



# OpenShift Container Platform 4.3

## Serverless applications

OpenShift Serverless installation, usage, and release notes



# OpenShift Container Platform 4.3 Serverless applications

---

OpenShift Serverless installation, usage, and release notes

## Legal Notice

Copyright © 2020 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux<sup>®</sup> is the registered trademark of Linus Torvalds in the United States and other countries.

Java<sup>®</sup> is a registered trademark of Oracle and/or its affiliates.

XFS<sup>®</sup> is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL<sup>®</sup> is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js<sup>®</sup> is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack<sup>®</sup> Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

## Abstract

This document provides information on how to use OpenShift Serverless in OpenShift Container Platform

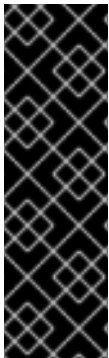
## Table of Contents

<b>CHAPTER 1. GETTING STARTED WITH OPENSIFT SERVERLESS</b>	<b>4</b>
1.1. HOW OPENSIFT SERVERLESS WORKS	4
1.2. APPLICATIONS ON OPENSIFT SERVERLESS	4
<b>CHAPTER 2. OPENSIFT SERVERLESS PRODUCT ARCHITECTURE</b>	<b>5</b>
2.1. KNATIVE SERVING	5
2.1.1. Knative Serving components	5
2.2. KNATIVE CLIENT	5
<b>CHAPTER 3. INSTALLING OPENSIFT SERVERLESS</b>	<b>6</b>
3.1. CLUSTER SIZE REQUIREMENTS	6
3.1.1. Scaling a MachineSet manually	6
3.2. INSTALLING THE OPENSIFT SERVERLESS OPERATOR	7
3.3. INSTALLING KNATIVE SERVING	7
3.4. UNINSTALLING KNATIVE SERVING	8
3.5. DELETING THE OPENSIFT SERVERLESS OPERATOR	9
3.6. DELETING KNATIVE SERVING CRDS FROM THE OPERATOR	9
<b>CHAPTER 4. GETTING STARTED WITH KNATIVE SERVICES</b>	<b>10</b>
4.1. CREATING A KNATIVE SERVICE	10
4.2. DEPLOYING A SERVERLESS APPLICATION	10
4.3. CONNECTING KNATIVE SERVICES TO EXISTING KUBERNETES DEPLOYMENTS	11
<b>CHAPTER 5. CREATING SERVERLESS APPLICATIONS</b>	<b>12</b>
5.1. IMPORTING A CODEBASE FROM GIT TO CREATE AN APPLICATION	12
<b>CHAPTER 6. SPLITTING TRAFFIC BETWEEN REVISIONS</b>	<b>16</b>
6.1. SPLITTING TRAFFIC BETWEEN REVISIONS USING THE DEVELOPER PERSPECTIVE	16
<b>CHAPTER 7. MONITORING OPENSIFT SERVERLESS COMPONENTS</b>	<b>18</b>
7.1. CONFIGURING CLUSTER FOR APPLICATION MONITORING	18
7.2. VERIFYING AN OPENSIFT CONTAINER PLATFORM MONITORING INSTALLATION FOR USE WITH KNATIVE SERVING	18
7.3. MONITORING KNATIVE SERVING USING THE OPENSIFT CONTAINER PLATFORM MONITORING STACK	19
<b>CHAPTER 8. USING METERING WITH OPENSIFT SERVERLESS</b>	<b>20</b>
8.1. INSTALLING METERING	20
8.2. DATASOURCES FOR KNATIVE SERVING METERING	20
8.2.1. Datasource for CPU usage in Knative Serving	20
8.2.2. Datasource for memory usage in Knative Serving	20
8.2.3. Applying Datasources for Knative Serving metering	21
8.3. QUERIES FOR KNATIVE SERVING METERING	21
8.3.1. Query for CPU usage in Knative Serving	21
8.3.2. Query for memory usage in Knative Serving	22
8.3.3. Applying Queries for Knative Serving metering	23
8.4. METERING REPORTS FOR KNATIVE SERVING	23
8.4.1. Running a metering report	24
<b>CHAPTER 9. CLUSTER LOGGING WITH OPENSIFT SERVERLESS</b>	<b>25</b>
9.1. ABOUT CLUSTER LOGGING	25
9.2. ABOUT DEPLOYING AND CONFIGURING CLUSTER LOGGING	25
9.2.1. Configuring and Tuning Cluster Logging	25
9.2.2. Sample modified Cluster Logging Custom Resource	28

9.3. USING CLUSTER LOGGING TO FIND LOGS FOR KNATIVE SERVING COMPONENTS	29
9.4. USING CLUSTER LOGGING TO FIND LOGS FOR SERVICES DEPLOYED WITH KNATIVE SERVING	29
<b>CHAPTER 10. CONFIGURING KNATIVE SERVING AUTOSCALING</b>	<b>31</b>
10.1. CONFIGURING CONCURRENT REQUESTS FOR KNATIVE SERVING AUTOSCALING	31
10.1.1. Configuring concurrent requests using the target annotation	32
10.1.2. Configuring concurrent requests using the containerConcurrency field	32
10.2. CONFIGURING SCALE BOUNDS KNATIVE SERVING AUTOSCALING	32
<b>CHAPTER 11. USING KNATIVE CLIENT</b>	<b>34</b>
11.1. INSTALLING THE CLI	34
11.1.1. Installing the kn CLI for Linux	34
11.1.2. Installing the kn CLI for Linux using an RPM	35
11.1.3. Installing the kn CLI for macOS	35
11.1.4. Installing the kn CLI for Windows	35
11.2. LOGGING IN TO THE CLI	36
11.3. BASIC WORKFLOW USING KNATIVE CLIENT	37
11.4. AUTOSCALING WORKFLOW USING KNATIVE CLIENT	38
11.5. TRAFFIC SPLITTING USING KNATIVE CLIENT	38
11.5.1. Assigning tag revisions	39
11.5.2. Unassigning tag revisions	40
11.5.3. Traffic flag operation precedence	41
11.5.4. Traffic splitting flags	41
<b>CHAPTER 12. OPENSIFT SERVERLESS RELEASE NOTES</b>	<b>42</b>
12.1. GETTING SUPPORT	42
12.2. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.3.0	42
12.2.1. New features	42
12.2.2. Fixed issues	42
12.2.3. Known issues	42
12.3. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.2.0	42
12.3.1. New features	42
12.3.2. Fixed issues	43
12.3.3. Known issues	43
12.4. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.1.0	43
12.4.1. New features	43
12.4.2. Fixed issues	44
12.4.3. Known issues	44
12.5. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.0.0	44
12.5.1. New features	44
12.5.2. Known issues	44
12.6. ADDITIONAL RESOURCES	45



# CHAPTER 1. GETTING STARTED WITH OPENSHIFT SERVERLESS



## IMPORTANT

OpenShift Serverless is a Technology Preview feature only. Technology Preview features are not supported with Red Hat production service level agreements (SLAs) and might not be functionally complete. Red Hat does not recommend using them in production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

For more information about the support scope of Red Hat Technology Preview features, see <https://access.redhat.com/support/offerings/techpreview/>.

OpenShift Serverless simplifies the process of delivering code from development into production by reducing the need for infrastructure set up or back-end development by developers.

## 1.1. HOW OPENSHIFT SERVERLESS WORKS

Developers on OpenShift Serverless can use the provided Kubernetes-native APIs, as well as familiar languages and frameworks, to deploy applications and container workloads. For information about installing OpenShift Serverless, see [Installing OpenShift Serverless](#).

OpenShift Serverless on OpenShift Container Platform enables stateful, stateless, and serverless workloads to all run on a single multi-cloud container platform with automated operations. Developers can use a single platform for hosting their microservices, legacy, and serverless applications.

OpenShift Serverless is based on the open source Knative project, which provides portability and consistency across hybrid and multi-cloud environments by enabling an enterprise-grade serverless platform.

## 1.2. APPLICATIONS ON OPENSHIFT SERVERLESS

Applications are created using Custom Resource Definitions (CRDs) and associated controllers in Kubernetes, and are packaged as OCI compliant Linux containers that can be run anywhere.

To deploy applications in OpenShift Serverless, you must create Knative Services. For more information see [Getting started with Knative Services](#).



## CHAPTER 2. OPENSIFT SERVERLESS PRODUCT ARCHITECTURE

### 2.1. KNATIVE SERVING

Knative Serving on OpenShift Container Platform builds on Kubernetes and Istio to support deploying and serving serverless applications.

It creates a set of Kubernetes Custom Resource Definitions (CRDs) that are used to define and control the behavior of serverless workloads on an OpenShift Container Platform cluster.

These CRDs can be used as building blocks to address complex use cases, such as rapid deployment of serverless containers, automatic scaling of Pods, routing and network programming for Istio components, or viewing point-in-time snapshots of deployed code and configurations.

#### 2.1.1. Knative Serving components

The components described in this section are the resources that Knative Serving requires to be configured and run correctly.

##### Knative service resource

The **service.serving.knative.dev** resource automatically manages the whole lifecycle of a serverless workload on a cluster. It controls the creation of other objects to ensure that an app has a route, a configuration, and a new revision for each update of the service. Services can be defined to always route traffic to the latest revision or to a pinned revision.

