



# **OpenShift Container Platform 3.6**

## **Cluster Administration**

OpenShift Container Platform 3.6 Cluster Administration



# OpenShift Container Platform 3.6 Cluster Administration

---

OpenShift Container Platform 3.6 Cluster Administration

## Legal Notice

Copyright © 2018 Red Hat, Inc.

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

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

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

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

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

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

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

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

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

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

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

All other trademarks are the property of their respective owners.

## Abstract

OpenShift Cluster Administration topics cover the day to day tasks for managing your OpenShift cluster and other advanced configuration topics.

# Table of Contents

|  |           |
|--|-----------|
| <b>CHAPTER 1. OVERVIEW</b>                                 | <b>12</b> |
| <b>CHAPTER 2. MANAGING NODES</b>                           | <b>13</b> |
| 2.1. OVERVIEW  | 13        |
| 2.2. LISTING NODES   | 13        |
| 2.3. ADDING NODES  | 15        |
| 2.4. DELETING NODES  | 15        |
| 2.5. UPDATING LABELS ON NODES                              | 16        |
| 2.6. LISTING PODS ON NODES                                 | 16        |
| 2.7. MARKING NODES AS UNSCHEDULABLE OR SCHEDULABLE         | 16        |
| 2.8. EVACUATING PODS ON NODES                              | 17        |
| 2.9. REBOOTING NODES                                       | 18        |
| 2.9.1. Infrastructure Nodes                                | 18        |
| 2.9.2. Using Pod Anti-affinity                             | 18        |
| 2.9.3. Handling Nodes Running Routers                      | 19        |
| 2.10. CONFIGURING NODE RESOURCES                           | 19        |
| 2.10.1. Setting Maximum Pods Per Node                      | 20        |
| 2.11. RESETTING DOCKER STORAGE                             | 21        |
| 2.12. CHANGING NODE TRAFFIC INTERFACE                      | 22        |
| <b>CHAPTER 3. MANAGING USERS</b>                           | <b>24</b> |
| 3.1. OVERVIEW  | 24        |
| 3.2. ADDING A USER   | 24        |
| 3.3. VIEWING USER AND IDENTITY LISTS                       | 24        |
| 3.4. MANAGING USER AND GROUP LABELS                        | 24        |
| 3.5. DELETING A USER                                       | 25        |
| <b>CHAPTER 4. MANAGING PROJECTS</b>                        | <b>26</b> |
| 4.1. OVERVIEW  | 26        |
| 4.2. SELF-PROVISIONING PROJECTS                            | 26        |
| 4.2.1. Modifying the Template for New Projects             | 26        |
| 4.2.2. Disabling Self-provisioning                         | 27        |
| 4.3. USING NODE SELECTORS                                  | 28        |
| 4.3.1. Setting the Cluster-wide Default Node Selector      | 28        |
| 4.3.2. Setting the Project-wide Node Selector              | 28        |
| 4.3.3. Developer-specified Node Selectors                  | 29        |
| 4.4. LIMITING NUMBER OF SELF-PROVISIONED PROJECTS PER USER | 29        |
| <b>CHAPTER 5. MANAGING PODS</b>                            | <b>31</b> |
| 5.1. OVERVIEW  | 31        |
| 5.2. LIMITING RUN-ONCE POD DURATION                        | 31        |
| 5.2.1. Configuring the RunOnceDuration Plug-in             | 31        |
| 5.2.2. Specifying a Custom Duration per Project            | 31        |
| 5.2.2.1. Deploying an Egress Router Pod                    | 32        |
| 5.2.2.2. Deploying an Egress Router Service                | 33        |
| 5.2.3. Limiting Pod Access with Egress Firewall            | 33        |
| 5.2.3.1. Configuring Pod Access Limits                     | 34        |
| 5.3. LIMITING THE BANDWIDTH AVAILABLE TO PODS              | 35        |
| 5.4. SETTING POD DISRUPTION BUDGETS                        | 36        |
| <b>CHAPTER 6. MANAGING NETWORKING</b>                      | <b>38</b> |
| 6.1. OVERVIEW  | 38        |
| 6.2. MANAGING POD NETWORKS                                 | 38        |

|  |           |
|--|-----------|
| 6.2.1. Joining Project Networks  | 38        |
| 6.3. ISOLATING PROJECT NETWORKS  | 38        |
| 6.3.1. Making Project Networks Global  | 38        |
| 6.4. DISABLING HOST NAME COLLISION PREVENTION FOR ROUTES AND INGRESS OBJECTS       | 39        |
| 6.5. CONTROLLING EGRESS TRAFFIC  | 40        |
| 6.5.1. Using an Egress Firewall to Limit Access to External Resources              | 40        |
| 6.5.2. Using an Egress Router to Allow External Resources to Recognize Pod Traffic | 43        |
| 6.5.2.1. Deploying an Egress Router Pod in Redirect Mode                           | 44        |
| 6.5.2.2. Redirecting to Multiple Destinations                                      | 45        |
| 6.5.2.3. Using a ConfigMap to specify EGRESS_DESTINATION                           | 46        |
| 6.5.2.4. Deploying an Egress Router HTTP Proxy Pod                                 | 47        |
| 6.5.2.5. Enabling Failover for Egress Router Pods                                  | 50        |
| 6.5.3. Using iptables Rules to Limit Access to External Resources                  | 51        |
| 6.6. ENABLING MULTICAST  | 51        |
| 6.7. ENABLING NETWORKPOLICY  | 52        |
| 6.7.1. NetworkPolicy and Routers   | 53        |
| 6.7.2. Setting a Default NetworkPolicy for New Projects                            | 55        |
| 6.8. TROUBLESHOOTING THROUGHPUT ISSUES   | 55        |
| <b>CHAPTER 7. CONFIGURING SERVICE ACCOUNTS</b>                                     | <b>57</b> |
| 7.1. OVERVIEW  | 57        |
| 7.2. USER NAMES AND GROUPS   | 57        |
| 7.3. MANAGING SERVICE ACCOUNTS   | 58        |
| 7.4. ENABLING SERVICE ACCOUNT AUTHENTICATION                                       | 58        |
| 7.5. MANAGED SERVICE ACCOUNTS  | 59        |
| 7.6. INFRASTRUCTURE SERVICE ACCOUNTS   | 60        |
| 7.7. SERVICE ACCOUNTS AND SECRETS  | 60        |
| <b>CHAPTER 8. MANAGING AUTHORIZATION POLICIES</b>                                  | <b>61</b> |
| 8.1. OVERVIEW  | 61        |
| 8.2. VIEWING ROLES AND BINDINGS  | 61        |
| 8.2.1. Viewing Cluster Policy  | 61        |
| 8.2.2. Viewing Local Policy  | 69        |
| 8.3. MANAGING ROLE BINDINGS  | 70        |
| 8.4. GRANTING USERS DAEMONSET PERMISSIONS  | 71        |
| 8.5. CREATING A LOCAL ROLE   | 72        |
| <b>CHAPTER 9. IMAGE POLICY</b>   | <b>74</b> |
| 9.1. OVERVIEW  | 74        |
| 9.2. CONFIGURING THE IMAGEPOLICY ADMISSION PLUG-IN                                 | 74        |
| 9.3. TESTING THE IMAGEPOLICY ADMISSION PLUG-IN                                     | 76        |
| <b>CHAPTER 10. IMAGE SIGNATURES</b>  | <b>78</b> |
| 10.1. OVERVIEW   | 78        |
| 10.2. SIGNING IMAGES USING ATOMIC CLI  | 78        |
| 10.3. VERIFYING IMAGE SIGNATURES USING OPENSIFT CLI                                | 79        |
| 10.4. ACCESSING IMAGE SIGNATURES USING REGISTRY API                                | 80        |
| 10.4.1. Writing Image Signatures via API   | 80        |
| 10.4.2. Reading Image Signatures via API   | 81        |
| <b>CHAPTER 11. SCOPED TOKENS</b>   | <b>82</b> |
| 11.1. OVERVIEW   | 82        |
| 11.2. EVALUATION   | 82        |
| 11.3. USER SCOPES  | 82        |

|  |           |
|--|-----------|
| 11.4. ROLE SCOPE   | 82        |
| <b>CHAPTER 12. MONITORING IMAGES</b>                               | <b>83</b> |
| 12.1. OVERVIEW   | 83        |
| 12.2. VIEWING IMAGES STATISTICS                                    | 83        |
| 12.3. VIEWING IMAGESTREAMS STATISTICS                              | 83        |
| 12.4. PRUNING IMAGES   | 84        |
| <b>CHAPTER 13. MANAGING SECURITY CONTEXT CONSTRAINTS</b>           | <b>85</b> |
| 13.1. OVERVIEW   | 85        |
| 13.2. LISTING SECURITY CONTEXT CONSTRAINTS                         | 85        |
| 13.3. EXAMINING A SECURITY CONTEXT CONSTRAINTS OBJECT              | 85        |
| 13.4. CREATING NEW SECURITY CONTEXT CONSTRAINTS                    | 86        |
| 13.5. DELETING SECURITY CONTEXT CONSTRAINTS                        | 87        |
| 13.6. UPDATING SECURITY CONTEXT CONSTRAINTS                        | 88        |
| 13.6.1. Example Security Context Constraints Settings              | 88        |
| 13.7. UPDATING THE DEFAULT SECURITY CONTEXT CONSTRAINTS            | 89        |
| 13.8. HOW DO I?  | 89        |
| 13.8.1. Grant Access to the Privileged SCC                         | 89        |
| 13.8.2. Grant a Service Account Access to the Privileged SCC       | 90        |
| 13.8.3. Enable Images to Run with USER in the Dockerfile           | 90        |
| 13.8.4. Enable Container Images that Require Root                  | 91        |
| 13.8.5. Use --mount-host on the Registry                           | 91        |
| 13.8.6. Provide Additional Capabilities                            | 91        |
| 13.8.7. Modify Cluster Default Behavior                            | 92        |
| 13.8.8. Use the hostPath Volume Plug-in                            | 92        |
| 13.8.9. Ensure That Admission Attempts to Use a Specific SCC First | 93        |
| 13.8.10. Add an SCC to a User, Group, or Project                   | 93        |
| <b>CHAPTER 14. SCHEDULING</b>                                      | <b>94</b> |
| 14.1. OVERVIEW   | 94        |
| 14.1.1. Overview   | 94        |
| 14.1.2. Default scheduling   | 94        |
| 14.1.3. Advanced scheduling  | 94        |
| 14.1.4. Custom scheduling  | 94        |
| 14.2. DEFAULT SCHEDULING   | 94        |
| 14.2.1. Overview   | 94        |
| 14.2.2. Generic Scheduler  | 94        |
| 14.2.3. Filter the Nodes   | 95        |
| 14.2.3.1. Prioritize the Filtered List of Nodes                    | 95        |
| 14.2.3.2. Select the Best Fit Node                                 | 95        |
| 14.2.4. Scheduler Policy   | 95        |
| 14.2.4.1. Modifying Scheduler Policy                               | 97        |
| 14.2.5. Available Predicates                                       | 98        |
| 14.2.5.1. Static Predicates  | 98        |
| 14.2.5.1.1. Default Predicates                                     | 98        |
| 14.2.5.1.2. Other Static Priorities                                | 99        |
| 14.2.5.2. General Predicates                                       | 100       |
| Non-critical general predicates                                    | 100       |
| Essential general predicates                                       | 100       |
| 14.2.5.3. Configurable Predicates                                  | 101       |
| 14.2.6. Available Priorities                                       | 102       |
| 14.2.6.1. Static Priorities  | 103       |
| 14.2.6.1.1. Default Priorities                                     | 103       |

|  |     |
|--|-----|
| 14.2.6.1.2. Other Static Priorities                          | 104 |
| 14.2.6.2. Configurable Priorities                            | 104 |
| 14.2.7. Use Cases  | 106 |
| 14.2.7.1. Infrastructure Topological Levels                  | 106 |
| 14.2.7.2. Affinity   | 106 |
| 14.2.7.3. Anti Affinity                                      | 106 |
| 14.2.8. Sample Policy Configurations                         | 106 |
| 14.3. CUSTOM SCHEDULING                                      | 109 |
| 14.3.1. Overview   | 109 |
| 14.3.2. Deploying the Scheduler                              | 109 |
| 14.4. CONTROLLING POD PLACEMENT                              | 111 |
| 14.4.1. Overview   | 111 |
| 14.4.2. Constraining Pod Placement Using Node Name           | 111 |
| 14.4.3. Constraining Pod Placement Using a Node Selector     | 111 |
| 14.4.4. Control Pod Placement to Projects                    | 113 |
| 14.5. ADVANCED SCHEDULING                                    | 115 |
| 14.5.1. Overview   | 115 |
| 14.5.2. Using Advanced Scheduling                            | 116 |
| 14.6. ADVANCED SCHEDULING AND NODE AFFINITY                  | 117 |
| 14.6.1. Overview   | 117 |
| 14.6.2. Configuring Node Affinity                            | 117 |
| 14.6.2.1. Configuring a Required Node Affinity Rule          | 119 |
| 14.6.2.2. Configuring a Preferred Node Affinity Rule         | 119 |
| 14.6.3. Examples   | 120 |
| 14.6.3.1. Node Affinity with Matching Labels                 | 120 |
| 14.6.3.2. Node Affinity with No Matching Labels              | 121 |
| 14.7. ADVANCED SCHEDULING AND POD AFFINITY AND ANTI-AFFINITY | 122 |
| 14.7.1. Overview   | 122 |
| 14.7.2. Configuring Pod Affinity and Anti-affinity           | 122 |
| 14.7.2.1. Configuring an Affinity Rule                       | 124 |
| 14.7.2.2. Configuring an Anti-affinity Rule                  | 125 |
| 14.7.3. Examples   | 126 |
| 14.7.3.1. Pod Affinity                                       | 126 |
| 14.7.3.2. Pod Anti-affinity                                  | 127 |
| 14.7.3.3. Pod Affinity with no Matching Labels               | 128 |
| 14.8. ADVANCED SCHEDULING AND NODE SELECTORS                 | 128 |
| 14.8.1. Overview   | 128 |
| 14.8.2. Configuring Node Selectors                           | 129 |
| 14.9. ADVANCED SCHEDULING AND TAINTS AND TOLERATIONS         | 130 |
| 14.9.1. Overview   | 130 |
| 14.9.2. Taints and Tolerations                               | 130 |
| 14.9.2.1. Using Multiple Taints                              | 131 |
| 14.9.3. Adding a Taint to an Existing Node                   | 132 |
| 14.9.4. Adding a Toleration to a Pod                         | 132 |
| 14.9.4.1. Using Toleration Seconds to Delay Pod Evictions    | 133 |
| 14.9.4.1.1. Setting a Default Value for Toleration Seconds   | 133 |
| 14.9.5. Preventing Pod Eviction for Node Problems            | 134 |
| 14.9.6. Daemonsets and Tolerations                           | 135 |
| 14.9.7. Examples   | 135 |
| 14.9.7.1. Dedicating a Node for a User                       | 136 |
| 14.9.7.2. Binding a User to a Node                           | 136 |
| 14.9.7.3. Nodes with Special Hardware                        | 136 |



|   |            |
|---|------------|
| <b>CHAPTER 15. SETTING QUOTAS</b> .....                           | <b>137</b> |
| 15.1. OVERVIEW  | 137        |
| 15.2. RESOURCES MANAGED BY QUOTA                                  | 137        |
| 15.3. QUOTA SCOPES  | 138        |
| 15.4. QUOTA ENFORCEMENT   | 139        |
| 15.5. REQUESTS VERSUS LIMITS                                      | 140        |
| 15.6. SAMPLE RESOURCE QUOTA DEFINITIONS                           | 140        |
| 15.7. CREATING A QUOTA  | 143        |
| 15.8. VIEWING A QUOTA   | 143        |
| 15.9. CONFIGURING QUOTA SYNCHRONIZATION PERIOD                    | 144        |
| 15.10. ACCOUNTING FOR QUOTA IN DEPLOYMENT CONFIGURATIONS          | 144        |
| 15.11. REQUIRE EXPLICIT QUOTA TO CONSUME A RESOURCE               | 144        |
| 15.12. KNOWN ISSUES   | 145        |
| <b>CHAPTER 16. SETTING MULTI-PROJECT QUOTAS</b> .....             | <b>146</b> |
| 16.1. OVERVIEW  | 146        |
| 16.2. SELECTING PROJECTS  | 146        |
| 16.3. VIEWING APPLICABLE CLUSTERRESOURCEQUOTAS                    | 147        |
| 16.4. SELECTION GRANULARITY                                       | 148        |
| <b>CHAPTER 17. SETTING LIMIT RANGES</b> .....                     | <b>149</b> |
| 17.1. OVERVIEW  | 149        |
| 17.1.1. Container Limits  | 150        |
| 17.1.2. Pod Limits  | 151        |
| 17.1.3. Image Limits  | 152        |
| 17.1.4. Image Stream Limits                                       | 153        |
| 17.1.4.1. Counting of Image References                            | 153        |
| 17.1.5. PersistentVolumeClaim Limits                              | 154        |
| 17.2. CREATING A LIMIT RANGE                                      | 155        |
| 17.3. VIEWING LIMITS  | 155        |
| 17.4. DELETING LIMITS   | 155        |
| <b>CHAPTER 18. PRUNING OBJECTS</b> .....                          | <b>156</b> |
| 18.1. OVERVIEW  | 156        |
| 18.2. BASIC PRUNE OPERATIONS                                      | 156        |
| 18.3. PRUNING DEPLOYMENTS   | 156        |
| 18.4. PRUNING BUILDS  | 157        |
| 18.5. PRUNING IMAGES  | 158        |
| 18.5.1. Image Prune Conditions                                    | 159        |
| 18.5.2. Using Secure or Insecure Connections                      | 160        |
| 18.5.3. Image Pruning Problems                                    | 161        |
| Images Not Being Pruned   | 161        |
| Using a Secure Connection Against Insecure Registry               | 161        |
| 18.5.3.1. Using an Insecure Connection Against a Secured Registry | 162        |
| Using the Wrong Certificate Authority                             | 162        |
| 18.6. HARD PRUNING THE REGISTRY                                   | 162        |
| 18.7. PRUNING CRON JOBS   | 165        |
| <b>CHAPTER 19. GARBAGE COLLECTION</b> .....                       | <b>167</b> |
| 19.1. OVERVIEW  | 167        |
| 19.2. CONTAINER GARBAGE COLLECTION                                | 167        |
| 19.2.1. Detecting Containers for Deletion                         | 168        |
| 19.3. IMAGE GARBAGE COLLECTION                                    | 168        |
| 19.3.1. Detecting Images for Deletion                             | 169        |

|  |            |
|--|------------|
| <b>CHAPTER 20. ALLOCATING NODE RESOURCES</b>                         | <b>170</b> |
| 20.1. OVERVIEW   | 170        |
| 20.2. CONFIGURING NODES FOR ALLOCATED RESOURCES                      | 170        |
| 20.3. COMPUTING ALLOCATED RESOURCES                                  | 170        |
| 20.4. VIEWING NODE ALLOCATABLE RESOURCES AND CAPACITY                | 171        |
| 20.5. SYSTEM RESOURCES REPORTED BY NODE                              | 171        |
| 20.6. NODE ENFORCEMENT   | 172        |
| 20.7. EVICTION THRESHOLDS  | 173        |
| 20.8. SCHEDULER  | 174        |
| <b>CHAPTER 21. OPAQUE INTEGER RESOURCES</b>                          | <b>175</b> |
| 21.1. OVERVIEW   | 175        |
| 21.2. CREATING OPAQUE INTEGER RESOURCES                              | 175        |
| <b>CHAPTER 22. OVERCOMMITTING</b>                                    | <b>178</b> |
| 22.1. OVERVIEW   | 178        |
| 22.2. REQUESTS AND LIMITS  | 178        |
| 22.2.1. Tune Buffer Chunk Limit                                      | 178        |
| 22.3. COMPUTE RESOURCES  | 179        |
| 22.3.1. CPU  | 179        |
| 22.3.2. Memory   | 179        |
| 22.4. QUALITY OF SERVICE CLASSES                                     | 180        |
| 22.5. CONFIGURING MASTERS FOR OVERCOMMITMENT                         | 180        |
| 22.6. CONFIGURING NODES FOR OVERCOMMITMENT                           | 181        |
| 22.6.1. Reserving Memory Across Quality of Service Tiers             | 182        |
| 22.6.2. Enforcing CPU Limits   | 182        |
| 22.6.3. Reserving Resources for System Processes                     | 183        |
| 22.6.4. Kernel Tunable Flags   | 184        |
| 22.6.5. Disabling Swap Memory  | 184        |
| <b>CHAPTER 23. ASSIGNING UNIQUE EXTERNAL IPS FOR INGRESS TRAFFIC</b> | <b>186</b> |
| 23.1. OVERVIEW   | 186        |
| 23.2. RESTRICTIONS   | 186        |
| 23.3. CONFIGURING THE CLUSTER TO USE UNIQUE EXTERNAL IPS             | 187        |
| 23.3.1. Configuring an Ingress IP for a Service                      | 187        |
| 23.4. ROUTING THE INGRESS CIDR FOR DEVELOPMENT OR TESTING            | 188        |
| 23.4.1. Service externalIPs  | 188        |
| <b>CHAPTER 24. HANDLING OUT OF RESOURCE ERRORS</b>                   | <b>190</b> |
| 24.1. OVERVIEW   | 190        |
| 24.2. CONFIGURING EVICTION POLICIES                                  | 190        |
| 24.2.1. Using the Node Configuration to Create a Policy              | 191        |
| 24.2.2. Understanding Eviction Signals                               | 192        |
| 24.2.3. Understanding Eviction Thresholds                            | 194        |
| 24.2.3.1. Understanding Hard Eviction Thresholds                     | 195        |
| 24.2.3.2. Understanding Soft Eviction Thresholds                     | 195        |
| 24.3. CONFIGURING THE AMOUNT OF RESOURCE FOR SCHEDULING              | 196        |
| 24.4. CONTROLLING NODE CONDITION OSCILLATION                         | 197        |
| 24.5. RECLAIMING NODE-LEVEL RESOURCES                                | 197        |
| With Imagefs   | 197        |
| Without Imagefs  | 198        |
| 24.6. UNDERSTANDING POD EVICTION                                     | 198        |
| 24.6.1. Understanding Quality of Service and Out of Memory Killer    | 199        |
| 24.7. UNDERSTANDING THE POD SCHEDULER AND OOR CONDITIONS             | 199        |

|  |            |
|--|------------|
| 24.8. EXAMPLE SCENARIO   | 200        |
| 24.9. RECOMMENDED PRACTICE   | 201        |
| 24.9.1. DaemonSets and Out of Resource Handling                            | 201        |
| <b>CHAPTER 25. MONITORING AND DEBUGGING ROUTERS</b>                        | <b>202</b> |
| 25.1. OVERVIEW   | 202        |
| 25.2. VIEWING STATISTICS   | 202        |
| 25.3. DISABLING STATISTICS VIEW  | 205        |
| 25.4. VIEWING LOGS   | 205        |
| 25.5. VIEWING THE ROUTER INTERNALS   | 206        |
| <b>CHAPTER 26. HIGH AVAILABILITY</b>                                       | <b>207</b> |
| 26.1. OVERVIEW   | 207        |
| 26.2. CONFIGURING IP FAILOVER  | 208        |
| 26.2.1. Virtual IP Addresses   | 209        |
| 26.2.2. Check and Notify Scripts   | 209        |
| 26.2.3. VRRP Preemption  | 211        |
| 26.2.4. Keepalived Multicast   | 211        |
| 26.2.5. Command Line Options and Environment Variables                     | 212        |
| 26.2.6. VRRP ID Offset   | 214        |
| 26.2.7. Configuring a Highly-available Service                             | 214        |
| 26.2.7.1. Deploy IP Failover Pod   | 216        |
| 26.2.8. Dynamically Updating Virtual IPs for a Highly-available Service    | 216        |
| 26.3. CONFIGURING SERVICE EXTERNALIP AND NODEPORT                          | 217        |
| 26.4. HIGH AVAILABILITY FOR INGRESSIP                                      | 217        |
| <b>CHAPTER 27. IPTABLES</b>  | <b>218</b> |
| 27.1. OVERVIEW   | 218        |
| 27.2. IPTABLES   | 218        |
| 27.3. IPTABLES.SERVICE   | 218        |
| <b>CHAPTER 28. SECURING BUILDS BY STRATEGY</b>                             | <b>220</b> |
| 28.1. OVERVIEW   | 220        |
| 28.2. DISABLING A BUILD STRATEGY GLOBALLY                                  | 220        |
| 28.3. RESTRICTING BUILD STRATEGIES TO A USER GLOBALLY                      | 221        |
| 28.4. RESTRICTING BUILD STRATEGIES TO A USER WITHIN A PROJECT              | 221        |
| <b>CHAPTER 29. RESTRICTING APPLICATION CAPABILITIES USING SECCOMP</b>      | <b>223</b> |
| 29.1. OVERVIEW   | 223        |
| 29.2. ENABLING SECCOMP   | 223        |
| 29.3. CONFIGURING OPENSIFT CONTAINER PLATFORM FOR SECCOMP                  | 223        |
| 29.4. CONFIGURING OPENSIFT CONTAINER PLATFORM FOR A CUSTOM SECCOMP PROFILE | 224        |
| <b>CHAPTER 30. SYSCTLS</b>   | <b>225</b> |
| 30.1. OVERVIEW   | 225        |
| 30.2. UNDERSTANDING SYSCTLS  | 225        |
| 30.3. NAMESPACE VERSUS NODE-LEVEL SYSCTLS                                  | 225        |
| 30.4. SAFE VERSUS UNSAFE SYSCTLS   | 226        |
| 30.5. ENABLING UNSAFE SYSCTLS  | 226        |
| 30.6. SETTING SYSCTLS FOR A POD  | 227        |
| <b>CHAPTER 31. ENCRYPTING DATA AT DATASTORE LAYER</b>                      | <b>228</b> |
| 31.1. OVERVIEW   | 228        |
| 31.2. CONFIGURATION AND DETERMINING WHETHER ENCRYPTION IS ALREADY ENABLED  | 228        |
| 31.3. UNDERSTANDING THE ENCRYPTION CONFIGURATION                           | 228        |

|   |            |
|---|------------|
| 31.3.1. Available Providers                                     | 229        |
| 31.4. ENCRYPTING DATA   | 230        |
| 31.5. VERIFYING THAT DATA IS ENCRYPTED                          | 231        |
| 31.6. ENSURE ALL SECRETS ARE ENCRYPTED                          | 232        |
| 31.7. ROTATING A DECRYPTION KEY                                 | 232        |
| 31.8. DECRYPTING DATA   | 232        |
| <b>CHAPTER 32. ENCRYPTING HOSTS WITH IPSEC</b>                  | <b>234</b> |
| 32.1. OVERVIEW  | 234        |
| 32.2. ENCRYPTING HOSTS  | 234        |
| 32.2.1. Step 1: Prerequisites                                   | 234        |
| 32.2.2. Step 2: Certificates                                    | 234        |
| 32.2.3. Step 3: libreswan IPsec Policy                          | 235        |
| 32.2.3.1. Opportunistic Group Configuration                     | 235        |
| 32.2.3.2. Explicit Connection Configuration                     | 236        |
| 32.3. IPSEC FIREWALL CONFIGURATION                              | 237        |
| 32.4. STARTING AND ENABLING IPSEC                               | 237        |
| 32.5. OPTIMIZING IPSEC  | 238        |
| 32.6. TROUBLESHOOTING   | 238        |
| <b>CHAPTER 33. BUILDING DEPENDENCY TREES</b>                    | <b>239</b> |
| 33.1. OVERVIEW  | 239        |
| 33.2. USAGE   | 239        |
| <b>CHAPTER 34. BACKUP AND RESTORE</b>                           | <b>240</b> |
| 34.1. OVERVIEW  | 240        |
| 34.2. PREREQUISITES   | 240        |
| 34.3. CLUSTER BACKUP  | 241        |
| 34.3.1. Master Backup   | 241        |
| 34.3.2. Etcd Backup   | 241        |
| 34.3.3. Registry Certificates Backup                            | 242        |
| 34.4. CLUSTER RESTORE FOR SINGLE-MEMBER ETCD CLUSTERS           | 242        |
| 34.5. CLUSTER RESTORE FOR MULTIPLE-MEMBER ETCD CLUSTERS         | 243        |
| 34.5.1. Embedded etcd   | 243        |
| 34.5.2. Separate etcd   | 244        |
| 34.5.2.1. Containerized etcd Deployments                        | 245        |
| 34.5.2.2. Non-Containerized etcd Deployments                    | 246        |
| 34.5.2.3. Adding Additional etcd Members                        | 247        |
| 34.6. ADDING NEW ETCD HOSTS                                     | 249        |
| 34.7. BRINGING OPENSIFT CONTAINER PLATFORM SERVICES BACK ONLINE | 253        |
| 34.8. PROJECT BACKUP  | 254        |
| 34.8.1. Role Bindings   | 254        |
| 34.8.2. Service Accounts  | 254        |
| 34.8.3. Secrets   | 254        |
| 34.8.4. Persistent Volume Claims                                | 254        |
| 34.9. PROJECT RESTORE   | 254        |
| 34.10. APPLICATION DATA BACKUP                                  | 255        |
| 34.11. APPLICATION DATA RESTORE                                 | 256        |
| <b>CHAPTER 35. TROUBLESHOOTING OPENSIFT SDN</b>                 | <b>257</b> |
| 35.1. OVERVIEW  | 257        |
| 35.2. NOMENCLATURE  | 257        |
| 35.3. DEBUGGING EXTERNAL ACCESS TO AN HTTP SERVICE              | 258        |
| 35.4. DEBUGGING THE ROUTER                                      | 259        |

|   |            |
|---|------------|
| 35.5. DEBUGGING A SERVICE                                   | 260        |
| 35.6. DEBUGGING NODE TO NODE NETWORKING                     | 261        |
| 35.7. DEBUGGING LOCAL NETWORKING                            | 262        |
| 35.7.1. The Interfaces on a Node                            | 263        |
| 35.7.2. SDN Flows Inside a Node                             | 263        |
| 35.7.3. Debugging Steps                                     | 263        |
| 35.7.3.1. Is IP Forwarding Enabled?                         | 263        |
| 35.7.3.2. Are your routes correct?                          | 263        |
| 35.7.4. Is the Open vSwitch configured correctly?           | 264        |
| 35.7.4.1. Is the iptables configuration correct?            | 265        |
| 35.7.4.2. Is your external network correct?                 | 265        |
| 35.8. DEBUGGING VIRTUAL NETWORKING                          | 265        |
| 35.8.1. Builds on a Virtual Network are Failing             | 265        |
| 35.9. DEBUGGING POD EGRESS                                  | 266        |
| 35.10. READING THE LOGS                                     | 266        |
| 35.11. DEBUGGING KUBERNETES                                 | 266        |
| 35.12. FINDING NETWORK ISSUES USING THE DIAGNOSTICS TOOL    | 267        |
| 35.13. MISCELLANEOUS NOTES                                  | 267        |
| 35.13.1. Other clarifications on ingress                    | 267        |
| 35.13.2. TLS Handshake Timeout                              | 267        |
| 35.13.3. Other debugging notes                              | 268        |
| <b>CHAPTER 36. DIAGNOSTICS TOOL</b>                         | <b>269</b> |
| 36.1. OVERVIEW  | 269        |
| 36.2. USING THE DIAGNOSTICS TOOL                            | 269        |
| 36.3. RUNNING DIAGNOSTICS IN A SERVER ENVIRONMENT           | 271        |
| 36.4. RUNNING DIAGNOSTICS IN A CLIENT ENVIRONMENT           | 272        |
| 36.5. ANSIBLE-BASED HEALTH CHECKS                           | 272        |
| 36.5.1. Running Health Checks via ansible-playbook          | 275        |
| 36.5.2. Running Health Checks via Docker CLI                | 275        |
| <b>CHAPTER 37. IDLING APPLICATIONS</b>                      | <b>277</b> |
| 37.1. OVERVIEW  | 277        |
| 37.2. IDLING APPLICATIONS                                   | 277        |
| 37.2.1. Idling Single Services                              | 277        |
| 37.2.2. Idling Multiple Services                            | 277        |
| 37.3. UNIDLING APPLICATIONS                                 | 277        |
| <b>CHAPTER 38. ANALYZING CLUSTER CAPACITY</b>               | <b>279</b> |
| 38.1. OVERVIEW  | 279        |
| 38.2. RUNNING CLUSTER CAPACITY ANALYSIS ON THE COMMAND LINE | 279        |
| 38.3. RUNNING CLUSTER CAPACITY AS A JOB INSIDE OF A POD     | 280        |
| <b>CHAPTER 39. REVISION HISTORY: CLUSTER ADMINISTRATION</b> | <b>283</b> |
| 39.1. TUES MAR 06 2018                                      | 283        |
| 39.2. FRI FEB 23 2018                                       | 283        |
| 39.3. TUES FEB 20 2018                                      | 283        |
| 39.4. FRI FEB 16 2018                                       | 283        |
| 39.5. TUE FEB 06 2018                                       | 283        |
| 39.6. TUE NOV 21 2017                                       | 284        |
| 39.7. FRI NOV 10 2017                                       | 284        |
| 39.8. FRI NOV 03 2017                                       | 284        |
| 39.9. TUE OCT 24 2017                                       | 284        |
| 39.10. WED OCT 11 2017                                      | 284        |

|                        |     |
|------------------------|-----|
| 39.11. MON OCT 02 2017 | 284 |
| 39.12. MON SEP 18 2017 | 285 |
| 39.13. FRI SEP 08 2017 | 285 |
| 39.14. TUE AUG 29 2017 | 285 |
| 39.15. TUE AUG 22 2017 | 285 |
| 39.16. MON AUG 14 2017 | 286 |
| 39.17. WED AUG 09 2017 | 286 |



## CHAPTER 1. OVERVIEW

These Cluster Administration topics cover the day-to-day tasks for managing your OpenShift Container Platform cluster and other advanced configuration topics.



## CHAPTER 2. MANAGING NODES

### 2.1. OVERVIEW

You can manage [nodes](#) in your instance using the [CLI](#).

When you perform node management operations, the CLI interacts with [node objects](#) that are representations of actual node hosts. The [master](#) uses the information from node objects to validate nodes with [health checks](#).

### 2.2. LISTING NODES

To list all nodes that are known to the master:

```
$ oc get nodes
NAME                                STATUS                                AGE
master.example.com                 Ready,SchedulingDisabled            165d
node1.example.com                  Ready                                165d
node2.example.com                  Ready                                165d
```

To only list information about a single node, replace **<node>** with the full node name:

```
$ oc get node <node>
```

The **STATUS** column in the output of these commands can show nodes with the following conditions:

**Table 2.1. Node Conditions**

| Condition                 | Description  |
|---------------------------|--|
| <b>Ready</b>              | The node is passing the health checks performed from the master by returning <b>StatusOK</b> . |
| <b>NotReady</b>           | The node is not passing the health checks performed from the master.                           |
| <b>SchedulingDisabled</b> | Pods cannot be <a href="#">scheduled for placement</a> on the node.                            |



#### NOTE

The **STATUS** column can also show **Unknown** for a node if the CLI cannot find any node condition.

To get more detailed information about a specific node, including the reason for the current condition:

```
$ oc describe node <node>
```

For example:

```
# oc describe node node1.example.com
Name:      node1.example.com 1
```

```

Role:
Labels:  beta.kubernetes.io/arch=amd64
         beta.kubernetes.io/os=linux
         kubernetes.io/hostname=node1.example.com
         region=infra
         zone=default
Annotations:  volumes.kubernetes.io/controller-managed-attach-detach=true
Taints:  <none>
CreationTimestamp:  Wed, 11 Apr 2018 03:00:25 +0530
Phase:
Conditions:
  Type          Status  LastHeartbeatTime   LastTransitionTime  ReasonMessage
  ----          -
  OutOfDisk     False   Wed, 30 May 2018 15:51:29 +0530   Wed, 11 Apr 2018 03:00:25 +0530   KubeletHasSufficientDisk   kubelet has sufficient disk space available
  MemoryPressure False   Wed, 30 May 2018 15:51:29 +0530   Wed, 11 Apr 2018 03:00:25 +0530   KubeletHasSufficientMemory kubelet has sufficient memory available
  DiskPressure  False   Wed, 30 May 2018 15:51:29 +0530   Wed, 11 Apr 2018 03:00:25 +0530   KubeletHasNoDiskPressure   kubelet has no disk pressure
  Ready         True    Wed, 30 May 2018 15:51:29 +0530   Mon, 21 May 2018 19:01:00 +0530   KubeletReady               kubelet is posting ready status
Addresses:  10.74.252.28,10.74.252.28,vm252-28.gsslab.pnq2.redhat.com
Capacity:
  cpu: 8
  memory: 32780308Ki
  pods: 80
Allocatable:
  cpu: 8
  memory: 32677908Ki
  pods: 80
System Info:
  Machine ID: 4cb86a52e94ab9a32d701689043
  System UUID: 4CB86A52-F94-AB9A-32D701689043
  Boot ID: 4a592874-d98-889a-6a58e3f3ae0a
  Kernel Version: 3.10.0-514.el7.x86_64
  OS Image: Employee SKU
  Operating System: linux
  Architecture: amd64
  Container Runtime Version: docker://1.12.6
  Kubelet Version: v1.6.1+5115d708d7
  Kube-Proxy Version: v1.6.1+5115d708d7
ExternalID:  node1.example.com
Non-terminated Pods: (9 in total)
  Namespace      Name              CPU Requests  CPU Limits  Memory Requests  Memory Limits
  -----
  default        docker-registry-4-7bs12  100m (1%)  0 (0%)  256Mi (0%)  0 (0%)
  default        registry-console-1-nnxdm  0 (0%)  0 (0%)  0 (0%)  0 (0%)
  default        router-1-d78c6  100m (1%)  0 (0%)  256Mi (0%)  0 (0%)
  openshift-infra hawkular-cassandra-1-29ctk  0 (0%)  0 (0%)  1G (2%)  2G (5%)

```

```

    openshift-infra   hawkular-metrics-9tlw3    0 (0%)  0 (0%)  1500M (4%)
2500M (7%)
    openshift-infra   heapster-bhpqq      4 (50%)  4 (50%)  937500k (2%) 2Gi
(6%)
    tcs      bgdemo-24-3shj5      2m (0%)  4m (0%)  0 (0%)  0 (0%)
    tcs      bgdemo-24-469d9      2m (0%)  4m (0%)  0 (0%)  0 (0%)
    tcs      httpd-1-x67vh      0 (0%)  0 (0%)  0 (0%)  0 (0%)
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted.)
  CPU Requests CPU Limits Memory Requests Memory Limits
  -----
    4204m (52%) 4008m (50%) 3974370912 (11%) 6647483648 (19%)
Events:  <none>

```

- 1 The name of the node.
- 2 The role of the node, either **master** or **compute**.
- 3 The [labels](#) applied to the node.
- 4 The annotations applied to the node.
- 5 The [taints](#) applied to the node.
- 6 [Node conditions](#).
- 7 The IP address and host name of the node.
- 8 The [pod resources and allocatable resources](#).
- 9 Information about the node host.
- 10 The fully-qualified domain name where the node can be reached.
- 11 The pods on the node.

## 2.3. ADDING NODES

To add nodes to your existing OpenShift Container Platform cluster, you can run an Ansible playbook that handles installing the node components, generating the required certificates, and other important steps. See the [advanced installation](#) method for instructions on running the playbook directly.

Alternatively, if you used the quick installation method, you can [re-run the installer to add nodes](#), which performs the same steps.

## 2.4. DELETING NODES

When you delete a node using the CLI, the node object is deleted in Kubernetes, but the pods that exist on the node itself are not deleted. Any bare pods not backed by a replication controller would be inaccessible to OpenShift Container Platform, pods backed by replication controllers would be rescheduled to other available nodes, and [local manifest pods](#) would need to be manually deleted.

To delete a node from the OpenShift Container Platform cluster:

1. [Evacuate pods](#) from the node you are preparing to delete.

2. Delete the node object:

```
$ oc delete node <node>
```

3. Check that the node has been removed from the node list:

```
$ oc get nodes
```

Pods should now be only scheduled for the remaining nodes that are in **Ready** state.

4. If you want to uninstall all OpenShift Container Platform content from the node host, including all pods and containers, continue to [Uninstalling Nodes](#) and follow the procedure using the ***uninstall.yml*** playbook. The procedure assumes general understanding of the [advanced installation method](#) using Ansible.

## 2.5. UPDATING LABELS ON NODES

To add or update [labels](#) on a node:

```
$ oc label node <node> <key_1>=<value_1> ... <key_n>=<value_n>
```

To see more detailed usage:

```
$ oc label -h
```

## 2.6. LISTING PODS ON NODES

To list all or selected pods on one or more nodes:

```
$ oc adm manage-node <node1> <node2> \
  --list-pods [--pod-selector=<pod_selector>] [-o json|yaml]
```

To list all or selected pods on selected nodes:

```
$ oc adm manage-node --selector=<node_selector> \
  --list-pods [--pod-selector=<pod_selector>] [-o json|yaml]
```

## 2.7. MARKING NODES AS UNSCHEDULABLE OR SCHEDULABLE

By default, healthy nodes with a **Ready** [status](#) are marked as schedulable, meaning that new pods are allowed for placement on the node. Manually marking a node as unschedulable blocks any new pods from being scheduled on the node. Existing pods on the node are not affected.

To mark a node or nodes as unschedulable:

```
$ oc adm manage-node <node1> <node2> --schedulable=false
```

For example:

```
$ oc adm manage-node node1.example.com --schedulable=false
NAME                                LABELS
```

**STATUS**

```
node1.example.com    kubernetes.io/hostname=node1.example.com
Ready,SchedulingDisabled
```

To mark a currently unschedulable node or nodes as schedulable:

```
$ oc adm manage-node <node1> <node2> --schedulable
```

Alternatively, instead of specifying specific node names (e.g., **<node1> <node2>**), you can use the **--selector=<node\_selector>** option to mark selected nodes as schedulable or unschedulable.

## 2.8. EVACUATING PODS ON NODES

Evacuating pods allows you to migrate all or selected pods from a given node or nodes. Nodes must first be [marked unschedulable](#) to perform pod evacuation.

Only pods backed by a [replication controller](#) can be evacuated; the replication controllers create new pods on other nodes and remove the existing pods from the specified node(s). Bare pods, meaning those not backed by a replication controller, are unaffected by default. You can evacuate a subset of pods by specifying a pod-selector. Pod selector is based on labels, so all the pods with the specified label will be evacuated.

To evacuate one or more nodes:

```
$ oc adm drain <node1> <node2>
```

You can force deletion of bare pods by using the **--force** option. When set to **true**, deletion continues even if there are pods not managed by a replication controller, ReplicaSet, job, daemonset, or StatefulSet:

```
$ oc adm drain <node1> <node2> --force=true
```

You can use **--grace-period** to set a period of time in seconds for each pod to terminate gracefully. If negative, the default value specified in the pod will be used:

```
$ oc adm drain <node1> <node2> --grace-period=-1
```

You can use **--ignore-daemonsets** and set it to **true** to ignore daemonset-managed pods:

```
$ oc adm drain <node1> <node2> --ignore-daemonset=true
```

You can use **--timeout** to set the length of time to wait before giving up. A value of **0** sets an infinite length of time:

```
$ oc adm drain <node1> <node2> --timeout=5s
```

You can use **--delete-local-data** and set it to **true** to continue deletion even if there are pods using emptyDir (local data that will be deleted when the node is drained):

```
$ oc adm drain <node1> <node2> --delete-local-data=true
```

## 2.9. REBOOTING NODES

To reboot a node without causing an outage for applications running on the platform, it is important to first [evacuate the pods](#). For pods that are made highly available by the routing tier, nothing else needs to be done. For other pods needing storage, typically databases, it is critical to ensure that they can remain in operation with one pod temporarily going offline. While implementing resiliency for stateful pods is different for each application, in all cases it is important to configure the scheduler to use [node anti-affinity](#) to ensure that the pods are properly spread across available nodes.

Another challenge is how to handle nodes that are running critical infrastructure such as the router or the registry. The same node evacuation process applies, though it is important to understand certain edge cases.

### 2.9.1. Infrastructure Nodes

Infrastructure nodes are nodes that are labeled to run pieces of the OpenShift Container Platform environment. Currently, the easiest way to manage node reboots is to ensure that there are at least three nodes available to run infrastructure. The scenario below demonstrates a common mistake that can lead to service interruptions for the applications running on OpenShift Container Platform when only two nodes are available.

- Node A is marked unschedulable and all pods are evacuated.
- The registry pod running on that node is now redeployed on node B. This means node B is now running both registry pods.
- Node B is now marked unschedulable and is evacuated.
- The service exposing the two pod endpoints on node B, for a brief period of time, loses all endpoints until they are redeployed to node A.

The same process using three infrastructure nodes does not result in a service disruption. However, due to pod scheduling, the last node that is evacuated and brought back in to rotation is left running zero registries. The other two nodes will run two and one registries respectively. The best solution is to rely on pod anti-affinity. This is an alpha feature in Kubernetes that is available for testing now, but is not yet supported for production workloads.

### 2.9.2. Using Pod Anti-affinity

[Pod anti-affinity](#) is slightly different than [node anti-affinity](#). Node anti-affinity can be violated if there are no other suitable locations to deploy a pod. Pod anti-affinity can be set to either required or preferred.

Using the `docker-registry` pod as an example, the first step in enabling this feature is to set the `scheduler.alpha.kubernetes.io/affinity` on the pod.

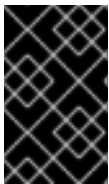
```
apiVersion: v1
kind: Pod
metadata:
  name: with-pod-antiaffinity
spec:
  affinity:
    podAntiAffinity: ❶
    preferredDuringSchedulingIgnoredDuringExecution: ❷
    - weight: 100 ❸
      podAffinityTerm:
```

```

labelSelector:
  matchExpressions:
    - key: docker-registry ④
      operator: In ⑤
      values:
        - default
  topologyKey: kubernetes.io/hostname

```

- ① Stanza to configure pod anti-affinity.
- ② Defines a preferred rule.
- ③ Specifies a weight for a preferred rule. The node with the highest weight is preferred.
- ④ Description of the pod label that determines when the anti-affinity rule applies. Specify a key and value for the label.
- ⑤ The operator represents the relationship between the label on the existing pod and the set of values in the **matchExpression** parameters in the specification for the new pod. Can be **In**, **NotIn**, **Exists**, or **DoesNotExist**.



### IMPORTANT

**scheduler.alpha.kubernetes.io/affinity** is internally stored as a string even though the contents are JSON. The above example shows how this string can be added as an annotation to a YAML deployment configuration.

This example assumes the Docker registry pod has a label of **docker-registry=default**. Pod anti-affinity can use any Kubernetes match expression.

The last required step is to enable the **MatchInterPodAffinity** scheduler predicate in **/etc/origin/master/scheduler.json**. With this in place, if only two infrastructure nodes are available and one is rebooted, the Docker registry pod is prevented from running on the other node. **oc get pods** reports the pod as unready until a suitable node is available. Once a node is available and all pods are back in ready state, the next node can be restarted.

### 2.9.3. Handling Nodes Running Routers

In most cases, a pod running an OpenShift Container Platform router will expose a host port. The **PodFitsPorts** scheduler predicate ensures that no router pods using the same port can run on the same node, and pod anti-affinity is achieved. If the routers are relying on [IP failover](#) for high availability, there is nothing else that is needed. For router pods relying on an external service such as AWS Elastic Load Balancing for high availability, it is that service's responsibility to react to router pod restarts.

In rare cases, a router pod may not have a host port configured. In those cases, it is important to follow the [recommended restart process](#) for infrastructure nodes.

## 2.10. CONFIGURING NODE RESOURCES

You can configure node resources by adding kubelet arguments to the node configuration file (**/etc/origin/node/node-config.yaml**). Add the **kubeletArguments** section and include any desired options:

```
kubeletArguments:
  max-pods: 1
    - "40"
  resolv-conf: 2
    - "/etc/resolv.conf"
  image-gc-high-threshold: 3
    - "90"
  image-gc-low-threshold: 4
    - "80"
```

- 1 Maximum number of pods that can run on this kubelet.
- 2 Resolver configuration file used as the basis for the container DNS resolution configuration.
- 3 The percent of disk usage after which image garbage collection is always run. Default: 90%
- 4 The percent of disk usage before which image garbage collection is never run. Lowest disk usage to garbage collect to. Default: 80%

To view all available kubelet options:

```
$ kubelet -h
```

This can also be set during an [advanced installation](#) using the `openshift_node_kubelet_args` variable. For example:

```
openshift_node_kubelet_args={'max-pods': ['40'], 'resolv-conf':
['/etc/resolv.conf'], 'image-gc-high-threshold': ['90'], 'image-gc-low-
threshold': ['80']}
```

### 2.10.1. Setting Maximum Pods Per Node

In the `/etc/origin/node/node-config.yaml` file, two parameters control the maximum number of pods that can be scheduled to a node: **pods-per-core** and **max-pods**. When both options are in use, the lower of the two limits the number of pods on a node. Exceeding these values can result in:

- Increased CPU utilization on both OpenShift Container Platform and Docker.
- Slow pod scheduling.
- Potential out-of-memory scenarios (depends on the amount of memory in the node).
- Exhausting the pool of IP addresses.
- Resource overcommitting, leading to poor user application performance.



#### NOTE

In Kubernetes, a pod that is holding a single container actually uses two containers. The second container is used to set up networking prior to the actual container starting. Therefore, a system running 10 pods will actually have 20 containers running.



**pods-per-core**  sets the number of pods the node can run based on the number of processor cores on the node. For example, if  **pods-per-core**  is set to  **10**  on a node with 4 processor cores, the maximum number of pods allowed on the node will be 40.

```
kubeletArguments:
  pods-per-core:
    - "10"
```



## NOTE

Setting  **pods-per-core**  to 0 disables this limit.

**max-pods**  sets the number of pods the node can run to a fixed value, regardless of the properties of the node.

```
kubeletArguments:
  max-pods:
    - "250"
```

Using the above example, the default value for  **pods-per-core**  is  **10**  and the default value for  **max-pods**  is  **250** . This means that unless the node has 25 cores or more, by default,  **pods-per-core**  will be the limiting factor.

## 2.11. RESETTING DOCKER STORAGE

As you download Docker images and run and delete containers, Docker does not always free up mapped disk space. As a result, over time you can run out of space on a node, which might prevent OpenShift Container Platform from being able to create new pods or cause pod creation to take several minutes.

For example, the following shows pods that are still in the  **ContainerCreating**  state after six minutes and the events log shows a  **FailedSync**  event.

```
$ oc get pod
NAME                                READY    STATUS
RESTARTS   AGE
cakephp-mysql-persistent-1-build    0/1      ContainerCreating    0
6m
mysql-1-9767d                       0/1      ContainerCreating    0
2m
mysql-1-deploy                      0/1      ContainerCreating    0
6m

$ oc get events
LASTSEEN   FIRSTSEEN   COUNT   NAME                                KIND
SUBJECT                                TYPE        REASON        MESSAGE
SOURCE
6m         6m         1       cakephp-mysql-persistent-1-build    Pod
Normal     Scheduled   default-scheduler
Successfully assigned cakephp-mysql-persistent-1-build to ip-172-31-71-
195.us-east-2.compute.internal
2m         5m         4       cakephp-mysql-persistent-1-build    Pod
Warning    FailedSync  kubelet, ip-172-31-71-195.us-
```

```

east-2.compute.internal    Error syncing pod
2m          4m          4          cakephp-mysql-persistent-1-build    Pod
Normal      SandboxChanged          kubelet, ip-172-31-71-195.us-
east-2.compute.internal    Pod sandbox changed, it will be killed and re-
created.

```

One solution to this problem is to reset Docker storage to remove artifacts not needed by Docker.

On the node where you want to restart Docker storage:

1. Run the following command to mark the node as unschedulable:

```
$ oc adm manage-node <node> --schedulable=false
```

2. Run the following command to shut down Docker and the **atomic-openshift-node** service:

```
$ systemctl stop docker atomic-openshift-node
```

3. Run the following command to remove the local volume directory:

```
$ rm -rf /var/lib/origin/openshift.local.volumes
```

This command clears the local image cache. As a result, images, including **ose-\*** images, will need to be re-pulled. This might result in slower pod start times while the image store recovers.

4. Remove the **/var/lib/docker** directory:

```
$ rm -rf /var/lib/docker
```

5. Run the following command to reset the Docker storage:

```
$ docker-storage-setup --reset
```

6. Run the following command to recreate the Docker storage:

```
$ docker-storage-setup
```

7. Recreate the **/var/lib/docker** directory:

```
$ mkdir /var/lib/docker
```

8. Run the following command to restart Docker and the **atomic-openshift-node** service:

```
$ systemctl start docker atomic-openshift-node
```

9. Run the following command to mark the node as schedulable:

```
$ oc adm manage-node <node> --schedulable=true
```

## 2.12. CHANGING NODE TRAFFIC INTERFACE

By default, DNS routes all node traffic. During node registration, the master receives the node IP addresses from the DNS configuration, and therefore accessing nodes via DNS is the most flexible solution for most deployments.

If your deployment is using a cloud provider, then the node gets the IP information from the cloud provider. However, **openshift-sdn** attempts to determine the IP through a variety of methods, including a DNS lookup on the nodeName (if set), or on the system hostname (if nodeName is not set).

However, you may need to change the node traffic interface. For example, where:

- OpenShift Container Platform is installed in a cloud provider where internal hostnames are not configured/resolvable by all hosts.
- The node's IP from the master's perspective is not the same as the node's IP from its own perspective.

Configuring the **openshift\_set\_node\_ip** Ansible variable forces node traffic through an interface other than the default network interface.

To change the node traffic interface:

1. Set the **openshift\_set\_node\_ip** Ansible variable to **true**.
2. Set the **openshift\_ip** to the IP address for the node you want to configure.

Although **openshift\_set\_node\_ip** can be useful as a workaround for the cases stated in this section, it is generally not suited for production environments. This is because the node will no longer function properly if it receives a new IP address.

## CHAPTER 3. MANAGING USERS

### 3.1. OVERVIEW

This topic describes the management of [user](#) accounts, including how new user accounts are created in OpenShift Container Platform and how they can be deleted.

### 3.2. ADDING A USER

After new users log in to OpenShift Container Platform, an account is created for that user per the [identity provider](#) configured on the master. The cluster administrator can [manage the access level of each user](#).

### 3.3. VIEWING USER AND IDENTITY LISTS

OpenShift Container Platform user configuration is stored in several locations within OpenShift Container Platform. Regardless of the identity provider, OpenShift Container Platform internally stores details like role-based access control (RBAC) information and group membership. To completely remove user information, this data must be removed in addition to the user account.

In OpenShift Container Platform, two object types contain user data outside the identification provider: **user** and **identity**.

To get the current list of users:

```
$ oc get user
```

| NAME               | UID                                  | FULL NAME | IDENTITIES |
|--------------------|--------------------------------------|-----------|------------|
| demo               | 75e4b80c-dbf1-11e5-8dc6-0e81e52cc949 |           |            |
| htpasswd_auth:demo |                                      |           |            |

To get the current list of identities:

```
$ oc get identity
```

| NAME                                 | IDP NAME      | IDP USER NAME | USER NAME | USER |
|--------------------------------------|---------------|---------------|-----------|------|
| htpasswd_auth:demo                   | htpasswd_auth | demo          | demo      |      |
| 75e4b80c-dbf1-11e5-8dc6-0e81e52cc949 |               |               |           |      |

Note the matching UID between the two object types. If you attempt to change the authentication provider after starting to use OpenShift Container Platform, the user names that overlap will not work because of the entries in the identity list, which will still point to the old authentication method.

### 3.4. MANAGING USER AND GROUP LABELS

To add a label to a user or group:

```
$ oc label user/<user_name> <label_name>
```

For example, if the user name is **theuser** and the label is **level=gold**:

```
$ oc label user/theuser level=gold
```

To remove the label:

```
$ oc label user/<user_name> <label_name>-
```

To show labels for a user or group:

```
$ oc describe user/<user_name>
```

## 3.5. DELETING A USER

To delete a user:

1. Delete the user record:

```
$ oc delete user demo
user "demo" deleted
```

2. Delete the user identity.

The identity of the user is related to the identification provider you use. Get the provider name from the user record in **oc get user**.

In this example, the identity provider name is **htpasswd\_auth**. The command is:

```
# oc delete identity htpasswd_auth:demo
identity "htpasswd_auth:demo" deleted
```

If you skip this step, the user will not be able to log in again.

After you complete these steps, a new account will be created in OpenShift Container Platform when the user logs in again.

If your intention is to prevent the user from being able to log in again (for example, if an employee has left the company and you want to permanently delete the account), you can also remove the user from your authentication back end (like **htpasswd**, **kerberos**, or others) for the configured identity provider.

For example, if you are using **htpasswd**, delete the entry in the **htpasswd** file that is configured for OpenShift Container Platform with the user name and password.

For external identification management like Lightweight Directory Access Protocol (LDAP) or Red Hat Identity Management (IdM), use the user management tools to remove the user entry.

