a workload. Tencent Cloud allocates computing resources to the workload and calculates the corresponding fees based on the specified specifications.

When you use the Kubernetes API or Kubectl to create a workload for EKS, you can use annotations to specify resource specifications. If annotations are not used, EKS will calculate the specifications based on the container parameters set for the workload, such as Request and Limit. For more information, see Specifying Resource Specifications.

Note：

- The resource specifications indicate the maximum amount of resources available for containers in a Pod.
- The following tables list the supported CPU and GPU specifications. Ensure that allocated resources do not exceed the supported specifications.
- The total amount of resources specified by Request for all the containers in a Pod cannot exceed the maximum pod specification.
- The amount of resources specified by Limit for any container in a Pod cannot exceed the maximum Pod specification.

CPU Specifications

The following table lists CPU specifications that EKS provides for Pods in all regions where CPU resources are supported. EKS also provides a set of CPU options. Different CPU sizes correspond to different memory ranges. Select the CPU specification as needed when creating a workload.

Intel

CPU (Cores)	Memory Range (GiB)	Granularity of Memory Range (GiB)
0.25	0.5, 1, 2	-

CPU (Cores)	Memory Range (GiB)	Granularity of Memory Range (GiB)
0.5	1, 2, 3, 4	-
1	1 - 8	1
2	4 - 16	1
4	8 - 32	1
8	16 - 32	1
12	24 - 48	1
16	32 - 64	1

Star Lake AMD

The AMD processor provides high performance with high reliability, security, and stability based on the Star Lake servers developed by Tencent Cloud. For more information, see CVM Standard SA2 Introduction.

CPU (Cores)	Memory Range (GiB)	Granularity of Memory Range (GiB)
1	1 - 4	1
2	2 - 8	1
4	4 - 16	1
8	8 - 32	1
16	32 - 64	1

GPU Specifications

The following table lists the GPU specifications that EKS provides for Pods. Different GPU card models and sizes map to different CPU and memory options. Select the GPU specification as needed when creating a workload.

Note：

If you create, manage, and use GPU workloads using a YAML file, see Annotation.

GPU Model	GPU (Cards)	CPU (Cores)	Memory (GiB)
Tesla V100-NVLINK-32G	1	8	40
Tesla V100-NVLINK-32G	2	18	80
Tesla V100-NVLINK-32G	4	36	160
Tesla V100-NVLINK-32G	8	72	320
1/4 NVIDIA T4	1	4	20
1/2 NVIDIA T4	1	10	40
NVIDIA T4	1	8	32
NVIDIA T4	1	20	80
NVIDIA T4	1	32	128
NVIDIA T4	2	40	160
NVIDIA T4	4	80	320

Spot Mode

Last updated：2022-05-11 17:06:24

Spot Mode Overview

You can purchase resources with low costs in the spot mode of Elastic Kubernetes Service. In some scenarios, you can pay a price lower than that of the pay-as-you-go instances to run Pods until they are interrupted and repossessed by Tencent Cloud, greatly reducing the costs.

When using the spot mode, you can deploy workloads in containers just like you are using other pay-as-you-go resources.

Spot Mode Policy

Price policy

A **fixed discount (80% off)** is used by the spot mode of Elastic Kubernetes Service. That means the Pod resources with all specifications are sold at the fixed discount (80% off) of the original prices defined in Product Pricing.

> Note：
> The discount only applies to the billable items (CPU, memory and GPU) of resources in the Pods. It is not applicable to resources such as network bandwidth, network traffic and persistent storage.

Interruption and repossessing mechanism

In the spot mode, when the resources in Tencent Cloud computing resource pool are insufficient, the assigned Pods are interrupted and repossessed randomly. Note that the cache data in the Pods will not be retained.

The following events may occur when the Pods are interrupted and repossessed:

```
EVENT REASON : “SpotPodInterruption”
EVENT MESSAGE : “Spot pod was interrupted, it will be killed and re-created”
```

Use Cases

Short-time emergent and periodic tasks

The spot mode is suitable for workloads that do not run for a long time. For example, video transcoding, video rendering, service stress testing, batch computing and crawlers.

Divisible computing tasks

The spot mode is suitable for systems that can divide long-term tasks into fine-grained tasks for computing based on the objects, such as EMR and other big data suite.

Stateless computing tasks or tasks supporting checkpoint restart

- It is suitable for workloads that put the intermediate computing results on the persistent storage and continue the computing after being restarted when Pods are repossessed.
- It is suitable for stateless workloads that support automatic load balancing and service discovery, and workloads that can be restarted when the Pods are repossessed.

Enabling the Spot Mode

You can enable the spot mode for a workload by adding the following Pod template annotation in the workload YAML.

```
eks.tke.cloud.tencent.com/spot-pod: "true"
```

For more information, see Annotation.