##### Knative route resource

The **route.serving.knative.dev** resource maps a network endpoint to one or more Knative revisions. You can manage the traffic in several ways, including fractional traffic and named routes.

##### Knative configuration resource

The **configuration.serving.knative.dev** resource maintains the required state for your deployment. Modifying a configuration creates a new revision.

##### Knative revision resource

The **revision.serving.knative.dev** resource is a point-in-time snapshot of the code and configuration for each modification made to the workload. Revisions are immutable objects and can be retained for as long as needed. Cluster administrators can modify the **revision.serving.knative.dev** resource to enable automatic scaling of Pods in your OpenShift Container Platform cluster.

### 2.2. KNATIVE CLIENT

The Knative Client (**kn**) extends the functionality of the **oc** or **kubectl** tools to enable interaction with Knative components on OpenShift Container Platform. **kn** allows developers to deploy and manage applications without editing YAML files directly.

## CHAPTER 3. INSTALLING OPENSIFT SERVERLESS



### IMPORTANT

OpenShift Serverless is not tested or supported for installation in a restricted network environment.

### 3.1. CLUSTER SIZE REQUIREMENTS

The cluster must be sized appropriately to ensure that OpenShift Serverless can run correctly. You can use the MachineSet API to manually scale your cluster up to the desired size.

An OpenShift cluster with 10 CPUs and 40 GB memory is the minimum requirement for getting started with your first serverless application. This usually means you must scale up one of the default MachineSets by two additional machines.



### NOTE

For this configuration, the requirements depend on the deployed applications. By default, each pod requests ~400m of CPU and recommendations are based on this value. In the given recommendation, an application can scale up to 10 replicas. Lowering the actual CPU request of the application further pushes the boundary.



### NOTE

The numbers given only relate to the pool of worker machines of the OpenShift cluster. Master nodes are not used for general scheduling and are omitted.

For more advanced use-cases, such as using OpenShift logging, monitoring, metering, and tracing, you must deploy more resources. Recommended requirements for such use-cases are 24 vCPUs and 96GB of memory.

### Additional resources

For more information on using the MachineSet API, see [Creating MachineSets](#).

#### 3.1.1. Scaling a MachineSet manually

If you must add or remove an instance of a machine in a MachineSet, you can manually scale the MachineSet.

#### Prerequisites

- Install an OpenShift Container Platform cluster and the **oc** command line.
- Log in to **oc** as a user with **cluster-admin** permission.

#### Procedure

1. View the MachineSets that are in the cluster:

```
$ oc get machinesets -n openshift-machine-api
```

The MachineSets are listed in the form of **<clusterid>-worker-<aws-region-az>**.

2. Scale the MachineSet:

```
$ oc scale --replicas=2 machineset <machineset> -n openshift-machine-api
```

Or:

```
$ oc edit machineset <machineset> -n openshift-machine-api
```

You can scale the MachineSet up or down. It takes several minutes for the new machines to be available.



### IMPORTANT

By default, the OpenShift Container Platform router pods are deployed on workers. Because the router is required to access some cluster resources, including the web console, do not scale the worker MachineSet to **0** unless you first relocate the router pods.

## 3.2. INSTALLING THE OPENSIFT SERVERLESS OPERATOR

The OpenShift Serverless Operator can be installed using the OpenShift Container Platform instructions for installing Operators.

You can install the OpenShift Serverless Operator in the host cluster by following the OpenShift Container Platform instructions on installing an Operator.



### NOTE

The OpenShift Serverless Operator only works for OpenShift Container Platform versions 4.1.13 and later.

For details, see the OpenShift Container Platform documentation on [adding Operators to a cluster](#).



### IMPORTANT

The OpenShift Serverless Operator automatically installs the Service Mesh Operator. If you already have a community version of Maistra installed, this will cause a conflict with the OpenShift Serverless Operator Service Mesh auto-install. In this case, the already existing community version of Maistra will be used instead.

## 3.3. INSTALLING KNATIVE SERVING

You must create a **KnativeServing** object to install Knative Serving using the OpenShift Serverless Operator.



### IMPORTANT

You must create the **KnativeServing** object in the **knative-serving** namespace, as shown in the sample YAML, or it is ignored.

## Sample serving.yaml

```
apiVersion: v1
kind: Namespace
metadata:
  name: knative-serving
---
apiVersion: serving.knative.dev/v1alpha1
kind: KnativeService
metadata:
  name: knative-serving
  namespace: knative-serving
```

### Prerequisite

- An account with cluster administrator access.
- Installed OpenShift Serverless Operator.

### Procedure

1. Copy the sample YAML file into **serving.yaml** and apply it using:

```
$ oc apply -f serving.yaml
```

2. Verify the installation is complete by using the command:

```
$ oc get knativeserving/knative-serving -n knative-serving --template='{{range .status.conditions}}{{printf "%s=%s\n" .type .status}}{{end}}'
```

Results should be similar to:

```
DeploymentsAvailable=True
InstallSucceeded=True
Ready=True
```

## 3.4. UNINSTALLING KNATIVE SERVING

To uninstall Knative Serving, you must remove its custom resource and delete the **knative-serving** namespace.

### Prerequisite

- Installed Knative Serving

### Procedure

1. To remove Knative Serving, use the following command:

```
$ oc delete knativeserving knative-serving -n knative-serving
```

2. After the command has completed and all pods have been removed from the **knative-serving** namespace, delete the namespace by using the command:

```
$ oc delete namespace knative-serving
```

## 3.5. DELETING THE OPENSIFT SERVERLESS OPERATOR

You can remove the OpenShift Serverless Operator from the host cluster by following the OpenShift Container Platform instructions on deleting an Operator.

For details, see the OpenShift Container Platform documentation on [deleting Operators from a cluster](#).

## 3.6. DELETING KNATIVE SERVING CRDS FROM THE OPERATOR

After uninstalling the OpenShift Serverless Operator, the Operator CRDs and API services remain on the cluster. Use this procedure to completely uninstall the remaining components.

### Prerequisite

- You have uninstalled Knative Serving and removed the OpenShift Serverless Operator using the previous procedure.

### Procedure

1. Run the following command to delete the remaining Knative Serving CRDs:

```
$ oc delete crd knativeservings.serving.knative.dev
```

## CHAPTER 4. GETTING STARTED WITH KNATIVE SERVICES

Knative services are Kubernetes services that a user creates to deploy a serverless application. Each Knative service is defined by a route and a configuration, contained in a **.yaml** file.

### 4.1. CREATING A KNATIVE SERVICE

To create a service, you must create the **service.yaml** file.

You can copy the sample below. This sample will create a sample go lang application called **helloworld-go** and allows you to specify the image for that application.

```
apiVersion: serving.knative.dev/v1alpha1 ❶
kind: Service
metadata:
  name: helloworld-go ❷
  namespace: default ❸
spec:
  template:
    spec:
      containers:
        - image: gcr.io/knative-samples/helloworld-go ❹
          env:
            - name: TARGET ❺
              value: "Go Sample v1"
```

- ❶ Current version of Knative
- ❷ The name of the application
- ❸ The namespace the application will use
- ❹ The URL to the image of the application
- ❺ The environment variable printed out by the sample application

### 4.2. DEPLOYING A SERVERLESS APPLICATION

To deploy a serverless application, you must apply the **service.yaml** file.

#### Procedure

1. Navigate to the directory where the **service.yaml** file is contained.
2. Deploy the application by applying the **service.yaml** file.

```
$ oc apply --filename service.yaml
```

Now that service has been created and the application has been deployed, Knative will create a new immutable revision for this version of the application.

Knative will also perform network programming to create a route, ingress, service, and load balancer for your application, and will automatically scale your pods up and down based on traffic, including inactive pods.



#### NOTE

The first time that a Knative service is created in a namespace, that namespace will automatically receive a new networking configuration. This might cause the initial service to take longer than is usually required for a service to become ready.

If the namespace has no existing NetworkPolicy configuration, an "allow all" type policy will be applied automatically. This policy will be removed automatically if all Knative Services are removed from that namespace and no other NetworkPolicy configurations have been applied.

## 4.3. CONNECTING KNATIVE SERVICES TO EXISTING KUBERNETES DEPLOYMENTS

Knative Services can call a Kubernetes deployment in any namespace, provided that there are no existing additional network barriers.

A Kubernetes deployment can call a Knative Service if:

- The Kubernetes deployment is in the same namespace as the target Knative Service.
- The Kubernetes deployment is in a namespace that was manually added to the ServiceMeshMemberRoll in **knative-serving-ingress**.
- The Kubernetes deployment uses the target Knative Service's public URL.



#### NOTE

Knative Services are accessed using a public URL by default. The target Knative Service must not be configured as a private, **cluster-local** visibility service if you want to connect it to your existing Kubernetes deploying using a public URL.