## CHAPTER 4. MANAGING PROJECTS

### 4.1. OVERVIEW

In OpenShift Container Platform, projects are used to group and isolate related objects. As an administrator, you can give developers access to certain projects, allow them to create their own, and give them administrative rights within individual projects.

### 4.2. SELF-PROVISIONING PROJECTS

You can allow developers to create their own projects. There is an endpoint that will provision a project according to a [template](#). The web console and `oc new-project` command use this endpoint when a developer [creates a new project](#).

#### 4.2.1. Modifying the Template for New Projects

The API server automatically provisions projects based on the template that is identified by the `projectRequestTemplate` parameter of the *master-config.yaml* file. If the parameter is not defined, the API server creates a default template that creates a project with the requested name, and assigns the requesting user to the "admin" role for that project.

To create your own custom project template:

1. Start with the current default project template:

```
$ oc adm create-bootstrap-project-template -o yaml > template.yaml
```

2. Use a text editor to modify the *template.yaml* file by adding objects or modifying existing objects.

3. Load the template:

```
$ oc create -f template.yaml -n default
```

4. Modify the *master-config.yaml* file to reference the loaded template:

```
...
projectConfig:
  projectRequestTemplate: "default/project-request"
...
```

When a project request is submitted, the API substitutes the following parameters into the template:

| Parameter                  | Description                                    |
|----------------------------|--|
| <b>PROJECT_NAME</b>        | The name of the project. Required.             |
| <b>PROJECT_DISPLAYNAME</b> | The display name of the project. May be empty. |
| <b>PROJECT_DESCRIPTION</b> | The description of the project. May be empty.  |

| Parameter                      | Description                              |
|--------------------------------|--|
| <b>PROJECT_ADMIN_USER</b>      | The username of the administrating user. |
| <b>PROJECT_REQUESTING_USER</b> | The username of the requesting user.     |

Access to the API is granted to developers with the **self-provisioner** role and the **self-provisioners** cluster role binding. This role is available to all authenticated developers by default.

### 4.2.2. Disabling Self-provisioning

You can prevent an authenticated user group from self-provisioning new projects.

1. Log in as a user with **cluster-admin** privileges.
2. Remove the **self-provisioners** cluster role from the group.

```
$ oc adm policy remove-cluster-role-from-group self-provisioner
system:authenticated system:authenticated:oauth
```

3. Set the **projectRequestMessage** parameter value in the **master-config.yaml** file to instruct developers how to request a new project. This parameter value is a string that will be presented to a user in the web console and command line when the user attempts to self-provision a project. You might use one of the following messages:
  - To request a project, contact your system administrator at [projectname@example.com](mailto:projectname@example.com).
  - To request a new project, fill out the project request form located at <https://internal.example.com/openshift-project-request>.

#### Example YAML file

```
...
projectConfig:
  ProjectRequestMessage: "message"
...
```

4. Edit the **self-provisioners** cluster role to prevent **automatic updates** to the role. Automatic updates reset the cluster roles to the default state.
  - To update the role from the command line:
    - i. Run the following command:

```
$ oc edit clusterrole self-provisioner
```

- ii. In the displayed role, set the **openshift.io/reconcile-protect** parameter value to **true**, as shown in the following example:

```
apiVersion: authorization.openshift.io/v1
kind: ClusterRole
```

```

metadata:
  annotations:
    authorization.openshift.io/system-only: "true"
    openshift.io/description: A user that can request project.
    openshift.io/reconcile-protect: "true"
  ...

```

- To update the role by using automation, use the following command:

```

$ oc patch clusterrole self-provisioner -p '{ "metadata": {
  "annotations": { "openshift.io/reconcile-protect": "true" } } }'

```

## 4.3. USING NODE SELECTORS

Node selectors are used in conjunction with labeled nodes to control pod placement.



### NOTE

Labels can be assigned [during an advanced installation](#), or [added to a node after installation](#).

### 4.3.1. Setting the Cluster-wide Default Node Selector

As a cluster administrator, you can set the cluster-wide default node selector to restrict pod placement to specific nodes.

Edit the master configuration file at `/etc/origin/master/master-config.yaml` and add a value for a default node selector. This is applied to the pods created in all projects without a specified `nodeSelector` value:

```

...
projectConfig:
  defaultNodeSelector: "type=user-node,region=east"
...

```

Restart the OpenShift service for the changes to take effect:

```

# systemctl restart atomic-openshift-master-api atomic-openshift-master-
  controllers

```

### 4.3.2. Setting the Project-wide Node Selector

To create an individual project with a node selector, use the `--node-selector` option when creating a project. For example, if you have an OpenShift Container Platform topology with multiple regions, you can use a node selector to restrict specific OpenShift Container Platform projects to only deploy pods onto nodes in a specific region.

The following creates a new project named `myproject` and dictates that pods be deployed onto nodes labeled `user-node` and `east`:

```

$ oc adm new-project myproject \
  --node-selector='type=user-node,region=east'

```



Once this command is run, this becomes the administrator-set node selector for all pods contained in the specified project.



## NOTE

While the **new-project** subcommand is available for both **oc adm** and **oc**, the cluster administrator and developer commands respectively, creating a new project with a node selector is only available with the **oc adm** command. The **new-project** subcommand is not available to project developers when self-provisioning projects.

Using the **oc adm new-project** command adds an **annotation** section to the project. You can edit a project, and change the **openshift.io/node-selector** value to override the default:

```
...
metadata:
  annotations:
    openshift.io/node-selector: type=user-node,region=east
...
```

You can also override the default value for an existing project namespace by using the following command:

```
# oc patch namespace myproject -p \
  '{"metadata":{"annotations":{"openshift.io/node-selector":"region=infra"}}}'
```

If **openshift.io/node-selector** is set to an empty string (**oc adm new-project --node-selector=""**), the project will not have an administrator-set node selector, even if the cluster-wide default has been set. This means that, as a cluster administrator, you can set a default to restrict developer projects to a subset of nodes and still enable infrastructure or other projects to schedule the entire cluster.

### 4.3.3. Developer-specified Node Selectors

OpenShift Container Platform developers [can set a node selector on their pod configuration](#) if they wish to restrict nodes even further. This will be in addition to the project node selector, meaning that you can still dictate node selector values for all projects that have a node selector value.

For example, if a project has been created with the above annotation (**openshift.io/node-selector: type=user-node,region=east**) and a developer sets another node selector on a pod in that project, for example **clearance=classified**, the pod will only ever be scheduled on nodes that have all three labels (**type=user-node**, **region=east**, and **clearance=classified**). If they set **region=west** on a pod, their pods would be demanding nodes with labels **region=east** and **region=west**, which cannot work. The pods will never be scheduled, because labels can only be set to one value.

## 4.4. LIMITING NUMBER OF SELF-PROVISIONED PROJECTS PER USER

The number of self-provisioned projects requested by a given user can be limited with the **ProjectRequestLimit**[admission control plug-in](#).



## IMPORTANT

If your project request template was created in OpenShift Container Platform 3.1 or earlier using the process described in [Modifying the Template for New Projects](#), then the generated template does not include the annotation **openshift.io/requester: \${PROJECT\_REQUESTING\_USER}**, which is used for the **ProjectRequestLimitConfig**. You must add the annotation.

In order to specify limits for users, a configuration must be specified for the plug-in within the master configuration file (**/etc/origin/master/master-config.yaml**). The plug-in configuration takes a list of user label selectors and the associated maximum project requests.

Selectors are evaluated in order. The first one matching the current user will be used to determine the maximum number of projects. If a selector is not specified, a limit applies to all users. If a maximum number of projects is not specified, then an unlimited number of projects are allowed for a specific selector.

The following configuration sets a global limit of 2 projects per user while allowing 10 projects for users with a label of **level=advanced** and unlimited projects for users with a label of **level=admin**.

```
admissionConfig:
  pluginConfig:
    ProjectRequestLimit:
      configuration:
        apiVersion: v1
        kind: ProjectRequestLimitConfig
        limits:
          - selector:
              level: admin ❶
          - selector:
              level: advanced ❷
            maxProjects: 10
          - maxProjects: 2 ❸
```

- ❶ For selector **level=admin**, no **maxProjects** is specified. This means that users with this label will not have a maximum of project requests.
- ❷ For selector **level=advanced**, a maximum number of 10 projects will be allowed.
- ❸ For the third entry, no selector is specified. This means that it will be applied to any user that doesn't satisfy the previous two rules. Because rules are evaluated in order, this rule should be specified last.



## NOTE

[Managing User and Group Labels](#) provides further guidance on how to add, remove, or show labels for users and groups.

Once your changes are made, restart OpenShift Container Platform for the changes to take effect.

```
# systemctl restart atomic-openshift-master-api atomic-openshift-master-controllers
```

## CHAPTER 5. MANAGING PODS

### 5.1. OVERVIEW

This topic describes the management of [pods](#), including limiting their run-once duration, and how much bandwidth they can use.

### 5.2. LIMITING RUN-ONCE POD DURATION

OpenShift Container Platform relies on run-once pods to perform tasks such as deploying a pod or performing a build. Run-once pods are pods that have a **RestartPolicy** of **Never** or **OnFailure**.

The cluster administrator can use the **RunOnceDuration** admission control plug-in to force a limit on the time that those run-once pods can be active. Once the time limit expires, the cluster will try to actively terminate those pods. The main reason to have such a limit is to prevent tasks such as builds to run for an excessive amount of time.

#### 5.2.1. Configuring the RunOnceDuration Plug-in

The plug-in configuration should include the default active deadline for run-once pods. This deadline is enforced globally, but can be superseded on a per-project basis.

```
admissionConfig:
  pluginConfig:
    RunOnceDuration:
      configuration:
        apiVersion: v1
        kind: RunOnceDurationConfig
        activeDeadlineSecondsOverride: 3600 ❶
    ....
```

- ❶ Specify the global default for run-once pods in seconds.

#### 5.2.2. Specifying a Custom Duration per Project

In addition to specifying a global maximum duration for run-once pods, an administrator can add an annotation (**openshift.io/active-deadline-seconds-override**) to a specific project to override the global default.

- For a new project, define the annotation in the project specification *.yaml* file.

```
apiVersion: v1
kind: Project
metadata:
  annotations:
    openshift.io/active-deadline-seconds-override: "1000" ❶
  name: myproject
```

- ❶ Overrides the default active deadline seconds for run-once pods to 1000 seconds. Note that the value of the override must be specified in string form.

- For an existing project,
  - Run **oc edit** and add the **openshift.io/active-deadline-seconds-override: 1000** annotation in the editor.

```
$ oc edit namespace <project-name>
```

Or

- Use the **oc patch** command:

```
$ oc patch namespace <project_name> -p '{"metadata":
{"annotations":{"openshift.io/active-deadline-seconds-
override":"1000"}}}'
```

### 5.2.2.1. Deploying an Egress Router Pod

#### Example 5.1. Example Pod Definition for an Egress Router

```
apiVersion: v1
kind: Pod
metadata:
  name: egress-1
  labels:
    name: egress-1
  annotations:
    pod.network.openshift.io/assign-macvlan: "true"
spec:
  containers:
  - name: egress-router
    image: openshift3/ose-egress-router
    securityContext:
      privileged: true
    env:
      - name: EGRESS_SOURCE 1
        value: 192.168.12.99
      - name: EGRESS_GATEWAY 2
        value: 192.168.12.1
      - name: EGRESS_DESTINATION 3
        value: 203.0.113.25
  nodeSelector:
    site: springfield-1 4
```

- 1 IP address on the node subnet reserved by the cluster administrator for use by this pod.
- 2 Same value as the default gateway used by the node itself.
- 3 Connections to the pod are redirected to 203.0.113.25, with a source IP address of 192.168.12.99
- 4 The pod will only be deployed to nodes with the label site **springfield-1**.

The `pod.network.openshift.io/assign-macvlan` annotation creates a Macvlan network interface on the primary network interface, and then moves it into the pod's network name space before starting the **egress-router** container.



## NOTE

Preserve the the quotation marks around **"true"**. Omitting them will result in errors.

The pod contains a single container, using the **openshift3/ose-egress-router** image, and that container is run privileged so that it can configure the Macvlan interface and set up **iptables** rules.

The environment variables tell the **egress-router** image what addresses to use; it will configure the Macvlan interface to use **EGRESS\_SOURCE** as its IP address, with **EGRESS\_GATEWAY** as its gateway.

NAT rules are set up so that connections to any TCP or UDP port on the pod's cluster IP address are redirected to the same port on **EGRESS\_DESTINATION**.

If only some of the nodes in your cluster are capable of claiming the specified source IP address and using the specified gateway, you can specify a **nodeName** or **nodeSelector** indicating which nodes are acceptable.

### 5.2.2.2. Deploying an Egress Router Service

Though not strictly necessary, you normally want to create a service pointing to the egress router:

```
apiVersion: v1
kind: Service
metadata:
  name: egress-1
spec:
  ports:
    - name: http
      port: 80
    - name: https
      port: 443
  type: ClusterIP
  selector:
    name: egress-1
```

Your pods can now connect to this service. Their connections are redirected to the corresponding ports on the external server, using the reserved egress IP address.

### 5.2.3. Limiting Pod Access with Egress Firewall

As an OpenShift Container Platform cluster administrator, you can use egress policy to limit the external addresses that some or all pods can access from within the cluster, so that:

- A pod can only talk to internal hosts, and cannot initiate connections to the public Internet.  
Or,
- A pod can only talk to the public Internet, and cannot initiate connections to internal hosts (outside the cluster).  
Or,

- A pod cannot reach specified internal subnets/hosts that it should have no reason to contact.

For example, you can configure projects with different egress policies, allowing **<project A>** access to a specified IP range, but denying the same access to **<project B>**.

## CAUTION

You must have the [ovs-multitenant plug-in](#) enabled in order to limit pod access via egress policy.

Project administrators can neither create **EgressNetworkPolicy** objects, nor edit the ones you create in their project. There are also several other restrictions on where **EgressNetworkPolicy** can be created:

1. The **default** project (and any other project that has been made global via **oc adm pod-network make-projects-global**) cannot have egress policy.
2. If you merge two projects together (via **oc adm pod-network join-projects**), then you cannot use egress policy in *any* of the joined projects.
3. No project may have more than one egress policy object.

Violating any of these restrictions will result in broken egress policy for the project, and may cause all external network traffic to be dropped.

### 5.2.3.1. Configuring Pod Access Limits

To configure pod access limits, you must use the **oc** command or the REST API. You can use **oc [create|replace|delete]** to manipulate **EgressNetworkPolicy** objects. The **api/swagger-spec/oapi-v1.json** file has API-level details on how the objects actually work.

To configure pod access limits:

1. Navigate to the project you want to affect.
2. Create a JSON file for the pod limit policy:

```
# oc create -f <policy>.json
```

3. Configure the JSON file with policy details. For example:

```
{
  "kind": "EgressNetworkPolicy",
  "apiVersion": "v1",
  "metadata": {
    "name": "default"
  },
  "spec": {
    "egress": [
      {
        "type": "Allow",
        "to": {
          "cidrSelector": "1.2.3.0/24"
        }
      },
      {

```

```

        "type": "Allow",
        "to": {
            "dnsName": "www.foo.com"
        }
    },
    {
        "type": "Deny",
        "to": {
            "cidrSelector": "0.0.0.0/0"
        }
    }
]
}

```

When the example above is added in a project, it allows traffic to IP range **1.2.3.0/24** and domain name **www.foo.com**, but denies access to all other external IP addresses. (Traffic to other pods is not affected because the policy only applies to *external* traffic.)

The rules in an **EgressNetworkPolicy** are checked in order, and the first one that matches takes effect. If the three rules in the above example were reversed, then traffic would not be allowed to **1.2.3.0/24** and **www.foo.com** because the **0.0.0.0/0** rule would be checked first, and it would match and deny all traffic.

Domain name updates are reflected within 30 minutes. In the above example, suppose **www.foo.com** resolved to **10.11.12.13**, but later it was changed to **20.21.22.23**. Then, OpenShift Container Platform will take up to 30 minutes to adapt to these DNS updates.

### 5.3. LIMITING THE BANDWIDTH AVAILABLE TO PODS

You can apply quality-of-service traffic shaping to a pod and effectively limit its available bandwidth. Egress traffic (from the pod) is handled by policing, which simply drops packets in excess of the configured rate. Ingress traffic (to the pod) is handled by shaping queued packets to effectively handle data. The limits you place on a pod do not affect the bandwidth of other pods.

To limit the bandwidth on a pod:

1. Write an object definition JSON file, and specify the data traffic speed using **kubernetes.io/ingress-bandwidth** and **kubernetes.io/egress-bandwidth** annotations. For example, to limit both pod egress and ingress bandwidth to 10M/s:

#### Limited Pod Object Definition

```

{
  "kind": "Pod",
  "spec": {
    "containers": [
      {
        "image": "openshift/hello-openshift",
        "name": "hello-openshift"
      }
    ]
  },
  "apiVersion": "v1",
  "metadata": {

```

```

    "name": "iperf-slow",
    "annotations": {
      "kubernetes.io/ingress-bandwidth": "10M",
      "kubernetes.io/egress-bandwidth": "10M"
    }
  }
}

```

2. Create the pod using the object definition:

```
oc create -f <file_or_dir_path>
```

## 5.4. SETTING POD DISRUPTION BUDGETS

A *pod disruption budget* is part of the [Kubernetes](#) API, which can be managed with **oc** commands like other [object types](#). They allow the specification of safety constraints on pods during operations, such as draining a node for maintenance.



### NOTE

Starting in OpenShift Container Platform 3.6, pod disruption budgets are now fully supported.

**PodDisruptionBudget** is an API object that specifies the minimum number or percentage of replicas that must be up at a time. Setting these in projects can be helpful during node maintenance (such as scaling a cluster down or a cluster upgrade) and is only honored on voluntary evictions (not on node failures).

A **PodDisruptionBudget** object's configuration consists of the following key parts:

- A label selector, which is a label query over a set of pods.
- An availability level, which specifies the minimum number of pods that must be available simultaneously.

The following is an example of a **PodDisruptionBudget** resource:

```

apiVersion: policy/v1beta1 ❶
kind: PodDisruptionBudget
metadata:
  name: my-pdb
spec:
  selector: ❷
    matchLabels:
      foo: bar
  minAvailable: 2 ❸

```

- ❶ **PodDisruptionBudget** is part of the **policy/v1beta1** API group.
- ❷ A label query over a set of resources. The result of **matchLabels** and **matchExpressions** are logically conjoined.

❸



The minimum number of pods that must be available simultaneously. This can be either an integer or a string specifying a percentage (for example, **20%**).

If you created a YAML file with the above object definition, you could add it to project with the following:

```
$ oc create -f </path/to/file> -n <project_name>
```

You can check for pod disruption budgets across all projects with the following:

```
$ oc get poddisruptionbudget --all-namespaces
```

| NAMESPACE       | NAME        | MIN-AVAILABLE | SELECTOR |
|-----------------|-------------|---------------|----------|
| another-project | another-pdb | 4             | bar=foo  |
| test-project    | my-pdb      | 2             | foo=bar  |

The **PodDisruptionBudget** is considered healthy when there are at least **minAvailable** pods running in the system. Every pod above that limit can be [evicted](#).

## CHAPTER 6. MANAGING NETWORKING

### 6.1. OVERVIEW

This topic describes the management of the overall [cluster network](#), including project isolation and outbound traffic control.

Pod-level networking features, such as per-pod bandwidth limits, are discussed in [Managing Pods](#).

### 6.2. MANAGING POD NETWORKS

When your cluster is configured to use [the ovs-multitenant SDN plug-in](#), you can manage the separate pod overlay networks for projects using the administrator CLI. See the [Configuring the SDN](#) section for plug-in configuration steps, if necessary.

#### 6.2.1. Joining Project Networks

To join projects to an existing project network:

```
$ oc adm pod-network join-projects --to=<project1> <project2> <project3>
```

In the above example, all the pods and services in **<project2>** and **<project3>** can now access any pods and services in **<project1>** and vice versa. Services can be accessed either by IP or fully-qualified DNS name (**<service>.<pod\_namespace>.svc.cluster.local**). For example, to access a service named **db** in a project **myproject**, use **db.myproject.svc.cluster.local**.

Alternatively, instead of specifying specific project names, you can use the **--selector=<project\_selector>** option.

### 6.3. ISOLATING PROJECT NETWORKS

To isolate the project network in the cluster and vice versa, run:

```
$ oc adm pod-network isolate-projects <project1> <project2>
```

In the above example, all of the pods and services in **<project1>** and **<project2>** can *not* access any pods and services from other non-global projects in the cluster and vice versa.

Alternatively, instead of specifying specific project names, you can use the **--selector=<project\_selector>** option.

#### 6.3.1. Making Project Networks Global

To allow projects to access all pods and services in the cluster and vice versa:

```
$ oc adm pod-network make-projects-global <project1> <project2>
```

In the above example, all the pods and services in **<project1>** and **<project2>** can now access any pods and services in the cluster and vice versa.

Alternatively, instead of specifying specific project names, you can use the `--selector=<project_selector>` option.

## 6.4. DISABLING HOST NAME COLLISION PREVENTION FOR ROUTES AND INGRESS OBJECTS

In OpenShift Container Platform, host name collision prevention for routes and ingress objects is enabled by default. This means that users without the **cluster-admin** role can set the host name in a route or ingress object only on creation and cannot change it afterwards. However, you can relax this restriction on routes and ingress objects for some or all users.



### WARNING

Because OpenShift Container Platform uses the object creation timestamp to determine the oldest route or ingress object for a given host name, a route or ingress object can hijack a host name of a newer route if the older route changes its host name, or if an ingress object is introduced.

As an OpenShift Container Platform cluster administrator, you can edit the host name in a route even after creation. You can also create a role to allow specific users to do so:

```
$ oc create -f - <<EOF
apiVersion: v1
kind: ClusterRole
metadata:
  name: route-editor
rules:
- apiGroups:
  - route.openshift.io
  - ""
  resources:
  - routes/custom-host
  verbs:
  - update
EOF
```

You can then bind the new role to a user:

```
$ oc adm policy add-cluster-role-to-user route-editor user
```

You can also disable host name collision prevention for ingress objects. Doing so lets users without the **cluster-admin** role edit a host name for ingress objects after creation. This is useful to OpenShift Container Platform installations that depend upon Kubernetes behavior, including allowing the host names in ingress objects be edited.

1. Add the following to the **master.yaml** file:

```
admissionConfig:
  pluginConfig:
```

```

openshift.io/IngressAdmission:
  configuration:
    apiVersion: v1
    allowHostnameChanges: true
    kind: IngressAdmissionConfig
    location: ""

```

2. Restart the master services for the changes to take effect:

```

$ systemctl restart atomic-openshift-master-api atomic-openshift-
master-controllers

```

## 6.5. CONTROLLING EGRESS TRAFFIC

As a cluster administrator you can allocate a number of static IP addresses to a specific node at the host level. If an application developer needs a dedicated IP address for their application service, they can request one during the process they use to ask for firewall access. They can then deploy an egress router from the developer's project, using a **nodeSelector** in the deployment configuration to ensure that the pod lands on the host with the pre-allocated static IP address.

The egress pod's deployment declares one of the source IPs, the destination IP of the protected service, and a gateway IP to reach the destination. After the pod is deployed, you can [create a service](#) to access the egress router pod, then add that source IP to the corporate firewall. The developer then has access information to the egress router service that was created in their project, for example, **service.project.cluster.domainname.com**.

When the developer needs to access the external, firewalled service, they can call out to the egress router pod's service (**service.project.cluster.domainname.com**) in their application (for example, the JDBC connection information) rather than the actual protected service URL.

As an OpenShift Container Platform cluster administrator, you can control egress traffic in these ways:

### Firewall

Using an egress firewall allows you to enforce the acceptable outbound traffic policies, so that specific endpoints or IP ranges (subnets) are the only acceptable targets for the dynamic endpoints (pods within OpenShift Container Platform) to talk to.

### Router

Using an egress router allows you to create identifiable services to send traffic to certain destinations, ensuring those external destinations treat traffic as though it were coming from a known source. This helps with security, because it allows you to secure an external database so that only specific pods in a namespace can talk to a service (the egress router), which proxies the traffic to your database.

### iptables

In addition to the above OpenShift Container Platform-internal solutions, it is also possible to create iptables rules that will be applied to outgoing traffic. These rules allow for more possibilities than the egress firewall, but cannot be limited to particular projects.

### 6.5.1. Using an Egress Firewall to Limit Access to External Resources

As an OpenShift Container Platform cluster administrator, you can use egress firewall policy to limit the external addresses that some or all pods can access from within the cluster, so that:

- A pod can only talk to internal hosts, and cannot initiate connections to the public Internet.
- Or,

- A pod can only talk to the public Internet, and cannot initiate connections to internal hosts (outside the cluster).  
Or,
- A pod cannot reach specified internal subnets/hosts that it should have no reason to contact.

You can configure projects to have different egress policies. For example, allowing **<project A>** access to a specified IP range, but denying the same access to **<project B>**. Or restrict application developers from updating from (Python) pip mirrors, and forcing updates to only come from desired sources.

## CAUTION

You must have the [ovs-multitenant plug-in](#) enabled in order to limit pod access via egress policy.

Project administrators can neither create **EgressNetworkPolicy** objects, nor edit the ones you create in their project. There are also several other restrictions on where **EgressNetworkPolicy** can be created:

- The **default** project (and any other project that has been made global via **oc adm pod-network make-projects-global**) cannot have egress policy.
- If you merge two projects together (via **oc adm pod-network join-projects**), then you cannot use egress policy in *any* of the joined projects.
- No project may have more than one egress policy object.

Violating any of these restrictions results in broken egress policy for the project, and may cause all external network traffic to be dropped.

Use the **oc** command or the REST API to configure egress policy. You can use **oc [create|replace|delete]** to manipulate **EgressNetworkPolicy** objects. The *api/swagger-spec/oapi-v1.json* file has API-level details on how the objects actually work.

To configure egress policy:

1. Navigate to the project you want to affect.
2. Create a JSON file with the desired policy details. For example:

```
{
  "kind": "EgressNetworkPolicy",
  "apiVersion": "v1",
  "metadata": {
    "name": "default"
  },
  "spec": {
    "egress": [
      {
        "type": "Allow",
        "to": {
          "cidrSelector": "1.2.3.0/24"
        }
      },
      {

```

```

        "type": "Allow",
        "to": {
            "dnsName": "www.foo.com"
        }
    },
    {
        "type": "Deny",
        "to": {
            "cidrSelector": "0.0.0.0/0"
        }
    }
]
}

```

When the example above is added to a project, it allows traffic to IP range **1.2.3.0/24** and domain name **www.foo.com**, but denies access to all other external IP addresses. Traffic to other pods is not affected because the policy only applies to *external* traffic.

The rules in an **EgressNetworkPolicy** are checked in order, and the first one that matches takes effect. If the three rules in the above example were reversed, then traffic would not be allowed to **1.2.3.0/24** and **www.foo.com** because the **0.0.0.0/0** rule would be checked first, and it would match and deny all traffic.

Domain name updates are polled based on the TTL (time to live) value of the domain returned by the local non-authoritative servers. The pod should also resolve the domain from the same local nameservers when necessary, otherwise the IP addresses for the domain perceived by the egress network policy controller and the pod will be different, and the egress network policy may not be enforced as expected. Since egress network policy controller and pod are asynchronously polling the same local nameserver, there could be a race condition where pod may get the updated IP before the egress controller. Due to this current limitation, domain name usage in **EgressNetworkPolicy** is only recommended for domains with infrequent IP address changes.



## NOTE

The egress firewall always allows pods access to the external interface of the node the pod is on for DNS resolution. If your DNS resolution is not handled by something on the local node, then you will need to add egress firewall rules allowing access to the DNS server's IP addresses if you are using domain names in your pods. The [default installer](#) sets up a local dnsmasq, so if you are using that setup you will not need to add extra rules.

3. Use the JSON file to create an EgressNetworkPolicy object:

```
# oc create -f <policy>.json
```

## CAUTION

Exposing services by creating [routes](#) will ignore **EgressNetworkPolicy**. Egress network policy service endpoint filtering is done at the node **kubeproxy**. When the router is involved, **kubeproxy** is bypassed and egress network policy enforcement is not applied. Administrators can prevent this bypass by limiting access to create routes.

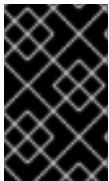
### 6.5.2. Using an Egress Router to Allow External Resources to Recognize Pod Traffic

The OpenShift Container Platform egress router runs a service that redirects traffic to a specified remote server, using a private source IP address that is not used for anything else. The service allows pods to talk to servers that are set up to only allow access from whitelisted IP addresses.



#### IMPORTANT

The egress router is not intended for every outgoing connection. Creating large numbers of egress routers can push the limits of your network hardware. For example, creating an egress router for every project or application could exceed the number of local MAC addresses that the network interface can handle before falling back to filtering MAC addresses in software.



#### IMPORTANT

Currently, the egress router is not compatible with Amazon AWS, Azure Cloud, or any other cloud platform that does not support layer 2 manipulations due to their incompatibility with macvlan traffic.

### Deployment Considerations

The Egress router adds a second IP address and MAC address to the node's primary network interface. If you are not running OpenShift Container Platform on bare metal, you may need to configure your hypervisor or cloud provider to allow the additional address.

#### Red Hat OpenStack Platform

If you are deploying OpenShift Container Platform on Red Hat OpenStack Platform, you need to whitelist the IP and MAC addresses on your OpenStack environment, otherwise [communication will fail](#):

```
neutron port-update $neutron_port_uuid \
  --allowed_address_pairs list=true \
  type=dict mac_address=<mac_address>,ip_address=<ip_address>
```

#### Red Hat Enterprise Virtualization

If you are using [Red Hat Enterprise Virtualization](#), you should set **EnableMACAntiSpoofingFilterRules** to **false**.

#### VMware vSphere

If you are using VMware vSphere, see the [VMWare documentation for securing vSphere standard switches](#). View and change VMWare vSphere default settings by selecting the host's virtual switch from the vSphere Web Client.

Specifically, ensure that the following are enabled:

- [MAC Address Changes](#)
- [Forged Transits](#)
- [Promiscuous Mode Operation](#)

### Egress Router Modes

The egress router can run in two different modes: [redirect mode](#) and [HTTP proxy mode](#). Redirect mode works for all services except for HTTP and HTTPS. For HTTP and HTTPS services, use HTTP proxy mode.

### 6.5.2.1. Deploying an Egress Router Pod in Redirect Mode

In *redirect mode*, the egress router sets up iptables rules to redirect traffic from its own IP address to one or more destination IP addresses. Client pods that want to make use of the reserved source IP address must be modified to connect to the egress router rather than connecting directly to the destination IP.

1. Create a pod configuration using the following:

```
apiVersion: v1
kind: Pod
metadata:
  name: egress-1
  labels:
    name: egress-1
  annotations:
    pod.network.openshift.io/assign-macvlan: "true" 1
spec:
  initContainers:
    - name: egress-router
      image: registry.access.redhat.com/openshift3/ose-egress-router
      securityContext:
        privileged: true
      env:
        - name: EGRESS_SOURCE 2
          value: 192.168.12.99
        - name: EGRESS_GATEWAY 3
          value: 192.168.12.1
        - name: EGRESS_DESTINATION 4
          value: 203.0.113.25
        - name: EGRESS_ROUTER_MODE 5
          value: init
  containers:
    - name: egress-router-wait
      image: registry.access.redhat.com/openshift3/ose-pod
  nodeSelector:
    site: springfield-1 6
```

- 1 The **pod.network.openshift.io/assign-macvlan** annotation creates a Macvlan network interface on the primary network interface, and then moves it into the pod's network name space before starting the **egress-router** container. Preserve the quotation marks around **"true"**. Omitting them results in errors.
- 2 IP address from the physical network that the node is on and is reserved by the cluster administrator for use by this pod.
- 3 Same value as the default gateway used by the node.
- 4 The external server to direct traffic to. Using this example, connections to the pod are redirected to 203.0.113.25, with a source IP address of 192.168.12.99.
- 5



This tells the egress router image that it is being deployed as an "init container". Previous versions of OpenShift Container Platform (and the egress router image) did not support this

- 6 The pod is only deployed to nodes with the label **site=springfield-1**.

2. Create the pod using the above definition:

```
$ oc create -f <pod_name>.json
```

To check to see if the pod has been created:

```
oc get pod <pod_name>
```

3. Ensure other pods can find the pod's IP address by creating a service to point to the egress router:

```
apiVersion: v1
kind: Service
metadata:
  name: egress-1
spec:
  ports:
    - name: http
      port: 80
    - name: https
      port: 443
  type: ClusterIP
  selector:
    name: egress-1
```

Your pods can now connect to this service. Their connections are redirected to the corresponding ports on the external server, using the reserved egress IP address.

The egress router setup is performed by an "init container" created from the **openshift3/ose-egress-router** image, and that container is run privileged so that it can configure the Macvlan interface and set up **iptables** rules. After it finishes setting up the **iptables** rules, it exits and the **openshift3/ose-pod** container will run (doing nothing) until the pod is killed.

The environment variables tell the **egress-router** image what addresses to use; it will configure the Macvlan interface to use **EGRESS\_SOURCE** as its IP address, with **EGRESS\_GATEWAY** as its gateway.

NAT rules are set up so that connections to any TCP or UDP port on the pod's cluster IP address are redirected to the same port on **EGRESS\_DESTINATION**.

If only some of the nodes in your cluster are capable of claiming the specified source IP address and using the specified gateway, you can specify a **nodeName** or **nodeSelector** indicating which nodes are acceptable.

### 6.5.2.2. Redirecting to Multiple Destinations

In the previous example, connections to the egress pod (or its corresponding service) on any port are redirected to a single destination IP. You can also configure different destination IPs depending on the port:

```

apiVersion: v1
kind: Pod
metadata:
  name: egress-multi
  labels:
    name: egress-multi
  annotations:
    pod.network.openshift.io/assign-macvlan: "true"
spec:
  initContainers:
  - name: egress-router
    image: registry.access.redhat.com/openshift3/ose-egress-router
    securityContext:
      privileged: true
  env:
  - name: EGRESS_SOURCE
    value: 192.168.12.99
  - name: EGRESS_GATEWAY
    value: 192.168.12.1
  - name: EGRESS_DESTINATION
    value: | 1
            80    tcp 203.0.113.25
            8080  tcp 203.0.113.26 80
            8443  tcp 203.0.113.26 443
            203.0.113.27
  - name: EGRESS_ROUTER_MODE
    value: init
  containers:
  - name: egress-router-wait
    image: registry.access.redhat.com/openshift3/ose-pod

```

**1** This uses the YAML syntax for a multi-line string; see below for details.

Each line of **EGRESS\_DESTINATION** can be one of three types:

- **<port> <protocol> <IP address>** - This says that incoming connections to the given **<port>** should be redirected to the same port on the given **<IP address>**. **<protocol>** is either **tcp** or **udp**. In the example above, the first line redirects traffic from local port 80 to port 80 on 203.0.113.25.
- **<port> <protocol> <IP address> <remote port>** - As above, except that the connection is redirected to a different **<remote port>** on **<IP address>**. In the example above, the second and third lines redirect local ports 8080 and 8443 to remote ports 80 and 443 on 203.0.113.26.
- **<fallback IP address>** - If the last line of **EGRESS\_DESTINATION** is a single IP address, then any connections on any other port will be redirected to the corresponding port on that IP address (eg, 203.0.113.27 in the example above). If there is no fallback IP address then connections on other ports would simply be rejected.)

### 6.5.2.3. Using a ConfigMap to specify EGRESS\_DESTINATION

For a large or frequently-changing set of destination mappings, you can use a ConfigMap to externally maintain the list, and have the egress router pod read it from there. This comes with the advantage of

project administrators being able to edit the ConfigMap, whereas they may not be able to edit the Pod definition directly, because it contains a privileged container.

1. Create a file containing the **EGRESS\_DESTINATION** data:

```
$ cat my-egress-destination.txt
# Egress routes for Project "Test", version 3

80    tcp 203.0.113.25

8080  tcp 203.0.113.26 80
8443  tcp 203.0.113.26 443

# Fallback
203.0.113.27
```

Note that you can put blank lines and comments into this file

2. Create a ConfigMap object from the file:

```
$ oc delete configmap egress-routes --ignore-not-found
$ oc create configmap egress-routes \
  --from-file=destination=my-egress-destination.txt
```

Here **egress-routes** is the name of the ConfigMap object being created and **my-egress-destination.txt** is the name of the file the data is being read from.

3. Create a egress router pod definition as above, but specifying the ConfigMap for **EGRESS\_DESTINATION** in the environment section:

```
...
env:
- name: EGRESS_SOURCE
  value: 192.168.12.99
- name: EGRESS_GATEWAY
  value: 192.168.12.1
- name: EGRESS_DESTINATION
  valueFrom:
    configMapKeyRef:
      name: egress-routes
      key: destination
- name: EGRESS_ROUTER_MODE
  value: init
...
```



#### NOTE

The egress router does not automatically update when the ConfigMap changes. Restart the pod to get updates.

#### 6.5.2.4. Deploying an Egress Router HTTP Proxy Pod

In *HTTP proxy mode*, the egress router runs as an HTTP proxy on port **8080**. This only works for clients talking to HTTP or HTTPS-based services, but usually requires fewer changes to the client pods to get them to work. Programs can be told to use an HTTP proxy by setting an environment variable.

1. Create the pod using the following as an example:

```
apiVersion: v1
kind: Pod
metadata:
  name: egress-http-proxy
  labels:
    name: egress-http-proxy
  annotations:
    pod.network.openshift.io/assign-macvlan: "true" 1
spec:
  initContainers:
  - name: egress-router-setup
    image: registry.access.redhat.com/openshift3/ose-egress-router
    securityContext:
      privileged: true
    env:
      - name: EGRESS_SOURCE 2
        value: 192.168.12.99
      - name: EGRESS_GATEWAY 3
        value: 192.168.12.1
      - name: EGRESS_ROUTER_MODE 4
        value: http-proxy
  containers:
  - name: egress-router-proxy
    image: registry.access.redhat.com/openshift3/ose-egress-router-http-proxy
    env:
      - name: EGRESS_HTTP_PROXY_DESTINATION 5
        value: |
          !*.example.com
          !192.168.1.0/24
          *
```

- 1 The **pod.network.openshift.io/assign-macvlan** annotation creates a Macvlan network interface on the primary network interface, then moves it into the pod's network name space before starting the **egress-router** container. Preserve the quotation marks around **"true"**. Omitting them results in errors.
- 2 An IP address from the physical network that the node itself is on and is reserved by the cluster administrator for use by this pod.
- 3 Same value as the default gateway used by the node itself.
- 4 This tells the egress router image that it is being deployed as part of an HTTP proxy, and so it should not set up iptables redirecting rules.
- 5 A string or YAML multi-line string specifying how to configure the proxy. Note that this is specified as an environment variable in the HTTP proxy container, not with the other environment variables in the init container.

You can specify any of the following for the **EGRESS\_HTTP\_PROXY\_DESTINATION** value. You can also use **\***, meaning "allow connections to all remote destinations". Each line in the configuration specifies one group of connections to allow or deny:

- An IP address (eg, **192.168.1.1**) allows connections to that IP address.
- A CIDR range (eg, **192.168.1.0/24**) allows connections to that CIDR range.
- A host name (eg, **www.example.com**) allows proxying to that host.
- A domain name preceded by **\***. (eg, **\*.example.com**) allows proxying to that domain and all of its subdomains.
- A **!** followed by any of the above denies connections rather than allowing them
- If the last line is **\***, then anything that hasn't been denied will be allowed. Otherwise, anything that hasn't been allowed will be denied.

2. Ensure other pods can find the pod's IP address by creating a service to point to the egress router:

```
apiVersion: v1
kind: Service
metadata:
  name: egress-1
spec:
  ports:
    - name: http-proxy
      port: 8080 1
  type: ClusterIP
  selector:
    name: egress-1
```

- 1 Ensure the **http** port is always set to **8080**.

3. Configure the client pod (not the egress proxy pod) to use the HTTP proxy by setting the **http\_proxy** or **https\_proxy** variables:

```
...
env:
  - name: http_proxy
    value: http://egress-1:8080/ 1
  - name: https_proxy
    value: http://egress-1:8080/
...
```

- 1 The service created in step 2.



## NOTE

Using the **http\_proxy** and **https\_proxy** environment variables is not necessary for all setups. If the above does not create a working setup, then consult the documentation for the tool or software you are running in the pod.

You can also specify the **EGRESS\_HTTP\_PROXY\_DESTINATION** using a ConfigMap, similarly to [the redirecting egress router example above](#).

### 6.5.2.5. Enabling Failover for Egress Router Pods

Using a replication controller, you can ensure that there is always one copy of the egress router pod in order to prevent downtime.

1. Create a replication controller configuration file using the following:

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: egress-demo-controller
spec:
  replicas: 1 1
  selector:
    name: egress-demo
  template:
    metadata:
      name: egress-demo
      labels:
        name: egress-demo
      annotations:
        pod.network.openshift.io/assign-macvlan: "true"
    spec:
      initContainers:
        - name: egress-demo-init
          image: registry.access.redhat.com/openshift3/ose-egress-
router
      env:
        - name: EGRESS_SOURCE
          value: 192.168.12.99
        - name: EGRESS_GATEWAY
          value: 192.168.12.1
        - name: EGRESS_DESTINATION
          value: 203.0.113.25
        - name: EGRESS_ROUTER_MODE
          value: init
      securityContext:
        privileged: true
      containers:
        - name: egress-demo-wait
          image: registry.access.redhat.com/openshift3/ose-pod
      nodeSelector:
        site: springfield-1
```

- 1 Ensure **replicas** is set to **1**, because only one pod can be using a given **EGRESS\_SOURCE** value at any time. This means that only a single copy of the router will be running, on a node with the label **site=springfield-1**.

2. Create the pod using the definition:

```
$ oc create -f <replication_controller>.json
```

3. To verify, check to see if the replication controller pod has been created:

```
oc describe rc <replication_controller>
```

### 6.5.3. Using iptables Rules to Limit Access to External Resources

Some cluster administrators may want to perform actions on outgoing traffic that do not fit within the model of **EgressNetworkPolicy** or the egress router. In some cases, this can be done by creating iptables rules directly.

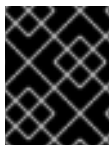
For example, you could create rules that log traffic to particular destinations, or to prevent more than a certain number of outgoing connections per second.

OpenShift Container Platform does not provide a way to add custom iptables rules automatically, but it does provide a place where such rules can be added manually by the administrator. Each node, on startup, will create an empty chain called **OPENSIFT-ADMIN-OUTPUT-RULES** in the **filter** table (assuming that the chain does not already exist). Any rules added to that chain by an administrator will be applied to all traffic going from a pod to a destination outside the cluster (and not to any other traffic).

There are a few things to watch out for when using this functionality:

1. It is up to you to ensure that rules get created on each node; OpenShift Container Platform does not provide any way to make that happen automatically.
2. The rules are not applied to traffic that exits the cluster via an egress router, and they run after **EgressNetworkPolicy** rules are applied (and so will not see traffic that is denied by an **EgressNetworkPolicy**).
3. The handling of connections from pods to nodes or pods to the master is complicated, because nodes have both "external" IP addresses and "internal" SDN IP addresses. Thus, some pod-to-node/master traffic may pass through this chain, but other pod-to-node/master traffic may bypass it.

## 6.6. ENABLING MULTICAST



### IMPORTANT

At this time, multicast is best used for low bandwidth coordination or service discovery and not a high-bandwidth solution.

Multicast traffic between OpenShift Container Platform pods is disabled by default. If you are using the **ovs-multitenant** or **ovs-networkpolicy** plugin, you can enable multicast on a per-project basis by setting an annotation on the project's corresponding **netnamespace** object:

```
# oc annotate netnamespace <namespace> \
    netnamespace.network.openshift.io/multicast-enabled=true
```

Disable multicast by removing the annotation:

```
# oc annotate netnamespace <namespace> \
    netnamespace.network.openshift.io/multicast-enabled-
```

When using the **ovs-multitenant** plugin:

1. In an isolated project, multicast packets sent by a pod will be delivered to all other pods in the project.
2. If you have [joined networks together](#), you will need to enable multicast in each project's **netnamespace** in order for it to take effect in any of the projects. Multicast packets sent by a pod in a joined network will be delivered to all pods in all of the joined-together networks.
3. To enable multicast in the **default** project, you must also enable it in all projects that have been [made global](#). Global projects are not "global" for purposes of multicast; multicast packets sent by a pod in a global project will only be delivered to pods in other global projects, not to all pods in all projects. Likewise, pods in global projects will only receive multicast packets sent from pods in other global projects, not from all pods in all projects.

When using the **ovs-networkpolicy** plugin:

1. Multicast packets sent by a pod will be delivered to all other pods in the project, regardless of **NetworkPolicy** objects. (Pods may be able to communicate over multicast even when they can't communicate over unicast.)
2. Multicast packets sent by a pod in one project will never be delivered to pods in any other project, even if there are **NetworkPolicy** objects allowing communication between the to projects.

## 6.7. ENABLING NETWORKPOLICY



### IMPORTANT

Enabling the Kubernetes **NetworkPolicy** 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/>.

Kubernetes **NetworkPolicy** is not currently fully supported by OpenShift Container Platform, and the **ovs-subnet** and **ovs-multitenant** plug-ins ignore **NetworkPolicy** objects. However, a Technology Preview of **NetworkPolicy** support is available by using the **ovs-networkpolicy** plug-in.

In a cluster [configured to use the ovs-networkpolicy plug-in](#), network isolation is controlled entirely by **NetworkPolicy** objects. By default, all pods in a project are accessible from other pods and network endpoints. To isolate one or more pods in a project, you can create **NetworkPolicy** objects in that project to indicate the allowed incoming connections. Project administrators can create and delete **NetworkPolicy** objects within their own project.

Pods that do not have **NetworkPolicy** objects pointing to them are fully accessible, whereas, pods that have one or more **NetworkPolicy** objects pointing to them are isolated. These isolated pods only accept connections that are accepted by at least one of their **NetworkPolicy** objects.

Following are a few sample **NetworkPolicy** object definitions supporting different scenarios:

- **Deny All Traffic**



To make a project "deny by default" add a **NetworkPolicy** object that matches all pods but accepts no traffic.

```
networkConfig:
  ...
  networkPluginName: "redhat/openshift-ovs-networkpolicy" 1
  ...
```

- 1 Set to **redhat/openshift-ovs-networkpolicy** for the **ovs-networkpolicy** plug-in

- **Only Accept connections from pods within project**

To make pods accept connections from other pods in the same project, but reject all other connections from pods in other projects:

```
networkConfig:
  ...
  networkPluginName: "redhat/openshift-ovs-networkpolicy" 1
```

- 1 Set to **redhat/openshift-ovs-networkpolicy** for the **ovs-networkpolicy** plug-in

- **Only allow HTTP and HTTPS traffic based on pod labels**

To enable only HTTP and HTTPS access to the pods with a specific label (**role=frontend** in following example), add a **NetworkPolicy** object similar to:

```
kind: NetworkPolicy
apiVersion: extensions/v1beta1
metadata:
  name: allow-http-and-https
spec:
  podSelector:
    matchLabels:
      role: frontend
  ingress:
    - ports:
      - protocol: TCP
        port: 80
      - protocol: TCP
        port: 443
```

**NetworkPolicy** objects are additive, which means you can combine multiple **NetworkPolicy** objects together to satisfy complex network requirements.

For example, for the **NetworkPolicy** objects defined in previous samples, you can define both **allow-same-namespace** and **allow-http-and-https** policies within the same project. Thus allowing the pods with the label **role=frontend**, to accept any connection allowed by each policy. That is, connections on any port from pods in the **same** namespace, and connections on ports **80** and **443** from pods in **any** namespace.

### 6.7.1. NetworkPolicy and Routers

When using the **ovs-multitenant** plug-in, router traffic is automatically allowed into all namespaces, because the routers are normally in the default namespace, and all namespaces allow connections from

pods in that namespace. This does not happen automatically when using the Networkpolicy plug-in, so if you have a policy that isolates a namespace by default, you will need to take additional steps to allow routers access.

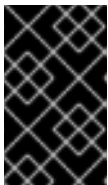
Create a policy for each service, allowing access from all sources:

```
kind: NetworkPolicy
apiVersion: extensions/v1beta1
metadata:
  name: allow-to-database-service
spec:
  podSelector:
    matchLabels:
      role: database
  ingress:
  - ports:
    - protocol: TCP
      port: 5432
```

This allows routers to access the service. However, this also allows pods in other users' namespaces to access it as well. In general, this should not be a problem, because those pods could normally access the service via the public router anyway.

Alternatively, you can create a policy allowing full access from the default namespace, as in the **ovs-multitenant** plug-in:

1. First, as a cluster administrator, add a label to the default namespace so it can be matched:



### IMPORTANT

If you labeled the default project with the **default** label in a previous procedure, then skip this step. The cluster administrator role is required to add labels to namespaces.

```
$ oc label namespace default name=default
```

2. Create policies allowing connections from that namespace.



### NOTE

Perform this step for each namespace you want to allow connections into. Users with the Project Administrator role can create policies.

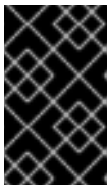
```
kind: NetworkPolicy
apiVersion: extensions/v1beta1
metadata:
  name: allow-from-default-namespace
spec:
  podSelector:
    ingress:
    - from:
```

```
- namespaceSelector:
  matchLabels:
    name: default
```

### 6.7.2. Setting a Default NetworkPolicy for New Projects

Cluster administrators can modify the default project template to enable automatic creation of default **NetworkPolicy** objects (one or more), whenever a new project is created. To do this:

1. Create a custom project template and configure the master to use it, as described in [Modifying the Template for New Projects](#).
2. Label the **default** project with the **default** label:



#### IMPORTANT

If you labeled the default project with the **default** label in a previous procedure, then skip this step. The cluster administrator role is required to add labels to namespaces.

```
$ oc label namespace default name=default
```

3. Edit the template to include the desired **NetworkPolicy** objects:

```
$ oc edit template project-request -n default
```



#### NOTE

To include **NetworkPolicy** objects into existing template, use the **oc edit** command. Currently, it is not possible to use **oc patch** to add objects to a **Template** resource.

- a. Add each default policy as an element in the **objects** array:

```
objects:
...
- apiVersion: extensions/v1beta1
  kind: NetworkPolicy
  metadata:
    name: allow-same-namespace
  spec:
    podSelector:
      ingress:
        - from:
            - podSelector: {}
...

```

## 6.8. TROUBLESHOOTING THROUGHPUT ISSUES

Sometimes applications deployed through OpenShift Container Platform can cause network throughput issues such as unusually high latency between specific services.

Use the following methods to analyze performance issues if pod logs do not reveal any cause of the problem:

- Use a packet analyzer, such as ping or [tcpdump](#) to analyze traffic between a pod and its node. For example, run the tcpdump tool on each pod while reproducing the behavior that led to the issue. Review the captures on both sides to compare send and receive timestamps to analyze the latency of traffic to/from a pod. Latency can occur in OpenShift Container Platform if a node interface is overloaded with traffic from other pods, storage devices, or the data plane.

```
$ tcpdump -s 0 -i any -w /tmp/dump.pcap host <podip 1> && host  
<podip 2> 1
```

- 1** **podip** is the IP address for the pod. Run the following command to get the IP address of the pods:

```
# oc get pod <podname> -o wide
```

tcpdump generates a file at **/tmp/dump.pcap** containing all traffic between these two pods. Ideally, run the analyzer shortly before the issue is reproduced and stop the analyzer shortly after the issue is finished reproducing to minimize the size of the file. You can also run a packet analyzer between the nodes (eliminating the SDN from the equation) with:

```
# tcpdump -s 0 -i any -w /tmp/dump.pcap port 4789
```

- Use a bandwidth measuring tool, such as iperf, to measure streaming throughput and UDP throughput. Run the tool from the pods first, then from the nodes to attempt to locate any bottlenecks. The iperf3 tool is included as part of RHEL 7.

For information on installing and using iperf3, see this [Red Hat Solution](#).

## CHAPTER 7. CONFIGURING SERVICE ACCOUNTS

### 7.1. OVERVIEW

When a person uses the OpenShift Container Platform CLI or web console, their API token authenticates them to the OpenShift Container Platform API. However, when a regular user's credentials are not available, it is common for components to make API calls independently. For example:

- Replication controllers make API calls to create or delete pods.
- Applications inside containers can make API calls for discovery purposes.
- External applications can make API calls for monitoring or integration purposes.

Service accounts provide a flexible way to control API access without sharing a regular user's credentials.

### 7.2. USER NAMES AND GROUPS

Every service account has an associated user name that can be granted roles, just like a regular user. The user name is derived from its project and name:

```
system:serviceaccount:<project>:<name>
```

For example, to add the **view** role to the **robot** service account in the **top-secret** project:

```
$ oc policy add-role-to-user view system:serviceaccount:top-secret:robot
```

#### IMPORTANT

If you want to grant access to a specific service account in a project, you can use the **-z** flag. From the project to which the service account belongs, use the **-z** flag and specify the **<serviceaccount\_name>**. This is highly recommended, as it helps prevent typos and ensures that access is granted only to the specified service account. For example:

```
$ oc policy add-role-to-user <role_name> -z
<serviceaccount_name>
```

If not in the project, use the **-n** option to indicate the project namespace it applies to, as shown in the examples below.

Every service account is also a member of two groups:

#### **system:serviceaccount**

Includes all service accounts in the system.

#### **system:serviceaccount:<project>**

Includes all service accounts in the specified project.

For example, to allow all service accounts in all projects to view resources in the **top-secret** project:

```
$ oc policy add-role-to-group view system:serviceaccount -n top-secret
```

To allow all service accounts in the **managers** project to edit resources in the **top-secret** project:

```
$ oc policy add-role-to-group edit system:serviceaccount:managers -n top-secret
```

## 7.3. MANAGING SERVICE ACCOUNTS

Service accounts are API objects that exist within each project. To manage service accounts, you can use the **oc** command with the **sa** or **serviceaccount** object type or use the web console.

To get a list of existing service accounts in the current project:

```
$ oc get sa
NAME          SECRETS  AGE
builder       2         2d
default       2         2d
deployer      2         2d
```

To create a new service account:

```
$ oc create sa robot
serviceaccount "robot" created
```

As soon as a service account is created, two secrets are automatically added to it:

- an API token
- credentials for the OpenShift Container Registry

These can be seen by describing the service account:

```
$ oc describe sa robot
Name: robot
Namespace: project1
Labels: <none>
Annotations: <none>

Image pull secrets: robot-dockercfg-qzbhb

Mountable secrets: robot-token-f4khf
                  robot-dockercfg-qzbhb

Tokens:
    robot-token-f4khf
    robot-token-z8h44
```

The system ensures that service accounts always have an API token and registry credentials.

The generated API token and registry credentials do not expire, but they can be revoked by deleting the secret. When the secret is deleted, a new one is automatically generated to take its place.

## 7.4. ENABLING SERVICE ACCOUNT AUTHENTICATION

Service accounts authenticate to the API using tokens signed by a private RSA key. The authentication layer verifies the signature using a matching public RSA key.

To enable service account token generation, update the **serviceAccountConfig** stanza in the */etc/origin/master/master-config.yml* file on the master to specify a **privateKeyFile** (for signing), and a matching public key file in the **publicKeyFiles** list:

```
serviceAccountConfig:
  ...
  masterCA: ca.crt ❶
  privateKeyFile: serviceaccount.private.key ❷
  publicKeyFiles:
  - serviceaccount.public.key ❸
  - ...
```

- ❶ CA file used to validate the API server's serving certificate.
- ❷ Private RSA key file (for token signing).
- ❸ Public RSA key files (for token verification). If private key files are provided, then the public key component is used. Multiple public key files can be specified, and a token will be accepted if it can be validated by one of the public keys. This allows rotation of the signing key, while still accepting tokens generated by the previous signer.

## 7.5. MANAGED SERVICE ACCOUNTS

Service accounts are required in each project to run builds, deployments, and other pods. The **managedNames** setting in the */etc/origin/master/master-config.yml* file on the master controls which service accounts are automatically created in every project:

```
serviceAccountConfig:
  ...
  managedNames: ❶
  - builder ❷
  - deployer ❸
  - default ❹
  - ...
```

- ❶ List of service accounts to automatically create in every project.
- ❷ A **builder** service account in each project is required by build pods, and is given the **system:image-builder** role, which allows pushing images to any image stream in the project using the internal container registry.
- ❸ A **deployer** service account in each project is required by deployment pods, and is given the **system:deployer** role, which allows viewing and modifying replication controllers and pods in the project.
- ❹ A **default** service account is used by all other pods unless they specify a different service account.

All service accounts in a project are given the **system:image-puller** role, which allows pulling images from any image stream in the project using the internal container registry.

## 7.6. INFRASTRUCTURE SERVICE ACCOUNTS

Several infrastructure controllers run using service account credentials. The following service accounts are created in the OpenShift Container Platform infrastructure project (**openshift-infra**) at server start, and given the following roles cluster-wide:

| Service Account               | Description   |
|-------------------------------|---|
| <b>replication-controller</b> | Assigned the <b>system:replication-controller</b> role  |
| <b>deployment-controller</b>  | Assigned the <b>system:deployment-controller</b> role   |
| <b>build-controller</b>       | Assigned the <b>system:build-controller</b> role. Additionally, the <b>build-controller</b> service account is included in the privileged security context constraint in order to create privileged build pods. |

To configure the project where those service accounts are created, set the **openshiftInfrastructureNamespace** field in the */etc/origin/master/master-config.yml* file on the master:

```
policyConfig:
  ...
  openshiftInfrastructureNamespace: openshift-infra
```

## 7.7. SERVICE ACCOUNTS AND SECRETS

Set the **limitSecretReferences** field in the */etc/origin/master/master-config.yml* file on the master to **true** to require pod secret references to be whitelisted by their service accounts. Set its value to **false** to allow pods to reference any secret in the project.

