OpenShift Container Platform 3.9

Architecture

OpenShift Container Platform 3.9 Architecture Information

Last Updated: 2019-08-16

OpenShift Container Platform 3.9 Architecture

OpenShift Container Platform 3.9 Architecture Information

Legal Notice

Abstract

Learn the architecture of OpenShift Container Platform 3.9 including the infrastructure and core components. These topics also cover authentication, networking and source code management.

Table of Contents

CHAPTER 1. OVERVIEW

OpenShift v3 is a layered system designed to expose underlying Docker-formatted container image and Kubernetes concepts as accurately as possible, with a focus on easy composition of applications by a developer. For example, install Ruby, push code, and add MySQL.

Unlike OpenShift v2, more flexibility of configuration is exposed after creation in all aspects of the model. The concept of an application as a separate object is removed in favor of more flexible composition of "services", allowing two web containers to reuse a database or expose a database directly to the edge of the network.

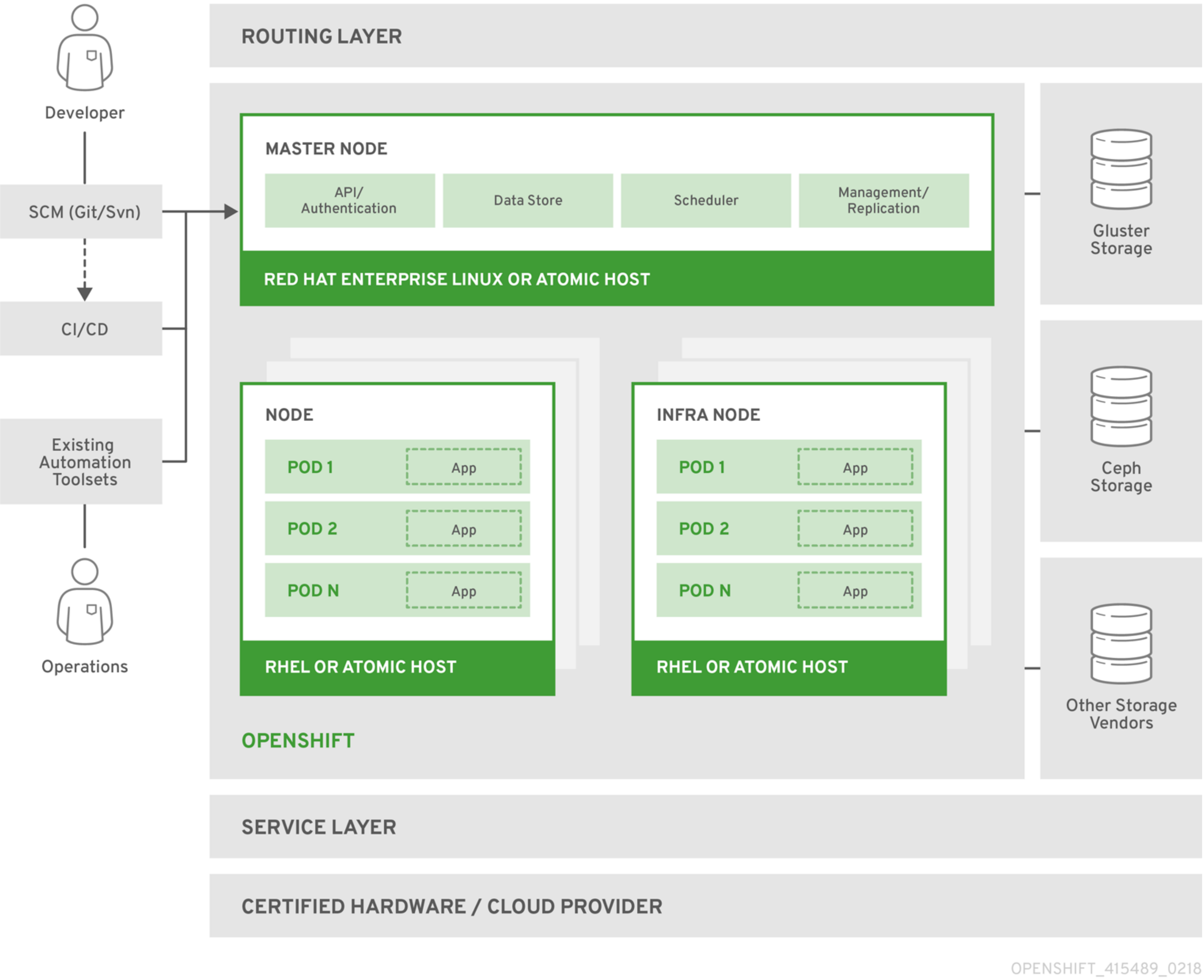
1.1. WHAT ARE THE LAYERS?

The Docker service provides the abstraction for packaging and creating Linux-based, lightweight container images. Kubernetes provides the cluster management and orchestrates containers on multiple hosts.

OpenShift Container Platform adds:

- Source code management, builds, and deployments for developers
- Managing and promoting images at scale as they flow through your system
- Application management at scale
- Team and user tracking for organizing a large developer organization
- Networking infrastructure that supports the cluster

Figure 1.1. OpenShift Container Platform Architecture Overview



1.2. WHAT IS THE OPENSHIFT CONTAINER PLATFORM ARCHITECTURE?

OpenShift Container Platform has a microservices-based architecture of smaller, decoupled units that work together. It runs on top of a Kubernetes cluster, with data about the objects stored in etcd, a reliable clustered key-value store. Those services are broken down by function:

- REST APIs, which expose each of the core objects.
- Controllers, which read those APIs, apply changes to other objects, and report status or write back to the object.

Users make calls to the REST API to change the state of the system. Controllers use the REST API to read the user's desired state, and then try to bring the other parts of the system into sync. For example, when a user requests a build they create a "build" object. The build controller sees that a new build has been created, and runs a process on the cluster to perform that build. When the build completes, the controller updates the build object via the REST API and the user sees that their build is complete.

The controller pattern means that much of the functionality in OpenShift Container Platform is extensible. The way that builds are run and launched can be customized independently of how images are managed, or how deployments happen. The controllers are performing the "business logic" of the system, taking user actions and transforming them into reality. By customizing those controllers or replacing them with your own logic, different behaviors can be implemented. From a system administration perspective, this also means the API can be used to script common administrative actions on a repeating schedule. Those scripts are also controllers that watch for changes and take action. OpenShift Container Platform makes the ability to customize the cluster in this way a first-class behavior.

To make this possible, controllers leverage a reliable stream of changes to the system to sync their view of the system with what users are doing. This event stream pushes changes from etcd to the REST API and then to the controllers as soon as changes occur, so changes can ripple out through the system very quickly and efficiently. However, since failures can occur at any time, the controllers must also be able to get the latest state of the system at startup, and confirm that everything is in the right state. This resynchronization is important, because it means that even if something goes wrong, then the operator can restart the affected components, and the system double checks everything before continuing. The system should eventually converge to the user's intent, since the controllers can always bring the system into sync.

1.3. HOW IS OPENSHIFT CONTAINER PLATFORM SECURED?

The OpenShift Container Platform and Kubernetes APIs authenticate users who present credentials, and then authorize them based on their role. Both developers and administrators can be authenticated via a number of means, primarily OAuth tokens and X.509 client certificates. OAuth tokens are signed with JSON Web Algorithm *RS256*, which is RSA signature algorithm PKCS#1 v1.5 with SHA-256.

Developers (clients of the system) typically make REST API calls from a client program like **oc** or to the web console via their browser, and use OAuth bearer tokens for most communications. Infrastructure components (like nodes) use client certificates generated by the system that contain their identities. Infrastructure components that run in containers use a token associated with their service account to connect to the API.

Authorization is handled in the OpenShift Container Platform policy engine, which defines actions like "create pod" or "list services" and groups them into roles in a policy document. Roles are bound to users or groups by the user or group identifier. When a user or service account attempts an action, the policy engine checks for one or more of the roles assigned to the user (e.g., cluster administrator or administrator of the current project) before allowing it to continue.

Since every container that runs on the cluster is associated with a service account, it is also possible to associate secrets to those service accounts and have them automatically delivered into the container. This enables the infrastructure to manage secrets for pulling and pushing images, builds, and the deployment components, and also allows application code to easily leverage those secrets.

1.3.1. TLS Support

All communication channels with the REST API, as well as between master components such as etcd and the API server, are secured with TLS. TLS provides strong encryption, data integrity, and authentication of servers with X.509 server certificates and public key infrastructure. By default, a new internal PKI is created for each deployment of OpenShift Container Platform. The internal PKI uses 2048 bit RSA keys and SHA-256 signatures. Custom certificates for public hosts are supported as well.

OpenShift Container Platform uses Golang's standard library implementation of **crypto/tls** and does not depend on any external crypto and TLS libraries. Additionally, the client depends on external libraries for GSSAPI authentication and OpenPGP signatures. GSSAPI is typically provided by either MIT Kerberos or Heimdal Kerberos, which both use OpenSSL's libcrypto. OpenPGP signature verification is handled by libgpgme and GnuPG.

The insecure versions SSL 2.0 and SSL 3.0 are unsupported and not available. The OpenShift Container Platform server and **oc** client only provide TLS 1.2 by default. TLS 1.0 and TLS 1.1 can be enabled in the server configuration. Both server and client prefer modern cipher suites with authenticated encryption algorithms and perfect forward secrecy. Cipher suites with deprecated and insecure algorithms such as RC4, 3DES, and MD5 are disabled. Some internal clients (for example, LDAP authentication) have less restrict settings with TLS 1.0 to 1.2 and more cipher suites enabled.

Table 1.1. Supported TLS Versions

TLS Version	OpenShift Container Platform Server	**oc** Client	Other Clients
SSL 2.0	Unsupported	Unsupported	Unsupported
SSL 3.0	Unsupported	Unsupported	Unsupported
TLS 1.0	No [a]	No [a]	Maybe [b]
TLS 1.1	No [a]	No [a]	Maybe [b]
TLS 1.2	**Yes**	**Yes**	**Yes**
TLS 1.3	N/A [c]	N/A [c]	N/A [c]

[a] Disabled by default, but can be enabled in the server configuration.

[b] Some internal clients, such as the LDAP client.

[c] TLS 1.3 is still under development.

The following list of enabled cipher suites of OpenShift Container Platform's server and **oc** client are sorted in preferred order:

- **TLS_ECDHE_ECDSA_WITH_CHACHA20_POLY1305**
- **TLS_ECDHE_RSA_WITH_CHACHA20_POLY1305**
- **TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256**
- **TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256**
- **TLS_ECDHE_ECDSA_WITH_AES_256_GCM_SHA384**
- **TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384**
- **TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA256**
- **TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256**
- **TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA**
- **TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA**
- **TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA**
- **TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA**

- **TLS_RSA_WITH_AES_128_GCM_SHA256**
- **TLS_RSA_WITH_AES_256_GCM_SHA384**
- **TLS_RSA_WITH_AES_128_CBC_SHA**
- **TLS_RSA_WITH_AES_256_CBC_SHA**

CHAPTER 2. INFRASTRUCTURE COMPONENTS

2.1. KUBERNETES INFRASTRUCTURE

2.1.1. Overview

Within OpenShift Container Platform, Kubernetes manages containerized applications across a set of containers or hosts and provides mechanisms for deployment, maintenance, and application-scaling. The Docker service packages, instantiates, and runs containerized applications. A Kubernetes cluster consists of one or more masters and a set of nodes.

You can optionally configure your masters for high availability (HA) to ensure that the cluster has no single point of failure.

NOTE

OpenShift Container Platform uses Kubernetes 1.9 and Docker 1.13.

2.1.2. Masters

The master is the host or hosts that contain the master components, including the API server, controller manager server, and etcd. The master manages nodes in its Kubernetes cluster and schedules pods to run on nodes.

Table 2.1. Master Components

Component	Description
API Server	The Kubernetes API server validates and configures the data for pods, services, and replication controllers. It also assigns pods to nodes and synchronizes pod information with service configuration. Can be run as a standalone process.
etcd	etcd stores the persistent master state while other components watch etcd for changes to bring themselves into the desired state. etcd can be optionally configured for high availability, typically deployed with 2n+1 peer services.
Controller Manager Server	The controller manager server watches etcd for changes to replication controller objects and then uses the API to enforce the desired state. Can be run as a standalone process. Several such processes create a cluster with one active leader at a time.
HAProxy	Optional, used when configuring highly-available masters with the **native** method to balance load between API master endpoints. The advanced installation method can configure HAProxy for you with the**native** method. Alternatively, you can use the **native** method but pre-configure your own load balancer of choice.

2.1.2.1. High Availability Masters

You can optionally configure your masters for high availability (HA) to ensure that the cluster has no single point of failure.

To mitigate concerns about availability of the master, two activities are recommended:

1. A runbook entry should be created for reconstructing the master. A runbook entry is a necessary backstop for any highly-available service. Additional solutions merely control the frequency that the runbook must be consulted. For example, a cold standby of the master host can adequately fulfill SLAs that require no more than minutes of downtime for creation of new applications or recovery of failed application components.

2. Use a high availability solution to configure your masters and ensure that the cluster has no single point of failure. The advanced installation method provides specific examples using the **native** HA method and configuring HAProxy. You can also take the concepts and apply them towards your existing HA solutions using the **native** method instead of HAProxy.

NOTE

In production OpenShift Container Platform clusters, you must maintain high availability of the API Server load balancer. If the API Server load balancer is not available, nodes cannot report their status, all their pods are marked dead, and the pods' endpoints are removed from the service.

In addition to configuring HA for OpenShift Container Platform, you must separately configure HA for the API Server load balancer. To configure HA, it is much preferred to integrate an enterprise load balancer (LB) such as an F5 Big-IP™ or a Citrix Netscaler™ appliance. If such solutions are not available, it is possible to run multiple HAProxy load balancers and use Keepalived to provide a floating virtual IP address for HA. However, this solution is not recommended for production instances.

When using the **native** HA method with HAProxy, master components have the following availability:

Table 2.2. Availability Matrix with HAProxy

Role	Style	Notes
etcd	Active-active	Fully redundant deployment with load balancing. Can be installed on separate hosts or collocated on master hosts.
API Server	Active-active	Managed by HAProxy.
Controller Manager Server	Active-passive	One instance is elected as a cluster leader at a time.
HAProxy	Active-passive	Balances load between API master endpoints.

While clustered etcd requires an odd number of hosts for quorum, the master services have no quorum or requirement that they have an odd number of hosts. However, since you need at least two master services for HA, it is common to maintain a uniform odd number of hosts when collocating master services and etcd.

2.1.3. Nodes

A node provides the runtime environments for containers. Each node in a Kubernetes cluster has the required services to be managed by the master. Nodes also have the required services to run pods, including the Docker service, a kubelet, and a service proxy.

OpenShift Container Platform creates nodes from a cloud provider, physical systems, or virtual systems. Kubernetes interacts with node objects that are a representation of those nodes. The master uses the information from node objects to validate nodes with health checks. A node is ignored until it passes the health checks, and the master continues checking nodes until they are valid. The Kubernetes documentation has more information on node management.

Administrators can manage nodes in an OpenShift Container Platform instance using the CLI. To define full configuration and security options when launching node servers, use dedicated node configuration files.

IMPORTANT

See the cluster limits section for the recommended maximum number of nodes.

2.1.3.1. Kubelet

Each node has a kubelet that updates the node as specified by a container manifest, which is a YAML file that describes a pod. The kubelet uses a set of manifests to ensure that its containers are started and that they continue to run.

A container manifest can be provided to a kubelet by:

- A file path on the command line that is checked every 20 seconds.
- An HTTP endpoint passed on the command line that is checked every 20 seconds.
- The kubelet watching an etcd server, such as ***/registry/hosts/$(hostname -f)***, and acting on any changes.
- The kubelet listening for HTTP and responding to a simple API to submit a new manifest.

2.1.3.2. Service Proxy

Each node also runs a simple network proxy that reflects the services defined in the API on that node. This allows the node to do simple TCP and UDP stream forwarding across a set of back ends.

2.1.3.3. Node Object Definition

The following is an example node object definition in Kubernetes:

```
apiVersion: v1 ①
kind: Node ②
metadata:
  creationTimestamp: null
  labels: ③
    kubernetes.io/hostname: node1.example.com
  name: node1.example.com ④
spec:
  externalID: node1.example.com ⑤
status:
  nodeInfo:
    bootID: ""
    containerRuntimeVersion: ""
    kernelVersion: ""
```

```
kubeProxyVersion: ""
kubeletVersion: ""
machineID: ""
osImage: ""
systemUUID: ""
```

1. **apiVersion** defines the API version to use.
2. **kind** set to **Node** identifies this as a definition for a node object.
3. **metadata.labels** lists any labels that have been added to the node.
4. **metadata.name** is a required value that defines the name of the node object. This value is shown in the **NAME** column when running the **oc get nodes** command.
5. **spec.externalID** defines the fully-qualified domain name where the node can be reached. Defaults to the **metadata.name** value when empty.

2.2. CONTAINER REGISTRY

2.2.1. Overview

OpenShift Container Platform can utilize any server implementing the Docker registry API as a source of images, including the Docker Hub, private registries run by third parties, and the integrated OpenShift Container Platform registry.

2.2.2. Integrated OpenShift Container Registry

OpenShift Container Platform provides an integrated container registry called *OpenShift Container Registry* (OCR) that adds the ability to automatically provision new image repositories on demand. This provides users with a built-in location for their application builds to push the resulting images.

Whenever a new image is pushed to OCR, the registry notifies OpenShift Container Platform about the new image, passing along all the information about it, such as the namespace, name, and image metadata. Different pieces of OpenShift Container Platform react to new images, creating new builds and deployments.

OCR can also be deployed as a stand-alone component that acts solely as a container registry, without the build and deployment integration. See Installing a Stand-alone Deployment of OpenShift Container Registry for details.

2.2.3. Third Party Registries

OpenShift Container Platform can create containers using images from third party registries, but it is unlikely that these registries offer the same image notification support as the integrated OpenShift Container Platform registry. In this situation OpenShift Container Platform will fetch tags from the remote registry upon imagestream creation. Refreshing the fetched tags is as simple as running **oc import-image <stream>**. When new images are detected, the previously-described build and deployment reactions occur.

2.2.3.1. Authentication

OpenShift Container Platform can communicate with registries to access private image repositories using credentials supplied by the user. This allows OpenShift to push and pull images to and from private repositories. The Authentication topic has more information.

2.3. WEB CONSOLE

2.3.1. Overview

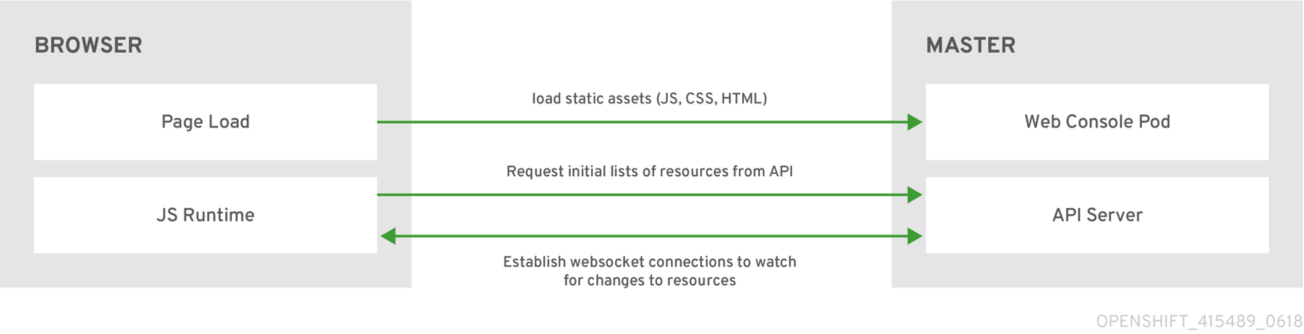
The OpenShift Container Platform web console is a user interface accessible from a web browser. Developers can use the web console to visualize, browse, and manage the contents of projects.

> **NOTE**
>
> JavaScript must be enabled to use the web console. For the best experience, use a web browser that supports WebSockets.

The web console runs as a pod on the master. The static assets required to run the web console are served by the pod. Administrators can also customize the web console using extensions, which let you run scripts and load custom stylesheets when the web console loads.

When you access the web console from a browser, it first loads all required static assets. It then makes requests to the OpenShift Container Platform APIs using the values defined from the **openshift start** option **--public-master**, or from the related parameter **masterPublicURL** in the **webconsole-config** config map defined in the **openshift-web-console** namespace. The web console uses WebSockets to maintain a persistent connection with the API server and receive updated information as soon as it is available.

Figure 2.1. Web Console Request Architecture


BROWSER
Page Load
JS Runtime
MASTER
Web Console Pod
API Server
load static assets (JS, CSS, HTML)
Request initial lists of resources from API
Establish websocket connections to watch for changes to resources
OPENSHIFT_415489_0618

The configured host names and IP addresses for the web console are whitelisted to access the API server safely even when the browser would consider the requests to be cross-origin. To access the API server from a web application using a different host name, you must whitelist that host name by specifying the **--cors-allowed-origins** option on **openshift start** or from the related master configuration file parameter **corsAllowedOrigins**.

The **corsAllowedOrigins** parameter is controlled by the configuration field. No pinning or escaping is done to the value. The following is an example of how you can pin a host name and escape dots:

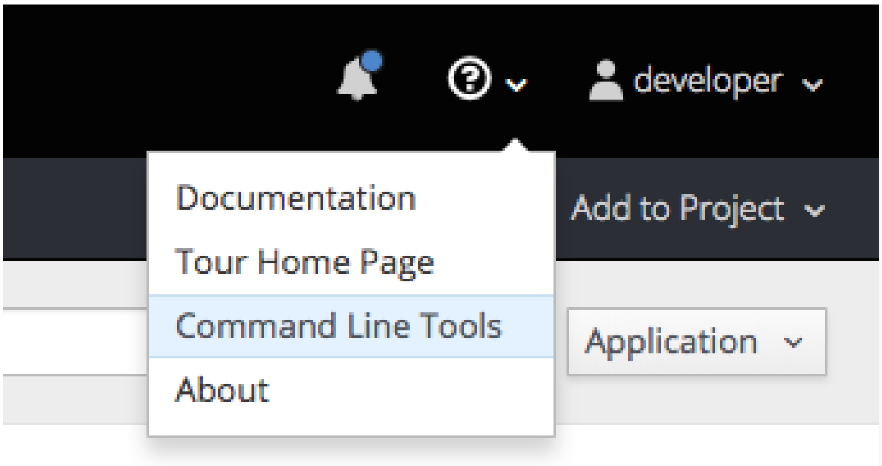
```
corsAllowedOrigins:
- (?i)//my\.subdomain\.domain\.com(:|\z)
```

- The **(?i)** makes it case-insensitive.

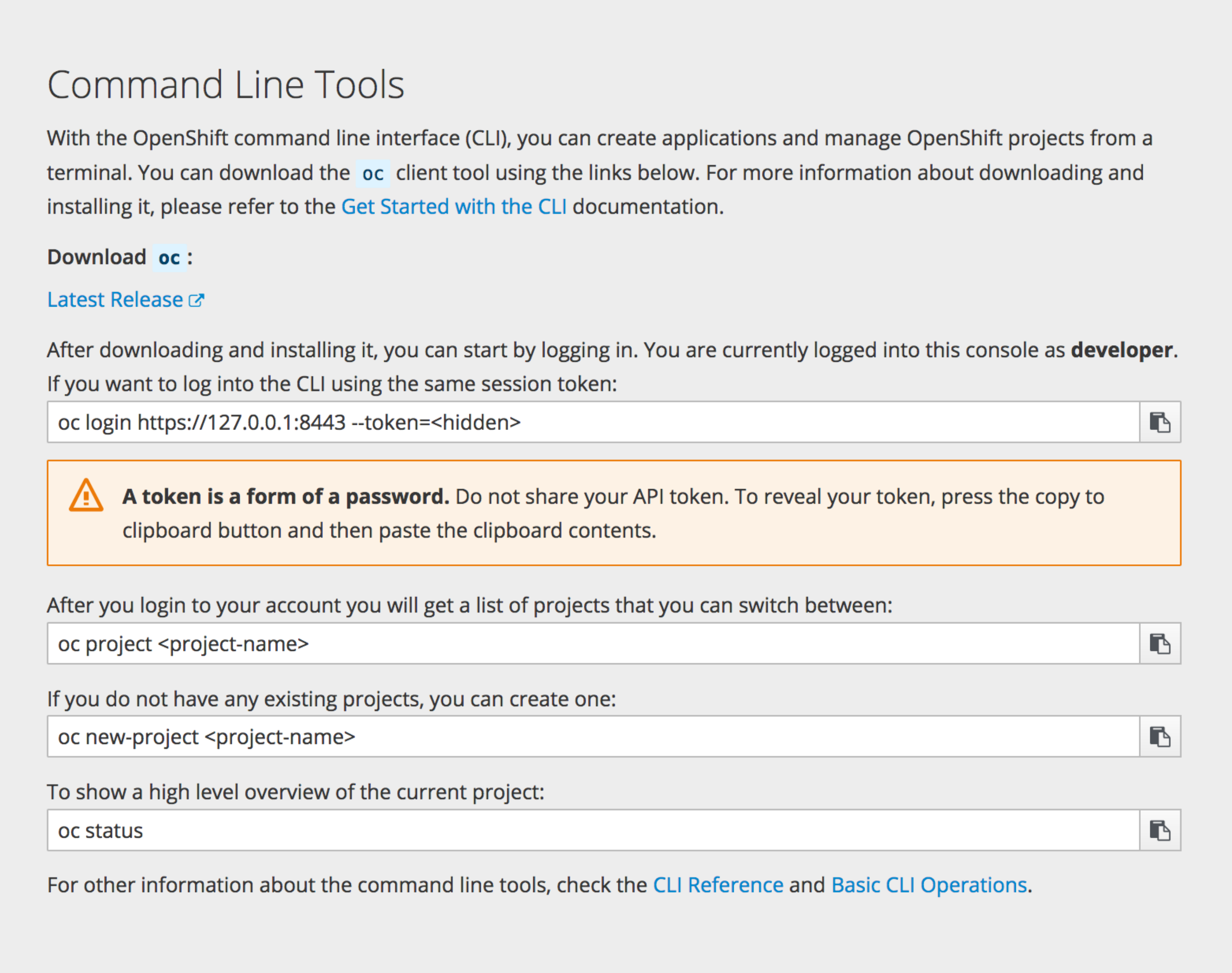
- The // pins to the beginning of the domain (and matches the double slash following **http:** or **https:**).
- The \. escapes dots in the domain name.
- The **(:|\z)** matches the end of the domain name **(\z)** or a port separator **(:)**.

2.3.2. CLI Downloads

You can access CLI downloads from the Help icon in the web console:



Cluster administrators can customize these links further.



Command Line Tools

With the OpenShift command line interface (CLI), you can create applications and manage OpenShift projects from a terminal. You can download the `oc` client tool using the links below. For more information about downloading and installing it, please refer to the Get Started with the CLI documentation.

Download `oc`:

Latest Release

After downloading and installing it, you can start by logging in. You are currently logged into this console as **developer**. If you want to log into the CLI using the same session token:

```
oc login https://127.0.0.1:8443 --token=<hidden>
```

A token is a form of a password. Do not share your API token. To reveal your token, press the copy to clipboard button and then paste the clipboard contents.

After you login to your account you will get a list of projects that you can switch between:

```
oc project <project-name>
```

If you do not have any existing projects, you can create one:

```
oc new-project <project-name>
```

To show a high level overview of the current project:

```
oc status
```

For other information about the command line tools, check the CLI Reference and Basic CLI Operations.

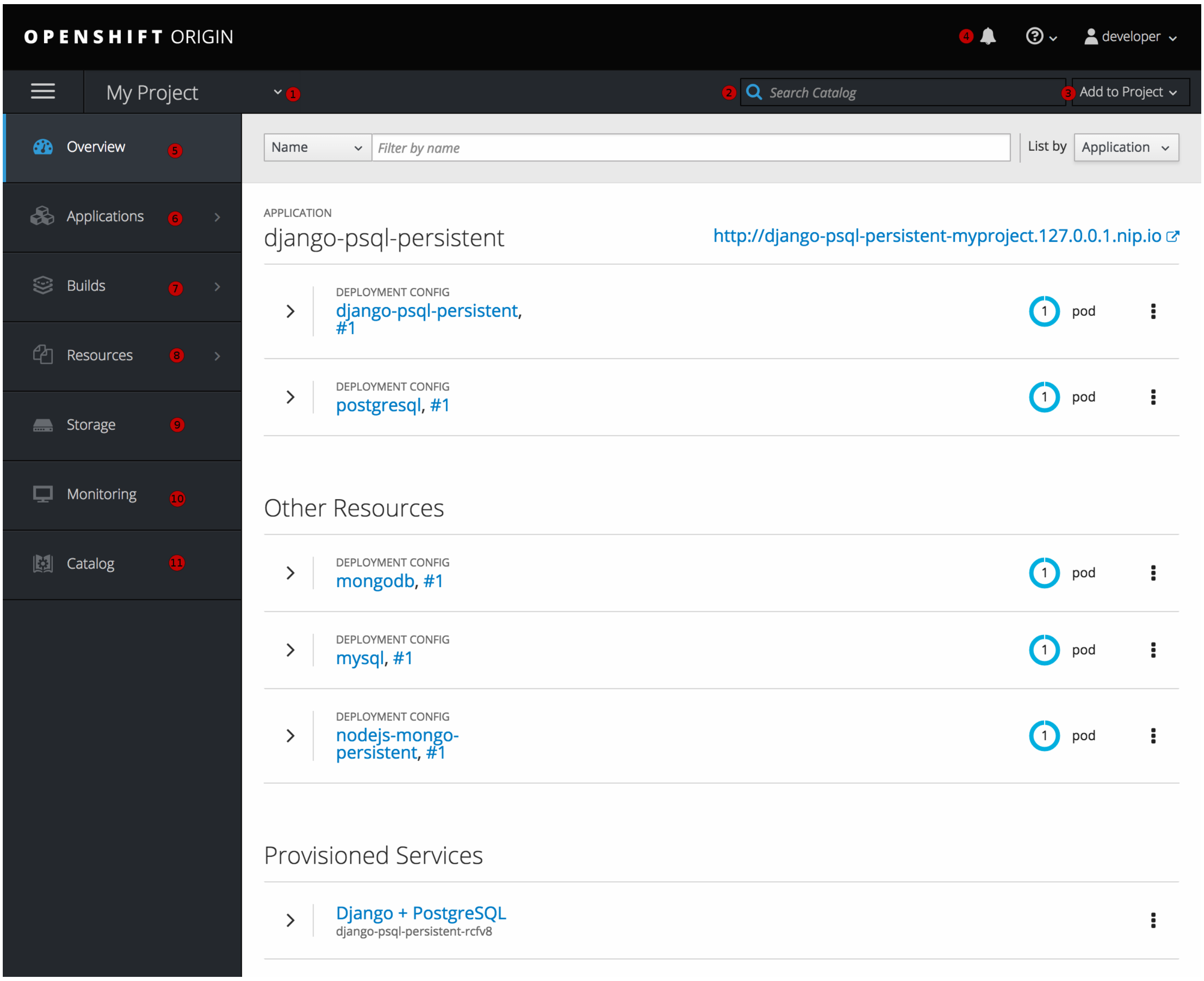
2.3.3. Browser Requirements

Review the tested integrations for OpenShift Container Platform.

2.3.4. Project Overviews

After logging in, the web console provides developers with an overview for the currently selected project:

Figure 2.2. Web Console Project Overview



The project selector allows you to switch between projects you have access to.

To quickly find services from within project view, type in your search criteria

Create new applications using a source repository or service from the service catalog.

Notifications related to your project.

The **Overview** tab (currently selected) visualizes the contents of your project with a high-level view of each component.

Applications tab: Browse and perform actions on your deployments, pods, services, and routes.

Builds tab: Browse and perform actions on your builds and image streams.

Resources tab: View your current quota consumption and other resources.

Storage tab: View persistent volume claims and request storage for your applications.

Monitoring tab: View logs for builds, pods, and deployments, as well as event notifications for all objects in your project.

Catalog tab: Quickly get to the catalog from within a project.

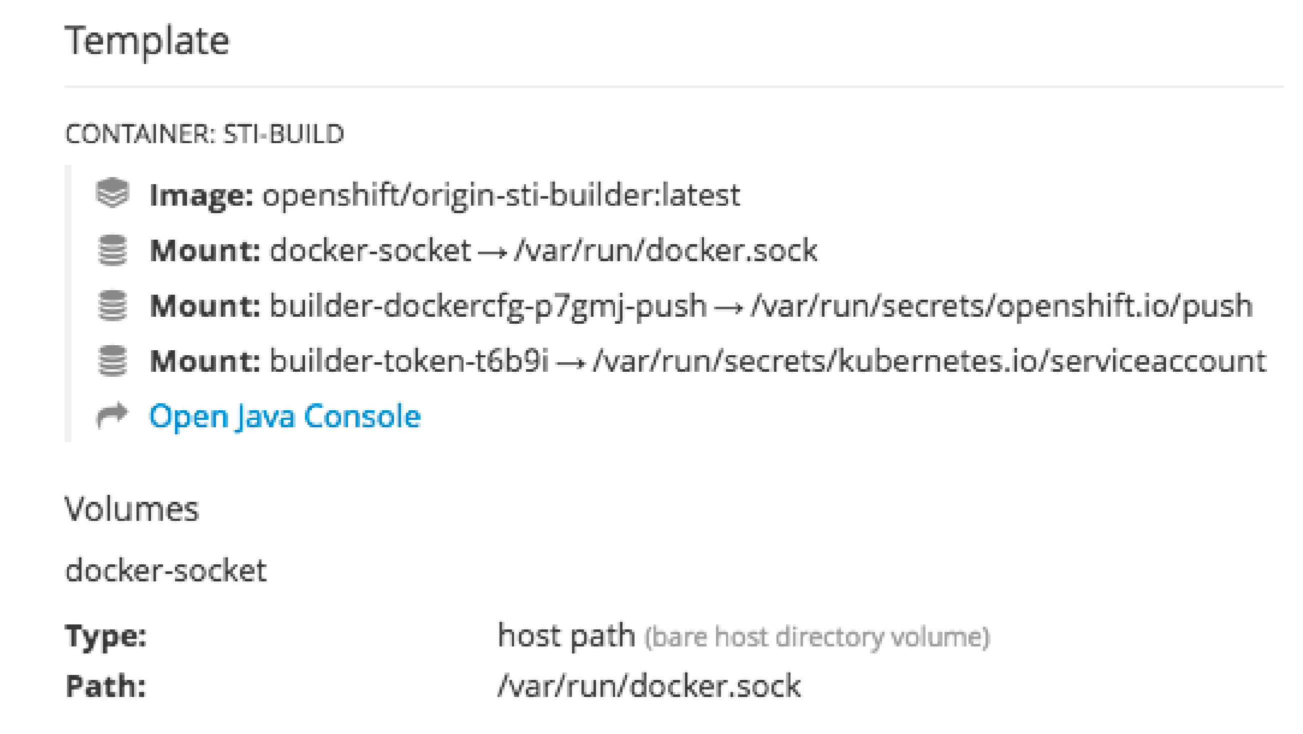
NOTE

Cockpit is automatically installed and enabled in OpenShift Container Platform 3.1 and later to help you monitor your development environment. Red Hat Enterprise Linux Atomic Host: Getting Started with Cockpit provides more information on using Cockpit.

2.3.5. JVM Console

For pods based on Java images, the web console also exposes access to a hawt.io-based JVM console for viewing and managing any relevant integration components. A **Connect** link is displayed in the pod's details on the *Browse* → *Pods* page, provided the container has a port named **jolokia**.

Figure 2.3. Pod with a Link to the JVM Console



Template

CONTAINER: STI-BUILD

Image: openshift/origin-sti-builder:latest

Mount: docker-socket → /var/run/docker.sock

Mount: builder-dockercfg-p7gmj-push → /var/run/secrets/openshift.io/push

Mount: builder-token-t6b9i → /var/run/secrets/kubernetes.io/serviceaccount

Open Java Console

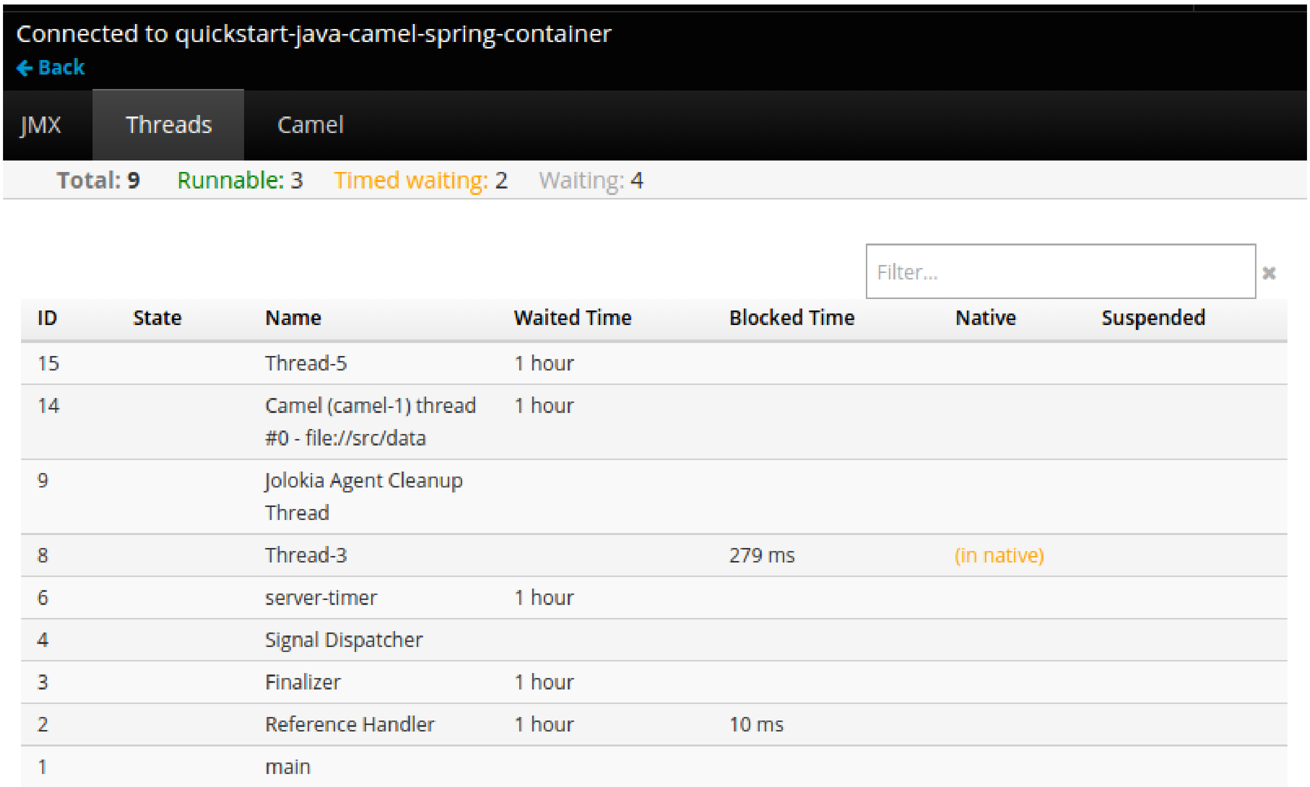
Volumes

docker-socket

Type:	host path (bare host directory volume)
Path:	/var/run/docker.sock

After connecting to the JVM console, different pages are displayed depending on which components are relevant to the connected pod.