## CHAPTER 5. CREATING SERVERLESS APPLICATIONS

### Prerequisites

To create serverless applications using the **Developer** perspective ensure that:

- You have [logged in to the web console](#).
- You are in the **Developer** perspective.
- You have the appropriate [roles and permissions](#) in a project to create applications and other workloads in OpenShift Container Platform.
- You have [installed the Openshift Serverless Operator](#).
- You have [created a knative-serving namespace and a KnativeServing resource in the knative-serving namespace](#).

### 5.1. IMPORTING A CODEBASE FROM GIT TO CREATE AN APPLICATION

The following procedure walks you through the **Import from Git** option in the **Developer** perspective to create an application.

Create, build, and deploy an application on OpenShift Container Platform using an existing codebase in GitHub as follows:

#### Procedure

1. In the **Add** view, click **From Git** to see the **Import from git** form.



**Import from git**












**Git**

Git Repo URL \*

[Show Advanced Git Options](#)

**Builder**

Builder Image \*

 Perl	 PHP	 Nginx	 Modern Webapp	 Httpd	 .NET Core	 Go	 Ruby	 Python	 Java	 Node.js
--	---	---	---	---	---	--	--	--	--	---

**General**

Application

nodejs-ex-app

Select an application for your grouping or Unassigned to not use an application grouping.

Name \*

A unique name given to the component that will be used to name associated resources.

**Pipelines** Dev Preview

Select a builder image to see if there is a pipeline template available for this runtime.

**Resources**

Select the resource type to generate

- Deployment
  - apps/Deployment
  - A Deployment enables declarative updates for Pods and ReplicaSets.
- Deployment Config
  - apps.openshift.io/DeploymentConfig
  - A Deployment Config defines the template for a pod and manages deploying new images or configuration changes
- Knative Service High Preview
  - serving.knative.dev/Service
  - A Knative Service enables scaling to zero when idle

**Advanced Options**

☒ Create a route to the application

Exposes your application at a public URL.

Click on the names to access advanced options for [Routing](#), [Build Configuration](#), [Deployment](#), [Scaling](#), [Resource Limits](#) and [Labels](#).

Create Cancel

- In the **Git** section, enter the Git repository URL for the codebase you want to use to create an application. For example, enter the URL of this sample Node.js application <https://github.com/sclorg/nodejs-ex>. The URL is then validated.
- Optionally, you can click **Show Advanced Git Options** to add details such as:
  - Git Reference** to point to code in a specific branch, tag, or commit to be used to build the application.
  - Context Dir** to specify the subdirectory for the application source code you want to use to build the application.
  - Source Secret** to create a **Secret Name** with credentials for pulling your source code from a private repository.
- In the **Builder** section, after the URL is validated, an appropriate builder image is detected, indicated by a star, and automatically selected. For the <https://github.com/sclorg/nodejs-ex> Git URL, the Node.js builder image is selected by default. If required, you can change the version using the **Builder Image Version** drop-down list.
- In the **General** section:
  - In the **Application** field, enter a unique name for the application grouping, for example, **myapp**. Ensure that the application name is unique in a namespace.
  - The **Name** field to identify the resources created for this application is automatically populated based on the Git repository URL.

**NOTE**

The resource name must be unique in a namespace. Modify the resource name if you get an error.

6. In the **Resources** section, select:

- **Deployment**, to create an application in plain Kubernetes style.
- **Deployment Config**, to create an OpenShift style application.
- **Knative Service**, to create a microservice.

**NOTE**

The **Knative Service** option is displayed in the **Import from git** form only if the **Serverless Operator** is installed in your cluster. For further details refer to documentation on installing OpenShift Serverless.

7. In the **Advanced Options** section, the **Create a route to the application** is selected by default so that you can access your application using a publicly available URL. You can clear the check box if you do not want to expose your application on a public route.
8. Optionally, you can use the following advanced options to further customize your application:

**Routing**

Click the **Routing** link to:

- Customize the hostname for the route.
- Specify the path the router watches.
- Select the target port for the traffic from the drop-down list.
- Secure your route by selecting the **Secure Route** check box. Select the required TLS termination type and set a policy for insecure traffic from the respective drop-down lists.

For serverless applications, the Knative Service manages all the routing options above. However, you can customize the target port for traffic, if required. If the target port is not specified, the default port of **8080** is used.

**Build and Deployment Configuration**

Click the **Build Configuration** and **Deployment Configuration** links to see the respective configuration options. Some of the options are selected by default; you can customize them further by adding the necessary triggers and environment variables. For serverless applications, the **Deployment Configuration** option is not displayed as the Knative configuration resource maintains the desired state for your deployment instead of a DeploymentConfig.

**Scaling**

Click the **Scaling** link to define the number of Pods or instances of the application you want to deploy initially.

For serverless applications, you can:

- Set the upper and lower limit for the number of pods that can be set by the autoscaler. If the lower limit is not specified, it defaults to zero.
- Define the soft limit for the required number of concurrent requests per instance of the application at a given time. It is the recommended configuration for autoscaling. If not specified, it takes the value specified in the cluster configuration.
- Define the hard limit for the number of concurrent requests allowed per instance of the application at a given time. This is configured in the revision template. If not specified, it defaults to the value specified in the cluster configuration.

### Resource Limit

Click the **Resource Limit** link to set the amount of **CPU** and **Memory** resources a container is guaranteed or allowed to use when running.

### Labels

Click the **Labels** link to add custom labels to your application.

9. Click **Create** to create the application and see its build status in the **Topology** view.

## CHAPTER 6. SPLITTING TRAFFIC BETWEEN REVISIONS

### 6.1. SPLITTING TRAFFIC BETWEEN REVISIONS USING THE DEVELOPER PERSPECTIVE

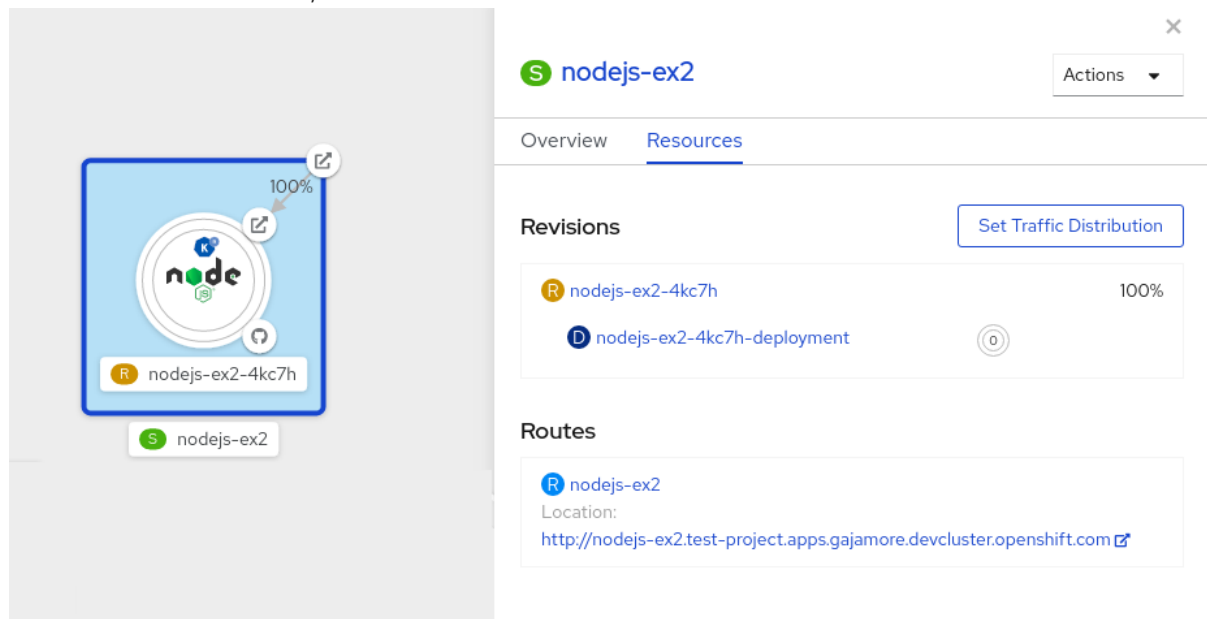
After you create a serverless application, the serverless application is displayed in the **Topology** view of the **Developer** perspective. The application revision is represented by the node and the serverless resource service is indicated by a quadrilateral around the node.

Any new change in the code or the service configuration triggers a revision, a snapshot of the code at a given time. For a service, you can manage the traffic between the revisions of the service by splitting and routing it to the different revisions as required.

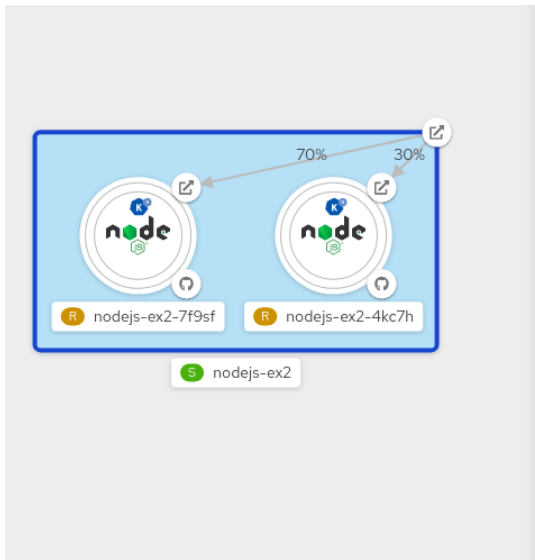
#### Procedure

To split traffic between multiple revisions of an application in the **Topology** view:

1. Click the serverless resource service, indicated by the quadrilateral, to see its overview in the side panel.
2. Click the **Resources** tab, to see a list of **Revisions** and **Routes** for the service.



