Alibaba Cloud Aliyun Container for Kubernetes

User Guide

Issue: 20190525

MORE THAN JUST CLOUD | C-J Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

- 1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
- 2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
- 3. The content of this document may be changed due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
- 4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

- 5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified , reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates . The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
- 6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1:	Style co	nventions
-----------	----------	-----------

Style	Description	Example
•	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning informatio n, supplementary instructions, and other content that the user must understand.	• Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
Courier font	It is used for commands.	Run the cd / d C :/ windows command to enter the Windows system folder.
Italics	It is used for parameters and variables.	bae log list instanceid Instance_ID
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	ipconfig [-all -t]

Style	Description	Example
{} or {a b}	It indicates that it is a required value, and only one item can be selected.	<pre>swich {stand slave}</pre>

Contents

Generic conventions.I1 Kubernetes cluster.11.1 Security bulletins.11.1.1 Vulnerability fix: CVE - 2019 - 5736 in runc.11.1.2 Vulnerability fix: CVE - 2018 - 18264 for Kubernetesdashboard.31.1.3 Vulnerability fix: CVE - 2018 - 18264 for Kubernetesdashboard.61.2.1 Overview.61.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes.71.3 Authorization management.111.3.1 Role authorization.1.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.291.4.1 View cluster overview.291.4.2 Create a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster311.4.8 Use a ServiceAccount token to access a managed Kubernetescluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetescluster.651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.701.4.11 Create a managed Kubernetes cluster.711.4.11 Create a managed Kubernetes cluster.72
1 Kubernetes cluster 1 1.1 Security bulletins 1 1.1.1 Vulnerability fix: CVE - 2019 - 5736 in runc 1 1.1.2 Vulnerability fix: CVE - 2018 - 18264 for Kubernetes dashboard 3 1.1.3 Vulnerability fix: CVE - 2018 - 18264 for Kubernetes dashboard 3 1.1.3 Vulnerability fix: CVE - 2018 - 1002105 5 1.2 Introduction 6 1.2.1 Overview 6 1.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes. 7 1.3 Authorization management. 11 1.3.1 Role authorization 11 1.3.2 Use the Container Service console as a RAM user 15 1.3.3 Create custom authorization policies 17 1.3.4 Grant Kubernetes permissions to a RAM user 20 1.4 Cluster management 29 1.4.1 View cluster overview 29 1.4.2 Create a Kubernetes cluster 31 1.4.3 Configure a Kubernetes Cluster to support GPU scheduling 41 1.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node 47 1.4.5 Create a multi-zone Kubernetes cluster by using kubectl 62 1.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster 63 1.4.8 Use a ServiceAcco
1.1 Security bulletins. 1 1.1.1 Vulnerability fix: CVE - 2019 - 5736 in runc. 1 1.1.2 Vulnerability fix: CVE - 2018 - 18264 for Kubernetes 3 1.1.3 Vulnerability fix: CVE - 2018 - 1002105 3 1.1.3 Vulnerability fix: CVE - 2018 - 1002105 6 1.2.1 Overview. 6 1.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes. 7 1.3 Authorization management. 11 1.3.1 Role authorization. 11 1.3.2 Use the Container Service console as a RAM user. 15 1.3.3 Create custom authorization policies. 17 1.3.4 Grant Kubernetes permissions to a RAM user. 20 1.4 Cluster management. 29 1.4.1 View cluster overview. 29 1.4.2 Create a Kubernetes cluster. 31 1.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling 41 1.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node
11.1.1 Vulnerability fix: <i>CVE</i> - 2019 - 5736 in runc
1.1.1Vulnerabilityfix:CVE201818264forKubernetesdashboard
dashboard
1.1.3 Vulnerability fix:CVE - 2018 - 100210551.2 Introduction61.2.1 Overview61.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes71.3 Authorization management111.3.1 Role authorization111.3.2 Use the Container Service console as a RAM user151.3.3 Create custom authorization policies171.3.4 Grant Kubernetes permissions to a RAM user201.4 Cluster management291.4.1 View cluster overview291.4.2 Create a Kubernetes cluster311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node471.4.5 Create a multi-zone Kubernetes cluster531.4.6 Connect to a Kubernetes cluster by using kubect!621.4.7 Use kubect! on Cloud Shell to manage a Kubernetes cluster631.4.8 Use a ServiceAccount token to access a managed Kubernetes621.4.9 Access a Kubernetes cluster by using SSH681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster701.4.11 Create a managed Kubernetes cluster72
1.2 Introduction
1.2.1 Overview
1.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes.71.3 Authorization management.111.3.1 Role authorization111.3.2 Use the Container Service console as a RAM user.151.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling.411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.72
1.3 Authorization management.111.3.1 Role authorization111.3.2 Use the Container Service console as a RAM user.151.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.72
1.3.1 Role authorization.111.3.2 Use the Container Service console as a RAM user.151.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.72
1.3.2 Use the Container Service console as a RAM user.151.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.701.4.11 Create a managed Kubernetes cluster.72
1.3.3 Create custom authorization policies.171.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.701.4.11 Create a managed Kubernetes cluster.72
1.3.4 Grant Kubernetes permissions to a RAM user.201.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.701.4.11 Create a managed Kubernetes cluster.72
1.4 Cluster management.291.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling411.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node.471.4.5 Create a multi-zone Kubernetes cluster.531.4.6 Connect to a Kubernetes cluster by using kubectl.621.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster.631.4.8 Use a ServiceAccount token to access a managed Kubernetes651.4.9 Access a Kubernetes cluster by using SSH.681.4.10 Set the SSH key pair logon mode for a Kubernetes cluster.701.4.11 Create a managed Kubernetes cluster.72
1.4.1 View cluster overview.291.4.2 Create a Kubernetes cluster.311.4.2 Create a Kubernetes GPU cluster to support GPU scheduling
1.4.2 Create a Kubernetes cluster
 1.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling 41 1.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node 47 1.4.5 Create a multi-zone Kubernetes cluster
 1.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node
 1.4.5 Create a multi-zone Kubernetes cluster
 1.4.6 Connect to a Kubernetes cluster by using kubectl
 1.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster
 1.4.8 Use a ServiceAccount token to access a managed Kubernetes cluster
cluster
1.4.9 Access a Kubernetes cluster by using SSH
1.4.10 Set the SSH key pair logon mode for a Kubernetes cluster70 1.4.11 Create a managed Kubernetes cluster
1.4.11 Create a managed Kubernetes cluster
1 4 10 Unaverale a Vich annatae alexatar 70
1.4.12 Upgrade a Kubernetes cluster
1.4.13 Upgrade a system component
ovpiro
1 4 15 Scale a Kubernetes cluster 86
1 4 16 Autoscale a Kubernetes cluster 87
1 4 17 Delete a Kubernetes cluster 98
1.4.18 Ungrade the Heapster components to the metrics-server
component
1.4.19 Create a Kubernetes cluster that supports Windows
1.5 Node management 116

1.5.1 Add an existing ECS instance	116
1.5.2 View node list	
1.5.3 Node monitoring	121
1.5.4 Manage node labels	122
1.5.5 Set node scheduling	
1.5.6 Remove a node	
1.5.7 Use Alibaba Cloud Kubernetes GPU node labels for scl	neduling 128
1.5.8 View resource request and limit on nodes	134
1.5.9 Mount a disk to a Kubernetes cluster node	135
1.5.10 Mount a disk to the Docker data directory	140
1.6 Namespace management	
1.6.1 Create a namespace	
1.6.2 Set resource quotas and limits for a namespace	
1.6.3 Edit a namespace	
1.6.4 Delete a namespace	153
1.7 Service catalog management	
1.7.1 Overview	154
1.7.2 Enable service catalog function	155
1.8 Application management	157
1.8.1 Create a deployment application by using an image	
1.8.2 Create a StatefulSet application by using an image	
1.8.3 Create a Job application by using an image	
	01/
1.8.4 Create an application in Kubernetes dashboard	
1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t	emplate218
1.8.4 Create an application in Kubernetes dashboard1.8.5 Create a Linux application by using an orchestration t1.8.6 Create a Windows application by using an or	emplate216 chestration
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 	
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 	
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 	216 emplate218 chestration 222 227 Helm227
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 	216 emplate218 chestration 222 227 ; Helm227 236
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 	216 emplate218 chestration 222 227 Helm227 236 238
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 	216 emplate218 chestration 222 227 ; Helm227 236 238 239
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 	216 emplate218 chestration 222 227 Helm227 236 238 239 244
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.13 View a service 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244 248
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244 246 248 252
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244 248 248 252 253
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 	216 emplate218 chestration 222 227 227 236 238 238 239 244 246 246 248 252 253 254
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.18 View image list 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244 246 248 252 253 254 256
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.19 Use an image Secret 	216 emplate218 chestration 222 227 Helm227 3 Helm227 236 238 239 238 239 244 246 248 252 253 253 254 254 256 257
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.19 Use an image Secret 1.8.20 Pull a private image without a password 	216 emplate218 chestration 222 227 ; Helm227 236 238 238 239 244 246 248 252 253 254 256 257 261
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.18 View image list 1.8.19 Use an image Secret 1.9 Network management 	216 emplate218 chestration 222 227 Helm227 236 238 238 239 244 246 248 252 253 254 256 257 261 266
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.18 View image list 1.8.19 Use an image Secret 1.8.20 Pull a private image without a password 1.9 Networks supported by Alibaba Cloud Container 	216 emplate218 chestration 222 227 ; Helm227 236 238 238 239 244 246 248 252 253 254 256 257 261 266 Service for
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.18 View image list 1.8.20 Pull a private image without a password 1.9 Network management 1.9.1 Networks supported by Alibaba Cloud Container Kubernetes 	216 emplate218 chestration 222 227 Helm227 236 238 238 238 239 244 246 248 252 253 254 256 257 266 Service for 266
 1.8.4 Create an application in Kubernetes dashboard 1.8.5 Create a Linux application by using an orchestration t 1.8.6 Create a Windows application by using an or template 1.8.7 Manage applications by using commands 1.8.7 Manage applications by using commands 1.8.8 Simplify Kubernetes application deployment by using 1.8.8 Simplify Kubernetes application deployment by using 1.8.9 Use an application trigger 1.8.10 View pods 1.8.11 Create a service 1.8.12 Scale a service 1.8.13 View a service 1.8.14 Update a service 1.8.15 Delete a service 1.8.16 Change container configurations 1.8.17 Schedule a pod to a specified node 1.8.18 View image list 1.8.19 Use an image Secret 1.8.20 Pull a private image without a password 1.9.1 Networks supported by Alibaba Cloud Container Kubernetes 1.9.2 Terway network plugin	216 emplate218 chestration 222 227 ; Helm227 236 238 238 239 244 246 246 248 252 253 254 256 257 261 266 Service for 266 267

1.9.4 Use a network policy	271
1.10 Server Load Balancer and Ingress management	277
1.10.1 Overview	277
1.10.2 Access services by using Server Load Balancer	277
1.10.3 Support for Ingress	300
1.10.4 Configure Ingress monitoring	305
1.10.5 Ingress configurations	308
1.10.6 Create an Ingress in the Container Service console	311
1.10.7 Update an Ingress	322
1.10.8 View Ingress details	324
1.10.9 Deleting a route	324
1.11 Config map and Secret management	325
1.11.1 Create a Config Map	325
1.11.2 Use a config map in a pod	330
1.11.3 View a ConfigMap	335
1.11.4 Update a config map	336
1.11.5 Delete a config map	339
1.11.6 Create a secret	340
1.11.7 View secret details	343
1.11.8 Update a secret	344
1.11.9 Delete a secret	345
1.12 Storage management	346
1.12.1 Overview	346
1.12.2 Install the plug-in	347
1.12.3 Use Alibaba Cloud cloud disk volumes	351
1.12.4 Use NAS file systems of Alibaba Cloud	357
1.12.5 Use Alibaba Cloud OSS volumes	366
1.12.6 Create a Persistent Volume Claim	371
1.12.7 Use a persistent volume claim	374
1.13 Log management	375
1.13.1 Application log management	375
1.13.2 View cluster logs	376
1.13.3 Use Log Service to collect Kubernetes cluster logs	376
1.13.4 A solution to log collection problems of Kubernetes clusters	by
using log-pilot, Elasticsearch, and Kibana	388
1.13.5 Configure Log4jAppender for Kubernetes and Log Service	395
1.14 Monitoring management	400
1.14.1 Deploy the Prometheus monitoring system	400
1.14.2 Group-based monitoring and alarms	405
1.14.3 Integration and usage with CloudMonitor	410
1.14.4 Use Gratana to display monitoring data	416
1.14.5 Use an HPA auto scaling container	424
1.14.6 Monitor a Kubernetes cluster and send alarm notifications	by
using DingTalk	428
1.15 Security management	435

1.15.1 Security	436
1.15.2 Kube-apiserver audit logs	437
1.15.3 Implement secure access through HTTPS in Kubernetes	451
1.16 Release management	462
1.16.1 Manage a Helm-based release	
1.16.2 Use batch release on Alibaba Cloud Container Service	e for
Kubernetes	
1.17 Istio management	471
1.17.1 Overview	471
1.17.2 Deploy Istio	
1.17.3 Update Istio	
1.17.4 Delete Istio	
1.17.5 Upgrade Istio components	492
1.18 Template management	495
1.18.1 Create an orchestration template	495
1.18.2 Edit an orchestration template	
1.18.3 Save an existing orchestration template as a new one	501
1.18.4 Download an orchestration template	502
1.18.5 Delete an orchestration template	503
1.19 App catalog management	504
1.19.1 App catalog overview	504
1.19.2 View app catalog list	505

1 Kubernetes cluster

1.1 Security bulletins

1.1.1 Vulnerability fix: CVE - 2019 - 5736 in runc

The security vulnerability *CVE* – 2019 – 5736 in runc has been fixed for Alibaba Cloud Container Service for Kubernetes. This topic describes the impacts of this vulnerability and how to remove it.

Background

The security vulnerability may occur with Docker, containerd, or any other containers that use runc. This vulnerability gives attackers the ability to use a specific container image or run the exec command to obtain the file handle used by the running host runc. Attackers can overwrite the host runc binary file, then obtain root permission to access the host, and execute commands as with root permission.

For more information about security vulnerability *CVE* - 2019 - 5736 , see https:// cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-5736.

Affected clusters

- · Alibaba Cloud Container Service clusters affected by the vulnerability:
 - All Docker Swarm clusters from versions earlier than Docker v18.09.02.
 - All Kubernetes clusters except for Serverless Kubernetes clusters.
- · Self-built Docker and Kubernetes clusters affected by the vulnerability:
 - All clusters that use Docker versions earlier than v18.09.2.
 - All clusters that use runc v1.0-rc6 or earlier.

Note:

For both of the preceding vulnerability cases, we recommend that you consult your Docker or runc vendor for a solution.

Resolution

- Method 1: Create new Kubernetes clusters of v1.11 or v1.12. Kubernetes clusters of these two versions run the latest version of Docker, which are protected from this vulnerability.
- Method 2: Upgrade the Docker version of all existing clusters to v18.09.2 or later.
 Using this method will interrupt your cluster containers and services.
- Method 3: Only upgrade runc. This method is applicable to clusters running Docker v17.06. We recommend that you upgrade the runc binary file of each cluster node individually to avoid a service interruption caused by upgrading the Docker engine. To upgrade a runc binary file, complete the following steps:
 - 1. Run the following command to locate docker-runc:

```
Note:
Usually, docker-runc is located in / usr / bin / docker - runc .
```

which docker - runc

2. Run the following command to back up the original runc:

```
mv / usr / bin / docker - runc / usr / bin / docker - runc .
orig .$( date - Iseconds )
```

3. Run the following command to download the fixed runc:

```
curl – o / usr / bin / docker – runc – sSL https :// acs –
public – mirror . oss – cn – hangzhou . aliyuncs . com / runc /
docker – runc – 17 . 06 – amd64
```

4. Run the following command to set permission availability for docker-runc:

chmod + x / usr / bin / docker - runc

5. Run the following command to test whether runc works normally:

```
docker - runc - v
# runc version 1.0.0 - rc3
# commit : fc48a25bde 6fb041aae0 977111ad81 41ff396438
# spec : 1.0.0 - rc5
docker run - it -- rm ubuntu echo OK
```

- 6. To upgrade the runc binary file of a Kubernetes cluster GPU node, you must also install nvidia-runtime by completing the following steps:
 - a. Run the following command to locate nvidia-container-runtime:



Usually, nvidia-container-runtime is located in / usr / bin / nvidia -

```
container - runtime .
```

which nvidia - container - runtime

b. Run the following command to back up the original nvidia-containerruntime:

```
mv / usr / bin / nvidia - container - runtime / usr / bin /
nvidia - container - runtime . orig .$( date - Iseconds )
```

c. Run the following command to download the fixed nvidia-container-runtime:

```
curl - o / usr / bin / nvidia - container - runtime - sSL
https :// acs - public - mirror . oss - cn - hangzhou .
aliyuncs . com / runc / nvidia - container - runtime - 17 . 06
- amd64
```

d. Run the following command to set permission availability for nvidia-

container-runtime:

chmod + x / usr / bin / nvidia - container - runtime

e. Run the following command to test whether nvidia-container-runtime works normally:

```
nvidia - container - runtime - v
# runc version 1.0.0 - rc3
# commit : fc48a25bde 6fb041aae0 977111ad81 41ff396438 -
dirty
# spec : 1.0.0 - rc5
docker run - it -- rm - e NVIDIA_VIS IBLE_DEVIC ES =
all ubuntu nvidia - smi - L
# GPU 0 : Tesla P100 - PCIE - 16GB ( UUID : GPU -
122e199c - 9aa6 - 5063 - 0fd2 - da009017e6 dc )
```

Note:

This test is performed on a node of the GPU P100 model. Test outputs vary by GPU models.

1.1.2 Vulnerability fix: *cvE* - 2018 - 18264 **for Kubernetes** dashboard

Alibaba Cloud Container Service for Kubernetes has fixed dashboard vulnerability *CVE* - 2018 - 18264 . This topic describes the dashboard versions affected by the vulnerability and how to fix the vulnerability. The Kubernetes dashboards that are built in Alibaba Cloud Container Service for Kubernetes are not affected by this vulnerability because they work in the hosted form and their security settings were upgraded before the vulnerability occurred.

Background information

A security vulnerability, that is, *CVE* – 2018 – 18264 , was discovered in Kubernetes dashboards of V1.10 and earlier versions. This vulnerability allowed attackers to bypass identity authentication and read secrets within the cluster by using the dashboard logon account.

The Kubernetes dashboards that are built in Alibaba Cloud Container Service for Kubernetes are not affected by this vulnerability because they work in the hosted form and their security settings were upgraded before the vulnerability occurred.

For more information about security vulnerability CVE - 2018 - 18264 , see:

- https://github.com/kubernetes/dashboard/pull/3289
- https://github.com/kubernetes/dashboard/pull/3400
- https://github.com/kubernetes/dashboard/releases/tag/v1.10.1

Conditions required to determine that a Kubernetes dashboard is vulnerable

Your dashboard is vulnerable if you have independently deployed Kubernetes dashboard V1.10 or earlier versions (V1.7.0 to V1.10.0) that supports the logon function in your Kubernetes cluster, and you have used custom certificates.

Resolution

• If you do not need a dashboard that is deployed independently, run the following command to remove the Kubernetes dashboard from your cluster:

kubectl -- namespace kube - system delete deployment kubernetes - dashboard

- If you need an independently deployed dashboard, upgrade your dashboard to V1.10.1. For more information, see https://github.com/kubernetes/dashboard/releases/tag/v1.10.1.
- If you use the dashboard hosted by Alibaba Cloud Container Service for Kubernetes, you can continue to use your dashboard in the Container Service console because the dashboard was upgraded before the vulnerability occurred.

1.1.3 Vulnerability fix: CVE - 2018 - 1002105

Alibaba Cloud has fixed system vulnerability *CVE* - 2018 - 1002105 . This topic describes the impacts of this vulnerability and how to remove it.

This vulnerability does not affect Serverless Kubernetes clusters. Serverless Kubernetes was upgraded before the vulnerability occurred.

Background information

Engineers of the Kubernetes community have found security vulnerability CVE

- 2018 - 1002105 . Kubernetes users can gain access to the backend service by forging the request and escalating the permission on the established API Server connection. Alibaba Cloud has fixed this vulnerability. To remove the vulnerability, you need to log on to the Container Service console and upgrade Kubernetes to the latest version.

For more information about the vulnerability CVE - 2018 - 1002105 , see https:// github.com/kubernetes/kubernetes/issues/71411.

Affected Kubernetes versions:

- Kubernetes v1.0.x-1.9.x
- Kubernetes v1.10.0-1.10.10 (fixed in v1.10.11)
- Kubernetes v1.11.0-1.11.4 (fixed in v1.11.5)
- Kubernetes v1.12.0-1.12.2 (fixed in v1.12.3)

Affected configurations:

- Kubernetes cluster, which runs on Container Service and uses an extension API server. Furthermore, the extension API server network is directly accessible to the cluster component, kube-apiserver.
- Kubernetes cluster, which runs on Container Service and has opened permission s to interfaces such as pod exec, attach, and portforward. Then, users can use the vulnerability to obtain permissions to access all kubelet APIs of the cluster.

Cluster configuration of Alibaba Cloud Container Service for Kubernetes

• The API server of a Kubernetes cluster that runs on Container Service has RBAC enabled by default. That is, the API server denies anonymous user access through primary account authorization. Furthermore, the starting parameter of Kubelet is anonymous - auth = false , providing security access control against external attacks.

- If your Kubernetes cluster has multiple RAM users, the RAM users may gain unauthorized access to the backend service through interfaces such as pod exec , attach, and portforward. If your cluster has no RAM users, you do not need to worry about the vulnerability.
- RAM users do not have access to aggregate API resources by default without custom authorization from the primary account.

Solution

Log on to the Container Service console to upgrade your cluster. For more information, see *Upgrade a Kubernetes cluster*.

- If your cluster is V1.11.2, upgrade it to V1.11.5.
- If your cluster is V1.10.4, upgrade it to V1.10.11 or V1.11.5.
- If your cluster is V1.9 or earlier, upgrade it to V1.10.11 or V1.11.5. When you upgrade the cluster from V1.9 to V1.10 or V1.11, upgrade the flexvolume plugin through the console if your cluster uses cloud disk volumes.



In the Container Service console, select the target cluster and choose More > Addon Upgrade. In the Addon Upgrade dialog box, select flexvolume and click Upgrade.

1.2 Introduction

1.2.1 Overview

Kubernetes is a popular open-source container orchestration technology. To allow you to use Kubernetes to manage container applications in Alibaba Cloud, Alibaba Cloud Container Service provides support for Kubernetes clusters.

You can create a safe and high-availability Kubernetes cluster in the Container Service console. The Kubernetes cluster integrates with the virtualization, storage , network, and security capabilities of Alibaba Cloud to provide scalable, highperformance container application management, simplify cluster creation and expansion, and focus on the development and management of containerized applications.

Kubernetes supports the deployment, expansion, and management of containerized applications, and provides the following features:

- Elastic expansion and self-reparation.
- Service discovery and server load balancing.
- Service release and rollback.
- Secret and configuration management.

Limits

- Currently, Kubernetes clusters only support Linux containers. The support for Kubernetes Windows containers is in the works.
- Currently, Kubernetes clusters only support Virtual Private Cloud (VPC). You can select to create a VPC or use an existing VPC when creating a Kubernetes cluster.

Related open-source projects

- Alibaba Cloud Kubernetes Cloud Provider: https://github.com/AliyunContainerService/ kubernetes.
- Alibaba Cloud VPC network drive for Flannel: https://github.com/coreos/flannel/blob/ master/Documentation/alicloud-vpc-backend.md.

If you have any questions or suggestions regarding a specific project, you are welcome to raise an issue or pull a request in the community.

1.2.2 Alibaba Cloud Kubernetes vs. self-built Kubernetes

Advantages of Alibaba Cloud Kubernetes

Easy to use

- Supports creating a Kubernetes cluster with one click in the Container Service console.
- Supports upgrading Kubernetes clusters with one click in the Container Service console.

You may have to deal with self-built Kubernetes clusters of different versions at the same time, including version 1.8.6, 1.9.4, and 1.10 in the future. Upgrading clusters each time brings you great adjustments and Operation & Maintenance (O&M) costs . Container Service upgrade solution performs rolling update by using images and uses the backup policy of complete metadata, which allows you to conveniently roll back to the previous version.

- Supports expanding or contracting Kubernetes clusters conveniently in the Container Service console.
 - Container Service Kubernetes clusters allow you to expand or contract the capacity vertically with one click to respond to the peak of the data analysis business quickly.

Powerful

Function	Description
Network	 High-performance Virtual Private Cloud (VPC) network plug-in. Supports network policy and flow control. Container Service provides you with continuous network integration and the best network optimization.
Server Load Balancer	Supports creating Internet or intranet Server Load Balancer instances. If your self-built Kubernetes clusters are implemented by using the self- built Ingress, releasing the business frequently may cause pressure on Ingress configuration and higher error probabilities. The Server Load Balancer solution of Container Service supports Alibaba Cloud native high- availability Server Load Balancer, and can automatically modify and update the network configurations. This solution has been used by a large number of users for a long time, which is more stable and reliable than self-built Kubernetes.

Function	Description
Storage	Container Service integrates with Alibaba Cloud cloud disk, Network Attached Storage (NAS), and block storage, and provides the standard FlexVolume drive. Self-built Kubernetes clusters cannot
	. Alibaba Cloud Container Service provides the best seamless integration.
O&M	 Integrates with Alibaba Cloud Log Service and CloudMonitor. Supports auto scaling.
Image repository	 High availability. Supports high concurrency. Supports speeding up the pull of images. Supports P2P distribution.
	The self-built image repository may crash if you pull images from millions of clients at the same time. Enhance the reliability of the image repository by using the image repository of Container Service, which reduces the O&M burden and upgrade pressure.

Function	Description
Stability	 The dedicated team guarantees the stability of the container. Each Linux version and Kubernetes version are provided to you after strict tests.
	Container Service provides the Docker
	CE to reveal all the details and promotes
	the repair capabilities of Docker. If
	you have issues such as Docker Engine
	hang, network problems, and kernel
	compatibility, Container Service
	provides you with the best practices.
High availability	 Supports multiple zones. Supports backup and disaster recovery.
Technical support	 Provides the Kubernetes upgrade capabilities. Supports upgrading a Kubernetes cluster to the latest version with one click. Alibaba Cloud container team is responsible for solving problems about containers in your environment.

Costs and risks of self-built Kubernetes

Building clusters is complicated

You must manually configure the components, configuration files, certificates, keys, plug-ins, and tools related to Kubernetes. It takes several days or weeks for professional personnel to build the cluster.

• For public cloud, it takes you significant costs to integrate with cloud products.

You must devote your own money to integrate with other products of Alibaba Cloud, such as Log Service, monitoring service, and storage management.

• The container is a systematic project, involving network, storage, operating system , orchestration, and other technologies, which requires the devotion of profession al personnel.

• The container technology is continuously developing with fast version iteration, which requires continuous upgrade and test.

1.3 Authorization management

1.3.1 Role authorization

Grant the system default roles AliyunCSDefaultRole and AliyunCSClusterRole to the service account when you activate Container Service. Only after the roles are correctly granted, Container Service can normally call services such as Elastic Compute Service (ECS), Object Storage Service (OSS), Network Attached Storage (NAS), and Server Load Balancer (SLB), create clusters, and store logs.

Instructions

- If you have used Container Service before 15 January 2018, the system completes the role authorization by default. For the detailed granted permissions, see the following Default role permissions section. If you used Container Service with a Resource Access Management (RAM) user before, upgrade the authorization policy for the RAM user. For more information, see *Create custom authorization policies*.
- On 15 January 2018, Container Service is fully accessed to the cross-service authorization. New users who use the primary account can use Container Service only after having the cross-service authorization completed. If new users need to authorize RAM users to use Container Service, go to the RAM console to authorize the RAM users. For more information, see *Use the Container Service console as a RAM user*.

Procedure

1. If you have not granted the default roles to the service account correctly, the Cloud Resource Access Authorization page appears after you log on to the Container Service console. Click Confirm Authorization Policy.

e: If you need to modify role permissions, please go to the F nissions.	AM Console. Role Management. If you do not configure it correctly, the following role: CS will not be able to obtain the required
needs your permission to access your cloud resources.	
e system has created roles for the following user. These role	s can be found below. CS. After authorization, CS will have access to your cloud resources.
AliyunCSDefaultRole	
Description: The Container Service will use this role to acces	s your resources in other services.
Permission Description: The policy for AliyunCSDefaultRole.	
AliyunCSClusterRole	
Description: The clusters of Container Service will use this re	le to access your resources in other services.
Permission Description: The policy for AliyunCSClusterRole.	



Container Service has configured the default role permissions. To modify the role permissions, go to the User Management page of the RAM console. Note that incorrect configurations might cause Container Service cannot obtain the required permissions.

2. After completing the authorization, refresh the Container Service console and then perform the operations.

To view the policy details of the roles AliyunCSDefaultRole and AliyunCSClusterRole, log on to the *RAM console*.

Default role permissions

For more information about permissions of each role, see the API documents of each product.

AliyunCSDefaultRole permissions

The default role AliyunCSDefaultRole contains the following main permissions:

• ECS-related permissions

Action	Description
ecs:RunInstances	Query ECS instance information.
ecs:RenewInstance	Renew ECS instances.

Action	Description
ecs:Create*	Create ECS-related resources, such as instances and disks.
ecs:AllocatePublicIpAddress	Allocate public IP addresses.
ecs:AllocateEipAddress	Allocate Elastic IP (EIP) addresses.
ecs:Delete*	Delete ECS instances.
ecs:StartInstance	Start ECS-related resources.
ecs:StopInstance	Stop ECS instances.
ecs:RebootInstance	Restart ECS instances.
ecs:Describe*	Query ECS-related resources.
ecs:AuthorizeSecurityGroup	Configure inbound security group rules.
ecs:RevokeSecurityGroup	Revoke security group rules.
ecs:AuthorizeSecurityGroupEgress	Configure outbound security group rules.
ecs:AttachDisk	Add disks.
ecs:DetachDisk	Clean up disks.
ecs:AddTags	Add tags.
ecs:ReplaceSystemDisk	Change system disks of ECS instances.
ecs:ModifyInstanceAttribute	Modify ECS instance attributes.
ecs:JoinSecurityGroup	Add ECS instances to specified security groups.
ecs:LeaveSecurityGroup	Remove ECS instances from specified security groups.
ecs:UnassociateEipAddress	Unbind EIP addresses.
ecs:ReleaseEipAddress	Release EIP addresses.

· Virtual Private Cloud (VPC)-related permissions

Permission name (Action)	Permission description
vpc:Describe*	Query information of VPC-related resources.
vpc:DescribeVpcs	Query VPC information.
vpc:AllocateEipAddress	Allocate EIP addresses.
vpc:AssociateEipAddress	Associate with EIP addresses.

Permission name (Action)	Permission description
vpc:UnassociateEipAddress	Do not associate with EIP addresses.
vpc:ReleaseEipAddress	Release EIP addresses.
vpc:CreateRouteEntry	Create router interfaces.
vpc:DeleteRouteEntry	Delete router interfaces.

· SLB-related permissions

Action	Description
slb:Describe*	Query information related to Server Load Balancer.
slb:CreateLoadBalancer	Create Server Load Balancer instances.
slb:DeleteLoadBalancer	Delete Server Load Balancer instances.
slb:RemoveBackendServers	Unbind Server Load Balancer instances.
slb:StartLoadBalancerListener	Start specified listeners.
slb:StopLoadBalancerListener	Stop specified listeners.
slb:CreateLoadBalancerTCPListener	Create TCP-based listening rules for Server Load Balancer instances.
slb:AddBackendServers	Add backend servers.

AliyunCSClusterRole permissions

The default role AliyunCSClusterRole contains the following main permissions:

· OSS-related permissions

Action	Description
oss: PutObject	Upload file or folder objects.
oss: GetObject	Get file or folder objects.
oss: ListObjects	Query file list information.

• NAS-related permissions

Action	Description
nas:Describe*	Return NAS-related information.
nas:CreateAccessRule	Create permission rules.

· SLB-related permissions

Action	Description
slb:Describe*	Query information related to Server Load Balancer.
slb:CreateLoadBalancer	Create Server Load Balancer instances.
slb:DeleteLoadBalancer	Delete Server Load Balancer instances.
slb:RemoveBackendServers	Unbind Server Load Balancer instances.
slb:StartLoadBalancerListener	Start specified listeners.
slb:StopLoadBalancerListener	Stop specified listeners.
slb:CreateLoadBalancerTCPListener	Create TCP-based listening rules for Server Load Balancer instances.
slb:AddBackendServers	Add backend servers.
slb:DeleteLoadBalancerListener	Delete listening rules of Server Load Balancer instances.
slb:CreateVServerGroup	Create VServer groups and add backend servers.
slb:ModifyVServerGroupBackendServers	Change backend servers in VServer groups.
slb:CreateLoadBalancerHTTPListener	Create HTTP-based listeners for Server Load Balancer instances.
slb:SetBackendServers	Configure backend servers and set the weight for a group of ECS instances at the Server Load Balancer instance backend.
slb:AddTags	Add tags for Server Load Balancer instances.

1.3.2 Use the Container Service console as a RAM user

You can log on to and perform operations in the Container Service console as a RAM user.

Before you can log on to the Container Service console and perform operations as a RAM user, you must grant related permissions to the RAM user.

Step 1: Create a RAM user and enable console logon

- 1. Log on to the RAM console.
- 2. In the left-side navigation bar, click Users. Then, click Create User.

- 3. Enter a user name for the RAM user and then click OK.
- 4. On the Users page, select the created RAM user and click Manage.
- 5. In the Web Console Logon Management area, click Enable Console Logon.
- 6. Enter a logon password and click OK.

Step 2: Grant the RAM user permissions to access Container Service

1. On the Users page, select the created RAM user and click Authorize.

RAM	User Management		Create User CRefresh
Dashboard Users	User Name V Search by User Name Search		
Groups	User Name/Display Name Description	Created At	Actions
Policies	thangtiven transplaied	2017-11-01 11:26:17	Manage Authorize Delete Join Group
Settings	Altxx1003TokenQementer/law	2017-11-27 11:55:21	Manage Authorize Delete Join Group
	yangsun yangsun	2017-11-01 11:24:54	Manage Authorize Delete Join Group

2. Select the required policies to attach them to the RAM user.

Edit User-Level Authorization					×
Members added to this group have a group more than once.	II the perr	niss	ions of this grou	up. A member cannot be added to the sa	me
Available Authorization Policy Names	Туре	;		Selected Authorization Policy Name	Туре
CS EcsRamRoleDocument		Q •		AdministratorAccess Provides full acce	System
AliyunACSResourcesAccess_yangx aliyun container s	Custom		>	AliyunACSResourcesAccess_xingy aliyun container s	Custom
AliyunCSReadOnlyAccess Provides read-only	System		<		
AliyunCSFullAccess Provides full acce	System				
		Ŧ			
				ОК	Close

You can use the following system policies:

- · AliyunCSFullAccess: Provides full access to Container Service.
- · AliyunCSReadOnlyAccess: Provides read-only access to Container Service.

You can also create custom policies as you need and attach them to the RAM user. For more information, see *Create custom authorization policies*.

Step 3: Log on to the Container Service console as a RAM user

• If you have granted the AliyunCSDefaultRole and AliyunCSClusterRole roles to the Alibaba Cloud account, you can log on to the Container Service console and perform operations as a RAM role directly.

Log on to the Container Service console as a RAM user.

• If you have not granted the AliyunCSDefaultRole and AliyunCSClusterRole roles to the Alibaba Cloud account, you must log on to the Container Service console using the account credentials and

click Confirm Authorization Policy on the authorization page to grant the account the following permissions.

Cloud Resource Access Authorization			
Note: If you need to modify role permissions, please go to the RAM Console. Role Management. If you do not configure it correctly, the following role: CS will not be able to obtain the required permissions.	×		
CS needs your permission to access your cloud resources.			
The system has created roles for the following user. These roles can be round below. CS. After authorization, CS will have access to your cloud resources.			
AliyunCSDefaultRole	~		
Description: The Container Service will use this role to access your resources in other services.			
Permission Description: The policy for AliyunCSDefaultRole.			
AliyunCSClusterRole	~		
Description: The clusters of Container Service will use this role to access your resources in other services.			
Permission Description: The policy for AliyunCSClusterRole.			
Confirm Authorization Policy Cancel			

After you grant the preceding permissions to the account, you can log on to the Container Service and perform related operations as a RAM user.

1.3.3 Create custom authorization policies

The authorization granularity of the system authorization policies provided by Container Service is coarse. If these authorization policies with coarse granularit y cannot satisfy your requirements, create the custom authorization policies. For example, to control the permissions to a specific cluster, you must use the custom authorization policy to meet the requirements with fine granularity.

Create custom authorization policies

Get to know the basic structure and syntax of the authorization policy language before creating custom authorization policies. For more information, see *Authorization policy language descriptions*. This document introduces how to grant Resource Access Management (RAM) users permissions to query, expand, and delete clusters.

Procedure

- 1. Log on to the RAM console with the primary account.
- 2. Click Policies in the left-side navigation pane. Click Create Authorization Policy in the upper-right corner.
- 3. Select a template. Enter the authorization policy name and the policy content.

Create Authorization Policy		\times		
Step 1: Select an authorizatio	n policy Step 2: Edit permissions and submit. Policy creation complete.			
* Authorization Policy Name :	clusterpolicy Names must be 1-128 characters long. They may only contain the letters A-Z, numbers 0-9, and hyphens.			
Description :				
Policy Content :	1 { "Statement": [{ "Action": ["cs:ScaleCluster", "cs:DeleteCluster", "cs:DeleteCluster" "Bffect": "Allow", "Resource": ["acs:cs:*:*:cluster/cb2f4c] "acs:cs:*:*:cluster/cb2f4c			
	Previous Create Authorization Policy Cano	el		
<pre>{ Statement ": [{ " Action ": [" cs : Get *", " cs : ScaleClust er ", " cs : DeleteClus ter "], " Effect ": " Allow ", " Resource ": [" acs : cs :*:*: cluster / cluster ID "] } </pre>				

}], " Version ": " 1 " }

where:

• Action : Enter the permission that you want to grant.

Note:

All the Actions support wildcards.

• Resource supports the following configuration methods.

- Grant permissions of a single cluster

```
" Resource ": [
        " acs : cs :*:*: cluster / cluster ID "
]
```

- Grant permissions of multiple clusters

```
" Resource ": [
    " acs : cs :*:*: cluster / cluster ID ",
    " acs : cs :*:*: cluster / cluster ID "
]
```

- Grant permissions of all your clusters

```
" Resource ": [
"*"
]
```

You must replace cluster ID with your actual cluster ID.

4. Click Create Authorization Policy after completing the configurations.

Table 1-1: Container Service RAM action

Action	Description
CreateCluster	Create clusters.
AttachInstances	Add existing Elastic Compute Service (ECS) instances to clusters.
ScaleCluster	Expand clusters.
GetClusters	View cluster list.
GetClusterById	View cluster details.
ModifyClusterName	Modify cluster names.
DeleteCluster	Delete clusters.
UpgradeClusterAgent	Upgrade cluster Agent.

Action	Description
GetClusterLogs	View cluster operation logs.
GetClusterEndpoint	View cluster access point.
GetClusterCerts	Download cluster certificate.
RevokeClusterCerts	Revoke cluster certificate.
BindSLB	Bind Server Load Balancer instances to clusters.
UnBindSLB	Unbind Server Load Balancer instances from clusters.
ReBindSecurityGroup	Rebind security groups to clusters.
CheckSecurityGroup	Check existing security group rules of clusters.
FixSecurityGroup	Fix cluster security group rules.
ResetClusterNode	Reset cluster nodes.
DeleteClusterNode	Delete cluster nodes.
CreateAutoScale	Create node auto scaling rules.
UpdateAutoScale	Update node auto scaling rules.
DeleteAutoScale	Delete node auto scaling rules.
GetClusterProjects	View applications in clusters.
CreateTriggerHook	Create triggers for applications.
GetTriggerHook	View application trigger list.
RevokeTriggerHook	Delete application triggers.
CreateClusterToken	Create tokens.

1.3.4 Grant Kubernetes permissions to a RAM user

This topic describes how to configure Kubernetes Resource Access Management (RAM) cluster permissions and the corresponding Kubernetes RBAC application permissions within the cluster by using the Container Service console.

Before you begin

• Make sure that you have an Alibaba Cloud account and that you have created one or more RAM users.

- If you grant a RAM user the corresponding permissions for a Container Service Kubernetes cluster, you must use an account to complete the task if further operations in the RAM console is required.
- If a RAM user is granted the preset admin role or the custom cluster-admin role in a cluster or namespace, then the RAM user can grant permissions to other RAM users in the same cluster or namespace.
- If RAM authorization is involved in configuring permissions through the Container Service console, you must manually perform authorization in the RAM console for the target RAM user according to the reference policy and operation instructions on the page due the security restrictions of RAM.

Authorization policy upgrade notice

Container Service will upgrade the cluster authorization policy (in April 2019) to enhance the security of Kubernetes clusters. After the upgrade, any RAM users in a Kubernetes cluster that are not granted the required permissions will not be able to access the cluster resources.

Therefore, we recommend that you grant required permissions to the RAM users in each of your Kubernetes clusters. After you complete this process, your managed RAM users will only have the specified permissions to access the their corresponding authorized cluster.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service Kubernetes, choose Clusters > Authorization.
- 3. On the right of the target RAM user, click Authorize.



If you want to use a RAM user for the preceding operations, make sure that the target RAM user has been granted the corresponding admin permissions or the cluster-admin role.

Container Service - Kubernetes +	Authorization 🗨 Back to List				
Overview	Calest Cubassount	Dec	ourse Authorization	Lindata Auth	orization Doligy
 Clusters 	Select Subaccount	Res	ource Authonzauon		Unzation Policy
	User Name	Display Name	Comment	Time Created	
Clusters		100,000		09/12/2018,14:01:28	Authorize
Nodes	10.00	1010-001	test user	09/04/2018,16:03:59	Authorize
Volumes	10.0	100		08/24/2018,13:43:17	Authorize
Namespace					
Authorization				08/09/2018,14:35:15	Authorize
 Application 	HENCE .	10100		06/29/2018,13:17:47	Authorize 2

4. In the upper-left corner of the Resource Authorization tab page, click the plus sign, set the cluster and namespace where the permissions to be granted apply, set a role for the RAM user, and then click Next Step.



• You can grant one preset role and multiple custom roles to a RAM user in a cluster or a namespace.

• You can click the minus sign to remove a group of permission settings (that is, the settings of the cluster, the namespace, and the role).

Authorizatio	n <u>* Back to List</u>	
	Select Subaccount	Resource Authorization Update Authorization Policy
0	Cluster/Namespace	Role
•	Clusters k8s-cluster v Namespace all namespace v	* Admin Operation Developer Restricted User Custom
•	Clusters k8s-cluster Namespace all namespace	Admin Operation Developer Restricted User Custom alibaba-loa-controller
•	Clusters k8s-cluster Namespace all namesoace	Admin Operation Developer Restricted User Custom alicloud-disk-controller-runner
•	Clusters k8s-cluster Namespace default	Developer Restricted User Custom
Permission d	escription	
	Cluster management permissions	Application management permissions
Admin	Read and write permissions of clusters. Can delete, scale clusters, add nodes	Read and write permissions of resources in all namespaces, Read and write permissions of nodes, volumes, namespaces and quotas
Operation	Read and write permissions of clusters. Can delete, scale clusters, add nodes	Read and write permissions of resources in all namespaces, Read only permissions of nodes, volumes, namespaces and quotas
Developer	Read only permissions of clusters	Read and write permissions of resources in all namespaces or specified namespace
Restricted User	Read only permissions of clusters	Read only permissions of resources in all namespaces or specified namespace
Custom	The permissions are determined by the specified cluster roles. Before you a prevent the RAM user from obtaining unnecessary permissions.	uthorize the RAM user, make sure that you are aware of all the resource access permissions of the selected cluster roles. This will
		Prev Step Next Step

The role permissions that apply to clusters and namespaces are shown in the following table.

Role	Cluster management permissions	Application management permissions
Admin	The read and write permissions of clusters . Specifically, the permissions to delete clusters, scale clusters, and add nodes.	The read and write permissions of resources in all namespaces, and those of nodes, volumes, namespaces, and quotas.
Operation	The read and write permissions of clusters . Specifically, the permissions to delete clusters, scale clusters, and add nodes.	The read and write permissions of resources in all namespaces, and the ready permission s of nodes, volumes, namespaces, and quotas.

Table 1-2: Roles and permissions

Role	Cluster management permissions	Application management permissions
Developer	The read permissions of clusters.	The read and write permissions of the resources in all namespaces or specified namespaces.
Restricted User	The read permissions of clusters.	The read permission s of resources in all namespaces or specified namespaces.
Custom	The read and write permissions of clusters . Specifically, the permissions to delete clusters, scale clusters, and add nodes.	The permissions of the RAM user depend on the cluster role you select. Confirm the permissions that your selected cluster role has on resources before authorization, to avoid inappropriate permissions granted to the RAM user.

- 5. If Authorized success is displayed, this means that you have granted the target RAM user the permissions. If Add authorization policies is displayed, follow these steps:
 - If you are using a RAM user, you need to log out, and then use an account to log on to the console to grant required permissions to the target RAM user.
 - If you are using a primary account, you can directly perform the next step.
6. Click Copy, and then click Policy Management.

	Container Service - Kubernetes 👻	Authorization t Back to List
	Overview	
		Select Subaccount Resource Authorization Update Authorization Policy
•	Clusters	Add authorization policies
	Clusters	To keep your account secure, you need to manually add authorization policies to the sub-account by accessing the RAM console.
	Nodes	1 Open Policy Management to add or update custom policies starting with AliyunACSResourcesAccess-mengguogiang. The policy name and content are as follows:
	Volumes	
	Namespace	Policy Name Policy Content
	Hamospace	AllyunACSResourcesAccess-mengguoqiang
	Authorization	Statement** [Copy
	Application	
		"Action": [
	Deployment	″xan:Get≭″,
	StatefulSet	″ram:List≭″,
		<pre>"cs:GetUserPermissions",</pre>
	DaemonSet	"cs:GetSubUsers",
	Job	"cs:GrantPermission"
		1,
	CronJob	10 million
	Pods	2 Open User Management, and perform the "authorization" operation on the sub-account mengguogiang to authorize the policy created above.
	Volumes Claim	3 Authorized success
	Release	

7. Choose Permissions > Policies, and then click Create Policy.

RAM		RAM / Policies				
Overview		Policies				
Identities /	~					
Groups		A policy describes a permission set. Alibaba Cloud uses a s RAM supports two types of policies: system polici	simple language specification to describe permission sets. For more information, cles managed by Alibaba Cloud and custom policies managed by yo	see Policy syntax structu DU.	ire.	×
Users		 You can use but cannot modify the system policies mana You can create, modify, or delete the custom policies. In 	aged by Alibaba Cloud. Alibaba Cloud maintains and updates the system policy v addition, you need to maintain the policy versions by yourself.	ersions.		
Settings		3				
Permissions	• 5	Create Policy Enter a policy name or note	Q Policy Type All \checkmark			c
Grants	«	Policy Name	Note	Policy Type	Used Times	Actions
Policies 2		and the second sec	Provides full access to Alibaba Cloud services and resources.	System Policy	9	
RAM Roles			Provides full access to Object Storage Service(OSS) via Management	System Policy	0	
OAuth Applications			Console.			

8. Enter a Policy Name, select the Script configuration mode, use the hot key Ctrl+V to paste the content that was copied in step 6 in the Policy Document area, and then click OK. For more information, see *Create custom authorization policies*.

RAM	RAM / Policies / Create Custom Policy
Overview	← Create Custom Policy
Identities ^	
Groups	Policy Name
Users	AliyunCS-001
Settings	Note
Permissions \land	
Grants	Configuration Mode
Policies	Visualized
- Oncies	Script
RAM Roles	V Policy Document
OAuth Applications	Import an existing system policy
	<pre> 1</pre>

9. In the left-side navigation pane, choose Identities > Users. On the right of the target user, click Add Permissions.

RAM		RAM / Users		
Overview		Users		
Identities 1 Groups	^	A RAM user is an identity entity. It represents a user or application in your organization that needs to access cloud You can manage users in the following steps:	resources.	×
Users 2		1.Create a RAM user, and set a password for this user to log on to the console or create an AccessKey for the appli 2.Add the user to a group. To perform this operation, you must have created a group and granted permissions to i	cation to call APIs. t.	
Settings				
Permissions	^	Create User User Logon Name 🗡 Enter Q		C
Grants		User Logon Name/Display Name Note	Created	Actions 3
Policies RAM Roles			Feb 18, 2019, 16:03:11	Add to Group Add Permissions Delete
OAuth Applications	«		Feb 15, 2019, 14:21:00	Add to Group Add Permissions Delete
			Jan 30, 2019, 17:09:53	Add to Group Add Permissions Delete
			Jan 22, 2019, 13:06:46	Add to Group Add Permissions Delete

10.Select Custom Policy, search for or manually look for the customized policy, click the policy name to add the policy to the Selected area on the right, and then click OK.

RAM	RAM / Users	Add Permissions	×
Overview	Users	. mengguoqiang@containerdoc.onaliyun.com $ imes$	
Identities ^ Groups	A RAM user is an identity entity. It represents a user or applica You can manage users in the following steps:	Select Policy 2	
Users	1.Create a RAM user, and set a password for this user to log or 2.Add the user to a group. To perform this operation, you mus	Custom Policy V Enter	Q Selected (1) 4 Clear
Settings Permissions	Create User User Logon Name 🗡 Enter	k8sWorkerRolePolicy- 672d4e42-16eb-43d2-b157- 42ec63c1f5e8	AliyunCS-001 ×
Grants	User Logon Name/Display Name	k8sMasterRolePolicy- 67204442-16eb-43d2-b157- 42ec63c1f5e8	
RAM Roles		k8sWorkerRolePolicy- 14d5de66-e1f7-4738-b33e-	
OAuth Applications	- MR	k8sMasterRolePolicy-	
		6b33802f1ddb	
		K8sMasterRolePolicy- 65edaba2-7c72-4812-bdfe- 7c3b40473a4b	_
	- Martin	k8sWorkerRolePolicy- 65edaba2-7c72-4812-bdfe- 7c3b40473a4b	
	C The second second second	alarmtest mgq	
		Ok Cancel	×



Note:

After you complete these steps, you can use the target RAM user to log on to the Container Service console and perform the operations allowed by the granted permissions. Typically, if the target RAM user is granted the admin permissions, you can use it to grant permissions to other RAM users.

Custom permissions

Alibaba Cloud Container Service offers four types of permissions by pre-setting four types of roles: Admin, Operation, Developer, and Restricted User. These types of permissions can meet the needs of most users in the Container Service console. However, if you want to customize the access permissions to clusters, you can also use the custom permissions.

Note:

Alibaba Cloud Container Service provides several custom permission. Among them, the cluster-admin permission is a super administrator permission with the permissions to access and operate on all resources.

Authoriza	tion 🔹 Back to List	
	Select Subaccount	Resource Authorization Vpdate Authorization Policy
0	Cluster/Namespace	Role
•	Clusters xuntest2 v Namespace all namespace v	Admin Operation Developer Restricted User Custom admin alibaba-log-controller
Permission	description	alicloud-disk-controller-runner duster-admin
	Cluster management permissions	Application management per flannel
Admin	Read and write permissions of clusters. Can delete, scale clusters, add nodes	Read and write permissions of resources in all namespaces, Read and write permissions of nodes, volumes, namespaces and quotas
Operation	Read and write permissions of clusters. Can delete, scale clusters, add nodes	Read and write permissions of resources in all namespaces, Read only permissions of nodes, volumes, namespaces and quotas
Developer	Read only permissions of clusters	Read and write permissions of resources in all namespaces or specified namespace
Restricted User	Read only permissions of clusters	Read only permissions of resources in all namespaces or specified namespace
Custom	The permissions are determined by the specified cluster roles. Before y roles. This will prevent the RAM user from obtaining unnecessary perm	ou authorize the RAM user, make sure that you are aware of all the resource access permissions of the selected cluster issions.
		Prev Step Next Step

You can log on to the cluster Master node and run the following command to view the details of the custom permissions.

```
Note:
```

Only some of the cluster roles are displayed.

```
kubectl
#
             get
                   clusterrol e
NAME
    AGE
admin
   13d
alibaba - log - controller
       13d
alicloud - disk - controller - runner
         13d
cluster - admin
      13d
cs : admin
      13d
edit
    13d
 flannel
   13d
kube - state - metrics
        22h
node - exporter
      22h
prometheus - k8s
      22h
prometheus - operator
      22h
system : aggregate - to - admin
          13d
```

```
....
system : volume - scheduler
13d
view
13d
```

To view the permission details of the super administrator cluster-admin, run the following command.

Note:

After the RAM user is granted the cluster-admin role, the RAM user can be regarded as a super administrator that has the same privileges as the Alibaba Cloud account, and it can perform operations on any resources in the cluster. Execute caution when you grant the cluster-admin role.

```
kubectl get clusterrol e cluster - admin - apiVersion : rbac . authorizat ion . k8s . io / v1
                                                        - o
#
                                                               yaml
kind : ClusterRol e
metadata :
   annotation s:
     rbac . authorizat ion . kubernetes . io / autoupdate : " true "
   creationTi mestamp : 2018 - 10 - 12T08 : 31 : 15Z
   labels :
     kubernetes . io / bootstrapp ing : rbac - defaults
  name : cluster - admin
   resourceVe rsion : " 57 "
   selfLink : / apis / rbac . authorizat ion . k8s . io / v1 /
clusterrol es / cluster - admin
   uid : 2f29f9c5 - cdf9 - 11e8 - 84bf - 00163e0b2f 97
 rules :
 apiGroups :
    1+1
  resources :
  - '*'
  verbs :
  - '*'
  nonResourc eURLs :
  - '*'
  verbs :
  - '*'
```

1.4 Cluster management

1.4.1 View cluster overview

This topic describes how to view the health status of a Kubernetes cluster. Alibaba Cloud Container Service for Kubernetes provides each Kubernetes cluster with the application status, component status, and resource monitoring charts.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Overview.
- 3. Select the target cluster and namespace. You can view the application status, component status, and resource monitoring charts.
 - Application status: The status of deployments, pods, and replica sets that are currently running. Green indicates the normal status and orange indicates an exception.
 - · Node status: Displays the node status of the current cluster.
 - Component status: The components of Kubernetes clusters are generally deployed under the kube-system namespace, including the core components such as scheduler, controller-manager, and etcd.
 - Resource monitor: Provides the monitoring charts of CPU and memory. CPU is measured in cores and is accurate to three decimal places. The minimum unit is millicores, that is, one thousandth of one core. Memory is measured in G and is accurate to three decimal places. For more information, see *Meaning of CPU* and *Meaning of memory*.
 - Event: Displays event information of the cluster, such as warnings and error events.



1.4.2 Create a Kubernetes cluster

This topic describes how to quickly create a Kubernetes cluster in the Container Service console.

Detail analysis

During cluster creation, Container Service performs the following operations:

- Creates Elastic Compute Service (ECS) instances, sets the public key used for SSH logon from the management node to other nodes, and installs and configures the Kubernetes cluster by using cloud-init.
- Creates a security group to allow inbound access to all ICMP ports in a Virtual Private Cloud (VPC).
- Creates a new VPC and VSwitch (if no existing VPC is selected), and creates an SNAT entry for the VSwitch.
- · Creates VPC routing rules.
- · Creates a NAT gateway and a shared bandwidth package or Elastic IP (EIP).
- Creates a Resource Access Management (RAM) user and an AccessKey. The RAM user has the permissions to query, create, and delete an ECS instance, add and delete a cloud disk, and all permissions to perform required actions on Server Load Balancer (SLB) instances, CloudMonitor, VPC, Log Service, and the Network Attached Storage (NAS) service. The Kubernetes cluster dynamically creates SLB instances, cloud disks, and VPC routing rules according to your configurations.
- · Creates an intranet SLB instance and allows access to port 6443.
- Creates an Internet SLB instance and allows access to port 6443. (If you enable the SSH logon for Internet access when creating the cluster, port 22 is opened. Otherwise, port 22 is closed.)

Prerequisites

You have activated the following services: Container Service, Resource Orchestration Service (ROS), RAM, and Auto Scaling service.

If you have not activated one or some of these services, log on to the *Container Service console*, *ROS console*, *RAM console*, and *Auto Scaling console* to activate the corresponding service or services.



The deployment of Kubernetes clusters in Alibaba Cloud Container Service for Kubernetes is dependent on the application deployment capabilities of Alibaba Cloud ROS. Therefore, you need to activate ROS before creating a Kubernetes cluster.

Limits

- The SLB instances created with the cluster support only the Pay-As-You-Go billing method.
- The Kubernetes cluster supports only the VPC network type.
- By default, each account has a set quota of cloud resources that can be used to create clusters. If you exceed this number when creating a cluster, the cluster creation fails. However, you can open a ticket to increase your quota of cloud resources.
 - By default, each account can create up to 5 clusters across all regions and add up to 40 nodes to each cluster. You can open a ticket to increase your quota to create more clusters or nodes.

Note:

In a Kubernetes cluster, the maximum number of default routes a VPC can have is 48, which means the number of nodes the Kubernetes cluster can use is also a maximum of 48. To increase the number of nodes, you need to first open a ticket for the target VPC to increase the number of VPC routes, and then open a ticket for Container Service to apply the increase.

- By default, each account can create up to 100 security groups.
- By default, each account can create up to 60 Pay-As-You-Go SLB instances.
- By default, each account can create up to 20 EIPs.
- The limits for ECS instances are as follows:
 - Only the CentOS operating system is supported.
 - Both Pay-As-You-Go and Subscription ECS instances can be created.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Clusters.

3. In the upper-right corner, click Create Kubernetes Cluster.

Container Service - Kubernetes 👻	Cluster List		You can create up to 5 clusters and can add up to	40 nodes in each cluster. Refresh	Create Kubernetes Cluster
Overview	Help: Ø Create cluster Ø Scale cluster Ø Co	onnect to Kubernetes cluster via kubectl $ \mathscr{O} $ Mar	nage applications with commands		2
Clusters	Name 🔻				
Clusters	Cluster Name/ID Clus	ter Type Region (All) 👻 Netwo	rk Type Cluster Status Time Created Kut	pernetes Version	Action
Nodes					
Create Kube	rnetes Cluster 🔹 Back	to Cluster List	Convertises Kulturentes (beta)		
Kubernetes	Managed Kubernetes	Multi-AZ Kudernetes	Serveriess Kubernetes (beta)		
* Cluster Nam	e k8s-cluster The cluster name s hyphens.	hould be 1-63 characters l	ong, and can contain numbers, Chir	nese characters, Engli	ish letters and

4. Enter the cluster name.

The cluster name must be 1 to 63 characters in length and can contain letters, numbers, Chinese characters, letters, and hyphens (-).

5. Select the region and the zone where you want to locate the cluster.

Region	China North 2	China North 3	China East 1	China East 2	China South 1		Asia Pacific SE 1	Asia Pacific SE 3	Asia Pacific SE 5	Asia Pacific SOU 1
	(Beijing)	(Zhangjiakou)	(Hangzhou)	(Shanghai)	(Shenzhen)	Hong Kong	(Singapore)	(Kuala Lumpur)	(Jakarta)	(Mumbai)
		US West 1 (Silicon	Middle East 1	EU Central 1						
	US East 1 (Virginia)	Valley)	(Dubai)	(Frankfurt)						
Zone	China North 2 Zone	A	•							

6. Set the cluster network type. Kubernetes clusters support only the VPC network type.

VPC: You can click Auto Create to create a VPC together with the Kubernetes cluster, or click Use Existing to use an existing VPC. If you click Use Existing, you can select a VPC and a VSwitch from the two displayed drop-down lists.

- If you click Auto Create, the system automatically creates a NAT gateway for your VPC when creating the cluster.
- If you click Use Existing and the selected VPC has an existing NAT gateway, the system then uses the already created NAT gateway. Otherwise, the system automatically creates a NAT gateway by default. Alternatively, if you do not want the system to automatically create a NAT gateway, deselect the Configure SNAT for VPC check box.



If you set the system to not automatically create a NAT gateway, you need to manually set a NAT gateway, or set an SNAT entry to ensure that your selected VPC is accessible to the Internet. Otherwise, instances in the VPC cannot access the Internet, which result in a cluster creation failure.

VPC	Auto Create	Use Existing			
	VPC123 (vpc-2zercq4	lpyanzxsfiiclyl)	•	VSwitch123 (vsw-2zeydhl5uwh1ej522lauo) ZoneA	-

- 7. Set the required billing method for the node type. Pay-As-You-Go and Subscription are supported.
- 8. Set the Master node configuration by selecting the Master node instance type and the system disk.



- · Only the CentOS operating system is supported.
- By default, the system creates three Master nodes for each cluster. This number cannot be modified.
- A system disk is attached to Master nodes by default. Available system disks are SSD disks and Ultra disks.

MASTER Configuration	on							
Instance Type	4 Core(s) 8 G (ecs.n1.large)				•	T	Quantity 3unit(s)	
System Disk	Ultra Disk	•	120	GiB	*			

9. Set Worker nodes. You can create Worker node instances or add existing Worker node instances.



- Only the CentOS operating system is supported.
- Each cluster can contain up to 37 Worker nodes. To create more Worker nodes, open a ticket.
- A system disk is attached to Worker nodes by default. Available system disks are SSD disks and Ultra disks.

- You can manually attach a data disk to Worker nodes. Available data disks are SSD disks, Ultra disks, and basic disks.
- To create Worker node instances, select the instance type and set the number of Worker nodes. In this example, one Worker node is created.

WORKER Configuration									
Instance Type	4 Core(s) 8 G (ecs.n1.large)				• 7	Quantity	1	unit(s) 🌲	
System Disk	Ultra Disk	•	40	GiB	*				
🗸 Attach Data Disk	Ultra Disk	-	100	GiB	*				

• To add existing Worker node instances, you must create ECS instances in the target region in advance.

Worker Instance	Create	Add	
	You can now convert a	paid instance to an ex	xample of an annual subscription through the ECS Management Console. View details
[Add Existing Instance		

10.Set the logon mode.

• Set a key pair.

Click Create a new key pair to create a key pair in the ECS console, and then set the key pair as the credentials for logging on to the cluster. For more information, see *Create an SSH key pair*.

Login	Key Pair	Password		
Key Pair Name			Ŧ	c
	You can visit ECS conso	le to Create a new key	y pair	
	 Please select a key 	pair		

- Set a password.
 - Logon Password: Set the node logon password.
 - Confirm Password: Confirm your node logon password.
- 11.Set the Pod Network CIDR and Service CIDR.



You need to set these two parameters only if you choose to use an existing VPC.

The two Classless Inter-Domain Routing (CIDR) blocks cannot overlap with each other, or with your selected VPC, or with the CIDR blocks used by any other existing Kubernetes clusters in the selected VPC. Furthermore, they cannot be modified after the cluster is created. For more information, see *Plan Kubernetes CIDR blocks under VPC*.

12.Select a Kubernetes version as needed. The system displays the Kubernetes version and the Docker versions.

Docker Version	17.06.2-ce-3	
Kubernetes Version	1.11.2	1.10.4

13.Configure an SNAT gateway for the VPC.



We recommend that you configure this feature. Otherwise, instances in the VPC cannot access the Internet, which will result in a cluster creation failure.

- If you set the system to automatically create a VPC, you must configure an SNAT gateway for the VPC.
- If you set the system to use an existing VPC, you can perform either of the following operations:
 - Set the system to automatically configure an SNAT gateway.
 - Manually configure a NAT gateway for the VPC or manually configure an SNAT gateway for the VPC.

Configure SNAT	Configure SNAT for VPC
	If the VPC you choose does not have access to Internet, NAT gateway and EIP will be used to configure SNAT for the VPC. During this period, NAT gateway, EIP, and other resources may be created.

14.Enable the Use Public SLB to Expose API Server function.

The API server provides add, delete, edit, check, watch, and other HTTP REST interfaces for resource objects, such as pods and services.

• If you enable this function by selecting the check box, an Internet SLB is created and the Master node port (namely, port 6443) is opened. In this case, you can use kubeconfig to connect to and operate the cluster through the Internet.

Note:

The API server uses the Master node port.

• If you do not enable this function, no Internet SLB is created. In this case, you can only use kubeconfig to connect to and operate the cluster within the VPC.

Public SLB	Expose API SERVER with public SLB
	When the selection is not open, the cluster API SERVER can not be accessed out of the VPC.

15.Enable the Internet SSH logon.



To enable this function, you must enable the Use Public SLB to Expose API Server function.

- · If you enable this function, you can use SSH to access the cluster.
- If you do not enable this function, the cluster cannot be accessed by using SSH or kubectl. To access the cluster by using SSH in such a case, you must manually associate an EIP to the ECS instance, set a security group rule, and open the SSH port (port 22). For more information, see *Access a Kubernetes cluster by using SSH*.

SSH Login	Enable SSH access for Internet
	If this function is disabled, see SSH access to Kubernetes clusters to manually enable SSH access.

16.Install cloud monitoring plug-ins on your ECS node instances.

If you install cloud monitoring plug-ins on the ECS nodes, you can view the ECS instance monitoring information in the CloudMonitor console.

Monitoring Plug-in	✓ Install cloud monitoring plug-in on your ECS.
	Installing a cloud monitoring plug-in on the node allows you to view the monitoring information of the created ECS instance in the CloudMonitor console

17.Enable Log Service. You can select an existing project or create a new project.

If you select the Using SLS check box, a Log Service plugin is automatically installed in the cluster. Then, when you create an application in the cluster, you can immediately use Log Service with only a few configurations required. For more information, see *Use Log Service to collect Kubernetes cluster logs*.

Log Service	Vsing SLS				
	Select Project	Create Project			
	A SLS Project named k	8s-log-{ClusterID} will	be created automatically		

18.Set advanced configurations.

- a. Select a network plugin. Available network plugins are Flannel and Terway. For more information, see *Do I select the Terway or Flannel plugin for my Kubernetes cluster network?*.
- b. Set the number of pods for a node. This parameter specifies the maximum number of pods that can be run by a single node. We recommend that you retain the default setting.



c. Enable the Custom Cluster CA function. A CA certificate ensures secure information exchanges between the server and the client. To enable this function, select the check box to add the CA certificate to the Kubernetes cluster.

Cluster CA	Custom Cluster CA

19.Click Create in the upper-right corner.



A Kubernetes cluster that contains multiple nodes typically takes ten minutes to be created.

View the cluster deployment result

After the cluster is created, you can view the cluster in the cluster list of the Container Service console.

Container Service - Kubernetes +	Cluster List			You can	create up to 5 clu	isters and can add up to	40 nodes in each cluster.	Refresh	Create Kubernetes Cluster
Overview	Help: & Create cluster & Scale cluster	𝖉 Connect to Kuberr	etes cluster via kubectl 🔗 I	Manage applications w	ith commands				
 Clusters 	Name 🔻								
Clusters	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
Nodes Volumes	k8s-cluster	Kubernetes	China East 1 (Hangzhou)	VPC vpc-bp1lkyevdjj	Running	09/26/2018,17:41:26	1.11.2	Manage	View Logs Dashboard Scale Cluster More →

• To view the cluster logs, click View Logs on the right of the cluster. To view more details, click Stack Events.

Detailed resource deployment logs: Stack Events	
Time	Information
06/22/2018,11:13:50	c06eadb6387ec4c6080d495e243649a7b Start to DescribeK8sUserCertConfig
06/22/2018,11:03:39	c06eadb6387ec4c6080d495e243649a7b Set up k8s DNS configuration successfully
06/22/2018,11:02:30	c06eadb6387ec4c6080d495e243649a7b Stack CREATE completed successfully:o
06/22/2018,11:02:30	c06eadb6387ec4c6080d495e243649a7b Start describeStackInfo
06/22/2018,11:02:29	c06eadb6387ec4c6080d495e243649a7b Start describeStackInfo
06/22/2018,10:46:55	c06eadb6387ec4c6080d495e243649a7b Successfully to CreateStack
06/22/2018,10:46:55	c06eadb6387ec4c6080d495e243649a7b Start to wait stack ready
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to create cluster task
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to CreateK8sCluster
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to CreateStack
06/22/2018,10:46:50	c06eadb6387ec4c6080d495e243649a7b Start to validateCIDR
06/22/2018,10:46:41	c06eadb6387ec4c6080d495e243649a7b Start create cluster certificate

 \cdot To view the basic information, the connection information, and other information, return to the cluster list page and click Manage in the action column of the cluster.

Basic Information				
Cluster ID:	VPC	Running	Region: China East 1 (Hangzhou)	
Connection Information				
API Server Internet endpoint	https: :6443			
API Server Intranet endpoint	https:// :6443			
Master node SSH IP address				
Service Access Domain	n-hangzhou.ali	container.com		
Cluster resource				
ROS	k8s-for-cs			
Internet SLB	lb-			
VPC	vpc-			
NAT Gateway	ngw			
Connect to Kubernetes cluster via kubectl				
1. Download the latest kubecti client from the Kubernetes Edition page .				
2. Install and set up the kubecti client. For more information, see Installing and Setting Up kubecti				
3. Configure the cluster credentials:				
KubeConfig SSH				

Cluster information

- API Server Internet endpoint: the IP address and the port through which the Kubernetes API server provides services to the Internet. It enables you to manage the cluster by using kubectl or other tools on your terminal.
- API Server Intranet endpoint: the IP address and the port through which the Kubernetes API server provides services within the cluster. This IP address is

the SLB instance IP address. The three Master nodes on its backend provide services.

- Master node SSH IP address: You can directly log on to Master nodes by using SSH to maintain the cluster.
- Service Access Domain: provides the services in the cluster with access domain name for testing. The service access domain name has the following suffix: <

cluster_id >.< region_id >. alicontain er . com .

For example, you can log on to the Master node by using SSH, and run the kubectl get node command to view the cluster nodes.

login as: root root@1B's password:				
Welcome to Alibaba Cloud Elastic Com	pute Servi	ce !		
[root@iZbp1d7yvpa3j183u0ur11Z ~]# ku	bectl get :	node		
NAME	STATUS	ROLES	AGE	VERSION
cn-hangzhou.i	Ready	<none></none>	17m	v1.8.4
cn-hangzhou.i	Ready	master	19m	v1.8.4
cn-hangzhou.i	Ready	master	24m	v1.8.4
cn-hangzhou.i	Ready	master	22m	v1.8.4
[root@iZbp1d7yvpa3j183u0urllZ ~]#				

In this example, the system displays four nodes, which are the three Master nodes and one Worker node.

1.4.3 Configure a Kubernetes GPU cluster to support GPU scheduling

From version 1.8, Kubernetes will support hardware acceleration devices such as NVIDIA GPU, InfiniBand, and FPGA, by using *device plugins*. Furthermore, GPU solutions of Kubernetes open source communities will be deprecated in version 1.10, and removed from the master code in version 1.11.

We recommend that you use an Alibaba Cloud Kubernetes cluster combined with GPU to run highly dense computational tasks such as machine learning and image processing. With this method, you can implement one-click deployment, elastic scaling, and other functions, without needing to install NVIDIA drivers or Compute Unified Device Architecture (CUDA) beforehand.

Background information

During cluster creation, Container Service performs the following operations:

- Creates Elastic Compute Service (ECS) instances, sets the public key used for SSH logon from the management node to other nodes, and installs and configures the Kubernetes cluster by using CloudInit.
- Creates a security group to allow inbound access to all ICMP ports in a VPC.
- Creates a new VPC and VSwitch if you do not use the existing VPC, and also creates an SNAT entry for the VSwitch.
- Creates VPC routing rules.
- Creates a NAT gateway and Elastic IP (EIP).
- Creates a Resource Access Management (RAM) user and AccessKey (AK). This RAM user has the permissions to query, create, and delete ECS instances, add and delete cloud disks, and all relevant access permissions for Server Load Balancer (SLB) instances, CloudMonitor, VPC, Log Service, and Network Attached Storage (NAS) services. The Kubernetes cluster dynamically creates the SLB instances, cloud disks, and VPC routing rules according to your configurations.
- · Creates an intranet SLB instance and exposes port 6443.
- Creates an Internet SLB instance and exposes ports 6443, 8443, and 22. (If you enable the SSH logon for Internet access when creating the cluster, port 22 is exposed. Otherwise, port 22 is not exposed.)

Prerequisites

You have activated Container Service, Resource Orchestration Service (ROS), and RAM.

You have logged on to the *Container Service console*, *ROS console*, and *RAM console* to activate the corresponding services.

Note:

The deployment of Container Service Kubernetes clusters depends on the application deployment capabilities of Alibaba Cloud ROS. Therefore, you need to activate ROS before creating a Kubernetes cluster.

Limits

- The SLB instance created with the Kubernetes cluster only supports the Pay-As-You -Go billing method.
- The Kubernetes cluster supports only Virtual Private Cloud (VPC).

- By default, each account has a specified quota of the number of cloud resources that it can create. If the number of cloud resources has reached the quota limit, the account cannot create a cluster. Make sure you have sufficient resource quota to create a cluster. You can open a ticket to increase your quota.
 - By default, each account can create up to 5 clusters across all regions and add up to 40 nodes to each cluster. You can open a ticket to create more clusters or nodes.
 - By default, each account can create up to 100 security groups.
 - By default, each account can create up to 60 Pay-As-You-Go SLB instances.
 - By default, each account can create up to 20 EIPs.
- The limits for ECS instances are as follows:
 - Only the CentOS operating system is supported.
 - Only Pay-As-You-Go ECS instances can be created.

Note:

After creating an instance, you can *Switch from Pay-As-You-Go to Subscription billing* in the ECS console.

Create a GN5 Kubernetes cluster

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. Click Create Kubernetes Cluster in the upper-right corner.

Container Service - Kubernetes 🕶		Cluster List		You	can create up to 5 cluster	and can add up	to 40 nodes in each cluster.	Refresh	Create Kubernetes Cluster
Overview	*	Help: & Create cluster & Create GPU clus creation failures & Authorization manageme	ters 🔗 Scale cluster nt	🔗 Connect to Ki	ubernetes cluster via kube	cti 🔗 Manage	applications with commands	🔗 Cluster plannin	g & Trenoot cluster
Clusters	l	Name T							
Nodes	l	Cluster Name/ID Cluster Type Regi	n Network Type	Cluster Status	Number of Nodes	Time Created	Kubernetes Version		Action

By default, the Create Kubernetes Cluster page is displayed.



Worker nodes are set to use GPU ECS instances to create a GPU cluster. For information about other parameter settings, see *Create a Kubernetes cluster*.



- 4. Set the Worker nodes. In this example, the gn5 GPU instance type is selected to set Worker nodes as GPU working nodes.
 - a. If you choose to create Worker instances, you must select the instance type and the number of Worker nodes. In this example, two GPU nodes are created.

WORKER Configuration								
Instance Type	4 Core(s) 30 G (ecs.gn5-c4g1.xlarge)		▼	Ŧ	Quantity	2	unit(s) 🌲	
System Disk	SSD Disk 👻	40	GiB 🌲					
Attach Data Disk								

- b. If you choose to add existing instances, you need to have already created GPU cloud servers in the same region where the cluster is to be created.
- 5. After you have completed all required settings, click Create to start cluster deployment.
- 6. After the cluster is created, choose Clusters > Nodes in the left-side navigation pane.
- 7. To view the GPU devices mounted to either of the created nodes, select the created cluster from the clusters drop-down list, select one of the created Worker nodes, and choose More > Details in the action column.

Create a GPU experimental environment to run TensorFLow

Jupyter is a popular tool used by data scientists for the experimental environment TensorFlow. This topic describes an example of how to deploy a Jupyter application.

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Applications > Deployments.

- 3. Click Create by Template in the upper-right corner.
- 4. Select the target cluster and namespace and then select a sample template or the custom template from the resource type drop-down list. After you orchestrate your template, click DEPLOY.



In this example, a Jupyter application template is orchestrated. The template includes a deployment and a service.

```
# Define
                  tensorflow
                                deployment
            the
apiVersion : apps / v1
kind : Deployment
metadata :
   name : tf - notebook
   labels :
     app : tf - notebook
spec :
   replicas :
               1
   selector : # define
                                 the
                                       deployment
                                                     finds
                                                             the
                           how
pods
       it
            manages
     matchLabel s :
   app : tf - notebook
template : # define the
                                        specificat ions
                                 pods
     metadata :
       labels :
         app : tf - notebook
     spec :
       containers :
         name : tf - notebook
         image : tensorflow / tensorflow : 1 . 4 . 1 - gpu - py3
         resources :
           limits :
```

```
nvidia . com / gpu : 1
                                                          # specify
  the
                      NVIDIA
        number of
                               GPUs
                                      that
                                             are
                                                   called
                                                            by
the
      applicatio
                 n
        ports :
          containerP
                      ort : 8888
                      8888
          hostPort :
        env :
            name : PASSWORD
                                                      # specify
      password used to
the
                                      the
                                            Jupyter
                                                      service .
                             access
            modify
You
      can
                     the password
                                      as
                                           needed .
            value : mypassw0rd
 Define
           the
                 tensorflow
                              service
#
apiVersion : v1
kind : Service
metadata :
  name : tf - notebook
spec :
  ports :
    port : 80
    targetPort : 8888
  name : jupyter
selector :
    app : tf - notebook
  type :
         LoadBalanc er
                                                            Alibaba
                                                    # set
  Cloud
                                                           that
          SLB service
                          for
                              the
                                      applicatio n so
                                    from
its
      services
                 are
                       accessible
                                           the
                                                 Internet .
```

If you use a GPU deployment solution of Kubernetes earlier than 1.9.3, you must define the following volumes in which the NVIDIA drivers reside:

```
volumes :
    - hostPath :
    path : / usr / lib / nvidia - 375 / bin
    name : bin
    hostPath :
    path : / usr / lib / nvidia - 375
    name : lib
```

When you orchestrate your deployment template in a cluster by using the GPU deployment solution of Kubernetes earlier than 1.9.3, your template must be highly dependent on the cluster. As a result, portability of the template is not achievable. However, in Kubernetes version 1.9.3 and later, you do not need to specify these hostPaths because the NIVEA plugins automatically discover the library links and execution files required by the drivers.

5. In the left-side navigation pane under Container Service-Kubernetes, choose Discovery and Load Balancing > Services. Then, select the target cluster and namespace, and then view the external endpoint of the tf-notebook service.

	Container Service - Kubernetes →	Service								Refresh	Create
	Overview	🖉 Canary re	lease								
×	Clusters	Clusters k8	s-test 🔻 M	Namespaces def	fault 🔹 🙎					Search By Name	Q
•	Applications	Name	Label	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint			Action
•	Service Mesh	kubernetes	component:apiserver provider:kubernetes	ClusterIP	05/17/2019,18:12:33		kubernetes:443 TCP	-	Details	Update View YAML	Delete
~	Discovery and Load B Services	tf-notebook		LoadBalancer	05/23/2019,10:46:02		tf-notebook:80 TCP tf-notebook:30708 TCP	3	Details	Update View YAML	Delete
	Ingresses										

6. Access the Jupyter application in a browser. The access address is http://

EXTERNAL - IP . You need to enter the password set in the template.

7. By running the following program, you can verify that this Jupyter application can use GPU, and the program is able to list all devices that can be used by Tensorflow:

```
from
           tensorflow . python . client
                                                          import
                                                                       device_lib
          get_availa ble_device s ():
 def
       local_devi ce_protos = device_lib . list_local _devices ()
       return [x.name for x in local_devi ce_protos]
 print ( get_availa ble_device s ())
File Edit View Insert Cell Kernel Widgets Help
                                                                                      Trusted
                                                                                            Kernel O
🖺 🕂 😹 🖆 🗈 🛧 🕹 🕅 Run 🔳 C 🇭 Code
                                        v
   In [2]: from tensorflow.python.client import device_lib
         def get_available_devices():
           get_available_devices():
local_device_protos = device_lib.list_local_devices()
return [x.name for x in local_device_protos]
         print(get_available_devices())
```

1.4.4 Upgrade the NVIDIA driver of a Kubernetes cluster GPU node

['/device:CPU:0', '/device:GPU:0']

This topic describes how to upgrade the NVIDIA driver of a Kubernetes cluster GPU node where services are running or no service runs.

Prerequisites

In []:

• You have created a Kubernetes GPU cluster. For more information, see *Configure a Kubernetes GPU cluster to support GPU scheduling*.

• You have connected to the Kubernetes GPU cluster by using kubectl, see *Connect to* a *Kubernetes cluster by using kubectl*.

Upgrade the NVIDIA driver of a GPU node where services are running

1. Run the following command to disable scheduling for the target GPU node:

kubectl cordon node-name

Note:

- Only the NVIDIA drivers of Worker nodes can be upgraded.
- The node-name parameter must be in the format of your-region-name.node-id.
 - your-region-name indicates the name of the region where your cluster is located.
 - node-id indicates the ID of the ECS instance where the target node is located.

You can run the following command to view node-name:

kubectl get node

2. Run the following command to migrate the pods on the target node to other nodes:

kubectl drain node-name --grace-period=120 --ignore-daemonsets=true

[root@gpu-test ~]# kubectl drain cn-hangzhou.i- --grace-period=120 --ignore-daemonsets=true node/cn-hangzhou.i- cordoned ARRUING: Ignoring DaemonSet-managed pods: flexvolume- , kube-flannel-ds- , kube-proxy-worker- , logtail-dspod/domain-nginx- evicted pod/old-nginx- evicted pod/old-nginx- evicted pod/old-nginx- evicted pod/old reginx- evicted nvidia - smi

3. Run the following command to log on to the target node:

ssh root @ xxx . xxx . x . xx

4. On the target node, run the following command to view the driver version:

[root@~]# nvidia-smi Fri Jan 18 16:44:52 2019	
NVIDIA-SMI 384.111 Driver Version: 384.111	
GPU Name Persistence-M Bus-Id Disp.A Volatile Fan Temp Perf Pwr:Usage/Cap Memory-Usage GPU-Util	Uncorr. ECC Compute M.
0 Tesla P4 0n 00000000:00:08.0 0ff N/A 24C P8 6W / 75W 0MiB / 7606MiB 0%	0 Default
· · · · · · · · · · · · · · · · · · ·	
Processes: GPU PID Type Process name	GPU Memory Usage
No running processes found	 +

5. Run the following commands to remove the existing driver:

Note:

- If the existing driver is v 384 . 111 , you can directly run the commands in this step.
- If the existing driver is not v 384 . 111 , you must download the correct driver version from NVIDIA Website before running the commands in this step.

```
cd / tmp
```

```
curl – 0 https://cn.download.nvidia.cn/tesla/384.
111 / NVIDIA – Linux – x86_64 – 384.111.run
```

```
chmod u + x NVIDIA - Linux - x86_64 - 384 . 111 . run
```

```
. / NVIDIA - Linux - x86_64 - 384 . 111 . run -- uninstall - a -
s - q
```

6. Run the following command to restart the target node:

reboot

7. Download the driver version that you want from the NVIDIA Website. This example uses v 410 . 79 .

8. Run the following command to install the downloaded NVIDIA driver in the directory where the driver is downloaded:

sh . / NVIDIA - Linux - x86_64 - 410 . 79 . run - a - s - q

9. Run the following commands to add the following settings to the NVIDIA driver:

nvidia - smi - pm 1 || true nvidia - smi - acp 0 || true

10.Run the following command to update two device plugins:

```
mv / etc / kubernetes / manifests / nvidia - device - plugin . yml
/
mv / nvidia - device - plugin . yml / etc / kubernetes /
manifests /
```

11.In any path of the Master node, run the following command to enable scheduling for the target node:

kubectl uncordon node-name

Verify the results

Run the following command on the Master node. Then check the driver version for the target GPU node. The system displays that the driver is $v \ 410$. 79, indicating the node driver has been upgraded.