Figure 2.4. JVM Console



Connected to quickstart-java-camel-spring-container

Total: 9 Runnable: 3 Timed waiting: 2 Waiting: 4

ID	State	Name	Waited Time	Blocked Time	Native	Suspended
15		Thread-5	1 hour			
14		Camel (camel-1) thread #0 - file://src/data	1 hour			
9		Jolokia Agent Cleanup Thread				
8		Thread-3		279 ms	(in native)	
6		server-timer	1 hour			
4		Signal Dispatcher				
3		Finalizer	1 hour			
2		Reference Handler	1 hour	10 ms		
1		main				

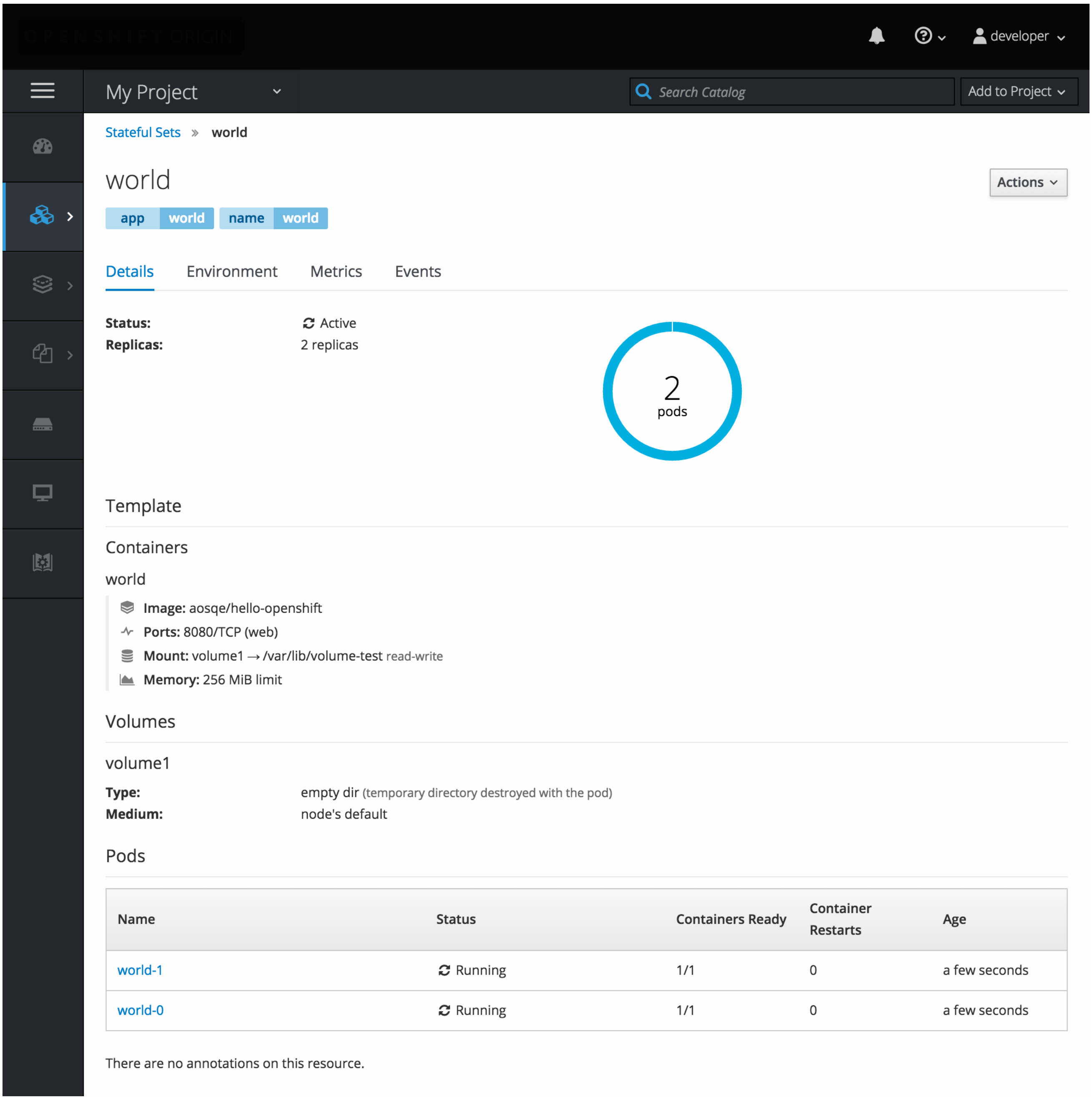
The following pages are available:

Page	Description
JMX	View and manage JMX domains and mbeans.
Threads	View and monitor the state of threads.
ActiveMQ	View and manage Apache ActiveMQ brokers.
Camel	View and manage Apache Camel routes and dependencies.
OSGi	View and manage the JBoss Fuse OSGi environment.

2.3.6. StatefulSets

A **StatefulSet** controller provides a unique identity to its pods and determines the order of deployments and scaling. **StatefulSet** is useful for unique network identifiers, persistent storage, graceful deployment and scaling, and graceful deletion and termination.

Figure 2.5. StatefulSet in OpenShift Container Platform



Stateful Sets » world

world

app world name world

Status: Active
Replicas: 2 replicas

2 pods

Template

Containers

world

- **Image:** aosqe/hello-openshift
- **Ports:** 8080/TCP (web)
- **Mount:** volume1 → /var/lib/volume-test read-write
- **Memory:** 256 MiB limit

Volumes

volume1

Type: empty dir (temporary directory destroyed with the pod)
Medium: node's default

Pods

Name	Status	Containers Ready	Container Restarts	Age
world-1	Running	1/1	0	a few seconds
world-0	Running	1/1	0	a few seconds

There are no annotations on this resource.

CHAPTER 3. CORE CONCEPTS

3.1. OVERVIEW

The following topics provide high-level, architectural information on core concepts and objects you will encounter when using OpenShift Container Platform. Many of these objects come from Kubernetes, which is extended by OpenShift Container Platform to provide a more feature-rich development lifecycle platform.

- Containers and images are the building blocks for deploying your applications.
- Pods and services allow for containers to communicate with each other and proxy connections.
- Projects and users provide the space and means for communities to organize and manage their content together.
- Builds and image streams allow you to build working images and react to new images.
- Deployments add expanded support for the software development and deployment lifecycle.
- Routes announce your service to the world.
- Templates allow for many objects to be created at once based on customized parameters.

3.2. CONTAINERS AND IMAGES

3.2.1. Containers

The basic units of OpenShift Container Platform applications are called *containers*. Linux container technologies are lightweight mechanisms for isolating running processes so that they are limited to interacting with only their designated resources.

Many application instances can be running in containers on a single host without visibility into each others' processes, files, network, and so on. Typically, each container provides a single service (often called a "micro-service"), such as a web server or a database, though containers can be used for arbitrary workloads.

The Linux kernel has been incorporating capabilities for container technologies for years. More recently the Docker project has developed a convenient management interface for Linux containers on a host. OpenShift Container Platform and Kubernetes add the ability to orchestrate Docker-formatted containers across multi-host installations.

Though you do not directly interact with the Docker CLI or service when using OpenShift Container Platform, understanding their capabilities and terminology is important for understanding their role in OpenShift Container Platform and how your applications function inside of containers. The **docker** RPM is available as part of RHEL 7, as well as CentOS and Fedora, so you can experiment with it separately from OpenShift Container Platform. Refer to the article Get Started with Docker Formatted Container Images on Red Hat Systems for a guided introduction.

3.2.1.1. Init Containers

A pod can have init containers in addition to application containers. Init containers allow you to reorganize setup scripts and binding code. An init container differs from a regular container in that it always runs to completion. Each init container must complete successfully before the next one is started.

For more information, see Pods and Services.

3.2.2. Images

Containers in OpenShift Container Platform are based on Docker-formatted container *images*. An image is a binary that includes all of the requirements for running a single container, as well as metadata describing its needs and capabilities.

You can think of it as a packaging technology. Containers only have access to resources defined in the image unless you give the container additional access when creating it. By deploying the same image in multiple containers across multiple hosts and load balancing between them, OpenShift Container Platform can provide redundancy and horizontal scaling for a service packaged into an image.

You can use the Docker CLI directly to build images, but OpenShift Container Platform also supplies builder images that assist with creating new images by adding your code or configuration to existing images.

Because applications develop over time, a single image name can actually refer to many different versions of the "same" image. Each different image is referred to uniquely by its hash (a long hexadecimal number e.g. **fd44297e2ddb050ec4f…**) which is usually shortened to 12 characters (e.g. **fd44297e2ddb**).

Image Version Tag Policy

Rather than version numbers, the Docker service allows applying tags (such as **v1**, **v2.1**, **GA**, or the default **latest**) in addition to the image name to further specify the image desired, so you may see the same image referred to as **centos** (implying the **latest** tag), **centos:centos7**, or **fd44297e2ddb**.

> **WARNING**
>
> Do not use the **latest** tag for any official OpenShift Container Platform images. These are images that start with **openshift3/**. **latest** can refer to a number of versions, such as **3.4**, or **3.5**.

How you tag the images dictates the updating policy. The more specific you are, the less frequently the image will be updated. Use the following to determine your chosen OpenShift Container Platform images policy:

vX.Y

The vX.Y tag points to X.Y.Z-<number>. For example, if the **registry-console** image is updated to v3.4, it points to the newest 3.4.Z-<number> tag, such as 3.4.1-8.

X.Y.Z

Similar to the vX.Y example above, the X.Y.Z tag points to the latest X.Y.Z-<number>. For example, 3.4.1 would point to 3.4.1-8

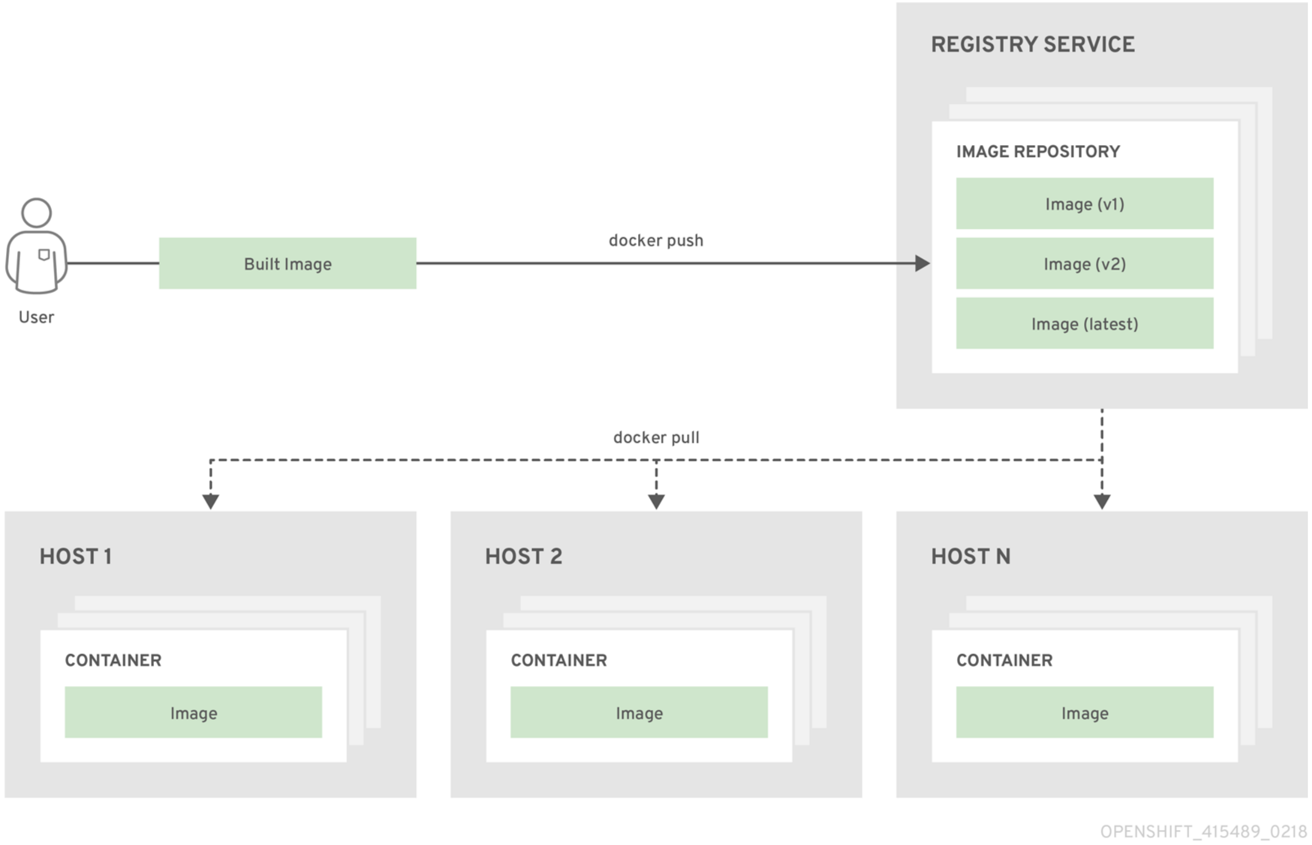
X.Y.Z-<number>

The tag is unique and does not change. When using this tag, the image does not update if an image is updated. For example, the 3.4.1-8 will always point to 3.4.1-8, even if an image is updated.

3.2.3. Container Registries

A container registry is a service for storing and retrieving Docker-formatted container images. A registry contains a collection of one or more image repositories. Each image repository contains one or more tagged images. Docker provides its own registry, the Docker Hub, and you can also use private or third-party registries. Red Hat provides a registry at **registry.access.redhat.com** for subscribers. OpenShift Container Platform can also supply its own internal registry for managing custom container images.

The relationship between containers, images, and registries is depicted in the following diagram:



3.3. PODS AND SERVICES

3.3.1. Pods

OpenShift Container Platform leverages the Kubernetes concept of a *pod*, which is one or more containers deployed together on one host, and the smallest compute unit that can be defined, deployed, and managed.

Pods are the rough equivalent of a machine instance (physical or virtual) to a container. Each pod is allocated its own internal IP address, therefore owning its entire port space, and containers within pods can share their local storage and networking.

Pods have a lifecycle; they are defined, then they are assigned to run on a node, then they run until their container(s) exit or they are removed for some other reason. Pods, depending on policy and exit code, may be removed after exiting, or may be retained in order to enable access to the logs of their containers.

OpenShift Container Platform treats pods as largely immutable; changes cannot be made to a pod definition while it is running. OpenShift Container Platform implements changes by terminating an existing pod and recreating it with modified configuration, base image(s), or both. Pods are also treated as expendable, and do not maintain state when recreated. Therefore pods should usually be managed by higher-level controllers, rather than directly by users.

NOTE

For the maximum number of pods per OpenShift Container Platform node host, see the Cluster Limits.

WARNING

Bare pods that are not managed by a replication controller will be not rescheduled upon node disruption.

Below is an example definition of a pod that provides a long-running service, which is actually a part of the OpenShift Container Platform infrastructure: the integrated container registry. It demonstrates many features of pods, most of which are discussed in other topics and thus only briefly mentioned here:

Example 3.1. Pod Object Definition (YAML)

```
apiVersion: v1
kind: Pod
metadata:
  annotations: { ... }
  labels:                  ❶
    deployment: docker-registry-1
    deploymentconfig: docker-registry
    docker-registry: default
  generateName: docker-registry-1-      ❷
spec:
  containers:              ❸
  - env:                   ❹
    - name: OPENSHIFT_CA_DATA
      value: ...
    - name: OPENSHIFT_CERT_DATA
      value: ...
    - name: OPENSHIFT_INSECURE
      value: "false"
    - name: OPENSHIFT_KEY_DATA
      value: ...
    - name: OPENSHIFT_MASTER
      value: https://master.example.com:8443
    image: openshift/origin-docker-registry:v0.6.2 ❺
    imagePullPolicy: IfNotPresent
    name: registry
    ports:                 ❻
    - containerPort: 5000
      protocol: TCP
    resources: {}
    securityContext: { ... }    ❼
    volumeMounts:                  (8)
    - mountPath: /registry
      name: registry-storage
    - mountPath: /var/run/secrets/kubernetes.io/serviceaccount
      name: default-token-br6yz
      readOnly: true
  dnsPolicy: ClusterFirst
  imagePullSecrets:
  - name: default-dockercfg-at06w
  restartPolicy: Always            (9)
  serviceAccount: default          (10)
  volumes:                         (11)
  - emptyDir: {}
    name: registry-storage
  - name: default-token-br6yz
    secret:
      secretName: default-token-br6yz
```

1. Pods can be "tagged" with one or more labels, which can then be used to select and manage groups of pods in a single operation. The labels are stored in key/value format in the **metadata** hash. One label in this example is **docker-registry=default**.

2. Pods must have a unique name within their namespace. A pod definition may specify the basis of a name with the **generateName** attribute, and random characters will be added automatically to generate a unique name.

3. **containers** specifies an array of container definitions; in this case (as with most), just one.

4. Environment variables can be specified to pass necessary values to each container.

5. Each container in the pod is instantiated from its own Docker-formatted container image.

6. The container can bind to ports which will be made available on the pod's IP.

7. OpenShift Container Platform defines a security context for containers which specifies whether they are allowed to run as privileged containers, run as a user of their choice, and more. The default context is very restrictive but administrators can modify this as needed.

8. The container specifies where external storage volumes should be mounted within the container. In this case, there is a volume for storing the registry's data, and one for access to credentials the registry needs for making requests against the OpenShift Container Platform API.

9. The pod restart policy with possible values **Always**, **OnFailure**, and **Never**. The default value is **Always**.

10. Pods making requests against the OpenShift Container Platform API is a common enough pattern that there is a **serviceAccount** field for specifying which service account user the pod should authenticate as when making the requests. This enables fine-grained access control for custom infrastructure components.

11. The pod defines storage volumes that are available to its container(s) to use. In this case, it provides an ephemeral volume for the registry storage and a **secret** volume containing the service account credentials.

NOTE

This pod definition does not include attributes that are filled by OpenShift Container Platform automatically after the pod is created and its lifecycle begins. The Kubernetes pod documentation has details about the functionality and purpose of pods.

3.3.1.1. Pod Restart Policy

A pod restart policy determines how OpenShift Container Platform responds when containers in that pod exit. The policy applies to all containers in that pod.

The possible values are:

- **Always** - Tries restarting a successfully exited container on the pod continuously, with an exponential back-off delay (10s, 20s, 40s) until the pod is restarted. The default is **Always**.
- **OnFailure** - Tries restarting a failed container on the pod with an exponential back-off delay (10s, 20s, 40s) capped at 5 minutes.
- **Never** - Does not try to restart exited or failed containers on the pod. Pods immediately fail and exit.

Once bound to a node, a pod will never be bound to another node. This means that a controller is necessary in order for a pod to survive node failure:

Condition	Controller Type	Restart Policy
Pods that are expected to terminate (such as batch computations)	Job	**OnFailure** or **Never**
Pods that are expected to not terminate (such as web servers)	Replication Controller	**Always**.
Pods that need to run one-per-machine	Daemonset	Any

If a container on a pod fails and the restart policy is set to **OnFailure**, the pod stays on the node and the container is restarted. If you do not want the container to restart, use a restart policy of **Never**.

If an entire pod fails, OpenShift Container Platform starts a new pod. Developers need to address the possibility that applications might be restarted in a new pod. In particular, applications need to handle temporary files, locks, incomplete output, and so forth caused by previous runs.

NOTE

Kubernetes architecture expects reliable endpoints from cloud providers. When a cloud provider is down, the kubelet prevents OpenShift Container Platform from restarting.

If the underlying cloud provider endpoints are not reliable, do not install a cluster using cloud provider integration. Install the cluster as if it was in a no-cloud environment. It is not recommended to toggle cloud provider integration on or off in an installed cluster.

For details on how OpenShift Container Platform uses restart policy with failed containers, see the Example States in the Kubernetes documentation.

3.3.1.2. Injecting Information into Pods Using Pod Presets

A *pod preset* is an object that injects user-specified information into pods as they are created.

IMPORTANT

As of OpenShift Container Platform 3.7, pod presets are no longer supported.

Using pod preset objects you can inject:

- secret objects
- **ConfigMap** objects
- storage volumes
- container volume mounts
- environment variables

Developers need to ensure the pod labels match the label selector on the PodPreset in order to add all that information to the pod. The label on a pod associates the pod with one or more pod preset objects that have a matching label selectors.

Using pod presets, a developer can provision pods without needing to know the details about the services the pod will consume. An administrator can keep configuration items of a service invisible from a developer without preventing the developer from deploying pods.

NOTE

The Pod Preset feature is available only if the Service Catalog has been installed.

You can exclude specific pods from being injected using the **podpreset.admission.kubernetes.io/exclude: "true"** parameter in the pod specification. See the example pod specification.

For more information, see Injecting Information into Pods Using Pod Presets .

3.3.2. Init Containers

An init container is a container in a pod that is started before the pod app containers are started. Init containers can share volumes, perform network operations, and perform computations before the remaining containers start. Init containers can also block or delay the startup of application containers until some precondition is met.

When a pod starts, after the network and volumes are initialized, the init containers are started in order. Each init container must exit successfully before the next is invoked. If an init container fails to start (due to the runtime) or exits with failure, it is retried according to the pod restart policy.

A pod cannot be ready until all init containers have succeeded.

See the Kubernetes documentation for some init container usage examples.

The following example outlines a simple pod which has two init containers. The first init container waits for **myservice** and the second waits for **mydb**. Once both containers succeed, the Pod starts.

Example 3.2. Sample Init Container Pod Object Definition (YAML)

```
apiVersion: v1
kind: Pod
metadata:
  name: myapp-pod
  labels:
    app: myapp
spec:
  containers:
  - name: myapp-container
    image: busybox
    command: ['sh', '-c', 'echo The app is running! && sleep 3600']
  initContainers:
  - name: init-myservice ①
    image: busybox
    command: ['sh', '-c', 'until nslookup myservice; do echo waiting for myservice; sleep 2; done;']
  - name: init-mydb ②
    image: busybox
    command: ['sh', '-c', 'until nslookup mydb; do echo waiting for mydb; sleep 2; done;']
```

① Specifies the **myservice** container.

② Specifies the **mydb** container.

Each init container has all of the fields of an app container except for **readinessProbe**. Init containers must exit for pod startup to continue and cannot define readiness other than completion.

Init containers can include **activeDeadlineSeconds** on the pod and **livenessProbe** on the container to prevent init containers from failing forever. The active deadline includes init containers.

3.3.3. Services

A Kubernetes service serves as an internal load balancer. It identifies a set of replicated pods in order to proxy the connections it receives to them. Backing pods can be added to or removed from a service arbitrarily while the service remains consistently available, enabling anything that depends on the service to refer to it at a consistent address. The default service clusterIP addresses are from the OpenShift Container Platform internal network and they are used to permit pods to access each other.

To permit external access to the service, additional **externalIP** and **ingressIP** addresses that are external to the cluster can be assigned to the service. These **externalIP** addresses can also be virtual IP addresses that provide highly available access to the service.

Services are assigned an IP address and port pair that, when accessed, proxy to an appropriate backing pod. A service uses a label selector to find all the containers running that provide a certain network service on a certain port.

Like pods, services are REST objects. The following example shows the definition of a service for the pod defined above:

Example 3.3. Service Object Definition (YAML)

```
apiVersion: v1
kind: Service
metadata:
  name: docker-registry      ❶
spec:
  selector:                  ❷
    docker-registry: default
  clusterIP: 172.30.136.123  ❸
  ports:
  - nodePort: 0
    port: 5000               ❹
    protocol: TCP
    targetPort: 5000         ❺
```

❶ The service name **docker-registry** is also used to construct an environment variable with the service IP that is inserted into other pods in the same namespace. The maximum name length is 63 characters.

❷ The label selector identifies all pods with the **docker-registry=default** label attached as its backing pods.

❸ Virtual IP of the service, allocated automatically at creation from a pool of internal IPs.

❹ Port the service listens on.

❺ Port on the backing pods to which the service forwards connections.

The Kubernetes documentation has more information on services.

3.3.3.1. Service externalIPs

In addition to the cluster's internal IP addresses, the user can configure IP addresses that are external to the cluster. The administrator is responsible for ensuring that traffic arrives at a node with this IP.

The externalIPs must be selected by the cluster adminitrators from the **externalIPNetworkCIDRs** range configured in ***master-config.yaml*** file. When ***master-config.yaml*** is changed, the master services must be restarted.

Example 3.4. Sample externalIPNetworkCIDR /etc/origin/master/master-config.yaml

```
networkConfig:
  externalIPNetworkCIDRs:
  - 192.0.1.0.0/24
```

Example 3.5. Service externalIPs Definition (JSON)

```
{
    "kind": "Service",
    "apiVersion": "v1",
    "metadata": {
        "name": "my-service"
    },
    "spec": {
        "selector": {
            "app": "MyApp"
        },
        "ports": [
            {
                "name": "http",
                "protocol": "TCP",
                "port": 80,
                "targetPort": 9376
            }
        ],
        "externalIPs" : [
            "192.0.1.1"         1
        ]
    }
}
```

1 List of external IP addresses on which the **port** is exposed. This list is in addition to the internal IP address list.

3.3.3.2. Service ingressIPs

In non-cloud clusters, externalIP addresses can be automatically assigned from a pool of addresses. This eliminates the need for the administrator manually assigning them.

The pool is configured in ***/etc/origin/master/master-config.yaml*** file. After changing this file, restart the master service.

The **ingressIPNetworkCIDR** is set to **172.29.0.0/16** by default. If the cluster environment is not already using this private range, use the default range or set a custom range.

NOTE

If you are using high availability, then this range must be less than 256 addresses.

Example 3.6. Sample ingressIPNetworkCIDR /etc/origin/master/master-config.yaml

```
networkConfig:
  ingressIPNetworkCIDR: 172.29.0.0/16
```

3.3.3.3. Service NodePort

Setting the service **type=NodePort** will allocate a port from a flag-configured range (default: 30000-32767), and each node will proxy that port (the same port number on every node) into your service.

The selected port will be reported in the service configuration, under **spec.ports[*].nodePort**.

To specify a custom port just place the port number in the nodePort field. The custom port number must be in the configured range for nodePorts. When '**master-config.yaml**' is changed the master services must be restarted.

Example 3.7. Sample servicesNodePortRange /etc/origin/master/master-config.yaml

```
kubernetesMasterConfig:
  servicesNodePortRange: ""
```

The service will be visible as both the **<NodeIP>:spec.ports[].nodePort** and **spec.clusterIp:spec.ports[].port**

NOTE

Setting a nodePort is a privileged operation.

3.3.3.4. Service Proxy Mode

OpenShift Container Platform has two different implementations of the service-routing infrastructure. The default implementation is entirely **iptables**-based, and uses probabilistic **iptables** rewriting rules to distribute incoming service connections between the endpoint pods. The older implementation uses a user space process to accept incoming connections and then proxy traffic between the client and one of the endpoint pods.

The **iptables**-based implementation is much more efficient, but it requires that all endpoints are always able to accept connections; the user space implementation is slower, but can try multiple endpoints in turn until it finds one that works. If you have good readiness checks (or generally reliable nodes and pods), then the **iptables**-based service proxy is the best choice. Otherwise, you can enable the user space-based proxy when installing, or after deploying the cluster by editing the node configuration file.

3.3.3.5. Headless services

If your application does not need load balancing or single-service IP addresses, you can create a headless service. When you create a headless service, no load-balancing or proxying is done and no cluster IP is allocated for this service. For such services, DNS is automatically configured depending on whether the service has selectors defined or not.

Services with selectors: For headless services that define selectors, the endpoints controller creates **Endpoints** records in the API and modifies the DNS configuration to return **A** records (addresses) that point directly to the pods backing the service.

Services without selectors: For headless services that do not define selectors, the endpoints controller does not create **Endpoints** records. However, the DNS system looks for and configures the following records:

- For **ExternalName** type services, **CNAME** records.
- For all other service types, **A** records for any endpoints that share a name with the service.

3.3.3.5.1. Creating a headless service

Creating a headless service is similar to creating a standard service, but you do not declare the **ClusterIP** address. To create a headless service, add the **clusterIP: None** parameter value to the service YAML definition.

For example, for a group of pods that you want to be a part of the same cluster or service.

List of pods

```
$ oc get pods -o wide
NAME               READY  STATUS    RESTARTS   AGE    IP            NODE
frontend-1-287hw   1/1    Running   0          7m     172.17.0.3    node_1
frontend-1-68km5   1/1    Running   0          7m     172.17.0.6    node_1
```

You can define the headless service as:

Headless service definition

```
apiVersion: v1
kind: Service
metadata:
  labels:
    app: ruby-helloworld-sample
    template: application-template-stibuild
  name: frontend-headless 1
spec:
  clusterIP: None 2
  ports:
  - name: web
    port: 5432
    protocol: TCP
    targetPort: 8080
  selector:
    name: frontend 3
  sessionAffinity: None
  type: ClusterIP
status:
  loadBalancer: {}
```

1 Name of the headless service.

2 Setting **clusterIP** variable to **None** declares a headless service.

3 Selects all pods that have **frontend** label.

Also, headless service does not have any IP address of its own.

```
$ oc get svc
NAME                TYPE        CLUSTER-IP      EXTERNAL-IP   PORT(S)    AGE
frontend            ClusterIP   172.30.232.77   <none>        5432/TCP   12m
frontend-headless   ClusterIP   None            <none>        5432/TCP   10m
```

3.3.3.5.2. Endpoint discovery by using a headless service

The benefit of using a headless service is that you can discover a pod's IP address directly. Standard services act as load balancer or proxy and give access to the workload object by using the service name. With headless services, the service name resolves to the set of IP addresses of the pods that are grouped by the service.

When you look up the DNS **A** record for a standard service, you get the loadbalanced IP of the service.

```
$ dig frontend.test A +search +short
172.30.232.77
```

But for a headless service, you get the list of IPs of individual pods.

```
$ dig frontend-headless.test A +search +short
172.17.0.3
172.17.0.6
```

NOTE

For using a headless service with a StatefulSet and related use cases where you need to resolve DNS for the pod during initialization and termination, set **publishNotReadyAddresses** to **true** (the default value is **false**). When **publishNotReadyAddresses** is set to **true**, it indicates that DNS implementations must publish the **notReadyAddresses** of subsets for the Endpoints associated with the Service.

3.3.4. Labels

Labels are used to organize, group, or select API objects. For example, pods are "tagged" with labels, and then services use label selectors to identify the pods they proxy to. This makes it possible for services to reference groups of pods, even treating pods with potentially different containers as related entities.

Most objects can include labels in their metadata. So labels can be used to group arbitrarily-related objects; for example, all of the pods, services, replication controllers, and deployment configurations of a particular application can be grouped.

Labels are simple key/value pairs, as in the following example:

```
labels:
  key1: value1
  key2: value2
```

Consider:

- A pod consisting of an **nginx** container, with the label **role=webserver**.

- A pod consisting of an **Apache httpd** container, with the same label **role=webserver**.

A service or replication controller that is defined to use pods with the **role=webserver** label treats both of these pods as part of the same group.

The Kubernetes documentation has more information on labels.

3.3.5. Endpoints

The servers that back a service are called its endpoints, and are specified by an object of type **Endpoints** with the same name as the service. When a service is backed by pods, those pods are normally specified by a label selector in the service specification, and OpenShift Container Platform automatically creates the Endpoints object pointing to those pods.

In some cases, you may want to create a service but have it be backed by external hosts rather than by pods in the OpenShift Container Platform cluster. In this case, you can leave out the **selector** field in the service, and create the Endpoints object manually .

Note that OpenShift Container Platform will not let most users manually create an Endpoints object that points to an IP address in the network blocks reserved for pod and service IPs. Only cluster admins or other users with permission to **create** resources under **endpoints/restricted** can create such Endpoint objects.

3.4. PROJECTS AND USERS

3.4.1. Users

Interaction with OpenShift Container Platform is associated with a user. An OpenShift Container Platform user object represents an actor which may be granted permissions in the system by adding roles to them or to their groups.

Several types of users can exist:

Regular users	This is the way most interactive OpenShift Container Platform users will be represented. Regular users are created automatically in the system upon first login, or can be created via the API. Regular users are represented with the **User** object. Examples: **joe alice**
System users	Many of these are created automatically when the infrastructure is defined, mainly for the purpose of enabling the infrastructure to interact with the API securely. They include a cluster administrator (with access to everything), a per-node user, users for use by routers and registries, and various others. Finally, there is an **anonymous** system user that is used by default for unauthenticated requests. Examples: **system:admin system:openshift-registry system:node:node1.example.com**
Service accounts	These are special system users associated with projects; some are created automatically when the project is first created, while project administrators can create more for the purpose of defining access to the contents of each project. Service accounts are represented with the **ServiceAccount** object. Examples: **system:serviceaccount:default:deployer system:serviceaccount:foo:builder**

Every user must authenticate in some way in order to access OpenShift Container Platform. API requests with no authentication or invalid authentication are authenticated as requests by the **anonymous** system user. Once authenticated, policy determines what the user is authorized to do.

3.4.2. Namespaces

A Kubernetes namespace provides a mechanism to scope resources in a cluster. In OpenShift Container Platform, a project is a Kubernetes namespace with additional annotations.

Namespaces provide a unique scope for:

- Named resources to avoid basic naming collisions.
- Delegated management authority to trusted users.
- The ability to limit community resource consumption.

Most objects in the system are scoped by namespace, but some are excepted and have no namespace, including nodes and users.

The Kubernetes documentation has more information on namespaces.

3.4.3. Projects

A project is a Kubernetes namespace with additional annotations, and is the central vehicle by which access to resources for regular users is managed. A project allows a community of users to organize and manage their content in isolation from other communities. Users must be given access to projects by administrators, or if allowed to create projects, automatically have access to their own projects.

Projects can have a separate **name**, **displayName**, and **description**.

- The mandatory **name** is a unique identifier for the project and is most visible when using the CLI tools or API. The maximum name length is 63 characters.
- The optional **displayName** is how the project is displayed in the web console (defaults to **name**).
- The optional **description** can be a more detailed description of the project and is also visible in the web console.

Each project scopes its own set of:

Objects	Pods, services, replication controllers, etc.
Policies	Rules for which users can or cannot perform actions on objects.
Constraints	Quotas for each kind of object that can be limited.
Service accounts	Service accounts act automatically with designated access to objects in the project.

Cluster administrators can create projects and delegate administrative rights for the project to any member of the user community. Cluster administrators can also allow developers to create their own projects.

Developers and administrators can interact with projects using the CLI or the web console.

3.4.3.1. Projects provided at installation

OpenShift Container Platform comes with a number of projects out of the box, and **openshift** is the most essential to users:

openshift A user-facing project, mainly for housing objects for day-to-day tasks. These include any application objects for access by multiple projects, such as templates and images. These objects should be those that do not require communication between the pods.

3.5. BUILDS AND IMAGE STREAMS

3.5.1. Builds

A *build* is the process of transforming input parameters into a resulting object. Most often, the process is used to transform input parameters or source code into a runnable image. A **BuildConfig** object is the definition of the entire build process.

OpenShift Container Platform leverages Kubernetes by creating Docker-formatted containers from build images and pushing them to a container registry.

Build objects share common characteristics: inputs for a build, the need to complete a build process, logging the build process, publishing resources from successful builds, and publishing the final status of the build. Builds take advantage of resource restrictions, specifying limitations on resources such as CPU usage, memory usage, and build or pod execution time.

The OpenShift Container Platform build system provides extensible support for *build strategies* that are based on selectable types specified in the build API. There are three primary build strategies available:

- Docker build
- Source-to-Image (S2I) build
- Custom build

By default, Docker builds and S2I builds are supported.

The resulting object of a build depends on the builder used to create it. For Docker and S2I builds, the resulting objects are runnable images. For Custom builds, the resulting objects are whatever the builder image author has specified.

Additionally, the Pipeline build strategy can be used to implement sophisticated workflows:

- continuous integration
- continuous deployment

For a list of build commands, see the Developer's Guide.

For more information on how OpenShift Container Platform leverages Docker for builds, see the upstream documentation.

3.5.1.1. Docker Build

The Docker build strategy invokes the docker build command, and it therefore expects a repository with a ***Dockerfile*** and all required artifacts in it to produce a runnable image.

3.5.1.2. Source-to-Image (S2I) Build

Source-to-Image (S2I) is a tool for building reproducible, Docker-formatted container images. It produces ready-to-run images by injecting application source into a container image and assembling a new image. The new image incorporates the base image (the builder) and built source and is ready to use with the **docker run** command. S2I supports incremental builds, which re-use previously downloaded dependencies, previously built artifacts, etc.

The advantages of S2I include the following:

Image flexibility	S2I scripts can be written to inject application code into almost any existing Docker-formatted container image, taking advantage of the existing ecosystem. Note that, currently, S2I relies on **tar** to inject application source, so the image needs to be able to process tarred content.
Speed	With S2I, the assemble process can perform a large number of complex operations without creating a new layer at each step, resulting in a fast process. In addition, S2I scripts can be written to re-use artifacts stored in a previous version of the application image, rather than having to download or build them each time the build is run.
Patchability	S2I allows you to rebuild the application consistently if an underlying image needs a patch due to a security issue.
Operational efficiency	By restricting build operations instead of allowing arbitrary actions, as a ***Dockerfile*** would allow, the PaaS operator can avoid accidental or intentional abuses of the build system.
Operational security	Building an arbitrary ***Dockerfile*** exposes the host system to root privilege escalation. This can be exploited by a malicious user because the entire Docker build process is run as a user with Docker privileges. S2I restricts the operations performed as a root user and can run the scripts as a non-root user.
User efficiency	S2I prevents developers from performing arbitrary **yum install** type operations, which could slow down development iteration, during their application build.
Ecosystem	S2I encourages a shared ecosystem of images where you can leverage best practices for your applications.
Reproducibility	Produced images can include all inputs including specific versions of build tools and dependencies. This ensures that the image can be reproduced precisely.

3.5.1.3. Custom Build

The Custom build strategy allows developers to define a specific builder image responsible for the entire build process. Using your own builder image allows you to customize your build process.

A Custom builder image is a plain Docker-formatted container image embedded with build process logic, for example for building RPMs or base images. The **openshift/origin-custom-docker-builder** image is available on the Docker Hub registry as an example implementation of a Custom builder image.

3.5.1.4. Pipeline Build

The Pipeline build strategy allows developers to define a *Jenkins pipeline* for execution by the Jenkins pipeline plugin. The build can be started, monitored, and managed by OpenShift Container Platform in the same way as any other build type.