3. Click the service, indicated by the **S** icon at the top of the side panel, to see an overview of the service details.
4. Click the **YAML** tab and modify the service configuration in the YAML editor, and click **Save**. For example, change the **timeoutseconds** from 300 to 301. This change in the configuration triggers a new revision. In the **Topology** view, the latest revision is displayed and the **Resources** tab for the service now displays the two revisions.
5. In the **Resources** tab, click the **Set Traffic Distribution** button to see the traffic distribution dialog box:
  - a. Add the split traffic percentage portion for the two revisions in the **Splits** field.
  - b. Add tags to create custom URLs for the two revisions.
  - c. Click **Save** to see two nodes representing the two revisions in the Topology view.



S **nodejs-ex2**

Actions

Overview
Resources

Revisions

Set Traffic Distribution

<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"><span style="color: orange; font-weight: bold;">R</span> nodejs-ex2-4kc7h</div> <div style="margin-right: 10px;"><span style="color: blue; font-weight: bold;">D</span> nodejs-ex2-4kc7h-deployment</div> <div style="border: 1px solid #ccc; border-radius: 50%; padding: 2px 5px; text-align: center;">0</div> </div>	30%
<div style="display: flex; align-items: center;"> <div style="margin-right: 10px;"><span style="color: orange; font-weight: bold;">R</span> nodejs-ex2-7f9sf</div> <div style="margin-right: 10px;"><span style="color: blue; font-weight: bold;">D</span> nodejs-ex2-7f9sf-deployment</div> <div style="border: 1px solid #ccc; border-radius: 50%; padding: 2px 5px; text-align: center;">0</div> </div>	70%

Routes

R nodejs-ex2

Location:  
<http://nodejs-ex2.test-project.apps.gajamore.devcluster.openshift.com>

## CHAPTER 7. MONITORING OPENSIFT SERVERLESS COMPONENTS

As an OpenShift Container Platform cluster administrator, you can deploy the OpenShift Container Platform monitoring stack and monitor the metrics of OpenShift Serverless components.

When using the OpenShift Serverless Operator, the required ServiceMonitor objects are created automatically for monitoring the deployed components.

OpenShift Serverless components, such as Knative Serving, expose metrics data. Administrators can monitor this data by using the OpenShift Container Platform web console.

### 7.1. CONFIGURING CLUSTER FOR APPLICATION MONITORING

Before application developers can monitor their applications, the human operator of the cluster needs to configure the cluster accordingly. This procedure shows how to.

#### Prerequisites

- You must log in as a user that belongs to a role with administrative privileges for the cluster.

#### Procedure

1. In the OpenShift Container Platform web console, navigate to the **Operators → OperatorHub** page and install the Prometheus Operator in the namespace where your application is.
2. Navigate to the **Operators → Installed Operators** page and install Prometheus, Alertmanager, Prometheus Rule, and Service Monitor in the same namespace.

### 7.2. VERIFYING AN OPENSIFT CONTAINER PLATFORM MONITORING INSTALLATION FOR USE WITH KNAIVE SERVING

Manual configuration for monitoring by an administrator is not required, but you can carry out these steps to verify that monitoring is installed correctly.

#### Procedure

1. Verify that the ServiceMonitor objects are deployed.

```
$ oc get servicemonitor -n knative-serving
NAME      AGE
activator 11m
autoscaler 11m
controller 11m
```

2. Verify that the **openshift.io/cluster-monitoring=true** label has been added to the Knative Serving namespace:

```
$ oc get namespace knative-serving --show-labels
NAME      STATUS AGE LABELS
knative-serving Active 4d istio-injection=enabled,openshift.io/cluster-monitoring=true,serving.knative.dev/release=v0.7.0
```

## 7.3. MONITORING KNATIVE SERVING USING THE OPENSIFT CONTAINER PLATFORM MONITORING STACK

This section provides example instructions for the visualization of Knative Serving Pod autoscaling metrics by using the OpenShift Container Platform monitoring tools.

### Prerequisites

- You must have the OpenShift Container Platform monitoring stack installed.

### Procedure

1. Navigate to the OpenShift Container Platform web console and authenticate.
2. Navigate to **Monitoring → Metrics**.
3. Enter the **Expression** and select **Run queries**. To monitor Knative Serving autoscaler Pods, use this example expression.

```
autoscaler_actual_pods
```

You will now see monitoring information for the Knative Serving autoscaler Pods in the console.

## CHAPTER 8. USING METERING WITH OPENSIFT SERVERLESS

As an OpenShift Container Platform cluster administrator, you can use metering to analyze what is happening in your OpenShift Serverless cluster.

For more information about metering on OpenShift Container Platform, see [About metering](#).

### 8.1. INSTALLING METERING

For information about installing metering on OpenShift Container Platform, see [Installing Metering](#).

### 8.2. DATASOURCES FOR KNATIVE SERVING METERING

The following **ReportDataSources** are examples of how Knative Serving can be used with OpenShift Container Platform metering.

#### 8.2.1. Datasource for CPU usage in Knative Serving

This datasource provides the accumulated CPU seconds used per Knative service over the report time period.

YAML file

```
apiVersion: metering.openshift.io/v1
kind: ReportDataSource
metadata:
  name: knative-service-cpu-usage
spec:
  prometheusMetricsImporter:
    query: >
      sum
        by(namespace,
          label_serving_knative_dev_service,
          label_serving_knative_dev_revision)
      (
        label_replace(rate(container_cpu_usage_seconds_total{container_name!="POD",container_name!="",pod_name!=""}[1m]), "pod", "$1", "pod_name", "(.*)")
        *
        on(pod, namespace)
        group_left(label_serving_knative_dev_service, label_serving_knative_dev_revision)
        kube_pod_labels{label_serving_knative_dev_service!=""}
      )
```

#### 8.2.2. Datasource for memory usage in Knative Serving

This datasource provides the average memory consumption per Knative service over the report time period.

YAML file



```

apiVersion: metering.openshift.io/v1
kind: ReportDataSource
metadata:
  name: knative-service-memory-usage
spec:
  prometheusMetricsImporter:
    query: >
      sum
        by(namespace,
          label_serving_knative_dev_service,
          label_serving_knative_dev_revision)
        (
          label_replace(container_memory_usage_bytes{container_name!="POD",
container_name!="",pod_name!=""}, "pod", "$1", "pod_name", "(.*)")
          *
          on(pod, namespace)
          group_left(label_serving_knative_dev_service, label_serving_knative_dev_revision)
          kube_pod_labels{label_serving_knative_dev_service!=""}
        )

```

### 8.2.3. Applying Datasources for Knative Serving metering

You can apply the **ReportDataSources** by using the following command:

```
$ oc apply -f <datasource-name>.yaml
```

#### Example

```
$ oc apply -f knative-service-memory-usage.yaml
```

## 8.3. QUERIES FOR KNATIVE SERVING METERING

The following **ReportQuery** resources reference the example **DataSources** provided.

### 8.3.1. Query for CPU usage in Knative Serving

#### YAML file

```

apiVersion: metering.openshift.io/v1
kind: ReportQuery
metadata:
  name: knative-service-cpu-usage
spec:
  inputs:
    - name: ReportingStart
      type: time
    - name: ReportingEnd
      type: time
    - default: knative-service-cpu-usage
      name: KnativeServiceCpuUsageDataSource
      type: ReportDataSource
  columns:

```

```

- name: period_start
  type: timestamp
  unit: date
- name: period_end
  type: timestamp
  unit: date
- name: namespace
  type: varchar
  unit: kubernetes_namespace
- name: service
  type: varchar
- name: data_start
  type: timestamp
  unit: date
- name: data_end
  type: timestamp
  unit: date
- name: service_cpu_seconds
  type: double
  unit: cpu_core_seconds
query: |
  SELECT
    timestamp '{| default .Report.ReportingStart .Report.Inputs.ReportingStart| prestoTimestamp |}'
  AS period_start,
    timestamp '{| default .Report.ReportingEnd .Report.Inputs.ReportingEnd | prestoTimestamp |}' AS
  period_end,
    labels['namespace'] as project,
    labels['label_serving_knative_dev_service'] as service,
    min("timestamp") as data_start,
    max("timestamp") as data_end,
    sum(amount * "timeprecision") AS service_cpu_seconds
  FROM {| dataSourceTableName .Report.Inputs.KnativeServiceCpuUsageDataSource |}
  WHERE "timestamp" >= timestamp '{| default .Report.ReportingStart .Report.Inputs.ReportingStart
| prestoTimestamp |}'
  AND "timestamp" < timestamp '{| default .Report.ReportingEnd .Report.Inputs.ReportingEnd |
prestoTimestamp |}'
  GROUP BY labels['namespace'],labels['label_serving_knative_dev_service']