Purchase Channels

Last updated：2021-04-19 17:15:54

Product Purchase

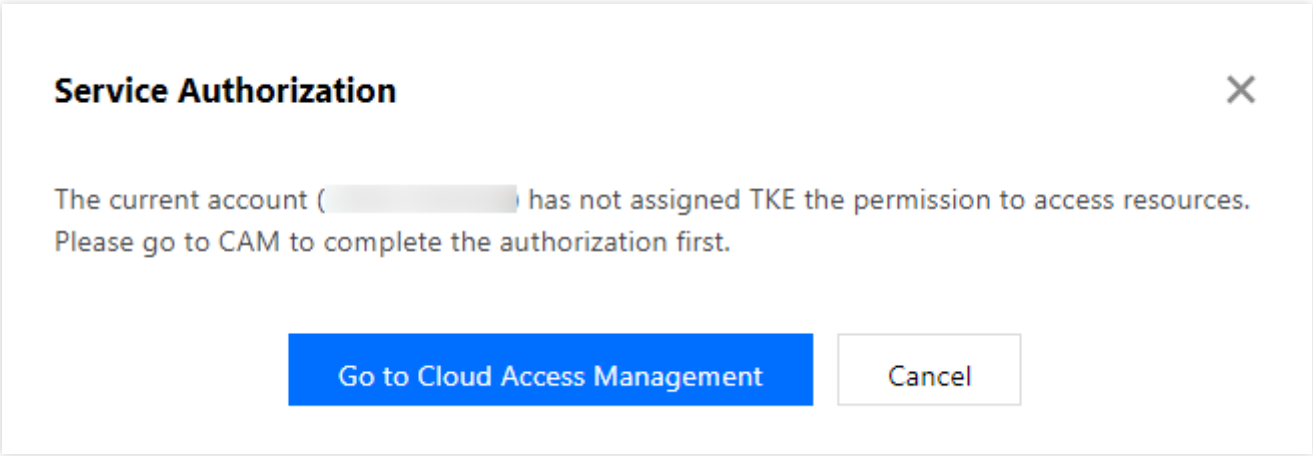
Log in to the Tencent Cloud, choose **Products** > **Container** > **Tencent Kubernetes Engine**, and click **Get Started** to go to the TKE console. When logging in to this page for the first time, you need to perform the following steps to complete service authorization before you can purchase TKE products.

Service Authorization

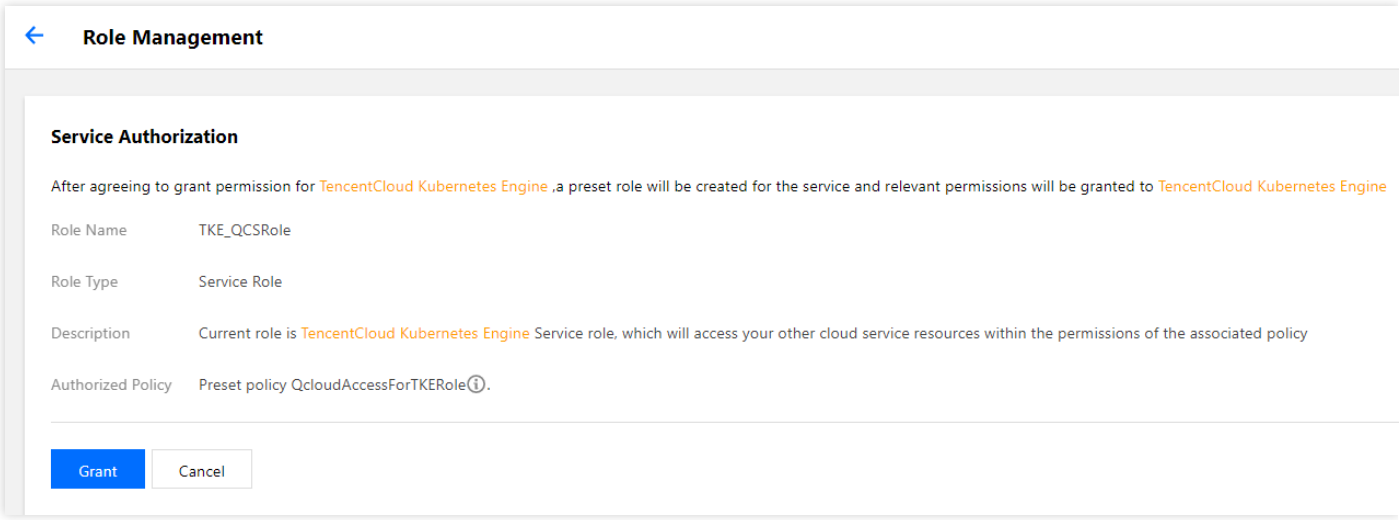
Note：

Ignore the following procedure if you have completed service authorization.

1. View information in the displayed **Service Authorization** dialog box, and click **Go to Cloud Access Management**, as shown in the following figure.



2. On the "Role Management" page, read information related to the role, as shown in the following figure.



3. Click **Grant** to grant authorization. Now you can go to the TKE console to create clusters and purchase related products.