Pipeline workflows are defined in a Jenkinsfile, either embedded directly in the build configuration, or supplied in a Git repository and referenced by the build configuration.

The first time a project defines a build configuration using a Pipeline strategy, OpenShift Container Platform instantiates a Jenkins server to execute the pipeline. Subsequent Pipeline build configurations in the project share this Jenkins server.

For more details on how the Jenkins server is deployed and how to configure or disable the autoprovisioning behavior, see Configuring Pipeline Execution.

NOTE

The Jenkins server is not automatically removed, even if all Pipeline build configurations are deleted. It must be manually deleted by the user.

For more information about Jenkins Pipelines, see the Jenkins documentation.

3.5.2. Image Streams

An image stream and its associated tags provide an abstraction for referencing Docker images from within OpenShift Container Platform. The image stream and its tags allow you to see what images are available and ensure that you are using the specific image you need even if the image in the repository changes.

Image streams do not contain actual image data, but present a single virtual view of related images, similar to an image repository.

You can configure Builds and Deployments to watch an image stream for notifications when new images are added and react by performing a Build or Deployment, respectively.

For example, if a Deployment is using a certain image and a new version of that image is created, a Deployment could be automatically performed to pick up the new version of the image.

However, if the image stream tag used by the Deployment or Build is not updated, then even if the Docker image in the Docker registry is updated, the Build or Deployment will continue using the previous (presumably known good) image.

The source images can be stored in any of the following:

- OpenShift Container Platform's integrated registry
- An external registry, for example **registry.access.redhat.com** or **hub.docker.com**
- Other image streams in the OpenShift Container Platform cluster

When you define an object that references an image stream tag (such as a Build or Deployment configuration), you point to an image stream tag, not the Docker repository. When you Build or Deploy your application, OpenShift Container Platform queries the Docker repository using the image stream tag to locate the associated ID of the image and uses that exact image.

The image stream metadata is stored in the etcd instance along with other cluster information.

The following image stream contains two tags: **34** which points to a Python v3.4 image and **35** which points to a Python v3.5 image:

```
oc describe is python
Name:   python
Namespace:  imagestream
Created:  25 hours ago
Labels:   app=python
Annotations:  openshift.io/generated-by=OpenShiftWebConsole
  openshift.io/image.dockerRepositoryCheck=2017-10-03T19:48:00Z
Docker Pull Spec: docker-registry.default.svc:5000/imagestream/python
```

```
Image Lookup:  local=false
Unique Images:  2
Tags:   2

34
  tagged from centos/python-34-centos7

  * centos/python-34-
centos7@sha256:28178e2352d31f240de1af1370be855db33ae9782de737bb005247d8791a54d0
      14 seconds ago

35
  tagged from centos/python-35-centos7

  * centos/python-35-
centos7@sha256:2efb79ca3ac9c9145a63675fb0c09220ab3b8d4005d35e0644417ee552548b10
      7 seconds ago
```

Using image streams has several significant benefits:

- You can tag, rollback a tag, and quickly deal with images, without having to re-push using the command line.
- You can trigger Builds and Deployments when a new image is pushed to the registry. Also, OpenShift Container Platform has generic triggers for other resources (such as Kubernetes objects).
- You can mark a tag for periodic re-import . If the source image has changed, that change is picked up and reflected in the image stream, which triggers the Build and/or Deployment flow, depending upon the Build or Deployment configuration.
- You can share images using fine-grained access control and quickly distribute images across your teams.
- If the source image changes, the image stream tag will still point to a known-good version of the image, ensuring that your application will not break unexpectedly.
- You can configure security around who can view and use the images through permissions on the image stream objects.
- Users that lack permission to read or list images on the cluster level can still retrieve the images tagged in a project using image streams.

For a curated set of image streams, see the OpenShift Image Streams and Templates library.

When using image streams, it is important to understand what the image stream tag is pointing to and how changes to tags and images can affect you. For example:

- If your image stream tag points to a Docker image tag, you need to understand how that Docker image tag is updated. For example, a Docker image tag **docker.io/ruby:2.4** will probably always point to a v2.4 ruby image. But, a Docker image tag **docker.io/ruby:latest** will probably change with major versions. So, the Docker image tag that a image stream tag points to can tell you how stable the image stream tag will be, if you choose to reference it.
- If your image stream tag follows another image stream tag (it does not point directly to a docker image tag), it is possible that the image stream tag will be updated to follow a different image stream tag in the future. Again, this could result in picking up an incompatible version

change.

3.5.2.1. Important terms

Docker repository

A collection of related docker images and tags identifying them. For example, the OpenShift Jenkins images are in a Docker repository:

```
docker.io/openshift/jenkins-2-centos7
```

Docker registry

A content server that can store and service images from Docker repositories. For example:

```
registry.access.redhat.com
```

Docker image

A specific set of content that can be run as a container. Usually associated with a particular tag within a Docker repository.

Docker image tag

A label applied to a Docker image in a repository that distinguishes a specific image. For example, here **3.6.0** is a tag:

```
docker.io/openshift/jenkins-2-centos7:3.6.0
```

> **NOTE**
>
> A Docker image tag can be updated to point to new Docker image content at any time.

Docker image ID

A SHA (Secure Hash Algorithm) code that can be used to pull an image. For example:

```
docker.io/openshift/jenkins-2-centos7@sha256:ab312bda324
```

> **NOTE**
>
> A SHA image ID cannot change. A specific SHA identifier always references the exact same docker image content.

Image stream

An OpenShift Container Platform object that contains pointers to any number of Docker-formatted container images identified by tags. You can think of an image stream as equivalent to a Docker repository.

Image stream tag

A named pointer to an image in an image stream. An image stream tag is similar to a Docker image tag. See Image Stream Tag below.

Image stream image

An image that allows you to retrieve a specific Docker image from a particular image stream where it is tagged. An image stream image is an API resource object that pulls together some metadata about a particular image SHA identifier. See Image Stream Images below.

Image stream trigger

A trigger that causes a specific action when an image stream tag changes. For example, importing can cause the value of the tag to change, which causes a trigger to fire when there are Deployments, Builds, or other resources listening for those. See Image Stream Triggers below.

3.5.2.2. Configuring Image Streams

An image stream object file contains the following elements.

NOTE

See the Developer Guide for details on managing images and image streams.

Image Stream Object Definition

```
apiVersion: v1
kind: ImageStream
metadata:
  annotations:
    openshift.io/generated-by: OpenShiftNewApp
  creationTimestamp: 2017-09-29T13:33:49Z
  generation: 1
  labels:
    app: ruby-sample-build
    template: application-template-stibuild
  name: origin-ruby-sample ①
  namespace: test
  resourceVersion: "633"
  selflink: /oapi/v1/namespaces/test/imagestreams/origin-ruby-sample
  uid: ee2b9405-c68c-11e5-8a99-525400f25e34
spec: {}
status:
  dockerImageRepository: 172.30.56.218:5000/test/origin-ruby-sample ②
  tags:
  - items:
    - created: 2017-09-02T10:15:09Z
      dockerImageReference: 172.30.56.218:5000/test/origin-ruby-
sample@sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d ③
      generation: 2
      image: sha256:909de62d1f609a717ec433cc25ca5cf00941545c83a01fb31527771e1fab3fc5 ④
    - created: 2017-09-29T13:40:11Z
      dockerImageReference: 172.30.56.218:5000/test/origin-ruby-
sample@sha256:909de62d1f609a717ec433cc25ca5cf00941545c83a01fb31527771e1fab3fc5
      generation: 1
      image: sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d
    tag: latest ⑤
```

① The name of the image stream.

② Docker repository path where new images can be pushed to add/update them in this image stream.

3. The SHA identifier that this image stream tag currently references. Resources that reference this image stream tag use this identifier.

4. The SHA identifier that this image stream tag previously referenced. Can be used to rollback to an older image.

5. The image stream tag name.

For a sample build configuration that references an image stream, see What Is a BuildConfig? in the **Strategy** stanza of the configuration.

For a sample deployment configuration that references an image stream, see Creating a Deployment Configuration in the **Strategy** stanza of the configuration.

3.5.2.3. Image Stream Images

An *image stream image* points from within an image stream to a particular image ID.

Image stream images allow you to retrieve metadata about an image from a particular image stream where it is tagged.

Image stream image objects are automatically created in OpenShift Container Platform whenever you import or tag an image into the image stream. You should never have to explicitly define an image stream image object in any image stream definition that you use to create image streams.

The image stream image consists of the image stream name and image ID from the repository, delimited by an @ sign:

```
<image-stream-name>@<image-id>
```

To refer to the image in the image stream object example above, the image stream image looks like:

```
origin-ruby-
sample@sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d
```

3.5.2.4. Image Stream Tags

An *image stream tag* is a named pointer to an image in an *image stream*. It is often abbreviated as *istag*. An image stream tag is used to reference or retrieve an image for a given image stream and tag.

Image stream tags can reference any local or externally managed image. It contains a history of images represented as a stack of all images the tag ever pointed to. Whenever a new or existing image is tagged under particular image stream tag, it is placed at the first position in the history stack. The image previously occupying the top position will be available at the second position, and so forth. This allows for easy rollbacks to make tags point to historical images again.

The following image stream tag is from the image stream object example above:

Image Stream Tag with Two Images in its History

```
tags:
- items:
  - created: 2017-09-02T10:15:09Z
    dockerImageReference: 172.30.56.218:5000/test/origin-ruby-
```

```
sample@sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d
    generation: 2
    image: sha256:909de62d1f609a717ec433cc25ca5cf00941545c83a01fb31527771e1fab3fc5
  - created: 2017-09-29T13:40:11Z
    dockerImageReference: 172.30.56.218:5000/test/origin-ruby-
sample@sha256:909de62d1f609a717ec433cc25ca5cf00941545c83a01fb31527771e1fab3fc5
    generation: 1
    image: sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d
  tag: latest
```

Image stream tags can be *permanent* tags or *tracking* tags.

- *Permanent tags* are version-specific tags that point to a particular version of an image, such as Python 3.5.

- *Tracking tags* are reference tags that follow another image stream tag and could be updated in the future to change which image they follow, much like a symlink. Note that these new levels are not guaranteed to be backwards-compatible.
  For example, the **latest** image stream tags that ship with OpenShift Container Platform are tracking tags. This means consumers of the **latest** image stream tag will be updated to the newest level of the framework provided by the image when a new level becomes available. A **latest** image stream tag to **v3.6** could be changed to **v3.7** at any time. It is important to be aware that these **latest** image stream tags behave differently than the Docker **latest** tag. The **latest** image stream tag, in this case, does not point to the latest image in the Docker repository. It points to another image stream tag, which might not be the latest version of an image. For example, if the **latest** image stream tag points to **v3.2** of an image, when the **3.3** version is released, the **latest** tag is not automatically updated to **v3.3**, and remains at **v3.2** until it is manually updated to point to a **v3.3** image stream tag.

NOTE

Tracking tags are limited to a single image stream and cannot reference other image streams.

You can create your own image stream tags for your own needs. See the Recommended Tagging Conventions.

The image stream tag is composed of the name of the image stream and a tag, separated by a colon:

```
<image stream name>:<tag>
```

For example, to refer to the **sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d** image in the image stream object example above , the image stream tag would be:

```
origin-ruby-sample:latest
```

3.5.2.5. Image Stream Change Triggers

Image stream triggers allow your Builds and Deployments to be automatically invoked when a new version of an upstream image is available.

For example, Builds and Deployments can be automatically started when an image stream tag is modified. This is achieved by monitoring that particular image stream tag and notifying the Build or Deployment when a change is detected.

The **ImageChange** trigger results in a new replication controller whenever the content of an image stream tag changes (when a new version of the image is pushed).

Example 3.8. An ImageChange Trigger

```
triggers:
  - type: "ImageChange"
    imageChangeParams:
      automatic: true 1
      from:
        kind: "ImageStreamTag"
        name: "origin-ruby-sample:latest"
        namespace: "myproject"
      containerNames:
        - "helloworld"
```

1 If the **imageChangeParams.automatic** field is set to **false**, the trigger is disabled.

With the above example, when the **latest** tag value of the **origin-ruby-sample** image stream changes and the new image value differs from the current image specified in the deployment configuration's **helloworld** container, a new replication controller is created using the new image for the **helloworld** container.

NOTE

If an **ImageChange** trigger is defined on a deployment configuration (with a **ConfigChange** trigger and **automatic=false**, or with **automatic=true**) and the **ImageStreamTag** pointed by the **ImageChange** trigger does not exist yet, then the initial deployment process will automatically start as soon as an image is imported or pushed by a build to the **ImageStreamTag**.

3.5.2.6. Image Stream Mappings

When the integrated registry receives a new image, it creates and sends an image stream mapping to OpenShift Container Platform, providing the image's project, name, tag, and image metadata.

NOTE

Configuring image stream mappings is an advanced feature.

This information is used to create a new image (if it does not already exist) and to tag the image into the image stream. OpenShift Container Platform stores complete metadata about each image, such as commands, entry point, and environment variables. Images in OpenShift Container Platform are immutable and the maximum name length is 63 characters.

NOTE

See the Developer Guide for details on manually tagging images.

The following image stream mapping example results in an image being tagged as **test/origin-ruby-sample:latest**:

Image Stream Mapping Object Definition

```
apiVersion: v1
kind: ImageStreamMapping
metadata:
  creationTimestamp: null
  name: origin-ruby-sample
  namespace: test
tag: latest
image:
  dockerImageLayers:
  - name: sha256:5f70bf18a086007016e948b04aed3b82103a36bea41755b6cddfaf10ace3c6ef
    size: 0
  - name: sha256:ee1dd2cb6df21971f4af6de0f1d7782b81fb63156801cfde2bb47b4247c23c29
    size: 196634330
  - name: sha256:5f70bf18a086007016e948b04aed3b82103a36bea41755b6cddfaf10ace3c6ef
    size: 0
  - name: sha256:5f70bf18a086007016e948b04aed3b82103a36bea41755b6cddfaf10ace3c6ef
    size: 0
  - name: sha256:ca062656bff07f18bff46be00f40cfbb069687ec124ac0aa038fd676cfaea092
    size: 177723024
  - name: sha256:63d529c59c92843c395befd065de516ee9ed4995549f8218eac6ff088bfa6b6e
    size: 55679776
  - name: sha256:92114219a04977b5563d7dff71ec4caa3a37a15b266ce42ee8f43dba9798c966
    size: 11939149
  dockerImageMetadata:
    Architecture: amd64
    Config:
      Cmd:
      - /usr/libexec/s2i/run
      Entrypoint:
      - container-entrypoint
      Env:
      - RACK_ENV=production
      - OPENSHIFT_BUILD_NAMESPACE=test
      - OPENSHIFT_BUILD_SOURCE=https://github.com/openshift/ruby-hello-world.git
      - EXAMPLE=sample-app
      - OPENSHIFT_BUILD_NAME=ruby-sample-build-1
      - PATH=/opt/app-root/src/bin:/opt/app-
root/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin
      - STI_SCRIPTS_URL=image:///usr/libexec/s2i
      - STI_SCRIPTS_PATH=/usr/libexec/s2i
      - HOME=/opt/app-root/src
      - BASH_ENV=/opt/app-root/etc/scl_enable
      - ENV=/opt/app-root/etc/scl_enable
      - PROMPT_COMMAND=. /opt/app-root/etc/scl_enable
      - RUBY_VERSION=2.2
      ExposedPorts:
        8080/tcp: {}
      Labels:
        build-date: 2015-12-23
        io.k8s.description: Platform for building and running Ruby 2.2 applications
        io.k8s.display-name: 172.30.56.218:5000/test/origin-ruby-sample:latest
        io.openshift.build.commit.author: Ben Parees <bparees@users.noreply.github.com>
        io.openshift.build.commit.date: Wed Jan 20 10:14:27 2016 -0500
        io.openshift.build.commit.id: 00cadc392d39d5ef9117cbc8a31db0889eedd442
        io.openshift.build.commit.message: 'Merge pull request #51 from php-coder/fix_url_and_sti'
        io.openshift.build.commit.ref: master
        io.openshift.build.image: centos/ruby-22-
centos7@sha256:3a335d7d8a452970c5b4054ad7118ff134b3a6b50a2bb6d0c07c746e8986b28e
        io.openshift.build.source-location: https://github.com/openshift/ruby-hello-world.git
        io.openshift.builder-base-version: 8d95148
        io.openshift.builder-version: 8847438ba06307f86ac877465eadc835201241df
        io.openshift.s2i.scripts-url: image:///usr/libexec/s2i
        io.openshift.tags: builder,ruby,ruby22
        io.s2i.scripts-url: image:///usr/libexec/s2i
        license: GPLv2
        name: CentOS Base Image
        vendor: CentOS
      User: "1001"
      WorkingDir: /opt/app-root/src
    Container: 86e9a4a3c760271671ab913616c51c9f3cea846ca524bf07c04a6f6c9e103a76
    ContainerConfig:
      AttachStdout: true
      Cmd:
      - /bin/sh
      - -c
      - tar -C /tmp -xf - && /usr/libexec/s2i/assemble
      Entrypoint:
      - container-entrypoint
      Env:
      - RACK_ENV=production
      - OPENSHIFT_BUILD_NAME=ruby-sample-build-1
      - OPENSHIFT_BUILD_NAMESPACE=test
      - OPENSHIFT_BUILD_SOURCE=https://github.com/openshift/ruby-hello-world.git
      - EXAMPLE=sample-app
      - PATH=/opt/app-root/src/bin:/opt/app-
root/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin
      - STI_SCRIPTS_URL=image:///usr/libexec/s2i
      - STI_SCRIPTS_PATH=/usr/libexec/s2i
      - HOME=/opt/app-root/src
      - BASH_ENV=/opt/app-root/etc/scl_enable
      - ENV=/opt/app-root/etc/scl_enable
      - PROMPT_COMMAND=. /opt/app-root/etc/scl_enable
      - RUBY_VERSION=2.2
      ExposedPorts:
        8080/tcp: {}
      Hostname: ruby-sample-build-1-build
      Image: centos/ruby-22-
centos7@sha256:3a335d7d8a452970c5b4054ad7118ff134b3a6b50a2bb6d0c07c746e8986b28e
      OpenStdin: true
      StdinOnce: true
      User: "1001"
      WorkingDir: /opt/app-root/src
    Created: 2016-01-29T13:40:00Z
    DockerVersion: 1.8.2.fc21
    Id: 9d7fd5e2d15495802028c569d544329f4286dcd1c9c085ff5699218dbaa69b43
    Parent: 57b08d979c86f4500dc8cad639c9518744c8dd39447c055a3517dc9c18d6fccd
    Size: 441976279
    apiVersion: "1.0"
    kind: DockerImage
  dockerImageMetadataVersion: "1.0"
  dockerImageReference: 172.30.56.218:5000/test/origin-ruby-
sample@sha256:47463d94eb5c049b2d23b03a9530bf944f8f967a0fe79147dd6b9135bf7dd13d
```

3.5.2.7. Working with Image Streams

The following sections describe how to use image streams and image stream tags. For more information on working with image streams, see Managing Images.

3.5.2.7.1. Getting Information about Image Streams

To get general information about the image stream and detailed information about all the tags it is pointing to, use the following command:

```
oc describe is/<image-name>
```

For example:

```
oc describe is/python

Name:   python
Namespace:  default
Created:  About a minute ago
Labels:   <none>
Annotations:  openshift.io/image.dockerRepositoryCheck=2017-10-02T17:05:11Z
Docker Pull Spec: docker-registry.default.svc:5000/default/python
Image Lookup:  local=false
Unique Images:  1
Tags:   1

3.5
  tagged from centos/python-35-centos7

  * centos/python-35-centos7@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
      About a minute ago
```

To get all the information available about particular image stream tag:

```
oc describe istag/<image-stream>:<tag-name>
```

For example:

```
oc describe istag/python:latest

Image Name: sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
Docker Image: centos/python-35-
centos7@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
Name:  sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
Created: 2 minutes ago
Image Size: 251.2 MB (first layer 2.898 MB, last binary layer 72.26 MB)
Image Created: 2 weeks ago
Author:  <none>
```

```
Arch:  amd64
Entrypoint: container-entrypoint
Command: /bin/sh -c $STI_SCRIPTS_PATH/usage
Working Dir: /opt/app-root/src
User:  1001
Exposes Ports: 8080/tcp
Docker Labels: build-date=20170801
```

NOTE

More information is output than shown.

3.5.2.7.2. Adding Additional Tags to an Image Stream

To add a tag that points to one of the existing tags, you can use the **oc tag** command:

```
oc tag <image-name:tag> <image-name:tag>
```

For example:

```
oc tag python:3.5 python:latest

Tag python:latest set to
python@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25.
```

Use the **oc describe** command to confirm the image stream has two tags, one (**3.5**) pointing at the external Docker image and another tag (**latest**) pointing to the same image because it was created based on the first tag.

```
oc describe is/python

Name:   python
Namespace:  default
Created:  5 minutes ago
Labels:   <none>
Annotations:  openshift.io/image.dockerRepositoryCheck=2017-10-02T17:05:11Z
Docker Pull Spec: docker-registry.default.svc:5000/default/python
Image Lookup:  local=false
Unique Images:  1
Tags:   2

latest
  tagged from python@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25

  * centos/python-35-centos7@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
      About a minute ago

3.5
  tagged from centos/python-35-centos7

  * centos/python-35-centos7@sha256:49c18358df82f4577386404991c51a9559f243e0b1bdc366df25
      5 minutes ago
```

3.5.2.7.3. Adding Tags for an External Image

Use the **oc tag** command for all tag-related operations, such as adding tags pointing to internal or external images:

```
oc tag <repositiory/image> <image-name:tag>
```

For example, this command maps the **docker.io/python:3.6.0** image to the **3.6** tag in the **python** image stream.

```
oc tag docker.io/python:3.6.0 python:3.6
Tag python:3.6 set to docker.io/python:3.6.0.
```

If the external image is secured, you will need to create a secret with credentials for accessing that registry. See Importing Images from Private Registries for more details.

3.5.2.7.4. Updating an Image Stream Tag

To update a tag to reflect another tag in an image stream:

```
oc tag <image-name:tag> <image-name:latest>
```

For example, the following updates the **latest** tag to reflect the **3.6** tag in an image stream:

```
oc tag python:3.6 python:latest
Tag python:latest set to
python@sha256:438208801c4806548460b27bd1fbcb7bb188273d13871ab43f.
```

3.5.2.7.5. Removing Image Stream Tags from an Image Stream

To remove old tags from an image stream:

```
oc tag -d <image-name:tag>
```

For example:

```
oc tag -d python:3.5

Deleted tag default/python:3.5.
```

3.5.2.7.6. Configuring Periodic Importing of Tags

When working with an external Docker registry, to periodically re-import an image (such as, to get latest security updates), use the **--scheduled** flag:

```
oc tag <repositiory/image> <image-name:tag> --scheduled
```

For example:

```
oc tag docker.io/python:3.6.0 python:3.6 --scheduled

Tag python:3.6 set to import docker.io/python:3.6.0 periodically.
```

This command causes OpenShift Container Platform to periodically update this particular image stream tag. This period is a cluster-wide setting set to 15 minutes by default.

To remove the periodic check, re-run above command but omit the **--scheduled** flag. This will reset its behavior to default.

```
oc tag <repositiory/image> <image-name:tag>
```

3.6. DEPLOYMENTS

3.6.1. Replication controllers

A replication controller ensures that a specified number of replicas of a pod are running at all times. If pods exit or are deleted, the replication controller acts to instantiate more up to the defined number. Likewise, if there are more running than desired, it deletes as many as necessary to match the defined amount.

A replication controller configuration consists of:

1. The number of replicas desired (which can be adjusted at runtime).
2. A pod definition to use when creating a replicated pod.
3. A selector for identifying managed pods.

A selector is a set of labels assigned to the pods that are managed by the replication controller. These labels are included in the pod definition that the replication controller instantiates. The replication controller uses the selector to determine how many instances of the pod are already running in order to adjust as needed.

The replication controller does not perform auto-scaling based on load or traffic, as it does not track either. Rather, this would require its replica count to be adjusted by an external auto-scaler.

A replication controller is a core Kubernetes object called **ReplicationController**.

The following is an example **ReplicationController** definition:

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: frontend-1
spec:
  replicas: 1 ①
  selector: ②
    name: frontend
  template: ③
    metadata:
      labels: ④
        name: frontend ⑤
    spec:
      containers:
      - image: openshift/hello-openshift
        name: helloworld
        ports:
```

```
        - containerPort: 8080
          protocol: TCP
      restartPolicy: Always
```

1. The number of copies of the pod to run.
2. The label selector of the pod to run.
3. A template for the pod the controller creates.
4. Labels on the pod should include those from the label selector.
5. The maximum name length after expanding any parameters is 63 characters.

3.6.2. Replica set

Similar to a replication controller, a replica set ensures that a specified number of pod replicas are running at any given time. The difference between a replica set and a replication controller is that a replica set supports set-based selector requirements whereas a replication controller only supports equality-based selector requirements.

> **NOTE**
>
> Only use replica sets if you require custom update orchestration or do not require updates at all, otherwise, use Deployments. Replica sets can be used independently, but are used by deployments to orchestrate pod creation, deletion, and updates. Deployments manage their replica sets automatically, provide declarative updates to pods, and do not have to manually manage the replica sets that they create.

A replica set is a core Kubernetes object called **ReplicaSet**.

The following is an example **ReplicaSet** definition:

```
apiVersion: apps/v1
kind: ReplicaSet
metadata:
  name: frontend-1
  labels:
    tier: frontend
spec:
  replicas: 3
  selector: 1
    matchLabels: 2
      tier: frontend
    matchExpressions: 3
      - {key: tier, operator: In, values: [frontend]}
  template:
    metadata:
      labels:
        tier: frontend
    spec:
      containers:
      - image: openshift/hello-openshift
        name: helloworld
        ports:
        - containerPort: 8080
          protocol: TCP
      restartPolicy: Always
```

1 A label query over a set of resources. The result of **matchLabels** and **matchExpressions** are logically conjoined.

2 Equality-based selector to specify resources with labels that match the selector.

3 Set-based selector to filter keys. This selects all resources with key equal to **tier** and value equal to **frontend**.

3.6.3. Jobs

A job is similar to a replication controller, in that its purpose is to create pods for specified reasons. The difference is that replication controllers are designed for pods that will be continuously running, whereas jobs are for one-time pods. A job tracks any successful completions and when the specified amount of completions have been reached, the job itself is completed.

The following example computes π to 2000 places, prints it out, then completes:

```
apiVersion: extensions/v1
kind: Job
metadata:
  name: pi
spec:
  selector:
    matchLabels:
      app: pi
  template:
    metadata:
      name: pi
      labels:
        app: pi
    spec:
      containers:
      - name: pi
        image: perl
        command: ["perl",  "-Mbignum=bpi", "-wle", "print bpi(2000)"]
      restartPolicy: Never
```

See the Jobs topic for more information on how to use jobs.

3.6.4. Deployments and Deployment Configurations

Building on replication controllers, OpenShift Container Platform adds expanded support for the software development and deployment lifecycle with the concept of deployments. In the simplest case, a deployment just creates a new replication controller and lets it start up pods. However, OpenShift Container Platform deployments also provide the ability to transition from an existing deployment of an image to a new one and also define hooks to be run before or after creating the replication controller.

The OpenShift Container Platform **DeploymentConfig** object defines the following details of a deployment:

1. The elements of a **ReplicationController** definition.
2. Triggers for creating a new deployment automatically.
3. The strategy for transitioning between deployments.
4. Life cycle hooks.

Each time a deployment is triggered, whether manually or automatically, a deployer pod manages the deployment (including scaling down the old replication controller, scaling up the new one, and running hooks). The deployment pod remains for an indefinite amount of time after it completes the deployment in order to retain its logs of the deployment. When a deployment is superseded by another, the previous replication controller is retained to enable easy rollback if needed.

For detailed instructions on how to create and interact with deployments, refer to Deployments.

Here is an example **DeploymentConfig** definition with some omissions and callouts:

```
apiVersion: v1
kind: DeploymentConfig
metadata:
  name: frontend
spec:
  replicas: 5
  selector:
    name: frontend
  template: { ... }
  triggers:
  - type: ConfigChange ❶
  - imageChangeParams:
      automatic: true
      containerNames:
      - helloworld
      from:
        kind: ImageStreamTag
        name: hello-openshift:latest
    type: ImageChange  ❷
  strategy:
    type: Rolling      ❸
```

❶ A **ConfigChange** trigger causes a new deployment to be created any time the replication controller template changes.

❷ An **ImageChange** trigger causes a new deployment to be created each time a new version of the backing image is available in the named image stream.

❸ The default **Rolling** strategy makes a downtime-free transition between deployments.

3.7. TEMPLATES

3.7.1. Overview

A template describes a set of objects that can be parameterized and processed to produce a list of objects for creation by OpenShift Container Platform. The objects to create can include anything that users have permission to create within a project, for example services, build configurations, and deployment configurations. A template may also define a set of labels to apply to every object defined in the template.

See the template guide for details about creating and using templates.

CHAPTER 4. ADDITIONAL CONCEPTS

4.1. AUTHENTICATION

4.1.1. Overview

The authentication layer identifies the user associated with requests to the OpenShift Container Platform API. The authorization layer then uses information about the requesting user to determine if the request should be allowed.

As an administrator, you can configure authentication using a master configuration file.

4.1.2. Users and Groups

A *user* in OpenShift Container Platform is an entity that can make requests to the OpenShift Container Platform API. Typically, this represents the account of a developer or administrator that is interacting with OpenShift Container Platform.

A user can be assigned to one or more *groups*, each of which represent a certain set of users. Groups are useful when managing authorization policies to grant permissions to multiple users at once, for example allowing access to objects within a project, versus granting them to users individually.

In addition to explicitly defined groups, there are also system groups, or *virtual groups*, that are automatically provisioned by OpenShift. These can be seen when viewing cluster bindings.

In the default set of virtual groups, note the following in particular:

Virtual Group	Description
system:authenticated	Automatically associated with all authenticated users.
system:authenticated:oauth	Automatically associated with all users authenticated with an OAuth access token.
system:unauthenticated	Automatically associated with all unauthenticated users.

4.1.3. API Authentication

Requests to the OpenShift Container Platform API are authenticated using the following methods:

OAuth Access Tokens

- Obtained from the OpenShift Container Platform OAuth server using the ***<master>*/oauth/authorize** and ***<master>*/oauth/token** endpoints.
- Sent as an **Authorization: Bearer…** header
- Sent as an **access_token=…** query parameter for websocket requests prior to OpenShift Container Platform server version 3.6.

- Sent as a websocket subprotocol header in the form **base64url.bearer.authorization.k8s.io.<base64url-encoded-token>** for websocket requests in OpenShift Container Platform server version 3.6 and later.

X.509 Client Certificates

- Requires a HTTPS connection to the API server.
- Verified by the API server against a trusted certificate authority bundle.
- The API server creates and distributes certificates to controllers to authenticate themselves.

Any request with an invalid access token or an invalid certificate is rejected by the authentication layer with a 401 error.

If no access token or certificate is presented, the authentication layer assigns the **system:anonymous** virtual user and the **system:unauthenticated** virtual group to the request. This allows the authorization layer to determine which requests, if any, an anonymous user is allowed to make.

4.1.3.1. Impersonation

A request to the OpenShift Container Platform API can include an **Impersonate-User** header, which indicates that the requester wants to have the request handled as though it came from the specified user. You impersonate a user by adding the **--as=<user>** flag to requests.

Before User A can impersonate User B, User A is authenticated. Then, an authorization check occurs to ensure that User A is allowed to impersonate the user named User B. If User A is requesting to impersonate a service account, **system:serviceaccount:namespace:name**, OpenShift Container Platform confirms that User A can impersonate the **serviceaccount** named **name** in **namespace**. If the check fails, the request fails with a 403 (Forbidden) error code.

By default, project administrators and editors can impersonate service accounts in their namespace. The **sudoers** role allows a user to impersonate **system:admin**, which in turn has cluster administrator permissions. The ability to impersonate **system:admin** grants some protection against typos, but not security, for someone administering the cluster. For example, running **oc delete nodes --all** fails, but running **oc delete nodes --all --as=system:admin** succeeds. You can grant a user that permission by running this command:

```
$ oc create clusterrolebinding <any_valid_name> --clusterrole=sudoer --user=<username>
```

If you need to create a project request on behalf of a user, include the **--as=<user> --as-group=<group1> --as-group=<group2>** flags in your command. Because **system:authenticated:oauth** is the only bootstrap group that can create project requests, you must impersonate that group, as shown in the following example:

```
$ oc new-project <project> --as=<user> \
--as-group=system:authenticated --as-group=system:authenticated:oauth
```

4.1.4. OAuth

The OpenShift Container Platform master includes a built-in OAuth server. Users obtain OAuth access tokens to authenticate themselves to the API.

When a person requests a new OAuth token, the OAuth server uses the configured identity provider to determine the identity of the person making the request.

It then determines what user that identity maps to, creates an access token for that user, and returns the token for use.

4.1.4.1. OAuth Clients

Every request for an OAuth token must specify the OAuth client that will receive and use the token. The following OAuth clients are automatically created when starting the OpenShift Container Platform API:

OAuth Client	Usage
openshift-web-console	Requests tokens for the web console.
openshift-browser-client	Requests tokens at **<*master*>/oauth/token/request** with a user-agent that can handle interactive logins.
openshift-challenging-client	Requests tokens with a user-agent that can handle **WWW-Authenticate** challenges.

To register additional clients:

```
$ oc create -f <(echo '
kind: OAuthClient
apiVersion: oauth.openshift.io/v1
metadata:
 name: demo (1)
secret: "..." (2)
redirectURIs:
 - "http://www.example.com/" (3)
grantMethod: prompt (4)
')
```

1. The **name** of the OAuth client is used as the **client_id** parameter when making requests to **<*master*>/oauth/authorize** and **<*master*>/oauth/token**.

2. The **secret** is used as the **client_secret** parameter when making requests to **<*master*>/oauth/token**.

3. The **redirect_uri** parameter specified in requests to **<*master*>/oauth/authorize** and **<*master*>/oauth/token** must be equal to (or prefixed by) one of the URIs in **redirectURIs**.

4. The **grantMethod** is used to determine what action to take when this client requests tokens and has not yet been granted access by the user. Uses the same values seen in Grant Options.

4.1.4.2. Service Accounts as OAuth Clients

A service account can be used as a constrained form of OAuth client. Service accounts can only request a subset of scopes that allow access to some basic user information and role-based power inside of the service account's own namespace:

- **user:info**
- **user:check-access**
- **role:<any_role>:<serviceaccount_namespace>**
- **role:<any_role>:<serviceaccount_namespace>:!**

When using a service account as an OAuth client:

- **client_id** is **system:serviceaccount:<serviceaccount_namespace>:<serviceaccount_name>**.
- **client_secret** can be any of the API tokens for that service account. For example:

  ```
  $ oc sa get-token <serviceaccount_name>
  ```

- To get **WWW-Authenticate** challenges, set an **serviceaccounts.openshift.io/oauth-want-challenges** annotation on the service account to **true**.
- **redirect_uri** must match an annotation on the service account. Redirect URIs for Service Accounts as OAuth Clients provides more information.

4.1.4.3. Redirect URIs for Service Accounts as OAuth Clients

Annotation keys must have the prefix **serviceaccounts.openshift.io/oauth-redirecturi.** or **serviceaccounts.openshift.io/oauth-redirectreference.** such as:

```
serviceaccounts.openshift.io/oauth-redirecturi.<name>
```

In its simplest form, the annotation can be used to directly specify valid redirect URIs. For example:

```
"serviceaccounts.openshift.io/oauth-redirecturi.first":  "https://example.com"
"serviceaccounts.openshift.io/oauth-redirecturi.second": "https://other.com"
```

The **first** and **second** postfixes in the above example are used to separate the two valid redirect URIs.

In more complex configurations, static redirect URIs may not be enough. For example, perhaps you want all ingresses for a route to be considered valid. This is where dynamic redirect URIs via the **serviceaccounts.openshift.io/oauth-redirectreference.** prefix come into play.

For example:

```
"serviceaccounts.openshift.io/oauth-redirectreference.first": "
{\"kind\":\"OAuthRedirectReference\",\"apiVersion\":\"v1\",\"reference\":
{\"kind\":\"Route\",\"name\":\"jenkins\"}}"
```

Since the value for this annotation contains serialized JSON data, it is easier to see in an expanded format:

```
{
```

```
  "kind": "OAuthRedirectReference",
  "apiVersion": "v1",
  "reference": {
    "kind": "Route",
    "name": "jenkins"
  }
}
```

Now you can see that an **OAuthRedirectReference** allows us to reference the route named **jenkins**. Thus, all ingresses for that route will now be considered valid. The full specification for an **OAuthRedirectReference** is:

```
{
  "kind": "OAuthRedirectReference",
  "apiVersion": "v1",
  "reference": {
    "kind": ..., 1
    "name": ..., 2
    "group": ... 3
  }
}
```

1. **kind** refers to the type of the object being referenced. Currently, only **route** is supported.

2. **name** refers to the name of the object. The object must be in the same namespace as the service account.

3. **group** refers to the group of the object. Leave this blank, as the group for a route is the empty string.

Both annotation prefixes can be combined to override the data provided by the reference object. For example:

```
"serviceaccounts.openshift.io/oauth-redirecturi.first":  "custompath"
"serviceaccounts.openshift.io/oauth-redirectreference.first": "
{\"kind\":\"OAuthRedirectReference\",\"apiVersion\":\"v1\",\"reference\":
{\"kind\":\"Route\",\"name\":\"jenkins\"}}"
```

The **first** postfix is used to tie the annotations together. Assuming that the **jenkins** route had an ingress of ***https://example.com***, now ***https://example.com/custompath*** is considered valid, but ***https://example.com*** is not. The format for partially supplying override data is as follows:

Type	Syntax
Scheme	"https://"
Hostname	"//website.com"
Port	"//:8000"
Path	"examplepath"

NOTE

Specifying a host name override will replace the host name data from the referenced object, which is not likely to be desired behavior.

Any combination of the above syntax can be combined using the following format:

<scheme:>//<hostname><:port>/<path>

The same object can be referenced more than once for more flexibility:

```
"serviceaccounts.openshift.io/oauth-redirecturi.first":  "custompath"
"serviceaccounts.openshift.io/oauth-redirectreference.first": "
{\"kind\":\"OAuthRedirectReference\",\"apiVersion\":\"v1\",\"reference\":
{\"kind\":\"Route\",\"name\":\"jenkins\"}}"
"serviceaccounts.openshift.io/oauth-redirecturi.second":  "//:8000"
"serviceaccounts.openshift.io/oauth-redirectreference.second": "
{\"kind\":\"OAuthRedirectReference\",\"apiVersion\":\"v1\",\"reference\":
{\"kind\":\"Route\",\"name\":\"jenkins\"}}"
```

Assuming that the route named **jenkins** has an ingress of ***https://example.com***, then both ***https://example.com:8000*** and ***https://example.com/custompath*** are considered valid.

Static and dynamic annotations can be used at the same time to achieve the desired behavior:

```
"serviceaccounts.openshift.io/oauth-redirectreference.first": "
{\"kind\":\"OAuthRedirectReference\",\"apiVersion\":\"v1\",\"reference\":
{\"kind\":\"Route\",\"name\":\"jenkins\"}}"
"serviceaccounts.openshift.io/oauth-redirecturi.second": "https://other.com"
```

4.1.4.3.1. API Events for OAuth

In some cases the API server returns an **unexpected condition** error message that is difficult to debug without direct access to the API master log. The underlying reason for the error is purposely obscured in order to avoid providing an unauthenticated user with information about the server’s state.

A subset of these errors is related to service account OAuth configuration issues. These issues are captured in events that can be viewed by non-administrator users. When encountering an **unexpected condition** server error during OAuth, run **oc get events** to view these events under **ServiceAccount**.

The following example warns of a service account that is missing a proper OAuth redirect URI:

```
$ oc get events | grep ServiceAccount
1m         1m          1         proxy                  ServiceAccount                          Warning
NoSAOAuthRedirectURIs   service-account-oauth-client-getter
system:serviceaccount:myproject:proxy has no redirectURIs; set serviceaccounts.openshift.io/oauth-
redirecturi.<some-value>=<redirect> or create a dynamic URI using
serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>
```

Running **oc describe sa/<service-account-name>** reports any OAuth events associated with the given service account name.