```
serviceAccountConfig:
  ...
  limitSecretReferences: false
```



## CHAPTER 8. MANAGING AUTHORIZATION POLICIES

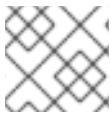
### 8.1. OVERVIEW

You can use [the CLI](#) to view [authorization policies](#) and the administrator CLI to manage the [roles and bindings](#) within a policy.

### 8.2. VIEWING ROLES AND BINDINGS

[Roles](#) grant various levels of access in the system-wide [cluster policy](#) as well as project-scoped [local policies](#). [Users and groups](#) can be associated with, or *bound* to, multiple roles at the same time. You can view details about the roles and their bindings using the **oc describe** command.

Users with the **cluster-admin** [default role](#) in the cluster policy can view cluster policy and all local policies. Users with the **admin** [default role](#) in a given local policy can view that project-scoped policy.



#### NOTE

Review a full list of verbs in the [Evaluating Authorization](#) section.

#### 8.2.1. Viewing Cluster Policy

To view the cluster roles and their associated rule sets in the cluster policy:

```
$ oc describe clusterPolicy default
```

#### Example 8.1. Viewing Cluster Roles

```
$ oc describe clusterPolicy default
Name:      default
Created:    5 days ago
Labels:     <none>
Annotations: <none>
Last Modified: 2016-03-17 13:25:27 -0400 EDT
admin      Verbs      Non-Resource URLs  Extension  Resource Names
API Groups  Resources
[create delete deletecollection get list patch update watch] []
[] [] [configmaps endpoints persistentvolumeclaims pods pods/attach
pods/exec pods/log pods/portforward pods/proxy replicationcontrollers
replicationcontrollers/scale secrets serviceaccounts services
services/proxy]
[create delete deletecollection get list patch update watch] []
[] [] [buildconfigs buildconfigs/instantiate
buildconfigs/instantiatebinary buildconfigs/webhooks buildlogs builds
builds/clone builds/custom builds/docker builds/log builds/source
deploymentconfigrollbacks deploymentconfigs deploymentconfigs/log
deploymentconfigs/scale deployments generateddeploymentconfigs
imagestreamimages imagestreamimports imagestreammappings imagestreams
imagestreams/secrets imagestreamtags localresourceaccessreviews
localsubjectaccessreviews processedtemplates projects
resourceaccessreviews rolebindings roles routes subjectaccessreviews
templateconfigs templates]
[create delete deletecollection get list patch update watch] []
```

```

[] [autoscaling] [horizontalpodautoscalers]
[create delete deletecollection get list patch update watch] []
[] [batch] [jobs]
[create delete deletecollection get list patch update watch] []
[] [extensions] [daemonsets horizontalpodautoscalers jobs
replicationcontrollers/scale]
[get list watch] [] [] [] [bindings configmaps
endpoints events imagestreams/status limitranges minions namespaces
namespaces/status nodes persistentvolumeclaims persistentvolumes pods
pods/log pods/status policies policybindings replicationcontrollers
replicationcontrollers/status resourcequotas resourcequotas/status
resourcequotausages routes/status securitycontextconstraints
serviceaccounts services]
[get update] [] [] [] [imagestreams/layers]
[update] [] [] [] [routes/status]
basic-user Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[get] [] [~] [] [users]
[list] [] [] [] [projectrequests]
[get list] [] [] [] [clusterroles]
[list] [] [] [] [projects]
[create] [] IsPersonalSubjectAccessReview [] []
[localsubjectaccessreviews subjectaccessreviews]
cluster-admin Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[*] [] [] [*] [*]
[*] [*] [] [] []
cluster-reader Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[get list watch] [] [] [] [bindings buildconfigs
buildconfigs/instantiate buildconfigs/instantiatebinary
buildconfigs/webhooks buildlogs builds builds/clone builds/details
builds/log clusternetworks clusterpolicies clusterpolicybindings
clusterrolebindings clusterroles configmaps deploymentconfigrollbacks
deploymentconfigs deploymentconfigs/log deploymentconfigs/scale
deployments endpoints events generateddeploymentconfigs groups
hostsubnets identities images imagestreamimages imagestreamimports
imagestreammappings imagestreams imagestreams/status imagestreamtags
limitranges localresourceaccessreviews localsubjectaccessreviews minions
namespaces netnamespaces nodes oauthclientauthorizations oauthclients
persistentvolumeclaims persistentvolumes pods pods/log policies
policybindings processedtemplates projectrequests projects
replicationcontrollers resourceaccessreviews resourcequotas
resourcequotausages rolebindings roles routes routes/status
securitycontextconstraints serviceaccounts services subjectaccessreviews
templateconfigs templates useridentitymappings users]
[get list watch] [] [] [autoscaling]
[horizontalpodautoscalers]
[get list watch] [] [] [batch] [jobs]
[get list watch] [] [] [extensions] [daemonsets
horizontalpodautoscalers jobs replicationcontrollers/scale]
[create] [] [] [] [resourceaccessreviews
subjectaccessreviews]
[get] [] [] [] [nodes/metrics]
[create get] [] [] [] [nodes/stats]
[get] [*] [] [] []

```

```

cluster-status      Verbs      Non-Resource URLs  Extension  Resource
Names API Groups  Resources
[get]               [/api /api/* /apis /apis/* /healthz /healthz/* /oapi
/oapi/* /osapi /osapi/ /version] [] [] []
edit      Verbs      Non-Resource URLs  Extension  Resource Names
API Groups  Resources
[create delete deletecollection get list patch update watch] []
[] [] [configmaps endpoints persistentvolumeclaims pods pods/attach
pods/exec pods/log pods/portforward pods/proxy replicationcontrollers
replicationcontrollers/scale secrets serviceaccounts services
services/proxy]
[create delete deletecollection get list patch update watch] []
[] [] [buildconfigs buildconfigs/instantiate
buildconfigs/instantiatebinary buildconfigs/webhooks buildlogs builds
builds/clone builds/custom builds/docker builds/log builds/source
deploymentconfigrollbacks deploymentconfigs deploymentconfigs/log
deploymentconfigs/scale deployments generateddeploymentconfigs
imagestreamimages imagestreamimports imagestreammappings imagestreams
imagestreams/secrets imagestreamtags processedtemplates routes
templateconfigs templates]
[create delete deletecollection get list patch update watch] []
[] [autoscaling] [horizontalpodautoscalers]
[create delete deletecollection get list patch update watch] []
[] [batch] [jobs]
[create delete deletecollection get list patch update watch] []
[] [extensions] [daemonsets horizontalpodautoscalers jobs
replicationcontrollers/scale]
[get list watch] [] [] [] [bindings configmaps
endpoints events imagestreams/status limitranges minions namespaces
namespaces/status nodes persistentvolumeclaims persistentvolumes pods
pods/log pods/status projects replicationcontrollers
replicationcontrollers/status resourcequotas resourcequotas/status
resourcequotausages routes/status securitycontextconstraints
serviceaccounts services]
[get update] [] [] [] [imagestreams/layers]
registry-admin      Verbs      Non-Resource URLs  Extension  Resource
Names API Groups  Resources
[create delete deletecollection get list patch update watch] []
[] [] [imagestreamimages imagestreamimports imagestreammappings
imagestreams imagestreams/secrets imagestreamtags]
[create delete deletecollection get list patch update watch] []
[] [] [localresourceaccessreviews localsubjectaccessreviews
resourceaccessreviews rolebindings roles subjectaccessreviews]
[get update] [] [] [] [imagestreams/layers]
[get list watch] [] [] [] [policies policybindings]
[get] [] [] [] [namespaces projects]
registry-editor      Verbs      Non-Resource URLs  Extension  Resource
Names API Groups  Resources
[get] [] [] [] [namespaces projects]
[create delete deletecollection get list patch update watch] []
[] [] [imagestreamimages imagestreamimports imagestreammappings
imagestreams imagestreams/secrets imagestreamtags]
[get update] [] [] [] [imagestreams/layers]
registry-viewer      Verbs      Non-Resource URLs  Extension  Resource
Names API Groups  Resources
[get list watch] [] [] [] [imagestreamimages

```

```

imagestreamimports imagestreammappings imagestreams imagestreamtags]
[get] [] [] [] [imagestreams/layers namespaces
projects]
self-provisioner Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[create] [] [] [] [projectrequests]
system:build-controller Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[get list watch] [] [] [] [builds]
[update] [] [] [] [builds]
[create] [] [] [] [builds/custom builds/docker
builds/source]
[get] [] [] [] [imagestreams]
[create delete get list] [] [] [] [pods]
[create patch update] [] [] [] [events]
system:daemonset-controller Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[list watch] [] [] [extensions] [daemonsets]
[list watch] [] [] [] [pods]
[list watch] [] [] [] [nodes]
[update] [] [] [extensions] [daemonsets/status]
[create delete] [] [] [] [pods]
[create] [] [] [] [pods/binding]
[create patch update] [] [] [] [events]
system:deployer Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[get list] [] [] [] [replicationcontrollers]
[get update] [] [] [] [replicationcontrollers]
[create get list watch] [] [] [] [pods]
[get] [] [] [] [pods/log]
[update] [] [] [] [imagestreamtags]
system:deployment-controller Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[list watch] [] [] [] [replicationcontrollers]
[get update] [] [] [] [replicationcontrollers]
[create delete get list update] [] [] [] [pods]
[create patch update] [] [] [] [events]
system:discovery Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[get] [/api /api/* /apis /apis/* /oapi /oapi/* /osapi
/osapi/ /version] [] [] []
system:hpa-controller Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[get list watch] [] [] [extensions autoscaling]
[horizontalpodautoscalers]
[update] [] [] [extensions autoscaling]
[horizontalpodautoscalers/status]
[get update] [] [] [extensions ]
[replicationcontrollers/scale]
[get update] [] [] [] [deploymentconfigs/scale]
[create patch update] [] [] [] [events]
[list] [] [] [] [pods]
[proxy] [] [https:heapster:] [] [services]
system:image-builder Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[get update] [] [] [] [imagestreams/layers]

```

```

    [update]          [] [] [] [builds/details]
system:image-pruner  Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [delete]          [] [] [] [images]
    [get list]        [] [] [] [buildconfigs builds
deploymentconfigs images imagestreams pods replicationcontrollers]
    [update]          [] [] [] [imagestreams/status]
system:image-puller  Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [get]             [] [] [] [imagestreams/layers]
system:image-pusher  Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [get update]      [] [] [] [imagestreams/layers]
system:job-controller Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [list watch]      [] [] [extensions batch] [jobs]
    [update]           [] [] [extensions batch] [jobs/status]
    [list watch]       [] [] [] [pods]
    [create delete]    [] [] [] [pods]
    [create patch update] [] [] [] [events]
system:master        Verbs          Non-Resource URLs  Extension  Resource
Names  API Groups  Resources
    [*]              [] [] [*] [*]
system:namespace-controller Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [delete get list watch] [] [] [] [namespaces]
    [update]         [] [] [] [namespaces/finalize
namespaces/status]
    [delete deletecollection get list] [] [] [*] [*]
system:node          Verbs          Non-Resource URLs  Extension  Resource
Names  API Groups  Resources
    [create]          [] [] [] [localsubjectaccessreviews
subjectaccessreviews]
    [get list watch]   [] [] [] [services]
    [create get list watch] [] [] [] [nodes]
    [update]           [] [] [] [nodes/status]
    [create patch update] [] [] [] [events]
    [get list watch]   [] [] [] [pods]
    [create delete get] [] [] [] [pods]
    [update]           [] [] [] [pods/status]
    [get]              [] [] [] [configmaps secrets]
    [get]              [] [] [] [persistentvolumeclaims
persistentvolumes]
    [get]              [] [] [] [endpoints]
system:node-admin     Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [get list watch]   [] [] [] [nodes]
    [proxy]            [] [] [] [nodes]
    [*]               [] [] [] [nodes/log nodes/metrics nodes/proxy
nodes/stats]
system:node-proxier    Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [list watch]       [] [] [] [endpoints services]
system:node-reader     Verbs          Non-Resource URLs  Extension
Resource Names  API Groups  Resources
    [get list watch]   [] [] [] [nodes]

```

```

[get] [] [] [] [nodes/metrics]
[create get] [] [] [] [nodes/stats]
system:oauth-token-deleter Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[delete] [] [] [] [oauthaccesstokens]
oauthauthorizetokens]
system:pv-binder-controller Verbs Non-Resource URLs Extension
Resource Names API Groups Resources
[list watch] [] [] [] [persistentvolumes]
[create delete get update] [] [] []
[persistentvolumes]
[update] [] [] [] [persistentvolumes/status]
[list watch] [] [] [] [persistentvolumeclaims]
[get update] [] [] [] [persistentvolumeclaims]
[update] [] [] [] [persistentvolumeclaims/status]
system:pv-provisioner-controller Verbs Non-Resource URLs
Extension Resource Names API Groups Resources
[list watch] [] [] [] [persistentvolumes]
[create delete get update] [] [] []
[persistentvolumes]
[update] [] [] [] [persistentvolumes/status]
[list watch] [] [] [] [persistentvolumeclaims]
[get update] [] [] [] [persistentvolumeclaims]
[update] [] [] [] [persistentvolumeclaims/status]
system:pv-recycler-controller Verbs Non-Resource URLs
Extension Resource Names API Groups Resources
[list watch] [] [] [] [persistentvolumes]
[create delete get update] [] [] []
[persistentvolumes]
[update] [] [] [] [persistentvolumes/status]
[list watch] [] [] [] [persistentvolumeclaims]
[get update] [] [] [] [persistentvolumeclaims]
[update] [] [] [] [persistentvolumeclaims/status]
[list watch] [] [] [] [pods]
[create delete get] [] [] [] [pods]
[create patch update] [] [] [] [events]
system:registry Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[delete get] [] [] [] [images]
[get] [] [] [] [imagestreamimages imagestreams]
imagestreams/secrets imagestreamtags]
[update] [] [] [] [imagestreams]
[create] [] [] [] [imagestreammappings]
[list] [] [] [] [resourcequotas]
system:replication-controller Verbs Non-Resource URLs
Extension Resource Names API Groups Resources
[list watch] [] [] [] [replicationcontrollers]
[get update] [] [] [] [replicationcontrollers]
[update] [] [] [] [replicationcontrollers/status]
[list watch] [] [] [] [pods]
[create delete] [] [] [] [pods]
[create patch update] [] [] [] [events]
system:router Verbs Non-Resource URLs Extension Resource
Names API Groups Resources
[list watch] [] [] [] [endpoints routes]
[update] [] [] [] [routes/status]

```

```

system:sdn-manager  Verbs      Non-Resource URLs  Extension
Resource Names  API Groups  Resources
[create delete get list watch]  []  []  []  [hostsubnets]
[create delete get list watch]  []  []  []
[netnamespaces]
[get list watch]  []  []  []  [nodes]
[create get]  []  []  []  [clusternetworks]
system:sdn-reader  Verbs      Non-Resource URLs  Extension
Resource Names  API Groups  Resources
[get list watch]  []  []  []  [hostsubnets]
[get list watch]  []  []  []  [netnamespaces]
[get list watch]  []  []  []  [nodes]
[get]  []  []  []  [clusternetworks]
[get list watch]  []  []  []  [namespaces]
system:webhook  Verbs      Non-Resource URLs  Extension  Resource
Names  API Groups  Resources
[create get]  []  []  []  [buildconfigs/webhooks]
view  Verbs      Non-Resource URLs  Extension  Resource Names
API Groups  Resources
[get list watch]  []  []  []  [bindings buildconfigs
buildconfigs/instantiate buildconfigs/instantiatebinary
buildconfigs/webhooks buildlogs builds builds/clone builds/log
configmaps deploymentconfigrollbacks deploymentconfigs
deploymentconfigs/log deploymentconfigs/scale deployments endpoints
events generatedeploymentconfigs imagestreamimages imagestreamimports
imagestreammappings imagestreams imagestreams/status imagestreamtags
limitranges minions namespaces namespaces/status nodes
persistentvolumeclaims persistentvolumes pods pods/log pods/status
processedtemplates projects replicationcontrollers
replicationcontrollers/status resourcequotas resourcequotas/status
resourcequotausages routes routes/status securitycontextconstraints
serviceaccounts services templateconfigs templates]
[get list watch]  []  []  [autoscaling]
[horizontalpodautoscalers]
[get list watch]  []  []  [batch]  [jobs]
[get list watch]  []  []  [extensions]  [daemonsets]
horizontalpodautoscalers jobs]

```

To view the current set of cluster bindings, which shows the users and groups that are bound to various roles:

```
$ oc describe clusterPolicyBindings :default
```

### Example 8.2. Viewing Cluster Bindings

```

$ oc describe clusterPolicyBindings :default
Name:      :default
Created:    4 hours ago
Labels:     <none>
Last Modified: 2015-06-10 17:22:26 +0000 UTC
Policy:     <none>
RoleBinding[basic-users]:
  Role: basic-user
  Users: []

```

```
    Groups: [system:authenticated]
RoleBinding[cluster-admins]:
  Role: cluster-admin
  Users: []
  Groups: [system:cluster-admins]
RoleBinding[cluster-readers]:
  Role: cluster-reader
  Users: []
  Groups: [system:cluster-readers]
RoleBinding[cluster-status-binding]:
  Role: cluster-status
  Users: []
  Groups: [system:authenticated system:unauthenticated]
RoleBinding[self-provisioners]:
  Role: self-provisioner
  Users: []
  Groups: [system:authenticated]
RoleBinding[system:build-controller]:
  Role: system:build-controller
  Users: [system:serviceaccount:openshift-infra:build-controller]
  Groups: []
RoleBinding[system:deployment-controller]:
  Role: system:deployment-controller
  Users: [system:serviceaccount:openshift-infra:deployment-
controller]
  Groups: []
RoleBinding[system:masters]:
  Role: system:master
  Users: []
  Groups: [system:masters]
RoleBinding[system:node-proxiers]:
  Role: system:node-proxier
  Users: []
  Groups: [system:nodes]
RoleBinding[system:nodes]:
  Role: system:node
  Users: []
  Groups: [system:nodes]
RoleBinding[system:oauth-token-deleters]:
  Role: system:oauth-token-deleter
  Users: []
  Groups: [system:authenticated system:unauthenticated]
RoleBinding[system:registries]:
  Role: system:registry
  Users: []
  Groups: [system:registries]
RoleBinding[system:replication-controller]:
  Role: system:replication-controller
  Users: [system:serviceaccount:openshift-infra:replication-
controller]
  Groups: []
RoleBinding[system:routers]:
  Role: system:router
  Users: []
  Groups: [system:routers]
RoleBinding[system:sdn-readers]:
```



```

Role: system:sdn-reader
Users: []
Groups: [system:nodes]
RoleBinding[system:webhooks]:
  Role: system:webhook
  Users: []
  Groups: [system:authenticated system:unauthenticated]

```

### 8.2.2. Viewing Local Policy

While the list of local roles and their associated rule sets are not viewable within a local policy, all of the [default roles](#) are still applicable and can be added to users or groups, other than the **cluster-admin** default role. The local bindings, however, are viewable.

To view the current set of local bindings, which shows the users and groups that are bound to various roles:

```
$ oc describe policyBindings :default
```

By default, the current project is used when viewing local policy. Alternatively, a project can be specified with the **-n** flag. This is useful for viewing the local policy of another project, if the user already has the [admin](#) default role in it.

#### Example 8.3. Viewing Local Bindings

```

$ oc describe policyBindings :default -n joe-project
Name:          :default
Created:       About a minute ago
Labels:        <none>
Last Modified: 2015-06-10 21:55:06 +0000 UTC
Policy:        <none>
RoleBinding[admins]:
  Role: admin
  Users: [joe]
  Groups: []
RoleBinding[system:deployers]:
  Role: system:deployer
  Users: [system:serviceaccount:joe-project:deployer]
  Groups: []
RoleBinding[system:image-builders]:
  Role: system:image-builder
  Users: [system:serviceaccount:joe-project:builder]
  Groups: []
RoleBinding[system:image-pullers]:
  Role: system:image-puller
  Users: []
  Groups: [system:serviceaccounts:joe-project]

```

By default in a local policy, only the binding for the **admin** role is immediately listed. However, if other [default roles](#) are added to users and groups within a local policy, they become listed as well.

## 8.3. MANAGING ROLE BINDINGS

Adding, or *binding*, a [role](#) to [users](#) or [groups](#) gives the user or group the relevant access granted by the role. You can add and remove roles to and from users and groups using **oc adm policy** commands.

When managing a user or group's associated roles for a local policy using the following operations, a project may be specified with the **-n** flag. If it is not specified, then the current project is used.

**Table 8.1. Local Policy Operations**

| Command   | Description   |
|---|---|
| <b>\$ oc adm policy who-can &lt;verb&gt; &lt;resource&gt;</b>                 | Indicates which users can perform an action on a resource.              |
| <b>\$ oc adm policy add-role-to-user &lt;role&gt; &lt;username&gt;</b>        | Binds a given role to specified users in the current project.           |
| <b>\$ oc adm policy remove-role-from-user &lt;role&gt; &lt;username&gt;</b>   | Removes a given role from specified users in the current project.       |
| <b>\$ oc adm policy remove-user &lt;username&gt;</b>                          | Removes specified users and all of their roles in the current project.  |
| <b>\$ oc adm policy add-role-to-group &lt;role&gt; &lt;groupname&gt;</b>      | Binds a given role to specified groups in the current project.          |
| <b>\$ oc adm policy remove-role-from-group &lt;role&gt; &lt;groupname&gt;</b> | Removes a given role from specified groups in the current project.      |
| <b>\$ oc adm policy remove-group &lt;groupname&gt;</b>                        | Removes specified groups and all of their roles in the current project. |

You can also manage role bindings for the cluster policy using the following operations. The **-n** flag is not used for these operations because the cluster policy uses non-namespaced resources.

**Table 8.2. Cluster Policy Operations**

| Command   | Description  |
|---|--|
| <b>\$ oc adm policy add-cluster-role-to-user &lt;role&gt; &lt;username&gt;</b>      | Binds a given role to specified users for all projects in the cluster.     |
| <b>\$ oc adm policy remove-cluster-role-from-user &lt;role&gt; &lt;username&gt;</b> | Removes a given role from specified users for all projects in the cluster. |
| <b>\$ oc adm policy add-cluster-role-to-group &lt;role&gt; &lt;groupname&gt;</b>    | Binds a given role to specified groups for all projects in the cluster.    |

| Command   | Description   |
|---|---|
| <b>\$ oc adm policy remove-cluster-role-from-group &lt;role&gt; &lt;groupname&gt;</b> | Removes a given role from specified groups for all projects in the cluster. |

For example, you can add the **admin** role to the **alice** user in **joe-project** by running:

```
$ oc adm policy add-role-to-user admin alice -n joe-project
```

You can then view the local bindings and verify the addition in the output:

```
$ oc describe policyBindings :default -n joe-project
Name:          :default
Created:       5 minutes ago
Labels:        <none>
Last Modified: 2015-06-10 22:00:44 +0000 UTC
Policy:        <none>
RoleBinding[admins]:
  Role: admin
  Users: [alice joe] 1
  Groups: []
RoleBinding[system:deployers]:
  Role: system:deployer
  Users: [system:serviceaccount:joe-project:deployer]
  Groups: []
RoleBinding[system:image-builders]:
  Role: system:image-builder
  Users: [system:serviceaccount:joe-project:builder]
  Groups: []
RoleBinding[system:image-pullers]:
  Role: system:image-puller
  Users: []
  Groups: [system:serviceaccounts:joe-project]
```

**1** The **alice** user has been added to the **admins RoleBinding**.

## 8.4. GRANTING USERS DAEMONSET PERMISSIONS

By default, project developers do not have the permission to create [daemonsets](#). As a cluster administrator, you can grant them the abilities.

1. Define a **ClusterRole** file:

```
apiVersion: v1
kind: ClusterRole
metadata:
  name: daemonset-admin
rules:
  - resources:
    - daemonsets
  apiGroups:
  - extensions
```

```
verbs:
- create
- get
- list
- watch
- delete
- update
```

2. Create the role:

```
$ oc adm policy add-role-to-user daemonset-admin <user>
```

## 8.5. CREATING A LOCAL ROLE

To create a local role for a project, you can either copy and modify an existing role or build a new role from scratch. It is recommended that you build it from scratch so that you understand each of the permissions assigned.

To copy the cluster role **view** to use as a local role, run:

```
$ oc get clusterrole view -o yaml > clusterrole_view.yaml
$ cp clusterrole_view.yaml localrole_exampleview.yaml
$ vim localrole_exampleview.yaml
# 1. Update kind: ClusterRole to kind: Role
# 2. Update name: view to name: exampleview
# 3. Remove resourceVersion, selfLink, uid, and creationTimestamp
$ oc create -f path/to/localrole_exampleview.yaml -n
<project_you_want_to_add_the_local_role_exampleview_to>
```

To create a new role from scratch, save this snippet into the file **role\_exampleview.yaml**:

### Example Role Named exampleview

```
apiVersion: v1
kind: Role
metadata:
  name: exampleview
rules:
- apiGroups: null
  attributeRestrictions: null
  resources:
  - pods
  - builds
  verbs:
  - get
  - list
  - watch
```

Then, to add the role to your project, run:

```
$ oc project <project_you_want_to_add_the_local_role_exampleview_to>
```

Optionally, annotate it with a description.

Save the following role binding in the `policybinding.yaml` file:

```
apiVersion: v1
kind: PolicyBinding
metadata:
  name: <string>
policyRef:
  name: <role-name>
  namespace: <project-name>
roleBindings: null
```

To create the `PolicyBinding`, run:

```
$ oc create -f policybinding.yaml -n <project-name>
```

To create the role, run:

```
$ oc create -f localrole_exampleview.yaml -n <project-name>
```

To use the new role, run:

```
$ oadm policy add-role-to-user customview <new-user> --role-namespace=
<project-name>
```

## NOTE

A **clusterrolebinding** is a role binding that exists at the cluster level. A **rolebinding** exists at the project level. This can be confusing. The **clusterrolebinding** *view* must be assigned to a user within a project for that user to view the project. Local roles are only created if a cluster role does not provide the set of permissions needed for a particular situation, which is unlikely.

Some cluster role names are initially confusing. The **clusterroleclusteradmin** can be assigned to a user within a project, making it appear that this user has the privileges of a cluster administrator. This is not the case. The **clusteradmin** cluster role bound to a certain project is more like a super administrator for that project, granting the permissions of the cluster role **admin**, plus a few additional permissions like the ability to edit rate limits. This can appear especially confusing via the web console UI, which does not list cluster policy (where cluster administrators exist). However, it does list local policy (where a locally bound **clusteradmin** may exist).

Within a project, project administrators should be able to see **rolebindings**, not **clusterrolebindings**.

## CHAPTER 9. IMAGE POLICY

### 9.1. OVERVIEW

You can control which images are allowed to run on your cluster using the ImagePolicy admission plug-in (currently considered beta). It allows you to control:

- **The source of images:** which registries can be used to pull images
- **Image resolution:** force pods to run with immutable digests to ensure the image does not change due to a re-tag
- **Container image label restrictions:** force an image to have or not have particular labels
- **Image annotation restrictions:** force an image in the integrated container registry to have or not have particular annotations

### 9.2. CONFIGURING THE IMAGEPOLICY ADMISSION PLUG-IN

To configure which images can run on your cluster, configure the ImagePolicy Admission plug-in in the *master-config.yaml* file. You can set one or more rules as required.

- **Reject images with a particular annotation:**  
Use this rule to reject all images that have a specific annotation set on them. The following rejects all images using the `images.openshift.io/deny-execution` annotation:

```
- name: execution-denied
  onResources:
    - resource: pods
    - resource: builds
  reject: true
  matchImageAnnotations:
    - key: images.openshift.io/deny-execution 1
      value: "true"
  skipOnResolutionFailure: true
```

- 1 If a particular image has been deemed harmful, administrators can set this annotation to flag those images.

- **Enable user to run images from Docker Hub:**  
Use this rule to allow users to use images from Docker Hub:

```
- name: allow-images-from-dockerhub
  onResources:
    - resource: pods
    - resource: builds
  matchRegistries:
    - docker.io
```

Following is an example configuration for setting multiple ImagePolicy admission plugin rules in the *master-config.yaml* file:

## Annotated Example File

```

admissionConfig:
  pluginConfig:
    openshift.io/ImagePolicy:
      configuration:
        kind: ImagePolicyConfig
        apiVersion: v1
        resolveImages: AttemptRewrite ❶
        executionRules: ❷
        - name: execution-denied
          # Reject all images that have the annotation
          images.openshift.io/deny-execution set to true.
          # This annotation may be set by infrastructure that wishes to
          flag particular images as dangerous
        onResources: ❸
        - resource: pods
        - resource: builds
        reject: true ❹
        matchImageAnnotations: ❺
        - key: images.openshift.io/deny-execution
          value: "true"
        skipOnResolutionFailure: true ❻
        - name: allow-images-from-internal-registry
          # allows images from the internal registry and tries to resolve
          them
          onResources:
            - resource: pods
            - resource: builds
          matchIntegratedRegistry: true
        - name: allow-images-from-dockerhub
          onResources:
            - resource: pods
            - resource: builds
          matchRegistries:
            - docker.io
        resolutionRules: ❼
        - targetResource:
            resource: pods
            localNames: true
            policy: AttemptRewrite
        - targetResource: ❽
            group: batch
            resource: jobs
            localNames: true ❾
            policy: AttemptRewrite

```

❶ Try to resolve images to an immutable image digest and update the image pull specification in the pod.

❷ Array of rules to evaluate against incoming resources. If you only have **reject: true** rules, the default is **allow all**. If you have any accept rule, that is **reject: false** in any of the rules, the default behaviour of the ImagePolicy switches to **deny-all**.

- 3 Indicates which resources to enforce rules upon. If nothing is specified, the default is **pods**.
- 4 Indicates that if this rule matches, the pod should be rejected.
- 5 List of annotations to match on the image object's metadata.
- 6 If you are not able to resolve the image, do not fail the pod.
- 7 Array of rules allowing use of image streams in Kubernetes resources. The default configuration allows pods, replicationcontrollers, replicaset, statefulsets, daemonsets, deployments, and jobs to use same-project image stream tag references in their image fields.
- 8 Identifies the group and resource to which this rule applies. If resource is **\***, this rule will apply to all resources in that group.
- 9 **LocalNames** will allow single segment names (for example, **ruby:2.4**) to be interpreted as namespace-local image stream tags, but only if the resource or target image stream has **local name resolution** enabled.



## NOTE

If you normally rely on infrastructure images being pulled using a default registry prefix (such as **docker.io** or **registry.access.redhat.com**), those images will not match to any **matchRegistries** value since they will have no registry prefix. To ensure infrastructure images have a registry prefix that can match your image policy, set the **imageConfig.format** value in your **master-config.yaml** file.

## 9.3. TESTING THE IMAGEPOLICY ADMISSION PLUG-IN

1. Use the **openshift/image-policy-check** to test your configuration.  
For example, use the information above, then test like this:

```
oc import-image openshift/image-policy-check:latest --confirm
```

2. Create a pod using this YAML. The pod should be created.

```
apiVersion: v1
kind: Pod
metadata:
  generateName: test-pod
spec:
  containers:
  - image: docker.io/openshift/image-policy-check:latest
    name: first
```

3. Create another pod pointing to a different registry. The pod should be rejected.

```
apiVersion: v1
kind: Pod
metadata:
  generateName: test-pod
spec:
```



```
containers:
- image: different-registry/openshift/image-policy-check:latest
  name: first
```

4. Create a pod pointing to the internal registry using the imported image. The pod should be created and if you look at the image specification, you should see a digest in place of the tag.

```
apiVersion: v1
kind: Pod
metadata:
  generateName: test-pod
spec:
  containers:
  - image: <internal registry IP>:5000/<namespace>/image-policy-check:latest
    name: first
```

5. Create a pod pointing to the internal registry using the imported image. The pod should be created and if you look at the image specification, you should see the tag unmodified.

```
apiVersion: v1
kind: Pod
metadata:
  generateName: test-pod
spec:
  containers:
  - image: <internal registry IP>:5000/<namespace>/image-policy-check:v1
    name: first
```

6. Get the digest from **oc get istag/image-policy-check:latest** and use it for **oc annotate images/<digest> images.openshift.io/deny-execution=true**. For example:

```
$ oc annotate
images/sha256:09ce3d8b5b63595ffca6636c7daefb1a615a7c0e3f8ea68e5db044
a9340d6ba8 images.openshift.io/deny-execution=true
```

7. Create this pod again, and you should see the pod rejected:

```
apiVersion: v1
kind: Pod
metadata:
  generateName: test-pod
spec:
  containers:
  - image: <internal registry IP>:5000/<namespace>/image-policy-check:latest
    name: first
```

## CHAPTER 10. IMAGE SIGNATURES

### 10.1. OVERVIEW

Container image signing on Red Hat Enterprise Linux (RHEL) systems provides a means of:

- Validating where a container image came from,
- Checking that the image has not been tampered with, and
- Setting policies to determine which validated images can be pulled to a host.

For a more complete understanding of the architecture of container image signing on RHEL systems, see the [Container Image Signing Integration Guide](#).

The OpenShift Container Registry allows the ability to store signatures via REST API. The **oc** CLI can be used to verify image signatures, with their validated displayed in the web console or CLI.



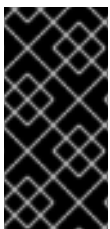
#### NOTE

Initial support for storing image signatures was added in OpenShift Container Platform 3.3. Initial support for verifying image signatures was added in OpenShift Container Platform 3.6.

### 10.2. SIGNING IMAGES USING ATOMIC CLI

OpenShift Container Platform does not automate image signing. Signing requires a developer's private GPG key, typically stored securely on a workstation. This document describes that workflow.

The **atomic** command line interface (CLI), version 1.12.5 or greater, provides commands for signing container images, which can be pushed to an OpenShift Container Registry. The **atomic** CLI is available on Red Hat-based distributions: RHEL, Centos, and Fedora. The **atomic** CLI is pre-installed on RHEL Atomic Host systems. For information on installing the **atomic** package on a RHEL host, see [Enabling Image Signature Support](#).



#### IMPORTANT

The **atomic** CLI uses the authenticated credentials from **oc login**. Be sure to use the same user on the same host for both **atomic** and **oc** commands. For example, if you execute **atomic** CLI as **sudo**, be sure to log in to OpenShift Container Platform using **sudo oc login**.

In order to attach the signature to the image, the user must have the **image-signer** cluster role. Cluster administrators can add this using:

```
$ oc adm policy add-cluster-role-to-user system:image-signer <user_name>
```

Images may be signed at push time:

```
$ atomic push [--sign-by <gpg_key_id>] --type atomic <image>
```

Signatures are stored in OpenShift Container Platform when the **atomic** transport type argument is specified. See [Signature Transports](#) for more information.

For full details on how to set up and perform image signing using the **atomic** CLI, see the [RHEL Atomic Host Managing Containers: Signing Container Images](#) documentation or the **atomic push --help** output for argument details.

A specific example workflow of working with the **atomic** CLI and an OpenShift Container Registry is documented in the [Container Image Signing Integration Guide](#).

## 10.3. VERIFYING IMAGE SIGNATURES USING OPENSIFT CLI

You can verify the signatures of an image imported to an OpenShift Container Registry using the **oc adm verify-image-signature** command. This command verifies if the image identity contained in the image signature can be trusted by using the public GPG key to verify the signature itself then match the provided expected identity with the identity (the pull spec) of the given image.

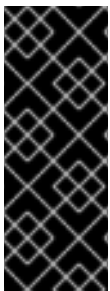
By default, this command uses the public GPG keyring located in **\$GNUPGHOME/pubring.gpg**, typically in path **~/.gnupg**. By default, this command does not save the result of the verification back to the image object. To do so, you must specify the **--save** flag, as shown below.



### NOTE

In order to verify the signature of an image, the user must have the **image-auditor** cluster role. Cluster administrators can add this using:

```
$ oc adm policy add-cluster-role-to-user system:image-auditor
<user_name>
```



### IMPORTANT

Using the **--save** flag on already verified image together with invalid GPG key or invalid expected identity causes the saved verification status and all signatures to be removed, and the image will become unverified.

In order to avoid deleting all signatures by mistake, you can run the command without the **--save** flag first and check the logs for potential issues.

To verify an image signature use the following format:

```
$ oc adm verify-image-signature <image> --expected-identity=<pull_spec> [-
-save] [options]
```

The **<pull\_spec>** can be found by describing the image stream. The **<image>** may be found by describing the image stream tag. See the following example command output.

### Example Image Signature Verification

```
$ oc describe is nodejs -n openshift
Name:                nodejs
Namespace:           openshift
Created:              2 weeks ago
Labels:              <none>
```

```

Annotations:      openshift.io/display-name=Node.js
                  openshift.io/image.dockerRepositoryCheck=2017-07-
05T18:24:01Z
Docker Pull Spec: 172.30.1.1:5000/openshift/nodejs
...

$ oc describe istag nodejs:latest -n openshift
Image Name:
sha256:2bba968aedb7dd2aafe5fa8c7453f5ac36a0b9639f1bf5b03f95de325238b288
...

$ oc adm verify-image-signature \

sha256:2bba968aedb7dd2aafe5fa8c7453f5ac36a0b9639f1bf5b03f95de325238b288 \
--expected-identity 172.30.1.1:5000/openshift/nodejs:latest \
--public-key /etc/pki/rpm-gpg/RPM-GPG-KEY-redhat-release \
--save

```

## 10.4. ACCESSING IMAGE SIGNATURES USING REGISTRY API

The OpenShift Container Registry provides an **extensions** endpoint that allows you to write and read image signatures. The image signatures are stored in the OpenShift Container Platform key-value store via the Docker Registry API.



### NOTE

This endpoint is experimental and not supported by the upstream Docker Registry project. See the [upstream API documentation](#) for general information about the Docker Registry API.

### 10.4.1. Writing Image Signatures via API

In order to add a new signature to the image, you can use the HTTP **PUT** method to send a JSON payload to the **extensions** endpoint:

```

PUT /extensions/v2/<namespace>/<name>/signatures/<digest>

$ curl -X PUT --data @signature.json http://<user>:
<token>@<registry_endpoint>:5000/extensions/v2/<namespace>/<name>/signatur
es/sha256:<digest>

```

The JSON payload with the signature content should have the following structure:

```

{
  "version": 2,
  "type":    "atomic",
  "name":
"sha256:4028782c08eae4a8c9a28bf661c0a8d1c2fc8e19dbaae2b018b21011197e1484@c
ddeb7006d914716e2728000746a0b23",
  "content": "<cryptographic_signature>"
}

```

The **name** field contains the name of the image signature, which must be unique and in the format

**<digest>@<name>**. The **<digest>** represents an image name and the **<name>** is the name of the signature. The signature name must be 32 characters long. The **<cryptographic\_signature>** must follow the specification documented in the [containers/image](#) library.

### 10.4.2. Reading Image Signatures via API

Assuming a signed image has already been pushed into the OpenShift Container Registry, you can read the signatures using the following command:

```
GET /extensions/v2/<namespace>/<name>/signatures/<digest>
```

```
$ curl http://<user>:
<token>@<registry_endpoint>:5000/extensions/v2/<namespace>/<name>/signatur
es/sha256:<digest>
```

The **<namespace>** represents the OpenShift Container Platform project name or registry repository name and the **<name>** refers to the name of the image repository. The **digest** represents the SHA-256 checksum of the image.

If the given image contains the signature data, the output of the command above should produce following JSON response:

```
{
  "signatures": [
    {
      "version": 2,
      "type": "atomic",
      "name":
"sha256:4028782c08eae4a8c9a28bf661c0a8d1c2fc8e19dbaae2b018b21011197e1484@c
ddeb7006d914716e2728000746a0b23",
      "content": "<cryptographic_signature>"
    }
  ]
}
```

The **name** field contains the name of the image signature, which must be unique and in the format **<digest>@<name>**. The **<digest>** represents an image name and the **<name>** is the name of the signature. The signature name must be 32 characters long. The **<cryptographic\_signature>** must follow the specification documented in the [containers/image](#) library.

## CHAPTER 11. SCOPED TOKENS

### 11.1. OVERVIEW

A user may want to give another entity the power to act as they have, but only in a limited way. For example, a project administrator may want to delegate the power to create pods. One way to do this is to create a scoped token.

A scoped token is a token that identifies as a given user, but is limited to certain actions by its scope. Right now, only a **cluster-admin** can create scoped tokens.

### 11.2. EVALUATION

Scopes are evaluated by converting the set of scopes for a token into a set of **PolicyRules**. Then, the request is matched against those rules. The request attributes must match at least one of the scope rules to be passed to the "normal" authorizer for further authorization checks.

### 11.3. USER SCOPES

User scopes are focused on getting information about a given user. They are intent-based, so the rules are automatically created for you:

- **user:full** - Allows full read/write access to the API with all of the user's permissions.
- **user:info** - Allows read-only access to information about the user: name, groups, and so on.
- **user:check-access** - Allows access to **self-localsubjectaccessreviews** and **self-subjectaccessreviews**. These are the variables where you pass an empty user and groups in your request object.
- **user:list-projects** - Allows read-only access to list the projects the user has access to.

### 11.4. ROLE SCOPE

The role scope allows you to have the same level of access as a given role filtered by namespace.

- **role:<cluster-role name>:<namespace or \* for all>** - Limits the scope to the rules specified by the cluster-role, but only in the specified namespace .



#### NOTE

Caveat: This prevents escalating access. Even if the role allows access to resources like secrets, rolebindings, and roles, this scope will deny access to those resources. This helps prevent unexpected escalations. Many people do not think of a role like **edit** as being an escalating role, but with access to a secret it is.

- **role:<cluster-role name>:<namespace or \* for all>:!** - This is similar to the example above, except that including the bang causes this scope to allow escalating access.

## CHAPTER 12. MONITORING IMAGES

### 12.1. OVERVIEW

You can monitor [images](#) in your instance using the [CLI](#).

### 12.2. VIEWING IMAGES STATISTICS

OpenShift Container Platform can display several usage statistics about all the images it manages. In other words, all the images pushed to the internal registry either [directly](#) or through a [build](#).

To view the usage statistics:

```
$ oc adm top images
NAME                                IMAGESTREAMTAG      PARENTS
USAGE                                METADATA           STORAGE
sha256:80c985739a78b openshift/python (3.5)
yes                                303.12MiB
sha256:64461b5111fc7 openshift/ruby (2.2)
yes                                234.33MiB
sha256:0e19a0290ddc1 test/ruby-ex (latest)    sha256:64461b5111fc71ec
Deployment: ruby-ex-1/test    yes            150.65MiB
sha256:a968c61adad58 test/django-ex (latest)  sha256:80c985739a78b760
Deployment: django-ex-1/test  yes            186.07MiB
```

The command displays the following information:

- image ID
- project, name, and tag of the accompanying **ImageStreamTag**
- potential parents of the image, using their ID
- information about where the image is being used
- flag informing whether the image contains proper Docker metadata information
- size of the image

### 12.3. VIEWING IMAGESTREAMS STATISTICS

OpenShift Container Platform can display several usage statistics about all the **ImageStreams**.

To view the usage statistics:

```
$ oc adm top imagestreams
NAME                STORAGE    IMAGES  LAYERS
openshift/python    1.21GiB    4        36
openshift/ruby       717.76MiB  3        27
test/ruby-ex         150.65MiB  1        10
test/django-ex       186.07MiB  1        10
```

The command displays the following information:

- project and name of the **ImageStream**
- size of the entire **ImageStream** stored in the internal [Red Hat Container Registry](#)
- number of images this particular **ImageStream** is pointing to
- number of layers **ImageStream** consists of

## 12.4. PRUNING IMAGES

The information returned from the above commands is helpful when performing [image pruning](#).



## CHAPTER 13. MANAGING SECURITY CONTEXT CONSTRAINTS

### 13.1. OVERVIEW

Security context constraints allow administrators to control permissions for pods. To learn more about this API type, see the [security context constraints \(SCCs\)](#) architecture documentation. You can manage SCCs in your instance as normal API [objects](#) using [the CLI](#).



#### NOTE

You must have [cluster-admin privileges](#) to manage SCCs.



#### IMPORTANT

Do not modify the default SCCs. Customizing the default SCCs can lead to issues when upgrading. Instead, [create new SCCs](#).

### 13.2. LISTING SECURITY CONTEXT CONSTRAINTS

To get a current list of SCCs:

```
$ 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 |
| 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]               |                       |

### 13.3. EXAMINING A SECURITY CONTEXT CONSTRAINTS OBJECT

To examine a particular SCC, use **oc get**, **oc describe**, **oc export**, or **oc edit**. For example, to examine the **restricted** SCC:

```
$ oc describe scc restricted
Name:      restricted
Priority:   <none>
Access:
  Users:    <none>
  Groups:   system:authenticated
Settings:
  Allow Privileged:  false
  Default Add Capabilities:  <none>
  Required Drop Capabilities:  KILL,MKNOD,SYS_CHROOT,SETUID,SETGID
  Allowed Capabilities:  <none>
  Allowed Seccomp Profiles:  <none>
  Allowed Volume Types:
configMap,downwardAPI,emptyDir,persistentVolumeClaim,projected,secret
  Allow Host Network:  false
  Allow Host Ports:    false
  Allow Host PID:      false
  Allow Host IPC:      false
  Read Only Root Filesystem:  false
  Run As User Strategy: MustRunAsRange
    UID:    <none>
    UID Range Min:  <none>
    UID Range Max:  <none>
  SELinux Context Strategy: MustRunAs
    User:    <none>
    Role:    <none>
    Type:    <none>
    Level:   <none>
  FSGroup Strategy: MustRunAs
    Ranges:  <none>
  Supplemental Groups Strategy: RunAsAny
    Ranges:  <none>
```

**NOTE**

In order to preserve customized SCCs during upgrades, do not edit settings on the default SCCs other than priority, users, groups, labels, and annotations.

## 13.4. CREATING NEW SECURITY CONTEXT CONSTRAINTS

To create a new SCC:

1. Define the SCC in a JSON or YAML file:

### Security Context Constraint Object Definition

```
kind: SecurityContextConstraints
apiVersion: v1
metadata:
  name: scc-admin
allowPrivilegedContainer: true
runAsUser:
  type: RunAsAny
seLinuxContext:
```

```

    type: RunAsAny
  fsGroup:
    type: RunAsAny
  supplementalGroups:
    type: RunAsAny
  users:
  - my-admin-user
  groups:
  - my-admin-group

```

Optionally, you can add drop capabilities to an SCC by setting the **requiredDropCapabilities** field with the desired values. Any specified capabilities will be dropped from the container. For example, to create an SCC with the **KILL**, **MKNOD**, and **SYS\_CHROOT** required drop capabilities, add the following to the SCC object:

```

requiredDropCapabilities:
- KILL
- MKNOD
- SYS_CHROOT

```

You can see the list of possible values in the [Docker documentation](#).

## TIP

Because capabilities are passed to the Docker, you can use a special **ALL** value to drop all possible capabilities.

2. Then, run **oc create** passing the file to create it:

```

$ oc create -f scc_admin.yaml
securitycontextconstraints "scc-admin" created

```

3. Verify that the SCC was created:

```

$ oc get scc scc-admin
NAME          PRIV          CAPS          SELINUX     RUNASUSER   FSGROUP
SUPGROUP     PRIORITY     READONLYROOTFS  VOLUMES
scc-admin    true         []            RunAsAny    RunAsAny    RunAsAny
RunAsAny     <none>       false         [awsElasticBlockStore
azureDisk azureFile cephFS cinder configMap downwardAPI emptyDir fc
flexVolume flocker gcePersistentDisk gitRepo glusterfs iscsi nfs
persistentVolumeClaim photonPersistentDisk quobyte rbd secret
vsphere]

```

## 13.5. DELETING SECURITY CONTEXT CONSTRAINTS

To delete an SCC:

```

$ oc delete scc <scc_name>

```

**NOTE**

If you delete a default SCC, it will be regenerated upon restart.

## 13.6. UPDATING SECURITY CONTEXT CONSTRAINTS

To update an existing SCC:

```
$ oc edit scc <scc_name>
```

**NOTE**

In order to preserve customized SCCs during upgrades, do not edit settings on the default SCCs other than priority, users, and groups.

### 13.6.1. Example Security Context Constraints Settings

#### Without Explicit runAsUser Setting

```
apiVersion: v1
kind: Pod
metadata:
  name: security-context-demo
spec:
  securityContext: ❶
  containers:
  - name: sec-ctx-demo
    image: gcr.io/google-samples/node-hello:1.0
```

- ❶ When a container or pod does not request a user ID under which it should be run, the effective UID depends on the SCC that emits this pod. Because restricted SCC is granted to all authenticated users by default, it will be available to all users and service accounts and used in most cases. The restricted SCC uses **MustRunAsRange** strategy for constraining and defaulting the possible values of the **securityContext.runAsUser** field. The admission plug-in will look for the **openshift.io/sa.scc.uid-range** annotation on the current project to populate range fields, as it does not provide this range. In the end, a container will have **runAsUser** equal to the first value of the range that is hard to predict because every project has different ranges. See [Understanding Pre-allocated Values and Security Context Constraints](#) for more information.

#### With Explicit runAsUser Setting

```
apiVersion: v1
kind: Pod
metadata:
  name: security-context-demo
spec:
  securityContext:
    runAsUser: 1000 ❶
  containers:
  - name: sec-ctx-demo
    image: gcr.io/google-samples/node-hello:1.0
```

- 1 A container or pod that requests a specific user ID will be accepted by OpenShift Container Platform only when a service account or a user is granted access to a SCC that allows such a user ID. The SCC can allow arbitrary IDs, an ID that falls into a range, or the exact user ID specific to the request.

This works with SELinux, fsGroup, and Supplemental Groups. See [Volume Security](#) for more information.

## 13.7. UPDATING THE DEFAULT SECURITY CONTEXT CONSTRAINTS

Default SCCs will be created when the master is started if they are missing. To reset SCCs to defaults, or update existing SCCs to new default definitions after an upgrade you may:

1. Delete any SCC you would like to be reset and let it be recreated by restarting the master
2. Use the `oc adm policy reconcile-sccs` command

The `oc adm policy reconcile-sccs` command will set all SCC policies to the default values but retain any additional users, groups, labels, and annotations as well as priorities you may have already set. To view which SCCs will be changed you may run the command with no options or by specifying your preferred output with the `-o <format>` option.

After reviewing it is recommended that you back up your existing SCCs and then use the `--confirm` option to persist the data.



### NOTE

If you would like to reset priorities and grants, use the `--additive-only=false` option.



### NOTE

If you have customized settings other than priority, users, groups, labels, or annotations in an SCC, you will lose those settings when you reconcile.

## 13.8. HOW DO I?

The following describe common scenarios and procedures using SCCs.

### 13.8.1. Grant Access to the Privileged SCC

In some cases, an administrator might want to allow users or groups outside the administrator group access to create more *privileged pods*. To do so, you can:

1. Determine the user or group you would like to have access to the SCC.

**WARNING**

Granting access to a user only works when the user directly creates a pod. For pods created on behalf of a user, **in most cases** by the system itself, **access should be given to a service account** under which related controller is operated upon. Examples of resources that create pods on behalf of a user are Deployments, StatefulSets, DaemonSets, etc.

2. Run:

```
$ oc adm policy add-scc-to-user <scc_name> <user_name>
$ oc adm policy add-scc-to-group <scc_name> <group_name>
```

For example, to allow the **e2e-user** access to the **privileged** SCC, run:

```
$ oc adm policy add-scc-to-user privileged e2e-user
```

3. Modify **SecurityContext** of a container to request a privileged mode.

### 13.8.2. Grant a Service Account Access to the Privileged SCC

First, create a [service account](#). For example, to create service account **mysvcacct** in project **myproject**:

```
$ oc create serviceaccount mysvcacct -n myproject
```

Then, add the service account to the **privileged** SCC.

```
$ oc adm policy add-scc-to-user privileged
system:serviceaccount:myproject:mysvcacct
```

Then, ensure that the resource is being created on behalf of the service account. To do so, set the **spec.serviceAccountName** field to a service account name. Leaving the service account name blank will result in the **default** service account being used.

Then, ensure that at least one of the pod's containers is requesting a privileged mode in the security context.

### 13.8.3. Enable Images to Run with USER in the Dockerfile

To relax the security in your cluster so that images are not forced to run as a pre-allocated UID, without granting everyone access to the **privileged** SCC:

1. Grant all authenticated users access to the **anyuid** SCC:

```
$ oc adm policy add-scc-to-group anyuid system:authenticated
```

**WARNING**

This allows images to run as the root UID if no **USER** is specified in the *Dockerfile*.

### 13.8.4. Enable Container Images that Require Root

Some container images (examples: **postgres** and **redis**) require root access and have certain expectations about how volumes are owned. For these images, add the service account to the **anyuid** SCC.

```
$ oc adm policy add-scc-to-user anyuid
system:serviceaccount:myproject:mysvcacct
```

### 13.8.5. Use --mount-host on the Registry

It is recommended that [persistent storage](#) using **PersistentVolume** and **PersistentVolumeClaim** objects be used for [registry deployments](#). If you are testing and would like to instead use the **oc adm registry** command with the **--mount-host** option, you must first create a new [service account](#) for the registry and add it to the **privileged** SCC. See the [Administrator Guide](#) for full instructions.

### 13.8.6. Provide Additional Capabilities

In some cases, an image may require capabilities that Docker does not provide out of the box. You can provide the ability to request additional capabilities in the pod specification which will be validated against an SCC.

**IMPORTANT**

This allows images to run with elevated capabilities and should be used only if necessary. You should not edit the default **restricted** SCC to enable additional capabilities.

When used in conjunction with a non-root user, you must also ensure that the file that requires the additional capability is granted the capabilities using the **setcap** command. For example, in the *Dockerfile* of the image:

```
setcap cap_net_raw,cap_net_admin+p /usr/bin/ping
```

Further, if a capability is provided by default in Docker, you do not need to modify the pod specification to request it. For example, **NET\_RAW** is provided by default and capabilities should already be set on **ping**, therefore no special steps should be required to run **ping**.

To provide additional capabilities:

1. Create a new SCC
2. Add the allowed capability using the **allowedCapabilities** field.

3. When creating the pod, request the capability in the `securityContext.capabilities.add` field.

### 13.8.7. Modify Cluster Default Behavior

When you grant access to the **anyuid** SCC for everyone, your cluster:

- Does not pre-allocate UIDs
- Allows containers to run as any user
- Prevents privileged containers

```
$ oc adm policy add-scc-to-group anyuid system:authenticated
```

To modify your cluster so that it does not pre-allocate UIDs and does not allow containers to run as root, grant access to the **nonroot** SCC for everyone:

```
$ oc adm policy add-scc-to-group nonroot system:authenticated
```



#### WARNING

Be very careful with any modifications that have a cluster-wide impact. When you grant an SCC to all authenticated users, as in the previous example, or modify an SCC that applies to all users, such as the **restricted** SCC, it also affects Kubernetes and OpenShift Container Platform components, including the web console and integrated docker registry. Changes made with these SCCs can cause these components to stop functioning.

Instead, create a custom SCC and target it to only specific users or groups. This way potential issues are confined to the affected users or groups and do not impact critical cluster components.

### 13.8.8. Use the hostPath Volume Plug-in

To relax the security in your cluster so that pods are allowed to use the **hostPath** volume plug-in without granting everyone access to more privileged SCCs such as **privileged**, **hostaccess**, or **hostmount-anyuid**, perform the following actions:

1. [Create a new SCC](#) named **hostpath**
2. Set the `allowHostDirVolumePlugin` parameter to **true** for the new SCC:

```
$ oc patch scc hostpath -p '{"allowHostDirVolumePlugin": true}'
```

3. Grant access to this SCC to all users:

```
$ oc adm policy add-scc-to-group hostpath system:authenticated
```



Now, all the pods that request **hostPath** volumes are admitted by the **hostpath** SCC.

### 13.8.9. Ensure That Admission Attempts to Use a Specific SCC First

You may control the sort ordering of SCCs in admission by setting the **Priority** field of the SCCs. See the [SCC Prioritization](#) section for more information on sorting.

### 13.8.10. Add an SCC to a User, Group, or Project

Before adding an SCC to a user or group, you can first use the **scc-review** option to check if the user or group can create a pod. See the [Authorization](#) topic for more information.

SCCs are not granted directly to a project. Instead, you add a service account to an SCC and either specify the service account name on your pod or, when unspecified, run as the **default** service account.

To add an SCC to a user:

```
$ oc adm policy add-scc-to-user <scc_name> <user_name>
```

To add an SCC to a service account:

```
$ oc adm policy add-scc-to-user <scc_name> \
    system:serviceaccount:<serviceaccount_namespace>:<serviceaccount_name>
```

If you are currently in the project to which the service account belongs, you can use the **-z** flag and just specify the **<serviceaccount\_name>**.

```
$ oc adm policy add-scc-to-user <scc_name> -z <serviceaccount_name>
```



#### IMPORTANT

Usage of the **-z** flag as described above is highly recommended, as it helps prevent typos and ensures that access is granted only to the specified service account. If not in the project, use the **-n** option to indicate the project namespace it applies to.