You need to replace the *node-name* parameter with your target node name.

```
kubectl exec -n kube-system -t nvidia-device-plugin-node-name nvidia-
smi
```

[root@ ia-smi Mon Ja	lgpu-te in 21 0	st ~]# 3:14:4	kubect 8 2019	l exec	-n kube-syste	em -t nvid:	ia-device-p	olugin-cn-	nvid
NVID	IA-SMI	410.7	9	Driver	Version: 410	79 (CUDA Versio	on: N/A	1
GPU Fan	Name Temp	Perf	Persist Pwr:Usa	ence-M ge/Cap	Bus-Id Memo	Disp.A pry-Usage	Volatile GPU-Util	Uncorr. ECC Compute M.	
0 N/A	Tesla 21C	P4 P8	6W /	0n 75W	00000000:00: 0MiB /	08.0 Off 7611MiB	0%	0 Default	
Proc GPU	esses:	PID	Туре	Process	name			GPU Memory Usage	
No	runnin	g proc	esses fo	und					+

Upgrade the NVIDIA driver of a GPU node where no service runs

1. Run the following command to log on to the target GPU node:

ssh root@xxx . xxx . x . xx

2. On the target node, run the following command to view the driver version:

nvidia - smi

[root@ Fri Jan 18 16:44:5	, ~]# 2 2019	nvidia-smi		
NVIDIA-SMI 384.1	11	Driver Version: 384	111	
GPU Name Fan Temp Perf	Persistence-M Pwr:Usage/Cap	Bus-Id Disp.A Memory-Usage	Volatile GPU-Util	Uncorr. ECC Compute M.
 0 Tesla P4 N/A 24C P8	0n 6W / 75W	00000000:00:08.0 Off 0MiB / 7606MiB	0%	0 Default
•				
Processes: GPU PID	Type Process	name		GPU Memory Usage
No running proc	esses found			

3. Run the following commands to remove the existing driver:



• If the existing driver is not v 384 . 111 , you must download the correct driver version from NVIDIA Website before running the commands in this step.

cd / tmp curl - 0 https :// cn . download . nvidia . cn / tesla / 384 . 111 / NVIDIA - Linux - x86_64 - 384 . 111 . run chmod u + x NVIDIA - Linux - x86_64 - 384 . 111 . run . / NVIDIA - Linux - x86_64 - 384 . 111 . run -- uninstall - a s - q

4. Run the following command to restart the target node:

reboot

- 5. Download the driver version that you want from the NVIDIA Website. This example uses v 410 . 79 .
- 6. Run the following command to install the downloaded NVIDIA driver in the directory where the driver is downloaded:

sh . / NVIDIA - Linux - x86_64 - 410 . 79 . run - a - s - q

7. Run the following commands to add the following settings to the NVIDIA driver:

nvidia - smi - pm 1 || true nvidia - smi - acp 0 || true

Verify the results

Run the following command on the Master node. Then check the driver version for the target GPU node. The system displays that the driver is $v 410 \cdot 79$, indicating the node driver has been upgraded.



You need to replace the *node-name* parameter with your target node name.

kubectl exec -n kube-system -t nvidia-device-plugin-node-name nvidiasmi

[root@gpu-test ~]# kubect ia-smi Mon Jan 21 03:14:48 2019	tl exec -n kube-system -t	nvidia-device-plugin-cn-	nvid
NVIDIA-SMI 410.79	Driver Version: 410.79	CUDA Version: N/A	+
GPU Name Persist Fan Temp Perf Pwr:Usa	tence-M Bus-Id Di age/Cap Memory-U	sp.A Volatile Uncorr. ECC sage GPU-Util Compute M.	
0 Tesla P4 N/A 21C P8 6W /	On 00000000:00:08.0 / 75W 0MiB / 761	0ff 0 1MiB 0% Default	
+			•
Processes: GPU PID Type 	Process name	GPU Memory Usage	
No running processes fo	ound		

1.4.5 Create a multi-zone Kubernetes cluster

You can create a multi-zone Kubernetes cluster to guarantee high availability.

Prerequisites

 You have activated the following services: Container Service, Resource Orchestration Service (ROS), Resource Access Management (RAM), and Auto Scaling service.

Log on to the *Container Service console*, *ROS console*, *RAM console*, and *Auto Scaling console* to activate the corresponding services.



The deployment of Container Service Kubernetes clusters depends on the application deployment capabilities of Alibaba Cloud ROS. Therefore, activate ROS before creating a Kubernetes cluster.

- You must create a Virtual Private Cloud (VPC) and create at least three VSwitches in the VPC. To achieve high availability, we recommend that you create VSwitches in different availability zones.
- You need to manually configure SNAT for each VSwitch in the VPC. Otherwise, instances in the VPC cannot access the Internet normally.

Context

You can create Kubernetes clusters with ECS instances in different availability zones by using the Container Service console to achieve high availability.

Context

During cluster creation, Container Service performs the following operations:

- Create Elastic Compute Service (ECS) instances and configure to log on to other nodes from management nodes with the SSH public key. Install and configure the Kubernetes cluster by using CloudInit.
- Create a security group. This security group allows the VPC inbound access of all the ICMP ports.
- Create a RAM user and an AccessKey. The RAM user has the permissions for querying, creating, and deleting ECS instances, the permissions for adding and deleting cloud disks, and all permissions for the operations on Server Load Balancer (SLB), CloudMonitor, VPC, Log Service, and Network Attached Storage (NAS). The Kubernetes cluster dynamically creates SLB instances, cloud disks, and VPC routing rules according to your configurations.
- · Create an intranet SLB instance and expose the port 6443.
- Create an Internet SLB instance and expose the port 6443. (If you enable the SSH logon for Internet when creating the cluster, port 22 is exposed. Otherwise, port 22 is not exposed.)

Limits

- The SLB instances created with the cluster support only the Pay-As-You-Go billing method.
- Kubernetes clusters support only the Virtual Private Cloud (VPC) network type.
- By default, each account has a specified quota for the cloud resources it can create. If the number of cloud resources exceeds the quota, the account cannot create a cluster. Make sure you have enough quota before creating a cluster. To increase your quota, open a ticket.
 - By default, each account can create up to 5 clusters in all regions and add up to 40 nodes to each cluster. To create more clusters or nodes, open a ticket.
 - By default, each account can create up to 100 security groups.
 - By default, each account can create up to 60 Pay-As-You-Go SLB instances.
- The limits for ECS instances are as follows:
 - Only the CentOS operating system is supported.
 - The Pay-As-You-Go and Subscription ECS instances can be created.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane. Click Create Kubernetes Cluster in the upper-right corner.
- 3. On the Create Kubernetes Cluster page, click Multi-AZ Kubernetes.
- 4. Enter the cluster name.

The cluster name can be 1–63 characters long and contain numbers, Chinese characters, English letters, and hyphens (-).

- 5. Select the region where the cluster is located.
- 6. Select a VPC.

Select a VPC from the existing VPC drop-down list and select three VSwitches under the VPC. To achieve high availability, we recommend that you select the VSwitches located in different zones.

VPC	VPC123	(vpc-2zercq4p)	vanzxsfiiclyl) 🔻		
VSwitch	Select thre	ee VSwitches. T	o ensure high availability, switches	in different zones are recommended.	
	N	Name	ID	Zone	CIDR
	V V	/s-01-k8s	vsw-2zey8crxysjn3pdrbykcl	China North 2 (Beijing) ZoneB	192.168.35.0/24
	🔽 t	est	vsw-2zeva6cj8lotckx1b9fc1	China North 2 (Beijing) ZoneC	192.168.18.0/24
	 V 	/Switch123	vsw-2zeydhl5uwh1ej522lauo	China North 2 (Beijing) ZoneA	192.168.0.0/24

- 7. Configure the Master nodes and Worker nodes.
 - a) Select a node payment type from Pay-As-You-Go and Subscription.
 - b) Select instance types of the Master nodes and Worker nodes, and set the number of Worker nodes.

Note:

- Currently, only the CentOS operating system is supported.
- · Currently, only three Master nodes can be created.
- Each cluster can contain up to 37 Worker nodes. To create more nodes, open a ticket.
- System disks are attached to Master nodes and Worker nodes by default. Available system disks include SSD Cloud Disks and Ultra Cloud Disks.

• You can also manually attach a data disk to the Worker node. The data disk can be an Ultra Cloud Disk or an SSD Cloud Disk.

Node Type	Pay-As-You-Go	Subscription		
Master Configuration				
Instance Type	Zone	Туре		Quantity
	Zone E	2 Core(s) 4 G (ecs.sn1ne.large)	• 7	1 unit(s)
	Zone D	2 Core(s) 4 G (ecs.n1.medium)	• 7	1 unit(s)
	Zone A	4 Core(s) 16 G (ecs.i1.xlarge)	▼ Ţ	1 unit(s)
System Disk	Ultra Cloud Disk	▼ 120 GiB ♣		
Worker Configuration				
Instance Type	Zone	Туре		Quantity
	Zone E	2 Core(s) 4 G (ecs.sn1ne.large)	▼ Ţ	1 unit(s) 🌲
	Zone D	2 Core(s) 4 G (ecs.n1.medium)	▼ Ţ	1 unit(s) 🌲
	Zone A	4 Core(s) 16 G (ecs.i1.xlarge)	▼ Ţ	1 unit(s) 🚔
System Disk	Ultra Cloud Disk	▼ 40 GiB ♣		
🗸 Attach Data Disk	Ultra Cloud Disk	▼ 100 GiB 🚔		

- 8. Configure the logon mode.
 - $\cdot\,$ Set the key pair.

When creating a cluster, select the key pair logon mode and click Create a new key pair. In the ECS console, create a key pair. For details, see *Create an SSH key*

pair. After the key pair is created, set the key pair as the credentials for logging on to the cluster.

Login	Key Pair	Password		
Key Pair Name			T	C
	You can visit ECS conso	ole to <mark>Create a new ke</mark>	y pair	~
	 Please select a key 	pair		

- Set the password.
 - Logon Password: Set the node logon password.
 - Confirm Password: Confirm your node logon password.
- 9. Specify the Pod Network CIDR and Service CIDR parameters.

Both of them cannot overlap with the Classless Inter-Domain Routing (CIDR) block used by VPC and the existing Kubernetes clusters in VPC. The values cannot be modified after the cluster is created. In addition, service CIDR cannot overlap with pod network CIDR. For more information about how to plan Kubernetes CIDR blocks, see *Plan Kubernetes CIDR blocks under VPC*.

10.Select whether to enable Use Public SLB to Expose API Server.

API server provides add, delete, edit, check, watch, and other HTTP Rest interfaces for a variety of resource objects (such as pods and services).

- a. If you select to enable this option, the Internet SLB is created and the port 6443 of the Master nodes is exposed. The port corresponds to the API server. Then you can use kubeconfig to connect to and operate the clusters through the Internet.
- b. If you select not to enable this option, the Internet SLB is not created. You can only use kubeconfig to connect to and operate the clusters inside the VPC.



11.Select whether to enable SSH logon for Internet.



To enable SSH access for Internet, you must select Use Public SLB to Expose API Server.

- If you select to enable SSH access for Internet, you can use SSH to access a cluster.
- If you select not to enable SSH access for Internet, you cannot access a cluster by using SSH or connect to a cluster by using kubectl. To access a cluster instance by using SSH, manually bind an EIP to the ECS instance, configure security group rules, and open the SSH port (22). For details, see *Access Kubernetes clusters by using SSH*.

SSH Login	Enable SSH access for Internet
	If you choose not to open it, please refer to SSH access to Kubernetes cluster to manually enable SSH access.

12.Select whether to install a cloud monitoring plug-in on your ECS.

You can install a cloud monitoring plug-in on the ECS node to view the monitoring information of the created ECS instances in the CloudMonitor console.



13.Select whether to use Log Service. You can select an existing project or create a project.

If you select Using SLS, the Log Service plug-in is automatically configured in the cluster. When creating an application, you can quickly use Log Service with a simple configuration. For details, see *Use Log Service to collect Kubernetes cluster logs*.

Log Service	Vsing SLS	
	Select Project	Create Project
	A SLS Project named k	8s-log-{ClusterID} will

14.Select whether to show advance config.

- a. Select a network plug-in. Available network plug-ins are Flannel and Terway. For details, see *Do I select the Terway or Flannel plugin for my Kubernetes cluster network?*.
 - Flannel: a simple and stable community Flannel CNI plug-in. It provides only a few simple features. For example, it does not support the Kubernetes Network Policy.
 - Terway: a network plug-in developed by Alibaba Cloud Container service. It can allocate Alibaba Cloud Elastic Network Interfaces (ENIs) to containers. It can also define the access policies between containers according to the Kubernetes Network Policy . In addition, it supports bandwidth limiting for individual containers.
- b. Set the number of pods for a node, that is, the maximum number of pods that can be run by a single node.



c. Select whether to use Custom Image. The ECS instance installs the default CentOS version if no custom image is selected.

Currently, you can only select an image based on CentOS custom version to quickly deploy the environment you need.

d. Sets whether to use Custom Cluster CA. If this option is selected, the CA certificate can be added to the Kubernetes cluster, which enhances the security of information exchange between the server and client.

CA Custom Cluster CA

15.Click Create, confirm the Internet access for VPC in the displayed dialog box, and click OK to start the deployment.

nfirm Cluster Configuration		
Item	Status	Detail
Account status check	Success	
Products Activation Status	Success	
Product Quota Check	Success	
Internet access for VPC	Confir	Ensure that the NAT gateway has been configured for the VPC, or configure SNA Check Agai T manually. Otherwise, instances in the VPC can not access the Internet the clust er will fail.
		OK Cancel



A multi-node Kubernetes cluster typically takes 10 minutes to be created.

Result

View cluster deployment results.

After the cluster is successfully created, you can view the cluster in the Cluster List of the Container Service console.

Cluster List		You c	an create up to 5 c	lusters and car	add up to 40 nodes in e	ach cluster.	Refresh	reate Kubernetes Cluster
Help: 🔗 Create cluster	& Scale cluster & Connect to Kuber	netes cluster via kubec	ti 🔗 Manage app	plications with o	commands			
Name 🔻								
Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
k8s-cluster		China East 1 (Hangzhou)	VPC vpc- bp1lkyevdjj	Running	09/26/2018,17:41:26	1.11.2	М	anage View Logs Dashboard Scale Cluster More ↓

What's next
• Click View Logs at the right of the cluster to view the cluster logs. To view more detailed information, click Stack Events.

Detailed resource deployment logs: Stack Events	
Time	Information
06/22/2018,11:13:50	c06eadb6387ec4c6080d495e243649a7b Start to DescribeK8sUserCertConfig
06/22/2018,11:03:39	c06eadb6387ec4c6080d495e243649a7b Set up k8s DNS configuration successfully
06/22/2018,11:02:30	c06eadb6387ec4c6080d495e243649a7b Stack CREATE completed successfully:o
06/22/2018,11:02:30	c06eadb6387ec4c6080d495e243649a7b Start describeStackInfo
06/22/2018,11:02:29	c06eadb6387ec4c6080d495e243649a7b Start describeStackInfo
06/22/2018,10:46:55	c06eadb6387ec4c6080d495e243649a7b Successfully to CreateStack
06/22/2018,10:46:55	c06eadb6387ec4c6080d495e243649a7b Start to wait stack ready
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to create cluster task
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to CreateK8sCluster
06/22/2018,10:46:53	c06eadb6387ec4c6080d495e243649a7b Start to CreateStack
06/22/2018,10:46:50	c06eadb6387ec4c6080d495e243649a7b Start to validateCIDR
06/22/2018,10:46:41	c06eadb6387ec4c6080d495e243649a7b Start create cluster certificate

• You can also click Manage on the right of the cluster to view the basic information and connection information about this cluster.

Basic Information					
Cluster ID: VPC Region: China East 1 (Hangzhou)					
Connection Information					
API Server Internet endpoint	No. 077 - 0794 - 84				
API Server Intranet endpoint	H22-111-80-10-041				
Master node SSH IP address	ar node SSH IP address				
Service Access Domain					

In the Cluster Info section:

- API Server Internet endpoint: The IP address and port through which the Kubernetes API server provides services for the Internet. It enables you to manage the cluster by using kubectl or other tools on your terminal.
- API Server Intranet endpoint: The IP address and port through which the Kubernetes API server provides services inside the cluster. This IP address is the

address of the SLB instance, and three Master nodes in the backend provide the services.

- Master node SSH IP address: You can directly log on to the Master nodes by using SSH to perform routine maintenance for the cluster.
- Service Access Domain: Provides the services in the cluster with access domain name for testing. The service access domain name suffix is < cluster_id >.
 region_id >. alicontain er . com .

1.4.6 Connect to a Kubernetes cluster by using kubectl

To connect to a Kubernetes cluster from a client computer, use the Kubernetes command line client *kubectl*.

Procedure

- 1. Download the latest kubectl client from the Kubernetes release page.
- 2. Install and set the kubectl client.

For more information, see Install and set kubectl.

3. Configure the cluster credentials.

You can use the scp command to safely copy the master node configurations from the / etc / kubernetes / kube . conf file on the master virtual machine of the Kubernetes cluster to the \$ HOME /. kube / config file (where the kubectl expected credentials reside) of the local computer.

• If you select Password in the Login field when creating the cluster, copy the kubectl configuration file in the following method:

• If you select Key Pair in the Login field when creating the cluster, copy the kubectl configuration file in the following method:

```
machine ] root @:/ etc / kubernetes / kube . conf  \ HOME /. kube / config
```

You can check the cluster master – public – ip on the cluster information page.

- a) Log on to the Container Service console.
- b) Under Kubernetes, click Clusters in the left-side navigation pane.
- c) Click Manage at the right of the cluster.

In the Connection Information section, view the Master node SSH IP address.

1.4.7 Use kubectl on Cloud Shell to manage a Kubernetes cluster

This topic describes how to use kubectl on Cloud Shell to manage a Kubernetes cluster after you log on to the console of Alibaba Cloud Container Service for Kubernetes.

Prerequisites

You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Context

If you want to use kubectl to manage a Kubernetes cluster of Container Servcie, you can download kubectl to your local host. For more information, see *Connect to a Kubernetes cluster by using kubectl*. Additionally, you can also start Cloud Shell on the console of Container Service for Kubernetes, and then use kubectl on Cloud Shell to manage a Kubernetes cluster.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.

3. In the Action column of the target cluster, choose More > cos.cls.cloudshell.

		_									
Containe Kuber	rnetes +		Cluster List			You can create up	to 100 cluste	rs and can add i	up to 1000 nodes in each	cluster. Refresh	Create Kubernetes Cluster
Overvier	w	*	Help: Ø Create cluster Ø Create (failures Ø Authorization manageme	GPU clusters 🔗 Scale clus int	ter 🔗 Connect to K	ubernetes cluster via	a kubecti 🔗	Manage applicat	ions with commands 🛭 🔗	Cluster planning	
 Clusters Cluster 	3		Name 🔻								
Nodes	-		Cluster Name/ID	Cluster Type	Region (All) +	Network Type	Cluster Status	Number of Nodes	Time Created	Kubernetes Version	Action
Volume Names	es ipace		kubernetes-test	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1frndpc9b	Running	6	12/18/2018,15:52:19	1.11.5	Manage View Logs Dashboard Scale Cluster More→
 Author Applicat Deploy 	ization ion ment	8	k8s-delete-node	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1nr6ohb0c	Running	5	12/17/2018,10:26:28	1.11.5	Manage View Logs Dashboard Scale Cluster More~
Statefi. Daemo	ulSet onSet		k8s-test	Kubernetes	China North 2 (Beijing)	VPC vpc- 2zefxdn1wdl	Running	6	12/05/2018,13:27:46	1.11.5	Manage View Logs Dashboard Scale Cluster <u>More</u> +
Job CronJo	b		k8s-managed-cluster	ManagedKubernetes	China East 1 (Hangzhou)	VPC vpc- bp1kd7yn4qn	Running	3	11/01/2018,11:21:13	1.11.5	Delete Add Existing Instance Upgrade Cluster
Pods Service	2		test-mia	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1lkyevdjj	Running	7	09/17/2018,11:37:55	1.11.5	Automatic Scaling Addon Upgrade
Ingros											Deploy ISU0



Do the following:

- On the Authorization page, click OK to obtain a temporary access key that expires within one hour.
- On the Storage Space page, click Create Now or Skip as needed.
- 4. On Cloud Shell, you can use kubectl to manage a Kubernetes cluster of Container Service.



When you start Cloud Shell associated with the Kubernetes cluster, the system loads the *kubeconfig* file of the cluster onto Cloud Shell. Then you can use kubectl to manage your cluster.

└ ◎ ? >				
Requesting a Cloud Shell <i>Succe</i>	eded.			
Connecting terminal				
Welcome to Alibaba Cloud Shell				
Type "aliyun" to use Alibaba Clo	oud CLI			
shell@Alicloud:~\$ use-k8s-cluste	er			
Switched to context "kubernetes-	-admin-			······································
Type "kubectl" to manage your ku	lbenetes	cluster		
shell@Alicloud:~\$ kubectl get po	bd			
NAME	READY	STATUS	RESTARTS	AGE
nginx-dynamic-5b4bdb64c4-gxqs5	1/1	Running	0	21h
web-0	1/1	Running	0	4h
web-1	1/1	Running	0	4h
shell@Alicloud:~\$				

1.4.8 Use a ServiceAccount token to access a managed Kubernetes cluster

This topic describes how to use a ServiceAccount token to access a managed Kubernetes cluster.

Context

- You have created a managed Kubernetes cluster. For more information, see *Create a managed Kubernetes cluster*.
- You have connected to the managed Kubernetes cluster by using kubectl, see Connect to a Kubernetes cluster by using kubectl.

Procedure

1. Run the following command to obtain the API server intranet endpoint:

\$	kubectl	get	endpoints	kubernetes				
ubu	ntu-mia@ub	untumia	a-VirtualBox	:~\$ kubectl	get	endpoints	kubernetes	
NAM	E	ENDPOI	NTS	AGE				
kub	ernetes		:6443	13d				

2. Create a file named kubernetes - public - service . yaml and set the ip

parameter to the intranet endpoint obtained in step 1.

```
kind : Service
apiVersion : v1
metadata :
  name : kubernetes - public
spec :
 type :
         LoadBalanc er
 ports :
 - name : https
   port : 443
   protocol : TCP
   targetPort : 6443
apiVersion : v1
kind : Endpoints
metadata :
 name : kubernetes - public
 namespace : default
subsets :
addresses :
 - ip : < API
               Service
                         address > # Set
                                           this
                                                  parameter
                                                             to
       intranet endpoint obtained in
                                           step
 the
                                                  1.
 ports :
   name : https
   port : 6443
   protocol : TCP
```

3. Run the following command to deploy the API server Internet endpoint:

\$ kubectl apply - f kubernetes - public - service . yaml

- 4. Run the following command to obtain the Internet SLB address, namely, EXTERNAL
 - IP :

\$ kubectl get service name

Note:

The name parameter in the command and the name parameter in the *kubernetes* - *public* - *service* . *yaml* file of step 2 must be set to the same value. In this example, this parameter is set to *kubernetes* - *public* .

ubuntu-mia@ubuntumi	a-VirtualBox:~\$	kubectl get	service kubernete	s-public	
NAME	TYPE	CLUSTER-IP	EXTERNAL-IP	PORT(S)	AGE
kubernetes-public	LoadBalancer			443: /TCP	7d

5. Run the following command to view the corresponding secret of the ServiceAccount (in this example, the *namespace* parameter is set to default):

services de la champre, die namespace parameter is serve de la

```
ubuntu-mia@ubuntumia-VirtualBox:~$ kubectl get secret --namespace=default
NAME
                           TYPE
                                                                   DATA
                                                                           AGE
aliyun-acr-credential-a
                           kubernetes.io/dockerconfigjson
                                                                   1
                                                                           13d
                           kubernetes.io/dockerconfigjson
aliyun-acr-credential-b
                                                                   1
                                                                           13d
                           kubernetes.io/service-account-token
                                                                   3
                                                                           13d
```

6. Run the following command to obtain a token value:

\$ kubectl get secret --namespace=namespace

```
$ kubectl get secret -n --namespace=namespace -o
jsonpath={.data.token} | base64 -d
```



The *namespace* parameter in this command and the *namespace* parameter in step 5 must be set to the same value.

7. Run the following command to access the managed Kubernetes cluster:

```
$ curl -k -H 'Authorization: Bearer token' https://service-ip
```

Note:

- The value of token is the token value obtained in step 6.
- The value of service ip is the Internet SLB address obtained in step 4, that is, EXTERNAL IP.

Result

After you run the command, the following message is displayed, indicating that you have connected to the cluster.

ubuntu-mia@ubuntumia-VirtualBox:~\$ curl -k -H 'Authorization: Bearer	
	AND A DESCRIPTION OF THE OWNER OF
	CONTRACTOR OF THE OWNER.
https://	Contraction of the state of the
{	
}, "status": "Failure", "message": "forbidden: User \"system:serviceaccount:default:default\" cannot get path \"/\"", "reason": "Forbidden", "details": {	
}, "code": 403 }ubuntu-mia@ubuntumia-VirtualBox:~\$ ^C	

1.4.9 Access a Kubernetes cluster by using SSH

If you select not to enable SSH access for Internet when creating the Kubernetes cluster, you cannot access the Kubernetes cluster by using SSH or connect to the Kubernetes cluster by using kubectl. To access the cluster by using SSH after creating the cluster, manually bind Elastic IP (EIP) to the Elastic Compute Service (ECS) instance, configure security group rules, and open the SSH port (22).

Procedure

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Clusters in the left-side navigation pane.
- 3. Click Manage at the right of the cluster.
- 4. In Cluster Resource, click the ID of the Internet SLB. Then, you are redirected to the Instance Details page of your Internet Server Load Balancer instance.

Cluster:k8s-cluster				
Basic Information				
Cluster ID:	VPC	Running	Region: China East 1 (Hangzhou)	
Connection Information				
API Server Internet endpoint	The Property of the second sec			
API Server Intranet endpoint	The residue and the residue an			
Master node SSH IP address	225.0.0			
Service Access Domain	The second se			
Cluster Resource				
ROS	0.0.1.010.00000000000000000000000000000			
Internet SLB	1.			
VPC	a survey as have a			
NAT Gateway	the second se			

5. Select Instances > Server Load Balancer, and click Add Listener.

6. Add the SSH listening rule.

- a. Front-end Protocol [Port]: Select TCP and enter 22.
- b. Backend Protocol [Port]: Enter 22.
- c. Turn on the Use Server Group switch and select VServer Group.
- d. Server Group ID: Select sshVirtualGroup.
- e. Click Next and then click Confirm to create the listener.

Front-end Protocol [Port]:*	TCP : 22 Port range is 1-65535.
Backend Protocol [Port]:*	TCP : 22 Port range is 1-65535. -
Peak Bandwidth:	No Limits Configure Instances charged by traffic are not limited by peak bandwidth. Peak bandwidth range is 1-5000.
Scheduling Algorithm:	Weighted F 🔻
Use Server Group: 🕜	
Server Group Type:	VServer Group Aster-Slave Server Group
Server Group ID:	sshVirtualGn 🔻
Automatically Enable Listener After Creation:	Enable
Show Advanced Options	
	Next Cancel

7. Then, you can use the Server Load Balancer instance IP address to access your cluster by using SSH.

Basic Information	^
Server Load Balancer ID: 8-1udbig:5012803hgaub	Status: OR Running
Server Load Balancer Name: King and Soliton	Region: China East 1 (Hangzhou)
Instance IP Type: Public IP	Zone: cn-hangzhou-b(Master)/cn-hangzhou-d(Slave)
Network Type: Classic Network	
Billing Information	Billing Details Release ^
Billing Method: Pay by Traffic	Created At: 2018-01-24 11:13:01
Instance IP Address: 1142 100 (Public IP)	Automatic Release Time: -

1.4.10 Set the SSH key pair logon mode for a Kubernetes cluster

This topic describes how to set the SSH key pair logon mode for a Kubernetes cluster, and how to log on to the cluster by using an SSH key pair. Alibaba Cloud Container Service for Kubernetes provides this mode to secure SSH logon to a Kubernetes cluster.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Clusters.
- 3. In the upper-right corner, click Create Kubernetes Cluster.

Container Service - Kubernetes +	Cluster List	You can create up to 5 clusters and can add up to 40 nodes in each cluster. Refresh Create Kubernetes Cluster
Overview	Help: & Create cluster & Create GPU clusters & Scale cluster creation failures & Authorization management	\mathscr{S} Connect to Kubernetes cluster via kubect) \mathscr{S} Manage applications with commands \mathscr{S} Cluster planning \mathscr{S} Tr. 2 oot cluster
Clusters	Name •	
Nodes	Cluster Name/ID Cluster Type Region Network Type	Cluster Status Number of Nodes Time Created Kubernetes Version Action You currently do not have any clusters. You can Create Cluster
Volumes		
Namespace 🖸		

- 4. Set the key pair logon mode for the cluster, and then set other cluster parameters. For more information, see *Create a Kubernetes cluster*. Then, click Create.
 - If you have created key pairs in the Elastic Compute Service (ECS) console, select a key pair from the drop-down list.
 - If you have not created any key pairs, click Create a new key pair to create one in the ECS console. For more information, see *Create an SSH key pair*.



5. After the cluster is created, it is displayed in the cluster list. Click Manage on the right of the target cluster and record the Master node SSH IP address in the cluster information area.

Basic Information	Basic Information
	Cluster ID: cm. 2004 at de54401 as 2 bat 4477 at ditte
	Connection Information
	API Server Internet endpoint
	API Server Intranet endpoint
	Master node SSH IP address
	Service Access Domain
_	

- 6. Download the . *pem* private key file and complete the cluster logon settings according to your operating system (which can be either Windows or Linux). This topic only describes one type of the cluster logon settings in a Linux operating system as follows:
 - a) Find the directory where you have stored the downloaded .pem private key file. For example, / root / xxx . pem .
 - b) Run the following command to modify the private key file attribute: chmod

400 [the directory where the . pem private key file is stored on your local host]. In this example, chmod 400 / root / xxx . pem is run.

c) Run the following command to connect to the cluster: `ssh - i [the directory where the . pem private key file is stored on your local host] root @[Internet IP address]. In the command, the Internet IP address is the Master node SSH IP address. For example, ssh - i / root / xxx . pem root @ 10 . 10 . 10 . 100 is run in this example.

For more information about the cluster logon settings in a Window or Linux operating system, see *Connect to a Linux instance by using an SSH key pair*.

1.4.11 Create a managed Kubernetes cluster

You can create a managed Kubernetes cluster quickly and easily in the Container Service console.

Prerequisites

You have activated the following services: Container Service, Resource Orchestration Service (ROS), Resource Access Management (RAM), and Auto Scaling service.

Log on to the *Container Service console*, *ROS console*, *RAM console*, and *Auto Scaling console* to activate the corresponding services.

Note:

The deployment of Container Service managed Kubernetes clusters depends on the application deployment capabilities of Alibaba Cloud ROS. Therefore, you need to activate ROS before creating a managed Kubernetes cluster.

Context

- The SLB instances created with the cluster support only the Pay-As-You-Go billing method.
- Kubernetes clusters support only the Virtual Private Cloud (VPC) network type.
- By default, each account has a specified quota for the cloud resources it can create. If the number of cloud resources exceeds the quota, the account cannot create a cluster. Make sure you have enough quota before creating a cluster. To increase your quota, open a ticket.
 - By default, each account can create up to 100 security groups.
 - By default, each account can create up to 60 Pay-As-You-Go SLB instances.
 - By default, each account can create up to 20 EIPs.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane. The Cluster List page is displayed. Click Create Kubernetes Cluster in the upper-right corner.
- 3. On the Create Kubernetes Cluster page, click Managed Kubernetes (beta).

4. Enter the cluster name.

The cluster name can be 1–63 characters long and contain numbers, Chinese characters, English letters, and hyphens (-).

* Cluster Name	cluster-managed
	The cluster name should be 1-63 characters long, and can contain numbers, Chinese characters, English letters and hyphens.

5. Select the region and zone where the cluster is located.

Region	China North 2	China East 1	Asia Pacific SE 1	
	(Beijing)	(Hangzhou)	(Singapore)	
Zone	China North 2 Zone A	A	•	

6. Set the cluster network type.



VPC: You can select Auto Create to create a VPC together with the Kubernetes cluster, or selectUse Existing to use an existing VPC. If you select Use Existing, you can select a VPC and VSwitch from the two displayed drop-down lists.

- Auto Create: The system automatically creates a NAT Gateway for your VPC when a cluster is created.
- Use Existing: If the selected VPC has a NAT Gateway, Container Service uses the NAT Gateway. Otherwise, the system automatically creates a NAT Gateway by default. If you do not want the system to automatically create a NAT Gateway, deselect the Configure SNAT for VPC check box.

Note:

If you deselect the check box, configure the NAT Gateway on your own to implement the VPC Internet environment with secure access, or manually configure the SNAT. Otherwise, instances in the VPC cannot access the Internet normally, which leads to cluster creation failure.

v_{PC} / v_{PC} for c_{PC} 2 / d_{C} 0.01 ho4574ff68of76h5odh22c4c2	VPC	Auto Create	Use Existing						
vpc-kus-tut-cs-czuuc-suibe-ta/+ttubet/vb3eub32c+cz *		vpc-k8s-for-cs-c2ddc901be4a74ff68ef76b5edb32c4c2 🔻							
(vsw-2zeezdtsxxwf2omlsdw23) ZoneA 🔹		(vsw-2zeezdtsxxwf2omlsdw23) ZoneA 🔹							

7. Set the node type.



Pay-As-You-Go and Subscription types are supported.

Node Type	Pay-As-You-Go

8. Configure the instance.



- Currently, only the CentOS operating system is supported.
- Each cluster contains at least two nodes.
- Each cluster contains up to 48 nodes. To create more nodes, open a ticket.
- System disks are attached to the instances by default. Available system disks are Ultra Disks and SSD Disks.
- You can attach a data disk to the instances. The data disk can be an Ultra Disk or an SSD Disk.

Instance Configuration									
Instance Type	4 Core(s) 8 G (ecs.sn1ne.xlarge)				• 7	Quantity	3	unit(s)	h. 7
System Disk	Ultra Disk	•	40	GiB	*				
🗸 Attach Data Disk	Ultra Disk	•	100	GiB	*				

9. Set the logon mode.

• Set the key pair.

When creating a cluster, select the key pair logon mode and click Create a new key pair. In the ECS console, create a key pair. For details, see *Create an SSH key pair*. After the key pair is created, set the key pair as the credentials for logging on to the cluster.

- · Set the password.
 - Logon Password: Set the node logon password.
 - Confirm Password: Confirm your node logon password.

Login	Key Pair	Password	
Key Pair Name	test You can visit ECS cons	ole to <mark>Create a new ke</mark> y	r pair

10.Set the Pod Network CIDR and Service CIDR parameters.



- $\cdot~$ These two parameters are available only when you select to Use Existing VPC.
- Both Pod Network CIDR and Service CIDR cannot overlap with the Classless Inter-Domain Routing (CIDR) block used by the VPC and the existing Kubernetes clusters in the VPC. The values cannot be modified after the cluster is created. In addition, service CIDR cannot overlap with pod network CIDR. For more information about how to plan Kubernetes CIDR blocks, see *Plan*

Kubernetes CIDR blocks under VPC.

Pod Network CIDR	172.20.0.0/16
	Please fill in a valid private CIDR, namely the following CIDR and its subnets: 10.0.0.0/8, 172.16-31.0.0/12-16, 192.168.0.0/16 Cannot be duplicated with the VPC CIDR and CIDR used by Kubernetes cluster in VPC, cannot be modified after
	creation
Service CIDR	172.21.0.0/20
	Optional range: 10.0.0.0/16-24, 172.16-31.0.0/16-24, 192.168.0.0/16-24
	Cannot be duplicated with the VPC CLDR and CLDR used by Kubernetes cluster in VPC, cannot be modified after creation

11 Select whether to configure a SNAT Gateway for the VPC.



- If you select Auto Create, you must configure a SNAT Gateway.
- If you select Use Existing, you can select whether to automatically configure a SNAT Gateway. If you select not to automatically configure a SNAT Gateway, you can configure a NAT Gateway for VPC instances to securely access the Internet, or you can configure a SNAT Gateway manually. Otherwise, the instances in the VPC cannot access the Internet, and the cluster fails to be created.

Configure SNAT Configure SNAT for VPC

If the VPC you choose does not have access to Internet, NAT gateway and EIP will be used to configure SNAT for the VPC. During this period, NAT gateway, EIP, and other resources may be created.

12.Select whether to install a cloud monitoring plug-in on your ECS.

You can install a cloud monitoring plug-in on the ECS node to view the monitoring information of the created ECS instances in the CloudMonitor console.

Monitoring Plug-in Install cloud monitoring plug-in on your ECS. Installing a cloud monitoring plug-in on the node allows you to view the monitoring information of the created ECS instance in the CloudMonitor console

13.Select a network plug-in. Available network plug-ins are Flannel and Terway. For details, see *Do I select the Terway or Flannel plugin for my Kubernetes cluster network?*.

- Flannel: a simple and stable community Flannel plug-in. It provides only a few simple features. For example, it does not support the Kubernetes Network Policy
- Terway: a network plug-in developed by Alibaba Cloud Container service. It can allocate Alibaba Cloud Elastic Network Interfaces (ENIs) to containers. It can also define the access policies between containers according to the Kubernetes Network Policy . In addition, it supports bandwidth limiting for individual containers.

Network Plugin	Flannel	Terway	

14.Set the RDS whitelist.

Add the IP addresses of the ECS instances to the RDS instance whitelist.

Note:

This option is available only when you select to Use Existing VPC.



15.Click Create in the upper-right corner.



Result

After the cluster is successfully created, you can view the cluster on the Cluster List page of the Container Service console.

Container Service - Kubernetes +	Cluster List			You can create	up to 5 cluster	s and can ad	d up to 40 nodes in each	cluster. Refresh	Create Kubernetes Cluster
Overview	Help: & Create cluster & Create GPU creation failures & Authorization manag	clusters 🔗 Scale clus ement	ster 🔗 Conne	ect to Kubernetes o	luster via kube	ecti 🔗 Mani	age applications with cor	nmands 🔗 Cluster plann	ning 🔗 Troubleshoot cluster
 Clusters 	Name 🔻								
Clusters			Region (All)		Chustor	Number		Kubamatas	
Nodes	Cluster Name/ID	Cluster Type	▼	Network Type	Status	of Nodes	Time Created	Version	Action
Volumes Namespace	managed-cluster	ManagedKubernetes	China North 2 (Beijing)	VPC vpc- 2zeweju129a	Running	3	11/16/2018,11:19:26	1.11.2	Manage View Logs Dashboard Scale Cluster More↓

Click View Logs on the right of the cluster to view the cluster logs on the Cluster Logs page. To view more information, click Stack Events.

Cluster Logs: mana	aged-cluster 🗈 Back to Cluster List	fresh
Detailed resource dep	ployment logs <mark>: Stack Events</mark>	
Time	Information	
11/16/2018,11:24:54	Set up k8s DNS configuration successfully	
11/16/2018,11:24:54	start to update cluster status CREATE_COMPLETE	
11/16/2018,11:24:54	Successfully to create managed kubernetes cluster	
11/16/2018,11:22:50	Stack CREATE completed successfully:	
11/16/2018,11:20:19	Successfully to CreateStack with response &ros.CreateStackResponse{Id: ", Name: "k8s "}	-for-cs-
11/16/2018,11:20:19	Start to wait stack ready	
11/16/2018,11:20:18	Start to CreateStack	
11/16/2018,11:19:32	Successfully to startLoadBalancerListener (lb-dj17u17byskq0vne2vrbs)	

On the Cluster List page, find the created cluster and click Manage to view the basic information and connection information about this cluster.

Basic Information								
Cluster ID:	VPC	Running	Region: China North 2 (Beijing)					
-								
Cluster Info								
API Server Internet endpoint	https://www.international.org	.k8s-g1.cn-beijing.aliyuncs.con	n:6443					
Pod Network CIDR	/16							
Service CIDR	/20	/20						
Service Access Domain	All second and second	.cn-beijing.alicontainer.com						
Cluster Resource								
ROS	k8s-for-cs-							
VPC	VPC vpc							
Connect to Kubernetes cluster via kubectl								
1. Download the latest kubectl client from the Kubernet	es Edition page .							
2. Install and set up the kubectl client. For more inform	ation, see Installing and Setting Up kubec	tl						
3. Configure the cluster credentials:								

In the Cluster Info section:

- API Server Internet endpoint: The IP address and port through which the Kubernetes API server provides services for the Internet. With the API Server Internet endpoint, you can manage the cluster by using kubectl or other tools on your terminal.
- Service Access Domain: Provides the services in the cluster with access domain name for testing. The service access domain name suffix is < cluster_id >.<

region_id >. alicontain er . com .

You can see Connect to a Kubernetes cluster by using kubectl and run kubectl get node to view the node information of the cluster.

😣 🚍 💷 ubuntu-mia@ubuntumia-VirtualBox: ~							
ubuntu-mia@ubuntumia-VirtualBox:~\$ kubectl get nodes							
NAME	STATUS	ROLES	AGE	VERSION			
cn-hangzhou.i-	Ready	<none></none>	5h	v1.11.2			
cn-hangzhou.i-	Ready	<none></none>	5h	v1.11.2			
cn-hangzhou.i-	Ready	<none></none>	5h	v1.11.2			
ubuntu-mia@ubuntumia-VirtualBox:~\$							

1.4.12 Upgrade a Kubernetes cluster

This topic describes how to upgrade the Kubernetes version of your cluster in the Alibaba Cloud Container Service for Kubernetes console.

On the cluster list page, you can view the Kubernetes version of your cluster.

Name 🔻						
Cluster Name/ID	Region	Network Type	Cluster Status	Time Created	Kubernetes Version	Action
test c5def192cdbcc4ef1e515407hef154e44	China East 1 (Hangzhou)	VPC repo-lopitimili Potolimp	Running	04/24/2018,09:58:28	1.9.3	Manage View Logs Dashboard Scale Cluster More+

Limits

- The cluster node instances must be able to access the Internet so that the system can download the required software package.
- We recommend that you create a snapshot for your cluster to guarantee your data security before upgrading the cluster. For more information, see *Create a snapshot*.
- If you are upgrading a Kubernetes cluster of version 1.8.1 or 1.8.4 to version 1.
 9.3, all cluster pods will be restarted. This means that applications running on the cluster will be affected. If you are upgrading a Kubernetes cluster of another version, cluster applications will not be affected. However, if a cluster application is highly dependent on the API server, the application may be temporarily affected by the upgrade.
- When a cluster is being upgraded, the network is reset, which means OSS volumes will be remounted to the cluster. As a result, you need to re-create the pods that use OSS volumes after the upgrade.

Preparations

You must make sure that the target cluster is in the healthy status before the upgrade. To do so, follow these steps:

- 1. Log on to the Master node. For more information, see Access Kubernetes clusters by using SSH and Connect to a Kubernetes cluster by using kubectl.
- 2. Run the kubectl get cs command to verify that all cluster modules are in the healthy status.

NAME	STATUS	MESSAGE	ERROR
scheduler	Healthy	ok	
controller - manager etcd - 0 etcd - 1	Healthy Healthy Healthy Healthy	ok {" health " {" health "	: " true "} : " true "}

3. Run the kubectl get nodes command to verify that all nodes are in the

ready status.



All nodes must be in the ready status.

kubectl get NAME	nodes	STATUS	ROLES	AGE
VERSION				
cn – shanghai	. – хххх	xx Ready	master	380
v1.9.3				_
cn - shanghai	. i - xxxx	xx Ready	< none	> 38d
v1.9.3				
cn – shanghai	. i – xxxx	xx Ready	< none	> 38d
v1.9.3				
cn – shanghai	. i - xxxx	xx Ready	< none	> 38d
v1.9.3		-		
cn – shanghai	. i - xxxx	xx Ready	master	. 38d
v1.9.3		,		
cn – shanghai	. i - xxxx	xx Readv	master	. 38d
v1.9.3		,		

If a node is abnormal, you can either repair it manually or open a ticket.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Clusters > Clusters.
- 3. On the right of the target cluster, choose More > Upgrade Cluster.

Container Service - Kubernetes +	Cluster List		Ŷ	ou can create up to	100 clusters a	nd can add up	p to 1000 nodes in each	cluster.	Refresh Create Kubernetes Cluster
Overview	Help: & Create cluster & Create creation failures & Authorization ma	GPU clusters 🔗 Sca anagement	le cluster 🛛 🔗 Conn	ect to Kubernetes o	luster via kube	cti 🔗 Mana	ge applications with com	imands	& Cluster planning & Troubleshoot cluster
Clusters	Name 🔻								
Nodes	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Number of Nodes	Time Created	Versi	ion Action
Volumes Namespace	k8s-cluster	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1nejoq3u8	Running	5	02/05/2019,15:51:15		Manage View Log 2 Dashboa Cluster Expansion More -
Authorization	test01	Serverless Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1q1bze1re	Running		01/31/2019,16:49:30	1.9	Delete Add Existing Instance Upgrade Cluster
Deployment StatefulSet	k8s-test	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1jh64hx00	Running	6	01/31/2019,14:28:32	1.1	Automatic Scaling Addon Upgrade
DaemonSet Job	kubernetes-test	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1frndpc9b	Running	6	02/21/2018,20:56:15	1.1	Deploy Istio
CronJob									Use Cloud Shell Collect Kubernetes diagnostics information

4. In the displayed dialog box, click Upgrade.

1.4.13 Upgrade a system component

This topic describes how to upgrade a system component.

Prerequisites

You have created a Kubernetes cluster. For more information, see Create a Kubernetes

cluster.

Context

The following procedure is for if you need to independently upgrade one or multiple system components of a Kubernetes cluster even if the cluster is of the latest version.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. In the Action column of the target cluster, choose More > Addon Upgrade.

Container Service Kubernetes +		Cluster List			You can create up	to 100 cluste	rs and can add u	p to 1000 nodes in each	cluster. Refresh	Create Kubernetes Cluster
Overview	^	Help: & Create cluster & Create GPU failures & Authorization management	clusters 🔗 Scale clus	ter 🔗 Connect to Ku	bernetes cluster via	kubecti 🔗 I	Manage applicati	ons with commands 🔗	Cluster planning	S Troubleshoot cluster creation
 Clusters 	L	Name 🔻								
Nodes	L	Cluster Name/ID	Cluster Type	Region (All) -	Network Type	Cluster Status	Number of Nodes	Time Created	Kubernetes Version	Action
Volumes Namespace	l	kubernetes-test	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1frndpc9b	Running	6	12/18/2018,15:52:19	1.11.5	Manage View Logs Dashboard Scale Cluster More→
Authorization Application Deployment	Ш	k8s-delete-node	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1nr6ohb0c	Running	5	12/17/2018,10:26:28	1.11.5	Manage View Logs Dashboard Scale Cluster More -
StatefulSet DaemonSet		k8s-test	Kubernetes	China North 2 (Beljing)	VPC vpc- 2zefxdn1wdl	Running	6	12/05/2018,13:27:46	1.11.5	Manage View Logs Dashboard Scale Cluster <u>More</u> -
Job CronJob		k8s-managed-cluster	ManagedKubernetes	China East 1 (Hangzhou)	VPC vpc- bp1kd7yn4qn	Running	3	11/01/2018,11:21:13	1.11.5	Delete Add Existing Instance Upgrade Cluster
Pods Service Ingress		test-mia	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1lkyevdjj	Running	7	09/17/2018,11:37:55	1.11.5	Automatic Scaling Addon Upgrade

4. Select the target system component, and click Upgrade in the Action column. Upgrading is then displayed in the Status column.

Ad	don Upgrade						\times
	Component	Current Version	Upgradeable Version	Consistency Check	Action	Status	
	alicloud-application-controller	v0.1.0.1-f832bed-aliyun	v0.1.0.1-f832bed-aliyun	Success	Latest		
	alicloud-disk-controller	v1.11.2.2-a390cfb- aliyun	v1.11.2.2-a390cfb- aliyun	Success	Latest		
	Cloud Controller Manager S Readme S Version Information	v1.9.3.16-gcc144c7- aliyun	v1.9.3.16-gcc144c7- aliyun	Success	Latest		
	flexvolume	v1.11.2.5-85c062f- aliyun	v1.11.2.32-af2d48c- aliyun	Success	Upgrade		
	Nginx Ingress Controller	v0.20.0.1-4597ce2- aliyun	v0.20.0.1-4597ce2- aliyun	Success	Latest		
					Refresh	Close	è

Result

On the Addon Upgrade page, Latest is displayed in the Action column of the target system component.

Ad	don Upgrade						\times
	Component	Current Version	Upgradeable Version	Consistency Check	Action	Status	
	alicloud-application-controller	v0.1.0.1-f832bed-aliyun	v0.1.0.1-f832bed-aliyun	Success	Latest		
	alicloud-disk-controller	v1.11.2.2-a390cfb-aliyun	v1.11.2.2-a390cfb-aliyun	Success	Latest		
	Cloud Controller Manager & Readme & Version Information	v1.9.3.16-gcc144c7- aliyun	v1.9.3.16-gcc144c7- aliyun	Success	Latest		
	flexvolume	v1.11.2.32-af2d48c- aliyun	v1.11.2.32-af2d48c- aliyun	Success	Latest		
	Nginx Ingress Controller S Readme S Version Information	v0.20.0.1-4597ce2-aliyun	v0.20.0.1-4597ce2-aliyun	Success	Latest		
					Refresh	Close	

1.4.14 Update the Kubernetes cluster certificates that are about to expire

This topic describes how to update the Kubernetes cluster certificates that are about to expire through the Container Service console.

Prerequisites

You have created a Kubernetes cluster and the system has already prompted you to update the cluster certificates. For more information, see *Create a Kubernetes cluster*.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. Click Update Certificate on the right of the target cluster.



If cluster certificates are about to expire in about two months, the system displays the Update Certificate prompt for the cluster.

Cluster List		
Help: S Create cluster S Create GPU Authorization management	clusters 🔗 S	Scale cluster 🔗
Name 🔻		
Cluster Name/ID	Cluster Type	Region (All) 🔻
k8s-test	Kubernetes	China North 2 (Beijing)
test-mia	Kubernetes	China East 1 (Hangzhou)
kubernetes-test	Kubernetes	China East 1 (Hangzhou)

4. Click Update Certificate.

Update Certificate	×	<
The certificate of the current cluster will expire at 02/21/2019,20:5 cause the cluster to become unavailable. Update the certificate bef Update Certificate	i6:15. The expiration will fore the expiration date.	
	Refresh Close	

5. Click Confirm.

Note			\times
	Are you sure you want to update the cluster certificate?		
		Confirm	Cancel

Result

• The Update Certificate page displays Success.

Update Certificate		\times
cos.k8s.cluster.cert.renew.success		
	Refresh	Close

• On the Cluster List page, the Update Certificate prompt of the target cluster has been removed.

1.4.15 Scale a Kubernetes cluster

This topic describes how to scale a Kubernetes cluster (by increasing or decreasing the number of Worker nodes in the cluster) in the Alibaba Cloud Container Service for Kubernetes console.

Limits

- The number of Master nodes in any Kubernetes cluster cannot be changed.
- You can only decrease the number of Worker nodes that are added when you create or scale out the cluster. The number of Worker nodes that are added through the *Add an existing ECS instance* feature cannot be decreased.
- Worker nodes cannot be removed either through the kubectl delete command or through node removal operations in the Container Service console.
- When you scale in a cluster, Worker nodes are removed from the cluster in the sequence that they were added when you scaled out the cluster.
- You can scale in a cluster only if the cluster has more than one node that was created automatically.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Clusters > Clusters.
- 3. On the right of the target cluster, click Scale Cluster.



4. Click Scale out or Scale in and then set the number of Worker nodes.

In this example, the number of Worker nodes of the cluster is increased from 1 to 4.

Cluster Name k8s-tet Region China East 1 (Hangzhou) ZoneG Existing 1 Scale Scale out Scale in Instance Type Scale out Scale in Scaling Number 3 unit(s) \$ Number of workers after scaling: 4 Scolop Password Selet RDS Instances RDS Whitelist Selet RDS Instances Submit Submit		
Region China East 1 (Hangzhou) ZoneG Existing 1 Scale Scale out Scale in Instance Type 2 Corc(s) 4 G (ecs.n4.large) Scaling Number 3 unit(s) Number of workers after scaling: 4 Logon Password Image: Please enter the login password used when creating the cluster RDS Whitelist Select RDS Instances Submit Submit	Cluster Name	k8s-test
Existing 1 Scale Scale out Scale out Scale in Instance Type 2 Core(s) 4 G (ecs.n4.large) Scaling Number 3 unit(s) Scaling Number 3 unit(s) Number of workers after scaling: 4 * Logon Password select RDS Instances RDS Whitelist Select RDS Instances Submit	Region	China East 1 (Hangzhou) ZoneG
Scale Scale out Instance Type 2 Core(s) 4 G (ecs.n4.large) Scaling Number 3 unit(s) Scaling Number of workers after scaling: 4 Number of workers after scaling: 4 Logon Password ••••••••• Please enter the login password used when creating the cluster RDS Whitelist Select RDS Instances Submit	Existing	1
Instance Type 2 Core(s) 4 G (ecs.n4.large) Scaling Number 3 unit(s) Number of workers after scaling: 4 * Logon Password •••••••• Please enter the login password used when creating the cluster RDS Whitelist Select RDS Instances Submit	Scale	Scale out Scale in
Scaling Number 3 unit(s) Number of workers after scaling: 4 * Logon Password •••••••• Please enter the login password used when creating the cluster RDS Whitelist Select RDS Instances Submit	Instance Type	2 Core(s) 4 G (ecs.n4.large)
Number of workers after scaling: 4 * Logon Password •••••••• Please enter the login password used when creating the cluster RDS Whitelist Select RDS Instances Submit	Scaling Number	3 unit(s)
* Logon Password ••••••• RDS Whitelist Select RDS Instances Submit Submit	Number of workers after scaling:	4
RDS Whitelist Select RDS Instances	* Logon Password	Please enter the login password used when creating the cluster
Submit	RDS Whitelist	Select RDS Instances
		Submit

- 5. Enter the node logon password.
- 6. Click Submit.

In the left-side navigation pane, choose Clusters > Node to verify that the current number of Worker nodes is changed to 4.

1.4.16 Autoscale a Kubernetes cluster

This topic describes how to autoscale a Kubernetes cluster to meet the requirements of your Kubernetes cluster workload. Alibaba Cloud Container Service for Kubernetes (ACK) provides the capability to autoscale a Kubernetes cluster through using the cluster autoscaler program.

Background information

You can set the the cluster autoscaler to add different ECS instance types to your Kubernetes cluster, such as the general, GPU, and preemptive instance types. You can set multiple zones, instance specifications, and autoscaling modes.

Cluster autoscaler overview

The cluster autoscaler changes the size of a Kubernetes cluster based on the use of resource in the nodes of a pod in a Kubernetes cluster. Resource usage is calculated based on the pod resource requests. When a pod requests more resources than what the associated node can provide, the pod enters the pending status. At which time, the autoscaler calculates the change to the size of cluster. It does so by calculating the number of nodes necessary to provide the requested resource with regard to the resource specification and threshold that you set for an autoscaling group.

For example, if you set a low threshold value for the number of nodes in an autoscalin g group, the cluster autoscaler deletes a node, which reduces the amount of resources that the pod can request.

Notes

- By default, your account can use up to 30 Pay-As-You-Go ECS instances in all your clusters, and the route table of one VPC can contain up to 50 entries. To increase the number of available ECS instances or entries in a route table of one VPC, open a ticket.
- For a single type of ECS instances, the number of ECS instances of one specification that is permitted at one time varies frequently. Therefore, we recommend that you set multiple ECS instance types of one ECS specification.
- When a node for which you set the fast scaling mode is shut down and reclaimed, it is in the NotReady status. When the node is reused by the cluster autoscaler, the node enters the Ready status.
- When a node for which you set the fast scaling mode is shut down and reclaimed , only the disks attached to the node are charged (except that the node uses local disks, for example, ecs.d1ne.2xlarge).

Enable cluster autoscaling

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Clusters > Clusters.

3. Find the target cluster. Then, in the Action column, choose More > Auto Scaling.

Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
mymanagedk8s1	ManagedKubernetes	China North 2 (Beijing)	VPC vpc-2zetk0nc8jr	Running	09/07/2018,16:01:46	1.10.4	Manage	View Logs Dashboard Scale Cluster More ▼
myserverlessk8s1	Serverless Kubernetes	China East 1 (Hangzhou)	VPC vpc-bp1bw6xxicy	Running	09/04/2018,20:13:20	1.9.7	Man	age View Logs Delete
mytest1	Kubernetes	China North 2 (Beijing)	VPC vpc-2ze7c50cjcu	Running	09/03/2018,10:06:49	1.10.4	Manage	View Logs Dashboard Scale Cluster <u>More</u> •
myk8s1	Kubernetes	China East 1 (Hangzhou)	VPC vpc-bp1faadnm5o	Running	08/26/2018,11:53:17	1.10.4	Manage	Delete Add Existing Instance
								Auto-scaling Addon Upgrade Upgrade monitoring service
								Deploy Istio

Grant required permissions for the Auto Scaling service and the cluster

- · Activate the Auto Scaling service
 - 1. Click ESS in the displayed dialog box.



2. Click Activate Auto Scaling service.

	S	Auto Scaling service You have not activated Auto Scaling service
	Overview • Auto : busin	Activate Auto scaling service
Elastic Scale-In: Auto adjust elastic compute resources. Decrease the number of instances when business needs decrease to save costs. Elastic Self-Heak When an unhealthy instance has been detected, the auto-scaling service automatically replaces the instance with a new one to ensure uninterrupted service. Documentation Product Page> Quick Start> User Cade> Product Page> Product Pag	Auto Features Flasti	caling is free to use. However, resources affiliated with autoccaling incur fees.
Documentation Product Page> Quick Start> User Guide> Developer Guide>	 Elasti Elasti 	Salah- Out- man angan basis, company resources increases in instances of immarces and man panets and assummany panets (saman resources). Self-Heal: When an unhealthy instance has been detected, the auto-scaling service automatically replaces the instance with a new one to ensure uninterrupted service.
	Documenta Produ Quick User Devel EAO2	tion t Page> Sat> sate> per Guide>

3. Read and confirm that you agree to the conditions by selecting the I agree with Auto Scaling Service Agreement of Service check box, and then click Enable Now.

E	Enable Service				
	Auto Scaling Service				
sic Configuratio	Product Auto Scaling Service				
	I agree with Auto Scaling Service Agreement of Service				
	Enable Now				

4. Click Console.

Confirm Order
Order com
The service you ordered is being opened.
Please wait 1
Console
Purchase histo

5. Click Authorize.



6. Click Confirm Authorization Policy to grant ESS the permission to access your cloud resources.

Cloud Resource Access Authorization	
Note: If you need to modify role permissions, please go to the RAM Console. Role Management. If you do not configure it correctly, the following role: ESS will not be able to obtain the required permissions.	×
ESS needs your permission to access your cloud resources. Authorize ESS to use the following roles to access your cloud resources.	V
AliyunESSDefaultRole	П
Description: The ESS service will use this role to run ELS instances. Permission Description: The policy for AllyunESSDefaultRole, including the permission for ECS.	
Confirm Authorization Policy Cancel	

Verify the result

If the page automatically redirects to the Auto Scaling console, the activation is successful.

- · Add ESS authorization policies to the cluster
 - 1. Click the Worker RAM role (Kubernetes WorkerRole [xxx]) in the following dialog box.



You need to use the primary account to log on to the console before perform this operation.

Note		\times
Auto-so	caling relies on the ESS service. Before enabling auto-scaling, you need to:	
1	Enable the service and complete the default role authorization: ESS	
2	Jump to RAM to add an ESS authorization policy to the current cluster: View detailed steps KubernetesWorkerRole	
Please	confirm the above steps, otherwise the Auto-scaling will not be enabled.	
	Confirm	n

2. Click View Permissions on the right of the target authorization policy.

<	KubernetesWorkerRole-627d8853-3009-40c9-add1-3f378613c5d1			Edit Authorization Policy
Role Details				
Dale Authorization D	Authorization Policy Name	Description	Type	Actions
NOR PLATALEDUATE	k8sWorkerRolePolicy-827d8853-3009-40c9-add1-3f378613c5d1		Custom	View Permissions Revoke Authorization

3. In the upper-right corner of the page, click Modify Authorization Policy.

<	k8sWorkerRolePolicy-827d8853-3009-40c9-add1-3f378613c5d1 (Custom)				
Authorization Policy	Policy Details				
Versions	Name k8sWorkerRolePolicy-827d8853-3009-40c9-add1-3f378613c5d1 Type Custom				
References	Description				
1	1 "Version": "1", *Statement": ["Statement": [6 "Action": [7 "ecs:AttachDisk", 7 "ecs:DescribeDisks", 9 "ecs:CreateDisk", 10 "ecs:CreateDisk", 11 "ecs:CreateBapehot", 12 "ecs:CreateStampehot", 13 "ecs:CreateStampehot", 14 "ecs:AttachNetworkInterface", 15 "ecs:DetachDisk", 16 "ecs:AttachNetworkInterface", 16 "ecs:AttachNetworkInterface", 17 "ecs:DetachDisk", 18 "ecs:DetachDisk", 19 "ecs:DetachDisk", 10 "ecs:AttachNetworkInterface", 13 "ecs:DetachDisk", 14 "ecs:AttachNetworkInterface", 15 "ecs:DetachNetworkInterface", 16 "ecs:DetachNetworkInterface", 17 "ecs:DetachNetworkInterface", 18 "ecs:DetachNetworkInterface", 19 "ecs:DetachNetworkInterface", 10 "ecs:DetachNetworkInterface", 11 "ecs:DetachNetworkInterface",				

4. In the Action field of the Policy Document area, add the following policies:

```
" ess : Describe *",
" ess : CreateScal ingRule ",
" ess : ModifyScal ingGroup ",
```

```
" ess : RemoveInst ances ",
" ess : ExecuteSca lingRule ",
" ess : ModifyScal ingRule ",
" ess : DeleteScal ingRule ",
" ecs : DescribeIn stanceType s ",
" ess : DetachInst ances "
```

Note:

You must add a comma (,) to the end of the last line in the Action field before adding these policies.

5. Click Modify Authorization.

Set cluster auto scaling parameters

1. On the Automatic Scaling page, set the following parameters.

Configuration	Description
Cluster	Target cluster name.
Scale-in Threshold	The ratio of the amount of resources requested by the cluster workload to the amount of cluster resources. When the amount of resources requested by the cluster workload is less than or equal to the threshold, the system automatically scales in the cluster. The default value is 50%.
Defer Scale-in For	The number of minutes for which the system must wait to automatically scale in the cluster after the scale-in threshold is reached. The default value is 10 minutes.
Cool-Down Time	The period (in minutes) during which the system does not automatically scale in or scale out a cluster after the number of cluster nodes increases or decreases. The default is 10 minutes.

2. Click Create on the right of the target type of resource (which can be CPU or GPU) that you want to autoscale.

Auto-scaling t	Auto-scaling t Back to Cluster List							Refresh		
Overview								Modify	Disable	
Cluster ID: c68b3c	Cluster ID: c68b3c5a4273c4fba8151f06276bcc224			Cluster Name: mytest1			Scaling Status: 🔴 Inactive			
Shrinkage Thresho	ld: 50%		Shrinkage Trigger D	elay: 10Min.		Cooldown Time: 1	0Min.			
Scaling Group	Status	Instance Type	Total Instance Number	Min Instance Number	Max Instance	e Number	Pending	Removing	Action	
CPU									Create	
GPU									Create	

On the Scaling Group Configuration page, set the following parameters to create a scaling group:

Configuration	Description
Region	The region to which the scaling group is deployed . You must ensure that the scaling group and the cluster where it is located share the same region. This region cannot be modified.
Zone	The zone where the scaling group is created.
VPC	The network where the scaling group is created. You must ensure that the scaling group and the cluster where it is located are in the same region.

Set worker nodes.

Configuration	Description
Instance Type	Set the specifications of instances in the scaling group
System Disk	Set the system disks of the scaling group.
Attach Data Disk	Mount a data disk when you create a scaling group. By default, no data disk is mounted.
Instance Quantity	Set the number of instances in the scaling group.
	 Note: The number does not include the existing instances in the cluster. By default, this parameter value is 0, and the cluster adds instances to the scaling group and the Kubernetes cluster where the scaling group is located when this parameter exceeds 0.

Configuration	Description
Key Pair	Set the key pair used to log on to the node added through autoscaling. You can create a new key pair in the Elastic Compute Service (ECS) console.
RDS Whitelist	Set the Relational Database Service (RDS) instances that can be accessed by the node added through autoscaling.

3. Click OK.

Verify the result

• You can directly verify that a scaling group under CPU is displayed on the Auto Scaling page.

Overview				
Cluster ID: c68b3c5a4273c4fba8151f06276bcc224			Cluster Name: mytest1	
Shrinkage Threshold: 50%			Shrinkage Trigger Delay: 10Min.	
caling Group	Status	Instance Type	Total Instance Number	Min Instance Number
PU so-2ze9npi7u9efzddlmrzx	Active	ecs.sn1ne.xlarge	0	0
- To verify the created autoscaling component, follow these steps:
 - 1. In the left-side navigation pane, choose Application > Deployment.
 - 2. Select the target cluster and the kube-system namespace to view the created component named cluster-autoscaler.

Container Service - Kubernetes +		Deployment				Refresh	ireate by image	Gre	ate by template
Overview	l	Clusters mytest1 • Nam	nespace kube-system 🔹 🙎						
✓ Clusters		Name	Tag	PodsQuantity	Time Created				Action
Clusters Nodes	Clusters albaba-log-controller Nodes		component.version:0.1.0 component.revision:a66d1d2 component.index:0 k8s-app:ailbaba-log-controller	1/1	09/03/2018,10:15:51	De	tails Edit	Scale	Monitor More +
Volumes Namespace		alicloud-application-controller	app:alicloud-application-controller owner:aliyun	1/1	09/03/2018,10:15:50	Di	etails Edit	Scale	Monitor More +
Authorization	Ξ.	alicloud-disk-controller	app:alicloud-disk-controller	1/1	09/03/2018,10:15:50	De	etails Edit	Scale	Monitor More •
Application		alicloud-monitor-controller	task:monitoring k8s-app:alicloud-monitor-controller	1/1	09/03/2018,10:15:50	De	tails Edit	Scale	Monitor More +
Pods		cluster-autoscaler	app:cluster-autoscaler	1/1 3	09/11/2018,11:17:10	De	etails Edit	Scale	Monitor More -

Troubleshooting

- If the cluster autoscaler cannot add nodes to a pod that requested more resources, you can perform the following checks:
 - Make sure that the amount of resources provided by the ECS instances that you set for the autoscaling group is greater than the amount of resources requested by the pod.
 - Make sure that you have granted the required permissions by following the preceding steps. You must grant the required permissions for each target cluster
 - Make sure that the target Kubernetes cluster are connected to the Internet. The cluster autoscaler calls an API action from Alibaba Cloud, therefore you must ensure that the cluster nodes can be accessed through the Internet.

- If the cluster autoscaler cannot delete nodes from the autoscaling group, you can perform the following checks:
 - Make sure that the resource request threshold of pods on all nodes is not greater than that of used to scale in the cluster.
 - Make sure that no node run the pods that belong to the kube system namespace.
 - Make sure that no node runs the pod for which any constrained scheduling policies are set. This is because a constrained scheduling policy can limit a pod to a fixed node.
 - Make sure that no pod contains a PodDisrupt ionBudget object that has reached the minimum value allowed. For more information, see How do Disruption Budgets work. For more information, see *How do Disruption Budgets work*.

For more information, see *Cluster autoscaler*.

1.4.17 Delete a Kubernetes cluster

This topic describes how to delete a Kubernetes cluster in the Alibaba Cloud Container Service for Kubernetes console.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Clusters > Clusters.

Container Service - Kubernetes 👻	Cluster List			You can creat	e up to 5 cluste	ers and can add	up to 40 nodes in each o	cluster.	Refresh Create Kubernetes Cluster
Overview	Help: & Create cluster & Cr creation failures & Authorization	eate GPU clusters 🔗 S in management	icale cluster 🔗 (Connect to Kubernetes	s cluster via ku	bectl 🔗 Mana	age applications with com	imands 💰	Cluster planning 🔗 Troubleshoot cluster
Clusters	Name 🔻	Charles			Charter	North and			
Nodes	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Status	Number of Nodes	Time Created	Version	Action
Volumes	test-01 🖌	Kubernetes	China North 2 (Beijing)	VPC vpc- 2zekpsau550	Running	4	02/28/2019,14:44:09	1.11.5	3 Manage View Log 2 Dashbo
Authorization				VPC				De	lete
 Application 	k8s-cluster	Kubernetes	China East 1 (Hangzhou)	vpc- bp1nejoq3u8	Running	5	02/05/2019,15:51:15	1.1 Ad Up	d Existing Instance Igrade Cluster
Deployment	k8s-test	Kubernetes	China East 1 (Hangzhou)	VPC vpc-	Running	6	01/31/2019,14:28:32	Au 1.1 Ad	tomatic Scaling don Upgrade
DaemonSet			(Hangenoa)	bp1jh64hx00				De	ploy Istio
Job								Clu	uster Audit
CronJob								Us	e Cloud Shell
Pods									

3. On the right of the target cluster, choose More > Delete .

What's next

Troubleshoot a cluster deletion failure

If you manually add cloud resources into resources created by Resource Orchestrat ion Service (ROS), ROS does not have the permission to delete the manually added resources. For example, if you manually add a VSwitch in the Virtual Private Cloud (VPC) created by ROS, ROS cannot delete this VPC when you delete the Kubernetes cluster. As a result, the cluster deletion will fail.

To solve this issue, Container Service allows you to forcibly delete the cluster. Specifically, you can delete the cluster record and the ROS stack if the cluster fails to be deleted. However, you must release the created resources manually.

If a cluster fails to be deleted, the cluster status is displayed as follows.

Cluster List			You ca	an create up to 5 clust	ters and can add up to 4	0 nodes in each cluster.	Refresh	Create Kubernetes Cluster
Help: & Create cluster & Scale cluster	Connect t	o Kubernetes cluster via kul	becti 🔗 Manage ap	plications with comma	ands			
Name 🔻								
Cluster Name/ID	Cluster Type	Region (All) -	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
test <5%7440x39404x67428x36x268511x8	Kubernetes	China East 1 (Hangzhou)	VPC spc-lapzbrighologi	Failed to delete	05/18/2018,11:11:05	1.9.3	Manage	View Logs Dashboard Scale Cluster More -

On the right of the cluster that failed to be deleted, choose More > Delete . In the displayed dialog box, select the Force Delete check box, and then click OK.

Note:

You must manually release the resources that failed to be deleted. For information, see Failed to delete Kubernetes clusters: ROS stack cannot be deleted.

Delete Clu	ster - test			×
0	Are you sure to delete the c	luster test ?		
	Force Delete Delete the following resources	e cluster record and stack or	nly, you need t	o manually release the
	Resource ID	Resource Type	Status	Updated At
	rac-bacichofrangrydiouszt tza	ALIYUN::ECS::VPC	 Delete F ailed 	2018-05-18 13:07:2 4
				OK Cancel

1.4.18 Upgrade the Heapster components to the metrics-server component

This topic describes how to upgrade the Heapster components to the metrics-server component without upgrading the target Kubernetes cluster.

Prerequisites

- A Kubernetes cluster is created. For more information, see *Create a Kubernetes cluster*.
- The Kubernetes cluster version is earlier than v1.12.

Procedure

You must reset the data collection component, reset the monitoring data link, and then modify component compatibility settings. • Reset the data collection component.

To change the data collection component from Heapster to metrics-server, follow these steps:

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Application > Deployment.

Deployment				Refresh	Create by Image	Create by Ten	mplate
Help: & How to use private Container monitoring & B	te images 🛛 Ø Cre lue-green release	ate applications	${\mathscr S}$ Schedule a pod to the specified node ${\mathscr S}$ Create a La	yer-4 Ingress 🔗 Create a Layer-7	Ingress 🔗 Configure	e pod auto scalin	ig 🔗
Clusters k8s-cluster	 Namespace 	default	Ŧ		Search By Nam	e	Q
Name	Tag	PodsQuantity	Image	Time Created			Action
new-nginx	run:new-nginx	1/1	registry.cn-hangzhou.aliyuncs.com/xianlu/new-nginx	02/17/2019,09:26:21 D	etails Edit Sc	ale Monitor	More 🗸
nginx-deployment-basic	app:nginx	2/2	nginx:1.7.9	02/16/2019,14:57:25 D	etails Edit Sc	ale Monitor	More -
old-nginx	run:old-nginx	2/2	registry.cn-hangzhou.aliyuncs.com/xianlu/old-nginx	02/17/2019,09:25:20 D	etails Edit Sc	ale Monitor	More -
	Deployment Help: Θ How to use priva container monitoring Θ B Clusters k8s-cluster Name new-nginx nginx-deployment-basic old-nginx	Deployment Help: Ø How to use private images Ø Cre- Container monitoring Ø Blue-green release Clusters Køs-cluster • Namespace Name Tag new-nginx run:new-nginx nglinx-deployment-basic app:nglinx app:nglinx old-nginx run:old-nginx run:old-nginx	Deployment Help: Ø How to use private images Ø Create applications Container monitoring Ø Blue-green release Ø Create applications Clusters køs-duster • Namespace default Name Tag Name Tag PodsQuantity new-nginx run:new-nginx 1/1 nginx-deployment-basic app:nginx 2/2 old-nginx run:old-nginx 2/2	Deployment Help: Ø How to use private images Ø create applications Ø Schedule a pod to the specified node Ø create a La Container monitoring Clusters Køs-duster Namespace default • Name Tag PodsQuantity Image new-nginx run:new-nginx 1/1 registry.cn-hangzhou.aliyuncs.com/xianlu/new-nginx nginx-deployment-basic app:nginx 2/2 nginx:17.9 old-nginx run:lod-nginx 2/2 registry.cn-hangzhou.aliyuncs.com/xianlu/neld-nginx	Deployment Refresh Help: Ø How to use private images Ø Create applications Ø Schedule a pod to the specified node Ø Create a Layer-4 Ingress Ø Create a Layer-7 Incontainer monitoring Ø Blue-green release Ø Create a Layer-4 Ingress Ø Create a Layer-7 Incontainer monitoring Ø Blue-green release Clusters Name Tag PodsQuantity Image Time Created Name Tag PodsQuantity Image Time Created new-nginx run:new-nginx 1/1 registry.cn-hangzhou.aliyuncs.com/xtanlu/new-nginx 02/17/2019,09:26:21 D nginx-deployment-basic app:nginx 2/2 nginx:1.7.9 02/16/2019,14:57:25 D old-nginx run:nold-nginx 2/2 registry.cn-hangzhou.aliyuncs.com/xtanlu/loid-nginx 02/17/2019,09:25:20 D	Deployment Refresh Create by Image Help: Ø How to use private images Ø Create applications Ø Schedule a pod to the specified node Ø Create a Layer-7 Ingress Ø Configure Clusters Namespace default • Search By Nam Name Tag PodsQuantity Image Time Created new-nginx run:new-nginx 1/1 registry.cn-hangzhou.aliyuncs.com/xianlu/new-nginx 02/10/2019,04:26:21 Details Edit 5 cond-nginx ndiv-deployment-basic app:nginx 2/2 nginx:1.7.9 02/16/2019,14:57:25 Details Edit 5 cond-nginx old-nginx run:lod-nginx 2/2 registry.cn-hangzhou.aliyuncs.com/xianlu/nek-nginx 02/17/2019,09:25:20 Details Edit 5 cond-nginx	Deployment Refresh Create by Image Create by Ima

3. In the upper-right corner, click Create by Template.

Deploy templates		
Only Kubernetes versions 1.8.4 ar	nd above are supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list	
Clusters	kßs-cluster v	
Namespace	kube-system v	
Resource Type	Custom	
Template	<pre>1 apiVersion: v1 2 kind: ServiceAccount 3 metadota: 4 name: admin 5 namespace: kube-system 6 7 apiVersion: rbac.authorization.k8s.io/v1beta1 8 kind: ClusterRoleBinding 9 metadota: 10 name: admin 11 roleRef: 12 apiVersion: rbac.authorization.k8s.io 13 kind: ClusterRole 14 name: clusterr-admin 15 subjects: 16 - kind: ServiceAccount 17 name: admin 18 namespace: kube-system 19 20 apiVersion: v1 21 kind: Service 22 metadota: 23 labels: 24 task: monitoring 25 # For use as a Cluster add-on (https://github.com/kubernetes/kubernetes/tree/master //cluster/addons) 26 # If you are NOT using this as an addon, you should comment out this line. 27 kuberneter io/cames metrict.service. 28 * Vuberneter io/cames metrict.service.</pre>	Add Deployment Deploy with exist template

- 4. Select the target cluster from the Clusters drop-down list, and then select kubesystem from the Namespace drop-down list.
- 5. Select Custom from the Resource Type, copy the following code and paste it to the Template area, then click DEPLOY.

```
apiVersion : v1
kind : ServiceAcc ount
metadata :
    name : admin
```

```
namespace : kube - system
 apiVersion : rbac . authorizat ion . k8s . io / v1beta1
 kind : ClusterRol eBinding
 metadata :
   name : admin
 roleRef :
   apiGroup : rbac . authorizat ion . k8s . io
   kind : ClusterRol e
name : cluster - admin
 subjects :
   kind : ServiceAcc ount
name : admin
   namespace : kube - system
 apiVersion : v1
 kind : Service
 metadata :
   labels :
    task : monitoring
# For use as a Cluster add - on ( https :// github
 . com / kubernetes / kubernetes / tree / master / cluster /
addons )
 # If you are NOT using this
you should comment out this line
                                                  an
                                               as
                                                         addon ,
     kubernetes . io / cluster - service : ' true '
     kubernetes . io / name : metrics - server
   name : heapster
   namespace : kube - system
 spec :
   ports :
  - port : 80
     targetPort : 8082
   selector :
     k8s - app : metrics - server
 apiVersion : v1
 kind : Service
 metadata :
   name : metrics - server
   namespace : kube - system
   labels :
     kubernetes . io / name : metrics - server
 spec :
   selector :
     k8s - app : metrics - server
   ports :
     port : 443
     protocol : TCP
     targetPort: 443
 apiVersion : apiregistr ation . k8s . io / v1beta1
 kind : APIService
 metadata :
   name : v1beta1 . metrics . k8s . io
 spec :
   service :
     name : metrics - server
     namespace : kube - system
   group : metrics . k8s . io
   version : v1beta1
   insecureSk ipTLSVerif y : true
   groupPrior ityMinimum : 100
   versionPri ority : 100
```

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
  name : metrics - server
  namespace : kube - system
  labels :
    k8s - app : metrics - server
spec :
  selector :
    matchLabel s :
      k8s - app : metrics - server
  template :
    metadata :
      name : metrics - server
      labels :
        k8s - app : metrics - server
    spec :
      serviceAcc ountName : admin
      containers :
        name : metrics - server
image : registry .## REGION ##. aliyuncs . com / acs /
metrics - server : v0 . 2 . 1 - 9dd9511 - aliyun
        imagePullP olicy : Always
        command :
       - / metrics - server
       - '-- source = kubernetes : https :// kubernetes . default
I
       - '-- sink = socket : tcp :// monitor . csk .## REGION ##.
aliyuncs . com : 8093 ? clusterId =## CLUSTER_ID ##& public =
true '
```

Note:

You need to replace ## REGION ## with the region (for example, China East 1:cn-hangzhou) to which your target cluster is located, and replace ## CLUSTER_ID ## with your target cluster ID.

- Reset the monitoring data link.
 - 1. In the left-side navigation pane, choose Clusters > Nodes.
 - 2. Select the target Cluster.
 - 3. Click the instance ID of one Master node. In this step, the master-01 node is used.

No	de List					Ref	iresh Labe	l Management	Cluster	Expansion Add I	Existing Instance
Hel	p: 🔗 Postpay instance to Prej	pay 🔗 Node	exception 🔗 N	lode monitorin	g and alarms 🔞	Collect Kuberr	netes diagnostics i	nformation			
Cluste	ers k8s-cluster 🔻	Filter by La	abels 👻								
	Instance ID/Name/IP Address	Role/Status	Configuration	Pods (Allocated)	CPU Request/Limit	CPU Usage	Memory Request/Limit	Memory Usage	Kubelet Version	Update Time	Action
	master-01-	Master Running	Pay-As-You- Go ecs.n1.large	14	30.00% 0.00%	5.15 %	8.90% 14.73%	53.66 %	v1.11.5	02/05/2019,15:58:0	7 Monitor More -

4. Click Connect. On the displayed page, enter the remote connection password and click OK. After you log on to the ECS instance, run the following command:

```
sed - i ' s /-- horizontal - pod - autoscaler - use - rest -
clients = false /-- horizontal - pod - autoscaler - use - rest
- clients = true /' / etc / kubernetes / manifests / kube -
controller - manager . yaml
```

CentOS Linux 7 (Core) Kernel 3.10.0-957.5.1.el7.	Copy and Paste Commands	\times	ite the system.	Enter Copy Commands	Mo
iZbp10bft22o8zu188wqq6Z lo Password: Welcome to Alibaba Cloud E [root@iZbp10bft22o8zu188wq	Copy and paste the commands to the text box. Up to 2,000 characters are allowed. Non-standarkeyboard characters are not supported (such as Chinese characters). * Commands: \$\$\frac{\text{Sed} -1 's/horizontal-pod-autoscaleruse-rest-clients=false/horizontal-pod-autoscaleruse-rest-clients=true/'(etc/kubernetes/manifests/kube-controller-manager.yam)	rd			
	OK	ncel			

5. Repeat step c to step d on the master-02 and master-03 nodes.

After you complete the preceding operations, kubelet automatically starts and updates the kube-controller-manager component.

- Reset the component compatibility settings.
 - 1. In the left-side navigation pane, choose Discovery and Load Balancing > Service.
 - 2. Select the target Cluster and the kube-system namespace. Then click View YAML on the right of heapster.

Container Service - Kubernetes 🗸	I	Service								Refresh	Cr	reate
DaemonSet	L	Help: 🖉 Canary rel	lease		_							
Job	¢	Clusters k8s-cluste	er v Nan	nespace kube-system 🔻					Search By N	lame		Q
CronJob		Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint					Action
Pods		heapster	ClusterIP	02/25/2019,10:35:52		heapster:80 TCP	-	Details	Update	View YAML	(Delete
Volumes Claim		kube-dns	ClusterIP	02/05/2019,15:58:11	10.00	kube-dns:53 UDP kube-dns:53 TCP	-	Details	Update	View YAML	(Delete
 Discovery and Load B 		monitoring-influxdb	ClusterIP	02/25/2019,10:35:52		metrics-server:443 TCP	-	Details	Update	View YAML	(Delete
Service		nginx-ingress-lb	LoadBalancer	02/05/2019,16:02:10		nginx-ingress-lb:80 TCP nginx-ingress-lb:30106 TCP	10.001	Details	Update	View YAML	(Delete Us
Ingress						nginx-ingress-lb:443 TCP nginx-ingress-lb:31653 TCP	and the second second					
▼ Configuration		tiller-deploy	ClusterIP	02/05/2019,16:02:14		tiller-deploy:44134 TCP	-	Details	Update	View YAML	[Delete
Config Maps												

3. In the displayed dialog box, reset the k8s-app parameter of the selector field to metrics-server. Then click Update.

Edit YAML	×
<pre>1 ppiversion: v1 2 kind: Service 3 metadata: 4 creationTimestamp: '2019-02-25T02:35:52Z' 5 labels: 6 kubernetes.io/cluster-service: 'true' 7 kubernetes.io/name: metrics-server 7 task: monitoring 9 name: heapster 10 namespace: kube-system 11 resourceVersion: '4191616' 12 selflink: /api/V1/namespaces/kube-system/services/heapster 13 uid: 117ba8e1-38a6-11e9-82a9-00163e0ecfb9 14 spec: 15 clusterIP: 16 ports: 7</pre>	
	Update Cancel

- 4. In the left-side navigation pane, choose Application > Deployment.
- 5. Select the target Cluster and the kube-system Namespace.

Contain Kube	er Service - ernetes +		Deployn	nent					Refresh Create by Imag	e Create I	oy Template
Overvie	ew	^	Help: § monitorin	How to use privil	vate images 🛭 🖉 n release	Create applicat	ions 🔗 Sched	ule a pod to the specified node \mathscr{S} Create a Layer-4 Ingress \mathscr{S} Create a Lay	er-7 Ingress 🔗 Configure pod	auto scaling 🔞	Container
 Cluster: 	rs	L	Clusters	k8s-cluster 🔻	Namespace	kube-system	v		Search B	y Name	Q
Cluste	ers	L	Name		Tag		PodsQuantity	Image	Time Created		Action
Volum Name Autho	s nes espace orization	l	alibaba-	log-controller	component.ve component.re component.ine k8s-app:alibat controller	rsion:0.1.3 vision:527ff4d dex:0 pa-log-	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/log-controller:0.1.3.0-527ff4d-aliyun	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More≁
 Applica Deploy 	ation	l	alicloud	-application- er	app:alicloud-a controller owner:aliyun	pplication-	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-app-lifecycle- manager:v0.1.0.1-f832bed-aliyun	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More+
Statef	fulSet 10nSet		alicloud	-disk-controller	app:alicloud-d	isk-controller	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/alicloud-disk-controller:v1.11.2.2- a390cfb-aliyun	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More~
Job CronJo	lob	l	alicloud	-monitor- er	task:monitorir k8s-app:aliclo controller	ng ud-monitor-	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/alicloud-monitor-controller:v1.0.0	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More+
Pods		L	aliyun-a helper	cr-credential-	app:aliyun-acı helper	-credential-	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-acr-credential- helper:v18.10.29.0-1a28f02-aliyun	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More+
Releas	nes claim Ise		coredns		k8s-app:kube	-dns	2/2	registry-vpc.cn-beijing.aliyuncs.com/acs/coredns:1.1.3	02/28/2019,14:50:57	Details Ed Mo	it Scale nitor More+
 Discove Service 	ery and Loa ce		heapste	r	task:monitorir k8s-app:heap	ng ster	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/heapster-amd64:v1.5.1.1	02/28/2019,14:55:36	Details Ed Mo	it Scale nitor More+
Ingres	ss uration		monitor	ing-influxdb	task:monitorir k8s-app:influx	ng idb	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/heapster-influxdb-amd64:v1.1.1	02/28/2019,14:55:36	Details Ed Mo	it Scale nitor More~
Config	g Maps	-	nginx-in	gress-controller	app:ingress-n	ginx	2/2	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-ingress-controller:v0.20.0.1- 4597ce2-aliyun	02/28/2019,14:55:37	Details Ed Mo	it Scale nitor More+

- 6. Delete the components related to Heapster. In this example, you need to delete the heapster and monitoring-influxdb components.
 - On the right of the heapster component, choose More > Delete. In the displayed dialog box, click OK.
 - On the right of the monitoring-influxdb component, choose More > Delete.
 In the displayed dialog box, select the Delete associated services monitoringinfluxdb check box, and then click OK.

Deployment				Refresh Create by Imag	e Create by Template
Help: I How to use pr monitoring I Blue-gre	ivate images 🔗 Create applica en release	tions 🔗 Sched	ule a pod to the specified node \mathscr{O} Create a Layer-4 Ingress \mathscr{O} Create a Layer	er-7 Ingress 🔗 Configure pod	auto scaling 🔗 Container
Clusters k8s-cluster	Namespace kube-system	Ŧ		Search B	y Name Q
Name	Tag	PodsQuantity	Image	Time Created	Action
alibaba-log-controller	component.version:0.1.3 component.revision:527ff4d component.index:0 k8s-app:alibaba-log- controller	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/log-controller:0.1.3.0-527ff4d-aliyun	02/28/2019,14:55:37	Details Edit Scale Monitor More -
alicloud-application- controller	app:alicloud-application- controller owner:aliyun	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-app-lifecycle- manager:v0.1.0.1-f832bed-aliyun	02/28/2019,14:55:37	Details Edit Scale Monitor More -
alicloud-disk-controller	app:alicloud-disk-controller	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/alicloud-disk-controller:v1.11.2.2- a390cfb-aliyun	02/28/2019,14:55:37	Details Edit Scale Monitor More -
alicloud-monitor- controller	task:monitoring k8s-app:alicloud-monitor- controller	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/alicloud-monitor-controller:v1.0.0	02/28/2019,14:55:37	Details Edit Scale Monitor More -
aliyun-acr-credential- helper	app:aliyun-acr-credential- helper	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-acr-credential- helper:v18.10.29.0-1a28f02-aliyun	02/28/2019,14:55:37	Details Edit Scale Monitor More -
coredns	k8s-app:kube-dns	2/2	registry-vpc.cn-beijing.aliyuncs.com/acs/coredns:1.1.3	02/28/2019,14:50:57	Details Edit Scale Monitor More-
heapster 1	task:monitoring k8s-app:heapster	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/heapster-amd64:v1.5.1.1	02/28/2019,14:55:36	Details Edit Scare Monitor More
monitoring-influxdb	task:monitoring k8s-app:influxdb	1/1	registry-vpc.cn-beijing.aliyuncs.com/acs/heapster-influxdb-amd64:v1.1.1	02/28/2019,14:55:36	Details YAML Delete
nginx-ingress-controller	app:ingress-nginx	2/2	registry-vpc.cn-beijing.aliyuncs.com/acs/aliyun-ingress-controller:v0.20.0.1- 4597ce2-aliyun	02/28/2019,14:55:37	Details Edit Scale Monitor More→

7. Check the new data link status.

Data link initialization takes about three minutes.

In the left-side navigation pane, choose Application > Pods. You can verify that the CPU and memory columns show normal values. This means the data link has been reset.



If both the displayed CPU and memory columns of each component are 0, it indicates an exception.

Container Service - Kubernetes 🔻	Pod		Refresh
 Application 	Clusters k8s-cluster v Namespace kube-syste	m 🔻	Search By Name Q
Deployment	Max Nama	ntr Dod 10 A Noda A Timo Crasted A Chu	Margany
StatefulSet		pis routr y houe y finite created y CPO	тенногу
DaemonSet	alibaba-log-controller-888499bc4- 9qdlx Padax	02/05/2019,16:02:12 0.	002 35.059 Mi More -
Job			
CronJob	alicloud-application-controller- b76c68764-xc2vk Running 1	02/05/2019,16:02:10	001 Details 36.574 Mi More -
Pods			
Volumes Claim	alicloud-disk-controller-84d64c55cd- 6lkn7	02/05/2019,16:02:11 0	Details 52.039 Mi More -
Release			
 Discovery and Load B 	alicloud-monitor-controller- 84f86d5cd5-rfmcd	02/05/2019,16:02:11	004 43.082 Mi More -
Service			
Ingress	aliyun-acr-credential-helper- 56764b6cfc-p99rj O	02/05/2019,16:02:11 0	Details 36.672 Mi More ▼
 Configuration 			
Config Maps	cloud-controller-manager-2f49v Running 1	02/05/2019,16:02:10 0	Details 41.383 Mi More ▼
Secret			
▼ Store	cloud-controller-manager-Shppl	02/05/2019,16:02:42	Details

1.4.19 Create a Kubernetes cluster that supports Windows

This topic describes how to use the Container Service console to create a Kubernetes cluster that supports Windows.

Prerequisites

The following services are activated: Container Service, Resource Orchestration Service (ROS), Resource Access Management (RAM), and Auto Scaling service.

Log on to the Container Service console, ROS console, RAM console, and Auto Scaling console to activate the corresponding services.



The deployment of Kubernetes clusters that support Windows in Container Service is dependent on the application deployment capabilities of Alibaba Cloud ROS.

Therefore, you must activate ROS before you create a Kubernetes cluster that supports Windows.

Limits

- An SLB instance created with the cluster supports only the Pay-As-You-Go billing method.
- Kubernetes clusters that support Windows support only the VPC network type.
- Each account has a set quota of resources that can be used to create clusters. The default numbers are as follows:
 - Each account can create up to 100 security groups.
 - Each account can create up to 60 SLB instances of the Pay-As-You-Go billing method.
 - Each account can create up to 20 EIPs.

Note:

If any of the preceding quotas are exceeded when you create a cluster, the cluster fails to be created. In the case that you need a larger quota for any of these resources, you can open a ticket.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Clusters > Clusters. Then click Create Kubernetes Cluster in the upper-right corner.
- 3. On the displayed page, find Windows Cluster, and then click Create.
- 4. Enter the cluster name.



The cluster name must be 1 to 63 characters in length, and can contain letters, numbers, Chinese characters, and hyphens (-).

	* Cluster Name	k8s-windows-cluster	
The cluster name should be 1-63 character hyphens.		The cluster name should be 1-63 character hyphens.	ers long, and can contain numbers, Chinese characters, English letters and

Region	China North 2	China North 3	China North 5	China East 1	China East 2
	(Beijing)	(Zhangjiakou)	(Huhehaote)	(Hangzhou)	(Shanghai)
	China South 1		Asia Pacific SE 1	Asia Pacific SE 3	Asia Pacific SE 5
	(Shenzhen)	Hong Kong	(Singapore)	(Kuala Lumpur)	(Jakarta)
	Asia Pacific SOU 1		US West 1 (Silicon	EU Central 1	
	(Mumbai)	US East 1 (Virginia)	Valley)	(Frankfurt)	
Zone	China East 1 Zone G		-		

5. Select the region and zone where you want to locate the cluster.

6. Select a VPC.

Note:

Kubernetes clusters support only the VPC network type.

VPC: You can click Auto Create to create a VPC together with the Kubernetes cluster, or click Use Existing to use an existing VPC. If you click Use Existing, you can select a VPC and a VSwitch from the two displayed drop-down lists.

- If you click Auto Create, the system automatically creates a NAT gateway for a VPC when you create a cluster.
- If you click Use Existing and the selected VPC is associated with an existing NAT gateway, the system then uses the preexisting NAT gateway. If the selected VPC is not associated with any existing NAT gateway, the system automatically creates a NAT gateway for the selected VPC. Furthermore, if you do not want the system to automatically create a NAT gateway for the selected VPC, clear the Configure SNAT for VPC check box.

Dote:

If you set the system not to create a NAT gateway automatically, then you need to manually set a NAT gateway, or set an SNAT entry to ensure that your selected VPC is accessible to the Internet. Otherwise, instances in the VPC cannot access the Internet, which results in a cluster creation failure.

VPC	Auto Create	Use Existing	
	vpc-k8s-for-cs-c7098	88d6aa0c44af8de3817(c78706e0 🔻
	(vsw-bp149pxcw4v9tdwzclr83) ZoneG		

7. Set the node type by selecting a node billing method.



Node Type	Pay-As-You-Go	Subscription	
	View the differences be	etween the two billing r	nethods Billing method comparison

8. Set the Worker instance.



- Each cluster must contain at least 2 Worker nodes.
- Each cluster can contain up to 48 Worker nodes. To create more Worker nodes, open a ticket.
- By default, a system disk is attached to Worker nodes. For the system disk, available system disk types are SSD disks and Ultra disks.

• You can manually attach a data disk to Worker nodes. For data disks, available data types are SSD disks and Ultra disks.

Instance Configuration					
	x86-Architecture	Heterogeneous Computir	g ECS Bare	Metal Instance	Super Computing Cluster
Instance Type	4 Core(s) 8 G (ecs	s.c5.xlarge)	•	T	
Quantity	3 unit(s) 🌲				
System Disk	Ultra Disk	▼ 12	GiB		
🛃 Attach Data Disk	Ultra Disk	▼ 10	GiB 🌲		

9. Select the Kubernetes version that supports the Windows operating system. By default, the Kubernetes version with this support is selected.

Note:
The system does not display the Operating Stem option if you select the
Kubernetes version that does not support the Windows operating system.

Docker Version	18.09.2		
Kubernetes	1.12.6-aliyun.1	1.11.5	
Version			

10.Select the Windows operating system.

Operating	Linux	Windows (Beta)
System		

11.Set the logon password.

- · Logon Password: Set the node logon password.
- · Confirm Password: Confirm your node logon password.

* Logon	
Password	
	The password should be 8-30 characters long and contain three types of characters (uppercase/lowercase letters, numbers and special characters).
* Confirm	
Password	
	Please fill in a valid password

12.Set the Pod Network CIDR and Service CIDR.



- If you choose to use an existing VPC, we recommend that you set these two parameters.
- The two Classless Inter-Domain Routing (CIDR) blocks cannot overlap with each other, or with your selected VPC, or with the CIDR blocks used by any other existing Kubernetes clusters in the selected VPC. Furthermore, they cannot be modified after the cluster is created. For more information, see *Plan Kubernetes CIDR blocks under VPC*.

Pod Network	172.20.0.0/16
CIDR	
	Please fill in a valid private CIDR, namely the following CIDR and its subnets: 10.0.0.0/8, 172.16-31.0.0/12-16, 192.168.0.0/16
	Cannot be duplicated with the VPC CIDR and CIDR used by Kubernetes cluster in VPC, cannot be modified after creation
	For more information about CIDR block planning for a cluster, see Plan Kubernetes CIDR blocks under a VPC.
Service CIDR	172.21.0.0/20
	Optional range: 10.0.0.0/16-24, 172.16-31.0.0/16-24, 192.168.0.0/16-24
	Cannot be duplicated with the VPC CIDR and CIDR used by Kubernetes cluster in VPC, cannot be modified after creation

13.Configure an SNAT gateway for the VPC.

Note:		

We recommend that you configure this setting. Otherwise, instances in the VPC cannot access the Internet, which will result in a cluster creation failure.

- If you set the system to automatically create a VPC, you must configure an SNAT gateway for the VPC.
- If you set the system to use an existing VPC, you can perform one of the following two operations:
 - Set the system to automatically configure an SNAT gateway.
 - Manually configure a NAT gateway or SNAT gateway for the VPC.

Configure SNAT	Configure SNAT for VPC
	If the VPC you choose does not have access to Internet, NAT gateway and EIP will be used to configure SNAT for the VPC. During this period, NAT gateway, EIP, and other resources may be created.

14.Enable the Use Public SLB to Expose API Server function.

The Kubernetes API supports such RESTful API actions as retrieving, creating, querying, updating, and deleting resources, such as pods and services.

• If you enable this function by selecting the check box, an Internet SLB instance is created and the Master node port (namely, port 6443) is opened. In this case, you can use kubeconfig to connect to and operate the cluster through the Internet.



The API server uses the Master node port.

• If you do not enable this function, no Internet SLB instance is created. In this case, you can only use kubeconfig to connect to and operate the cluster within the VPC.



15.Enable Log Service.

You can select an existing project or create a new project.

If you select the Using Log Service check box, a Log Service plugin is automatically installed in the cluster. Then, when you create an application in the cluster, you

can immediately use Log Service with only a few configurations required. For more information, see *Use Log Service to collect Kubernetes cluster logs*.

Log Service	Vsing Log Service		
	Select Project	Create Project	
	A SLS Project named k	8s-log-{ClusterID} will	be created automatically

16.Set the RDS instance whitelist.

Add the IP address of the node to the RDS instance whitelist.



If you choose to use an existing VPC, we recommend that you set this parameter.

RDS Whitelist	Select RDS Instances

17.Set advanced configurations.

Set the number of pods for a node. This parameter specifies the maximum number of pods that can be run by a single node. We recommend that you retain the default setting.



18.Click Create, and then click Create in the displayed dialog box.



A Kubernetes cluster that contains multiple nodes typically takes ten minutes to be created.

Verify the result

After the cluster is created, you can view the cluster in the cluster list of the Container Service console.

Container Service - Kubernetes 🔻		Cluster List	You can cr	eate up	o to 100 cluster	s and add up to) 1000 nodes in ea	ach cluster. To i	request a q	uota increase, submit a t	icket.	Refresh Create	Kubernetes Cluster
Overview	*	Help: & Create cluster duster cluster creation failures	Create GPU of Authorization r	clusters nanage	s 🔗 Scale clu ement	ster 🔗 Conne	ect to Kubernetes	cluster via kube	ecti 🔗 M	anage applications with c	ommands	& Cluster planning	
 Clusters 		Name 🔻			Tags								
Clusters													
Nodes		Cluster Name/ID		Tags	Cluster Type	Region (All)	Network Type	Cluster Status	of Nodes	Time Created	Version		Action
Volumes Namespaces	IJ	111		۲	Serverless Kubernetes	China North 2 (Beijing)	VPC	Running		04/20/2019,18:27:30	1.11.5	Manag Delete	e View Logs Use Cloud Shell
Authorization Application		k8s-windows-cluster		۲	Windows Kubernetes	China East 1 (Hangzhou)	VPC	Running	3	04/20/2019,15:46:00	1.12.6- aliyun.1	Manag	e View Logs Scale Out More↓

• To view the cluster logs, click View Logs on the right of the cluster. To view more details, click Stack Events.

Container Service - Kubernetes 🔻		Cluster Logs: k8s-windows-cluster tist								
Overview	^	Detailed resource deployment logs <mark>. Stack Events</mark>								
 Clusters 		Time Information								
Clusters		04/22/2019,10:51:01 Start to DescribeK8sUserCertConfig								
Nodes		04/22/2019,10:49:30 Start to DescribeK8sUserCertConfig								
Volumes		04/22/2019,10:32:43 Start to DescribeK8sUserCertConfig								
Namespaces	-	04/20/2019,16:03:53 start to update cluster status CREATE_COMPLETE								
Authorization		04/20/2019,16:03:53 Successfully to create managed kubernetes cluster								
 Application 		04/20/2019,15:58:49 Install addons successfully								
Deployments		04/20/2019,15:58:17 Start to install addons								
Stateful Sets		04/20/2019,15:57:15 Stack CREATE completed successfully:								
Daemon Sets Jobs		04/20/2019,15:47:14 Successfully to CreateStack with response &ros.CreateStackResponse(Id: "3 ", Name	e:"k8s-for-							

• To view the basic information, the connection information, and other information, return to the cluster list page and click Manage in the action column of the cluster.

Cluster:k8s-windows-cluster			Refn	Use Cloud Shell		
Basic Information						
Cluster ID: c	VPC	Running	Region: China East 1 (Han	gzhou)		
Cluster Information						
API Server Internet endpoint	https://-					
API Server Intranet endpoint	https://					
Pod Network CIDR						
Service CIDR	10.0.000					
Service Access Domain	-	the second second second				
Cluster Resource						
ROS						
VPC	-					
Worker RAM Role	Worker RAM Role					
Connect to Kubernetes cluster via kubectl (Use Cloud	Shell)					
1. Download the latest kubect client from the Kubernetes Edition page .						
2. Install and set up the kubectl client. For more informa	tion, see Installing and Setting Up kubect	I				
3. Configure the cluster credentials:						

Cluster information

- API Server Internet endpoint: the IP address and the port through which the Kubernetes API server provides services to the Internet. It enables you to manage the cluster by using kubectl or other tools on your terminal.
- API Server Intranet endpoint: the IP address and the port through which the Kubernetes API server provides services within the cluster. This IP address is the SLB instance IP address. The three Master nodes on its backend provide services.
- Pod Network CIDR: the CIDR block where the pods of a Kubernetes cluster are located.
- Service CIDR: the CIDR block where the services of a Kubernetes cluster are located.

For example, you can log on to the Master node by using SSH (for more information, see *Connect to a Kubernetes cluster by using kubectl*), and run the kubectl get node command to view the cluster nodes.

<pre>shell@Alicloud:~\$ kubectl get node</pre>				
NAME	STATUS	ROLES	AGE	VERSION
cn-hangzhou.	Ready	<none></none>	26m	v1.12.6-aliyun.1
cn-hangzhou.	Ready	<none></none>	26m	v1.12.6-aliyun.1
cn-hangzhou.	Ready	<none></none>	26m	v1.12.6-aliyun.1
shell@Alicloud:~\$				

1.5 Node management

1.5.1 Add an existing ECS instance

You can add existing Elastic Compute Service (ECS) instances to a created Kubernetes cluster. Currently, Kubernetes clusters only support adding worker nodes.

Prerequisites

- If you have not created a cluster before, create a cluster first. For how to create a cluster, see *Create a Kubernetes cluster*.
- · Add the ECS instance to the security group of the Kubernetes cluster first.

Context

- By default, each cluster can contain up to 40 nodes. To add more nodes, open a ticket.
- The ECS instance to be added must be in the same Virtual Private Cloud (VPC) region as the cluster.

- When adding an existing instance, make sure that your instance has an Elastic IP (EIP) for the VPC network type, or the corresponding VPC is already configured with the NAT gateway. In short, make sure the corresponding node can access public network normally. Otherwise, the ECS instance fails to be added.
- The ECS instance to be added must be under the same account as the cluster.
- Only the ECS instance whose operating system is CentOS can be added.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.
- 3. Select the target cluster and click More > Add Existing Instance.

The Add Existing ECS Instance page appears. All the available ECS instances under the current account are displayed on this page. Select to add existing ECS instances automatically or manually.

If Automatically Add is selected, select the ECS instances to add them to the cluster automatically. If Manually Add is selected, you must obtain the command and then log on to the corresponding ECS instance to add the ECS instance to this cluster. You can only add one ECS instance at a time.

	Container Service - Kubernetes +	Cluster List			r'ou can create up	to 5 clusters and can add	up to 40 nodes in each cluster.	Refresh	Create Kubernetes Cluster
	Overview	Help: & Create cluster & Scale cluste	Connect to Kubernetes cluster via kubectl	& Manage applications	with commands				
-	Clusters 1	Name 🔻							
l	Clusters	Cluster Name/ID	Cluster Type Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		2
	Nodes	k8s-cluster 🖌	Kubernetes China East 1 (Hangzhou)	VPC vpc-bp1kd7yn4qn	Running	08/24/2018,09:57:11	1.10.4	Manage	View Logs Dashboard Scale Cluster More -
	Namespace							3	Delete Add Existing Instance

- 4. Select Automatically Add to add multiple ECS instances at a time.
 - a) In the list of existing cloud servers, select the target ECS instance, and then click Next Step.

Add Existing ECS Instance - test 🔹 Back to Cluster List							
Select Existing ECS Instance(s) Adding Method: Automatically Add Manually Add	Enta	er Instance Information	Added S	iuccessfully			
Adding an existing ECS instance to the Container Service v	vill change the instance system disk.	The disk ID will change and the previous s	ystem disk will be released.				
Note: 1. Your previous system disk's user snapshots will be retained. Automatic snapshots will be retained or deleted according to the setting of the "Delete automatic snapshots when releasing disk" option. You can go to the disk list and click "Modify Attributes" to view and modify attribute values. 2. To retain enough snapshot quota for the periodic automatic snapshot policy, you can delete unwanted user and automatic snapshots. 3. Back up data before performing this operation to avoid data loss.							
Select Existing ECS Instance(s):							
Instance ID Enter the instance ID for exact search	. Search						
 Instance ID Instance Name 	IP Address	Zone	Network Type	Instance Type			
e test 2	47.57.108.17 (Elastic) 192.188.245.000 (Private)	cn-hangzhou-g	VPC	ecs.sn2ne.large			
				3 Next Step			

b) Enter the instance information, set the logon password, and then click Next Step.

Add Existing ECS Instance - test 🛛 🖻 🛨 Back to Clu	ster List		
Select Existing ECS Instance(s)	Enter Inst	ance Information	Added Successfully
Cluster ID/Name :	Information of the cluster to which to add the E	CS instance(s).	
Instance Information :	Instance ID	Instance Name	
	Hapt.diviserapbacayr%a	test	
			Dro Choo Maut Choo

c) Click Confirm in the displayed dialog box. The selected ECS instances are automatically added to this cluster.

Add Existing ECS Instance - test * Back to Clust	ter List		
Select Existing ECS Instance(s)	Enter Instance Informa	tion	Added Successfully
Cluster ID/Name :	Information of the cluster to which to add the ECS instance(s).	
Instance Information :	Instance ID	Instance Name	
	Hapt.divlaevapbecapi%a	test	
			Pre Step Next Step

- 5. Optional: You can also select Manually Add to manually add an existing ECS instance to the cluster.
 - a) Select the ECS instance to be added and then click Next Step. You can add only one ECS instance at a time.

Confirm	adding existing instance to cluster	\times
0	Are you sure you want to add the selected ECS instance to the clu ster remove the selected ECS instance to the clu	
	 Adding an existing ECS instance will change the system disk. The disk ID will change and the previous system disk will be re leased. 1. After the system disk is changed, the user snapshots on the previous system disk will be retained, but the automatic snaps hots will be released with the disk. 2. To retain enough snapshot quota for the automatic snapshot policy of the new disk, you can delete unwanted snapshots. 3.Back up data before performing this operation to avoid data loss. Alibaba Cloud is not liable for any data losses caused by your failure to back up personal data in the system. 	
	Confirm Cance	

b) Confirm the information and then click Next Step.

Cluster ID/Name :	/ KBs-cluster Information of the cluster to which to add the ECS instance(s).				
Instance Information :	Instance ID	Instance Name			
		7da6e71d6ae8d29			
				Prev Step	Next Step

c) Copy the command.

Select Existing ECS Ins	tance(s)	Enter Instance Information	Add	ed Successfully
Only supports adding nodes in the same VPC with CentOS operating system				
Log in to the node you v	want to add, execute the follov	ving command:		
	,			

- d) Click Done.
- e) Log on to the *ECS console* and click Instances in the left-side navigation pane. Select the region in which the cluster resides and the ECS instance to be added.

f) Click Connect at the right of the ECS instance to be added. The Enter VNC Password dialog box appears. Enter the VNC password and then click OK. Enter the copied command and then click OK to run the script.



g) After the script is successfully run, the ECS instance is added to the cluster. You can click the cluster ID on the Cluster List page to view the node list of the cluster and check if the ECS instance is successfully added to the cluster.

1.5.2 View node list

You can view the node list of the Kubernetes cluster by using commands, in the Container Service console, or in the Kubernetes dashboard.

View node list by using commands



Before using commands to view the node list of the Kubernetes cluster, *#unique_57* first.

After connecting to the Kubernetes cluster by using kubectl, run the following command to view the nodes in the cluster:

kubectl get nodes

Sample output:

```
$ kubectl
            get
                  nodes
                      VERSION
NAME
       STATUS
                AGE
                                           v1.6.1 - 2 +
iz2ze2n6ep
           53tch701yh
                        9zz
                              Ready
                                     19m
ed9e3d33a0
            7093
            2wibijx39e
                                          v1.6.1 - 2 +
                              Ready
iz2zeafr76
                       5az
                                     7m
ed9e3d33a0
           7093
                                          v1.6.1 - 2 +
 iz2zeafr76
           2wibijx39e 5bz
                              Ready
                                     7m
ed9e3d33a0
           7093
                                           v1.6.1 - 2 +
 iz2zef4dnn
           9nos8elyr3 2kz
                              Ready
                                     14m
ed9e3d33a0
           7093
                                           v1.6.1 - 2 +
           8enoreufst
                              Ready
 iz2zeitvvo
                        kmz
                                     11m
ed9e3d33a0
            7093
```

View node list in Container Service console

- 1. Log on to the Container Service console.
- 2. Click Kubernetes >Clusters > > Nodesin the left-side navigation pane.
- 3. Select the cluster from the Cluster drop-down list and then view the node list of this cluster.

View node list in Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. Click Kubernetes > Clusters in the left-side navigation pane.
- 3. Click Dashboard at the right of the cluster to enter the Kubernetes dashboard.
- 4. In the Kubernetes dashboard, click Nodes in the left-side navigation pane to view the node list of this cluster.

1.5.3 Node monitoring

Kubernetes clusters integrate with the Alibaba Cloud monitoring service seamlessly. You can view the monitoring information of Kubernetes nodes and get to know the node monitoring metrics of the Elastic Compute Service (ECS) instances under Kubernetes clusters.

Procedure

1. Log on to the Container Service console.

- 2. Under Kubernetes, click Clusters > Nodes to enter the Node List page.
- 3. Select the target cluster and node under the cluster.
- 4. Click Monitor at the right of the node to view the monitoring information of this node.
- 5. You are redirected to the CloudMonitor console. View the basic monitoring information of the corresponding ECS instance, including the CPU usage, network inbound bandwidth, network outbound bandwidth, disk BPS, and disk IOPS.

What's next

To view the monitoring metrics at the operating system level, install the CloudMonitor component. For more information, see *Host monitoring overview*.

Kubernetes clusters can now monitor resources by using application groups. For more information, see *#unique_60*.

1.5.4 Manage node labels

You can manage node labels in the Container Service console, including adding node labels in batches, filtering nodes by using a label, and deleting a node label quickly.

For how to use node labels to schedule pods to specified nodes, see *#unique_62*.

Prerequisite

You have successfully created a Kubernetes cluster. For more information, see *#unique_63*.

Add node labels in batches

- 1. Log on to the Container Service console.
- 2. Click Kubernetes Clusters > Nodes in the left-side navigation pane.
- 3. Select the cluster from the Clusters drop-down list and then click Label Management in the upper-right corner.
- 4. Select one or more nodes by selecting the corresponding check boxes and then click Add Tag.

5. Ener the name and value of the label in the displayed dialog box and then click OK.

Nodes with the same label are displayed on the Label Management page.

Filter nodes by using a label

- 1. Log on to the Container Service console.
- 2. Click Kubernetes Clusters > Nodes in the left-side navigation pane.
- 3. Select the cluster from the Clusters drop-down list and then click Label Management in the upper-right corner.
- 4. Click the label at the right of a node to filter nodes by using the label. In this example, click group : worker .

Nodes with the label group : worker are filtered.

Delete a node label

- 1. Log on to the Container Service console.
- 2. Click Kubernetes Clusters > Nodes in the left-side navigation pane.
- 3. Select the cluster from the Clusters drop-down list and then click Label Management in the upper-right corner.
- 4. Click the delete (x) button of a node label, for example, group : worker .

Click Confirm in the displayed dialog box. The node label is deleted.

1.5.5 Set node scheduling

You can set node scheduling through the web interface so that you can allocate loads to each node properly.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters > Nodes to enter the Node List page.

3. Select a cluster, select a node under the cluster, and click Schedule Settings on the right.

Container Service - Kubernetes 👻	Node List	Refresh Label Management Scale Cluster Add Existing Instance
Overview	Help: & Pos 3 stance to Prepay	
Clusters 1	Clusters k8s-cluster Filter by labels	
Clusters	IP Address Role Instance ID/Name Configuration Pods(Allocated) Limit)	Memory(Request Limit) Update Time Action
Volumes	Pay-As-You- Go ecs.c5.large	O8/24/2018,09:57:00 Details Monitor Remove Scheduling Settings
Application	Vorker Pay-&-You- Go ecs.c5.large	- 08/24/2018,10:06:00 Details Monitor Remove Scheduling Settings
Deployment Pods	Pay-As-You- Go ecs.cS.large	- 08/24/2018,09:59:00 Details Monitor Remove Scheduling Settings
Service Ingress	Pay-As-You- Go ecs.c5.large	- 08/24/2018,09:58:00 Details Monitor Remove Scheduling Settings

4. Set node scheduling in the displayed dialog box. In this example, click Change to Unschedulable to set the node to unschedulable.



The scheduling status of the current node is displayed in the Scheduling Settings dialog box, which is schedulable by default. You can change the status.

5	cheduling Settings			\times
	Scheduling Status:	Schedulable	Change to Unschedulable	
				Close

After the status is set, the scheduling status of the node changes in the dialog box.

4	Scheduling Settings			\times
	Scheduling Status:	Unschedulable	Change to Schedulable	
				Close

What's next

When you deploy your application later, you can find that pods are not scheduled to the node.

1.5.6 Remove a node

Before you restart or release an ECS instance in a Kubernetes cluster, you need to remove the ECS node from the cluster. This topic describes how to remove a node from a Kubernetes cluster.

Prerequisites

• You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

• You have connected to the Kubernetes cluster by using kubectl, see Connect to a Kubernetes cluster by using kubectl.

Context

- Removing a node causes pod migration. This may affect the services provided by the pods running on the node. Therefore, we recommend that you remove a node only when fewer services are in demand.
- Removing a node may cause unintended risks. We recommend that you back up your data in advance and exercise caution when performing this action.
- Only Worker nodes can be removed.

Procedure

1. Run the following command to migrate the pods on the target node to other nodes:



You must ensure that other nodes in the Kubernetes cluster have sufficient resources to run the pods that you want to migrate.

```
kubectl drain node-name
```

Note:

The node-name parameter must be in the format of your-region-name.node-id.

- your-region-name indicates the name of the region where your cluster resides.
- node-id indicates the ID of the ECS instance in which the node to be removed
 resides. For example, cn hanghzou . i xxx .
- 2. Set the node to be removed as the non-schedulable node.

Method 1: Use a command

• Run the following command to set the node to be removed as the nonschedulable node:

kubectl cordon *node-name*

Method 2: Use the Container Service console

For more information, see Set node scheduling.

3. In the left-side navigation pane under Kubernetes, choose Clusters > Nodes.

4. Under the target cluster, select the target node, and choose More > Remove in the Action column.

<	Node List	Refresh Label Management Scale Cluster Add Existing Inst	tance
Basic Information	Help: & Postpay instance to Prepay & Node exception & Node monitoring and alarms & Collect Kubernetes diagnostics in	nformation	
Node List	Clusters kas-delete-node v Filter by Labels v		
Event List	IP Address Role Instance ID/Name Status Configuration Pods(Allocated) CPU(Reque	est Limit) Memory(Request Limit) Update Time Action	
	Worker worker-k8s-for-c Running Pay-As-You-Go 10 5.00% [2.85 % 5.83% 50.52 % 12/17/2018.10:36:00 Scheduling Settings Monitor	ore •
	Master master-03-k8s-f0 Running Pay-As-You-Go 8 23.75%	5.75 % 5.83% 54.69 % 12/17/2018.10:28:00 Scheduling Settings Monitor	Remove
	Muster master-02-k0is-fo PRunning Pay-As You-Go 14 30.00%	5,45 % 8,90% 64,63 % 12/17/2018,10:27:00 Scheduling Settings Monitor Mc	Group Mor

Note:

If you want to remove multiple nodes at a time, you can select the target cluster on the Node List page, select all the nodes to be removed, and then click Remove.

<	Node List Refresh Label Management Scale Cluster Add E	xisting Instance
Basic Information	Help: ∂ Postpay instance to Prepay ∂ Node exception ∂ Node monitoring and alarms ∂ Collect Kubernetes diagnostics information	
Node List	Clusters kills-delete-node v Filter by Labels v	
Event List	IP Address Role Instance ID/Name Status Configuration Pods(Allocated) CPU(Request Limit) Memory(Request Limit) Update Time Action	
	192.168.0.135 Worker Hop1thrags/keh5tp?vy0 Pay-As-You-Go 10 5.00% \$2.95 % 5.83% \$2.95 % 5.83% 12/17/2018,10:36:00 Scheduling Settings Pay-As-You-Go	fonitor More -
	192.168.0.132 Mester Hot Influence Pay-A6 You-So 8 23.75% S.68 % S.68 % S.83% 12/17/2018,10:28:00 Scheduling Settings Market no.	tonitor More -
	I 192.168.0.131 Mester Hop177/R9xyfm8514w862 • Running es.nt.large Pay-A6 Your-Go es.nt.large 14 30.00% imaster - 02-48% / 0 12/17/2018,10:27:00 Scheduling Settings Mester	fonitor More -
-	I 192.168.0.130 Mester Hop Ladyky/SHidrophyne Pay-As You-Go 8 23.75% 12/17/2018,10:26:00 Scheduling Settings Market of Hop Setings Market of Hop Settings Market of Hop Setti	fonitor More -
	Isol.168.0.134 Worker HopIthrags/kebSp7xw • Running Pay-A6-You-So es.niLlarge 4 5.00% 1.75 % 5.83% 28.28 % 12/17/2018,10:26:00 Scheduling Settings //	tonitor More -
	International state Internatinget Internationa state	tonitor More -
	Total: 6 item(s), Per Page: 20 item(s) < <	1 > >

5. Optional: Select the Release ECS at the Same Time check box to permanently release the ECS instance where the node resides.



Note:

- Only Pay-As-You-Go ECS instances can be released.
- · A Subscription ECS instance will be released automatically when it expires.
- If you do not select the Release ECS at the Same Time check box, the ECS instance in which the node resides will continue to be charged.
- 6. Click OK.

1.5.7 Use Alibaba Cloud Kubernetes GPU node labels for scheduling

When you implement GPU computing through a Kubernetes cluster, you can schedule an application to the node installed with GPU devices as needed by using GPU node labels.

Prerequisites

- You have created a Kubernetes cluster that has GPU nodes. For more information, see *Configure a Kubernetes GPU cluster to support GPU scheduling*.
- You have connected to the Master node, which makes it easier to view node labels and other information. For more information, see *Connect to a Kubernetes cluster by using kubectl*.

Context

When deploying NVIDIA GPU nodes, Kubernetes that runs on Alibaba Cloud discovers the GPU attribute and exposes it as the node label information. Node labels provide the following benefits:

- 1. Node labels help you filter GPU nodes.
- 2. Node labels can be used as the scheduling conditions for application deployment.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Nodes.



In this example, the cluster has three Worker nodes of which two Worker nodes are mounted with GPU devices. You need to view the node IP addresses for verification.

Container Service - Kubernetes +	Node List Refresh Label Management Scale Cluster Ad	d Existing Instance
Overview	Help: Ø Postpay instantion O Prepay Ø Node exception Ø Node monitoring and alarms Ø Collect Kubernetes diagnostics information	
Clusters	Clusters k8s-test v Filter by Labels v	
Clusters	IP Address Role Instance ID/Name Configuration Pods(Allocated) CPU(Reguest Limit) Memory(Reguest Limit) Update Time Action	
Nodes 2 Volumes	Pay-As-You-Go 5 30.00% 1.23 % 2.52% 11.20 % 12/05/2018,13:40:00 Scheduling Setting	ngs Monitor More+
 Application Deployment 	166 Worker Pay-As-You-Go ecs.gn4- c4g1.xlarge 11 30.00% 1.88 % 2.52% 16.33 % 12/05/2018,13:40:00 Scheduling Setting	ngs Monitor More+
StatefulSet Job	Inference Pay-As-You-Go ecs.n1.large 8 23.75% 4.68 % 5.83% 12/05/2018,13:30:00 Scheduling Setting	ngs Monitor More+
CronJob Pods	Pay-As-You-Go 8 23.75% 4.30 % 5.83% 39.02 % 12/05/2018,13:29:00 Scheduling Setting	ngs Monitor More+
Service	Info: Master Pay-As-You-Go ecs.n1.large 14 30.00% 6.25 % 8.90% 12/05/2018,13:28:00 Scheduling Setting	ngs Monitor More+
Volumes Claim Release	167 Worker Pay-As-You-Go ecs.n4.Jarge 4 10.00% 12.03% 52.84 % 12/05/2018,13:55:00 Scheduling Settin	ngs Monitor More+
Config Maps Secret	Total: 6 item(s) , Per Page : [20 v]item(s)	« < 1 > »

3. Select a GPU node, and choose More > Details in the action column. Then, you can view the GPU node label on the Kubernetes dashboard.



You can also log on to a Master node and run the following command to view the GPU node label:

# kubectl	get	nodes				
NAME	0			STATI	JS RO	LES AGE
VERSIC	DN					
cn - beijing	g.i-	2ze2dy2h9w	97v65	uuaft	Ready	master
2d	v1 .	11 . 2				
cn - beijing	g.i-	2ze8o1a45q	dv5q8	a7luz	Ready	< none >
2d	v1	. 11 . 2		#	Compare	this node
with the	node	displayed	in	the	console	to determine
the GPU	node	•				
cn - beijing	g.i-	2ze8o1a45q	dv5q8	a7lv0	Ready	< none >
2d	v1 .	11 . 2				
cn - beijing	g.i-	2ze9xylyn1	1vop7	g5bwe	Ready	master
2d	v1 .	11 . 2				
cn - beijing	g.i-	2zed5sw8sn	jniq6	mf5e5	Ready	master
2d	v1 .	11 . 2				
cn - beijing	g.i-	2zej9s0zij	ykp9p	wf7lu	Ready	< none >
2d	v1 .	11 . 2			-	

Select a GPU node and run the following command to view the GPU node label:

<pre># kubectl dv5g8a7luz</pre>	describe node cn - beijing .i - 2ze8o1a45q
Name :	cn – beijing . i – 2ze8o1a45q dv5q8a7luz
Roles :	< none >
Labels :	aliyun . accelerato r / nvidia_cou nt = 1 # This field is important
	aliyun . accelerato r / nvidia_mem = 12209MiB
	aliyun . accelerato r / nvidia_nam e = Tesla
- M40	
	beta . kubernetes . io / arch = amd64

In this example, the GPU node contains the following three node labels:

Key	Value
aliyun . accelerato r / nvidia_cou nt	Number of GPU cores
aliyun . accelerato r / nvidia_mem	GPU memory in MiB
aliyun . accelerato r / nvidia_nam e	Name of the GPU computing card of the NVIDIA device

The GPU cloud servers of the same type share the same GPU computing card name . Therefore, you can use this label to filter nodes.

```
aliyun . accelerato r / nvidia_nam e
                      no - l
# kubectl
              get
 = Tesla - M40
 NAME
                                              STATUS
                                                             ROLES
                                                                           AGE
        VERSION
 cn - beijing . i - 2ze8o1a45q
2d v1 . 11 . 2
                                     dv5q8a7luz
                                                                     < none >
                                                        Ready
2d vi . 11 . 2
cn - beijing . i - 2ze8o1a45q
2d v1 . 11 . 2
                                     dv5q8a7lv0
                                                        Ready
                                                                     < none >
```

- 4. Return to the Container Service console home page, choose Application > Deployment in the left-side navigation pane, and click Create by Template in the upper-right corner.
 - a) Create a TensorFlow application and schedule this application to the GPU node.



In this example, the YAML template is orchestrated as follows:

```
#
                  tensorflow
                               deployment
 Define
            the
apiVersion : apps / v1
kind : Deployment
metadata :
   name : tf - notebook
   labels :
     app : tf - notebook
spec :
   replicas :
               1
   selector : # define
                          how
                                the
                                       deployment
                                                    finds
                                                            the
pods
      it
            manages
    matchLabel s :
  app : tf - notebook
template : # Define the
                                     specificat ions .
                               pod
    metadata :
       labels :
         app : tf - notebook
     spec :
       nodeSelect or :
               field is important .
       # This
         aliyun . accelerato r / nvidia_nam e : Tesla - M40
       containers :
```
```
- name : tf - notebook
image : tensorflow / tensorflow : 1 . 4 . 1 - gpu - py3
resources :
    limits :
        nvidia . com / gpu : 1
        # This field is important .
    ports :
        containerP ort : 8888
        hostPort : 8888
env :
        - name : PASSWORD
        value : mypassw0rd v
```

b) You can also avoid deploying an application to a GPU node. The following deploys an Nginx pod and schedules it by using the node affinity feature. For more information about node affinity, see *Create a deployment application by using an image*.

The example YAML template is orchestrated as follows:

5. In the left-side navigation pane, choose Application > Pods, and select the target cluster and namespace.

Container Service - Kubernetes +	Pod	s											Refresh
Overview	Cluste	ers k8s-test	 Namespace 	default	• 3								
 Clusters 	Nam	ie 🖕	Status 🖕 ma	ax attempts	Pod IP 💠	Node 🔶	Time Created 🖕	CPU		Memory			
Clusters	not-	in-gpu-node	Running	0	-		12/06/2018,11:04:30	0			1.453 Mi	Details	More +
Nodes	tf-n	tebook-6b89fdb95-it6nh	Running	0			12/06/2018.11:14:35		0		40.191 Mi	Details	More -
Volumes			•						-				
	-												
Deployment													
StatefulSet													Cont
Job													tu
CronJob													
Pods 2													

Result

In the pod list, you can see that the two example pods have been scheduled to the target nodes, indicating you have implemented flexible scheduling by using GPU node labels.

1.5.8 View resource request and limit on nodes

The Container Service Console allows you to view resource usage of each node in a Kubernetes cluster.

Prerequisites

You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters > Nodes.

You can view the resource usage for the CPU and memory of each node, namely, the request and limit, which are calculated as follows:

- CPU request = sum (CPU request value from all pods on the current node) /total CPU of the current node.
- CPU limit= sum (actual CPU usage of all pods on the current node)/total CPU of the current node.
- Memory request = sum (memory request value from all pods on the current node) /total memory of the current node.
- Memory limit= sum (actual memory usage of all pods on the current node)/total memory of the current node.



• You can allocate loads to a node based on the resource usage on the node. For more information, see *Set node scheduling*.

• When both the request and limit on a node is 100%, no new pod is scheduled to the node.

Container Service - Kubernetes 🔻	Node List	Refresh Label Management Scale Ouster Add Existing Instance
Overview	Help: & Post 3 tance to Prepay	
Clusters	Clusters k8s-test Filter by labels	
Clusters	IP Address Role Instance ID/Name Configuration Pods(Allocated) Limit)	Memory(Request Limit) Update Time Action
Volumes	Pay-As-You- Go 12 26.25% 3.15 4 ecs.n4.xlarge	% 4.35% 53.87 % 09/18/2018,14:02:00 Scheduling Settings Monitor More →
Namespace	Pay-As-You- Go 7 21.25% 4.05	% 2.56% 53.01 % 09/18/2018,14:04:00 Scheduling Settings Monitor More→
 Application Deployment 	Master Pay-As-You- Go 7 21.25% 3.88 4 ecs.n4.xlarge	% 2.56% 54.15 % 09/18/2018,14:01:00 Scheduling Settings Monitor More→
StatefulSet Pods	Pay-As-You- Go 18 52.75% 2.13 Worker ecs.n4.xlarge 52.75% 2.13 52.75% 52.75	% 2.81% 30.70 % 09/18/2018,14:13:00 Scheduling Settings Monitor More→

1.5.9 Mount a disk to a Kubernetes cluster node

This topic describes how to mount a disk to a Kubernetes cluster node. Mounting a disk allows you to expand the Docker data directory and maintain a sufficient disk capacity when the number of containers or images that run on a node increases.

Prerequisites

Your Kubernetes cluster version must be v1.10.4 or later.

You can mount a disk to an existing Kubernetes cluster node by using either of the following methods:

- · If no disk is mounted to the existing node, see Mount a disk to the Docker data directory.
- If you have created a disk for the existing node, but you have failed to mount the disk to the node, you can follow these steps.



Note:

- We recommend that you create a snapshot of the target node or back up node data to avoid data loss.
- Additionally, you must ensure that you can schedule your cluster applications to other nodes.
- We recommend that you perform this operation during off-peak service hours to avoid disruptions to your business.

• Draining a node reschedules pods on the node to other nodes. Therefore, you must ensure that your Kubernetes cluster has sufficient nodes. We recommend that you add cluster nodes in advance as needed.

Before performing the operation, you need to determine whether a disk is already mounted to the target cluster node. To do so, run the df command on the target Worker node, and then check whether / var / lib / docker has been mounted to / dev / vdb1 . If the disk mounting operation failed, you can mount the disk by following these steps.

[root@	<u>C (MITTAN) A</u>	~]7	# df tn/feat	ure/v	alidate-slb-quo	tg _∈ and-op
Filesystem	1K-blocks	Used	Available	Use%	Mounted on	
/dev/vda1 ng:t	ro 41151808	2273772	36764604	6%	/	
devtmpfsing:t	roop 3995592 a	obing\$0	1 3995592	0%	/dev	
(tmpfsanch; feat	ure 4005096	easlb-Q	ot 4005096	0%	/dev/shmte-time	
tmpfsonanchais	up 4005096	wit 508	4004588	ur 1% /	drunate-slb-quo	
tmpfsnelions	4005096	0	4005096	0%	/sys/fs/cgroup	Music
/dev/vdb1 com	101441464	g 61668	96120584	1%	/var/lib/docke	^r Pictures
tmpfslaobing:t	roop 801020 1	obina\$0	801020	0%	/run/user/0	Public

- 1. Set the target node as unschedulable. For more information, see *Mark node as unschedulable*.
- 2. Drain the target node. For more information, see Safely drain a node.

3. Remove the target node. This topic uses the Container Service console as an example.

- a. Log on to the Container Service console.
- b. In the left-side navigation pane, click Node.
- c. Select the target node, and click Remove or choose More > Remove.

<	Node List Refresh Label Management Scale Cluster Add Existing Instance
Basic Information	Help: & Postpay Instance to Prepay & Node exception & Node monitoring and alarms & Collect Kubernetes diagnostics information
Node List	Clusters k8s-test * Filter by Labels -
Event List Cluster Audit	Instance ID/Name/IP Address Role/Status Configuration (Allocated) Request/Limit CPU Usage Request/Limit Memory Usage Kubelet Version Update Time Action
	Master Pay-As-You-Go 8 23.75% 5.83% y1.11.5 01/31/2019,14:35:31 Monitor More ● Running ecs.n1.large 8 0.00% 6.43 % 9.10% 65.14 % v1.11.5 01/31/2019,14:35:31 Monitor More
	Master Pay-As-You-Go ● Running ecs.n1.large 14 30.00% 5.63 % 8.90% 14.73% 72.25 % V1.11.5 01/31/2019,14:37:48 Monitor More-
3	Master Pay-As-You-Go 8 23.75% 5.83% √1.11.5 01/31/2019,14:39:52 Monitor More €
	Worker Pay-As-You-Go 10 5.00% 5.83% V1.11.5 01/31/2019,14:42:26 Monitor Monitor • Running • ecs.n1.large 10 5.00% 2.93 % 9.10% 52.34 % V1.11.5 01/31/2019,14:42:26 Monitor Monitor
	Worker Pay-As-You-Go 5 10.00% 9.10% 9.10% V1.11.5 01/31/2019.14* © Running ecs.n1.large 5 1.78 % 9.10% 56.75 % V1.11.5 01/31/2019.14* Scheduling Settings
	Worker Pay-As-You-Go 5 10.00% 9.10% 9.10% V1.11.5 01/31/2019,14:42:27 Monitor More Running ecs.n1.large 5 0.00% 2.38 % 9.10% 58.07 % V1.11.5 01/31/2019,14:42:27 Monitor More
	Remove Total: 6 item(s), Per Page: 25 item(s) « < 1 > >

d. In the displayed Remove Node dialog box, click OK.

Remove Node	\times
Are you sure you want to delete the following 1 nodes?	
Notes:	
 This operation may affect your business. We recommend that you perform this operation during the non-rush hours. For more information, see Help Documentation. After you have removed a node, its containers will not be automatically migrated. Make sure that you have backed up the required data. The drain operation may cause the pods in the node to be automatically recovered in other nodes in the cluster. Make sure the cluster resources are sufficient. Only Pay-As-You-Go-based ECS instances will be released. Other ECS instances will still be hilled. 	
OK Cancel	



We recommend that you do not select the Release ECS at the same time check box. Otherwise, the ECS instance used by the target node will be released.

- 4. Add the removed node to the cluster.
 - a. In the left-side navigation pane, click Clusters.
 - b. On the right of the target cluster, choose More > Add Existing Instance.

Container Service - Kubernetes 🕶	Cluster List		Y	'ou can create up t	o 100 clusters a	ind can add u	p to 1000 nodes in each	cluster. Refresh	Create Kubernetes Cluster	
Overview	Help: Ø Create cluster Ø Create G creation failures Ø Authorization man	PU clusters 🔗 Sc agement	ale cluster 🔗 Conn	ect to Kubernetes	cluster via kube	ecti 🔗 Mana	ge applications with com	imands 🔗 Cluste	r planning 🛛 🔗 Troubleshoot cluste	er
▼ Clusters	Name 🔻									
Clusters 1 Nodes	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Number of Nodes	Time Created	Version	Action	n
Volumes	test01	Serverless Kubernetes	China East 1 (Hangzhou)	VPC	Running		01/31/2019,16:49:30	1.9.7	Manage View Logs Delete Use Cloud Shell	
Namespace										
Authorization	k8s-test	Kubernetes	China East 1 (Hangzhou)	VPC	Running	6	01/31/2019,14:28:32	1.11.5	Manage View Log 2 Dashboar Scale Cluster <u>More</u>	
Deployment	kubernetes-test	Kubernetes	China East 1 (Hangzhou)	VPC	Running		02/21/2018,20:56:15	1.1 Delete	ng Instance 3	
								Automatic	Scaling	0
DaemonSet								Addon Up	trade	ontad
Job								Doplay Ist	ia	ŝ
CronJob								Cluster Ch	eck	
Pods								Cluster Au	dit	
Valueses Claim								Use Cloud	Shell	
volumes Claim								Collect Ku	bernetes diagnostics information	
Release										

c. Select Automatically Add or Manually Add. In this example, the instance is added automatically.

Container Service - Kubernetes 👻	Add Existing ECS	5 Instance - k8s-te	est € Back to Clust	er List					
Overview	Sele	ct Existing ECS In	stance(s)	Enter Inst	ance Information	Ac	ded Successfully		
 Clusters 	Adding Method:								
Clusters	Automatically #	Add 🔍 Manua	lly Add						
Nodes	Adding an existing EC	CS instance to the Co	ntainer Service will cha	nge the instance system disk. The	disk ID will change and the	previous system disk will be release	d.		
Volumes	Note: 1. Your previous sy	/stem disk's user sna	pshots will be retained.	Automatic snapshots will be reta	ned or deleted according to	the setting of the "Delete automatic	snapshots when releasi	ng disk" option.	
Namespace	You can go to the 2. To retain enough	disk list and click "Mo h snapshot quota for	odify Attributes" to view the periodic automatic	and modify attribute values. snapshot policy, you can delete u	nwanted user and automatic	: snapshots.			
Authorization	3. Back up data be	3. Back up data before performing this operation to avoid data loss.							
 Application 	Select Existing ECS	Instance(s):							
Deployment	Instance ID 🔹	Enter the instance	e ID for exact search.	Search					
StatefulSet	Instance ID	1	instance Name		IP Address	Zone	Network Type	Instance Type	
DaemonSet									
Job	does not have	nat the node public network an attached ETP				China East 1 (Hangzhou) Zon	-B VPC	ers n1 Jarne	
CronJob	Adding an inst cluster will fail	tance to the as you are not			10.000.000	China Last 1 (Hang2100) 201	to vic	contraininge	
Pods	connected to t	the network.]							
Volumes Claim									
Release								Next Step	

- d. Select the existing instance and then click Next Step.
- e. Turn on the Format Data Disk switch.



f. Complete other required settings.

After the node has been added to the cluster, you can log on to the node to run the df command to check whether a disk has been mounted to the target node.

The following figure shows the disk has been amounted to the target node.

[root@	<u>n hann for in</u>	~]#	fr df in/feat			
Filesystem	1K-blocks	Used	Available	Use%	Mounted on	
/dev/vda1_ng	tro 41151808	2273772	36764604	6%	/	
devtmpfsing	troo 3995592	obing\$ 0	3995592	0%	/dev	
tmpfisanch; fea	ture 4005096	e-slb-Qu	+ 4005096	0%	/dev/shm e-tim	
tmpfsmanchai	s up 4005096	1wit 508	4004588	ur 1% /	⟨ run ate-slb-qu	
tmpfsmel	4005096	0	4005096	0%	/sys/fs/cgroup) Music
/dev/vdb1 col	101441464	g 61668	96120584	1%	/var/lib/docke	er Picture
tmpfs.aobina:	troop 801020	obing\$0	801020	0%	/run/user/0	Public

1.5.10 Mount a disk to the Docker data directory

This topic describes how to mount a disk to the Docker data directory. If the number of containers or images that run on an ECS instance increases constantly, the ECS instance disk capacity may be insufficient. In this case, you can expand the Docker data directory by mounting a disk to the ECS instance.

Docker data directory

Docker data is stored in disks through a union file system (UnionFS). The default container data and image data of Docker is stored in the / var / lib / docker directory. You can run the du command to view the disk space size occupied by this directory.

du - h -- max - depth = 0 / var / lib / docker

7.9G / var / lib / docker

Scenarios

Generally, a Docker image occupies a large amount of disk space. If you want to use multiple Docker images or a large number of containers, you must mount a disk to the Docker data directory to ensure sufficient disk capacity is available.

Mount a disk

To mount a disk to the Docker data directory, follow these steps:

- 1. Create a disk and mount it to the target ECS instance for which you want to expand the disk capacity.
 - a. Log on to the ECS console to create a disk.
 - b. In the left-side navigation pane, click Instances.
 - c. Click the target ECS instance ID.
 - d. In the left-side navigation pane, click Disks.
 - e. In the upper-right corner, click Mount.
 - f. In the displayed dialog box, select the created disk from the target disk dropdown list, and then click OK.
 - g. Click Mount to mount the new disk to the target ECS instance, and record the new disk mounting point which is in the format of / dev / xvd * or / dev /

vd *.

- 2. Log on to the target ECS instance to format the new disk.
 - a. Run the ls l / dev / xvd * or ls l / dev / vd * command to verify whether a disk that has the recorded mounting point has been mounted to the ECS instance.
 - b. Run the fdisk command to partition the new disk, and then run the mkfs . ext4 command to format the new disk.

root@c836831d69e4040e797eff4d3c4dcd983-node2:~# ll /dev/xvd* brw-rw---- 1 root disk 202, 0 May 26 15:44 /dev/xvda brw-rw---- 1 root disk 202, 1 May 26 15:44 /dev/xvda1 brw-rw---- 1 root disk 202, 16 May 27 13:03 /dev/xvdb root@c836831d69e4040e797eff4d3c4dcd983-node2:~# fdisk -S 56 /dev/xvdb Device contains neither a valid DOS partition table, nor Sun, SGI or OSF disklabel Building a new DOS disklabel with disk identifier 0x446953ae. Changes will remain in memory only, until you decide to write them. After that, of course, the previous content won't be recoverable. Warning: invalid flag 0x0000 of partition table 4 will be corrected by w(rite) Command (m for help): n Partition type: primary (0 primary, 0 extended, 4 free) р extended e Select (default p): p Partition number (1-4, default 1): 1 First sector (2048-62914559, default 2048): Using default value 2048 Last sector, +sectors or +size{K,M,G} (2048-62914559, default 62914559): Using default value 62914559 Command (m for help): wq The partition table has been altered! Calling ioctl() to re-read partition table. Syncing disks. root@c836831d69e4040e797eff4d3c4dcd983-node2:~# ll /dev/xvd* brw-rw---- 1 root disk 202, 0 May 26 15:44 /dev/xvda brw-rw---- 1 root disk 202, 1 May 26 15:44 /dev/xvda1 brw-rw---- 1 root disk 202, 16 May 27 13:08 /dev/xvdb brw-rw---- 1 root disk 202, 17 May 27 13:08 /dev/xvdb1 root@c836831d69e4040e797eff4d3c4dcd983-node2:~# mkfs.ext4 /dev/xvdb1 ke2fs 1.42.9 (4-Feb-2014) ilesystem label= OS type: Linux Block size=4096 (log=2) Fragment size=4096 (log=2) Stride=0 blocks, Stripe width=0 blocks 1966080 inodes, 7864064 blocks 393203 blocks (5.00%) reserved for the super user First data block=0 Maximum filesystem blocks=4294967296 240 block groups 32768 blocks per group, 32768 fragments per group 8192 inodes per group Superblock backups stored on blocks: 32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208, 4096000 Allocating group tables: done Vriting inode tables: done Creating journal (32768 blocks): done Writing superblocks and filesystem accounting information: done

3. Migrate the Docker data to the new disk.

If you do not want to suspend the applications that run on the target ECS instance, you must migrate the applications. For how to migrate applications on a Swarm cluster, see *Schedule an application to specified nodes*. For how to migrate applications on a Kubernetes cluster, see *Safely drain a node while respecting application SLOs*.

- a. To ensure that data can be migrated, run the service docker stop command to stop Docker daemon, and run the service kubelet stop command to stop kubelet.
- b. Migrate the Docker directory data to a backup directory. For example, mv /
 var / lib / docker / var / lib / docker_dat a .
- c. Mount the new disk to the / var / lib / docker and / var / lib /
 kubelet directories. For example,

```
echo "/ dev / xvdb1 / var / lib / container / ext4
defaults 0 0 " >>/ etc / fstab
echo "/ var / lib / container / kubelet / var / lib / kubelet
none defaults , bind 0 0 " >>/ etc / fstab
echo "/ var / lib / container / docker / var / lib / docker
none defaults , bind 0 0 " >>/ etc / fstab
mkdir / var / lib / docker
mount - a
```

d. Migrate the backed up Docker data to the new disk. For example, mv / var /

lib / docker_dat a /* / var / lib / docker /.

- 4. Start the Docker daemon and kubelet, and check the data location.
 - a. Run the service docker start command to start the Docker daemon, and run the service kubelet start command to start kubelet.
 - b. Run the df command to verify whether / var / lib / docker has been mounted to the new disk. If you need to start the Kubernetes cluster, skip this step.

root@c836831d6	9e4040e797e	eff4d3c4	dcd983-node	e2:/v	ar/lib# df
Filesystem	1K-blocks	Used	Available	Use%	Mounted on
udev	497280	4	497276	1%	/dev
tmpfs	101628	712	100916	1%	/run
/dev/xvda1	41151808	1928420	37109960	5%	1
none	4	0	4	0%	/sys/fs/cgroup
none	5120	0	5120	0%	/run/lock
none	508136	288	507848	1%	/run/shm
none	102400	0	102400	0%	/run/user

c. Run the docker ps command to check whether containers are lost. Restart containers as needed. For example, you can restart a container that has not been set the restart : always label.

root@c836831d69e4040)e797eff4d3c4dcd983-node2:/var/lib# docker ps			
CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS
4f564091bffa	registry.aliyuncs.com/acs/logspout:0.1-41e0e21	"/bin/logspout"	21 hours ago	Up 3 minutes
gspout_2				
a5aba5fbedae	registry.aliyuncs.com/acs/ilogtail:0.9.9	"/bin/sh -c 'sh /usr/"	21 hours ago	Up 3 minutes
gtail_2				
5e3d8fe154bb	registry.aliyuncs.com/acs/monitoring-agent:0.7-1cf85e6	"acs-mon-run.shhel"	21 hours ago	Up 3 minutes
_acs-monitoring-agen	1t_1			
fb72c2388b0e	registry.aliyuncs.com/acs/volume-driver:0.7-252cb09	"acs-agent volume_exe"	21 hours ago	Up 3 minutes
er_volumedriver_2				
604fcb4ad720	registry.aliyuncs.com/acs/routing:0.7-c8c15f0	"/opt/run.sh"	21 hours ago	Up 3 minutes
uting_1				
8fe1d6ed15b5	registry.aliyuncs.com/acs/agent:0.7-6967e86	"acs-agent joinnod"	21 hours ago	Up 3 minutes
999da3883264	registry.aliyuncs.com/acs/tunnel-agent:0.21	"/acs/agent -config=c"	21 hours ago	Up 3 minutes

5. If a container has been migrated to other nodes, you can schedule it back to the target node to which you mounted the new disk.

For more information, see Container Service.

1.6 Namespace management

1.6.1 Create a namespace

This topic describes how to create a namespace.

Prerequisites

You have created a Kubernetes cluster. For more information, see Create a Kubernetes

cluster.

Context

In a Kubernetes cluster, you can use namespaces to create multiple virtual spaces. When a large number of users share a cluster, multiple namespaces can be used to effectively divide different work spaces and assign cluster resources to different tasks. Furthermore, you can use *resource quotas* to assign resources to each namespace.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Namespace.
- 3. Select the target cluster, and then click Create in the upper-right corner.

Container Service - Kubernetes +	Namespace				Refr	esh	Create
Overview	Help: ${\mathscr S}$ Configure resource quotas for namespaces						4
▼ Clusters	Clusters k8s-cluster •						
Clusters 1	Name Tag	Status	Time Created				Action
Nodes	default	Ready	02/05/2019,15:58:07	ResourceQuota and LimitRang	e	Edit	Delete
Volumes	kube-public	Ready	02/05/2019,15:58:07	ResourceQuota and LimitRang	e l	Edit	Delete
Namespace 2	kube-system	Ready	02/05/2019,15:58:07	ResourceQuota and LimitRang	ə İ	Edit	Delete

4. In the displayed dialog box, set a namespace.

Create Namespace							
Name	test						
	1-63 charact letters, and " beginning an	ers, can only contain number -", and can only be letters or d end	s, lower case numbers at the				
Tags	Variable Na	me Variable Value	Action				
	env	test	Edit Delete				
	Name	Value	Add				
			OK Cancel				

- Name: Enter a name for the namespace name. The name must be 1 to 63 characters in length and can contain numbers, letters, and hyphens (-). It must start and end with a letter or number. In this example, test is used as the name.
- Tags: Add one or multiple tags to the namespace to identify the characteristics of the namespace. For example, you can set a tag to identify that this namespace is used for the test environment.

You can enter a variable name and a variable value, and then click Add on the right to add a tag to the namespace.

- 5. Click OK.
- 6. The namespace named test is displayed in the namespace list.

Container Service - Kubernetes +	Namespace		Refresh Create
Overview	Help: 🔗 Configure resource quotas for namespaces		
 Clusters 	Clusters k8s-cluster •		
Clusters	Name Tag	Status Time Created	Action
Nodes	default	Ready 02/05/2019,15:58:07 Re	sourceQuota and LimitRange Edit Delete
Volumes	kube-public	Ready 02/05/2019,15:58:07 Re	sourceQuota and LimitRange Edit Delete
Namespace	kube-system	Ready 02/05/2019,15:58:07 Re	sourceQuota and LimitRange Edit Delete
Authorization • Application	test env: test	Ready 02/11/2019,16:50:04 Re	sourceQuota and LimitRange Edit Delete

1.6.2 Set resource quotas and limits for a namespace

This topic describes how to set resource quotas and limits for a namespace through the Container Services console.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a namespace. In this topic, a namespace named test is used. For more information, see *Create a namespace*.
- You have connected to the Master node of the cluster. For more information, see Connect to a Kubernetes cluster by using kubectl.

Context

By default, a running pod uses the CPU and memory resources of nodes without limit . That is, any pod can use the computing resources of the cluster without restraint. Therefore, pods of a namespace may deplete the cluster resources.

Namespaces can be used as virtual clusters to serve multiple users. Therefore, setting resource quotas for a namespace is regarded as a best practice.

For a namespace, you can set the quotas of resources, such as CPU, memory, and number of pods. For more information, see *Resource quotas*.

Procedure

- 1. Log on to the Container Service console.
- In the left-side navigation pane under Kubernetes, choose Clusters > Namespace. Select the target cluster and click ResourceQuota and LimitRange on the right of the test namespace.

Container Service - Kubernetes -	Namespace	Ref	resh Create
Overview	Help: ${\mathscr O}$ Configure resource quotas for namespaces		
Clusters	Clusters k8s-cluster		
Clusters	Name Tag	Status Time Created	Action
Nodes	default	Ready 02/05/2019,15:58:07 ResourceQuota and LimitRange	Edit Delete
Volumes	kube-public	Ready 02/05/2019,15:58:07 ResourceQuota and LimitRange	Edit Delete
Namespace 2	kube-system	Ready 02/05/2019,15:58:07 ResourceQuota and LimitRange	Edit Delete
Authorization Application	test env: test	Ready 02/11/2019,16:50:04 ResourceQuota and LimitRange	Edit Delete

3. In the displayed dialog box, set resource quotas and default resource limits.



After setting CPU/memory quotas for a namespace, you must specify CPU/memory resource limits or set the default resource limits for the namespace when creating a pod. For more information, see *Resource quotas*.

a) Set resource quotas for the namespace.

esourceQuota and	l LimitRange				2
Tip: After setting the configuring Pods, o	ne CPU/memory quota or configure the defau	a (ResourceQue It resource limi	ota) for the namespace t (LimitRange) for the r	, you must specify the CPU/memory resource limit when namespace. For details, please refer to: Resource Quotas	
ResourceQuota	LimitRange				
^ Compute Resou	rce Quota				
🗹 CPU Limit		Total	2	Core(s)	
 Memory Limit 		Total	4Gi	0	
^ Storage Resourc	ce Quota				
🖉 storage		Total	1024Gi	0	
🖉 persistentvolu	meclaims	Total	50		
^ Object Count Qu	iota				
 configmaps 		Total	100		
🖉 pods		Total	50		
services		Total	20		
 services.loadb 	alancers	Total	5		
 secrets 		Total	10		
				04	
				Caller Caller	-

b) To control the amount of resources consumed by containers, set resource limits and resource requests for containers in this namespace. For more information, see https://kubernetes.io//memory-default-namespace/.

ResourceQue	ota and LimitRange		×
Tip: After s configuring	etting the CPU/memory quota (Pods, or configure the default (ResourceQuota) for the namespa resource limit (LimitRange) for th	ce, you must specify the CPU/memory resource limit when e namespace. For details, please refer to: Resource Quotas
ResourceQ	uota LimitRange		
	CPU		Memory 🔞
Limit	0.5	Core(s)	512Mi
Request	0.1	Core(s)	256Mi
			OK Cancel

4. Connect to the Master node and then run the following commands to view the resources of the test namespace:

```
get limitrange, ResourceQu ota - n test
# kubectl
NAME AGE
limitrange / limits
                   8m
NAME AGE
resourcequ ota / quota 8m
          describe limitrange / limits resourcequ ota /
# kubectl
quota – n test
Name : limits
Namespace : test
Type Resource Min Max Default Request Default
                                                   Limit
Max
     Limit / Request Ratio
--- ----- --- --- ------ ------
 _____
Container cpu - - 100m 500m -
Container memory - - 256Mi 512Mi -
Name : quota
Namespace : test
Resource Used Hard
 _____ ____
configmaps 0 100
limits . cpu 0 2
limits . memory 0 4Gi
persistent volumeclai ms 0 50
pods 0 50
requests . storage 0 1Ti
secrets 1 10
services 0 20
```

services . loadbalanc ers 0 5

1.6.3 Edit a namespace

This topic describes how to edit a namespace.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a namespace. In this topic, a namespace named test is used. For more information, see *Create a namespace*.

Context

Editing a namespace means to add, modify, or delete the details of a namespace tag.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Namespace.
- 3. Select the target cluster and then click Edit on the right of the target namespace tag.

Namespace		Refresh Create
Clusters k8s-cluster 🔻 3		
Name Tag	Status Time Created	Action
default	Ready 08/24/2018,10:02:37	ResourceQuota and LimitRange Edit Delete
Istio-system istio-injection: disabled	Ready 08/30/2018,18:03:28	ResourceQuota and LimitRange Edit Delete
kube-public	Ready 08/24/2018,10:02:37	ResourceQuota and LimitRange Edit Delete
kube-system	Ready 08/24/2018,10:02:37	ResourceQuota and LimitRange 4 Delete
test env: test	Ready 09/04/2018,10:47:21	ResourceQuota and LimitRange Edit Delete
	Namespace Custers Kas-cluster Tag default Istio-system istio-injection: disabled kube-public kube-system test	Namespace Cutters Satus Time Created Mane Tag Status Time Created default • Ready 08/34/2018,10:02:37 isto-system • Beady 08/30/2018,18:03:28 kube-system • Ready 08/24/2018,10:02:37 test • Ready 08/24/2018,10:02:37

4. In the displayed dialog box, click Edit to modify the namespace tag. For example, change the tag to env : test - V2 and click Save.

Edit Namespa	ce	\times
Name	test	
	1-63 characters, can only contain numbers, lower case letters, and "-", and can only be letters or numbers at the beginning and end	
Tags	Variable Name Variable Value Action	
	env test-V2 Save Delete	
	Name Value Add	
	OK Cano	el

5. Click OK. The edited namespace tag is then displayed in the namespace list.

Namespace				Refr	esh	Create
Clusters k8s-cluster •						
Name Tag	Status	Time Created				Action
default	Ready	08/24/2018,10:02:37	ResourceQuota and LimitRang	e	Edit	Delete
istio-system istio-injection: disabled	Ready	08/30/2018,18:03:28	ResourceQuota and LimitRang	e	Edit	Delete
kube-public	Ready	08/24/2018,10:02:37	ResourceQuota and LimitRang	e	Edit	Delete
kube-system	Ready	08/24/2018,10:02:37	ResourceQuota and LimitRang	e	Edit	Delete
test env: test-V2	Ready	09/04/2018,10:47:21	ResourceQuota and LimitRang	e	Edit	Delete

1.6.4 Delete a namespace

This topic describes how to delete a namespace you no longer require.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a namespace. In this topic, a namespace named test is used. For more information, see *Create a namespace*.

Context



Deleting a namespace also deletes all of its resource objects. Exercise caution when performing this action.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Namespace.
- 3. Select the target cluster and then click Delete on the right of the cluster.

Container Service - Kubernetes +	Namespace		Refresh Create
Overview	Help: ${\mathscr S}$ Configure resource quotas for namespaces		
Clusters	Clusters k8s-cluster V		
Clusters	Name Tag	Status Time Created	Action
Nodes	default	Ready 02/05/2019,15:58:07	ResourceQuota and LimitRange Edit Delete
Volumes	kube-public	Ready 02/05/2019,15:58:07	ResourceQuota and LimitRange Edit Delete
Namespace 2	kube-system	Ready 02/05/2019,15:58:07	ResourceQuota and LimitRange Edit Delete
Authorization Application	test env: test-v2	Ready 02/11/2019,16:50:04	ResourceQuota and LimitRange Edit Delete

4. In the displayed dialog box, click Confirm.

Note		×	
0	Are you sure to delete test ?		
	Confirm	Cancel	

5. The namespace is then deleted from the namespace list, and its resource objects are also deleted.

1.7 Service catalog management

1.7.1 Overview

Applications running on the cloud platform need some basic services such as databases, application servers, and other generic basic softwares. For example, a WordPress application, as a Web application, needs a database service (such as MariaDB) in the backend. Traditionally, you can create the MariaDB service on which the application depends in the WordPress application orchestration, and integrate the MariaDB service with the Web application. To develop applications on the cloud in this way, developers must spend time and energy deploying and configuring the dependent infrastructure softwares, which increases the costs of hosting and migrating applications.

Alibaba Cloud Container Service supports and integrates with the service catalog function. The service catalog function aims to access and manage the service brokers , which allows applications running in Kubernetes clusters to use the managed services offered by service brokers. A series of infrastructure softwares are supported by the service catalog function, which allows the developers to use these softwares as services and focus on the applications, the core of the development, without concerning about the availability and scalability of the softwares or managing the softwares.

The service catalog uses the Open service broker API of Kubernetes to communicate with service brokers, acting as an intermediary for the Kubernetes API server to negotiate the initial provisioning and obtain the credentials necessary for the applications to use the managed services. For more information about the implementation principle of the service catalog, see *Service catalog*.

1.7.2 Enable service catalog function

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > Service Catalog in the left-side navigation pane. Select the cluster from the Cluster drop-down list in the upper-right corner.
- 3. If you have not deployed the service catalog, click to install the service catalog as instructed on the page.

Container Service - Kubernetes 🝷	Service Catalog
Overview	Clusters test-ma 7 3
Clusters	
Application	 Service Catalog has not been deployed in this cluster, Please click to deploy
▼ Store 1	4
Docker Images	
Orchestration Templa	
App Catalog	
Service Catalog 2	

4. After the installation, the service broker, which is installed by default, is displayed on the Service Catalog page. You can click the mariadb-broker to view the details.



The service catalog is implemented as an extension API server and a controller. After Alibaba Cloud Container Service installs the service catalog function, the namespace catalog is created.



5. Click Clusters in the left-side navigation pane. Click Dashboard at the right of a cluster.

Container Service - Kubernetes ▼	Cluster List			You can create up	to 5 clusters an	d can add up to 40 nodes	s in each cluster.	Refresh	Create Kubernetes Cluster
Overview	Help: 🖉 Create cluster 🛛 & Scale clust	er 🔗 Connec	t to Kubernetes cluster vi	a kubecti 🔗 Manag	ge applications	with commands			
▼ Clusters	Name 🔻								
Clusters	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
Volumes	test-yxm cl	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1lkyevdjj	Running	09/17/2018,15:03:04	1.11.2		Manage View Logs Dashboard Scale Cluster More→
Authorization	test-mia Ca	Kubernetes	China East 1 (Hangzhou)	VPC vpc- bp1lkyevdjj	Running	09/17/2018,11:37:55	1.11.2		Manage View Logs Dashboard Scale Cluster More

6. In the Kubernetes dashboard, select catalog as the Namespace in the left-side navigation pane. You can see the resource objects related to catalog apiserver and controller are installed under this namespace.

<	= Overview		+ CREATE
Cluster	Workloads		
Namespaces	Workloads Statuses		
Nodes Persistent Volumes Roles Storage Classes	NO DOM	100.00%	100 00%
Namespace	Deployments	Pods	Replica Sets
catalog - 1	Deployments		
Workloads	Name Labels	Pods Age	a ≑ Images
Cron Jobs	 app: aliacs-service-catalog-catalog-apiserver chart: catalog-0.1.9 release: aliacs-service 	talog-catalog-apiserver heritage: Tiller 1 / 1 09/ e-catalog	/17/2018, 15:15:08 registry.cn-hangzhou.aliyuncs.com/service-ca : registry.cn-hangzhou.aliyuncs.com/service-ca :
Deployments Jobs	 app: aliacs-service-catalog-catalog-controller-mana; chart: catalog-0.1.9 release: aliacs-service 	talog-catalog heritage: Tiller 1 / 1 09/ e-catalog	/17/2018, 15:15:08 registry.cn-hangzhou.aliyuncs.com/service-ca :
Pods Replica Sets	Pods		
Replication Controllers	Name \$ Node	Status \$ Restarts #	Age CPU (cores) Memory (bytes)
Stateful Sets	 aliacs-service-catalog-catalog-apiserver-7b7b7b9f78 cn-hangzhou.i- bp1hk1m08e5rx 	kgrae8c Running 0	09/17/2 15:15:0 0.013 26.871 Mi ≡ :
Discovery and Load Bal	 aliacs-service-catalog-catalog-controller-manager-5c cn-hangzhou.i- bp1hk1m08e5rx 	kgrae8a Running O	09/17/2 15:15:0 0.006 ■ 11.652 Mi = :

What's next

Then, you have successfully enabled the service catalog function. You can create a managed service by using the service broker in the service catalog, and apply the managed service to your applications.

1.8 Application management

1.8.1 Create a deployment application by using an image

This topic describes how to use an image to create a deployment application. In this topic, an Nginx application that is accessible to the Internet is created.

Prerequisites

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Application > Deployment, and then click Create by Image in the upper-right corner.

3. Set Name, Cluster, Namespace, Replicas, Type, Tag, and Annotation. The replicas parameter indicates the number of pods contained in the application. Then click Next.



In this example, you need to select the Deployment type.

If you do not set Namespace, the system automatically uses the default namespace.

eate Applicatior	1					
Basic	Information	Container	>	Advanced	>	Done
Name:	nginx		han lawar nan Fastisi		tent with a lovelage	
Clusters	k8s-test	aracters iong, and can contain hun	bers, lower case Englis	netters and hypnens, but cannot s	uart with a hypnen.	
Namespaces	default	٣				
Replicas:	2					
Туре	Deployments	¥				
Tag	• Add					
Annotation	O Add					
						Back Next

4. Configure a container.

Note:

You can configure multiple containers for the pod of the application.

- a) Set general container parameters.
 - Image Name: Click Select image to select the image in the displayed dialog box and then click OK. In this example, select the Nginx image.

You can also enter a private registry in the format of domainname / namespace / imagename : tag to specify an image.

- Image Version: Click Select image version to select a version. If you do not select an image version, the system uses the latest version by default.
- Always pull image: Container Service caches the image to improve deployment efficiency. During deployment, if the tag of the newly specified image is the same as that of the cached image, Container Service reuses the cached image, instead of re-pulling the same image. Therefore, if you do not

modify the image tag when changing your code and image, the early image in the local cache is used in the application deployment. If you select this check box, Container Service ignores the cached image and re-pulls an image when deploying the application to make sure the latest image and code are always used.

- Image pull secret: Create a Secret for the image. A secret is required to pull a image from a private image repository. For more information, see *Use an image Secret*.
- Resource Limit: Specify the upper limit for the resources (CPU and memory) that can be used by this application to avoid occupying excessive resources.
 CPU is measured in millicores, that is, one thousandth of one core. Memory is measured in bytes, which can be Gi, Mi, or Ki.
- Resource Request: Specify how many resources (CPU and memory) are reserved for the application. These resources can be set to be exclusive to the container by using this parameter. If you do not set this parameter, other

services or processes will compete for resources. Then the application may become unavailable due to resource shortage.

• Init Container: Select this check box to create an Init Container that contains useful tools. For more information, see *Init containers*.

	Basic Informa	tion Con	tainer	Advanced
С	ontainer1 O Add Cont	ainer		
	Image Name:	nginx	Select image	
	Image Version:	latest	Select image version	
		Always pull image Image pull secret		
General	Resource Limit:	CPU eg : 500m Core Memory	eg : 128Mi MiB	
	Resource Request:	CPU 500m Core Memory	eg : 128Mi MiB	
	Init Container			

b) Optional: Set environment variables.

You can use key-value pairs to set environment variables for the pods. Environment variables are used to add environment labels or pass configurations for the pods. For more information, see *Pod variable*.

c) Optional: Set health checks.

You can set liveness probes and readiness probes. Liveness probes are used to detect when to restart the container. Readiness probes determine if the container is ready to receive traffic. For more information about health checks, see *Configure liveness and readiness probes*.

	Liveness	Enable				
		нттр		тср	Command	~
		Protocol	HTTP	*		
		path				
		Port				
		Http Header	name			
			value			
		Initial Delay	3			
		Period	10			
		Timeout	1			
		Success Threshold	1			
		Failure Threshold	3			
eck			5			
Ith ch	Dendingen					
Hea	Readiness	Enable				
		НТТР		ТСР	Command	*
		Protocol	HTTP	•		
		path				
		Port				
		Http Header	name			
			value			
		Initial Delay	3			
		Period	10			
		Timeout	1			
		Timeout	1			

Request method	Description
HTTP request	With this health check method, you can send an HTTP GET request to the container. The following parameters are supported:
	• Protocol: HTTP/HTTPS.
	 Protocol: HTP/HTPS. Path: the path to access the HTTP server. Port: the number or name of the access port exposed by the container. The port number must be in the range of 1 to 65535. HTTP Header: custom headers in the HTTP request. HTTP allows repeated headers. You can use a key -value pair to set an HTTP Header. Initial Delay (in seconds): the initialDelaySeconds parameter, indicating the number of seconds for which the first probe must wait after the container is started. The default value is 3. Period (in seconds): the periodseco nds parameter, indicating the number of seconds for which the first probes are performed. The default value is 10. The minimum value is 1. Timeout (in seconds): the timeoutSeconds parameter, indicating the number of time that the probe has timed out. The default value is 1 and the minimum value is 1. Success Threshold: The minimum number of consecutive successful probes needed for determining a probe success after a failed probe. The default value is 1 and the minimum value is 1. Failure Threshold: The minimum number of consecutive failed probe secutive failed probes.
	probe. The default value is 3. The minimum value is 1.

Request method	Description
TCP connection	If you use this health check method, a TCP socket is sent to the container. The kubelet then attempts to open the socket of the container on a specified port. If a connection can be established, the container is considered healthy. If not, it is considered unhealthy. The following parameters are supported:
	 Port: the number or name of the access port exposed by the container. The port number must be in the range of 1 to 65535. Initial Delay (in seconds): the initialDelaySeconds parameter, indicating the seconds for the first liveness or readiness probe must wait for after the container is
	 started. The default value is 15. Period (in seconds): the periodseco nds parameter, indicating the interval at which probes are performed. The default value is 10. The minimum value is 1. Timeout (in seconds): the timeoutSeconds parameter, indicating the number of time that the probe has timed out. The default value is 1 and the minimum value is 1
	 Success Threshold: The minimum number of consecutive successful probes needed for determining a probe success after a failed probe. The default value is 1 and the minimum value is 1. It must be set to 1 for a liveness probe. Failure Threshold: The minimum number of consecutive failed probes needed for determining a probe failure after a successful probe. The default value is 3. The minimum value is 1.

Request method	Description
Command line	With this heath check method, you can detect the container health by executing a probe detection command in the container. The following parameters are supported:
	 Command: a probe command used to detect the container health. Initial Delay (in seconds): the initialDelaySeconds parameter, indicating the number of seconds for which the first liveness or readiness probe must wait after the container is started. The default value is 5.
	 Period (in seconds): the periodseco nds parameter, indicating the interval at which probes are performed. The default value is 10. The minimum value 1. Timeout (in seconds): the timeoutSeconds parameter, indicating the number of time that the probe has timed out. The default value is 1 and the minimum
	 Success Threshold: The minimum number of consecutive successful probes needed for determining a probe success after a failed probe. The default value is 1 and the minimum value is 1. It must be set to 1 for a liveness probe. Failure Threshold: The minimum number of consecutive failed probes needed for determining a probe failure after a successful probe. The default value is 3. The minimum value is 1.

d) Set life cycle rules.

You can set the following parameters for the container life cycle: start, post start, and pre-stop. For more information, see *Attach handlers to container lifecycle events*.

- · Start: Set a pre-start command and parameter for the container.
- Post Start: Set a post-start command for the container.
- Pre Stop: Set a pre-stop command for the container.

	Start:	Command ["/bin/sh","-c,"echo Hello > /user/share/message"]
ycle		Parameter
Life c	Post Start:	Command
	Pre Stop:	Command ["/user/sbin/nginx","-s","quit"]

e) Optional: Set volumes.

You can configure local storage and cloud storage.

- local storage: Supported storage types include HostPath, ConfigMap, Secret, and EmptyDir. By setting a type of local storage, you can mount its mount source to the container path. For more information, see *Volumes*.
- cloud storage: Supported types of cloud storage include disks, Network Attached Storage (NAS), and Object Storage Service (OSS).

This example sets a disk as the volume and mounts the disk to the / tmp container path. Then container data generated in this path is stored to the disk.

Data Volume:	• Add local storage						
	Storage type	Mount source	Container Path				
	• Add cloud storage						
	Storage type	Mount source	Container Path				
	Disk	pvc-disk	▼ /tmp	•			

f) Optional: Set Log Service. You can set collection parameters and customize tags.



Note:

Make sure that you have deployed a Kubernetes cluster and installed the log plugin on the cluster.

Set the following log collection parameters:

- Log Store: Set a Logstore. After you specify the Logstore name, the Logstore is generated in Log Service to store collected logs.
- Log path in the container: Set this parameter to stdout or set a log path.
 - stdout: If you set the log path parameter to stdout, you can collect the standard output logs of the container.
 - text log: If you specify a container log path, you can collect the text logs of the path. Wildcards can be used in setting the log file name for a log path. In this example, text logs in the path of /var/log/nginx are collected.

You can also customize log tags. The customized log tags can be collected together with container output logs and can benefit log analysis actions such as collecting log statistics and filtering specific logs.

	Log Service:	Note: please ensure that	at cluster has deployed log plug-ins.	
		😌 Configuration		
		Log Store	Log path in the container (can be set to stdout)	
uo		catalina	stdout	•
mfigurati		access	/var/log/nginx	•
Log Co		Custom Tag		
		Name Of Tag	Value Of Tag	
		арр	nginx	•

5. Click Next.

- 6. Configure advanced settings.
 - a) Set Access Control.

You can set the methods to expose the application pod and then click Create. In this example, a cluster IP service and an Ingress are set to create an Nginx application that is accessible for the Internet.



You can set access methods according to the communication requirements of your application.

• Internal applicatio n : an application that works only inside the cluster. You can create a cluster IP service or a node port service as needed for communication within the cluster.

External applicatio n : an application that needs to be exposed to the Internet. You can set how the application is accessed by using either of the following two methods:

- Create a Server Load Balancer service. This method uses Alibaba Cloud Server Load Balancer (SLB) to provide Internet accessibility for the application.
- Create a cluster IP service or a node port service, and create an Ingress.
 This method provides Internet accessibility through the Ingress. For more information, see *Ingress*.

	Basic Inform	ation	\geq	Container	Advanced	Done
Access Control	Service(Service) Ingress(Ingress)	Create				
Scale	HPA	Enable				
	Node Affinity	Add				
Scheduling	Pod Affinity	Add				
	Pod Anti Affinity	Add				
						Prev Create

A. Click Create on the right of Service. Configure a service in the displayed dialog box, and then click Create.
Create Service		\times
Name:	nginx-svc	
Type:	Server Load Balancer 🔻 public 🔻	
Port Mapping:	OAdd Name service port Container Port Protocol nginx-svc 80 80 TCP<▼	
Annotation:	Add Annotations for load balancer	
Tag:	O Add	
	Create Car	ncel

- Name: Enter the service name. The default is applicatio nname svc
 .
- Type: Select one service type.
 - Cluster IP: Exposes the service by using the internal IP address of your cluster. If you select this service type, the service is accessible only within the cluster.
 - Node port: Exposes the service by using the IP address and the static port (NodePort) of each node. A node port service routes to a cluster IP service that is automatically created. You can access the node port service from outside the cluster by requesting < NodeIP >:< NodePort >.
 - Server Load Balancer: Alibaba Cloud Server Load Balancer service. With this type of service, you can set an Internet or intranet access method

for your application. SLB can route to a node port service and a cluster IP service.

- Port Mapping: Add a service port and a container port, and select the TCP or UDP protocol. If you select the node port Type, you must add a node port to avoid port conflict.
- annotation: Add an annotation to the service. You can set SLB parameters. For more information, see *Access services by using Server Load Balancer*.
- Tag: Add a tag to the service to identify the service.
- B. Click Create on the right of Ingress. In the displayed dialog box, configure an Ingress rule for the application pod, and then click Create. For more information, see *Ingress configurations*.

Note:

When you create an application by using an image, you can create an Ingress rule for only one service. In this example, a virtual host name is used as the test domain name. You need to add a record to the host. You must use a filing domain name when you create your application.

101 .	37 . 224 .	146	fc	o . bar . com	# This	is	the
IP	address	of	the	Ingress .			

Create		×
Name:	nginx-ingress	
Rule:	• Add	
	Domain foo.bar.com Select * path e.g./ Services • Add Name port nginx-svc 80	
	EnableTLS	
Service weight:	Enable	
Grayscale release:	Add After the gray rule is set, the request meeting the rule will be routed to the new service. If you set a weight other than 100, the request to satisfy the gamma rule will continue to be routed to the new and old version services according to the weights.	
Annotation:	• Add rewrite annotation	
Tag:	O Add	
	Create	Cancel

C. In the access control area, the created service and Ingress are displayed. You can perform further configurations by clicking Update or Delete.

Create Application						
	Basic	Information	Container	Advanced		Done
	Services(Service	Update Delete				
		service port	Container P	ort	Protocol	
itrol		80	80		ТСР	
ccess Cor	Ingresses(Ingres	ss Update Delete				
4		Domain	path	Name	service port	
		foo.bar.com		nginx-svc	80	
Scale	HPA	Enable				
ling	Update Method	 Enable RollingUpdate OnDelete Max Unavailable 25 % ▼ Max Surge 25 % ▼ 				
Schedu	Node Affinity	Add				
	Pod Affinity	Add				
	Pod Anti Affinity	Add				

b) Optional: Set Horizontal Pod Autoscaling (HPA).

You enable HPA by selecting the Enable check box. Alibaba Cloud Container Service for Kubernetes provides pod auto scaling to deal with different application workloads. That is, you can change the number of pods according to the container CPU and memory usage.

	НРА	Enable
		Metric: CPU Usage
Scale		Condition: Usage 70 %
		Maximum Replicas: 10 Range : 2-100
		Minimum Replicas: 1 Range : 1-100
	Note:	

To use this function, you must set required resources for the pod. Otherwise, pod auto scaling cannot take effect. For more information, see general container settings.

- Metric: resource type. CPU or memory is available. This parameter must be specified with a resource type that is the same as the required resource type.
- Condition: the percentage value of resource usage. The number of containers increases when the resource usage exceeds this value.
- Maximum Replicas: the maximum number of the containers that the deployment can include.
- Minimum Replicas: the minimum number of the containers that the deployment can include.
- c) Optional: Set Scheduling.

You can set an update method, node affinity, pod affinity, and pod anti affinity. For more information, see *Affinity and anti-affinity*.

Note:

Affinity scheduling depends on node tags and pod tags. You can use built-in or customized tags to schedule nodes or pods.

A. Set Update Method.

You can select the RollingUpd ate or Recreate (OnDelete) method to replace old pods with new ones. For more information, see *Deployments*.

B. Set Node Affinity by using node tags.

Create		\times
Required:	• Add Rule	
	Selector 🛇 Add	•
	Tag Name Operator Tag Value	
	kubernetes.io/hostname 🔻 In 🔻 a	
	kubernetes.io/hostname 🔻 In 🔻 c	
Preferred:	Add Pula	
	Weight	
	100	
	Selector 📀 Add	
	Tag Name Operator Tag Value	
	kubernetes.io/hostname 🔻 NotIn 🔻 c	
	ОК	Cancel

Required rules and preferred rules are supported, and available operators include In , NotIn , Exists , DoesNotExi st , Gt , and Lt .

• Required rules must be satisfied and correspond to requiredDu ringSchedu lingIgnore dDuringExe cution . The required rules

have the same effect as NodeSelect or . In this example, the pod can be scheduled to only a node with the specified tags.

You can add multiple required rules, but only one required rule needs to be satisfied for pod scheduling.

 Preferred rules can be unnecessarily satisfied and correspond to preferredD uringSched ulingIgnor edDuringEx ecution. With the scheduling setting in this example, the system tries not to schedule the pod to the nodes with the specified tag.

You can also set Weight for each preferred rule. If multiple nodes satisfies the preferred rules, the system schedules the pod to a node with the highest weight.

You can add multiple preferred rules, and all the rules must be satisfied for pod scheduling.

C. Set Pod Affinity to deploy the application pod in a topology domain together with other pods. For example, to reduce network latency between the services

that communicate with each other, you can deploy their pods to a topology domain (for example, a host).

Create	\times
Required: 📀 Add Rule	
1 Namespace default	 ⊗
Topology Key kubernetes.io/hostname	
2 Selector • Add View Application List	
Tag Name Operator Tag Value	
3 app In r nginx G	
Preferred: • Add Rule	
Weight	8
100	
Namespace	_
default	v
Topology Key	_
kubernetes.io/hostname	
Selector • Add View Application List	
Tag Name Operator Tag Value	
app NotIn 🔻 wordpress	
ОК	Cancel

You can schedule pods according to tags of pods running on nodes. Required rules and preferred rules are supported, and available operators include In

, NotIn , Exists , DoesNotExi st .

- Required rules must be satisfied and correspond to requiredDu ringSchedu lingIgnore dDuringExe cution. All specified conditions of required rules must be met for pod affinity scheduling.
 - Namespace: Set a namespace. This parameter is required because the scheduling policy is based on pod tags.

- Topology Key: Set a topology domain to which pods are scheduled. This parameter takes effect through node tags. For example, if you set kubernetes . io / hostname as the topology key, a node is used to identify a topology. If you set beta . kubernetes . io / os as the topology key, a node operating system is used to identify a topology.
- Selector: Click this button to add a required rule.
- View Application List: Click View Application List, a dialog box is displayed. In the dialog box, you can view applications in each namespace and export application tags to the dialog box in which you set pod affinity.
- Required rule tag: Set a tag name, its operator, and the tag value for existing applications. This example schedules the application to be created to a host on which applications tagged with app : nginx run.
- Preferred rules can be unnecessarily satisfied and correspond to preferredD uringSched ulingIgnor edDuringEx ecution.
 Specified conditions of required rules will be met as many as possible for pod affinity scheduling.

You can set Weight for each preferred rule. The weight value range is 1 to 100. If multiple nodes satisfies the preferred rules, the system schedules the pod to a node with the highest weight. Other parameters are the same with the required rule setting.

- D. Set Pod Anti Affinity to deploy the application pods in a topology domain that excludes other pods. Scenarios that use pod anti affinity scheduling include:
 - Distribute the pods of a service to different topology domains (for example , different hosts) to improve the service stability.
 - Grant a pod the exclusive access to a node so as to guarantee that no other pods use the resources of the node.
 - · Distribute pods of the services that may affect each other to different hosts.

Note:

You can set pod anti affinity scheduling by using the same method as setting pod affinity scheduling. But the same scheduling rules have different meanings for these two types of scheduling. You need to select appropriate scheduling rules as needed.

- 7. Click Create.
- 8. After you create the application, anew page is displayed by default to prompt that you have created the application and lists objects included in the application.You can click View detail to view the deployment details.

Crea	ete Success	
Create deployment	nginx	Succeeded
Create service	nginx-svc	Succeeded
Create ingress	nginx-ingress	Succeeded
View deta	il Create again	

The nginx-deployment page is displayed by default.

Deployment nginx 主 Back	Deployment nginx 🔹 Back to List						
Overview							
Name:	nginx						
Namespace:	default						
Time Created:	10/11/2018,10:26:08						
Label:	app:nglnx						
annotation:	deployment.kubernetes.io/revision:1						
Selector:	app:nginx						
Strategy:	RollingUpdate						
Status:	Updated: , Unavailable:0 , Replica:2						
Trigger 1. You can only have	one of each trigger type.		Create Trigger 🔺				
No trigger is available at the mome	nt. Click "Create Trigger" in the upper-right corner.						
Pods Access Events	Horizontal Pod Autoscaler						
Name		Status	Image				
nginx-884c7fc54-rh26t		Running	nginx:latest				
nginx-884c7fc54-tc9rc		Running	nginx:latest				

9. Choose Discovery and Load Balancing > Ingress to verify that a rule is displayed in the Ingress list.

	Container Service - Kubernetes 🔻	Ingress					Ref	resh	Create
	Overview		sis and monitoring 🔞	Blue-green release					
ı	Clusters	Clusters k8s-test	▼ Nam	espaces default 🔻			Search By Name		Q
)	Applications	Name	Endpoint	Rule	Time Created				Action
•	 Service Mesh 	nginx-ingress		foo.bar.com/ -> nginx	-svc 10/10/2018,22:12:43	Details	Update Vie	ew YAML	Delete
	Virtual Services								
•	Discovery and Load B								
	Services								8
	Ingresses								

10.Access the test domain name in your browser to verify that you can visit the Nginx

welcome page.

foo.bar.com/?spm=5176.2020520152.0.0.704061b1K4UgO					
	Welcome to nginx!				
	If you see this page, the nginx web server is successfully installed and working. Further configuration is required.				
	For online documentation and support please refer to <u>nginx.org</u> . Commercial support is available at <u>nginx.com</u> .				
	Thank you for using nginx.				

1.8.2 Create a StatefulSet application by using an image

Kubernetes clusters of Alibaba Cloud Container Service allows you to quickly create applications of the StatefultSet type through the web interface. In this example, create a StatefultSet Nginx application and show features of a StatefultSet application.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have successfully created a cloud disk storage volume claim. For more information, see *Create a persistent volume claim*.
- You have successfully connected to the master node of the Kubernetes cluster. For more information, see *Connect to a Kubernetes cluster by using kubectl*.

Context

StatefulSet features are as follows:

Scenarios	Description
Pod consistency	Contains order (such as startup and stop order) and network consistency. This consistency is related to pods and has nothing to do with the node to which the pods are to be scheduled.
Stable persistent storage	Create a PV for each pod through VolumeClaimTemplate. Deleting or reducing replicas does not delete relevant volumes.
Stable network marker	The hostname mode for a pod is: (statefulse t name)-(sequence number).
Stable order	For StatefulSet of N replicas, each pod is assigned a unique order number within the range of 0 to N.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment in the left-side navigation pane, and then click Create by image in the upper-right corner.
- 3. Configure the basic parameters and then click Next.
 - Name: Enter the application name.
 - Cluster: Select a cluster to which the application is deployed.
 - Namespace: Select a namespace in which the application deployment is located. By default, the default namespace is used.
 - Replicas: Set the number of pods included in the application.
 - Type: Deployment type and StatefulSet type are available.



In this example, select the StatefulSet type.

Basic	Information	Container	Advanced	>	Done
Name:	nginx				
	The name should be 1-64 charact	ers long, and can contain numbers, lowe	r case English letters and hyphens, but canno	ot start with a hyphen.	
Cluster:	test-mia	v			
Namespace :	default	Ŧ			
Replicas:	2				
Туре	StatefulSet	•			
					Back Next

4. Configure containers.



You can configure multiple containers for the pod of the application.

- a) Configure the general settings for the application.
 - Image Name: Click Select image to select the image in the displayed dialog box and then click OK. In this example, select the nginx image.

You can also enter the private registry in the format of domainname / namespace / imagename : tag to specify an image.

- Image Version: Click Select image version to select a version. If the image version is not specified, the system uses the latest version by default.
- Always pull image: Container Service caches the image to improve deployment efficiency. During deployment, if the image tag is found consistent with that on the local cache, the image on the local cache is reused and is not pulled again. Therefore, if you do not modify the image tag when changing your codes and image for convenience of upper-layer business, the early image on the local cache is used in the application deployment. With this check box selected, Container Service ignores the cached image and repulls the image from the repository when deploying the application to make sure the latest image and codes are always used.
- Resource Limit: Specify the upper limit for the resources (CPU and memory) that can be used by this application to avoid occupying excessive resources.

CPU is measured in millicores, that is, one thousandth of one core. Memory is measured in bytes, which can be Gi, Mi, or Ki.

- Resource Request: Specify how many resources (CPU and memory) are reserved for the application, that is, these resources are exclusive to the container. Other services or processes will compete for resources when the resources are insufficient. By specifying the Resource Request, the application will not become unavailable because of insufficient resources.
- Init Container: Selecting this check box creates an Init Container which contains useful tools. For more information, see https://kubernetes.io/docs/concepts/workloads/pods/init-containers/.

	Basic Informa	ation Container Adva	nced
0	Container1 O Add Cont	tainer	
	Image Name:	nginx Select image	
	Image Version:	latest Select image version	
_		Always pull image Image pull secret	
Genera	Resource Limit:	CPU eg : 500m Core Memory eg : 128Mi MiB	
	Resource Request:	CPU eg : 500m Core Memory eg : 128Mi MiB	
	Init Container		

b) Optional: Configure Environment.

You can configure environment variables for the pod by using key-value pairs. Environment variables are used to add environment labels or pass configurations for the pod. For more information, see *Pod variable*.

c) Optional: Configure Health Check.

The health check function includes liveness probes and readiness probes. Liveness probes are used to detect when to restart the container. Readiness probes determine if the container is ready for receiving traffic. For more information about health check, see https://kubernetes.io/docs/tasks/configure-pod-

container/configure-liveness-readiness-probes.

Liveness	Enable				
	нттр		ТСР	Command	~
	Protocol	НТТР	•		
	path				
	Port				
	Http Header	name			
		value			
	Initial Delay	3			
	Period	10			
	Timeout	1			
	Success Threshold	1			
	Failure Threshold	3			
r Check					
Readiness	Enable				
	НТТР		ТСР	Command	~
	Protocol	НТТР	v		
	path				
	Port				
	Http Header	name			
		value			
	Initial Delay	3			
	Period	10			
	Timeout	1			

Request method	Description
HTTP request	An HTTP GET request is sent to the container. The following are supported parameters:
	 supported parameters: Protocol: HTTP/HTTPS Path: Path to access the HTTP server Port: Number or name of the port exposed by the container. The port number must be in the range of 1 to 65535. HTTP Header: Custom headers in the HTTP request. HTTP allows repeated headers. Supports the correct configuration of key values. Initial Delay (in seconds): Namely, the initialDelaySeconds. Seconds for the first probe has to wait after the container is started. The default is 3. Period (in seconds): Namely, the periodseconds. Intervals at which the probe is performed. The default value is 10. The minimum value is 1. Timeout (in seconds): Namely, the timeoutSeconds. The time of probe timeout. The default value is 1 and the minimum value is 1.
	probes that are considered as successful after a failed probe. The default is 1 and the minimum is 1. It must be 1 for a liveness probe.
	 Failure Threshold: The minimum number of consecutive failed probes that are considered as failed after a successful probe. The default value is 3. The minimum value is 1.

Request method	Description
TCP connection	 A TCP socket is send to the container. The kubelet attempts to open a socket to your container on the specified port. If a connection can be established, the container is considered healthy. If not, it is considered as a failure. The following are supported parameters: Port: Number or name of the port exposed by the container. The port number must be in the range of 1 to 65535. Initial Delay (in seconds): Namely, the initialDelaySeconds. Seconds for the first liveness or readiness probe has to wait after the container is started. The default is 15. Period (in seconds): Namely, the periodseconds. Intervals at which the probe is performed. The default value is 10. The minimum value is 1. Timeout (in seconds): Namely, the timeoutSeconds. The time of probe timeout. The default value is 1 and the minimum value is 1. Success Threshold: The minimum number of consecutive successful probes that are considered as successful after a failed probe. The default is 1 and the minimum number of consecutive failed probes that are considered as failed after a successful probe. The default value is 3. The minimum number of consecutive failed probes that are considered as failed after a successful probe. The default value is 3. The minimum number of consecutive failed probes that are considered as failed after a successful probe. The default value is 3. The minimum number of consecutive failed probes that are considered as failed after a successful probe. The default value is 3. The minimum value is 1.

Request method	Description
Command line	Detect the health of the container by executing probe detection commands in the container. The following are supported parameters:
	 supported parameters: Command: A probe command used to detect the health of the container Initial Delay (in seconds): Namely, the initialDelaySeconds. Seconds for the first liveness or readiness probe has to wait after the container is started. The default is Period (in seconds): Namely, the periodseconds. Intervals at which the probe is performed. The default value is 10. The minimum value 1. Timeout (in seconds): Namely, the timeoutSeconds. The time of probe timeout. The default value is 1 and the minimum value is 1. Success Threshold: The minimum number of consecutive successful probes that are considered as successful after a failed probe. The default is 1 and the minimum is 1. It must be 1 for a liveness probe. Failure Threshold: The minimum number of consecutive failed probes that are considered as
	failed after a successful probe. The default value is 3. The minimum value is 1.

d) Optional: Configure the lifecycle rule.

You can configure the following parameters for the container lifecycle: container config start, post start, and pre-stop. For more information, see https:// kubernetes.io/docs/tasks/configure-pod-container/attach-handler-lifecycle-event/.

• Container Config: Select the stdin check box to enable standard input for the container. Select the tty check box to assign an virtual terminal to for the

container to send signals to the container. These two options are usually used together, which indicates to bind the terminal (tty) to the container standard

input (stdin). For example, an interactive program obtains standard input from you and then displays the obtained standard input in the terminal.

- Start: Configure a pre-start command and parameter for the container.
- Post Start: Configure a post-start command for the container.
- Pre Stop: Configure a pre-end command for the container.

	Start:	Command ["/bin/sh","-c,"echo Hello > /user/share/message"]
ycle		Parameter
Life c	Post Start:	Command
	Pre Stop:	Command ["/user/sbin/nginx","-s","quit"]

e) Configure data volumes.

Local storage and cloud storage can be configured.

- Local storage: Supports hostPath, configmap, secret, and temporary directory. The local data volumes mount the corresponding mount source to the container path. For more information, see *Volumes*.
- Cloud storage: Supports three types of cloud storage: cloud disks, Network Attached Storage (NAS), and Object Storage Service (OSS).

In this example, configure a data volume claim named disk-ssd of cloud disk type and mount it to the / data path.

Data volume:	Add local storage			
	Storage type	Mount source	Container Path	
	Add cloud storage	e		
	Storage type	Mount source	Container Path	
	Disk	 disk-ssd 	▼ /data	

f) Optional: Configure Log Service. You can configure collection methods and customize tags for this service.



Make sure that a Kubernetes cluster is deployed and that the log plug-in is installed on the cluster.

Configure log collection methods as follows:

- Log Store: Configure a Logstore generated in Log Service which is used to store collected logs.
- Log path in the container: Supports stdout and text logs.
 - stdout: Collects standard output logs of containers.
 - text log: Collects logs in the specified path in the container. In this example, collect text logs in the path of /var/log/nginx. Wildcards are also supported.

You can also set custom tags. The customized tags are collected to the container output logs. A custom tag can help you tag container logs, providing convenienc e to log analysis such as log statistics and filter.

	Log Service:	Note: please ensure that	at cluster has deployed log plug-ins.	
		📀 Configuration		
		Log Store	Log path in the container (can be set to stdout)	
u		catalina	stdout	•
nfigurati		access	/var/log/nginx	•
Log Col				
		😌 Custom Tag		
		Name Of Tag	Value Of Tag	
		арр	nginx	•

- 5. Click Next after completing the configurations.
- 6. Configure advanced settings. In this example, configure only access settings.
 - a) Set Access Control.

You can set the methods to expose the application pod and then click Create. In this example, a cluster IP service and an Ingress are set to create an Nginx application that is accessible for the Internet.



You can set access methods according to the communication requirements of your application.

• Internal applicatio n : an application that works only inside the cluster. You can create a cluster IP service or a node port service as needed for communication within the cluster.

External applicatio n : an application that needs to be exposed to the Internet. You can set how the application is accessed by using either of the following two methods:

- Create a Server Load Balancer service. This method uses Alibaba Cloud Server Load Balancer (SLB) to provide Internet accessibility for the application.
- Create a cluster IP service or a node port service, and create an Ingress.
 This method provides Internet accessibility through the Ingress. For more information, see *Ingress*.

	Basic Inform	ation	\geq	Container	Advanced	Done	
Access Control	Service(Service) Ingress(Ingress)	Create					
Scale	HPA	Enable					
	Node Affinity	Add					
Scheduling	Pod Affinity	Add					
	Pod Anti Affinity	Add					
						Prev Create	

A. Click Create on the right of Service. Configure a service in the displayed dialog box, and then click Create.

Create Service		\times
Name:	nginx-svc	
Type:	Server Load Balancer 🔻 public 🔻	
Port Mapping:	Order Service port Container Protocol Port nginx-svc 80 80 TCP ●	
Annotation:	Add Annotations for load balancer	
Tag:	O Add	
	Create Car	ncel

- Name: Enter the service name. The default is applicatio nname svc
 .
- Type: Select one service type.
 - Cluster IP: Exposes the service by using the internal IP address of your cluster. If you select this service type, the service is accessible only within the cluster.
 - Node port: Exposes the service by using the IP address and the static port (NodePort) of each node. A node port service routes to a cluster IP service that is automatically created. You can access the node port service from outside the cluster by requesting < NodeIP >:< NodePort >.
 - Server Load Balancer: Alibaba Cloud Server Load Balancer service. With this type of service, you can set an Internet or intranet access method

for your application. SLB can route to a node port service and a cluster IP service.

- Port Mapping: Add a service port and a container port, and select the TCP or UDP protocol. If you select the node port Type, you must add a node port to avoid port conflict.
- annotation: Add an annotation to the service. You can set SLB parameters. For more information, see *Access services by using Server Load Balancer*.
- Tag: Add a tag to the service to identify the service.
- B. Click Create on the right of Ingress. In the displayed dialog box, configure an Ingress rule for the application pod, and then click Create. For more information, see *Ingress configurations*.

Note:

When you create an application by using an image, you can create an Ingress rule for only one service. In this example, a virtual host name is used as the test domain name. You need to add a record to the host. You must use a filing domain name when you create your application.

101 .	37 . 224 .	146	fc	o . bar . com	# This	is	the
IP	address	of	the	Ingress .			

Create		\times
Name:	nginx-ingress	
Rule:	• Add	
	Domain foo.bar.com Select * path e.g./ Services • Add Name port nginx-svc	
	EnableTLS	
Service weight:	Enable	
Grayscale release:	Add After the gray rule is set, the request meeting the rule will be routed to the new service. If you set a weight other than 100, the request to satisfy the gamma rule will continue to be routed to the new and old version services according to the weights.	
Annotation:	Add rewrite annotation	
Tag:	Add	
	Create	Cancel

C. In the access control area, the created service and Ingress are displayed. You can perform further configurations by clicking Update or Delete.

Crea	Create Application							
	Basic	Information >	Container	Advanced		Done		
	Services(Service)	Update Delete						
		service port	Container Por	t	Protocol			
trol		80	80		TCP			
ccess Con	Ingresses(Ingres	is Update Delete						
~		Domain	path	Name	service port			
		foo.bar.com		nginx-svc	80			
Scale	HPA	Enable						
lding	Update Method	Enable RollingUpdate OnDelete Max Unavailable 25 96 Max Surge 25 96						
Sched	Node Affinity	Add						
	Pod Affinity	Add						
	Pod Anti Affinity	Add						

b) Optional: Set Horizontal Pod Autoscaling (HPA).

You enable HPA by selecting the Enable check box. Alibaba Cloud Container Service for Kubernetes provides pod auto scaling to deal with different application workloads. That is, you can change the number of pods according to the container CPU and memory usage.

	HPA	C Enable
		Metric: CPU Usage
Scale		Condition: Usage 70 %
		Maximum Replicas: 10 Range : 2-100
		Minimum Replicas: 1 Range : 1-100
	Note:	

To use this function, you must set required resources for the pod. Otherwise, pod auto scaling cannot take effect. For more information, see general container settings.

- Metric: resource type. CPU or memory is available. This parameter must be specified with a resource type that is the same as the required resource type.
- Condition: the percentage value of resource usage. The number of containers increases when the resource usage exceeds this value.
- Maximum Replicas: the maximum number of the containers that the deployment can include.
- Minimum Replicas: the minimum number of the containers that the deployment can include.
- c) Optional: Set Scheduling.

You can set an update method, node affinity, pod affinity, and pod anti affinity. For more information, see *Affinity and anti-affinity*.

Note:

Affinity scheduling depends on node tags and pod tags. You can use built-in or customized tags to schedule nodes or pods.

A. Set Update Method.

You can select the RollingUpd ate or Recreate (OnDelete) method to replace old pods with new ones. For more information, see *Deployments*.

B. Set Node Affinity by using node tags.

Create		\times
Required:	• Add Rule	
	Selector 🖸 Add	
	Tag Name Operator Tag Value	
	kubernetes.io/hostname 🔻 In 🔻 a	
	kubernetes.io/hostname • In • c	
Preferred:	• Add Rule	
	Weight 8	
	Selector 📀 Add	
	Tag Name Operator Tag Value	
	kubernetes.io/hostname 🔻 NotIn 🔻 c	
	ОК	Cancel

Required rules and preferred rules are supported, and available operators include In , NotIn , Exists , DoesNotExi st , Gt , and Lt .

• Required rules must be satisfied and correspond to requiredDu ringSchedu lingIgnore dDuringExe cution . The required rules

have the same effect as NodeSelect or . In this example, the pod can be scheduled to only a node with the specified tags.

You can add multiple required rules, but only one required rule needs to be satisfied for pod scheduling.

 Preferred rules can be unnecessarily satisfied and correspond to preferredD uringSched ulingIgnor edDuringEx ecution. With the scheduling setting in this example, the system tries not to schedule the pod to the nodes with the specified tag.

You can also set Weight for each preferred rule. If multiple nodes satisfies the preferred rules, the system schedules the pod to a node with the highest weight.

You can add multiple preferred rules, and all the rules must be satisfied for pod scheduling.

C. Set Pod Affinity to deploy the application pod in a topology domain together with other pods. For example, to reduce network latency between the services

that communicate with each other, you can deploy their pods to a topology domain (for example, a host).

Create	\times
Required: 📀 Add Rule	
Namespace	8
Topology Key kubernetes.io/hostname	
2 Selector • Add View Application List	
Tag Name Operator Tag Value	
3 app In r nginx O	
Preferred: O Add Rule	
Weight	8
100	
Namespace	
default 🔻	
Topology Key	
kubernetes.io/hostname	
Selector • Add View Application List	
Tag Name Operator Tag Value	
app NotIn 🔻 wordpress 🗢	
ОК	Cancel

You can schedule pods according to tags of pods running on nodes. Required rules and preferred rules are supported, and available operators include In

, NotIn , Exists , DoesNotExi st .

- Required rules must be satisfied and correspond to requiredDu ringSchedu lingIgnore dDuringExe cution. All specified conditions of required rules must be met for pod affinity scheduling.
 - Namespace: Set a namespace. This parameter is required because the scheduling policy is based on pod tags.

- Topology Key: Set a topology domain to which pods are scheduled. This parameter takes effect through node tags. For example, if you set kubernetes . io / hostname as the topology key, a node is used to identify a topology. If you set beta . kubernetes . io / os as the topology key, a node operating system is used to identify a topology.
- Selector: Click this button to add a required rule.
- View Application List: Click View Application List, a dialog box is displayed. In the dialog box, you can view applications in each namespace and export application tags to the dialog box in which you set pod affinity.
- Required rule tag: Set a tag name, its operator, and the tag value for existing applications. This example schedules the application to be created to a host on which applications tagged with app : nginx run.
- Preferred rules can be unnecessarily satisfied and correspond to preferredD uringSched ulingIgnor edDuringEx ecution.
 Specified conditions of required rules will be met as many as possible for pod affinity scheduling.

You can set Weight for each preferred rule. The weight value range is 1 to 100. If multiple nodes satisfies the preferred rules, the system schedules the pod to a node with the highest weight. Other parameters are the same with the required rule setting.

- D. Set Pod Anti Affinity to deploy the application pods in a topology domain that excludes other pods. Scenarios that use pod anti affinity scheduling include:
 - Distribute the pods of a service to different topology domains (for example , different hosts) to improve the service stability.
 - Grant a pod the exclusive access to a node so as to guarantee that no other pods use the resources of the node.
 - Distribute pods of the services that may affect each other to different hosts.

Note:

You can set pod anti affinity scheduling by using the same method as setting pod affinity scheduling. But the same scheduling rules have different meanings for these two types of scheduling. You need to select appropriate scheduling rules as needed.

- 7. Click Create.
- 8. After you create the application, the create success page is displayed by default and objects contained in the application are listed. You can click View detail to view the deployment details.

	Create	Success	
Create d	eployment	nginx	Succeeded
Create	service	nginx-svc	Succeeded
Create	ingress	nginx-ingress	Succeeded
	View detail	Create again	

The StatefulSet page is displayed by default.

StatefulSet nginx	+ Back	to List					Refresh
Overview							
Name:	ng	ginx					
Namespace:	de	efault					
Time Created:	10	0/11/2018,15:08:16					
Label:	ap	p:nginx					
annotation:							
Selector:	ap	op:nginx					
Strategy:	Ro	ollingUpdate					
Status:	U	pdated:0 , Unavailable:0 , Re	plica:0				
Trigger 1. You No trigger is availa	can only ha	ve one of each trigger type. noment. Click "Create Trigger"	in the upper-right corner.				Create Trigger 🔨
Pods Access	Events	Horizontal Pod Autoscaler					
CPU				Memory			
0				0.00			
15:10:00				0			15:10:00
		 CPU usa; 	(Cores)			Memory usage(GI)	
Name			Status		Image		
nginx-0			Running		nginx:latest		
nginx-1			Running		nginx:latest		

9. Then click Back to list in the upper-left corner to view the created StatefulSet application in the StatefulSet list page.

Statefu	ulSet				Refresh	Create by Image	Create by Template
Clusters	test-mia	▼ Namespace default	•				
Name	Tag	PodsQuantity	Image	Time Created			Action
nginx		2/2	nginx	10/11/2018,15:55:57		Details E	dit Scale More▼

- 10.Optional: To verify service scalability, click Scale at the right of a target nginx application.
 - a) In the displayed dialog box, set the number of pod to 3. You can see that when you expand pods, the pods are in the increment order; when you contract pods, the pods are in the descending order. This shows the order stability of pods in StatefulSet.

Name	Status	Image
nginx-0	Running	nginx:latest
nginx-1	Running	nginx:latest
nginx-2	Running	nginx:latest

b) Click Application > Volumes Claim in the left-side navigation pane, you can see that as the application expands, new cloud disk volumes are created with pods; if the application contracts, created PV/PVC will not be deleted.

Volumes Claims						Refresh	Create
Clusters test-mia	• Name	espace default	v				
Name	Capacity	Access Mode	Status	Storage Class Name	Relate Volume	Time Created	Action
disk-ssd-nginx-0	20Gi	ReadWriteOnce	Bound	alicloud-disk-ssd	d-bp1cy8o56jfgodpst28f	10/11/2018,15:55:57	Delete
disk-ssd-nginx-1	20Gi	ReadWriteOnce	Bound	alicloud-disk-ssd	d-bp1gdkenf5ki7pt5pct9	10/11/2018,15:56:09	Delete
disk-ssd-nginx-2	20Gi	ReadWriteOnce	Bound	alicloud-disk-ssd	d-bp1f2xopk3sz02ug12ls	10/11/2018,15:57:02	Delete

What's next

Connect to the master node and run following commands to verify the persistent storage feature.

Create a temporary file on a cloud disk:

```
kubectl
            exec
                  nginx - 1
                              ls
                                / tmp
                                                   # list
                                                           files
#
under this
              directory
 lost + found
 kubectl
            exec
                  nginx - 1
                             touch / tmp / statefulse
                                                      t
 # add
      а
            temporty file
                             named
                                    statefulse
                                               t
```

#

kubectl exec nginx - 1 ls / tmp lost + found statefulse t

Remove the pod to verify the data persistence:

```
# kubectl delete pod nginx - 1
pod " nginx - 1 " deleted
# kubectl exec nginx - 1 ls / tmp
data persistenc e storage
lost + found
statefulse t
```

In addition, you can also find that after you delete a pod, the pod automatically restarts after a period of time, which indicates the high availability of the StatefulSet application.

1.8.3 Create a Job application by using an image

By running a Kubernetes cluster with Alibaba Cloud Container Service, you can create a Job application through the Web interface. This example creates a Job application named busybox to describe features of the Job application features.

Prerequisites

You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Context

A Job processes short-lived one-off tasks in batches to guarantee that one or multiple pods in the batch tasks successfully terminate.

Kubernetes supports the following types of Jobs:

- Non-parallel Job: A Job of this type creates only one pod. The Job is completed when the pod terminates successfully.
- Job with a fixed completion count: A Job of this type has . spec . completion s
 set to create multiple pods. The Job is completed when the number of these pods
 reaches the . spec . completion s value.
- Parallel Job with a work queue: A Job of this type has . spec . Parallelis m set but has . spec . completion s not set. The Job is completed when at least one pod has terminated with success, and all pods are terminated.

Parallel Job with a fixed completion count: A Job of this type has both . spec .
 completion s and . spec . Parallelis m set. Multiple pods of the Job process the work queue at the same time.

According to the . spec . completion s and . spec . Parallelis m settings, Jobs can be classified into the following patterns.



The Job created in this example is a parallel Job with a fixed completion count.

Job pattern	Usage example	Action	Completion	Parallelism
One-off Job	Database migration	A Job creates a pod and the Job is completed when the pod terminates successfully.	1	1
Job with a fixed completion count	Pod that processes the work queue	A Job creates pods one by one. When the pods terminate successful ly and the number of the terminated pods reaches the completion s value, the Job is completed.	2+	1
Parallel Job with a fixed completion count	Multiple pods process work queues at the same time	A Job creates pods one by one. When the number of pods reaches the completion s value, the Job is completed.	2+	2+
Job pattern	Usage example	Action	Completion	Parallelism
--------------	---	--	------------	-------------
Parallel Job	Multiple pods process work queues at the same time	A Job creates one or multiple pods. When at least one pod terminates successful ly, the Job is completed.	1	2+

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Application > Job, and then click Create by Image in the upper-right corner.
- 3. Set the basic parameters and then click Next.
 - Name: Enter a name for the application.
 - · Cluster: Select a cluster to which the application is deployed.
 - Namespace: Select a namespace in which the application deployment is located. You can also choose to use the default namespace.
 - Type: Select the Job type.

Note:

In this example, select the Job type.

Basic I	Information	Container	\rightarrow	Advanced	\rightarrow	Done
Name:	busybox					
	The name should be 1-64 characters	long, and can contain numbers, l	ower case English le	tters and hyphens, but cannot	start with a hyphen.	
Cluster:	k8s-test	•				
Namespace :	default	¥				

4. Configure containers.



You can configure multiple containers for the pods of the application.

a) Set the container parameters.

• Image Name: Click Select image to select an image in the displayed dialog box and then click OK. In this example, select the busybox image.

You can also enter a private registry in the format of domainname / namespace / imagename : tag to specify an image.

- Image Version: Click Select image version to select a version. If you do not specify any image version, the system uses the latest version by default.
- Always pull image: Container Service caches the image to improve deployment efficiency. During deployment, if the tag of the newly specified image is the same as that of the cached image, Container Service reuses the cached image rather than pulls the same image again. Therefore, if you do not modify the image tag during scenarios where you are changing your code and image, the earlier image in the local cache is used in the application deployment. If you select this check box, Container Service ignores the cached image and re-pulls the image when deploying the application to make sure the latest image and code are always used.
- Image pull secret: If you use a private image, we recommend that you use a secret to guarantee the security of your image. For more information, see *Use an image Secret*.
- Resource Limit: Specify the upper limit for the resources (CPU and memory) that can be used by this application to avoid occupying excessive resources.
 CPU is measured in millicores, that is, one thousandth of one core. Memory is measured in bytes, which can be Gi, Mi, or Ki.
- Resource Request: Specify how many resources (CPU and memory) are reserved for the application (that is, these resources become exclusive to the container). If you do not set this parameter, other services or processes will

compete for resources, which means the application may become unavailable due to resource shortage.

 Init Container: Select this check box to create an Init Container that contains useful tools. For more information, see https://kubernetes.io/docs/concepts/

workloads/pod	ls/init-co	ntainers/.

Con	tainer1 O Add Conta	ainer
	Image Name:	busybox Select image
	Image Version:	latest Select image version
		Always pull image Image pull secret
General	Resource Limit:	CPU eg : 500m Core Memory eg : 128Mi MiB
	Resource Request:	CPU eg : 500m Core Memory eg : 128Mi MiB
	Init Container	

b) Optional: Set Environment .

You can use key-value pairs to set environment variables for the pods. Environment variables are used to add environment labels or pass configurations for the pods. For more information, see *Pod variable*.

c) Optional: Set Health Check.

You can set liveness probes and readiness probes. Liveness probes are used to detect when to restart the container. Readiness probes determine if the container is ready to receive traffic. For more information about health check,

see https://kubernetes.io/docs/tasks/configure-pod-container/configure-liveness-readiness-

probes.

	Liveness	Enable				
		нттр		ТСР	Command	~
		Protocol	НТТР	•		
		path				
		Port				
		Http Header	name			
			value			
		Initial Delay	3			
		Period	10			
		Timeout	1			
		Success Threshold	1			
			T			
ĸ		Failure Threshold	3			
h Che						
Healt	Readiness	Enable				
		нттр		ТСР	Command	~
		Protocol				
			HTTP	•		
		path				
		Port				
		Http Header	name			
			value			
			2			
		Initial Delay	3			
		Initial Delay Period	10			

Request method	Description
HTTP request	With this health check method, you can send an HTTP GET request to the container. The following parameters are supported:
	• Protocol: HTTP/HTTPS.
	 Path: path to access the HTTP server. Port: number or name of the access port exposed by the container. The port number must be in the range of 1 to 65535. HTTP Header: custom headers in
	the HTTP request. HTTP allows repeated headers. You can use a key -value pair to set an HTTP Header.
	 Initial Delay (in seconds): namely, the initialDelaySeconds, indicating the number of seconds for which the first probe must wait after the container is started. The default yalue is 3
	 Period (in seconds): namely, the periodseconds, indicating the interval at which probes are performed. The default value is 10. The minimum value is 1.
	• Timeout (in seconds): namely, the timeoutSeconds, indicating the number of time that the probe has timed out. The default value is 1 and the minimum value is 1.
	 Success Threshold: The minimum number of consecutive successful probes needed for determining a probe success after a failed probe The default value is 1 and the
	minimum value is 1. It must be set
	 Failure Threshold: The minimum number of consecutive failed probes needed for determining a probe failure after a successful
	probe. The default value is 3. The
525	minimum value is 1.

Request method	Description
TCP connection	If you use this health check method, a TCP socket is sent to the container. The kubelet then attempts to open the socket of the container on a specified port. If a connection can be established, the container is considered healthy. If not, it is considered unhealthy. The following parameters are supported:
	 Port: number or name of the access port exposed by the container. The port number must be in the range of 1 to 65535. Initial Delay (in seconds): namely, the initialDelaySeconds, indicating the seconds for the first liveness or readiness probe must wait for after the container is started. The default is is 15. Period (in seconds): namely, the periodseconds, indicating the interval at which probes are performed. The default value is 10. The minimum value is 1. Timeout (in seconds): namely, the timeoutSeconds, indicating the number of time that the probe has timed out. The default value is 1 and the minimum value is 1. Success Threshold: The minimum number of consecutive successful probes needed for determining a probe success after a failed probe. The default value is 1 and the minimum value is 1. It must be set to 1 for a liveness probe. Failure Threshold: The minimum number of consecutive failed probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe failure after a successful probes needed for determining a probe fail
	minimum value is 1.

Request method	Description
Command line	With this heath check method, you can detect the container health by executing a probe detection command in the container. The following parameters are supported:
	 Command: a probe command used to detect the health of the container Initial Delay (in seconds): namely, the initialDelaySeconds, indicating the number of seconds for which the first liveness or readiness probe must wait after the container is started. The default value is 5. Period (in seconds): namely, the periodseconds, indicating the interval at which probes are performed. The default value is 10. The minimum value 1. Timeout (in seconds): namely, the timeoutSeconds, indicating the number of time that the probe has timed out. The default value is 1 and the minimum value is 1.
	 probes needed for determining a probe success after a failed probe The default value is 1 and the minimum value is 1. It must be set to 1 for a liveness probe. Failure Threshold: The minimum number of consecutive failed probes needed for determining a probe failure after a successful probe. The default value is 3. The minimum value is 1.

d) Optional: Set the life cycle.

You can set the following parameters for the container life cycle: container config, start, post start, and pre-stop. For more information, see https://kubernetes.io/docs/tasks/configure-pod-container/attach-handler-lifecycle-event/.

- Container Config: You can select the stdin check box to enable standard input for the container, or select the tty check box to assign a virtual terminal to the container to send signals to the container. You can also select the two options at the same time. That is, you can bind the terminal (tty) to the container standard input (stdin). For example, an interactive program can obtain standard input from you and then display the obtained standard input in the terminal.
- Start: Set a pre-start command and parameter for the container.
- Post Start: Set a post-start command for the container.
- Pre Stop: Set a pre-stop command for the container.

	Container Config:	stdin 🔲 tty
	Start:	Command echo hello world
cycle		Parameter
Life	Post Start:	Command
	Pre Stop:	Command

e) Optional: Set data volumes.

You can configure local storage and cloud storage.

- Local storage: Supported storage types include HostPath, ConfigMap, Secret, and EmptyDir. By setting a type of local storage, you can mount its mount source to the container path. For more information, see *Volumes*.
- Cloud storage: Supported types of cloud storage include cloud disks, Network Attached Storage (NAS), and Object Storage Service (OSS).
- f) Optional: Set Log Service. You can set collection parameters and customize tags.



Make sure that you have deployed a Kubernetes cluster and installed the log plugin on the cluster.

Set the following log collection parameters:

- Log Store: Set a Logstore. After you specify the Logstore name, the Logstore is generated in Log Service to store collected logs.
- Log path in the container: You can set this parameter to stdout or set a log path.
 - stdout: If you set a log path to stdout, you can collect the standard output logs of the container.
 - text log: If you set a container log path, you can collect the text logs of the path. Wildcards can be used in setting the log file name for a log path.

You can also set custom tags. The custom tags are collected to the container output logs. A custom tag can help you tag container logs, making it easy to collect log statistics, filter logs, and analyze logs by using other methods.

- 5. After you complete the container configuration, click Next.
- 6. Configure advanced settings.

You can configure Job Settings.

Parameter	Description
Completions	Number of pods that must be run successfully by the configured Job. The default value is 1.
Parallelism	Number of pods that must be run in parallel by the configured Job at any time. The default value is 1.
ActiveDeadlineSeconds	Operating time limit of the configured Job. If the Job is not completed within the time limit, the system tries to terminate the Job.

Parameter	Description
BackoffLimit	Number of retries performed by the configured Job to create pods after a failure. The default is 6. Each time the Job fails, the failed pods associated with the Job are recreated with time delay . The time delay grows exponentially each time. The upper limit of the time delay is six minutes.
Restart	Only Never and OnFailure restart policies are supported.

	Basic Informatio	on		Container	Advanced	Done	
	Completions	6					
	Parallelism	2					
Settings	ActiveDeadlineSeconds	600					
dot	BackoffLimit	6					
	Restart	never	•				🚺 Contact L
						Prev Crea	te

7. Click Create.

8. After you create the Job application, a new page is displayed by default to prompt that you have created the application with the objects included.



You can click View Details to view the Job details.

During the creation process, you can view the creation status of the pods in the Status column. In this example, two pods are created in parallel according to the Job definition.

Job busybox 🛛 🕆 Back to L	ist		Refresh
Overview			
Name:	busybox		
Namespace:	default		
Time Created:	11/21/2018,15:56:54		
Label:	job-name:busybox controller-uid:0340c7b1-ed63-11e8-89e5-00163e15	887a	
annotation:			
Status:	Active2 , Succeeded3, Failed0		
Pods Events			
Name	Status	Image	Action
busybox-c2k6s	Succeeded	busybox:latest	Details More -
busybox-f5pfs	Pending	busybox:latest	Details More -
busybox-gnjkg	Pending	busybox:latest	Details More -
busybox-pgq7f	Succeeded	busybox:latest	Details More 🚽
busybox-phnvc	Succeeded	busybox:latest	Details More -

Wait until all pods are created.

Job busybox 🗧 🛧 Back to List]		Refresh
Overview			
Name:	busybox		
Namespace:	default		
Time Created:	11/21/2018,15:56:54		
Label:	job-name:busybox controller-uid:0340c7b1-ed63-11e8-8	9e5-00163e15887a	
annotation:			
Status:	Active0, Succeeded6, Failed0		
Pods Events			
Name	Status	Image	Action
busybox-c2k6s	Succeeded	busybox:latest	Details More 🗸
busybox-f5pfs	Succeeded	busybox:latest	Details More 🗸
busybox-gnjkg	Succeeded	busybox:latest	Details More 🗸
busybox-pgq7f	Succeeded	busybox:latest	Details More 🗸
busybox-phnvc	Succeeded	busybox:latest	Details More 🗸
busybox-wdrlj	Succeeded	busybox:latest	Details More 🗸

9. In the upper-left corner, click Back to List. On the Jog page, the Job completion time is displayed.



If the Job has not created all the pods, the page does not display the Job completion time.

Job					Refresh	Create by Image	Create by Template
Help: 🔗 H Container m	How to use private images of Create applications of Schedule onitoring of Blue-green release	a pod to the s	pecified node	🔗 Create a Layer-4	Ingress 🔗 Create a Laye	er-7 Ingress 🔗 Configur	e pod auto scaling 🛛 🔗
Clusters	▼ Namespace default ▼						
Name	Tag	Status	Pod Status	Image	Time Created	Completion Time	Action
busybox	job-name:busybox controller-uid:0340c7b1-ed63-11e8-89e5-00163e15887a	Succeeded	Active0 Succeeded6 Failed0	busybox:latest	11/21/2018,15:56:54	11/21/2018,15:57:15	Details More 🗸

1.8.4 Create an application in Kubernetes dashboard

You can create an application in the Kubernetes dashboard.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.

- 3. Click Dashboard at the right of the cluster to enter the Kubernetes dashboard.
- 4. In the Kubernetes dashboard, click CREATE in the upper-right corner to create an application.
- 5. The Resource creation page appears. Configure the application information.

Create an application in any of the following three ways:

- CREATE FROM TEXT INPUT: Directly enter the orchestration codes in the YAML or JSON format to create an application. You must know the corresponding orchestration format.
- CREATE AN APP: Complete the following configurations to create an application.
 - App name: Enter the name of the application you are about to create. In this example, enter nginx test .
 - Container image: Enter the URL of the image to be used. In this example, use Docker *Nginx*.
 - Number of pods: Configure the number of pods for this application.
 - Service: Select External or Internal. External indicates to create a service that can be accessed from outside the cluster. Internal indicates to create a service that can be accessed from within the cluster.
 - Advanced options: To configure the information such as labels and environment variables, click SHOW ADVANCED OPTIONS. This configuration distributes the traffic load evenly to three pods.
- CREATE FROM FILE: Upload an existing YAML or JSON configuration file to create an application.
- 6. Click UPLOAD or DEPLOY to deploy the containers and services.

You can also click SHOW ADVANCED OPTIONS to configure more parameters.

What's next

After clicking UPLOAD or DEPLOY, you can view the services and containers of the application.

Click Pods in the left-side navigation pane. You can check the status of each Kubernetes object according to the icon on the left. indicates the object is still being deployed. indicates the object has completed the deployment.

1.8.5 Create a Linux application by using an orchestration template

In a Container Service Kubernetes orchestration template, you must define resource objects required for running an application, and combine the resource objects into a complete application by using label selector.

Prerequisites

Create a Kubernetes cluster. For more information, see Create a Kubernetes cluster.

Context

Create an Nginx application in this example. Firstly, create a backend pod resource object by creating the deployment. Then, deploy the service to bind it to the backend pod, forming a complete Nginx application.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment in the left-side navigation pane.
- 3. Click Create by Template in the upper-right corner.

Container Service - Kubernetes +	Deployment					Refresh Create by image Create by template
Overview	Clusters k8s-test v Nan	nespace default	Ŧ			2
 Clusters 	Name	Tag	PodsQuantity	Image	Time Created	Action
Clusters	nginx-deployment-basic	app:nginx	2/2	nginx:1.7.9	09/19/2018,11:59:31	Details Edit Scale Monitor More -
Nodes						
Volumes						
Namespace						
Authorization						001
Application Deployment						

- 4. Configure the template and then click DEPLOY.
 - Clusters: Select the cluster in which in which the resource object is to be deployed.
 - Namespace: Select a namespace to which resource object belongs. The default namespace is default. Except for the underlying computing resources such as

nodes and persistent storage volumes, most of the resource objects must act on a namespace.

- Resource Type: Alibaba Cloud Container Service provides Kubernetes YAML sample templates of many resource types for you to deploy resource objects quickly. You can write your own template based on the format requirements of Kubernetes YAML orchestration to describe the resource type you want to define.
- · Add Deployment: You can quickly define a YAML template with this feature.
- Deploy with exist template: You can import an existing template into the template configuration page.



The following is a sample orchestration for an Nginx application. The orchestration is based on an orchestration template built in Container Service. By using this orchestration template, you can create a deployment that belongs to an Nginx application quickly.



Container Service supports Kubernetes YAML orchestration in which you can

use the --- symbol to separate resource objects so as to create multiple resource

objects through a single template.

```
apiVersion : apps / v1beta2 # for
                                      versions
                                                 before
                                                         1.8.
0 use apps / v1beta1
kind : Deployment
metadata :
    name : nginx - deployment
    labels :
      app : nginx
spec :
    replicas : 2
    selector :
      matchLabel s :
       app : nginx
    template :
      metadata :
        labels :
          app : nginx
      spec :
       containers :
       – name : nginx
          image : nginx : 1 . 7 . 9 # replace
                                                 it
                                                      with
                                                             your
  exactly < image_name : tags >
          ports :
         - containerP ort : 80
___
                     # for
                                         before
apiVersion : v1
                              versions
                                                 1.8.0
                                                             use
apps / v1beta1
kind : Service
metadata :
  name : my - service1
rvice name
                               # TODO : to
                                               specify
                                                        your
service
   labels :
     app : nginx
spec :
   selector :
                              # TODO : change
                                                label
     app : nginx
                                                        selector
  to match your backend
                               pod
   ports :
     protocol : TCP
     name : http
     port : 30080
                              # TODO : choose
                                                 an
                                                     unique
      on each
                                      port conflict
port
                  node
                         to
                              avoid
     targetPort : 80
```

```
type : LoadBalanc er  ## In this example , change
the type from NodePort to LoadBalanc er .
```

5. After you click DEPLOY, a message indicating the deployment status is displayed. After the deployment succeeds, click Kubernetes Dashboard in the message to go to the dashboard and check the deployment progress.

<								+ CREATE
Cluster Namespaces	CPU usage			Memory usage ()			
Nodes Persistent Volumes	0.001 0.001			25.7 Mi 22.9 Mi 25.1 T.2 Mi				_
Roles Storage Classes	0.0005 0.0003 14:02 14:03 14:06	14:10	14:13 14:16	5.72 Mi 0 14:02 14:03	14:06	14:10	14:13	14:16
Namespace		Time				Time		
default 👻	Deployments							-
Overview	Name 🌩	Labels	Pods	Age ≑	Imag	jes		
Workloads	oginx-deployment	app: nginx	0 / 2	10/10/	2018, 14:13:15 ngir	1.7.9		:
Cron Jobs	oginx-deployment-basic	app: nginx	2 / 2	09/19/	2018, 11:59:31 ngir	nx:1.7.9		:
Daemon Sets								
Deployments								

6. In the Kubernetes dashboard, you can see that the service named my-service1 is successfully deployed and its external endpoint is exposed. Click the access address under External endpoints.

Ser	vices						Ŧ
	Name 🌲	Labels	Cluster IP	Internal endpoints	External endpoints	Age 🌲	
0	my-service1	app: nginx	172.19.0.114	my-service1:30080 TCP my-service1:30327 TCP	:30080	10/10/2018, 14:13:15	:

7. You can access the Nginx service welcome page in the browser.



What's next

You can also go back to the home page of Container Services and then click Discovery and Load Balancing > Services in the left-side navigation pane to view the Nginx service.

1.8.6 Create a Windows application by using an orchestration template

This topic describes how to create a Windows application by using an orchestration template. Such a template is used to customize the resources required by a Windows application to operate.

Prerequisites

A Kubernetes cluster that supports Windows is created. For more information, see *Create a Windows application by using an orchestration template.*

Context

In this topic, an application named aspnet is created by using an orchestration template. This application contains a deployment and a service. On the backend, the deployment creates pods according to settings. Then, the service is associated with the pods.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Application > Deployment.
- 3. In the upper-right corner, click Create by Template.



- 4. To set the orchestration template, set the following:
 - Clusters: Select the target cluster. The resources required by the application are deployed in a cluster.
 - Namespace: Select the target namespace. The default namespace is preset.
 Except for nodes, persistent volumes, and other underlying resource types, most resources required by the application are deployed in a namespace.
 - Sample Template: Select the target sample template. Alibaba Cloud Container Service for Kubernetes provides many built-in YAML orchestration templates for

different types of resources. You can customize an orchestration template to set a type of resource according to the YAML orchestration requirements.

- · Add Deployment: Edit a YAML template quickly by using this function.
- Using Existing Template: Import an existing template to the template setting area.

Then, click DEPLOY.



The following is an orchestration template for the Windows application named aspnet . With such an orchestration template, you can quickly create a deployment and a service for the application.

Note:

If you want to create multiple resources in a template, you can use --- to separate different resources.

```
apiVersion : v1
kind : Service
metadata :
name : aspnet - svc
spec :
ports :
- port : 80
```

```
protocol : TCP
      targetPort : 80
  selector :
   app : aspnet
  type : LoadBalanc er
apiVersion : apps / v1beta2
kind : Deployment
metadata :
  name : aspnet
spec :
  selector :
   matchLabel s :
     app : aspnet
  template :
    metadata :
     labels :
        app : aspnet
    spec :
      containers :
          image : ' microsoft / dotnet - samples : aspnetapp '
          name : aspnet
```

5. A message is displayed on the bottom of the Template area to show the result. If a success message is displayed (shown in the preceding figure), click Kubernetes Dashboard at the end of the message to view the progress.

<		Deployments				+ CREATE
Cluster	Deployments					Ŧ
Namespaces						
Nodes	Name 🗘	Labels	Pods	Age ≑	Images	
Persistent Volumes	aspnet	-	1 / 1	04/22/2019, 16:05:06	microsoft/dotnet-samples:aspnetapp	:
Roles						
Storage Classes						
Namespace						
default 🔻						
Overview						
Workloads						
Cron Jobs						Contact
Daemon Sets						S
Deployments						
Jobs						
Pods						

6. In the left-side navigation pane, choose Discovery and Load Balancing > Service. Then, in the External endpoints column, click the IP address of the created service named aspnet - svc to visit the home page of the aspnet application in your browser.



In the Kubernetes dashboard, you can view that a service named aspnet - svc and its external endpoint are created.

Namespace	■ Discovery and load	balancing Services					+ CREATE
default 👻							
Overview	Services						Ŧ
Workloads	Name 🜲	Labels	Cluster IP	Internal endpoints	External endpoints	Age ≑	
Cron Jobs	aspnet-svc	-	-	100000000	80 🖾	04/22/2019, 16:05:06	:
Daemon Sets	kubernetes	component: apiserver provider: kubernetes	10.000	10000		04/20/2019, 15:46:52	:
lobs							
Pods							
Replica Sets							
Replication Controllers							
Stateful Sets							
Discovery and Load Balancing							
Ingresses							
Services							

The following figure shows the home page of the aspnet application.

Home Page - aspnetapp X	+				10 100		i er i	~
\leftrightarrow \rightarrow C (0)						Q 🕁	0) :
Use th	is space to summarize your privacy a	nd cookie use policy.			Learn More Accept			
	ASP.NET (Core Wind	dows Linux that can run anywhere. Leam M 00	OSX ^{xe}	>			
Applica • Sample Core M • Themin	ation uses Ho pages using ASP.NET VC g using Bootstrap	Add a Controller and View Manage User Secrets using Secret Manager. Use logging to log a message. Add packages using NuGet. Target development, staging or production environment.	Overview • Conceptual overview of what is ASPNET Core • Fundamentals of ASPNET Core • such as Startup and middleware. • Working with Data • Security • Client side development • Develop on different platforms • Read more on the documentation site	Run & E - Run your: - Run tools - Run too	Deploy app such as EF migrations Microsoft Azure Web			

What's next

You can also return to the home page of Container Service-Kubernetes, and then choose Discovery and Load Balancing > Service in the left-side navigation pane to view the service of the aspnet application.

1.8.7 Manage applications by using commands

You can create applications or view containers in applications by using commands.

Prerequisites

Before using commands to manage applications, *#unique_57*.

Create an application by using commands

Run the following statements to run a simple container (a Nginx Web server in this example).

```
root @ master # kubectl run - it nginx -- image = registry .
aliyuncs . com / spacexnice / netdia : latest
```

This command creates a service portal for this container. Specify -- type =

LoadBalancer and an Alibaba Cloud Server Load Balancer route will be created

to the Nginx container.

```
root @ master # kubectl expose deployment nginx -- port = 80
    -- target - port = 80 -- type = LoadBalanc er
```

View containers by using commands

Run the following command to list all the running containers in the default namespaces.

root @ master # kubectl get pods NAME READY STATUS RESTARTS AGE nginx - 2721357637 - dvwq3 1 / 1 Running 1 9h

1.8.8 Simplify Kubernetes application deployment by using Helm

In Kubernetes, app management is the most challenging and in demand field. The Helm project provides a uniform software packaging method which supports version control and greatly simplifies Kubernetes app distribution and deployment complexity.

Alibaba Cloud Container Service integrates the app catalog management function with the Helm tool, extends the functions, and supports official repository, allowing you to deploy the application quickly. You can deploy the application in the Container Service console or by using command lines. This document introduces the basic concepts and usage of Helm and demonstrat es how to use Helm to deploy the sample applications WordPress and Spark on an Alibaba Cloud Kubernetes cluster.

Basic concepts of Helm

Helm is an open-source tool initiated by Deis and helps to simplify the deployment and management of Kubernetes applications.

You can understand Helm as a Kubernetes package management tool that facilitate s discovery, sharing and use of apps built for Kubernetes. It involves several basic concepts.

- Chart: A Helm package containing the images, dependencies, and resource definitions required for running an application. It may also contain service definitions in a Kubernetes cluster, similar to the formula of Homebrew, the dpkg of APT, or the rpm file of Yum.
- Release: A chart running on a Kubernetes cluster. A chart can be installed multiple times on the same cluster. A new release will be created every time a chart is installed. For example, to run two databases on the server, you can install the MySQL chart twice. Each installation will generate its own release with its own release name.
- Repository: The repository for publishing and storing charts.

Helm components

Helm adopts a client/server architecture composed of the following components:

- Helm CLI is the Helm client and can be run locally or on the master nodes of the Kubernetes cluster.
- Tiller is the server component and runs on the Kubernetes cluster. It manages the lifecycles of Kubernetes applications.
- Repository is the chart repository. The Helm client accesses the chart index files and packages in the repository by means of the HTTP protocol.

Use Helm to deploy applications

Prerequisites

• Before using Helm to deploy an application, create a Kubernetes cluster in Alibaba Cloud Container Service. For more information, see *Create a Kubernetes cluster*.

Tiller is automatically deployed to the cluster when the Kubernetes cluster is created. Helm CLI is automatically installed on all the master nodes and the configuration points to the Alibaba Cloud chart repository.

• Check the Kubernetes version of your cluster.

Only clusters whose Kubernetes version is 1.8.4 or later are supported. For clusters whose Kubernetes version is 1.8.1, upgrade the cluster on the Cluster List page.

Deploy applications in Container Service console

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > App Catalog in the left-side navigation pane.
- 3. On the App Catalog page, click a chart (WordPress in this example) to enter the chart details page.

Overview	0°	6 0	8 0	6 0
Clusters	¥0	¥o	¥o	¥o
Nodes	hdfs-namenode-k8s stable	jenkins 2.67 stable	kafka incubator	monocular v0.5.3 stable
Namespace	Ø ^O	Øo	Øo	Øo
Deployment Pods Service	mssql-linux 14.0.3023.8 stable	mysql stable	mysqlha 5.7.13 incubator	nodered Incubator
Ingress Volumes Claim	Øo	Øo	Øo	Øo
Release Config Maps	postgresql stable	redis 3.2.9 stable	redis-ha 4.0.8-r0 incubator	spark stable
Secret Store 1 Docker Images	Øo	Øo	Øo	Øo
App Catalog Service Catalog	spark-oss incubator	tiller-rest-proxy 0.7.3 stable	3 wordpress 4.8.2 stable	zookeeper incubator

- 4. Enter the basic information for the deployment on the right.
 - · Clusters: Select the cluster in which the application is to be deployed.
 - · Namespace: Select the namespace. default is selected by default.
 - Release Name: Enter the release name for the application. Enter test in this example.

Readme Values	Deploy
WordPress	Only Kybernatas yersians 1.8.4 and above are
WordPress is one of the most versatile open source content management systems on the market. A publishing platform for building blogs and websites.	supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list
TL;DR;	Clusters k8s-cluster
\$ helm install stable/wordpress	Namespace default •
Introduction	Release Name wordpress-default
This chart bootstraps a WordPress deployment on a Kubernetes cluster using the Helm package manager. It also packages the Bitnami MariaDB chart which is required for bootstrapping a MariaDB deployment for the database requirements of the WordPress application.	DEPLOY

5. Click the Values tab to modify the configurations.

In this example, bind dynamic data volumes of the cloud disk to a persistent storage volume claim (PVC). For more information, see .



You need to create a persistent storage volume (PV) of cloud disk in advance. The capacity of the PV cannot be less than the value defined by the PVC.

Readm	Values	Deploy
72 73 74 75 ∽ 76 77 78	<pre>## mariadbDatabase: bitnami_wordpress ## Create a database user ## crei https://github.com/bitnami/bitnami-docker-mariadb/blob/master/README.md#creating-a -database-user-on-first-run ## mariadbUser: bn_wordpress</pre>	Only Kubernetes versions 1.8.4 and above are supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list
79 80 - 81	<pre>## Password for mariadbUser ## ref: https://github.com/bitnami/bitnami-docker-mariadb/blob/master/README.md#creating-a -database-user-on-first-run ##</pre>	Clusters k8s-cluster
82 83 84 85 - 86 87	<pre># mariadbPassword: # mariadbPassword: ## Enable persistence using Persistent Volume Claims ## ref: http://kubernetes.io/docs/user-guide/persistent-volumes/ ##</pre>	Namespace default •
87 88 √ 89 90 91	<pre>## persistence: enabled: true ## mariadb data Persistent Volume Storage Class ## If defined, storageClassName: <storageclass> ## If defined, storageClassName: <storageclass></storageclass></storageclass></pre>	Release Name wordpress-default
92 93 94 95 96	<pre>## If set to -, storageclassmame: , which utsades dynamic provisioning ## If undefined (the default) or set to multi, no storageclassimame spec is ## set, choosing the default provisioner. (gp2 on AWS, standard on ## GKE, AWS & OpenStack) ## force_flow</pre>	DEPLOY
97 98 99 100 101 ~	## Kubernetes configuration	Version
102 103 104 105	## For minikube, set this to NodePort, elsewhere use LoadBalancer ## serviceType: LoadBalancer	0.6.13 Project Homepage

6. Click DEPLOY after completing the configurations. After the successful deployment, you are redirected to the release page of this application.

Container Service - Kubernetes 💌	R	elease List - wordpress-default				Refresh
Overview		Current Version				
 Clusters 						
Clusters		Release Name : wordpress-default	Namespace : defau	lt	Deployed at : 07/10/2018,17:11:24	
Nodes	C	Current Version: 1			Time Updated : 0	7/10/2018,17:11:24
				_		
Volumes		Resource			Values	
Namespace		Resource	Kind			
Application		wordpress-default-mariadb	Secret			View YAML
Deployment		wordpress-default-wordpress	Secret			View YAML
Pods		wordpress-default-mariadb	ConfigM	1ap		View YAML
Service		wordpress-default-mariadb	Persiste	entVolumeClaim		View YAML
		wordpress-default-wordpress	Persiste	entVolumeClaim		View YAML
Ingress		wordpress-default-mariadb	Service			View YAML
Volumes Claim		wordpress-default-wordpress	Service			View YAML
Release 2		wordpress-default-mariadb	Deploy	ment		View YAML
Config Maps		wordpress-default-wordpress	Deploy	ment		View YAML

7. Click Application > Service in the left-hand navigation pane. Select the target cluster and namespace and find the corresponding service. You can obtain the HTTP/HTTPS external endpoint address.

Container Service - Kubernetes ▼	Service List						Refresh Create
Overview	Clusters k8s-cluster •	Namespace	default 🔹	3			
Clusters	Name	Туре	Time Created	ClustersIP	internalendpoint	externalendpoint	Action
Clusters	kubernetes	ClusterIP	06/27/2018,17:53:49	-	kubernetes:443 TCP		Details Update View YAML Delete
Volumes	wordpress-default-mariadb	ClusterIP	07/10/2018,17:36:16	11140	wordpress-default-mariadb:3306 TCP	-	Details Update View YAML Delete
Namespace	wordpress-default- wordpress	LoadBalancer	07/10/2018,17:36:16		wordpress-default-wordpress:80 TCP wordpress-default-wordpress:32109 TCP wordpress-default-wordpress:443 TCP wordpress-default-wordpress:32094 TCP	:80 :443	Details Update View YAML Delete
Service 2							

8. Click the preceding access address to enter the WordPress blog publishing page.

Deploy applications by using command lines

You can use SSH to log on to the master node of the Kubernetes cluster when deploying applications by using command lines (Helm CLI is automatically installed and has configured the repository). For more information, see *Access Kubernetes clusters by using SSH*. You can also install and configure the kubectl and Helm CLI locally.

In this example, install and configure the kubectl and Helm CLI locally and deploy the applications WordPress and Spark.

Install and configure kubectl and Helm CLI

1. Install and configure kubectl on a local computer.

For more information, see Connect to a Kubernetes cluster by using kubectl.

To view information of the target Kubernetes cluster, enter the command kubectl

cluster - info .

2. Install Helm on a local computer.

For the installation method, see Install Helm.

3. Configure the Helm repository. Here the charts repository provided by Alibaba Cloud Container Service is used.

```
helm init -- client - only -- stable - repo - url https ://
aliacs - app - catalog . oss - cn - hangzhou . aliyuncs . com /
charts /
helm repo add incubator https :// aliacs - app - catalog .
oss - cn - hangzhou . aliyuncs . com / charts - incubator /
helm repo update
```

Basic operations of Helm

• To view the list of charts installed on the cluster, enter the following command:

helm list

Or you can use the abbreviated version:

helm ls

• To view the repository configurations, enter the following command:

helm repo list

• To view or search for the Helm charts in the repository, enter one of the following commands:

```
helm
       search
helm
       search
                repository
                              name # For
                                            example ,
                                                      stable
                                                                 or
incubator .
helm
       search
                chart
                             # For
                                       example ,
                        name
                                                  wordpress
                                                               or
spark .
```

 $\cdot\;$ To update the chart list to get the latest version, enter the following command:

helm repo update

For more information about how to use Helm, see Helm document.

Deploy WordPress by using Helm

Use Helm to deploy a WordPress blog website.

Enter the following command.

helm install -- name wordpress - test stable / wordpress



The Alibaba Cloud Kubernetes service provides the support for dynamic storage volumes of block storage (cloud disk). You need to create a storage volume of cloud disk in advance.

The result is as follows:

```
NAME : wordpress - test
LAST DEPLOYED : Mon Nov 20 19 : 01 : 55 2017
NAMESPACE : default
STATUS : DEPLOYED
...
```

Use the following command to view the release and service of WordPress.

helm list kubectl get svc

Use the following command to view the WordPress related pods and wait until the status changes to Running.

kubectl get pod

Use the following command to obtain the WordPress access address:

echo http://\$(kubectl get svc wordpress - test - wordpress - o jsonpath ='{. status . loadBalanc er . ingress [0]. ip }')

Access the preceding URL in the browser, and you can see the familiar WordPress website.

You can also follow the chart instructions and use the following command to obtain the administrator account and password of the WordPress website:

echo Username : user

```
echo Password : $( kubectl get secret -- namespace default
wordpress - test - wordpress - o jsonpath ="{. data . wordpress -
password }" | base64 -- decode )
```

To completely delete the WordPress application, enter the following command:

helm delete -- purge wordpress - test

Deploy Spark by using Helm

Use Helm to deploy Spark for processing big data.

Enter the following command:

helm install -- name myspark stable / spark

The result is as follows:

```
NAME : myspark
LAST DEPLOYED : Mon Nov 20 19 : 24 : 22 2017
NAMESPACE : default
STATUS : DEPLOYED
...
```

Use the following commands to view the release and service of Spark.

helm list kubectl get svc

Use the following command to view the Spark related pods and wait until the status changes to Running. Pulling images takes some time because the Spark related images are large.

kubectl get pod

Use the following command to obtain the Spark Web UI access address:

```
echo http ://$( kubectl get svc myspark - webui - o
jsonpath ='{. status . loadBalanc er . ingress [ 0 ]. ip }'): 8080
```

Access the preceding URL in the browser, and you can see the Spark Web UI, on which indicating currently three worker instances exist.

Then, use the following command to use Helm to upgrade the Spark application and change the number of worker instances from three to four. The parameter name is case sensitive.

helm upgrade myspark -- set "Worker . Replicas = 4 " stable / spark

The result is as follows:

```
Release "myspark "has been upgraded Happy Helming !
LAST DEPLOYED : Mon Nov 20 19 : 27 : 29 2017
NAMESPACE : default
STATUS : DEPLOYED
```

Use the following command to view the newly added pods of Spark and wait until the status changes to Running.

kubectl get pod

Refresh the Spark Web UI in the browser. The number of worker instances changes to four.

To completely delete the Spark application, enter the following command:

helm delete -- purge myspark

Use third-party chart repository

Besides the preset Alibaba Cloud chart repository, you can also use the third-party chart repository (make sure the network is accessible). Add the third-party chart repository in the following command format:

helm repo add repository name repository URL helm repo update

For more information about the Helm related commands, see Helm document.

References

Helm boosts the growth of communities. More and more software providers, such as Bitnami, have begun to provide high-quality charts. You can search for and discover existing charts at https://kubeapps.com/.

1.8.9 Use an application trigger

Alibaba Cloud Container Service Kubernetes supports the application trigger

function. You can use an application trigger in many ways.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created an application that is used to create an application trigger and test the trigger. In this example, create an nginx application.

Procedure

- 1. Log on to the Container Service console.
- 2. Click Application > Deployment and select a cluster and namespace. Click Details at the right of the target nginx application.

Container Service - Kubernetes 🔻	Deployment					Refresh Create by image Create by template
Overview	Clusters k8s-test v Name	space default	· 2			
 Clusters 	Name	Tag	PodsQuantity	Image	Time Created	Action
Clusters	nginx-deployment-basic	app:nginx	2/2	nginx:1.7.9	09/19/2018,11:59:31	Details Edit Scale Monitor More -
Nodes						3
Volumes						
Namespace						_
Authorization						
Application						a B
Deployment						

3. On the nginx application details page, click Create Trigger on the right side of the trigger bar.

Deployment ng	jinx-deployment-basic 🗨 Back to List	Refresh						
Overview								
Name:	nginx-deployment-basic							
Namespace:	default							
Time Created:	09/19/2018,11:59:31							
Label:	app:nginx							
annotation:	deployment.kubernetes.io/revision:1							
Selector:	app:nginx							
Strategy:	RollingUpdate							
Status:	Updated:2 , Unavailable:0 , Replica:2							
Trigger 1. Yo	Trigger 1. You can only have one of each trigger type.							
No trigger is availab	No trigger is available at the moment. Click "Create Trigger" in the upper-right corner.							

4. In the pop-up dialog box, click Redeploy and click Confirm.



Currently, only the redeploy action is supported.

Create Trigger				×
* Action :	Redeploy	v		
			Confirm	Cancel

After the trigger is created, a trigger link is displayed in the trigger bar on the nginx application detail page.

Trigger 1. You can only have one of each trigger type.	Cre	ate Trigger	^
Trigger Link (move mouse over to copy)	Туре	Action	
https://cs.console.aliyun.com/hook/trigger?token=eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCJ9.eyJjbHVzdGVySWQiOIJjNjJkN2NiZTYyODQ0NDMyMWFjYTNIZjkyYjQ3OGQx	Redeploy	Delete Trigge	er

5. Copy the trigger link and visit it in the browser. A message is returned on the web page, containing information such as the request ID.

🗋 https://	/cs.cor	nsole.aliyun.com/ ×	+	di contra e	A		C) all and a second sec	
$\leftrightarrow \rightarrow c$	С	https://cs.consol	e.aliyun.com/hook/tr	rigger	NUMBER OF STREET	and a support of		
"code":"200","message":"","requestId":"6e75bec1-69ce-4228-956b-56461da134db"}								

6. Back to the nginx application detail page, you can see that a new pod appears.

Pods	Access	Events	Horizontal Pod Autoscaler		
Name				Status	Image
nginx-de	ployment-bas	ic-6898cc69f	b-9726v	Running	nginx:1.7.9
nginx-deployment-basic-6898cc69fb-9nlns				e Running	nginx:1.7.9

After a period of time, the nginx application removes the old pod and keeps only the new pod.

What's next

You can call a trigger by using GET or POST in a third-party system. For example, you can run the curl command to call a trigger.

Call the redeploy trigger as follows:

```
curl https :// cs . console . aliyun . com / hook / trigger ? token =
xxxxxxxx
```

1.8.10 View pods

You can view the pods of a Kubernetes cluster in the Container Service console or in the Kubernetes dashboard.

View pods in Container Service console

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Pods in the left-side navigation pane to go to the Pods page.
- 3. Select the target cluster and namespace, the target pod, and click Details on the right.



You can update or delete a pod. For pods created by using deployments, we recommend that you manage these pods by using deployments.

4. View the pod details.

View pods in Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.
- 3. Click Dashboard at the right of the cluster to enter the Kubernetes dashboard.
- 4. In the Kubernetes dashboard, click Pods in the left-side navigation pane to view the pods in the cluster.

You can also click Services in the left-side navigation pane and then click the service name to view the pods in this service.

- 5. You can check the status of each Kubernetes object according to the icon on the left. indicates the object is still being deployed. indicates the object has completed the deployment.
- 6. Click the pod name to view the details, CPU usage, and memory usage of the pod.
- 7. Click LOGS in the upper-right corner to view the pod logs.
- 8. You can also click the icon at the right of the pod and then select Delete to delete the pod.

1.8.11 Create a service

This topic describes how to create a Kubernetes service in Alibaba Cloud Container Service for Kubernetes.

A Kubernetes service, known as a service in this and related topicsof Alibaba Cloud Container Service for Kubernetes, is an abstract object that defines a logical set of pods and a policy through which to access the pods. Usually, a label selector determines which set of pods are targeted by a service.

In a Kubernetes cluster, each pod has its own IP address, and the pods of a deployment can be removed at any time. However, this action changes the IP addresses of the pods. As a result, directly using IP addresses of pods is ineffective as the scenario does not provide high availability. By comparison, a Kubernetes service decouples the relationship between the frontend and the backend. Specifically, a Kubernetes service is a loose coupling service solution where the operations of the backend do not impact the frontend.

For more information, see *Kubernetes service*.

Prerequisite

A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

Step 1: Create a deployment

1. Log on to the Container Service console.

2. In the left-side navigation pane under Kubernetes, choose Applications > Deployments. Then click Create by Template in the upper-right corner.

Container Service - Kubernetes 🔻	Deployment Create by Image	Create by Template
Overview	How to use private images O Create applications O Schedule a pod to the specified node O Create a Layer-4 Ingress O Create a Layer-7 Ingress O Configure pod auto scaling	2
 Clusters 	Ø Container monitoring Ø Blue-green release	
Clusters	Clusters k8s-test Namespaces default Search By Name	Q
Nodes	Name Tag PodsQuantity Image Time Created	Action
Volumes	O Could not find any record that met the condition.	
Namespaces		_
Authorization		Con
Applications		act Us
Deployments		

3. Select the target cluster and namespace, and select a custom template or a sample template from the Resource Type drop-down list. Then, click DEPLOY.

Clusters	k8s-cluster 🔹	
Namespace	default	
Resource Type	Custom	
Template	<pre>1 apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1 2 kind: Deplayment 3 metadata: 4 name: nginx-deployment-basic 5 labels: 6 app: nginx 7 spec: 7 replicas: 2 9 selector: 10 metadata: 14 labels: 15 app: nginx 16 spec: 17 containers: 18 - name: nginx: 17 linage: nginx: 19 linage: nginx: 17 forts: 20 ports: 21 containerPort: 80</pre>	Add Deployment Deploy with exist template
	Deployed successfully. Go to Dashboard to see the deployment progress: Kubernetes Dashboard	
	Save Template DEPLOY	

In this example, the sample template specifies an Nginx deployment.