```
$ oc describe sa/proxy | grep -A5 Events
Events:
  FirstSeen     LastSeen       Count   From                              SubObjectPath   Type          Reason
Message
  ---------     --------       -----   ----                              -------------   --------      ------           -------
  3m            3m             1       service-account-oauth-client-getter                Warning
NoSAOAuthRedirectURIs   system:serviceaccount:myproject:proxy has no redirectURIs; set
serviceaccounts.openshift.io/oauth-redirecturi.<some-value>=<redirect> or create a dynamic URI
using serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>
```

The following is a list of the possible event errors:

No redirect URI annotations or an invalid URI is specified

```
Reason                  Message
NoSAOAuthRedirectURIs   system:serviceaccount:myproject:proxy has no redirectURIs; set
serviceaccounts.openshift.io/oauth-redirecturi.<some-value>=<redirect> or create a dynamic URI
using serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>
```

Invalid route specified

```
Reason                  Message
NoSAOAuthRedirectURIs   [routes.route.openshift.io "<name>" not found,
system:serviceaccount:myproject:proxy has no redirectURIs; set serviceaccounts.openshift.io/oauth-
redirecturi.<some-value>=<redirect> or create a dynamic URI using
serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>]
```

Invalid reference type specified

```
Reason                  Message
NoSAOAuthRedirectURIs   [no kind "<name>" is registered for version "v1",
system:serviceaccount:myproject:proxy has no redirectURIs; set serviceaccounts.openshift.io/oauth-
redirecturi.<some-value>=<redirect> or create a dynamic URI using
serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>]
```

Missing SA tokens

```
Reason                  Message
NoSAOAuthTokens         system:serviceaccount:myproject:proxy has no tokens
```

4.1.4.3.1.1. Sample API Event Caused by a Possible Misconfiguration

The following steps represent one way a user could get into a broken state and how to debug or fix the issue:

1. Create a project utilizing a service account as an OAuth client.

   a. Create YAML for a proxy service account object and ensure it uses the route **proxy**:

   ```
   vi serviceaccount.yaml
   ```

   Add the following sample code:

   ```
   apiVersion: v1
   kind: ServiceAccount
   metadata:
   ```

```
  name: proxy
  annotations:
    serviceaccounts.openshift.io/oauth-redirectreference.primary:
'{"kind":"OAuthRedirectReference","apiVersion":"v1","reference":
{"kind":"Route","name":"proxy"}}'
```

b. Create YAML for a route object to create a secure connection to the proxy:

```
vi route.yaml
```

Add the following sample code:

```
apiVersion: route.openshift.io/v1
kind: Route
metadata:
  name: proxy
spec:
  to:
    name: proxy
  tls:
    termination: Reencrypt
apiVersion: v1
kind: Service
metadata:
  name: proxy
  annotations:
    service.alpha.openshift.io/serving-cert-secret-name: proxy-tls
spec:
  ports:
  - name: proxy
    port: 443
    targetPort: 8443
  selector:
    app: proxy
```

c. Create a YAML for a deployment configuration to launch a proxy as a sidecar:

```
vi proxysidecar.yaml
```

Add the following sample code:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: proxy
spec:
  replicas: 1
  selector:
    matchLabels:
      app: proxy
  template:
    metadata:
      labels:
        app: proxy
    spec:
      serviceAccountName: proxy
      containers:
      - name: oauth-proxy
        image: openshift3/oauth-proxy
        imagePullPolicy: IfNotPresent
        ports:
        - containerPort: 8443
          name: public
        args:
        - --https-address=:8443
        - --provider=openshift
        - --openshift-service-account=proxy
        - --upstream=http://localhost:8080
        - --tls-cert=/etc/tls/private/tls.crt
        - --tls-key=/etc/tls/private/tls.key
        - --cookie-secret=SECRET
        volumeMounts:
        - mountPath: /etc/tls/private
          name: proxy-tls

      - name: app
        image: openshift/hello-openshift:latest
      volumes:
      - name: proxy-tls
        secret:
          secretName: proxy-tls
```

d. Create the objects

```
oc create -f serviceaccount.yaml
oc create -f route.yaml
oc create -f proxysidecar.yaml
```

2. Run **oc edit sa/proxy** to edit the service account and change the **serviceaccounts.openshift.io/oauth-redirectreference** annotation to point to a Route that does not exist.

```
apiVersion: v1
imagePullSecrets:
- name: proxy-dockercfg-08d5n
kind: ServiceAccount
metadata:
  annotations:
    serviceaccounts.openshift.io/oauth-redirectreference.primary:
'{"kind":"OAuthRedirectReference","apiVersion":"v1","reference":
{"kind":"Route","name":"notexist"}}'
...
```

3. Review the OAuth log for the service to locate the server error:

```
The authorization server encountered an unexpected condition that prevented it from fulfilling
the request.
```

4. Run **oc get events** to view the **ServiceAccount** event:

```
oc get events | grep ServiceAccount

23m        23m        1        proxy                ServiceAccount                              Warning
NoSAOAuthRedirectURIs   service-account-oauth-client-getter   [routes.route.openshift.io
"notexist" not found, system:serviceaccount:myproject:proxy has no redirectURIs; set
serviceaccounts.openshift.io/oauth-redirecturi.<some-value>=<redirect> or create a dynamic
URI using serviceaccounts.openshift.io/oauth-redirectreference.<some-value>=<reference>]
```

4.1.4.4. Integrations

All requests for OAuth tokens involve a request to ***<master>*/oauth/authorize**. Most authentication integrations place an authenticating proxy in front of this endpoint, or configure OpenShift Container Platform to validate credentials against a backing identity provider. Requests to ***<master>*/oauth/authorize** can come from user-agents that cannot display interactive login pages, such as the CLI. Therefore, OpenShift Container Platform supports authenticating using a **WWW-Authenticate** challenge in addition to interactive login flows.

If an authenticating proxy is placed in front of the ***<master>*/oauth/authorize** endpoint, it should send unauthenticated, non-browser user-agents **WWW-Authenticate** challenges, rather than displaying an interactive login page or redirecting to an interactive login flow.

NOTE

To prevent cross-site request forgery (CSRF) attacks against browser clients, Basic authentication challenges should only be sent if a **X-CSRF-Token** header is present on the request. Clients that expect to receive Basic **WWW-Authenticate** challenges should set this header to a non-empty value.

If the authenticating proxy cannot support **WWW-Authenticate** challenges, or if OpenShift Container Platform is configured to use an identity provider that does not support WWW-Authenticate challenges, users can visit ***<master>*/oauth/token/request** using a browser to obtain an access token manually.

4.1.4.5. OAuth Server Metadata

Applications running in OpenShift Container Platform may need to discover information about the built-in OAuth server. For example, they may need to discover what the address of the **<master>** server is without manual configuration. To aid in this, OpenShift Container Platform implements the IETF OAuth 2.0 Authorization Server Metadata draft specification.

Thus, any application running inside the cluster can issue a **GET** request to ***https://openshift.default.svc/.well-known/oauth-authorization-server*** to fetch the following information:

```
{
  "issuer": "https://<master>", 1
  "authorization_endpoint": "https://<master>/oauth/authorize", 2
  "token_endpoint": "https://<master>/oauth/token", 3
  "scopes_supported": [ 4
    "user:full",
    "user:info",
    "user:check-access",
    "user:list-scoped-projects",
    "user:list-projects"
  ],
  "response_types_supported": [ ❺
    "code",
    "token"
  ],
  "grant_types_supported": [ ❻
    "authorization_code",
    "implicit"
  ],
  "code_challenge_methods_supported": [ ❼
    "plain",
    "S256"
  ]
}
```

1. The authorization server's issuer identifier, which is a URL that uses the **https** scheme and has no query or fragment components. This is the location where **.well-known** RFC 5785 resources containing information about the authorization server are published.

2. URL of the authorization server's authorization endpoint. See RFC 6749.

3. URL of the authorization server's token endpoint. See RFC 6749.

4. JSON array containing a list of the OAuth 2.0 RFC 6749 scope values that this authorization server supports. Note that not all supported scope values are advertised.

5. JSON array containing a list of the OAuth 2.0 **response_type** values that this authorization server supports. The array values used are the same as those used with the **response_types** parameter defined by "OAuth 2.0 Dynamic Client Registration Protocol" in RFC 7591.

6. JSON array containing a list of the OAuth 2.0 grant type values that this authorization server supports. The array values used are the same as those used with the **grant_types** parameter defined by **OAuth 2.0 Dynamic Client Registration Protocol** in RFC 7591.

7. JSON array containing a list of PKCE RFC 7636 code challenge methods supported by this authorization server. Code challenge method values are used in the **code_challenge_method** parameter defined in Section 4.3 of RFC 7636 . The valid code challenge method values are those registered in the IANA **PKCE Code Challenge Methods** registry. See IANA OAuth Parameters.

4.1.4.6. Obtaining OAuth Tokens

The OAuth server supports standard authorization code grant and the implicit grant OAuth authorization flows.

Run the following command to request an OAuth token by using the authorization code grant method:

```
$ curl -H "X-Remote-User: <username>" \
     --cacert /etc/origin/master/ca.crt \
     --cert /etc/origin/master/admin.crt \
     --key /etc/origin/master/admin.key \
     -I https://<master-address>/oauth/authorize?response_type=token\&client_id=openshift-
challenging-client | grep -oP "access_token=\K[^&]*"
```

When requesting an OAuth token using the implicit grant flow (**response_type=token**) with a client_id configured to request **WWW-Authenticate challenges** (like **openshift-challenging-client**), these are the possible server responses from **/oauth/authorize**, and how they should be handled:

Status	Content	Client response
302	**Location** header containing an **access_token** parameter in the URL fragment (RFC 4.2.2)	Use the **access_token** value as the OAuth token
302	**Location** header containing an **error** query parameter (RFC 4.1.2.1)	Fail, optionally surfacing the **error** (and optional **error_description**) query values to the user
302	Other **Location** header	Follow the redirect, and process the result using these rules
401	**WWW-Authenticate** header present	Respond to challenge if type is recognized (e.g. **Basic**, **Negotiate**, etc), resubmit request, and process the result using these rules
401	**WWW-Authenticate** header missing	No challenge authentication is possible. Fail and show response body (which might contain links or details on alternate methods to obtain an OAuth token)
Other	Other	Fail, optionally surfacing response body to the user

To request an OAuth token using the implicit grant flow:

```
$ curl -u <username>:<password>
'https://<master-address>:8443/oauth/authorize?client_id=openshift-challenging-
client&response_type=token' -skv / 1
/ -H "X-CSRF-Token: xxx" 2
*   Trying 10.64.33.43...
* Connected to 10.64.33.43 (10.64.33.43) port 8443 (#0)
* found 148 certificates in /etc/ssl/certs/ca-certificates.crt
* found 592 certificates in /etc/ssl/certs
* ALPN, offering http/1.1
* SSL connection using TLS1.2 / ECDHE_RSA_AES_128_GCM_SHA256
*        server certificate verification SKIPPED
*        server certificate status verification SKIPPED
*        common name: 10.64.33.43 (matched)
*        server certificate expiration date OK
*        server certificate activation date OK
*        certificate public key: RSA
*        certificate version: #3
*        subject: CN=10.64.33.43
*        start date: Thu, 09 Aug 2018 04:00:39 GMT
*       expire date: Sat, 08 Aug 2020 04:00:40 GMT
*       issuer: CN=openshift-signer@1531109367
*       compression: NULL
* ALPN, server accepted to use http/1.1
* Server auth using Basic with user 'developer'
> GET /oauth/authorize?client_id=openshift-challenging-client&response_type=token HTTP/1.1
> Host: 10.64.33.43:8443
> Authorization: Basic ZGV2ZWxvcGVyOmRzc2Zkcw==
> User-Agent: curl/7.47.0
> Accept: */*
> X-CSRF-Token: xxx
>
< HTTP/1.1 302 Found
< Cache-Control: no-cache, no-store, max-age=0, must-revalidate
< Expires: Fri, 01 Jan 1990 00:00:00 GMT
< Location:
https://10.64.33.43:8443/oauth/token/implicit#access_token=gzTwOq_mVJ7ovHliHBTgRQEEXa1aCZ
D9lnj7lSw3ekQ&expires_in=86400&scope=user%3Afull&token_type=Bearer 3
< Pragma: no-cache
< Set-Cookie:
ssn=MTUzNTk0OTc1MnxIckVfNW5vNFlLSlF5MF9GWEF6Zm55Vl95bi1ZNE41S1NCbFJMYnN1TWV
wR1hwZmlLMzFQRklzVXRkc0RnUGEzdnBEa0NZZndXV2ZUVzN1dmFPM2dHSUlzUmVXakQ3Q09rV
XpxNlRoVmVkQU5DYmdLTE9SUWlyNkJJTm1mSDQ0N2pCV09La3gzMkMzckwxc1V1QXpybFlXT2ZY
SmI2R2FTVEZsdDBzRjJ8vk6zrQPjQUmoJCqb8Dt5j5s0b4wZlITgKlho9wlKAZI=; Path=/; HttpOnly;
Secure
< Date: Mon, 03 Sep 2018 04:42:32 GMT
< Content-Length: 0
< Content-Type: text/plain; charset=utf-8
<
* Connection #0 to host 10.64.33.43 left intact
```

1. **client-id** is set to **openshift-challenging-client** and **response-type** is set to **token**.
2. Set **X-CSRF-Token** header to a non-empty value.
3. The token is returned in the **Location** header of the **302** response as **access_token=gzTwOq_mVJ7ovHliHBTgRQEEXa1aCZD9lnj7lSw3ekQ**.

To view only the OAuth token value, run the following command:

```
$ curl -u <username>:<password>
'https://<master-address>:8443/oauth/authorize?client_id=openshift-challenging-
client&response_type=token' 1
-skv -H "X-CSRF-Token: xxx" --stderr - |  grep -oP "access_token=\K[^&]*" 2

hvqxe5aMlAzvbqfM2WWw3D6tR0R2jCQGKx0viZBxwmc
```

1. **client-id** is set to **openshift-challenging-client** and **response-type** is set to **token**.
2. Set **X-CSRF-Token** header to a non-empty value.

You can also use the **Code Grant** method to request a token

4.1.4.7. Authentication Metrics for Prometheus

OpenShift Container Platform captures the following Prometheus system metrics during authentication attempts:

- **openshift_auth_basic_password_count** counts the number of **oc login** user name and password attempts.
- **openshift_auth_basic_password_count_result** counts the number of **oc login** user name and password attempts by result (success or error).
- **openshift_auth_form_password_count** counts the number of web console login attempts.
- **openshift_auth_form_password_count_result** counts the number of web console login attempts by result (success or error).
- **openshift_auth_password_total** counts the total number of **oc login** and web console login attempts.

4.2. AUTHORIZATION

4.2.1. Overview

Role-based Access Control (RBAC) objects determine whether a user is allowed to perform a given action within a project.

This allows platform administrators to use the cluster roles and bindings to control who has various access levels to the OpenShift Container Platform platform itself and all projects.

It allows developers to use local roles and bindings to control who has access to their projects. Note that authorization is a separate step from authentication, which is more about determining the identity of who is taking the action.

Authorization is managed using:

Rules	Sets of permitted verbs on a set ofobjects. For example, whether something can**create** pods.
Roles	Collections of rules. Users and groups can be associated with, or*bound* to, multiple roles at the same time.
Bindings	Associations between users and/or groups with a role.

Cluster administrators can visualize rules, roles, and bindings using the CLI.

For example, consider the following excerpt that shows the rule sets for the **admin** and **basic-user** default cluster roles:

```
$ oc describe clusterrole.rbac admin basic-user
```

```
Name:  admin
Labels:  <none>
Annotations: openshift.io/description=A user that has edit rights within the project and can change the project's membership.
```

```
  rbac.authorization.kubernetes.io/autoupdate=true
PolicyRule:
  Resources       Non-Resource URLs Resource Names Verbs
  ---------       ----------------- -------------- -----
  appliedclusterresourcequotas     []   []  [get list watch]
  appliedclusterresourcequotas.quota.openshift.io  []   []  [get list watch]
  bindings       []   []  [get list watch]
  buildconfigs       []   []  [create delete deletecollection get list patch update watch]
  buildconfigs.build.openshift.io    []   []  [create delete deletecollection get list patch update watch]
  buildconfigs/instantiate     []   []  [create]
  buildconfigs.build.openshift.io/instantiate   []   []  [create]
  buildconfigs/instantiatebinary    []   []  [create]
  buildconfigs.build.openshift.io/instantiatebinary  []   []  [create]
  buildconfigs/webhooks      []   []  [create delete deletecollection get list patch update watch]
  buildconfigs.build.openshift.io/webhooks   []   []  [create delete deletecollection get list patch update
watch]
  buildlogs       []   []  [create delete deletecollection get list patch update watch]
  buildlogs.build.openshift.io     []   []  [create delete deletecollection get list patch update watch]
  builds       []   []  [create delete deletecollection get list patch update watch]
  builds.build.openshift.io     []   []  [create delete deletecollection get list patch update watch]
  builds/clone       []   []  [create]
  builds.build.openshift.io/clone    []   []  [create]
  builds/details      []   []  [update]
  builds.build.openshift.io/details    []   []  [update]
  builds/log       []   []  [get list watch]
  builds.build.openshift.io/log     []   []  [get list watch]
  configmaps       []   []  [create delete deletecollection get list patch update watch]
  cronjobs.batch      []   []  [create delete deletecollection get list patch update watch]
  daemonsets.extensions      []   []  [get list watch]
  deploymentconfigrollbacks     []   []  [create]
  deploymentconfigrollbacks.apps.openshift.io   []   []  [create]
  deploymentconfigs      []   []  [create delete deletecollection get list patch update watch]
  deploymentconfigs.apps.openshift.io    []   []  [create delete deletecollection get list patch update
watch]
  deploymentconfigs/instantiate     []   []  [create]
  deploymentconfigs.apps.openshift.io/instantiate  []   []  [create]
  deploymentconfigs/log      []   []  [get list watch]
  deploymentconfigs.apps.openshift.io/log   []   []  [get list watch]
  deploymentconfigs/rollback     []   []  [create]
  deploymentconfigs.apps.openshift.io/rollback   []   []  [create]
  deploymentconfigs/scale     []   []  [create delete deletecollection get list patch update watch]
  deploymentconfigs.apps.openshift.io/scale   []   []  [create delete deletecollection get list patch
update watch]
  deploymentconfigs/status     []   []  [get list watch]
  deploymentconfigs.apps.openshift.io/status   []   []  [get list watch]
  deployments.apps      []   []  [create delete deletecollection get list patch update watch]
  deployments.extensions      []   []  [create delete deletecollection get list patch update watch]
  deployments.extensions/rollback    []   []  [create delete deletecollection get list patch update watch]
  deployments.apps/scale      []   []  [create delete deletecollection get list patch update watch]
  deployments.extensions/scale     []   []  [create delete deletecollection get list patch update watch]
  deployments.apps/status     []   []  [create delete deletecollection get list patch update watch]
  endpoints       []   []  [create delete deletecollection get list patch update watch]
  events       []   []  [get list watch]
  horizontalpodautoscalers.autoscaling    []   []  [create delete deletecollection get list patch update
watch]
  horizontalpodautoscalers.extensions    []   []  [create delete deletecollection get list patch update
```

```
watch]
  imagestreamimages      []   []  [create delete deletecollection get list patch update watch]
  imagestreamimages.image.openshift.io    []   []  [create delete deletecollection get list patch update
watch]
  imagestreamimports      []   []  [create]
  imagestreamimports.image.openshift.io    []   []  [create]
  imagestreammappings      []   []  [create delete deletecollection get list patch update watch]
  imagestreammappings.image.openshift.io   []   []  [create delete deletecollection get list patch update
watch]
  imagestreams       []   []  [create delete deletecollection get list patch update watch]
  imagestreams.image.openshift.io    []   []  [create delete deletecollection get list patch update watch]
  imagestreams/layers      []   []  [get update]
  imagestreams.image.openshift.io/layers   []   []  [get update]
  imagestreams/secrets      []   []  [create delete deletecollection get list patch update watch]
  imagestreams.image.openshift.io/secrets   []   []  [create delete deletecollection get list patch update
watch]
  imagestreams/status      []   []  [get list watch]
  imagestreams.image.openshift.io/status   []   []  [get list watch]
  imagestreamtags      []   []  [create delete deletecollection get list patch update watch]
  imagestreamtags.image.openshift.io    []   []  [create delete deletecollection get list patch update
watch]
  jenkins.build.openshift.io     []   []  [admin edit view]
  jobs.batch       []   []  [create delete deletecollection get list patch update watch]
  limitranges       []   []  [get list watch]
  localresourceaccessreviews     []   []  [create]
  localresourceaccessreviews.authorization.openshift.io  []   []  [create]
  localsubjectaccessreviews     []   []  [create]
  localsubjectaccessreviews.authorization.k8s.io  []   []  [create]
  localsubjectaccessreviews.authorization.openshift.io  []   []  [create]
  namespaces       []   []  [get list watch]
  namespaces/status      []   []  [get list watch]
  networkpolicies.extensions     []   []  [create delete deletecollection get list patch update watch]
  persistentvolumeclaims      []   []  [create delete deletecollection get list patch update watch]
  pods        []   []  [create delete deletecollection get list patch update watch]
  pods/attach       []   []  [create delete deletecollection get list patch update watch]
  pods/exec       []   []  [create delete deletecollection get list patch update watch]
  pods/log       []   []  [get list watch]
  pods/portforward      []   []  [create delete deletecollection get list patch update watch]
  pods/proxy       []   []  [create delete deletecollection get list patch update watch]
  pods/status       []   []  [get list watch]
  podsecuritypolicyreviews     []   []  [create]
  podsecuritypolicyreviews.security.openshift.io  []   []  [create]
  podsecuritypolicyselfsubjectreviews     []   []  [create]
  podsecuritypolicyselfsubjectreviews.security.openshift.io []   []  [create]
  podsecuritypolicysubjectreviews     []   []  [create]
  podsecuritypolicysubjectreviews.security.openshift.io  []   []  [create]
  processedtemplates      []   []  [create delete deletecollection get list patch update watch]
  processedtemplates.template.openshift.io   []   []  [create delete deletecollection get list patch update
watch]
  projects       []   []  [delete get patch update]
  projects.project.openshift.io     []   []  [delete get patch update]
  replicasets.extensions     []   []  [create delete deletecollection get list patch update watch]
  replicasets.extensions/scale     []   []  [create delete deletecollection get list patch update watch]
  replicationcontrollers     []   []  [create delete deletecollection get list patch update watch]
  replicationcontrollers/scale     []   []  [create delete deletecollection get list patch update watch]
  replicationcontrollers.extensions/scale   []   []  [create delete deletecollection get list patch update
```

```
watch]
  replicationcontrollers/status     []   []  [get list watch]
  resourceaccessreviews      []   []  [create]
  resourceaccessreviews.authorization.openshift.io  []   []  [create]
  resourcequotas      []   []  [get list watch]
  resourcequotas/status      []   []  [get list watch]
  resourcequotausages      []   []  [get list watch]
  rolebindingrestrictions     []   []  [get list watch]
  rolebindingrestrictions.authorization.openshift.io  []   []  [get list watch]
  rolebindings       []   []  [create delete deletecollection get list patch update watch]
  rolebindings.authorization.openshift.io   []   []  [create delete deletecollection get list patch update
watch]
  rolebindings.rbac.authorization.k8s.io   []   []  [create delete deletecollection get list patch update
watch]
  roles        []   []  [create delete deletecollection get list patch update watch]
  roles.authorization.openshift.io    []   []  [create delete deletecollection get list patch update watch]
  roles.rbac.authorization.k8s.io    []   []  [create delete deletecollection get list patch update watch]
  routes       []   []  [create delete deletecollection get list patch update watch]
  routes.route.openshift.io     []   []  [create delete deletecollection get list patch update watch]
  routes/custom-host      []   []  [create]
  routes.route.openshift.io/custom-host    []   []  [create]
  routes/status       []   []  [get list watch update]
  routes.route.openshift.io/status    []   []  [get list watch update]
  scheduledjobs.batch      []   []  [create delete deletecollection get list patch update watch]
  secrets       []   []  [create delete deletecollection get list patch update watch]
  serviceaccounts      []   []  [create delete deletecollection get list patch update watch impersonate]
  services       []   []  [create delete deletecollection get list patch update watch]
  services/proxy      []   []  [create delete deletecollection get list patch update watch]
  statefulsets.apps      []   []  [create delete deletecollection get list patch update watch]
  subjectaccessreviews      []   []  [create]
  subjectaccessreviews.authorization.openshift.io  []   []  [create]
  subjectrulesreviews      []   []  [create]
  subjectrulesreviews.authorization.openshift.io  []   []  [create]
  templateconfigs      []   []  [create delete deletecollection get list patch update watch]
  templateconfigs.template.openshift.io    []   []  [create delete deletecollection get list patch update
watch]
  templateinstances      []   []  [create delete deletecollection get list patch update watch]
  templateinstances.template.openshift.io   []   []  [create delete deletecollection get list patch update
watch]
  templates      []   []  [create delete deletecollection get list patch update watch]
  templates.template.openshift.io    []   []  [create delete deletecollection get list patch update watch]


Name:  basic-user
Labels:  <none>
Annotations: openshift.io/description=A user that can get basic information about projects.
  rbac.authorization.kubernetes.io/autoupdate=true
PolicyRule:
  Resources      Non-Resource URLs Resource Names Verbs
  ---------      ----------------- -------------- -----
  clusterroles      []   []  [get list]
  clusterroles.authorization.openshift.io  []   []  [get list]
  clusterroles.rbac.authorization.k8s.io  []   []  [get list watch]
  projectrequests     []   []  [list]
  projectrequests.project.openshift.io   []   []  [list]
  projects      []   []  [list watch]
  projects.project.openshift.io    []   []  [list watch]
  selfsubjectaccessreviews.authorization.k8s.io  []   []  [create]
  selfsubjectrulesreviews    []   []  [create]
  selfsubjectrulesreviews.authorization.openshift.io []   []  [create]
  storageclasses.storage.k8s.io    []   []  [get list]
  users       []   [~]  [get]
  users.user.openshift.io    []   [~]  [get]
```

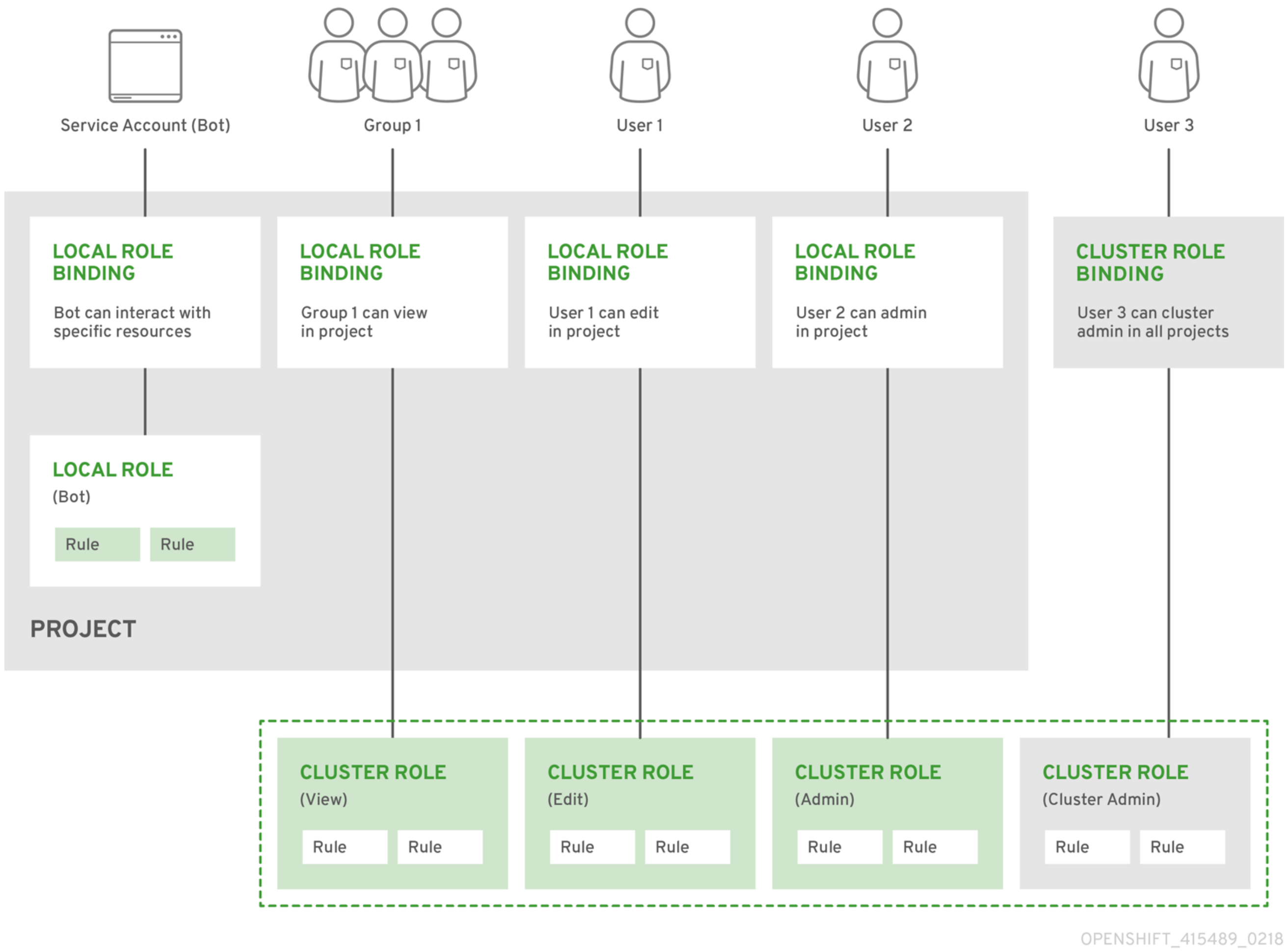
The following excerpt from viewing local role bindings shows the above roles bound to various users and groups:

```
oc describe rolebinding.rbac admin basic-user -n alice-project
```

```
Name:  admin
Labels:  <none>
Annotations: <none>
Role:
  Kind: ClusterRole
  Name: admin
Subjects:
  Kind Name  Namespace
  ---- ----  ---------
  User system:admin
  User alice


Name:  basic-user
Labels:  <none>
Annotations: <none>
Role:
  Kind: ClusterRole
  Name: basic-user
Subjects:
  Kind Name Namespace
  ---- ---- ---------
  User joe
  Group devel
```

The relationships between cluster roles, local roles, cluster role bindings, local role bindings, users, groups and service accounts are illustrated below.



4.2.2. Evaluating Authorization

Several factors are combined to make the decision when OpenShift Container Platform evaluates authorization:

Identity	In the context of authorization, both the user name and list of groups the user belongs to.
Action	The action being performed. In most cases, this consists of:
	Project: The project being accessed.
	Verb: Can be **get**, **list**, **create**, **update**, **delete**, **deletecollection** or **watch**.
	Resource Name: The API endpoint being accessed.
Bindings	The full list of bindings.

OpenShift Container Platform evaluates authorizations using the following steps:

1. The identity and the project-scoped action is used to find all bindings that apply to the user or their groups.

2. Bindings are used to locate all the roles that apply.

3. Roles are used to find all the rules that apply.

4. The action is checked against each rule to find a match.

5. If no matching rule is found, the action is then denied by default.

4.2.3. Cluster and Local RBAC

There are two levels of RBAC roles and bindings that control authorization:

Cluster RBAC	Roles and bindings that are applicable across all projects. Roles that exist cluster-wide are considered *cluster roles*. Cluster role bindings can only reference cluster roles.
Local RBAC	Roles and bindings that are scoped to a given project. Roles that exist only in a project are considered *local roles*. Local role bindings can reference both cluster and local roles.

This two-level hierarchy allows re-usability over multiple projects through the cluster roles while allowing customization inside of individual projects through local roles.

During evaluation, both the cluster role bindings and the local role bindings are used. For example:

1. Cluster-wide "allow" rules are checked.

2. Locally-bound "allow" rules are checked.

3. Deny by default.

4.2.4. Cluster Roles and Local Roles

Roles are collections of policy rules, which are sets of permitted verbs that can be performed on a set of resources. OpenShift Container Platform includes a set of default cluster roles that can be bound to users and groups cluster wide or locally.

Default Cluster Role	Description
admin	A project manager. If used in a local binding, an **admin** user will have rights to view any resource in the project and modify any resource in the project except for quota.
basic-user	A user that can get basic information about projects and users.
cluster-admin	A super-user that can perform any action in any project. When bound to a user with a local binding, they have full control over quota and every action on every resource in the project.
cluster-status	A user that can get basic cluster status information.
edit	A user that can modify most objects in a project, but does not have the power to view or modify roles or bindings.

Default Cluster Role	Description
self-provisioner	A user that can create their own projects.
view	A user who cannot make any modifications, but can see most objects in a project. They cannot view or modify roles or bindings.

TIP

Remember that users and groups can be associated with, or *bound* to, multiple roles at the same time.

Project administrators can visualize roles, including a matrix of the verbs and resources each are associated using the CLI to view local roles and bindings .

IMPORTANT

The cluster role bound to the project administrator is limited in a project via a local binding. It is not bound cluster-wide like the cluster roles granted to the **cluster-admin** or **system:admin**.

Cluster roles are roles defined at the cluster level, but can be bound either at the cluster level or at the project level.

Learn how to create a local role for a project .

4.2.4.1. Updating Cluster Roles

After any OpenShift Container Platform cluster upgrade , the default roles are updated and automatically reconciled when the server is started. During reconciliation, any permissions that are missing from the default roles are added. If you added more permissions to the role, they are not removed.

If you customized the default roles and configured them to prevent automatic role reconciliation, you must manually update policy definitions when you upgrade OpenShift Container Platform.

4.2.4.2. Applying Custom Roles and Permissions

To add or update custom roles and permissions, it is strongly recommended to use the following command:

```
# oc auth reconcile -f FILE
```

This command ensures that new permissions are applied properly in a way that will not break other clients. This is done internally by computing logical covers operations between rule sets, which is something you cannot do via a JSON merge on RBAC resources.

4.2.4.3. Cluster Role Aggregation

The default admin, edit, and view cluster roles support cluster role aggregation, where the cluster rules for each role are dynamically updated as new rules are created. This feature is relevant only if you extend the Kubernetes API by creating custom resources.

Learn how to use cluster role aggregation .

4.2.5. Security Context Constraints

In addition to the RBAC resources that control what a user can do, OpenShift Container Platform provides *security context constraints* (SCC) that control the actions that a pod can perform and what it has the ability to access. Administrators can manage SCCs using the CLI.

SCCs are also very useful for managing access to persistent storage.

SCCs are objects that define a set of conditions that a pod must run with in order to be accepted into the system. They allow an administrator to control the following:

1. Running of privileged containers.
2. Capabilities a container can request to be added.
3. Use of host directories as volumes.
4. The SELinux context of the container.
5. The user ID.
6. The use of host namespaces and networking.
7. Allocating an **FSGroup** that owns the pod’s volumes
8. Configuring allowable supplemental groups
9. Requiring the use of a read only root file system
10. Controlling the usage of volume types
11. Configuring allowable seccomp profiles

Seven SCCs are added to the cluster by default, and are viewable by cluster administrators using the CLI:

```
$ oc get scc
NAME               PRIV      CAPS      SELINUX     RUNASUSER          FSGROUP     SUPGROUP
PRIORITY   READONLYROOTFS   VOLUMES
anyuid             false     []        MustRunAs   RunAsAny           RunAsAny    RunAsAny    10         false
[configMap downwardAPI emptyDir persistentVolumeClaim secret]
hostaccess         false     []        MustRunAs   MustRunAsRange     MustRunAs   RunAsAny    <none>
false              [configMap downwardAPI emptyDir hostPath persistentVolumeClaim secret]
hostmount-anyuid   false     []        MustRunAs   RunAsAny           RunAsAny    RunAsAny    <none>
false              [configMap downwardAPI emptyDir hostPath nfs persistentVolumeClaim secret]
hostnetwork        false     []        MustRunAs   MustRunAsRange     MustRunAs   MustRunAs   <none>
false              [configMap downwardAPI emptyDir persistentVolumeClaim secret]
nonroot            false     []        MustRunAs   MustRunAsNonRoot   RunAsAny    RunAsAny    <none>
false              [configMap downwardAPI emptyDir persistentVolumeClaim secret]
privileged         true      [*]       RunAsAny    RunAsAny           RunAsAny    RunAsAny    <none>
```

```
false          [*]
restricted       false     []        MustRunAs   MustRunAsRange     MustRunAs   RunAsAny    <none>
false          [configMap downwardAPI emptyDir persistentVolumeClaim secret]
```

IMPORTANT

Do not modify the default SCCs. Customizing the default SCCs can lead to issues when OpenShift Container Platform is upgraded. Instead, create new SCCs.