To add an SCC to a group:

```
$ oc adm policy add-scc-to-group <scc_name> <group_name>
```

To add an SCC to all service accounts in a namespace:

```
$ oc adm policy add-scc-to-group <scc_name> \
    system:serviceaccounts:<serviceaccount_namespace>
```

## CHAPTER 14. SCHEDULING

### 14.1. OVERVIEW

#### 14.1.1. Overview

Pod scheduling is an internal process that determines placement of new pods onto nodes within the cluster.

The scheduler code has a clean separation that watches new pods as they get created and identifies the most suitable node to host them. It then creates bindings (pod to node bindings) for the pods using the master API.

#### 14.1.2. Default scheduling

OpenShift Container Platform comes with a default scheduler that serves the needs of most users. The default scheduler uses both inherent and customizable tools to determine the best fit for a pod.

For information on how the default scheduler determines pod placement and available customizable parameters, see [Default Scheduling](#).

#### 14.1.3. Advanced scheduling

In situations where you might want more control over where new pods are placed, the OpenShift Container Platform advanced scheduling features allow you to configure a pod so that the pod is required to (or has a preference to) run on a particular node or alongside a specific pod. Advanced scheduling also allows you to prevent pods from being placed on a node or with another pod.

For information about advanced scheduling, see [Advanced Scheduling](#).

#### 14.1.4. Custom scheduling

OpenShift Container Platform also allows you to use your own or third-party schedulers by editing the pod specification.

For more information, see [Custom Schedulers](#).

### 14.2. DEFAULT SCHEDULING

#### 14.2.1. Overview

The default OpenShift Container Platform pod scheduler is responsible for determining placement of new pods onto nodes within the cluster. It reads data from the pod and tries to find a node that is a good fit based on configured policies. It is completely independent and exists as a standalone/pluggable solution. It does not modify the pod and just creates a binding for the pod that ties the pod to the particular node.

#### 14.2.2. Generic Scheduler

The existing generic scheduler is the default platform-provided scheduler *engine* that selects a node to host the pod in a three-step operation:

1. The scheduler [filters out inappropriate nodes using predicates](#).

2. The scheduler [prioritizes the filtered list of nodes](#).
3. The scheduler [selects the highest priority node](#) for the pod.

### 14.2.3. Filter the Nodes

The available nodes are filtered based on the constraints or requirements specified. This is done by running each node through the list of filter functions called [predicates](#).

#### 14.2.3.1. Prioritize the Filtered List of Nodes

This is achieved by passing each node through a series of [priority functions](#) that assign it a score between 0 - 10, with 0 indicating a bad fit and 10 indicating a good fit to host the pod. The scheduler configuration can also take in a simple *weight* (positive numeric value) for each priority function. The node score provided by each priority function is multiplied by the weight (default weight for most priorities is 1) and then combined by adding the scores for each node provided by all the priorities. This weight attribute can be used by administrators to give higher importance to some priorities.

#### 14.2.3.2. Select the Best Fit Node

The nodes are sorted based on their scores and the node with the highest score is selected to host the pod. If multiple nodes have the same high score, then one of them is selected at random.

### 14.2.4. Scheduler Policy

The selection of the [predicate](#) and [priorities](#) defines the policy for the scheduler.

The scheduler configuration file is a JSON file that specifies the predicates and priorities the scheduler will consider.

In the absence of the scheduler policy file, the default configuration file, */etc/origin/master/scheduler.json*, gets applied.



#### IMPORTANT

The predicates and priorities defined in the scheduler configuration file completely override the default scheduler policy. If any of the default predicates and priorities are required, you must explicitly specify the functions in the scheduler configuration file.

#### Default scheduler configuration file

```
{
  "apiVersion": "v1",
  "kind": "Policy",
  "predicates": [
    {
      "name": "NoVolumeZoneConflict"
    },
    {
      "name": "MaxEBSVolumeCount"
    },
    {
      "name": "MaxGCEPDVolumeCount"
    }
  ],
}
```

```

    {
      "name": "MaxAzureDiskVolumeCount"
    },
    {
      "name": "MatchInterPodAffinity"
    },
    {
      "name": "NoDiskConflict"
    },
    {
      "name": "GeneralPredicates"
    },
    {
      "name": "PodToleratesNodeTaints"
    },
    {
      "name": "CheckNodeMemoryPressure"
    },
    {
      "name": "CheckNodeDiskPressure"
    },
    {
      "name": "NoVolumeNodeConflict"
    },
    {
      "argument": {
        "serviceAffinity": {
          "labels": [
            "region"
          ]
        }
      },
      "name": "Region"
    }
  ],
  "priorities": [
    {
      "name": "SelectorSpreadPriority",
      "weight": 1
    },
    {
      "name": "InterPodAffinityPriority",
      "weight": 1
    },
    {
      "name": "LeastRequestedPriority",
      "weight": 1
    },
    {
      "name": "BalancedResourceAllocation",
      "weight": 1
    },
    {
      "name": "NodePreferAvoidPodsPriority",
      "weight": 10000
    }
  ]
}

```

```

    },
    {
      "name": "NodeAffinityPriority",
      "weight": 1
    },
    {
      "name": "TaintTolerationPriority",
      "weight": 1
    },
    {
      "argument": {
        "serviceAntiAffinity": {
          "label": "zone"
        }
      },
      "name": "Zone",
      "weight": 2
    }
  ]
}

```

#### 14.2.4.1. Modifying Scheduler Policy

The scheduler policy is defined in a file on the master, named */etc/origin/master/scheduler.json* by default, unless overridden by the `kubernetesMasterConfig.schedulerConfigFile` field in the [master configuration file](#).

#### Sample modified scheduler configuration file

```

kind: "Policy"
version: "v1"
"predicates": [
  {
    "name": "PodFitsResources"
  },
  {
    "name": "NoDiskConflict"
  },
  {
    "name": "MatchNodeSelector"
  },
  {
    "name": "HostName"
  },
  {
    "name": "NoVolumeNodeConflict"
  },
  {
    "argument": {
      "serviceAffinity": {
        "labels": [
          "region"
        ]
      }
    }
  },

```

```

        "name": "Region"
    },
    ],
    "priorities": [
        {
            "name": "LeastRequestedPriority",
            "weight": 1
        },
        {
            "name": "BalancedResourceAllocation",
            "weight": 1
        },
        {
            "name": "ServiceSpreadingPriority",
            "weight": 1
        },
        {
            "argument": {
                "serviceAntiAffinity": {
                    "label": "zone"
                }
            },
            "name": "Zone",
            "weight": 2
        }
    ]
]

```

To modify the scheduler policy:

1. Edit the scheduler configuration file to configure the desired [default predicates and priorities](#). You can create a custom configuration, or use and modify one of the [sample policy configurations](#).
2. Add any [configurable predicates](#) and [configurable priorities](#) you require.
3. Restart the OpenShift Container Platform for the changes to take effect.

```
# systemctl restart atomic-openshift-master-api atomic-openshift-
master-controllers
```

## 14.2.5. Available Predicates

Predicates are rules that filter out unqualified nodes.

There are several predicates provided by default in OpenShift Container Platform. Some of these predicates can be customized by providing certain parameters. Multiple predicates can be combined to provide additional filtering of nodes.

### 14.2.5.1. Static Predicates

These predicates do not take any configuration parameters or inputs from the user. These are specified in the scheduler configuration using their exact name.

#### 14.2.5.1.1. Default Predicates

The default scheduler policy includes the following predicates:

**NoVolumeZoneConflict** checks that the volumes a pod requests are available in the zone.

```
{ "name" : "NoVolumeZoneConflict" }
```

**MaxEBSVolumeCount** checks the maximum number of volumes that can be attached to an AWS instance.

```
{ "name" : "MaxEBSVolumeCount" }
```

**MaxGCEPDVolumeCount** checks the maximum number of Google Compute Engine (GCE) Persistent Disks (PD).

```
{ "name" : "MaxGCEPDVolumeCount" }
```

**MatchInterPodAffinity** checks if the pod affinity/antiaffinity rules permit the pod.

```
{ "name" : "MatchInterPodAffinity" }
```

**NoDiskConflict** checks if the volume requested by a pod is available.

```
{ "name" : "NoDiskConflict" }
```

**PodToleratesNodeTaints** checks if a pod can tolerate the node taints.

```
{ "name" : "PodToleratesNodeTaints" }
```

**CheckNodeMemoryPressure** checks if a pod can be scheduled on a node with a memory pressure condition.

```
{ "name" : "CheckNodeMemoryPressure" }
```

**CheckNodeDiskPressure** checks if a pod can be scheduled on a node with a disk pressure condition.

```
{ "name" : "CheckNodeDiskPressure" }
```

#### 14.2.5.1.2. Other Static Priorities

OpenShift Container Platform also supports the following priorities:

**NoVolumeNodeConflict**

```
{ "name" : "NoVolumeNodeConflict" }
```

**CheckVolumeBinding** evaluates if a pod can fit based on the volumes, it requests, for both bound and unbound PVCs. \* For PVCs that are bound, the predicate checks that the corresponding PV's node affinity is satisfied by the given node. \* For PVCs that are unbound, the predicate searched for available PVs that can satisfy the PVC requirements and that the PV node affinity is satisfied by the given node.

The predicate returns true if all bound PVCs have compatible PVs with the node, and if all unbound PVCs can be matched with an available and node-compatible PV.

```
{ "name" : "CheckVolumeBinding" }
```

The **CheckVolumeBinding** predicate must be enabled in non-default schedulers.

**CheckNodeCondition** checks if a pod can be scheduled on a node reporting **out of disk**, **network unavailable**, or **not ready** conditions.

```
{ "name" : "CheckNodeCondition" }
```

**PodToleratesNodeNoExecuteTaints** checks if a pod tolerations can tolerate a node **NoExecute** taints.

```
{ "name" : "PodToleratesNodeNoExecuteTaints" }
```

**CheckNodeLabelPresence** checks if all of the specified labels exist on a node, regardless of their value.

```
{ "name" : "CheckNodeLabelPresence" }
```

**checkServiceAffinity** checks that ServiceAffinity labels are homogeneous for pods that are scheduled on a node.

```
{ "name" : "checkServiceAffinity" }
```

**MaxAzureDiskVolumeCount** checks the maximum number of Azure Disk Volumes.

```
{ "name" : "MaxAzureDiskVolumeCount" }
```

#### 14.2.5.2. General Predicates

The following general predicates check whether non-critical predicates and essential predicates pass. Non-critical predicates are the predicates that only non-critical pods need to pass and essential predicates are the predicates that all pods need to pass.

*The default scheduler policy includes the general predicates.*

##### Non-critical general predicates

**PodFitsResources** determines a fit based on resource availability (CPU, memory, GPU, and so forth). The nodes can declare their resource capacities and then pods can specify what resources they require. Fit is based on requested, rather than used resources.

```
{ "name" : "PodFitsResources" }
```

##### Essential general predicates

**PodFitsHostPorts** determines if a node has free ports for the requested pod ports (absence of port conflicts).

```
{ "name" : "PodFitsHostPorts" }
```

**HostName** determines fit based on the presence of the Host parameter and a string match with the name of the host.

```
{ "name" : "HostName" }
```



**MatchNodeSelector** determines fit based on [node selector](#) (`nodeSelector`) queries defined in the pod.

```
{ "name" : "MatchNodeSelector" }
```

### 14.2.5.3. Configurable Predicates

You can configure these predicates in the scheduler configuration, by default `/etc/origin/master/scheduler.json`, to add labels to affect how the predicate functions.

Since these are configurable, multiple predicates of the same type (but different configuration parameters) can be combined as long as their user-defined names are different.

For information on using these priorities, see [Modifying Scheduler Policy](#).

**ServiceAffinity** places pods on nodes based on the service running on that pod. Placing pods of the same service on the same or co-located nodes can lead to higher efficiency.

This predicate attempts to place pods with specific labels in its [node selector](#) on nodes that have the same label.

If the pod does not specify the labels in its node selector, then the first pod is placed on any node based on availability and all subsequent pods of the service are scheduled on nodes that have the same label values as that node.

```
"predicates": [
  {
    "name": "<name>", ①
    "argument": {
      "serviceAffinity": {
        "labels": [
          "<label>" ②
        ]
      }
    }
  }
],
```

① Specify a name for the predicate.

② Specify a label to match.

For example:

```
"name": "ZoneAffinity",
"argument": {
  "serviceAffinity": {
    "labels": [
      "rack"
    ]
  }
}
```

For example, if the first pod of a service had a node selector `rack` was scheduled to a node with label `region=rack`, all the other subsequent pods belonging to the same service will be scheduled on nodes with the same `region=rack` label. For more information, see [Controlling Pod Placement](#).

Multiple-level labels are also supported. Users can also specify all pods for a service to be scheduled on nodes within the same region and within the same zone (under the region).

The **labelsPresence** parameter checks whether a particular node has a specific label. The labels create node *groups* that the **LabelPreference** priority uses. Matching by label can be useful, for example, where nodes have their physical location or status defined by labels.

```
"predicates":[
  {
    "name": "<name>", ❶
    "argument":{
      "labelsPresence":{
        "labels":[
          "<label>" ❷
        ],
        "presence": true ❸
      }
    }
  }
],
```

❶ Specify a name for the predicate.

❷ Specify a label to match.

❸ Specify whether the labels are required, either **true** or **false**.

- For **presence: false**, if any of the requested labels are present in the node labels, the pod cannot be scheduled. If the labels are not present, the pod can be scheduled.
- For **presence: true**, if all of the requested labels are present in the node labels, the pod can be scheduled. If all of the labels are not present, the pod is not scheduled.

For example:

```
"name": "RackPreferred",
"argument":{
  "labelsPresence":{
    "labels":[
      "rack"
    ],
    "labelsPresence":{
      "labels":["region"]
    },
    "presence": true
  }
}
],
```

### 14.2.6. Available Priorities

Priorities are rules that rank remaining nodes according to preferences.

A custom set of priorities can be specified to configure the scheduler. There are several priorities provided by default in OpenShift Container Platform. Other priorities can be customized by providing certain parameters. Multiple priorities can be combined and different weights can be given to each in order to impact the prioritization.

### 14.2.6.1. Static Priorities

Static priorities do not take any configuration parameters from the user, except weight. A weight is required to be specified and cannot be 0 or negative.

These are specified in the scheduler configuration, by default `/etc/origin/master/scheduler.json`.

#### 14.2.6.1.1. Default Priorities

The default scheduler policy includes the following priorities. Each of the priority function has a weight of **1** except **NodePreferAvoidPodsPriority**, which has a weight of **10000**.

**SelectorSpreadPriority** looks for services, replication controllers (RC), replication sets (RS), and stateful sets that match the pod, then finds existing pods that match those selectors. The scheduler favors nodes that have fewer existing matching pods. Then, it schedules the pod on a node with the smallest number of pods that match those selectors as the pod being scheduled.

```
{"name" : "SelectorSpreadPriority", "weight" : 1}
```

**InterPodAffinityPriority** computes a sum by iterating through the elements of **weightedPodAffinityTerm** and adding *weight* to the sum if the corresponding **PodAffinityTerm** is satisfied for that node. The node(s) with the highest sum are the most preferred.

```
{"name" : "InterPodAffinityPriority", "weight" : 1}
```

**LeastRequestedPriority** favors nodes with fewer requested resources. It calculates the percentage of memory and CPU requested by pods scheduled on the node, and prioritizes nodes that have the highest available/remaining capacity.

```
{"name" : "LeastRequestedPriority", "weight" : 1}
```

**BalancedResourceAllocation** favors nodes with balanced resource usage rate. It calculates the difference between the consumed CPU and memory as a fraction of capacity, and prioritizes the nodes based on how close the two metrics are to each other. This should always be used together with **LeastRequestedPriority**.

```
{"name" : "BalancedResourceAllocation", "weight" : 1}
```

**NodePreferAvoidPodsPriority** ignores pods that are owned by a controller other than a replication controller.

```
{"name" : "NodePreferAvoidPodsPriority", "weight" : 10000}
```

**NodeAffinityPriority** prioritizes nodes according to node affinity scheduling preferences

```
{"name" : "NodeAffinityPriority", "weight" : 1}
```

**TaintTolerationPriority** prioritizes nodes that have a fewer number of *intolerable* taints on them for a pod. An intolerable taint is one which has key **PreferNoSchedule**.

```
{ "name" : "TaintTolerationPriority", "weight" : 1 }
```

#### 14.2.6.1.2. Other Static Priorities

OpenShift Container Platform also supports the following priorities:

**EqualPriority** gives an equal weight of **1** to all nodes, if no priority configurations are provided. We recommend using this priority only for testing environments.

```
{ "name" : "EqualPriority", "weight" : 1 }
```

**MostRequestedPriority** prioritizes nodes with most requested resources. It calculates the percentage of memory and CPU requested by pods scheduled on the node, and prioritizes based on the maximum of the average of the fraction of requested to capacity.

```
{ "name" : "MostRequestedPriority", "weight" : 1 }
```

**ImageLocalityPriority** prioritizes nodes that already have requested pod container's images.

```
{ "name" : "ImageLocalityPriority", "weight" : 1 }
```

**ServiceSpreadingPriority** spreads pods by minimizing the number of pods belonging to the same service onto the same machine.

```
{ "name" : "ServiceSpreadingPriority", "weight" : 1 }
```

#### 14.2.6.2. Configurable Priorities

You can configure these priorities in the scheduler configuration, by default **/etc/origin/master/scheduler.json**, to add labels to affect how the priorities.

The type of the priority function is identified by the argument that they take. Since these are configurable, multiple priorities of the same type (but different configuration parameters) can be combined as long as their user-defined names are different.

For information on using these priorities, see [Modifying Scheduler Policy](#).

**ServiceAntiAffinity** takes a label and ensures a good spread of the pods belonging to the same service across the group of nodes based on the label values. It gives the same score to all nodes that have the same value for the specified label. It gives a higher score to nodes within a group with the least concentration of pods.

```
"priorities": [
  {
    "name": "<name>", 1
    "weight" : "1" 2
    "argument": {
      "serviceAntiAffinity": {
        "label": [
          "<label>" 3
        ]
      }
    }
  }
]
```

```

    ]
  }
}
]

```

- 1 Specify a name for the priority.
- 2 Specify a weight. Enter a non-zero positive value.
- 3 Specify a label to match.

For example:

```

"name": "RackSpread", 1
"weight" : "1" 2
"argument": {
  "serviceAffinity": {
    "label": [ 3
      "rack"
    ]
  }
}

```

- 1 Specify a name for the priority.
- 2 Specify a weight. Enter a non-zero positive value.
- 3 Specify a label to match.



## NOTE

In some situations using **ServiceAntiAffinity** based on custom labels does not spread pod as expected. See [this Red Hat Solution](#).

\*The **labelPreference** parameter gives priority based on the specified label. If the label is present on a node, that node is given priority. If no label is specified, priority is given to nodes that do not have a label.

```

"priorities": [
  {
    "name": "<name>", 1
    "weight" : "1" 2
    "argument": {
      "labelPreference": {
        "label": [ 3
          "<label>"
        ]
      }
    }
  }
],

```

- 1 Specify a name for the priority.

2 Specify a weight. Enter a non-zero positive value.

3 Specify a label to match.

## 14.2.7. Use Cases

One of the important use cases for scheduling within OpenShift Container Platform is to support flexible affinity and anti-affinity policies.

### 14.2.7.1. Infrastructure Topological Levels

Administrators can define multiple topological levels for their infrastructure (nodes) by specifying [labels on nodes](#) (e.g., **region=r1**, **zone=z1**, **rack=s1**).

These label names have no particular meaning and administrators are free to name their infrastructure levels anything (eg, city/building/room). Also, administrators can define any number of levels for their infrastructure topology, with three levels usually being adequate (such as: **regions** → **zones** → **racks**). Administrators can specify affinity and anti-affinity rules at each of these levels in any combination.

### 14.2.7.2. Affinity

Administrators should be able to configure the scheduler to specify affinity at any topological level, or even at multiple levels. Affinity at a particular level indicates that all pods that belong to the same service are scheduled onto nodes that belong to the same level. This handles any latency requirements of applications by allowing administrators to ensure that peer pods do not end up being too geographically separated. If no node is available within the same affinity group to host the pod, then the pod is not scheduled.

If you need greater control over where the pods are scheduled, see [Using Node Affinity](#) and [Using Pod Affinity and Anti-affinity](#). These advanced scheduling features allow administrators to specify which node a pod can be scheduled on and to force or reject scheduling relative to other pods.

### 14.2.7.3. Anti Affinity

Administrators should be able to configure the scheduler to specify anti-affinity at any topological level, or even at multiple levels. Anti-affinity (or 'spread') at a particular level indicates that all pods that belong to the same service are spread across nodes that belong to that level. This ensures that the application is well spread for high availability purposes. The scheduler tries to balance the service pods across all applicable nodes as evenly as possible.

If you need greater control over where the pods are scheduled, see [Using Node Affinity](#) and [Using Pod Affinity and Anti-affinity](#). These advanced scheduling features allow administrators to specify which node a pod can be scheduled on and to force or reject scheduling relative to other pods.

## 14.2.8. Sample Policy Configurations

The configuration below specifies the default scheduler configuration, if it were to be specified via the scheduler policy file.

```
kind: "Policy"
version: "v1"
predicates:
...
```

```

- name: "RegionZoneAffinity" ❶
  argument:
    serviceAffinity: ❷
      labels: ❸
        - "region"
        - "zone"
priorities:
...
- name: "RackSpread" ❹
  weight: 1
  argument:
    serviceAntiAffinity: ❺
      label: "rack" ❻

```

❶ The name for the predicate.

❷ The [type of predicate](#).

❸ The labels for the predicate.

❹ The name for the priority.

❺ The [type of priority](#).

❻ The labels for the priority.

In all of the sample configurations below, the list of predicates and priority functions is truncated to include only the ones that pertain to the use case specified. In practice, a complete/meaningful scheduler policy should include most, if not all, of the default predicates and priorities listed above.

The following example defines three topological levels, region (affinity) → zone (affinity) → rack (anti-affinity):

```

kind: "Policy"
version: "v1"
predicates:
...
- name: "RegionZoneAffinity"
  argument:
    serviceAffinity:
      labels:
        - "region"
        - "zone"
priorities:
...
- name: "RackSpread"
  weight: 1
  argument:
    serviceAntiAffinity:
      label: "rack"

```

The following example defines three topological levels, city (affinity) → building (anti-affinity) → room (anti-affinity):

```
kind: "Policy"
version: "v1"
predicates:
...
- name: "CityAffinity"
  argument:
    serviceAffinity:
      labels:
        - "city"
priorities:
...
- name: "BuildingSpread"
  weight: 1
  argument:
    serviceAntiAffinity:
      label: "building"
- name: "RoomSpread"
  weight: 1
  argument:
    serviceAntiAffinity:
      label: "room"
```

The following example defines a policy to only use nodes with the 'region' label defined and prefer nodes with the 'zone' label defined:

```
kind: "Policy"
version: "v1"
predicates:
...
- name: "RequireRegion"
  argument:
    labelsPresence:
      labels:
        - "region"
      presence: true
priorities:
...
- name: "ZonePreferred"
  weight: 1
  argument:
    labelPreference:
      label: "zone"
      presence: true
```

The following example combines both static and configurable predicates and also priorities:

```
kind: "Policy"
version: "v1"
predicates:
...
- name: "RegionAffinity"
  argument:
    serviceAffinity:
      labels:
        - "region"
```



```

- name: "RequireRegion"
  argument:
    labelsPresence:
      labels:
        - "region"
      presence: true
- name: "BuildingNodesAvoid"
  argument:
    labelsPresence:
      labels:
        - "building"
      presence: false
- name: "PodFitsPorts"
- name: "MatchNodeSelector"
priorities:
...
- name: "ZoneSpread"
  weight: 2
  argument:
    serviceAntiAffinity:
      label: "zone"
- name: "ZonePreferred"
  weight: 1
  argument:
    labelPreference:
      label: "zone"
      presence: true
- name: "ServiceSpreadingPriority"
  weight: 1

```

## 14.3. CUSTOM SCHEDULING

### 14.3.1. Overview

You can run multiple, custom schedulers alongside the default scheduler and configure which scheduler to use for each pods.

To schedule a given pod using a specific scheduler, [specify the name of the scheduler in that pod specification](#).

### 14.3.2. Deploying the Scheduler

The steps below are the general process for deploying a scheduler into your cluster.



#### NOTE

Information on how to create/deploy a scheduler is outside the scope of this document. For an example, see [plugin/pkg/scheduler](#) in the Kubernetes source directory.

1. Create or edit a pod configuration and specify the name of the scheduler with the **schedulerName** parameter. The name must be unique.

#### Sample pod specification with scheduler

■

```

apiVersion: v1
kind: Pod
metadata:
  name: custom-scheduler
  labels:
    name: multischeduler-example
spec:
  schedulerName: custom-scheduler 1
  containers:
  - name: pod-with-second-annotation-container
    image: docker.io/ocpqe/hello-pod

```

- 1** The name of the scheduler to use. When no scheduler name is supplied, the pod is automatically scheduled using the default scheduler.

2. Run the following command to create the pod:

```
$ oc create -f scheduler.yaml
```

3. Run the following command to check that the pod was created with the custom scheduler:

```
$ oc get pod custom-scheduler -o yaml
```

4. Run the following command to check the status of the pod:

```
$ oc get pod
```

The pod should not be running.

| NAME             | READY | STATUS  | RESTARTS | AGE |
|------------------|-------|---------|----------|-----|
| custom-scheduler | 0/1   | Pending | 0        | 2m  |

5. Deploy the custom scheduler.

6. Run the following command to check the status of the pod:

```
$ oc get pod
```

The pod should be running.

| NAME             | READY | STATUS  | RESTARTS | AGE |
|------------------|-------|---------|----------|-----|
| custom-scheduler | 1/1   | Running | 0        | 4m  |

7. Run the following command to check that the scheduler was used:

```
$ oc describe pod custom-scheduler
```

The name of the scheduler is listed, as shown in the following truncated output:

```

[...]
Events:
  FirstSeen    LastSeen    Count    From                                     SubObjectPath    Type

```

```
Reason Message
-----
1m          1m          1          my-scheduler          Normal
Scheduled Successfully assigned custom-scheduler to <$node1>
[...]
```

## 14.4. CONTROLLING POD PLACEMENT

### 14.4.1. Overview

As a cluster administrator, you can set a policy to prevent application developers with certain roles from targeting specific nodes when scheduling pods.

The Pod Node Constraints admission controller ensures that pods are deployed onto only specified node hosts using labels] and prevents users without a specific role from using the **nodeSelector** field to schedule pods.

### 14.4.2. Constraining Pod Placement Using Node Name

Use the Pod Node Constraints admission controller to ensure a pod is deployed onto only a specified node host by assigning it a label and specifying this in the **nodeName** setting in a pod configuration.

1. Ensure you have the desired labels (see [Updating Labels on Nodes](#) for details) and **node selector** set up in your environment.

For example, make sure that your pod configuration features the **nodeName** value indicating the desired label:

```
apiVersion: v1
kind: Pod
spec:
  nodeName: <value>
```

2. Modify the master configuration file, **/etc/origin/master/master-config.yaml**, to add **PodNodeConstraints** to the **admissionConfig** section:

```
...
admissionConfig:
  pluginConfig:
    PodNodeConstraints:
      configuration:
        apiversion: v1
        kind: PodNodeConstraintsConfig
...
```

3. Restart OpenShift Container Platform for the changes to take effect.

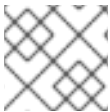
```
# systemctl restart atomic-openshift-master
```

### 14.4.3. Constraining Pod Placement Using a Node Selector

Using [node selectors](#), you can ensure that pods are only placed onto nodes with specific labels. As a cluster administrator, you can use the Pod Node Constraints admission controller to set a policy that prevents users without the **pods/binding** permission from using node selectors to schedule pods.

The **nodeSelectorLabelBlacklist** field of a master configuration file gives you control over the labels that certain roles can specify in a pod configuration's **nodeSelector** field. Users, service accounts, and groups that have the **pods/binding** permission [role](#) can specify any node selector. Those without the **pods/binding** permission are prohibited from setting a **nodeSelector** for any label that appears in **nodeSelectorLabelBlacklist**.

For example, an OpenShift Container Platform cluster might consist of five data centers spread across two regions. In the U.S., **us-east**, **us-central**, and **us-west**; and in the Asia-Pacific region (APAC), **apac-east** and **apac-west**. Each node in each geographical region is labeled accordingly. For example, **region: us-east**.



## NOTE

See [Updating Labels on Nodes](#) for details on assigning labels.

As a cluster administrator, you can create an infrastructure where application developers should be deploying pods only onto the nodes closest to their geographical location. You can create a node selector, grouping the U.S. data centers into **superregion: us** and the APAC data centers into **superregion: apac**.

To maintain an even loading of resources per data center, you can add the desired **region** to the **nodeSelectorLabelBlacklist** section of a master configuration. Then, whenever a developer located in the U.S. creates a pod, it is deployed onto a node in one of the regions with the **superregion: us** label. If the developer tries to target a specific region for their pod (for example, **region: us-east**), they receive an error. If they try again, without the node selector on their pod, it can still be deployed onto the region they tried to target, because **superregion: us** is set as the project-level node selector, and nodes labeled **region: us-east** are also labeled **superregion: us**.

1. Ensure you have the desired labels (see [Updating Labels on Nodes](#) for details) and [node selector](#) set up in your environment.

For example, make sure that your pod configuration features the **nodeSelector** value indicating the desired label:

```
apiVersion: v1
kind: Pod
spec:
  nodeSelector:
    <key>: <value>
  ...
```

2. Modify the master configuration file, **/etc/origin/master/master-config.yaml**, to add **nodeSelectorLabelBlacklist** to the **admissionConfig** section with the labels that are assigned to the node hosts you want to deny pod placement:

```
...
admissionConfig:
  pluginConfig:
    PodNodeConstraints:
      configuration:
        apiversion: v1
```

```

kind: PodNodeConstraintsConfig
nodeSelectorLabelBlacklist:
  - kubernetes.io/hostname
  - <label>
...

```

3. Restart OpenShift Container Platform for the changes to take effect.

```
# systemctl restart atomic-openshift-master
```

#### 14.4.4. Control Pod Placement to Projects

The Pod Node Selector admission controller allows you to force pods onto nodes associated with a specific project and prevent pods from being scheduled in those nodes.

The Pod Node Selector admission controller determines where a pod can be placed using [labels on projects](#) and node selectors specified in pods. A new pod will be placed on a node associated with a project only if the node selectors in the pod match the labels in the project.

After the pod is created, the node selectors are merged into the pod so that the pod specification includes the labels originally included in the specification and any new labels from the node selectors. The example below illustrates the merging effect.

The Pod Node Selector admission controller also allows you to create a list of labels that are permitted in a specific project. This list acts as a *whitelist* that lets developers know what labels are acceptable to use in a project and gives administrators greater control over labeling in a cluster.

To activate the **Pod Node Selector** admission controller:

1. Configure the **Pod Node Selector** admission controller and whitelist, using one of the following methods:

- Add the following to the master configuration file, */etc/origin/master/master-config.yaml*:

```

admissionConfig:
  pluginConfig:
    PodNodeSelector:
      configuration:
        podNodeSelectorPluginConfig: ❶
        clusterDefaultNodeSelector: "k3=v3" ❷
        ns1: region=west,env=test,infra=fedora,os=fedora ❸

```

- ❶ Adds the **Pod Node Selector** admission controller plug-in.
- ❷ Creates default labels for all nodes.
- ❸ Creates a whitelist of permitted labels in the specified project. Here, the project is **ns1** and the labels are the **key=value** pairs that follow.

- Create a file containing the admission controller information:

```

podNodeSelectorPluginConfig:
  clusterDefaultNodeSelector: "k3=v3"
  ns1: region=west,env=test,infra=fedora,os=fedora

```

Then, reference the file in the master configuration:

```
admissionConfig:
  pluginConfig:
    PodNodeSelector:
      location: <path-to-file>
```



#### NOTE

If a project does not have node selectors specified, the pods associated with that project will be merged using the default node selector (**clusterDefaultNodeSelector**).

- Restart OpenShift Container Platform for the changes to take effect.

```
# systemctl restart atomic-openshift-master
```

- Create a project object that includes the **scheduler.alpha.kubernetes.io/node-selector** annotation and labels.

```
apiVersion: v1
kind: Namespace
metadata:
  name: ns1
  annotations:
    scheduler.alpha.kubernetes.io/node-selector:
      env=test,infra=fedora ❶
spec: {},
status: {}
```

- Annotation to create the labels to match the project label selector. Here, the key/value labels are **env=test** and **infra=fedora**.



#### NOTE

When using the **Pod Node Selector** admission controller, you cannot use **oc adm new-project <project-name>** for setting project node selector. When you set the project node selector using the **oc adm new-project myproject --node-selector='type=user-node,region=<region>'** command, OpenShift Container Platform sets the **openshift.io/node-selector** annotation, which is processed by **NodeEnv** admission plugin.

- Create a pod specification that includes the labels in the node selector, for example:

```
apiVersion: v1
kind: Pod
metadata:
  labels:
    name: hello-pod
spec:
  containers:
```

```

- image: "docker.io/ocpqe/hello-pod:latest"
  imagePullPolicy: IfNotPresent
  name: hello-pod
  ports:
    - containerPort: 8080
      protocol: TCP
  resources: {}
  securityContext:
    capabilities: {}
    privileged: false
  terminationMessagePath: /dev/termination-log
  dnsPolicy: ClusterFirst
  restartPolicy: Always
  nodeSelector: ❶
    env: test
    os: fedora
  serviceAccount: ""
  status: {}

```

- ❶ Node selectors to match project labels.

5. Create the pod in the project:

```
# oc create -f pod.yaml --namespace=ns1
```

6. Check that the node selector labels were added to the pod configuration:

```

get pod pod1 --namespace=ns1 -o json

nodeSelector": {
  "env": "test",
  "infra": "fedora",
  "os": "fedora"
}

```

The node selectors are merged into the pod and the pod should be scheduled in the appropriate project.

If you create a pod with a label that is not specified in the project specification, the pod is not scheduled on the node.

For example, here the label **env: production** is not in any project specification:

```

nodeSelector:
  "env: production"
  "infra": "fedora",
  "os": "fedora"

```

If there is a node that does not have a node selector annotation, the pod will be scheduled there.

## 14.5. ADVANCED SCHEDULING

### 14.5.1. Overview

Advanced scheduling involves configuring a pod so that the pod is required to run on particular nodes or has a preference to run on particular nodes.

Generally, advanced scheduling is not necessary, as the OpenShift Container Platform automatically places pods in a reasonable manner. For example, the default scheduler attempts to distribute pods across the nodes evenly and considers the available resources in a node. However, you might want more control over where a pod is placed.

If a pod needs to be on a machine with a faster disk speed (or prevented from being placed on that machine) or pods from two different services need to be located so they can communicate, you can use advanced scheduling to make that happen.

To ensure that appropriate new pods are scheduled on a dedicated group of nodes and prevent other new pods from being scheduled on those nodes, you can combine these methods as needed.

### 14.5.2. Using Advanced Scheduling

There are several ways to invoke advanced scheduling in your cluster:

#### Pod Affinity and Anti-affinity

Pod affinity allows a **pod** to specify an affinity (or anti-affinity) towards a group of **pods** (for an application's latency requirements, due to security, and so forth) it can be placed with. The node does not have control over the placement.

Pod affinity uses labels on nodes and label selectors on pods to create rules for pod placement. Rules can be mandatory (required) or best-effort (preferred).

See [Using Pod Affinity and Anti-affinity](#).

#### Node Affinity

Node affinity allows a **pod** to specify an affinity (or anti-affinity) towards a group of **nodes** (due to their special hardware, location, requirements for high availability, and so forth) it can be placed on. The node does not have control over the placement.

Node affinity uses labels on nodes and label selectors on pods to create rules for pod placement. Rules can be mandatory (required) or best-effort (preferred).

See [Using Node Affinity](#).

#### Node Selectors

Node selectors are the simplest form of advanced scheduling. Like node affinity, node selectors also use labels on nodes and label selectors on pods to allow a **pod** to control the **nodes** on which it can be placed. However, node selectors do not have required and preferred rules that node affinities have.

See [Using Node Selectors](#).

#### Taints and Tolerations

Taints/Tolerations allow the **node** to control which **pods** should (or should not) be scheduled on them. Taints are labels on a node and tolerations are labels on a pod. The labels on the pod must match (or tolerate) the label (taint) on the node in order to be scheduled.

Taints/tolerations have one advantage over affinities. For example, if you add to a cluster a new group of nodes with different labels, you would need to update affinities on each of the pods you want to access the node and on any other pods you do not want to use the new nodes. With taints/tolerations, you would only need to update those pods that are required to land on those new nodes, because other pods would be repelled.



See [Using Taints and Tolerations](#).

## 14.6. ADVANCED SCHEDULING AND NODE AFFINITY

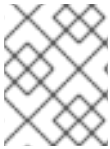
### 14.6.1. Overview

*Node affinity* is a set of rules used by the scheduler to determine where a pod can be placed. The rules are defined using custom [labels on nodes](#) and label selectors specified in pods. Node affinity allows a **pod** to specify an affinity (or anti-affinity) towards a group of **nodes** it can be placed on. The node does not have control over the placement.

For example, you could configure a pod to only run on a node with a specific CPU or in a specific availability zone.

There are two types of node affinity rules: *required* and *preferred*.

Required rules **must** be met before a pod can be scheduled on a node. Preferred rules specify that, if the rule is met, the scheduler tries to enforce the rules, but does not guarantee enforcement.



#### NOTE

If labels on a node change at runtime that results in an node affinity rule on a pod no longer being met, the pod continues to run on the node.

### 14.6.2. Configuring Node Affinity

You configure node affinity through the pod specification file. You can specify a [required rule](#), a [preferred rule](#), or both. If you specify both, the node must first meet the required rule, then attempts to meet the preferred rule.

The following example is a pod specification with a rule that requires the pod be placed on a node with a label whose key is **e2e-az-NorthSouth** and whose value is either **e2e-az-North** or **e2e-az-South**:

#### Sample pod configuration file with a node affinity required rule

```
apiVersion: v1
kind: Pod
metadata:
  name: with-node-affinity
spec:
  affinity:
    nodeAffinity: ❶
      requiredDuringSchedulingIgnoredDuringExecution: ❷
        nodeSelectorTerms:
          - matchExpressions:
              - key: e2e-az-NorthSouth ❸
                operator: In ❹
                values:
                  - e2e-az-North ❺
                  - e2e-az-South ❻
  containers:
    - name: with-node-affinity
```

```
image: docker.io/ocpqe/hello-pod
```

- 1 The stanza to configure node affinity.
- 2 Defines a required rule.
- 3 5 6 The key/value pair (label) that must be matched to apply the rule.
- 4 The [operator](#) represents the relationship between the label on the node and the set of values in the **matchExpression** parameters in the pod specification. This value can be **In**, **NotIn**, **Exists**, or **DoesNotExist**, **Lt**, or **Gt**.

The following example is a node specification with a preferred rule that a node with a label whose key is **e2e-az-EastWest** and whose value is either **e2e-az-East** or **e2e-az-West** is preferred for the pod:

### Sample pod configuration file with a node affinity preferred rule

```
apiVersion: v1
kind: Pod
metadata:
  name: with-node-affinity
spec:
  affinity:
    nodeAffinity: 1
      preferredDuringSchedulingIgnoredDuringExecution: 2
        - weight: 1 3
          preference:
            matchExpressions:
              - key: e2e-az-EastWest 4
                operator: In 5
                values:
                  - e2e-az-East 6
                  - e2e-az-West 7
  containers:
    - name: with-node-affinity
      image: docker.io/ocpqe/hello-pod
```

- 1 The stanza to configure node affinity.
- 2 Defines a preferred rule.
- 3 Specifies a weight for a preferred rule. The node with highest weight is preferred.
- 4 6 7 The key/value pair (label) that must be matched to apply the rule.
- 5 The operator represents the relationship between the label on the node and the set of values in the **matchExpression** parameters in the pod specification. This value can be **In**, **NotIn**, **Exists**, or **DoesNotExist**, **Lt**, or **Gt**.

There is no explicit *node anti-affinity* concept, but using the **NotIn** or **DoesNotExist** operator replicates that behavior.



## NOTE

If you are using node affinity and [node selectors](#) in the same pod configuration, note the following:

- If you configure both **nodeSelector** and **nodeAffinity**, both conditions must be satisfied for the pod to be scheduled onto a candidate node.
- If you specify multiple **nodeSelectorTerms** associated with **nodeAffinity** types, then the pod can be scheduled onto a node if one of the **nodeSelectorTerms** is satisfied.
- If you specify multiple **matchExpressions** associated with **nodeSelectorTerms**, then the pod can be scheduled onto a node only if all **matchExpressions** are satisfied.

### 14.6.2.1. Configuring a Required Node Affinity Rule

Required rules **must** be met before a pod can be scheduled on a node.

The following steps demonstrate a simple configuration that creates a node and a pod that the scheduler is required to place on the node.

1. Add a label to a node by editing the node configuration or by using the **oc label node** command:

```
$ oc label node node1 e2e-az-name=e2e-az1
```

2. In the pod specification, use the **nodeAffinity** stanza to configure the **requiredDuringSchedulingIgnoredDuringExecution** parameter:
  - a. Specify the key and values that must be met. If you want the new pod to be scheduled on the node you edited, use the same **key** and **value** parameters as the label in the node.
  - b. Specify an **operator**. The [operator can be In, NotIn, Exists, DoesNotExist, Lt, or Gt](#). For example, use the operator **In** to require the label to be in the node:

```
spec:
  affinity:
    nodeAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        nodeSelectorTerms:
          - matchExpressions:
              - key: e2e-az-name
                operator: In
              values:
                - e2e-az1
                - e2e-az2
```

3. Create the pod:

```
$ oc create -f e2e-az2.yaml
```

### 14.6.2.2. Configuring a Preferred Node Affinity Rule

Preferred rules specify that, if the rule is met, the scheduler tries to enforce the rules, but does not guarantee enforcement.

The following steps demonstrate a simple configuration that creates a node and a pod that the scheduler tries to place on the node.

1. Add a label to a node by editing the node configuration or by executing the **oc label node** command:

```
$ oc label node node1 e2e-az-name=e2e-az3
```

2. In the pod specification, use the **nodeAffinity** stanza to configure the **preferredDuringSchedulingIgnoredDuringExecution** parameter:
  - a. Specify a weight for the node, as a number 1-100. The node with highest weight is preferred.
  - b. Specify the key and values that must be met. If you want the new pod to be scheduled on the node you edited, use the same **key** and **value** parameters as the label in the node:

```
preferredDuringSchedulingIgnoredDuringExecution:
- weight: 1
  preference:
    matchExpressions:
    - key: e2e-az-name
      operator: In
      values:
      - e2e-az3
```

3. Specify an **operator**. The [operator can be In, NotIn, Exists, DoesNotExist, Lt, or Gt](#). For example, use the operator **In** to require the label to be in the node.
4. Create the pod.

```
$ oc create -f e2e-az3.yaml
```

### 14.6.3. Examples

The following examples demonstrate node affinity.

#### 14.6.3.1. Node Affinity with Matching Labels

The following example demonstrates node affinity for a node and pod with matching labels:

- The **Node1** node has the label **zone:us**:

```
$ oc label node node1 zone=us
```

- The pod **pod-s1** has the **zone** and **us** key/value pair under a required node affinity rule:

```
$ cat pod-s1.yaml
apiVersion: v1
kind: Pod
metadata:
```

```

    name: pod-s1
  spec:
    containers:
      - image: "docker.io/ocpqe/hello-pod"
        name: hello-pod
    affinity:
      nodeAffinity:
        requiredDuringSchedulingIgnoredDuringExecution:
          nodeSelectorTerms:
            - matchExpressions:
                - key: "zone"
                  operator: In
                  values:
                    - us

```

- Create the pod using the standard command:

```

$ oc create -f pod-s1.yaml
pod "pod-s1" created

```

- The pod **pod-s1** can be scheduled on **Node1**:

```

oc get pod -o wide

```

| NAME   | READY | STATUS  | RESTARTS | AGE | IP  | NODE  |
|--------|-------|---------|----------|-----|-----|-------|
| pod-s1 | 1/1   | Running | 0        | 4m  | IP1 | node1 |

### 14.6.3.2. Node Affinity with No Matching Labels

The following example demonstrates node affinity for a node and pod without matching labels:

- The **Node1** node has the label **zone: emea**:

```

$ oc label node node1 zone=emea

```

- The pod **pod-s1** has the **zone** and **us** key/value pair under a required node affinity rule:

```

$ cat pod-s1.yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-s1
spec:
  containers:
    - image: "docker.io/ocpqe/hello-pod"
      name: hello-pod
  affinity:
    nodeAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        nodeSelectorTerms:
          - matchExpressions:
              - key: "zone"
                operator: In
                values:
                  - us

```

- The pod **pod-s1** cannot be scheduled on **Node1**:

```
oc describe pod pod-s1
<---snip--->
Events:
  FirstSeen    LastSeen  Count    From                                      SubObjectPath  Type
Reason
-----
1m            33s       8        default-scheduler Warning
FailedScheduling    No nodes are available that match all of the
following predicates: MatchNodeSelector (1).
```

## 14.7. ADVANCED SCHEDULING AND POD AFFINITY AND ANTI-AFFINITY

### 14.7.1. Overview

*Pod affinity* and *pod anti-affinity* allow you to specify rules about how pods should be placed relative to other pods. The rules are defined using custom [labels on nodes](#) and label selectors specified in pods. Pod affinity/anti-affinity allows a **pod** to specify an affinity (or anti-affinity) towards a group of **pods** it can be placed with. The node does not have control over the placement.

For example, using affinity rules, you could spread or pack pods within a service or relative to pods in other services. Anti-affinity rules allow you to prevent pods of a particular service from scheduling on the same nodes as pods of another service that are known to interfere with the performance of the pods of the first service. Or, you could spread the pods of a service across nodes or availability zones to reduce correlated failures.

Pod affinity/anti-affinity allows you to constrain which nodes your pod is eligible to be scheduled on based on the labels on other pods. A [label](#) is a key/value pair.

- Pod affinity can tell the scheduler to locate a new pod on the same node as other pods if the label selector on the new pod matches the label on the current pod.
- Pod anti-affinity can prevent the scheduler from locating a new pod on the same node as pods with the same labels if the label selector on the new pod matches the label on the current pod.

There are two types of pod affinity rules: *required* and *preferred*.

Required rules **must** be met before a pod can be scheduled on a node. Preferred rules specify that, if the rule is met, the scheduler tries to enforce the rules, but does not guarantee enforcement.

### 14.7.2. Configuring Pod Affinity and Anti-affinity

You configure pod affinity/anti-affinity through the pod specification files. You can specify a [required rule](#), a [preferred rule](#), or both. If you specify both, the node must first meet the required rule, then attempts to meet the preferred rule.

The following example shows a pod specification configured for pod affinity and anti-affinity.

In this example, the pod affinity rule indicates that the pod can schedule onto a node only if that node has at least one already-running pod with a label that has the key **security** and value **S1**. The pod anti-affinity rule says that the pod prefers to not schedule onto a node if that node is already running a pod

with label having key **security** and value **S2**.

### Sample pod config file with pod affinity

```
apiVersion: v1
kind: Pod
metadata:
  name: with-pod-affinity
spec:
  affinity:
    podAffinity: ❶
      requiredDuringSchedulingIgnoredDuringExecution: ❷
        - labelSelector:
            matchExpressions:
              - key: security ❸
                operator: In ❹
                values:
                  - S1 ❺
            topologyKey: failure-domain.beta.kubernetes.io/zone
  containers:
    - name: with-pod-affinity
      image: docker.io/ocpqe/hello-pod
```

❶ Stanza to configure pod affinity.

❷ Defines a required rule.

❸ ❺ The key and value (label) that must be matched to apply the rule.

❹ The **operator** represents the relationship between the label on the existing pod and the set of values in the **matchExpression** parameters in the specification for the new pod. Can be **In**, **NotIn**, **Exists**, or **DoesNotExist**.

### Sample pod config file with pod anti-affinity

```
apiVersion: v1
kind: Pod
metadata:
  name: with-pod-antiaffinity
spec:
  affinity:
    podAntiAffinity: ❶
      preferredDuringSchedulingIgnoredDuringExecution: ❷
        - weight: 100 ❸
          podAffinityTerm:
            labelSelector:
              matchExpressions:
                - key: security ❹
                  operator: In ❺
                  values:
                    - S2
            topologyKey: kubernetes.io/hostname
```

```
containers:
- name: with-pod-affinity
  image: docker.io/ocpqe/hello-pod
```

- 1 Stanza to configure pod anti-affinity.
- 2 Defines a preferred rule.
- 3 Specifies a weight for a preferred rule. The node with the highest weight is preferred.
- 4 Description of the pod label that determines when the anti-affinity rule applies. Specify a key and value for the label.
- 5 The operator represents the relationship between the label on the existing pod and the set of values in the **matchExpression** parameters in the specification for the new pod. Can be **In**, **NotIn**, **Exists**, or **DoesNotExist**.



## NOTE

If labels on a node change at runtime such that the affinity rules on a pod are no longer met, the pod continues to run on the node.

### 14.7.2.1. Configuring an Affinity Rule

The following steps demonstrate a simple two-pod configuration that creates pod with a label and a pod that uses affinity to allow scheduling with that pod.

1. Create a pod with a specific label in the pod specification:

```
$ cat team4.yaml
apiVersion: v1
kind: Pod
metadata:
  name: security-s1
  labels:
    security: S1
spec:
  containers:
  - name: security-s1
    image: docker.io/ocpqe/hello-pod
```

2. When creating other pods, edit the pod specification as follows:
  - a. Use the **podAffinity** stanza to configure the **requiredDuringSchedulingIgnoredDuringExecution** parameter or **preferredDuringSchedulingIgnoredDuringExecution** parameter:
  - b. Specify the key and value that must be met. If you want the new pod to be scheduled with the other pod, use the same **key** and **value** parameters as the label on the first pod.

```
podAffinity:
  requiredDuringSchedulingIgnoredDuringExecution:
  - labelSelector:
      matchExpressions:
```



```

- key: security
  operator: In
  values:
  - S1
topologyKey: failure-domain.beta.kubernetes.io/zone

```

- c. Specify an **operator**. The [operator](#) can be **In**, **NotIn**, **Exists**, or **DoesNotExist**. For example, use the operator **In** to require the label to be in the node.
  - d. Specify a **topologyKey**, which is a prepopulated [Kubernetes label](#) that the system uses to denote such a topology domain.
3. Create the pod.

```
$ oc create -f <pod-spec>.yaml
```

### 14.7.2.2. Configuring an Anti-affinity Rule

The following steps demonstrate a simple two-pod configuration that creates pod with a label and a pod that uses an anti-affinity preferred rule to attempt to prevent scheduling with that pod.

1. Create a pod with a specific label in the pod specification:

```

$ cat team4.yaml
apiVersion: v1
kind: Pod
metadata:
  name: security-s2
  labels:
    security: S2
spec:
  containers:
  - name: security-s2
    image: docker.io/ocpqe/hello-pod

```

2. When creating other pods, edit the pod specification to set the following parameters:
3. Use the **podAffinity** stanza to configure the **requiredDuringSchedulingIgnoredDuringExecution** parameter or **preferredDuringSchedulingIgnoredDuringExecution** parameter:
  - a. Specify a weight for the node, 1-100. The node that with highest weight is preferred.
  - b. Specify the key and values that must be met. If you want the new pod to not be scheduled with the other pod, use the same **key** and **value** parameters as the label on the first pod.

```

podAntiAffinity:
  preferredDuringSchedulingIgnoredDuringExecution:
  - weight: 100
    podAffinityTerm:
      labelSelector:
        matchExpressions:
        - key: security
          operator: In

```

```

      values:
      - S2
    topologyKey: kubernetes.io/hostname

```

- c. For a preferred rule, specify a weight, 1-100.
- d. Specify an **operator**. The **operator** can be **In**, **NotIn**, **Exists**, or **DoesNotExist**. For example, use the operator **In** to require the label to be in the node.
4. Specify a **topologyKey**, which is a prepopulated [Kubernetes label](#) that the system uses to denote such a topology domain.
5. Create the pod.

```
$ oc create -f <pod-spec>.yaml
```

### 14.7.3. Examples

The following examples demonstrate pod affinity and pod anti-affinity.

#### 14.7.3.1. Pod Affinity

The following example demonstrates pod affinity for pods with matching labels and label selectors.

- The pod **team4** has the label **team: 4**.

```

$ cat team4.yaml
apiVersion: v1
kind: Pod
metadata:
  name: team4
  labels:
    team: "4"
spec:
  containers:
  - name: ocp
    image: docker.io/ocpqe/hello-pod

```

- The pod **team4a** has the label selector **team: 4** under **podAffinity**.

```

$ cat pod-team4a.yaml
apiVersion: v1
kind: Pod
metadata:
  name: team4a
spec:
  affinity:
    podAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
      - labelSelector:
          matchExpressions:
          - key: team
            operator: In
            values:

```

```

      - "4"
      topologyKey: kubernetes.io/hostname
    containers:
      - name: pod-affinity
        image: docker.io/ocpqe/hello-pod

```

- The **team4a** pod is scheduled on the same node as the **team4** pod.

### 14.7.3.2. Pod Anti-affinity

The following example demonstrates pod anti-affinity for pods with matching labels and label selectors.

- The pod **pod-s1** has the label **security:s1**.

```

cat pod-s1.yaml
apiVersion: v1
kind: Pod
metadata:
  name: s1
  labels:
    security: s1
spec:
  containers:
    - name: ocp
      image: docker.io/ocpqe/hello-pod

```

- The pod **pod-s2** has the label selector **security:s1** under **podAntiAffinity**.

```

cat pod-s2.yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-s2
spec:
  affinity:
    podAntiAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        - labelSelector:
            matchExpressions:
              - key: security
                operator: In
                values:
                  - s1
            topologyKey: kubernetes.io/hostname
  containers:
    - name: pod-antiaffinity
      image: docker.io/ocpqe/hello-pod

```

- The pod **pod-s2** is not scheduled unless there is a node with a pod that has the **security:s2** label. If there is no other pod with that label, the new pod remains in a pending state:

| NAME   | READY | STATUS  | RESTARTS | AGE | IP     | NODE |
|--------|-------|---------|----------|-----|--------|------|
| pod-s2 | 0/1   | Pending | 0        | 32s | <none> |      |

### 14.7.3.3. Pod Affinity with no Matching Labels

The following example demonstrates pod affinity for pods without matching labels and label selectors.

- The pod **pod-s1** has the label **security:s1**.

```
$ cat pod-s1.yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-s1
  labels:
    security: s1
spec:
  containers:
  - name: ocp
    image: docker.io/ocpqe/hello-pod
```

- The pod **pod-s2** has the label selector **security:s2**.

```
$ cat pod-s2.yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-s2
spec:
  affinity:
    podAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
      - labelSelector:
          matchExpressions:
          - key: security
            operator: In
            values:
            - s2
        topologyKey: kubernetes.io/hostname
  containers:
  - name: pod-affinity
    image: docker.io/ocpqe/hello-pod
```

- The pod **pod-s2** cannot be scheduled on the same node as **pod-s1**.

## 14.8. ADVANCED SCHEDULING AND NODE SELECTORS

### 14.8.1. Overview

A *node selector* specifies a map of key-value pairs. The rules are defined using custom [labels on nodes](#) and selectors specified in pods.

For the pod to be eligible to run on a node, the pod must have the indicated key-value pairs as the label on the node.

If you are using node affinity and [node selectors](#) in the same pod configuration, see the [important considerations](#) below.

## 14.8.2. Configuring Node Selectors

Using **nodeSelector** in a pod configuration, you can ensure that pods are only placed onto nodes with specific labels.

1. Ensure you have the desired labels (see [Updating Labels on Nodes](#) for details) and **node selector** set up in your environment.

For example, make sure that your pod configuration features the **nodeSelector** value indicating the desired label:

```
apiVersion: v1
kind: Pod
spec:
  nodeSelector:
    <key>: <value>
  ...
```

2. Modify the master configuration file, */etc/origin/master/master-config.yaml*, to add **nodeSelectorLabelBlacklist** to the **admissionConfig** section with the labels that are assigned to the node hosts you want to deny pod placement:

```
...
admissionConfig:
  pluginConfig:
    PodNodeConstraints:
      configuration:
        apiVersion: v1
        kind: PodNodeConstraintsConfig
        nodeSelectorLabelBlacklist:
          - kubernetes.io/hostname
          - <label>
  ...
```

3. Restart OpenShift Container Platform for the changes to take effect.

```
# systemctl restart atomic-openshift-master
```

### NOTE

If you are using node selectors and **node affinity** in the same pod configuration, note the following:

- If you configure both **nodeSelector** and **nodeAffinity**, both conditions must be satisfied for the pod to be scheduled onto a candidate node.
- If you specify multiple **nodeSelectorTerms** associated with **nodeAffinity** types, then the pod can be scheduled onto a node if one of the **nodeSelectorTerms** is satisfied.
- If you specify multiple **matchExpressions** associated with **nodeSelectorTerms**, then the pod can be scheduled onto a node only if all **matchExpressions** are satisfied.