```
apiVersion : apps / v1beta2
                             #
                                for
                                      versions
                                                before
                                                         1.8.
   use apps / v1beta1
0
kind : Deployment
metadata :
  name : nginx - deployment - basic
   labels :
    app : nginx
spec :
   replicas :
              2
   selector :
```
```
matchLabel s :
     app : nginx
 template :
   metadata :
     labels :
       app : nginx
   spec :
     containers :
       name : nginx
                                                     replace
       image : nginx : 1 . 7 . 9
                                                  #
                                                                it
with
             exactly < image_name : tags >
       your
       ports :
         containerP ort :
                            80
                                            in
                                                     service .
     ## This
               port
                      must
                             be
                                  exposed
                                                 а
```

4. Click Kubernetes Dashboard to view the running status of this deployment.

<							+ CREATE
Cluster	CPU usage		Me	mory usage 🛈			
Nodes	0.001 0.001		(tes)	150 Mi 134 Mi			
Persistent Volumes Roles	0.0005 0.0003		Memory (b)	36.8 Mi			
Storage Classes	0 14:50 14:53	14:56 15:00 Time	15:04	0 14:50 14:53	14:56 Time	15:00	15:04
Namespace							
default 👻	Deployments						Ŧ
Overview	Name 🗢	Labels	Pods	Age 🜲	Images		
Workloads	nginx-deployment-basic	app: nginx	2 / 2	08/31/2018, 11:39:51	nginx:1.7.9		:
Cron Jobs	wordpress	app: wordpress	2/2	08/28/2018, 17:09:06	wordpress:4		:
Daemon Sets							
Deployments	1						

Step 2: Create a service

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Discovery and Load Balancing > Services.
- 3. Select the target cluster and namespace. Then, click Create in the upper-right corner.

Container Service - Kubernetes +		Service							Refresh
Overview	^	& Canary re	lease						3
Clusters	l	Clusters k8	is-test 🔻 M	lamespaces de	fault 🔹 🙎				Search By Name Q
 Applications 	L	Name	Label	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint	Action
 Service Mesh 	l	kubernetes	component:apiserver provider:kubernetes	ClusterIP	05/17/2019,18:12:33	-	kubernetes:443 TCP		Details Update View YAML Delete
Virtual Services	9	ngnix	run:nginx	LoadBalancer	05/23/2019,11:11:03		ngnix:80 TCP ngnix:31298 TCP		Details Update View YAML Delete
Services Ingresses	1								Contactus

4. In the displayed dialog box, set service parameters.

Create Service		\times
Name:	nginx-svc	
Type:	Server Load Balancer 🔻 public 🔻	
Related:	nginx-deployment-basic 🔻	
Port Mapping:	 Add service port Ro Container Port Protocol TCP ▼ 	
annotation:	Add Annotations for load balancer Name Value service.beta.kubernetes.ic 20	
Tag:	 Add Name Value app nginx 	
	Create Can	cel

- Name: Enter the service name. In this example, the service name is set to nginxsvc.
- Type: Select the service type, namely, the service access method.
 - Cluster IP: Exposes the service by using the internal IP address of your cluster. If you select this service type, the service is accessible only within the cluster. This is the default service type.

- Node port: Exposes the service by using the IP address and the static port (NodePort) of each node. A node port service routes to a cluster IP service that is automatically created. You can access the node port service from outside the cluster by requesting < NodeIP >:< NodePort >.
- Server Load Balancer: Alibaba Cloud Server Load Balancer service. With this type of service, you can set an Internet or intranet access method for your application. SLB can route to a node port service and a cluster IP service.
- Related: Select the backend object to associate with the service. In this example, the nginx-deployment-basic deployment created in the preceding step is associated with the service. If you do not associate the service with any objects, the system does not create any corresponding endpoint objects. In this case, you can manually associate the service with your own specific endpoints. For more information, see *Services without selectors*.
- Port Mapping: Add a service port number and a container port number. The container port number that you set must be the same as the port number of the container exposed by the pod.
- annotation: Add an annotation to the service. You can set SLB parameters.
 For example, to control the service traffic, you can set the peak bandwidth of the service to 20 Mbit/s by setting this parameter as service . beta .
 kubernetes . io / alicloud loadbalanc er bandwidth : 20 . For more information, see Access services by using Server Load Balancer.
- Tag: Add a tag to the service to identify the service.
- 5. Click Create. The nginx-svc service is then displayed in the service list.

Service						Refresh Create
Help: 🔗 Can	ary release					
Clusters k8	s-cluster 🔻	Namespace default	v			Search By Name Q
Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint	Action
kubernetes	ClusterIP	02/05/2019,15:58:07	10.00	kubernetes:443 TCP	-	Details Update View YAML Delete
nginx-svc	LoadBalancer	02/16/2019,15:17:13	$(0,1) \in [0,1]$	nginx-svc:80 TCP nginx-svc:31200 TCP		Details Update View YAML Delete

6. Enter the external endpoint of the nginx-svc service in your browser to access the service.

~		
G	() 1 m m m //spm=51/6.2020520152.0.0.49fd16ddLtg1Ua	
		Welcome to nginx!
		If you see this page, the nginx web server is successfully installed and working. Further configuration is required.
		For online documentation and support please refer to <u>nginx.org</u> . Commercial support is available at <u>nginx.com</u> .
		Thank you for using nginx.

1.8.12 Scale a service

This topic describes how to scale out or scale in an application service as needed after an application is created.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Clusters.
- 3. On the right of the target cluster, click Dashboard.

Container Service - Kubernetes 🔻	Cluster List		Yo	ou can create up to	100 clusters a	nd can add up	to 1000 nodes in each o	cluster. Refresh	Create Kubernetes Cluste	er
Overview	Help: & Create cluster & Create GP creation failures & Authorization man	U clusters 🛛 🔗 Sca agement	e cluster 🔗 Conne	ect to Kubernetes c	luster via kube	cti 🔗 Manaj	ge applications with com	mands 🔗 Cluste	er planning 🔗 Troubleshoot clu	ster
Clusters	Name 🔻									
Nodes	Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Number of Nodes	Time Created	Version	2 Act	ion
Volumes	k8s-cluster	Kubernetes	China East 1 (Hangzhou)	VPC vpc-	Running	5	02/05/2019,15:51:15	1.11.5	Manage View Logs Dashboar	d
Namespace				ppinej0q3u8					Cluster Expansion Mon	8 -

4. In the left-side navigation pane, click Deployments.





6. In the displayed dialog box, change the value of Desired number of pods to the number you require. Here, the example number of desired pods is 2. Then, click OK.

This action adds a new pod. The number of replicas becomes 2.



What's next

You can check the status of each Kubernetes object according to the icon on the left of the deployment list. indicates the object is being deployed.

object has been deployed.

Additionally, you can click a deployment name to view the details of the running Web service. Specifically, you can view the replica sets included in the deployment, and the CPU usage and memory usage of these replica sets.



If no resources are displayed, we recommend that you wait a few minutes and then refresh the page.

<	Workloads Deployments nginx-deployment-basic		+++ SCALE	🖍 EDIT	DELETE	+ CREATE
Cluster Namespaces	CPU usage	Memory usage (i)				
Nodes Persistent Volumes Roles Storage Classes	0.001 0.001 0.0006 0.0006	3.22 Mi 2.86 Mi (84 (6) (2.15 Mi) 732 Ki				/
Namespace default 👻	138 12:00 12:03 12:06 12:10 12:12 Time	11:58 12:00	12.03 Time	12:08	12:10	12:12
Overview Workloads	Details					
Cron Jobs	Name: nginx-deployment-basic					
Daemon Sets	Namespace: default Labels: app: ngi					
Deployments	Annotations: deployment.kubernetes.io/revision:					
Jobs	Creation Time: 2018-04-24T10:46 UTC					
Pods	Selector: app: ngi					
Replica Sets	Min ready seconds: 0					
Replication Controllers	Revision history limit: 10					
Stateful Sets	Rolling update strategy: Max surge: 25%, Max unavailable: 25% Status: 2 updated, 2 total, 2 available, 0 unavailable					

1.8.13 View a service

This topic describes how to view a Kubernetes service in Alibaba Cloud Container Service for Kubernetes.

Context

If you set an external service when you create an application, the Kubernetes dashboard creates the external service, in addition to running containers. The service is used to pre-set a Server Load Balancer to distribute traffic to the containers.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Discovery and Load Balancing > Service.

3. Select the target cluster and namespace, and then click Details on the right of the

Service **Container Service -**Kubernetes -Help: S Canary release Overview Clusters Namespace k8s-cluster default v Clusters Clusters Time Created Name Type kubernetes Nodes ClusterIP 02/05/2019,15:58: Volumes NodePort 02/17/2019,09:26: new-nginx -Namespace LoadBalancer 02/16/2019,16:41: nginx-svc Authorization Application) old-nginx 02/17/2019,09:25: NodePort Discovery and Load B .. Service Ingress

target service.

You can view the service name, service type, service creation time, cluster IP address, external endpoint, and other information. In this example, the external

endpoint (IP address) assigned to the service is displayed. To access the Nginx application, click this IP address.

nginx-svc 🗧 Back to List	
Quantieur	
Overview	
Name:	nginx-svc
Namespace:	default
Time Created:	02/16/2019,16:41:42
Label:	app:nginx-v2
annotation:	service.beta.kubernetes.io/alicloud-loadbalancer-bandwidth:20
Туре:	LoadBalancer
ClustersIP:	10.000
InternalEndpoint:	nginx-svc:80 TCP nginx-svc:32397 TCP
ExternalEndpoint:	Contract on the second

You can also open the Kubernetes dashboard of the target cluster and click Services in the left-side navigation pane to view all services of the cluster.

1.8.14 Update a service

This topic describes how to update a Kubernetes service in the Container Service console or the Kubernetes dashboard.

Update a service in the Container Service console

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Discovery and Load Balancing > Service.

3. Select the target cluster and namespace, and then click Update on the right of the target service. In this example, the target service is named nginx-svc.

Container Service - Kubernetes →		Service							Refresh	Cre	eate
Overview	*	Help: 🔗 Cana	ary release								
 Clusters 		Clusters k8s	-cluster 🔻 l	Namespace default	· 3			Search By N	lame		Q
Clusters		Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint				Action
Nodes		kubernetes	ClusterIP	02/05/2019,15:58:07	-	kubernetes:443 TCP		Details Update	View YAML	D	elete
Volumes	ţ	new-nginx	NodePort	02/17/2019,09:26:21		new-nginx:80 TCP new-nginx:32580 TCP		Details Undete	View YAML	D	elete
Authorization		nginx-svc	LoadBalancer	02/16/2019,16:41:42		nginx-svc:80 TCP nginx-svc:32397 TCP	100.00	Details Update	View YAML	D	elete 🖸
Application		old-nginx	NodePort	02/17/2019,09:25:20		old-nginx:80 TCP old-nginx:30675 TCP	-	Details Update	View YAML	D	elete Ç
Service 2 Ingress											

4. In the displayed dialog box, update the service parameters. Then, click Update. In this example, the service tag is updated from app:nginx-v1 to app:nginx-v2.

Update Service		\times
Name:	nginx-svc	
Type:	Server Load Balancer 🔻 public 🔻	
Port Mapping:	⊙Add	
	service port Container Port Protocol	
	80 80 TCP • •	
annotation:	• Add Annotations for load balancer	
	Name Value	
	service.beta.kubernetes.ic 20	
Tag:	• Add	
	Name Value	
	app nginx-v1 🗢	
	Update Car	ncel

5. In the service list, click Details on the right of the target service to view the service updates. In this example, the updated service tag app:nginx-v2 is displayed.

nginx-svc 🔁 Back to List	
Overview	
Name:	nginx-svc
Namespace:	default
Time Created:	02/16/2019,16:41:42
Label:	app:nginx-v2
annotation:	service.beta.kubernetes.io/alicloud-loadbalancer-bandwidth:20
Туре:	LoadBalancer
ClustersIP:	
InternalEndpoint:	nginx-svc:80 TCP nginx-svc:32397 TCP
ExternalEndpoint:	:80

Update a service in the Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. Click Dashboard on the right of the target cluster.



4. In the Kubernetes dashboard, select the target namespace and then click Services in the left-side navigation pane.

5. Click the icon on the right of the target service and then click View/edit YAML.

■ Discovery and load I	balancing Services					+ CREATE
Services						Ŧ
Name 🗢	Labels	Cluster IP	Internal endpoints	External endpoints	Age 🗢	
onew-nginx	-	11,000	new-nginx:80 TCP new-nginx:32580 TCP	-	02/17/2019, 09:26:21	:
old-nginx	-	10,000	old-nginx:80 TCP old-nginx:30675 TCP	-	02/17/2019, 09:25:20	:
orginx-svc	app: nginx-v2		nginx-svc:80 TCP nginx-svc:32397 TCP	0	02/16/2019, 16:41:42 D	elete
kubernetes	component: apiserver provider: kubernetes	1.01	kubernetes:443 TCP	-	02/05/2019, V 15:58:07	iew/edit YAML

6. In the displayed dialog box, modify the service settings. Then, click UPDATE. In this example, the nodePort is changed to 31000.



1.8.15 Delete a service

This topic describes how to delete a Kubernetes service in the Alibaba Cloud Container Service console.

Prerequisites

- A Kubernetes cluster is created. For more information, see *Create a Kubernetes cluster*.
- A service is created. For more information, see *Create a service*.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Discovery and Load Balancing > Service.
- 3. Select the target cluster and namespace, and then click Delete on the right of the target service. In this example, the target service is named nginx-svc.

Service						Refresh Create
Help: 🔗 Can	ary release					
Clusters k8	-cluster 🔻	Namespace default	v 3			Search By Name Q
Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint	Action
kubernetes	ClusterIP	02/05/2019,15:58:07	1.00	kubernetes:443 TCP	-	Details Update View YAML Delete
new-nginx	NodePort	02/17/2019,09:26:21	10.000	new-nginx:80 TCP new-nginx:32580 TCP	-	Details Update View YAML Delete
nginx-svc	LoadBalancer	02/16/2019,16:41:42	1.000	nginx-svc:80 TCP nginx-svc:32397 TCP	100.00	Details Update View YAML Delete
old-nginx	NodePort	02/17/2019,09:25:20	1.00	old-nginx:80 TCP old-nginx:30675 TCP	-	Details Update View YAML Delete
	Service Help: & Can Clusters kas Name kubernetes new-nginx nginx-svc old-nginx	Service Help: & Canary release Clusters k8s-cluster Name Type kubernetes Cluster3P new-nginx NodePort nginx-svc LoadBalancer old-nginx NodePort	Service Help: & Canary release Clusters k8s-cluster Name Type Time Created kubernetes ClusterIP 02/05/2019,15:58:07 new-nginx NodePort 02/17/2019,09:26:21 nginx-svc LoadBalancer 02/16/2019,16:41:42 old-nginx NodePort 02/17/2019,09:25:20	Service Help:	Service Help:	Service Help: IP Canary release Clusters kBs-cluster Name Type Time Created ClustersIP InternalEndpoint ExternalEndpoint kubernetes ClusterIP 02/05/2019,15:58:07 kubernetes:443 TCP new-nginx: NodePort 02/17/2019,09:26:21 new-nginx:30 TCP nginx-svc LoadBalancer 02/16/2019,16:41:42 nginx-svc:80 TCP old-nginx NodePort 02/17/2019,09:25:20 old-nginx:80 TCP old-nginx NodePort

4. In the displayed dialog box, click Confirm.

Note	×	r L
•	Are you sure to delete nginx-svc ?	
	Confirm Cancel	

1.8.16 Change container configurations

You can change the container configurations in the Container Service console.

Procedure

- 1. Log on to the Container Service console.
- 2. Click Kubernetes > Clusters in the left-side navigation pane.
- 3. Click Dashboard at the right of the cluster to enter the Kubernetes dashboard.

- 4. In the Kubernetes dashboard, click Pods in the left-side navigation pane.
- 5. Click the icon at the right of the pod and then select View/edit YAML.
- 6. The Edit a Pod dialog box appears. Change the container configurations and then click UPDATE.

1.8.17 Schedule a pod to a specified node

You can add a node label and then configure the nodeSelect or to schedule a pod to a specified node. For more information about the implementation principle of nodeSelector, see *nodeselector*.

For business scenario needs, to deploy a service used for management and control to a master node, or deploy services to a machine with an SSD disk, you can use this method to schedule pods to specified nodes.

Prerequisites

You have successfully created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Step 1 Add a node label

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Clusters > Nodes in the left-side navigation pane.
- 3. Select the cluster from the Cluster drop-down list and then click Label Management in the upper-right corner.
- 4. Select one or more nodes by selecting the corresponding check boxes and then click Add Tag. In this example, select a worker node.
- 5. Ener the name and value of the label in the displayed dialog box and then click OK.

The node label group : worker is displayed on the Label Management page.

You can also add a node label by running the command kubectl label nodes < node - name > < label - key >=< label - value >.

Step 2 Deploy a pod to a specified node

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Applications > Deployment in the left-side navigation pane.
- 3. Click Create by template in the upper-right corner.
- 4. Configure the template to deploy a pod. After completing the configurations, click DEPLOY.
 - · Clusters: Select a cluster.
 - Namespace: Select the namespace to which the resource object belongs. In this example, use default as the namespace.
 - Resource Type: Select Custom in this example.

The orchestration template in this example is as follows:

```
apiVersion :
              v1
 kind : Pod
metadata :
   labels :
    name : hello - pod
   name : hello - pod
 spec :
   containers :
      image : nginx
       imagePullP olicy :
                            IfNotPrese nt
       name : hello - pod
       ports :
                      ort :
           containerP
                              8080
           protocol :
                       TCP
       resources : {}
       securityCo ntext :
         capabiliti es : {}
privileged : false
       terminatio nMessagePa
                              th : / dev / terminatio
                                                        n – log
   dnsPolicy : ClusterFir
                           st
   restartPol icy : Always
   nodeSelect or :
     group : worker
                       ## The
                                                         label
                                            the
                                                  node
                                same
                                       as
configured in the preceding step.
```

status :{}

- 5. A message indicating the deployment status is displayed after you click DEPLOY . After the successful deployment, click Kubernetes Dashboard in the message to go to the dashboard and check the deployment status.
- 6. Click the pod name to view the pod details.

You can view the information such as the pod label and node ID, which indicates the pod is successfully deployed to a node with the label group : worker .

1.8.18 View image list

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > Docker Images in the left-side navigation pane.

Container Service - Kubernetes +	Im	age List Popular O	fficial Alibaba Cloud User Vser Favorites Search	Create Image Container Registry Console
Overview	He	lp: 🔗 Create image		
 Clusters 			mongo	
Clusters		• mongoDB	Type:Public Source:DOCKER_HUB ★822 MongoDB document databases provide high availability and easy scalability.	Create Application →
Volumes			Details Add to Collection	
Namespace		n.	mysql Type:Public_Source:DOCKER_HUB ★1314	Crate Application -
Authorization		MySQL.	MySQL is a widely used, open-source relational database management system (RDBMS).	
 Application 				
Deployment			nginx	
StatefulSet 🔤		NGINX	Official build of Nginx.	Create Application →
Pods			Details Add to Collection	
Service			php	
Ingress			Type:Public_Source:DOCKER_HUB ★396 While designed for web development, the PHP scripting language also provides general-purpose use.	Create Application →
Volumes Claim			Details Add to Collection	
Helm			postgres	
Release		(F)	Type:Public Source:DOCKER_HUB #233	Create Application →
Config Maps		PostgreSQL.	The PostgreSQL object-relational database system provides reliability and data integrity. Details Add to Collection	
Secret				
▼ Store		• "	redis	
Docker Images	/	🥌 redis	Redis is an open source key-value store that functions as a data structure server.	Create Application ->
Orchestration Te			Details Add to Collection	

You can view the image category.

- · Popular: Some common images recommended by Container Service.
- · Official: Official images provided by Docker Hub.

1.8.19 Use an image Secret

Container Service Kubernetes clusters support using image secrets through the web interface. You can create an image secret and use an existing image secret.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have built a private image repository and uploaded your image to the repository. In this example, use Alibaba Cloud Container Registry. For more information, see *Use a private image repository to create an application*.

Context

When you use a private image to create an application, you have to configure a secret for the image to secure the image. In the Container Service console, you can deliver the identity authentication information of the private image repository to Kubernetes through a secret of the docker-registry type.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Applications > Deployments. Then, click Create by Image in the upper-right corner.
- 3. Configure Name, Cluster, Namespace, Replicas, and Type. The configured value of the replicas parameter specifies the number of pods contained in the application. Click Next.



In this example, select the Deployment type.

If you do not configure Namespace, the system uses the default namespace by default.

Basic	Information	Container	>	Advanced	\rangle	Done
Name:	tomcat					
	The name should be 1-64 characters long, an	d can contain numbers, lowe	r case English letters a	nd hyphens, but cannot start wit	h a hyphen.	
Cluster:	-	v				
Namespace :	default	Ŧ				
Replicas:	2					
Туре	Deployment	Ŧ				
						Back Next

4. Configure containers.



This example describes only the configuration of the container image secret. For more information about container configuration, see *Create a deployment application by using an image*.

5. On the container configuration page, configure the image name first. Enter the private image address in the Image Name box. The format is domainname /

namespace / imagename .



Public images do not require image secrets.

6. In the image version box, enter the private image address version.

Container1	 Add Container 		
Image	e Name:	registry.cn-hangzhou.aliyuncs.com/dev-	Select image
Image	e Version:	V1	Select image version

- 7. Click Image pull secret.
 - · Select Create secret.
 - Name: Specifies the secret name. You can define it by yourself.
 - Repository Domain Name: Specified the Docker repository address. If you enter the Alibaba Cloud Container Service image repository in the image name box, the system automatically adds the repository address by default.
 - Username: Specifies the user name of the Docker repository. If you use Alibaba Cloud Container Registry, the username is your Alibaba Cloud account name.
 - Password: Specifies the logon password of the Docker repository. If you use Alibaba Cloud Container Registry, the password is the independent logon password for Container Registry.

Image pull secret		\times
	\odot Create secret \bigcirc Exist secret	
Name*	tomcat-secret	
Repository Domain Name*	registry.cn-hangzhou.aliyuncs.com	
Username*		
Password*		
Email		
	ок	Cancel

- Email: Specifies an email address. This is optional.

Click OK. The created secrete is displayed on the page.

Image Version:	V1	Select image version
	Always pull image Image pull secret	
	tomcat-secret	

· You can also click Exist secret. You can pre-create a container image secret by using command lines or a YAML file. For information, see How to use private

images in Kubernetes clusters and Use a private image repository to create an application.

Image pull secret		\times
Exist secret*	Create secret Exist secret	•
	aliyun-acr-credential-a aliyun-acr-credential-b tomcat-secret	OK Cancel

- 8. After you complete the container configuration, click Next.
- 9. Follow the page guide to complete other configurations, and then click Create.
- 10.Click Applications > Deployments in the left-side navigation pane, and select the cluster and namespace in which the application is created to view the status of the tomcat application.



Note:

The system shows that the tomcat application runs properly, which indicates that you have used the tomcat private image through the secret.

Container Service - Kubernetes -	Deployment Refresh Create by Image Create by Template	1
Overview	Help: If How to use private images If Create applications If Schedule a pod to the specified node If Create a Layer-4 Ingress If Create a Layer-7 Ingress If Configure pod auto scaling If Container monitoring If Blue-green release	
 Clusters 	Clusters Namespace default •	
Clusters	Name Tag PodsQuantity Image Time Created Action	
Nodes	tomcat app:tomcat 0/2 registry.cn-hangzhou.aliyuncs.com/dev-testcs/tomcat-private:V1 10/25/2018,22:20:20 Details Edit Scale Monitor More	
Volumes		
Namespace		_
Authorization		
✓ Applications		
Deployments		

1.8.20 Pull a private image without a password

This topic describes how to pull a private image without a password from the Alibaba Cloud container image repository.

Prerequisites

You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Context

Function overview

- You can only pull a private image from an Alibaba Cloud container image repository that belongs to your account.
- You can pull a private image from a cross-region Alibaba Cloud container image repository.
- You can only perform this operation in the default namespace.
- Kubernetes clusters that support this function include:
 - Dedicated Kubernetes clusters
 - Managed Kubernetes clusters
 - Serverless Kubernetes clusters

- The following are Kubernetes cluster versions that support this function:
 - Dedicated Kubernetes cluster versions that are not earlier than v1.11.2 support this function by default. If the dedicated Kubernetes cluster version is earlier than v1.11.2, follow the procedures described in this topic.
 - All versions of managed Kubernetes clusters support this function.
 - All versions of serverless Kubernetes clusters support this function.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. Click the target cluster name to view the cluster details.
- 4. In the Cluster Resources area, click Worker RAM Role.

Cluster Resource		
ROS	Reference for the second se	
Internet SLB	No. of Concession, Name of Con Name of Concession, Name of Concess	
VPC	and the standard in the standard standards	
NAT Gateway	the Country of Country	
Master RAM Role	NUMBER OF STREET, STREE	
Worker RAM Role	reliantiate delate a result of the delated residence.	



This topic uses the latest version of the RAM console.

If you use an earlier version of the RAM console, you can modify the target policy document by using either of the following two methods:

Method 1

a. In the left-side navigation pane, click Roles, and then enter the Worker RAM Role name in the Role Name box. Click the target Role Name.

Role Management		Create Role	C Refresh
Role Name K Searce	th		
Role Name	Created At		Actions
And a second state of a large state state of a large state stat	2018-12-28 15:46:49	Manage Authorize) Delete

b. In the Basic Information area, click Edit Basic Information in the upper-right corner.

Basic Information	Edit Basic Information
Role Name Ku	Description Grant ecs with kubernetes worker role.
Created At 2018-12-28 15:46:49	

Method 2

In the lower-right corner of the RAM dashboard page, click New Version to switch to the latest version of the RAM console. In the Container Service console, click Worker RAM Role to log on to the RAM console.

RAM	RAM Overview			
Dashboard Users	Welcome to Resource Acc	ess Management (RAM)		
Groups Policies Boles	RAM User Logon Link: http://signin.aliyun.com/cont	tainerdoc/login.htm		
Settings	User Overview	Group Overview	Permission Policy Overview	Role Overview
ActionTrail	You have 6 users	You have 0 groups	You have 15 custom authorization policies	You have 28 roles
	Operations Guide			
	 Manage authorization policies. Create and authorize groups. Create users and add them to groups. Authorization complete. 	RAM Pol Deathboard (By Uters 1.Choose Putces Aut Putces Aut	Iter Management term Paler Custom Paler Cu	We Alterate Park mpolicy according to your needs. No. of these references Automatical according to your needs. No. of these references No. o

- 5. On the RAM Roles page, click the policy name in the Permission area to view the policy details.
- 6. On the Policies page, click Modify Policy Document in the Policy Document area.

RAM	RAM / Policies / k8sWorkerRolePolicy-8b896e4c-64bf-4222-b7a9-9bb0d75b2fae
Overview	← k8sWc
Identities ^	Basic Information
Groups	Policy Name k8sWork Version Number v1
Users	Policy Type Custom Policy Note
Settings	
Permissions ^	Policy Document Versions References
Grants	« Modify Policy Document
Policies	1
RAM Roles	2 "Version": "1", 3 "Statement": [
OAuth Applications	4 { 5 "Action": [
	6 "ecs:AttachDisk", 7 "ecs:DetachDisk", 8 "ecs:DescribeDisks", 9 "ers:CreateDisk",

7. In the Policy Document area, add the following fields and then click OK.

```
{
    " Action ": [
    " cr : Get *",
    " cr : List *",
    " cr : PullReposi tory "
    ],
    " Resource ": "*",
    " Effect ": " Allow "
}
```

96 "Resource": [97 "*" 98], 99 "Effect": "Allow" 100], 101 { 102 "cr:Get*", 103 "cr:List*", 104 "cr:PullRepository" 106], 107 "Resource": "*", 108 "Effect": "Allow" 109] 110]	Policy Name k8sWork		
<pre>100 101 101 102 103 104 104 105 105 106 1, 106 1, 107 107 108 109 110 111]</pre>	96 97 98 99	<pre>"Resource": ["*"], "Effect": "Allow"</pre>	
110] 111]	100 101 102 103 104 105 106 107 108 109	<pre>{ "Action": ["cr:Get*", "cr:List*", "cr:PullRepository"], "Resource": "*", "Effect": "Allow" }</pre>	
	110] 111 }		

8. Create the aliyun - acr - credential - helper service to refresh the

temporary token of Container Registry at intervals.

```
apiVersion : v1
kind : ServiceAcc ount
metadata :
     name : aliyun - acr - credential - helper
     namespace : kube - system
apiVersion : rbac . authorizat ion . k8s . io / v1beta1
kind : ClusterRol eBinding
metadata :
     name : aliyun - acr - credential - helper - rolebindin g
     namespace : kube - system
roleRef :
     apiGroup : rbac . authorizat ion . k8s . io
     kind : ClusterRol e
     name : cluster - admin
subjects :
     kind : ServiceAcc ount
       name : aliyun - acr - credential - helper
       namespace : kube - system
# kubectl
           create secret docker - registry acr - image -
pull - secret - public -- docker - server = cr - tmp - xxx --
docker - username = cr - temp - xxx -- docker - password = cr -
temp - xxx -- docker - email = cr - temp - xxx
apiVersion : v1
data :
    . dockerconf igjson : eyJhdXRocy I6eyJjci10 bXAteHh4Ij
p7InVzZXJu YW1lIjoiY3 ItdGVtcC14 eHgiLCJwYX Nzd29yZCI6
ImNyLXRlbX AteHh4Iiwi ZW1haWwiOi Jjci10ZW1w LXh4eCIsIm
F1dGgi0iJZ M0l0ZEdWdG NDMTRlSGc2 WTNJdGRHVn RjQzE0ZUhn PSJ9fX0
kind : Secret
metadata :
     name : aliyun - acr - credential - a
     namespace : default
type : kubernetes . io / dockerconf igjson
# kubectl create secret docker - registry acr - image -
pull - secret - vpc -- docker - server = cr - tmp - xxx -- docker
- username = cr - temp - xxx -- docker - password = cr - temp -
xxx -- docker - email = cr - temp - xxx
apiVersion : v1
data :
    . dockerconf igjson : eyJhdXRocy I6eyJjci10 bXAteHh4Ij
p7InVzZXJu YW1lIjoiY3 ItdGVtcC14 eHgiLCJwYX Nzd29yZCI6
ImNyLXRlbX AteHh4Iiwi ZW1haWwiOi Jjci10ZW1w LXh4eCIsIm
F1dGgiOiJZ M0l0ZEdWdG NDMTRlSGc2 WTNJdGRHVn RjQzE0ZUhn PSJ9fX0
kind : Secret
metadata :
     name :
            aliyun - acr - credential - b
     namespace : default
type : kubernetes . io / dockerconf igjson
apiVersion : apps / v1beta2
kind : Deployment
metadata :
     name : aliyun - acr - credential - helper
     namespace : kube - system
```

```
labels :
        app : aliyun - acr - credential - helper
spec :
     replicas : 1
     selector :
       matchLabel s :
          app : aliyun - acr - credential - helper
     template :
       metadata :
          labels :
            app : aliyun - acr - credential - helper
       spec :
          serviceAcc ount : aliyun - acr - credential - helper
          containers :
            name : aliyun - acr - credential - helper
image : registry . cn - shanghai . aliyuncs . com / acs
/ aliyun - acr - credential - helper : v18 . 10 . 29 . 0 - 1a28f02
 aliyun
            imagePullP olicy : Always
          terminatio nGracePeri odSeconds : 0
```

1.9 Network management

1.9.1 Networks supported by Alibaba Cloud Container Service for Kubernetes

This topic describes the networks supported by Alibaba Cloud Container Service for Kubernetes.

Container networks

Container Service provides a stable and high-performance container network through its deep integration of the Kubernetes network and Alibaba Cloud Virtual Private Cloud (VPC). Container Service supports the following types of interconnections:

- · Pods within a container cluster can access each other.
- A pod can access a service within a container cluster.
- An Elastic Compute Service (ECS) instance can access a service within a container cluster.
- A pod can directly access an ECS instance (*) in the same VPC.
- An ECS instance can directly access a pod (*) in the same VPC.

Note:

The asterisk (*) indicates that you need to set a valid security group rule.

1.9.2 Terway network plugin

This topic describes how to use the Terway network plugin in a Kubernetes cluster that runs on Alibaba Cloud Container Service.

Terway network plugin

Terway, a network plugin developed by Alibaba Cloud Container Service, is fully compatible with Flannel, and provides the following features:

- · Allocates Alibaba Cloud Elastic Network Interfaces (ENIs) to containers.
- Defines the access policies for containers according to the Kubernetes Network Policy. This network plugin is also compatible with the Calico Network Policy.



If you install the Terway network plugin in a Kubernetes cluster, each pod then has its own network stack and an IP address. Packets between pods on one ECS instance are forwarded directly by the instance. Packets between pods on different ECS instances are forwarded through the VRouter of a VPC. The Terway network plugin delivers high communication performance because it does not use tunneling technologies such as VXLAN to encapsulate packets.

Use the Terway network plugin

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, click Clusters.

3. In the upper-right corner, click Create Kubernetes Cluster.

	Container Service - Kubernetes 👻	Cluster List			You can create up to 5 clust	ers and can add up to 40 nodes in each clu	uster. Refresh	Create Kubernetes Cluster
	Overview	Help: & Create cluster	${\mathscr S}$ Scale cluster ${\mathscr S}$ Connect to Kubernetes clu	ister via kubectl 🔗 Manage applic	ations with commands			2
•	Clusters	Name 🔻						
	Clusters	Cluster Name/ID	Cluster Type Region (All)	 Network Type 	Cluster Status Time Cr	reated Kubernetes Version		Action
	Nodes							

By default, the Create Kubernetes Cluster page is displayed.

Note:
In this example, a dedicated Kubernetes cluster is created. For more information,
see Create a Kubernetes cluster.

Create Kubernet	es Cluster 🛧 Bac	k to Cluster List	
Kubernetes	Multi-AZ Kubernetes		
* Cluster Name	k8s-cluster The cluster name : hyphens.	should be 1-63 characters long, and can contain nu	mbers, Chinese characters, English letters and

4. Select the Terway network plugin.

Network Plugin Flannel Terway (Compatible with Calico NetworkPolicy)							
How to choose network plugin of Kubernetes clusters							

Flannel and Terway

Alibaba Cloud Container Service for Kubernetes provides two types of network plugins for you to create a Kubernetes cluster: Terway and Flannel.



For how to select a network plugin, see *Do I* select the Terway or Flannel plugin for my *Kubernetes cluster network*?

 Flannel: a simple and stable community *Flannel* CNI plugin. Flannel can interoperate with the high-speed network of Alibaba Cloud VPC to provide a highperformance and stable container network for clusters. However, it provides a limited amount of features. For example, it does not support the Kubernetes Network Policy. Terway: a network plugin developed by Alibaba Cloud Container service. It is fully compatible with Flannel, and can allocate Alibaba Cloud Elastic Network Interfaces (ENIs) to containers. It can also define the access policies between containers according to the Kubernetes Network Policy. In addition, you can use this network plugin to limit the bandwidth traffic of a single container. If you do not need to use the Network Policy, we recommend that you select Flannel. In other cases, we recommend that you select Terway.

Note:

- Terway provides the same Network Policy as Calico because Terway is integrated with the Felix component of Calico. If you create a cluster to use Calico, you can use Terway to switch to Alibaba Cloud Container Service for Kubernetes.
- Terway is integrated with the Felix component V2.6.6.

1.9.3 Allocate an ENI to a pod

This topic describes how to allocate an Elastic Network Interface (ENI) to a pod.

Context

- When you create a Kubernetes cluster, you need to select Network Plugin as Terway. For more information, see *Create a Kubernetes cluster*.
- If you use a Kubernetes cluster that is installed with the Terway network plugin, you must make sure that the Terway plugin is V1.0.0.1 or later.

Note:

- 1. Log on to the Container Service console, click Clusters under the Kubernetes menu.
- 2. In the action column of the target cluster, choose More > Addon Upgrade.
- 3. On the Addon Upgrade page, view your current version of Terway.

4. Determine whether to upgrade according to Current Version and Upgradeable Version. If you want to upgrade Terway, click Upgrade in the action column.

Component	Current Version	Upgradeable Version	Consistency Check	Action	Statu
alicloud-application-controller	v0.1.0.1-f832bed-aliyun	v0.1.0.1-f832bed-aliyun	Success	Latest	
alicloud-disk-controller	v1.11.2.2-a390cfb-aliyun	v1.11.2.2-a390cfb-aliyun	Success	Latest	
Cloud Controller Manager & Readme & Version Information	v1.9.3.59-ge3bc999- aliyun	v1.9.3.59-ge3bc999- aliyun	Success	Latest	
flexvolume	v1.11.2.32-af2d48c-aliyun	v1.11.2.32-af2d48c-aliyun	Success	Latest	
Nginx Ingress Controller & Readme & Version Information	v0.20.0.1-4597ce2-aliyun	v0.20.0.1-4597ce2-aliyun	Success	Latest	
terway	v1.0.8.11-g323b1f3- aliyun	v1.0.8.11-g323b1f3- aliyun	Success	Latest	

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Application > Deployment.
- 3. In the upper-right corner, click Create by Template.

You can use the following YAML template to create a pod:

```
apiVersion : v1
kind : Pod
metadata :
    name : terway - pod
    labels :
        app : nginx
spec :
        containers :
        name : nginx
        image : nginx
        ports :
            containerP ort : 80
        resources :
            limits :
```



Result

1. In the left-side navigation pane under Kubernetes, choose Application > Pods. The pod named terway-pod is displayed.

Container Service - Kubernetes +	Pod Re	fresh
 Application 	Clusters CSK Namespace default	
Deployment	Name (c) Status (c) max attempts Pod IP (c) Node (c) Time Created (c) CPU Memory	
StatefulSet	terway-pod Running 0 12/28/2018,16:14:07 0 ODetails	More 🗸
Job	Total: 1 item(s), Per Page: 20 item(s)	> >>
CronJob		
Pods		

- 2. In the left-side navigation pane under Kubernetes, click Clusters.
- 3. Click the name of the target cluster to view the cluster details.
- 4. In the Cluster Resource area, click VPC to view the VPC CIDR block of the cluster.
- 5. Run the following command to obtain the IP address of the deployed pod and verify that the IP address is within the VPC CIDR block of the cluster:

\$ kubectl get pod - o wide

1.9.4 Use a network policy

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have selected the Terway network plugin when creating the Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have connected to the Kubernetes cluster by using kubectl, see Connect to a Kubernetes cluster by using kubectl.

Verify that an Nginx service is accessible to pods

1. Run the following command to create an Nginx application and expose it through a service named Nginx:

```
kubectl
            run
                  nginx -- image = nginx
$
deployment . apps / nginx created
                  pod
  kubectl
            get
NAME
                           READY
                                     STATUS
                                                 RESTARTS
                                                              AGE
nginx - 64f497f8fd - znbxb
                             1 / 1
                                           Running
45s
```

```
deployment nginx -- port = 80
$ kubectl
            expose
service / nginx
                  exposed
$ kubectl
                  service
            get
NAME
               TYPE
                              CLUSTER - IP
                                               EXTERNAL - IP
PORT (S)
               AGE
kubernetes
               ClusterIP
                             172 . 19 . 0 . 1
                                                   < none >
443 / TCP
               3h
                             172 . 19 . 8 . 48
nginx
               ClusterIP
                                                   < none >
80 / TCP
               10s
```

2. Run the following command to create a pod named busybox and use the pod to

access the Nginx service created in step 1:

\$ kubectl run busybox -- rm - ti -- image = busybox / bin / sh run -- generator = deployment / apps . v1beta1 kubectl is will be removed DEPRECATED and version future in а . Use kubectl create instead . If you don't command prompt , try see а pressing enter / # wget nginx nginx (172.19.8.48:80) Connecting to 100 % index . html 612 0 : 00 : 00 ETA / #

Use a network policy to set the Nginx service to be accessible only to a specifically labeled application

1. Run the following command to create a policy . yaml file:

```
$ vim
        policy . yaml
kind :
        NetworkPol icy
apiVersion : networking . k8s . io / v1
metadata :
  name : access - nginx
spec :
  podSelecto r :
    matchLabel
                s :
      run :
             nginx
  ingress :
   from :
      podSelecto r :
        matchLabel s :
          access : " true "
```

2. Run the following command to create a network policy according to the policy.

yaml file created in step 1:

\$ kubectl apply - f policy . yaml

networkpol icy . networking . k8s . io / access - nginx created

3. Run the following command to verify that the Nginx service cannot be accessed if you do not define any access label in the command:

```
busybox -- rm - ti -- image = busybox / bin /
$ kubectl
           run
sh
          don ' t
Ιf
     you
                   see
                        а
                            command
                                     prompt, try
                                                   pressing
enter .
/ # wget
          nginx
Connecting to
                nginx (172.19.8.48:80)
wget: can ' t
                connect to
                                      host (172.19.8.48
                              remote
): Connection
               timed
                      out
/ #
```

4. Run the following command to verify that the Nginx service can be accessed if an access label is defined in the command:

```
busybox -- rm - ti -- labels =" access = true
kubectl
        run
" -- image = busybox / bin / sh
If you
        don ' t
               see a command
                             prompt, try
                                        pressing
enter .
        nginx
/ # wget
             nginx (172.19.8.48:80)
Connecting
         to
index . html
                 100 % |
0 : 00 : 00 ETA
  612
/ #
```

Use a network policy to specify a source IP CIDR block that can access a service exposed by an SLB service over the Internet

 Run the following command to create an Alibaba Cloud SLB service for the preceding Nginx application, that is, specify type = LoadBalanc er to expose the Nginx service to the Internet:

```
$ vim
        nginx - service . yaml
apiVersion : v1
kind : Service
metadata :
  labels :
    run : nginx
  name : nginx - slb
spec :
  externalTr afficPolic y : Local
  ports :
    port : 80
    protocol :
               ТСР
    targetPort :
                 80
  selector :
    run : nginx
  type : LoadBalanc er
$ kubectl apply - f nginx - service . yaml
service / nginx - slb
                       created
$ kubectl get service
                           nginx - slb
```

NAME	TYPE	CLI	JSTER - IP		EXTERNAL		IP
PORT (S) AGE						
nginx - slb	LoadBalanc	er	172 . 19 .	12.	254	47	. 110
. 200 . 119	80 : 32240	/ TCP	8m				

2. Run the following command to verify that the IP address of the created SLB

service, that is, 47.110.200.119, cannot be accessed:

```
$ wget 47 . 110 . 200 . 119
-- 2018 - 11 - 21 11 : 46 : 05 -- http :// 47 . 110 . 200 . 119
/
Connecting to 47 . 110 . 200 . 119 : 80 ... failed :
Connection refused .
```



Access failure occurs due to the following reasons:

- You have configured access to the Nginx service only for the applications labeled with access = true .
- You have attempted to access the IP address of the SLB instance from outside the Kubernetes system. This is different from Use a network policy to set the Nginx service to be accessible only to a specifically labeled application.

Solution: Modify the network policy and add a source IP CIDR block that is allowed to access the Nginx service.

3. Run the following command to view your local IP address:

```
$ curl myip.ipip.net
IP address: 10.0.0.1 from: China Beijing Beijing
# The local IP address varies by devices.
```

4. Run the following command to modify the created *policy* . yaml file:

```
policy . yaml
$ vim
kind : NetworkPol icy
apiVersion : networking . k8s . io / v1
metadata :
  name : access - nginx
spec :
  podSelecto r :
    matchLabel s :
      run : nginx
  ingress :
   from :
      podSelecto r :
        matchLabel s :
          access : " true "
      ipBlock :
        cidr: 100.64.0.0/10
   - ipBlock :
```

```
cidr : 10 . 0 . 0 . 1 / 24 # Set the CIDR
block to which the local IP address belongs . This
is an example . Set the required parameters according
to your device .
$ kubectl apply - f policy . yaml
networkpol icy . networking . k8s . io / access - nginx
unchanged
```

Note:

- The outgoing interface of a network may have multiple IP addresses. We recommend that you specify an entire CIDR block.
- The SLB health check address belongs to the 100 . 64 . 0 . 0 / 10 CIDR block. Therefore, you must specify the 100 . 64 . 0 . 0 / 10 CIDR block.
- 5. Run the following command to verify that the Nginx service can be accessed:

```
busybox -- rm - ti -- labels =" access = true
  kubectl
          run
" -- image = busybox / bin / sh
If you
        don't see a command
                                prompt , try
                                            pressing
enter .
       47 . 110 . 200 . 119
/ # wget
Connecting to 47 . 110 . 200 . 119 ( 47 . 110 . 200 . 119 : 80
                   100 % |
index . html
612
  0 : 00 : 00
            ETA
 #
```

Use a network policy to set a pod that can access only www . aliyun . com

1. Run the following command to obtain the IP address list resolved from the domain

name of www . aliyun . com :

\$ dig + short www . aliyun . com www - jp - de - intl - adns . aliyun . com . www - jp - de - intl - adns . aliyun . com . gds . alibabadns . com . v6wagbridg e . aliyun . com . v6wagbridg e . aliyun . com . gds . alibabadns . com . 106 . 11 . 93 . 21 140 . 205 . 32 . 4 140 . 205 . 230 . 13 140 . 205 . 34 . 3

2. Run the following command to create a busybox - policy file:

```
$ vim busybox - policy . yaml
kind : NetworkPol icy
apiVersion : networking . k8s . io / v1
metadata :
    name : busybox - policy
spec :
    podSelecto r :
```

```
matchLabel s :
   run : busybox
egress :
 to :
   ipBlock :
     cidr :
             106 . 11 . 93 . 21 / 32
   ipBlock :
     cidr : 140 . 205 . 32 . 4 / 32
   ipBlock :
     cidr : 140 . 205 . 230 . 13 / 32
   ipBlock :
     cidr : 140 . 205 . 34 . 3 / 32
 to :
   ipBlock :
     cidr: 0.0.0.0/0
 ports :
   protocol : UDP
   port : 53
```

Note:

In the preceding *busybox* – *policy* file, an egress rule is set to specify the CIDR blocks that can be accessed by cluster applications. You need to set the condition that UDP requests are allowed. Otherwise, DNS resolution will fail.

- 3. Run the following command to create a network policy according to the busybox
 - policy file:

```
$ kubectl apply - f busybox - policy . yaml
networkpol icy . networking . k8s . io / busybox - policy
created
```

- 4. Run the following command to verify that no website (for example, www . google
 - . com) can be accessed except for www . aliyun . com :

\$ kubectl busybox -- rm - ti -- image = busybox / bin / run sh you Ιf don 't see а command prompt, try pressing enter . Connecting to www.google.com (64.13.192.74:80) wget: can't connect to remote bost (64.13.192.74:80) / # wget www . google . com 74): Connection timed out

5. Run the following command to verify that www . aliyun . com can be accessed:

/ # wget www . aliyun . com Connecting to www.aliyun.com (140.205.34.3:80) www . aliyun . com (140 . 205 . 34 . 3 : 443) Connecting to wget : note : TLS certificat e validation not implemente d index . html 100 % 462k 0:00:00 ETA
/ #

1.10 Server Load Balancer and Ingress management

1.10.1 Overview

Kubernetes clusters provide a diversity of approaches to access container applications, and support accessing internal services and realizing load balancing by means of Alibaba Cloud Server Load Balancer or Ingress.

1.10.2 Access services by using Server Load Balancer

This topic describes how to access services by using Alibaba Cloud Server Load Balancer (SLB).

Check the cloud-controller-manager version

If you specify an existing SLB in a cluster that has a cloud-controller-manager component of v1.9.3 or later versions, the system does not process listeners for this SLB by default. You must manually configure listeners for this SLB.

To view the cloud-controller-manager version, run the following command:

```
root @ master # kubectl get po - n kube - system - o yaml
| grep image :| grep cloud - con | uniq
image : registry - vpc . cn - hangzhou . aliyuncs . com / acs /
cloud - controller - manager - amd64 : v1 . 9 . 3
```

Use a command-line tool

Method 1

1. Create an Nginx application by using a command-line tool.

root @ master # kubectl nginx -- image = registry . run aliyuncs . com / acs / netdia : latest root @ master # kubectl get po NAME READY **STATUS** RESTARTS AGE nginx - 2721357637 - dvwq3 1 / 1 Running 1 65

2. Create an SLB service for the Nginx application and specify type = LoadBalanc

er to expose the Nginx service to the Internet.

```
root @ master # kubectl expose deployment nginx -- port =
80 -- target - port = 80 -- type = LoadBalanc er
root @ master # kubectl get svc
```

NAME	CLUSTER - IP	EXTERNAL - IP
PORT (S)	AGE	
nginx	172 . 19 . 10 .	209 101 . 37 . 192 . 20
80 : 31891 / TCP		4s

3. Visit http://101.37.192.20 in a browser to access your Nginx service.

Method 2

1. Save the following yml code to the nginx - svc . yml file:

```
apiVersion : v1
kind : Service
metadata :
    labels :
    run : nignx
name : nginx - 01
namespace : default
spec :
    ports :
    - port : 80
    protocol : TCP
    targetPort : 80
    selector :
    run : nginx
    type : LoadBalanc er
```

2. Run the kubectl apply - f nginx - svc . yml command.

root @ master # kubectl apply - f nginx - svc . yml root @ master # kubectl service get NAME CLUSTER - IP EXTERNAL - IP TYPE PORT (S) AGE9d ngi - 01nx LoadBalanc er 80 : 32325 / TCP 37 . 192 . 129 172 . 19 . 9 . 243 101 . 3h

3. Visit http://101.37.192.129 in a browser to access your Nginx

service.

Use the Kubernetes dashboard

1. Save the following yml code to the nginx - svc . yml file:

```
apiVersion : v1
kind : Service
metadata :
   labels :
    run : nginx
   name : http - svc
   namespace : default
spec :
   ports :
    port : 80
    protocol : TCP
    targetPort : 80
   selector :
    run : nginx
```

type : LoadBalanc er

- 2. Log on to the *Container Service console* and click Dashboard on the right of the target cluster.
- 3. Click CREATE in the upper-right corner to create an application.

<	≡ Overview								+ CREATE
Cluster Namespaces	CPU usage				Memory usage	(i)			
Nodes	0.001 0.001				257 Mi				
Persistent Volumes Roles	0.0003				Ag 172 Mi Ag 114 Mi By 57.2 Mi				
Storage Classes	0 18:27 18:30	18:33 Time	18:36	18:41	0 18:27	18:30	18:33 Time	18:36	18:41
default 👻	Workloads								Contact
Overview Workloads	Workloads Statuses								R

- 4. Click the CREATE FROM FILE tab. Select the nginx svc . yml file you saved.
- 5. Click UPLOAD.

An SLB instance that points to the created Nginx application is created. The service name is http-svc .

6. In the left-side navigation pane on the dashboard page, select the default namespace, and then click Services .

You can view the created Nginx service http - svc and the SLB address http :// 114 . 55 . 79 . 24 : 80 .

<	\equiv Discovery and load bala
Cluster	Services
Namespaces	Name 🚖
Nodes	
Persistent Volumes	Mttp-svc
Roles	kubernetes
Storage Classes	
Namespace	
default 👻	
Overview	
Workloads	
Cron Jobs	
Daemon Sets	
Deployments	
Jobs	

Issue: 20190525 Pods 7. Open this address in your browser to access the service.

Use the Container Service console

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Applications > Deployments.
- 3. Select the target cluster and namespace, and then click Create by Template in the upper-right corner.

Container Service - Kubernetes 🕶	Deployment Refresh Create by Image Create by Template
Overview	How to use private images O Create applications O Schedule a pod to the specified node O Create a Layer-4 Ingress O Create a Layer-7 Ingress O Configure pod auto scaling
 Clusters 	Container monitoring Blue-green release Clusters Kec-teet Namespaces Idefault
Clusters	
Nodes	Name Tag PodsQuantity Image Time Created Action
Volumes	O Could not find any record that met the condition.
Namespaces	
Authorization	
 Applications 	
Deployments	
Stateful Sets	
Daemon Sets	

4. Select the custom Resource Type and then copy the following code to the Template.

```
apiVersion : v1
kind : Service
metadata :
    labels :
    run : nginx
    name : ngnix
    namespace : default
spec :
    ports :
    port : 80
    protocol : TCP
    targetPort : 80
    selector :
    run : nginx
    type : LoadBalanc er
```

5. Click DEPLOY.

6. Click Kubernetes Dashboard to check the deployment progress on the dashboard

page.

Namespace	≡ Overview						+ CREATE
default 👻	Discovery and Load Ba	lancing					
Overview	,	5					
Workloads	Services						Ŧ
Cron Jobs	Name 🖨	Labels	Cluster IP	Internal endpoints	External endpoints	Age 🜲	
Daemon Sets	ngnix	run: nginx	12125844	ngnix:80 TCP ngnix:		12/03/2018, 17:02:18	:
Deployments Jobs	kubernetes	component: apiserver provider: kubernetes	1203003	kubernetes: TCP		11/22/2018, 20:42:57	:

Alternatively, choose Discovery and Load Balancing > Services in the left-side navigation pane, and select the target cluster and namespace to view the deployed service.

Container Service - Kubernetes 👻	Service							Refresh Create
Overview	Canary r	release						
Clusters	Clusters	k8s-test 🔻	Namespaces de	fault 🔻				Search By Name Q
 Applications 	Name	Label	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint	Action
 Service Mesh 	kubernetes	component:apiserver provider:kubernetes	ClusterIP	05/17/2019,18:12:33		kubernetes:443 TCP	-	Details Update View YAML Delete
Virtual Services Discovery and Load B	ngnix	run:nginx	LoadBalancer	05/23/2019,11:11:03		ngnix:80 TCP ngnix: TCP	11111	Details Update View YAML Delete
Services	tf-notebook	k -	LoadBalancer	05/23/2019,10:46:02		tf-notebook:80 TCP tf-notebook:30708 TCP		Details Update View YAML Delete
Ingresses								

More information

Alibaba Cloud SLB also supports a lot of parameters such as health checks, billing methods, and SLB types. For more information, see SLB configuration parameters.

Annotations

Alibaba Cloud supports plenty of SLB features by using annotation s.

Use an existing intranet SLB instance

You must specify three annotations. Replace "your-loadbalancer-id" with your SLB instance ID.



Note:

Multiple Kubernetes services can reuse the same SLB instance.

• The SLB instances created by Kubernetes through a service cannot be reused. Otherwise, the reused SLB instances may be removed incidentally. Only the SLB instances manually created in the console or created by calling API can be reused.

- The multiple services that reuse the same SLB instance cannot have the same frontend listening port. Otherwise, port conflicts will occur.
- If you reuse an SLB instance, the listener name and the virtual server group name are used to identify the SLB instance in Kubernetes. We recommend that you do not modify the listener name or virtual server group name.
- You can modify SLB instance names.
- · SLB instances cannot be reused across clusters.

```
apiVersion :
              v1
kind : Service
metadata :
  annotation s:
    service . beta . kubernetes . io / alicloud - loadbalanc er -
address - type : " intranet "
service . beta . kubernetes . io / alicloud - loadbalanc er - id
: " your - loadbalanc er - id "
service . beta . kubernetes . io / alicloud - loadbalanc er -
force - override - listeners : " true "
  labels :
  run : nginx
name : nginx
  namespace : default
spec :
  ports :
   port : 80
    protocol : TCP
    targetPort : 80
  selector :
    run : nginx
  sessionAff inity : None
  type : LoadBalanc
                         er
```

Create an HTTP-type SLB instance

```
apiVersion : v1
kind : Service
metadata :
 annotation s:
   service . beta . kubernetes . io / alicloud - loadbalanc er -
protocol - port : " http : 80 "
 name : nginx
 namespace : default
spec :
 ports :
   port: 80
   protocol : TCP
   targetPort: 80
  selector :
   run : nginx
  type : LoadBalanc er
```

Create an HTTPS-type SLB instance

You must first create a certificate in the Alibaba Cloud console before creating an HTTPS-type SLB instance by using the following template (the certificate ID is required by the annotations in the template):

```
apiVersion : v1
kind : Service
metadata :
    annotation s :
    service . beta . kubernetes . io / alicloud - loadbalanc er -
cert - id : " your - cert - id "
    service . beta . kubernetes . io / alicloud - loadbalanc er -
protocol - port : " https : 443 "
    name : nginx
    namespace : default
spec :
    ports :
    - port : 443
    protocol : TCP
    targetPort : 443
    selector :
    run : nginx
    sessionAff inity : None
    type : LoadBalanc er
```

Limit SLB instance bandwidth

```
apiVersion : v1
kind : Service
metadata :
  annotation s:
   service . beta . kubernetes . io / alicloud - loadbalanc er -
charge - type : " paybybandw idth "
service . beta . kubernetes . io / alicloud - loadbalanc er -
bandwidth : " 100 "
  name : nginx
  namespace : default
spec :
  ports :
    port : 443
    protocol : TCP
    targetPort : 443
  selector :
    run : nginx
          LoadBalanc er
  type :
```

Specify the SLB instance specification

```
apiVersion : v1
kind : Service
metadata :
    annotation s :
        service . beta . kubernetes . io / alicloud - loadbalanc er -
spec : " slb . s1 . small "
    name : nginx
    namespace : default
spec :
    ports :
        - port : 443
        protocol : TCP
```

```
targetPort : 443
selector :
  run : nginx
type : LoadBalanc er
```

Use an existing SLB instance

By default, listeners are not overridden if you use an existing SLB instance. To forcibly override the existing listeners, set service . beta . kubernetes . io / alicloud - loadbalanc er - force - override - listeners to true .



Note:

Multiple Kubernetes services can reuse the same SLB instance.

- The SLB instances created by Kubernetes through a service cannot be reused. Otherwise, the reused SLB instances may be removed incidentally. Only the SLB instances manually created in the console or created by calling API can be reused.
- The multiple services that reuse the same SLB instance cannot have the same frontend listening port. Otherwise, port conflicts will occur.
- If you reuse an SLB instance, the listener name and the virtual server group name are used to identify the SLB instance in Kubernetes. We recommend that you do not modify the listener name or virtual server group name.
- You can modify SLB instance names.
- SLB instances cannot be reused across clusters.

```
apiVersion : v1
kind : Service
metadata :
 annotation s:
   service . beta . kubernetes . io / alicloud - loadbalanc er - id
: " your_loadb alancer_id "
 name : nginx
 namespace : default
spec :
 ports :
   port : 443
   protocol : TCP
   targetPort : 443
 selector :
    run : nginx
         LoadBalanc er :
                           LoadBalanc
  type :
                                       er
```

Use an existing SLB instance and forcibly override existing listeners

If you forcibly override the existing listeners, they are removed.

Note:

Multiple Kubernetes services can reuse the same SLB instance.

- The SLB instances created by Kubernetes through a service cannot be reused. Otherwise, the reused SLB instances may be removed incidentally. Only the SLB instances manually created in the console or created by calling API can be reused.
- The multiple services that reuse the same SLB instance cannot have the same frontend listening port. Otherwise, port conflicts will occur.
- If you reuse an SLB instance, the listener name and the virtual server group name are used to identify the SLB instance in Kubernetes. We recommend that you do not modify the listener name or virtual server group name.
- You can modify SLB instance names.
- · SLB instances cannot be reused across clusters.

```
apiVersion : v1
kind : Service
metadata :
 annotation s:
   service . beta . kubernetes . io / alicloud - loadbalanc er - id
: " your_loadb alancer_id "
   service . beta . kubernetes . io / alicloud - loadbalanc er -
force - override - listeners : " true "
 name : nginx
 namespace : default
spec :
 ports :
  port : 443
   protocol : TCP
   targetPort : 443
 selector :
   run : nginx
  type : LoadBalanc ere : LoadBalanc
                                        er
```

Use the Worker node with specified labels as a backend server

Use a comma (,) to separate two labels, for example, K1: V1, K2: V2.

The relationship between multiple labels is and .

```
apiVersion : v1
kind : Service
metadata :
    annotation s :
        service . beta . kubernetes . io / alicloud - loadbalanc er -
backend - label : " failure - domain . beta . kubernetes . io / zone
: ap - southeast - 5a "
    name : nginx
    namespace : default
spec :
    ports :
    - port : 443
    protocol : TCP
    targetPort : 443
    selector :
```

run : nginx type : LoadBalanc er

Set the session persistence timeout for a TCP-type SLB instance

The parameter service . beta . kubernetes . io / alicloud - loadbalanc er - persistenc e - tim applies only to TCP listeners.

If the SLB instance is configured with multiple TCP listener ports, this parameter setting applies to all the ports by default.

```
apiVersion : v1
kind : Service
metadata :
  annotation
              s:
service . beta . kubernetes . io / alicloud - loadbalanc er - persistenc e - timeout : " 1800 "
  name : nginx
  namespace : default
spec :
  ports :
    port : 443
    protocol : TCP
    targetPort : 443
  selector :
    run : nginx
  type :
           LoadBalanc
                        er
```

Set session persistence for HTTP-type and HTTPS-type SLB instances (insert cookie)

Only HTTP-type and HTTPS-type SLB instances support this setting.

If an instance is configured with multiple HTTP or HTTPS listener ports, the session persistence setting applies to all the HTTP or HTTPS listener ports by default.

```
apiVersion : v1
kind : Service
metadata :
  annotation s:
    service . beta . kubernetes . io / alicloud - loadbalanc er -
sticky - session : " on "
    service . beta . kubernetes . io / alicloud - loadbalanc er -
sticky - session - type : " insert "
    service . beta . kubernetes . io / alicloud - loadbalanc er -
cookie - timeout : " 1800 "
service . beta . kubernetes . io / alicloud - loadbalanc er -
protocol - port : " http : 80 "
  name : nginx
  namespace : default
spec :
  ports :
   port : 80
    protocol :
               TCP
    targetPort : 80
  selector :
    run : nginx
```

type : LoadBalanc er

Set session persistence for HTTP-type and HTTPS-type SLB instances (server cookie)

Only HTTP-type and HTTPS-type SLB instances support this setting.

If an instance is configured with multiple HTTP or HTTPS listener ports, the session persistence setting applies to all the HTTP or HTTPS listener ports by default.

```
apiVersion :
                 v1
kind : Service
metadata :
  annotation
                 s:
     service . beta . kubernetes . io / alicloud - loadbalanc er -
sticky - session : " on "
     service . beta . kubernetes . io / alicloud - loadbalanc er -
sticky - session - type : " server "
service . beta . kubernetes . io / alicloud - loadbalanc er -
cooyour_co okie : " your_cooki e "
    service . beta . kubernetes . io / alicloud - loadbalanc er -
protocol - port : " http : 80 "
    name : nginx
  namespace :
                   default
spec :
  ports :
     port :
              80
                     TCP
     protocol :
     targetPort : 80
  selector :
     run : nginx
   type :
             LoadBalanc
                             er
```

Specify the primary and secondary zones when creating an SLB instance

Support for the primary and secondary zones varies according to region, for example , the ap-southeast-5.

Once created, the primary and secondary zones cannot be changed.

```
apiVersion : v1
kind : Service
metadata :
  annotation s:
    service . beta . kubernetes . io / alicloud - loadbalanc er -
master - zoneid : " ap - southeast - 5a "
    service . beta . kubernetes . io / alicloud - loadbalanc er -
slave - zoneid : " ap - southeast - 5a "
  name : nginx
  namespace : default
spec :
  ports :
   port: 80
              TCP
    protocol :
    targetPort : 80
  selector :
    run : nginx
```

type : LoadBalanc er

Use the node where the pod is located as a backend server

```
apiVersion : v1
kind : Service
metadata :
   name : nginx
   namespace : default
spec :
   externalTr afficPolic y : Local
   ports :
        port : 80
        protocol : TCP
        targetPort : 80
        selector :
        run : nginx
   type : LoadBalanc er
```



The annotations are case sensitive.

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - protocol - port	Use a comma (,) to separate two values, for example, https:443,http:80	None
service . beta . kubernetes . io / alicloud - loadbalanc er - address - type	Valid values: internet or intranet .	internet
service . beta . kubernetes . io / alicloud - loadbalanc er - slb - network - type	SLB instance network type. Valid values: classic or vpc .	classic
service . beta . kubernetes . io / alicloud - loadbalanc er - charge - type	Valid values: paybytraff ic or paybybandw idth .	paybytraff ic

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - id	SLB instance ID. You can specify an existing SLB instance by using service . beta . kubernetes . io / alicloud - loadbalanc er - id , and existing listeners will be overridden. Note that the SLB instance will not be deleted if you delete the service.	None
service . beta . kubernetes . io / alicloud - loadbalanc er - backend - label	Use labels to specify the Worker nodes to be mounted to the backend of the SLB instance.	None
service . beta . kubernetes . io / alicloud - loadbalanc er - spec	SLB instance specification. For more information, see <i>#unique_120</i> .	None
service . beta . kubernetes . io / alicloud - loadbalanc er - persistenc e - timeout	Session persistence timeout (in seconds). This parameter setting applies only to TCP listeners and the value can be 0 to 3600. The default value is 0, indicating that the session remains disabled. For more information, see #unique_121.	0

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc	Whether to enable session persistence. Valid value: on off .	off
er - sticky - session	Note: It applies only to HTTP and HTTPS listeners.	
	For more information,	
	#unique_123.	
service . beta . kubernetes . io / alicloud - loadbalanc er - sticky - session - type	Method used to handle the cookie. Valid values: · insert : Insert the cookie. · server : Rewrite the cookie.	None
	 Note: It applies only to HTTP and HTTPS listeners. If you set the value of the parameter service . beta kubernetes io / alicloud - loadbalanc er - sticky - session to on , you must specify this parameter. For more information, see #unique_122 and #unique_123. 	

Annotation	Description	Default value
Annotation service . beta . kubernetes . io / alicloud - loadbalanc er - cookie - timeout	Description Cookie timeout period (in seconds). Value range: 1 to 86400. Note: If the service . beta . kubernetes . io / alicloud - loadbalanc er - sticky - session parameter is set to on and the service . beta . kubernetes . io / alicloud - loadbalanc er - sticky - session - type parameter	Default value None
	is set to insert, this parameter is mandatory.	
	For more information, see #unique_122 and #unique_123.	

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc	Cookie configured on the server.	None
er – cookie	The cookie must be	
	a string of 1 to 200	
	contain ASCII letters and	
	numeric characters. It	
	cannot contain commas	
	(,), semicolons (;), or	
	spaces, and it cannot start	
	with a dollar sign (\$).	
	Note:	
	If the service .	
	beta . kubernetes	
	. io / alicloud -	
	loadbalanc er -	
	sticky - session	
	parameter is set to on	
	and the service .	
	beta . kubernetes	
	. io / alicloud -	
	loadbalanc er -	
	sticky - session	
	- type parameter	
	is set to server, this	
	parameter is mandatory.	
	For more information,	
	see #unique_122 and	
	#unique_123.	

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - master - zoneid	Zone ID of the primary backend server.	None
service . beta . kubernetes . io / alicloud - loadbalanc er - slave - zoneid	Zone ID of the secondary backend server.	None
externalTr afficPolic y	Nodes that can be used as backend servers. Valid values: · Cluster : Use all backend nodes as back- end servers. · Local : Use the nodes where pods are located as backend servers.	Cluster
service . beta . kubernetes . io / alicloud - loadbalanc er - force - override - listeners	Determines whether to override the listeners when you specify an existing SLB instance.	false : Do not override.
service . beta . kubernetes . io / alicloud - loadbalanc er - region	Region where the SLB instance is located.	None
service . beta . kubernetes . io / alicloud - loadbalanc er - bandwidth	SLB instance bandwidth.	50
service . beta . kubernetes . io / alicloud - loadbalanc er - cert - id	ID of a certificate on Alibaba Cloud. You must upload a certificate first.	None

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - flag	Valid values: on off	The default value is off. Modifying this parameter is not required for TCP, because the health check function is enabled for TCP by default and this parameter cannot be set.
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - type	Health check type. Valid values: tcp http . For more information, see #unique_121.	tcp
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - uri	URI used for health checks. Note: If the health check type is TCP, you do not need to set this parameter. For more information, see #unique_121.	None
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - connect - port	 Port used for health checks. Valid values: - 520 : The backend port configured for the listener is used by default. 1 - 65535 : The port opened on the backend server for health checks is used. For more information, see #unique_121. 	None

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - healthy - threshold	For more information, see <i>#unique_121</i> .	None
service . beta . kubernetes . io / alicloud - loadbalanc er - unhealthy - threshold	The number of consecutive health check successes before the backend server is determined healthy (from failure to success). Value range: 2 to 10 For more information, see <i>#unique_121</i> .	None
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - interval	Time interval between two consecutive health checks (seconds). Value range: 1 to 50 For more information, see <i>#unique_121</i> .	None

Annotation	Description	Default value
<pre>service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - connect - timeout</pre>	Time period required by waiting for a health check response (in seconds). If the backend ECS instance does not send a valid response within a specified period of time, the system determines that the health check has failed. Value range: 1 to	None
	300 .	
	If the value of the	
	parameter service .	
	beta . kubernetes	
	. 10 / alicloud -	
	loadbalanc er -	
	nealth - check -	
	is loss then the value of	
	the parameter sorvice	
	heta kubernetes	
	io / alicloud -	
	loadbalanc er -	
	health - check -	
	interval, service	
	. beta . kubernetes	
	. io / alicloud -	
	, loadbalanc er –	
	health - check -	
	connect - timeout	
	is invalid and the	
	timeout period equals	Issue: 20190525
	the value of service	

Annotation	Description	Default value
service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - timeout	Time period required by waiting for a health check response (in seconds). If the backend ECS instance does not send a valid response within a specified period of time, the system determines that the health check has failed.	None
	Value range: 1 to	
	<pre>Note: If the value of the parameter service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - timeout is less than that of the parameter service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - interval , service . beta . kubernetes . io / alicloud - loadbalanc er - health - check - timeout is invalid, and the timeout period equals the value of the parameter service . beta . kubernetes . io / alicloud -</pre>	
	loadbalanc er - health - check - interval .	
: 20190525	For more information, see	299
	#unique_122.	

1.10.3 Support for Ingress

In Kubernetes clusters, Ingress is a collection of rules that authorize inbound connection to the cluster services and provides you with Layer-7 Server Load Balancer capabilities. You can provide the Ingress configuration with externally accessible URL, Server Load Balancer, SSL, and name-based virtual host.

Prerequisites

To test the complex routing service, create an Nginx application in this example. You must create the Nginx deployment and multiple services in advance to observe the routing effect. Replace with your own service in the actual test. In the actual test enter your own service.

root @ master # kubectl nginx -- image = registry . cn run hangzhou . aliyuncs . com / acs / netdia : latest root @ master # kubectl nginx -- name = http expose deploy svc -- port = 80 -- target - port = 80 root @ master # kubectl expose nginx -- name = http deploy svc1 -- port = 80 -- target - port = 80 root @ master # kubectl expose nginx -- name = http deploy svc2 -- port = 80 -- target - port = 80 root @ master # kubectl expose nginx -- name = http deploy svc3 -- port = 80 -- target - port = 80

Simple routing service

Create a simple Ingress service by using the following commands. All the accesses to the / svc path are routed to the Nginx service. nginx . ingress . kubernetes . io / rewrite - target : / redirects the path / svc to the path / that can be recognized by backend services.

```
root @ master # cat << EOF |
apiVersion : extensions / v1beta1</pre>
                                      kubectl
                                                 create - f -
kind : Ingress
metadata :
  name : simple
  annotation s:
    nginx . ingress . kubernetes . io / rewrite - target : /
spec :
  rules :
   http :
      paths :
         path : / svc
         backend :
           serviceNam e :
                              http - svc
           servicePor t:
                              80
EOF
root @ master #
                    kubectl
                               get
                                      ing
                    HOSTS
                                      ADDRESS
                                                           PORTS
NAME
```

AGE

simple * 101.37.192.211 80 11s

Now visit http://101.37.192.211/svc to access the Nginx service.

Simple fanout routing based on domain names

If you have multiple domain names providing different external services, you can generate the following configuration to implement a simple fanout effect based on domain names:

```
create - f -
root @ master # cat << EOF | kubectl
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
         simple – fanout
  name :
spec :
  rules :
   host :
           foo . bar . com
    http
         •
      paths :
        path : / foo
        backend :
                          http - svcl
          serviceNam e :
          servicePor
                     t :
                           80
       path : / bar
        backend :
          serviceNam e : http - svc2
          servicePor
                     t :
                           80
    host :
           foo . example . com
    http :
     paths :
        path : / film
        backend :
          serviceNam e :
                           http - svc3
          servicePor t:
                           80
EOF
root @ master # kubectl
                                  ing
                            get
NAME
                  HOSTS
                                  ADDRESS
                                                     PORTS
                                                                 AGE
simple - fanout
                                  101 . 37 . 192 . 211
                                                           80
                   *
  11s
```

Then, you can access the http - svc1 service by using http :// foo . bar
. com / foo , access the http - svc2 service by using http :// foo . bar
. com / bar , and access the http - svc3 service by using http :// foo .
example . com / film .

Note:

 In a production environment, point the domain name to the preceding returned address 101 . 37 . 192 . 211 . • - In a testing environment, you can modify the hosts file to add a domain name mapping rule.

101 . 37 . 192 . 211 foo . bar . com 101 . 37 . 192 . 211 foo . example . com

Default domain name of simple routing

It does not matter if you do not have the domain name address. Container Service binds a default domain name for Ingress service. You can use this default domain name to access the services. The domain name is in the format of *.[cluster - id].[region - id]. alicontain er . com . You can obtain the address on the cluster Basic Information page in the console.

Use the following configuration to expose two services with the default domain name.

```
root @ master # cat << EOF | kubectl
                                            create - f -
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
  name : shared - dns
spec :
  rules :
 - host : foo .[ cluster - id ].[ region - id ]. alicontain er .
     ## Replace with the default service
                                                  access
                                                             domain
com
     of
name
           your
                   cluster .
    http :
     paths :
       path : /
        backend :
          serviceNam e : http - svc1
          servicePor t: 80
    host : bar .[ cluster - id ].[ region - id ]. alicontain er .
_
      ## Replace with the
                               default
                                        service
com
                                                   access
                                                             domain
      of
                  cluster .
name
           your
    http :
     paths :
        path : /
        backend :
          serviceNam e :
                          http - svc2
          servicePor t:
                           80
EOF
root @ master # kubectl
                                  ing
                            get
                                  ADDRESS
                                                     PORTS
                                                                 AGE
NAME
                  HOSTS
                 foo .[ cluster - id ].[ region - id ]. alicontain
shared - dns
er . com , bar .[ cluster - id ].[ region - id ]. alicontain er .
com 47 . 95 . 160 . 171 80 40m
```

Then, you can access the http - svc1 service by using http :// foo .[cluster - id].[region - id]. alicontain er . com /and access the http - svc2 service by using http :// bar .[cluster - id].[region - id]. alicontain er . com .

Configure a safe routing service

Management of multiple certificates is supported to provide security protection for your services.

1. Prepare your service certificate.

If no certificate is available, generate a test certificate in the following method:



The domain name must be consistent with your Ingress configuration.

```
root @ master # openssl req - x509 - nodes - days 365
- newkey rsa : 2048 - keyout tls . key - out tls . crt -
subj "/ CN = foo . bar . com / 0 = foo . bar . com "
```

The above command generates a certificate file tls . crt and a private key file

tls . key .

Create a Kubernetes secret named *foo*. *bar* using the certificate and private key. The secret must be referenced when you create the Ingress.

root @ master # kubectl create secret tls foo . bar -key tls . key -- cert tls . crt

2. Create a safe Ingress service.

```
root @ master # cat << EOF
                                  kubectl
                                            create - f -
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
  name : tls - fanout
spec :
 tls :
 - hosts :
   - foo . bar . com
    secretName : foo . bar
  rules :
    host :
          foo . bar . com
    http :
     paths :
       path : / foo
       backend :
                           http - svc1
          serviceNam e :
          servicePor t:
                           80
       path : / bar
       backend :
          serviceNam e :
                          http - svc2
          servicePor t:
                           80
EOF
root @ master # kubectl
                                 ing
                           get
                                  ADDRESS
                                                     PORTS
NAME
                 HOSTS
AGE
```

```
tls - fanout * 101 . 37 . 192 . 211 80
11s
```

3. Follow the notes in Simple fanout routing based on domain names to configure the hosts file or set the domain name to access the TLS service.

```
You can access the http - svc1 service by using http :// foo . bar . com / foo and access the http - svc2 service by using http :// foo . bar . com / bar .
```

You can also access the HTTPS service by using HTTP. By default, Ingress redirects HTTP access configured with HTTPS to the HTTPS address. Therefore, access to

http :// foo . bar . com / foo will be automatically redirected to https
:// foo . bar . com / foo .

Deploy Ingress in Kubernetes dashboard

1. 1. Save the following yml code to the nginx - ingress . yml file.

```
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
   name : simple
spec :
   rules :
    - http :
        paths :
        - path : / svc
        backend :
            serviceNam e : http - svc
            servicePor t : 80
```

2. Log on to the Container Service console. In the left-side navigation pane under

Kubernetes, click Clusters. Then click Dashboardon the right of the target cluster.

- Overview + CREATE < Cluster CPU usage Memory usage (i) Namespaces 0.011 3.35 Gi 2.98 Gi 2.30 Gi CPU (cores) 0.007 -+ G 1.49 Gi -+ G -+ G Persistent Volume 0.00 Roles 0.003 0 20:24 20:24 20:26 20:30 20:33 20:36 20:38 20:26 20:30 20:33 20:36 20:38 Storage Classe Time Namespace default Workloads Overview Workloads Statuses
- 3. Click CREATE in the upper-right corner to create an application.

4. Click the CREATE FROM FILE tab. Select the nginx - ingress . yml file you saved.

5. Click UPLOAD.

Then an Ingress Layer-7 proxy route will be created to the http - svc service.

6. Click default under Namespace in the left-side navigation pane. Click Ingresses in the left-side navigation pane.

You can view the created Ingress resource and its access address http:// 118 .

Cluster Namespaces	Ingresses			Ŧ
Nodes	Name 🗢	Endpoints	Age 🌲	
Persistent Volumes	simple	11. 11. 207.8 🖾	6 days	
Roles Storage Classes	simple-fanout	114 55 207.8 🖾	6 days	
	simple-routing	1 14 23.207.8 🖾	2 days	
Namespace				
Overview				
Workloads				
Daemon Sets				
Deployments				
Jobs				
Pods				
Replica Sets				
Replication Controllers				
Stateful Sets				
Discovery and Load Balanci Ingresses Services				
Config and Storage				

178 . 174 . 161 / svc .

7. Enter the address in the browser to access the created http - svc service.

1.10.4 Configure Ingress monitoring

You can view the Ingress monitoring data by enabling the default VTS module of Ingress.

Enable VTS module by running commands

1. Modify the Ingress ConfigMap configuration to add the configuration item enable

- vts - status : " true ".

```
root @ master # kubectl edit configmap nginx - configurat
ion - n kube - system
configmap " nginx - configurat ion " edited
```

After the modification, the contents of the Ingress ConfigMap are as follows:

```
apiVersion : v1
data :
    enable - vts - status : " true " # Enable VTS module
    proxy - body - size : 20m
kind : ConfigMap
metadata :
    Annotation s :
```

2. Verify if Ingress Nginx has enabled the VTS module normally.

```
get
root @ master # kubectl
                                   pods -- selector = app = ingress
– nginx – n
               kube - system
NAME READY
             STATUS RESTARTS
                                    AGE
nginx - ingress - controller - 79877595c8 - 78gq8
                                                     1 / 1
                                                              Running
 0
     1h
root @ master # kubectl exec - it
controller - 79877595c8 - 78gq8 - n
                                         nginx - ingress -
                                        kube - system -- cat
/ etc / nginx / nginx . conf
                              grep vhost_traf fic_status
_display
         vhost_traf
                     fic_status
                                  _display ;
         vhost_traf
                     fic_status
                                  display f
                                              ormat
                                                       html ;
```

3. Locally access the Ingress Nginx monitoring console.

```
Note:
```

By default, the VTS port is not opened for security considerations. Here use the port-forward method to access the console.

```
root @ master # kubectl port - forward nginx - ingress -
controller - 79877595c8 - 78gq8 - n kube - system 18080
Forwarding from 127 . 0 . 0 . 1 : 18080 -> 18080
```

Handling connection for	18080
-------------------------	-------

4. Use http://localhost: 18080 / nginx_stat us to access the VTS

monitoring console.

Nginx Vhost Traffic Status

Server main

Heat							lavala		lma		Co	nneo	ctions			Requests							:	Shared n	nemo	ory		
		по	sı				v	ersic	on Opt	ime	active	readi	ng ۱	writin	g wai	ting	accepte	dl	handled	Total	Req/s		name		maxS	zeι	usedSize	usedNode
nginx-	ingress-c	ontrol	ler-79	987759	95c8-7	78g	q8	1.13	3.7 32m	41s	7		0		1	6	9356	66	93566	1428	i 1	vhos	t_traffic	_statu	s 10.0 l	ИiВ	2.4 KiB	1
Ser	Server zones																											
7	Requ	uests			Re	espo	onse	s				Traf	fic								Cach	he						
Zone	Total Re	eq/s Ti	ime	1xx 2	xx 3	хx	4xx	5xx	Total	Ser	nt Re	cvd	Sent	t/s R	cvd/s	Miss	Bypas	s E	Expired	Stale	Updati	ng Re	evalidate	ed Hi	t Scarce	e To	tal	
_	660	1	0ms	0 6	660	0	0	0	660	1.7 N	AiB 145	4 KiB	1.1 1	КiВ	503 B		0	0	0	0		0		0	0	0	0	
*	660	1	0ms	0 6	660	0	0	0	660	1.7 N	AiB 145.	4 KiB	1.1 H	КiВ	503 B	1	0	0	0	0		0		0	0	0	0	
upst	Jpstreams Jpstream-default-backend																											
Se	erver	State	Res	ponse	e Time	e W	/eigł	nt Ma	axFails	Fai	ITimeo	ut To	tal F	Rea/s	Time	1xx	2xx 3	a	4xx 5xx	Tota	Sent	Rcvd	Sent/s	Reve	t/s			
172.16	.3.6:8080	up			0m	IS		1		0		0	0	0	Oms	s (0 0	0	0 0	0	0 0 B	0 E	3 O E	3	0 B			
updat <u>Json</u> I	Total Req/s Time 1xx 2xx 3xx 4xx 5xx Total Sent Rcvd/s 172.16.3.6:8080 up 0ms 1 0 <td></td>																											

Enable VTS module by using the Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. On the Cluster List page of Kubernetes clusters, click Dashboard at the right of a cluster to enter the Kubernetes dashboard page.
- 3. Select kube-system under Namespace in the left-side navigation pane. Click Config Maps in the left-side navigation pane. Click the icon at the right of nginxconfiguration and then select View/edit YAML. Edit the config map to add the configuration item enable - vts - status : " true ".

The contents of the saved Ingress ConfigMap are as follows:

```
{
    "kind ": " ConfigMap ",
    "apiVersion ": " v1 ",
    "metadata ": {
        "name ": "nginx - configurat ion ",
        "namespace ": "kube - system ",
        "selfLink ": "/api / v1 / namespaces / kube - system /
        configmaps / nginx - configurat ion ",
        "creationTi mestamp ": "2018 - 03 - 20T07 : 10 : 18Z ",
        "labels ": {
        "app ": "ingress - nginx "
        },
        "annotation s ": {
        "kubectl . kubernetes . io / last - applied - configurat
        ion ": "{\ "apiVersion \":\" v1 \",\" data \":{\" proxy - body -
        size \":\" 20m \"},\" kind \":\" ConfigMap \",\" metadata \":{\"
        annotation s \":{}," labels \":{\" app \":\" ingress - nginx \"},
        \" name \":\" nginx - configurat ion \",\" namespace \":\" kube -
        system \"}\" n "
```

```
}
}
,
" data ": {
    " proxy - body - size ": " 20m ",
    " enable - vts - status ": " true "
}
```

4. Locally access the Ingress Nginx monitoring console.

Note:

By default, the VTS port is not opened for security considerations. Here use the port-forward method to access the console.

```
root @ master # kubectl port - forward nginx - ingress -
controller - 79877595c8 - 78gq8 - n kube - system 18080
Forwarding from 127 . 0 . 0 . 1 : 18080 -> 18080
Handling connection for 18080
```

5. Use http://localhost: 18080 / nginx_stat us to access the VTS

monitoring console.

Nginx Vhost Traffic Status

Server main

11												Con	nection	S		Requests						Shared memory						
		r	1051				ve	ersior		me a	ctive	readin	g writi	ng wa	iting	accepte	d han	dled	Total	Req/s		name		maxSiz	e used	Size u	usedNode	
nginx-	nginx-ingress-controller-79877595c8-78gq8					q8	1.13.	7 32m	41s	7		0	1	6	9356	6 9	93566	1428	1	vhost	_traffic_	status	10.0 M	B 2.4	KiB	1		
Server zones																												
7000	Req	ueste	3		F	Resp	onses	5				Traffi	;							Cach	е							
Zone	Total Re	eq/s	Time	1xx	2xx	3xx	4xx	5xx	Total	Sent	Rc	vd S	ent/s I	Rcvd/s	Miss	Bypas	s Exp	ired \$	Stale	Updatir	ng Re	evalidate	d Hit	Scarce	Total			
_	660	1	0ms	0	660	0	0	0	660 1	.7 Mi	B 145.4	4 KiB 1	.1 KiB	503 E		D	0	0	0		0		0 0) 0	0			
*	660	1	0ms	0	660	0	0	0	660 1	.7 Mi	B 145.4	4 KiB 1	.1 KiB	503 E		D	0	0	0		0		0 () 0	0			
Ups upst	Jpstreams upstream-default-backend																											
S	erver	Stat	e Re	spon	se Tin	ne V	Veight	t Ma	xFails	Fail	Timeou	.t	Reque	sts		Responses Traffic					-							
170.10	0.0000				0							IOta	Heq/	s IIme	e 1xx	2XX 33	x 4xx	5xx	Iotal	Sent	HCVO	Sent/s	нсча	/s				
172.10	.3.0.0000		μ		0	115			0			0	0	0 011	5 (, ,	0 1	0 0		00	0 6	0 0 0	, c	Б				
upda <u>Json</u> I	ו72.16.3.6:8080 up 0ms 1 0 0 0 0ms 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0																											

1.10.5 Ingress configurations

Alibaba Cloud Container Service provides the highly reliable Ingress controller components and integrates with Alibaba Cloud Server Load Balancer to provide the flexible and reliable Ingress service for your Kubernetes clusters.

See the following Ingress orchestration example. You must configure the annotation when creating an Ingress in the Container Service console. Some configurations must create dependencies. For more information, see *Create an Ingress in the Container Service console, Support for Ingress, and Kubernetes Ingress.* Ingress also supports the

configuration of configmap. For more information, see https://kubernetes.github.io/

ingress-nginx/user-guide/nginx-configuration/configmap/.