The definition for each SCC is also viewable by cluster administrators using the CLI. For example, for the privileged SCC:

```
# oc export scc/privileged
allowHostDirVolumePlugin: true
allowHostIPC: true
allowHostNetwork: true
allowHostPID: true
allowHostPorts: true
allowPrivilegedContainer: true
allowedCapabilities: 1
- '*'
apiVersion: v1
defaultAddCapabilities: [] 2
fsGroup: 3
  type: RunAsAny
groups: 4
- system:cluster-admins
- system:nodes
kind: SecurityContextConstraints
metadata:
  annotations:
    kubernetes.io/description: 'privileged allows access to all privileged and host
      features and the ability to run as any user, any group, any fsGroup, and with
      any SELinux context.  WARNING: this is the most relaxed SCC and should be used
      only for cluster administration. Grant with caution.'
  creationTimestamp: null
  name: privileged
priority: null
readOnlyRootFilesystem: false
requiredDropCapabilities: [] 5
runAsUser: 6
  type: RunAsAny
seLinuxContext: 7
  type: RunAsAny
seccompProfiles:
- '*'
supplementalGroups: 8
  type: RunAsAny
users: 9
- system:serviceaccount:default:registry
- system:serviceaccount:default:router
- system:serviceaccount:openshift-infra:build-controller
volumes:
- '*'
```

1. A list of capabilities that can be requested by a pod. An empty list means that none of capabilities can be requested while the special symbol * allows any capabilities.

2. A list of additional capabilities that will be added to any pod.

3. The **FSGroup** strategy which dictates the allowable values for the Security Context.

4. The groups that have access to this SCC.

5. A list of capabilities that will be dropped from a pod.

6. The run as user strategy type which dictates the allowable values for the Security Context.

7. The SELinux context strategy type which dictates the allowable values for the Security Context.

8. The supplemental groups strategy which dictates the allowable supplemental groups for the Security Context.

9. The users who have access to this SCC.

The **users** and **groups** fields on the SCC control which SCCs can be used. By default, cluster administrators, nodes, and the build controller are granted access to the privileged SCC. All authenticated users are granted access to the restricted SCC.

Docker has a default list of capabilities that are allowed for each container of a pod. The containers use the capabilities from this default list, but pod manifest authors can alter it by requesting additional capabilities or dropping some of defaulting. The **allowedCapabilities**, **defaultAddCapabilities**, and **requiredDropCapabilities** fields are used to control such requests from the pods, and to dictate which capabilities can be requested, which ones must be added to each container, and which ones must be forbidden.

The privileged SCC:

- allows privileged pods.
- allows host directories to be mounted as volumes.
- allows a pod to run as any user.
- allows a pod to run with any MCS label.
- allows a pod to use the host’s IPC namespace.
- allows a pod to use the host’s PID namespace.
- allows a pod to use any FSGroup.
- allows a pod to use any supplemental group.
- allows a pod to use any seccomp profiles.
- allows a pod to request any capabilities.

The restricted SCC:

- ensures pods cannot run as privileged.
- ensures pods cannot use host directory volumes.
- requires that a pod run as a user in a pre-allocated range of UIDs.
- requires that a pod run with a pre-allocated MCS label.
- allows a pod to use any FSGroup.
- allows a pod to use any supplemental group.

NOTE

For more information about each SCC, see the **kubernetes.io/description** annotation available on the SCC.

SCCs are comprised of settings and strategies that control the security features a pod has access to. These settings fall into three categories:

Controlled by a boolean	Fields of this type default to the most restrictive value. For example, **AllowPrivilegedContainer** is always set to **false** if unspecified.
Controlled by an allowable set	Fields of this type are checked against the set to ensure their value is allowed.
Controlled by a strategy	Items that have a strategy to generate a value provide: ● A mechanism to generate the value, and ● A mechanism to ensure that a specified value falls into the set of allowable values.

4.2.5.1. SCC Strategies

4.2.5.1.1. RunAsUser

1. **MustRunAs** - Requires a **runAsUser** to be configured. Uses the configured **runAsUser** as the default. Validates against the configured **runAsUser**.
2. **MustRunAsRange** - Requires minimum and maximum values to be defined if not using pre-allocated values. Uses the minimum as the default. Validates against the entire allowable range.
3. **MustRunAsNonRoot** - Requires that the pod be submitted with a non-zero **runAsUser** or have the **USER** directive defined in the image. No default provided.
4. **RunAsAny** - No default provided. Allows any **runAsUser** to be specified.

4.2.5.1.2. SELinuxContext

1. **MustRunAs** - Requires **seLinuxOptions** to be configured if not using pre-allocated

1. **MustRunAs** - Requires **seLinuxOptions** to be configured if not using pre-allocated values. Uses **seLinuxOptions** as the default. Validates against **seLinuxOptions**.

2. **RunAsAny** - No default provided. Allows any **seLinuxOptions** to be specified.

4.2.5.1.3. SupplementalGroups

1. **MustRunAs** - Requires at least one range to be specified if not using pre-allocated values. Uses the minimum value of the first range as the default. Validates against all ranges.

2. **RunAsAny** - No default provided. Allows any **supplementalGroups** to be specified.

4.2.5.1.4. FSGroup

1. **MustRunAs** - Requires at least one range to be specified if not using pre-allocated values. Uses the minimum value of the first range as the default. Validates against the first ID in the first range.

2. **RunAsAny** - No default provided. Allows any **fsGroup** ID to be specified.

4.2.5.2. Controlling Volumes

The usage of specific volume types can be controlled by setting the **volumes** field of the SCC. The allowable values of this field correspond to the volume sources that are defined when creating a volume:

- azureFile
- azureDisk
- flocker
- flexVolume
- hostPath
- emptyDir
- gcePersistentDisk
- awsElasticBlockStore
- gitRepo
- secret
- nfs
- iscsi
- glusterfs
- persistentVolumeClaim
- rbd
- cinder

- cephFS
- downwardAPI
- fc
- configMap
- vsphereVolume
- quobyte
- photonPersistentDisk
- projected
- portworxVolume
- scaleIO
- storageos
- ***** (a special value to allow the use of all volume types)
- **none** (a special value to disallow the use of all volumes types. Exist only for backwards compatibility)

The recommended minimum set of allowed volumes for new SCCs are **configMap**, **downwardAPI**, **emptyDir**, **persistentVolumeClaim**, **secret**, and **projected**.

NOTE

The list of allowable volume types is not exhaustive because new types are added with each release of OpenShift Container Platform.

NOTE

For backwards compatibility, the usage of **allowHostDirVolumePlugin** overrides settings in the **volumes** field. For example, if **allowHostDirVolumePlugin** is set to false but allowed in the **volumes** field, then the **hostPath** value will be removed from **volumes**.

4.2.5.3. Restricting Access to FlexVolumes

OpenShift Container Platform provides additional control of FlexVolumes based on their driver. When SCC allows the usage of FlexVolumes, pods can request any FlexVolumes. However, when the cluster administrator specifies driver names in the **AllowedFlexVolumes** field, pods must only use FlexVolumes with these drivers.

Example of Limiting Access to Only Two FlexVolumes

```
volumes:
- flexVolume
allowedFlexVolumes:
- driver: example/lvm
- driver: example/cifs
```

4.2.5.4. Seccomp

SeccompProfiles lists the allowed profiles that can be set for the pod or container's seccomp annotations. An unset (nil) or empty value means that no profiles are specified by the pod or container. Use the wildcard ***** to allow all profiles. When used to generate a value for a pod, the first non-wildcard profile is used as the default.

Refer to the seccomp documentation for more information about configuring and using custom profiles.

4.2.5.5. Admission

Admission control with SCCs allows for control over the creation of resources based on the capabilities granted to a user.

In terms of the SCCs, this means that an admission controller can inspect the user information made available in the context to retrieve an appropriate set of SCCs. Doing so ensures the pod is authorized to make requests about its operating environment or to generate a set of constraints to apply to the pod.

The set of SCCs that admission uses to authorize a pod are determined by the user identity and groups that the user belongs to. Additionally, if the pod specifies a service account, the set of allowable SCCs includes any constraints accessible to the service account.

Admission uses the following approach to create the final security context for the pod:

1. Retrieve all SCCs available for use.
2. Generate field values for security context settings that were not specified on the request.
3. Validate the final settings against the available constraints.

If a matching set of constraints is found, then the pod is accepted. If the request cannot be matched to an SCC, the pod is rejected.

A pod must validate every field against the SCC. The following are examples for just two of the fields that must be validated:

> **NOTE**
>
> These examples are in the context of a strategy using the preallocated values.

A FSGroup SCC Strategy of MustRunAs

If the pod defines a **fsGroup** ID, then that ID must equal the default **fsGroup** ID. Otherwise, the pod is not validated by that SCC and the next SCC is evaluated.

If the **SecurityContextConstraints.fsGroup** field has value **RunAsAny** and the pod specification omits the **Pod.spec.securityContext.fsGroup**, then this field is considered valid. Note that it is possible that during validation, other SCC settings will reject other pod fields and thus cause the pod to fail.

A SupplementalGroups SCC Strategy of MustRunAs

If the pod specification defines one or more **supplementalGroups** IDs, then the pod's IDs must equal one of the IDs in the namespace's **openshift.io/sa.scc.supplemental-groups** annotation. Otherwise, the pod is not validated by that SCC and the next SCC is evaluated.

If the **SecurityContextConstraints.supplementalGroups** field has value **RunAsAny** and the pod specification omits the **Pod.spec.securityContext.supplementalGroups**, then this field is considered valid. Note that it is possible that during validation, other SCC settings will reject other pod fields and thus cause the pod to fail.

4.2.5.5.1. SCC Prioritization

SCCs have a priority field that affects the ordering when attempting to validate a request by the admission controller. A higher priority SCC is moved to the front of the set when sorting. When the complete set of available SCCs are determined they are ordered by:

1. Highest priority first, nil is considered a 0 priority
2. If priorities are equal, the SCCs will be sorted from most restrictive to least restrictive
3. If both priorities and restrictions are equal the SCCs will be sorted by name

By default, the anyuid SCC granted to cluster administrators is given priority in their SCC set. This allows cluster administrators to run pods as any user by without specifying a **RunAsUser** on the pod's **SecurityContext**. The administrator may still specify a **RunAsUser** if they wish.

4.2.5.5.2. Understanding Pre-allocated Values and Security Context Constraints

The admission controller is aware of certain conditions in the security context constraints that trigger it to look up pre-allocated values from a namespace and populate the security context constraint before processing the pod. Each SCC strategy is evaluated independently of other strategies, with the pre-allocated values (where allowed) for each policy aggregated with pod specification values to make the final values for the various IDs defined in the running pod.

The following SCCs cause the admission controller to look for pre-allocated values when no ranges are defined in the pod specification:

1. A **RunAsUser** strategy of **MustRunAsRange** with no minimum or maximum set. Admission looks for the **openshift.io/sa.scc.uid-range** annotation to populate range fields.
2. An **SELinuxContext** strategy of **MustRunAs** with no level set. Admission looks for the **openshift.io/sa.scc.mcs** annotation to populate the level.
3. A **FSGroup** strategy of **MustRunAs**. Admission looks for the **openshift.io/sa.scc.supplemental-groups** annotation.
4. A **SupplementalGroups** strategy of **MustRunAs**. Admission looks for the **openshift.io/sa.scc.supplemental-groups** annotation.

During the generation phase, the security context provider will default any values that are not specifically set in the pod. Defaulting is based on the strategy being used:

1. **RunAsAny** and **MustRunAsNonRoot** strategies do not provide default values. Thus, if the pod needs a field defined (for example, a group ID), this field must be defined inside the pod specification.
2. **MustRunAs** (single value) strategies provide a default value which is always used. As an example, for group IDs: even if the pod specification defines its own ID value, the namespace's default field will also appear in the pod's groups.
3. **MustRunAsRange** and **MustRunAs** (range-based) strategies provide the minimum value of the range. As with a single value **MustRunAs** strategy, the namespace's default value will

appear in the running pod. If a range-based strategy is configurable with multiple ranges, it will provide the minimum value of the first configured range.

NOTE

FSGroup and **SupplementalGroups** strategies fall back to the **openshift.io/sa.scc.uid-range** annotation if the **openshift.io/sa.scc.supplemental-groups** annotation does not exist on the namespace. If neither exist, the SCC will fail to create.

NOTE

By default, the annotation-based **FSGroup** strategy configures itself with a single range based on the minimum value for the annotation. For example, if your annotation reads **1/3**, the **FSGroup** strategy will configure itself with a minimum and maximum of **1**. If you want to allow more groups to be accepted for the **FSGroup** field, you can configure a custom SCC that does not use the annotation.

NOTE

The **openshift.io/sa.scc.supplemental-groups** annotation accepts a comma delimited list of blocks in the format of **<start>/<length** or **<start>-<end>**. The **openshift.io/sa.scc.uid-range** annotation accepts only a single block.

4.2.6. Determining What You Can Do as an Authenticated User

From within your OpenShift Container Platform project, you can determine what verbs you can perform against all namespace-scoped resources (including third-party resources). Run:

```
$ oc policy can-i --list --loglevel=8
```

The output will help you to determine what API request to make to gather the information.

To receive information back in a user-readable format, run:

```
$ oc policy can-i --list
```

The output will provide a full list.

To determine if you can perform specific verbs, run:

```
$ oc policy can-i <verb> <resource>
```

User scopes can provide more information about a given scope. For example:

```
$ oc policy can-i <verb> <resource> --scopes=user:info
```

4.3. PERSISTENT STORAGE

4.3.1. Overview

Managing storage is a distinct problem from managing compute resources. OpenShift Container Platform leverages the Kubernetes persistent volume (PV) framework to allow cluster administrators to provision persistent storage for a cluster. Using persistent volume claims (PVCs), developers can request PV resources without having specific knowledge of the underlying storage infrastructure.

PVCs are specific to a project and are created and used by developers as a means to use a PV. PV resources on their own are not scoped to any single project; they can be shared across the entire OpenShift Container Platform cluster and claimed from any project. After a PV is *bound* to a PVC, however, that PV cannot then be bound to additional PVCs. This has the effect of scoping a bound PV to a single namespace (that of the binding project).

PVs are defined by a **PersistentVolume** API object, which represents a piece of existing, networked storage in the cluster that was provisioned by the cluster administrator. It is a resource in the cluster just like a node is a cluster resource. PVs are volume plug-ins like **Volumes** but have a lifecycle that is independent of any individual pod that uses the PV. PV objects capture the details of the implementation of the storage, be that NFS, iSCSI, or a cloud-provider-specific storage system.

IMPORTANT

High availability of storage in the infrastructure is left to the underlying storage provider.

PVCs are defined by a **PersistentVolumeClaim** API object, which represents a request for storage by a developer. It is similar to a pod in that pods consume node resources and PVCs consume PV resources. For example, pods can request specific levels of resources (e.g., CPU and memory), while PVCs can request specific storage capacity and access modes (e.g, they can be mounted once read/write or many times read-only).

4.3.2. Lifecycle of a Volume and Claim

PVs are resources in the cluster. PVCs are requests for those resources and also act as claim checks to the resource. The interaction between PVs and PVCs have the following lifecycle.

4.3.2.1. Provisioning

In response to requests from a developer defined in a PVC, a cluster administrator configures one or more dynamic provisioners that provision storage and a matching PV.

Alternatively, a cluster administrator can create a number of PVs in advance that carry the details of the real storage that is available for use. PVs exist in the API and are available for use.

4.3.2.2. Binding

When you create a PVC, you request a specific amount of storage, specify the required access mode, and can create a storage class to describe and classify the storage. The control loop in the master watches for new PVCs and binds the new PVC to an appropriate PV. If an appropriate PV does not exist, a provisioner for the storage class creates one.

The PV volume might exceed your requested volume. This is especially true with manually provisioned PVs. To minimize the excess, OpenShift Container Platform binds to the smallest PV that matches all other criteria.

Claims remain unbound indefinitely if a matching volume does not exist or cannot be created with any available provisioner servicing a storage class. Claims are bound as matching volumes become available. For example, a cluster with many manually provisioned 50Gi volumes would not match a PVC requesting 100Gi. The PVC can be bound when a 100Gi PV is added to the cluster.

4.3.2.3. Using

Pods use claims as volumes. The cluster inspects the claim to find the bound volume and mounts that volume for a pod. For those volumes that support multiple access modes, you must specify which mode applies when you use the claim as a volume in a pod.

After you have a claim and that claim is bound, the bound PV belongs to you for as long as you need it. You can schedule pods and access claimed PVs by including **persistentVolumeClaim** in the pod's volumes block. See below for syntax details.

4.3.2.4. Persistent Volume Claim Protection

NOTE

PVC protection is an alpha feature and may change in a future release of OpenShift Container Platform.

The purpose of PVC protection is to ensure that PVCs in active use by a pod are not removed from the system, as this may result in data loss.

NOTE

A PVC is in active use by a pod when the pod status is **Pending**, and the pod is assigned to a node or the pod status is **Running**.

When the PVC protection feature is enabled, if you delete a PVC that is in active use by a pod, the PVC is not immediately removed. PVC removal is postponed until the PVC is no longer actively used by any pods.

You can see that a PVC is protected when the PVC's status is **Terminating** and the **Finalizers** list includes **kubernetes.io/pvc-protection**:

```
oc describe pvc hostpath
Name:          hostpath
Namespace:     default
StorageClass:  example-hostpath
Status:        Terminating
Volume:
Labels:        <none>
Annotations:   volume.beta.kubernetes.io/storage-class=example-hostpath
               volume.beta.kubernetes.io/storage-provisioner=example.com/hostpath
Finalizers:    [kubernetes.io/pvc-protection]
...
```

To enable PVC protection, see Configuring Persistent Volume Claim Protection .

4.3.2.5. Releasing

When you are done with a volume, you can delete the PVC object from the API, which allows reclamation of the resource. The volume is considered "released" when the claim is deleted, but it is not yet available for another claim. The previous claimant's data remains on the volume and must be handled according to policy.

4.3.2.6. Reclaiming

The reclaim policy of a **PersistentVolume** tells the cluster what to do with the volume after it is released. Volumes reclaim policy can either be **Retain**, **Recycle**, or **Delete**.

Retain reclaim policy allows manual reclamation of the resource for those volume plug-ins that support it. **Delete** reclaim policy deletes both the **PersistentVolume** object from OpenShift Container Platform and the associated storage asset in external infrastructure, such as AWS EBS, GCE PD, or Cinder volume.

NOTE

Dynamically provisioned volumes have a default **ReclaimPolicy** value of **Delete**. Manually provisioned volumes have a default **ReclaimPolicy** value of **Retain**.

4.3.2.6.1. Recycling

If supported by appropriate volume plug-in, recycling performs a basic scrub (**rm -rf /thevolume/***) on the volume and makes it available again for a new claim.

WARNING

The **recycle** reclaim policy is deprecated in favor of dynamic provisioning and it will be removed in future releases.

You can configure a custom recycler pod template using the controller manager command line arguments as described in the ControllerArguments section. The custom recycler pod template must contain a **volumes** specification, as shown in the example below:

```
apiVersion: v1
kind: Pod
metadata:
  name: pv-recycler-
  namespace: openshift-infra 1
spec:
  restartPolicy: Never
  serviceAccount: pv-recycler-controller 2
  volumes:
  - name: nfsvol
    nfs:
      server: any-server-it-will-be-replaced 3
      path: any-path-it-will-be-replaced 4
  containers:
  - name: pv-recycler
    image: "gcr.io/google_containers/busybox"
    command: ["/bin/sh", "-c", "test -e /scrub && rm -rf /scrub/..?* /scrub/.[!.]* /scrub/*  && test -z \"$(ls -A /scrub)\" || exit 1"]
    volumeMounts:
    - name: nfsvol
      mountPath: /scrub
```

1 Namespace where the recycler pod runs. **openshift-infra** is the recommended namespace, as it already has a **pv-recycler-controller** service account that can recycle volumes.

2 Name of service account that is allowed to mount NFS volumes. It must exist in the specified namespace. A **pv-recycler-controller** account is recommended, as it is automatically created in **openshift-infra** namespace and has all the required permissions.

3 4 The particular **server** and **path** values specified in the custom recycler pod template in the **volumes** part is replaced with the particular corresponding values from the PV that is being recycled.

4.3.3. Persistent Volumes

Each PV contains a **spec** and **status**, which is the specification and status of the volume.

Persistent Volume Object Definition

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv0003
spec:
  capacity:
    storage: 5Gi
  accessModes:
    - ReadWriteOnce
  persistentVolumeReclaimPolicy: Recycle
  nfs:
    path: /tmp
    server: 172.17.0.2
```

4.3.3.1. Types of Persistent Volumes

OpenShift Container Platform supports the following **PersistentVolume** plug-ins:

- NFS
- HostPath
- GlusterFS
- Ceph RBD
- OpenStack Cinder
- AWS Elastic Block Store (EBS)
- GCE Persistent Disk
- iSCSI
- Fibre Channel
- Azure Disk

- Azure File
- VMWare vSphere
- Local

4.3.3.2. Capacity

Generally, a PV has a specific storage capacity. This is set using the PV's **capacity** attribute.

Currently, storage capacity is the only resource that can be set or requested. Future attributes may include IOPS, throughput, and so on.

4.3.3.3. Access Modes

A **PersistentVolume** can be mounted on a host in any way supported by the resource provider. Providers will have different capabilities and each PV's access modes are set to the specific modes supported by that particular volume. For example, NFS can support multiple read/write clients, but a specific NFS PV might be exported on the server as read-only. Each PV gets its own set of access modes describing that specific PV's capabilities.

Claims are matched to volumes with similar access modes. The only two matching criteria are access modes and size. A claim's access modes represent a request. Therefore, you might be granted more, but never less. For example, if a claim requests RWO, but the only volume available is an NFS PV (RWO+ROX+RWX), then the claim would match NFS because it supports RWO.

Direct matches are always attempted first. The volume's modes must match or contain more modes than you requested. The size must be greater than or equal to what is expected. If two types of volumes (NFS and iSCSI, for example) both have the same set of access modes, then either of them can match a claim with those modes. There is no ordering between types of volumes and no way to choose one type over another.

All volumes with the same modes are grouped, then sorted by size (smallest to largest). The binder gets the group with matching modes and iterates over each (in size order) until one size matches.

The access modes are:

Access Mode	CLI Abbreviation	Description
ReadWriteOnce	**RWO**	The volume can be mounted as read-write by a single node.
ReadOnlyMany	**ROX**	The volume can be mounted read-only by many nodes.
ReadWriteMany	**RWX**	The volume can be mounted as read-write by many nodes.

IMPORTANT

A volume’s **AccessModes** are descriptors of the volume’s capabilities. They are not enforced constraints. The storage provider is responsible for runtime errors resulting from invalid use of the resource.

For example, Ceph offers **ReadWriteOnce** access mode. You must mark the claims as **read-only** if you want to use the volume’s ROX capability. Errors in the provider show up at runtime as mount errors.

iSCSI and Fibre Channel volumes do not have any fencing mechanisms yet. You must ensure the volumes are only used by one node at a time. In certain situations, such as draining a node, the volumes can be used simultaneously by two nodes. Before draining the node, first ensure the pods that use these volumes are deleted.

The table below lists the access modes supported by different persistent volumes:

Table 4.1. Supported Access Modes for Persistent Volumes

Volume Plug-in	ReadWriteOnce	ReadOnlyMany	ReadWriteMany
AWS EBS	▮	-	-
Azure File	▮	▮	▮
Azure Disk	▮	-	-
Ceph RBD	▮	▮	-
Fibre Channel	▮	▮	-
GCE Persistent Disk	▮	-	-
GlusterFS	▮	▮	▮
HostPath	▮	-	-
iSCSI	▮	▮	-
NFS	▮	▮	▮
Openstack Cinder	▮	-	-

Volume Plug-in	ReadWriteOnce	ReadOnlyMany	ReadWriteMany
VMWare vSphere	▮	-	-
Local	▮	-	-

NOTE

- If pods rely on AWS EBS, GCE Persistent Disks, or Openstack Cinder PVs, use a recreate deployment strategy

4.3.3.4. Reclaim Policy

The current reclaim policies are:

Reclaim Policy	Description
Retain	Manual reclamation
Recycle	Basic scrub (e.g, **rm -rf /<volume>/***)

NOTE

Currently, only NFS and HostPath support the 'Recycle' reclaim policy.

WARNING

The **recycle** reclaim policy is deprecated in favor of dynamic provisioning and it will be removed in future releases.

4.3.3.5. Phase

Volumes can be found in one of the following phases:

Phase	Description
Available	A free resource that is not yet bound to a claim.
Bound	The volume is bound to a claim.
Released	The claim was deleted, but the resource is not yet reclaimed by the cluster.

Phase	Description
Failed	The volume has failed its automatic reclamation.

The CLI shows the name of the PVC bound to the PV.

4.3.3.6. Mount Options

You can specify mount options while mounting a persistent volume by using the annotation **volume.beta.kubernetes.io/mount-options**.

For example:

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv0001
  annotations:
    volume.beta.kubernetes.io/mount-options: rw,nfsvers=4,noexec 1
spec:
  capacity:
    storage: 1Gi
  accessModes:
  - ReadWriteOnce
  nfs:
    path: /tmp
    server: 172.17.0.2
  persistentVolumeReclaimPolicy: Recycle
  claimRef:
    name: claim1
    namespace: default
```

1 Specified mount options are then used while mounting the persistent volume to the disk.

The following persistent volume types support mount options:

- NFS
- GlusterFS
- Ceph RBD
- OpenStack Cinder
- AWS Elastic Block Store (EBS)
- GCE Persistent Disk
- iSCSI
- Azure Disk
- Azure File

- VMWare vSphere

NOTE

Fibre Channel and HostPath persistent volumes do not support mount options.

4.3.4. Persistent Volume Claims

Each PVC contains a **spec** and **status**, which is the specification and status of the claim.

Persistent Volume Claim Object Definition

```
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: myclaim
spec:
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 8Gi
  storageClassName: gold
```

4.3.4.1. Storage Class

Claims can optionally request a specific storage class by specifying the storage class's name in the **storageClassName** attribute. Only PVs of the requested class, ones with the same **storageClassName** as the PVC, can be bound to the PVC. The cluster administrator can configure dynamic provisioners to service one or more storage classes. The cluster administrator can create a PV on demand that matches the specifications in the PVC.

The cluster administrator can also set a default storage class for all PVCs. When a default storage class is configured, the PVC must explicitly ask for **StorageClass** or **storageClassName** annotations set to **""** to be bound to a PV without a storage class.

4.3.4.2. Access Modes

Claims use the same conventions as volumes when requesting storage with specific access modes.

4.3.4.3. Resources

Claims, like pods, can request specific quantities of a resource. In this case, the request is for storage. The same resource model applies to both volumes and claims.

4.3.4.4. Claims As Volumes

Pods access storage by using the claim as a volume. Claims must exist in the same namespace as the pod using the claim. The cluster finds the claim in the pod's namespace and uses it to get the **PersistentVolume** backing the claim. The volume is then mounted to the host and into the pod:

```
kind: Pod
apiVersion: v1
```

```
metadata:
  name: mypod
spec:
  containers:
    - name: myfrontend
      image: dockerfile/nginx
      volumeMounts:
      - mountPath: "/var/www/html"
        name: mypd
  volumes:
    - name: mypd
      persistentVolumeClaim:
        claimName: myclaim
```

4.3.5. Block Volume Support

IMPORTANT

Block Volume Support is a Technology Preview feature and it is only available for manually provisioned persistent volumes. Technology Preview features are not supported with Red Hat production service level agreements (SLAs), might not be functionally complete, and Red Hat does not recommend to use them for 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 on Red Hat Technology Preview features support scope, see https://access.redhat.com/support/offerings/techpreview/.

You can statically provision raw block volumes by including some new API fields in your PV and PVC specifications.

To use block volume, you must first enable the **BlockVolume** feature gate. To enable the feature gates for master(s), add **feature-gates** to **apiServerArguments** and **controllerArguments**. To enable the feature gates for node(s), add **feature-gates** to **kubeletArguments**. For example:

```
kubeletArguments:
   feature-gates:
     - BlockVolume=true
```

For more information, see Configuring for Local Volume.

Example Persistent Volume

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: block-pv
spec:
  capacity:
    storage: 10Gi
  accessModes:
    - ReadWriteOnce
  volumeMode: Block 1
  persistentVolumeReclaimPolicy: Retain
  fc:
    targetWWNs: ["50060e801049cfd1"]
    lun: 0
    readOnly: false
```

① **volumeMode** field indicating that this PV is a raw block volume.

Example Persistent Volume Claim

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: block-pvc
spec:
  accessModes:
    - ReadWriteOnce
  volumeMode: Block ①
  resources:
    requests:
      storage: 10Gi
```

① **volumeMode** field indicating that a raw block persistent volume is requested.

Example Pod Specification

```
apiVersion: v1
kind: Pod
metadata:
  name: pod-with-block-volume
spec:
  containers:
    - name: fc-container
      image: fedora:26
      command: ["/bin/sh", "-c"]
      args: [ "tail -f /dev/null" ]
      volumeDevices: ①
        - name: data
          devicePath: /dev/xvda ②
  volumes:
    - name: data
      persistentVolumeClaim:
        claimName: block-pvc ③
```

① **volumeDevices** (similar to **volumeMounts**) is used for block devices and can only be used with **PersistentVolumeClaim** sources.

② **devicePath** (similar to **mountPath**) represents the path to the physical device.

③ The volume source must be of type **persistentVolumeClaim** and should match the name of the PVC as expected.

Table 4.2. Accepted Values for VolumeMode

Value	Default
Filesystem	Yes
Block	No

Table 4.3. Binding Scenarios for Block Volumes

PV VolumeMode	PVC VolumeMode	Binding Result
Filesystem	Filesystem	Bind
Unspecified	Unspecified	Bind
Filesystem	Unspecified	Bind
Unspecified	Filesystem	Bind
Block	Block	Bind
Unspecified	Block	No Bind
Block	Unspecified	No Bind
Filesystem	Block	No Bind
Block	Filesystem	No Bind

IMPORTANT

Unspecified values result in the default value of **Filesystem**.

4.4. SOURCE CONTROL MANAGEMENT

OpenShift Container Platform takes advantage of preexisting source control management (SCM) systems hosted either internally (such as an in-house Git server) or externally (for example, on GitHub, Bitbucket, etc.). Currently, OpenShift Container Platform only supports Git solutions.

SCM integration is tightly coupled with builds, the two points being:

- Creating a **BuildConfig** using a repository, which allows building your application inside of OpenShift Container Platform. You can create a **BuildConfig**manually or let OpenShift Container Platform create it automatically by inspecting your repository.
- Triggering a build upon repository changes.

4.5. ADMISSION CONTROLLERS

4.5.1. Overview

Admission control plug-ins intercept requests to the master API prior to persistence of a resource, but after the request is authenticated and authorized.

Each admission control plug-in is run in sequence before a request is accepted into the cluster. If any plug-in in the sequence rejects the request, the entire request is rejected immediately, and an error is returned to the end-user.

Admission control plug-ins may modify the incoming object in some cases to apply system configured defaults. In addition, admission control plug-ins may modify related resources as part of request processing to do things such as incrementing quota usage.

> **WARNING**
>
> The OpenShift Container Platform master has a default list of plug-ins that are enabled by default for each type of resource (Kubernetes and OpenShift Container Platform). These are required for the proper functioning of the master. Modifying these lists is not recommended unless you strictly know what you are doing. Future versions of the product may use a different set of plug-ins and may change their ordering. If you do override the default list of plug-ins in the master configuration file, you are responsible for updating it to reflect requirements of newer versions of the OpenShift Container Platform master.

4.5.2. General Admission Rules

Starting in 3.3, OpenShift Container Platform uses a single admission chain for Kubernetes and OpenShift Container Platform resources. This changed from 3.2, and before where we had separate admission chains. This means that the top-level **admissionConfig.pluginConfig** element can now contain the admission plug-in configuration, which used to be contained in **kubernetesMasterConfig.admissionConfig.pluginConfig**.

The **kubernetesMasterConfig.admissionConfig.pluginConfig** should be moved and merged into **admissionConfig.pluginConfig**.

Also, starting in 3.3, all the supported admission plug-ins are ordered in the single chain for you. You should no longer set **admissionConfig.pluginOrderOverride** or the **kubernetesMasterConfig.admissionConfig.pluginOrderOverride**. Instead, you should enable plug-ins that are off by default by either adding their plug-in-specific configuration, or adding a **DefaultAdmissionConfig** stanza like this:

```
admissionConfig:
  pluginConfig:
    AlwaysPullImages: 1
      configuration:
        kind: DefaultAdmissionConfig
        apiVersion: v1
        disable: false 2
```

1 Admission plug-in name.

2 Indicates that a plug-in should be enabled. It is optional and shown here only for reference.

Setting **disable** to **true** will disable an admission plug-in that defaults to on.

WARNING

Admission plug-ins are commonly used to help enforce security on the API server. Be careful when disabling them.

NOTE

If you were previously using **admissionConfig** elements that cannot be safely combined into a single admission chain, you will get a warning in your API server logs and your API server will start with two separate admission chains for legacy compatibility. Update your **admissionConfig** to resolve the warning.

4.5.3. Customizable Admission Plug-ins

Cluster administrators can configure some admission control plug-ins to control certain behavior, such as:

- Limiting Number of Self-Provisioned Projects Per User
- Configuring Global Build Defaults and Overrides
- Controlling Pod Placement
- Managing Role Bindings

4.5.4. Admission Controllers Using Containers

Admission controllers using containers also support init containers.

4.6. CUSTOM ADMISSION CONTROLLERS

4.6.1. Overview

In addition to the default admission controllers, you can use *admission webhooks* as part of the admission chain.

Admission webhooks call webhook servers to either mutate pods upon creation, such as to inject labels, or to validate specific aspects of the pod configuration during the admission process.

Admission webhooks intercept requests to the master API prior to the persistence of a resource, but after the request is authenticated and authorized.

4.6.2. Admission Webhooks

In OpenShift Container Platform you can use admission webhook objects that call webhook servers during the API admission chain.

There are two types of admission webhook objects you can configure:

- Mutating admission webhooks allow for the use of mutating webhooks to modify resource content before it is persisted.
- Validating admission webhooks allow for the use of validating webhooks to enforce custom admission policies.

Configuring the webhooks and external webhook servers is beyond the scope of this document. However, the webhooks must adhere to an interface in order to work properly with OpenShift Container Platform.

IMPORTANT

Admission webhooks is a Technology Preview feature only. Technology Preview features are not supported with Red Hat production service level agreements (SLAs), might not be functionally complete, and Red Hat does not recommend to use them for 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 on Red Hat Technology Preview features support scope, see https://access.redhat.com/support/offerings/techpreview/.

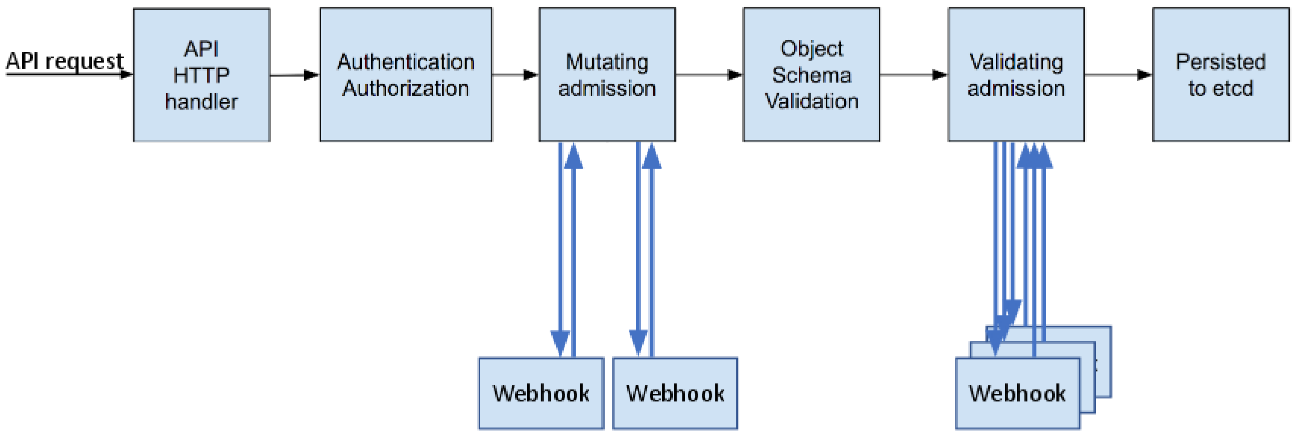
When an object is instantiated, OpenShift Container Platform makes an API call to admit the object. During the admission process, a *mutating admission controller* can invoke webhooks to perform tasks, such as injecting affinity labels. At the end of the admissions process, a *validating admission controller* can invoke webhooks to make sure the object is configured properly, such as verifying affinity labels. If the validation passes, OpenShift Container Platform schedules the object as configured.

When the API request comes in, the mutating or validating admission controller uses the list of external webhooks in the configuration and calls them in parallel:

- If **all** of the webhooks approve the request, the admission chain continues.
- If **any** of the webhooks deny the request, the admission request is denied, and the reason for doing so is based on the *first* webhook denial reason.
  If more than one webhook denies the admission request, only the first will be returned to the user.
- If there is an error encountered when calling a webhook, that request is either denied or the webhook is ignored.

The communication between the admission controller and the webhook server needs to be secured using TLS. Generate a CA certificate and use the certificate to sign the server certificate used by your webhook server. The PEM-formatted CA certificate is supplied to the admission controller using a mechanism, such as Service Serving Certificate Secrets.

The following diagram illustrates this process with two admission webhooks that call multiple webhooks.



A simple example use case for admission webhooks is syntactical validation of resources. For example, you have an infrastructure that requires all pods to have a common set of labels, and you do not want any pod to be persisted if the pod does not have those labels. You could write a webhook to inject these labels and another webhook to verify that the labels are present. The OpenShift Container Platform will then schedule pod that have the labels and pass validation and reject pods that do not pass due to missing labels.

Some common use-cases include:

- Mutating resources to inject side-car containers into pods.
- Restricting projects to block some resources from a project.
- Custom resource validation to perform complex validation on dependent fields.

4.6.2.1. Types of Admission Webhooks

Cluster administrators can include *mutating admission webhooks* or *validating admission webhooks* in the admission chain of the API server.

Mutating admission webhooks are invoked during the mutation phase of the admission process, which allows modification of the resource content before it is persisted. One example of a mutating admission webhook is the Pod Node Selector feature, which uses an annotation on a namespace to find a label selector and add it to the pod specification.

Sample mutating admission webhook configuration:

```
apiVersion: admissionregistration.k8s.io/v1beta1
  kind: MutatingWebhookConfiguration 1
  metadata:
    name: <controller_name> 2
  webhooks:
  - name: <webhook_name> 3
    clientConfig: 4
      service:
        namespace: 5
        name: 6
       path: <webhook_url> 7
      caBundle: <cert> 8
```

```
rules: 9
- operations: 10
  - <operation>
  apiGroups:
  - ""
  apiVersions:
  - "*"
  resources:
  - <resource>
failurePolicy: <policy> 11
```

1. Specifies a mutating admission webhook configuration.
2. The name for the admission webhook object.
3. The name of the webhook to call.
4. Information about how to connect to, trust, and send data to the webhook server.
5. The project where the front-end service is created.
6. The name of the front-end service.
7. The webhook URL used for admission requests.
8. A PEM-encoded CA certificate that signs the server certificate used by the webhook server.
9. Rules that define when the API server should use this controller.
10. The operation(s) that triggers the API server to call this controller:
    - create
    - update
    - delete
    - connect
11. Specifies how the policy should proceed if the webhook admission server is unavailable. Either **Ignore** (allow/fail open) or **Fail** (block/fail closed).