## 14.9. ADVANCED SCHEDULING AND TAINTS AND TOLERATIONS

### 14.9.1. Overview

Taints and tolerations allow the **node** to control which **pods** should (or should not) be scheduled on them.

### 14.9.2. Taints and Tolerations

A *taint* allows a node to refuse pod to be scheduled unless that pod has a matching *toleration*.

You apply taints to a node through the node specification (**NodeSpec**) and apply tolerations to a pod through the pod specification (**PodSpec**). A taint on a node instructs the node to repel all pods that do not tolerate the taint.

Taints and tolerations consist of a key, value, and effect. An operator allows you to leave one of these parameters empty.

**Table 14.1. Taint and toleration components**

| Parameter               | Description  |                   |   |                         |  |                  |   |
|-------------------------|--|-------------------|---|-------------------------|--|------------------|---|
| <b>key</b>              | The <b>key</b> is any string, up to 253 characters. The key must begin with a letter or number, and may contain letters, numbers, hyphens, dots, and underscores.  |                   |   |                         |  |                  |   |
| <b>value</b>            | The <b>value</b> is any string, up to 63 characters. The value must begin with a letter or number, and may contain letters, numbers, hyphens, dots, and underscores.   |                   |   |                         |  |                  |   |
| <b>effect</b>           | <p>The effect is one of the following:</p> <table> <tr> <td><b>NoSchedule</b></td><td> <ul style="list-style-type: none"> <li>New pods that do not match the taint are not scheduled onto that node.</li> <li>Existing pods on the node remain.</li> </ul> </td></tr> <tr> <td><b>PreferNoSchedule</b></td><td> <ul style="list-style-type: none"> <li>New pods that do not match the taint might be scheduled onto that node, but the scheduler tries not to.</li> <li>Existing pods on the node remain.</li> </ul> </td></tr> <tr> <td><b>NoExecute</b></td><td> <ul style="list-style-type: none"> <li>New pods that do not match the taint cannot be scheduled onto that node.</li> <li>Existing pods on the node that do not have a matching toleration are removed.</li> </ul> </td></tr> </table> | <b>NoSchedule</b> | <ul style="list-style-type: none"> <li>New pods that do not match the taint are not scheduled onto that node.</li> <li>Existing pods on the node remain.</li> </ul> | <b>PreferNoSchedule</b> | <ul style="list-style-type: none"> <li>New pods that do not match the taint might be scheduled onto that node, but the scheduler tries not to.</li> <li>Existing pods on the node remain.</li> </ul> | <b>NoExecute</b> | <ul style="list-style-type: none"> <li>New pods that do not match the taint cannot be scheduled onto that node.</li> <li>Existing pods on the node that do not have a matching toleration are removed.</li> </ul> |
| <b>NoSchedule</b>       | <ul style="list-style-type: none"> <li>New pods that do not match the taint are not scheduled onto that node.</li> <li>Existing pods on the node remain.</li> </ul>  |                   |   |                         |  |                  |   |
| <b>PreferNoSchedule</b> | <ul style="list-style-type: none"> <li>New pods that do not match the taint might be scheduled onto that node, but the scheduler tries not to.</li> <li>Existing pods on the node remain.</li> </ul>   |                   |   |                         |  |                  |   |
| <b>NoExecute</b>        | <ul style="list-style-type: none"> <li>New pods that do not match the taint cannot be scheduled onto that node.</li> <li>Existing pods on the node that do not have a matching toleration are removed.</li> </ul>  |                   |   |                         |  |                  |   |

| Parameter       | Description   |  |
|-----------------|---------------|--|
| <b>operator</b> | <b>Equal</b>  | The <b>key/value/effect</b> parameters must match. This is the default.  |
|                 | <b>Exists</b> | The <b>key/effect</b> parameters must match. You must leave a blank <b>value</b> parameter, which matches any. |

A toleration matches a taint:

- If the **operator** parameter is set to **Equal**:
  - the **key** parameters are the same;
  - the **value** parameters are the same;
  - the **effect** parameters are the same.
- If the **operator** parameter is set to **Exists**:
  - the **key** parameters are the same;
  - the **effect** parameters are the same.

#### 14.9.2.1. Using Multiple Taints

You can put multiple taints on the same node and multiple tolerations on the same pod. OpenShift Container Platform processes multiple taints and tolerations as follows:

1. Process the taints for which the pod has a matching toleration.
2. The remaining unmatched taints have the indicated effects on the pod:
  - If there is at least one unmatched taint with effect **NoSchedule**, OpenShift Container Platform cannot schedule a pod onto that node.
  - If there is no unmatched taint with effect **NoSchedule** but there is at least one unmatched taint with effect **PreferNoSchedule**, OpenShift Container Platform tries to not schedule the pod onto the node.
  - If there is at least one unmatched taint with effect **NoExecute**, OpenShift Container Platform evicts the pod from the node (if it is already running on the node), or the pod is not scheduled onto the node (if it is not yet running on the node).
    - Pods that do not tolerate the taint are evicted immediately.
    - Pods that tolerate the taint without specifying **tolerationSeconds** in their toleration specification remain bound forever.

- Pods that tolerate the taint with a specified **tolerationSeconds** remain bound for the specified amount of time.

For example:

- The node has the following taints:

```
$ oc adm taint nodes node1 key1=value1:NoSchedule
$ oc adm taint nodes node1 key1=value1:NoExecute
$ oc adm taint nodes node1 key2=value2:NoSchedule
```

- The pod has the following tolerations:

```
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoSchedule"
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoExecute"
```

In this case, the pod cannot be scheduled onto the node, because there is no toleration matching the third taint. The pod continues running if it is already running on the node when the taint is added, because the third taint is the only one of the three that is not tolerated by the pod.

### 14.9.3. Adding a Taint to an Existing Node

You add a taint to a node using the **oc adm taint** command with the parameters described in the [Taint and toleration components](#) table:

```
$ oc adm taint nodes <node-name> <key>=<value>:<effect>
```

For example:

```
$ oc adm taint nodes node1 key1=value1:NoSchedule
```

The example places a taint on **node1** that has key **key1**, value **value1**, and taint effect **NoSchedule**.

### 14.9.4. Adding a Toleration to a Pod

To add a toleration to a pod, edit the pod specification to include a **tolerations** section:

#### Sample pod configuration file with Equal operator

```
tolerations:
- key: "key1" 1
  operator: "Equal" 2
  value: "value1" 3
  effect: "NoExecute" 4
  tolerationSeconds: 3600 5
```



- 1 2 3 4 The toleration parameters, as described in the [Taint and toleration components](#) table.
- 5 The **tolerationSeconds** parameter specifies how long a pod can remain bound to a node before being evicted. See [Using Toleration Seconds to Delay Pod Evictions](#) below.

### Sample pod configuration file with Exists operator

```
tolerations:
- key: "key1"
  operator: "Exists"
  effect: "NoExecute"
  tolerationSeconds: 3600
```

Both of these tolerations match the [taint created by the `oc adm taint` command above](#). A pod with either toleration would be able to schedule onto **node1**.

#### 14.9.4.1. Using Toleration Seconds to Delay Pod Evictions

You can specify how long a pod can remain bound to a node before being evicted by specifying the **tolerationSeconds** parameter in the pod specification. If a taint with the **NoExecute** effect is added to a node, any pods that do not tolerate the taint are evicted immediately (pods that do tolerate the taint are not evicted). However, if a pod that to be evicted has the **tolerationSeconds** parameter, the pod is not evicted until that time period expires.

For example:

```
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoExecute"
  tolerationSeconds: 3600
```

Here, if this pod is running but does not have a matching taint, the pod stays bound to the node for 3,600 seconds and then be evicted. If the taint is removed before that time, the pod is not evicted.

##### 14.9.4.1.1. Setting a Default Value for Toleration Seconds

This plug-in sets the default forgiveness toleration for pods, to tolerate the **node.alpha.kubernetes.io/not-ready:NoExecute** and **node.alpha.kubernetes.io/unreachable:NoExecute** taints for five minutes.

If the pod configuration provided by the user already has either toleration, the default is not added.

To enable Default Toleration Seconds:

1. Modify the master configuration file (*/etc/origin/master/master-config.yaml*) to Add **DefaultTolerationSeconds** to the admissionConfig section:

```
admissionConfig:
  pluginConfig:
    DefaultTolerationSeconds:
      configuration:
```

```
kind: DefaultAdmissionConfig
apiVersion: v1
disable: false
```

2. Restart OpenShift for the changes to take effect:

```
# systemctl restart atomic-openshift-master-api atomic-openshift-
master-controllers
```

3. Verify that the default was added:

- a. Create a pod:

```
$ oc create -f </path/to/file>
```

For example:

```
$ oc create -f hello-pod.yaml
pod "hello-pod" created
```

- b. Check the pod tolerations:

```
$ oc describe pod <pod-name> |grep -i toleration
```

For example:

```
$ oc describe pod hello-pod |grep -i toleration
Tolerations:      node.alpha.kubernetes.io/not-
ready=:Exists:NoExecute for 300s
```

### 14.9.5. Preventing Pod Eviction for Node Problems

OpenShift Container Platform can be configured to represent **node unreachable** and **node not ready** conditions as taints. This allows per-pod specification of how long to remain bound to a node that becomes unreachable or not ready, rather than using the default of five minutes.

When the Taint Based Evictions feature is enabled, the taints are automatically added by the node controller and the normal logic for evicting pods from **Ready** nodes is disabled.

- If a node enters a not ready state, the **node.alpha.kubernetes.io/not-ready:NoExecute** taint is added and pods cannot be scheduled on the node. Existing pods remain for the toleration seconds period.
- If a node enters a not reachable state, the **node.alpha.kubernetes.io/unreachable:NoExecute** taint is added and pods cannot be scheduled on the node. Existing pods remain for the toleration seconds period.

To enable Taint Based Evictions:

1. Modify the master configuration file (*/etc/origin/master/master-config.yaml*) to add the following to the **kubernetesMasterConfig** section:

```
kubernetesMasterConfig:
  controllerArguments:
```

```
feature-gates:
- "TaintBasedEvictions=true"
```

2. Check that the taint is added to a node:

```
oc describe node $node | grep -i taint

Taints: node.alpha.kubernetes.io/not-ready:NoExecute
```

3. Restart OpenShift for the changes to take effect:

```
# systemctl restart atomic-openshift-master-api atomic-openshift-
master-controllers
```

4. Add a toleration to pods:

```
tolerations:
- key: "node.alpha.kubernetes.io/unreachable"
  operator: "Exists"
  effect: "NoExecute"
  tolerationSeconds: 6000
```

or

```
tolerations:
- key: "node.alpha.kubernetes.io/not-ready"
  operator: "Exists"
  effect: "NoExecute"
  tolerationSeconds: 6000
```



## NOTE

To maintain the existing [rate limiting](#) behavior of pod evictions due to node problems, the system adds the taints in a rate-limited way. This prevents massive pod evictions in scenarios such as the master becoming partitioned from the nodes.

### 14.9.6. Daemonsets and Tolerations

[DaemonSet](#) pods are created with **NoExecute** tolerations for **node.alpha.kubernetes.io/unreachable** and **node.alpha.kubernetes.io/not-ready** with no **tolerationSeconds** to ensure that DaemonSet pods are never evicted due to these problems, even when the Default Toleration Seconds feature is disabled.

### 14.9.7. Examples

Taints and tolerations are a flexible way to steer pods away from nodes or evict pods that should not be running on a node. A few of typical scenarios are:

- [Dedicating a node for a user](#)
- [Binding a user to a node](#)
- [Dedicating nodes with special hardware](#)

### 14.9.7.1. Dedicating a Node for a User

You can specify a set of nodes for exclusive use by a particular set of users.

To specify dedicated nodes:

1. Add a taint to those nodes:

For example:

```
$ oc adm taint nodes node1 dedicated=groupName:NoSchedule
```

2. Add a corresponding toleration to the pods by writing a custom [admission controller](#).  
Only the pods with the tolerations are allowed to use the dedicated nodes.

### 14.9.7.2. Binding a User to a Node

You can configure a node so that particular users can use only the dedicated nodes.

To configure a node so that users can use only that node:

1. Add a taint to those nodes:

For example:

```
$ oc adm taint nodes node1 dedicated=groupName:NoSchedule
```

2. Add a corresponding toleration to the pods by writing a custom [admission controller](#).  
The admission controller should add a node affinity to require that the pods can only schedule onto nodes labeled with the **key:value** label (**dedicated=groupName**).
3. Add a label similar to the taint (such as the **key:value** label) to the dedicated nodes.

### 14.9.7.3. Nodes with Special Hardware

In a cluster where a small subset of nodes have specialized hardware (for example GPUs), you can use taints and tolerations to keep pods that do not need the specialized hardware off of those nodes, leaving the nodes for pods that do need the specialized hardware. You can also require pods that need specialized hardware to use specific nodes.

To ensure pods are blocked from the specialized hardware:

1. Taint the nodes that have the specialized hardware using one of the following commands:

```
$ oc adm taint nodes <node-name> disktype=ssd:NoSchedule  
$ oc adm taint nodes <node-name> disktype=ssd:PreferNoSchedule
```

2. Adding a corresponding toleration to pods that use the special hardware using an [admission controller](#).

For example, the admission controller could use some characteristic(s) of the pod to determine that the pod should be allowed to use the special nodes by adding a toleration.

To ensure pods can only use the specialized hardware, you need some additional mechanism. For example, you could label the nodes that have the special hardware and use node affinity on the pods that need the hardware.

## CHAPTER 15. SETTING QUOTAS

### 15.1. OVERVIEW

A resource quota, defined by a **ResourceQuota** object, provides constraints that limit aggregate resource consumption per project. It can limit the quantity of objects that can be created in a project by type, as well as the total amount of compute resources and storage that may be consumed by resources in that project.

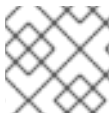


#### NOTE

See the [Developer Guide](#) for more on compute resources.

### 15.2. RESOURCES MANAGED BY QUOTA

The following describes the set of compute resources and object types that may be managed by a quota.



#### NOTE

A pod is in a terminal state if **status.phase in (Failed, Succeeded)** is true.

**Table 15.1. Compute Resources Managed by Quota**

| Resource Name          | Description   |
|------------------------|---|
| <b>cpu</b>             | The sum of CPU requests across all pods in a non-terminal state cannot exceed this value. <b>cpu</b> and <b>requests.cpu</b> are the same value and can be used interchangeably.          |
| <b>memory</b>          | The sum of memory requests across all pods in a non-terminal state cannot exceed this value. <b>memory</b> and <b>requests.memory</b> are the same value and can be used interchangeably. |
| <b>requests.cpu</b>    | The sum of CPU requests across all pods in a non-terminal state cannot exceed this value. <b>cpu</b> and <b>requests.cpu</b> are the same value and can be used interchangeably.          |
| <b>requests.memory</b> | The sum of memory requests across all pods in a non-terminal state cannot exceed this value. <b>memory</b> and <b>requests.memory</b> are the same value and can be used interchangeably. |
| <b>limits.cpu</b>      | The sum of CPU limits across all pods in a non-terminal state cannot exceed this value.   |
| <b>limits.memory</b>   | The sum of memory limits across all pods in a non-terminal state cannot exceed this value.  |

**Table 15.2. Storage Resources Managed by Quota**

| Resource Name  | Description  |
|--|--|
| <b>requests.storage</b>  | The sum of storage requests across all persistent volume claims in any state cannot exceed this value.                                     |
| <b>persistentvolumeclaims</b>  | The total number of persistent volume claims that can exist in the project.  |
| <b>&lt;storage-class-name&gt;.storageclass.storage.k8s.io/requests.storage</b>       | The sum of storage requests across all persistent volume claims in any state that have a matching storage class, cannot exceed this value. |
| <b>&lt;storage-class-name&gt;.storageclass.storage.k8s.io/persistentvolumeclaims</b> | The total number of persistent volume claims with a matching storage class that can exist in the project.                                  |

Table 15.3. Object Counts Managed by Quota

| Resource Name                    | Description   |
|----------------------------------|---|
| <b>pods</b>                      | The total number of pods in a non-terminal state that can exist in the project. |
| <b>replicationcontrollers</b>    | The total number of replication controllers that can exist in the project.      |
| <b>resourcequotas</b>            | The total number of resource quotas that can exist in the project.              |
| <b>services</b>                  | The total number of services that can exist in the project.                     |
| <b>secrets</b>                   | The total number of secrets that can exist in the project.                      |
| <b>configmaps</b>                | The total number of <b>ConfigMap</b> objects that can exist in the project.     |
| <b>persistentvolumeclaims</b>    | The total number of persistent volume claims that can exist in the project.     |
| <b>openshift.io/imagestreams</b> | The total number of image streams that can exist in the project.                |

## 15.3. QUOTA SCOPES

Each quota can have an associated set of *scopes*. A quota will only measure usage for a resource if it matches the intersection of enumerated scopes.

Adding a scope to a quota restricts the set of resources to which that quota can apply. Specifying a resource outside of the allowed set results in a validation error.

| Scope                 | Description  |
|-----------------------|--|
| <b>Terminating</b>    | Match pods where <code>spec.activeDeadlineSeconds &gt;= 0</code> .   |
| <b>NotTerminating</b> | Match pods where <code>spec.activeDeadlineSeconds</code> is <code>nil</code> .   |
| <b>BestEffort</b>     | Match pods that have best effort quality of service for either <b>cpu</b> or <b>memory</b> .<br>See the <a href="#">Quality of Service Classes</a> for more on committing compute resources. |
| <b>NotBestEffort</b>  | Match pods that do not have best effort quality of service for <b>cpu</b> and <b>memory</b> .  |

A **BestEffort** scope restricts a quota to limiting the following resources:

- **pods**

A **Terminating**, **NotTerminating**, and **NotBestEffort** scope restricts a quota to tracking the following resources:

- **pods**
- **memory**
- **requests.memory**
- **limits.memory**
- **cpu**
- **requests.cpu**
- **limits.cpu**

## 15.4. QUOTA ENFORCEMENT

After a resource quota for a project is first created, the project restricts the ability to create any new resources that may violate a quota constraint until it has calculated updated usage statistics.

After a quota is created and usage statistics are updated, the project accepts the creation of new content. When you create or modify resources, your quota usage is incremented immediately upon the request to create or modify the resource.

When you delete a resource, your quota use is decremented during the next full recalculation of quota statistics for the project.

A [configurable amount of time](#) determines how long it takes to reduce quota usage statistics to their current observed system value.

If project modifications exceed a quota usage limit, the server denies the action, and an appropriate error message is returned to the user explaining the quota constraint violated, and what their currently observed usage stats are in the system.

## 15.5. REQUESTS VERSUS LIMITS

When allocating [compute resources](#), each container may specify a request and a limit value each for CPU and memory. Quotas can restrict any of these values.

If the quota has a value specified for **requests.cpu** or **requests.memory**, then it requires that every incoming container make an explicit request for those resources. If the quota has a value specified for **limits.cpu** or **limits.memory**, then it requires that every incoming container specify an explicit limit for those resources.

## 15.6. SAMPLE RESOURCE QUOTA DEFINITIONS

### *core-object-counts.yaml*

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: core-object-counts
spec:
  hard:
    configmaps: "10" ❶
    persistentvolumeclaims: "4" ❷
    replicationcontrollers: "20" ❸
    secrets: "10" ❹
    services: "10" ❺
```

- ❶ The total number of **ConfigMap** objects that can exist in the project.
- ❷ The total number of persistent volume claims (PVCs) that can exist in the project.
- ❸ The total number of replication controllers that can exist in the project.
- ❹ The total number of secrets that can exist in the project.
- ❺ The total number of services that can exist in the project.

### *openshift-object-counts.yaml*

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: openshift-object-counts
spec:
  hard:
    openshift.io/imagestreams: "10" ❶
```

- ❶ The total number of image streams that can exist in the project.

### *compute-resources.yaml*

```
apiVersion: v1
```



```

kind: ResourceQuota
metadata:
  name: compute-resources
spec:
  hard:
    pods: "4" ❶
    requests.cpu: "1" ❷
    requests.memory: 1Gi ❸
    limits.cpu: "2" ❹
    limits.memory: 2Gi ❺

```

- ❶ The total number of pods in a non-terminal state that can exist in the project.
- ❷ Across all pods in a non-terminal state, the sum of CPU requests cannot exceed 1 core.
- ❸ Across all pods in a non-terminal state, the sum of memory requests cannot exceed 1Gi.
- ❹ Across all pods in a non-terminal state, the sum of CPU limits cannot exceed 2 cores.
- ❺ Across all pods in a non-terminal state, the sum of memory limits cannot exceed 2Gi.

### ***besteffort.yaml***

```

apiVersion: v1
kind: ResourceQuota
metadata:
  name: besteffort
spec:
  hard:
    pods: "1" ❶
  scopes:
    - BestEffort ❷

```

- ❶ The total number of pods in a non-terminal state with **BestEffort** quality of service that can exist in the project.
- ❷ Restricts the quota to only matching pods that have **BestEffort** quality of service for either memory or CPU.

### ***compute-resources-long-running.yaml***

```

apiVersion: v1
kind: ResourceQuota
metadata:
  name: compute-resources-long-running
spec:
  hard:
    pods: "4" ❶
    limits.cpu: "4" ❷
    limits.memory: "2Gi" ❸
  scopes:
    - NotTerminating ❹

```

- 1 The total number of pods in a non-terminal state.
- 2 Across all pods in a non-terminal state, the sum of CPU limits cannot exceed this value.
- 3 Across all pods in a non-terminal state, the sum of memory limits cannot exceed this value.
- 4 Restricts the quota to only matching pods where **spec.activeDeadlineSeconds** is set to **nil**. Build pods will fall under **NotTerminating** unless the **RestartNever** policy is applied.

### *compute-resources-time-bound.yaml*

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: compute-resources-time-bound
spec:
  hard:
    pods: "2" 1
    limits.cpu: "1" 2
    limits.memory: "1Gi" 3
  scopes:
    - Terminating 4
```

- 1 The total number of pods in a non-terminal state.
- 2 Across all pods in a non-terminal state, the sum of CPU limits cannot exceed this value.
- 3 Across all pods in a non-terminal state, the sum of memory limits cannot exceed this value.
- 4 Restricts the quota to only matching pods where **spec.activeDeadlineSeconds**  $\geq 0$ . For example, this quota would charge for build or deployer pods, but not long running pods like a web server or database.

### *storage-consumption.yaml*

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: storage-consumption
spec:
  hard:
    persistentvolumeclaims: "10" 1
    requests.storage: "50Gi" 2
    gold.storageclass.storage.k8s.io/requests.storage: "10Gi" 3
    silver.storageclass.storage.k8s.io/requests.storage: "20Gi" 4
    silver.storageclass.storage.k8s.io/persistentvolumeclaims: "5" 5
    bronze.storageclass.storage.k8s.io/requests.storage: "0" 6
    bronze.storageclass.storage.k8s.io/persistentvolumeclaims: "0" 7
```

- 1 The total number of persistent volume claims in a project
- 2

Across all persistent volume claims in a project, the sum of storage requested cannot exceed this value.

- 3 Across all persistent volume claims in a project, the sum of storage requested in the gold storage class cannot exceed this value.
- 4 Across all persistent volume claims in a project, the sum of storage requested in the silver storage class cannot exceed this value.
- 5 Across all persistent volume claims in a project, the total number of claims in the silver storage class cannot exceed this value.
- 6 Across all persistent volume claims in a project, the sum of storage requested in the bronze storage class cannot exceed this value. When this is set to **0**, it means bronze storage class cannot request storage.
- 7 Across all persistent volume claims in a project, the sum of storage requested in the bronze storage class cannot exceed this value. When this is set to **0**, it means bronze storage class cannot create claims.

## 15.7. CREATING A QUOTA

To create a quota, first define the quota to your specifications in a file, for example as seen in [Sample Resource Quota Definitions](#). Then, create using that file to apply it to a project:

```
$ oc create -f <resource_quota_definition> [-n <project_name>]
```

For example:

```
$ oc create -f resource-quota.json -n demoproject
```

## 15.8. VIEWING A QUOTA

You can view usage statistics related to any hard limits defined in a project's quota by navigating in the web console to the project's **Quota** page.

You can also use the CLI to view quota details:

1. First, get the list of quotas defined in the project. For example, for a project called **demoproject**:

```
$ oc get quota -n demoproject
NAME                AGE
besteffort          11m
compute-resources   2m
core-object-counts  29m
```

2. Then, describe the quota you are interested in, for example the **core-object-counts** quota:

```
$ oc describe quota core-object-counts -n demoproject
Name:      core-object-counts
Namespace: demoproject
Resource   Used Hard
-----

```

```

configmaps 3 10
persistentvolumeclaims 0 4
replicationcontrollers 3 20
secrets 9 10
services 2 10

```

## 15.9. CONFIGURING QUOTA SYNCHRONIZATION PERIOD

When a set of resources are deleted, the synchronization time frame of resources is determined by the **resource-quota-sync-period** setting in the `/etc/origin/master/master-config.yaml` file.

Before quota usage is restored, a user may encounter problems when attempting to reuse the resources. You can change the **resource-quota-sync-period** setting to have the set of resources regenerate at the desired amount of time (in seconds) and for the resources to be available again:

```

kubernetesMasterConfig:
  apiLevels:
    - v1beta3
    - v1
  apiServerArguments: null
  controllerArguments:
    resource-quota-sync-period:
      - "10s"

```

After making any changes, restart the master services to apply them.

```
# systemctl restart atomic-openshift-master-api atomic-openshift-master-controllers
```

Adjusting the regeneration time can be helpful for creating resources and determining resource usage when automation is used.



### NOTE

The **resource-quota-sync-period** setting is designed to balance system performance. Reducing the sync period can result in a heavy load on the master.

## 15.10. ACCOUNTING FOR QUOTA IN DEPLOYMENT CONFIGURATIONS

If a quota has been defined for your project, see [Deployment Resources](#) for considerations on any deployment configurations.

## 15.11. REQUIRE EXPLICIT QUOTA TO CONSUME A RESOURCE



### NOTE

This feature is tech preview and subject to change in future releases.

If a resource is not managed by quota, a user has no restriction on the amount of resource that can be consumed. For example, if there is no quota on storage related to the gold storage class, the amount of gold storage a project can create is unbounded.

For high-cost compute or storage resources, administrators may want to require an explicit quota be granted in order to consume a resource. For example, if a project was not explicitly given quota for storage related to the gold storage class, users of that project would not be able to create any storage of that type.

In order to require explicit quota to consume a particular resource, the following stanza should be added to the master-config.yaml.

```
admissionConfig:
  pluginConfig:
    ResourceQuota:
      configuration:
        apiVersion: resourcequota.admission.k8s.io/v1alpha1
        kind: Configuration
        limitedResources:
          - resource: persistentvolumeclaims ❶
          matchContains:
            - gold.storageclass.storage.k8s.io/requests.storage ❷
```

❶ The group/resource to whose consumption is limited by default.

❷ The name of the resource tracked by quota associated with the group/resource to limit by default.

In the above example, the quota system will intercept every operation that creates or updates a **PersistentVolumeClaim**. It checks what resources understood by quota would be consumed, and if there is no covering quota for those resources in the project, the request is denied. In this example, if a user creates a **PersistentVolumeClaim** that uses storage associated with the gold storage class, and there is no matching quota in the project, the request is denied.

## 15.12. KNOWN ISSUES

- Invalid objects can cause quota resources for a project to become exhausted. Quota is incremented in admission prior to validation of the resource. As a result, quota can be incremented even if the pod is not ultimately persisted. This will be resolved in a future release. ([BZ1485375](#))

## CHAPTER 16. SETTING MULTI-PROJECT QUOTAS

### 16.1. OVERVIEW

A multi-project quota, defined by a **ClusterResourceQuota** object, allows [quotas](#) to be shared across multiple projects. Resources used in each selected project will be aggregated and that aggregate will be used to limit resources across all the selected projects.

### 16.2. SELECTING PROJECTS

You can select projects based on annotation selection, label selection, or both. For example, to select projects based on annotations, run the following command:

```
$ oc create clusterquota for-user \
  --project-annotation-selector openshift.io/requester=<user-name> \
  --hard pods=10 \
  --hard secrets=20
```

It creates the following **ClusterResourceQuota** object:

```
apiVersion: v1
kind: ClusterResourceQuota
metadata:
  name: for-user
spec:
  quota: 1
  hard:
    pods: "10"
    secrets: "20"
  selector:
    annotations: 2
    openshift.io/requester: <user-name>
    labels: null 3
status:
  namespaces: 4
  - namespace: ns-one
    status:
      hard:
        pods: "10"
        secrets: "20"
      used:
        pods: "1"
        secrets: "9"
  total: 5
  hard:
    pods: "10"
    secrets: "20"
  used:
    pods: "1"
    secrets: "9"
```

**1** The **ResourceQuotaSpec** object that will be enforced over the selected projects.

- 2 A simple key/value selector for annotations.
- 3 A label selector that can be used to select projects.
- 4 A per-namespace map that describes current quota usage in each selected project.
- 5 The aggregate usage across all selected projects.

This multi-project quota document controls all projects requested by **<user-name>** using the default project request endpoint. You are limited to 10 pods and 20 secrets.

Similarly, to select projects based on labels, run this command:

```
$ oc create clusterresourcequota for-name \ 1
  --project-label-selector=name=frontend \ 2
  --hard=pods=10 --hard=secrets=20
```

- 1 Both **clusterresourcequota** and **clusterquota** are aliases of the same command. **for-name** is the name of the **clusterresourcequota** object.
- 2 To select projects by label, provide a key-value pair by using the format **--project-label-selector=key=value**.

It creates the following **ClusterResourceQuota** object definition:

```
apiVersion: v1
kind: ClusterResourceQuota
metadata:
  creationTimestamp: null
  name: for-name
spec:
  quota:
    hard:
      pods: "10"
      secrets: "20"
  selector:
    annotations: null
    labels:
      matchLabels:
        name: frontend
```

## 16.3. VIEWING APPLICABLE CLUSTERRESOURCEQUOTAS

A project administrator is not allowed to create or modify the multi-project quota that limits his or her project, but the administrator is allowed to view the multi-project quota documents that are applied to his or her project. The project administrator can do this via the **AppliedClusterResourceQuota** resource.

```
$ oc describe AppliedClusterResourceQuota
```

produces:

```
Name:      for-user
Namespace:  <none>
Created:    19 hours ago
Labels:     <none>
Annotations: <none>
Label Selector: <null>
AnnotationSelector: map[openshift.io/requester:<user-name>]
Resource    Used    Hard
-----
pods         1      10
secrets      9      20
```

## 16.4. SELECTION GRANULARITY

Because of the locking consideration when claiming quota allocations, the number of active projects selected by a multi-project quota is an important consideration. Selecting more than 100 projects under a single multi-project quota may have detrimental effects on API server responsiveness in those projects.



## CHAPTER 17. SETTING LIMIT RANGES

### 17.1. OVERVIEW

A limit range, defined by a **LimitRange** object, enumerates [compute resource constraints](#) in a [project](#) at the pod, container, image, image stream, and persistent volume claim level, and specifies the amount of resources that a pod, container, image, image stream, or persistent volume claim can consume.

All resource create and modification requests are evaluated against each **LimitRange** object in the project. If the resource violates any of the enumerated constraints, then the resource is rejected. If the resource does not set an explicit value, and if the constraint supports a default value, then the default value is applied to the resource.

#### Example 17.1. Limit Range Object Definition

```
apiVersion: "v1"
kind: "LimitRange"
metadata:
  name: "core-resource-limits" ❶
spec:
  limits:
    - type: "Pod"
      max:
        cpu: "2" ❷
        memory: "1Gi" ❸
      min:
        cpu: "200m" ❹
        memory: "6Mi" ❺
    - type: "Container"
      max:
        cpu: "2" ❻
        memory: "1Gi" ❼
      min:
        cpu: "100m" ❽
        memory: "4Mi" ❾
      default:
        cpu: "300m" ❿
        memory: "200Mi" ⓫
      defaultRequest:
        cpu: "200m" ⓫
        memory: "100Mi" ⓫
      maxLimitRequestRatio:
        cpu: "10" ⓫
```

- ❶ The name of the limit range object.
- ❷ The maximum amount of CPU that a pod can request on a node across all containers.
- ❸ The maximum amount of memory that a pod can request on a node across all containers.
- ❹ The minimum amount of CPU that a pod can request on a node across all containers.
- ❺ The minimum amount of memory that a pod can request on a node across all containers.

- 6 The maximum amount of CPU that a single container in a pod can request.
- 7 The maximum amount of memory that a single container in a pod can request.
- 8 The minimum amount of CPU that a single container in a pod can request.
- 9 The minimum amount of memory that a single container in a pod can request.
- 10 The default amount of CPU that a container will be limited to use if not specified.
- 11 The default amount of memory that a container will be limited to use if not specified.
- 12 The default amount of CPU that a container will request to use if not specified.
- 13 The default amount of memory that a container will request to use if not specified.
- 14 The maximum amount of CPU burst that a container can make as a ratio of its limit over request.

For more information on how CPU and memory are measured, see [Compute Resources](#).

### Example 17.2. OpenShift Container Platform Limit Range Object Definition

```
apiVersion: "v1"
kind: "LimitRange"
metadata:
  name: "openshift-resource-limits"
spec:
  limits:
    - type: openshift.io/Image
      max:
        storage: 1Gi 1
    - type: openshift.io/ImageStream
      max:
        openshift.io/image-tags: 20 2
        openshift.io/images: 30 3
```

- 1 The maximum size of an image that can be pushed to an internal registry.
- 2 The maximum number of unique image tags per image stream's spec.
- 3 The maximum number of unique image references per image stream's status.

Both core and OpenShift Container Platform resources can be specified in just one limit range object. They are separated here into two examples for clarity.

#### 17.1.1. Container Limits

##### Supported Resources:

- CPU

- Memory

### Supported Constraints:

Per container, the following must hold true if specified:

**Table 17.1. Container**

| Constraint                  | Behavior  |
|-----------------------------|---|
| <b>Min</b>                  | <p><b>Min[resource]</b> less than or equal to <b>container.resources.requests[resource]</b> (required) less than or equal to <b>container/resources.limits[resource]</b> (optional)</p> <p>If the configuration defines a <b>min</b> CPU, then the request value must be greater than the CPU value. A limit value does not need to be specified.</p>   |
| <b>Max</b>                  | <p><b>container.resources.limits[resource]</b> (required) less than or equal to <b>Max[resource]</b></p> <p>If the configuration defines a <b>max</b> CPU, then you do not need to define a request value, but a limit value does need to be set that satisfies the maximum CPU constraint.</p>   |
| <b>MaxLimitRequestRatio</b> | <p><b>MaxLimitRequestRatio[resource]</b> less than or equal to (<b>container.resources.limits[resource]</b> / <b>container.resources.requests[resource]</b>)</p> <p>If a configuration defines a <b>maxLimitRequestRatio</b> value, then any new containers must have both a request and limit value. Additionally, OpenShift Container Platform calculates a limit to request ratio by dividing the limit by the request. This value should be a non-negative integer greater than 1.</p> <p>For example, if a container has <b>cpu: 500</b> in the <b>limit</b> value, and <b>cpu: 100</b> in the <b>request</b> value, then its limit to request ratio for <b>cpu</b> is <b>5</b>. This ratio must be less than or equal to the <b>maxLimitRequestRatio</b>.</p> |

### Supported Defaults:

#### Default[resource]

Defaults **container.resources.limit[resource]** to specified value if none.

#### Default Requests[resource]

Defaults **container.resources.requests[resource]** to specified value if none.

### 17.1.2. Pod Limits

#### Supported Resources:

- CPU
- Memory

#### Supported Constraints:

Across all containers in a pod, the following must hold true:

**Table 17.2. Pod**

| Constraint                  | Enforced Behavior  |
|-----------------------------|--|
| <b>Min</b>                  | <b>Min[resource]</b> less than or equal to <b>container.resources.requests[resource]</b> (required) less than or equal to <b>container.resources.limits[resource]</b> (optional) |
| <b>Max</b>                  | <b>container.resources.limits[resource]</b> (required) less than or equal to <b>Max[resource]</b>  |
| <b>MaxLimitRequestRatio</b> | <b>MaxLimitRequestRatio[resource]</b> less than or equal to ( <b>container.resources.limits[resource]</b> / <b>container.resources.requests[resource]</b> )                      |

### 17.1.3. Image Limits

#### Supported Resources:

- Storage

#### Resource type name:

- `openshift.io/Image`

Per image, the following must hold true if specified:

**Table 17.3. Image**

| Constraint | Behavior   |
|------------|--|
| <b>Max</b> | <b>image.dockerimagemetadata.size</b> less than or equal to <b>Max[resource]</b> |



#### NOTE

To prevent blobs exceeding the limit from being uploaded to the registry, the registry must be configured to enforce quota. An environment variable **REGISTRY\_MIDDLEWARE\_REPOSITORY\_OPENSHIFT\_ENFORCEQUOTA** must be set to **true** which is done by default for new deployments. To update older deployment configuration, refer to [Enforcing quota in the Registry](#).

**WARNING**

The image size is not always available in the manifest of an uploaded image. This is especially the case for images built with Docker 1.10 or higher and pushed to a v2 registry. If such an image is pulled with an older Docker daemon, the image manifest will be converted by the registry to schema v1 lacking all the size information. No storage limit set on images will prevent it from being uploaded.

[The issue](#) is being addressed.

### 17.1.4. Image Stream Limits

#### Supported Resources:

- `openshift.io/image-tags`
- `openshift.io/images`

#### Resource type name:

- `openshift.io/ImageStream`

Per image stream, the following must hold true if specified:

**Table 17.4. ImageStream**

| Constraint                                | Behavior  |
|---|---|
| <code>Max[openshift.io/image-tags]</code> | <p><code>length( uniqueimagetags( imagestream.spec.tags ) )</code> less than or equal to <code>Max[openshift.io/image-tags]</code></p> <p><code>uniqueimagetags</code> returns unique references to images of given spec tags.</p>                  |
| <code>Max[openshift.io/images]</code>     | <p><code>length( uniqueimages( imagestream.status.tags ) )</code> less than or equal to <code>Max[openshift.io/images]</code></p> <p><code>uniqueimages</code> returns unique image names found in status tags. The name equals image's digest.</p> |

#### 17.1.4.1. Counting of Image References

Resource `openshift.io/image-tags` represents unique [image references](#). Possible references are an `ImageStreamTag`, an `ImageStreamImage` and a `DockerImage`. They may be created using commands `oc tag` and `oc import-image` or by using [tag tracking](#). No distinction is made between internal and external references. However, each unique reference tagged in the image stream's specification is counted just once. It does not restrict pushes to an internal container registry in any way, but is useful for tag restriction.

Resource `openshift.io/images` represents unique image names recorded in image stream status. It allows for restriction of a number of images that can be pushed to the internal registry. Internal and external references are not distinguished.

### 17.1.5. PersistentVolumeClaim Limits

#### Supported Resources:

- Storage

#### Supported Constraints:

Across all persistent volume claims in a project, the following must hold true:

**Table 17.5. Pod**

| Constraint | Enforced Behavior   |
|------------|---|
| <b>Min</b> | $\text{Min}[\text{resource}] \Leftarrow \text{claim.spec.resources.requests}[\text{resource}]$ (required)   |
| <b>Max</b> | $\text{claim.spec.resources.requests}[\text{resource}]$ (required) $\Leftarrow \text{Max}[\text{resource}]$ |

#### Example 17.3. Limit Range Object Definition

```
{
  "apiVersion": "v1",
  "kind": "LimitRange",
  "metadata": {
    "name": "pvcs" 1
  },
  "spec": {
    "limits": [{
      "type": "PersistentVolumeClaim",
      "min": {
        "storage": "2Gi" 2
      },
      "max": {
        "storage": "50Gi" 3
      }
    }]
  }
}
```

1 The name of the limit range object.

2 The minimum amount of storage that can be requested in a persistent volume claim

3 The maximum amount of storage that can be requested in a persistent volume claim



## CHAPTER 18. PRUNING OBJECTS

### 18.1. OVERVIEW

Over time, [API objects](#) created in OpenShift Container Platform can accumulate in the [etcd data store](#) through normal user operations, such as when building and deploying applications.

As an administrator, you can periodically prune older versions of objects from your OpenShift Container Platform instance that are no longer needed. For example, by pruning images you can delete older images and layers that are no longer in use, but are still taking up disk space.

### 18.2. BASIC PRUNE OPERATIONS

The CLI groups prune operations under a common parent command.

```
$ oc adm prune <object_type> <options>
```

This specifies:

- The **<object\_type>** to perform the action on, such as **builds**, **deployments**, or **images**.
- The **<options>** supported to prune that object type.

### 18.3. PRUNING DEPLOYMENTS

In order to prune deployments that are no longer required by the system due to age and status, administrators may run the following command:

```
$ oc adm prune deployments [<options>]
```

**Table 18.1. Prune Deployments CLI Configuration Options**

| Option                                      | Description  |
|---|--|
| <b>--confirm</b>                            | Indicate that pruning should occur, instead of performing a dry-run.   |
| <b>--orphans</b>                            | Prune all deployments whose deployment config no longer exists, status is complete or failed, and replica count is zero.   |
| <b>--keep-complete=&lt;N&gt;</b>            | Per deployment config, keep the last N deployments whose status is complete and replica count is zero. (default <b>5</b> )   |
| <b>--keep-failed=&lt;N&gt;</b>              | Per deployment config, keep the last N deployments whose status is failed and replica count is zero. (default <b>1</b> )   |
| <b>--keep-younger-than=&lt;duration&gt;</b> | Do not prune any object that is younger than <b>&lt;duration&gt;</b> relative to the current time. (default <b>60m</b> ) Valid units of measurement include nanoseconds ( <b>ns</b> ), microseconds ( <b>us</b> ), milliseconds ( <b>ms</b> ), seconds ( <b>s</b> ), minutes ( <b>m</b> ), and hours ( <b>h</b> ). |



To see what a pruning operation would delete:

```
$ oc adm prune deployments --orphans --keep-complete=5 --keep-failed=1 \
  --keep-younger-than=60m
```

To actually perform the prune operation:

```
$ oc adm prune deployments --orphans --keep-complete=5 --keep-failed=1 \
  --keep-younger-than=60m --confirm
```

## 18.4. PRUNING BUILDS

In order to prune builds that are no longer required by the system due to age and status, administrators may run the following command:

```
$ oc adm prune builds [<options>]
```

**Table 18.2. Prune Builds CLI Configuration Options**

| Option                                      | Description  |
|---|--|
| <b>--confirm</b>                            | Indicate that pruning should occur, instead of performing a dry-run.   |
| <b>--orphans</b>                            | Prune all builds whose build config no longer exists, status is complete, failed, error, or canceled.                    |
| <b>--keep-complete=&lt;N&gt;</b>            | Per build config, keep the last N builds whose status is complete. (default <b>5</b> )                                   |
| <b>--keep-failed=&lt;N&gt;</b>              | Per build config, keep the last N builds whose status is failed, error, or canceled (default <b>1</b> )                  |
| <b>--keep-younger-than=&lt;duration&gt;</b> | Do not prune any object that is younger than <b>&lt;duration&gt;</b> relative to the current time. (default <b>60m</b> ) |

To see what a pruning operation would delete:

```
$ oc adm prune builds --orphans --keep-complete=5 --keep-failed=1 \
  --keep-younger-than=60m
```

To actually perform the prune operation:

```
$ oc adm prune builds --orphans --keep-complete=5 --keep-failed=1 \
  --keep-younger-than=60m --confirm
```



### NOTE

Developers can enable [automatic build pruning](#) by modifying their build configuration.

## 18.5. PRUNING IMAGES

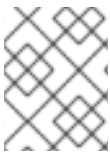
In order to prune images that are no longer required by the system due to age, status, or exceed limits, administrators may run the following command:

```
$ oc adm prune images [<options>]
```



### NOTE

Currently, to prune images you must first [log in to the CLI](#) as a user with an [access token](#). The user must also have the [cluster role](#) **system:image-pruner** or greater (for example, **cluster-admin**).



### NOTE

Pruning images removes data from the integrated registry. For this operation to work properly, ensure your [registry is configured](#) with **storage:delete:enabled** set to **true**.



### NOTE

Pruning images with the **--namespace** flag does not remove images, only image streams. Images are non-namespaced resources. Therefore, limiting pruning to a particular namespace makes it impossible to calculate their current usage.

**Table 18.3. Prune Images CLI Configuration Options**

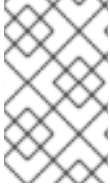
| Option                                | Description   |
|---------------------------------------|---|
| <b>--all</b>                          | Include images that were not pushed to the registry, but have been mirrored by pullthrough. This is on by default. To limit the pruning to images that were pushed to the integrated registry, pass <b>--all=false</b> .  |
| <b>--certificate-authority</b>        | The path to a certificate authority file to use when communicating with the OpenShift Container Platform-managed registries. Defaults to the certificate authority data from the current user's configuration file. If provided, secure connection will be initiated. |
| <b>--confirm</b>                      | Indicate that pruning should occur, instead of performing a dry-run. This requires a valid route to the integrated Docker registry. If this command is run outside of the cluster network, the route needs to be provided using <b>--registry-url</b> .               |
| <b>--force-insecure</b>               | <b>Use caution with this option.</b> Allow an insecure connection to the Docker registry that is hosted via HTTP or has an invalid HTTPS certificate. See <a href="#">Using Secure or Insecure Connections</a> for more information.                                  |
| <b>--keep-tag-revisions=&lt;N&gt;</b> | For each image stream, keep up to at most N image revisions per tag. (default <b>3</b> )  |

| Option                                      | Description   |
|---|---|
| <b>--keep-younger-than=&lt;duration&gt;</b> | Do not prune any image that is younger than <b>&lt;duration&gt;</b> relative to the current time. Do not prune any image that is referenced by any other object that is younger than <b>&lt;duration&gt;</b> relative to the current time. (default <b>60m</b> )  |
| <b>--prune-over-size-limit</b>              | Prune each image that exceeds the smallest <b>limit</b> defined in the same project. This flag cannot be combined with <b>--keep-tag-revisions</b> nor <b>--keep-younger-than</b> .   |
| <b>--registry-url</b>                       | The address to use when contacting the registry. The command will attempt to use a cluster-internal URL determined from managed images and image streams. In case it fails (the registry cannot be resolved or reached), an alternative route that works needs to be provided using this flag. The registry host name may be prefixed by <b>https://</b> or <b>http://</b> which will enforce particular connection protocol. |

### 18.5.1. Image Prune Conditions

- Remove any image "managed by OpenShift Container Platform" (images with the annotation **openshift.io/image.managed**) that was created at least **--keep-younger-than** minutes ago and is not currently referenced by:
  - any pod created less than **--keep-younger-than** minutes ago.
  - any image stream created less than **--keep-younger-than** minutes ago.
  - any running pods.
  - any pending pods.
  - any replication controllers.
  - any deployment configurations.
  - any build configurations.
  - any builds.
  - the **--keep-tag-revisions** most recent items in **stream.status.tags[].items**.
- Remove any image "managed by OpenShift Container Platform" (images with the annotation **openshift.io/image.managed**) that is exceeding the smallest **limit** defined in the same project and is not currently referenced by:
  - any running pods.
  - any pending pods.
  - any replication controllers.
  - any deployment configurations.

- any build configurations.
- any builds.
- There is no support for pruning from external registries.
- When an image is pruned, all references to the image are removed from all image streams that have a reference to the image in **status.tags**.
- Image layers that are no longer referenced by any images are removed as well.

**NOTE**

**--prune-over-size-limit** cannot be combined with **--keep-tag-revisions** nor **--keep-younger-than** flags. Doing so will return an information that this operation is not allowed.

To see what a pruning operation would delete:

1. Keeping up to three tag revisions, and keeping resources (images, image streams and pods) younger than sixty minutes:

```
$ oc adm prune images --keep-tag-revisions=3 --keep-younger-than=60m
```

2. Pruning every image that exceeds defined limits:

```
$ oc adm prune images --prune-over-size-limit
```

To actually perform the prune operation for the previously mentioned options accordingly:

```
$ oc adm prune images --keep-tag-revisions=3 --keep-younger-than=60m --confirm
```

```
$ oc adm prune images --prune-over-size-limit --confirm
```

### 18.5.2. Using Secure or Insecure Connections

The secure connection is the preferred and recommended approach. It is done over HTTPS protocol with a mandatory certificate verification. The **prune** command always attempts to use it if possible. If not possible, in some cases it can fall-back to insecure connection, which is dangerous. In this case, either certificate verification is skipped or plain HTTP protocol is used.

The fall-back to insecure connection is allowed in the following cases unless **--certificate-authority** is specified:

1. The **prune** command is run with the **--force-insecure** option.
2. The provided **registry-url** is prefixed with the **http://** scheme.
3. The provided **registry-url** is a local-link address or localhost.
4. The configuration of the current user allows for an insecure connection. This may be caused by the user either logging in using **--insecure-skip-tls-verify** or choosing the insecure connection when prompted.



## IMPORTANT

If the registry is secured by a certificate authority different from the one used by OpenShift Container Platform, it needs to be specified using the **--certificate-authority** flag. Otherwise, the **prune** command will fail with an error similar to those listed in [Using the Wrong Certificate Authority](#) or [Using an Insecure Connection Against a Secured Registry](#).

### 18.5.3. Image Pruning Problems

#### Images Not Being Pruned

If your images keep accumulating and the **prune** command removes just a small portion of what you expect, ensure that you understand [the conditions](#) that must apply for an image to be considered a candidate for pruning.

Especially ensure that images you want removed occur at higher positions in each [tag history](#) than your chosen tag revisions threshold. For example, consider an old and obsolete image named **sha:abz**. By running the following command in namespace **N**, where the image is tagged, you will see the image is tagged three times in a single image stream named **myapp**:

```
$ image_name="sha:abz"
$ oc get is -n N -o go-template='{{range $isi, $is := .items}}{{range $ti,
$tag := $is.status.tags}}'\
  '{{range $ii, $item := $tag.items}}{{if eq $item.image
  ""$image_name}}\
    '$''}}{{$is.metadata.name}}:{{$tag.tag}} at position {{$ii}} out of {{len
$tag.items}}\n'\
  '{{end}}{{end}}{{end}}{{end}}'
myapp:v2 at position 4 out of 5
myapp:v2.1 at position 2 out of 2
myapp:v2.1-may-2016 at position 0 out of 1
```

When default options are used, the image will not ever be pruned because it occurs at position **0** in a history of **myapp:v2.1-may-2016** tag. For an image to be considered for pruning, the administrator must either:

1. Specify **--keep-tag-revisions=0** with the **oc adm prune images** command.

## CAUTION

This action will effectively remove all the tags from all the namespaces with underlying images, unless they are younger or they are referenced by objects younger than the specified threshold.

2. Delete all the [istags](#) where the position is below the revision threshold, which means **myapp:v2.1** and **myapp:v2.1-may-2016**.
3. Move the image further in the history, either by running new builds pushing to the same *istag*, or by tagging other image. Unfortunately, this is not always desirable for old release tags.

Tags having a date or time of a particular image's build in their names should be avoided, unless the image needs to be preserved for undefined amount of time. Such tags tend to have just one image in its history, which effectively prevents them from ever being pruned. [Learn more about istag naming](#).

#### Using a Secure Connection Against Insecure Registry

If you see a message similar to the following in the output of the **oc adm prune images**, then your registry is not secured and the **oc adm prune images** client will attempt to use secure connection:

```
error: error communicating with registry: Get
https://172.30.30.30:5000/healthz: http: server gave HTTP response to
HTTPS client
```

1. The recommended solution is to [secure the registry](#). If that is not desired, you can force the client to use an insecure connection by appending **--force-insecure** to the command (**not recommended**).

### 18.5.3.1. Using an Insecure Connection Against a Secured Registry

If you see one of the following errors in the output of the **oc adm prune images** command, it means that your registry is secured using a certificate signed by a certificate authority other than the one used by **oc adm prune images** client for connection verification.

```
error: error communicating with registry: Get
http://172.30.30.30:5000/healthz: malformed HTTP response
"\x15\x03\x01\x00\x02\x02"
error: error communicating with registry: [Get
https://172.30.30.30:5000/healthz: x509: certificate signed by unknown
authority, Get http://172.30.30.30:5000/healthz: malformed HTTP response
"\x15\x03\x01\x00\x02\x02"]
```

By default, the certificate authority data stored in user's configuration file are used — the same for communication with the master API.

Use the **--certificate-authority** option to provide the right certificate authority for the Docker registry server.

#### Using the Wrong Certificate Authority

The following error means that the certificate authority used to sign the certificate of the secured Docker registry is different than the authority used by the client.

```
error: error communicating with registry: Get https://172.30.30.30:5000/:
x509: certificate signed by unknown authority
```

Make sure to provide the right one with the flag **--certificate-authority**.

As a work-around, the **--force-insecure** flag can be added instead (**not recommended**).

## 18.6. HARD PRUNING THE REGISTRY

The OpenShift Container Registry can accumulate blobs that are not referenced by the OpenShift Container Platform cluster's etcd. The basic [Pruning Images](#) procedure, therefore, is unable to operate on them. These are called *orphaned blobs*.

Orphaned blobs can occur from the following scenarios:

- Manually deleting an image with **oc delete image <sha256:image-id>** command, which only removes the image from etcd, but not from the registry's storage.

- Pushing to the registry initiated by **docker** daemon failures, which causes some blobs to get uploaded, but the image manifest (which is uploaded as the very last component) does not. All unique image blobs become orphans.
- OpenShift Container Platform refusing an image because of quota restrictions.
- The standard image pruner deleting an image manifest, but is interrupted before it deletes the related blobs.
- A bug in the registry pruner, which fails to remove the intended blobs, causing the image objects referencing them to be removed and the blobs becoming orphans.

*Hard pruning* the registry, a separate procedure from basic image pruning, allows you to remove orphaned blobs. You should hard prune if you are running out of storage space in your OpenShift Container Registry and believe you have orphaned blobs.

This should be an infrequent operation and is necessary only when you have evidence that significant numbers of new orphans have been created. Otherwise, you can perform standard image pruning at regular intervals, for example, once a day (depending on the number of images being created).

To hard prune orphaned blobs from the registry:

1. **Log in:** Log in using [the CLI](#) as a user with an [access token](#).
2. **Run a basic image prune:** Basic image pruning removes additional images that are no longer needed. The hard prune does not remove images on its own. It only removes blobs stored in the registry storage. Therefore, you should run this just before the hard prune. See [Pruning Images](#) for steps.
3. **Switch the registry to read-only mode:** If the registry is not running in read-only mode, any pushes happening at the same time as the prune will either:
  - fail and cause new orphans, or
  - succeed although the images will not be pullable (because some of the referenced blobs were deleted).

Pushes will not succeed until the registry is switched back to read-write mode. Therefore, the hard prune must be carefully scheduled.

To switch the registry to read-only mode:

- a. Set the following environment variable:

```
$ oc env -n default \
    dc/docker-registry \
    'REGISTRY_STORAGE_MAINTENANCE_READONLY={"enabled":true}'
```

- b. By default, the registry should automatically redeploy when the previous step completes; wait for the redeployment to complete before continuing. However, if you have disabled these triggers, you must manually redeploy the registry so that the new environment variables are picked up:

```
$ oc rollout -n default \
    latest dc/docker-registry
```

4. **Add the `system:image-pruner` role:** The service account used to run the registry instances requires additional permissions in order to list some resources.