```
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
  annotation
               s :
     nginx . ingress . kubernetes . io / service - match : ' new -
nginx : header (" foo ", /^ bar $/)'  # Gray release
rule . Header is used in this example .
    nginx . ingress . kubernetes . io / service - weight : ' new -
nginx : 50 , old - nginx : 50 ' # Traffic weight
                                                                       weight
annotation
  creationTi mestamp : null
  generation : 1
  name : nginx - ingress
  selfLink : / apis / extensions / v1beta1 / namespaces / default /
ingresses / nginx - ingress
spec :
  rules :
                                                                            ##
           rule
Ingress
    host : foo . bar . com
    http :
       paths :
         backend :
            serviceNam e : new - nginx
            servicePor t:
                               80
         path : /
         backend :
           serviceNam e : old - nginx
            servicePor t: 80
         path : /
tls :
                                                                         ##
Enable
         TLS
               to
                       set a
                                   secure
                                              Ingress .
   hosts :
   - *. xxxxxx . cn - hangzhou . alicontain er . com
   - foo . bar . com
    secretName : nginx - ingress - secret
## Secret
              name
status :
  loadBalanc er : {}
```

Annotation

You can configure an ingress annotation, specifying the ingress controller to use, rules for routing, such as routing weight rules, grayscale publish, and rewrite rules. For more information, see https://kubernetes.github.io/ingress-nginx/user-guide/nginx-configuration/annotations/.

For example, a typical rewrite annotation nginx . ingress . kubernetes . io / rewrite - target : / redirects the path / path to the path / that can be recognized by the backend services.

Rules

The rules indicate those that authorize the inbound access to the cluster and are generally the HTTP rules, including the domain name (virtual hostname), URL access path, service name, and port.

You must complete the following configurations for each HTTP rule:

- Host: Enter the testing domain name of an Alibaba Cloud Kubernetes cluster or a virtual hostname, such as foo . bar . com .
- Path: Specify the URL path of the service access. Each path is associated with a backend service. Before Alibaba Cloud Server Load Balancer forwards the traffic to the backend, all inbound requests must match with the domain name and path.
- Backend configuration: Service configuration that is a combination of service : port and traffic weight. The Ingress traffic is forwarded to the matched backend services based on the traffic weight.
 - Name: The name of the backend service forwarded by Ingress.
 - Port: The port exposed by the service.
 - Weight: The weight rate of each service in a service group.

Note:

- 1. The service weight is calculated in relative values. For example, if both service weights are set to 50, the weight ratio of both services is 50%.
- 2. A service group (a service with the same Host and Path in the same ingress yaml) has a default weight value of 100 and the weight is not explicitly set.

Grayscale publish

Container Service supports different traffic segmentation methods for grayscale publish and AB test scenarios.

Note:

Currently, the Alibaba Cloud Container Service Kubernetes Ingress Controller requires 0 . 12 . 0 – 5 and above to support the traffic segmentation feature.

- 1. Traffic segmentation based on the request header.
- 2. Traffic segmentation based on cookie.
- 3. Traffic segmentation based on query (request) parameters.

After the grayscale rule is configured, the request that matches the grayscale publish rule can be routed to the set service. If the service sets a weight rate of less than 100%, requests that match the grayscale publish rule continue to be routed to the corresponding service based on the weight rate.

TLS

You can encrypt the Ingress by specifying a secret that contains the TLS private key and certificate to implement the secure Ingress access. The TLS secret must contain the certificate named tls.crt and private key named tls.key. For more information about the TLS principles, see *TLS*. For how to create a secret, see *Configure a safe routing service*.

Label

You can add tags for Ingress to indicate the characteristics of the Ingress.

1.10.6 Create an Ingress in the Container Service console

Alibaba Cloud Container Service console integrates with the Ingress service, which allows you to quickly create an Ingress service in the Container Service console to build the flexible and reliable traffic access layer.

Prerequisites

- You have successfully created a Kubernetes cluster and Ingress controller is running normally in the cluster. For how to create a Kubernetes cluster, see *Create a Kubernetes cluster*.
- Log on to the master node by using SSH. For more information, see Access *Kubernetes clusters by using SSH*.

Step 1. Create a deployment and a service

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Applications > Deployments.

3. Click Create by template in the upper-right corner.

Container Service - Kubernetes 👻		Deployr	ment									Refresh	Create	by Image	Create by Te	mplate
Overview		8 How t	o use private imag	es 🔗 Cr	eate application	ns 🔗 Sched	lule a pod	to the specified node	🔗 Create a Layer-	-4 Ingress 👌	Create a Layer-7 1	ingress 🔗 Co	onfigure poo	d auto scaling		
▼ Clusters	L	🖉 Conta	iner monitoring (Blue-gro	een release											
Clusters	L	Clusters	k8s-test	•	Namespaces	default	•							Search By Nan	ne	Q
Nodes	L	Name		Tag	Po	odsQuantity			Image		Time Created					Action
Volumes							(1) Could not find	any record that	t met the	condition.					
Namespaces																
Authorization) Conta
 Applications 																ctUs
Deployments																
Stateful Sets																
Daemon Sets																

4. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

In this example, three nginx applications are created. One for the old application (old-nginx), one for the new (new-nginx), and an application for testing the cluster access domain name (domain-nginx).



The orchestration template for old-nginx is as follows:

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
   name : old - nginx
spec :
   replicas : 2
   selector :
      matchLabel s :
```
```
run : old - nginx
  template :
   metadata :
     labels :
       run : old - nginx
   spec :
     containers :
       image : registry . cn - hangzhou . aliyuncs . com / xianlu
/ old - nginx
       imagePullP olicy : Always
       name : old - nginx
       ports :
       - containerP ort : 80
         protocol : TCP
      restartPol icy : Always
apiVersion : v1
kind : Service
metadata :
 name : old - nginx
spec :
 ports :
 - port : 80
   protocol : TCP
   targetPort : 80
  selector :
   run : old - nginx
  sessionAff inity : None
  type : NodePort
```

The orchestration template for new-nginx is as follows:

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
 name : new - nginx
spec :
  replicas : 1
  selector :
   matchLabel s :
     run : new - nginx
  template :
    metadata :
      labels :
        run : new - nginx
    spec :
     containers :
     - image : registry . cn - hangzhou . aliyuncs . com / xianlu
/ new - nginx
       imagePullP olicy :
                            Always
       name : new - nginx
       ports :
         containerP ort :
                            80
         protocol : TCP
      restartPol icy : Always
apiVersion : v1
kind : Service
metadata :
 name : new - nginx
spec :
 ports :
 - port : 80
```

```
protocol : TCP
targetPort : 80
selector :
run : new - nginx
sessionAff inity : None
type : NodePort
```

The orchestration template for domain-nginx is as follows:

```
apiVersion : apps / v1beta2 #
                                For
                                      versions
                                                 before
                                                          1.8.
0 use apps / v1beta1
kind : Deployment
metadata :
 name : domain - nginx
 labels :
   app : nginx
spec :
  replicas : 2
  selector :
   matchLabel s :
     app : nginx
  template :
   metadata :
     labels :
       app : nginx
   spec :
     containers :
     - name : nginx
image : nginx : 1 . 7 . 9 # replace it with
                                                            your
exactly < image_name : tags >
       ports :
       - containerP ort : 80
apiVersion : v1
kind : Service
metadata :
 name : domain - nginx
spec :
 ports :
   port: 80
   protocol : TCP
   targetPort : 80
  selector :
   app : nginx
  sessionAff inity : None
```

type : NodePort

5. In the left-side navigation pane under Container Service-Kubernetes, choose Discovery and Load Balancing > Services.

After the service is created, you can see it on the Service List page.

Service List							Refresh	Create
Clusters k8s-clus	ster 🔻 Na	mespace default 🔻						
Name	Туре	Time Created	ClustersIP	internalendpoint	externalendpoint			Action
domain-nginx	NodePort	07/11/2018,17:43:32	10.004	domain-nginx:80 TCP domain-nginx:32347 TCP	-	Details Update	View YAML	Delete
kubernetes	ClusterIP	07/11/2018,17:35:35	1000	kubernetes:443 TCP	-	Details Update	View YAML	Delete
new-nginx	NodePort	07/11/2018,17:37:01	1000	new-nginx:80 TCP new-nginx:32637 TCP	-	Details Update	View YAML	Delete
old-nginx	NodePort	07/11/2018,17:37:01	1000	old-nginx:80 TCP old-nginx:32039 TCP	-	Details Update	View YAML	Delete

Step 2. Create an Ingress

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Discovery and Load Balancing > Ingresses.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Then click Create in the upper-right corner.

Container Service - Kubernetes 👻	Ingress	Ingress Overview Refresh	Create
Overview	🔗 Ingress log analysis and monitoring 🔗 Blue-green release		4
Clusters	Clusters kas-test v Namespaces default v	Search By Name	Q
Applications	Name Endpoint Rule Time Created		Action
 Service Mesh 	① Could not find any record that met the condition.		
Virtual Services			
Discovery and Load B			
Services) Cant
Ingresses 2			act Us
 Configuration 			

4. In the displayed dialog box, enter the Ingress name. In this example, enter nginxingress.

Name:	nginx-ingress

5. Configure the rules.

The Ingress rules are the rules that authorize the inbound access to the cluster and are generally the HTTP rules. Configure the domain name (virtual hostname), URL path, service name, and port. For more information, see *Ingress configurations*.

In this example, add a complicated Ingress rule. Configure the default test domain name and virtual hostname of the cluster to display the Ingress service based on the domain names.

Rule:	O Add	
	Domain	۲
	test.classicontainer.com	
	Select *	
	path	
	/	
	Service 📀 Add	
	Name Port Weight Percent of Weight	
	domain-nginx • 80 • 100 100.0%	•
	Domain	۲
	foo.bar.com	
	Select *.c	
	path	
	/	
	Service 📀 Add	
	Name Port Weight Percent of Weight	
	new-nginx • 80 • 50 50.0%	•
	old-nginx • 80 • 50 50.0%	•

- The simple Ingress based on the default domain name, that is, provide the access service externally by using the default domain name of the cluster.
 - Domain: Enter the default domain name of the cluster. In this example, use test .[cluster id].[region id]. alicontain er . com .

The default domain name of this cluster is displayed in the Create dialog box, in the *. [cluster - id]. [region - id]. alicontain er . com format. You can also obtain the default domain name on the Basic Information page of the cluster.

- Service: Configure the access path, name, and port of the service.
 - Path: Specify the URL path of the service access. The default is the root path /, which is not configured in this example. Each path is associated with a backend service. Before Alibaba Cloud Server Load Balancer forwards the traffic to the backend, all inbound requests must match with the domain name and path.
 - Service configuration: The backend configuration, which is a combinatio n of service name, port, and service weight. The configuration of multiple services in the same access path is supported, and Ingress traffic is split and is forwarded to the matched backend services.
- The simple fanout Ingress based on the domain name. In this example, use a virtual hostname as the testing domain name to provide the access service externally. You can use the recorded domain name in the production environment to provide the access service. You can use the recorded domain name in the production environment to provide the access service.
 - Domain: In this example, use the testing domain name foo . bar . com .

You must modify the hosts file to add a domain name mapping rule.

118 . 178 . 108 . 143 foo . bar . com # Ingress IP address

- Service: Configure the access path, name, and port of the service.
 - Path: Specify the URL path of the service access. Path is not configured in this example, and the root path is /.
 - Name: In this example, set up both new and old services, nginx-new and nginx-old.
 - Port: Expose 80 port.
 - Weight settings: Set the weight of multiple services under this path. The service weight is calculated by relative value. The default value is 100. As shown in this example, the service weight values of both the old and new versions are 50, which means that the weight rate of both services is 50%.

6. Grayscale publish configuration.

Note:

Currently, the Alibaba Cloud Container Service Kubernetes Ingress Controller requires 0 . 12 . 0 - 5 and above to support the traffic segmentation feature.

Container Service supports different traffic segmentation methods for grayscale publish and AB test scenarios.

- a. Traffic segmentation based on the request header.
- b. Traffic segmentation based on cookie.
- c. Traffic segmentation based on query (request) parameters.

After the grayscale rule is configured, the request that matches the grayscale publish rule can be routed to the new service version new-nginx. If the service sets a weight rate of less than 100%, requests that match the grayscale publish rule continue to be routed to the corresponding service based on the weight rate.

In this case, set the request header to meet a grayscale publish rule of foo =^ bar

\$, only requests with the request header can access the new-nginx service.

Grayscale release:	Add After the gray set a weight other that and old version service	rule is set, the request n 100, the request to sa according to the weig	t meeting the rule atisfy the gamma ahts.	e will rule
	Service	Туре	Name	I
	new-nginx	▼ Header ▼	foo	

- · Service: Routing rule configuration service.
- Type: matching request header, cookie, and query (request) parameters are supported.
- Name and match value: User-defined request field, name and match value are key-value pairs.
- Match rules: Regular and exact matches are supported.

7. Configure the annotations.

Click rewrite annotation, a typical redirection annotation can be added to the route. nginx . ingress . kubernetes . io / rewrite - target : / indicates that the / *path* is redirected to the root path / that the backend service can recognize.



In this example, the access path is not configured, so no need to configure rewrite annotations. The purpose of the rewrite annotation is to enable Ingress to forward to the backend as the root path, avoiding 404 errors caused by incorrect access path configuration.

You can also click Add to enter the annotation name and value, which is the annotation key-value pair for Ingress. For more information, see *https://kubernetes.github.io/ingress-nginx/user-guide/nginx-configuration/annotations/*.

annotation:	• Add rewrite annotation		
	Name	Value	
	nginx.ingress.kubernetes.io/rewri	/	•

- 8. Configure TLS. Select Enable and configure the secure Ingress service. For more information, see *Configure a safe routing service*.
 - You can select to use an existing secret.

a. Log on to the master node and create tls . key and tls . crt .

```
openssl req - x509 - nodes - days 365 - newkey rsa:
2048 - keyout tls.key - out tls.crt - subj "/ CN =
foo .bar .com / 0 = foo .bar .com "
```

b. Create a secret.

```
kubectl create secret tls foo . bar -- key tls .
key -- cert tls . crt
```

- c. Run the kubectl get secret command to see that secret has been successfully created. You can use the secret that you have created in the Web interface, foo . bar .
- You can create the secret with one click by using the created TLS private key and certificate.



a. Log on to the master node and create tls . key and tls . crt .

```
openssl req - x509 - nodes - days 365 - newkey rsa:
2048 - keyout tls.key - out tls.crt - subj "/ CN =
foo.bar.com / 0 = foo.bar.com "
```

- b. Run the vim tls . key and vim tls . crt to get the generated private key and certificate.
- c. Copy the generated certificate and private key to the Cert and Key fields.

9. Adding the tags.

Add the corresponding tags for Ingress to indicate the characteristics of the Ingress.

Tag:	O Add		
	Name	Value	
	node-role.kubernetes.io/ingress	true	•

10.Click Create.

The Ingress nginx-ingress is displayed on the Ingress page.

Ingress	Refresh	Create
Clusters k8s-cluster v Namespace default v		
Name endpoint Rule	Time Created	Action
nginx-ingress foo.bar.com/ -> new-nginx foo.bar.com/ -> new-nginx foo.bar.com/ -> old-nginx	07/11/2018,17:49:46 Details Update View YAML	Delete

11.Click on the access domain name test . [cluster - id]. [region - id

]. alicontain er . com in the route, and foo . bar . com to access the

welcome page of nginx.

() test.c0, ca, accessed accessed d5.cn-hangzhou.alicontainer.com/?spm=5176.2020520152.0.0.509c16dd0LCNzI						
	Welcome to nginx!					
	If you see this page, the nginx web server is successfully installed and working. Further configuration is required.					
	For online documentation and support please refer to <u>nginx.org</u> . Commercial support is available at <u>nginx.com</u> .					
	Thank you for using nginx.					

Click on the route address pointing the new-nginx service and find the page that points the old-nginx application.



Access the route address in the browser. By default, the request header does not have the $foo =^{bar}$ s, so the traffic is directed to the old-nginx application.

```
← → C ① foo.bar.com/?spm=5176.2020520152.0.0.509c61b1iW1N16
old
```

12.Log on to the master node by using SSH. Run the following command to simulate the access result with a specific request header.

```
curl - H " Host : foo . bar . com " http :// 47 . 107 . 20 .
35
old
 curl - H " Host : foo . bar . com " http :// 47 . 107 . 20 .
35
old
       - H " Host :
 curl
                       foo . bar . com "
                                           http://47.107.20.
35 #
      Similar
                 to
                       browser
                                  access
                                           requests
old
curl - H " Host : foo . bar . com " - H " foo :
:// 47 . 107 . 20 . 35 # Simulate an access re
a unique header , returning results based
                                                         bar "
                                                                 http
                                                       request
                                                                  with
               header , returning
                                                          on
                                                                routing
  weight
new
 curl - H "Host: foo.bar.com"-H "foo: bar"
                                                                 http
:// 47 . 107 . 20 . 35
old
 curl - H " Host : foo . bar . com " - H " foo :
                                                         bar "
                                                                 http
:// 47 . 107 . 20 . 35
old
 curl - H "Host: foo.bar.com"-H "foo: bar"
                                                                 http
:// 47 . 107 . 20 . 35
new
```

1.10.7 Update an Ingress

Prerequisites

- You have successfully created a Kubernetes cluster and Ingress controller is running normally in the cluster. For how to create a Kubernetes cluster, see *Create a Kubernetes cluster*.
- You have successfully created an Ingress. For more information, see *Create an Ingress in the Container Service console*.

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Ingress in the left-side navigation pane.

3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Update at the right of the Ingress.

Container Service	Ingress		Create	Refresh
Overview	Clusters test v Namespace default v 4			
 Clusters 	Name endpoint Rule	Time Created	5	Action
Clusters Nodes	nginx-ingress 40.82.217.40 foo.bar.com/svc2 -> http-svc2 foo.bar.com/svc3 -> http-svc3	05/02/2018,18:41:55	Details Update	Delete
Storage Application Deployment Pods Service				
Release Config Maps				

4. Update the Ingress parameters in the displayed dialog box and then click OK.

change test .[cluster - id].[region - id]. alicontain er . com to testv2 .[cluster - id].[region - id]. alicontain er . com 。

Update		\times
Template:	<pre>1 spec: 2 tls: 3 - hosts: 4 testv2 contained and a secret and a secretName: nginx-ingress - secret 7 rules: 8 - host: testv2 contained and a secret and a se</pre>	
	OK Cance	el

What's next

On the Ingress page, you can see a rule of this Ingress is changed.

Ingress				Create	Refresh
Clusters test •	Namespace	default •			
Name	endpoint	Rule	Time Created		Action
nginx-ingress	0.07.237.40	testv2> http-svc1 foo.bar.com/svc2 -> http-svc2 foo.bar.com/svc3 -> http-svc3	05/02/2018,18:41:55	Details Update	e Delete

1.10.8 View Ingress details

Prerequisites

- You have successfully created a Kubernetes cluster and Ingress controller is running normally in the cluster. For how to create a Kubernetes cluster, see *Create a Kubernetes cluster*.
- You have successfully created an Ingress. For more information, see *Create an Ingress in the Container Service console*.

Procedure

- 1. Log on to the Container Service console.
- 2. Click Kubernetes Application > Ingress in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Details at the right of the Ingress.

Container Service	Ingress		Create	Refresh
Overview 1	Clusters test Namespace default			
 Clusters 	Name endpoint Rule	Time Created	5	Action
Clusters Nodes	nginx-ingress 40 10 2010 foo.bar.com/svc2 -> http-svc2 foo.bar.com/svc3 -> http-svc3	05/02/2018,18:41:55	Details Update	Delete
Storage Application 2 Deployment Pods = Service Ingress 3 Release Config Maps				

On the details page, you can view the overview and rules of the Ingress.

nginx-ingress 🔹 t Back to List				
Overview				
Name:	nginx-ingress			
Namespace:	default			
Time Created:	2018-05-02T10:41:55Z			
Label:	node-role.kubernetes.io/ingress:true			
annotation:	nginx.ingress.kubernetes.io/rewrite-targetc/			
endpoint:	47.07.122.40			
Rule				
Domain		path	Name	service port
test. Codic Moch Microsoft	is the state of th	/svc1	http-svc1	80
foo.bar.com			http-svc2	80
foo.bar.com			http-svc3	80

1.10.9 Deleting a route

Prerequisites

- You have successfully created a Kubernetes cluster and Ingress controller is running normally in the cluster. For how to create a Kubernetes cluster, see *Create a Kubernetes cluster*.
- You have successfully created an Ingress. For more information, see *Create an Ingress in the Container Service console*.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Ingress in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Delete at the right of the Ingress.

Container Service	Ingress		Create Refresh
Overview 1 m	Clusters test • Namespace default •		
 Clusters 	Name endpoint Rule	Time Created	Action
Clusters Nodes	nginx-ingress 47 00 2017 test 10 000 1000 10000 1000 1000 1000 1000 1000 1000 10000 1000 1000 10	05/02/2018,18:41:55	Details Update Delete
Storage			5
Application 2 Deployment			
Pods 🔤			
Service Ingress Release			

4. Click Confirm in the displayed dialog box.

Confirm			×
Deletenginx-ingr	ess ?		
		Confirm	Cancel

1.11 Config map and Secret management

1.11.1 Create a Config Map

In the Container Service console, you can create a Config Map on the Config Maps page or by using a template.

Create a Config Map on Config Maps page

1. Log on to the Container Service console.

- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Configuration > Config Maps.
- 3. Select the target cluster and namespace. Then, in the upper-right corner, click Create.

Container Service - Kubernetes 🕶	Config Maps	Refresh
Overview	${\mathscr S}$ Use a config map in a pod	3
Clusters	Clusters k8s-test v Namespaces default v 2	
 Applications 	Config Map Name Namespaces Time Created	Operation
Service Mesh		
 Discovery and Load B 		
Services 🔄		_
Ingresses		
▼ Configuration		act Us
Config Maps		
Secrets		

- 4. Complete the settings and then click OK.
 - Namespace: Select the namespace to which the Config Map belongs. Config Map is a Kubernetes resource object that must be applied to the namespace.
 - Config Map Name: Enter the Config Map name, which can contain lowercase letters, numbers, hyphens (-), and periods (.). The name cannot be empty.

Other resource objects must reference the Config Map name to obtain the configuration information.

• Configuration: Enter the Variable Name and the Variable Value. Then, click Add on the right. You can also click Edit, complete the configuration in the displayed dialog box, and click OK.

* Namespace:	default		•
* Config Map Name:	test-config Name must consist of empty.	lowercase alphanumeric characters, '-' or '.'. Nam	ie cannot be
Configuration:	Variable Name	Variable Value	Action
	enemies	aliens	Edit Delete
	lives	3	Edit Delete
	Name	Value	Add
	Variable key must be	unique. Variable key and value cannot be empty.	
	Edit yaml the	OK Cancel	

In this example, configure the variables enemies and lives to pass the parameters aliens and 3 respectively.

YAML format
<pre>1 data: 2 enemies: aliens 3 lives; '3' 4 metadata: 5 name: test-config 6 namespace: default 7 * Configuration must be in YAML format.</pre>
OK Cancel

5. You can view the Config Map test-config on the Config Maps page after clicking OK.

Config Maps			Refresh	Create
Cluster: test 🔻				
Config Map Name	Namespace	Time Created		Operation
test	default	2018-02-09 03:30:31	Delete	Modify
test-config	default	2018-02-09 05:56:47	Delete	Modify

Create a Config Map by using a template

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Applications > Deployments.

3. Click Create by template in the upper-right corner.

Container Service - Kubernetes 👻		Deploy	ment				Refresh	Create by Image Create	e by Template
Overview	l	S How	& How to use private images & Create applications & Schedule a pod to the specified node & Create a Layer-4 Ingress & Create a Layer-7 Ingress & Configure pod auto scaling						
▼ Clusters		🔗 Conta	ainer monitoring 🔗 Blue-gro	en release					
Clusters		Clusters	k8s-test v	Namespaces default	*			Search By Name	Q
Nodes		Name	Tag	PodsQuantity		Image	Time Created		Action
Volumes					(i) Could r	ot find any record that met t	he condition.		
Namespaces									_
Authorization									U Conta
 Applications 									ict Us
Deployments 1									
Stateful Sets									
Daemon Sets									

- 4. On the Deploy templates page, complete the settings and then click DEPLOY.
 - Clusters: Select the cluster in which the Config Map is to be created.
 - Namespace: Select the namespace to which the Config Map belongs. Config map is a Kubernetes resource object that must be applied to the namespace.
 - Resource Type: You can write your own Config Map based on the Kubernetes YAML syntax rules, or select the sample template resource-ConfigMap. In the sample template, the Config Map is named as aliyun-config and includes two variable files game . properties and ui . properties . You can make modifications based on the sample template. Then, click DEPLOY.

Deploy temp	olates						
Only Kuberr	Only Kubernetes versions 1.8.4 and above are supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list						
	Clusters	test v					
	Namespace	default 🔹					
	Resource Type	Resource - ConfigMap 🔹					
	Template	<pre>1 # kubectl create configmap game-configfrom-file=docs/user-guide/configmap/kubectl 2 apiVersion: v1 3 kind: ConfigMap 4 metadata: 7 name: aliyun-config 6 data: 7 # replace your filename of properties configration and contents here. 8 game.properties: 9 enemies=aliens 10 lives=3 11 enemies.cheat=true 12 enemies.cheat.level=noGoodRotten 13 secret.code.allowed=true 15 secret.code.allowed=true 15 secret.code.allowed=true 16 color.good=purple 17 color.good=purple 18 color.bad=yellow 19 allow.textmode=true 20 how.nice.to.look=fairlyNice 21 special.env: env_value 22 DEFLOY </pre>					

5. After the deployment, you can view the Config Map aliyun-config on the Config Maps page.

Config Maps			Refresh Create
Cluster: test 🔻			
Config Map Name	Namespace	Time Created	Operation
aliyun-config	default	04/24/2018,15:41:32	Delete Modify

1.11.2 Use a config map in a pod

You can use a config map in a pod in the following scenarios:

- Use a config map to define the pod environment variables.
- Use a config map to configure command line parameters.
- Use a config map in data volumes.

For more information, see Configure a pod to use a ConfigMap.

Limits

To use a config map in a pod, make sure the config map and the pod are in the same cluster and namespace.

Create a config map

In this example, create a config map special-config, which includes two key-value pairs: SPECIAL_LE VEL : very and SPECIAL_TY PE : charm .

Create a config map by using an orchestration template

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment. Click Create by template in the upper-right corner.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

You can use the following YAML sample template to create a config map.

```
apiVersion : v1
kind : ConfigMap
metadata :
    name : special - config
    namespace : default
data :
    SPECIAL_LE VEL : very
```

SPECIAL_TY PE : charm

Create a config map on Config Maps page

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, choose Configuration > Config Maps in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Create in the upper-right corner.
- 4. Enter a config map name, click the plus icon , set the name and value for each

Config Map Clusters ch3 Namespace default special-config Config Map Name: Name must consist of lowercase alphanumeric characters, '-' or '.'. Name cannot be empty Configuration: Value Name SPECIAL_LEVEL Verv SPECIAL_TYPE charm Names can only contain numbers, letters, "_", "-" and "." Choose a file Cancel

entry of the config map, and then click OK.

Use a config map to define pod environment variables

Use config map data to define pod environment variables

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, clickApplication > Deployment . Click Create by template in the upper-right corner.

3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

You can define the environment variables in a pod. Use valueFrom to reference the value of SPECIAL_LEVEL to define the pod environment variables.

See the following orchestration example:

```
apiVersion : v1
kind : Pod
metadata :
  name : config - pod - 1
spec :
  containers :
      name : test - container
      image : busybox
      command : [ "/ bin / sh ", "- c ", " env " ]
      env :
          name : SPECIAL_LE VEL_KEY
          valueFrom :
                                                 ## Use
valueFrom
          to
                specify
                          env
                               to
                                  reference
                                                the value
                                                              of
  the config
                map .
            configMapK eyRef :
                                                     ## The
              name : special - config
referenced
            config map name
                    SPECIAL_LE VEL
                                                     ## The
            key :
referenced config
                     map
                           key .
  restartPol icy :
                     Never
```

Similarly, to define the values of multiple config maps to the environment variable values of the pod, add multiple env parameters in the pod.

Configure all key-value pairs of a config map to pod environment variables

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment. Click Create by template in the upper-right corner.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

To configure all the key-value pairs of a config map to the environment variables of a pod, use the envFrom parameter. The key in a config map becomes the environment variable name in the pod.

See the following orchestration example:

```
apiVersion : v1
kind : Pod
```

```
metadata :
  name : config - pod - 2
spec :
  containers :
      name : test - container
      image : busybox
      command : [ "/ bin / sh ", "- c ", " env " ]
      envFrom :
                             ## Reference all
                                                  the
                                                        kev -
                   the
value pairs in
                         config map special - config.
        configMapR ef :
          name : special - config
  restartPol icy : Never
```

Use a config map to configure command line parameters

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment . Click Create by template in the upper-right corner.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

You can use the config map to configure the commands or parameter values in the container by using the environment variable replacement syntax \$(VAR_NAME).

See the following orchestration example:

```
apiVersion :
             v1
kind : Pod
metadata :
  name : config - pod - 3
spec :
   containers :
      name : test - container
      image : busybox
      command : [ "/ bin / sh ", "- c ", " echo $( SPECIAL_LE
VEL_KEY ) $( SPECIAL_TY PE_KEY )" ]
       env :
          name : SPECIAL_LE VEL_KEY
          valueFrom :
            configMapK eyRef :
              name : special - config
              key : SPECIAL_LE VEL
          name : SPECIAL_TY PE_KEY
          valueFrom :
            configMapK eyRef :
              name : special - config
              key : SPECIAL_TY PE
```

restartPol icy : Never

The output after running the pod is as follows:

very charm

Use a config map in data volumes

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Application Deployment in the left-side navigation pane. Click Create by template in the upper-right corner.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Select a sample template or Custom from the Resource Type drop-down list. Click DEPLOY.

You can also use a config map in data volumes. Specifying the config map name under volumes stores the key-value pair data to the mountPath directory (/ etc / config in this example). It finally generates a configuration file with key as the file name and values as the contents of the file.

Then, the configuration file with key as the name and value as the contents is generated.

```
apiVersion :
               v1
kind : Pod
metadata :
   name : config - pod - 4
spec :
   containers :
    - name : test - container
       image : busybox
       command : [ "/ bin / sh ", "- c ", " ls / etc / config /" ]
st the file names under this directory .
  ## List the
       volumeMoun ts :
      - name : config - volume
         mountPath : / etc / config
   volumes :
    - name : config - volume
       configMap :
         name : special - config
   restartPol icy : Never
```

Keys of the config map are output after running the pod.

SPECIAL_TY PE

SPECIAL_LE VEL

1.11.3 View a ConfigMap

This topic describes how to view a created ConfigMap by using the Container Service console of Alibaba Cloud.

Prerequisites

· A Kubernetes cluster is created. For more information, see Create a Kubernetes cluster.

· A ConfigMap is created. For more information, see Create a ConfigMap.

Procedure

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Configuration > Config Maps.
- 3. Select the target cluster and namespace, find the target ConfigMap, and then click Details in the Operation column.

Container Service - Kubernetes +	Config Maps	Config Maps				
Service Mesh	Help: 🖉 Use a config map in a pod					
Virtual Services	Clusters k8s-cluster •	Namespaces default 🔻 3				
 Discovery and Load B 	Config Map Name	Namespaces	Time Created	4 Operation		
Services	test-config	default	02/06/2019,15:46:37	Details Edit Delete		
Config Maps 2						
Secrets						
 Store 				03		
Docker Images						

Then, you can view the details of the ConfigMap.

Config Maps - test-config			Refresh	Edit
Overview				
Name : test-config		Namespaces : default		
Time Created : 02/06/2019,15:46:37	7			
Tag :				
Annotation :				
Details				
partner	A			

1.11.4 Update a config map

You can modify the configurations of a config map.



Updating a config map affects applications that use this config map.

Update a config map on Config Maps page

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Config Maps in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Modify at the right of the config map.

Container Service	Config Maps			Refresh Create
Clusters	Clusters test v Namespace default v	4		
Clusters	Config Map Name	Namespace	Time Created	Operatio
Nodes	aliyun-config	default	04/24/2018,15:41:32	Delete 5
Storage	test	default	04/24/2018,17:01:14	Delete Modify
Application Deployment	test-mariadb	default	04/25/2018,20:50:20	Delete Modify
Pods				
Service				
Ingress				
Release Config Maps 3				

4. Click Confirm in the displayed dialog box.



5. Modify the configurations.

- Click Edit on the right of the configuration you want to modify. Update the configuration and then click Save.
- You can also click Edit YAML file. Click OK after making the modifications.

Clusters	c8def192cdbco4ef1e5c50078cd184e44				
Namespace	default	default			
* Config Map Name:	test				
Configuration:	Variable Name	Variable Value	Action		
	enemies	aliens	Edit Delete		
	lives	3	Edit Delete		
	Name	Value	Add		
	Variable key must be un Edit YAML file	ique. Variable key and value cannot be empty.			
		OK Cancel			

6. After modifying the configurations, click OK.

Update a config map in Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.
- 3. Click Dashboardat the right of the cluster.

Container Service	Cluster List			You can create up	to 5 clusters and can add	up to 20 nodes in each cluster.	Refresh	Create Kubernetes Cluster
Clusters	Help: ${\mathscr O}$ Create cluster ${\mathscr O}$ Scale cluster	${\mathscr S}$ Connect to Kubernetes clu	ster via kubectl 🔗 Mar	nage applications w	ith commands			
Clusters	Name 🔻							
Nodes	Cluster Name/ID	Region	Network Type	Cluster Status	Time Created	Kubernetes Version		Action
Storage	test clinin't Ricciboo kafta Sicki Officit Bika Ak	China East 1 (Hangzhou)	VPC	Running	04/24/2018,09:58:28	1.9.3	Manage	View Logs Dashboard Scale Cluster More -
Deployment								

4. Under Kubernetes, select a namespace, click Config and Storage > Secrets in the left-side navigation pane. Select the target secret and clickActions > View/edit YAML.

Storage Classes	Config and storage Config Ma	aps
Namespace	Config Maps	
default 👻		
Overview	Name 👻	
Workloads	test	
Cron Jobs	aliuma confin	
Daemon Sets	aliyun-coniig	
Deployments		
Jobs		
Pods		
Replica Sets		
Replication Controllers		
Stateful Sets		
Discovery and Load Baland		
Ingresses		
Services 1		
Config and Storage		
Config Maps 2		

5. The Edit a Secret dialog box appears. Modify the configurations and then click UPDATE.



1.11.5 Delete a config map

You can delete a config map that is no longer in use.

Delete a config map on Config Maps page

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Config Maps in the left-side navigation pane.

3. Select the target cluster from the Cluster drop-down list. Click Delete at the right of the config map.

Container Service	Config Maps			Refresh Create
Clusters	Cluster: test 🔻 4			
Clusters	Config Map Name	Namespace	Time Created	Operation
Nodes	aliyun-config	default	04/24/2018,15:41:32	Delete Modify
Storage	test	default	04/24/2018,17:01:14	5 Delete Modify
Application 2	cluster-info	kube-public	04/24/2018,10:05:36	Delete Modify
Service	extension-apiserver-authentication	kube-system	04/24/2018,10:05:34	Delete Modify
Release	ingress-controller-leader-nginx	kube-system	04/24/2018,10:15:55	Delete Modify
Config Maps 3	kube-flannel-cfg	kube-system	04/24/2018,10:05:44	Delete Modify

Delete a config map in Kubernetes dashboard

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.
- 3. Click Clusters in the left-side navigation pane, select the target cluster, and click Dashboard on the right.



4. Under Kubernetes, select a namespace, click Config and Storage > Secrets in the left-side navigation pane. Click the actions button on the right and click Delete in the drop-down list.

Storage Classes	Config and storage Config Maps		+ CREATE
Namespace	Config Maps		Ŧ
Quaniau	Name 🗢	Labels	Age 🗢
Workloads	test		4 04 Delete
Cron Jobs Daemon Sets	aliyun-config		20: View/edit YAML 04- 15:41:32
Deployments			
Jobs			
Pods			
Replica Sets			
Replication Controllers			
Stateful Sets			
Discovery and Load Balan(Gal
Ingresses			act Us
Services			
Config and Storage			
Config Maps 3			

5. Click Delete in the displayed dialog box.

1.11.6 Create a secret

Prerequisites

You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Context

We recommend that you use secrets for sensitive configurations in Kubernetes clusters, such as passwords and certificates.

Secrets have many types. For example:

- Service Account: Automatically created by Kubernetes, which is used to access Kubernetes APIs and is automatically mounted to the pod directory / run / secrets / kubernetes . io / serviceacc ount .
- Opaque: Secret in the base64 encoding format, which is used to store sensitive information such as passwords and certificates.

By default, you can only create secrets of the Opaque type in the Container Service console. Opaque data is of the map type, which requires the value to be in the base64 encoding format. Alibaba Cloud Container Service supports creating secrets with one click and automatically encoding the clear data to base64 format.

You can also create secrets manually by using command lines. For more information, see *Kubernetes secrets* .

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Secrets in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Create in the upper-right corner.

Container Service - Kubernetes 🕶	Secrets				Refresh Create
Overview	Clusters k8s-test v Namespaces default	3			4
Clusters	Name	Туре	Namespaces	Time Created	Action
 Applications 	acr-credential-be5ac8be6a88c42ac1d831b85135a585	kubernetes.io/dockerconfigjson	default	05/17/2019,18:21:02	Details Edit Delete
Service Mesh	acr-credential-c6c63a12436031a13361512e545490fd	kubernetes.io/dockerconfigjson	default	05/17/2019,18:21:02	Details Edit Delete
Discovery and Load B	default-token-v7]46	kubernetes.io/service-account-token	default	05/17/2019,18:13:26	Details Edit Delete
Configuration					
Config Maps					
Secrets 2					
▼ Store					

4. Complete the configurations to create a secret.



To enter the clear data of the secret, select the Encode data values using Base64 check box.

* Name	account 1 Name must consist of lowercase alphanumeric characters, '-' or '.'. Name cannot be empty.			
* Type	Opaque Private Repository Logon Password TLS Certificate			
* Data	Name Value			
	password lqaz2wsx09			
	● username admin			
Names can only contain numbers, letters, "_", "-" and "." Cancel				

- a. Name: Enter the secret name, which must be 1–253 characters long, and can only contain lowercase letters, numbers, hyphens (-), and dots (.).
- b. Configure the secret data. Click the add icon next to Name and enter the name and value of the secret, namely, the key-value pair. In this example, the secret contains two values: username : admin and passwrod : 1f2d1e2e67 df.
- c. Click OK.

5. The Secret page appears. You can view the created secret in the secret list.

Secrets				Refresh Create
Clusters k8s-test v Namespaces default	¥			
Name	Туре	Namespaces	Time Created	Action
account	Opaque	default	05/20/2019,10:52:53	Details Edit Delete
acr-credential-be5ac8be6a88c42ac1d831b85135a585	kubernetes.io/dockerconfigjson	default	05/17/2019,18:21:02	Details Edit Delete
acr-credential-c6c63a12436031a13361512e545490fd	kubernetes.io/dockerconfigjson	default	05/17/2019,18:21:02	Details Edit Delete
default-token-v7j46	kubernetes.io/service-account-token	default	05/17/2019,18:13:26	Details Edit Delete

1.11.7 View secret details

You can view the details of a created secret in the Container Service console.

Prerequisites

- You have created an Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- · You have created a secret. For more information, see Create a secret.

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Secrets in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Detail at the right of the secret.

Container Service	Secret	-			Refrest	Create
Overview	Clusters test Namespace	default 🔻 3				
▼ Clusters	Name	Туре	Namespace	Time Created	4	Action
Clusters	account	Opaque	default	06/13/2018,11:39:06	Detail E	dit Delete
Nodes	default-token-6wbb6	kubernetes.io/service-account-token	default	06/13/2018,11:13:40	Detail E	dit Delete
Storage						
Namespace						
Application						
Deployment						
Pods 📼						
Service						
Ingress						
Release						
Config Maps Secret						

4. You can view the basic information of the secret, and the data that the secret contains.

Click the icon at the right of the data name under Detail to view the clear data.

Secret - account	Refresh Edit
Basic Information	
Name: account	Namespace : default
Time Created: 06/13/2018,11:39:06	Type: Opaque
Detail	
password 💿	
username	

1.11.8 Update a secret

You can update an existing secret directly in the Container Service console.

Prerequisites

- You have created an Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a secret. For more information, see Create a secret.

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Secrets in the left-side navigation pane.
- 3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Edit at the right of the secret.

Container Service	Secret					Refresh	Create
Overview	Clusters test Namespace	default 🔻 🛛					
▼ Clusters	Name	Туре	Namespace	Time Created		4	Action
Clusters	account	Opaque	default	06/13/2018,11:39:06	Detail	Edit	Delete
Nodes	default-token-6wbb6	kubernetes.io/service-account-token	default	06/13/2018,11:13:40	Detail	Edit	Delete
Storage							
Namespace							
Application							
Deployment							
Pods 🖆							
Service							
Ingress							
Release							
Config Maps							
Secret 2							

4. Update the secret data on the Edit Secret page.

Edit Secret				
Namespace	default			
* Name	account			
* Data				
	Name	Value		
		admin		
	Gaeriane			
		172018286/df		
	password			
	Names can only contain numbers, letters, "_", "-" and "."			
		OK Cancel		

5. Click OK.

1.11.9 Delete a secret

You can delete an existing secret directly in the Container Service console.

Prerequisites

- You have created an Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a secret. For more information, see Create a secret.

Context

Ê

Note:

Do not delete the secret generated when the cluster is created.

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Secrets in the left-side navigation pane.

3. Select the cluster and namespace from the Clusters and Namespace drop-down lists. Click Delete at the right of the secret.

Container Service	Secret				Refresh Create
Kubernetes Swarm	Churters Newsers				
Overview	Clusters test V Namespa				
▼ Clusters	Name	Туре	Namespace	Time Created	4 Action
Clusters	account	Opaque	default	06/13/2018,11:39:06	Detail Edit Delete
Nodes	default-token-6wbb6	kubernetes.io/service-account-token	default	06/13/2018,11:13:40	Detail Edit Delete
Storage					
Namespace					
Application					
Deployment					
Pods 🔤					
Service					
Ingress					
Release					
Config Maps					
Secret 2					

4. Click Confirm in the displayed dialog box.

Note			×
0	Are you sure to delete account ?		
		Confirm	Cancel

1.12 Storage management

1.12.1 Overview

Container Service supports automatically binding Kubernetes pods to Alibaba Cloud cloud disks, NAS, and Object Storage Service (OSS).

Currently, static storage volumes and dynamic storage volumes are supported. See the following table for how each type of data volumes supports the static data volumes and dynamic data volumes.

Alibaba Cloud storage	Static data volume	Dynamic data volume
Alibaba Cloud cloud disk	 You can use the cloud disk static storage volumes by: Using the volume method. Using PV/PVC. 	Supported.

Alibaba Cloud storage	Static data volume	Dynamic data volume
Alibaba Cloud NAS	You can use the NAS static storage volumes by:	Supported.
	 Using flexvolume plug- in. 	
	 Using the volume method. Using PV/PVC. Using NFS drive of Kubernetes. 	
Alibaba Cloud OSS	 You can use the OSS static storage volumes by: Using the volume method. Using PV/PVC. 	Not supported.

1.12.2 Install the plug-in

Deploy the Alibaba Cloud Kubernetes storage plug-in by using the following yaml configurations.



If your Kubernetes cluster is created before February 6th, 2018, install the Alibaba Cloud Kubernetes storage plug-in before using the data volumes. If your Kubernetes cluster is created after February 6th, 2018, you can directly use the data volumes without installing the Alibaba Cloud Kubernetes storage plug-in.

Limits

Currently, CentOS 7 operating system is supported.

Instructions

- Disable the -- enable controller attach detach option by using kubelet if you use the flexvolume. By default, Alibaba Cloud Kubernetes clusters have disabled this option.
- Deploy flexvolume in the kube-system user space.

Verify that the installation is complete

On the master node:

- Run the kubectl get pod n kube system | grep flexvolume
 command. Output is the list of running pods (number of nodes).
- Run the kubectl get pod n kube system | grep alicloud disk controller command. Output is the list of running pods.

Installation example

Install flexvolume

```
apiVersion : apps / v1 # for versions
    extensions / v1beta1
                                             before 1.8.0
                                                                  use
kind : DaemonSet
metadata :
  name : flexvolume
  namespace : kube - system
  labels :
    k8s - volume : flexvolume
spec :
  selector :
    matchLabel s :
      name : acs - flexvolume
  template :
    metadata :
      labels :
       name : acs - flexvolume
    spec :
      hostPID : true
      hostNetwor k : true
      toleration s :
      key : node - role . kubernetes . io / master
       operator : Exists
        effect :
                NoSchedule
      containers :
       name : acs - flexvolume
        image : registry . cn - hangzhou . aliyuncs . com / acs /
flexvolume : v1 . 9 . 7 - 42e8198
        imagePullP olicy : Always
        securityCo ntext :
          privileged : true
        env :

    name : ACS_DISK

          value : " true "
       - name : ACS_NAS
          value : " true "
       - name : ACS_OSS
          value : " true "
        resources :
          limits :
            memory : 200Mi
          requests :
            cpu : 100m
            memory : 200Mi
        volumeMoun ts :
          name : usrdir
          mountPath : / host / usr /
          name : etcdir
          mountPath : / host / etc /
          name : logdir
          mountPath : / var / log / alicloud /
```
```
volumes :
- name : usrdir
hostPath :
    path : / usr /
- name : etcdir
hostPath :
    path : / etc /
- name : logdir
hostPath :
    path : / var / log / alicloud /
```

Install Disk provisioner

```
kind : StorageCla ss
apiVersion : storage . k8s . io / v1beta1
metadata :
 name : alicloud - disk - common
provisione r : alicloud / disk
parameters :
  type : cloud
kind : StorageCla ss
apiVersion : storage . k8s . io / v1beta1
metadata :
  name : alicloud - disk - efficiency
provisione r : alicloud / disk
parameters :
 type : cloud_effi ciency
kind : StorageCla ss
apiVersion : storage . k8s . io / v1beta1
metadata :
  name : alicloud - disk - ssd
provisione r : alicloud / disk
parameters :
 type : cloud_ssd
kind : StorageCla ss
apiVersion : storage . k8s . io / v1beta1
metadata :
   name : alicloud - disk - available
provisione r : alicloud / disk
parameters :
 type : available
____
kind : ClusterRol e
apiVersion : rbac . authorizat ion . k8s . io / v1beta1
metadata :
  name : alicloud - disk - controller - runner
 rules :
 resources : [" persistent volumes "]
verbs : [" get ", " list ", " watch ", " create ", " delete "]
- apiGroups : [""]
    resources : [" persistent volumeclai ms "]
verbs : [" get ", " list ", " watch ", " update "]
 - apiGroups : [" storage . k8s . io "]
    resources : [" storagecla sses "]
verbs : [" get ", " list ", " watch "]
apiGroups : [""]
     resources : [" events "]
     verbs : [" list ", " watch ", " create ", " update ", " patch "]
```

```
apiVersion : v1
kind : ServiceAcc ount
metadata :
 name : alicloud - disk - controller
 namespace : kube - system
kind : ClusterRol eBinding
apiVersion : rbac . authorizat ion . k8s . io / v1beta1
metadata :
 name : run - alicloud - disk - controller
subjects :
    kind : ServiceAcc ount
    name : alicloud - disk - controller
    namespace : kube - system
roleRef :
  kind : ClusterRol e
 name : alicloud - disk - controller - runner
 apiGroup : rbac . authorizat ion . k8s . io
kind : Deployment
apiVersion : extensions / v1beta1
metadata :
  name : alicloud - disk - controller
  namespace : kube - system
spec :
  replicas : 1
  strategy :
   type : Recreate
  template :
   metadata :
      labels :
        app : alicloud - disk - controller
    spec :
     toleration s :
     - effect : NoSchedule
       operator : Exists
       key : node - role . kubernetes . io / master
      effect : NoSchedule
       operator : Exists
        key : node . cloudprovi der . kubernetes . io / uninitiali
zed
      nodeSelect or :
         node - role . kubernetes . io / master : ""
      serviceAcc ount : alicloud - disk - controller
      containers :
       - name : alicloud - disk - controller
         image : registry . cn - hangzhou . aliyuncs . com / acs /
alicloud - disk - controller : v1 . 9 . 3 - ed710ce
          volumeMoun ts :
           - name : cloud - config
             mountPath : / etc / kubernetes /
             name : logdir
             mountPath : / var / log / alicloud /
      volumes :
         name : cloud - config
         hostPath :
           path : / etc / kubernetes /
         name : logdir
         hostPath :
```

```
path : / var / log / alicloud /
```

1.12.3 Use Alibaba Cloud cloud disk volumes

You can use Alibaba Cloud cloud disk volumes in a Kubernetes cluster of Alibaba Cloud Container Service.

You can mount an Alibaba Cloud cloud disk to a Kubernetes cluster by using the following two methods:

Static volumes

You can use a static cloud disk volume in either of the following ways:

- Use a cloud disk through a volume.
- Use a cloud disk through a PV and PVC.
- Dynamic volumes



Depending on the type of cloud disk you create, the following requirements must be met:

- The minimum capacity of a basic cloud disk is 5 GiB.
- The minimum capacity of an Ultra disk is 20 GiB.
- The minimum capacity of an SSD disk is 20 GiB.

Static volumes

You can use an Alibaba Cloud cloud disk through a volume or through a PV and PVC.

Prerequisites

You have created a cloud disk in the ECS console. For more information, see *Create a cloud disk*.

Limits

- A cloud disk is a non-shared storage device and can be mounted to only one pod.
- You must have created a cloud disk and obtained the disk ID before using the cloud disk volume. For more information, see *Create a cloud disk*.
- The volumeId parameter indicates the ID of a mounted cloud disk. The volume name and PV name must be the same as the value of the volumeId parameter.

- In a Kubernetes cluster, a cloud disk can be mounted only to a node that resides in the same zone as the cloud disk.
- Only Pay-As-You-Go cloud disks can be mounted. In a Kubernetes cluster, the ECS instance billing method can be changed to Subscription, but the cloud disk billing method cannot be changed to Subscription. Otherwise, the cloud disks will fail to be mounted.

Use a cloud disk through a volume

Use the following disk - deploy . yaml file to create a pod:

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
  name : nginx - disk - deploy
spec :
  replicas : 1
  template :
   metadata :
      labels :
        app : nginx
    spec :
      containers :
      name : nginx - flexvolume - disk
       image : nginx
        volumeMoun ts :
            name : " d - bp1j17ifxf asvts3tf40 "
            mountPath : "/ data "
      volumes :
          name : " d - bp1j17ifxf asvts3tf40 "
          flexVolume :
            driver : " alicloud / disk "
            fsType : " ext4 "
            options :
              volumeId : " d - bp1j17ifxf asvts3tf40 "
```

Use a cloud disk through a PV and PVC

Step 1: Create a cloud disk PV

You can create a cloud disk PV in the Container Service console or by using a YAML file.

Create a PV by using a YAML file

Use the following disk - pv . yaml file to create a PV:

Note:

The PV name must be the same as the cloud disk ID.

apiVersion : v1 kind : Persistent Volume

```
metadata :
  name : d - bp1j17ifxf asvts3tf40
  labels :
    failure - domain . beta . kubernetes . io / zone : cn - hangzhou
- b
    failure - domain . beta . kubernetes . io / region : cn -
hangzhou
spec :
  capacity :
    storage : 20Gi
  storageCla ssName : disk
  accessMode s :
      ReadWrite0 nce
  flexVolume :
    driver : " alicloud / disk "
    fsType : " ext4 "
    options :
      volumeId : " d - bp1j17ifxf asvts3tf40 "
```

Create a cloud disk volume in the Container Service console

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Volumes.
- 3. Select the target cluster and then click Create in the upper-right corner.

Container Service - Kubernetes 👻	Volumes List					Refresh Create
Overview	Clusters k8s-cluster 🔻 3					4
Clusters	Name	Capacity Access Mode	Reclaim Policy	Status Storage Class Name	Time Created	Action
Clusters						
Nodes						
Volumes 2						
Namespace						

- 4. In the displayed dialog box, set the volume parameters.
 - Storage type: Cloud Disk is used in this example.
 - · Access Mode: By default, it is set to ReadWriteOnce.
 - Cloud Disk ID: We recommend that you select a cloud disk that is in the same region and zone as the cluster.
 - File System Type: Select a data type for the data to be stored. The available data types include ext4, ext3, xfs, and vfat. The default setting is ext4.
 - Tag: Add tags to the volume.

Create volume		\times
Storage type : Access Mode:	Cloud Disk NAS OSS ReadWriteOnce	
Cloud Disk ID:	Select Cloud Disk	
File System Type:	ext4 •	
Tag :	Add Tag	
	Create	Cancel

5. Click Create.

Step 2: Create a PVC

Use the following disk - pvc . yaml file to create a PVC:

```
kind : Persistent VolumeClai m
apiVersion : v1
metadata :
    name : pvc - disk
spec :
    accessMode s :
    - ReadWrite0 nce
    storageCla ssName : disk
    resources :
        requests :
        storage : 20Gi
```

Step 3: Create a pod

Use the following disk - pod . yaml file to create a pod:

```
apiVersion : v1
kind : Pod
metadata :
    name : " flexvolume - alicloud - example "
spec :
    containers :
        - name : " nginx "
        image : " nginx "
        volumeMoun ts :
            - name : pvc - disk
            mountPath : "/ data "
volumes :
        - name : pvc - disk
        persistent VolumeClai m :
        claimName : pvc - disk
```

Dynamic volumes

To use a dynamic volume, you need to manually create a StorageClass, and specify a cloud disk type through storageClassName in a PVC.

Create a StorageClass

```
kind : StorageCla ss
apiVersion : storage . k8s . io / v1beta1
metadata :
    name : alicloud - disk - ssd - hangzhou - b
provisione r : alicloud / disk
parameters :
    type : cloud_ssd
    regionid : cn - hangzhou
    zoneid : cn - hangzhou - b
reclaimPol icy : Retain
```

Parameter setting:

- provisioner: Set this parameter to alicloud/disk to indicate that the StorageClass creates a cloud disk by using the provisioner plugin of Alibaba Cloud cloud disks.
- type: Specify the type of a cloud disk by using one the following values: cloud, cloud_efficiency, cloud_ssd, and available. If you set this parameter to available, the system will cycle through cloud_efficiency, cloud_ssd, and cloud in order until one of them takes effect.
- · regionid: Set the region in which you want to create a cloud disk.
- reclaimPolicy: Set the policy to reclaim a cloud disk. The default setting is Delete.
 You can also set this parameter to Retain.
- zoneid: Set the zone in which you want to create a cloud disk.



If you want to create cloud disks in multiple zones, you can set multiple values for the zoneid parameter, for example,

```
zoneid : cn - hangzhou - a , cn - hangzhou - b , cn - hangzhou - c
```

 encrypted: (optional) Set whether to encrypt a cloud disk. The default value is false . That is, a cloud disk will not be encrypted.

Create a service

```
kind : Persistent VolumeClai
                                 m
apiVersion : v1
metadata :
  name : disk - ssd
spec :
  accessMode s:
    - ReadWriteO nce
  storageCla ssName : alicloud - disk - ssd - hangzhou - b
  resources :
    requests :
      storage : 20Gi
kind : Pod
apiVersion :
             v1
metadata :
 name :
         disk – pod – ssd
spec :
  containers :
   name : disk - pod
image : nginx
    volumeMoun ts :
- name : disk - pvc
        mountPath : "/ mnt "
  restartPol icy : " Never "
  volumes :
      name : disk - pvc
      persistent VolumeClai m :
        claimName : disk - ssd
```

Default options

By default, Kubernetes clusters provide the following StorageClasses that can be used in the single-zone clusters:

- · alicloud-disk-common, namely, a basic cloud disk.
- · alicloud-disk-efficiency, namely, an Ultra disk.
- alicloud-disk-ssd, namely, an SSD disk.
- alicloud-disk-available: This StorageClass provides a systematic method of disk selection. Specifically, the system first attempts to create an Ultra disk. If the Ultra disks in the specified zone are sold out, the system tries to create an SSD disk. If the SSD disks are sold out, the system tries to create a basic cloud disk.

Create a multi-instance StatefulSet by using a cloud disk

We recommend that you create a multi-instance StatefulSet through volumeClai mTemplates so that you can dynamically create multiple PVCs and PVs, and connect the PVCs and PVs together.

```
apiVersion : v1
kind : Service
metadata :
  name : nginx
  labels :
   app : nginx
spec :
 ports :
   port: 80
   name : web
  clusterIP : None
  selector :
    app : nginx
apiVersion : apps / v1beta2
kind : StatefulSe t
metadata :
  name : web
spec :
  selector :
   matchLabel s :
     app : nginx
  serviceNam e : " nginx "
  replicas :
            2
  template :
   metadata :
      labels :
       app : nginx
    spec :
     containers :

    name : nginx

       image : nginx
       ports :
       - containerP
                     ort : 80
         name : web
       volumeMoun ts :
         name : disk - ssd
         mountPath : / data
  volumeClai mTemplates :
   metadata :
      name : disk - ssd
    spec :
      accessMode s : [ " ReadWriteO nce " ]
      storageCla ssName : " alicloud - disk - ssd "
      resources :
        requests :
                    20Gi
          storage :
```

1.12.4 Use NAS file systems of Alibaba Cloud

You can use Alibaba Cloud NAS volumes in a Kubernetes cluster of Container Service.

You can mount a NAS file system of Alibaba Cloud to a Kubernetes cluster as either of the following two types of volumes:

Static volumes

You can use a static volume in either of the following two ways:

- Use a static volume through the flexvolume plugin.
 - Use a static volume directly.
 - Use a static volume through a Persistent Volume (PV) and a Persistent Volume Claim (PVC).
- Use a static volume through the NFS driver of Kubernetes.
- Dynamic volumes

Prerequisites

You have created a NAS file system in the NAS console and added a mount point for a Kubernetes cluster in the file system. You must make sure that the NAS file system and your cluster are in the same VPC.

Static volumes

You can use the Alibaba Cloud NAS file storage service by using the flexvolume plugin provided by Alibaba Cloud or the NFS driver of Kubernetes.

Use a static volume through the flexvolume plugin

With a flexvolume plugin, you can use an Alibaba Cloud NAS volume directly or through a PV and a PVC.



Note:

- NAS: a shared storage system that can provide storage services for multiple pods at the same time.
- server: defines the mount point of a NAS file system.
- path: defines the mount directory that connects to the NAS volume. You can specify a NAS sub-directory and mount it to your NAS volume. If the NAS subdirectory specified by you does not exist, the system automatically creates the NAS sub-directory and mounts it to your NAS volume.
- vers: defines the version number of the NFS mount protocol. NFS file system versions 3.0 and 4.0 are supported.

 mode:defines the access permission to a mount directory. When the mount directory is the root directory of a NAS file system, the access permission to the root directory cannot be set. If you set the mode parameter for a NAS file system that stores a large amount of data, the process of mounting the NAS file system to a cluster may take an excessive amount of time or even fail.

Use a static volume directly

Use a nas - deploy . yaml file to create a pod as follows:

```
apiVersion : v1
kind : Pod
metadata :
  name : " flexvolume - nas - example "
spec :
  containers :
   - name : " nginx "
      image : " nginx "
      volumeMoun ts :
      - name : " nas1 "
          mountPath : "/ data "
  volumes :
      name : " nas1 "
      flexVolume :
        driver : " alicloud / nas "
        options :
          server : " 0cd8b4a576 - grs79 . cn - hangzhou . nas .
aliyuncs . com "
          path : "/ k8s "
          vers : " 3 "
```

Use a static volume through a PV and a PVC

Step 1: Create a PV

You can create a NAS volume by using a YAML file or create a NAS volume in the Alibaba Cloud Container Service console.

· Create a PV by using a YAML file.

Use a nas - pv . yaml file to create a PV as follows:

```
apiVersion : v1
kind : Persistent Volume
metadata :
   name : pv - nas
spec :
   capacity :
    storage : 5Gi
   storageCla ssName : nas
   accessMode s :
        - ReadWriteM any
   flexVolume :
        driver : " alicloud / nas "
        options :
```

```
server : " 0cd8b4a576 - uih75 . cn - hangzhou . nas .
aliyuncs . com "
path : "/ k8s "
```

vers : " 3 "

- · Create a NAS volume in the Container Service console.
 - 1. Log on to the Container Service console.
 - 2. In the left-side navigation pane under Kubernetes, choose Clusters > Volumes.
 - 3. Select the target cluster from the cluster drop-down list and then click Create in the upper-right corner.

Container Service - Kubernetes +	Volumes List	:						Refresh	Create
Overview	Clusters k8s-t	test	•						
 Clusters 	Name C	Capacity A	Access Mode	Reclaim Policy	Status	Storage Class Name	Binding Volume Claim	Time Created	Action
Clusters									
Nodes				(i) Cor	uld not find	any record that met th	e condition.		
Volumes									
Namespace									

- 4. In the displayed dialog box, set the volume parameters.
 - Storage type: NAS is selected in this example.
 - Name: Customize a volume name. The volume name must be unique in the cluster. In this example, pv-nas is set as the volume name.
 - Capacity: Set the volume capacity. Make sure that the volume capacity does not exceed the NAS file system capacity.
 - Access Mode: By default, it is set to ReadWriteOnce.
 - Mount Point Domain Name: Enter the mount address of the mount point that is used to mount the NAS file system to the Kubernetes cluster.
 - Path: sub-directory under the NAS path, which starts with a forward slash (
 /). If you specify a sub-directory, your volume will be mounted to the sub-directory.
 - If no sub-directory exists in the root directory of a NAS file system, the system automatically creates a sub-directory by default.
 - This parameter is optional. A NAS volume is mounted to the root directory of a NAS file system by default.
 - Privilege: Set the access permission to the mount directory. For example, you can set this parameter to 755, 644, or 777.
 - You can set this parameter only if you mount a NAS volume to the NAS sub -directory. This parameter cannot be set if you mount a NAS volume to the NAS root directory.

- This parameter is optional. By default, the original access permission to a NAS file system is used.
- Tag: Add tags to the volume.

Create Data Volume	e	\times
Type: Name:	○ Cloud Disk	
Capacity :	20Gi	
Access Mode :	ReadWriteMany	
Mount Point Domain Name:	additional DDS district, nas.aliyuncs.com	
Path :	Example : /xxx	
Privilege :	Example : 755	
Tag :	OAdd Tag	
	Create	ancel

5. Click Create.

Step 2: Create a PVC

Use a nas - pvc . yaml file to create a PVC as follows:

```
apiVersion : v1
kind : Persistent VolumeClai m
metadata :
    name : pvc - nas
spec :
    accessMode s :
    - ReadWriteM any
    storageCla ssName : nas
    resources :
        requests :
            storage : 5Gi
```

Step 3: Create a pod

Use a *nas - pod*. *yaml* file to create a pod as follows:

```
apiVersion : v1
kind : Pod
metadata :
```

```
name : " flexvolume - nas - example "
spec :
    containers :
        - name : " nginx "
        image : " nginx "
        volumeMoun ts :
            - name : pvc - nas
            mountPath : "/ data "
volumes :
        - name : pvc - nas
        persistent VolumeClai m :
        claimName : pvc - nas
```

Use the Kubernetes NFS driver



Note:

Alibaba Cloud NAS supports NFS 3.0 and NFS 4.0. You must specify a valid NFS version when you create a NAS volume.

Step 1: Create a NAS file system

Log on to the NAS console to create a NAS file system.



You must ensure that the NAS file system and your cluster are in the same region.

For example, assume that the mount point of your NAS file system is 055f84ad83 -

ixxxx . cn - hangzhou . nas . aliyuncs . com .

Step 2: Create a PV

You can create a NAS volume by using an orchestration template or the Alibaba Cloud Container Service console.

· Use an orchestration template to create a NAS volume

Use a nas - pv . yaml file to create a PV.

Run the following command to create a NAS PV:

```
root @ master # cat << EOF | kubectl apply - f -
apiVersion : v1
kind : Persistent Volume
metadata :
    name : nas
spec :
    capacity :
    storage : 8Gi
    accessMode s :
    - ReadWriteM any
mountOptio ns :
    - noresvport
    - nfsvers = 3</pre>
```

```
persistent VolumeRecl aimPolicy : Retain
nfs :
    path : /
    server : 055f84ad83 - ixxxx . cn - hangzhou . nas . aliyuncs .
com
EOF
```

· Create a NAS volume in the Container Service console

For more information, see Use a PV and a PVC.

Step 2: Create a PVC

Create a PVC to request to bind the PV.

```
apply - f -
root @ master # cat << EOF
                               kubectl
apiVersion : v1
kind : Persistent VolumeClai
                               m
metadata :
    name : nasclaim
spec :
     accessMode s :

    ReadWriteM

                   any
     resources :
      requests :
        storage : 8Gi
EOF
```

Step 3: Create a pod

Create an application to declare to mount and use the volume.

```
cat << EOF
                               | kubectl
root @ master #
                                           apply - f -
apiVersion :
            v1
kind : Pod
metadata :
     name : mypod
spec :
     containers :
       name : myfrontend
        image : registry . aliyuncs . com / spacexnice / netdia :
latest
        volumeMoun ts :
        - mountPath : "/ var / www / html "
          name : mypd
     volumes :
       name : mypd
        persistent VolumeClai m :
          claimName : nasclaim
EOF
```

The NAS file system is successfully mounted to the application that runs on the pod.

Dynamic volumes

To use a dynamic NAS volume, you need to manually install a driver plugin and configure a NAS mount point.

Note:

To dynamically generate a NAS volume is to automatically generate a directory in an existing NAS file system. This directory is defined as the target volume.

Install a plugin

```
apiVersion : storage . k8s . io / v1
kind : StorageCla ss
metadata :
  name : alicloud - nas
mountOptio ns :
vers = 3
          r : alicloud / nas
provisione
reclaimPol icy : Retain
kind : Deployment
apiVersion : extensions / v1beta1
metadata :
  name : alicloud - nas - controller
  namespace : kube - system
spec :
  replicas : 1
  strategy :
    type : Recreate
  template :
   metadata :
      labels :
        app : alicloud - nas - controller
    spec :
     toleration s :

    effect : NoSchedule

       operator : Exists
        key : node - role . kubernetes . io / master
     - effect : NoSchedule
       operator : Exists
        key : node . cloudprovi der . kubernetes . io / uninitiali
zed
      nodeSelect or :
        node - role . kubernetes . io / master : ""
      serviceAcc ount : admin
      containers :
         name : alicloud - nas - controller
         image : registry . cn - hangzhou . aliyuncs . com / acs /
alicloud - nas - controller : v3 . 1 . 0 - k8s1 . 11
         volumeMoun ts :

    mountPath : / persistent volumes

           name : nfs - client - root
         env :
             name : PROVISIONE R_NAME
             value : alicloud / nas
             name : NFS_SERVER
             value : 0cd8b4a576 - mmi32 . cn - hangzhou . nas .
aliyuncs . com
             name : NFS_PATH
             value : /
     volumes :
       name : nfs - client - root
        flexVolume :
         driver : alicloud / nas
```

Use the dynamic volume

```
apiVersion : apps / v1beta1
kind : StatefulSe t
metadata :
  name : web
spec :
  serviceNam e : " nginx "
  replicas :
              2
            ∠
mTemplates :
 volumeClai
   metadata :
     name : html
    spec :
      accessMode s :
          ReadWrite0 nce
      storageCla ssName : alicloud - nas
      resources :
        requests :
          storage : 2Gi
  template :
    metadata :
      labels :
        app : nginx
    spec :
     containers :
     – name : nginx
        image : nginx : alpine
       volumeMoun ts :
       - mountPath : "/ usr / share / nginx / html /"
          name : html
```

1.12.5 Use Alibaba Cloud OSS volumes

You can use Alibaba Cloud OSS volumes in a Kubernetes cluster of Alibaba Cloud Container Service.

Specifically, you can only use static OSS volumes. Dynamic OSS volumes are not supported. You can use a static OSS volume in either of the following two ways:

- Use an OSS bucket through a volume.
- Use an OSS bucket through a Persistent Volume (PV) and a Persistent Volume Claim (PVC).

Prerequisites

You have created a bucket in the OSS console.

OSS parameter setting

- OSS: OSS is a shared storage system that can provide storage services to multiple pods at the same time.
- bucket: Only buckets can be mounted to a Kubernetes cluster. The sub-directories or files under a bucket cannot be mounted to a Kubernetes cluster.
- url: Specify an OSS endpoint, namely, the domain name used to mount an OSS bucket to a cluster.
- akId: Enter your Access Key ID.
- · akSecret: Enter your Access Key Secret.
- otherOpts: Customize other parameters in the format of o *** o ***.

Notices

- If your Kubernetes cluster is created before February 6th, 2018, *Install the plug-in* before using a volume. Before you can use the OSS volume, you must first create a secret and then enter your Access Key information into the secret when you deploy the flexvolume service.
- If you use the flexvolume component of an earlier version, we recommend that you upgrade it to the latest version.

Note:

When you upgrade a Kubernetes cluster or restart a kubelet, an OSS volume mounted by the flexvolume component of an earlier version will cause the OSSFS driver to restart. To solve the exception caused by this event, you must recreate the pod that used the OSS volume to remount the OSS volume to the cluster. However, you can solve this issue more easily by upgrading the flexvolume component to the latest version.

Use a static OSS volume

Use an OSS bucket through a volume

Use a oss - deploy . yaml file to create a pod.

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
   name : nginx - oss - deploy
spec :
   replicas : 1
   template :
      metadata :
```

```
labels :
        app : nginx
    spec :
      containers :
       name : nginx - flexvolume - oss
        image : nginx
        volumeMoun ts :
    - name : " oss1 "
            mountPath : "/ data "
        livenessPr obe :
          exec :
            command :
           – sh
           - - c
           - cd / data
          initialDel aySeconds : 30
          periodSeco nds: 30
      volumes :
          name : " oss1 "
          flexVolume :
            driver : " alicloud / oss "
            options :
              bucket : " docker "
              url : " oss - cn - hangzhou . aliyuncs . com "
              akId : ***
              akSecret : ***
              otherOpts : "- o max_stat_c ache_size = 0 - o
            r "
allow_othe
```

Use a PV and a PVC

Step 1: Create a PV

You can create a PV by using a YAML file or the Container Service console.

Use a YAML file to create a PV

Use a oss - pv . yaml file to create a PV as follows:

```
apiVersion :
             v1
kind : Persistent Volume
metadata :
 name : pv - oss
spec :
  capacity :
   storage : 5Gi
  accessMode s :
     ReadWriteM any
  storageCla ssName :
                       055
  flexVolume :
    driver : " alicloud / oss "
    options :
      bucket : " docker "
     url : " oss - cn - hangzhou . aliyuncs . com "
      akId : ***
      akSecret : ***
      otherOpts : "- o
                         max_stat_c ache_size = 0 - o
                                                          allow_othe
r "
```

Create an OSS volume in the Container Service console

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Kubernetes, choose Clusters > Volumes.
- 3. Select the target cluster from the cluster drop-down list and then click Create in the upper-right corner.

Container Service - Kubernetes +	Volumes and Volumes Claim							
Overview	Volumes Volumes Claim							4
Clusters	Clusters k8s-cluster 🔻	3						Refresh Create
Clusters	Name C	Capacity Access Mode	Reclaim Policy	Status	Storage Class Name	Binding Volume Claim	Time Created	Action
Nodes Volumes 2	d-bp1g1m7s7snofbflvfrd 2	20Gi ReadWriteOnce	Delete	Bound	alicloud-disk-efficiency	Namespace: test Name:jenkins-demo-jenkins	02/18/2019,15:29:08	Edit Labels YAML Delete

- 4. In the displayed dialog box, set the volume parameters.
 - Storage type: OSS is selected in this example.
 - Name: Customize a volume name. The volume name must be unique in the cluster. In this example, pv-oss is set as the volume name.
 - Capacity: Set the volume capacity.
 - · Access Mode: By default, it is set to ReadWriteMany.
 - AccessKey ID and AccessKey Secret: Use these two parameters to specify the Access Key used to access OSS.
 - Bucket ID: Select an OSS bucket name. Click Select Bucket. In the displayed dialog box, select the target bucket and clickSelect.
 - Access Domain Name. If the selected bucket and the cluster ECS instances are in different regions, you need to selectInternet. If they are in the same region, your choice is dependent on your cluster network type. If your cluster uses a VPC,

you need to select VPC; if your cluster uses a classic network, you need to select Intranet.

Create Data Volume	• ×
Type:	○ Cloud Disk ○ NAS ◎ OSS
Capacity :	20Gi
Access Mode :	ReadWriteMany
Access Key ID:	wapowinista
Access Key Secret:	adghafafah
Optional Parameters:	For the formats of other parameters, refer to this document. Example: -o allow_other -o default_permission=666 -onoxattr
Bucket ID:	Select Bucket
Access Domain Name:	○ Intranet ○ Internet ⑧ VPC 🖗
Tag :	●Add Tag
	Create Cancel

• Tag: Add tags to the volume.

5. Click Create.

Step 2: Create a PVC

Use a oss - pvc . yaml file to create a PVC as follows:

```
kind : Persistent VolumeClai m
apiVersion : v1
metadata :
   name : pvc - oss
spec :
   storageCla ssName : oss
   accessMode s :
        - ReadWriteM any
   resources :
        requests :
```

storage : 5Gi

Step 3: Create a pod

Use a oss - pod . yaml file to create a pod.

```
apiVersion :
               v1
kind : Pod
metadata :
  name : " flexvolume - oss - example "
spec :
  containers :
    - name : " nginx "
image : " nginx "
      volumeMoun ts :
- name : pvc - oss
             mountPath : "/ data "
      livenessPr obe :
        exec :
          command :
            sh
          - - c
        - cd / data
initialDel aySeconds :
                                     30
         periodSeco nds : 30
  volumes :
    name : pvc - oss
    persistent VolumeClai m :
        claimName : pvc - oss
```

1.12.6 Create a Persistent Volume Claim

You can create a Persistent Volume Claim (PVC) by using the Container Service console.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a volume. In this example, use a cloud disk to create a cloud storage volume. For more information, see *Use Alibaba Cloud cloud disk volumes*.

By default, the storage claim is bound to the storage volume depending on the label alicloud – pvname . When the data volume is created by using the Container Service console, the storage volume is labeled by default. If the storage volume label does not exist, you must add a label before you select to bound this storage volume.

Context

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Volumes Claim in the left-side navigation pane to enter the Volumes Claims list page.
- 3. Select the target cluster and namespace, and click Create in the upper-right corner.



4. Complete the configurations in the Create Volume Claim dialog box, and click

Create.

Create Volume Claims		\times
Volume type : Name :	Cloud Disk NAS OSS pv-disk	
	Name must start with a lowercase letter and can only contain lowercase letters, numbers, "." and "-"	
Allocate mode :	Existing volume	
Existing volume :	20Gi Select Volume	
Capacity :	20Gi	
	Create	Cancel

- Volume claim type: Consistent with volume, including cloud disk, NAS, and OSS types.
- Name: Enter the storage volume claim name.
- · Distribution mode: Currently, only existing storage volumes are supported.
- Existing storage volume: Select to bind the storage volume of this type.
- · Total: Claim usage, cannot be greater than the total amount of storage volumes.

Note:

If a storage volume already exists in your cluster and is not used, but cannot be found in Select Existing Storage Volume, maybe the alicloud – pvname label is not defined.

If you cannot find an available storage volume, you can click Clusters > Volumes in the left-side navigation pane. Find the target storage volume, click Label Management on the right. Add the corresponding label alicloud - pvname , the value is the name of the storage volume. The cloud storage volume defaults to the cloud disk ID as the name of the storage volume.

Edit Labels	\times
Add Tag	
Name	Value
alicloud-pvname	d-bp+ TJSOCOOCT CTTATE TWO
failure-domain.beta.kubernetes.io/zone	cn-hangzhou-g
failure-domain.beta.kubernetes.io/region	cn-hangzhou $igodot$
	OK Close

5. Return to the Volumes Claims list, you can see that the newly created storage claim appears in the list.

1.12.7 Use a persistent volume claim

In the Container Service console, use an image or a template to deploy an application, so that you can use a persistent volume claim (PVC). In this example, an image is used to create an application. If you want to use a PVC through a template, see *Use Alibaba Cloud cloud disk volumes*.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have created a PVC. In this topic, a pvc-disk PVC is created by using a disk. For more information, see *Create a persistent volume claim*.

Procedure

1. Log on to the Container Service console.

2. In the left-side navigation pane under Kubernetes, choose Application > Deployment. Then click Create by Image in the upper-right corner.

Container Service - Kubernetes 👻	Deployment				Refresh Create by image Create by template
Overview	Clusters k8s-cluster v Namespace	default 🔻			2
 Clusters 	Name	Тад	PodsQuantity	Time Created	Action
Clusters	new-nginx	run:new-nginx	1/1	07/11/2018,17:00:19	Details Edit Monitor More 🗸
Nodes	nginx-deployment-basic	app:nginx	2/2	07/11/2018,18:00:12	Details Edit Monitor More 🗸
Volumes	old-nginx	run:old-nginx	2/2	07/11/2018,17:00:19	Details Edit Monitor More 🗸
Namespace					
Application Deployment					

- 3. Set the application name and select the target cluster and the namespace. Then click Next.
- 4. Select an image, set the data volume of disk type, and then click Next. Disk, NAS, and OSS types are available. In this example, an existing disk PVC is used.
- 5. Set services for the test-nginx application, and then click Create.
- 6. In the left-side navigation pane, choose Application > Pods. Then click Details on the right of the pod to which the test-nginx application belongs.
- 7. Click Volumes to verify that the target pod is associated with the pvc-disk.

1.13 Log management

1.13.1 Application log management

A Kubernetes cluster that runs on Alibaba Cloud Container Service provides you with multiple methods to manage application logs.

- Following the instructions of *Use Log Service to collect Kubernetes cluster logs*, you can make the best use of the functions provided by Alibaba Cloud Log Service, such as log statistics and analysis.
- With Log-pilot, an open source project provided by Alibaba Cloud Container
 Service, and A solution to log collection problems of Kubernetes clusters by using log-pilot,
 Elasticsearch, and Kibana, you can easily build your own application log clusters.

1.13.2 View cluster logs

Context

You can view the cluster operation logs by using the simple log service of Container Service.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Clusters in the left-side navigation pane.
- 3. Click View Logs at the right of the cluster.

Container Service	Cluster List			You can create up	to 5 clusters and can add	up to 20 nodes in each cluster.	Refresh	Create Kuberr	netes Cluster
Clusters	Help: ${\mathscr O}$ Create cluster ${\mathscr O}$ Scale cluster	& Connect to Kubernetes clus	ster via kubectl 🔗 Mana	ge applications wi	ith commands				
Clusters 2	Name 🔻								
Nodes	Cluster Name/ID	Region	Network Type	Cluster Status	Time Created	Kubernetes Version		3	Action
Storage	test chtell:SlocbecAella/SLSelCTeel284e44	China East 1 (Hangzhou)	VPC vpe-lape.nettacong	Running	04/24/2018,09:58:28	1.9.3	Manage	View Logs Scale Clus	Dashboard ster More +

View the cluster operation information.

Cluster Logs: test + Back to Cluster List		Refresh
Detailed resource deployment logs: Stack Ev	ients	
Time	Information	
04/24/2018,13:55:38	c8def192cdbcc4af1a515d07fed184e44 Start to client.DescribeTemplate	
04/24/2018,13:55:35	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,11:27:38	c8def192cdbcc4af1a515d07fed184e44 Start to client.DescribeTemplate	
04/24/2018,11:27:36	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,11:27:08	c8def192cdbcc4af1a515d07fed184e44 Start to client.DescribeTemplate	
04/24/2018,11:27:06	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,11:26:55	c8def192cdbcc4af1a515d07fed184e44 Start to client.DescribeTemplate	
04/24/2018,11:26:54	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,11:25:02	c8def192cdbcc4af1a515d07fed184e44 Start to client.DescribeTemplate	
04/24/2018,11:25:00	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,10:16:42	c8def192cdbcc4af1a515d07fed184e44 Set up k8s DNS configuration successfully	
04/24/2018,10:15:36	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	
04/24/2018,10:15:36	c8def192cdbcc4af1a515d07fed184e44 Stack CREATE completed successfully:	
04/24/2018,10:15:34	c8def192cdbcc4af1a515d07fed184e44 Start describeStackInfo	

1.13.3 Use Log Service to collect Kubernetes cluster logs

Log Service is integrated with Kubernetes clusters of Alibaba Cloud Container Service. You can enable Log Service when creating a cluster to quickly collect container logs for the Kubernetes cluster, such as the standard output of the container and text files of the container.

Enable Log Service when creating a Kubernetes cluster

If you have not created any Kubernetes clusters, follow steps in this section to enable Log Service:

- 1. Log on to the Container Service console.
- 2. Click Clusters in the left-side navigation pane and click Create Kubernetes Cluster in the upper-right corner.
- 3. For how to configure a cluster on the creation page, see Create a Kubernetes cluster.
- 4. Drag to the bottom of the page and select the Using Log Service check box. The log plug-in will be installed in the newly created Kubernetes cluster.
- 5. When you select the Using Log Service check box, project options are displayed. A project is the unit in Log Service to manage logs. For more information about projects, see *Project*. Currently, two ways of using a project are available:
 - Select an existing project to manage collected logs.

Log Service	Vsing Log Service			
	Select Project	Create Project	ta igainin fan anis fairean a'	

The system automatically creates a new project to manage collected logs. The project is automatically named k8s - log -{ ClusterID }, where ClusterID represents the unique identifier of the created Kubernetes cluster.



6. After you complete the configurations, click Create in the upper-right corner. In the displayed dialog box, click OK.

After the cluster creation is completed, the newly created Kubernetes cluster is displayed on the cluster list page.

Container Service - Kubernetes -	1	Cluster List			You can create u	p to 5 clusters and c	an add up to 40 nodes in	each cluster.	Refresh	Create Kubernetes Clus	ster
Overview		Help: Ø Create cluster Ø Scale clust	er 🔗 Connec	t to Kubernetes cluster	via kubectl 🛛 🔗 Man	age applications wit	h commands				
Clusters	[Name 🔻									
Clusters		Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Time Created	Kubernetes Version		A	ction
Nodes											
Volumes		k8s-test	Kubernetes	China North 2 (Beijing)	VPC vpc- 2ze8sk2raka	Running	09/18/2018,14:00:45	1.11.2		Manage View Logs Dashbo Scale Cluster Mo	3 I ard ore∙
Namespace											

Manually install Log Service components in a created Kubernetes cluster

If you have created a Kubernetes cluster, following instructions in this section to use Log Service:

- · Log Service components are not installed. Manually install the components.
- Log Service components are installed but in an earlier version. Upgrade the components. If you do not upgrade the components, you can only use the Log Service console or custom resource definition (CRD) to configure log collection.

Check the Log Service component version

1. Configure the local kubeconfig to connect to the Kubernetes cluster through kubectl.

For information about the configuration, see *Connect to a Kubernetes cluster by using kubectl*.

2. Run the following command to fast determine whether an upgrade or migration operation is required:

\$ kubectl describe daemonsets - n kube - system logtail ds | grep ALICLOUD_L OG_DOCKER_ ENV_CONFIG

- If ALICLOUD_L OG_DOCKER_ ENV_CONFIG : true is output, the components can be used directly without requiring upgrade or migration.
- If other results are output, check the components further.
- 3. Run the following command to determine whether Helm is used to install the components.

\$ helm get alibaba - log - controller | grep CHART

CHART : alibaba - cloud - log - 0 . 1 . 1

- 0.1.1 in the output indicates the version of the Log Service components. Please use the version of 0.1.1 and later. If the version is too early, please see *Upgrade Log Service components* to upgrade the components. If you have used Helm to install the components of a valid version, you can skip next steps.
- If no results are output, the components are installed by using Helm. But the DaemonSet installation method might be used. Follow the next step to check further.
- 4. DaemonSet can be an old one or a new one:
 - \$ kubectl get daemonsets n kube system logtail
 - If no result is output or No resources found . is output, the Log Service components are not installed. For information about the installation method, see *Manually install Log Service components*.
 - If the correct result is output, an old DaemonSet is used to install the components which require upgrade. For information about upgrading the components, see *Upgrade Log Service components*.

Manually install Log Service components

1. Configure the local kubeconfig to connect to the Kubernetes cluster through kubectl.

For information about the configuration, see *Connect to a Kubernetes cluster by using kubectl*.

2. Replace a parameter and run the following command.

Replace \${your_k8s_cluster_id} in the following command with your Kubernetes cluster ID, and then run the command.

wget https :// acs - logging . oss - cn - hangzhou . aliyuncs . com / alicloud - k8s - log - installer . sh - 0 alicloud k8s - log - installer . sh ; chmod 744 ./ alicloud - k8s log - installer . sh ; ./ alicloud - k8s - log - installer . sh -- cluster - id \${ your_k8s_c luster_id } -- ali - uid \${ your_ali_u id } -- region - id \${ your_k8s_c luster_reg ion_id
}

Parameter descriptions:

- your_k8s_cluster_id: You Kubernetes cluster ID.
- your_ali_uid: You account ID of Alibaba Cloud, which can be viewed in the user info.
- your_k8s_cluster_region_id: The region in which you Kubernetes cluster resides, which can be found in*Regions and zones*. For example, if the cluster resides in Hangzhou, the value of this parameter cn-hangzhou.

Installation example

```
[ root @ iZbp ***** biaZ ~]# wget https :// acs - logging . oss

uid 19 ******** 19 -- region - id cn - hangzhou
- 2018 - 09 - 28 15 : 25 : 33 -- https :// acs - logging . oss - cn - hangzhou . aliyuncs . com / alicloud - k8s - log - installer .

 sh
Resolving acs - logging . oss - cn - hangzhou . aliyuncs . com ...
118 . 31 . 219 . 217 , 118 . 31 . 219 . 206
Connecting to acs - logging . oss - cn - hangzhou . aliyuncs .
com | 118 . 31 . 219 . 217 |: 443 ... connected .
HTTP request sent , awaiting response ... 200 OK
Length : 2273 ( 2 . 2K ) [ text / x - sh ]
Saving to : ' alicloud - k8s - log - installer . sh '
 alicloud - k8s - log - installer . sh
                                                                        100
 2.22K --.-KB/s
                                    in
                                           0s
 2018 - 09 - 28 15 : 25 : 33 (13.5
                                                  MB / s ) - ' alicloud - k8s
 - log - installer . sh ' saved [ 2273 / 2273 ]
-- 2018 - 09 - 28 15 : 25 : 33 -- http :// logtail - release -
 cn - hangzhou . oss - cn - hangzhou . aliyuncs . com / kubernetes /
 alibaba - cloud - log . tgz
 Resolving logtail - release - cn - hangzhou . oss - cn - hangzhou .
 aliyuncs. com ... 118 . 31 . 219 . 49
Connecting to logtail - release - cn - hangzhou . oss - cn -
 hangzhou . aliyuncs . com | 118 . 31 . 219 . 49 |: 80 ... connected
 HTTP request sent, awaiting response ... 200
Length : 2754 (2.7K) [applicatio n/x-gzip]
                                                                  0K
         to : ' alibaba - cloud - log . tgz'
 Saving
 alibaba - cloud - log . tgz
                                                                      100
 --. - KB / s
    2.69K
                                     in
                                           0s
 2018 - 09 - 28 15 : 25 : 34 ( 79 . 6 MB / s ) - 'alibaba -
 cloud - log . tgz ' saved [ 2754 / 2754 ]
```

[INFO] your project : k8s - log k8s is using c77a92ec5a 3ce4e64a1b f13bde1820 106 NAME : alibaba - log - controller LAST DEPLOYED : Fri Sep 28 15 : 25 : 34 2018 NAMESPACE : default STATUS : DEPLOYED **RESOURCES** : ==> v1beta1 / CustomReso urceDefini tion NAME AGE aliyunlogc onfigs.log.alibabaclo ud.com 0s ==> v1beta1 / ClusterRol e alibaba - log - controller 0s ==> v1beta1 / ClusterRol eBinding NAME AGE alibaba - log - controller 0s ==> v1beta1 / DaemonSet CURRENT READY UP - TO - DATE NAME DESIRED AVAILABLE NODE SELECTOR AGE logtail - ds 2 0 2 2 0 < none > 05 ==> v1beta1 / Deployment UP - TO - DATE NAME CURRENT DESIRED AVAILABLE AGE alibaba - log - controller 1 1 1 0 0s ==> v1 / Pod (related) NAME READY STATUS RESTARTS AGE logtail - ds - 6v979 0 / 1 ContainerC reating 0 05 logtail - ds - 7ccqv 0 / 1 ContainerC reating 0 05 alibaba - log - controller - 84d8b6b8cf - nkrkx 0 / 1 ContainerC reating 0 0s ==> v1 / ServiceAcc ount NAME SECRETS AGE alibaba - log - controller 1 0s [SUCCESS] install helm package : alibaba - log - controller success .

Upgrade Log Service components

If you have installed Log Service components of an early version through Helm or DaemonSet, upgrade or migrate the components as follows.

Note:

To perform the following operations, first log on to the master node of your Kubernetes cluster of Alibaba Cloud Container Service. For information about the logon method, see *Connect to a Kubernetes cluster by using kubectl*.

Use Helm to upgrade Log Service components (recommended)

1. Run the following command to download the latest Helm package of Log Service components:

```
wget http://logtail - release - cn - hangzhou . oss - cn -
hangzhou . aliyuncs . com / kubernetes / alibaba - cloud - log .
tgz - 0 alibaba - cloud - log . tgz
```

2. Upgrade the components by using helm upgrade. The command is as follows:

```
helm get values alibaba - log - controller -- all > values
. yaml && helm upgrade alibaba - log - controller alibaba -
cloud - log . tgz -- recreate - pods - f values . yaml
```

Use DaemonSet to upgrade Log Service components

You can upgrade Log Service components by modifying the DaemonSet template. If your image account is acs, upgrade the image tag to the latest version that can be viewed in *Container Registry*. If your image account is acs, upgrade the image tag to the latest version that can be viewed in *Container Registry*.

Note:

- If upgrading the tag has not enabled a rolling update of Logtail, you must manually remove the Logtail pod to trigger a Logtail update.
- You need to check whether Logtail runs on all nodes, including Master nodes. If Logtail does not run on all nodes, you must set *tolerations* for Logtail.