Validating admission webhooks are invoked during the validation phase of the admission process. This phase allows the enforcement of invariants on particular API resources to ensure that the resource does not change again. The Pod Node Selector is also an example of a validation admission, by ensuring that all **nodeSelector** fields are constrained by the node selector restrictions on the project.

Sample validating admission webhook configuration:

```
apiVersion: admissionregistration.k8s.io/v1beta1
 kind: ValidatingWebhookConfiguration 1
 metadata:
   name: <controller_name> 2
 webhooks:
 - name: <webhook_name> 3
```

```
  clientConfig: 4
    service:
      namespace: default 5
      name: kubernetes 6
     path: <webhook_url> 7
    caBundle: <cert> 8
  rules: 9
  - operations: 10
    - <operation>
    apiGroups:
    - ""
    apiVersions:
    - "*"
    resources:
    - <resource>
  failurePolicy: <policy> 11
```

1 Specifies a validating admission webhook configuration.

2 The name for the webhook admission object.

3 The name of the webhook to call.

4 Information about how to connect to, trust, and send data to the webhook server.

5 The project where the front-end service is created.

6 The name of the front-end service.

7 The webhook URL used for admission requests.

8 A PEM-encoded CA certificate that signs the server certificate used by the webhook server.

9 Rules that define when the API server should use this controller.

10 The operation that triggers the API server to call this controller.

- create
- update
- delete
- connect

11 Specifies how the policy should proceed if the webhook admission server is unavailable. Either **Ignore** (allow/fail open) or **Fail** (block/fail closed).

NOTE

Fail open can result in unpredictable behavior for all clients.

4.6.2.2. Create the Admission Webhook

First deploy the external webhook server and ensure it is working properly. Otherwise, depending whether the webhook is configured as **fail open** or **fail closed**, operations will be unconditionally accepted or rejected.

1. Configure a mutating or validating admission webhook object in a YAML file.

2. Run the following command to create the object:

   ```
   oc create -f <file-name>.yaml
   ```

   After you create the admission webhook object, OpenShift Container Platform takes a few seconds to honor the new configuration.

3. Create a front-end service for the admission webhook:

   ```
   apiVersion: v1
   kind: Service
   metadata:
     labels:
       role: webhook 1
     name: <name>
   spec:
     selector:
      role: webhook 2
   ```

   1 2 Free-form label to trigger the webhook.

4. Run the following command to create the object:

   ```
   oc create -f <file-name>.yaml
   ```

5. Add the admission webhook name to pods you want controlled by the webhook:

   ```
   apiVersion: v1
   kind: Pod
   metadata:
     labels:
       role: webhook 1
     name: <name>
   spec:
     containers:
       - name: <name>
         image: myrepo/myimage:latest
         imagePullPolicy: <policy>
         ports:
          - containerPort: 8000
   ```

   1 Label to trigger the webhook.

NOTE

See the kubernetes-namespace-reservation projects for an end-to-end example of how to build your own secure and portable webhook admission server and generic-admission-apiserver for the library.

4.6.2.3. Admission Webhook Example

The following is an example admission webhook that will not allow namespace creation if the namespace is reserved:

```
apiVersion: admissionregistration.k8s.io/v1beta1
  kind: ValidatingWebhookConfiguration
  metadata:
    name: namespacereservations.admission.online.openshift.io
  webhooks:
  - name: namespacereservations.admission.online.openshift.io
    clientConfig:
      service:
        namespace: default
        name: webhooks
       path: /apis/admission.online.openshift.io/v1beta1/namespacereservations
      caBundle: KUBE_CA_HERE
    rules:
    - operations:
      - CREATE
      apiGroups:
      - ""
      apiVersions:
      - "b1"
      resources:
      - namespaces
    failurePolicy: Ignore
```

The following is an example pod that will be evaluated by the admission webhook named *webhook*:

```
apiVersion: v1
kind: Pod
metadata:
  labels:
    role: webhook
  name: webhook
spec:
  containers:
    - name: webhook
      image: myrepo/myimage:latest
      imagePullPolicy: IfNotPresent
      ports:
- containerPort: 8000
```

The following is the front-end service for the webhook:

```
apiVersion: v1
kind: Service
metadata:
  labels:
    role: webhook
  name: webhook
spec:
  ports:
    - port: 443
      targetPort: 8000
  selector:
role: webhook
```

4.7. OTHER API OBJECTS

4.7.1. LimitRange

A limit range provides a mechanism to enforce min/max limits placed on resources in a Kubernetes namespace.

By adding a limit range to your namespace, you can enforce the minimum and maximum amount of CPU and Memory consumed by an individual pod or container.

4.7.2. ResourceQuota

Kubernetes can limit both the number of objects created in a namespace, and the total amount of resources requested across objects in a namespace. This facilitates sharing of a single Kubernetes cluster by several teams, each in a namespace, as a mechanism of preventing one team from starving another team of cluster resources.

See Cluster Administration for more information on **ResourceQuota**.

4.7.3. Resource

A Kubernetes **Resource** is something that can be requested by, allocated to, or consumed by a pod or container. Examples include memory (RAM), CPU, disk-time, and network bandwidth.

See the Developer Guide for more information.

4.7.4. Secret

Secrets are storage for sensitive information, such as keys, passwords, and certificates. They are accessible by the intended pod(s), but held separately from their definitions.

4.7.5. PersistentVolume

A persistent volume is an object (**PersistentVolume**) in the infrastructure provisioned by the cluster administrator. Persistent volumes provide durable storage for stateful applications.

4.7.6. PersistentVolumeClaim

A **PersistentVolumeClaim** object is a request for storage by a pod author . Kubernetes matches the claim against the pool of available volumes and binds them together. The claim is then used as a volume by a pod. Kubernetes makes sure the volume is available on the same node as the pod that requires it.

4.7.6.1. Custom Resources

A *custom resource* is an extension of the Kubernetes API that extends the API or allows you to introduce your own API into a project or a cluster.

> See link:https://access.redhat.com/documentation/en-us/openshift_container_platform/3.9/html-single/cluster_administration/#admin-guide-custom-resources[Extend the Kubernetes API with Custom Resources].

4.7.7. OAuth Objects

4.7.7.1. OAuthClient

An **OAuthClient** represents an OAuth client, as described in RFC 6749, section 2 .

The following **OAuthClient** objects are automatically created:

openshift-web-console	Client used to request tokens for the web console
openshift-browser-client	Client used to request tokens at /oauth/token/request with a user-agent that can handle interactive logins
openshift-challenging-client	Client used to request tokens with a user-agent that can handle WWW-Authenticate challenges

OAuthClient Object Definition

```
kind: "OAuthClient"
accessTokenMaxAgeSeconds: null 1
apiVersion: "oauth.openshift.io/v1"
metadata:
  name: "openshift-web-console" 2
  selflink: "/oapi/v1/oAuthClients/openshift-web-console"
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:01Z"
respondWithChallenges: false 3
secret: "45e27750-a8aa-11e4-b2ea-3c970e4b7ffe" 4
redirectURIs:
  - "https://localhost:8443" 5
```

1. The lifetime of access tokens in seconds (see the description below).
2. The **name** is used as the **client_id** parameter in OAuth requests.
3. When **respondWithChallenges** is set to **true**, unauthenticated requests to **/oauth/authorize** will result in **WWW-Authenticate** challenges, if supported by the configured authentication methods.
4. The value in the **secret** parameter is used as the **client_secret** parameter in an authorization code flow.

5 One or more absolute URIs can be placed in the **redirectURIs** section. The **redirect_uri** parameter sent with authorization requests must be prefixed by one of the specified **redirectURIs**.

The **accessTokenMaxAgeSeconds** value overrides the default **accessTokenMaxAgeSeconds** value in the master configuration file for individual OAuth clients. Setting this value for a client allows long-lived access tokens for that client without affecting the lifetime of other clients.

- If **null**, the default value in the master configuration file is used.
- If set to **0**, the token will not expire.
- If set to a value greater than **0**, tokens issued for that client are given the specified expiration time. For example, **accessTokenMaxAgeSeconds: 172800** would cause the token to expire 48 hours after being issued.

4.7.7.2. OAuthClientAuthorization

An **OAuthClientAuthorization** represents an approval by a **User** for a particular **OAuthClient** to be given an **OAuthAccessToken** with particular scopes.

Creation of **OAuthClientAuthorization** objects is done during an authorization request to the **OAuth** server.

OAuthClientAuthorization Object Definition

```
kind: "OAuthClientAuthorization"
apiVersion: "oauth.openshift.io/v1"
metadata:
  name: "bob:openshift-web-console"
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:01-00:00"
clientName: "openshift-web-console"
userName: "bob"
userUID: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe"
scopes: []
```

4.7.7.3. OAuthAuthorizeToken

An **OAuthAuthorizeToken** represents an **OAuth** authorization code, as described in RFC 6749, section 1.3.1.

An **OAuthAuthorizeToken** is created by a request to the /oauth/authorize endpoint, as described in RFC 6749, section 4.1.1.

An **OAuthAuthorizeToken** can then be used to obtain an **OAuthAccessToken** with a request to the **/oauth/token** endpoint, as described in RFC 6749, section 4.1.3.

OAuthAuthorizeToken Object Definition

```
kind: "OAuthAuthorizeToken"
apiVersion: "oauth.openshift.io/v1"
metadata:
  name: "MDAwYjM5YjMtMzM1MC00NDY4LTkxODItOTA2OTE2YzE0M2Fj" 1
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:01-00:00"
clientName: "openshift-web-console" 2
expiresIn: 300 3
scopes: []
redirectURI: "https://localhost:8443/console/oauth" 4
userName: "bob" 5
userUID: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe" 6
```

1 **name** represents the token name, used as an authorization code to exchange for an OAuthAccessToken.

2 The **clientName** value is the OAuthClient that requested this token.

3 The **expiresIn** value is the expiration in seconds from the creationTimestamp.

4 The **redirectURI** value is the location where the user was redirected to during the authorization flow that resulted in this token.

5 **userName** represents the name of the User this token allows obtaining an OAuthAccessToken for.

6 **userUID** represents the UID of the User this token allows obtaining an OAuthAccessToken for.

4.7.7.4. OAuthAccessToken

An **OAuthAccessToken** represents an **OAuth** access token, as described in RFC 6749, section 1.4 .

An **OAuthAccessToken** is created by a request to the **/oauth/token** endpoint, as described in RFC 6749, section 4.1.3.

Access tokens are used as bearer tokens to authenticate to the API.

OAuthAccessToken Object Definition

```
kind: "OAuthAccessToken"
apiVersion: "oauth.openshift.io/v1"
metadata:
  name: "ODliOGE5ZmMtYzczYi00Nzk1LTg4MGEtNzQyZmUxZmUwY2Vh" 1
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:02-00:00"
clientName: "openshift-web-console" 2
expiresIn: 86400 3
scopes: []
redirectURI: "https://localhost:8443/console/oauth" 4
userName: "bob" 5
userUID: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe" 6
authorizeToken: "MDAwYjM5YjMtMzM1MC00NDY4LTkxODItOTA2OTE2YzE0M2Fj" 7
```

1 **name** is the token name, which is used as a bearer token to authenticate to the API.

2 The **clientName** value is the OAuthClient that requested this token.

3 The **expiresIn** value is the expiration in seconds from the creationTimestamp.

4 The **redirectURI** is where the user was redirected to during the authorization flow that resulted in this token.

5 **userName** represents the User this token allows authentication as.

6 **userUID** represents the User this token allows authentication as.

7 **authorizeToken** is the name of the OAuthAuthorizationToken used to obtain this token, if any.

4.7.8. User Objects

4.7.8.1. Identity

When a user logs into OpenShift Container Platform, they do so using a configured identity provider. This determines the user's identity, and provides that information to OpenShift Container Platform.

OpenShift Container Platform then looks for a **UserIdentityMapping** for that **Identity**:

NOTE

If the identity provider is configured with the **lookup** mapping method, for example, if you are using an external LDAP system, this automatic mapping is not performed. You must create the mapping manually. For more information, see Lookup Mapping Method.

- If the **Identity** already exists, but is not mapped to a **User**, login fails.
- If the **Identity** already exists, and is mapped to a **User**, the user is given an **OAuthAccessToken** for the mapped **User**.
- If the **Identity** does not exist, an **Identity**, **User**, and **UserIdentityMapping** are created, and the user is given an **OAuthAccessToken** for the mapped **User**.

Identity Object Definition

```
kind: "Identity"
apiVersion: "user.openshift.io/v1"
metadata:
  name: "anypassword:bob" 1
  uid: "9316ebad-0fde-11e5-97a1-3c970e4b7ffe"
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:01-00:00"
providerName: "anypassword" 2
providerUserName: "bob" 3
user:
  name: "bob" 4
  uid: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe" 5
```

1 The identity name must be in the form providerName:providerUserName.

2 **providerName** is the name of the identity provider.

3 **providerUserName** is the name that uniquely represents this identity in the scope of the identity provider.

4 The **name** in the **user** parameter is the name of the user this identity maps to.

5 The **uid** represents the UID of the user this identity maps to.

4.7.8.2. User

A **User** represents an actor in the system. Users are granted permissions by adding roles to users or to their groups.

User objects are created automatically on first login, or can be created via the API.

NOTE

OpenShift Container Platform user names containing /, **:**, and **%** are not supported.

User Object Definition

```
kind: "User"
apiVersion: "user.openshift.io/v1"
metadata:
  name: "bob" 1
  uid: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe"
  resourceVersion: "1"
  creationTimestamp: "2015-01-01T01:01:01-00:00"
identities:
  - "anypassword:bob" 2
fullName: "Bob User" 3
```

1 **name** is the user name used when adding roles to a user.

2 The values in **identities** are Identity objects that map to this user. May be **null** or empty for users that cannot log in.

3 The **fullName** value is an optional display name of user.

4.7.8.3. UserIdentityMapping

A **UserIdentityMapping** maps an **Identity** to a **User**.

Creating, updating, or deleting a **UserIdentityMapping** modifies the corresponding fields in the **Identity** and **User** objects.

An **Identity** can only map to a single **User**, so logging in as a particular identity unambiguously determines the **User**.

A **User** can have multiple identities mapped to it. This allows multiple login methods to identify the same **User**.

UserIdentityMapping Object Definition

```
kind: "UserIdentityMapping"
apiVersion: "user.openshift.io/v1"
```

```
metadata:
  name: "anypassword:bob" ❶
  uid: "9316ebad-0fde-11e5-97a1-3c970e4b7ffe"
  resourceVersion: "1"
identity:
  name: "anypassword:bob"
  uid: "9316ebad-0fde-11e5-97a1-3c970e4b7ffe"
user:
  name: "bob"
  uid: "9311ac33-0fde-11e5-97a1-3c970e4b7ffe"
```

❶ **UserIdentityMapping** name matches the mapped **Identity** name

4.7.8.4. Group

A **Group** represents a list of users in the system. Groups are granted permissions by adding roles to users or to their groups.

Group Object Definition

```
kind: "Group"
apiVersion: "user.openshift.io/v1"
metadata:
  name: "developers" ❶
  creationTimestamp: "2015-01-01T01:01:01-00:00"
users:
  - "bob" ❷
```

❶ **name** is the group name used when adding roles to a group.

❷ The values in **users** are the names of User objects that are members of this group.

CHAPTER 5. NETWORKING

5.1. NETWORKING

5.1.1. Overview

Kubernetes ensures that pods are able to network with each other, and allocates each pod an IP address from an internal network. This ensures all containers within the pod behave as if they were on the same host. Giving each pod its own IP address means that pods can be treated like physical hosts or virtual machines in terms of port allocation, networking, naming, service discovery, load balancing, application configuration, and migration.

Creating links between pods is unnecessary, and it is not recommended that your pods talk to one another directly using the IP address. Instead, it is recommended that you create a service, then interact with the service.

5.1.2. OpenShift Container Platform DNS

If you are running multiple services, such as frontend and backend services for use with multiple pods, in order for the frontend pods to communicate with the backend services, environment variables are created for user names, service IPs, and more. If the service is deleted and recreated, a new IP address can be assigned to the service, and requires the frontend pods to be recreated in order to pick up the updated values for the service IP environment variable. Additionally, the backend service has to be created before any of the frontend pods to ensure that the service IP is generated properly, and that it can be provided to the frontend pods as an environment variable.

For this reason, OpenShift Container Platform has a built-in DNS so that the services can be reached by the service DNS as well as the service IP/port. OpenShift Container Platform supports split DNS by running SkyDNS on the master that answers DNS queries for services. The master listens to port 53 by default.

When the node starts, the following message indicates the Kubelet is correctly resolved to the master:

```
0308 19:51:03.118430    4484 node.go:197] Started Kubelet for node
openshiftdev.local, server at 0.0.0.0:10250
I0308 19:51:03.118459    4484 node.go:199]   Kubelet is setting 10.0.2.15 as a
DNS nameserver for domain "local"
```

If the second message does not appear, the Kubernetes service may not be available.

On a node host, each container’s nameserver has the master name added to the front, and the default search domain for the container will be **.*<pod_namespace>*.cluster.local**. The container will then direct any nameserver queries to the master before any other nameservers on the node, which is the default behavior for Docker-formatted containers. The master will answer queries on the **.cluster.local** domain that have the following form:

Table 5.1. DNS Example Names

Object Type	Example
Default	<pod_namespace>.cluster.local

Object Type	Example
Services	<service>.<pod_namespace>.svc.cluster.local
Endpoints	<name>.<namespace>.endpoints.cluster.local

This prevents having to restart frontend pods in order to pick up new services, which would create a new IP for the service. This also removes the need to use environment variables, because pods can use the service DNS. Also, as the DNS does not change, you can reference database services as **db.local** in configuration files. Wildcard lookups are also supported, because any lookups resolve to the service IP, and removes the need to create the backend service before any of the frontend pods, since the service name (and hence DNS) is established upfront.

This DNS structure also covers headless services, where a portal IP is not assigned to the service and the kube-proxy does not load-balance or provide routing for its endpoints. Service DNS can still be used and responds with multiple A records, one for each pod of the service, allowing the client to round-robin between each pod.

5.2. OPENSHIFT SDN

5.2.1. Overview

OpenShift Container Platform uses a software-defined networking (SDN) approach to provide a unified cluster network that enables communication between pods across the OpenShift Container Platform cluster. This pod network is established and maintained by the OpenShift SDN, which configures an overlay network using Open vSwitch (OVS).

OpenShift SDN provides three SDN plug-ins for configuring the pod network:

- The **ovs-subnet** plug-in is the original plug-in, which provides a "flat" pod network where every pod can communicate with every other pod and service.

- The **ovs-multitenant** plug-in provides project-level isolation for pods and services. Each project receives a unique Virtual Network ID (VNID) that identifies traffic from pods assigned to the project. Pods from different projects cannot send packets to or receive packets from pods and services of a different project.
  However, projects that receive VNID 0 are more privileged in that they are allowed to communicate with all other pods, and all other pods can communicate with them. In OpenShift Container Platform clusters, the **default** project has VNID 0. This facilitates certain services, such as the load balancer, to communicate with all other pods in the cluster and vice versa.

- The **ovs-networkpolicy** plug-in allows project administrators to configure their own isolation policies using NetworkPolicy objects.

NOTE

Information on configuring the SDN on masters and nodes is available in Configuring the SDN.

5.2.2. Design on Masters

On an OpenShift Container Platform master, OpenShift SDN maintains a registry of nodes, stored in

etcd. When the system administrator registers a node, OpenShift SDN allocates an unused subnet from the cluster network and stores this subnet in the registry. When a node is deleted, OpenShift SDN deletes the subnet from the registry and considers the subnet available to be allocated again.

In the default configuration, the cluster network is the **10.128.0.0/14** network (i.e. **10.128.0.0 - 10.131.255.255**), and nodes are allocated **/23** subnets (i.e., **10.128.0.0/23**, **10.128.2.0/23**, **10.128.4.0/23**, and so on). This means that the cluster network has 512 subnets available to assign to nodes, and a given node is allocated 510 addresses that it can assign to the containers running on it. The size and address range of the cluster network are configurable, as is the host subnet size.

NOTE

If the subnet extends into the next higher octet, it is rotated so that the subnet bits with 0s in the shared octet are allocated first. For example, if the network is 10.1.0.0/16, with **hostsubnetlength=6**, then the subnet of 10.1.0.0/26 and 10.1.1.0/26, through to 10.1.255.0/26 are allocated before 10.1.0.64/26, 10.1.1.64/26 are filled. This ensures that the subnet is easier to follow.

Note that the OpenShift SDN on a master does not configure the local (master) host to have access to any cluster network. Consequently, a master host does not have access to pods via the cluster network, unless it is also running as a node.

When using the **ovs-multitenant** plug-in, the OpenShift SDN master also watches for the creation and deletion of projects, and assigns VXLAN VNIDs to them, which are later used by the nodes to isolate traffic correctly.

5.2.3. Design on Nodes

On a node, OpenShift SDN first registers the local host with the SDN master in the aforementioned registry so that the master allocates a subnet to the node.

Next, OpenShift SDN creates and configures three network devices:

- **br0**: the OVS bridge device that pod containers will be attached to. OpenShift SDN also configures a set of non-subnet-specific flow rules on this bridge.

- **tun0**: an OVS internal port (port 2 on **br0**). This gets assigned the cluster subnet gateway address, and is used for external network access. OpenShift SDN configures **netfilter** and routing rules to enable access from the cluster subnet to the external network via NAT.

- **vxlan_sys_4789**: The OVS VXLAN device (port 1 on **br0**), which provides access to containers on remote nodes. Referred to as **vxlan0** in the OVS rules.

Each time a pod is started on the host, OpenShift SDN:

1. assigns the pod a free IP address from the node’s cluster subnet.

2. attaches the host side of the pod’s veth interface pair to the OVS bridge **br0**.

3. adds OpenFlow rules to the OVS database to route traffic addressed to the new pod to the correct OVS port.

4. in the case of the **ovs-multitenant** plug-in, adds OpenFlow rules to tag traffic coming from the pod with the pod’s VNID, and to allow traffic into the pod if the traffic’s VNID matches the pod’s VNID (or is the privileged VNID 0). Non-matching traffic is filtered out by a generic rule.

OpenShift SDN nodes also watch for subnet updates from the SDN master. When a new subnet is added, the node adds OpenFlow rules on **br0** so that packets with a destination IP address in the remote subnet go to **vxlan0** (port 1 on **br0**) and thus out onto the network. The **ovs-subnet** plug-in sends all packets across the VXLAN with VNID 0, but the **ovs-multitenant** plug-in uses the appropriate VNID for the source container.

5.2.4. Packet Flow

Suppose you have two containers, A and B, where the peer virtual Ethernet device for container A's **eth0** is named **vethA** and the peer for container B's **eth0** is named **vethB**.

NOTE

If the Docker service's use of peer virtual Ethernet devices is not already familiar to you, see Docker's advanced networking documentation .

Now suppose first that container A is on the local host and container B is also on the local host. Then the flow of packets from container A to container B is as follows:

eth0 *(in A's netns) →* ***vethA*** *→* ***br0*** *→* ***vethB*** *→* ***eth0*** *(in B's netns)*

Next, suppose instead that container A is on the local host and container B is on a remote host on the cluster network. Then the flow of packets from container A to container B is as follows:

eth0 *(in A's netns) →* ***vethA*** *→* ***br0*** *→* ***vxlan0*** *→ network* [1] *→* ***vxlan0*** *→* ***br0*** *→* ***vethB*** *→* ***eth0*** *(in B's netns)*

Finally, if container A connects to an external host, the traffic looks like:

eth0 *(in A's netns) →* ***vethA*** *→* ***br0*** *→* ***tun0*** *→ (NAT) →* ***eth0*** *(physical device) → Internet*

Almost all packet delivery decisions are performed with OpenFlow rules in the OVS bridge **br0**, which simplifies the plug-in network architecture and provides flexible routing. In the case of the **ovs-multitenant** plug-in, this also provides enforceable network isolation.

5.2.5. Network Isolation

You can use the **ovs-multitenant** plug-in to achieve network isolation. When a packet exits a pod assigned to a non-default project, the OVS bridge **br0** tags that packet with the project's assigned VNID. If the packet is directed to another IP address in the node's cluster subnet, the OVS bridge only allows the packet to be delivered to the destination pod if the VNIDs match.

If a packet is received from another node via the VXLAN tunnel, the Tunnel ID is used as the VNID, and the OVS bridge only allows the packet to be delivered to a local pod if the tunnel ID matches the destination pod's VNID.

Packets destined for other cluster subnets are tagged with their VNID and delivered to the VXLAN tunnel with a tunnel destination address of the node owning the cluster subnet.

As described before, VNID 0 is privileged in that traffic with any VNID is allowed to enter any pod assigned VNID 0, and traffic with VNID 0 is allowed to enter any pod. Only the **default** OpenShift Container Platform project is assigned VNID 0; all other projects are assigned unique, isolation-enabled VNIDs. Cluster administrators can optionally control the pod network for the project using the administrator CLI.

5.3. AVAILABLE SDN PLUG-INS

OpenShift Container Platform supports the Kubernetes Container Network Interface (CNI) as the interface between the OpenShift Container Platform and Kubernetes. Software defined network (SDN) plug-ins match network capabilities to your networking needs. Additional plug-ins that support the CNI interface can be added as needed.

5.3.1. OpenShift SDN

OpenShift SDN is installed and configured by default as part of the Ansible-based installation procedure. See the OpenShift SDN section for more information.

5.3.2. Third-Party SDN plug-ins

5.3.2.1. Flannel SDN

flannel is a virtual networking layer designed specifically for containers. OpenShift Container Platform can use it for networking containers instead of the default software-defined networking (SDN) components. This is useful if running OpenShift Container Platform within a cloud provider platform that also relies on SDN, such as OpenStack, and you want to avoid encapsulating packets twice through both platforms.

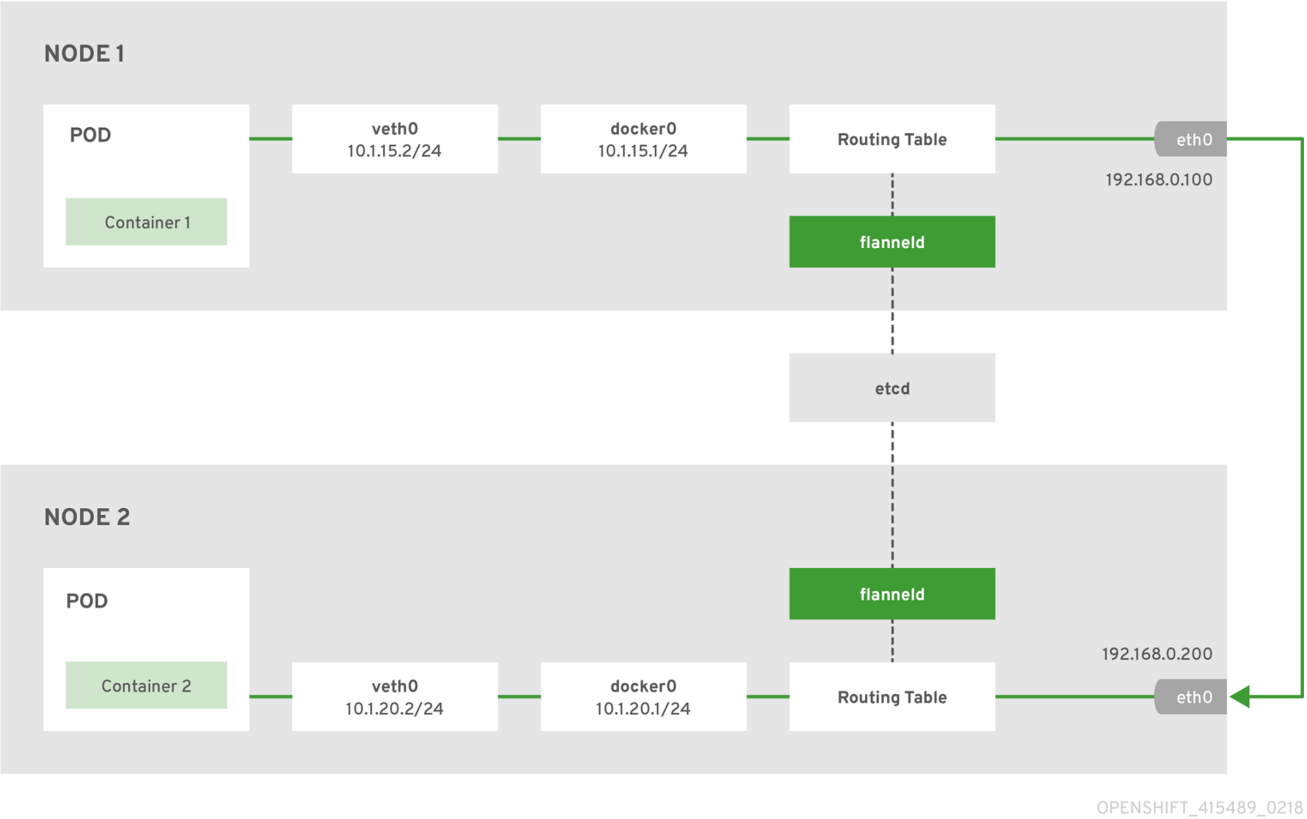
Architecture

OpenShift Container Platform runs **flannel** in **host-gw** mode, which maps routes from container to container. Each host within the network runs an agent called **flanneld**, which is responsible for:

- Managing a unique subnet on each host
- Distributing IP addresses to each container on its host
- Mapping routes from one container to another, even if on different hosts

Each **flanneld** agent provides this information to a centralized **etcd** store so other agents on hosts can route packets to other containers within the **flannel** network.

The following diagram illustrates the architecture and data flow from one container to another using a **flannel** network:



Node 1 would contain the following routes:

```
default via 192.168.0.100 dev eth0 proto static metric 100
10.1.15.0/24 dev docker0 proto kernel scope link src 10.1.15.1
10.1.20.0/24 via 192.168.0.200 dev eth0
```

Node 2 would contain the following routes:

```
default via 192.168.0.200 dev eth0 proto static metric 100
10.1.20.0/24 dev docker0 proto kernel scope link src 10.1.20.1
10.1.15.0/24 via 192.168.0.100 dev eth0
```

5.3.2.2. Nuage SDN

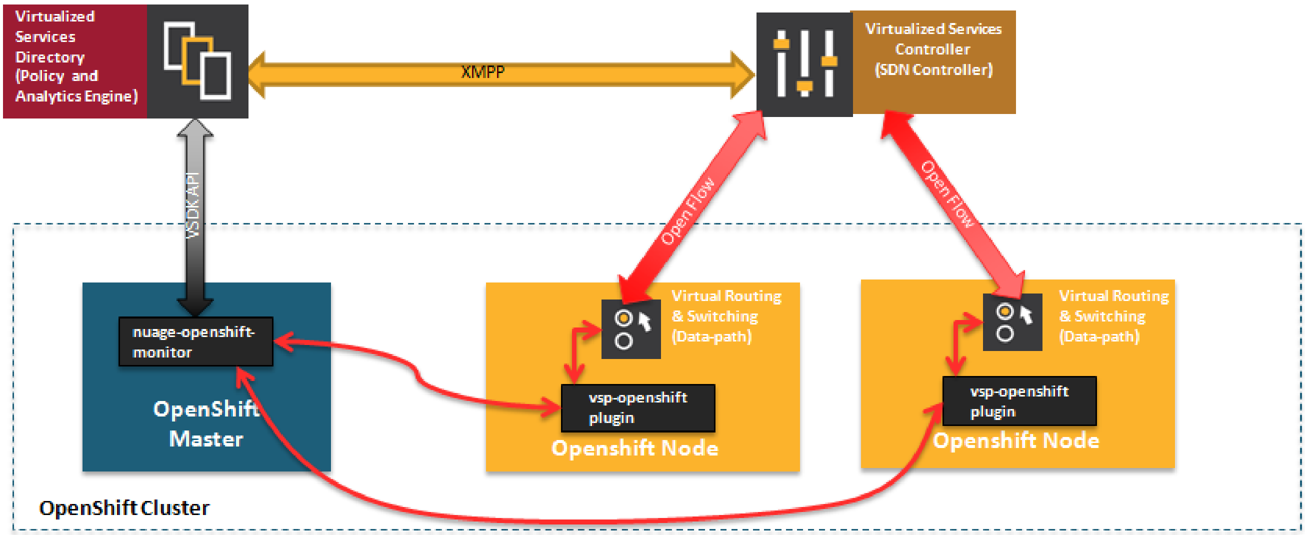
Nuage Networks' SDN solution delivers highly scalable, policy-based overlay networking for pods in an OpenShift Container Platform cluster. Nuage SDN can be installed and configured as a part of the Ansible-based installation procedure. See the Advanced Installation section for information on how to install and deploy OpenShift Container Platform with Nuage SDN.

Nuage Networks provides a highly scalable, policy-based SDN platform called Virtualized Services Platform (VSP). Nuage VSP uses an SDN Controller, along with the open source Open vSwitch for the data plane.

Nuage uses overlays to provide policy-based networking between OpenShift Container Platform and other environments consisting of VMs and bare metal servers. The platform’s real-time analytics engine enables visibility and security monitoring for OpenShift Container Platform applications.

Nuage VSP integrates with OpenShift Container Platform to allows business applications to be quickly turned up and updated by removing the network lag faced by DevOps teams.

Figure 5.1. Nuage VSP Integration with OpenShift Container Platform


Virtualized Services Directory (Policy and Analytics Engine)
XMPP
Virtualized Services Controller (SDN Controller)
VSDK API
Open Flow
Open Flow
nuage-openshift-monitor
OpenShift Master
Virtual Routing & Switching (Data-path)
vsp-openshift plugin
Openshift Node
Virtual Routing & Switching (Data-path)
vsp-openshift plugin
Openshift Node
OpenShift Cluster

There are two specific components responsible for the integration.

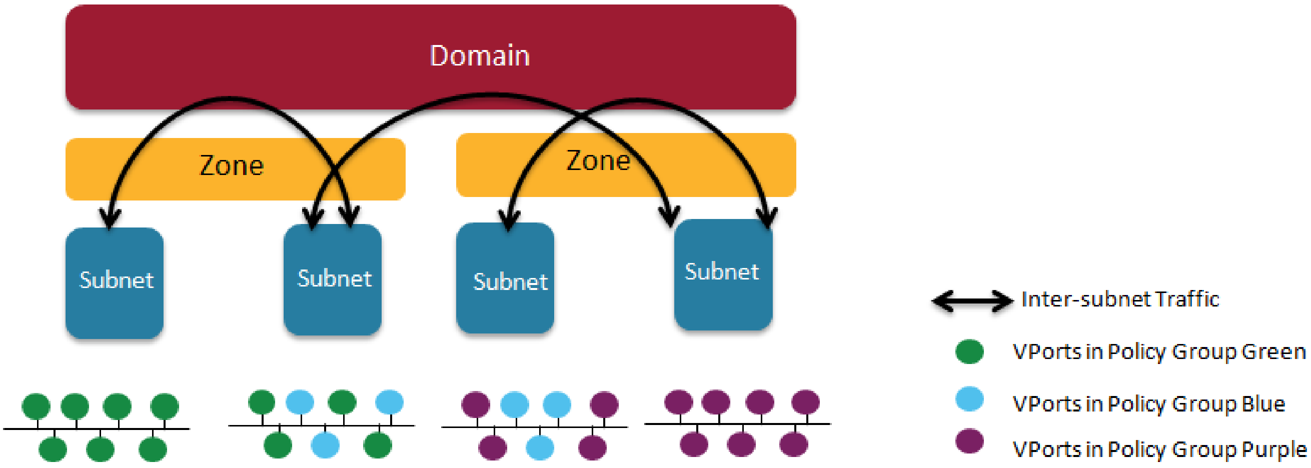
1. The **nuage-openshift-monitor** service, which runs as a separate service on the OpenShift Container Platform master node.

2. The **vsp-openshift** plug-in, which is invoked by the OpenShift Container Platform runtime on each of the nodes of the cluster.

Nuage Virtual Routing and Switching software (VRS) is based on open source Open vSwitch and is responsible for the datapath forwarding. The VRS runs on each node and gets policy configuration from the controller.

Nuage VSP Terminology

Figure 5.2. Nuage VSP Building Blocks


Domain
Zone
Zone
Subnet
Subnet
Subnet
Subnet
Inter-subnet Traffic
VPorts in Policy Group Green
VPorts in Policy Group Blue
VPorts in Policy Group Purple

1. Domains: An organization contains one or more domains. A domain is a single "Layer 3" space. In standard networking terminology, a domain maps to a VRF instance.

2. Zones: Zones are defined under a domain. A zone does not map to anything on the network directly, but instead acts as an object with which policies are associated such that all endpoints in the zone adhere to the same set of policies.

3. Subnets: Subnets are defined under a zone. A subnet is a specific Layer 2 subnet within the domain instance. A subnet is unique and distinct within a domain, that is, subnets within a Domain are not allowed to overlap or to contain other subnets in accordance with the standard

IP subnet definitions.

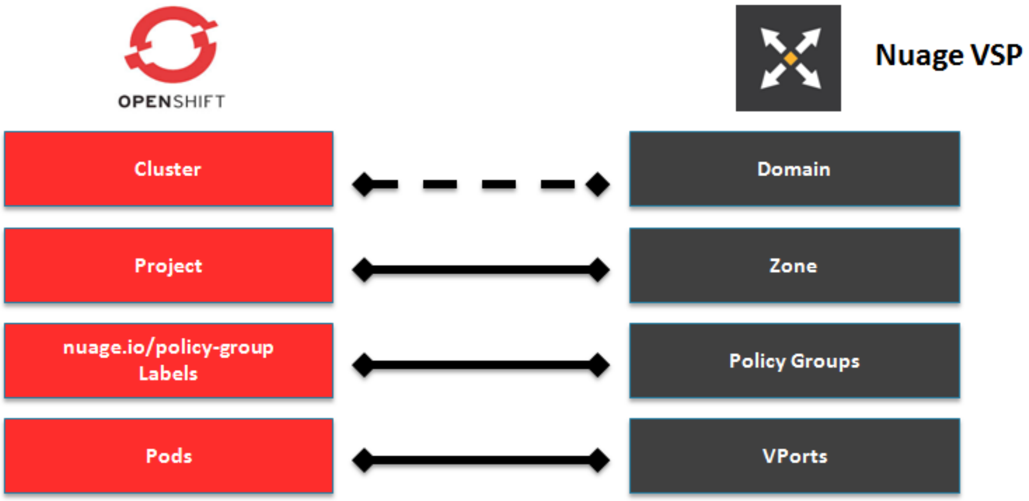
4. VPorts: A VPort is a new level in the domain hierarchy, intended to provide more granular configuration. In addition to containers and VMs, VPorts are also used to attach Host and Bridge Interfaces, which provide connectivity to Bare Metal servers, Appliances, and Legacy VLANs.