- a. Get the service account name:

```
$ service_account=$(oc get -n default \
  -o jsonpath='{$system:serviceaccount:{.metadata.namespace}:
  .spec.template.spec.serviceAccountName}\n' \
  dc/docker-registry)
```

- b. Add the **`system:image-pruner`** cluster role to the service account:

```
$ oc adm policy add-cluster-role-to-user \
  system:image-pruner \
  ${service_account}
```

5. **(Optional) Run the pruner in dry-run mode:** To see how many blobs would be removed, run the hard pruner in dry-run mode. No changes are actually made:

```
$ oc -n default \
  exec -i -t "$(oc -n default get pods -l deploymentconfig=docker-
  registry \
  -o jsonpath='{$.items[0].metadata.name}\n')" \
  -- /usr/bin/dockerregistry -prune=check
```

Alternatively, to get the exact paths for the prune candidates, increase the logging level:

```
$ oc -n default \
  exec "$(oc -n default get pods -l deploymentconfig=docker-
  registry \
  -o jsonpath='{$.items[0].metadata.name}\n')" \
  -- /bin/sh \
  -c 'REGISTRY_LOG_LEVEL=info /usr/bin/dockerregistry -
  prune=check'
```

### Sample Output (Truncated)

```
$ oc exec docker-registry-3-vhndw \
  -- /bin/sh -c 'REGISTRY_LOG_LEVEL=info /usr/bin/dockerregistry -
  prune=check'

time="2017-06-22T11:50:25.066156047Z" level=info msg="start prune
(dry-run mode)" distribution_version="v2.4.1+unknown"
kubernetes_version=v1.6.1+$Format:%h$ openshift_version=unknown
time="2017-06-22T11:50:25.092257421Z" level=info msg="Would delete
blob:
sha256:00043a2a5e384f6b59ab17e2c3d3a3d0a7de01b2cabeb606243e468acc663
fa5" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
time="2017-06-22T11:50:25.092395621Z" level=info msg="Would delete
blob:
sha256:0022d49612807cb348cab562c072ef34d756adfe0100a61952cbcb87ee65
78a" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
```



```

time="2017-06-22T11:50:25.092492183Z" level=info msg="Would delete
blob:
sha256:0029dd4228961086707e53b881e25eba0564fa80033fbbb2e27847a28d16a
37c" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
time="2017-06-22T11:50:26.673946639Z" level=info msg="Would delete
blob:
sha256:ff7664dfc213d6cc60fd5c5f5bb00a7bf4a687e18e1df12d349a1d07b2cf7
663" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
time="2017-06-22T11:50:26.674024531Z" level=info msg="Would delete
blob:
sha256:ff7a933178ccd931f4b5f40f9f19a65be5eeec207e4fad2a5bafd28afbef
57e" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
time="2017-06-22T11:50:26.674675469Z" level=info msg="Would delete
blob:
sha256:ff9b8956794b426cc80bb49a604a0b24a1553aae96b930c6919a6675db3d5
e06" go.version=go1.7.5 instance.id=b097121c-a864-4e0c-ad6c-
cc25f8fdf5a6
...
Would delete 13374 blobs
Would free up 2.835 GiB of disk space
Use -prune=delete to actually delete the data

```

6. **Run the hard prune:** Execute the following command inside one running instance of **docker-registry** pod to run the hard prune:

```

$ oc -n default \
  exec -i -t "$(oc -n default get pods -l deploymentconfig=docker-
registry -o jsonpath='{$.items[0].metadata.name}\n')" \
  -- /usr/bin/dockerregistry -prune=delete

```

### Sample Output

```

$ oc exec docker-registry-3-vhndw \
  -- /usr/bin/dockerregistry -prune=delete

Deleted 13374 blobs
Freed up 2.835 GiB of disk space

```

7. **Switch the registry back to read-write mode:** After the prune is finished, the registry can be switched back to read-write mode by executing:

```

$ oc env -n default dc/docker-registry
REGISTRY_STORAGE_MAINTENANCE_READONLY-

```

## 18.7. PRUNING CRON JOBS



## IMPORTANT

Cron Jobs 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/>.

At this time, the results of cron jobs are not automatically pruned. Therefore, cluster administrator should perform regular [cleanup of jobs](#), manually. We also recommend to [restrict the access](#) to cron jobs to a small group of trusted users and set appropriate [quota](#) to prevent the cron job from creating too many jobs and pods.

## CHAPTER 19. GARBAGE COLLECTION

### 19.1. OVERVIEW

The OpenShift Container Platform node performs two types of garbage collection:

- [Container garbage collection](#): Removes terminated containers.
- [Image garbage collection](#): Removes images not referenced by any running pods.

### 19.2. CONTAINER GARBAGE COLLECTION

Container garbage collection is enabled by default and happens automatically in response to eviction thresholds being reached. The node tries to keep any container for any pod accessible from the API. If the pod has been deleted, the containers will be as well. Containers are preserved as long the pod is not deleted and the eviction threshold is not reached. If the node is under disk pressure, it will remove containers and their logs will no longer be accessible via **oc logs**.

The policy for container garbage collection is based on three node settings:

| Setting                                      | Description   |
|--|---|
| <b>minimum-container-ttl-duration</b>        | The minimum age that a container is eligible for garbage collection. The default is <b>0</b> . Use <b>0</b> for no limit. Values for this setting can be specified using unit suffixes such as <b>h</b> for hour, <b>m</b> for minutes, <b>s</b> for seconds. |
| <b>maximum-dead-containers-per-container</b> | The number of instances to retain per pod container. The default is <b>1</b> .  |
| <b>maximum-dead-containers</b>               | The maximum number of total dead containers in the node. The default is <b>-1</b> , which means unlimited.  |

The **maximum-dead-containers** setting takes precedence over the **maximum-dead-containers-per-container** setting when there is a conflict. For example, if retaining the number of **maximum-dead-containers-per-container** would result in a total number of containers that is greater than **maximum-dead-containers**, the oldest containers will be removed to satisfy the **maximum-dead-containers** limit.

When the node removes the dead containers, all files inside those containers are removed as well. Only containers created by the node will be garbage collected.

You can specify values for these settings in the **kubeletArguments** section of the **/etc/origin/node/node-config.yaml** file on node hosts. Add the section if it does not already exist:

#### Container Garbage Collection Settings

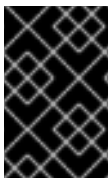
```
kubeletArguments:
  minimum-container-ttl-duration:
    - "10s"
  maximum-dead-containers-per-container:
    - "2"
```

```
maximum-dead-containers:
  - "240"
```

### 19.2.1. Detecting Containers for Deletion

Each spin of the garbage collector loop goes through the following steps:

1. Retrieve a list of available containers.
2. Filter out all containers that are running or are not alive longer than the **minimum-container-ttl-duration** parameter.
3. Classify all remaining containers into equivalence classes based on pod and image name membership.
4. Remove all unidentified containers (containers that are managed by kubelet but their name is malformed).
5. For each class that contains more containers than the **maximum-dead-containers-per-container** parameter, sort containers in the class by creation time.
6. Start removing containers from the oldest first until the **maximum-dead-containers-per-container** parameter is met.
7. If there are still more containers in the list than the **maximum-dead-containers** parameter, the collector starts removing containers from each class so the number of containers in each one is not greater than the average number of containers per class, or **<all\_remaining\_containers>/<number\_of\_classes>**.
8. If this is still not enough, sort all containers in the list and start removing containers from the oldest first until the **maximum-dead-containers** criterion is met.



#### IMPORTANT

Update the default settings to meet your needs.

Garbage collection only removes the containers that do not have a pod associated with it.

## 19.3. IMAGE GARBAGE COLLECTION

Image garbage collection relies on disk usage as reported by **cAdvisor** on the node to decide which images to remove from the node. It takes the following settings into consideration:

| Setting                        | Description  |
|--------------------------------|--|
| <b>image-gc-high-threshold</b> | The percent of disk usage (expressed as an integer) which triggers image garbage collection. The default is <b>85</b> .        |
| <b>image-gc-low-threshold</b>  | The percent of disk usage (expressed as an integer) to which image garbage collection attempts to free. Default is <b>80</b> . |

You can specify values for these settings in the **kubeletArguments** section of the */etc/origin/node/node-config.yaml* file on node hosts. Add the section if it does not already exist:

### Image Garbage Collection Settings

```
kubeletArguments:
  image-gc-high-threshold:
    - "85"
  image-gc-low-threshold:
    - "80"
```

#### 19.3.1. Detecting Images for Deletion

Two lists of images are retrieved in each garbage collector run:

1. A list of images currently running in at least one pod
2. A list of images available on a host

As new containers are run, new images appear. All images are marked with a time stamp. If the image is running (the first list above) or is newly detected (the second list above), it is marked with the current time. The remaining images are already marked from the previous spins. All images are then sorted by the time stamp.

Once the collection starts, the oldest images get deleted first until the stopping criterion is met.

## CHAPTER 20. ALLOCATING NODE RESOURCES

### 20.1. OVERVIEW

To provide more reliable scheduling and minimize node resource overcommitment, each node can reserve a portion of its resources for use by all underlying [node components](#) (e.g., kubelet, kube-proxy, Docker) and the remaining system components (e.g., **sshd**, **NetworkManager**) on the host. Once specified, the scheduler has more information about the resources (e.g., memory, CPU) a node has allocated for pods.

### 20.2. CONFIGURING NODES FOR ALLOCATED RESOURCES

Resources reserved for node components are based on two node settings:

| Setting                | Description  |
|------------------------|--|
| <b>kube-reserved</b>   | Resources reserved for node components. Default is none.                 |
| <b>system-reserved</b> | Resources reserved for the remaining system components. Default is none. |

You can set these in the **kubeletArguments** section of the [node configuration file](#) (the `/etc/origin/node/node-config.yaml` file by default) using a set of **<resource\_type>=<resource\_quantity>** pairs (e.g., **cpu=200m,memory=512Mi**). Add the section if it does not already exist:

#### Example 20.1. Node Allocatable Resources Settings

```
kubeletArguments:
  kube-reserved:
    - "cpu=200m,memory=512Mi"
  system-reserved:
    - "cpu=200m,memory=512Mi"
```

Currently, the **cpu** and **memory** resource types are supported. For **cpu**, the resource quantity is specified in units of cores (e.g., 200m, 0.5, 1). For **memory**, it is specified in units of bytes (e.g., 200Ki, 50Mi, 5Gi).

See [Compute Resources](#) for more details.

If a flag is not set, it defaults to **0**. If none of the flags are set, the allocated resource is set to the node's capacity as it was before the introduction of allocatable resources.

### 20.3. COMPUTING ALLOCATED RESOURCES

An allocated amount of a resource is computed based on the following formula:

```
[Allocatable] = [Node Capacity] - [kube-reserved] - [system-reserved] -
[Hard-Eviction-Thresholds]
```

**NOTE**

The withholding of **Hard-Eviction-Thresholds** from allocatable is a change in behavior to improve system reliability now that allocatable is enforced for end-user pods at the node level. The **experimental-allocatable-ignore-eviction** setting is available to preserve legacy behavior, but it will be deprecated in a future release.

If **[Allocatable]** is negative, it is set to **0**.

## 20.4. VIEWING NODE ALLOCATABLE RESOURCES AND CAPACITY

To see a node's current capacity and allocatable resources, you can run:

```
$ oc get node/<node_name> -o yaml
...
status:
...
  allocatable:
    cpu: "4"
    memory: 8010948Ki
    pods: "110"
  capacity:
    cpu: "4"
    memory: 8010948Ki
    pods: "110"
...
```

## 20.5. SYSTEM RESOURCES REPORTED BY NODE

Starting with OpenShift Container Platform 3.3, each node reports system resources utilized by the container runtime and kubelet. To better aid your ability to configure **--system-reserved** and **--kube-reserved**, you can introspect corresponding node's resource usage using the node summary API, which is accessible at **<master>/api/v1/nodes/<node>/proxy/stats/summary**.

For instance, to access the resources from **cluster.node22** node, you can run:

```
$ curl <certificate details>
https://<master>/api/v1/nodes/cluster.node22/proxy/stats/summary
{
  "node": {
    "nodeName": "cluster.node22",
    "systemContainers": [
      {
        "cpu": {
          "usageCoreNanoSeconds": 929684480915,
          "usageNanoCores": 190998084
        },
        "memory": {
          "rssBytes": 176726016,
          "usageBytes": 1397895168,
          "workingSetBytes": 1050509312
        },
        "name": "kubelet"
      }
    ]
  },
}
```

```

    {
      "cpu": {
        "usageCoreNanoSeconds": 128521955903,
        "usageNanoCores": 5928600
      },
      "memory": {
        "rssBytes": 35958784,
        "usageBytes": 129671168,
        "workingSetBytes": 102416384
      },
      "name": "runtime"
    }
  ]
}

```

See [REST API Overview](#) for more details about certificate details.

## 20.6. NODE ENFORCEMENT

The node is able to limit the total amount of resources that pods may consume based on the configured allocatable value. This feature significantly improves the reliability of the node by preventing pods from starving system services (for example: container runtime, node agent, etc.) for resources. It is strongly encouraged that administrators reserve resources based on the desired node utilization target in order to improve node reliability.

The node enforces resource constraints using a new cgroup hierarchy that enforces quality of service. All pods are launched in a dedicated cgroup hierarchy separate from system daemons.

To configure this ability, the following kubelet arguments are provided.

### Example 20.2. Node Cgroup Settings

```

kubeletArguments:
  cgroups-per-qos:
    - "true" ❶
  cgroup-driver:
    - "systemd" ❷
  enforce-node-allocatable:
    - "pods" ❸

```

- ❶ Enable or disable the new cgroup hierarchy managed by the node. Any change of this setting requires a full drain of the node. This flag must be true to allow the node to enforce node allocatable. We do not recommend users change this value.
- ❷ The cgroup driver used by the node when managing cgroup hierarchies. This value must match the driver associated with the container runtime. Valid values are **systemd** and **cgroupfs**. The default is **systemd**.
- ❸ A comma-delimited list of scopes for where the node should enforce node resource constraints. Valid values are **pods**, **system-reserved**, and **kube-reserved**. The default is **pods**. We do not recommend users change this value.



Optionally, the node can be made to enforce kube-reserved and system-reserved by specifying those tokens in the `enforce-node-allocatable` flag. If specified, the corresponding `--kube-reserved-cgroup` or `--system-reserved-cgroup` needs to be provided. In future releases, the node and container runtime will be packaged in a common cgroup separate from `system.slice`. Until that time, we do not recommend users change the default value of `enforce-node-allocatable` flag.

Administrators should treat system daemons similar to Guaranteed pods. System daemons can burst within their bounding control groups and this behavior needs to be managed as part of cluster deployments. Enforcing system-reserved limits can lead to critical system services being CPU starved or OOM killed on the node. The recommendation is to enforce system-reserved only if operators have profiled their nodes exhaustively to determine precise estimates and are confident in their ability to recover if any process in that group is OOM killed.

As a result, we strongly recommended that users only enforce node allocatable for **pods** by default, and set aside appropriate reservations for system daemons to maintain overall node reliability.

## 20.7. EVICTION THRESHOLDS

If a node is under memory pressure, it can impact the entire node and all pods running on it. If a system daemon is using more than its reserved amount of memory, an OOM event may occur that can impact the entire node and all pods running on it. To avoid (or reduce the probability of) system OOMs the node provides [Out Of Resource Handling](#).

By reserving some memory via the `--eviction-hard` flag, the node attempts to evict pods whenever memory availability on the node drops below the absolute value or percentage. If system daemons did not exist on a node, pods are limited to the memory **capacity - eviction-hard**. For this reason, resources set aside as a buffer for eviction before reaching out of memory conditions are not available for pods.

Here is an example to illustrate the impact of node allocatable for memory:

- Node capacity is **32Gi**
- `--kube-reserved` is **2Gi**
- `--system-reserved` is **1Gi**
- `--eviction-hard` is set to **<100Mi**.

For this node, the effective node allocatable value is **28.9Gi**. If the node and system components use up all their reservation, the memory available for pods is **28.9Gi**, and kubelet will evict pods when it exceeds this usage.

If we enforce node allocatable (**28.9Gi**) via top level cgroups, then pods can never exceed **28.9Gi**. Evictions would not be performed unless system daemons are consuming more than **3.1Gi** of memory.

If system daemons do not use up all their reservation, with the above example, pods would face memcg OOM kills from their bounding cgroup before node evictions kick in. To better enforce QoS under this situation, the node applies the hard eviction thresholds to the top-level cgroup for all pods to be **Node Allocatable + Eviction Hard Thresholds**.

If system daemons do not use up all their reservation, the node will evict pods whenever they consume more than **28.9Gi** of memory. If eviction does not occur in time, a pod will be OOM killed if pods consume **29Gi** of memory.

## 20.8. SCHEDULER

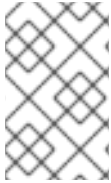
The scheduler now uses the value of **node.Status.Allocatable** instead of **node.Status.Capacity** to decide if a node will become a candidate for pod scheduling.

By default, the node will report its machine capacity as fully schedulable by the cluster.

## CHAPTER 21. OPAQUE INTEGER RESOURCES

### 21.1. OVERVIEW

Opaque integer resources allow cluster operators to provide new node-level resources that would be otherwise unknown to the system. Users can consume these resources in pod specifications, similar to CPU and memory. The scheduler performs resource accounting so that no more than the available amount is simultaneously allocated to pods.



#### NOTE

Opaque integer resources are Alpha currently, and only resource accounting is implemented. There is no resource quota or limit range support for these resources, and they have no impact on QoS.

Opaque integer resources are called *opaque* because OpenShift Container Platform does not know what the resource is, but will schedule a pod on a node only if enough of that resource is available. They are called *integer resources* because they must be available, or *advertised*, in integer amounts. The API server restricts quantities of these resources to whole numbers. Examples of *valid* quantities are **3**, **3000m**, and **3Ki**.

Opaque integer resources can be used to allocate:

- Last-level cache (LLC)
- Graphics processing unit (GPU) devices
- Field-programmable gate array (FPGA) devices
- Slots for sharing bandwidth to a parallel file system.

For example, if a node has 800 GiB of a special kind of disk storage, you could create a name for the special storage, such as ***opaque-int-resource-special-storage***. You could advertise it in chunks of a certain size, such as 100 GiB. In that case, your node would advertise that it has eight resources of type ***opaque-int-resource-special-storage***.

Opaque integer resource names must begin with the prefix **`pod.alpha.kubernetes.io/opaque-int-resource-`**.

### 21.2. CREATING OPAQUE INTEGER RESOURCES

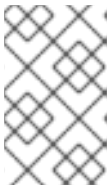
There are two steps required to use opaque integer resources. First, the cluster operator must name and advertise a per-node opaque resource on one or more nodes. Second, application developer must request the opaque resource in pods.

To make opaque integer resources available:

1. Allocate the resource and assign a name starting with **`pod.alpha.kubernetes.io/opaque-int-resource-`**
2. Advertise a new opaque integer resource by submitting a PATCH HTTP request to the API server that specifies the available quantity in the **`status.capacity`** for a node in the cluster. For example, the following HTTP request advertises five **`foo`** resources on the **`openshift-node-1`** node.

```
PATCH /api/v1/nodes/openshift-node-1/status HTTP/1.1
Accept: application/json
Content-Type: application/json-patch+json
Host: openshift-master:8080

[
  {
    "op": "add",
    "path": "/status/capacity/pod.alpha.kubernetes.io~1opaque-int-
resource-foo",
    "value": "5"
  }
]
```

**NOTE**

The **~1** in the **path** is the encoding for the character **/**. The operation path value in the JSON-Patch is interpreted as a JSON-Pointer. For more details, refer to [IETF RFC 6901, section 3](#).

After this operation, the node **status.capacity** includes a new resource. The **status.allocatable** field is updated automatically with the new resource asynchronously.

**NOTE**

Since the scheduler uses the node **status.allocatable** value when evaluating pod fitness, there might be a short delay between patching the node capacity with a new resource and the first pod that requests the resource to be scheduled on that node.

The application developer can then consume the opaque resources by editing the pod config to include the name of the opaque resource as a key in the **spec.containers[].resources.requests** field.

For example: The following pod requests two CPUs and one **foo** (an opaque resource).

```
apiVersion: v1
kind: Pod
metadata:
  name: my-pod
spec:
  containers:
  - name: my-container
    image: myimage
    resources:
      requests:
        cpu: 2
        pod.alpha.kubernetes.io/opaque-int-resource-foo: 1
```

The pod will be scheduled only if all of the resource requests are satisfied (including CPU, memory, and any opaque resources). The pod will remain in the **PENDING** state while the resource request cannot be met by any node.

Conditions:

```

    Type      Status
    PodScheduled  False
...
Events:
  FirstSeen    LastSeen  Count  From              SubObjectPath Type      Reason
Message
-----
14s          0s      6  default-scheduler  Warning    FailedScheduling  No nodes
are available that match all of the following predicates:: Insufficient
pod.alpha.kubernetes.io/opaque-int-resource-foo (1).

```

This information can also be found in the Developer Guide under [Quotas and Limit Ranges](#).

## CHAPTER 22. OVERCOMMITTING

### 22.1. OVERVIEW

Containers can specify [compute resource requests and limits](#). Requests are used for scheduling your container and provide a minimum service guarantee. Limits constrain the amount of compute resource that may be consumed on your node.

The [scheduler](#) attempts to optimize the compute resource use across all nodes in your cluster. It places pods onto specific nodes, taking the pods' compute resource requests and nodes' available capacity into consideration.

Requests and limits enable administrators to allow and manage the overcommitment of resources on a node, which may be desirable in development environments where a tradeoff of guaranteed performance for capacity is acceptable.

### 22.2. REQUESTS AND LIMITS

For each compute resource, a container may specify a resource request and limit. Scheduling decisions are made based on the request to ensure that a node has enough capacity available to meet the requested value. If a container specifies limits, but omits requests, the requests are defaulted to the limits. A container is not able to exceed the specified limit on the node.

The enforcement of limits is dependent upon the compute resource type. If a container makes no request or limit, the container is scheduled to a node with no resource guarantees. In practice, the container is able to consume as much of the specified resource as is available with the lowest local priority. In low resource situations, containers that specify no resource requests are given the lowest quality of service.

#### 22.2.1. Tune Buffer Chunk Limit

If Fluentd logger is unable to keep up with a high number of logs, you will need to increase the compute resource values.

The memory limit is used to calculate the Fluentd **buffer\_queue\_limit** as follows:

```
buffer_queue_limit = resource memory limit / (number of output *
buffer_chunk_size)
```

By default, **buffer\_chunk\_size** is 1 MB.

The following steps allow you to adjust the available resources.

1. Edit the daemonset of Fluentd:

```
$ oc edit daemonset logging-fluentd

resources:
  limits:
    cpu: 100m
    memory: 512Mi
```

2. Increase the values according to available resources. For example:

```
resources:
```

```
limits:
  cpu: 150m
  memory: 1Gi
```

If the **mux** server is behind the incoming logs, the same configuration is available. The memory limit is used to calculate the **mux buffer\_queue\_limit** as follows:

```
buffer_queue_limit = resource memory limit / (number of output *
buffer_chunk_size)
```

By default, **buffer\_chunk\_size** is 1 MB.

1. Edit the deploymentconfig of **mux**:

```
$ oc edit deploymentconfig logging-mux

resources:
  limits:
    cpu: 500m
    memory: 2Gi
```

2. Increase the values according to available resources. For example:

```
resources:
  limits:
    cpu: 600m
    memory: 2.5Gi
```

## 22.3. COMPUTE RESOURCES

The node-enforced behavior for compute resources is specific to the resource type.

### 22.3.1. CPU

A container is guaranteed the amount of CPU it requests and is additionally able to consume excess CPU available on the node, up to any limit specified by the container. If multiple containers are attempting to use excess CPU, CPU time is distributed based on the amount of CPU requested by each container.

For example, if one container requested 500m of CPU time and another container requested 250m of CPU time, then any extra CPU time available on the node is distributed among the containers in a 2:1 ratio. If a container specified a limit, it will be throttled not to use more CPU than the specified limit.

CPU requests are enforced using the CFS shares support in the Linux kernel. By default, CPU limits are enforced using the CFS quota support in the Linux kernel over a 100ms measuring interval, though [this can be disabled](#).

### 22.3.2. Memory

A container is guaranteed the amount of memory it requests. A container may use more memory than requested, but once it exceeds its requested amount, it could be killed in a low memory situation on the node.

If a container uses less memory than requested, it will not be killed unless system tasks or daemons need more memory than was accounted for in the node's resource reservation. If a container specifies a limit on memory, it is immediately killed if it exceeds the limit amount.

## 22.4. QUALITY OF SERVICE CLASSES

A node is *overcommitted* when it has a pod scheduled that makes no request, or when the sum of limits across all pods on that node exceeds available machine capacity.

In an overcommitted environment, it is possible that the pods on the node will attempt to use more compute resource than is available at any given point in time. When this occurs, the node must give priority to one pod over another. The facility used to make this decision is referred to as a Quality of Service (QoS) Class.

For each compute resource, a container is divided into one of three QoS classes with decreasing order of priority:

**Table 22.1. Quality of Service Classes**

| Priority    | Class Name        | Description  |
|-------------|-------------------|--|
| 1 (highest) | <b>Guaranteed</b> | If limits and optionally requests are set (not equal to 0) for all resources and they are equal, then the container is classified as <b>Guaranteed</b> .     |
| 2           | <b>Burstable</b>  | If requests and optionally limits are set (not equal to 0) for all resources, and they are not equal, then the container is classified as <b>Burstable</b> . |
| 3 (lowest)  | <b>BestEffort</b> | If requests and limits are not set for any of the resources, then the container is classified as <b>BestEffort</b> .   |

Memory is an incompressible resource, so in low memory situations, containers that have the lowest priority are killed first:

- **Guaranteed** containers are considered top priority, and are guaranteed to only be killed if they exceed their limits, or if the system is under memory pressure and there are no lower priority containers that can be evicted.
- **Burstable** containers under system memory pressure are more likely to be killed once they exceed their requests and no other **BestEffort** containers exist.
- **BestEffort** containers are treated with the lowest priority. Processes in these containers are first to be killed if the system runs out of memory.

## 22.5. CONFIGURING MASTERS FOR OVERCOMMITMENT

Scheduling is based on resources requested, while quota and hard limits refer to resource limits, which can be set higher than requested resources. The difference between request and limit determines the level of overcommit; for instance, if a container is given a memory request of 1Gi and a memory limit of 2Gi, it is scheduled based on the 1Gi request being available on the node, but could use up to 2Gi; so it is 200% overcommitted.

If OpenShift Container Platform administrators would like to control the level of overcommit and manage



container density on nodes, masters can be configured to override the ratio between request and limit set on developer containers. In conjunction with a [per-project LimitRange](#) specifying limits and defaults, this adjusts the container limit and request to achieve the desired level of overcommit.

This requires configuring the **ClusterResourceOverride** admission controller in the **master-config.yaml** as in the following example (reuse the existing configuration tree if it exists, or introduce absent elements as needed):

```
admissionConfig:
  pluginConfig:
    ClusterResourceOverride: ❶
    configuration:
      apiVersion: v1
      kind: ClusterResourceOverrideConfig
      memoryRequestToLimitPercent: 25 ❷
      cpuRequestToLimitPercent: 25 ❸
      limitCPUToMemoryPercent: 200 ❹
```

- ❶ This is the plug-in name; case matters and anything but an exact match for a plug-in name is ignored.
- ❷ (optional, 1-100) If a container memory limit has been specified or defaulted, the memory request is overridden to this percentage of the limit.
- ❸ (optional, 1-100) If a container CPU limit has been specified or defaulted, the CPU request is overridden to this percentage of the limit.
- ❹ (optional, positive integer) If a container memory limit has been specified or defaulted, the CPU limit is overridden to a percentage of the memory limit, with a 100 percentage scaling 1Gi of RAM to equal 1 CPU core. This is processed prior to overriding CPU request (if configured).

After changing the master configuration, a master restart is required.

Note that these overrides have no effect if no limits have been set on containers. [Create a LimitRange object](#) with default limits (per individual project, or in the [project template](#)) in order to ensure that the overrides apply.

Note also that after overrides, the container limits and requests must still be validated by any [LimitRange](#) objects in the project. It is possible, for example, for developers to specify a limit close to the minimum limit, and have the request then be overridden below the minimum limit, causing the pod to be forbidden. This unfortunate user experience should be addressed with future work, but for now, configure this capability and [LimitRanges](#) with caution.

When configured, overrides can be disabled per-project (for example, to allow infrastructure components to be configured independently of overrides) by editing the project and adding the following annotation:

```
quota.openshift.io/cluster-resource-override-enabled: "false"
```

## 22.6. CONFIGURING NODES FOR OVERCOMMITMENT

In an overcommitted environment, it is important to properly configure your node to provide best system behavior.

## 22.6.1. Reserving Memory Across Quality of Service Tiers

You can use the **experimental-qos-reserved** parameter to specify a percentage of memory to be reserved by a pod in a particular QoS level. This feature attempts to reserve requested resources to exclude pods from lower QoS classes from using resources requested by pods in higher QoS classes.

By reserving resources for higher QoS levels, pods that don't have resource limits are prevented from encroaching on the resources requested by pods at higher QoS levels.

### IMPORTANT

The **experimental-qos-reserved** parameter 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/>.

To configure **experimental-qos-reserved**, edit the `/etc/origin/node/node-config.yaml` file for the node.

```
kubeletArguments:
  cgroups-per-qos:
    - true
  cgroup-driver:
    - 'systemd'
  cgroup-root:
    - '/'
  experimental-qos-reserved: 1
    - 'memory=50%'
```

- 1 Specifies how pod resource requests are reserved at the QoS level.

OpenShift Container Platform uses the **experimental-qos-reserved** parameter as follows:

- A value of **experimental-qos-reserved=memory=100%** will prevent the **Burstable** and **BestEffort** QoS classes from consuming memory that was requested by a higher QoS class. This increases the risk of inducing OOM on **BestEffort** and **Burstable** workloads in favor of increasing memory resource guarantees for **Guaranteed** and **Burstable** workloads.
- A value of **experimental-qos-reserved=memory=50%** will allow the **Burstable** and **BestEffort** QoS classes to consume half of the memory requested by a higher QoS class.
- A value of **experimental-qos-reserved=memory=0%** will allow a **Burstable** and **BestEffort** QoS classes to consume up to the full node allocatable amount if available, but increases the risk that a **Guaranteed** workload will not have access to requested memory. This condition effectively disables this feature.

## 22.6.2. Enforcing CPU Limits

Nodes by default enforce specified CPU limits using the CPU CFS quota support in the Linux kernel. If you do not want to enforce CPU limits on the node, you can disable its enforcement by modifying the [node configuration file](#) (the **node-config.yaml** file) to include the following:

```
kubeletArguments:
  cpu-cfs-quota:
    - "false"
```

If CPU limit enforcement is disabled, it is important to understand the impact that will have on your node:

- If a container makes a request for CPU, it will continue to be enforced by CFS shares in the Linux kernel.
- If a container makes no explicit request for CPU, but it does specify a limit, the request will default to the specified limit, and be enforced by CFS shares in the Linux kernel.
- If a container specifies both a request and a limit for CPU, the request will be enforced by CFS shares in the Linux kernel, and the limit will have no impact on the node.

### 22.6.3. Reserving Resources for System Processes

The [scheduler](#) ensures that there are enough resources for all pods on a node based on the pod requests. It verifies that the sum of requests of containers on the node is no greater than the node capacity. It includes all containers started by the node, but not containers or processes started outside the knowledge of the cluster.

It is recommended that you reserve some portion of the node capacity to allow for the system daemons that are required to run on your node for your cluster to function (**sshd**, **docker**, etc.). In particular, it is recommended that you reserve resources for incompressible resources such as memory.

If you want to explicitly reserve resources for non-pod processes, there are two ways to do so:

- The preferred method is to allocate node resources by specifying resources available for scheduling. See [Allocating Node Resources](#) for more details.
- Alternatively, you can create a **resource-reserver** pod that does nothing but reserve capacity from being scheduled on the node by the cluster. For example:

#### Example 22.1. resource-reserver Pod Definition

```
apiVersion: v1
kind: Pod
metadata:
  name: resource-reserver
spec:
  containers:
    - name: sleep-forever
      image: gcr.io/google_containers/pause:0.8.0
      resources:
        limits:
          cpu: 100m 1
          memory: 150Mi 2
```

- 1 The amount of CPU to reserve on a node for host-level daemons unknown to the cluster.

- 2 The amount of memory to reserve on a node for host-level daemons unknown to the cluster.

You can save your definition to a file, for example ***resource-reserver.yaml***, then place the file in the node configuration directory, for example ***/etc/origin/node/*** or the **`--config=<dir>`** location if otherwise specified.

Additionally, the node server needs to be configured to read the definition from the node configuration directory, by naming the directory in the **`kubeletArguments.config`** field of the [node configuration file](#) (usually named ***node-config.yaml***):

```
kubeletArguments:
  config:
    - "/etc/origin/node" 1
```

- 1 If **`--config=<dir>`** is specified, use **`<dir>`** here.

With the ***resource-reserver.yaml*** file in place, starting the node server also launches the **sleep-forever** container. The scheduler takes into account the remaining capacity of the node, adjusting where to place cluster pods accordingly.

To remove the **resource-reserver** pod, you can delete or move the ***resource-reserver.yaml*** file from the node configuration directory.

#### 22.6.4. Kernel Tunable Flags

When the node starts, it ensures that the kernel tunable flags for memory management are set properly. The kernel should never fail memory allocations unless it runs out of physical memory.

To ensure this behavior, the node instructs the kernel to always overcommit memory:

```
$ sysctl -w vm.overcommit_memory=1
```

The node also instructs the kernel not to panic when it runs out of memory. Instead, the kernel OOM killer should kill processes based on priority:

```
$ sysctl -w vm.panic_on_oom=0
```



#### NOTE

The above flags should already be set on nodes, and no further action is required.

#### 22.6.5. Disabling Swap Memory

You can disable swap by default on your nodes in order to preserve quality of service guarantees. Otherwise, physical resources on a node can oversubscribe, affecting the resource guarantees the Kubernetes scheduler makes during pod placement.

For example, if two guaranteed pods have reached their memory limit, each container could start using swap memory. Eventually, if there is not enough swap space, processes in the pods can be terminated due to the system being oversubscribed.

To disable swap:

```
$ swapoff -a
```

Failing to disable swap results in nodes not recognizing that they are experiencing **MemoryPressure**, resulting in pods not receiving the memory they made in their scheduling request. As a result, additional pods are placed on the node to further increase memory pressure, ultimately increasing your risk of experiencing a system out of memory (OOM) event.



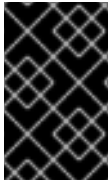
### IMPORTANT

If swap is enabled, any [out of resource handling](#) eviction thresholds for available memory will not work as expected. Take advantage of out of resource handling to allow pods to be evicted from a node when it is under memory pressure, and rescheduled on an alternative node that has no such pressure.

## CHAPTER 23. ASSIGNING UNIQUE EXTERNAL IPS FOR INGRESS TRAFFIC

### 23.1. OVERVIEW

One approach to getting external traffic into the cluster is by using ExternalIP or IngressIP addresses.



#### IMPORTANT

This feature is only supported in non-cloud deployments. For cloud (GCE, AWS, and OpenStack) deployments, use the Load Balancer services for automatic deployment of a cloud load balancer to target the service's endpoints.

OpenShift Container Platform supports two pools of IP addresses:

- IngressIP uses by the Loadbalancer when choosing an external IP address for the service.
- ExternalIP is used when the user selects a specific IP from the configured pool.



#### NOTE

Both have to be configured to a device on an OpenShift Container Platform host to be used, whether with network interface controller (NIC) or virtual ethernet, as well as external routing. Ipfailover is recommended for this, because it selects the host and configures the NIC.

IngressIP and ExternalIP both allow external traffic access to the cluster, and, if routed correctly, external traffic can reach that service's endpoints via any TCP/UDP port the service exposes. This can be simpler than having to manage the port space of a limited number of shared IP addresses when manually assigning external IPs to services. Also, these addresses can be used as virtual IPs (VIPs) when configuring [high availability](#).

OpenShift Container Platform supports both the automatic and manual assignment of IP addresses, and each address is guaranteed to be assigned to a maximum of one service. This ensures that each service can expose its chosen ports regardless of the ports exposed by other services.

### 23.2. RESTRICTIONS

To use an **ExternalIP**, you can:

- Select an IP address from the [externalIPNetworkCIDRs](#) range.
- Have an IP address assigned from the [ingressIPNetworkCIDR](#) pool in the master configuration file. In this case, OpenShift Container Platform implements a non-cloud version of the load balancer service type and assigns IP addresses to the services.

#### CAUTION

You must ensure that the IP address pool you assign terminates at one or more nodes in your cluster. You can use the existing [oc adm ipfailover](#) to ensure that the external IPs are highly available.

For manually-configured external IPs, potential port clashes are handled on a first-come, first-served basis. If you request a port, it is only available if it has not yet been assigned for that IP address. For example:

### Port clash example for manually-configured external IPs

Two services have been manually configured with the same external IP address of 172.7.7.7.

**MongoDB service A** requests port 27017, and then **MongoDB service B** requests the same port; the first request gets the port.

However, port clashes are not an issue for external IPs assigned by the ingress controller, because the controller assigns each service a unique address.

## 23.3. CONFIGURING THE CLUSTER TO USE UNIQUE EXTERNAL IPS

In non-cloud clusters, `ingressIPNetworkCIDR` is set by default to `172.29.0.0/16`. If your cluster environment is not already using this private range, you can use the default. However, if you want to use a different range, then you must set `ingressIPNetworkCIDR` in the `/etc/origin/master/master-config.yaml` file before you assign an ingress IP. Then, restart the master service.

### CAUTION

External IPs assigned to services of type **LoadBalancer** will always be in the range of `ingressIPNetworkCIDR`. If `ingressIPNetworkCIDR` is changed such that the assigned external IPs are no longer in range, the affected services will be assigned new external IPs compatible with the new range.



### NOTE

If you are using [high availability](#), then this range must be less than 255 IP addresses.

### Sample `/etc/origin/master/master-config.yaml`

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

### 23.3.1. Configuring an Ingress IP for a Service

To assign an ingress IP:

1. Create a YAML file for a LoadBalancer service that requests a specific IP via the `loadBalancerIP` setting:

#### Sample LoadBalancer Configuration

```
apiVersion: v1
kind: Service
metadata:
  name: egress-1
spec:
  ports:
    - name: db
```

```

    port: 3306
    loadBalancerIP: 172.29.0.1
    type: LoadBalancer
    selector:
      name: my-db-selector

```

2. Create a LoadBalancer service on your pod:

```
$ oc create -f loadbalancer.yaml
```

3. Check the service for an external IP. For example, for a service named **myservice**:

```
$ oc get svc myservice
```

When your LoadBalancer-type service has an external IP assigned, the output displays the IP:

| NAME      | CLUSTER-IP    | EXTERNAL-IP | PORT(S)  | AGE |
|-----------|---------------|-------------|----------|-----|
| myservice | 172.30.74.106 | 172.29.0.1  | 3306/TCP | 30s |

## 23.4. ROUTING THE INGRESS CIDR FOR DEVELOPMENT OR TESTING

Add a static route directing traffic for the ingress CIDR to a node in the cluster. For example:

```
# route add -net 172.29.0.0/16 gw 10.66.140.17 eth0
```

In the example above, **172.29.0.0/16** is the **ingressIPNetworkCIDR**, and **10.66.140.17** is the node IP.

### 23.4.1. Service externalIPs

In addition to the cluster's internal IP addresses, the application developer can configure IP addresses that are external to the cluster. As the OpenShift Container Platform administrator, you are responsible for ensuring that traffic arrives at a node with this IP.

The externalIPs must be selected by the administrator from the **externalIPNetworkCIDRs** range configured in the [master-config.yaml](#) file. When **master-config.yaml** changes, the master services must be restarted.

```
# systemctl restart atomic-openshift-master-api atomic-openshift-master-controllers
```

#### Sample externalIPNetworkCIDR /etc/origin/master/master-config.yaml

```
networkConfig:
  externalIPNetworkCIDR: 172.47.0.0/24
```

#### 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" : [
      "80.11.12.10"
    ]
  }
}
```

- 1** List of External IP addresses on which the **port** is exposed. In addition to the internal IP addresses)

## CHAPTER 24. HANDLING OUT OF RESOURCE ERRORS

### 24.1. OVERVIEW

This topic discusses best-effort attempts to prevent OpenShift Container Platform from experiencing out-of-memory (OOM) and out-of-disk-space conditions.

A node must maintain stability when available compute resources are low. This is especially important when dealing with incompressible resources such as memory or disk. If either resource is exhausted, the node becomes unstable.

Administrators can proactively monitor nodes for and prevent against situations where the node runs out of compute and memory resources using configurable [eviction policies](#).

This topic also provides information on how OpenShift Container Platform handles out-of-resource conditions and provides an [example scenario](#) and [recommended practices](#):

- [Resource reclaiming](#)
- [Pod eviction](#)
- [Pod scheduling](#)
- [Out of Resource and Out of Memory Killer](#)



#### WARNING

If swap memory is enabled for a node, that node cannot detect that it is under **MemoryPressure**.

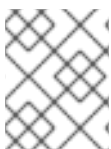
To take advantage of memory based evictions, operators must [disable swap](#).

### 24.2. CONFIGURING EVICTION POLICIES

An *eviction policy* allows a node to fail one or more pods when the node is running low on available resources. Failing a pod allows the node to reclaim needed resources.

An eviction policy is a combination of an [eviction trigger signal](#) with a specific [eviction threshold value](#) that is set in the node configuration file or through the [command line](#). Evictions can be either [hard](#), where a node takes immediate action on a pod that exceeds a threshold, or [soft](#), where a node allows a grace period before taking action.

By using well-configured eviction policies, a node can proactively monitor for and prevent against total starvation of a compute resource.



#### NOTE

When the node fails a pod, it terminates all containers in the pod, and the **PodPhase** is transitioned to **Failed**.

When detecting disk pressure, the node supports the **nodefs** and **imagefs** file system partitions.

The **nodefs**, or **rootfs**, is the file system that the node uses for local disk volumes, daemon logs, emptyDir, and so on (for example, the file system that provides `/`). The **rootfs** contains **openshift.local.volumes**, by default `/var/lib/origin/openshift.local.volumes`.

The **imagefs** is the file system that the container runtime uses for storing images and individual container-writable layers. Eviction thresholds are at 85% full for **imagefs**. The **imagefs** file system depends on the runtime and, in the case of Docker, which storage driver you are using.

- For Docker:
    - If you are using the **devicemapper** storage driver, the **imagefs** is thin pool. You can limit the read/write layer for the container by setting the `--storage-opt dm.basesize` flag in the Docker daemon.
- ```
$ sudo dockerd --storage-opt dm.basesize=50G
```
- If you are using the **overlay2** storage driver, the **imagefs** is the file system that contains `/var/lib/docker/overlay2`.
- For CRI-O, which uses the overlay driver, the **imagefs** is `/var/lib/containers/storage` by default.



## NOTE

If you do not use local storage isolation (ephemeral storage) and not using XFS quota (volumeConfig), you cannot limit local disk usage by the pod.

### 24.2.1. Using the Node Configuration to Create a Policy

To configure an eviction policy, edit the node configuration file (the `/etc/origin/node/node-config.yaml` file) to specify the eviction thresholds under the **eviction-hard** or **eviction-soft** parameters.

For example:

#### Example 24.1. Sample Node Configuration file for a hard eviction

```
kubeletArguments:
  eviction-hard: 1
    - memory.available<100Mi 2
    - nodefs.available<10%
    - nodefs.inodesFree<5%
    - imagefs.available<15%
    - imagefs.inodesFree<10%
```

- 1 The type of eviction: Use this parameter for a [hard eviction](#).
- 2 Eviction thresholds based on a specific eviction trigger signal.

**NOTE**

You must provide percentage values for the **inodesFree** parameters. You can provide a percentage or a numerical value for the other parameters.

**Example 24.2. Sample Node Configuration file for a soft eviction**

```
kubeletArguments:
  eviction-soft: ❶
    - memory.available<100Mi ❷
    - nodefs.available<10%
    - nodefs.inodesFree<5%
    - imagefs.available<15%
    - imagefs.inodesFree<10%
  eviction-soft-grace-period: ❸
    - memory.available=1m30s
    - nodefs.available=1m30s
    - nodefs.inodesFree=1m30s
    - imagefs.available=1m30s
    - imagefs.inodesFree=1m30s
```

- ❶ The type of eviction: Use this parameter for a [soft eviction](#).
- ❷ An eviction threshold based on a specific eviction trigger signal.
- ❸ The grace period for the soft eviction. Leave the default values for optimal performance.

1. Restart the OpenShift Container Platform service for the changes to take effect:

```
# systemctl restart atomic-openshift-node
```

**24.2.2. Understanding Eviction Signals**

You can configure a node to trigger eviction decisions on any of the signals described in the table below. You add an eviction signal to an [eviction threshold](#) along with a threshold value.

The value of each signal is described in the **Description** column based on the node summary API.

To view the signals:

```
curl <certificate details> \
  https://<master>/api/v1/nodes/<node>/proxy/stats/summary
```

**Table 24.1. Supported Eviction Signals**

| Node Condition | Eviction Signal | Value | Description |
|----------------|-----------------|-------|-------------|
|----------------|-----------------|-------|-------------|

| Node Condition        | Eviction Signal           | Value                                                                                 | Description                                                                                                       |
|-----------------------|---------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| <b>MemoryPressure</b> | <b>memory.available</b>   | <b>memory.available = node.status.capacity[memory] - node.stats.memory.workingSet</b> | Available memory on the node has exceeded an eviction threshold.                                                  |
| <b>DiskPressure</b>   | <b>nodefs.available</b>   | <b>nodefs.available = node.stats.fs.available</b>                                     | Available disk space on either the node root file system or image file system has exceeded an eviction threshold. |
|                       | <b>nodefs.inodesFree</b>  | <b>nodefs.inodesFree = node.stats.fs.inodesFree</b>                                   |                                                                                                                   |
|                       | <b>imagefs.available</b>  | <b>imagefs.available = node.stats.runtime.imagefs.available</b>                       |                                                                                                                   |
|                       | <b>imagefs.inodesFree</b> | <b>imagefs.inodesFree = node.stats.runtime.imagefs.inodesFree</b>                     |                                                                                                                   |

Each of the above signals supports either a literal or percentage-based value. The percentage-based value is calculated relative to the total capacity associated with each signal.

A script derives the value for **memory.available** from your cgroup driver using the same set of steps that the kubelet performs. The script excludes inactive file memory (that is, the number of bytes of file-backed memory on inactive LRU list) from its calculation as it assumes that inactive file memory is reclaimable under pressure.



#### NOTE

Do not use tools like **free -m**, because **free -m** does not work in a container.

OpenShift Container Platform monitors these file systems every 10 seconds.

If you store volumes and logs in a dedicated file system, the node will not monitor that file system.



#### NOTE

As of OpenShift Container Platform 3.4, the node supports the ability to trigger eviction decisions based on disk pressure. Operators must opt-in to enable disk-based evictions. Prior to evicting pods due to disk pressure, the node also performs [container and image garbage collection](#). In future releases, garbage collection will be deprecated in favor of a pure disk-eviction based configuration.

### 24.2.3. Understanding Eviction Thresholds

You can configure a node to specify eviction thresholds, which triggers the node to reclaim resources, by adding a threshold to the [node configuration file](#).

If an eviction threshold is met, independent of its associated grace period, the node reports a condition indicating that the node is under memory or disk pressure. This prevents the scheduler from scheduling any additional pods on the node while attempts to reclaim resources are made.

The node continues to report node status updates at the frequency specified by the **node-status-update-frequency** argument, which defaults to **10s** (ten seconds).

Eviction thresholds can be [hard](#), for when the node takes immediate action when a threshold is met, or [soft](#), for when you allow a grace period before reclaiming resources.



#### NOTE

Soft eviction usage is more common when you are targeting a certain level of utilization, but can tolerate temporary spikes. We recommended setting the soft eviction threshold lower than the hard eviction threshold, but the time period can be operator-specific. The system reservation should also cover the soft eviction threshold.

The soft eviction threshold is an advanced feature. You should configure a hard eviction threshold before attempting to use soft eviction thresholds.

Thresholds are configured in the following form:

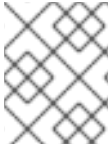
```
<eviction_signal><operator><quantity>
```

- the **eviction-signal** value can be any [supported eviction signal](#).

- the **operator** value is `<`.
- the **quantity** value must match the [quantity representation](#) used by Kubernetes and can be expressed as a percentage if it ends with the `%` token.

For example, if an operator has a node with 10Gi of memory, and that operator wants to induce eviction if available memory falls below 1Gi, an eviction threshold for memory can be specified as either of the following:

```
memory.available<1Gi
memory.available<10%
```



#### NOTE

The node evaluates and monitors eviction thresholds every 10 seconds and the value can not be modified. This is the housekeeping interval.

### 24.2.3.1. Understanding Hard Eviction Thresholds

A hard eviction threshold has no grace period and, if observed, the node takes immediate action to reclaim the associated starved resource. If a hard eviction threshold is met, the node kills the pod immediately with no graceful termination.

To configure hard eviction thresholds, add eviction thresholds to the [node configuration file](#) under **eviction-hard**, as shown in [Using the Node Configuration to Create a Policy](#).

#### Sample Node Configuration file with hard eviction thresholds

```
kubeletArguments:
  eviction-hard:
    - memory.available<500Mi
    - nodefs.available<500Mi
    - nodefs.inodesFree<100Mi
    - imagefs.available<100Mi
    - imagefs.inodesFree<100Mi
```

This example is a general guideline and not recommended settings.

### 24.2.3.2. Understanding Soft Eviction Thresholds

A soft eviction threshold pairs an eviction threshold with a required administrator-specified grace period. The node does not reclaim resources associated with the eviction signal until that grace period is exceeded. If no grace period is provided in the node configuration the node errors on startup.

In addition, if a soft eviction threshold is met, an operator can specify a maximum allowed pod termination grace period to use when evicting pods from the node. If **eviction-max-pod-grace-period** is specified, the node uses the lesser value among the **pod.Spec.TerminationGracePeriodSeconds** and the maximum-allowed grace period. If not specified, the node kills pods immediately with no graceful termination.

For soft eviction thresholds the following flags are supported:

- **eviction-soft**: a set of eviction thresholds (for example, **memory.available<1.5Gi**) that, if met over a corresponding grace period, triggers a pod eviction.

- **eviction-soft-grace-period:** a set of eviction grace periods (for example, **memory.available=1m30s**) that correspond to how long a soft eviction threshold must hold before triggering a pod eviction.
- **eviction-max-pod-grace-period:** the maximum-allowed grace period (in seconds) to use when terminating pods in response to a soft eviction threshold being met.

To configure soft eviction thresholds, add eviction thresholds to the [node configuration file](#) under **eviction-soft**, as shown in [Using the Node Configuration to Create a Policy](#).

### Sample Node Configuration files with soft eviction thresholds

```
kubeletArguments:
  eviction-soft:
    - memory.available<500Mi
    - nodefs.available<500Mi
    - nodefs.inodesFree<100Mi
    - imagefs.available<100Mi
    - imagefs.inodesFree<100Mi
  eviction-soft-grace-period:
    - memory.available=1m30s
    - nodefs.available=1m30s
    - nodefs.inodesFree=1m30s
    - imagefs.available=1m30s
    - imagefs.inodesFree=1m30s
```

This example is a general guideline and not recommended settings.

## 24.3. CONFIGURING THE AMOUNT OF RESOURCE FOR SCHEDULING

You can control how much of a node resource is made available for scheduling in order to allow the scheduler to fully allocate a node and to prevent evictions.

Set **system-reserved** equal to the amount of resource you want available to the scheduler for deploying pods and for system-daemons. Evictions should only occur if pods use more than their requested amount of an allocatable resource.

A node reports two values:

- **Capacity:** How much resource is on the machine
- **Allocatable:** How much resource is made available for scheduling.

To configure the amount of allocatable resources:

1. Edit the node configuration file (the **/etc/origin/node/node-config.yaml** file) to add or modify the **system-reserved** parameter for **eviction-hard** or **eviction-soft**.

```
kubeletArguments:
  eviction-hard: 1
    - "memory.available<500Mi"
  system-reserved:
    - "memory=1.5Gi"
```



- 1 This threshold can either be **eviction-hard** or **eviction-soft**.

2. Restart the OpenShift Container Platform service for the changes to take effect:

```
# systemctl restart atomic-openshift-node
```

## 24.4. CONTROLLING NODE CONDITION OSCILLATION

If a node is oscillating above and below a soft eviction threshold, but not exceeding its associated grace period, the corresponding node condition oscillates between **true** and **false**, which can cause problems for the scheduler.

To prevent this oscillation, set the **eviction-pressure-transition-period** parameter to control how long the node must wait before transitioning out of a pressure condition.

1. Edit or add the parameter to the **kubeletArguments** section of the node configuration file (the */etc/origin/node/node-config.yaml*) using a set of **<resource\_type>=<resource\_quantity>** pairs.

```
kubeletArguments:
  eviction-pressure-transition-period="5m"
```

+ The node toggles the condition back to **false** when the node has not observed an eviction threshold being met for the specified pressure condition for the specified period.

+



### NOTE

Use the default value (5 minutes) before doing any adjustments. The default choice is intended to allow the system to stabilize, and to prevent the scheduler from assigning new pods to the node before it has settled.

1. Restart the OpenShift Container Platform services for the changes to take effect:

```
# systemctl restart atomic-openshift-node
```

## 24.5. RECLAIMING NODE-LEVEL RESOURCES

If an eviction criteria is satisfied, the node initiates the process of reclaiming the pressured resource until the signal goes below the defined threshold. During this time, the node does not support scheduling any new pods.

The node attempts to reclaim node-level resources prior to evicting end-user pods, based on whether the host system has a dedicated **imagefs** configured for the container runtime.

### With Imagefs

If the host system has **imagefs**:

- If the **nodefs** file system meets eviction thresholds, the node frees up disk space in the following order:

- Delete dead pods/containers
- If the **imagefs** file system meets eviction thresholds, the node frees up disk space in the following order:
  - Delete all unused images

### Without Imagefs

If the host system does not have **imagefs**:

- If the **nodefs** file system meets eviction thresholds, the node frees up disk space in the following order:
  - Delete dead pods/containers
  - Delete all unused images

## 24.6. UNDERSTANDING POD EVICTION

If an eviction threshold is met and the grace period is passed, the node initiates the process of evicting pods until the signal goes below the defined threshold.

The node ranks pods for eviction by their [quality of service](#), and, among those with the same quality of service, by the consumption of the starved compute resource relative to the pod's scheduling request.

Each QOS level has an OOM score, which the Linux out-of-memory tool (OOM killer) uses to determine which pods to kill. See [Understanding Quality of Service and Out of Memory Killer](#) below.

The following table lists each QOS level and the associated OOM score.

**Table 24.2. Quality of Service Levels**

| Quality of Service | Description                                                                                                                                                                                                                       |
|--------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Guaranteed</b>  | Pods that consume the highest amount of the starved resource relative to their request are failed first. If no pod has exceeded its request, the strategy targets the largest consumer of the starved resource.                   |
| <b>Burstable</b>   | Pods that consume the highest amount of the starved resource relative to their request for that resource are failed first. If no pod has exceeded its request, the strategy targets the largest consumer of the starved resource. |
| <b>BestEffort</b>  | Pods that consume the highest amount of the starved resource are failed first.                                                                                                                                                    |

A **Guaranteed** pod will never be evicted because of another pod's resource consumption unless a system daemon (such as node, **docker**, **journald**) is consuming more resources than were reserved using **system-reserved**, or **kube-reserved** allocations or if the node has only **Guaranteed** pods remaining.

If the node has only **Guaranteed** pods remaining, the node evicts a **Guaranteed** pod that least impacts node stability and limits the impact of the unexpected consumption to other **Guaranteed** pods.

Local disk is a **BestEffort** resource. If necessary, the node evicts pods one at a time to reclaim disk when **DiskPressure** is encountered. The node ranks pods by quality of service. If the node is

responding to inode starvation, it will reclaim inodes by evicting pods with the lowest quality of service first. If the node is responding to lack of available disk, it will rank pods within a quality of service that consumes the largest amount of local disk, and evict those pods first.

### 24.6.1. Understanding Quality of Service and Out of Memory Killer

If the node experiences a system out of memory (OOM) event before it is able to reclaim memory, the node depends on the OOM killer to respond.

The node sets a **oom\_score\_adj** value for each container based on the quality of service for the pod.

**Table 24.3. Quality of Service Levels**

| Quality of Service | oom_score_adj Value                                                                                 |
|--------------------|-----------------------------------------------------------------------------------------------------|
| <b>Guaranteed</b>  | -998                                                                                                |
| <b>Burstable</b>   | $\min(\max(2, 1000 - (1000 * \text{memoryRequestBytes}) / \text{machineMemoryCapacityBytes}), 999)$ |
| <b>BestEffort</b>  | 1000                                                                                                |

If the node is unable to reclaim memory prior to experiencing a system OOM event, the **oom\_killer** calculates an **oom\_score**:

```
% of node memory a container is using + `oom_score_adj` = `oom_score`
```

The node then kills the container with the highest score.

Containers with the lowest quality of service that are consuming the largest amount of memory relative to the scheduling request are failed first.

Unlike pod eviction, if a pod container is OOM failed, it can be restarted by the node based on the node [restart policy](#).

## 24.7. UNDERSTANDING THE POD SCHEDULER AND OOR CONDITIONS

The scheduler views node conditions when placing additional pods on the node. For example, if the node has an eviction threshold like the following:

```
eviction-hard is "memory.available<500Mi"
```

and available memory falls below 500Mi, the node reports a value in **Node.Status.Conditions** as **MemoryPressure** as true.

**Table 24.4. Node Conditions and Scheduler Behavior**

| Node Condition | Scheduler Behavior |
|----------------|--------------------|
|----------------|--------------------|

| Node Condition        | Scheduler Behavior                                                                                  |
|-----------------------|-----------------------------------------------------------------------------------------------------|
| <b>MemoryPressure</b> | If a node reports this condition, the scheduler will not place <b>BestEffort</b> pods on that node. |
| <b>DiskPressure</b>   | If a node reports this condition, the scheduler will not place any additional pods on that node.    |

## 24.8. EXAMPLE SCENARIO

Consider the following scenario.

An operator:

- has a node with a memory capacity of **10Gi**;
- wants to reserve 10% of memory capacity for system daemons (kernel, node, etc.);
- wants to evict pods at 95% memory utilization to reduce thrashing and incidence of system OOM.

Implicit in this configuration is the understanding that **system-reserved** should include the amount of memory covered by the eviction threshold.

To reach that capacity, either some pod is using more than its request, or the system is using more than **1Gi**.

If a node has 10 Gi of capacity, and you want to reserve 10% of that capacity for the system daemons (**system-reserved**), perform the following calculation:

```
capacity = 10 Gi
system-reserved = 10 Gi * .1 = 1 Gi
```

The amount of allocatable resources becomes:

```
allocatable = capacity - system-reserved = 9 Gi
```

This means by default, the scheduler will schedule pods that request 9 Gi of memory to that node.

If you want to turn on eviction so that eviction is triggered when the node observes that available memory falls below 10% of capacity for 30 seconds, or immediately when it falls below 5% of capacity, you need the scheduler to see allocatable as 8Gi. Therefore, ensure your system reservation covers the greater of your eviction thresholds.