```
toleration s :
- operator : " Exists "
```

For more information, see Latest Helm package configurations.

DaemonSet migrate

This upgrade method is applicable to the situation that you find the components are installed through the old DaemonSet when you check the Log Service component version. This method does not support configuring Log Service in Container Service. You can upgrade the components as follows: 1. At the end of the installation command, add a parameter which is the name of the project of Log Service used by your Kubernetes cluster.

For example, if the project name is k8s-log-demo and the cluster ID is c12ba2028cxxxxxx6939f0b, then the installation command is:

wget https :// acs - logging . oss - cn - hangzhou . aliyuncs . com / alicloud - k8s - log - installer . sh - 0 alicloud k8s - log - installer . sh ; chmod 744 ./ alicloud - k8s - log - installer . sh ; ./ alicloud - k8s - log - installer . sh -cluster - id c12ba2028c xxxxxxxx 6939f0b -- ali - uid 19 ********* 19 -- region - id cn - hangzhou -- log - project k8s - log - demo

- 2. After you complete the installation, log on the Log Service console.
- 3. After you complete the installation, log on the Log Service console.
- 4. In the Log service console, apply the history collection configuration of the project and Logstore to the new machine group k8s group -\${ your_k8s_c luster_id }.
- 5. After one minute, the history collection configuration is unbound from the history machine group.
- 6. When log collection is normal, you can delete the previously installed Logtail DaemonSet.



During the upgrade, some logs are duplicated. The CRD configuration management takes effect only for the configuration created by using CRD. The history configuration does not support the CRD management because the history configuration is created by using the non-CRD mode.

Configure Log Service when creating an application

Container Service allows you to configure Log Service to collect container logs when creating an application. Currently, you can use the console or a YAML template to create an application.

Create an application by using the console

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Application > Deployment in the left-side navigation pane, and then click Create by Image in the upper-right corner.

3. Configure Name, Cluster, Namespace, Replicas, and Type, and then click Next.

	Information	Container	Advanced	>	Done
Name:	tomcat				
	The name should be 1-64 characte	ers long, and can contain numbers, lower ca	se English letters and hyphens, but cannot	start with a hyphen.	
Cluster:	log-test	T			
Namespace :	default	T			
Doplicaci	2				
Replicas.					

4. On the Container page, select the Tomcat image and configure container log collection.

The following describes only configurations related to Log Service. For information about other application configurations, see *Create a deployment application by using an image*.

- 5. Configure Log Service. Click the + sign to create a configuration which consists of a Logstore name and a log path.
 - Logstore name: Specify a Logstore in which collected logs are stored. If your specified Logstore does not exist, the system automatically creates the Logstore in the project of Log Service with which the cluster is associated .

Note:

A Logstore name cannot contain underscores (_). You can use hyphens (-) instead.

Log path: Specify the path where logs to be collected reside. For example, use / usr / local / tomcat / logs / catalina . *. log to collect text logs of tomcat.

Note:

If you specify the log path as stdout, the container standard output and standard error output will be collected.

Each configuration is automatically created as a configuration for the corresponding Logstore. By default, the simple mode (by row) is used to collect
logs. To use more collection modes, log on go to the Log Service console, and enter the corresponding project (prefixed with k8s-log by default) and Logstore to modify the configuration.

2	Note, please ensure that cluster has deployed log plug-ins.					
	📀 Configuration					
	Log Store	Log path in the container (can be set to stdout)				
	access	/usr/local/tomcat/logs/catalina.*.log	•			
	catalina	stdout	•			

6. Custom tag. Click the + sign to create a new custom tag. Each custom tag is a keyvalue pair which will be added to collected logs. You can use a custom tag to mark container logs. For example, you can create a custom tag as a version number.

Custom Tag		
Name Of Tag	Value Of Tag	
release	1.0.0	•

7. When you complete all the configurations of the container, click Next in the upperright corner to perform further configurations. For more information, see *Create a*

deployment application by using an image.

Create an application by using a YAML template

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Application > Deployment in the left-side navigation pane, and then click Create by Template in the upper-right corner.
- 3. The syntax of the YAML template is the same as the Kubernetes syntax. To specify the collection configuration for the container, you need to use env to add collection configuration and custom tag for the container, and create corresponding volumeMounts and volumns. The following is a simple pod example:

```
apiVersion : v1
kind : Pod
metadata :
name : my - demo
spec :
containers :
- name : my - demo - app
```

```
image : ' registry . cn - hangzhou . aliyuncs . com / log -
service / docker - log - test : latest
   env :
  ######## Configure
                       environmen t
                                     variables
                                                ###########
     name : aliyun_log s_log - stdout
     value : stdout
     name : aliyun_log s_log - varlog
     value : / var / log /*. log
     name : aliyun_log s_mytag1_t ags
value : tag1 = v1
  ######### Configure
                      vulume
                               mount ############
   volumeMoun ts :
     name : volumn - sls - mydemo
     mountPath : / var / log
 volumes :
   name : volumn - sls - mydemo
   emptyDir : {}
```

- · Configure three parts in order based on your needs.
- In the first part, use environment variables to create your collection configurat ion and custom tag. All environment variables related to configuration are prefixed with aliyun_log s_.
- Rules for creating the collection configuration are as follows:

- name : aliyun_log s_ { Logstore name }
value : { log path }

In the example, create two collection configurations. The aliyun_log s_log

- stdout env creates a configuration that contains a Logstore named logstdout and the log path of stdout. The standard output of the container is collected and stored to the Logstore named log-stdout.

Note:

A Logstore name cannot contain underscores (_). You can use hyphens (-) instead.

Rules for creating a custom tag are as follows:

- name : aliyun_log s_ { a name without ' _ '} _tags

value : { Tag name }={ Tag value }

After a tag is configured, when logs of the container are collected, fields corresponding to the tag are automatically attached to Log Service.

• If you specify a non-stdout log path in your collection configuration, create corresponding volumnMounts in this part.

In the example, the / var / log /*. log log path is added to the collection configuration, therefore, the / var / log volumeMounts is added.

4. When you complete a YAML template, click DEPLOY to deliver the configurations in the template to the Kubernetes cluster to execute.

View logs

In this example, view logs of the tomcat application created in the console. After you complete the application configuration, logs of the tomcat application are collected and stored to Log Service. You can view your logs as follows:

- 1. Log on to the Log Service console.
- 2. Log on to the Log Service console.
- 3. In the console, select the project (k8s-log-{Kubernetes cluster ID} by default) corresponding to the Kubernetes cluster.
- 4. In the Logstore list, locate the Logstore specified in your configuration and click Search.

<		k8s-log-c3730910447504a	≇ Back to	Project List				Re	gion : China (Beijin:	ng)
Logstores 1	L	Logstores Learning Path Enc							its Create	
 LogHub - Collect 	Fr	nter a Logstore name to Search								
[Doc] Collection Hel						Lo	a Consumption Mode			
Logtail Config	Ŀ.	ogstore Name	Data Import	Monitor	Log Collection Mode	Log	goonsampuon noue		Actions	
Logtail Machine Grou						Consumption	Log Shipper	LogSearch		
 LogHub - Consume 	а	ccess	8	ĸ	Logtail Config (Manage) Diagnose More -	Preview More -	MaxCompute OSS	Search	Modify Delete	_
[Doc] Consumption H Consumer Group	a	udit- 3730910447504a6692ca160df6d9f24d		ĸ	Logtail Config (Manage) Diagnose More →	Preview More -	MaxCompute OSS	Search	Modify Delete	Contact L
 LogSearch/Analytics 	c	atalina		ĸ	Logtail Config (Manage) Diagnose More ↓	Preview More -	MaxCompute OSS	Search	Modify Delete	2
Saved Search Alarm	0	onfig-operation-log	8	Ľ	Logtail Config (Manage) Diagnose More -	Preview More +	MaxCompute OSS	Search	Modify Delete	

5. In this example, on the log search page, you can view the standard output logs of the tomcat application and text logs in the container, and you can find your custom tag is attached to log fields.

🗟 catalina						
🚯 catalina			O1Hour(Relative) ▼ St	hare Index Attributes	Save Search Sav	ved as Alarm
1 1.0.0					© 🕜 Searc	th & Analysis
100						
0						
16:31:31	16:40:3	0	16:49:20 16:58:20 17:07:20 17:	16:20	17:25:80	
Raw Logs Log	Reduce	LiveTail Graph	Log Entries:88 Search Status: I he results are accurate.	Display Content Colum	n Column Setting	gs [J]
Quick Analysis	<	Time 🛋	Content			
tag: 💿	1 Q	01-14 17:30:42	source: 172.16.2.130 tag:hostname: logtail-ds-nbr7f			
tag: ⊙			tagnode_np_: 192.168.0.21 tag:_node_name_: cn-beijing.i-2ze9pjyf12hdsyzeco41 topic :			
tag: ③	Container_jo_: 172.16.2.132 _container_name_: torocat					
tag: ③			_mage_name_: tomcatlatest _namespace_: default pod_name_: tomcatld59585f5f-5b8c4			_
tag: ③			_pod_uid_: 128ae02a-17de-11e9-be60-00163e036a23 _source_: stderr			Cont
tag: ③			_time_: 2019-01-14T09:30:42.086789817Z content: 14-Jan-2019 09:30:42.086 INFO [main] org.apache.catalina.startup.Catalina.start Server startu release: 4.0.0	ip in 1047 ms		act Us
tag: ③	2 0	01-14 17:30:42	source: 172.16.2.130			
tag: ③			tagnossname: logtal-ds-nor/f tag:_node_ip_: 192.168.0.21 tag:_node_name_:_co_belling.127a9niuf12bdsvzeco/11			
_containe						-

More information

- 1. By default, the system use the simple mode to collect your data, that is, to collect data by row without parsing. To perform more complex configurations, see the following Log Service documents and log on to the Log Service console to modify configurations.
 - Container text logs
 - Container stdout
- 2. Currently, Log Service uses plug-ins to collect the standard output logs of containers. You can configure more plug-ins to process collected logs further, such as to filter and extract fields.
- 3. In addition to configuring log collection through the console, you can also directly collect logs of the Kubernetes cluster through the CRD configuration. For more information, see *Configure Kubernetes log collection on CRD*.
- 4. For troubleshooting exceptions, see *Troubleshoot collection errors*.

1.13.4 A solution to log collection problems of Kubernetes clusters by using log-pilot, Elasticsearch, and Kibana

Requirements for logs of distributed Kubernetes clusters always bother developers . This is mainly because of the characteristics of containers and the defects of log collection tools.

- · Characteristics of containers:
 - Many collection targets: The characteristics of containers cause the number of collection targets is large, which requires to collect the container logs and container stdout. Currently, no good tool can collect file logs from containers dynamically. Different data sources have different collection softwares. However, no one-stop collection tool exists.
 - Difficulty caused by auto scaling: Kubernetes clusters are in the distributed mode. The auto scaling of services and the environment brings great difficulty to log collection. You cannot configure the log collection path in advance, the same as what you do in the traditional virtual machine (VM) environment. The dynamic collection and data integrity are great challenges.
- Defects of current log collection tools:
 - Lack the capability to dynamically configure log collection: The current log collection tools require you to manually configure the log collection method and path in advance. These tools cannot dynamically configure the log collection because they cannot automatically detect the lifecycle changes or dynamic migration of containers.
 - Log collection problems such as logs are duplicate or lost: Some of the current log collection tools collect logs by using the tail method. Logs may be lost in this way. For example, the application is writing logs when the log collection tool is being restarted. Logs written during this period may be lost. Generally, the conservative solution is to collect logs of 1 MB or 2 MB previous to the current log by default. However, this may cause the duplicate log collection.
 - Log sources without clear marks: An application may have multiple containers that output the same application logs. After all the application logs are collected to a unified log storage backend, you cannot know a log is generated on which application container of which node when querying logs.

This document introduces log-pilot, a tool to collect Docker logs, and uses the tool together with Elasticsearch and Kibana to provide a one-stop solution to log collection problems in the Kubernetes environment.

Introduction on log-pilot

Log-pilot is an intelligent tool used to collect container logs, which not only collects container logs and outputs these logs to multiple types of log storage backends

efficiently and conveniently, but also dynamically discovers and collects log files from containers.

Log-pilot uses declarative configuration to manage container events strongly and obtain the stdout and file logs of containers, which solves the problem of auto scaling. Besides, log-pilot has the functions of automatic discovery, maintenance of checkpoint and handle, and automatic tagging for log data, which effectively deals with the problems such as dynamic configuration, duplicate logs, lost logs, and log source marking.

Currently, log-pilot is completely open-source in GitHub. The project address is *https://github.com/AliyunContainerService/log-pilot*. You can know more implementation principles about it.

Declarative configuration for container logs

Log-pilot supports managing container events, can dynamically listen to the event changes of containers, parse the changes according to the container labels, generate the configuration file of log collection, and then provide the file to collection plug-in to collect logs.

For Kubernetes clusters, log-pilot can dynamically generate the configuration file of log collection according to the environment variable aliyun_log s_ \$ name = \$ path . This environment variable contains the following two variables:

- One variable is \$name, a custom string which indicates different meanings in different scenarios. In this scenario, \$name indicates index when collecting logs to Elasticsearch.
- The other is \$path which supports two input modes, stdout and paths of log files within containers, respectively corresponding to the standard output of logs and log files within containers.
 - Stdout indicates to collect standard output logs from containers. In this example, to collect Tomcat container logs, configure the label aliyun . logs
 . catalina = stdout to collect standard output logs of Tomcat.
 - The path of a log file within a container also supports wildcards. To collect logs within the Tomcat container, configure the environment variable aliyun_log
 s_access =/ usr / local / tomcat / logs /*. log . To not use the keyword aliyun, you can use the environment variable PILOT_LOG_PREFIX,

which is also provided by log-pilot, to specify the prefix of your declarative log configuration. For example, PILOT_LOG_ PREFIX : " aliyun , custom ".

Besides, log-pilot supports multiple log parsing formats, including none, JSON, CSV, Nginx, apache2, and regxp. You can use the aliyun_log s_ \$ name_format t =< format > label to tell log-pilot to use what format to parse logs when collecting logs.

Log-pilot also supports custom tags. If you configure $aliyun_log s_ $ name_tags$ =" K1 = V1 , K2 = V2 " in the environment variable, K1=V1 and K2=V2 are collected to log output of the container during the log collection. Custom tags help you tag the log generation environment for convenient statistics, routing, and filter of logs.

Log collection mode

In this document, deploy a log-pilot on each machine and collect all the Docker application logs from the machines.

Compared with deploying a logging container on each pod, the most obvious advantage of this solution is less occupied resources. The larger the cluster scale is, the more obvious the advantage is. This solution is also recommended in the community.

Prerequisites

You have activated Container Service and created a Kubernetes cluster. In this example, create a Kubernetes cluster in China East 1 (Hangzhou).

Step 1 Deploy Elasticsearch

- Connect to your Kubernetes cluster. For more information, see #unique_63 or #unique_168.
- 2. Deploy the resource object related to Elasticsearch first. Then, enter the following orchestration template. This orchestration template includes an elasticsearch-api

service, an elasticsearch-discovery service, and a status set of Elasticsearch. All of these objects are deployed under the namespace kube-system.

kubectl apply - f https :// acs - logging . oss - cn hangzhou . aliyuncs . com / elasticsea rch . yml

3. After the successful deployment, corresponding objects are under the namespace kube-system. Run the following commands to check the running status:

```
t svc , StatefulSe t - n = kube - system
CLUSTER - IP EXTERNAL - IP PORT (S) AGE
$
  kubectl
         tl get
TYPE (
 NAME
  svc / elasticsea rch - api ClusterIP
                                               172 . 21 . 5 . 134
                                                                      <
none > 9200 / TCP
                      22h
  svc / elasticsea rch - discovery ClusterIP
                                                      172 . 21 . 13 . 91
  < none > 9300 / TCP
                           22h
 NAME
         DESIRED
                    CURRENT
                               AGE
                                        3
                                            3
                                                 22h
  statefulse ts / elasticsea rch
```

Step 2 Deploy log-pilot and the Kibana service

1. Deploy the log-pilot log collection tool. The orchestration template is as follows:

```
kubectl apply - f https :// acs - logging . oss - cn -
hangzhou . aliyuncs . com / log - pilot . yml
```

2. Deploy the Kibana service. The sample orchestration template contains a service and a deployment.

kubectl apply - f https :// acs - logging . oss - cn hangzhou . aliyuncs . com / kibana . yml

Step 3 Deploy the test application Tomcat

After deploying the log tool set of Elasticsearch + log-pilot + Kibana, deploy a test application Tomcat to test whether or not logs can be successfully collected, indexed, and displayed.

The orchestration template is as follows:

```
apiVersion :
              v1
kind : Pod
metadata :
  name : tomcat
  namespace : default
  labels :
    name : tomcat
spec :
  containers :
   image : tomcat
    name : tomcat - test
   volumeMoun ts :
     mountPath : / usr / local / tomcat / logs
      name : accesslogs
   env :
```

```
name : aliyun_log s_catalina
   _
      value : " stdout " ## Collect
                                      standard
                                                output
                                                         logs .
      name : aliyun_log s_access
      value : "/ usr / local / tomcat / logs / catalina . *. log "
             log files within
  Collect
                                  the
                                         container .
##
 volumes :
     name :
            accesslogs
     emptyDir : {}
```

The Tomcat image is a Docker image that both uses stdout and file logs. In the preceding orchestration, the log collection configuration file is dynamically generated by defining the environment variable in the pod. See the following descriptions for the environment variable:

```
• aliyun_log s_catalina = stdout indicates to collect stdout logs from the container.
```

- aliyun_log s_access =/ usr / local / tomcat / logs / catalina .
 - *. log indicates to collect all the log files whose name matches catalina
 - . *. log under the directory / usr / local / tomcat / logs / from the container.

In the Elasticsearch scenario of this solution, the \$ name in the environment variable indicates index. In this example, \$ name is catalina and access.

Step 4 Expose the Kibana service to Internet

The Kibana service deployed in the preceding section is of the NodePort type, which cannot be accessed from the Internet by default. Therefore, create an Ingress in this document to access the Kibana service from Internet and test whether or not logs are successfully indexed and displayed.

 Create an Ingress to access the Kibana service from Internet. In this example, use the simple routing service to create an Ingress. For more information, see *#unique_169*. The orchestration template of the Ingress is as follows:

```
extensions / v1beta1
apiVersion :
 kind : Ingress
metadata :
  name : kibana - ingress
  namespace : kube - system
                              # Make
                                       sure
                                              the
                                                    namespace
                                                                is
                   that
                         of
                               the
                                     Kibana
                                              service .
  the
       same
              as
 spec :
  rules :
    http :
      paths :
        path : /
        backend :
```

```
kibana # Enter
         serviceNam e :
                                             the
                                                    name
                                                           of
                                                                the
Kibana
         service .
                          80 # Enter
         servicePor
                     t :
                                         the
                                                port
                                                       exposed
                                                                 by
the
      Kibana
               service .
```

2. After the Ingress is successfully created, run the following commands to obtain the access address of the Ingress:

\$ kubectl get ingress - n = kube - system
NAME HOSTS ADDRESS PORTS AGE
shared - dns * 120 . 55 . 150 . 30 80 5m

- 3. Access the address in the browser as follows.
- 4. Click Management in the left-side navigation pane. Then, click Index Patterns > Create Index Pattern. The detailed index name is the \$ name variable suffixed with a time string. You can create an index pattern by using the wildcard *. In this example, use \$ name * to create an index pattern.

You can also run the following commands to enter the corresponding pod of Elasticsearch and list all the indexes of Elasticsearch:

Find \$ kubectl get pods - n = kube - system the correspond ing pod of Elasticsea rch . exec - it \$ kubectl elasticsea rch - 1 bash # Enter а Elasticsea rch . pod of curl ' localhost : 9200 / _cat / indices ? v ' ## List all indexes . the index docs . health uuid pri rep docs . count status pri . store . deleted store . size size open . kibana x06jj19PS4 Cim6Ajo51P Wg 4 1 1 green 53 . 6kb 26 . 8kb access - 2018 . 03 . 19 txd3tG - NR6 open green 143 0 823 . 5kb 411 . 7kb guqmMEKKzE w 5 1 catalina - 2018 . 03 . 19 ZgtWd16FQ7 qqJNNWXxFP open green 915 . 5kb 1 457 . 5kb c0 5 143 0

5. After successfully creating the indexes, click Discover in the left-side navigation pane, select the created index and the corresponding time range, and then enter the related field in the search box to query logs.

Then, you have successfully tested the solution to log collection problems of Alibaba Cloud Kubernetes clusters based on log-pilot, Elasticsearch, and Kibana. By using this solution, you can deal with requirements for logs of distributed Kubernetes clusters effectively, improve the Operation and Maintenance and operational efficiencies, and guarantee the continuous and stable running of the system.

1.13.5 Configure Log4jAppender for Kubernetes and Log Service

Log4j is an open-source project of Apache, which consists of three important components: log level, log output destination, and log output format. By configurin g Log4jAppender, you can set the log output destination to console, file, GUI component, socket server, NT event recorder, or UNIX Syslog daemon.

This document introduces how to configure a YAML file to output Alibaba Cloud Container Service Kubernetes cluster logs to Alibaba Cloud Log Service, without modifying the application codes. In this document, deploy a sample API application in the Kubernetes cluster for demonstration.

Prerequisites

· You have activated Container Service and created a Kubernetes cluster.

In this example, create a Kubernetes cluster in the region of China East 1 (Hangzhou).

• Enable AccessKey or Resource Access Management (RAM). Make sure you have sufficient access permissions. Use the AccessKey in this example.

Step 1 Configure Log4jAppender in Alibaba Cloud Log Service

- 1. Log on to the Log Service console.
- 2. On the Project List page, click Create Project in the upper-right corner. Complete the configurations and then click Confirm to create the project.

In this example, create a project named k8s-log4j and select the same region (China East 1 (Hangzhou)) as the Kubernetes cluster.

Note:

Generally, create a Log Service project in the same region as the Kubernetes cluster. When the Kubernetes cluster and Log Service project are in the same region, log data is transmitted by using the intranet, which saves the Internet bandwidth cost and time of data transmission because of different regions, and implements the best practice of real-time collection and quick query.

Create Project	\times
* Project Name: k8s-log4j	
Description: k8s-log4j demo	
<>"'\ are not supported, and the description cannot exceed 512 characters. * Region: China East 1 (Han	
Confirm	Cancel

- 3. After being created, the project k8s-log4j is displayed on the Project List page. Click the project name.
- 4. The Logstore List page appears. Click Create in the upper-right corner.

Logstores	Logstore List					Endpoint List	Create
 LogHub - Collect 							
[Doc] Collection	Searching by logstore nam	Search					
Logtail Config	Logstore Name	Data Import	Monitor Log Collection Mode	Log Consumption Mode			Action
Logtail Machine	Logstore nume	Wizard	Homor Log conceron Hore	LogHub	LogShipper	LogSearch	Action
 LogHub - Consume 							
[Doc] Consumpt			You don't have any Logstore, pleaseCreate				

5. Complete the configurations and then click Confirm.

Create Logstore		\times
* Logstore Name: Logstore Attributes	k8s-logstore	
* WebTracking:	WebTracking supports the collection of various types of access logs in web browsers or mobile phone apps (iOS/Android). By default, it is disabled. (Help Link)	
* Data Retention Time:	30 Data retention time for LogHub and LogSearch is unified. The data lifecycle is determined by the LogHub setting (the unit is in days).	
* Number of Shards:	2 ▼ What is shard?	
* Billing:	Refer to pricing	
	Confirm Can	cel

In this example, create a Logstore named k8s-logstore.

6. Then, a dialog box asking you to use the data import wizard appears.

Create	×	
0	You have created a logstore, use the data import wizard to learn ab out collecting logs, analysis and more.	
	Data Import Wizard Cancel	

7. Click Data Import Wizard. In the Select Data Source step, select log4jAppender under Other Sources and then complete the configurations as instructed on the page.

Use the default configurations in this example. Configure the settings according to the specific scenarios of log data.

Other Sources							
	Text	SysLog	Web Tracking	API/SDK	🕑 log4jAppender	Producer	
							Next

Step 2 Configure Log4jAppender in the Kubernetes cluster

In this example, use the sample YAML files *demo-deployment* and *demo-service* for demonstration.

1. Connect to your Kubernetes cluster.

For more information, see Access Kubernetes clusters by using SSH or Connect to a Kubernetes cluster by using kubectl.

2. Obtain the *demo* - *deployment* . *yaml* file and configure the environment variable JAVA_OPTS to collect logs from the Kubernetes cluster.

The sample orchestrat ion of the demo – deployment. yaml file is as follows:

```
apiVersion : apps / v1beta2
kind : Deployment
metadata :
  name : log4j - appender - demo - spring - boot
  labels :
    app : log4j - appender
spec :
  replicas :
             1
  selector :
    matchLabel s :
      app : log4j - appender
  template :
    metadata :
    labels :
     app :
            log4j – appender
  spec :
    containers :
      name : log4j - appender - demo - spring - boot
      image : registry . cn - hangzhou . aliyuncs . com /
jaegertrac ing / log4j - appender - demo - spring - boot : 0 . 0 .
2
      env :
     - name : JAVA_OPTS ## Note
```

```
value : "- Dproject ={ your_proje ct } - Dlogstore ={
your_logst ore } - Dendpoint ={ your_endpo int } - Daccess_ke
y_id ={ your_acces s_key_id } - Daccess_ke y ={ your_acces
s_key_secr et }"
    ports :
    - containerP ort : 8080
```

Wherein:

- - Dproject : The name of the used Alibaba Cloud Log Service project. In this example, it is k8s-log4j.
- - Dlogstore : The name of the used Alibaba Cloud Log Service Logstore. In this example, it is k8s-logstore.
- Dendpoint : The service endpoint of Log Service. You must configure your service endpoint according to the region where the Log Service project resides. For more information, see *Service endpoint*. In this example, it is cnhangzhou.log.aliyuncs.com.
- - Daccess_ke y_id : Your AccessKey ID.
- - Daccess_ke y : Your AccessKey Secret.
- 3. Run the following command in the command line to create the deployment:

kubectl create - f demo - deployment . yaml

4. Obtain the *demo* - *service* . *yaml* file and run the following command to create the service.

No need to modify the configurations in the demo - service . yaml file.

kubectl create - f demo - service . yaml

Step 3 Test to generate Kubernetes cluster logs

You can run the kubectl get command to view the deployment status of the resource object. Wait until the deployment and the service are successfully deployed. Then, run the kubectl get svc command to view the external access IP of the service, that is, the EXTERNAL-IP.

```
$ kubectl get svc
NAME TYPE CLUSTER - IP EXTERNAL - IP PORT (S) AGE
log4j - appender - demo - spring - boot - svc LoadBalanc er 172
. 21 . XX . XX 120 . 55 . XXX . XXX 8080 : 30398 / TCP 1h
```

In this example, test to generate Kubernetes cluster logs by running the login command, wherein, K8S_SERVIC E_IP is the EXTERNAL - IP .



See GitHub log4j-appender-demo to view the complete collection of APIs.

```
curl http ://${ K8S_SERVIC E_IP }: 8080 / login ? name = bruce
```

Step 4 View logs in Alibaba Cloud Log Service

Log on to the Log Service console.

Click the project name and click Search at the right of the Logstore k8s-logstore to view the output logs of the Kubernetes cluster.



The output content of the log corresponds to the preceding command. This example demonstrates how to output the logs of the sample application to Alibaba Cloud Log Service. By completing the preceding steps, you can configure Log4JAppender in Alibaba Cloud and implement advanced functions such as collecting logs in real time , filtering data, and querying logs by using Alibaba Cloud Log Service.

1.14 Monitoring management

1.14.1 Deploy the Prometheus monitoring system

Prometheus is an open source monitoring tool for cloud native applications. This topic describes how to deploy the Prometheus monitoring system by using Alibaba Cloud Container Service for Kubernetes.

Background information

A monitoring system monitors the following two types of objects:

• Resource, namely, the resource usage of a node or application. The monitoring system of Container Service for Kubernetes monitors node resource usage, cluster resource usage, and pod resource usage.

 Application, namely, internal metrics of an application. For example, The monitoring system collects statistics regarding the number of online users that use an application in real time, and performs service-level monitoring and alarming for the application by exposing ports.

The following are the objects monitored in a Kubernetes cluster:

- System components, which are built-in components of the Kubernetes cluster, such as apiserver, controller-manager, and etcd.
- · Static resource entities, which include node resource status and kernel events.
- Dynamic resource entities, which are abstract workload entities of Kubernetes, such as deployment, DaemonSet, and pods.
- Customized application objects, which includes the data and metrics that require customization within an application.

To monitor system components and static resource entities, you need to specify monitoring methods for them in the configuration file.

To monitor dynamic resource entities, we recommend that you deploy the Prometheus monitoring system.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have connected to the Master node so that you can view node labels and other information. For more information, see *Connect to a Kubernetes cluster by using kubectl*.

Deploy the Prometheus monitoring system

1. Run the following command to download the prometheus-operator code:

```
git clone https :// github . com / AliyunCont ainerServi ce /
prometheus - operator
```

2. Run the following command to deploy the Prometheus monitoring system:

Note:

Some Prometheus components may fail to be deployed when you run this command for the first time because Prometheus components require a specific

sequence to be deployed. If any exceptions occur during your first deployment, you need to run the command again.

```
cd prometheus - operator / contrib / kube - prometheus
kubectl apply - f manifests
```

3. Run the following command to set the access method for Prometheus:

```
kubectl -- namespace monitoring port - forward svc / prometheus - k8s 9090
```

- 4. View the deployment result
 - a. To view Prometheus, access localhost : 9090 in a browser.

Note:

By default, Prometheus cannot be accessed through the Internet. You must use your local proxy to access it.

b. Select Targets under the Status menu to view all collection tasks.

Prometheus	Alerts	Graph	Status -	Help
C Enable query history			Runtime 8	Build Information
Expression (pre	ss Shift+Fr	ter for new	Command	-Line Flags
Expression (pre	Expression (press Smit+Enter for ner		Configura	ion
Execute		ric at curso	Rules	
Execute	Graph Console Console		Targets	
Graph Conse			Service Di	scovery
Element				Value
no data				
				Remove Graph
Add Graph				

If the status of all tasks is UP, all collection tasks are running properly.

Prometheus Alerts Graph Sta	atus - Help			
Targets				
alertmanager-main (3/3 up) 🛚	now less			
Endpoint	State	Labels	Last Scrape	Error
http:// :9093/metrics	UP	endpoint="web" instance=":9093" namespace="monitoring" i pod="alertmanager-ma in-2" i service="alertmanager-main"	23.222s ago	
http://	UP	endpoint="web" instance=' :9093" namespace="monitoring" i pod="alertmanager-ma in-1" i service="alertmanager-main"	27.703s ago	
http://	UP	endpoint="web" instance="9093" namespace="monitoring" pod="alertmanager-ma in-0" iservice="alertmanager-main"	16.792s ago	
apiserver (3/3 up) show less				
Endpoint	State	Labels	Last Scrape	Error
https://	UP	endpoint="https" instance="6443" namespace="default" service="kubernetes"	26.006s ago	

View and display data aggregation

1. Run the following command to access Grafana:

```
kubectl -- namespace monitoring port - forward svc / grafana
3000
```

2. Access localhost : 3000 in your browser and then select a dashboard to view data aggregation.

Note:

The default user name and password are both admin.



View alerting rules and set alert silencing

• View alerting rules

Access localhost : 9090 in your browser and click the Alerts menu to view the current alerting rules.

- Red: indicates that an alert is triggered.
- Green: indicates the normal status.

Prometheus Alerts Graph Status - Help
Alerts
Show annotations
CPUThrottlingHigh (12 active)
DeadMansSwitch (1 active)
KubeControllerManagerDown (1 active)
KubePodNotReady (1 active)
KubeSchedulerDown (1 active)
KubeDeploymentReplicasMismatch (1 active)
KubePodCrashLooping (1 active)
AlertmanagerConfigInconsistent (0 active)
AlertmanagerDown (0 active)
AlertmanagerFailedReload (0 active)
KubeAPIDown (0 active)
KubeAPIErrorsHigh (0 active)
KubeAPIErrorsHigh (0 active)
KubeAPILatencyHigh (0 active)

· Set alert silencing

Run the following command, open localhost : 9093 in your browser, and select Silenced to set alert silencing:

kubectl -- namespace monitoring port - forward svc / alertmanag er - main 9093

Filter Group	Receiver: All Silenced Inhibited
	+
Custom matcher, e.g. env="production"	
alertname="CPUThrottlingHigh" +	
08:14:17, 2018-10-29 + Info 🗠 Source 🔀 Silence	
severity="werping" + prometheus="monitoring/k9s" + pod_pame="kube_state_matrice_775dd59046.9yzif"	+ namespace="monitoring" +

1.14.2 Group-based monitoring and alarms

Alibaba Cloud Container Service is interoperable with CloudMonitor to enable groupbased monitoring and alarms.

Prerequisites

- Create a Kubernetes cluster if you do not have one.
- The Kubernetes version must be 1.8.4 or later. Otherwise, you must first upgrade the cluster.

Context

In the Operation & Maintenance (O&M) of IT infrastructure, monitoring and alarms facilitate daily O&M, system monitoring, troubleshooting, and debugging, and guarantee the reliability and security of O&M.

The traditional container monitoring solution that uses a statically configured monitoring agent or a centralized server for monitoring and alarms may not be suitable for the Kubernetes scenario because it can cause some problems . For example, the information required to identify the monitoring objects is missing because containers are mostly scheduled in the resource pool whereas the monitoring agent is deployed on the host. Also, containers have shorter lives than applications. The monitoring and alarm rules, and such monitoring data as ReplicaSet and Deployment for a single container cannot be used for the correspond ing application.

Alibaba Cloud Container Service for Kubernetes is deeply integrated with CloudMonit or to use application groups to unify the monitoring objects and metrics. In addition , CloudMonitor of Alibaba Cloud is equipped with many functions and custom tools, which provide you with the best practice to monitor your Kubernetes resources and manage the alarms.

Procedure

- 1. Log on to the CloudMonitor console.
- 2. In the left-side navigation pane, click Application Groups. The Kubernetes groups with cluster IDs are displayed.

CloudMonitor	Application Groups							Creat	e Group
Overview	Type Type Type Type Type Type Type Type			Search					
Dashboard	Group Name / Group ID	Health Status 🔞	Туре	Total Server Number 🔞	Resource Types 🕜	Unhealthy Instances	Created At		Actions
Host Monitoring	g20 / 23939	•	Custom	2	1	0	2017-09-08 13:45:22	Manage	More +
Custom Monitoring	k8s-rief i fondor in i noorfeel intern- master / 64553	•	kubernetes	3	2	0	2018-04-24 10:16:42	Manage	More +
Cloud Service Mo	k8s-million and a 515 ft field sector worker / 64554	•	kubernetes	1	1	0	2018-04-24 10:16:42	Manage	More +
 Alarms 							Total 3	10 T « < 1	>

3. Click the Group Name to go to the group details page. You can view the resources contained in the group. For example, in a Master group of Kubernetes, you can see such resources as Elastic Compute Service (ECS) instances and Server Load Balancer (SLB) instances.

Kubernetes has two types of nodes: Worker nodes and Master nodes. Master nodes generally contain management and control applications and the resources are required to be highly robust. Worker nodes are generally responsible for scheduling pods and the overall requirement on the resources focuses on scheduling capability. When you create a group, Container Service automatically creates two resource groups, a Master group and a Worker group. The Master group includes the Master nodes and the related SLB instances. The Worker group includes all the Worker nodes.

<	k8s-chef Kreiter ef 12 Hold-Wert Bield-master + Back to Application Group
Group Resource	
 Dashboards 	C Refresh Create Alam Rule Modify Group Delete Group Apply Template to Group
Fault List	
Availability Monit	
Custom Monitoring	
Alarm Logs	Server Load Balancer
Alarm Rule	
Ξ	ECS
	View Charts • by Line 3 Charts • Add Monitoring Chart

4. You can view the details of other cloud products, such as SLB, in the group.

<	ľ	k8s-dieler".Slotbooler".all [5dl79ed.84e4	-master ± Back	to Application	Group	1h 2h 6h 12h 1days	3days 7days 1	4days 2018-04-25	5 03:27:33 - 2018-04-25 🗮									
Group Resource Dashboards Fault List		Δ Ξ				Create Alarm Rule	Modify Group	Delete Group	Apply Template to Group									
Availability Monit		ECS (3)		Enter EC	Server Load Balancer (2)	Enter Server Load Balancer Console												
Custom Monitoring		Instance Name	Health Status 📀	Resource D	e Instance Name	Health Status 🖉 Resource Description												
Alarm Logs Alarm Rule		tano har-es- cilitari Microsoftania (Scill/Rect. 84644- menter)	©ок	199.159.22	b- Indcovellage.wi884	Ок												
=	=	Nde-fer-ca- caciety sociales we final a second	ок	241.144.12	lb- tudelintectedskabite	OK Total 2 , Per page10 < 1 →												
		klis-for-co- cdol/1905/bco4/1s803/bc184e44- mader1	©ок	192.268.22	1.163													
												Total 3 , Per	page10 <	1 >				
	`	/iew Charts				View All	t	oy Line 3 Charts 🔹	Add Monitoring Chart									

5. In the left-side navigation pane, click Dashboards to view the detailed monitoring metrics of each cloud product in the group.



- 6. In the left-side navigation pane, click Alarm Rule. A list of existing alarm rules in the group is displayed. By default, the health of the core components of all nodes in the Mater group is checked.
 - a. Click Create Alarm Rule to create an alarm rule for the group according to your business requirements.

<	k8	8s-ther Concorder at 1800 and 1944 master 🔹 Back to Application Group							
Group Resource • Dashboards	Ent	ter the alarm rule name.			Search	Create Alarm Rule	Modify Group D	elete Group	Apply Template to Group
Fault List		Rule Name	Status (All) +	Enable	Dimensions (All) +	Alarm Rules	Product Name (All)	Notification Contact	Actions
Custom Monitoring	٥	kube-controller- TelnetStatus.Value	©ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Value>400 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable
Alarm Rule		kube-apiserver- TelnetStatus.Value	⊘ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Value>400 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable
=		kubelet- TelnetLatency.Average	©ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Average>10000 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable
		kube-proxy- TelnetStatus.Value	©ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Value>400 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable
		kube-schedule- TelnetLatency.Average	⊘ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Average>10000 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable
		kubelet- TelnetStatus.Value	ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Value>400 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	View Modify Disable

- b. On the displayed Create Alarm Rule page, set the alarm rules.
 - Select the related resource, such as ECS.
 - Select whether to use a template to create the alarm rule. If yes, select an alarm template from the Select Template drop-down list. You can also click

Create Alarm Template to create a new custom alarm template. For more information, see .

• Set the notification method. For example, you can know the Kubernetes cluster status through DingTalk, email, and SMS.

Create Alarm Rule 🔹 Ba	ck to										
Related Resource											
Products :	k8s -										
Resource Range :	Application Group										
Group Name :	k8s-c2174629b4ea049d887da6e71d										
2 Set Alarm Rules	2 Set Alarm Rules										
Use Template :	● Yes ○ No										
Select Template :	常用基础模板 🗙		• Create	Alarm Template							
	常用基础模板_cpu_total	Host.cpu.total	• 1m •	it alar 🔻	> •	90	%				
	常用基础模板_diskusage_utili	Host.disk.utilization	• 1m •	it alar 🝷	> *	90	%				
	常用基础模板_memory_usedu	Host.mem.usedutilization I	• 1m •	it alar 🔻	> •	90	%				
	常用基础模板_InternetOutRat	Internet Outbound Bandwi	• 1m •	it alar 🝷	> -	90	%				
	常用基础模板_agent_heartbe	Heartbeat not Detected	• 1m •				•				
Mute for :	24h	- 0									
Effective Period :	00:00 • To: 23:59	•									
3 Notification Me	thod										
Notification	Contact Group	All Selected	Groups 1 count	Д	JI						
Contact :	Search	Q 云账号	报警联系人								
		-									
		4									
	Quickly create a contact	group									

c. Click Confirm. The created alarm rule is displayed on the Alarm Rule page.

<	k8s-clideft.K2cdbco4eft	ut51.5d87Net30NeH4-m	aster 🔹 B	ack to Application Group	1h 2h 6h 12h 1days 3d	ays 7days 14days	2018-04-2	5 19:26:17 - 2018-04-2 菌
Group Resource	Enter the alarm rule nar	ne.	Sea	ch	Create Alarm Rule	Modify Group De	ete Group	Apply Template to Group
 Dashboards Fault List 						Product Name (All)	Notification	
Availability Monit	Rule Name	Status (All) -	Enable	Dimensions (All) +	Alarm Rules	•	Contact	Actions
Custom Monitoring Alarm Logs	test_CPUUtilization	⊘ок	Enabled	Group Dimension:kks- c8def192cdbcc4af1a515d07fed184e44- master	1minute CPU Usage Average>=90 % it alarms 1 times To alarm	ECS	Default Co View	Modify Disable Delete
Alarm Rule	 kube-controller- TelnetStatus.Value 	⊘ок	Enabled	Group Dimension:k8s- c8def192cdbcc4af1a515d07fed184e44- master	1minute Value>400 it alarms 3 times To alarm	CloudMonitor- Availability Monitoring	Default Co View	Modify Disable

What's next

More features are provided to meet your resource monitoring requirements, such as fault list, event monitoring, availability monitoring, and log monitoring. You can find them in the left-side navigation pane.

1.14.3 Integration and usage with CloudMonitor

Prerequisites

Check whether alicloud - monitor - controller has been deployed in the kube - system namespace. If not, upgrade the version of the cluster.

Procedure

1. Log on to the Container Service console.

2. Under Kubernetes, click Deployment in the left-side navigation pane.

3. Select the target deployment, click Monitor on the right. You can also click Monitor on the Deployment page of the built-in kubernetes dashboard.

Container Service - Kubernetes 🔻	Deployment	_		Refi	resh Create by image Create by template
Overview	Clusters k8s-cluster v Name	ipace kube-system 🔻 🙎			
▼ Clusters	Name	Tag	PodsQuantity	Time Created	3 Action
Clusters	alicloud-disk-controller	app:alicloud-disk-controller	1/1	06/27/2018,17:53:56	Details Edit Monitor More -
Nodes	alicloud-monitor-controller	k8s-app:alicloud-monitor-controller task:monitoring	1/1	06/28/2018,14:06:30	Details Edit Monitor More 🗸
Namespace	default-http-backend	app:default-http-backend	1/1	06/27/2018,17:53:56	Details Edit Monitor More -
Application	heapster	k8s-app:heapster task:monitoring	1/1	06/27/2018,17:53:56	Details Edit Monitor More -
Deployment	kube-dns	k8s-app:kube-dns	1/1	06/27/2018,17:53:51	Details Edit Monitor More -
Cluster	E Workloads D Deployments	eployments			+ CREATE
Namespaces	Name 🗢	Labels	Pods	Age 🗢	Images
Nodes Persistent Volumes	monitoring-grafana	k8s-app: grafana task: monitoring	1/1	2018-07-10 16:04:34	registry.cn-hangzhou.aliyuncs.com/acs/
Roles	alicloud-monitor-cont	roller k8s-app: alicloud-monitor-co	ontroller 1 / 1	2018-06-28 14:06:30	registry.cn-hangzhou.aliyuncs.com/acs/
Namosaasa	Stiller-deploy	app: helm name: tiller	1 / 1	2018-06-27 17:53:58	registry-vpc.cn-hangzhou.aliyuncs.com/
kube-system	alicloud-disk-controll	app: alicloud-disk-controller	1/1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyur Monitoring
Overview	leapster	k8s-app: heapster task: monitoring	1/1	2018-06-27 17:53:56	registry.cn-hangzhou.aliyuncs.c Scale
Workloads Cron Jobs	monitoring-influxdb	k8s-app: influxdb task: monitoring	1/1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyur
Daemon Sets	orginx-ingress-control	er app: ingress-nginx	1 / 1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyur
Deployments 2	default-http-backend	app: default-http-backend	1 / 1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyuncs.com/

In this case, you jump to the corresponding Application group details page of CloudMonitor.

<	k8s-ct → Back to Application Group							
	k8s Customchart							
Group Resource	The 3h 6h 10h 1days 3days 7days 1ddays 🗎 Auth Refrech (Chart relevance)							
 Dashboards 	Dimension: Group Name Instances							

4. Application group supports monitoring in two dimensions: group and instance.

<	k8s-c07ea7a5639924917af756e7d3f95e0d5-master	9
-	ECS Server Load Balancer Customchart	
Group Resource	Ih 3h 6h 12h Idays 3days 7days 14days Auto Refresh :	Chart relevance : Chart releva
Dashboards	Dimension Group Name Instances	
Fault List		
Event Monitor	Disk Metric	
Availability Monitor	Disk bps(Bytes/s)	Disk iops(Count/Second)
Log Monitoring	116.13K	23.85
Custom Monitoring		20.00
Alarm Logs	48.83K	10.00
Alarm Rule	8.02K 16:13:00 16:26:40 16:43:20 17:00:00 17:11:00	1.83 16:13:00 16:26:40 16:43:20 17:00:00 17:11:0
=	(ECS) Group DimensionDisk Read BPS—Average—k8s-c	(ECS) Group DimensionDisk Read IOPS—Average—k8s
	• (ECS) Group DimensionDisk Write BPS—Average—k8s	 (ECS) Group DimensionDisk Write IOPS—Average—k8s
	100 c07co7aE620024017aF7E6a7d2f0Ea0dE lasha autam D ×	
<	k8c Outemplant	
Group Rocourco		
Group Resource	1h 3h 6h 12h 1days 3days 7days 14days 🗮 Chart relevance :	◯ C Refresh
 Dashboards 	Dimension: Group Name Instances allcloud-disk-controller-7f8c59bbb7-9kbgv	•
Fault List	InstanceId	×
Event Monitor	CPU Usage(%) alicloud-disk-controller-7f8c59bbb7-9kbgv 2	Net rate(bps)
Availability Monitor	100.00	
Log Monitoring	10/00	9.02
Custom Monitoring	50.00	
Alarm Logs		07-01 07-02 6.00 MW W Y Y W Y W Y Y Y W Y Y Y W Y Y Y W Y Y Y W Y Y Y W Y Y Y W Y Y Y Y W Y Y Y W Y Y Y Y W Y Y Y W Y Y Y Y Y W Y
Alarm Rule	0.00	aue 06-28 07-01 07-02 Average e e e e e
Œ	pod.cpu.usage_rate Average	pod.network.rx_rate—Average

5. For alarm settings, the index of group level starts with group , and the instance level index starts with pod .

Products :	k8s -		
Resource Range :	Application Group	When selecting an application group, you can use an	alarm template. Click View alarm template best practices.
Group Name :	k8s-c07ea7a5639924917af756e7d3f95e 🔻		
Set Alarm Rule	5		
Use Template :	○ Yes ● No		
Alarm Rule :	usage of Deployment]	No Data
Rule Describe :	group.cpu.usage_rate 5mins	▼ Total ▼ >= ▼ 80	
+Add Alarm	Ri group.cpu.limit		
Mute for :	group.cpu.request		
Triggered when threshold is exceeded for :	group.cpu.usage group.cpu.usage_rate group.disk.io_read_bytes		The line chart indicates the average aggregate value trend of instances under the application

Upgrade cluster version

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment in the left-side navigation pane to enter the Deployment List page. Click Create by template in the upper-right corner.

Container Service - Kubernetes +	Deployment				Refresh Create by image Create by template
Overview	Clusters k8s-cluster v Namespace	default 🔻			2
 Clusters 	Name	Tag	PodsQuantity	Time Created	Action
Clusters	new-nginx	run:new-nginx	1/1	07/11/2018,17:00:19	Details Edit Monitor More -
Nodes	nginx-deployment-basic	app:nginx	2/2	07/11/2018,18:00:12	Details Edit Monitor More -
Volumes	old-nginx	run:old-nginx	2/2	07/11/2018,17:00:19	Details Edit Monitor More -
Namespace	test-nginx-deployment	app:test-nginx	1/1	07/12/2018,15:34:56	Details Edit Monitor More -
Deployment					

3. Select the target cluster, kube-system namespace, and use the following sample template. Then click Create.



Replace REGION and CLUSTER_ID with your actual cluster information, and redeploy heapster yaml template.

Clusters	k8s-cluster v
Namespace	kube-system v
Resource Type	Resource - basic Deployment
Template	<pre>1 aplVersion: extensions/v1beta1 2 kind: DepLoyment 3 metadata: 1 name: heapster 1 name: heapster 1 task: monitoring 1 k8s-app: heapster 1 amotations: 1 task: monitoring 1 k8s-app: heapster 1 amotations: 1 scheduler.alpha.kubernetes.io/critical-pod: '' 5 spec: 1</pre>
	DEPLOY

An example of heapster template is as follows. If you have an earlier version of the heapster in the cluster, you can log on to the Kubernetes cluster and run the

```
kubectl apply - f xxx . yaml command to upgrade it.
```

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
    name : heapster
    namespace : kube - system
spec :
    replicas : 1
    template :
        metadata :
        labels :
        task : monitoring
        k8s - app : heapster
        annotation s :
            scheduler . alpha . kubernetes . io / critical - pod : ''
        spec :
```

```
serviceAcc ount : admin
containers :
        - name : heapster
        image : registry . ## REGION ##. aliyuncs . com / acs /
heapster - amd64 : v1 . 5 . 1 . 1
        imagePullP olicy : IfNotPrese nt
        command :
        - / heapster
        --- source = kubernetes : https :// kubernetes . default
        --- historical - source = influxdb : http :// monitoring -
influxdb : 8086
        --- sink = influxdb : http :// monitoring - influxdb :
8086
        --- sink = socket : tcp :// monitor . csk . ## REGION ##.
aliyuncs . com : 8093 ? clusterId =## CLUSTER_ID ##& public = true
```

The example layout of alicloud-monitor-controller is as follows. Run the kubectl

create - f xxx . yaml command to deploy alicloud-monitor-controller.

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
  name : alicloud - monitor - controller
  namespace : kube - system
spec :
  replicas : 1
  template :
   metadata :
     labels :
        task : monitoring
       k8s - app : alicloud - monitor - controller
     annotation s :
       scheduler . alpha . kubernetes . io / critical - pod : ''
   spec :
     hostNetwor k : true
     toleration s :
     - effect : NoSchedule
       operator : Exists
       key : node - role . kubernetes . io / master
     - effect : NoSchedule
       operator : Exists
       key : node . cloudprovi der . kubernetes . io /
uninitiali zed
     serviceAcc ount : admin
     containers :
      name : alicloud - monitor - controller
       image : registry . ## REGION ##. aliyuncs . com / acs /
alicloud - monitor - controller : v1 . 0 . 0
       imagePullP olicy : IfNotPrese nt
       command :
       - / alicloud - monitor - controller
      - agent
      - -- regionId =## REGION ##
       - -- clusterId =## CLUSTER_ID ##
      - -- logtostder r
```

- - - v = 4

4. Go to the Kubernetes console. In the kube-system namespace, you can see that the two deployments are running, and the upgrade is complete.

<		S				+ CREATE
Cluster	Deployments					Ŧ
Namespaces	Name ≑	Labels	Pods	Age ≑	Images	
Nodes Persistent Volumes	Monitoring-grafana	k8s-app: grafana task: monitoring	1/1	2018-07-10 16:04:34	registry.cn-hangzhou.aliyuncs.com/acs/	:
Roles Storage Classes	alicloud-monitor-controller	k8s-app: alicloud-monitor-controller task: monitoring	1/1	2018-06-28 14:06:30	registry.cn-hangzhou.aliyuncs.com/acs/	:
Namasaaca 1	Stiller-deploy	app: helm name: tiller	1/1	2018-06-27 17:53:58	registry-vpc.cn-hangzhou.aliyuncs.com/	:
kube-system 🔻	alicloud-disk-controller	app: alicloud-disk-controller	1/1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyuncs.com/	:
Overview	heapster	k8s-app: heapster task: monitoring	1 / 1	2018-06-27 17:53:56	registry.cn-hangzhou.aliyuncs.com/acs/	:
Workloads Cron Jobs	monitoring-influxdb	k8s-app: influxdb task: monitoring	1 / 1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyuncs.com/	:
Daemon Sets	oginx-ingress-controller	app: ingress-nginx	1/1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyuncs.com/	:
Deployments 2	default-http-backend	app: default-http-backend	1/1	2018-06-27 17:53:56	registry-vpc.cn-hangzhou.aliyuncs.com/	:

If you do not know the REGION information, you can go to the ECS console and select the region where your cluster resides. The last segment of the page URL address is REGION.

×	× ■ 安全 https://ecs.console.aliyun.com/#/server/region/ <u>/us-east-1</u>									
Home	Products 🗸	US (Virginia) 🗸 2	Search Q							
Elastic	You can switch Region here. X									
Overview		 Select the instance attribute, or directly enter the keyword 	Q, Tag							
Instan Launc	h Template	Instance ID/Name Tags Monitor Zone IP Address	Status - Network Type							

1.14.4 Use Grafana to display monitoring data

Prerequisites

- You have successfully created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- In this example, use the Grafana with built-in monitoring templates and the image address is registry . cn hangzhou . aliyuncs . com / acs / grafana :
 5 . 0 . 4 .

Context

Among Kubernetes monitoring solutions, compared with open-source solutions such as Prometheus, the combination of Heapster + InfluxDB + Grafana is more simple and direct. Heapster not only collects monitoring data in Kubernetes, but also is relied on by the monitoring interface of the console and the POD auto scaling of HPA . Therefore, Heapster is an essential component of Kubernetes. An Alibaba Cloud Kubernetes cluster has the built-in Heapster + InfluxDB combination. To display the monitoring data, you must configure an available Grafana and the corresponding dashboard.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment in the left-side navigation pane.
- 3. Click Create by template in the upper-right corner.



- 4. Configure the template to create the deployment and service of Grafana. After completing the configurations, click DEPLOY.
 - · Clusters: Select a cluster.
 - Namespace: Select the namespace to which the resource object belongs, which must be kube system .
 - Resource Type: Select Custom in this example. The template must contain a deployment and a service.

Deploy temp	lates			
Only Kuberr	etes versions 1.8.4 ar	nd above are supported.	For clusters of version 1.8.1, you can perform "upgrade cluster" operatio	n in the cluster list
	Clusters	test		٣
	Namespace	kube-system	1	•
	Resource Type	Custom		*
	Template	<pre>1 apiVersion: 0 2 kind: Deploy 3 metadata: 4 name: mon 5 namespace 6 spec: 7 replicas: 8 template: 9 metadat 10 label 11 tas 12 k8s 13 spec: 14 conta 15 - nam 16 ima 17 por 18 - cc 19 p 20 vol 21 - m 22 mn 22 env 24 - n 25 v 26 volum 27 - nam</pre>	extensions/v1beta1 ment itoring-grafana : kube-system 1 a: s: k: monitoring -app: grafana ge: registry.cn-hangzhou.aliyuncs.com/acs/grafana:5.0.4 ts: ontainerPort: 3000 rotocol: TCP umeMounts: ountPath: /var ame: grafana-storage : ame: INFLUXDB_HOST alue: monitoring-influxdb es: e: grafana-storage	

The orchestration template in this example is as follows:

```
apiVersion : extensions / v1beta1
kind : Deployment
metadata :
   name : monitoring - grafana
namespace : kube - system
spec :
   replicas : 1
   template :
     metadata :
       labels :
         task : monitoring
         k8s - app : grafana
     spec :
       containers :
         name : grafana
         image : registry . cn - hangzhou . aliyuncs . com / acs /
grafana : 5 . 0 . 4
         ports :
          containerP ort : 3000
```

```
protocol : TCP
        volumeMoun ts :
        - mountPath : / var
           name : grafana - storage
         env :
        - name : INFLUXDB_H OST
           value : monitoring - influxdb
      volumes :
        name : grafana - storage
emptyDir : {}
apiVersion : v1
kind : Service
metadata :
   name : monitoring - grafana
   namespace : kube - system
spec :
  ports :
   port : 80
     targetPort : 3000
   type : LoadBalanc er
   selector :
     k8s - app : grafana
```

5. Go back to the Deployment page after the successful deployment. Select the cluster from the Clusters drop-down list and then select kube-system from the Namespace drop-down list to view the deployed applications.

Container Service - Kubernetes 🔻	Deployment			Refresh	Create by image	Create by template
Overview	Clusters k8s-cluster Vam	espace kube-system 🔻 2				
 Clusters 	Name	Tag	PodsQuantity	Time Created		Action
Clusters	alicloud-disk-controller	app:alicloud-disk-controller	1/1	06/27/2018,17:53:56	Details	Edit Monitor More -
Nodes Volumes	alicloud-monitor-controller	k8s-app:alicloud-monitor-controller task:monitoring	1/1	06/28/2018,14:06:30	Details	Edit Monitor More 🗸
Namespace	default-http-backend	app:default-http-backend	1/1	06/27/2018,17:53:56	Details	Edit Monitor More -
Application	heapster	k8s-app:heapster task:monitoring	1/1	06/27/2018,17:53:56	Details	Edit Monitor More 🗸
Deployment	kube-dns	k8s-app:kube-dns	1/1	06/27/2018,17:53:51	Details	Edit Monitor More -
Pods Service	monitoring-grafana	k8s-app:grafana task:monitoring	1/1	07/10/2018,16:04:34	Details	Edit Monitor More 🗸

6. Click the name monitoring-grafana to view the deployment status. Wait until the running status changes to Running.

Deploymentmonitoring-grafana e Back to List						
Overview						
Name:	monitoring-grafana					
Namespace:	kube-system					
Time Created:	2018-04-27 17:54:26					
Label:	k8s-app:grafana task:monitoring					
annotation:	deployment.kubernetes.io/revision:1					
Selector:	k8s-app:grafana task:monitoring					
Strategy:	RollingUpdate					
Status:	Updated:1 , Unavailable:0 , Replica:1					
Pods RelatedService						
Name		Status	Image	Events		
monitoring-grafana-675d	c8448c-drfkq	Running	registry.cn-hangzhou.aliyuncs.com/acs/grafana:5.0.4			

7. Click Application > Service in the left-side navigation pane. Select the cluster from the Clusters drop-down list and kube-system from the Namespace drop-down list to view the external endpoint.

The external endpoint is automatically created by using the LoadBalancer type service. For developers who require more secure access policies, we recommend that you increase the security by adding the external endpoint to the IP whitelist or configuring the certificate.

Container Service - Kubernetes 🝷	Service List							Refresh	Create
Overview	Clusters k8s-cluster	 Namespa 	ce kube-system 🔻						
▼ Clusters	Name	Туре	Time Created	ClustersIP	internalendpoint	externalendpoint			Action
Clusters	default-http-backend	ClusterIP	06/27/2018,17:53:56	10.0	default-http-backend:80 TCP	-	Details Update	View YAML	Delete
Nodes	heapster	ClusterIP	06/27/2018,17:53:56	-	heapster:80 TCP	-	Details Update	View YAML	Delete
Volumes Namespace	kube-dns	ClusterIP	06/27/2018,17:53:51	-	kube-dns:53 UDP kube-dns:53 TCP	-	Details Update	View YAML	Delete
✓ Application	monitoring-grafana	LoadBalancer	07/10/2018,16:04:34	1000	monitoring-grafana:80 TCP monitoring-grafana:32746 TCP		Details Update	View YAML	Delete
Deployment	monitoring-influxdb	ClusterIP	06/27/2018,17:53:56	1000	monitoring-influxdb:8086 TCP		Details Update	View YAML	Delete
Pods Service	nginx-ingress-lb	LoadBalancer	06/27/2018,17:53:56	1000	nginx-ingress-lb:80 TCP nginx-ingress-lb:30883 TCP nginx-ingress-lb:443 TCP nginx-ingress-lb:32380 TCP	:80	Details Update	View YAML	Delete
8. Click the external endpoint at the right of the monitoring-grafana service to log on to the Grafana monitoring page.

By default, the username and password of Grafana are both admin. We recommend that you change the password after the logon.



9. Select the built-in monitoring templates to view the monitoring dashboards of the pod and node.

In this example, the Grafana has two built-in templates, one for displaying physical resources at the node level, and one for displaying resources related to the pod.

Developers can also perform more complex presentations by adding custom dashboards or configure resource alarms based on Grafana.



namespace default - pod_nam	e hello-pod ▼						
✓ Status	✓ Status						
Uptime	CPU Usage Rate	Memory Usage	Network Tx Rate	Network Rx Rate			
1.446 hour	0	1.516 MB	0 Bps	0 Bps			
✓ Resource Limit and Request							
CPU Limit	CPU Request	Memory Limit	Memory Request				
0	0	0 B	0 B				
✓ History							
	CPU Usage Rate			Memory Usage			
1.0 0.5 -0.5 -1.0			2.0 MB 1.8 MB 1.5 MB 1.3 MB 1.0 MB				
17:35 17:40 17:45 17 — cou/usage rate.mean	2:50 17:55 18:00 18:05 18:10	18:15 18:20 18:25 18:30	17:35 17:40 17:45 — memory/usage.mean	17:50 17:55 18:00 18:05 18:	10 18:15 18:20 18:25 18:30		
6 Bps	Network		1.67 hour	Uptime			

1.14.5 Use an HPA auto scaling container

Alibaba Cloud Container Service supports the rapid creation of HPA-enabled applications on the console interface to achieve auto scaling of container resources. You can also configure it by defining the yaml configuration of Horizontal Pod Autoscaling (HPA).

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have successfully connected to the master node of the Kubernetes cluster.

Method 1 Create an HPA application in the Container Service console

In Alibaba Cloud Container Service, HPA has been integrated. You can easily create it through the Container Service console.

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Deployment in the left-side navigation pane. Click Create by image in the upper-right corner.

Container Service - Kubernetes +	Deployment	Refresh Create by image Create by template
Overview	Clusters k8s-test • Namespace default •	2
 Clusters 	Name Tag PodsQuantity Time Created	Action
Clusters		
Nodes	⑦ Could not find any record that met the condition.	
Volumes		
Namespace		
Authorization		
Application Deployment		8

- 3. Enter the application name, select the cluster and namespace, and click Next.
- 4. Configure the application settings. Set the number of replicas, select the Enable box for Automatic Scaling, and configure the settings for scaling.
 - Metric: CPU and memory. Configure a resource type as needed.
 - Condition: The percentage value of resource usage. The container begins to expand when the resource usage exceeds this value.
 - Maximum Replicas: The maximum number of replicas that the deployment can expand to.
 - Minimum Replicas: The minimum number of replicas that the deployment can contract to.

Replicas: 1
Auto Scaling: 🗷 Enable
Metric: CPU Usage
Condition: Usage 50 %
Maximum Replicas: 10 Range : 2-100
Minimum Replicas: 1 Range : 1-100

5. Configure the container. Select an image and configure the required resources. Click Next.

Note:

You must configure the required resources for the deployment. Otherwise, container auto scaling cannot be achieved.

cor	tainer0		
	Image Name:	nginx	Image Version: latest Always pull image
		Select image	Select image version
sener al	Resource Limit:	CPU eg : 500m Memory eg : 128Mi	
0	Resource Request:	CPU 500m Memory eg : 128Mi	
	Init Container		

6. In the Access Control page, do not configure any settings in this example. Click Create directly.

Now a deployment that supports HPA has been created. You can view the auto scaling group information in the details of your deployment.

Overview					
Name:	nginx-deployment				
Namespace:	default				
Time Created:	2018-08-23 10:07:23				
Label:	app:nginx				
annotation:	deployment.kubernetes.io/revision:1				
Selector:	app:nginx				
Strategy:	RollingUpdate				
Status:	Updated:1 , Unavailable:1 , Replica:1				
Trigger 1. You can only have on	ne of each trigger type.				Create Trigger
No trigger is available at the moment.	Click "Create Trigger" in the upper-right corner.				
CPU			Memory		
0			0		
10:09:00			10:08:00		10:09:00
	 CPU usage(Cores) 			 Memory usage(Gi) 	
Pods Access Events	Horizontal Pod Autoscaler				
Name	Target Utilization	Minimum Replicas	Maximum Replicas	Created At	Action
nginx	cpu:70%	1	10	08/23/2018,10:07:23	Edit Delete

7. In the actual environment, the application scales according to the CPU load. You can also verify auto scaling in the test environment. By performing a CPU pressure test on the pod, you can find that the pod can complete the horizontal expansion in half a minute.

Method 2 Use kubectl commands to configure container auto scaling

You can also manually create an HPA by using an orchestration template and bind it to the deployment object to be scaled. Use the kubectl command to complete the container auto scaling configuration. The following is an example of an Nginx application. Execute the kubectl

create - f xxx . yml command to create an orchestration template for the

deployment as follows:

```
apiVersion : apps / v1beta2 # for
                                                before
                                     versions
                                                         1.8.0
       apps / v1beta1
 use
kind : Deployment
metadata :
 name : nginx
  labels :
   app : nginx
spec :
  replicas : 2
 selector :
   matchLabel s :
     app : nginx
  template :
   metadata :
     labels :
       app : nginx
    spec :
     containers :
       name : nginx
       image : nginx : 1 . 7 . 9 #
                                               it
                                                    with
                                     replace
                                                           your
exactly < image_name : tags >
       ports :
         containerP ort : 80
       resources :
                                           ## This
         requests :
                                                     parameter
           configured . Otherwise , the
                                           HPA
                                                          operate .
must
      be
                                                 cannot
                 500m
           cpu :
```

Create an HPA. Configure an object to which the current HPA is bound by using

scaleTarge tRef . In this example, the object is the deployment named nginx.

```
apiVersion : autoscalin g / v2beta1
kind : Horizontal PodAutosca ler
metadata :
 name : nginx - hpa
 namespace : default
spec :
                                               ## Bind
 scaleTarge tRef :
                                                              HPA
                                                        the
  to
     a deployment
                       named
                              nginx
    apiVersion : apps / v1beta2
    kind : Deployment
   name : nginx
 minReplica s:
                  1
 maxReplica s: 10
 metrics :
   type : Resource
    resource :
     name :
            cpu
     targetAver ageUtiliza tion :
                                    50
```



The HPA needs to configure the request resource for the pod.The HPA does not operate without the request resource.

Warnings similar to the following are displayed when you execute kubectl

describe hpa [name]: Warning FailedGetR esourceMet ric 2m (x6 over 4m) for horizontal - pod - autoscaler missing request cpu on container nginx in pod default / nginx - deployment - basic - 75675f5897 - mqzs7 Warning FailedComp uteMetrics Replicas 2m (x6 over horizontal - pod - autoscaler 4m) failed to get cpu сри utilizatio n : missing request for on container default / nginx - deployment - basic - 75675f5 nginx in pod

After creating the HPA, execute the kubectl describe hpa [name] command again. You can see the following message, which indicates that the HPA is running normally.

Normal Successful Rescale 39s horizontal - pod - autoscaler New size : 1 ; reason : All metrics below target

When the usage of Nginx pod exceeds 50% set in this example, the container expands horizontally. When the usage of Nginx pod drops below 50%, the container contracts.

1.14.6 Monitor a Kubernetes cluster and send alarm notifications by using DingTalk

After you deploy a robot in a DingTalk group, the cluster sends a notification of an exception event to the DingTalk group through the robot, implementing real-time monitoring and alarming for cluster exception events.

Context

- You have created a DingTalk group .
- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

Procedure

1. Click the icon in the upper-right corner of the DingTalk group.

428

2. Click ChatBot. On the ChatBot page, select a robot. Select a Custom robot.

ChatBot						
Please choose which robot to add						
	6600	0				
Weather Automatically push weather forecasts and warnings	Alibaba Cloud Code Code hosting service provided by Alibaba Cloud	GitHub Git-based code hosting service	GitLab ROR-based open source code hosting software			
×		D				
JIRA Excellent project and work tracking tool	Travis Excellent project and work tracking tool	Trello Real-time cards wall to manage everything	钉档 钉钉开放平台文档查 询机器人			
¢,						
CODING 以Git为基础的研发管 理平台	Custom Custom message services via Webbook					

3. On the Robot details page, click Add.



4. Configure the following parameters for a robot and then click Finished:

Configuration	Description
Edit profile picture	(Optional) Set a profile picture for the robot.
ChatBot Name	The robot name.
Add to Group	The DingTalk group to which the robot is added to.
Enable the outgoing function	(Optional) By perform the @robot operation, you can send messages to a specified external service as well as return response results of the external service to the group.
	Note: We recommend that you do not enable this function.

Configuration	Description
POST address	The HTTP service address that receives messages. Note: You can configure this parameter after you enable the outgoing function.
Token	The key used to verify that a request is from DingTalk. Note: You can configure this parameter after you enable the outgoing function.

5. Click Copy to copy the webhook address.

Add Robot	×
1. Add robot✓	
2. Set up webhook, click setting instruction and check how to make robot effective	
webhook : https://oapi.dingtalk.com/robot/send?access_token=	
Finished Setting ins	
Note:	
On the ChatBot page, click the icon at the right of a robot and ther	i you
can perform following operations:	

- Modify the profile picture and name of the robot.
- Open or Close notifications.
- Reset the webhook address.
- Remove the robot.

ChatBot		×
DingTalk ro	obot can automatically send out your notification and messages you need to group chat	Learn More
۵	Add Robot Currently, one chatting group can add 6 chat robots at most	+
Robot for	this group	
۵	Kubernetes Robot Added by accecpting groups:	Ø

- 6. Log on to the Container Service console.
- 7. Under the Kubernetes menu, click Application > Deployment in the left-side navigation pane.
- 8. Select a cluster, select the kube-system namespace, and click Create by Template in the upper-right corner.

Deployment	Refresh Create by Image C	reate by Template
Help: & How to use private images & Create applications & Schedule a pod to the specified node & Create a Layer-4 Ingress & Create a Layer indicate a Layer ind	ayer-7 Ingress 🔗 Configure pod auto scali	ng 3 tainer
Clusters Namespace kube-system		

9. Configure a template based on the following parameters, and then click Deploy.

Configuration	Description
Clusters	Select a cluster.

Configuration	Description
Namespace	Select a namespace to which resource object belongs. The default namespace is default. Select kube-system.
Sample template	Alibaba Cloud Container Service provides Kubernetes YAML sample templates of many resource types for you to deploy resource objects quickly. You can write your own template based on the format requirements of Kubernetes YAML orchestration to describe the resource type you want to define. Select Custom.

Configuration	Description
Template	Enter the following custom content:
Template	<pre>Enter the following custom content: apiVersion : extensions / vlbetal kind : Deployment metadata : name : eventer namespace : kube - system spec : replicas : 1 template : metadata : labels : task : monitoring k8s - app : eventer annotation s : scheduler . alpha . kubernetes . io / critical - pod : '' spec : serviceAcc ount : admin containers : - name : eventer image : registry . cn - hangzhou . aliyuncs . com / acs / eventer : vl . 6 . 0 imagePullP olicy : IfNotPrese nt command : - / eventer source = kubernetes : https :// kubernetes . default sink = dingtalk :[your_webho ok_url]& label =[your_clust er_id]& level =[Optional parameters are : Normal and Warning . The default is Warning .] # The level field can be set to Normal or Warning . The default is Warning . When the level field is set to Normal alarm notificati ons of the Normal alarm notificati ons of the Normal alarm potificati ons of the Normal Normal Alarm potificati ons of</pre>
	and Warning levels can be received in the DingTalk group. When you do not set the level field or set the level field to Warning, only alarm notificati ons of the Warning level can be received in the DingTalk group.

On the Cluster List page, click Dashboard at the right of the cluster. On the Dashboard, select kube-system from the drop-down list of Namespace, and click Deployments in the left-side navigation pane. The deployed eventer is displayed.

Namespace		ints									+ CREATE
kube-system *	CPU usage					Memory usage	(i)				
Workloads	0.027					901 Mi 801 Mi					
Cron Jobs Daemon Sets	80.018 0.012 0.000			~		5 601 Mi 2 401 Mi 2 200 Mi					
Deployments	19:09 19:10	19:13	19:16 Time	19:20	19:23	≥ 19:09 10:	10	19:13	19:16 Time	19:20	19:23
Jobs Pods											
Replica Sets	Deployments		Labela		De de		1 4		·		Ŧ
Replication Controllers Stateful Sets	eventer		k8s-app: eventer	task: monitoring	Pods		Age = 12/06/201	8, 19:23:25	registry.cn-ha	angzhou.aliyuncs.com/acs/eve	:

Result

The eventer takes effect 30 seconds after you complete the deployment. When an event exceeds the threshold level, you receive the following alarm notifications in the DingTalk group.



1.15 Security management

1.15.1 Security

Authorization

Kubernetes clusters support authorizing RAM users to perform operations on clusters.

For more information, see Use the Container Service console as a RAM user.

Full-link TLS certificates

The following communication links in Container Service Kubernetes clusters are verified by TLS certificates to prevent the communication from being eavesdropped or tampered:

- kubelet on worker nodes actively communicates with apiserver on master nodes
- apiserver on master nodes actively communicates with kubelet on worker nodes

During initialization, the master node uses SSH tunnels to connect to the SSH service of other nodes (port 22) for initialization.

Native secret & RBAC support

Kubernetes secrets are used to store sensitive information such as passwords, OAuth tokens, and SSH keys. Using plain text to write sensitive information to a pod YAML file or a Docker image may leak the information, while using secrets avoids such security risks effectively.

For more information, see Secret.

Role-Based Access Control (RBAC) uses the Kubernetes built-in API group to drive authorization and authentication, which allows you to use APIs to manage pods that correspond to different roles, and the access permissions of roles.

For more information, see Using RBAC authorization.

Network policy

In a Kubernetes cluster, pods on different nodes can communicate with each other by default. In some scenarios, to reduce risks, the network intercommunication among different business services is not allowed and you must introduce the network policy. In Kubernetes clusters, you can use the Canal network driver to implement the support for network policy.

Image security scan

Kubernetes clusters can use Container Registry to manage images, which allows you to perform image security scan.

Image security scan identifies the security risks in images quickly and reduces the possibility of applications running on your Kubernetes cluster being attacked.

For more information, see Image security scan.

Security group and Internet access

By default, each newly created Kubernetes cluster is assigned a new security group with the minimal security risk. This security group only allows ICMP for the Internet inbound.

By default, you cannot use Internet SSH to access your clusters. To use Internet SSH to connect to the cluster nodes, see *Access Kubernetes clusters by using SSH*.

The cluster nodes access the Internet by using the NAT Gateway, which further reduces the security risks.

1.15.2 Kube-apiserver audit logs

In a Kubernetes cluster, apiserver audit logs are important for cluster Operation & Maintenance (O&M) because they record daily operations of different users. This topic describes how to configure the apiserver audit logs of an Alibaba Cloud Kubernetes cluster, and how to collect and analyze audit logs through Log Service, and how to customize audit log alarm rules.

Configurations of apiserver audit logs

The apiserver audit function is enabled by default when you create a Kubernetes cluster. Relevant parameters and description are as follows:

Note:

Log on to the Master node, and the directory of the apiserver configuration files is / etc / kubernetes / manifests / kube - apiserver . yaml .

Configuration	Description
audit-log-maxbackup	The maximum fragment of audit logs stores 10 log files.
audit-log-maxsize	The maximum size of a single audit log is 100 MB.
audit-log-path	The audit log output path is / var / log / kubernetes / kubernetes . audit
audit-log-maxage	The longest storage period of audit logs is seven days.
audit-policy-file	Configuration policy file of audit logs. The directory is / etc / kubernetes / audit - policy . yml .

Log on to the Master node machine. The directory of the audit log configuration policy file is / etc / kubernetes / audit - policy . yml . The content of the file is as follows:

```
apiVersion : audit . k8s . io / v1beta1 # This
                                                    is
                                                         required .
kind : Policy
  We recommend that for all requests
                          you do
                                                      audit
# We
                   that
                                    not
                                           generate
                                                              events
                         in RequestRec eived stage.
omitStages :
 - " RequestRec eived "
 rules :
          following
                                 are manually
                                                  identified
 # The
                      requests
                                                               as
high - volume and
                      low - risk .
    Therefore ,
                we recommend
                                 that
                                               drop
 #
                                         you
                                                      them .
    level : None
    users : [" system : kube - proxy "]
    verbs : [" watch "]
    resources :
       group : "" # core
        resources : [" endpoints ", " services "]
    level : None
    users : [" system : unsecured "]
    namespaces : [" kube - system "]
    verbs : [" get "]
    resources :
        group : "" # core
        resources : [" configmaps "]
    level : None
    users : [" kubelet "] # legacy
                                      kubelet
                                                identity
    verbs : [" get "]
    resources :
        group : "" # core
        resources : [" nodes "]
    level : None
    userGroups : [" system : nodes "]
    verbs : [" get "]
    resources :
```

```
- group : "" # core
        resources : [" nodes "]
    level : None
    users :
     - system : kube - controller - manager
     - system : kube - scheduler
       system : serviceacc ount : kube - system : endpoint -
controller
    verbs : [" get ", " update "]
    namespaces : [" kube - system "]
    resources :
    - group : "" # core
        resources : [" endpoints "]
    level : None
users : [" system : apiserver "]
    verbs : [" get "]
resources :
     - group : "" # core
       resources : [" namespaces "]
recommend that you do
                                       not log
 # We
                                                    these read -
only URLs .
- level : None
    nonResourc eURLs :
     - / healthz *
     - / version
    - / swagger *
                    that you do not log events requests
   We recommend
 #
- level : None
    resources :
     - group : "" # core
       resources : [" events "]
# Secrets , ConfigMaps , and TokenRevie ws can contain sensitive and binary data .
# Therefore , they are logged only at
                                                     the
                                                           Metadata
level .
  level : Metadata
    resources :
      group : "" # core
       resources : [" secrets ", " configmaps "]
     - group : authentica tion . k8s . io
    resources : [" tokenrevie ws "]
Get repsonses can be large; skip them.
 #
    level : Request
    verbs : [" get ", " list ", " watch "]
    resources :
     - group : "" # core
       group : " admissionr egistratio n . k8s . io "
     _
       group : " apps "
     _
       group : " authentica tion . k8s . io "
     _
       group : " authorizat ion . k8s . io "
     _
       group : " autoscalin g "
     _
       group : " batch "
     _
       group : " certificat es . k8s . io "
     _
       group : " extensions "
       group : " networking . k8s . io "
     _
       group : " policy "
       group : " rbac . authorizat ion . k8s . io "
       group : " settings . k8s . io "
    - group : " storage . k8s . io "
Default level for known APIs .
 #
    level : RequestRes ponse
    resources :
     - group : "" # core
```