5. Policy Group: Policy Groups are collections of VPorts.

Mapping of Constructs

Many OpenShift Container Platform concepts have a direct mapping to Nuage VSP constructs:

Figure 5.3. Nuage VSP and OpenShift Container Platform mapping



A Nuage subnet is not mapped to an OpenShift Container Platform node, but a subnet for a particular project can span multiple nodes in OpenShift Container Platform.

A pod spawning in OpenShift Container Platform translates to a virtual port being created in VSP. The **vsp-openshift** plug-in interacts with the VRS and gets a policy for that virtual port from the VSD via the VSC. Policy Groups are supported to group multiple pods together that must have the same set of policies applied to them. Currently, pods can only be assigned to policy groups using the operations workflow where a policy group is created by the administrative user in VSD. The pod being a part of the policy group is specified by means of **nuage.io/policy-group** label in the specification of the pod.

Integration Components

Nuage VSP integrates with OpenShift Container Platform using two main components:

1. **nuage-openshift-monitor**

2. **vsp-openshift plugin**

nuage-openshift-monitor

nuage-openshift-monitor is a service that monitors the OpenShift Container Platform API server for creation of projects, services, users, user-groups, etc.

NOTE

In case of a Highly Available (HA) OpenShift Container Platform cluster with multiple masters, **nuage-openshift-monitor** process runs on all the masters independently without any change in functionality.

For the developer workflow, **nuage-openshift-monitor** also auto-creates VSD objects by exercising the VSD REST API to map OpenShift Container Platform constructs to VSP constructs. Each cluster instance maps to a single domain in Nuage VSP. This allows a given enterprise to potentially have multiple cluster installations - one per domain instance for that Enterprise in Nuage. Each OpenShift Container Platform project is mapped to a zone in the domain of the cluster on the Nuage VSP. Whenever **nuage-openshift-monitor** sees an addition or deletion of the project, it instantiates a zone using the VSDK APIs corresponding to that project and allocates a block of subnet for that zone. Additionally, the **nuage-openshift-monitor** also creates a network macro group for this project. Likewise, whenever **nuage-openshift-monitor** sees an addition ordeletion of a service, it creates a network macro corresponding to the service IP and assigns that network macro to the network macro group for that project (user provided network macro group using labels is also supported) to enable communication to that service.

For the developer workflow, all pods that are created within the zone get IPs from that subnet pool. The subnet pool allocation and management is done by **nuage-openshift-monitor** based on a couple of plug-in specific parameters in the master-config file. However the actual IP address resolution and vport policy resolution is still done by VSD based on the domain/zone that gets instantiated when the project is created. If the initial subnet pool is exhausted, **nuage-openshift-monitor** carves out an additional subnet from the cluster CIDR to assign to a given project.

For the operations workflow, the users specify Nuage recognized labels on their application or pod specification to resolve the pods into specific user-defined zones and subnets. However, this cannot be used to resolve pods in the zones or subnets created via the developer workflow by **nuage-openshift-monitor**.

NOTE

In the operations workflow, the administrator is responsible for pre-creating the VSD constructs to map the pods into a specific zone/subnet as well as allow communication between OpenShift entities (ACL rules, policy groups, network macros, and network macro groups). Detailed description of how to use Nuage labels is provided in the Nuage VSP Openshift Integration Guide.

vsp-openshift Plug-in

The vsp-openshift networking plug-in is called by the OpenShift Container Platform runtime on each OpenShift Container Platform node. It implements the network plug-in init and pod setup, teardown, and status hooks. The vsp-openshift plug-in is also responsible for allocating the IP address for the pods. In particular, it communicates with the VRS (the forwarding engine) and configures the IP information onto the pod.

5.4. AVAILABLE ROUTER PLUG-INS

A router can be assigned to a node to control traffic in an OpenShift Container Platform cluster. OpenShift Container Platform uses HAProxy as the default router, but options are available.

5.4.1. The HAProxy Template Router

The HAProxy template router implementation is the reference implementation for a template router plug-in. It uses the **openshift3/ose-haproxy-router** repository to run an HAProxy instance alongside the template router plug-in.

The template router has two components:

- A wrapper that watches endpoints and routes and causes a HAProxy reload based on changes
- A controller that builds the HAProxy configuration file based on routes and endpoints

NOTE

The HAProxy router uses version 1.8.1.

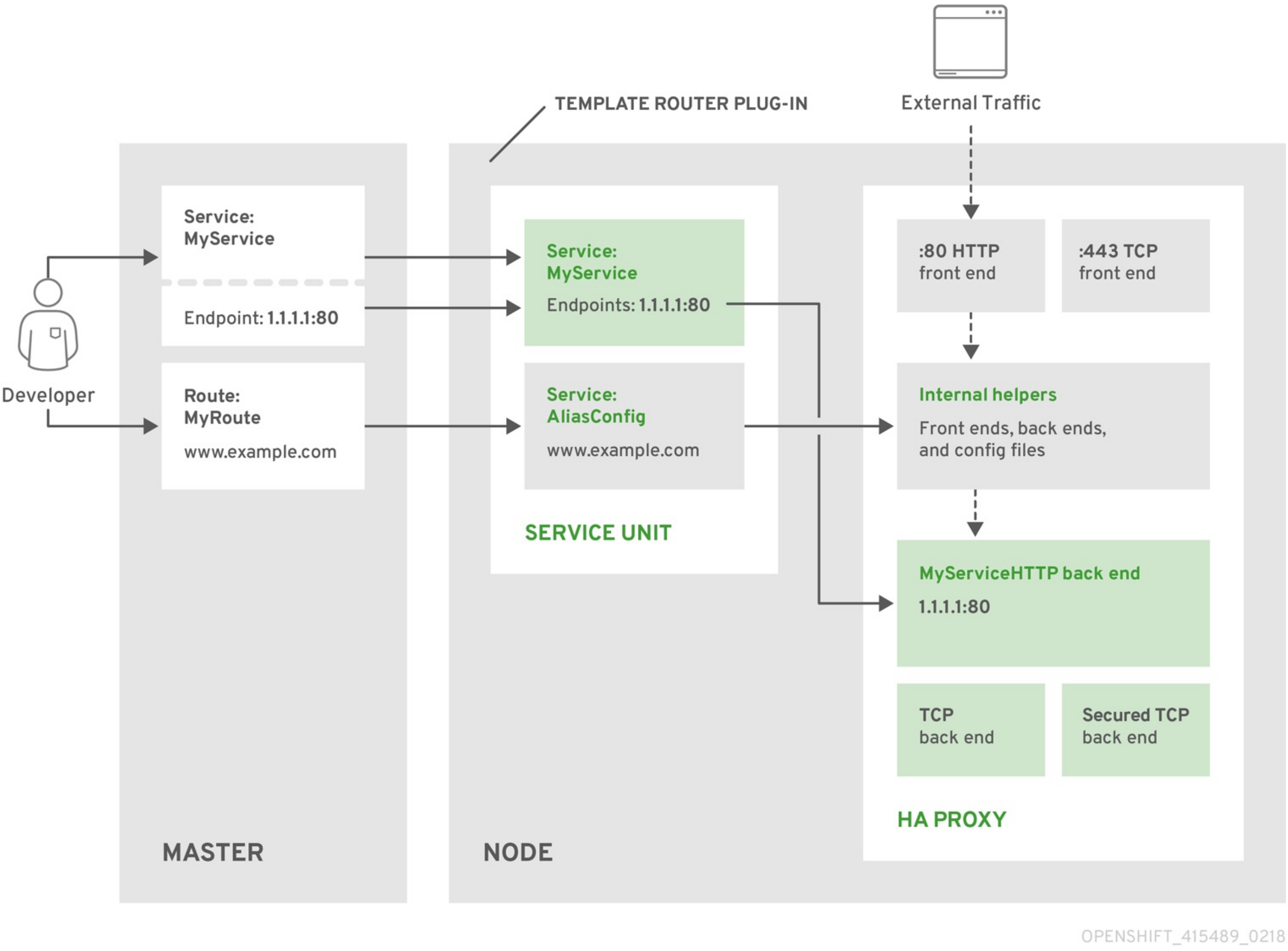
The controller and HAProxy are housed inside a pod, which is managed by a deployment configuration. The process of setting up the router is automated by the **oc adm router** command.

The controller watches the routes and endpoints for changes, as well as HAProxy's health. When a change is detected, it builds a new haproxy-config file and restarts HAProxy. The haproxy-config file is constructed based on the router's template file and information from OpenShift Container Platform.

The HAProxy template file can be customized as needed to support features that are not currently supported by OpenShift Container Platform. The HAProxy manual describes all of the features supported by HAProxy.

The following diagram illustrates how data flows from the master through the plug-in and finally into an HAProxy configuration:

Figure 5.4. HAProxy Router Data Flow



HAProxy Template Router Metrics

The HAProxy router exposes or publishes metrics in Prometheus format for consumption by external metrics collection and aggregation systems (e.g. Prometheus, statsd). The router can be configured to provide HAProxy CSV format metrics, or provide no router metrics at all.

The metrics are collected from both the router controller and from HAProxy every five seconds. The router metrics counters start at zero when the router is deployed and increase over time. The HAProxy metrics counters are reset to zero every time haproxy is reloaded. The router collects HAProxy statistics for each frontend, back end, and server. To reduce resource usage when there are more than 500 servers, the back ends are reported instead of the servers because a back end can have multiple servers.

The statistics are a subset of the available HAProxy statistics .

The following HAProxy metrics are collected on a periodic basis and converted to Prometheus format. For every front end the "F" counters are collected. When the counters are collected for each back end and the "S" server counters are collected for each server. Otherwise, the "B" counters are collected for each back end and no server counters are collected.

See router environment variables for more information.

In the following table:

Column 1 - Index from HAProxy CSV statistics

Column 2

F	Front end metrics
b	Back end metrics when not showing Server metrics due to the Server Threshold,
B	Back end metrics when showing Server metrics
S	Server metrics.

Column 3 - The counter

Column 4 - Counter description

Index	Usage	Counter	Description
2	bBS	current_queue	Current number of queued requests not assigned to any server.
4	FbS	current_sessions	Current number of active sessions.
5	FbS	max_sessions	Maximum observed number of active sessions.
7	FbBS	connections_total	Total number of connections.
8	FbS	bytes_in_total	Current total of incoming bytes.
9	FbS	bytes_out_total	Current total of outgoing bytes.
13	bS	connection_errors_total	Total of connection errors.
14	bS	response_errors_total	Total of response errors.
17	bBS	up	Current health status of the back end (1 = UP, 0 = DOWN).
21	S	check_failures_total	Total number of failed health checks.

24	S	downtime_seconds_total	Total downtime in seconds.", nil),
33	FbS	current_session_rate	Current number of sessions per second over last elapsed second.
35	FbS	max_session_rate	Maximum observed number of sessions per second.
40	FbS	http_responses_total	Total of HTTP responses, code 2xx
43	FbS	http_responses_total	Total of HTTP responses, code 5xx
60	bS	http_average_response_latency_milliseconds	of the last 1024 requests in milliseconds.

The router controller scrapes the following items. These are only available with Prometheus format metrics.

Name	Description
template_router_reload_seconds	Measures the time spent reloading the router in seconds.
template_router_write_config_seconds	Measures the time spent writing out the router configuration to disk in seconds.
haproxy_exporter_up	Was the last scrape of haproxy successful.
haproxy_exporter_csv_parse_failures	Number of errors while parsing CSV.
haproxy_exporter_scrape_interval	The time in seconds before another scrape is allowed, proportional to size of data.
haproxy_exporter_server_threshold	Number of servers tracked and the current threshold value.
haproxy_exporter_total_scrapes	Current total HAProxy scrapes.
http_request_duration_microseconds	The HTTP request latencies in microseconds.
http_request_size_bytes	The HTTP request sizes in bytes.

http_response_size_bytes	The HTTP response sizes in bytes.
openshift_build_info	A metric with a constant '1' value labeled by major, minor, git commit & git version from which OpenShift was built.
ssh_tunnel_open_count	Counter of SSH tunnel total open attempts
ssh_tunnel_open_fail_count	Counter of SSH tunnel failed open attempts

5.4.2. F5 BIG-IP® Router plug-in

The F5 BIG-IP® Router plug-in is one of the available router plugins.

NOTE

The F5 router plug-in is available starting in OpenShift Enterprise 3.0.2.

The F5 router plug-in integrates with an existing **F5 BIG-IP®** system in your environment. **F5 BIG-IP®** version 11.4 or newer is required in order to have the F5 iControl REST API. The F5 router supports unsecured, edge terminated, re-encryption terminated, and passthrough terminated routes matching on HTTP vhost and request path.

The F5 router has feature parity with the HAProxy template router, and has additional features over the **F5 BIG-IP®** support in OpenShift Enterprise 2. Compared with the **routing-daemon** used in earlier versions, the F5 router additionally supports:

- path-based routing (using policy rules),
- re-encryption (implemented using client and server SSL profiles)
- passthrough of encrypted connections (implemented using an iRule that parses the SNI protocol and uses a data group that is maintained by the F5 router for the servername lookup).

NOTE

Passthrough routes are a special case: path-based routing is technically impossible with passthrough routes because **F5 BIG-IP®** itself does not see the HTTP request, so it cannot examine the path. The same restriction applies to the template router; it is a technical limitation of passthrough encryption, not a technical limitation of OpenShift Container Platform.

5.4.2.1. Routing Traffic to Pods Through the SDN

Because **F5 BIG-IP®** is external to the OpenShift SDN, a cluster administrator must create a peer-to-peer tunnel between **F5 BIG-IP®** and a host that is on the SDN, typically an OpenShift Container Platform node host. This *ramp node* can be configured as unschedulable for pods so that it will not be doing anything except act as a gateway for the **F5 BIG-IP®** host. You can also configure multiple such hosts and use the OpenShift Container Platform **ipfailover** feature for redundancy; the **F5 BIG-IP®** host would then need to be configured to use the **ipfailover** VIP for its tunnel’s remote endpoint.

5.4.2.2. F5 Integration Details

The operation of the F5 router is similar to that of the OpenShift Container Platform **routing-daemon** used in earlier versions. Both use REST API calls to:

- create and delete pools,
- add endpoints to and delete them from those pools, and
- configure policy rules to route to pools based on vhost.

Both also use **scp** and **ssh** commands to upload custom TLS/SSL certificates to **F5 BIG-IP®**.

The F5 router configures pools and policy rules on virtual servers as follows:

- When a user creates or deletes a route on OpenShift Container Platform, the router creates a pool to **F5 BIG-IP®** for the route (if no pool already exists) and adds a rule to, or deletes a rule from, the policy of the appropriate vserver: the HTTP vserver for non-TLS routes, or the HTTPS vserver for edge or re-encrypt routes. In the case of edge and re-encrypt routes, the router also uploads and configures the TLS certificate and key. The router supports host- and path-based routes.

  NOTE

  Passthrough routes are a special case: to support those, it is necessary to write an iRule that parses the SNI ClientHello handshake record and looks up the servername in an F5 data-group. The router creates this iRule, associates the iRule with the vserver, and updates the F5 data-group as passthrough routes are created and deleted. Other than this implementation detail, passthrough routes work the same way as other routes.

- When a user creates a service on OpenShift Container Platform, the router adds a pool to **F5 BIG-IP®** (if no pool already exists). As endpoints on that service are created and deleted, the router adds and removes corresponding pool members.
- When a user deletes the route and all endpoints associated with a particular pool, the router deletes that pool.

5.4.2.3. F5 Native Integration

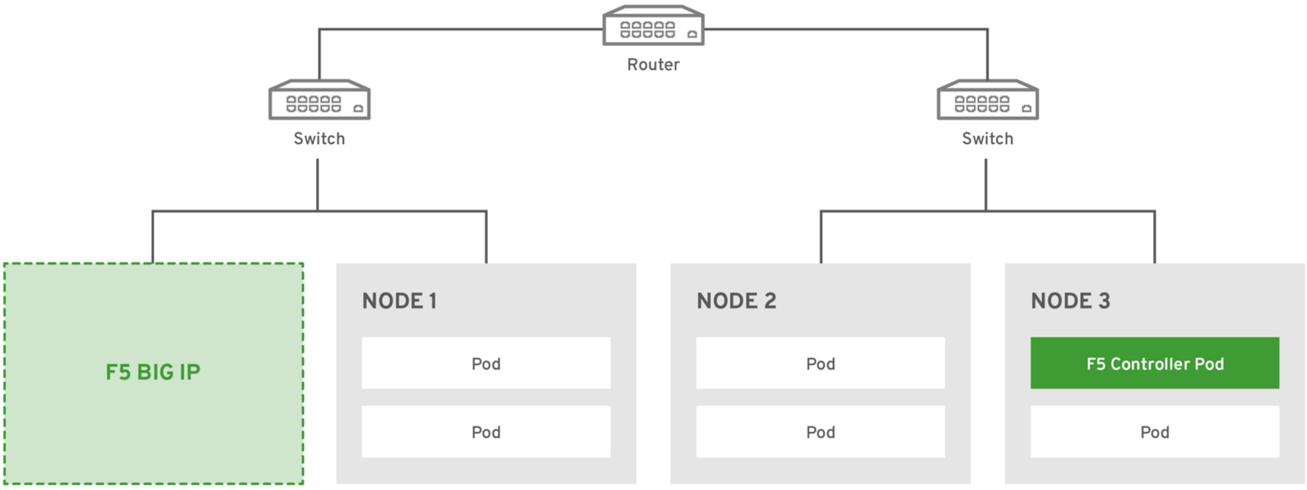
With native integration of F5 with OpenShift Container Platform , you do not need to configure a ramp node for F5 to be able to reach the pods on the overlay network as created by OpenShift SDN.

Also, only **F5 BIG-IP®** appliance version 12.x and above works with the native integration presented in this section. You also need **sdn-services** add-on license for the integration to work properly. For version 11.x, set up a ramp node.

Connection

The F5 appliance can connect to the OpenShift Container Platform cluster via an L3 connection. An L2 switch connectivity is not required between OpenShift Container Platform nodes. On the appliance, you can use multiple interfaces to manage the integration:

- Management interface - Reaches the web console of the F5 appliance.
- External interface - Configures the virtual servers for inbound web traffic.
- Internal interface - Programs the appliance and reaches out to the pods.



OPENSHIFT_415489_0218

An F5 controller pod has **admin** access to the appliance. The F5 image is launched within the OpenShift Container Platform cluster (scheduled on any node) that uses iControl REST APIs to program the virtual servers with policies, and configure the VxLAN device.

Data Flow: Packets to Pods

NOTE

This section explains how the packets reach the pods, and vice versa. These actions are performed by the F5 controller pod and the F5 appliance, not the user.

When natively integrated, The F5 appliance reaches out to the pods directly using VxLAN encapsulation. This integration works only when OpenShift Container Platform is using **openshift-sdn** as the network plug-in. The **openshift-sdn** plug-in employs VxLAN encapsulation for the overlay network that it creates.

To make a successful data path between a pod and the F5 appliance:

1. F5 needs to encapsulate the VxLAN packet meant for the pods. This requires the **sdn-services** license add-on. A VxLAN device needs to be created and the pod overlay network needs to be routed through this device.

2. F5 needs to know the VTEP IP address of the pod, which is the IP address of the node where the pod is located.

3. F5 needs to know which **source-ip** to use for the overlay network when encapsulating the packets meant for the pods. This is known as the *gateway address*.

4. OpenShift Container Platform nodes need to know where the F5 gateway address is (the VTEP address for the return traffic). This needs to be the internal interface's address. All nodes of the cluster must learn this automatically.

5. Since the overlay network is multi-tenant aware, F5 must use a VxLAN ID that is representative of an **admin** domain, ensuring that all tenants are reachable by the F5. Ensure that F5 encapsulates all packets with a **vnid** of **0** (the default **vnid** for the **admin** namespace in OpenShift Container Platform) by putting an annotation on the manually created **hostsubnet** - **pod.network.openshift.io/fixed-vnid-host: 0**.

A ghost **hostsubnet** is manually created as part of the setup, which fulfills the third and forth listed requirements. When the F5 controller pod is launched, this new ghost **hostsubnet** is provided so that the F5 appliance can be programmed suitably.

NOTE

The term *ghost* **hostsubnet** is used because it suggests that a subnet has been given to a node of the cluster. However, in reality, it is not a real node of the cluster. It is hijacked by an external appliance.

The first requirement is fulfilled by the F5 controller pod once it is launched. The second requirement is also fulfilled by the F5 controller pod, but it is an ongoing process. For each new node that is added to the cluster, the controller pod creates an entry in the VxLAN device's VTEP FDB. The controller pod needs access to the **nodes** resource in the cluster, which you can accomplish by giving the service account appropriate privileges. Use the following command:

```
$ oc adm policy add-cluster-role-to-user system:sdn-reader system:serviceaccount:default:router
```

Data Flow from the F5 Host

NOTE

These actions are performed by the F5 controller pod and the F5 appliance, not the user.

1. The destination pod is identified by the F5 virtual server for a packet.
2. VxLAN dynamic FDB is looked up with pod's IP address. If a MAC address is found, go to step 5.
3. Flood all entries in the VTEP FDB with ARP requests seeking the pod's MAC address. An entry is made into the VxLAN dynamic FDB with the pod's MAC address and the VTEP to be used as the value.
4. Encap an IP packet with VxLAN headers, where the MAC of the pod and the VTEP of the node is given as values from the VxLAN dynamic FDB.
5. Calculate the VTEP's MAC address by sending out an ARP or checking the host's neighbor cache.
6. Deliver the packet through the F5 host's internal address.

Data Flow: Return Traffic to the F5 Host

NOTE

These actions are performed by the F5 controller pod and the F5 appliance, not the user.

1. The pod sends back a packet with the destination as the F5 host's VxLAN gateway address.
2. The **openvswitch** at the node determines that the VTEP for this packet is the F5 host's internal interface address. This is learned from the ghost **hostsubnet** creation.
3. A VxLAN packet is sent out to the internal interface of the F5 host.

NOTE

During the entire data flow, the VNID is pre-fixed to be **0** to bypass multi-tenancy.

5.5. PORT FORWARDING

5.5.1. Overview

OpenShift Container Platform takes advantage of a feature built-in to Kubernetes to support port forwarding to pods. This is implemented using HTTP along with a multiplexed streaming protocol such as **SPDY** or **HTTP/2**.

Developers can use the CLI to port forward to a pod. The CLI listens on each local port specified by the user, forwarding via the described protocol.

5.5.2. Server Operation

The Kubelet handles port forward requests from clients. Upon receiving a request, it upgrades the response and waits for the client to create port forwarding streams. When it receives a new stream, it copies data between the stream and the pod's port.

Architecturally, there are options for forwarding to a pod's port. The supported implementation currently in OpenShift Container Platform invokes **nsenter** directly on the node host to enter the pod's network namespace, then invokes **socat** to copy data between the stream and the pod's port. However, a custom implementation could include running a "helper" pod that then runs **nsenter** and **socat**, so that those binaries are not required to be installed on the host.

5.6. REMOTE COMMANDS

5.6.1. Overview

OpenShift Container Platform takes advantage of a feature built into Kubernetes to support executing commands in containers. This is implemented using HTTP along with a multiplexed streaming protocol such as **SPDY** or **HTTP/2**.

Developers can use the CLI to execute remote commands in containers.

5.6.2. Server Operation

The Kubelet handles remote execution requests from clients. Upon receiving a request, it upgrades the response, evaluates the request headers to determine what streams (**stdin**, **stdout**, and/or **stderr**) to expect to receive, and waits for the client to create the streams.

After the Kubelet has received all the streams, it executes the command in the container, copying between the streams and the command's stdin, stdout, and stderr, as appropriate. When the command terminates, the Kubelet closes the upgraded connection, as well as the underlying one.

Architecturally, there are options for running a command in a container. The supported implementation currently in OpenShift Container Platform invokes **nsenter** directly on the node host to enter the container's namespaces prior to executing the command. However, custom implementations could include using **docker exec**, or running a "helper" container that then runs **nsenter** so that **nsenter** is not a required binary that must be installed on the host.

5.7. ROUTES

5.7.1. Overview

An OpenShift Container Platform route exposes a service at a host name, such as *www.example.com*, so that external clients can reach it by name.

DNS resolution for a host name is handled separately from routing. Your administrator may have configured a DNS wildcard entry that will resolve to the OpenShift Container Platform node that is running the OpenShift Container Platform router. If you are using a different host name you may need to modify its DNS records independently to resolve to the node that is running the router.

Each route consists of a name (limited to 63 characters), a service selector, and an optional security configuration.

5.7.2. Routers

An OpenShift Container Platform administrator can deploy *routers* to nodes in an OpenShift Container Platform cluster, which enable routes created by developers to be used by external clients. The routing layer in OpenShift Container Platform is pluggable, and several router plug-ins are provided and supported by default.

> **NOTE**
>
> See the Installation and Configuration guide for information on deploying a router.

A router uses the service selector to find the service and the endpoints backing the service. When both router and service provide load balancing, OpenShift Container Platform uses the router load balancing. A router detects relevant changes in the IP addresses of its services and adapts its configuration accordingly. This is useful for custom routers to communicate modifications of API objects to an external routing solution.

The path of a request starts with the DNS resolution of a host name to one or more routers. The suggested method is to define a cloud domain with a wildcard DNS entry pointing to one or more virtual IP (VIP) addresses backed by multiple router instances. Routes using names and addresses outside the cloud domain require configuration of individual DNS entries.

When there are fewer VIP addresses than routers, the routers corresponding to the number of addresses are *active* and the rest are *passive*. A passive router is also known as a *hot-standby* router. For example, with two VIP addresses and three routers, you have an "active-active-passive" configuration. See High Availability for more information on router VIP configuration.

Routes can be sharded among the set of routers. Administrators can set up sharding on a cluster-wide basis and users can set up sharding for the namespace in their project. Sharding allows the operator to define multiple router groups. Each router in the group serves only a subset of traffic.

OpenShift Container Platform routers provide external host name mapping and load balancing of service end points over protocols that pass distinguishing information directly to the router; the host name must be present in the protocol in order for the router to determine where to send it.

Router plug-ins assume they can bind to host ports 80 (HTTP) and 443 (HTTPS), by default. This means that routers must be placed on nodes where those ports are not otherwise in use. Alternatively, a router can be configured to listen on other ports by setting the **ROUTER_SERVICE_HTTP_PORT** and **ROUTER_SERVICE_HTTPS_PORT** environment variables.

Because a router binds to ports on the host node, only one router listening on those ports can be on each node if the router uses host networking (the default). Cluster networking is configured such that all routers can access all pods in the cluster.

Routers support the following protocols:

- HTTP
- HTTPS (with SNI)
- WebSockets
- TLS with SNI

NOTE

WebSocket traffic uses the same route conventions and supports the same TLS termination types as other traffic.

For a secure connection to be established, a cipher common to the client and server must be negotiated. As time goes on, new, more secure ciphers become available and are integrated into client software. As older clients become obsolete, the older, less secure ciphers can be dropped. By default, the router supports a broad range of commonly available clients. The router can be configured to use a selected set of ciphers that support desired clients and do not include the less secure ciphers.

5.7.2.1. Template Routers

A *template router* is a type of router that provides certain infrastructure information to the underlying router implementation, such as:

- A wrapper that watches endpoints and routes.
- Endpoint and route data, which is saved into a consumable form.
- Passing the internal state to a configurable template and executing the template.
- Calling a reload script.

5.7.3. Available Router Plug-ins

See the Available router plug-ins section for the verified available router plug-ins.

Instructions on deploying these routers are available in Deploying a Router .

5.7.4. Sticky Sessions

Implementing sticky sessions is up to the underlying router configuration. The template router plug-in provides the service name and namespace to the underlying implementation. This can be used for more advanced configuration, such as implementing stick-tables that synchronize between a set of peers.

Sticky sessions ensure that all traffic from a user’s session go to the same pod, creating a better user experience. While satisfying the user’s requests, the pod caches data, which can be used in subsequent requests. For example, for a cluster with five back-end pods and two load-balanced routers, you can ensure that the same pod receives the web traffic from the same web browser regardless of the router that handles it.

While returning routing traffic to the same pod is desired, it cannot be guaranteed. However, you can use HTTP headers to set a cookie to determine the pod used in the last connection. When the user sends another request to the application the browser re-sends the cookie and the router knows where to send the traffic.

Cluster administrators can turn off stickiness for passthrough routes separately from other connections, or turn off stickiness entirely.

By default, sticky sessions for passthrough routes are implemented using the **source** load-balancing strategy. However, the **roundrobin** load-balancing strategy is the default when there are active services with weights greater than **1**. You can change the default for all passthrough routes by using the **ROUTER_LOAD_BALANCE_ALGORITHM** environment variable, and for individual routes by using the **haproxy.router.openshift.io/balance** route specific annotation.

Other types of routes use the **leastconn** load-balancing strategy by default, which can be changed by using the **ROUTER_LOAD_BALANCE_ALGORITHM** environment variable. It can be changed for individual routes by using the **haproxy.router.openshift.io/balance** route specific annotation.

NOTE

Cookies cannot be set on passthrough routes, because the HTTP traffic cannot be seen. Instead, a number is calculated based on the source IP address, which determines the back-end.

If back-ends change, the traffic could head to the wrong server, making it less sticky, and if you are using a load-balancer (which hides the source IP) the same number is set for all connections and traffic is sent to the same pod.

In addition, the template router plug-in provides the service name and namespace to the underlying implementation. This can be used for more advanced configuration such as implementing stick-tables that synchronize between a set of peers.

Specific configuration for this router implementation is stored in the ***haproxy-config.template*** file located in the ***/var/lib/haproxy/conf*** directory of the router container. The file may be customized.

NOTE

The **source** load-balancing strategy does not distinguish between external client IP addresses; because of the NAT configuration, the originating IP address (HAProxy remote) is the same. Unless the HAProxy router is running with **hostNetwork: true**, all external clients will be routed to a single pod.

5.7.5. Router Environment Variables

For all the items outlined in this section, you can set environment variables in the **deployment config** for the router to alter its configuration, or use the **oc set env** command:

```
$ oc set env <object_type>/<object_name> KEY1=VALUE1 KEY2=VALUE2
```

For example:

```
$ oc set env dc/router ROUTER_SYSLOG_ADDRESS=127.0.0.1 ROUTER_LOG_LEVEL=debug
```

Table 5.2. Router Environment Variables

Variable	Default	Description
DEFAULT_CERTIFICATE		The contents of a default certificate to use for routes that don't expose a TLS server cert; in PEM format.
DEFAULT_CERTIFICATE_DIR		A path to a directory that contains a file named ***tls.crt***. If ***tls.crt*** is not a PEM file which also contains a private key, it is first combined with a file named tls.key in the same directory. The PEM-format contents are then used as the default certificate. Only used if **DEFAULT_CERTIFICATE** or **DEFAULT_CERTIFICATE_PATH** are not specified.
DEFAULT_CERTIFICATE_PATH		A path to default certificate to use for routes that don't expose a TLS server cert; in PEM format. Only used if **DEFAULT_CERTIFICATE** is not specified.
EXTENDED_VALIDATION	**true**	If **true**, the router confirms that the certificate is structurally correct. It does not verify the certificate against any CA. Set **false** to turn off the tests.
NAMESPACE_LABELS		A label selector to apply to namespaces to watch, empty means all.
PROJECT_LABELS		A label selector to apply to projects to watch, emtpy means all.
RELOAD_SCRIPT		The path to the reload script to use to reload the router.
ROUTER_ALLOWED_DOMAINS		A comma-separated list of domains that the host name in a route can only be part of. Any subdomain in the domain can be used. Option **ROUTER_DENIED_DOMAINS** overrides any values given in this option. If set, everything outside of the allowed domains will be rejected.
ROUTER_BACKEND_PROCESS_ENDPOINTS		String to specify how the endpoints should be processed while using the template function processEndpointsForAlias. Valid values are ["shuffle", ""]. "shuffle" will randomize the elements upon every call. Default behavior returns in pre-determined order.
ROUTER_BIND_PORTS_AFTER_SYNC	**false**	If set to **true** or **TRUE**, then the router does not bind to any ports until it has completely synchronized state. If not set to 'true' or 'TRUE', the router will bind to ports and start processing requests immediately, but there may be routes that are not loaded.
ROUTER_COOKIE_NAME		Specifies cookie name to override the internally generated default name. The name must consist of any combination of upper and lower case letters, digits, "_", and "-". The default is the hashed internal key name for the route.

Variable	Default	Description
ROUTER_COMPRESSION_MIME	"text/html text/plain text/css"	A space separated list of mime types to compress.
ROUTER_DENIED_DOMAINS		A comma-separated list of domains that the host name in a route can not be part of. No subdomain in the domain can be used either. Overrides option **ROUTER_ALLOWED_DOMAINS**.
ROUTER_ENABLE_COMPRESSION		If **true** or **TRUE**, compress responses when possible.
ROUTER_LISTEN_ADDR	0.0.0.0:1936	Sets the listening address for router metrics.
ROUTER_LOG_LEVEL	warning	The log level to send to the syslog server.
ROUTER_MAX_CONNECTIONS	20000	Maximum number of concurrent connections.
ROUTER_METRICS_HAPROXY_SERVER_THRESHOLD	500	
ROUTER_METRICS_HAPROXY_EXPORTED		Metrics collected in CSV format. For example, **defaultSelectedMetrics** = []int{2, 4, 5, 7, 8, 9, 13, 14, 17, 21, 24, 33, 35, 40, 43, 60}
ROUTER_METRICS_HAPROXY_BASE_SCRAPE_INTERVAL	5s	
ROUTER_METRICS_HAPROXY_TIMEOUT	5s	
ROUTER_METRICS_TYPE	haproxy	Generate metrics for the HAProxy router. (haproxy is the only supported value)
ROUTER_OVERRIDE_HOSTNAME		If set **true**, override the spec.host value for a route with the template in **ROUTER_SUBDOMAIN**.

Variable	Default	Description
ROUTER_SERVICE_HTTPS_PORT	443	Port to listen for HTTPS requests.
ROUTER_SERVICE_HTTP_PORT	80	Port to listen for HTTP requests.
ROUTER_SERVICE_NAME	public	The name that the router identifies itself in the in route status.
ROUTER_CANONICAL_HOSTNAME		The (optional) host name of the router shown in the in route status.
ROUTER_SERVICE_NAMESPACE		The namespace the router identifies itself in the in route status. Required if **ROUTER_SERVICE_NAME** is used.
ROUTER_SERVICE_NO_SNI_PORT	10443	Internal port for some front-end to back-end communication (see note below).
ROUTER_SERVICE_SNI_PORT	10444	Internal port for some front-end to back-end communication (see note below).
ROUTER_SUBDOMAIN		The template that should be used to generate the host name for a route without spec.host (e.g. ${name}-${namespace}.myapps.mycompany.com).
ROUTER_SYSLOG_ADDRESS		Address to send log messages. Disabled if empty.
ROUTER_SYSLOG_FORMAT		If set, override the default log format used by underlying router implementation. Its value should conform with underlying router implementation's specification.
ROUTER_TCP_BALANCE_SCHEME	source	load balancing strategy. for multiple endpoints for pass-through routes. Available options are **source**, **roundrobin**, or **leastconn**.
ROUTER_LOAD_BALANCE_ALGORITHM	leastconn	load balancing strategy. for routes with multiple endpoints. Available options are **source**, **roundrobin**, and **leastconn**.
ROUTE_LABELS		A label selector to apply to the routes to watch, empty means all.

Variable	Default	Description
STATS_PASSWORD		The password needed to access router stats (if the router implementation supports it).
STATS_PORT		Port to expose statistics on (if the router implementation supports it). If not set, stats are not exposed.
STATS_USERNAME		The user name needed to access router stats (if the router implementation supports it).
TEMPLATE_FILE	**/var/lib/haproxy/conf/custom/haproxy-config-custom.template**	The path to the HAProxy template file (in the container image).
ROUTER_USE_PROXY_PROTOCOL		When set to **true** or **TRUE**, HAProxy expects incoming connections to use the **PROXY** protocol on port 80 or port 443. The source IP address can pass through a load balancer if the load balancer supports the protocol, for example Amazon ELB.
ROUTER_ALLOW_WILDCARD_ROUTES		When set to **true** or **TRUE**, any routes with a wildcard policy of **Subdomain** that pass the router admission checks will be serviced by the HAProxy router.
ROUTER_DISABLE_NAMESPACE_OWNERSHIP_CHECK		Set to **true** to relax the namespace ownership policy.
ROUTER_STRICT_SNI		strict-sni
ROUTER_CIPHERS	intermediate	Specify the set of ciphers supported by bind.

NOTE

If you want to run multiple routers on the same machine, you must change the ports that the router is listening on, **ROUTER_SERVICE_SNI_PORT** and **ROUTER_SERVICE_NO_SNI_PORT**. These ports can be anything you want as long as they are unique on the machine. These ports will not be exposed externally.

Router timeout variables

TimeUnits are represented by a number followed by the unit: **us** *(microseconds), **ms** (milliseconds, default), **s** (seconds), **m** (minutes), **h** *(hours), **d** (days).

The regular expression is: [1-9][0-9]*(us\|ms\|s\|m\|h\|d)

ROUTER_BACKEND_CHECK_INTERVAL	5000ms	Length of time between subsequent liveness checks on backends.
ROUTER_CLIENT_FIN_TIMEOUT	1s	Controls the TCP FIN timeout period for the client connecting to the route. If the FIN sent to close the connection is not answered within the given time, HAProxy will close the connection. This is harmless if set to a low value and uses fewer resources on the router.
ROUTER_DEFAULT_CLIENT_TIMEOUT	30s	Length of time that a client has to acknowledge or send data.
ROUTER_DEFAULT_CONNECT_TIMEOUT	5s	The maximum connect time.
ROUTER_DEFAULT_SERVER_FIN_TIMEOUT	1s	Controls the TCP FIN timeout from the router to the pod backing the route.
ROUTER_DEFAULT_SERVER_TIMEOUT	30s	Length of time that a server has to acknowledge or send data.
ROUTER_DEFAULT_TUNNEL_TIMEOUT	1h	Length of time for TCP or WebSocket connections to remain open. If you have websockets/tcp connections (and any time HAProxy is reloaded), the old HAProxy processes will stay for that period.
ROUTER_SLOWLORIS_HTTP_KEEPALIVE	300s	Set the maximum time to wait for a new HTTP request to appear. If this is set too low, it can cause problems with browsers and applications not expecting a small **keepalive** value. Additive. See note box below for more information.
ROUTER_SLOWLORIS_TIMEOUT	10s	Length of time the transmission of an HTTP request can take.
RELOAD_INTERVAL	5s	The minimum frequency the router is allowed to reload to accept new changes.
ROUTER_METRICS_HAPROXY_TIMEOUT	5s	Timeout for the gathering of HAProxy metrics.

NOTE

Some effective timeout values can be the sum of certain variables, rather than the specific expected timeout.

For example: **ROUTER_SLOWLORIS_HTTP_KEEPALIVE** adjusts **timeout http-keep-alive**, and is set to **300s** by default, but haproxy also waits on **tcp-request inspect-delay**, which is set to **5s**. In this case, the overall timeout would be **300s** plus **5s**.