```
capacity = 10 Gi
eviction-threshold = 10 Gi * .1 = 1 Gi
system-reserved = (10Gi * .1) + eviction-threshold = 2 Gi
allocatable = capacity - system-reserved = 8 Gi
```

Enter the following in the **node-config.yaml**:

```
kubeletArguments:
```

```
system-reserved:
- "memory=2Gi"
eviction-hard:
- "memory.available<.5Gi"
eviction-soft:
- "memory.available<1Gi"
eviction-soft-grace-period:
- "memory.available=30s"
```

This configuration ensures that the scheduler does not place pods on a node that immediately induce memory pressure and trigger eviction assuming those pods use less than their configured request.

## 24.9. RECOMMENDED PRACTICE

### 24.9.1. DaemonSets and Out of Resource Handling

If a node evicts a pod that was created by a DaemonSet, the pod will immediately be recreated and rescheduled back to the same node, because the node has no ability to distinguish a pod created from a DaemonSet versus any other object.

In general, DaemonSets should not create **BestEffort** pods to avoid being identified as a candidate pod for eviction. Instead DaemonSets should ideally launch **Guaranteed** pods.

## CHAPTER 25. MONITORING AND DEBUGGING ROUTERS

### 25.1. OVERVIEW

Depending on the underlying implementation, you can monitor a running [router](#) in multiple ways. This topic discusses the HAProxy template router and the components to check to ensure its health.

### 25.2. VIEWING STATISTICS

The HAProxy router exposes a web listener for the HAProxy statistics. Enter the router's public IP address and the correctly configured port (**1936** by default) to view the statistics page. The administrator password and port are configured during the router installation, but they can be found by viewing the *haproxy.config* file on the container.

To view HAProxy router stats:

1. If needed, [create the router](#) using **--stats-port** to expose statistics on the specified port:

```
$ oc adm router <name> --replicas=<number> --service-account=router
--selector='<zone>' --stats-port=<port>
```

For example:

```
$ oc adm router router --replicas=1 --selector='zone=west' --stats-
port=1936
```



#### NOTE

**-replicas** is the replication factor of the router; commonly 2 when high availability is desired.

**-selector** is used to filter nodes on deployment. Used to run routers on a specific set of nodes.

If you receive the following error, run the **oc adm policy add-scc-to-user hostnetwork -z <name>** command as suggested.

```
error: router could not be created; service account "router" is not
allowed to access the host network on nodes, grant access with oc
adm policy add-scc-to-user hostnetwork -z router
```

2. Run the following command to get the router pod name:

```
oc get pod
NAME          READY    STATUS    RESTARTS   AGE
router-1-deploy 0/1      Pending   0           48s
```

3. If needed, run the following command to get the stats password that was assigned during router creation:

```
$ oc describe pod <pod-name> |grep STATS_PASSWORD
```

For example:

```
oc describe pod router-1-lt7xm |grep STATS_PASSWORD
STATS_PASSWORD:      C1dSdUff00
```

4. Run the following command to open iptables for the router stats port:

```
$ iptables -I OS_FIREWALL_ALLOW -p tcp -m tcp --dport 1936 -j ACCEPT
```

You can add this rule to **/etc/sysconfig/iptables** to keep the rule across reboots.

You can make this port accessible via [port forwarding](#), if desired.

5. Use one of the following methods to view statistics:

- To launch the stats window,
  - a. Enter the following command to unset the **ROUTER\_METRICS\_TYPE** environment variable:

```
oc env dc router ROUTER_METRICS_TYPE-
```

b. Use the following URL in a browser:

```
http://admin:<stats-password>@<master-ip>:1936/metrics
```

For example:

```
http://admin:C1dSdUff00@master.example.com:1936/metrics
```

The statistics display in a table similar to the following:

[illegible]

- To generate a CSV output of the statistics, execute the following command on master:

```
$ cmd="echo 'show stat' | socat - UNIX-
```

```
CONNECT:/var/lib/haproxy/run/haproxy.sock"
$ routerPod=$(oc get pods --selector="router=router" \
  --template="{{with index .items 0}}{{.metadata.name}}
  {{end}}")
$ oc exec $routerPod -- bash -c "$cmd"
```

Statistics are output similar to the following:

```
#
pxname,svname,qcur,qmax,scur,smax,slim,stot,bin,bout,dreq,dresp,e
req,econ,eresp,wretr,wredis,status,weight,act,bck,chkfail,chkdown
,lastchg,downtime,qlimit,pid,iid,sid,throttle,lbtot,tracked,type,
rate,rate_lim,rate_max,check_status,check_code,check_duration,hrs
p_1xx,hrsp_2xx,hrsp_3xx,hrsp_4xx,hrsp_5xx,hrsp_other,hanafail,req
_rate,req_rate_max,req_tot,cli_abrt,srv_abrt,comp_in,comp_out,com
p_byp,comp_rsp,lastsess,last_chk,last_agt,qtime,ctime,rtime,ttime
/
stats,FRONTEND,,,0,3,20000,10273,1254027,1600580,0,0,0,,,,,OPEN,,
,,,,,1,2,0,,,,0,0,0,2,,,,0,10274,0,1,0,0,,0,3,10275,,,0,0,0,0,,
,,,,,
stats,BACKEND,0,0,0,0,2000,0,1254027,1600580,0,0,,0,0,0,0,UP,0,0,
0,,0,51350,0,,1,2,0,,0,,1,0,,0,,,,0,0,0,0,0,0,,,,0,0,0,0,0,0,266
,,,0,0,0,1,
public,FRONTEND,,,0,0,20000,0,0,0,0,0,0,,,,,OPEN,,,,,,,,,1,3,0,,,
,0,0,0,0,,,,0,0,0,0,0,0,,0,0,0,,0,0,0,0,,,,,
public_ssl,FRONTEND,,,0,0,20000,0,0,0,0,0,0,,,,,OPEN,,,,,,,,,1,4,
0,,,0,0,0,0,,,,,0,0,0,,0,0,0,0,,,,,
be_sni,fe_sni,0,0,0,0,,0,0,0,,0,,0,0,0,0,no
check,1,1,0,,,,,1,5,1,,0,,2,0,,0,,,,,0,,,,0,0,,,,-
1,,,0,0,0,0,
be_sni,BACKEND,0,0,0,0,2000,0,0,0,0,0,,0,0,0,0,UP,1,1,0,,0,51350,
0,,1,5,0,,0,,1,0,,0,,,,,0,0,0,0,0,0,-1,,,0,0,0,0,
fe_sni,FRONTEND,,,0,0,20000,0,0,0,0,0,0,,,,,OPEN,,,,,,,,,1,6,0,,,
,0,0,0,0,,,,0,0,0,0,0,0,,0,0,0,,0,0,0,0,,,,,
be_no_sni,fe_no_sni,0,0,0,0,,0,0,0,,0,,0,0,0,0,no
check,1,1,0,,,,,1,7,1,,0,,2,0,,0,,,,,0,,,,0,0,,,,-
1,,,0,0,0,0,
be_no_sni,BACKEND,0,0,0,0,2000,0,0,0,0,0,,0,0,0,0,UP,1,1,0,,0,513
50,0,,1,7,0,,0,,1,0,,0,,,,,0,0,0,0,0,0,-1,,,0,0,0,0,
fe_no_sni,FRONTEND,,,0,0,20000,0,0,0,0,0,0,,,,,OPEN,,,,,,,,,1,8,0
,,,0,0,0,0,,,,0,0,0,0,0,0,,0,0,0,,0,0,0,0,,,,,
openshift_default,BACKEND,0,0,0,0,6000,0,0,0,0,0,,0,0,0,0,UP,0,0,
0,,0,51350,0,,1,9,0,,0,,1,0,,0,,,,0,0,0,0,0,0,,,,0,0,0,0,0,0,-
1,,,0,0,0,0,
be_tcp_default_docker-
registry,5b5da52795f08ded1114814facd16158,0,0,0,0,,0,0,0,,0,0,0,
0,0,UP,100,1,0,0,0,51350,0,,1,10,1,,0,,2,0,,0,L40K,,0,,,,,0,,,,
0,0,,,,-1,,,0,0,0,0,
be_tcp_default_docker-
registry,BACKEND,0,0,0,0,1,0,0,0,0,0,,0,0,0,0,UP,100,1,0,,0,51350
,0,,1,10,0,,0,,1,0,,0,,,,,0,0,0,0,0,0,-1,,,0,0,0,0,
be_tcp_default_registry-
console,12d16a9727abf73501f11715b016bffd,0,0,0,0,,0,0,0,,0,,0,0,0
,0,UP,100,1,0,0,0,51350,0,,1,11,1,,0,,2,0,,0,L40K,,0,,,,,0,,,,0
,0,,,,-1,,,0,0,0,0,
```



```
be_tcp_default_registry-
console,BACKEND,0,0,0,0,1,0,0,0,0,0,,0,0,0,0,UP,100,1,0,,0,51350,
0,,1,11,0,,0,,1,0,,0,,,,,,,,,,,,,0,0,0,0,0,0,-1,,0,0,0,0,
```



### IMPORTANT

For security purposes, the **oc exec** command does not work when accessing privileged containers. Instead, you can SSH into a node host, then use the **docker exec** command on the desired container.

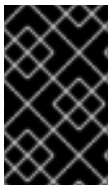
## 25.3. DISABLING STATISTICS VIEW

By default the HAProxy statistics are exposed on port **1936** (with a password protected account). To disable exposing the HAProxy statistics, specify **0** as the stats port number.

```
$ oc adm router hap --service-account=router --stats-port=0
```

Note: HAProxy will still collect and store statistics, it would just *not* expose them via a web listener. You can still get access to the statistics by sending a request to the HAProxy AF\_UNIX socket inside the HAProxy Router container.

```
$ cmd="echo 'show stat' | socat - UNIX-
CONNECT:/var/lib/haproxy/run/haproxy.sock"
$ routerPod=$(oc get pods --selector="router=router" \
  --template="{{with index .items 0}}{{.metadata.name}}{{end}}")
$ oc exec $routerPod -- bash -c "$cmd"
```



### IMPORTANT

For security purposes, the **oc exec** command does not work when accessing privileged containers. Instead, you can SSH into a node host, then use the **docker exec** command on the desired container.

## 25.4. VIEWING LOGS

To view a router log, run the **oc logs** command on the pod. Since the router is running as a plug-in process that manages the underlying implementation, the log is for the plug-in, not the actual HAProxy log.

To view the logs generated by HAProxy, start a syslog server and pass the location to a router pod using the following environment variables.

**Table 25.1. Router Syslog Variables**

| Environment Variable                    | Description                                                                                                                                                      |
|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>ROUTER_SYSLOG_ADDR</b><br><b>ESS</b> | The IP address of the syslog server. Port <b>514</b> is the default if no port is specified.                                                                     |
| <b>ROUTER_LOG_LEVEL</b>                 | Optional. Set to change the HAProxy log level. If not set, the default log level is <b>warning</b> . This can be changed to any log level that HAProxy supports. |

| Environment Variable        | Description                                                                                                               |
|-----------------------------|---------------------------------------------------------------------------------------------------------------------------|
| <b>ROUTER_SYSLOG_FORMAT</b> | Optional. Set to define customized HAProxy log format. This can be changed to any log format string that HAProxy accepts. |

To set a running router pod to send messages to a syslog server:

```
$ oc set env dc/router ROUTER_SYSLOG_ADDRESS=<dest_ip:dest_port>
ROUTER_LOG_LEVEL=<level>
```

For example, the following sets HAProxy to send logs to 127.0.0.1 with the default port **514** and changes the log level to **debug**.

```
$ oc set env dc/router ROUTER_SYSLOG_ADDRESS=127.0.0.1
ROUTER_LOG_LEVEL=debug
```

## 25.5. VIEWING THE ROUTER INTERNALS

### routes.json

Routes are processed by the HAProxy router, and are stored both in memory, on disk, and in the HAProxy configuration file. The internal route representation, which is passed to the template to generate the HAProxy configuration file, is found in the **`/var/lib/haproxy/router/routes.json`** file. When troubleshooting a routing issue, view this file to see the data being used to drive configuration.

### HAProxy configuration

You can find the HAProxy configuration and the backends that have been created for specific routes in the **`/var/lib/haproxy/conf/haproxy.config`** file. The mapping files are found in the same directory. The helper frontend and backends use mapping files when mapping incoming requests to a backend.

### Certificates

Certificates are stored in two places:

- Certificates for edge terminated and re-encrypt terminated routes are stored in the **`/var/lib/haproxy/router/certs`** directory.
- Certificates that are used for connecting to backends for re-encrypt terminated routes are stored in the **`/var/lib/haproxy/router/cacerts`** directory.

The files are keyed by the namespace and name of the route. The key, certificate, and CA certificate are concatenated into a single file. You can use [OpenSSL](#) to view the contents of these files.

## CHAPTER 26. HIGH AVAILABILITY

### 26.1. OVERVIEW

This topic describes setting up high availability for pods and services on your OpenShift Container Platform cluster.

IP failover manages a pool of Virtual IP (VIP) addresses on a set of nodes. Every VIP in the set will be serviced by a node selected from the set. As long as a single node is available, the VIPs will be served. There is no way to explicitly distribute the VIPs over the nodes, so there may be nodes with no VIPs and other nodes with many VIPs. If there is only one node, all VIPs will be on it.



#### NOTE

The VIPs must be routable from outside the cluster.

IP failover monitors a port on each VIP to determine whether the port is reachable on the node. If the port is not reachable, the VIP will not be assigned to the node. If the port is set to **0**, this check is suppressed. [The `check` script](#) does the needed testing.

IP failover uses [Keepalived](#) to host a set of externally accessible VIP addresses on a set of hosts. Each VIP is only serviced by a single host at a time. **Keepalived** uses the VRRP protocol to determine which host (from the set of hosts) will service which VIP. If a host becomes unavailable or if the service that **Keepalived** is watching does not respond, the VIP is switched to another host from the set. Thus, a VIP is always serviced as long as a host is available.

When a host running **Keepalived** passes the **check** script, the host can become in the **MASTER** state based on its priority and the priority of the current **MASTER**, as determined by the [preemption strategy](#).

The administrator can provide a script via the `--notify-script=` option, which is called whenever the state changes. **Keepalived** is in **MASTER** state when it is servicing the VIP, in **BACKUP** state when another node is servicing the VIP, or in **FAULT** state when the **check** script fails. The [notify script](#) is called with the new state whenever the state changes.

OpenShift Container Platform supports creation of IP failover deployment configuration, by running the `oc adm ipfailover` command. The IP failover deployment configuration specifies the set of VIP addresses, and the set of nodes on which to service them. A cluster can have multiple IP failover deployment configurations, with each managing its own set of unique VIP addresses. Each node in the IP failover configuration runs an IP failover pod, and this pod runs **Keepalived**.

When using VIPs to access a pod with host networking (e.g. a router), the application pod should be running on all nodes that are running the ipfailover pods. This enables any of the ipfailover nodes to become the master and service the VIPs when needed. If application pods are not running on all nodes with ipfailover, either some ipfailover nodes will never service the VIPs or some application pods will never receive any traffic. Use the same selector and replication count, for both ipfailover and the application pods, to avoid this mismatch.

While using VIPs to access a service, any of the nodes can be in the ipfailover set of nodes, since the service is reachable on all nodes (no matter where the application pod is running). Any of the ipfailover nodes can become master at any time. The service can either use external IPs and a service port or it can use a nodePort.

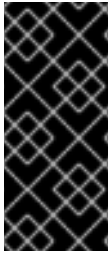
When using external IPs in the service definition the VIPs are set to the external IPs and the ipfailover monitoring port is set to the service port. A nodePort is open on every node in the cluster and the service will load balance traffic from whatever node currently supports the VIP. In this case, the ipfailover

monitoring port is set to the nodePort in the service definition.



### IMPORTANT

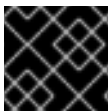
Setting up a nodePort is a privileged operation.



### IMPORTANT

Even though a service VIP is highly available, performance can still be affected. **keepalived** makes sure that each of the VIPs is serviced by some node in the configuration, and several VIPs may end up on the same node even when other nodes have none. Strategies that externally load balance across a set of VIPs may be thwarted when ipfailover puts multiple VIPs on the same node.

When you use ingressIP, you can set up ipfailover to have the same VIP range as the ingressIP range. You can also disable the monitoring port. In this case, all the VIPs will appear on same node in the cluster. Any user can set up a service with an ingressIP and have it highly available.

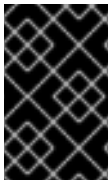


### IMPORTANT

There are a maximum of 255 VIPs in the cluster.

## 26.2. CONFIGURING IP FAILOVER

Use the **oc adm ipfailover** command with suitable [options](#), to create ipfailover deployment configuration.



### IMPORTANT

Currently, ipfailover is not compatible with cloud infrastructures. For AWS, an Elastic Load Balancer (ELB) can be used to make OpenShift Container Platform highly available, [using the AWS console](#).

As an administrator, you can configure ipfailover on an entire cluster, or on a subset of nodes, as defined by the label selector. You can also configure multiple IP failover deployment configurations in your cluster, where each one is independent of the others. The **oc adm ipfailover** command creates an ipfailover deployment configuration which ensures that a failover pod runs on each of the nodes matching the constraints or the label used. This pod runs **Keepalived** which uses VRRP (Virtual Router Redundancy Protocol) among all the **Keepalived** daemons to ensure that the service on the watched port is available, and if it is not, **Keepalived** will automatically float the VIPs.

For production use, make sure to use a **--selector=<label>** with at least two nodes to select the nodes. Also, set a **--replicas=<n>** value that matches the number of nodes for the given labeled selector.

The **oc adm ipfailover** command includes command line options that set environment variables that control **Keepalived**. The [environment variables](#) start with **OPENSIFT\_HA\_\*** and they can be changed as needed.

For example, the command below will create an IP failover configuration on a selection of nodes labeled **router=us-west-ha** (on 4 nodes with 7 virtual IPs monitoring a service listening on port 80, such as the router process).

```
$ oc adm ipfailover --selector="router=us-west-ha" \
  --virtual-ips="1.2.3.4,10.1.1.100-104,5.6.7.8" \
  --watch-port=80 --replicas=4 --create
```

### 26.2.1. Virtual IP Addresses

**Keepalived** manages a set of virtual IP addresses. The administrator must make sure that all these addresses:

- Are accessible on the configured hosts from outside the cluster.
- Are not used for any other purpose within the cluster.

**Keepalived** on each node determines whether the needed service is running. If it is, VIPs are supported and **Keepalived** participates in the negotiation to determine which node will serve the VIP. For a node to participate, the service must be listening on the watch port on a VIP or the check must be disabled.



#### NOTE

Each VIP in the set may end up being served by a different node.

### 26.2.2. Check and Notify Scripts

**Keepalived** monitors the health of the application by periodically running an optional user supplied check script. For example, the script can test a web server by issuing a request and verifying the response.

The script is provided through the `--check-script=<script>` option to the `oc adm ipfailover` command. The script must exit with `0` for **PASS** or `1` for **FAIL**.

By default, the check is done every two seconds, but can be changed using the `--check-interval=<seconds>` option.

When a check script is not provided, a simple default script is run that tests the [TCP connection](#). This default test is suppressed when the monitor port is `0`.

For each VIP, **keepalived** keeps the state of the node. The VIP on the node may be in **MASTER**, **BACKUP**, or **FAULT** state. All VIPs on the node that are not in the **FAULT** state participate in the negotiation to decide which will be **MASTER** for the VIP. All of the losers enter the **BACKUP** state. When the **check** script on the **MASTER** fails, the VIP enters the **FAULT** state and triggers a renegotiation. When the **BACKUP** fails, the VIP enters the **FAULT** state. When the **check** script passes again on a VIP in the **FAULT** state, it exits **FAULT** and negotiates for **MASTER**. The resulting state is either **MASTER** or **BACKUP**.

The administrator can provide an optional **notify** script, which is called whenever the state changes. **Keepalived** passes the following three parameters to the script:

- **\$1** - "GROUP"|"INSTANCE"
- **\$2** - Name of the group or instance
- **\$3** - The new state ("MASTER"|"BACKUP"|"FAULT")

These scripts run in the IP failover pod and use the pod's file system, not the host file system. The options require the full path to the script. The administrator must make the script available in the pod to

extract the results from running the **notify** script. The recommended approach for providing the scripts is to use a [ConfigMap](#).

The full path names of the **check** and **notify** scripts are added to the **keepalived** configuration file, **/etc/keepalived/keepalived.conf**, which is loaded every time **keepalived** starts. The scripts can be added to the pod with a ConfigMap as follows.

1. Create the desired script and create a ConfigMap to hold it. The script has no input arguments and must return **0** for **OK** and **1** for **FAIL**.

The check script, **mycheckscript.sh**:

```
#!/bin/bash
# Whatever tests are needed
# E.g., send request and verify response
exit 0
```

2. Create the ConfigMap:

```
$ oc create configmap mycustomcheck --from-file=mycheckscript.sh
```

3. There are two approaches to adding the script to the pod: use **oc** commands or edit the deployment configuration. In both cases, the **defaultMode** for the mounted **configMap** files must allow execution. A value of **0755** (**493** decimal) is typical.

- a. Using **oc** commands:

```
$ oc env dc/ipf-ha-router \
    OPENSIFT_HA_CHECK_SCRIPT=/etc/keepalive/mycheckscript.sh
$ oc volume dc/ipf-ha-router --add --overwrite \
    --name=config-volume \
    --mount-path=/etc/keepalive \
    --source='{ "configMap": { "name": "mycustomcheck",
    "defaultMode": 493}}'
```

- b. Editing the **ipf-ha-router** deployment configuration:

- i. Use **oc edit dc ipf-ha-router** to edit the router deployment configuration with a text editor.

```
...
spec:
  containers:
    - env:
      - name: OPENSIFT_HA_CHECK_SCRIPT ①
        value: /etc/keepalive/mycheckscript.sh
    ...
    volumeMounts: ②
      - mountPath: /etc/keepalive
        name: config-volume
    dnsPolicy: ClusterFirst
    ...
  volumes: ③
    - configMap:
        defaultMode: 0755 ④
```

```

        name: customrouter
        name: config-volume
    ...

```

- 1 In the `spec.container.env` field, add the `OPENSIFT_HA_CHECK_SCRIPT` environment variable to point to the mounted script file.
- 2 Add the `spec.container.volumeMounts` field to create the mount point.
- 3 Add a new `spec.volumes` field to mention the ConfigMap.
- 4 This sets execute permission on the files. When read back, it will be displayed in decimal (**493**).

ii. Save the changes and exit the editor. This restarts **ipf-ha-router**.

### 26.2.3. VRRP Preemption

When a host leaves the **FAULT** state by passing the check script, the host becomes a **BACKUP** if the new host has lower priority than the host currently in the **MASTER** state. However, if it has a higher priority, the preemption strategy determines its role in the cluster.

The **nopreempt** strategy does not move **MASTER** from the lower priority host to the higher priority host. With **preempt 300**, the default, **keepalived** waits the specified 300 seconds and moves **MASTER** to the higher priority host.

To specify preemption:

- a. When creating ipfailover using the **preemption-strategy**:

```

$ oc adm ipfailover --preempt-strategy=nopreempt \
...

```

- b. Setting the variable using the **oc set env** command:

```

$ oc set env dc/ipf-ha-router \
  --overwrite=true \
  OPENSIFT_HA_PREEMPTION=nopreempt

```

- c. Using **oc edit dc ipf-ha-router** to edit the router deployment configuration:

```

...
spec:
  containers:
  - env:
    - name: OPENSIFT_HA_PREEMPTION 1
      value: nopreempt
...

```

### 26.2.4. Keepalived Multicast

OpenShift Container Platform's IP failover internally uses **keepalived**.

**IMPORTANT**

Ensure that **multicast** is enabled on the nodes labeled above and they can accept network traffic for 224.0.0.18 (the VRRP multicast IP address).

Before starting the **keepalived** daemon, the startup script verifies the **iptables** rule that allows multicast traffic to flow. If there is no such rule, the startup script creates a new rule and adds it to the IP tables configuration. Where this new rule gets added to the IP tables configuration depends on the **--iptables-chain=** option. If there is an **--iptables-chain=** option specified, the rule gets added to the specified chain in the option. Otherwise, the rule is added to the **INPUT** chain.

**IMPORTANT**

The **iptables** rule must be present whenever there is one or more **keepalived** daemon running on the node.

The **iptables** rule can be removed after the last **keepalived** daemon terminates. The rule is not automatically removed.

You can manually manage the **iptables** rule on each of the nodes. It only gets created when none is present (as long as ipfailover is not created with the **--iptables-chain=""** option).

**IMPORTANT**

You must ensure that the manually added rules persist after a system restart.

Be careful since every **keepalived** daemon uses the VRRP protocol over multicast 224.0.0.18 to negotiate with its peers. There must be a different VRRP-id (in the range 0..255) for [each VIP](#).

```
$ for node in openshift-node-{5,6,7,8,9}; do    ssh $node <<EOF

export interface=${interface:-"eth0"}
echo "Check multicast enabled ... ";
ip addr show $interface | grep -i MULTICAST

echo "Check multicast groups ... "
ip maddr show $interface | grep 224.0.0

EOF
done;
```

## 26.2.5. Command Line Options and Environment Variables

Table 26.1. Command Line Options and Environment Variables

| Option | Variable Name | Default | Notes |
|--------|---------------|---------|-------|
|--------|---------------|---------|-------|



| Option                 | Variable Name                 | Default        | Notes                                                                                                                                                                                                                          |
|------------------------|-------------------------------|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| -- watch-port          | OPENSIFT_HA_MONITOR_PORT      | 80             | The ipfailover pod tries to open a TCP connection to this port on each VIP. If connection is established, the service is considered to be running. If this port is set to 0, the test always passes.                           |
| -- interface           | OPENSIFT_HA_NETWORK_INTERFACE |                | The interface name for ipfailover to use, to send VRRP traffic. By default, <b>eth0</b> is used.                                                                                                                               |
| -- replicas            | OPENSIFT_HA_REPLICA_COUNT     | 2              | Number of replicas to create. This must match <b>spec.replicas</b> value in ipfailover deployment configuration.                                                                                                               |
| -- virtual-ips         | OPENSIFT_HA_VIRTUAL_IPS       |                | The list of IP address ranges to replicate. This must be provided. (For example, 1.2.3.4-6,1.2.3.9.) See <a href="#">this discussion</a> for more details.                                                                     |
| -- vrrp-id-offset      | OPENSIFT_HA_VRRP_ID_OFFSET    | 0              | See <a href="#">VRRP ID Offset</a> discussion for more details.                                                                                                                                                                |
| -- iptables-chain      | OPENSIFT_HA_IPTABLES_CHAIN    | INPUT          | The name of the iptables chain, to automatically add an <b>iptables</b> rule to allow the VRRP traffic on. If the value is not set, an <b>iptables</b> rule will not be added. If the chain does not exist, it is not created. |
| -- check-script        | OPENSIFT_HA_CHECK_SCRIPT      |                | Full path name in the pod file system of a script that is periodically run to verify the application is operating. See <a href="#">this discussion</a> for more details.                                                       |
| -- check-interval      | OPENSIFT_HA_CHECK_INTERVAL    | 2              | The period, in seconds, that the check script is run.                                                                                                                                                                          |
| -- notify-script       | OPENSIFT_HA_NOTIFY_SCRIPT     |                | Full path name in the pod file system of a script that is run whenever the state changes. See <a href="#">this discussion</a> for more details.                                                                                |
| -- preemption-strategy | OPENSIFT_HA_PREEMPTION        | preempt<br>300 | Strategy for handling a new higher priority host. See <a href="#">the VRRP Preemption section</a> for more details.                                                                                                            |

## 26.2.6. VRRP ID Offset

Each ipfailover pod managed by the ipfailover deployment configuration (1 pod per node/replica) runs a **keepalived** daemon. As more ipfailover deployment configurations are configured, more pods are created and more daemons join into the common VRRP negotiation. This negotiation is done by all the **keepalived** daemons and it determines which nodes will service which VIPs.

Internally, **keepalived** assigns a unique vrrp-id to each VIP. The negotiation uses this set of vrrp-ids, when a decision is made, the VIP corresponding to the winning vrrp-id is serviced on the winning node.

Therefore, for every VIP defined in the ipfailover deployment configuration, the ipfailover pod must assign a corresponding vrrp-id. This is done by starting at **--vrrp-id-offset** and sequentially assigning the vrrp-ids to the list of VIPs. The vrrp-ids may have values in the range 1..255.

When there are multiple ipfailover deployment configuration care must be taken to specify **--vrrp-id-offset** so that there is room to increase the number of VIPS in the deployment configuration and none of the vrrp-id ranges overlap.

## 26.2.7. Configuring a Highly-available Service

The following example describes how to set up highly-available **router** and **geo-cache** network services with IP failover on a set of nodes.

1. Label the nodes that will be used for the services. This step can be optional if you run the services on all the nodes in your OpenShift Container Platform cluster and will use VIPs that can float within all nodes in the cluster.

The following example defines a label for nodes that are servicing traffic in the US west geography **ha-svc-nodes=geo-us-west**:

```
$ oc label nodes openshift-node-{5,6,7,8,9} "ha-svc-nodes=geo-us-west"
```

2. Create the service account. You can use ipfailover or when using a router (depending on your environment policies), you can either reuse the **router** service account created previously or a new ipfailover service account.

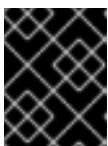
The following example creates a new service account with the name ipfailover in the **default** namespace:

```
$ oc create serviceaccount ipfailover -n default
```

3. Add the ipfailover service account in the **default** namespace to the **privileged** SCC:

```
$ oc adm policy add-scc-to-user privileged
system:serviceaccount:default:ipfailover
```

4. Start the **router** and the **geo-cache** services.



### IMPORTANT

Since the ipfailover runs on all nodes from step 1, it is recommended to also run the router/service on all the step 1 nodes.

- a. Start the router with the nodes matching the labels used in the first step. The following example runs five instances using the `ipfailover` service account:

```
$ oc adm router ha-router-us-west --replicas=5 \
  --selector="ha-svc-nodes=geo-us-west" \
  --labels="ha-svc-nodes=geo-us-west" \
  --service-account=ipfailover
```

- b. Run the **geo-cache** service with a replica on each of the nodes. See an [example configuration](#) for running a **geo-cache** service.



### IMPORTANT

Make sure that you replace the **myimages/geo-cache** Docker image referenced in the file with your intended image. Change the number of replicas to the number of nodes in the **geo-cache** label. Check that the label matches the one used in the first step.

```
$ oc create -n <namespace> -f ./examples/geo-cache.json
```

5. Configure `ipfailover` for the **router** and **geo-cache** services. Each has its own VIPs and both use the same nodes labeled with **ha-svc-nodes=geo-us-west** in the first step. Ensure that the number of replicas match the number of nodes listed in the label setup, in the first step.



### IMPORTANT

The **router**, **geo-cache**, and `ipfailover` all create deployment configuration and all must have different names.

6. Specify the VIPs and the port number that `ipfailover` should monitor on the desired instances. The `ipfailover` command for the **router**:

```
$ oc adm ipfailover ipf-ha-router-us-west \
  --replicas=5 --watch-port=80 \
  --selector="ha-svc-nodes=geo-us-west" \
  --virtual-ips="10.245.2.101-105" \
  --iptables-chain="INPUT" \
  --service-account=ipfailover --create
```

The following is the `oc adm ipfailover` command for the **geo-cache** service that is listening on port 9736. Since there are two **ipfailover** deployment configurations, the `--vrrp-id-offset` must be set so that each VIP gets its own offset. In this case, setting a value of **10** means that the **ipf-ha-router-us-west** can have a maximum of 10 VIPs (0-9) since **ipf-ha-geo-cache** is starting at 10.

```
$ oc adm ipfailover ipf-ha-geo-cache \
  --replicas=5 --watch-port=9736 \
  --selector="ha-svc-nodes=geo-us-west" \
  --virtual-ips=10.245.3.101-105 \
  --vrrp-id-offset=10 \
  --service-account=ipfailover --create
```

In the commands above, there are **ipfailover**, **router**, and **geo-cache** pods on each node. The set of VIPs for each ipfailover configuration must not overlap and they must not be used elsewhere in the external or cloud environments. The five VIP addresses in each example, **10.245.2.101-105** are served by the two ipfailover deployment configurations. IP failover dynamically selects which address is served on which node.

The administrator sets up external DNS to point to the VIP addresses knowing that all the **router** VIPs point to the same **router**, and all the **geo-cache** VIPs point to the same **geo-cache** service. As long as one node remains running, all the VIP addresses are served.

### 26.2.7.1. Deploy IP Failover Pod

Deploy the ipfailover router to monitor postgresql listening on node port 32439 and the external IP address, as defined in the **postgresql-ingress** service:

```
$ oc adm ipfailover ipf-ha-postgresql \
  --replicas=1 \ 1
  --selector="app-type=postgresql" \ 2
  --virtual-ips=10.9.54.100 \ 3
  --watch-port=32439 \ 4
  --service-account=ipfailover --create
```

- 1 1 Specifies the number of instances to deploy.
- 2 Restricts where the ipfailover is deployed.
- 3 Virtual IP address to monitor.
- 4 Port on which ipfailover will monitor on each node.

### 26.2.8. Dynamically Updating Virtual IPs for a Highly-available Service

The default deployment strategy for the IP failover service is to recreate the deployment. In order to dynamically update the VIPs for a highly available routing service with minimal or no downtime, you must:

- Update the IP failover service deployment configuration to use a rolling update strategy, and
- Update the **OPENSHIFT\_HA\_VIRTUAL\_IPS** environment variable with the updated list or sets of virtual IP addresses.

The following example shows how to dynamically update the deployment strategy and the virtual IP addresses:

1. Consider an IP failover configuration that was created using the following:

```
$ oc adm ipfailover ipf-ha-router-us-west \
  --replicas=5 --watch-port=80 \
  --selector="ha-svc-nodes=geo-us-west" \
  --virtual-ips="10.245.2.101-105" \
  --service-account=ipfailover --create
```

2. Edit the deployment configuration:

■

```
$ oc edit dc/ipf-ha-router-us-west
```

- Update the `spec.strategy.type` field from **Recreate** to **Rolling**:

```
spec:
  replicas: 5
  selector:
    ha-svc-nodes: geo-us-west
  strategy:
    recreateParams:
      timeoutSeconds: 600
    resources: {}
    type: Rolling ❶
```

- ❶ Set to **Rolling**.

- Update the `OPENSIFT_HA_VIRTUAL_IPS` environment variable to contain the additional virtual IP addresses:

```
- name: OPENSIFT_HA_VIRTUAL_IPS
  value: 10.245.2.101-105,10.245.2.110,10.245.2.201-205 ❶
```

- ❶ **10.245.2.110, 10.245.2.201-205** have been added to the list.

- Update the external DNS to match the set of VIPs.

## 26.3. CONFIGURING SERVICE EXTERNALIP AND NODEPORT

The user can assign VIPs as [ExternalIPs](#) in a service. **Keepalived** makes sure that each VIP is served on some node in the ipfailover configuration. When a request arrives on the node, the service that is running on all nodes in the cluster, load balances the request among the service's endpoints.

The [NodePorts](#) can be set to the ipfailover watch port so that **keepalived** can check the application is running. The NodePort is exposed on all nodes in the cluster, therefore it is available to **keepalived** on all ipfailover nodes.

## 26.4. HIGH AVAILABILITY FOR INGRESSIP

In non-cloud clusters, ipfailover and [ingressIP](#) to a service can be combined. The result is high availability services for users that create services using ingressIP.

The approach is to specify an `ingressIPNetworkCIDR` range and then use the same range in creating the ipfailover configuration.

Since, ipfailover can support up to a maximum of 255 VIPs for the entire cluster, the `ingressIPNetworkCIDR` needs to be `/24` or less.

## CHAPTER 27. IPTABLES

### 27.1. OVERVIEW

There are many system components including OpenShift Container Platform, containers, and software that manage local firewall policies that rely on the kernel iptables configuration for proper network operation. In addition, the iptables configuration of all nodes in the cluster must be correct for networking to work.

All components independently work with iptables without knowledge of how other components are using them. This makes it very easy for one component to break another component's configuration. Further, OpenShift Container Platform and the Docker service assume that iptables remains set up exactly as they have set it up. They may not detect changes introduced by other components and if they do there may be some lag in implementing the fix. In particular, OpenShift Container Platform does monitor and fix problems. However, the Docker service does not.



#### IMPORTANT

Ensure that any changes you make to the iptables configuration on a node do not impact the operation of OpenShift Container Platform and the Docker service. Also, changes will often need to be made on all nodes in the cluster. Use caution, as iptables is not designed to have multiple concurrent users, and is very easy to break OpenShift Container Platform and Docker networking.

OpenShift Container Platform provides several chains, one of which is specifically intended for administrators to use for their own purposes: **OPENSIFT-ADMIN-OUTPUT-RULES**.

See the discussion of [using iptables rules to limit access to external resources](#) for more information.

The chains, order of the chains, and rules in the kernel iptables must be properly set up on each node in the cluster for OpenShift Container Platform and Docker networking to work properly. There are several tools and services that are commonly used in the system that interact with the kernel iptables and can accidentally impact OpenShift Container Platform and the Docker service.

### 27.2. IPTABLES

The iptables tool can be used to set up, maintain, and inspect the tables of IPv4 packet filter rules in the Linux kernel.

Independent of other use, such as a firewall, OpenShift Container Platform and the the Docker service manage chains in some of the tables. The chains are inserted in specific order and the rules are specific to their needs.

#### CAUTION

`iptables --flush [chain]` can remove key required configuration. Do not execute this command.

### 27.3. IPTABLES.SERVICE

The iptables service supports a local network firewall. It assumes total control of the iptables configuration. When it starts, it flushes and restores the complete iptables configuration. The restored

rules are from its configuration file, `/etc/sysconfig/iptables`. The configuration file is not kept up to date during operation, so the dynamically added rules are lost during every restart.



### WARNING

Stopping and starting **iptables.service** will destroy configuration that is required by OpenShift Container Platform and Docker. OpenShift Container Platform and Docker are not notified of the change.

```
# systemctl disable iptables.service
# systemctl mask iptables.service
```

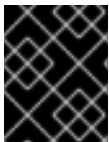
If you need to run **iptables.service**, keep a limited configuration in the configuration file and rely on OpenShift Container Platform and Docker to install their needed rules.

The **iptables.service** configuration is loaded from:

```
/etc/sysconfig/iptables
```

To make permanent rules changes, edit the changes into this file. Do not include Docker or OpenShift Container Platform rules.

After **iptables.service** is started or restarted on a node, the Docker service and **atomic-openshift-node.service** must be restarted to reconstruct the needed iptables configuration.



### IMPORTANT

Restarting the Docker service will cause all containers running on the node to be stopped and restarted.

```
# systemctl restart iptables.service
# systemctl restart docker
# systemctl restart atomic-openshift-node.service
```

## CHAPTER 28. SECURING BUILDS BY STRATEGY

### 28.1. OVERVIEW

**Builds** in OpenShift Container Platform are run in [privileged containers](#) that have access to the Docker daemon socket. As a security measure, it is recommended to limit who can run builds and the strategy that is used for those builds. [Custom builds](#) are inherently less safe than [Source builds](#), given that they can execute any code in the build with potentially full access to the node's Docker socket, and as such are disabled by default. [Docker build](#) permission should also be granted with caution as a vulnerability in the Docker build logic could result in a privileges being granted on the host node.

By default, all users that can create builds are granted permission to use the Docker and Source-to-Image build strategies. Users with [cluster-admin](#) privileges can enable the Custom build strategy, as referenced in the [Restricting Build Strategies to a User Globally](#) section of this page.

You can control who can build with what build strategy using an [authorization policy](#). Each build strategy has a corresponding build subresource. A user must have permission to create a build *and* permission to create on the build strategy subresource in order to create builds using that strategy. Default roles are provided which grant the **create** permission on the build strategy subresource.

**Table 28.1. Build Strategy Subresources and Roles**

| Strategy        | Subresource            | Role                                  |
|-----------------|------------------------|---------------------------------------|
| Docker          | builds/docker          | system:build-strategy-docker          |
| Source-to-Image | builds/source          | system:build-strategy-source          |
| Custom          | builds/custom          | system:build-strategy-custom          |
| JenkinsPipeline | builds/jenkinspipeline | system:build-strategy-jenkinspipeline |

### 28.2. DISABLING A BUILD STRATEGY GLOBALLY

To prevent access to a particular build strategy globally, log in as a user with [cluster-admin](#) privileges and remove the corresponding role from the **system:authenticated** group:

```
$ oc adm policy remove-cluster-role-from-group system:build-strategy-
custom system:authenticated
$ oc adm policy remove-cluster-role-from-group system:build-strategy-
docker system:authenticated
$ oc adm policy remove-cluster-role-from-group system:build-strategy-
source system:authenticated
$ oc adm policy remove-cluster-role-from-group system:build-strategy-
jenkinspipeline system:authenticated
```

In versions prior to 3.2, the build strategy subresources were included in the **admin** and **edit** roles. Ensure the build strategy subresources are also removed from these roles:

```
$ oc edit clusterrole admin
$ oc edit clusterrole edit
```



For each role, remove the line that corresponds to the resource of the strategy to disable.

### Example 28.1. Disable the Docker Build Strategy for admin

```
kind: ClusterRole
metadata:
  name: admin
...
rules:
- resources:
  - builds/custom
  - builds/docker 1
  - builds/source
...
...
```

- 1 Delete this line to disable Docker builds globally for users with the **admin** role.

## 28.3. RESTRICTING BUILD STRATEGIES TO A USER GLOBALLY

To allow only a set of specific users to create builds with a particular strategy:

1. [Disable global access to the build strategy.](#)
2. Assign the role corresponding to the build strategy to a specific user. For example, to add the **system:build-strategy-docker** cluster role to the user **devuser**:

```
$ oc adm policy add-cluster-role-to-user system:build-strategy-
docker devuser
```



### WARNING

Granting a user access at the cluster level to the **builds/docker** subresource means that the user will be able to create builds with the Docker strategy in any project in which they can create builds.

## 28.4. RESTRICTING BUILD STRATEGIES TO A USER WITHIN A PROJECT

Similar to granting the build strategy role to a user globally, to allow only a set of specific users within a project to create builds with a particular strategy:

1. [Disable global access to the build strategy.](#)

2. Assign the role corresponding to the build strategy to a specific user within a project. For example, to add the **system:build-strategy-docker** role within the project **devproject** to the user **devuser**:

```
$ oc adm policy add-role-to-user system:build-strategy-docker  
devuser -n devproject
```

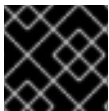
## CHAPTER 29. RESTRICTING APPLICATION CAPABILITIES USING SECCOMP

### 29.1. OVERVIEW

Seccomp (secure computing mode) is used to restrict the set of system calls applications can make, allowing cluster administrators greater control over the security of workloads running in OpenShift Container Platform.

Seccomp support is achieved via two annotations in the pod configuration:

- **seccomp.security.alpha.kubernetes.io/pod**: profile applies to all containers in the pod that do not override
- **container.seccomp.security.alpha.kubernetes.io/<container\_name>**: container-specific profile override



#### IMPORTANT

Containers are run with **unconfined** seccomp settings by default.

For detailed design information, refer to the [seccomp design document](#).

### 29.2. ENABLING SECCOMP

Seccomp is a feature of the Linux kernel. To ensure seccomp is enabled on your system, run:

```
$ cat /boot/config-`uname -r` | grep CONFIG_SECCOMP=
CONFIG_SECCOMP=y
```

### 29.3. CONFIGURING OPENSIFT CONTAINER PLATFORM FOR SECCOMP

A seccomp profile is a json file providing syscalls and the appropriate action to take when a syscall is invoked.

1. Create the seccomp profile.

The [default profile](#) is sufficient in many cases, but the cluster administrator must define the security constraints of an individual system.

To create your own custom profile, create a file on every node in the **seccomp-profile-root** directory.

If you are using the default **docker/default** profile, you do not need to create one.

2. Configure your nodes to use the **seccomp-profile-root** where your profiles will be stored. In the **node-config.yaml** via the **kubeletArguments**:

```
kubeletArguments:
  seccomp-profile-root:
    - "/your/path"
```

- Restart the node service to apply the changes:

```
# systemctl restart atomic-openshift-node
```

- In order to control which profiles may be used, and to set the default profile, [configure your SCC](#) via the **seccompProfiles** field. The first profile will be used as a default. The allowable formats of the **seccompProfiles** field include:

- **docker/default**: the default profile for the container runtime (no profile required)
- **unconfined**: unconfined profile, and disables seccomp
- **localhost/<profile-name>**: the profile installed to the node's local seccomp profile root  
For example, if you are using the default **docker/default** profile, configure your SCC with:

```
seccompProfiles:  
- docker/default
```

## 29.4. CONFIGURING OPENSIFT CONTAINER PLATFORM FOR A CUSTOM SECCOMP PROFILE

To ensure pods in your cluster run with a custom profile:

- Create the seccomp profile in **seccomp-profile-root**.
- Configure **seccomp-profile-root**:

```
kubeletArguments:  
  seccomp-profile-root:  
    - "/your/path"
```

- Restart the node service to apply the changes:

```
# systemctl restart atomic-openshift-node
```

- Configure your SCC:

```
seccompProfiles:  
- localhost/<profile-name>
```

## CHAPTER 30. SYSCTLs

### 30.1. OVERVIEW

Sysctl settings are exposed via Kubernetes, allowing users to modify certain kernel parameters at runtime for namespaces within a container. Only sysctls that are namespaced can be set independently on pods; if a sysctl is not namespaced (called *node-level*), it cannot be set within OpenShift Container Platform. Moreover, only those sysctls considered *safe* are whitelisted by default; other *unsafe* sysctls can be manually enabled on the node to be available to the user.



#### NOTE

As of OpenShift Container Platform 3.3.1, sysctl support is a feature in [Technology Preview](#).

### 30.2. UNDERSTANDING SYSCTLs

In Linux, the sysctl interface allows an administrator to modify kernel parameters at runtime. Parameters are available via the */proc/sys/* virtual process file system. The parameters cover various subsystems such as:

- kernel (common prefix: **kernel.**)
- networking (common prefix: **net.**)
- virtual memory (common prefix: **vm.**)
- MDADM (common prefix: **dev.**)

More subsystems are described in [Kernel documentation](#). To get a list of all parameters, you can run:

```
$ sudo sysctl -a
```

### 30.3. NAMESPACEd VERSUS NODE-LEVEL SYSCTLs

A number of sysctls are *namespaced* in today's Linux kernels. This means that they can be set independently for each pod on a node. Being namespaced is a requirement for sysctls to be accessible in a pod context within Kubernetes.

The following sysctls are known to be namespaced:

- **kernel.shm\***
- **kernel.msg\***
- **kernel.sem**
- **fs.mqueue.\***
- **net.\***

Sysctls that are not namespaced are called *node-level* and must be set manually by the cluster administrator, either by means of the underlying Linux distribution of the nodes (e.g., via */etc/sysctls.conf*) or using a DaemonSet with privileged containers.

**NOTE**

Consider marking nodes with special sysctls as tainted. Only schedule pods onto them that need those sysctl settings. Use the [Kubernetes taints and toleration feature](#) to implement this.

## 30.4. SAFE VERSUS UNSAFE SYSCTLS

Sysctls are grouped into *safe* and *unsafe* sysctls. In addition to proper namespacing, a safe sysctl must be properly isolated between pods on the same node. This means that setting a safe sysctl for one pod:

- must not have any influence on any other pod on the node,
- must not allow to harm the node's health, and
- must not allow to gain CPU or memory resources outside of the resource limits of a pod.

By far, most of the namespaced sysctls are not necessarily considered safe.

Currently, OpenShift Container Platform supports, or whitelists, the following sysctls in the safe set:

- ***kernel.shm\_rmid\_forced***
- ***net.ipv4.ip\_local\_port\_range***
- ***net.ipv4.tcp\_syncookies***

This list will be extended in future versions when the kubelet supports better isolation mechanisms.

All safe sysctls are enabled by default. All unsafe sysctls are disabled by default and must be allowed manually by the cluster administrator on a per-node basis. Pods with disabled unsafe sysctls will be scheduled, but will fail to launch.

**WARNING**

Due to their nature of being unsafe, the use of unsafe sysctls is at-your-own-risk and can lead to severe problems like wrong behavior of containers, resource shortage, or complete breakage of a node.

## 30.5. ENABLING UNSAFE SYSCTLS

With the warning above in mind, the cluster administrator can allow certain unsafe sysctls for very special situations, e.g., high-performance or real-time application tuning.

If you want to use unsafe sysctls, cluster administrators must enable them individually on nodes. Only namespaced sysctls can be enabled this way.

1. Use the **kubeletArguments** field in the **/etc/origin/node/node-config.yaml** file, as described in [Configuring Node Resources](#), to set the desired unsafe sysctls:

```
kubeletArguments:
  experimental-allowed-unsafe-sysctls:
    - "kernel.msg*,net.ipv4.route.min_pmtu"
```

2. Restart the node service to apply the changes:

```
# systemctl restart atomic-openshift-node
```

## 30.6. SETTING SYSCTLS FOR A POD

Sysctls are set on pods using annotations. They apply to all containers in the same pod.

Here is an example, with different annotations for safe and unsafe sysctls:

```
apiVersion: v1
kind: Pod
metadata:
  name: sysctl-example
  annotations:
    security.alpha.kubernetes.io/sysctls: kernel.shm_rmid_forced=1
    security.alpha.kubernetes.io/unsafe-sysctls:
      net.ipv4.route.min_pmtu=1000,kernel.msgmax=1 2 3
spec:
  ...
```



### NOTE

A pod with the unsafe sysctls specified above will fail to launch on any node that has not enabled those two unsafe sysctls explicitly. As with node-level sysctls, use the [taints and toleration feature](#) or [labels on nodes](#) to schedule those pods onto the right nodes.

## CHAPTER 31. ENCRYPTING DATA AT DATASTORE LAYER

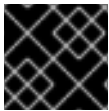
### 31.1. OVERVIEW

This topic reviews how to enable and configure encryption of secret data at the datastore layer. While the examples use the **secrets** resource, any resource can be encrypted, such as **configmaps**.



#### WARNING

This is an alpha feature and may change in future.



#### IMPORTANT

etcd v3 or later is required in order to use this feature.

### 31.2. CONFIGURATION AND DETERMINING WHETHER ENCRYPTION IS ALREADY ENABLED

To activate data encryption, pass the `--experimental-encryption-provider-config` argument to the Kubernetes API server:

#### Excerpt of *master-config.yaml*

```
kubernetesMasterConfig:
  apiServerArguments:
    experimental-encryption-provider-config:
      - /path/to/encryption-config.yaml
```

For more information about *master-config.yaml* and its format, see the [Master Configuration Files](#) topic.

### 31.3. UNDERSTANDING THE ENCRYPTION CONFIGURATION

#### Encryption configuration file with all available providers

```
kind: EncryptionConfig
apiVersion: v1
resources: ①
- resources: ②
  - secrets
providers: ③
- aescbc: ④
  keys:
    - name: key1 ⑤
      secret: c2VjcmV0IGlzlIHNlY3VyZQ== ⑥
    - name: key2
      secret: dGhpcyBpcyBwYXNzd29yZA==
```



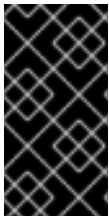
```

- secretbox:
  keys:
    - name: key1
      secret: YWJjZGVmZ2hpamtsbW5vcHFyc3R1dnd4eXoxMjM0NTY=
- aesgcm:
  keys:
    - name: key1
      secret: c2VjcmV0IGlzIHNlY3VyZQ==
    - name: key2
      secret: dGhpcyBpcyBwYXNzd29yZA==
- identity: {}

```

- 1 Each **resources** array item is a separate configuration and contains a complete configuration.
- 2 The **resources.resources** field is an array of Kubernetes resource names (**resource** or **resource.group**) that should be encrypted.
- 3 The **providers** array is an ordered [list of the possible encryption providers](#). Only one provider type can be specified per entry (**identity** or **aescbc** can be provided, but not both in the same item).
- 4 The first provider in the list is used to encrypt resources going into storage.
- 5 Arbitrary name of the secret.
- 6 Base64 encoded random key. Different providers have different key lengths. See instructions on [how to generate the key](#).

When reading resources from storage, each provider that matches the stored data attempts to decrypt the data in order. If no provider can read the stored data due to a mismatch in format or secret key, an error is returned, which prevents clients from accessing that resource.



### IMPORTANT

If any resource is not readable via the encryption configuration (because keys were changed), the only recourse is to delete that key from the underlying etcd directly. Calls attempting to read that resource will fail until it is deleted or a valid decryption key is provided.

#### 31.3.1. Available Providers

| Name            | Encryption                  | Strength  | Speed | Key Length | Other Considerations                                                                                                                  |
|-----------------|-----------------------------|-----------|-------|------------|---------------------------------------------------------------------------------------------------------------------------------------|
| <b>identity</b> | None                        | N/A       | N/A   | N/A        | Resources written as-is without encryption. When set as the first provider, the resource will be decrypted as new values are written. |
| <b>aescbc</b>   | AES-CBC with PKCS#7 padding | Strongest | Fast  | 32-byte    | The recommended choice for encryption, but may be slightly slower than <b>secretbox</b> .                                             |

| Name             | Encryption                                       | Strength                             | Speed   | Key Length         | Other Considerations                                                                                      |
|------------------|--------------------------------------------------|--------------------------------------|---------|--------------------|-----------------------------------------------------------------------------------------------------------|
| <b>secretbox</b> | XSalsa20 and Poly1305                            | Strong                               | Faster  | 32-byte            | A newer standard and may not be considered acceptable in environments that require high levels of review. |
| <b>aesgcm</b>    | AES-GCM with a random initialization vector (IV) | Must be rotated every 200,000 writes | Fastest | 16, 24, or 32-byte | <b>Is not recommended</b> for use except when an automated key rotation scheme is implemented.            |

Each provider supports multiple keys. The keys are tried in order for decryption. If the provider is the first provider, the first key is used for encryption.



## NOTE

Kubernetes has no proper nonce generator and uses a random IV as nonce for AES-GCM. Since AES-GCM requires a proper nonce to be secure, AES-GCM is not recommended. The 200,000 write limit just limits the possibility of a fatal nonce misuse to a reasonable low margin.

## 31.4. ENCRYPTING DATA

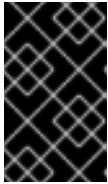
Create a new encryption configuration file.

```
kind: EncryptionConfig
apiVersion: v1
resources:
  - resources:
    - secrets
  providers:
    - aescbc:
        keys:
          - name: key1
            secret: <BASE 64 ENCODED SECRET>
    - identity: {}
```

To create a new secret:

1. Generate a 32-byte random key and base64 encode it. For example, on Linux and macOS use:

```
$ head -c 32 /dev/urandom | base64
```



## IMPORTANT

The encryption key must be generated with an appropriate cryptographically secure random number generator like `/dev/urandom`. For example, `math/random` from Golang or `random.random()` from Python are not suitable.

2. Place that value in the **secret** field.

3. Restart the API server:

```
# systemctl restart atomic-openshift-master-api
```



## IMPORTANT

The encryption provider configuration file contains keys that can decrypt content in etcd, so you must properly restrict permissions on masters so only the user who runs the master API server can read it.

## 31.5. VERIFYING THAT DATA IS ENCRYPTED

Data is encrypted when written to etcd. After restarting the API server, any newly created or updated secrets should be encrypted when stored. To check, you can use the **etcdctl** command line program to retrieve the contents of your secret.

1. Create a new secret called **secret1** in the **default** namespace:

```
$ oc create secret generic secret1 -n default --from-literal=mykey=mydata
```

2. Using the **etcdctl** command line, read that secret out of etcd:

```
$ ETCDCTL_API=3 etcdctl get /kubernetes.io/secrets/default/secret1 -w fields [...] | grep Value
```

[...] must be the additional arguments for connecting to the etcd server.

The final command will look similar to:

```
$ ETCDCTL_API=3 etcdctl get /kubernetes.io/secrets/default/secret1 -w fields \
--cacert=/var/lib/origin/openshift.local.config/master/ca.crt \
--key=/var/lib/origin/openshift.local.config/master/master.etcd-client.key \
--cert=/var/lib/origin/openshift.local.config/master/master.etcd-client.crt \
--endpoints 'https://127.0.0.1:4001' | grep Value
```

3. Verify that the output of the command above is prefixed with **k8s:enc:aescbc:v1**: which indicates the **aescbc** provider has encrypted the resulting data.
4. Verify the secret is correctly decrypted when retrieved via the API:

```
$ oc get secret secret1 -n default -o yaml | grep mykey
```

This should match **mykey: bXIkYXRh**.

## 31.6. ENSURE ALL SECRETS ARE ENCRYPTED

Since secrets are encrypted when written, performing an update on a secret will encrypt that content.