```
- group : " admissionr egistratio n . k8s . io "
- group : " apps "
- group : " authentica tion . k8s . io "
- group : " authorizat ion . k8s . io "
- group : " autoscalin g "
- group : " batch "
- group : " certificat es . k8s . io "
- group : " extensions "
- group : " networking . k8s . io "
- group : " rbac . authorizat ion . k8s . io "
- group : " settings . k8s . io "
- group : " storage . k8s . io "
# Default level for all other requests .
- level : Metadata
```

Note:

- Logs are not recorded immediately after requests are received. Log recording starts only after the response body header is sent.
- The following requests or operations are not audited: redundant kube-proxy watch requests, GET requests from kubelet and system:nodes for nodes, operations performed on endpoints by kube components in the kube-system, and GET requests from the apiserver for namespaces.
- Read-only urls such as / healthz *, / version *, and / swagger * are not audited.
- Logs of interfaces of secrets, configmaps, and tokenreviews are set to the metadata level because they might contain sensitive information or binary files
 For logs of this level, only the user, timestamp, request resources, and request actions of the request event are audited. The request body and the response body are not audited.
- For sensitive interfaces such as authentication, rbac, certificates, autoscaling, and storage, the corresponding request bodies and response bodies are audited according to the read and write requests.

View audit log reports

- A Kubernetes cluster that runs on Alibaba Cloud Container Service has three audit log reports that provide the following information:
- · Operations performed by all users and system components on the cluster
- The source IP address of each operation, the area to which a source IP addresses belongs, and the source IP address distribution
- · Detailed operation charts of all resources

- · Operation charts of each sub-account
- Charts of important operations such as logging on to a container, accessing a secret, and removing resources

Note:

- For Kubernetes clusters created after January 13, 2019, if you active Log Service for the clusters, the system automatically enables audit log report functions. If audit log report functions are disabled for a Kubernetes cluster, see *Manually enable audit log report functions*.
- We recommend that you do not modify audit log reports. If you want to customize audit log reports, you can create new reports in the *Log Service console*.

You can access audit log reports by using either of the following two methods:

• Log on to the *Container Service console*. In the action column of the target cluster, chooseMore > Cluster Audit.



• Log on to the *Container Service console*. Click the target cluster name, and then click Cluster Auditin the left-side navigation pane.

<	Cluster:						
Basic Information	Basic Information						
Node List	Cluster ID:						
Event List							
Cluster Audit	Cluster Information						
	API Server Internet endpoint						
-	API Server Intranet endpoint						
	Pod Network CIDR						
	Service CIDR						
	Master node SSH IP address						
	Service Access Domain						

Audit log report overview

The following three apiserver audit log reports are available: Audit Center Overview, Resource Operation Overview, and Resource Operation Detail.

Audit Center Overview

This report displays an overview of the Kubernetes cluster events and the detailed information about important events, such as public network visits, command execution, resource removal, and secret visits.

<	Audit Center Overview Resource Operation 0	enter Overview Resource Operation Overview Resource Operation Detail						
×	M Kubernetes Audit Center Overv	() Refresh I Title Reset Time						
Basic Information		NameSpace(optional): Search	Sub-account ID (optional): Search	Status Code (optional): Search				
Node List	XX							
Event List		Oursulation Delete Frante (D.) (D.) (D.)	Bullis Maturals Visite - (Dur(Dulutur)	lileest visite on which the				
Cluster Audit		Cumulative Delete Events TDay(Relative)	Public Network Visits 1 Day(Relative)	inegal visits 1Day(Relative)				
		0		0 7				
	Event count 1Day(Relative)	Ocount Today/Ring Yesterday	-93.75%	Count InfinityE18%				
	0.47414	iouuj/inig iouuluuj	roduy/ring restoredy	loudy, king resterauy				
	2.171Mil							
	Events level distribution 1Day(Relative)	Operation distribution 1Day(Time Frame) 100K		ApiServer load 1Day(Relative)				
		80К						
		60K	- Iist	30.74%				
	Request		• get	Total				



By default, this report displays statistics for one week. You can customize a statistics time range. In addition, you can filter events by specifying one or multiple factors, such as a namespace, a sub-account ID, and a status code.

· Resource Operation Overview

This report displays the operation statistics information about computing resources, network resources, and storage resources of a Kubernetes cluster. Operations include creation, update, removal, and access.

- Computing resources include deployment, StatefulSet, CronJob, DaemonSet, Job , and pod.
- Network resources include service and Ingress.
- Storage resources include ConfigMap, secret, and Persistent Volume Claim.





- By default, this report displays statistics for one week. You can customize a statistics time range. In addition, you can filter events by specifying one or both of the following factors: a namespace or a sub-account ID.
- If you want to view the detailed operation events of a resource, we recommend that you use Resource Operation Detail.
- **Resource Operation Detail**

This report displays detailed operation information of a Kubernetes cluster resource. You must select or enter a resource type to view detailed operation

information in time. This report displays the total number of operation events, namespace distribution, success rate, timing trend, and specific operation charts.

<	Audit Center Overview Resource Operation Overview Resource Operation Detail
•	Kubernetes Resource Operation Detail (Belong To k8s-log-c7b729e7ea122425182133bc2c3eab8c3) 🛈 30Days(Relative) 🔻 🎲 Refresh 🚳 Title Reset Time
Basic Information	Variable: ResourceType: Pods X
Node List	Resource Type: V Search Sub-Account ID(optional): Search NameSpace(optional): Search Status Code(optional): Search
Event List	
Cluster Audit	Create Event NameSpace Distribution Update Event NameSpace Distributio Delete Event NameSpace Distribution Event Access NameSpace Distribution
Ξ	No data Total 1.45% 0 default 5.1.21% 0 default 5.1.21% 0 default 0 d



- If you want to view a CRD resource registered in Kubernetes or any other resources not listed in the report, you can enter the plural form of the target resource. For example, to view a CRD resource named AliyunLogConfig, you can enter AliyunLogConfigs.
- By default, this report displays statistics for one week. You can customize a statistics time range. In addition, you can filter events by specifying one or multiple factors, such as a namespace, a sub-account ID, and a status code.

logs, you can log on to Log Service to view detailed log records.

- 1. Log on to the Log Service console.
- 2. In the left-side navigation pane, click Project Management, select the Project configured when you create the cluster, and then click the Project name.

Log Service	Projects				Learning Path	Create Project
Project Management	Search by project name Search					
	Project Name -	Description	Region	Created At-		Actions
	k8s-log	k8s log project, create	China (Beijing)	Sep 17, 2018, 10:06:58 AM	Modify Global Ac	celeration Delete

3. On the Logstores page, find the Logstore named audit-\${clusterid} and click Search at the right side of the Logstore. The audit logs of the cluster are stored in this Logstore.

<	() k8s-log-cfe0b30312db344	€ Back to	Project List	1 1			R	gion : China (Beijing)
Logstores 1	Logstores					Learning Path	Endpoir	ts Create
 LogHub - Collect [Deal Collection Hell 	Enter a Logstore name to Search Search							
		Data			Log	Consumption	Mode	
Logtail Machine Grou	Logstore Name	Import Wizard	Monitor	Log Collection Mode	Log Consumptio	Log on Shipper	LogSearch	Actions
 LogHub - Consume 	access		⊵	Logtail Config (Manage) Diagnose More ▼	Preview More •	OSS	Search	Modify Delete
[Doc] Consumption H	audit-		×	Logtail Config (Manage) Diagnose	Preview	OSS	Search	2 Modify Delete
Consumer Group	PROFESSION RESIDENCE			More -	More 🕶			

- Note:
- When you create a Kubernetes cluster, your specified log Project automatically creates a Logstore named audit-\${clusterid}.
- The audit log Logstore index is set by default. We recommend that you do not modify the index. Otherwise, the audit log reports become invalid.

To searchfor an audit log, you can use one of the following methods:

- To querya sub-account operation record, enter the sub-account ID and then click Search & Analysis.
- To query operations on a resource, enter the resource name and click Search & Analysis.
- To filter out operations performed by system components, enter NOT user .
 username : node NOT user . username : serviceacc ount NOT
 user . username : apiserver NOT user . username : kube scheduler NOT user . username : kube controller manager , and
 then click Search & Analysis.

For more information, see Log Service search and analysis methods.

Set resource alarms

You can use the alarm function of Log Service to set resource alarms. Alarm notificati ons can be sent through a DingTalk group robot, a customized Webhook, and the Message Center.



Note:

Audit log reports provide multiple query statements. On the Logstores page, click Dashboardin the left-side navigation pane and then click a dashboard (namely, an audit log report) to display all charts, Click the menu in the upper-right corner of a chart, and then click View Details.

Example 1: Set an alarm notification for running a command on a container

To prevent Kubernetes cluster users from logging on to any container to run a command, you must set an alarm notification for running a command on any container. Furthermore, the alarm notification must include detailed information such as the container to which the user logged on, commands, user name, the event ID, the operation time, and the user IP address.

• The query statement is as follows:

and objectRef . subresourc e : exec verb : create and stage : ResponseSt arted | SELECT auditID as " event ID , date_forma t (from_unixt ime (__time__), '% Y -% m -% d % T ') as "operation time ", regexp_ext ract (" requestURI ", '([^\?]*)/ exec \?.*', 1) as " resource ", regexp_ext ract (" requestURI ", '\?(.*)', 1) as " command "," responseSt atus . code " as " status code ", CASE WHEN " user . username " != ' kubernetes - admin ' then " user username " " user . username " = ' kubernetes - admin ' WHEN and / reason ", ' RoleBindin g ') then regexp_ext ract (" annotation s . authorizat ion . k8s . io / reason ", ' to "(\ w +)"', 1) User ' kubernetes - admin ' END ELSE " operation account ", as WHEN json_array _length (sourceIPs) = 1 CASE then json_forma t (json_array _get (sourceIPs , 0)) ELSE sourceIPs END " source IΡ address " limit as 100

• The condition expression is operation event =~ ".*".

Example 2: Set an alarm notification for failed Internet access to apiserver

To prevent malicious attacks on a Kubernetes cluster for which Internet access is enabled, you need to monitor the number of Internet access times and the failed access rate. Specifically, an alarm notification must be sent, when the number of Internet access times reaches a specified threshold and the failed access rate exceeds a specified threshold. Furthermore, the alarm notification must include detailed information such asto which the user IP address belongs, the user IP address, and the high risk IP address. For example, to receive an alarm notification when the number of Internet access times reaches 10 and the failed access rate exceeds 50%, configure the following settings:

· Query statement.

" " source address ", * | select ip IΡ total as as number of access times ", round (rate * 100, 2) as
"failed access rate %", failCount as "number of
illegal access times ", CASE when security_c heck_ip
(ip) = 1 then ' yes ' else ' no ' end as " high
risk IP address ", ip_to_coun try (ip) as " country ",
ip_to_prov ince (ip) as " province ", ip_to_city (ip) as
" city ", ip_to_prov ider (ip) as " network operator "
from (select CASE WHEN ison array length (sourceIPs) CASE WHEN json_array _length (sourceIPs) json_forma t (json_array _get (sourceIPs , 0 from (select = 1 then 0)) ELSE sourceIPs END total , ip, count (1) as as sum (CASE WHEN " responseSt atus . code " < 400</pre> then 0 ELSE 1 END) * 1 . 0 / count (1) as count_if (" responseSt atus . code " = 403) rate , failCount as group by ip limit 10000) from log where ip_to_doma in (ip) != ' intranet ' " number having of times " > 10 and " failed access rate %" > 50 access " number of times " desc ORDER by access limit 100

· Condition expression is source IP address =~ ".*".

Manually enable audit log report functions

You can manually enable audit log report functions.

1. Enable API server audit log.

View the API server pod settings of the three Master nodes. That is, check whether audit log settings are configured for the startup parameters, the policy file, the environment variable, and the mounting directory.

• Startup parameters

```
containers :
    command :
        kube - apiserver
        - -- audit - log - maxbackup = 10
        --- audit - log - maxsize = 100
        --- audit - log - path =/ var / log / kubernetes / kubernetes
. audit
        --- audit - log - maxage = 7
```

```
- -- audit - policy - file =/ etc / kubernetes / audit -
policy . yml
```

Policy file (stored in the / etc / kubernetes / audit - policy . yml directory)

For more information, see *Configure a policy file*.

Note:

```
If the / etc / kubernetes / directory does not have any policy file, you need
to run the vi audit - policy . yml command to create a file, and then
copy the content of the policy file and paste the content to the created file.
```

Environment variable

· Mounting directory

```
volumeMoun ts :
        - mountPath : / var / log / kubernetes
        name : k8s - audit
        - mountPath : / etc / kubernetes / audit - policy . yml
        name : audit - policy
        readOnly : true
volumes :
        - hostPath :
        path : / var / log / kubernetes
        type : Directory0 rCreate
        name : k8s - audit
        - hostPath :
        path : / etc / kubernetes / audit - policy . yml
        type : FileOrCrea te
        name : audit - policy
```

Backup the original YAML file and then restart the API server by using a new kube

- apiserver . yaml YAML file. This action will overwrite the original YAML file

stored in the / etc / kubernetes / manifests / kube - apiserver . yaml
directory.

If the API server pod settings does not contain the preceding settings, you must upgrade the Kubernete cluster to the latest version. For more information, see *Upgrade a Kubernetes cluster*.

- 2. Use the latest version of the Log Service component.
 - For how to install the Log Service component, see *Manually install the Log Service component*.
 - If you have installed the Log Service component, but the audit log function is disabled, you must upgrade the component to the latest version and you must ensure that your Logtail version is not earlier than v0.16.16 and can run on Master nodes. For more information, see *Upgrade the Log Service component*.
- 3. Update audit log parsing methods.
 - a. Log on to the Log Service console.
 - b. In the left-side navigation pane, click Project Management, and then click the name of the Project specified when creating your Kubernetes cluster.
 - c. The Logstores page is displayed by default. Click Manage on the right of the Logstore named audit-\${clustered}, and then click the configuration name. On the Specify Collection Mode tab page, select the JSON Mode.

Environment Variable Whitelist:	EnvKey +	EnvValue -					
	aliyun_logs_audit-c3730910447504a6692ca1	/var/log/kubernetes/kubernetes.audit					
	Collects log entries that contain the environment variables in the whitelist. If the whitelist is empty, all log entries will be collected.						
Environment Variable Blacklist:	EnvKey +	EnvValue -					
	Collects log entries that do not contain environment entries will be collected.	variables in the blacklist. If the blacklist is empty, all log					
Mode	JSON Mode						
	now to set 550W conliguration						
Use System Time:							

Use a thirty-party log solution

Log on to the Master node of the cluster, and you can find the source file of the audit logs in the path of / var / log / kubernetes / kubernetes . audit . The source

file is in standard JSON format. When deploying a cluster, you can use other log solutions to collect and search audit logs, instead of using Alibaba Cloud Log Service.

1.15.3 Implement secure access through HTTPS in Kubernetes

A Container Service Kubernetes cluster supports multiple application access methods. The most common methods include SLB : Port access, NodeIP : NodePort access, and domain name access. By default, a Kubernetes cluster does not support HTTPS access. To access applications through HTTPS, you can use the secure HTTPS access method provided by Container Service and Alibaba Cloud Server Load Balancer (SLB) service. This document explains how to configure a certificate in Container Service Kubernetes by using HTTPS access configuration as an example.

Depending on different access methods, your certificate can be configured with the following two methods:

- · Configure the certificate on the frontend SLB.
- · Configure the certificate on Ingress.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- You have connected to the Master node through SSH. For more information, see Access Kubernetes clusters by using SSH.
- After connecting to the Master node, you have created the server certificates for the cluster, including the public key certificate and the private key certificate by running the following commands :

```
$
  openssl
          genrsa – out tls.key
                               2048
Generating
          RSA
              private
                      key, 2048 bit
                                     long
                                          modulus
                             ..... +++
    +++
   is 65537 (0x10001)
е
$ openssl req - sha256 - new - x509 - days 365 - key
tls.key - out tls.crt
     are about
                            to enter informatio n
                   be asked
You
               to
that will be incorporat ed
. . .
Country Name ( 2 letter code ) [ XX ]: CN
State or Province Name ( full name ) []: zhejiang
```

(eg , city) [Default Name City]: hangzhou Locality Organizati company) [Default Ltd on Name (eg , Company]: alibaba Organizati onal Unit Name (eg , section) []: test Name (eg , Common name or your your server 's configure hostname) []: foo . bar . com # you must correctly the domain name Email Address []: a @ alibaba . com

Method 1: Configure the HTTPS certificate on SLB

This method has the following advantages and disadvantages:

- Advantages: The certificate is configured on SLB and it is the external access portal of applications. The access to applications in the cluster still uses the HTTP access method.
- Disadvantages: You need to maintain many associations between domain names and their corresponding IP addresses.
- Scenarios: This method is applicable to applications that use LoadBalancer service rather than Ingress to expose access methods.

Preparations

You have created a Tomcat application in the Kubernetes cluster. The application provides external access by using the LoadBalancer service. For more information, see *Create a service*.

Example

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane, click Application > Service, and select the cluster and the namespace to view the pre-created Tomcat application. As shown in the following figure, the created Tomcat application is named tomcat and the service name is tomcat-svc. The service type of the application is LoadBalancer, and the service port exposed by the application is 8080.

Container Service - Kubernetes -	Service List						Refresh Create
Overview	Clusters k8s	-cluster 🔻 Namespace	e default 🔹 3				
 Clusters 							
Clusters	Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint	Action
Volumes	kubernetes	ClusterIP	08/24/2018,10:02:37	172.21.0.1	kubernetes:443 TCP	-	Details Update View YAML Delete
Namespace	nginx-svc	LoadBalancer	08/24/2018,12:37:38	172.21.2.77	nginx-svc:80 TCP nginx-svc:30232 TCP	80	Details Update View YAML Delete
▼ Application 1	tomcat-svc	LoadBalancer	08/28/2018,16:43:44	172.21.14.24	tomcat-svc:8080 TCP tomcat-svc:32403 TCP	8080	Details Update View YAML Delete
Pods							
Service 2							

3. By clicking the external endpoint, you can access the Tomcat application through



- 4. Log on to the SLB console.
- 5. By default, the Server Load Balancer page is displayed. In the IP address column, find the server load balancer that corresponds to the external endpoint of the tomcat-svc service, and click Configure Listener in the actions column.

Server Load Balancer	Serve	er Load Balancer	Back to Old Version>>	Back to Old Version>> Product Updates			
▼ Instances	Crea	ite SLB Instance C' ≔ ⊥				Select a tag V Zones: All V Select an item V Enter a va	lue Q
Server Load Ba		Instance Name/ID	IP Address 77	Status 🏹	Monitoring	Port/Health Check/Backend Server ~	Actions
Expiring Instances Certificates Access Control			9 9 Public IPv4 Address)	Active	a	TCP: 443 • AbnormaDefault Server Group 3 V	Configure Listener Add Backend Servers More ~
Operation Logs Access Logs Health Check Logs			Address)	 Active 	al	TCP: 443 • Normal Default Server Group 3 v TCP: 80 • Normal Default Server Group 3 v	Configure Listener Add Backend Servers More V

6. Configure the server load balancer. Select a listener protocol first. Select HTTPS, set the listening port to 443, and then click Next.

- 7. Configure the SSL certificate.
 - a. Click Create Server Certificate.

onfigure Server Load Balan	DCER 🗇 Back			
Protocol and Listener	SSL Certificates	Backend Servers	Health Check	Submit
Configure SSL Certificates				
(i) Configure SSL certificates to ensure	e that your business is protected by e	encryptions and authenticated by a trusted cer	tificate authority.	
Select Server Certificate				
Select		✓ Create Server Ce	Ruy Certificate	
Advanced Modify ⊗				
Enable Mutual Authentication Disat	bled	CA Certificate	None Selected	
Previous Next	Cancel			

- b. On the displayed page, select a certificate source. In this example, select Upload Third-Party Certificate, and then click Next.
- c. On the uploading third-party certificate page, set the certificate name and select the region in which the certificate is deployed. In the Certificate Content and

the Private Key columns, enter the server public key certificate and private key created in *Prerequisites*, and then click OK.

Upload Third-Party Certificate ×									
Certificate Name Cert-tomcat									
Regions China East 1 (Hangzhou) ×	~								
Certificate Content C	•								
(NGINX-compatible) Upload View Sample Certificate									
• Private Key: ⑦ 22 €	•								
(NGINX-compatible) Upload View Sample Certificate									
Previous OK Cance	əl								
- d. From the Select Server Certificate drop-down list, select the created server certificate.
- e. Click Next.
- 8. Configure Backend Servers. By default, servers are added. You need to configure a port for each backend server to listen to the tomcat-svc service, and then click Next.

Note:

You need to find the NodePort number of this service in the Container Service Web interface, and configure the number as the port number of each backend server.

Protocol and Listener	SSL Certificates		Backend Servers		Health Check	\rightarrow	Submit		
Add Backend Servers									
Add backend servers to handle the acces	ss requests received by the SL	B instance.							
rward Requests To									
Default Server Group VSe	rver Group Active/S		D						
ervers Added									
ECS Instance ID/Name	Public/Internal IP Address	Port		Weight		Actions			
node-0003-k8s-for-cs- cad7e15c0784848a5be02443e9186ccb7 i-bp1a4u8paol71d36zfg1	192.168.0.78(Private) vpc-bp1lkyevdjjerjqs0u4vb vsw- bp1qb1yn2nbzm4kck66xx	32529		100		Delete			
node-0001-k8s-for-cs- cad7e15c0784848a5be02443e9186ccb7 i-bp1hk1m08e5rxkgrae8a	192.168.0.37(Private) vpc-bp1lkyevdjjerjqs0u4vb vsw- bp1qb1yn2nbzm4kck66xx	32529		100		Delete			
node-0002-k8s-for-cs- cad7e15c0784848a5be02443e9186ccb7 i-bp1hk1m08e5rxkgrae8b	192.168.0.38(Private) vpc-bp1lkyevdjjerjqs0u4vb vsw- bp1qb1yn2nbzm4kck66xx	32529		100		Delete			
4 servers have begin tided. D servers are to be added, and 1 servers are to be deleted. Add More									

- 9. Configure Health Check, and then click Next. In this example, use the default settings.
- 10.Confirm the Submit tab. When you make sure that all configurations are correct, click Submit.

11.After completing the configuration, click OK.

Configure Server Load Balance	r 5 Back		
Protocol and Listener	Backend Servers	Health Check	Submit
Submit			
Layer-7 listener	Success		
Start Listener	Success		
OK Cancel			

- 12.Return to the Server Load Balancer page to view the instance. The listening rule of HTTPS : 443 is generated.
- 13.Access the Tomcat application through HTTPS. In the address bar of the browser, enter https://slb_ip to access the application.

Note:

If the domain name authentication is included in the certificate, you can access the application by using the domain name. You can also access the application through slb_ip : 8080 because tcp : 8080 is not deleted.



Method 2: Configure the certificate on Ingress

This method has the following advantages and disadvantages:

- Advantages: You do not need to modify the SLB configuration. All applications can manage their own certificates through Ingress without interfering with each other.
- Disadvantages: Each application can be accessed by using a separate certificate or the cluster has applications that can be accessed by only using a certificate.

Preparations

You have created a Tomcat application in the Kubernetes cluster. The service of the application provides access through ClusterIP. In this example, use Ingress to provide the HTTPS access service.

Example

1. Log on to the Master node of the Kubernetes cluster and create a secret according to the prepared certificate.



You must set the domain name properly. Otherwise, you will encounter exceptions when accessing the application through HTTPS.

kubectl create secret tls secret - https -- key tls .
key -- cert tls . crt

- 2. Log on to the Container Service console.
- 3. In the left-side navigation pane, click Application > Ingress, select a cluster and namespace, and click Create in the upper-right corner.

4. In the displayed dialog box, configure the Ingress to make it accessible through HTTPS, and then click OK.

For more information about Ingress configuration, see *Create an Ingress in the Container Service console*. The configuration in this example is as follows:

- Name: Enter an Ingress name.
- Domain: Enter the domain name set in the preceding steps. It must be the same as that configured in the SSL certificate.
- Service: Select the service corresponding to the tomcat application. The service port is 8080.
- Enable TLS: After enabling TLS, select the existing secret.

Create		\times
Name:	tomcat-https	
Rule:	 Add Domain foo.bar.com Select *.cd5f29d03dcd544d3943a4c2cb45bb4ec.cn-hangzhou.alicontainer.com or Custom path 	
	e.g./ Service • Add Name Port Weight Percent of Weight	
	tomcat-svc ▼ 8080 ▼ 100 100.0% ● ✓ EnableTLS ● Exist secret ○ Create secret	
	secret-https •	
Service weight: Grayscale release:	 Enable Add After the gray rule is set, the request meeting the rule will be routed to the new service. If you set a weight other than 100, the request to satisfy the gamma rule will continue to be routed to the new and old version services according to the weights. 	
annotation:	 Add rewrite annotation 	
Tag:	Opprovide the second se	
	Create C	ancel

You can also use a YAML file to create an Ingress. In this example, the YAML sample file is as follows:

```
apiVersion : extensions / v1beta1
kind : Ingress
metadata :
   name : tomcat - https
spec :
 tls :
   hosts :
   - foo . bar . com
   secretName : secret - https
  rules :
   host: foo.bar.com
   http :
     paths :
      path : /
       backend :
         serviceNam e : tomcat - svc
         servicePor t: 8080
```

5. Return to the Ingress list to view the created Ingress, the endpoint, and the domain name. In this example, the domain name is foo . bar . com . You can also enter the Ingress detail page to view the Ingress.

Note:

In this example, foo . bar . com is used as a testing domain name, and you need to create a record in the hosts file.

```
47.110.119.203 foo.bar.com # where
, the IP address is the Ingress endpoint.
```

Ingress		Refresh Create
Help: 🔗 Blue-green release		
Clusters k8s-test111 v Namespace default v		
Name Endpoint Rule	Time Created	Action
tomcat-https foo.bar.com/ -> tomcat-svc	11/07/2018,15:37:00	Details Update View YAML Delete

6. In the browser, access https :// foo . bar . com .



You need to access the domain name by using HTTPS because you have created a TLS access certificate. This example uses foo . bar . com as a sample domain name to be parsed locally. In your specific configuration scenarios, you need to use the registered domain names.



1.16 Release management

1.16.1 Manage a Helm-based release

Alibaba Cloud Container Service for Kubernetes is integrated with the package management tool Helm to help you quickly deploy applications on the cloud. However, Helm charts can be released multiple times and the release version must be managed. Container Service for Kubernetes provides a release function, which allows you to manage the applications released by using Helm in the Container Service console.

Prerequisites

• You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.

• You have installed a Helm application by using the App Catalog function or Service Catalog function. For more information, see *Simplify Kubernetes application deployment by using Helm*. In this topic, the wordpress-default application is used as an example.

View release details

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane, select Container Service Kubernetes. Then, select Application > Release and click the Helm tab. Select the target cluster from the Clusters drop-down list.

In the displayed release list, you can view the applications and services released through Helm in the selected cluster.

Container Service -		Release									Refresh
Namespace		Batch Release Hein	n 4								
- Application		Clusters test-mia									
Deployment	Ŀ	Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time			Action
StatefulSet	L	wordpress-default	Deployed	default	wordpress	0.6.13	4.8.2	11/23/2018,11:43:51	Details	Update	Delete
Job	Ξ	wordpress-test	Deployed	default	wordpress	4.0.0	4.9.8	11/23/2018,10:49:40	Details	Update	Delete
CronJob		istio	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/26/2018,18:18:28	Details	Update	Delete
Pods	L	alibaba-log-controller	Deployed	default	alibaba-cloud-log	0.1.1	1.0	10/22/2018,15:20:34	Details	Update	Delete
Service											
Ingress											
Volumes Claim											
Release 3											

3. Find your target release (wordpress-default in this example) and click Details to view the release details.

You can view such release details as the current version and history version. In this example, the current version is 1 and no history version exists. On the Resource tab page, you can view the resource information of wordpress-default, such as the resource name and the resource type, and view the YAML information.



You can view the running status of the resource in details by clicking the resource name and going to the Kubernetes dashboard page.

Container Service	Release List - wordpress-default				Refresh
Clusters	Current Version				
Clusters Nodes	Release Name : wordpress-default		Namespace : default	Deployed at : 04/20/2018	15:27:55
Storage	Current Version: 1				Time Updated : 04/20/2018,15:27:55
 Application 		Resource		Value	35
Deployment	Resource		Kind		
Service	wordpress-default-mariadb		Secret		View YAML
Release	wordpress-default-wordpress		Secret		View YAML
Config Maps	wordpress-default-mariadb		ConfigMap		View YAML
✓ Store	wordpress-default-mariadb		PersistentVo	olumeClaim	View YAML
App Catalog	wordpress-default-wordpress		PersistentVo	olumeClaim	View YAML
Service Catalog	wordpress-default-mariadb		Service		View YAML
	wordpress-default-wordpress		Service		View YAML
	wordpress-default-mariadb		Deployment	t	View YAML
	wordpress-default-wordpress		Deployment	t	View YAML
	History Version				

4. Click the Values tab to view the release parameters.

Container Service	Release List - wordpress-default			Refresh						
✓ Clusters	Current Version									
Clusters Nodes	Release Name : wordpress-default	Namespace : default	Deployed at : 04/20/2018,15:27:55							
Storage	Current Version: 1			Time Updated : 04/20/2018,15:27:55						
 Application 	Resource		Values							
Deployment Service Release Config Maps • Store App Catalog Service Catalog	<pre>1 ## Bitnami MondPress image version 2 ## ref: https://hub.docker.com/r/bitnami/wordpress/tags 4 # 4 image: bitnami/wordpress:4.8.2-70 5 ## specify a imagePullPolicy 7 ## ref: http://kubernetes.io/docs/user-guide/images/#pn 8 # 9 imagePullPolicy: IfNotPresent 10 ## User of the application 11 ## User of the application 12 ## ref: https://github.com/bitnami/bitnami-docker-wordp 13 ## 14 wordpressUsername: user 15 ## Application password 17 ## Defaults to a random 10-character alphanumeric strin 18 ## ref: https://github.com/bitnami/bitnami-docker-wordp 19 ## 20 # wordpressPassword: 21 ## Admin email 23 ## ref: https://github.com/bitnami/bitnami-docker-wordp 24 ## admin email 25 ## ref: https://github.com/bitnami/bitnami-docker-wordp 24 ## comercisely wordpowerbe wordpowerb</pre>	/ e-pulling-images ress#environment-variable g if not set ress#environment-variable ress#environment-variable	s s							

Update a release version

1. Log on to the Container Service console.

2. In the left-side navigation pane, select Container Service - Kubernetes. Then, select Application > Release and click the Helm tab. Select the target cluster from the Clusters drop-down list.

In the displayed release list, you can view the applications and services released through Helm in the selected cluster.

Container Service - Kubernetes -	0	Release									Refresh
Namespace	1	Batch Release Hel	m 4								
- Application 2		Clusters test-mia	5 .								
Deployment	a.	Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time			Action
StatefulSet	L	wordpress-default	Deployed	default	wordpress	0.6.13	4.8.2	11/23/2018,11:43:51	Details	Update	Delete
Job	E	wordpress-test	Deployed	default	wordpress	4.0.0	4.9.8	11/23/2018,10:49:40	Details	Update	Delete
CronJob		istio	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/26/2018,18:18:28	Details	Update	Delete
Pods	L	alibaba-log-controller	Deployed	default	alibaba-cloud-log	0.1.1	1.0	10/22/2018,15:20:34	Details	Update	Delete
Service	P.										
Ingress											
Volumes Claim											
Release											

3. Find your target release (wordpress-default in this example). Click Update and the Update Release dialog box appears.

	Update Release X		
Release			Refresh
Batch Release Helm	1 + ## Bitnami NordPress image version 2 ## ref: 3 ## 4 image: bitnami/wordpress:4.8.2-r0		
Clusters test-mia	5 6 * ## Specify a imagePullPolicy 7 ## ref: http://kubernetes.io/docs/user-guide/images/#pre-pulling-images 8 ##		
Release Name	9 imagePullPolicy: IfNotPresent 10		Action
wordpress-default	11 * ## User of the application 12 ## ref: https://github.com/bitnami/bitnami-docker-wordpress#environment-variables 13 ## 14 \wordpressUsername: user	43:51 Details	Update Delete
wordpress-test	15 16 - ## Application password	49:40 Details	Update Delete
istio	17 ## Defaults to a random 10-character alphanumeric string if not set 18 ## ref: https://github.com/bitnami/bitnami-docker-wondpress#environment-variables 19 ## 20 # wondpressPassword:	18:28 Details	Update Delete
alibaba-log-controller	21 22 - ## Admin email 23 ## ref: https://github.com/bitnami/bitnami-docker-wordpress#environment-variables 24 ##	20:34 Details	Update Delete
	<pre>25 wordpressEmail: user@example.com 26 27 ## First name 28 ## ref: https://github.com/bitnami/docker-wordpress#environment-variables 29 ## 29 ## 29 ## 20 ## ist name 31 ## ref: https://github.com/bitnami/bitnami-docker-wordpress#environment-variables 33 ## ref: https://github.com/bitnami/bitnami-docker-wordpress#environment-variables 34 ## 35 wordpresslastName: LastName </pre>		
	Update Cancel		

4. Modify the parameters and then click Update.

Update Release	\times
<pre>## ## ## ## ## ## ## ## ## ## ## ## ##</pre>	•
Update	Cancel

On the release list page, you can see that the current version changes to 2. To roll back to version 1, click Details and in the History Version area, click Rollback.

Current Version		
Release Name : wordpress-default	Namespace : default	Deployed at : 04/20/2018,17:45:35
Current Version : 2		Time Updated : 04/20/2018,17:45:46
Resource		Values
Resource	Kind	
wordpress-default-mariadb	Secret	View YAML
wordpress-default-wordpress	Secret	View YAML
wordpress-default-mariadb	ConfigMap	View YAML
wordpress-default-mariadb	PersistentVolumeClaim	View YAML
wordpress-default-wordpress	PersistentVolumeClaim	View YAML
wordpress-default-mariadb	Service	View YAML
wordpress-default-wordpress	Service	View YAML
wordpress-default-mariadb	Deployment	View YAML
wordpress-default-wordpress	Deployment	View YAML
History Version		
Version : 1 Rollback		Time Updated : 04/20/2018,17:45:35

Delete a release

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane, select Container Service Kubernetes. Then, select Application > Release and click the Helm tab. Select the target cluster from the Clusters drop-down list.

In the displayed release list, you can view the applications and services released through Helm in the selected cluster.

Container Service - Kubernetes -	1	Release								I	Refresh
Namespace	*	Batch Release Heir	4								
✓ Application 2		Clusters test-mia	·								
Deployment	1	Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time			Action
StatefulSet		wordpress-default	Deployed	default	wordpress	0.6.13	4.8.2	11/23/2018,11:43:51	Details	Update	Delete
Job	Ę	wordpress-test	Deployed	default	wordpress	4.0.0	4.9.8	11/23/2018,10:49:40	Details	Update	Delete
CronJob		istio	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/26/2018,18:18:28	Details	Update	Delete
Pods	L	alibaba-log-controller	Deployed	default	alibaba-cloud-log	0.1.1	1.0	10/22/2018,15:20:34	Details	Update	Delete
Service											
Ingress											
Volumes Claim											
Release 3											

3. Find your target release (wordpress-default in this example). Click Delete and the Delete dialog box appears.

Release List	Delete X	Refrach
Trobase Ear		- Nor Gari
Clusters test 🔻	Are you sure to delete the release wordpress-delatit ? Purge	
Release Name Status		Update Time 1 Action
wordpress-default	2 OK Cancel	04/20/2018,17:52:59 Detail Update Delete

4. Select the Purge check box if you want to clear the release records, and then click OK. After you delete a release, the related resources such as the services and deployments are deleted too.

1.16.2 Use batch release on Alibaba Cloud Container Service for Kubernetes

You can use Alibaba Cloud Container Service for Kubernetes to release application versions in batches, achieving fast version verification and rapid iteration of applications.

Context



The latest Kubernetes cluster has installed alicloud-application-controller by default. For older versions of clusters, only versions of 1.9.3 and later are currently supported, and you can upgrade old versions of clusters through the prompt link on the console.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Application > Release in the left-side navigation pane. Click Create batch release in the upper-right corner.

Note:

If the button is gray, you can upgrade the cluster by following the upgrade link.



3. Configure batch release information, including the application name, cluster, namespace, and options. Click Next.

Create batch release	* Return distribution list				
	Batch release information		Batch publishing Configuration	>	Done
Name:	01-batch				
	The name should be 1-64 characters long, an	d can contain numbers, I	English letters and hyphens, but cannot start with a	hyphen.	
Cluster:	k8s-test v				
Namespace :	default 🔻				
Options :	Points 2 wholesale cloth, 1st batch Sus; *				
					Back Next

4. On the batch publishing configuration page, configure the backend pod and service, and then click Update to create an application.

	Batch release information	Batch publishing Configuration	>	Done
The	wizard mode			
_			_	
	Image Name: nginx	Image Version: latest Always pull image		
	Select image	Select image version		Overview
ral	Scale: 4			Gameral
Gene	Resource Limit: CPU eq : 500m Memory eq : 128Mi	0	-	Access Control
	og i soon			Container
	Resource Request: CPU eg : 500m Memory eg : 128Mi	0		Volume
				Environment
ſ	Service: O ClusterIp O NodePort ® Server Load Balancer	ad Balancer will be created		Prev Update
	public inner			3
Introl	Port Mapping:			
ess Co	Port t	arget port P	rotocol	
Acc	80	80	тср 🔹 💿 🚩	
L	O Add			
	VAdvar	ce Config		

5. Return to the release list, an application is displayed in the Not started status. Click Detail on the right.

Release List				Refresh	Create batch release
Batch release					
Clusters k8s-test v Namesp	ace default v				
Release Name	Namespace	Update Time	Status		Action
01-batch	default	2018-09-03 16:00:56	Not started	Deta	ail Update Delete

6. On the application detail page, you can view more information. Click Change Configuration in the upper-right corner of the page to make a batch release change.

Release List - nginx					Refresh	Change Configuration
Details History						
Overview						
Release Name:	nginx					
Release Type:	Batch Release					
Created At:	2018-09-28 16:24:04					
RelatedService:	batchrelease-nginx-svc					
Status:	Not Started					
Not Started In Progres	s Completed				Refresh Continue	Roll Back Complete
Name	App Version	Status	Pod IP	Time Created		Action
batchrelease-nginx-0	V1	Running	172.16.1.159	09/28/2018,16:24:04		Terminal Logs
batchrelease-nginx-1	v1	Running	172.16.1.160	09/28/2018,16:24:07		Terminal Logs
batchrelease-nginx-2	v1	Running	172.16.1.161	09/28/2018,16:24:10		Terminal Logs
batchrelease-nginx-3	v1	Running	172.16.1.162	09/28/2018,16:24:12		Terminal Logs 😝

Issue: 20190525

7. Configure changes for the new version of the application, and then click Update.

Rele	ase List - 01-batch				
The	e wizard mode				
	Image Name:	nginx	Image Version:	latest Always pull image	Overview
General	Scale: Resource Limit:	Select image 4 CPU eg : 500m Ø Memory eg : 128Ms	0	Select image version	General Access Control Container Volume Environment
	Resource Reques	O Memory eg : 128Mi	0		Prev Update
ner	Command:				
Contai	Args:				
Volume	Data Volume:	0			
ironmer	Environment:	0			

8. By default, you return to the release list page, where you can view the batch release status of the application. After completing the first deployment, click Detail.

Release List	Refresh	Create Batch	Release
Batch Release			
Clusters k8s-test V Namespace default V			
Release Name Namespace Update Time Status		1	Action
nginx default 2018-09-28 16:24:04 Wait for Deployment to Complete (Total Batches: 2. Currently, batch 0 is in process and the batch status is Deploying)		Details Update	Delete

9. You can see that the Not Started list is has two pods and the Completed list has two pods, which indicates that the first batch has been completed in batch release. Click Continue, you can release the second batch of pods. Click Roll Back to roll back to the previous version.

Details History							
Overview							
Release Name:	nginx						
Release Type:	Batch Release						
Created At:	2018-09-28 16:24:0	2018-09-28 16:24:04					
RelatedService:	batchrelease-nginx-	svc					
Status:	To be confirmed (To	otal Batches: 2. Currently	, batch 1 is in process and	the batch status is Completed)			
Not Started In Progress	Completed			Refresh	Continue Roll Back Complete		
Name	App Version	Status	Pod IP	Time Created	Action		
batchrelease-nginx-2	v2	Running	172.16.1.164	09/28/2018,16:26:57	Terminal Logs		
batchrelease-nginx-3	v2	Running	172.16.1.163	09/28/2018,16:26:53	Terminal Logs		

10.When completing the release, click Historyto roll back to history versions.

Release List - nginx		Refresh Change Configuration
Details History		
Release History		
v1	2018-09-28 16:24:04	Roll Back

What's next

You can use batch release to quickly verify your application version without traffic consumption. Batch release is more resource-saving than blue-green release. Currently, batch release can be performed on only web pages. The yaml file editing is to be opened later to support more complex operations.

1.17 Istio management

1.17.1 Overview

Istio is an open platform that provides connection, protection, control and monitors microservices.

Microservices are currently being valued by more and more IT enterprises. Microservices are multiple services divided from a complicated application. Each service can be developed, deployed, and scaled. Combining the microservices and container technology simplifies the delivery of microservices and improves the liability and scalability of applications.

As microservices are extensively used, the distributed application architecture composed of microservices becomes more complicated in dimensions of operation and maintenance, debugging, and security management. Developers have to deal with greater challenges, such as service discovery, load balancing, failure recovery, metric collection and monitoring, A/B testing, gray release, blue-green release, traffic limiting, access control, and end-to-end authentication.

Istio emerged. Istio is an open platform for connecting, protecting, controlling, and monitoring microservices. It provides a simple way to create microservices networks and provides capabilities such as load balancing, inter-service authentication, and monitoring. Besides, Istio can provide the preceding functions without modifying services.

Istio provides the following functions:

- Traffic management: Controls traffic and API calls between services to enhance the system reliability.
- Authentication and security protection: Provides authentication for services in meshes, and protects the traffic of services to enhance the system security.
- Policy execution: Controls access policies between services without requiring changes to the services.
- Observability: Obtains traffic distribution and call relationships between services to quickly locate problems.

Istio architecture

Istio is logically divided into a control plane and a data plane:

- Control plane: Administration proxy (the default is Envoy) for managing traffic routing, runtime policy execution, and more
- Data plane: Consists of a series of proxys (the default is Envoy) for managing and controlling network communication between services.



Istio is composed of the following components:

- Istio Pilot: Collects and validates configurations, and propagates them to various Istio components. It extracts environment-specific implementation details from the policy execution module (Mixer) and the intelligent proxy (Envoy), providing them with an abstract representation of user services, independent of the underlying platform. In addition, traffic management rules (that is, generic Layer-4 rules and Layer-7 HTTP/gRPC routing rules) can be programmed through Pilot at runtime.
- Policy execution module (Mixer): Executes access control and usage policies across the service mesh, and collects telemetry data from the intelligent proxy (Envoy) and other services. Mixer executes policies based on the Attributes provided by the intelligent proxy (Envoy).
- Istio security module: Provides inter-service and inter-user authentication to guarantee enhanced security between services without modifying service codes. Includes three components:
 - Identification: When Istio runs on Kubernetes, it identifies the principal that runs the service according to the service account provided by container Kubernetes.
 - Key management: Provides CA automated generation, and manages keys and certificates.
 - Communication security: Provides a tunnel between the client and the server through the intelligent proxy (Envoy) to secure services.
- Intelligent proxy (Envoy): Deployed as an independent component in the same Kubernetes pod along with relevant microservice, and provides a series of attributes to the policy execution module (Mixer). The policy execution module (Mixer) uses these attributes as the basis to execute policies, and sends them to monitoring systems.

1.17.2 Deploy Istio

The distributed application architecture composed of microservices has disadvantages in aspects such as operation and maintenance (O&M), debugging, and security management. To eliminate the disadvantages, you can deploy Istio to create microservice network and to provide load balancing, service-to-service authentication, monitoring, and other functions. Istio provides the functions without requiring any changes to services.

Prerequisites

- A Kubernetes cluster is created. For more information, see *Create a Kubernetes cluster*.
- You have logged on to the Container Service console by using the primary account or by using a sub-account that has been granted sufficient permissions. For example, if the cluster - admin permission is granted to a sub-account then Istio can be deployed. Other combinations of permissions are also sufficient. For more information, see *Grant Kubernetes permissions to a RAM user*.

Background information

- Alibaba Cloud Container Service for Kubernetes in versions of 1.10.4 and later support Istio deployment. If your Container Service for Kubernetes is in any version prior to 1.10.4, update the version to 1.10.4 or later.
- To guarantee sufficient resources, the number of Worker nodes in a cluster must be greater than or equal to 3.

Procedure

You can deploy Istio through the Clusters page or through the App Catalog page.

- Deploy Istio through the Clusters page
 - 1. Deploy Istio.
 - a. Log on to the Container Service console.
 - b. In the left-side navigation pane, choose Clusters > Clusters.
 - c. On the right of the target cluster, choose More > Deploy Istio.

Cluster Name/ID	Cluster Type	Region (All) 👻	Network Type	Cluster Status	Number of Nodes	Time Created	Kubernetes Version	Action
	Kubernetes	China North 2 (Beijing)	VPC vpc- 2ze6clfm15d	Running	6	10/22/2018,19:41:50	1.11.2	Manage View Logs Dashboard Scale Cluster <u>More</u> -
	Kubernetes	China North 2 (Beijing)	VPC vpc- 2ze8sk2raka	Failed to delete	4	09/18/2018,14:00:45	1.11.2	Delete Add Existing Instance Upgrade Cluster
								Automatic Scaling Addon Upgrade Deploy Istio

d. Set the following Istio parameters.

Configuration	Description
Clusters	Target cluster in which Istio is deployed.
Namespace	Namespace in which Istio is deployed.

Configuration	Description
Release Name	Name of Istio to be released.
Enable Prometheus for metrics/ logs collection	Whether to enable Prometheus for metrics/logs collection. Enabled by default.
Enable Grafana for metrics display	Whether to enable Grafana for metrics display. Enabled by default.
Enable automatic Istio Sidecar injection	Whether to enable automatic Istio Sidecar injection. Enabled by default.
Enable the Kiali Visualization Service Mesh	 Whether to enable the Kiali Visualization Service Mesh. Disabled by default. Username: Set a user name. The default is admin. Password: Set a password. The default is admin.

Configuration	Description
Tracing Analysis Settings	- Enable Distributed Tracing with Jaeger : indicates whether to enable Jaeger (the distributed tracing system). To use Jaeger, select this radio button, and activate Alibaba Cloud Log Service.
	Note: If you select this radio button, Log Service will automatically create a project named istio - tracing -{ ClusterID } that is used to store the tracking data.
	 Activate Tracing Analysis: inditcates whether to activate the Tracing Analysis service. To activate this service, select this radio button, and then click Activate now. Additionally, you need to enter an endpoint address in the format of http://tracing - analysis - dc - hz . aliyuncs . com // api / v1 / spans .
	 Note: An address of this format indicates an Internet or intranet endpoint used by a Zipkin client to transmit collected data in a region to the Tracing Analysis service that uses the API from v1 release. If you use an intranet endpoint, you must ensure that your Kubernetes cluster and the <i>Tracing</i> <i>Analysis</i> instance are in the same region to maintain stable network
Pilot Settings	Set the trace sampling percentage in the range of 0 to 100. The default value is 1.

Configuration	Description
Control Egress Traffic	 Permitted Addresses for External Access: range of IP addresses that can be used to directly access services in the Istio service mesh. By default, this field is left blank. Use commas (,) to separate multiple IP address ranges. Blocked Addresses for External Access: range of IP addresses that are blocked against external accesses. By default, this IP address range contains the cluster pod CIDR block and service CIDR block. Use commas (,) to separate multiple IP address ranges. ALL: Select this check box to block all the IP addresses used to access the Internet.
	Note:If the settings of these two parametersconflict with each other, the PermittedAddresses for External Access prevails.For example, if an IP address is listed inboth IP address ranges that you set forthese two parameters, the IP addresscan be still accessed. That is, the settingof Permitted Addresses for ExternalAccess prevails.

e. Click Deploy Istio.

At the bottom of the deployment page, you can view the deployment progress and status in real time.

Step	Status		
Create Istio Resource Definition	Succeeded	53 / 53	
Deploy Istio	₹. Running		

Verify the result

You can view your deployment results in the following ways:

- At the bottom of the Deploy Istio page, Deploy Istio is changed to Deployed.

Step	Status
Create Istio Resource Definition	pending
Deploy Istio	pending
	loyed

■ In the left-side navigation pane, choose Application > Pods.

Select the cluster and namespace in which Istio is deployed, and you can see the relevant pods in which Istio is deployed.

	Container Service - Kubernetes +		Pods							Refresh
	Overview	^	Clusters k8s v Namespace istio-	system 🔻						
•	Clusters	L	Name 🜲	Status 🖨	Pod IP 🜲	Node 🜲	Time Created 🜲	CPU	Memory	
	Clusters	l	grafana-6d786489b9-87288	Running	1.0		10/22/2018,20:14:53	0.006	24.57 Mi	Details More 🗸
	Volumes		istio-citadel-765964b585-nc4lg	Running			10/22/2018,20:14:53	0	10.598 Mi	Details More 🗸
	Namespace	=	istio-egressgateway-5ddcf6d6f4-6p5kw	Running	10,000,000		10/22/2018,20:14:53	0.002	27.918 Mi	Details More 🗸
	Application		istio-egressgateway-5ddcf6d6f4-wtw4t	Running			10/22/2018,20:18:38	0.002	26.758 Mi	Details More 👻
	StatefulSet		istio-galley-85666b6578-qlnfq	Running			10/22/2018,20:14:53	0.017	11.895 Mi	Details More +
	Pods		istio-ingressgateway-85689f5c5-g48q2	Running		10.00	10/22/2018,20:14:53	0.002	25.086 Mi	Details More +

- In the left-side navigation pane, choose Application > Service.

Select the cluster and namespace in which Istio is deployed, and you can see the access addresses provided by the relevant services in which Istio is deployed.

Container Service - Kubernetes -		Service List							Refresh	Create
Overview	^	Help: 🖉 Canary release								
 Clusters 	L	Clusters k8s 🔻 Nar	nespace istio-	system 🔻						
Clusters	I.	Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint			Action
Nodes	l	grafana	ClusterIP	10/22/2018,20:14:52		grafana:3000 TCP	-	Details Update	View YAML	Delete
Volumes	l	istio-citadel	ClusterIP	10/22/2018,20:14:53	10.000	istio-citadel:8060 TCP Istio-citadel:9093 TCP	-	Details Update	View YAML	Delete
Namespace		istio-egressgateway	ClusterIP	10/22/2018,20:14:52		istio-egressgateway:80 TCP istio-egressgateway:443 TCP	-	Details Update	View YAML	Delete
 Application 	E	istio-galley	ClusterIP	10/22/2018,20:14:52		istio-galley:443 TCP istio-galley:9093 TCP	-	Details Update	View YAML	Delete
Deployment StatefulSet Pods	l					istio-ingressgateway:80 TCP istio-ingressgateway:31380 TCP istio-ingressgateway:443 TCP istio-ingressgateway:31390 TCP istio-ingressgateway:31400 TCP				
Service Ingress	l	istio-ingressgateway	io-ingressgateway LoadBalancer 10/22/20		/2018,20:14:52	istio-ingressgateway:31400 TCP istio-ingressgateway:15011 TCP istio-ingressgateway:32009 TCP istio-ingressgateway:8060 TCP	Details Update V		View YAML Delete	
Volumes Claim						istio-ingressgateway:853 TCP istio-ingressgateway:32512 TCP istio-ingressgateway:15030 TCP				
Helm						istio-ingressgateway:30282 TCP istio-ingressgateway:15031 TCP istio-ingressgateway:31006 TCP				

2. Create the Istio Ingress gateway.



By default, the system does not create the Istio Ingress gateway after you have deployed Istio.

- a. In the left-side navigation pane, choose Store > App Catalog.
- b. Click ack-istio-ingressgateway.



c. Click the Values tab, and then set the parameters.



Note:

- For more information about the description, the values, and the default parameters, see the Configuration section on the Readme tab page.

- You can set customized parameters, including indicating whether to enable a specific port, or whether to use the intranet SLB or the Internet SLB by setting the serviceAnn otations parameter.
- d. In the Deploy area on the right, select the target Cluster from the drop-down list, and then click DEPLOY.



Namespace is fixed as istio-system, and Release Name is fixed as istioingressgateway.

Verify the result.

- a. In the left-side navigation pane, choose Application > Pod.
- b. Select the target cluster and the istio-system namespace to view the pod to which the Istio Ingress gateway has been deployed.

Overview Clusters Clusters Search By Name Q Clusters Name Status Mame Status Status Pod IP Node Grafana-6f5c47c6d-8h94s Grafana-5f5c47c6d-8h94s Grafana-5f5c4f5c47c6d-8h94s G	
 Clusters Name Status	
Clusters Name () Status () attempts Pod IP () Node () Memory Nodes grafana-65c4/2/c6d-8h84s () () () () () () () Volumes grafana-55/c4/2/c6d-8h84s () () () () () () ()	
Nodes grafana-6f5c4/2/c6d-8h84s Running 0 04/01/2019;16:24:12 0.003 Details Volumes grafana-15.2.3 Running 0 04/01/2019;16:24:12 0.003 22.297 Mi More	
Volumes	s I
Namespace istio-ctadel-6d658fcf88-r/pmt Details Details Details w ctadel:1.0.5 0 04/01/2019;16:24:12 0.36 9.441 Mi More	s
Authorization	
★ Application	5 •
Deployment	
StatefulSet Istio-grafana-post-install-world Succeeded 0 04/01/2019.16:24:13 Details	5
DaemonSet More	•
Job Istio-Ingressgateway- 84J4865fb-2tz55 Running 0 04/01/2019;16:21:52 0.002 29:48 Mil	s
Cronlob Note	_
Pods istio-init-crd-10-csqs4 Occeeded 0 04/01/2019.16:23:44	s
kubecti:1.0.5 More	÷

- Deploy Istio through the App Catalog page
 - 1. Deploy the CRD of Istio.
 - a. Log on to the Container Service console.
 - b. In the left-side navigation pane, choose Store > App Catalog.

Container Service - Kubernetes 🔻	App Catalog			
Overview				
 Clusters 	D O	D O	D O	D O
 Application 		U		
 Discovery and Load B 	ack-arms-pilot	ack-consul	ack-hyperledger-fabric	ack-istio
 Configuration 	1.0.1 incubator	0.5.0 incubator	1.4.0 incubator	1.0.5 incubator
▼ Store	•	•	~	•
Docker Images	O O	Q S	Q O	Q O
Orchestration Templa		0	0	0
App Catalog	ack-istio-certmanager	ack-istio-coredns	ack-istio-customgateway	ack-istio-egressgateway
Service Catalog		1.0.0 30000	100 0000	10.0 0000
			0	0
	Ω_0°	$\mathbf{Q}_{\mathbf{O}}^{U}$	Φö	QÖ
	ack-istio-ingressgateway	ack-istio-init	ack-istio-remote	ack-knative-build
	1.0.5 stable	1.0.5 stable	1.0.5 incubator	0.3.0 incubator

c. Click ack-istio-init.

	Container Service - Kubernetes 👻	App Catalog -	ack-istio-init	
*	Overview Clusters Application		ack-istio-init stable Helm chart to Initialize Istio CRDs	
•	Configuration	Readme V	alues	Deploy
•	Store Docker Images Orchestration Templa	1 - global: 2 # Def 3 # Rel 4 # Dai 5 hub: 6 7 # Def	ult hub for Istio images. ases are published to docker hub under 'istio' project. y bulks from spone are on ger.io, and nightly builds from circle on docker.io/istionightly ggistry.on-hangzhou.aliyunS.com/liecs-app-catalog ult mag for Istio images.	Only Kubernetes versions 1.8.4 and above are supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list
	App Catalog	8 tag: 9	.0.5	Clusters
	Service Catalog	11 # loc 12 # TOD 13 image 14	It tests require ifflotPresent, to avoid uploading to dockerhub. Switch to Always as default, and override in the local tests. ullPolicy: ifflotPresent	Vamespace Istio-system Release Name
				istio-init
				DEPLOY

d. In the Deploy area on the right, select the target Cluster from the drop-down list, and then click DEPLOY.



Namespace is fixed as istio-system, and Release Name is fixed as istio-init.

Verify the result.

- a. In the left-side navigation pane, choose Application > Pods.
- b. Select the target cluster and the istio-system namespace to view the pod to which the Istio CRD has been deployed.

Container Service - Kubernetes 👻		Pods							Refresh
Overview	1	Clusters k8s v Namespace istio-	system 🔻						
 Clusters 		Name 🗢	Status 🗢	Pod IP 🜲	Node 🜲	Time Created 🜲	CPU	Memory	
Clusters	I	grafana-6d786489b9-87288	Running	1.00		10/22/2018,20:14:53	0.006	24.57 Mi	Details More 🗸
Volumes		istio-citadel-765964b585-nc4lg	Running			10/22/2018,20:14:53	0	10.598 Mi	Details More -
Namespace Authorization	-	istio-egressgateway-5ddcf6d6f4-6p5kw	Running			10/22/2018,20:14:53	0.002	27.918 Mi	Details More 👻
 Application 		istio-egressgateway-5ddcf6d6f4-wtw4t	Running			10/22/2018,20:18:38	0.002	26.758 Mi	Details More -
StatefulSet		istio-galley-85666b6578-qlnfq	Running	-		10/22/2018,20:14:53	0.017	11.895 Mi	Details More 👻
Pods		istio-ingressgateway-85689f5c5-g48q2	Running		10.00	10/22/2018,20:14:53	0.002	25.086 Mi	Details More 🗸

- 2. Deploy Istio.
 - a. In the left-side navigation pane, choose Store > App Catalog.
 - b. Click ack-istio.



c. Click the Values tab, and then set the parameters.

Container Service - Kubernetes +	App Catalog - ack-istio	
Overview Clusters Application	ack-istio incubator Helm chart of all istio components for Kubernetes on Alibaba Cloud Container Service	
Configuration	Readme Values	Deploy
 Store Docker Images Orchestration Templa 	1 # Common settings. 2 < global: 3 # Default hub for Istic images. 4 hub: registry.cn-hangzhou.aljyuncs.com/aliacs-app-catalog 5 # Default tag for Istic images. 7 tag: 1.0.5	Only Kubernetes versions 1.8.4 and above are supported. For clusters of version 1.8.1, you can perform "upgrade cluster" operation in the cluster list
App Catalog	 9 # Gateway used for legacy k8s Ingress resources. By default it is 10 # using 'istioningress', to match 0.8 config. It requires that 11 # ingress.embled is set to true. You can also set it 	test-01
Service Catalog	<pre>12 * Conservation and other gateway you define in the "gateway' 13 # gateman' 14 k&ingressBetter: ingress 15 15 16 # KeOingressThey will add port 443 on the ingress and ingressgateway. 16 # It REQUIRES that the certificates are installed in the 18 # expected secrets - enabling this option without certificates 19 # will result in DS rejection and the ingress will not work. 28 &&SingressHtps: false 21 22 proxy: 23 image: proxy2 24 24 25 26 27 27 28 28 29 29 29 29 29 29 20 20 20 20 20 20 20 20 20 20 20 20 20</pre>	Namespace Istio-system Release Name Istio DEPLOY

Note:

- For more information about the description, the values, and the default parameters, see the Configuration section on the Readme tab page.
- You can set customized values for parameters, including the grafana, prometheus, tracing, kiali, and other parameters, to better meet your requirements.
- d. In the Deploy area on the right, select the target Cluster from the drop-down list, and then click DEPLOY.



Namespace is fixed as istio-system, and Release Name is fixed as istio.

e. Click DEPLOY.

Verify the result.

- Verify that Istio has been deployed to a pod.
 - a. In the left-side navigation pane, choose Application > Pods.
 - b. Select the target cluster and namespace to view the pod to which Istio has been deployed.

Container Service - Kubernetes ▼		Pods									Refresh
Overview	^	Clusters k8s v Namespace istio-	system 🔻								
 Clusters 	L	Name 🗢	Status 🔶	Pod IP 🜲	Node 🗢	Time Created 🜲	CPU		Memory		
Clusters	L	grafana-6d786489b9-87288	Running	1.00		10/22/2018,20:14:53		0.006	24	I.57 Mi	Details More +
Volumes		istio-citadel-765964b585-nc4lg	Running			10/22/2018,20:14:53		0	10	.598 Mi	Details More 👻
Namespace		istio-egressgateway-5ddcf6d6f4-6p5kw	Running	10,000,000		10/22/2018,20:14:53		0.002	27	.918 Mi	Details More 👻
 Application 		istio-egressgateway-5ddcf6d6f4-wtw4t	Running			10/22/2018,20:18:38		0.002	26	i.758 Mi	Details More 👻
StatefulSet		istio-galley-85666b6578-qlnfq	Running			10/22/2018,20:14:53		0.017	11	.895 Mi	Details More +
Pods		istio-ingressgateway-85689f5c5-g48q2	Running	-		10/22/2018,20:14:53		0.002	25	i.086 Mi	Details More 👻

- Verify that Istio has been deployed to a service.
 - a. In the left-side navigation pane, choose Application > Service.
 - b. Select the cluster and namespace in which Istio is deployed to view the IP addresses provided by the services to which Istio has been deployed.

	Container Service - Kubernetes -		Service List Refresh						Create		
	Overview	*	Help: 🖉 Canary release								
•	Clusters		Clusters k8s v Nam	iespace istio-s	system 🔻						
	Clusters		Name	Туре	Time Created	ClustersIP	InternalEndpoint	ExternalEndpoint			Action
	Nodes		grafana	ClusterIP	10/22/2018,20:14:52		grafana:3000 TCP	-	Details Update	View YAML	Delete
	Volumes	l	istio-citadel	ClusterIP	10/22/2018,20:14:53	1.100	istio-citadel:8060 TCP istio-citadel:9093 TCP	-	Details Update	View YAML	Delete
	Namespace		istio-egressgateway	ClusterIP	10/22/2018,20:14:52		istio-egressgateway:80 TCP istio-egressgateway:443 TCP	-	Details Update	View YAML	Delete
•	Application	E	istio-galley	ClusterIP	10/22/2018,20:14:52		istio-galley:443 TCP istio-galley:9093 TCP		Details Update	View YAML	Delete
	Deployment StatefulSet Pods Service Ingress Volumes Claim Helm		isto-Ingressgateway	LoadBalancer	10/22/2018,20:14:52		istlo-ingressgateway-80 TCP istlo-ingressgateway-3180 TCP istlo-ingressgateway-3180 TCP istlo-ingressgateway-3140 TCP istlo-ingressgateway-31400 TCP istlo-ingressgateway-31400 TCP istlo-ingressgateway-31400 TCP istlo-ingressgateway-31501 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP istlo-ingressgateway-3050 TCP		Details Update	View YAML	Delete
	Release						istio-ingressgateway:31006 TCP				

3. Create the Istio Ingress gateway.

Note:

By default, the system does not create the Istio Ingress gateway after you have deployed Istio.

- a. In the left-side navigation pane, choose Store > App Catalog.
- b. Click ack-istio-ingressgateway.



c. Click the Values tab, and then set the parameters.



Note:

- For more information about the description, the values, and the default of the parameters, see the Configuration section on the Readme tab page.

- You can set customized parameters. For example, you can enable a specific port, or use the intranet SLB or the Internet SLB by set the serviceAnn otations parameter.
- d. In the Deploy area on the right, select the target Cluster from the drop-down list, and then click DEPLOY.



Namespace is fixed as istio-system, and Release Name is fixed as istioingressgateway.

Verify the result.

- a. In the left-side navigation pane, choose Application > Pod.
- b. Select the target cluster and the istio-system namespace to view the pod to which the Istio Ingress gateway has been deployed.

	Container Service - Kubernetes 👻		Pod								Refresh
	Overview	•	Clusters test-01	Namespace	istio-syst	em 🔻				Search By Name	Q
•	Clusters	Ŀ			max						
	Clusters	L	Name 🗢	Status 🔶	attempts	Pod IP 🔶	Node 🔶	Time Created 🔶	CPU	Memory	
	Nodes	l	grafana-6f5c4c7c6d-8h84s grafana:5.2.3	Running	0			04/01/2019,16:24:12	0.003	22.297 Mi	Details
	Volumes	Ŀ									
	Namespace	L	istio-citadel-6d658fcf88-rlpmt citadel:1.0.5	Running	0	1.1		04/01/2019,16:24:12	0.36	9.441 Mi	Details
	Authorization										
•	Application	-	istio-galley-c87c85d69-87snf galley:1.0.5	Running	0	1.0.0		04/01/2019,16:24:12	0	7.664 Mi	Details
	Deployment	Ŀ									
	StatefulSet		istio-grafana-post-install-kvqhf	Succeeded	0			04/01/2019,16:24:13			Details
	DaemonSet		 Tryperkube.v1.7.0_coreos.o 								More 👻
	Job		istio-ingressgateway- 84b4868f6b-2tz55	Running	0		1.11	04/01/2019,16:21:52	0.002	29.48 Mi	Details
	CronJob		proxyv2:1:0:5								More -
	Pods		istio-init-crd-10-csqs4	Succeeded	0			04/01/2019,16:23:44			Details
	Volumes Claim										More 👻

1.17.3 Update Istio

You can modify the deployed Istio through updates.

Prerequisites

· You have created an Kubernetes cluster. For more information, see Create a

Kubernetes cluster.

· You have created an Istio. For more information, see *Deploy Istio*.

Procedure

1. Log on to the Container Service console.

- 2. Under the Kubernetes menu, click Application > Helm in the left-side navigation pane.
- 3. Select a cluster, select the Istio to be updated, and click Update in the action column.



- The release name of the Istio that is deployed through the cluster interface is istio. Configurations to be updated are the same as the options configured in deployment.
- The release name of the Istio that is deployed through the application catalog is the name specified when you create the Istio. Configurations to be updated are the same as the options configured in deployment.

Release List								Refresh
Clusters k8s 🔻								
Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time		Action
ack-istio-default	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:50:57	Details Update	Delete
istio	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:14:52	Details Update	Delete

4. In the displayed dialog box, modify parameters of the Istio, and then click Update.

In this example, update the Istio that is deployed through the cluster interface:



Result

You can view updated content in two ways:

- After you complete the update, the page automatically jumps to theRelease List page. On the Resource tab, you can view updated content.
- Under the Kubernetes menu, click Application > Pods, and select the target cluster and namespace to view updating results.

1.17.4 Delete Istio

You can delete a deployed Istio through the deletingoperation.

Prerequisites

- You have created a Kubernetes cluster. For more information, see *Create a Kubernetes cluster*.
- · You have created an Istio. For more information, see *Deploy Istio*.

Procedure

- 1. Log on to the Container Service console.
- 2. Under the Kubernetes menu, click Application > Helm in the left-side navigation pane.
- 3. Select a cluster, select the Istio to be deleted, and click Delete in the action column.

Release List									Refresh
Clusters k8s 🔻									
Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time			Action
ack-istio-default	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:50:57	Details	Update	Delete
istio	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:14:52	Details	Update	Delete

4. In the displayed dialog box, click OK.

Delete			×
	Are you sure to delete the release istio ? Purge		
		ОК	Cancel
Note:			

• Do not select the Purge box:

- Releasing records are not deleted:

Release List									Refresh
Clusters k8s v									
Release Name	Status	Namespace	Chart Name	Chart Version	App Version	Update Time			Action
ack-istio-default	Deployed	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:50:57	Details	Update	Delete
istio	Deleted	istio-system	ack-istio	1.0.2	1.0.2	10/22/2018,20:14:52	Details	Update	Delete

- The name of this Istio cannot be used again.

When you redeploy the Istio through the cluster interface, the deployment status is deployed.

Step	Status			
Create Istio Resource Definition	pending			
Deploy Istio	pending			

When you redeploy the Istio through the application catalog, the system prompts you that deployment or resource with the same name already exists and please modify the Istio name.



• Selecting the Purge box deletes all releasing records and the Istio name can be reused.

We recommend that you keep the Purge box selected.

Result

Back to the Release List page, you can see that the Istio is removed.

1.17.5 Upgrade Istio components

This topic describes how to upgrade Istio components.

Background information

- The Istio upgrade may install new binaries, and change configurations and API schemas.
- The upgrade process may cause service downtime.
- To minimize downtime, use multiple replicas to ensure that your Istio control plane components and your applications remain highly available.



In the following example, assume that the Istio components are installed and upgraded in the istio-system namespace.

Procedure

To complete the upgrade process, you need to upgrade CRD files, the control plane, and the data plane sidecar.

Upgrade CRD files

- 1. Log on to the Container Service console.
- 2. In the left-side navigation pane under Container Service-Kubernetes, choose Applications > Releases. Then, click the Helm tab, find the istio-init release, and delete it.

Note:

If no release named istio-init is displayed on the Helm tab page, you can directly perform the next step.

- 3. In the left-side navigation pane under Container Service-Kubernetes, choose Store > App Catalog. Then, click ack-istio-init.
- 4. In the Deploy area, select the target cluster. Then, click DEPLOY.



By default, the istio-system namespace is selected, and the release name is set to istio-init.

Upgrade the control plane
The Istio control plane components include the Citadel, Pilot, Policy, Telemetry, and Sidecar injector.

- 1. In the left-side navigation pane under Container Service-Kubernetes, choose Applications > Releases.
- 2. Select the target cluster, select the target Release Name, and click Upgrade in the Action column.
- 3. In the configurations of the deployed Istio, specify the version number of the Istio to be deployed.

Note:

On this page, you can also modify other parameters.

global :
 tag : < enter the version number >

4. Click Update.

Upgrade the data plane sidecar

Note that after you upgrade the control plane, the applications that have already run Istio will still use the sidecar of an earlier version. To upgrade the sidecar, you need to re-inject it.

Automatic sidecar injection

If you use automatic sidecar injection, you can upgrade the sidecar by performing a rolling update for all pods. Then, the sidecar of the new version will be automatically re-injected.

You can use the following script to trigger the rolling update by patching the termination grace period.

```
NAMESPACE =$ 1
     DEPLOYMENT
                   _LIST =$( kubectl - n $ NAMESPACE
                                                             get
deployment - o
                   jsonpath ='{. items [*]. metadata . name }')
     echo " Refreshing
                            pods in
                                          all
                                                 Deployment
                                                              s:$
             _LIST "
DEPLOYMENT
     for
            deployment __name
                                  in $ DEPLOYMENT
                                                       LIST
                                                                  do
# echo " get TERMINATIO N_
deployment : $ deployment _ name "
                                   N_GRACE_PE RIOD_SECON
                                                              ĎS
                                                                    from
     TERMINATIO N_GRACE_PE RIOD_SECON DS =$( kubectl - n
                                                                    $
NAMESPACE get deployment "$ deployment _name " - o jsonpath
='{. spec . template . spec . terminatio nGracePeri odSeconds }')
                                                                 jsonpath
     if [ "$ TERMINATIO N_GRACE_PE RIOD_SECON DS " - eq
                                                                     30
                                                                        ];
then
     TERMINATIO N GRACE PE RIOD SECON DS =' 31 '
     else
     TERMINATIO N GRACE PE
                                RIOD_SECON DS =' 30 '
```

```
fi
   patch_stri ng ="{\" spec \":{\" template \":{\" spec \":{\"
terminatio nGracePeri odSeconds \":$ TERMINATIO N_GRACE_PE
RIOD_SECON DS }}}"
   # echo $ patch_stri ng
   kubectl - n $ NAMESPACE patch deployment $ deployment
_name - p $ patch_stri ng
   done
   echo " done ."
```

Manual sidecar injection

Run the following command to manually upgrade the sidecar:

kubectl apply - f <(istioctl kube - inject - f \$ ORIGINAL_D EPLOYMENT_ YAML)

If the sidecar was previously injected with some customized injection configuration

files, run the following command to manually upgrade the sidecar:

```
kubectl apply - f <( istioctl kube - inject -- injectConf
igFile inject - config . yaml -- filename $ ORIGINAL_D
EPLOYMENT_ YAML )
```

Impacts caused by the Istio upgrade

Impacts caused by the CRD file upgrade

The upgrade process does not impact the calls between services within the cluster or the calls from the gateway to services.

Calls between services within the cluster.



Impacts caused by the control plane upgrade

If HA is enabled, that is, the replicas of Pilot is 2, the HPA setting of istio - pilot
 / istio - policy / istio - telemetry is minReplica s : 2.

If you have changed the Istio version multiple times by upgrading or rolling back the component version, testing results will indicate that the QPS of calls between services remains unchanged and the calls proceed normally.

Φορτίο (fortio) v1.0.1						
List of sa Filter:	ed results: 2018-10-09-072745.f1.to.f2 2018-10-09-072418.f1.to.f2 2018-10-09-072418.f1.to.f2 2018-10-09-071929.f1.to.f2 2018-10-09-071929.f1.to.f2 2018-10-09-071608.f1.to.f2 2018-10-09-071445.f1.to.f2	Graph link: 2018-10-09-072745_f1_t0_f2 son (son)				
		f1 to f2 - Response time histogram at max t min 1.292 ms, average 28.896 ms, p50 22.1	http://echosrv2:8080/echo - 2018-10- arget qps (1658.4 actual) 48 connectio 6 ms, p75 33.86 ms, p90 51.45 ms, p9	09 15:27:45 ns for 30s (actual time 30.4s), no em 9 130.93 ms, p99.9 369.31 ms, max	or 611.09 ms	
8000		-	Cumulative % Histogram:	Count		. 100
0000						100
7000						90
						80
6000	1					
	1					70
5000	1					60
ŧ						
3 4000- S	Н					50 %
3000-	I					40
0000	1					30
2000						50
						20
1000						10
						
0-	100	200	300 4	100	500 6	0 700
			Response time in ms			
Update:	ime axis min ms, max	ms, logarithmic: 🗌 - Count axis logarithmic: 🗌				

Impacts caused by the control plane sidecar upgrade

No obvious change occurs to both the QPS of the calls between services within the cluster and the QPS of the calls from the gateway to services. But these calls will terminate temporarily. We recommend that you use multiple replicas to upgrade the sidecar to reduce the impacts.

1.18 Template management

1.18.1 Create an orchestration template

You can use multiple methods to create orchestration templates through the Container Service console.

Procedure

1. Log on to the Container Service console.

2. Under Kubernetes, click Store > Orchestration Templates in the left-side navigation pane. Click Create in the upper-right corner.

Container Service - Kubernetes 🕶	Templates list	Refresh	Create
Overview	My Template		3
 Clusters 			
 Application 			
▼ Store 1			
Docker Images			
Orchestration Templa	2		
App Catalog			g
Service Catalog			act Us

- 3. In the displayed dialog box, configure the orchestration template, and then click Save. In this example, build a tomcat application template that contains a deployment and a service.
 - Name: Set the template name.
 - Description: Enter the description for the template. This parameter is optional.
 - Template: Configure the template that conforms to Kubernetes yaml syntax rules. The template can contain multiple resource objects that are separated by
 ---.

Create		×
Name:	tomcat The name should be 1-64 characters long, and can contain numbers, English letters, Chinese characters and hyphens.	
Description:	tomcat application	
Template:	<pre>1 apiVersion: apps/vlbeta2 # for versions before 1.8.0 use apps/vlbeta1 2 kind: Deployment 3 metadata: 4 name: tomcat-deployment 5 labels: 6 app: tomcat 7 spec: 8 replicas: 1 9 selector: 10 matchLabels: 11 app: tomcat 12 template: 13 metadata: 14 labels: 15 app: tomcat 16 spec: 17 containers: 18</pre>	
	Save	Cancel

4. After the template is created, the Template List page is displayed. You can see the template under My template.



- 5. Optional: You can also click Application > Deployment in the left-hand navigation pane, and click Create by template to enter the Deploy templates page. Save one of orchestration templates built-in Container Service as your custom template.
 - a) Select a built-in template and click Save Template.

Clusters	test-sis •	•
Namespace	default	•
Resource Type	Resource - basic Deployment	•
Template	<pre>1 apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1 2 kind: Deployment 3 metadata: 4 name: nginx-deployment-basic 5 labels: 6 app: nginx 7 spec: 8 replicas: 2 9 selector: 10 matchLabels: 11</pre>	Add Deployment Deploy with exist template

b) In the displayed dialog box, configure the name, description, and template. After completing the configurations, click Save.



c) Click Store > Orchestration Template, the created template is displayed under My Template.

Templates lis	st		Refresh Create
My Templat	te		
0°	tomcat tomcat application Details	Resource Type / Resource Name Deployment: tomcat-deployment Service: tomcat-svc	Create Application 🔶
Ø ₀	nginx Details	Resource Type / Resource Name Deployment: nginx-deployment-basic	Create Application →

What's next

You can quickly create an application by using the orchestration template under My Template.

1.18.2 Edit an orchestration template

You can edit an orchestration arrangement template.

Prerequisites

You have created an orchestration template, see Create an orchestration template.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > Orchestration Templates. Existing orchestration templates are displayed under My Template.
- 3. Select a template and click Details.



4. Click Edit in the upper-right corner.

Temp	ates list - tomcat Return to template list		Edit Save As	Download Delete
Ċ	tomcat tomcat application			
1 2 3 -	apiVersion: apps/vibeta2 # for versions before 1.8.0 use apps/vibeta1 kind: beployment metadata:	Deploy		
4	name: tomcat-deployment	Clusters		
6	app: tomcat	test-sls		
7 -	spec:	000000		
8		Namespace		
9 -	selector:	default		*
10 -	matchLabels:			
12.	app: tomcat		DEPLOY	
13	metadata:			
14 -	labels:			
15	app: tomcat			
16 -				
17	containers:			
18 -	- name: tomcat			
20	norts:			
21	- containerPort: 8080			

5. In the displayed dialog box, edit the name, description, and template, and click Save.

Modify template		×
Name:	tomcat-V2 The name should be 1-64 characters long, and can contain numbers, English letters, Chinese characters and hyphens.	
Description:	tomcat application	
Template:	<pre>1 apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1 2 kind: Deployment 3 metadata: 4 name: tomcat-deployment 5 labels: 6 app: tomcat 7 spec: 8 replicas: 1 9 selector: 10 matchLabels: 11 app: tomcat 12 template: 13 metadata: 14 labels: 15 app: tomcat 16 spec: 17 containers: 18 - name: tomcat 19 image: tomcat # replace it with your exactly <image_name:tags> 20 ports: 21 - containerPort: 8080</image_name:tags></pre>	
	Save Can	icel

6. Back to the Template List page, under My Template, you can see the template is changed.

Templates lis	t		Refresh Create
My Template			
00 00	tomcat-V2 tomcat application Details	Resource Type / Resource Name Deployment: tomcat-deployment Service: tomcat-svc	Create Application ->
Øo	nginx Details	Resource Type / Resource Name Deployment: nginx-deployment-basic	Create Application ->

1.18.3 Save an existing orchestration template as a new one

You can save an existing template as a new one.

Prerequisites

You have created an orchestration template, see Create an orchestration template.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > Orchestration Templates. Existing orchestration templates are displayed under My Template.
- 3. Select a template and click Details.



4. You can modify the template and click Save as in the upper-right corner.

Templates list - tomcat Return to template list		Edit	Save As	Download	Delete
tomcat tomcat application					
1 apiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1 2 kind: Deployment	Depk	v			
3 metadata:	Boph				
4 name: tomcat-deployment					
labels:	Clus	ters			
6 app: tomcat	test	vele			*
7 speri:		515			
8 replicas: 1	Nam	espace			
9 selector:	4-6				
10 matchLabels:	dera	auic			•
11 app: tomcat					
12 template:			DEPLOY		
13 metadata:					
14 labels:					

5. In the displayed dialog box, configure the template name and click OK.

Save Template As		\times
Name:	tomcat-V3 The name should be 1-64 characters long, and can contain numbers, English letters, Chinese characters and hyphens.]
	ОК	Cancel

6. Back to the Template List page, you can see that the saved template is displayed under My Template.

Templates lis	t		Refresh Create
My Template			
	tomcat-V3	Resource Type / Resource Name	Product and institution
¥0	tomcat application Details	Service: tomcat-svc	Create Application ->

1.18.4 Download an orchestration template

You can download an existing orchestration template.

Prerequisites

You have created an orchestration template, see Create an orchestration template.

Procedure

1. Log on to the Container Service console.

- 2. Under Kubernetes, click Store > Orchestration Templates. Existing orchestration templates are displayed under My Template.
- 3. Select a template and click Details.

Container Service - Kubernetes -	Templates list	:		Refresh Create
	My Template			
Clusters Nodes Volumes Namespace	Øo	tomcat tomcat application Details	Resource Type / Resource Name Deployment: tomcat-deployment Service: tomcat-svc	Create Application →
Authorization • Application	00	nginx Details	Resource Type / Resource Name Deployment: nginx-deployment-basic	Create Application →
Deployment StatefulSet Pods Service Ingress Volumes Claim				
Helm Release Config Maps Secret				Contact U
Orchestration	2			

4. Click Download in the upper-right corner, a template file with yml suffix is downloaded immediately.

Templates list - tomcat 🛛 + Return to template list	Edit Save As Download Delet
tomcat tomcat application	
<pre>1 apiversion: apps/vibeta2 # for versions before 1.8.0 use apps/vibeta1 2 kind: Denloyment</pre>	Baslay
3 metadata:	Depidy
4 name: tomcat-deployment	Shutan
5 labels:	Clusters
6 app: tomcat	test-sls
7 spec:	
8 replicas: 1	Namespace
9 selector:	default
10 matchLabels:	
11 app: comcat	DEPLOY
12 template:	
10 metadata.	

1.18.5 Delete an orchestration template

You can delete an orchestration template that is no longer needed.

Prerequisites

You have created an orchestration template, see Create an orchestration template.

Procedure

- 1. Log on to the Container Service console.
- 2. Under Kubernetes, click Store > Orchestration Template. Existing orchestration templates are displayed under My Template on the Template list page.

3. Select a template and click Detail.

Container Service -	Templates	list		Refresh Create
 Clusters 	My Templa	ate		
Clusters				
Nodes	0 ⁰	tomcat	Resource Type / Resource Name	Create Application 🖈
Volumes	₩ 0	Details	Service: tomcat-svc	
Namespace				
Authorization	¢°	nginx	Resource Type / Resource Name	Create Application ->
 Application 	Ū	Details	Deployment, nginx-deployment-basic	
Deployment				
StatefulSet				
Pods				
Service				
Ingress				
Volumes Claim				
Helm				
Release				
Config Maps				
Secret				
Docker Images				
Orchestration	2			

4. On the detail page of the template, you can click Delete in the upper-right corner.

Temp	ates list - tomcat 🔹 Return to template list		Edit	Save As	Download	Delete
4	tomcat tomcat application					
1 2 3 -	<pre>upiVersion: apps/v1beta2 # for versions before 1.8.0 use apps/v1beta1 cind: Deployment metadata:</pre>	Deploy				
4 5.	name: tomcat-deployment	Clusters				
6	app: tomcat	test-sls				•
7 · 8	spec: replicas: 1	Namespace				
9 · 10 ·	selector:	default				٣
10 11 12 13	app:tuncat template: metalata: metalata:		C	DEPLOY		

5. Click Confirm in the displayed dialog box.

1.19 App catalog management

1.19.1 App catalog overview

Microservice is the theme of container era. The application microservice brings great challenge to the deployment and management. By dividing a large single applicatio n into several microservices, the microservice can be independently deployed and extended so as to realize the agile development and fast iteration. Microservice brings great benefits to us. However, developers have to face the management issues of the microservices, such as the resource management, version management, and configuration management. The number of microservices is large because an application is divided into many components that correspond to many microservices. For the microservice management issues under Kubernetes orchestration, Alibaba Cloud Container Service introduces and integrates with the Helm open-source project to help simplify the deployment and management of Kubernetes applications.

Helm is an open-source subproject in the Kubernetes service orchestration field and a package management tool for Kubernetes applications. Helm supports managing and controlling the published versions in the form of packaging softwares, which simplifies the complexity of deploying and managing Kubernetes applications.

Alibaba Cloud app catalog feature

Alibaba Cloud Container Service app catalog feature integrates with Helm, provides the Helm-related features, and extends the features, such as providing graphic interface and Alibaba Cloud official repository.

The chart list on the App Catalog page includes the following information:

- Chart name: A Helm package corresponding to an application, which contains the image, dependencies, and resource definition required to run an application.
- Version: The version of the chart.
- Repository: The repository used to publish and store charts, such as the official repository stable and incubator.

The information displayed on the details page of each chart may be different and include the following items:

- · Chart introduction
- · Chart details
- Prerequisites for installing chart to the cluster, such as pre-configuring the persistent storage volumes (pv)
- · Chart installation commands
- · Chart uninstallation commands
- · Chart parameter configurations

Currently, you can deploy and manage the charts in the app catalog by using the Helm tool. For more information, see *Simplify Kubernetes application deployment by using Helm*.

1.19.2 View app catalog list

Procedure

1. Log on to the Container Service console.

2. Under Kubernetes, click Store > App Catalog in the left-side navigation pane.

View the charts on the App Catalog page, each of which corresponds to an application, containing some basic information such as the application name, version, and source repository.



What's next

You can click to enter a chart and get to know the detailed chart information. Deploy the application according to the corresponding information by using the Helm tool. For more information, see *Simplify Kubernetes application deployment by using Helm*.