5.7.6. Load-balancing Strategy

When a route has multiple endpoints, HAProxy distributes requests to the route among the endpoints based on the selected load-balancing strategy. This applies when no persistence information is available, such as on the first request in a session.

The strategy can be one of the following:

- **roundrobin**: Each endpoint is used in turn, according to its weight. This is the smoothest and fairest algorithm when the server's processing time remains equally distributed.

- **leastconn**: The endpoint with the lowest number of connections receives the request. Round-robin is performed when multiple endpoints have the same lowest number of connections. Use this algorithm when very long sessions are expected, such as LDAP, SQL, TSE, or others. Not intended to be used with protocols that typically use short sessions such as HTTP.

- **source**: The source IP address is hashed and divided by the total weight of the running servers to designate which server will receive the request. This ensures that the same client IP address will always reach the same server as long as no server goes down or up. If the hash result changes due to the number of running servers changing, many clients will be directed to different servers. This algorithm is generally used with passthrough routes.

The **ROUTER_TCP_BALANCE_SCHEME** environment variable sets the default strategy for passthorugh routes. The **ROUTER_LOAD_BALANCE_ALGORITHM** environment variable sets the default strategy for the router for the remaining routes. A route specific annotation, **haproxy.router.openshift.io/balance**, can be used to control specific routes.

5.7.7. HAProxy Strict SNI

By default, when a host does not resolve to a route in a HTTPS or TLS SNI request, the default certificate is returned to the caller as part of the 503 response. This exposes the default certificate and can pose security concerns because the wrong certificate is served for a site. The HAProxy **strict-sni** option to bind suppresses use of the default certificate.

The **ROUTER_STRICT_SNI** environment variable controls bind processing. When set to **true** or **TRUE**, **strict-sni** is added to the HAProxy bind. The default setting is **false**.

The option can be set when the router is created or added later.

```
$ oc adm router --strict-sni
```

This sets **ROUTER_STRICT_SNI=true**.

5.7.8. Router Cipher Suite

Each client (for example, Chrome 30, or Java8) includes a suite of ciphers used to securely connect with the router. The router must have at least one of the ciphers for the connection to be complete:

Table 5.3. Router Cipher Profiles

Profile	Oldest compatible client
modern	Firefox 27, Chrome 30, IE 11 on Windows 7, Edge, Opera 17, Safari 9, Android 5.0, Java 8
intermediate	Firefox 1, Chrome 1, IE 7, Opera 5, Safari 1, Windows XP IE8, Android 2.3, Java 7
old	Windows XP IE6, Java 6

See the Security/Server Side TLS reference guide for more information.

The router defaults to the **intermediate** profile. You can select a different profile using the **--ciphers** option when creating a route, or by changing the **ROUTER_CIPHERS** environment variable with the values **modern**, **intermediate**, or **old** for an existing router. Alternatively, a set of ":" separated ciphers can be provided. The ciphers must be from the set displayed by:

```
openssl ciphers
```

5.7.9. Route Host Names

In order for services to be exposed externally, an OpenShift Container Platform route allows you to associate a service with an externally-reachable host name. This edge host name is then used to route traffic to the service.

When multiple routes from different namespaces claim the same host, the oldest route wins and claims it for the namespace. If additional routes with different path fields are defined in the same namespace, those paths are added. If multiple routes with the same path are used, the oldest takes priority.

A consequence of this behavior is that if you have two routes for a host name: an older one and a newer one. If someone else has a route for the same host name that they created between when you created the other two routes, then if you delete your older route, your claim to the host name will no longer be in effect. The other namespace now claims the host name and your claim is lost.

Example 5.1. A Route with a Specified Host:

```
apiVersion: v1
kind: Route
metadata:
  name: host-route
spec:
  host: www.example.com 1
  to:
    kind: Service
    name: service-name
```

1 Specifies the externally-reachable host name used to expose a service.

Example 5.2. A Route Without a Host:

```
apiVersion: v1
kind: Route
metadata:
  name: no-route-hostname
spec:
  to:
    kind: Service
    name: service-name
```

If a host name is not provided as part of the route definition, then OpenShift Container Platform automatically generates one for you. The generated host name is of the form:

```
<route-name>[-<namespace>].<suffix>
```

The following example shows the OpenShift Container Platform-generated host name for the above configuration of a route without a host added to a namespace **mynamespace**:

Example 5.3. Generated Host Name

```
no-route-hostname-mynamespace.router.default.svc.cluster.local 1
```

1 The generated host name suffix is the default routing subdomain **router.default.svc.cluster.local**.

A cluster administrator can also customize the suffix used as the default routing subdomain for their environment.

5.7.10. Route Types

Routes can be either secured or unsecured. Secure routes provide the ability to use several types of TLS termination to serve certificates to the client. Routers support edge, passthrough, and re-encryption termination.

Example 5.4. Unsecured Route Object YAML Definition

```
apiVersion: v1
kind: Route
metadata:
  name: route-unsecured
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name
```

Unsecured routes are simplest to configure, as they require no key or certificates, but secured routes offer security for connections to remain private.

A secured route is one that specifies the TLS termination of the route. The available types of termination are described below.

5.7.10.1. Path Based Routes

Path based routes specify a path component that can be compared against a URL (which requires that the traffic for the route be HTTP based) such that multiple routes can be served using the same host name, each with a different path. Routers should match routes based on the most specific path to the least; however, this depends on the router implementation. The host name and path are passed through to the backend server so it should be able to successfully answer requests for them. For example: a request to http://example.com/foo/ that goes to the router will result in a pod seeing a request to http://example.com/foo/.

The following table shows example routes and their accessibility:

Table 5.4. Route Availability

Route	When Compared to	Accessible
www.example.com/test	*www.example.com/test*	Yes
	www.example.com	No
www.example.com/test and *www.example.com*	*www.example.com/test*	Yes
	www.example.com	Yes
www.example.com	*www.example.com/test*	Yes (Matched by the host, not the route)
	www.example.com	Yes

Example 5.5. An Unsecured Route with a Path:

```
apiVersion: v1
kind: Route
metadata:
  name: route-unsecured
spec:
  host: www.example.com
  path: "/test"  ①
  to:
    kind: Service
    name: service-name
```

① The path is the only added attribute for a path-based route.

NOTE

Path-based routing is not available when using passthrough TLS, as the router does not terminate TLS in that case and cannot read the contents of the request.

5.7.10.2. Secured Routes

Secured routes specify the TLS termination of the route and, optionally, provide a key and certificate(s).

NOTE

TLS termination in OpenShift Container Platform relies on SNI for serving custom certificates. Any non-SNI traffic received on port 443 is handled with TLS termination and a default certificate (which may not match the requested host name, resulting in validation errors).

Secured routes can use any of the following three types of secure TLS termination.

Edge Termination

With edge termination, TLS termination occurs at the router, prior to proxying traffic to its destination. TLS certificates are served by the front end of the router, so they must be configured into the route, otherwise the router's default certificate will be used for TLS termination.

Example 5.6. A Secured Route Using Edge Termination

```
apiVersion: v1
kind: Route
metadata:
  name: route-edge-secured ①
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name ②
  tls:
    termination: edge          ③
    key: |-                ④
      -----BEGIN PRIVATE KEY-----
      [...]
      -----END PRIVATE KEY-----
    certificate: |-            ⑤
      -----BEGIN CERTIFICATE-----
      [...]
      -----END CERTIFICATE-----
    caCertificate: |-          ⑥
      -----BEGIN CERTIFICATE-----
      [...]
      -----END CERTIFICATE-----
```

① ② The name of the object, which is limited to 63 characters.

③ The **termination** field is **edge** for edge termination.

4 The **key** field is the contents of the PEM format key file.

5 The **certificate** field is the contents of the PEM format certificate file.

6 An optional CA certificate may be required to establish a certificate chain for validation.

Because TLS is terminated at the router, connections from the router to the endpoints over the internal network are not encrypted.

Edge-terminated routes can specify an **insecureEdgeTerminationPolicy** that enables traffic on insecure schemes (**HTTP**) to be disabled, allowed or redirected. The allowed values for **insecureEdgeTerminationPolicy** are: **None** or empty (for disabled), **Allow** or **Redirect**. The default **insecureEdgeTerminationPolicy** is to disable traffic on the insecure scheme. A common use case is to allow content to be served via a secure scheme but serve the assets (example images, stylesheets and javascript) via the insecure scheme.

Example 5.7. A Secured Route Using Edge Termination Allowing HTTP Traffic

```
apiVersion: v1
kind: Route
metadata:
  name: route-edge-secured-allow-insecure 1
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name 2
  tls:
    termination:                   edge  3
    insecureEdgeTerminationPolicy: Allow  4
    [ ... ]
```

1 2 The name of the object, which is limited to 63 characters.

3 The **termination** field is **edge** for edge termination.

4 The insecure policy to allow requests sent on an insecure scheme **HTTP**.

Example 5.8. A Secured Route Using Edge Termination Redirecting HTTP Traffic to HTTPS

```
apiVersion: v1
kind: Route
metadata:
  name: route-edge-secured-redirect-insecure 1
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name 2
  tls:
    termination:                edge    ❸
    insecureEdgeTerminationPolicy: Redirect ❹
    [ ... ]
```

❶ ❷ The name of the object, which is limited to 63 characters.

❸ The **termination** field is **edge** for edge termination.

❹ The insecure policy to redirect requests sent on an insecure scheme **HTTP** to a secure scheme **HTTPS**.

Passthrough Termination

With passthrough termination, encrypted traffic is sent straight to the destination without the router providing TLS termination. Therefore no key or certificate is required.

Example 5.9. A Secured Route Using Passthrough Termination

```
apiVersion: v1
kind: Route
metadata:
  name: route-passthrough-secured ❶
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name ❷
  tls:
    termination: passthrough    ❸
```

❶ ❷ The name of the object, which is limited to 63 characters.

❸ The **termination** field is set to **passthrough**. No other encryption fields are needed.

The destination pod is responsible for serving certificates for the traffic at the endpoint. This is currently the only method that can support requiring client certificates (also known as two-way authentication).

NOTE

Passthrough routes can also have an **insecureEdgeTerminationPolicy**. The only valid values are **None** (or empty, for disabled) or **Redirect**.

Re-encryption Termination

Re-encryption is a variation on edge termination where the router terminates TLS with a certificate, then re-encrypts its connection to the endpoint which may have a different certificate. Therefore the full path of the connection is encrypted, even over the internal network. The router uses health checks to determine the authenticity of the host.

Example 5.10. A Secured Route Using Re-Encrypt Termination

```
apiVersion: v1
kind: Route
metadata:
  name: route-pt-secured 1
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name 2
  tls:
    termination: reencrypt        3
    key: [as in edge termination]
    certificate: [as in edge termination]
    caCertificate: [as in edge termination]
    destinationCACertificate: |-  4
      -----BEGIN CERTIFICATE-----
      [...]
      -----END CERTIFICATE-----
```

1 2 The name of the object, which is limited to 63 characters.

3 The **termination** field is set to **reencrypt**. Other fields are as in edge termination.

4 Required for re-encryption. **destinationCACertificate** specifies a CA certificate to validate the endpoint certificate, securing the connection from the router to the destination pods. If the service is using a service signing certificate, or the administrator has specified a default CA certificate for the router and the service has a certificate signed by that CA, this field can be omitted.

If the **destinationCACertificate** field is left empty, the router automatically leverages the certificate authority that is generated for service serving certificates, and is injected into every pod as **/var/run/secrets/kubernetes.io/serviceaccount/service-ca.crt**. This allows new routes that leverage end-to-end encryption without having to generate a certificate for the route. This is useful for custom routers or the F5 router, which might not allow the **destinationCACertificate** unless the administrator has allowed it.

NOTE

Re-encrypt routes can have an **insecureEdgeTerminationPolicy** with all of the same values as edge-terminated routes.

5.7.11. Router Sharding

In OpenShift Container Platform, each route can have any number of labels in its **metadata** field. A router uses *selectors* (also known as a *selection expression*) to select a subset of routes from the entire pool of routes to serve. A selection expression can also involve labels on the route's namespace. The selected routes form a *router shard*. You can create and modify router shards independently from the routes, themselves.

This design supports *traditional* sharding as well as *overlapped* sharding. In traditional sharding, the selection results in no overlapping sets and a route belongs to exactly one shard. In overlapped sharding, the selection results in overlapping sets and a route can belong to many different shards. For example, a single route may belong to a **SLA=high** shard (but not **SLA=medium** or **SLA=low** shards), as well as a **geo=west** shard (but not a **geo=east** shard).

Another example of overlapped sharding is a set of routers that select based on namespace of the route:

Router	Selection	Namespaces
router-1	**A* – J***	**A***, **B***, **C***, **D***, **E***, **F***, **G***, **H***, **I***, **J***
router-2	**K* – T***	**K***, **L***, **M***, **N***, **O***, **P***, **Q***, **R***, **S***, **T***
router-3	**Q* – Z***	**Q***, **R***, **S***, **T***, **U***, **V***, **W***, **X***, **Y***, **Z***

Both **router-2** and **router-3** serve routes that are in the namespaces **Q***, **R***, **S***, **T***. To change this example from overlapped to traditional sharding, we could change the selection of **router-2** to **K* – P***, which would eliminate the overlap.

When routers are sharded, a given route is bound to zero or more routers in the group. The route binding ensures uniqueness of the route across the shard. Uniqueness allows secure and non-secure versions of the same route to exist within a single shard. This implies that routes now have a visible life cycle that moves from created to bound to active.

In the sharded environment the first route to hit the shard reserves the right to exist there indefinitely, even across restarts.

During a green/blue deployment a route may be be selected in multiple routers. An OpenShift Container Platform application administrator may wish to bleed traffic from one version of the application to another and then turn off the old version.

Sharding can be done by the administrator at a cluster level and by the user at a project/namespace level. When namespace labels are used, the service account for the router must have **cluster-reader** permission to permit the router to access the labels in the namespace.

NOTE

For two or more routes that claim the same host name, the resolution order is based on the age of the route and the oldest route would win the claim to that host. In the case of sharded routers, routes are selected based on their labels matching the router's selection criteria. There is no consistent way to determine when labels are added to a route. So if an older route claiming an existing host name is "re-labelled" to match the router's selection criteria, it will replace the existing route based on the above mentioned resolution order (oldest route wins).

5.7.12. Alternate Backends and Weights

A route is usually associated with one service through the **to:** token with **kind: Service**. All of the requests to the route are handled by endpoints in the service based on the load balancing strategy.

It is possible to have as many as four services supporting the route. The portion of requests that are handled by each service is governed by the service **weight**.

The first service is entered using the **to:** token as before, and up to three additional services can be entered using the **alternateBackend:** token. Each service must be **kind: Service** which is the default.

Each service has a **weight** associated with it. The portion of requests handled by the service is **weight** / **sum_of_all_weights**. When a service has more than one endpoint, the service's weight is distributed among the endpoints with each endpoint getting at least 1. If the service **weight** is 0 each of the service's endpoints will get 0.

The **weight** must be in the range 0-256. The default is 1. When the **weight** is 0 no requests are passed to the service. If all services have **weight** 0, requests are returned with a 503 error. When a service has no endpoints, the weight is effectively 0.

When using **alternateBackends** also use the **roundrobin** load balancing strategy to ensure requests are distributed as expected to the services based on **weight**. **roundrobin** can be set for a route using a route annotation, or for the router in general using an environment variable.

The following is an example route configuration using alternate backends for A/B deployments.

A Route with alternateBackends and weights:

```
apiVersion: v1
kind: Route
metadata:
  name: route-alternate-service
  annotations:
    haproxy.router.openshift.io/balance: roundrobin 1
spec:
  host: www.example.com
  to:
    kind: Service
    name: service-name 2
    weight: 20        3
  alternateBackends:
  - kind: Service
    name: service-name2 4
    weight: 10        5
  - kind: Service
    name: service-name3 6
    weight: 10        7
```

1 This route uses **roundrobin**load balancing strategy.

2 The first service name is **service-name** which may have 0 or more pods

4 6 The alternateBackend services may also have 0 or more pods

3 5 7 The total **weight** is 40. **service-name** will get 20/40 or 1/2 of the requests, **service-name2** and **service-name3** will each get 1/4 of the requests, assuming each service has 1 or more endpoints.

5.7.13. Route-specific Annotations

Using environment variables, a router can set the default options for all the routes it exposes. An individual route can override some of these defaults by providing specific configurations in its annotations.

Route Annotations

For all the items outlined in this section, you can set annotations on the **route definition** for the route to alter its configuration

Table 5.5. Route Annotations

Variable	Description	Environment Variable Used as Default
haproxy.router.openshift.io/balance	Sets the load-balancing algorithm. Available options are **source**, **roundrobin**, and **leastconn**.	**ROUTER_TCP_BALANCE_SCHEME** for passthrough routes. Otherwise, use **ROUTER_LOAD_BALANCE_ALGORITHM**.
haproxy.router.openshift.io/disable_cookies	Disables the use of cookies to track related connections. If set to **true** or **TRUE**, the balance algorithm is used to choose which back-end serves connections for each incoming HTTP request.	
haproxy.router.openshift.io/cookie_name	Specifies an optional cookie to be used for this route. The name must consist of any combination of upper and lower case letters, digits, "_", and "-". The default is the hashed internal key name for the route.	
haproxy.router.openshift.io/rate-limit-connections	Setting **true** or **TRUE** to enables rate limiting functionality.	
haproxy.router.openshift.io/rate-limit-connections.concurrent-tcp	Limits the number of concurrent TCP connections shared by an IP address.	
haproxy.router.openshift.io/rate-limit-connections.rate-http	Limits the rate at which an IP address can make HTTP requests.	
haproxy.router.openshift.io/rate-limit-connections.rate-tcp	Limits the rate at which an IP address can make TCP connections.	
haproxy.router.openshift.io/timeout	Sets a server-side timeout for the route. (TimeUnits)	**ROUTER_DEFAULT_SERVER_TIMEOUT**

Variable	Description	Environment Variable Used as Default
router.openshift.io/haproxy.health.check.interval	Sets the interval for the back-end health checks. (TimeUnits)	**ROUTER_BACKEND_CHECK_INTERVAL**
haproxy.router.openshift.io/ip_whitelist	Sets a whitelist for the route.	
haproxy.router.openshift.io/hsts_header	Sets a Strict-Transport-Security header for the edge terminated or re-encrypt route.	

Example 5.11. A Route Setting Custom Timeout

```
apiVersion: v1
kind: Route
metadata:
  annotations:
    haproxy.router.openshift.io/timeout: 5500ms ①
[...]
```

① Specifies the new timeout with HAProxy supported units (us, ms, s, m, h, d). If unit not provided, ms is the default.

NOTE

Setting a server-side timeout value for passthrough routes too low can cause WebSocket connections to timeout frequently on that route.

5.7.14. Route-specific IP Whitelists

You can restrict access to a route to a select set of IP addresses by adding the **haproxy.router.openshift.io/ip_whitelist** annotation on the route. The whitelist is a space-separated list of IP addresses and/or CIDRs for the approved source addresses. Requests from IP addresses that are not in the whitelist are dropped.

Some examples:

When editing a route, add the following annotation to define the desired source IP's. Alternatively, use **oc annotate route <name>**.

Allow only one specific IP address:

```
metadata:
  annotations:
    haproxy.router.openshift.io/ip_whitelist: 192.168.1.10
```

Allow several IP addresses:

```
metadata:
  annotations:
    haproxy.router.openshift.io/ip_whitelist: 192.168.1.10 192.168.1.11 192.168.1.12
```

Allow an IP CIDR network:

```
metadata:
  annotations:
    haproxy.router.openshift.io/ip_whitelist: 192.168.1.0/24
```

Allow mixed IP addresses and IP CIDR networks:

```
metadata:
  annotations:
    haproxy.router.openshift.io/ip_whitelist: 180.5.61.153 192.168.1.0/24 10.0.0.0/8
```

5.7.15. Creating Routes Specifying a Wildcard Subdomain Policy

A wildcard policy allows a user to define a route that covers all hosts within a domain (when the router is configured to allow it). A route can specify a wildcard policy as part of its configuration using the **wildcardPolicy** field. Any routers run with a policy allowing wildcard routes will expose the route appropriately based on the wildcard policy.

Learn how to configure HAProxy routers to allow wildcard routes .

Example 5.12. A Route Specifying a Subdomain WildcardPolicy

```
apiVersion: v1
kind: Route
spec:
  host: wildcard.example.com  1
  wildcardPolicy: Subdomain  2
  to:
    kind: Service
    name: service-name
```

1 Specifies the externally reachable host name used to expose a service.

2 Specifies that the externally reachable host name should allow all hosts in the subdomain **example.com**. ***.example.com** is the subdomain for host name **wildcard.example.com** to reach the exposed service.

5.7.16. Route Status

The **route status** field is only set by routers. If changes are made to a route so that a router no longer serves a specific route, the status becomes stale. The routers do not clear the **route status** field. To remove the stale entries in the route status, use the clear-route-status script.

5.7.17. Denying or Allowing Certain Domains in Routes

A router can be configured to deny or allow a specific subset of domains from the host names in a route using the **ROUTER_DENIED_DOMAINS** and **ROUTER_ALLOWED_DOMAINS** environment variables.

ROUTER_DENIED_DOMAINS	Domains listed are not allowed in any indicated routes.
ROUTER_ALLOWED_DOMAINS	Only the domains listed are allowed in any indicated routes.

The domains in the list of denied domains take precedence over the list of allowed domains. Meaning OpenShift Container Platform first checks the deny list (if applicable), and if the host name is not in the list of denied domains, it then checks the list of allowed domains. However, the list of allowed domains is more restrictive, and ensures that the router only admits routes with hosts that belong to that list.

For example, to deny the **[*.]open.header.test**, **[*.]openshift.org** and **[*.]block.it** routes for the **myrouter** route:

```
$ oc adm router myrouter ...
$ oc set env dc/myrouter ROUTER_DENIED_DOMAINS="open.header.test, openshift.org, block.it"
```

This means that **myrouter** will admit the following based on the route's name:

```
$ oc expose service/<name> --hostname="foo.header.test"
$ oc expose service/<name> --hostname="www.allow.it"
$ oc expose service/<name> --hostname="www.openshift.test"
```

However, **myrouter** will deny the following:

```
$ oc expose service/<name> --hostname="open.header.test"
$ oc expose service/<name> --hostname="www.open.header.test"
$ oc expose service/<name> --hostname="block.it"
$ oc expose service/<name> --hostname="franco.baresi.block.it"
$ oc expose service/<name> --hostname="openshift.org"
$ oc expose service/<name> --hostname="api.openshift.org"
```

Alternatively, to block any routes where the host name is *not* set to **[*.]stickshift.org** or **[*.]kates.net**:

```
$ oc adm router myrouter ...
$ oc set env dc/myrouter ROUTER_ALLOWED_DOMAINS="stickshift.org, kates.net"
```

This means that the **myrouter** router will admit:

```
$ oc expose service/<name> --hostname="stickshift.org"
$ oc expose service/<name> --hostname="www.stickshift.org"
$ oc expose service/<name> --hostname="kates.net"
$ oc expose service/<name> --hostname="api.kates.net"
$ oc expose service/<name> --hostname="erno.r.kube.kates.net"
```

However, **myrouter** will deny the following:

```
$ oc expose service/<name> --hostname="www.open.header.test"
$ oc expose service/<name> --hostname="drive.ottomatic.org"
$ oc expose service/<name> --hostname="www.wayless.com"
$ oc expose service/<name> --hostname="www.deny.it"
```

To implement both scenarios, run:

```
$ oc adm router adrouter ...
$ oc env dc/adrouter ROUTER_ALLOWED_DOMAINS="openshift.org, kates.net" \
    ROUTER_DENIED_DOMAINS="ops.openshift.org, metrics.kates.net"
```

This will allow any routes where the host name is set to **[*.]openshift.org** or **[*.]kates.net**, and not allow any routes where the host name is set to **[*.]ops.openshift.org** or **[*.]metrics.kates.net**.

Therefore, the following will be denied:

```
$ oc expose service/<name> --hostname="www.open.header.test"
$ oc expose service/<name> --hostname="ops.openshift.org"
$ oc expose service/<name> --hostname="log.ops.openshift.org"
$ oc expose service/<name> --hostname="www.block.it"
$ oc expose service/<name> --hostname="metrics.kates.net"
$ oc expose service/<name> --hostname="int.metrics.kates.net"
```

However, the following will be allowed:

```
$ oc expose service/<name> --hostname="openshift.org"
$ oc expose service/<name> --hostname="api.openshift.org"
$ oc expose service/<name> --hostname="m.api.openshift.org"
$ oc expose service/<name> --hostname="kates.net"
$ oc expose service/<name> --hostname="api.kates.net"
```

5.7.18. Disabling the Namespace Ownership Check

Hosts and subdomains are owned by the namespace of the route that first makes the claim. Other routes created in the namespace can make claims on the subdomain. All other namespaces are prevented from making claims on the claimed hosts and subdomains. The namespace that owns the host also owns all paths associated with the host, for example ***www.abc.xyz/path1***.

For example, if the host ***www.abc.xyz*** is not claimed by any route. Creating route **r1** with host ***www.abc.xyz*** in namespace **ns1** makes namespace **ns1** the owner of host ***www.abc.xyz*** and subdomain **abc.xyz** for wildcard routes. If another namespace, **ns2**, tries to create a route with say a different path ***www.abc.xyz/path1/path2***, it would fail because a route in another namespace (**ns1** in this case) owns that host.

With wildcard routes the namespace that owns the subdomain owns all hosts in the subdomain. If a namespace owns subdomain **abc.xyz** as in the above example, another namespace cannot claim **z.abc.xyz**.

By disabling the namespace ownership rules, you can disable these restrictions and allow hosts (and subdomains) to be claimed across namespaces.

WARNING

If you decide to disable the namespace ownership checks in your router, be aware that this allows end users to claim ownership of hosts across namespaces. While this change can be desirable in certain development environments, use this feature with caution in production environments, and ensure that your cluster policy has locked down untrusted end users from creating routes.

For example, with **ROUTER_DISABLE_NAMESPACE_OWNERSHIP_CHECK=true**, if namespace **ns1** creates the oldest route **r1** ***www.abc.xyz***, it owns only the hostname (+ path). Another namespace can create a wildcard route even though it does not have the oldest route in that subdomain (**abc.xyz**) and we could potentially have other namespaces claiming other non-wildcard overlapping hosts (for example, **foo.abc.xyz**, **bar.abc.xyz**, **baz.abc.xyz**) and their claims would be granted.

Any other namespace (for example, **ns2**) can now create a route **r2** ***www.abc.xyz/p1/p2***, and it would be admitted. Similarly another namespace (**ns3**) can also create a route **wildthing.abc.xyz** with a subdomain wildcard policy and it can own the wildcard.

As this example demonstrates, the policy **ROUTER_DISABLE_NAMESPACE_OWNERSHIP_CHECK=true** is more lax and allows claims across namespaces. The only time the router would reject a route with the namespace ownership disabled is if the host+path is already claimed.

For example, if a new route **rx** tries to claim ***www.abc.xyz/p1/p2***, it would be rejected as route **r2** owns that host+path combination. This is true whether route **rx** is in the same namespace or other namespace since the exact host+path is already claimed.

This feature can be set during router creation or by setting an environment variable in the router's deployment configuration.

```
$ oc adm router ... --disable-namespace-ownership-check=true
```

```
$ oc env dc/router ROUTER_DISABLE_NAMESPACE_OWNERSHIP_CHECK=true
```

[1] After this point, device names refer to devices on container B's host.

CHAPTER 6. SERVICE CATALOG COMPONENTS

6.1. SERVICE CATALOG

6.1.1. Overview

When developing microservices-based applications to run on cloud native platforms, there are many ways to provision different resources and share their coordinates, credentials, and configuration, depending on the service provider and the platform.

To give developers a more seamless experience, OpenShift Container Platform includes a *service catalog*, an implementation of the Open Service Broker API (OSB API) for Kubernetes. This allows users to connect any of their applications deployed in OpenShift Container Platform to a wide variety of service brokers.

The service catalog allows cluster administrators to integrate multiple platforms using a single API specification. The OpenShift Container Platform web console displays the cluster service classes offered by service brokers in the service catalog, allowing users to discover and instantiate those services for use with their applications.

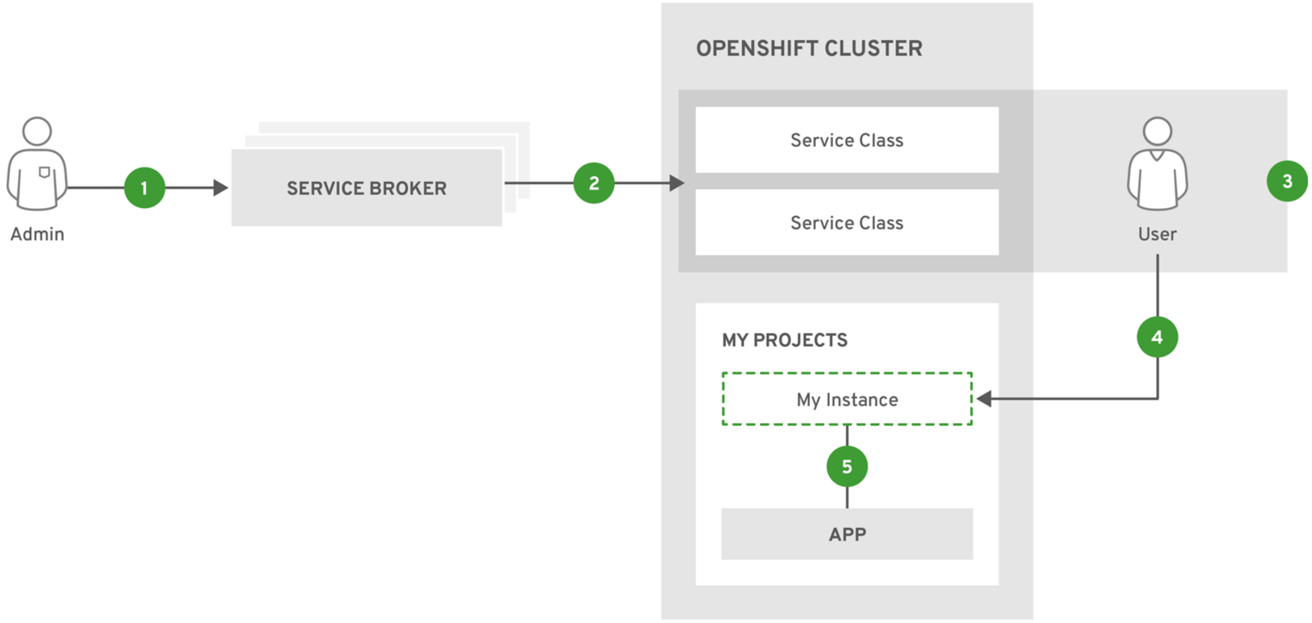
As a result, service users benefit from ease and consistency of use across different types of services from different providers, while service providers benefit from having one integration point that gives them access to multiple platforms.

6.1.2. Design

The design of the service catalog follows this basic workflow:

NOTE

New terms in the following are defined further in Concepts and Terminology.



OPENSHIFT_415489_0218

A cluster administrator registers one or more *cluster service brokers* with their OpenShift Container Platform cluster. This can be done automatically during installation for some default-provided service brokers or manually.

Each service broker specifies a set of *cluster service classes* and variations of those services (*service plans*) to OpenShift Container Platform that should be made available to users.

Using the OpenShift Container Platform web console or CLI, users discover the services that are available. For example, a cluster service class may be available that is a database-as-a-service called BestDataBase.

A user chooses a cluster service class and requests a new *instance* of their own. For example, a service instance may be a BestDataBase instance named **my_db**.

A user links, or *binds*, their service instance to a set of pods (their application). For example, the **my_db** service instance may be bound to the user's application called **my_app**.

When a user makes a request to provision or deprovision a resource, the request is made to the service catalog, which then sends a request to the appropriate cluster service broker. With some services, some operations such as **provision**, **deprovision**, and **update** are expected to take some time to fulfill. If the cluster service broker is unavailable, the service catalog will continue to retry the operation.

This infrastructure allows a loose coupling between applications running in OpenShift Container Platform and the services they use. This allows the application that uses those services to focus on its own business logic while leaving the management of these services to the provider.

6.1.2.1. Deleting Resources

When a user is done with a service (or perhaps no longer wishes to be billed), the service instance can be deleted. In order to delete the service instance, the service bindings must be removed first. Deleting the service bindings is known as *unbinding*. Part of the deletion process includes deleting the secret that references the service binding being deleted.

Once all the service bindings are removed, the service instance may be deleted. Deleting the service instance is known as *deprovisioning*.

If a project or namespace containing service bindings and service instances is deleted, the service catalog must first request the cluster service broker to delete the associated instances and bindings. This is expected to delay the actual deletion of the project or namespace since the service catalog must communicate with cluster service brokers and wait for them to perform their deprovisioning work. In normal circumstances, this may take several minutes or longer depending on the service.

NOTE

If you delete a service binding used by a deployment, you must also remove any references to the binding secret from the deployment. Otherwise, the next rollout will fail.

6.1.3. Concepts and Terminology

Cluster Service Broker

A *cluster service broker* is a server that conforms to the OSB API specification and manages a set of one or more services. The software could be hosted within your own OpenShift Container Platform cluster or elsewhere.
Cluster administrators can create **ClusterServiceBroker** API resources representing cluster service brokers and register them with their OpenShift Container Platform cluster. This allows cluster

administrators to make new types of managed services using that cluster service broker available within their cluster.

A **ClusterServiceBroker** resource specifies connection details for a cluster service broker and the set of services (and variations of those services) to OpenShift Container Platform that should then be made available to users. Of special note is the **authInfo** section, which contains the data used to authenticate with the cluster service broker.

Example **ClusterServiceBroker** Resource

```
apiVersion: servicecatalog.k8s.io/v1beta1
kind: ClusterServiceBroker
metadata:
  name: BestCompanySaaS
spec:
  url: http://bestdatabase.example.com
  authInfo:
    basic:
      secretRef:
        namespace: test-ns
        name: secret-name
```

Cluster Service Class

Also synonymous with "service" in the context of the service catalog, a *cluster service class* is a type of managed service offered by a particular cluster service broker. Each time a new cluster service broker resource is added to the cluster, the service catalog controller connects to the corresponding cluster service broker to obtain a list of service offerings. A new **ClusterServiceClass** resource is automatically created for each.

> **NOTE**
>
> OpenShift Container Platform also has a core concept called services, which are separate Kubernetes resources related to internal load balancing. These resources are not to be confused with how the term is used in the context of the service catalog and OSB API.

Example **ClusterServiceClass** Resource

```
apiVersion: servicecatalog.k8s.io/v1beta1
kind: ClusterServiceClass
metadata:
  name: smallDB
  brokerName: BestDataBase
  plans: [...]
```

Cluster Service Plan

A *cluster service plan* is represents tiers of a cluster service class. For example, a cluster service class may expose a set of plans that offer varying degrees of quality-of-service (QoS), each with a different cost associated with it.

Service Instance

A *service instance* is a provisioned instance of a cluster service class. When a user wants to use the capability provided by a service class, they can create a new service instance.

When a new **ServiceInstance** resource is created, the service catalog controller connects to the appropriate cluster service broker and instructs it to provision the service instance.

Example ServiceInstance Resource

```
apiVersion: servicecatalog.k8s.io/v1beta1
kind: ServiceInstance
metadata:
  name: my_db
  namespace: test-ns
spec:
  externalClusterServiceClassName: smallDB
  externalClusterServicePlanName: default
```

Application

The term *application* refers to the OpenShift Container Platform deployment artifacts, for example pods running in a user’s project, that will use a *service instance*.

Credentials

Credentials are information needed by an application to communicate with a service instance.

Service Binding

A *service binding* is a link between a service instance and an application. These are created by cluster users who wish for their applications to reference and use a service instance.
Upon creation, the service catalog controller creates a Kubernetes secret containing connection details and credentials for the service instance. Such secrets can be mounted into pods as usual. There is also integration with **PodPresets**, which allow you to express how the secret should be consumed, and in which pods.

Example ServiceBinding Resource

```
apiVersion: servicecatalog.k8s.io/v1beta1
kind: ServiceBinding
metadata:
  name: myBinding
  namespace: test-ns
spec:
  instanceRef:
    name: my_db
  parameters:
    securityLevel: confidential
  secretName: mySecret
```

Parameters

A *parameter* is a special field available to pass additional data to the cluster service broker when using either service bindings or service instances. The only formatting requirement is for the parameters to be valid YAML (or JSON). In the above example, a security level parameter is passed to the cluster service broker in the service binding request. For parameters that need more security, place them in a secret and reference them using **parametersFrom**.

Example Service Binding Resource Referencing a Secret

```
apiVersion: servicecatalog.k8s.io/v1beta1
kind: ServiceBinding
```

```
metadata:
  name: myBinding
  namespace: test-ns
spec:
  instanceRef:
    name: my_db
  parametersFrom:
    - secretKeyRef:
        name: securityLevel
        key: myKey
  secretName: mySecret
```

6.1.4. Provided Cluster Service Brokers

OpenShift Container Platform provides the following cluster service brokers for use with the service catalog.

- Template Service Broker
- OpenShift Ansible Broker

6.2. TEMPLATE SERVICE BROKER

The *template service broker* (TSB) gives the service catalog visibility into the default Instant App and Quickstart templates that have shipped with OpenShift Container Platform since its initial release. The TSB can also make available as a service anything for which an OpenShift Container Platform template has been written, whether provided by Red Hat, a cluster administrator or user, or a third party vendor.

By default, the TSB shows the objects that are globally available from the **openshift** project. It can also be configured to watch any other project that a cluster administrator chooses.

6.3. OPENSHIFT ANSIBLE BROKER

6.3.1. Overview

The OpenShift Ansible broker (OAB) is an implementation of the Open Service Broker (OSB) API that manages applications defined by *Ansible playbook bundles (APBs)* . APBs provide a new method for defining and distributing container applications in OpenShift Container Platform, consisting of a bundle of Ansible playbooks built into a container image with an Ansible runtime. APBs leverage Ansible to create a standard mechanism for automating complex deployments.

The design of the OAB follows this basic workflow:

1. A user requests list of available applications from the service catalog using the OpenShift Container Platform web console.
2. The service catalog requests the OAB for available applications.
3. The OAB communicates with a defined container registry to learn which APBs are available.
4. The user issues a request to provision a specific APB.

5. The provision request makes its way to the OAB, which fulfills the user's request by invoking the provision method on the APB.

6.3.2. Ansible Playbook Bundles

An Ansible playbook bundle (APB) is a lightweight application definition that allows you to leverage existing investment in Ansible roles and playbooks.

APBs use a simple directory with named playbooks to perform OSB API actions, such as provision and bind. Metadata defined in ***apb.yml*** spec file contains a list of required and optional parameters for use during deployment.

See the APB Development Guide for details on the overall design and how APBs are written.