```

### 8.3.2. Query for memory usage in Knative Serving

#### YAML file

```

apiVersion: metering.openshift.io/v1
kind: ReportQuery
metadata:
  name: knative-service-memory-usage
spec:
  inputs:
    - name: ReportingStart
      type: time
    - name: ReportingEnd
      type: time
  - default: knative-service-memory-usage
    name: KnativeServiceMemoryUsageDataSource
    type: ReportDataSource

```

```

columns:
- name: period_start
  type: timestamp
  unit: date
- name: period_end
  type: timestamp
  unit: date
- name: namespace
  type: varchar
  unit: kubernetes_namespace
- name: service
  type: varchar
- name: data_start
  type: timestamp
  unit: date
- name: data_end
  type: timestamp
  unit: date
- name: service_usage_memory_byte_seconds
  type: double
  unit: byte_seconds
query: |
  SELECT
    timestamp '{| default .Report.ReportingStart .Report.Inputs.ReportingStart| prestoTimestamp |}'
  AS period_start,
    timestamp '{| default .Report.ReportingEnd .Report.Inputs.ReportingEnd | prestoTimestamp |}' AS
  period_end,
    labels['namespace'] as project,
    labels['label_serving_knative_dev_service'] as service,
    min("timestamp") as data_start,
    max("timestamp") as data_end,
    sum(amount * "timeprecision") AS service_usage_memory_byte_seconds
  FROM {| dataSourceTableName .Report.Inputs.KnativeServiceMemoryUsageDataSource |}
  WHERE "timestamp" >= timestamp '{| default .Report.ReportingStart .Report.Inputs.ReportingStart
| prestoTimestamp |}'
  AND "timestamp" < timestamp '{| default .Report.ReportingEnd .Report.Inputs.ReportingEnd |
prestoTimestamp |}'
  GROUP BY labels['namespace'],labels['label_serving_knative_dev_service']

```

### 8.3.3. Applying Queries for Knative Serving metering

You can apply the **ReportQuery** by using the following command:

```
$ oc apply -f <query-name>.yaml
```

#### Example

```
$ oc apply -f knative-service-memory-usage.yaml
```

## 8.4. METERING REPORTS FOR KNATIVE SERVING

You can run metering reports against Knative Serving by creating **Report** resources. Before you run a report, you must modify the input parameter within the **Report** resource to specify the start and end dates of the reporting period.

## YAML file

```
apiVersion: metering.openshift.io/v1
kind: Report
metadata:
  name: knative-service-cpu-usage
spec:
  reportingStart: '2019-06-01T00:00:00Z' ❶
  reportingEnd: '2019-06-30T23:59:59Z' ❷
  query: knative-service-cpu-usage ❸
runImmediately: true
```

- ❶ Start date of the report, in ISO 8601 format.
- ❷ End date of the report, in ISO 8601 format.
- ❸ Either **knative-service-cpu-usage** for CPU usage report or **knative-service-memory-usage** for a memory usage report.

### 8.4.1. Running a metering report

Once you have provided the input parameters, you can run the report using the command:

```
$ oc apply -f <report-name>.yaml
```

You can then check the report as shown in the following example:

```
$ kubectl get report
```

NAME	QUERY	SCHEDULE	RUNNING	FAILED	LAST REPORT
TIME	AGE				
knative-service-cpu-usage	knative-service-cpu-usage		Finished		2019-06-30T23:59:59Z 10h

## CHAPTER 9. CLUSTER LOGGING WITH OPENSIFT SERVERLESS

### 9.1. ABOUT CLUSTER LOGGING

As an OpenShift Container Platform cluster administrator, you can deploy cluster logging to aggregate logs for a range of OpenShift Container Platform services.

The cluster logging components are based upon Elasticsearch, Fluentd or Rsyslog, and Kibana. The collector, [Fluentd](#), is deployed to each node in the OpenShift Container Platform cluster. It collects all node and container logs and writes them to [Elasticsearch](#) (ES). [Kibana](#) is the centralized, web UI where users and administrators can create rich visualizations and dashboards with the aggregated data.

OpenShift Container Platform cluster administrators can deploy cluster logging using a few CLI commands and the OpenShift Container Platform web console to install the Elasticsearch Operator and Cluster Logging Operator. When the operators are installed, create a Cluster Logging Custom Resource (CR) to schedule cluster logging pods and other resources necessary to support cluster logging. The operators are responsible for deploying, upgrading, and maintaining cluster logging.

OpenShift Container Platform cluster administrators can deploy cluster logging using a few CLI commands and the OpenShift Container Platform web console to install the Elasticsearch Operator and Cluster Logging Operator. When the operators are installed, create a Cluster Logging Custom Resource (CR) to schedule cluster logging pods and other resources necessary to support cluster logging. The operators are responsible for deploying, upgrading, and maintaining cluster logging.

You can configure cluster logging by modifying the Cluster Logging Custom Resource (CR), named **instance**. The CR defines a complete cluster logging deployment that includes all the components of the logging stack to collect, store and visualize logs. The Cluster Logging Operator watches the **ClusterLogging** Custom Resource and adjusts the logging deployment accordingly.

Administrators and application developers can view the logs of the projects for which they have view access.

### 9.2. ABOUT DEPLOYING AND CONFIGURING CLUSTER LOGGING

OpenShift Container Platform cluster logging is designed to be used with the default configuration, which is tuned for small to medium sized OpenShift Container Platform clusters.

The installation instructions that follow include a sample Cluster Logging Custom Resource (CR), which you can use to create a cluster logging instance and configure your cluster logging deployment.

If you want to use the default cluster logging install, you can use the sample CR directly.

If you want to customize your deployment, make changes to the sample CR as needed. The following describes the configurations you can make when installing your cluster logging instance or modify after installation. See the Configuring sections for more information on working with each component, including modifications you can make outside of the Cluster Logging Custom Resource.

#### 9.2.1. Configuring and Tuning Cluster Logging

You can configure your cluster logging environment by modifying the Cluster Logging Custom Resource deployed in the **openshift-logging** project.

You can modify any of the following components upon install or after install

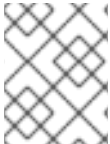
## Management state

The Cluster Logging Operator and Elasticsearch Operator can be in a *Managed* or *Unmanaged* state.

In managed state, the Cluster Logging Operator (CLO) responds to changes in the Cluster Logging Custom Resource (CR) and attempts to update the cluster to match the CR.

In order to modify certain components managed by the Cluster Logging Operator or the Elasticsearch Operator, you must set the operator to the *unmanaged* state.

In Unmanaged state, the operators do not respond to changes in the CRs. The administrator assumes full control of individual component configurations and upgrades when in unmanaged state.



### NOTE

The OpenShift Container Platform documentation indicates in a prerequisite step when you must set the cluster to Unmanaged.

```
spec:
  managementState: "Managed"
```

The OpenShift Container Platform documentation indicates in a prerequisite step when you must set the cluster to Unmanaged.



### IMPORTANT

An unmanaged deployment will not receive updates until the **ClusterLogging** custom resource is placed back into a managed state.

## Memory and CPU

You can adjust both the CPU and memory limits for each component by modifying the **resources** block with valid memory and CPU values:

```
spec:
  logStore:
    elasticsearch:
      resources:
        limits:
          cpu:
          memory:
        requests:
          cpu: 1
          memory: 16Gi
      type: "elasticsearch"
    collection:
      logs:
        fluentd:
          resources:
            limits:
              cpu:
              memory:
            requests:
              cpu:
              memory:
```

```

    type: "fluentd"
  visualization:
    kibana:
      resources:
        limits:
          cpu:
          memory:
        requests:
          cpu:
          memory:
      type: kibana
  curation:
    curator:
      resources:
        limits:
          memory: 200Mi
        requests:
          cpu: 200m
          memory: 200Mi
      type: "curator"

```

### Elasticsearch storage

You can configure a persistent storage class and size for the Elasticsearch cluster using the **storageClass name** and **size** parameters. The Cluster Logging Operator creates a **PersistentVolumeClaim** for each data node in the Elasticsearch cluster based on these parameters.

```

spec:
  logStore:
    type: "elasticsearch"
  elasticsearch:
    nodeCount: 3
    storage:
      storageClassName: "gp2"
      size: "200G"

```

This example specifies each data node in the cluster will be bound to a **PersistentVolumeClaim** that requests "200G" of "gp2" storage. Each primary shard will be backed by a single replica.

### NOTE

Omitting the **storage** block results in a deployment that includes ephemeral storage only.

```

spec:
  logStore:
    type: "elasticsearch"
  elasticsearch:
    nodeCount: 3
    storage: {}

```

### Elasticsearch replication policy

You can set the policy that defines how Elasticsearch shards are replicated across data nodes in the cluster:

- **FullRedundancy**. The shards for each index are fully replicated to every data node.

- **MultipleRedundancy.** The shards for each index are spread over half of the data nodes.
- **SingleRedundancy.** A single copy of each shard. Logs are always available and recoverable as long as at least two data nodes exist.
- **ZeroRedundancy.** No copies of any shards. Logs may be unavailable (or lost) in the event a node is down or fails.

### Curator schedule

You specify the schedule for Curator in the [cron format](<https://en.wikipedia.org/wiki/Cron>).

```
spec:
  curation:
    type: "curator"
  resources:
    curator:
      schedule: "30 3 * * *"
```

## 9.2.2. Sample modified Cluster Logging Custom Resource

The following is an example of a Cluster Logging Custom Resource modified using the options previously described.

### Sample modified Cluster Logging Custom Resource

```
apiVersion: "logging.openshift.io/v1"
kind: "ClusterLogging"
metadata:
  name: "instance"
  namespace: "openshift-logging"
spec:
  managementState: "Managed"
  logStore:
    type: "elasticsearch"
  elasticsearch:
    nodeCount: 2
    resources:
      limits:
        memory: 2Gi
      requests:
        cpu: 200m
        memory: 2Gi
    storage: {}
    redundancyPolicy: "SingleRedundancy"
  visualization:
    type: "kibana"
  kibana:
    resources:
      limits:
        memory: 1Gi
      requests:
        cpu: 500m
        memory: 1Gi
    replicas: 1
```



```

curation:
  type: "curator"
curator:
  resources:
    limits:
      memory: 200Mi
    requests:
      cpu: 200m
      memory: 200Mi
  schedule: "*/5 * * * *"
collection:
  logs:
    type: "fluentd"
  fluentd:
    resources:
      limits:
        memory: 1Gi
      requests:
        cpu: 200m
        memory: 1Gi

```

### 9.3. USING CLUSTER LOGGING TO FIND LOGS FOR KNATIVE SERVING COMPONENTS

#### Procedure

1. To open the Kibana UI, the visualization tool for Elasticsearch, use the following command to get the Kibana route:
- ```
$ oc -n openshift-logging get route kibana
```
2. Use the route's URL to navigate to the Kibana dashboard and log in.
  3. Ensure the index is set to **.all**. If the index is not set to **.all**, only the OpenShift system logs will be listed.
  4. You can filter the logs by using the **knative-serving** namespace. Enter **kubernetes.namespace\_name:knative-serving** in the search box to filter results.



#### NOTE

Knative Serving uses structured logging by default. You can enable the parsing of these logs by customizing the cluster logging Fluentd settings. This makes the logs more searchable and enables filtering on the log level to quickly identify issues.

### 9.4. USING CLUSTER LOGGING TO FIND LOGS FOR SERVICES DEPLOYED WITH KNATIVE SERVING

With OpenShift Cluster Logging, the logs that your applications write to the console are collected in Elasticsearch. The following procedure outlines how to apply these capabilities to applications deployed by using Knative Serving.

## Procedure

1. Use the following command to find the URL to Kibana:

```
$ oc -n cluster-logging get route kibana`
```

2. Enter the URL in your browser to open the Kibana UI.
3. Ensure the index is set to **.all**. If the index is not set to **.all**, only the OpenShift system logs will be listed.
4. Filter the logs by using the Kubernetes namespace your service is deployed in. Add a filter to identify the service itself: **kubernetes.namespace\_name:default AND kubernetes.labels.serving\_knative\_dev/service:{SERVICE\_NAME}**.



### NOTE

You can also filter by using **/configuration** or **/revision**.

5. You can narrow your search by using **kubernetes.container\_name:<user-container>** to only display the logs generated by your application. Otherwise, you will see logs from the queue-proxy.



### NOTE

Use JSON-based structured logging in your application to allow for the quick filtering of these logs in production environments.

## CHAPTER 10. CONFIGURING KNATIVE SERVING AUTOSCALING

OpenShift Serverless provides capabilities for automatic Pod scaling, including scaling inactive Pods to zero, by enabling the Knative Serving autoscaling system in an OpenShift Container Platform cluster.

To enable autoscaling for Knative Serving, you must configure concurrency and scale bounds in the revision template.



### NOTE

Any limits or targets set in the revision template are measured against a single instance of your application. For example, setting the **target** annotation to **50** will configure the autoscaler to scale the application so that each instance of it will handle 50 requests at a time.

### 10.1. CONFIGURING CONCURRENT REQUESTS FOR KNATIVE SERVING AUTOSCALING

You can specify the number of concurrent requests that should be handled by each instance of an application (revision container) by adding the **target** annotation or the **containerConcurrency** field in the revision template.

Here is an example of **target** being used in a revision template:

```
apiVersion: serving.knative.dev/v1alpha1
kind: Service
metadata:
  name: myapp
spec:
  template:
    metadata:
      annotations:
        autoscaling.knative.dev/target: 50
    spec:
      containers:
        - image: myimage
```

Here is an example of **containerConcurrency** being used in a revision template:

```
apiVersion: serving.knative.dev/v1alpha1
kind: Service
metadata:
  name: myapp
spec:
  template:
    metadata:
      annotations:
    spec:
      containerConcurrency: 100
      containers:
        - image: myimage
```

Adding a value for both **target** and **containerConcurrency** will target the **target** number of concurrent requests, but impose a hard limit of the **containerConcurrency** number of requests.

For example, if the **target** value is 50 and the **containerConcurrency** value is 100, the targeted number of requests will be 50, but the hard limit will be 100.

If the **containerConcurrency** value is less than the **target** value, the **target** value will be tuned down, since there is no need to target more requests than the number that can actually be handled.



#### NOTE

**containerConcurrency** should only be used if there is a clear need to limit how many requests reach the application at a given time. Using **containerConcurrency** is only advised if the application needs to have an enforced constraint of concurrency.

### 10.1.1. Configuring concurrent requests using the target annotation

The default target for the number of concurrent requests is **100**, but you can override this value by adding or modifying the **autoscaling.knative.dev/target** annotation value in the revision template.

Here is an example of how this annotation is used in the revision template to set the target to **50**.

```
autoscaling.knative.dev/target: 50
```

### 10.1.2. Configuring concurrent requests using the containerConcurrency field

**containerConcurrency** sets a hard limit on the number of concurrent requests handled.

```
containerConcurrency: 0 | 1 | 2-N
```

0

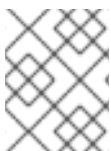
allows unlimited concurrent requests.

1

guarantees that only one request is handled at a time by a given instance of the revision container.

2 or more

will limit request concurrency to that value.



#### NOTE

If there is no **target** annotation, autoscaling is configured as if **target** is equal to the value of **containerConcurrency**.

## 10.2. CONFIGURING SCALE BOUNDS KNATIVE SERVING AUTOSCALING

The **minScale** and **maxScale** annotations can be used to configure the minimum and maximum number of Pods that can serve applications. These annotations can be used to prevent cold starts or to help control computing costs.

**minScale**

If the **minScale** annotation is not set, Pods will scale to zero (or to 1 if `enable-scale-to-zero` is false per the **ConfigMap**).

### **maxScale**

If the **maxScale** annotation is not set, there will be no upper limit for the number of Pods created.

**minScale** and **maxScale** can be configured as follows in the revision template:

```
spec:
  template:
    metadata:
      autoscaling.knative.dev/minScale: "2"
      autoscaling.knative.dev/maxScale: "10"
```

Using these annotations in the revision template will propagate this configuration to **PodAutoscaler** objects.



### **NOTE**

These annotations apply for the full lifetime of a revision. Even when a revision is not referenced by any route, the minimal Pod count specified by **minScale** will still be provided. Keep in mind that non-routeable revisions may be garbage collected, which enables Knative to reclaim the resources.

## CHAPTER 11. USING KNATIVE CLIENT

Knative Client (**kn**) is the Knative command line interface (CLI). The CLI exposes commands for managing your applications, as well as lower level tools to interact with components of OpenShift Container Platform. With **kn**, you can create applications and manage OpenShift Container Platform projects from the terminal.

Knative client does not have its own log in mechanism. To log in to the cluster you must install the **oc** CLI and use the **oc** login. Installation options for the CLI vary depending on your operating system.

### 11.1. INSTALLING THE CLI

You can install the CLI in order to interact with OpenShift Container Platform using a command-line interface.



#### IMPORTANT

If you installed an earlier version of **oc**, you cannot use it to complete all of the commands in OpenShift Container Platform 4.3. Download and install the new version of **oc**.

#### Procedure

1. From the [Infrastructure Provider](#) page on the Red Hat OpenShift Cluster Manager site, navigate to the page for your installation type and click **Download Command-line Tools**
2. Click the folder for your operating system and architecture and click the compressed file.



#### NOTE

You can install **oc** on Linux, Windows, or macOS.

3. Save the file to your file system.
4. Extract the compressed file.
5. Place it in a directory that is on your **PATH**.

After you install the CLI, it is available using the **oc** command:

```
$ oc <command>
```

#### 11.1.1. Installing the kn CLI for Linux

For Linux distributions, you can download the CLI directly as a **tar.gz** archive.

#### Procedure

1. Download the [CLI](#).
2. Unpack the archive:

```
$ tar -xf <file>
```

3. Move the **kn** binary to a directory on your PATH.
4. To check your path, run:

```
$ echo $PATH
```



#### NOTE

If you do not use RHEL or Fedora, ensure that **libc** is installed in a directory on your library path. If **libc** is not available, you might see the following error when you run CLI commands:

```
$ kn: No such file or directory
```

### 11.1.2. Installing the kn CLI for Linux using an RPM

For Red Hat Enterprise Linux (RHEL), you can install **kn** as an RPM if you have an active OpenShift Container Platform subscription on your Red Hat account.

#### Procedure

- Use the following command to install **kn**:

```
# subscription-manager register
# subscription-manager refresh
# subscription-manager attach --pool=<pool_id> 1
# subscription-manager repos --enable="openshift-serverless-1-for-rhel-8-x86_64-rpms"
# yum install openshift-serverless-clients
```

- 1** Pool ID for an active OpenShift Container Platform subscription

### 11.1.3. Installing the kn CLI for macOS

**kn** for macOS is provided as a **tar.gz** archive.

#### Procedure

1. Download the [CLI](#).
2. Unpack and unzip the archive.
3. Move the **kn** binary to a directory on your PATH.
4. To check your PATH, open a terminal window and run:

```
$ echo $PATH
```

### 11.1.4. Installing the kn CLI for Windows

The CLI for Windows is provided as a zip archive.

## Procedure

1. Download the [CLI](#).
2. Unzip the archive with a ZIP program.
3. Move the **kn** binary to a directory on your PATH.
4. To check your PATH, open the Command Prompt and run the command:

```
C:\> path
```

## 11.2. LOGGING IN TO THE CLI

You can log in to the **oc** CLI to access and manage your cluster.

### Prerequisites

- You must have access to an OpenShift Container Platform cluster.
- You must have installed the CLI.

### Procedure

- Log in to the CLI using the **oc login** command and enter the required information when prompted.

```
$ oc login
Server [https://localhost:8443]: https://openshift.example.com:6443 1
The server uses a certificate signed by an unknown authority.
You can bypass the certificate check, but any data you send to the server could be
intercepted by others.
Use insecure connections? (y/n): y 2

Authentication required for https://openshift.example.com:6443 (openshift)
Username: user1 3
Password: 4
Login successful.

You don't have any projects. You can try to create a new project, by running

    oc new-project <projectname>

Welcome! See 'oc help' to get started.
```

- 1** Enter the OpenShift Container Platform server URL.
- 2** Enter whether to use insecure connections.
- 3** Enter the user name to log in as.
- 4** Enter the user's password.

You can now create a project or issue other commands for managing your cluster.



## 11.3. BASIC WORKFLOW USING KNATIVE CLIENT

Use this basic workflow to create, read, update, delete (CRUD) operations on a service. The following example deploys a [simple Hello World service](#) that reads the environment variable **TARGET** and prints its output.

### Procedure

1. Create a service in the **default** namespace from an image.

```
$ kn service create hello --image gcr.io/knative-samples/helloworld-go --env
TARGET=Knative
Creating service 'hello' in namespace 'default':
```

```
0.085s The Route is still working to reflect the latest desired specification.
0.101s Configuration "hello" is waiting for a Revision to become ready.
11.590s ...
11.650s Ingress has not yet been reconciled.
11.726s Ready to serve.
```

```
Service 'hello' created with latest revision 'hello-gsdks-1' and URL:
http://hello.default.apps-crc.testing
```

2. List the service.

```
$ kn service list
NAME URL LATEST AGE CONDITIONS READY
REASON
hello http://hello.default.apps-crc.testing hello-gsdks-1 8m35s 3 OK / 3 True
```

3. Check if the service is working by using the **curl** service endpoint command:

```
$ curl http://hello.default.apps-crc.testing

Hello Knative!
```

4. Update the service.

```
$ kn service update hello --env TARGET=Kn
Updating Service 'hello' in namespace 'default':
```

```
10.136s Traffic is not yet migrated to the latest revision.
10.175s Ingress has not yet been reconciled.
10.348s Ready to serve.
```

```
Service 'hello' updated with latest revision 'hello-dghll-2' and URL:
http://hello.default.apps-crc.testing
```

The service's environment variable **TARGET** is now set to **Kn**.

5. Describe the service.

```
$ kn service describe hello
Name: hello
```

```

Namespace: default
Age:      13m
URL:      http://hello.default.apps-crc.testing
Address:  http://hello.default.svc.cluster.local

Revisions:
  100% @latest (hello-dghll-2) [2] (1m)
    Image: gcr.io/knative-samples/helloworld-go (pinned to 5ea96b)

Conditions:
  OK TYPE          AGE REASON
  ++ Ready         1m
  ++ ConfigurationsReady 1m
  ++ RoutesReady    1m

```

6. Delete the service.

```

$ kn service delete hello
Service 'hello' successfully deleted in namespace 'default'.

```

You can then verify that the **hello** service is deleted by attempting to **list** it.

```

$ kn service list hello
No services found.

```

## 11.4. AUTOSCALING WORKFLOW USING KNATIVE CLIENT

You can access autoscaling capabilities by using **kn** to modify Knative services without editing YAML files directly.

Use the **service create** and **service update** commands with the appropriate flags to configure the autoscaling behavior.

| Flag                            | Description                                                                                                                       |
|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| <b>--concurrency-limit int</b>  | Hard limit of concurrent requests to be processed by a single replica.                                                            |
| <b>--concurrency-target int</b> | Recommendation for when to scale up based on the concurrent number of incoming requests. Defaults to <b>--concurrency-limit</b> . |
| <b>--max-scale int</b>          | Maximum number of replicas.                                                                                                       |
| <b>--min-scale int</b>          | Minimum number of replicas.                                                                                                       |

## 11.5. TRAFFIC SPLITTING USING KNATIVE CLIENT

**kn** helps you control which revisions get routed traffic on your Knative service.

Knative service allows for traffic mapping, which is the mapping of revisions of the service to an allocated portion of traffic. It offers the option to create unique URLs for particular revisions and has the ability to assign traffic to the latest revision.

With every update to the configuration of the service, a new revision is created with the service route pointing all the traffic to the latest ready revision by default.

You can change this behavior by defining which revision gets a portion of the traffic.

## Procedure

- Use the **kn service update** command with the **--traffic** flag to update the traffic.



### NOTE

**--traffic RevisionName=Percent** uses the following syntax:

- The **--traffic** flag requires two values separated by separated by an equals sign (=).
- The **RevisionName** string refers to the name of the revision.
- **Percent** integer denotes the traffic portion assigned to the revision.
- Use identifier **@latest** for the RevisionName to refer to the latest ready revision of the service. You can use this identifier only once with the **--traffic** flag.
- If the **service update** command updates the configuration values for the service along with traffic flags, the **@latest** reference will point to the created revision to which the updates are applied.
- **--traffic** flag can be specified multiple times and is valid only if the sum of the **Percent** values in all flags totals 100.



### NOTE

For example, to route 10% of traffic to your new revision before putting all traffic on, use the following command:

```
$ kn service update svc --traffic @latest=10 --traffic svc-vwxyz=90
```

## 11.5.1. Assigning tag revisions

A tag in a traffic block of service creates a custom URL, which points to a referenced revision. A user can define a unique tag for an available revision of a service which creates a custom URL by using the format **http(s)://TAG-SERVICE.DOMAIN**.

A given tag must be unique to its traffic block of the service. **kn** supports assigning and unassigning custom tags for revisions of services as part of the **kn service update** command.



### NOTE

If you have assigned a tag to a particular revision, a user can reference the revision by its tag in the **--traffic** flag as **--traffic Tag=Percent**.

## Procedure

- Use the following command:

```
$ kn service update svc --tag @latest=candidate --tag svc-vwxyz=current
```

### NOTE

**--tag RevisionName=Tag** uses the following syntax:

- **--tag** flag requires two values separated by a **=**.
- **RevisionName** string refers to name of the **Revision**.
- **Tag** string denotes the custom tag to be given for this Revision.
- Use the identifier **@latest** for the RevisionName to refer to the latest ready revision of the service. You can use this identifier only once with the **--tag** flag.
- If the **service update** command is updating the configuration values for the Service (along with tag flags), **@latest** reference will be pointed to the created Revision after applying the update.
- **--tag** flag can be specified multiple times.
- **--tag** flag may assign different tags to the same revision.

## 11.5.2. Unassigning tag revisions

Tags assigned to revisions in a traffic block can be unassigned. Unassigning tags removes the custom URLs.

### NOTE

If a revision is untagged and it is assigned 0% of the traffic, it is removed from the traffic block entirely.

## Procedure

- A user can unassign the tags for revisions using the **kn service update** command:

```
$ kn service update svc --untag candidate
```

### NOTE

**--untag Tag** uses the following syntax:

- The **--untag** flag requires one value.
- The **tag** string denotes the unique tag in the traffic block of the service which needs to be unassigned. This also removes the respective custom URL.
- The **--untag** flag can be specified multiple times.

### 11.5.3. Traffic flag operation precedence

All traffic-related flags can be specified using a single **kn service update** command. **kn** defines the precedence of these flags. The order of the flags specified when using the command is not taken into account.

The precedence of the flags as they are evaluated by **kn** are:

1. **--untag**: All the referenced revisions with this flag are removed from the traffic block.
2. **--tag**: Revisions are tagged as specified in the traffic block.
3. **--traffic**: The referenced revisions are assigned a portion of the traffic split.

### 11.5.4. Traffic splitting flags

**kn** supports traffic operations on the traffic block of a service as part of the **kn service update** command.

The following table displays a summary of traffic splitting flags, value formats, and the operation the flag performs. The "Repetition" column denotes whether repeating the particular value of flag is allowed in a **kn service update** command.

| Flag             | Value(s)                    | Operation                                                      | Repetition |
|------------------|-----------------------------|----------------------------------------------------------------|------------|
| <b>--traffic</b> | <b>RevisionName=Percent</b> | Gives <b>Percent</b> traffic to <b>RevisionName</b>            | Yes        |
| <b>--traffic</b> | <b>Tag=Percent</b>          | Gives <b>Percent</b> traffic to the Revision having <b>Tag</b> | Yes        |
| <b>--traffic</b> | <b>@latest=Percent</b>      | Gives <b>Percent</b> traffic to the latest ready Revision      | No         |
| <b>--tag</b>     | <b>RevisionName=Tag</b>     | Gives <b>Tag</b> to <b>RevisionName</b>                        | Yes        |
| <b>--tag</b>     | <b>@latest=Tag</b>          | Gives <b>Tag</b> to the latest ready Revision                  | No         |
| <b>--untag</b>   | <b>Tag</b>                  | Removes <b>Tag</b> from Revision                               | Yes        |

## CHAPTER 12. OPENSIFT SERVERLESS RELEASE NOTES

For an overview of OpenShift Serverless functionality, see [Getting started with OpenShift Serverless](#).

### 12.1. GETTING SUPPORT

If you experience difficulty with a procedure described in this documentation, visit the [Customer Portal](#) to learn more about support for Technology Preview features.

### 12.2. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.3.0

#### 12.2.1. New features

- OpenShift Serverless has been updated to use Knative Serving 0.10.1.
- OpenShift Serverless has been updated to use Knative Client (**kn** CLI) 0.10.0.
- OpenShift Serverless 1.3.0 is available on OpenShift Container Platform 4.2 and newer versions.

#### 12.2.2. Fixed issues

- Fixed a bug which caused Routes to have incorrect cross-namespaced **OwnerReferences**.

#### 12.2.3. Known issues

- Connecting to a private, cluster local Knative Service from a namespace that is not part of the **knative-serving-ingress** Service Mesh fails on **i/o timeout**.

### 12.3. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.2.0

#### 12.3.1. New features

- OpenShift Serverless has been updated to use Knative Serving 0.9.0.
- OpenShift Serverless has been updated to use Knative Client (**kn** CLI) 0.9.0.
- OpenShift Serverless on OpenShift Container Platform 4.2 now uses the Operator Lifecycle Manager (OLM) dependency resolution mechanism to install the ServiceMesh Operator automatically. The required ServiceMeshControlPlane and ServiceMeshMemberRoll are also installed and managed for the user.
- Access to the KnativeServing resource is now restricted to **cluster-admin** roles to prevent any user from blocking the resource. Only **cluster-admin** roles can create KnativeServing CRs.
- The OpenShift Serverless Operator can now be found in the OperatorHub by searching for "knative".
- The OpenShift Container Platform web console now shows status conditions for the KnativeServing resource.

- In version 1.2.0, the OpenShift Serverless Operator inspects network policies for namespaces. If no network policy exists, the Operator automatically creates a wide open policy, to ensure that traffic can flow in and out of the namespace and OpenShift routes can be used.

If there is an existing network policy, OpenShift Serverless will not create a new policy. The Operator expects the user to continue managing their own network policies as needed for their applications. For example, the user must set policies that allow traffic to flow in and out of the namespace, and allow OpenShift routes to still be used after the namespace is added to a ServiceMeshMemberRoll.

### 12.3.2. Fixed issues

- In previous releases, using the same services or routes in different namespaces caused services to not work properly and caused OpenShift Container Platform routes to be overridden. This issue has been fixed.
- In previous releases, different traffic split targets required a mandatory tag. A single traffic split can now be defined with **untagged** traffic targets.
- Existing Knative Services and Routes which had been created with public visibility in OpenShift Serverless Operator version 1.1.0 were not able to be updated to cluster-local visibility. This issue is now fixed.
- The **Unknown Uninitialized : Waiting for VirtualService** error has been fixed.
- Knative service no longer returns a 503 status code when the cluster is running for a long time.

### 12.3.3. Known issues

- Installing the OpenShift Serverless Operator on OpenShift Container Platform versions older than 4.2.4 using OLM may incorrectly use community versions of the required dependencies. As a workaround, on OpenShift Container Platform versions older than 4.2.4, explicitly install the Red Hat provided versions of the Elastic Search, Jaeger, Kiali and ServiceMesh Operators before installing the OpenShift Serverless Operator.
- If you are upgrading the OpenShift Serverless from version 1.1.0 to version 1.2.0 and you have set up a ServiceMeshControlPlane and ServiceMeshMemberRoll to work with your Knative Serving instance, you must remove the **knative-serving** namespace and any other namespaces that contain Knative Services from the ServiceMeshMemberRoll in **istio-system**. You can also delete the ServiceMeshControlPlane from the namespace entirely if it is not required for other applications.

Once the upgrade starts, existing services will continue to work as before, but new Services will never become ready. Once you unblock the release by removing the **knative-serving** and any other relevant namespaces from the ServiceMeshMemberRoll, there will be a brief outage to all active Services. This will fix itself. Make sure that you remove all namespaces containing Knative Services from the original ServiceMeshMemberRoll.

- gRPC and HTTP2 do not work against routes. This is a known limitation of OpenShift routes.

## 12.4. RELEASE NOTES FOR RED HAT OPENSIFT SERVERLESS TECHNOLOGY PREVIEW 1.1.0

### 12.4.1. New features

- OpenShift Serverless has been updated to use Knative Serving 0.8.1.
- Enhanced Operator metadata now includes more information regarding support state and a link to the official installation documentation.
- A developer preview version of Knative Eventing is now available for use with OpenShift Serverless, however this is not included in the OpenShift Serverless Operator and is not currently supported as part of this Technology Preview. For more information, see [Knative Eventing on OpenShift Container Platform](#).

#### 12.4.2. Fixed issues

- Users who were not project administrators would previously see the following error when using OpenShift Serverless:

```
revisions.serving.knative.dev: User "sounds" cannot list resource "revisions"
```

This issue has now been fixed with the addition of new RBAC rules.

- A race condition was preventing Istio sidecar injection from working correctly. Istio did not consider the **knative-serving** namespace to be present in the ServiceMeshMemberRoll at the time of Pod creation. Istio now waits for status information from ServiceMeshMemberRoll which fixes this issue.

#### 12.4.3. Known issues

- Users may see the error **Unknown Uninitialized : Waiting for VirtualService to be ready** while waiting for a service in a newly created namespace to be ready, which can take several minutes. If a user allows enough time between the creation of a namespace and the creation of a service in the namespace (approximately one minute), this error may be avoided.
- Existing Knative Services and Routes which have been created with public visibility cannot be updated to cluster-local visibility. If you require cluster-local visibility on Knative Services and Routes, this must be configured at the time of creating these resources.
- Knative service returns a 503 status code when the cluster is running for a long time. The Knative Serving Pods do not show any errors. Restarting the **istio-pilot** Pod temporarily fixes the issue.
- gRPC and HTTP2 do not work against routes. This is a known limitation of OpenShift routes.

### 12.5. RELEASE NOTES FOR RED HAT OPENS SHIFT SERVERLESS TECHNOLOGY PREVIEW 1.0.0

#### 12.5.1. New features

This release of OpenShift Serverless introduces the OpenShift Serverless Operator, which supports Knative Serving 0.7.1 and is tested for OpenShift Service Mesh 1.0.

#### 12.5.2. Known issues

The following limitations exist in OpenShift Serverless at this time:

- The Knative Serving Operator should wait for ServiceMeshMemberRoll to include the **knative-**



**serving** namespace. The installation procedure recommends creating the **knative-serving** namespace and then installing the operator. Istio does not consider the **knative-serving** namespace to be in the ServiceMeshMemberRoll when the Knative Serving Pods are being created. Consequently, the sidecars are not injected.

- Knative service returns a 503 status code when the cluster is running for a long time. The Knative Serving Pods do not show any errors. Restarting the **istio-pilot** Pod temporarily fixes the issue.
- The gRPC and HTTP2 do not work against routes. This is a known limitation of OpenShift routes.

## 12.6. ADDITIONAL RESOURCES

OpenShift Serverless is based on the open source Knative project.

- For details about the latest Knative Serving release, see the [Knative Serving releases page](#).
- For details about the latest Knative Client release, see the [Knative Client releases page](#).
- For details about the latest Knative Eventing release, see the [Knative Eventing releases page](#).



### NOTE

Knative Eventing is currently available as a Developer Preview on OpenShift Container Platform. See the upstream [Knative Eventing on OpenShift Container Platform documentation](#).