```
$ oc adm migrate storage --include=secrets --confirm
```

This command reads all secrets, then updates them to apply server-side encryption. If an error occurs due to a conflicting write, retry the command.

For larger clusters, you can subdivide the secrets by namespace or script an update.

## 31.7. ROTATING A DECRYPTION KEY

Changing the secret without incurring downtime requires a multi-step operation, especially in the presence of a highly available deployment where multiple API servers are running.

1. Generate a new key and add it as the second key entry for the current provider on all servers.
2. Restart all API servers to ensure each server can decrypt using the new key.



### NOTE

If using a single API server, you can skip this step.

```
# systemctl restart atomic-openshift-master-api
```

3. Make the new key the first entry in the **keys** array so that it is used for encryption in the configuration.
4. Restart all API servers to ensure each server now encrypts using the new key.

```
# systemctl restart atomic-openshift-master-api
```

5. Run the following to encrypt all existing secrets with the new key:

```
$ oc adm migrate storage --include=secrets --confirm
```

6. After you back up etcd with the new key in use and update all secrets, remove the old decryption key from the configuration.

## 31.8. DECRYPTING DATA

To disable encryption at the datastore layer:

1. Place the **identity** provider as the first entry in the configuration:

```
kind: EncryptionConfig
apiVersion: v1
resources:
- resources:
```

```
- secrets
providers:
- identity: {}
- aescbc:
  keys:
  - name: key1
    secret: <BASE 64 ENCODED SECRET>
```

1. Restart all API servers:

```
# systemctl restart atomic-openshift-master-api
```

2. Run the following to force all secrets to be decrypted:

```
$ oc adm migrate storage --include=secrets --confirm
```

## CHAPTER 32. ENCRYPTING HOSTS WITH IPSEC

### 32.1. OVERVIEW

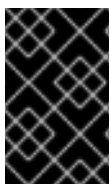
IPsec protects traffic in an OpenShift Container Platform cluster by encrypting the communication between all master and node hosts that communicate using the Internet Protocol (IP).

This topic shows how to secure communication of an entire IP subnet from which the OpenShift Container Platform hosts receive their IP addresses, including all cluster management and pod data traffic.



#### NOTE

Because OpenShift Container Platform management traffic uses HTTPS, enabling IPsec encrypts management traffic a second time.



#### IMPORTANT

This procedure should be repeated on each master host, then node host, in your cluster. Hosts that do not have IPsec enabled will not be able to communicate with a host that does.

### 32.2. ENCRYPTING HOSTS

#### 32.2.1. Step 1: Prerequisites

At this time, **libreswan** version 3.15 is the latest version supported on Red Hat Enterprise Linux 7. Ensure that **libreswan** 3.15 or later is installed on cluster hosts. If [opportunistic group functionality](#) is required, then **libreswan** version 3.19 or later is required.

[Configure the SDN MTU](#) to allow space for the IPsec header. In the configuration described here IPsec requires 62 bytes. If the cluster is operating on an ethernet network with an MTU of 1500 then the SDN MTU should be 1388, to allow for the overhead of IPsec and the SDN encapsulation.

After modifying the MTU in the OpenShift Container Platform configuration, the SDN must be made aware of the change by removing the SDN interface and restarting the OpenShift Container Platform node process.

```
# systemctl stop atomic-openshift-node
# ovs-vsctl del-br br0
# systemctl start atomic-openshift-node
```

#### 32.2.2. Step 2: Certificates

By default, OpenShift Container Platform secures cluster management communication with mutually authenticated HTTPS communication. This means that both the client (for example, an OpenShift Container Platform node) and the server (for example, an OpenShift Container Platform api-server) send each other their certificates, which are checked against a known certificate authority (CA). These certificates are generated at cluster set up time and typically live on each host.

These certificates can also be used to secure pod communications with IPsec. You need three files on each host:

- Cluster CA file
  - Host client certificate file
  - Host private key file
1. Determine what the certificate's nickname will be after it has been imported into the **libreswan** certificate database. The nickname is taken directly from the certificate's subject's Common Name (CN):

```
# openssl x509 \
-in /path/to/client-certificate -subject -noout | \
sed -n 's/.*CN=\.*/\1/p'
```

2. Use **openssl** to combine the client certificate, CA certificate, and private key files into a **PKCS#12** file, which is a common file format for multiple certificates and keys:

```
# openssl pkcs12 -export \
-in /path/to/client-certificate \
-inkey /path/to/private-key \
-certfile /path/to/certificate-authority \
-passout pass: \
-out certs.p12
```

3. Import the **PKCS#12** file into the **libreswan** certificate database. The **-w** option is left empty because no password is assigned to the **PKCS#12** file, as it is only temporary.

```
# ipsec initnss
# pk12util -i certs.p12 -d sql:/etc/ipsec.d -w ""
# rm certs.p12
```

### 32.2.3. Step 3: libreswan IPsec Policy

Now that the necessary certificates are imported into the **libreswan** certificate database, create a policy that uses them to secure communication between hosts in your cluster.

If you are using **libreswan** 3.19 or later, then [opportunistic group configuration](#) is recommended. Otherwise, explicit connections are required.

#### 32.2.3.1. Opportunistic Group Configuration

The following configuration creates two **libreswan** connections. The first encrypts traffic using the OpenShift Container Platform certificates, while the second creates exceptions to the encryption for cluster-external traffic.

1. Place the following into the **/etc/ipsec.d/openshift-cluster.conf** file:

```
conn private
left=%defaultroute
leftid=%fromcert
# our certificate
leftcert="NSS Certificate DB:<cert_nickname>"
right=%opportunisticgroup
rightid=%fromcert
```

1

```
# their certificate transmitted via IKE
rightca=%same
ikev2=insist
authby=rsasig
failureshunt=drop
negotiationshunt=hold
auto=ondemand

conn clear
left=%defaultroute
right=%group
authby=never
type=passthrough
auto=route
priority=100
```

1. Replace `<cert_nickname>` with the certificate nickname from step one.
2. Tell **libreswan** which IP subnets and hosts to apply each policy using policy files in `/etc/ipsec.d/policies/`, where each configured connection has a corresponding policy file. So, in the example above, the two connections, **private** and **clear**, each have a file in `/etc/ipsec.d/policies/`. `/etc/ipsec.d/policies/private` should contain the IP subnet of your cluster, which your hosts receive IP addresses from. By default, this causes all communication between hosts in the cluster subnet to be encrypted if the remote host's client certificate authenticates against the local host's Certificate Authority certificate. If the remote host's certificate does not authenticate, all traffic between the two hosts will be blocked.

For example, if all hosts are configured to use addresses in the **172.16.0.0/16** address space, your **private** policy file would contain **172.16.0.0/16**. Any number of additional subnets to encrypt may be added to this file, which results in all traffic to those subnets using IPsec as well.

3. Unencrypt the communication between all hosts and the subnet gateway to ensure that traffic can enter and exit the cluster. Add the gateway to the `/etc/ipsec.d/policies/clear` file:

```
172.16.0.1/32
```

Additional hosts and subnets may be added to this file, which will result in all traffic to these hosts and subnets being unencrypted.

### 32.2.3.2. Explicit Connection Configuration

In this configuration, each IPSec node configuration must explicitly list the configuration of every other node in the cluster. Using a configuration management tool such as Ansible to generate this file on each host is recommended.

1. This configuration also requires the full certificate subject of each node to be placed into the configuration for every other node. To read this subject from the node's certificate, use **openssl**:

```
# openssl x509 \
-in /path/to/client-certificate -text | \
grep "Subject:" | \
sed 's/[[:blank:]]*Subject: //'
```



- Place the following lines into the `/etc/ipsec.d/openshift-cluster.conf` file on each node for every other node in the cluster:

```
conn <other_node_hostname>
    left=<this_node_ip> 1
    leftid="CN=<this_node_cert_nickname>" 2
    lefttrsasigkey=%cert
    leftcert=<this_node_cert_nickname> 3
    right=<other_node_ip> 4
    rightid="<other_node_cert_full_subject>" 5
    righttrsasigkey=%cert
    auto=start
    keyingtries=%forever
```

- 1 Replace `<this_node_ip>` with the cluster IP address of this node.
- 2 3 Replace `<this_node_cert_nickname>` with the node certificate nickname from step one.
- 4 Replace `<other_node_ip>` with the cluster IP address of the other node.
- 5 Replace `<other_node_cert_full_subject>` with the other node's certificate subject from just above. For example: `"O=system:nodes,CN=openshift-node-45.example.com"`.

- Place the following in the `/etc/ipsec.d/openshift-cluster.secrets` file on each node:

```
: RSA "<this_node_cert_nickname>" 1
```

- 1 Replace `<this_node_cert_nickname>` with the node certificate nickname from step one.

## 32.3. IPSEC FIREWALL CONFIGURATION

All nodes within the cluster need to allow IPSec related network traffic. This includes IP protocol numbers 50 and 51 as well as UDP port 500.

For example, if the cluster nodes communicate over interface `eth0`:

```
-A OS_FIREWALL_ALLOW -i eth0 -p 50 -j ACCEPT
-A OS_FIREWALL_ALLOW -i eth0 -p 51 -j ACCEPT
-A OS_FIREWALL_ALLOW -i eth0 -p udp --dport 500 -j ACCEPT
```



### NOTE

IPSec also uses UDP port 4500 for NAT traversal, though this should not apply to normal cluster deployments.

## 32.4. STARTING AND ENABLING IPSEC

- Start the `ipsec` service to load the new configuration and policies, and begin encrypting:

```
# systemctl start ipsec
```

2. Enable the **ipsec** service to start on boot:

```
# systemctl enable ipsec
```

## 32.5. OPTIMIZING IPSEC

See the [Scaling and Performance Guide](#) for performance suggestions when encrypting with IPsec.

## 32.6. TROUBLESHOOTING

When authentication cannot be completed between two hosts, you will not be able to ping between them, because all IP traffic will be rejected. If the **clear** policy is not configured correctly, you will also not be able to SSH to the host from another host in the cluster.

You can use the **ipsec status** command to check that the **clear** and **private** policies have been loaded.

## CHAPTER 33. BUILDING DEPENDENCY TREES

### 33.1. OVERVIEW

OpenShift Container Platform uses [image change triggers](#) in a **BuildConfig** to detect when an [image stream tag](#) has been updated. You can use the **oc adm build-chain** command to build a dependency tree that identifies which [images](#) would be affected by updating an image in a specified [image stream](#).

The **build-chain** tool can determine which [builds](#) to trigger; it analyzes the output of those builds to determine if they will in turn update another [image stream tag](#). If they do, the tool continues to follow the dependency tree. Lastly, it outputs a graph specifying the image stream tags that would be impacted by an update to the top-level tag. The default output syntax for this tool is set to a human-readable format; the DOT format is also supported.

### 33.2. USAGE

The following table describes common **build-chain** usage and general syntax:

**Table 33.1. Common build-chain Operations**

| Description                                                                                                     | Syntax                                                                                       |
|-----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| Build the dependency tree for the <b>latest</b> tag in <b>&lt;image-stream&gt;</b> .                            | <pre>\$ oc adm build-chain &lt;image-stream&gt;</pre>                                        |
| Build the dependency tree for the <b>v2</b> tag in DOT format, and visualize it using the DOT utility.          | <pre>\$ oc adm build-chain &lt;image-stream&gt;:v2 \ -o dot \   dot -T svg -o deps.svg</pre> |
| Build the dependency tree across all projects for the specified image stream tag found the <b>test</b> project. | <pre>\$ oc adm build-chain &lt;image-stream&gt;:v1 \ -n test --all</pre>                     |



#### NOTE

You may need to install the **graphviz** package to use the **dot** command.

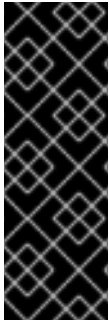
## CHAPTER 34. BACKUP AND RESTORE

### 34.1. OVERVIEW

In OpenShift Container Platform, you can *back up* (saving state to separate storage) and *restore* (recreating state from separate storage) at the cluster level. There is also some preliminary support for [per-project backup](#). The full state of a cluster installation includes:

- etcd data (v2 and v3) on each master
- API objects
- registry storage
- volume storage

This topic does not cover how to back up and restore [persistent storage](#), as those topics are left to the underlying storage provider. However, an example of how to perform a **generic** backup of [application data](#) is provided.

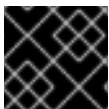


#### IMPORTANT

This topic only provides a generic way of backing up applications and the OpenShift Container Platform cluster. It can not take into account custom requirements. Therefore, you should create a full backup and restore procedure. To prevent data loss, necessary precautions should be taken.

Backup and restore procedures are not fully supported in OpenShift Container Platform 3.6 due to dependencies on cluster state.

Note that the etcd backup still has all the references to the storage volumes. When you restore etcd, OpenShift Container Platform starts launching the previous pods on nodes and reattaching the same storage. This is really no different than the process of when you remove a node from the cluster and add a new one back in its place. Anything attached to that node will be reattached to the pods on whatever nodes they get rescheduled to.



#### IMPORTANT

Backup and restore is not guaranteed. You are responsible for backing up your own data.

### 34.2. PREREQUISITES

1. Because the restore procedure involves a complete reinstallation, save all the files used in the initial installation. This may include:
  - `~/config/openshift/installer.cfg.yml` (from the [Quick Installation](#) method)
  - Ansible playbooks and inventory files (from the [Advanced Installation](#) method)
  - `/etc/yum.repos.d/ose.repo` (from the [Disconnected Installation](#) method)
2. Backup the procedures for post-installation steps. Some installations may involve steps that are not included in the installer. This may include changes to the services outside of the control of OpenShift Container Platform or the installation of extra services like monitoring agents.

Additional configuration that is not supported yet by the advanced installer might also be affected, for example when using multiple authentication providers.

3. Install packages that provide various utility commands:

```
# yum install etcd
```

4. If using a container-based installation, pull the etcd image instead:

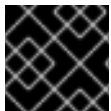
```
# docker pull rhel7/etcd
```

Note the location of the **etcd** data directory (or **\$ETCD\_DATA\_DIR** in the following sections), which depends on how **etcd** is deployed.

| Deployment Type | Description                                                                              | Data Directory                              |
|-----------------|------------------------------------------------------------------------------------------|---------------------------------------------|
| separate etcd   | etcd runs as a separate service, either co-located on master nodes or on separate nodes. | <i>/var/lib/etcd</i>                        |
| embedded etcd   | etcd runs as part of the master service.                                                 | <i>/var/lib/origin/openshift.local.etcd</i> |

## 34.3. CLUSTER BACKUP

### 34.3.1. Master Backup



#### IMPORTANT

You must perform the following step on each master node.

1. Create a backup of the master host configuration files:

```
$ MYBACKUPDIR=/backup/${hostname}/${date +%Y%m%d}
$ sudo mkdir -p ${MYBACKUPDIR}/etc/sysconfig
$ sudo cp -aR /etc/origin ${MYBACKUPDIR}/etc
$ sudo cp -aR /etc/sysconfig/atomic-* ${MYBACKUPDIR}/etc/sysconfig/
```

### 34.3.2. Etcd Backup

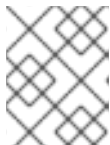
1. If **etcd** is running on more than one host, stop it on each host:

```
# sudo systemctl stop etcd
```

Although this step is not strictly necessary, doing so ensures that the **etcd** data is fully synchronized.

2. Capture **etcd** v2 data by creating a backup:

```
# etcdctl backup \
  --data-dir $ETCD_DATA_DIR \
  --backup-dir $ETCD_DATA_DIR.bak
```

**NOTE**

If **etcd** is running on more than one host, the various instances regularly synchronize their data, so creating a backup for one of them is sufficient.

**NOTE**

For a container-based installation, you must use **docker exec** to run **etcdctl** inside the container.

3. Capture **etcd** v3 data by copying the **db** file over to the backup you created:

```
# cp "$ETCD_DATA_DIR"/member/snap/db
  "$ETCD_DATA_DIR.bak"/member/snap/db
```

4. Back up the **/etc/etcd/ca/** directory on the master host that contains certificates in that folder:

```
$ sudo cp -aR /etc/etcd/ca/ ${MYBACKUPDIR}/etc
```

### 34.3.3. Registry Certificates Backup

1. Save all the registry certificates, on every master and node host.

```
# cd /etc/docker/certs.d/
# tar cf /tmp/docker-registry-certs-$(hostname).tar *
```

**NOTE**

When working with one or more [external secured registry](#), any host required to pull or push images must trust registry certificates in order to run pods.

## 34.4. CLUSTER RESTORE FOR SINGLE-MEMBER ETCD CLUSTERS

To restore the cluster:

1. Reinstall OpenShift Container Platform.  
This should be done in the [same way](#) that OpenShift Container Platform was previously installed.
2. Run all necessary post-installation steps.
3. Restore the certificates and keys, on each master:

```
# cd /etc/origin/master
# tar xvf /tmp/certs-and-keys-$(hostname).tar
```

4. Restore from the **etcd** backup:

```
# mv $ETCD_DATA_DIR $ETCD_DATA_DIR.orig
# cp -Rp $ETCD_DATA_DIR.bak $ETCD_DATA_DIR
# chcon -R --reference $ETCD_DATA_DIR.orig $ETCD_DATA_DIR
# chown -R etcd:etcd $ETCD_DATA_DIR
```

5. Create the new single node cluster using etcd's **--force-new-cluster** option. You can do this using the values from **/etc/etcd/etcd.conf**, or you can temporarily modify the **systemd** unit file and start the service normally.

To do so, edit the **/usr/lib/systemd/system/etcd.service** file, and add **--force-new-cluster**:

```
# sed -i '/ExecStart/s/"$/ --force-new-cluster"/'
/usr/lib/systemd/system/etcd.service
# systemctl show etcd.service --property ExecStart --no-pager

ExecStart=/bin/bash -c "GOMAXPROCS=$(nproc) /usr/bin/etcd --force-
new-cluster"
```

Then, restart the **etcd** service:

```
# systemctl daemon-reload
# systemctl start etcd
```

6. Verify the **etcd** service started correctly, then re-edit the **/usr/lib/systemd/system/etcd.service** file and remove the **--force-new-cluster** option:

```
# sed -i '/ExecStart/s/ --force-new-cluster//'
/usr/lib/systemd/system/etcd.service
# systemctl show etcd.service --property ExecStart --no-pager

ExecStart=/bin/bash -c "GOMAXPROCS=$(nproc) /usr/bin/etcd"
```

7. Restart the **etcd** service, then verify the etcd cluster is running correctly and displays OpenShift Container Platform's configuration:

```
# systemctl daemon-reload
# systemctl restart etcd
```

## 34.5. CLUSTER RESTORE FOR MULTIPLE-MEMBER ETCD CLUSTERS

When using a separate etcd cluster, you must first restore the etcd backup by creating a new, single node etcd cluster. If you run etcd as a stand-alone service on your master nodes, you can create the single node etcd cluster on a master node. If you use separate etcd with multiple members, you must then also add any additional etcd members to the etcd cluster one by one.

However, the details of the restoration process differ between [embedded](#) and [external](#) etcd. See the following section and follow the relevant steps before [Bringing OpenShift Services Back Online](#).

### 34.5.1. Embedded etcd

Restore your etcd backup and configuration:

1. Run the following on the master with the embedded etcd:

```
# ETCD_DIR=/var/lib/origin/openshift.local.etcd
# mv $ETCD_DIR /var/lib/etcd.orig
# cp -Rp /var/lib/origin/etcd-backup-<timestamp>/ $ETCD_DIR
# chcon -R --reference /var/lib/etcd.orig/ $ETCD_DIR
# chown -R etcd:etcd $ETCD_DIR
```



### WARNING

The **\$ETCD\_DIR** location differs between external and embedded etcd.

2. Create the new, single node etcd cluster:

```
# etcd -data-dir=/var/lib/origin/openshift.local.etcd \
    -force-new-cluster
```

Verify etcd has started successfully by checking the output from the above command, which should look similar to the following near the end:

```
[...]
2016-06-24 12:14:45.644073 I | etcdserver: starting server...
[version: 2.2.5, cluster version: 2.2]
[...]
2016-06-24 12:14:46.834394 I | etcdserver: published {Name:default
ClientURLs:[http://localhost:2379 http://localhost:4001]} to cluster
5580663a6e0002
```

3. Shut down the process by running the following from a separate terminal:

```
# pkill etcd
```

4. Continue to [Bringing OpenShift Container Platform Services Back Online](#).

## 34.5.2. Separate etcd

Choose a system to be the initial etcd member, and restore its etcd backup and configuration:

1. Run the following on the etcd host:

```
# ETCD_DIR=/var/lib/etcd/
# mv $ETCD_DIR /var/lib/etcd.orig
# cp -Rp /var/lib/origin/etcd-backup-<timestamp>/ $ETCD_DIR
# chcon -R --reference /var/lib/etcd.orig/ $ETCD_DIR
# chown -R etcd:etcd $ETCD_DIR
```



**WARNING**

The `$ETCD_DIR` location differs between external and embedded etcd.

2. Restore your `/etc/etcd/etcd.conf` file from backup or `.rpmsave`.
3. Depending on your environment, follow the instructions for [Containerized etcd Deployments](#) or [Non-Containerized etcd Deployments](#).

### 34.5.2.1. Containerized etcd Deployments

1. Create the new single node cluster using etcd's `--force-new-cluster` option. You can do this with a long, complex command using the values from `/etc/etcd/etcd.conf`, or you can temporarily modify the `systemd` unit file and start the service normally. To do so, edit the `/etc/systemd/system/etcd_container.service` file, and add `--force-new-cluster`:

```
# sed -i '/ExecStart=/s/$/ --force-new-cluster/'
/etc/systemd/system/etcd_container.service

ExecStart=/usr/bin/docker run --name etcd --rm -v \
/var/lib/etcd:/var/lib/etcd:z -v /etc/etcd:/etc/etcd:ro --env-
file=/etc/etcd/etcd.conf \
--net=host --entrypoint=/usr/bin/etcd rhel7/etcd:3.1.9 --force-new-
cluster
```

Then, restart the `etcd` service:

```
# systemctl daemon-reload
# systemctl start etcd_container
```

2. Verify the `etcd` service started correctly, then re-edit the `/etc/systemd/system/etcd_container.service` file and remove the `--force-new-cluster` option:

```
# sed -i '/ExecStart=/s/ --force-new-cluster//'
/etc/systemd/system/etcd_container.service

ExecStart=/usr/bin/docker run --name etcd --rm -v \
/var/lib/etcd:/var/lib/etcd:z -v \
/etc/etcd:/etc/etcd:ro --env-file=/etc/etcd/etcd.conf --net=host \
--entrypoint=/usr/bin/etcd rhel7/etcd:3.1.9
```

3. Restart the `etcd` service, then verify the etcd cluster is running correctly and displays OpenShift Container Platform's configuration:

```
# systemctl daemon-reload
# systemctl restart etcd_container
# etcdctl --cert-file=/etc/etcd/peer.crt \
```

```
--key-file=/etc/etcd/peer.key \
--ca-file=/etc/etcd/ca.crt \
--peers="https://172.16.4.18:2379,https://172.16.4.27:2379" \ 1
ls /
```

- 1 Ensure that you specify the URLs of only active etcd members in the **--peers** parameter value.

4. If you have additional etcd members to add to your cluster, continue to [Adding Additional etcd Members](#). Otherwise, if you only want a single node standalone etcd, continue to [Bringing OpenShift Container Platform Services Back Online](#).

### 34.5.2.2. Non-Containerized etcd Deployments

1. Create the new single node cluster using etcd's **--force-new-cluster** option. You can do this with a long, complex command using the values from */etc/etcd/etcd.conf*, or you can temporarily modify the **systemd** unit file and start the service normally. To do so, edit the */usr/lib/systemd/system/etcd.service* file, and add **--force-new-cluster**:

```
# sed -i '/ExecStart/s/"$/ --force-new-cluster"/'
/usr/lib/systemd/system/etcd.service
# systemctl show etcd.service --property ExecStart --no-pager

ExecStart=/bin/bash -c "GOMAXPROCS=$(nproc) /usr/bin/etcd --force-
new-cluster"
```

Then restart the **etcd** service:

```
# systemctl daemon-reload
# systemctl start etcd
```

2. Verify the **etcd** service started correctly, then re-edit the */usr/lib/systemd/system/etcd.service* file and remove the **--force-new-cluster** option:

```
# sed -i '/ExecStart/s/ --force-new-cluster//'
/usr/lib/systemd/system/etcd.service
# systemctl show etcd.service --property ExecStart --no-pager

ExecStart=/bin/bash -c "GOMAXPROCS=$(nproc) /usr/bin/etcd"
```

3. Restart the **etcd** service, then verify the etcd cluster is running correctly and displays OpenShift Container Platform's configuration:

```
# systemctl daemon-reload
# systemctl restart etcd
# etcdctl --cert-file=/etc/etcd/peer.crt \
--key-file=/etc/etcd/peer.key \
--ca-file=/etc/etcd/ca.crt \
--peers="https://172.16.4.18:2379,https://172.16.4.27:2379" \ 1
ls /
```

- 1 Ensure that you specify the URLs of only active etcd members in the **--peers** parameter value.
4. If you have additional etcd members to add to your cluster, continue to [Adding Additional etcd Members](#). Otherwise, if you only want a single node separate etcd cluster, continue to [Bringing OpenShift Container Platform Services Back Online](#).

### 34.5.2.3. Adding Additional etcd Members

To add additional etcd members to the cluster, you must first adjust the default **localhost** peer in the **peerURLs** value for the first member:

1. Get the member ID for the first member using the **member list** command:

```
# etcdctl --cert-file=/etc/etcd/peer.crt \
  --key-file=/etc/etcd/peer.key \
  --ca-file=/etc/etcd/ca.crt \
  --
peers="https://172.18.1.18:2379,https://172.18.9.202:2379,https://17
2.18.0.75:2379" \ 1
  member list
```

- 1 Ensure that you specify the URLs of only active etcd members in the **--peers** parameter value.
2. Obtain the IP address where etcd listens for cluster peers:

```
$ ss -ltn | grep 2380
```

3. Update the value of **peerURLs** using the **etcdctl member update** command by passing the member ID and IP address obtained from the previous steps:

```
# etcdctl --cert-file=/etc/etcd/peer.crt \
  --key-file=/etc/etcd/peer.key \
  --ca-file=/etc/etcd/ca.crt \
  --
peers="https://172.18.1.18:2379,https://172.18.9.202:2379,https://17
2.18.0.75:2379" \
  member update 511b7fb6cc0001 https://172.18.1.18:2380
```

Alternatively, you can use **curl**:

```
# curl --cacert /etc/etcd/ca.crt \
  --cert /etc/etcd/peer.crt \
  --key /etc/etcd/peer.key \
  https://172.18.1.18:2379/v2/members/511b7fb6cc0001 \
  -XPUT -H "Content-Type: application/json" \
  -d '{"peerURLs":["https://172.18.1.18:2380"]}'
```

4. Re-run the **member list** command and ensure the peer URLs no longer include **localhost**.
5. Now, add each additional member to the cluster one at a time.

**WARNING**

Each member must be fully added and brought online one at a time. When adding each additional member to the cluster, the **peerURLs** list must be correct for that point in time, so it will grow by one for each member added. The **etcdctl member add** command will output the values that need to be set in the **etcd.conf** file as you add each member, as described in the following instructions.

- a. For each member, add it to the cluster using the values that can be found in that system's **etcd.conf** file:

```
# etcdctl --cert-file=/etc/etcd/peer.crt \
  --key-file=/etc/etcd/peer.key \
  --ca-file=/etc/etcd/ca.crt \
  --peers="https://172.16.4.18:2379,https://172.16.4.27:2379" \
  member add 10.3.9.222 https://172.16.4.27:2380 1

Added member named 10.3.9.222 with ID 4e1db163a21d7651 to cluster

ETCD_NAME="10.3.9.222"
ETCD_INITIAL_CLUSTER="10.3.9.221=https://172.16.4.18:2380,10.3.9.222=https://172.16.4.27:2380"
ETCD_INITIAL_CLUSTER_STATE="existing"
```

- 1 In this line, **10.3.9.222** is a label for the etcd member. You can specify the host name, IP address, or a simple name.

- b. Using the environment variables provided in the output of the above **etcdctl member add** command, edit the **/etc/etcd/etcd.conf** file on the member system itself and ensure these settings match. If you previously used the member system as an etcd node, you must still overwrite the current values in the **/etc/etcd/etcd.conf** file.

- c. Now start etcd on the new member:

```
# rm -rf /var/lib/etcd/member
# systemctl enable etcd
# systemctl start etcd
```

- d. Ensure the service starts correctly and the etcd cluster is now healthy:

```
# etcdctl --cert-file=/etc/etcd/peer.crt \
  --key-file=/etc/etcd/peer.key \
  --ca-file=/etc/etcd/ca.crt \
  --peers="https://172.16.4.18:2379,https://172.16.4.27:2379" \
  member list

51251b34b80001: name=10.3.9.221 peerURLs=https://172.16.4.18:2380
clientURLs=https://172.16.4.18:2379
```

```
d266df286a41a8a4: name=10.3.9.222
peerURLs=https://172.16.4.27:2380
clientURLs=https://172.16.4.27:2379

# etcdctl --cert-file=/etc/etcd/peer.crt \
  --key-file=/etc/etcd/peer.key \
  --ca-file=/etc/etcd/ca.crt \
  --peers="https://172.16.4.18:2379,https://172.16.4.27:2379" \
  cluster-health

cluster is healthy
member 51251b34b80001 is healthy
member d266df286a41a8a4 is healthy
```

- e. Now repeat this process for the next member to add to the cluster.
6. After all additional etcd members have been added, continue to [Bringing OpenShift Container Platform Services Back Online](#).

## 34.6. ADDING NEW ETCD HOSTS

In cases where etcd members have failed and you still have a quorum of etcd cluster members running, you can use the surviving members to add additional etcd members without downtime.

### Suggested Cluster Size

Having a cluster with an odd number of etcd hosts can account for fault tolerance. Having an odd number of etcd hosts does not change the number needed for a quorum, but increases the tolerance for failure. For example, a cluster size of three members, quorum is two leaving a failure tolerance of one. This ensures the cluster will continue to operate if two of the members are healthy.

Having an in-production cluster of three etcd hosts is recommended.



### NOTE

The following presumes you have a backup of the **/etc/etcd** configuration for the etcd hosts.

1. If the new etcd members will also be OpenShift Container Platform nodes, see [Add the desired number of hosts to the cluster](#). The rest of this procedure presumes you have added just one host, but if adding multiple, perform all steps on each host.
2. Upgrade etcd and iptables on the surviving nodes:

```
# yum update etcd iptables-services
```

Ensure version **etcd-2.3.7-4.el7.x86\_64** or greater is installed, and that the same version is installed on each host.

3. Install etcd and iptables on the new host

```
# yum install etcd iptables-services
```

Ensure version **etcd-2.3.7-4.el7.x86\_64** or greater is installed, and that the same version is installed on the new host.

4. [Backup the etcd data store](#) on surviving hosts before making any cluster configuration changes.
5. If replacing a failed etcd member, remove the failed member *before* adding the new member.

```
# etcdctl -C https://<surviving host IP>:2379 \
--ca-file=/etc/etcd/ca.crt \
--cert-file=/etc/etcd/peer.crt \
--key-file=/etc/etcd/peer.key cluster-health

# etcdctl -C https://<surviving host IP>:2379 \
--ca-file=/etc/etcd/ca.crt \
--cert-file=/etc/etcd/peer.crt \
--key-file=/etc/etcd/peer.key member remove <failed member
identifier>
```

Stop the etcd service on the failed etcd member:

```
# systemctl stop etcd
```

6. On the new host, add the appropriate iptables rules:

```
# systemctl enable iptables.service --now
# iptables -N OS_FIREWALL_ALLOW
# iptables -t filter -I INPUT -j OS_FIREWALL_ALLOW
# iptables -A OS_FIREWALL_ALLOW -p tcp -m state \
--state NEW -m tcp --dport 2379 -j ACCEPT
# iptables -A OS_FIREWALL_ALLOW -p tcp -m state \
--state NEW -m tcp --dport 2380 -j ACCEPT
# iptables-save > /etc/sysconfig/iptables
```

7. Generate the required certificates for the new host. On a surviving etcd host:

- a. Make a backup of the **/etc/etcd/ca/** directory. Ensure that the directory contains the etcd certificates.
- b. Set the variables and working directory for the certificates, ensuring to create the **PREFIX** directory if one has not been created:

```
# cd /etc/etcd
# export NEW_ETCD=<NEW_HOST_NAME>

# export CN=$NEW_ETCD
# export SAN="IP:<NEW_HOST_IP>"
# export PREFIX="./generated_certs/etcd-$CN/"
```

- c. Create the \$PREFIX directory:

```
$ mkdir -p $PREFIX
```

- d. Create the **server.csr** and **server.crt** certificates:

```
# openssl req -new -keyout ${PREFIX}server.key \
-config ca/openssl.cnf \
-out ${PREFIX}server.csr \
-reqexts etcd_v3_req -batch -nodes \
-subj /CN=$CN

# openssl ca -name etcd_ca -config ca/openssl.cnf \
-out ${PREFIX}server.crt \
-in ${PREFIX}server.csr \
-extensions etcd_v3_ca_server -batch
```

- e. Create the **peer.csr** and **peer.crt** certificates:

```
# openssl req -new -keyout ${PREFIX}peer.key \
-config ca/openssl.cnf \
-out ${PREFIX}peer.csr \
-reqexts etcd_v3_req -batch -nodes \
-subj /CN=$CN

# openssl ca -name etcd_ca -config ca/openssl.cnf \
-out ${PREFIX}peer.crt \
-in ${PREFIX}peer.csr \
-extensions etcd_v3_ca_peer -batch
```

- f. Copy the **etcd.conf** and **ca.crt** files, and archive the contents of the directory:

```
# cp etcd.conf ${PREFIX}
# cp ca.crt ${PREFIX}
# tar -czvf ${PREFIX}${CN}.tgz -C ${PREFIX} .
```

- g. Transfer the files to the new etcd hosts:

```
# scp ${PREFIX}${CN}.tgz $CN:/etc/etcd/
```

8. While still on the surviving etcd host, add the new host to the cluster:

- a. Add the new host to the cluster:

```
# export ETCD_CA_HOST="<SURVIVING_ETCD_HOSTNAME>"
# export NEW_ETCD="<NEW_ETCD_HOSTNAME>"
# export NEW_ETCD_IP="<NEW_HOST_IP>"

# etcdctl -C https://${ETCD_CA_HOST}:2379 \
--ca-file=/etc/etcd/ca.crt \
--cert-file=/etc/etcd/peer.crt \
--key-file=/etc/etcd/peer.key member add ${NEW_ETCD}
https://${NEW_ETCD_IP}:2380

ETCD_NAME="<NEW_ETCD_HOSTNAME>"
ETCD_INITIAL_CLUSTER="
<NEW_ETCD_HOSTNAME>=https://<NEW_HOST_IP>:2380,
<SURVIVING_ETCD_HOST>=https://<SURVIVING_HOST_IP>:2380
ETCD_INITIAL_CLUSTER_STATE="existing"
```

Copy the three environment variables in the `etcdctl` member add output. They will be used later.

- b. On the new host, extract the copied configuration data and set the permissions:

```
# tar -xf /etc/etcd/<NEW_ETCD_HOSTNAME>.tgz -C /etc/etcd/ --  
overwrite  
# chown -R etcd:etcd /etc/etcd/*
```

- c. On the new host, remove any etcd data:

```
# rm -rf /var/lib/etcd/member  
# chown -R etcd:etcd /var/lib/etcd
```

9. On the new etcd host, update the ***etcd.conf*** file:

- a. Replace the following with the values generated in the previous step:

- `ETCD_NAME`
- `ETCD_INITIAL_CLUSTER`
- `ETCD_INITIAL_CLUSTER_STATE`

- b. Replace the IP address with the "NEW\_ETCD" value for:

- `ETCD_LISTEN_PEER_URLS`
- `ETCD_LISTEN_CLIENT_URLS`
- `ETCD_INITIAL_ADVERTISE_PEER_URLS`
- `ETCD_ADVERTISE_CLIENT_URLS`

- c. For replacing failed members, replace the failed hosts with the new hosts.

10. To ensure the etcd configuration does not use the failed host when the etcd service is restarted, modify the ***etcd.conf*** file on all remaining etcd hosts and remove the failed host in the value for the ***ETCD\_INITIAL\_CLUSTER*** variable.

11. On the node that hosts the installation files, update the **`[etcd]`** hosts group in the ***/etc/ansible/hosts*** inventory file. Remove the old etcd hosts and add the new ones.

12. Start etcd on the new host:

```
# systemctl enable etcd --now
```

13. To verify that the new member has been added successfully:

```
etcdctl -C https://${ETCD_CA_HOST}:2379 --ca-file=/etc/etcd/ca.crt \  
--cert-file=/etc/etcd/peer.crt \  
--key-file=/etc/etcd/peer.key cluster-health
```

14. Update the master configuration on all masters to point to the new etcd host

- a. On every master in the cluster, edit ***/etc/origin/master/master-config.yaml***



- b. Find the **etcdClientInfo** section.
  - c. Add the new etcd host to the **urls** list.
  - d. If a failed etcd host was replaced, remove it from the list.
  - e. Restart the master API service.
- On each master:

```
# systemctl restart atomic-openshift-master-api atomic-openshift-
master-controllers
```

The procedure to add an etcd member is complete.

## 34.7. BRINGING OPENSIFT CONTAINER PLATFORM SERVICES BACK ONLINE

On each OpenShift Container Platform master, restore your master and node configuration from backup and enable and restart all relevant services.

On the master in a single master cluster:

```
# cp ${MYBACKUPDIR}/etc/sysconfig/atomic-openshift-master
/etc/sysconfig/atomic-openshift-master
# cp ${MYBACKUPDIR}/etc/origin/master/master-config.yaml.<timestamp>
/etc/origin/master/master-config.yaml
# cp ${MYBACKUPDIR}/etc/origin/node/node-config.yaml.<timestamp>
/etc/origin/node/node-config.yaml
# systemctl enable atomic-openshift-master
# systemctl enable atomic-openshift-node
# systemctl start atomic-openshift-master
# systemctl start atomic-openshift-node
```

On each master in a multi-master cluster:

```
# cp ${MYBACKUPDIR}/etc/sysconfig/atomic-openshift-master-api
/etc/sysconfig/atomic-openshift-master-api
# cp ${MYBACKUPDIR}/etc/sysconfig/atomic-openshift-master-controllers
/etc/sysconfig/atomic-openshift-master-controllers
# cp ${MYBACKUPDIR}/etc/origin/master/master-config.yaml.<timestamp>
/etc/origin/master/master-config.yaml
# cp ${MYBACKUPDIR}/etc/origin/node/node-config.yaml.<timestamp>
/etc/origin/node/node-config.yaml
# systemctl enable atomic-openshift-master-api
# systemctl enable atomic-openshift-master-controllers
# systemctl enable atomic-openshift-node
# systemctl start atomic-openshift-master-api
# systemctl start atomic-openshift-master-controllers
# systemctl start atomic-openshift-node
```

On each OpenShift Container Platform node, restore your **node-config.yaml** file from backup and enable and restart the **atomic-openshift-node** service:

```
# cp /etc/origin/node/node-config.yaml.<timestamp> /etc/origin/node/node-
```

```
config.yaml
# systemctl enable atomic-openshift-node
# systemctl start atomic-openshift-node
```

Your OpenShift Container Platform cluster should now be back online.

## 34.8. PROJECT BACKUP

A future release of OpenShift Container Platform will feature specific support for per-project back up and restore.

For now, to back up API objects at the project level, use **oc export** for each object to be saved. For example, to save the deployment configuration **frontend** in YAML format:

```
$ oc export dc frontend -o yaml > dc-frontend.yaml
```

To back up all of the project (with the exception of cluster objects like namespaces and projects):

```
$ oc export all -o yaml > project.yaml
```

### 34.8.1. Role Bindings

Sometimes custom policy [role bindings](#) are used in a project. For example, a project administrator can give another user a certain role in the project and grant that user project access.

These role bindings can be exported:

```
$ oc get rolebindings -o yaml --export=true > rolebindings.yaml
```

### 34.8.2. Service Accounts

If custom service accounts are created in a project, these need to be exported:

```
$ oc get serviceaccount -o yaml --export=true > serviceaccount.yaml
```

### 34.8.3. Secrets

Custom secrets like source control management secrets (SSH Public Keys, Username/Password) should be exported if they are used:

```
$ oc get secret -o yaml --export=true > secret.yaml
```

### 34.8.4. Persistent Volume Claims

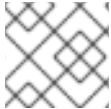
If the application within a project uses a persistent volume through a persistent volume claim (PVC), these should be backed up:

```
$ oc get pvc -o yaml --export=true > pvc.yaml
```

## 34.9. PROJECT RESTORE

To restore a project, recreate the project and recreate all of the objects that were exported during the backup:

```
$ oc new-project myproject
$ oc create -f project.yaml
$ oc create -f secret.yaml
$ oc create -f serviceaccount.yaml
$ oc create -f pvc.yaml
$ oc create -f rolebindings.yaml
```



## NOTE

Some resources can fail to be created (for example, pods and default service accounts).

## 34.10. APPLICATION DATA BACKUP

In many cases, application data can be backed up using the **oc rsync** command, assuming **rsync** is installed within the container image. The Red Hat **rhel7** base image does contain **rsync**. Therefore, all images that are based on **rhel7** contain it as well. See [Troubleshooting and Debugging CLI Operations - rsync](#).



## WARNING

This is a *generic* backup of application data and does not take into account application-specific backup procedures, for example, special export/import procedures for database systems.

Other means of backup may exist depending on the type of the persistent volume (for example, Cinder, NFS, Gluster, or others).

The paths to back up are also *application specific*. You can determine what path to back up by looking at the **mountPath** for volumes in the **deploymentconfig**.

### Example of Backing up a Jenkins Deployment's Application Data

1. Get the application data **mountPath** from the **deploymentconfig**:

```
$ oc get dc/jenkins -o jsonpath='{ .spec.template.spec.containers[?
(@.name=="jenkins")].volumeMounts[?(@.name=="jenkins-
data")].mountPath }'
/var/lib/jenkins
```

2. Get the name of the pod that is currently running:

```
$ oc get pod --selector=deploymentconfig=jenkins -o jsonpath='{
.metadata.name }'
jenkins-1-37nux
```

3. Use the **oc rsync** command to copy application data:

```
$ oc rsync jenkins-1-37nux:/var/lib/jenkins /tmp/
```



#### NOTE

This type of application data backup can only be performed while an application pod is currently running.

## 34.11. APPLICATION DATA RESTORE

The process for restoring application data is similar to the [application backup procedure](#) using the **oc rsync** tool. The same restrictions apply and the process of restoring application data requires a persistent volume.

### Example of Restoring a Jenkins Deployment's Application Data

1. Verify the backup:

```
$ ls -la /tmp/jenkins-backup/
total 8
drwxrwxr-x.  3 user      user   20 Sep  6 11:14 .
drwxrwxrwt. 17 root      root  4096 Sep  6 11:16 ..
drwxrwsrwx. 12 user      user  4096 Sep  6 11:14 jenkins
```

2. Use the **oc rsync** tool to copy the data into the running pod:

```
$ oc rsync /tmp/jenkins-backup/jenkins jenkins-1-37nux:/var/lib
```



#### NOTE

Depending on the application, you may be required to restart the application.

3. Restart the application with new data (*optional*):

```
$ oc delete pod jenkins-1-37nux
```

Alternatively, you can scale down the deployment to 0, and then up again:

```
$ oc scale --replicas=0 dc/jenkins
$ oc scale --replicas=1 dc/jenkins
```

## CHAPTER 35. TROUBLESHOOTING OPENSIFT SDN

### 35.1. OVERVIEW

As described in the [SDN documentation](#) there are multiple layers of interfaces that are created to correctly pass the traffic from one container to another. In order to debug connectivity issues, you have to test the different layers of the stack to work out where the problem arises. This guide will help you dig down through the layers to identify the problem and how to fix it.

Part of the problem is that OpenShift Container Platform can be set up many ways, and the networking can be wrong in a few different places. So this document will work through some scenarios that, hopefully, will cover the majority of cases. If your problem is not covered, the tools and concepts that are introduced should help guide debugging efforts.

### 35.2. NOMENCLATURE

#### Cluster

The set of machines in the cluster. *i.e.* the Masters and the Nodes.

#### Master

A controller of the OpenShift Container Platform cluster. Note that the master may not be a node in the cluster, and thus, may not have IP connectivity to the pods.

#### Node

Host in the cluster running OpenShift Container Platform that can host pods.

#### Pod

Group of containers running on a node, managed by OpenShift Container Platform.

#### Service

Abstraction that presents a unified network interface that is backed by one or more pods.

#### Router

A web proxy that can map various URLs and paths into OpenShift Container Platform services to allow external traffic to travel into the cluster.

#### Node Address

The IP address of a node. This is assigned and managed by the owner of the network to which the node is attached. Must be reachable from any node in the cluster (master and client).

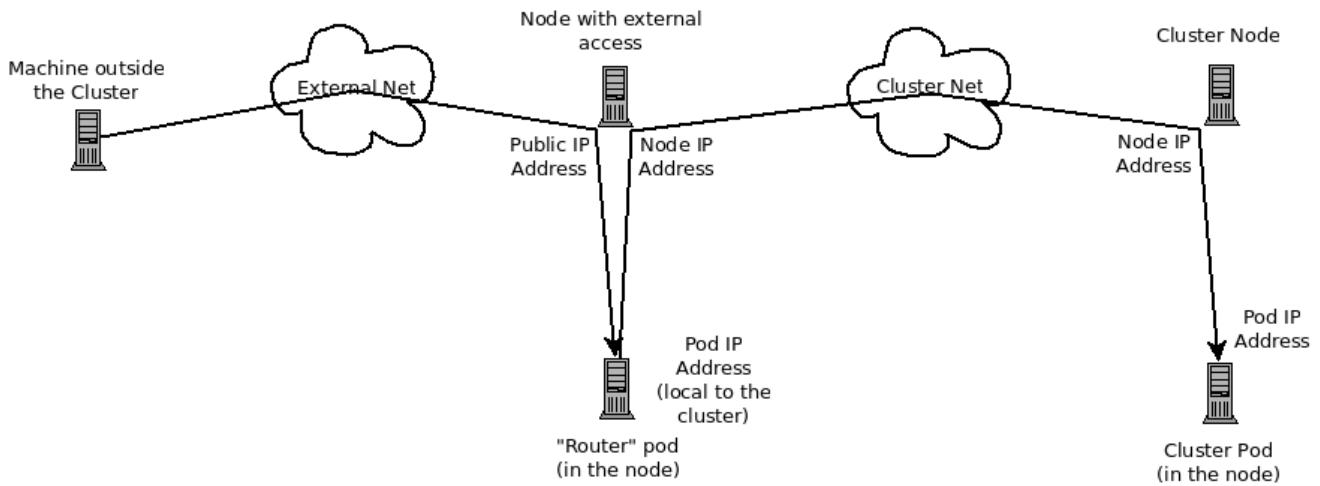
#### Pod Address

The IP address of a pod. These are assigned and managed by OpenShift Container Platform. By default they are assigned out of the 10.128.0.0/14 network (or, in older versions, 10.1.0.0/16). Only reachable from the client nodes.

#### Service Address

An IP address that represents the service, and is mapped to a pod address internally. These are assigned and managed by OpenShift Container Platform. By default they are assigned out of the 172.30.0.0/16 network. Only reachable from the client nodes.

The following diagram shows all of the pieces involved with external access.



### 35.3. DEBUGGING EXTERNAL ACCESS TO AN HTTP SERVICE

If you are on a machine outside the cluster and are trying to access a resource provided by the cluster there needs to be a process running in a pod that listens on a public IP address and "routes" that traffic inside the cluster. The [OpenShift Container Platform router](#) serves that purpose for HTTP, HTTPS (with SNI), WebSockets, or TLS (with SNI).

Assuming you can't access an HTTP service from the outside of the cluster, let's start by reproducing the problem on the command line of the machine where things are failing. Try:

```
curl -kv http://foo.example.com:8000/bar    # But replace the argument
with your URL
```

If that works, are you reproducing the bug from the right place? It is also possible that the service has some pods that work, and some that don't. So jump ahead to the [Section 35.4, "Debugging the Router"](#) section.

If that failed, then let's resolve the DNS name to an IP address (assuming it isn't already one):

```
dig +short foo.example.com                # But replace the hostname
with yours
```

If that doesn't give back an IP address, it's time to troubleshoot DNS, but that's outside the scope of this guide.



#### IMPORTANT

Make sure that the IP address that you got back is one that you expect to be running the router. If it's not, fix your DNS.

Next, use **ping -c address** and **tracpath address** to check that you can reach the router host. It is possible that they will not respond to ICMP packets, in which case those tests will fail, but the router machine may be reachable. In which case, try using the telnet command to access the port for the router directly:

```
telnet 1.2.3.4 8000
```

You may get:

■

```
Trying 1.2.3.4...
Connected to 1.2.3.4.
Escape character is '^]'.
```

If so, there's something listening on the port on the IP address. That's good. Hit **ctrl-]** then hit the *enter* key and then type **close** to quit telnet. Move on to the [Section 35.4, "Debugging the Router"](#) section to check other things on the router.

Or you could get:

```
Trying 1.2.3.4...
telnet: connect to address 1.2.3.4: Connection refused
```

Which tells us that the router is not listening on that port. See the [Section 35.4, "Debugging the Router"](#) section for more pointers on how to configure the router.

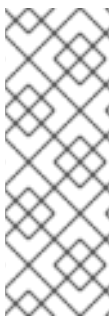
Or if you see:

```
Trying 1.2.3.4...
telnet: connect to address 1.2.3.4: Connection timed out
```

Which tells us that you can't talk to anything on that IP address. Check your routing, firewalls, and that you have a router listening on that IP address. To debug the router, see the [Section 35.4, "Debugging the Router"](#) section. For IP routing and firewall issues, debugging that is beyond the purview of this guide.

## 35.4. DEBUGGING THE ROUTER

Now that you have an IP address, we need to **ssh** to that machine and check that the router software is running on that machine and configured correctly. So let's **ssh** there and get administrative OpenShift Container Platform credentials.



### NOTE

If you have access to administrator credentials but are no longer logged in as the [default system user](#) **system:admin**, you can log back in as this user at any time as long as the credentials are still present in your [CLI configuration file](#). The following command logs in and switches to the **default** project:

```
$ oc login -u system:admin -n default
```

Check that the router is running:

```
# oc get endpoints --namespace=default --selector=router
NAMESPACE   NAME          ENDPOINTS
default     router        10.128.0.4:80
```

If that command fails, then your OpenShift Container Platform configuration is broken. Fixing that is outside the scope of this document.

You should see one or more router endpoints listed, but that won't tell you if they are running on the machine with the given external IP address, since the endpoint IP address will be one of the pod addresses that is internal to the cluster. To get the list of router host IP addresses, run:

```
# oc get pods --all-namespaces --selector=router --template='{{range
.items}}HostIP: {{.status.hostIP}} PodIP: {{.status.podIP}}{{end}}
{{"\n"}}'
```

HostIP: 192.168.122.202 PodIP: 10.128.0.4

You should see the host IP that corresponds to your external address. If you do not, refer to the [router documentation](#) to configure the router pod to run on the right node (by setting the affinity correctly) or update your DNS to match the IP addresses where the routers are running.

At this point in the guide, you should be on a node, running your router pod, but you still cannot get the HTTP request to work. First we need to make sure that the router is mapping the external URL to the correct service, and if that works, we need to dig into that service to make sure that all endpoints are reachable.

Let's list all of the routes that OpenShift Container Platform knows about:

```
# oc get route --all-namespaces
```

| NAME            | HOST/PORT       | PATH  | SERVICE      | LABELS |
|-----------------|-----------------|-------|--------------|--------|
| TLS TERMINATION |                 |       |              |        |
| route-unsecured | www.example.com | /test | service-name |        |

If the host name and path from your URL don't match anything in the list of returned routes, then you need to add a route. See the [router documentation](#).

If your route is present, then you need to debug access to the endpoints. That's the same as if you were debugging problems with a service, so continue on with the next [Section 35.5, "Debugging a Service"](#) section.

## 35.5. DEBUGGING A SERVICE

If you can't communicate with a service from inside the cluster (either because your services can't communicate directly, or because you are using the router and everything works until you get into the cluster) then you need to work out what endpoints are associated with a service and debug them.

First, let's get the services:

```
# oc get services --all-namespaces
```

| NAMESPACE | NAME            | LABELS                                  | SELECTOR                | IP(S)          | PORT(S)  |
|-----------|-----------------|-----------------------------------------|-------------------------|----------------|----------|
| default   | docker-registry | docker-registry=default                 | docker-registry=default | 172.30.243.225 | 5000/TCP |
| default   | kubernetes      | component=apiserver,provider=kubernetes | <none>                  | 172.30.0.1     | 443/TCP  |
| default   | router          | router=router                           | router=router           | 172.30.213.8   | 80/TCP   |

You should see your service in the list. If not, then you need to define your [service](#).

The IP addresses listed in the service output are the Kubernetes service IP addresses that Kubernetes will map to one of the pods that backs that service. So you should be able to talk to that IP address. But, unfortunately, even if you can, it doesn't mean all pods are reachable; and if you can't, it doesn't mean all pods aren't reachable. It just tells you the status of the *one* that kubeproxy hooked you up to.

Let's test the service anyway. From one of your nodes:

■



```
curl -kv http://172.30.243.225:5000/bar # Replace the
argument with your service IP address and port
```

Then, let's work out what pods are backing our service (replace **docker-registry** with the name of the broken service):

```
# oc get endpoints --selector=docker-registry
NAME                ENDPOINTS
docker-registry     10.128.2.2:5000
```

From this, we can see that there's only one endpoint. So, if your service test succeeded, and the router test succeeded, then something really odd is going on. But if there's more than one endpoint, or the service test failed, try the following *for each* endpoint. Once you identify what endpoints aren't working, then proceed to the next section.

First, test each endpoint (change the URL to have the right endpoint IP, port, and path):

```
curl -kv http://10.128.2.2:5000/bar
```

If that works, great, try the next one. If it failed, make a note of it and we'll work out why, in the next section.

If all of them failed, then it is possible that the local node is not working, jump to the [Section 35.7, "Debugging Local Networking"](#) section.

If all of them worked, then jump to the [Section 35.11, "Debugging Kubernetes"](#) section to work out why the service IP address isn't working.

## 35.6. DEBUGGING NODE TO NODE NETWORKING

Using our list of non-working endpoints, we need to test connectivity to the node.

1. Make sure that all nodes have the expected IP addresses:

```
# oc get hostsubnet
NAME                HOST                HOST IP
SUBNET
rh71-os1.example.com rh71-os1.example.com 192.168.122.46
10.1.1.0/24
rh71-os2.example.com rh71-os2.example.com 192.168.122.18
10.1.2.0/24
rh71-os3.example.com rh71-os3.example.com 192.168.122.202
10.1.0.0/24
```

If you are using DHCP they could have changed. Ensure the host names, IP addresses, and subnets match what you expect. If any node details have changed, use **oc edit hostsubnet** to correct the entries.

2. After ensuring the node addresses and host names are correct, list the endpoint IPs and node IPs:

```
# oc get pods --selector=docker-registry \
  --template='{{range .items}}HostIP: {{.status.hostIP}} PodIP:
  {{.status.podIP}}{{end}}\n''
```

HostIP: 192.168.122.202    PodIP: 10.128.0.4

- Find the endpoint IP address you made note of before and look for it in the **PodIP** entry, and find the corresponding **HostIP** address. Then test connectivity at the node host level using the address from **HostIP**:

- **ping -c 3 <IP\_address>**: No response could mean that an intermediate router is eating the ICMP traffic.
- **tracpath <IP\_address>**: Shows the IP route taken to the target, if ICMP packets are returned by all hops.  
If both **tracpath** and **ping** fail, then look for connectivity issues with your local or virtual network.

- For local networking, check the following:

- Check the route the packet takes out of the box to the target address:

```
# ip route get 192.168.122.202
192.168.122.202 dev ens3 src 192.168.122.46
cache
```

In the above example, it will go out the interface named **ens3** with the source address of **192.168.122.46** and go directly to the target. If that is what you expected, use **ip a show dev ens3** to get the interface details and make sure that is the expected interface.

An alternate result may be the following:

```
# ip route get 192.168.122.202
1.2.3.4 via 192.168.122.1 dev ens3 src 192.168.122.46
```

It will pass through the **via** IP value to route appropriately. Ensure that the traffic is routing correctly. Debugging route traffic is beyond the scope of this guide.

Other debugging options for node to node networking can be solved with the following:

- Do you have ethernet link on both ends? Look for **Link detected: yes** in the output from **ethtool <network\_interface>**.
- Are your duplex settings, and ethernet speeds right on both ends? Look through the rest of the **ethtool <network\_interface>** information.
- Are the cables plugged in correctly? To the correct ports?
- Are the switches configured correctly?

Once you have ascertained that the node to node connectivity is fine, we need to look at the SDN configuration on both ends.

## 35.7. DEBUGGING LOCAL NETWORKING

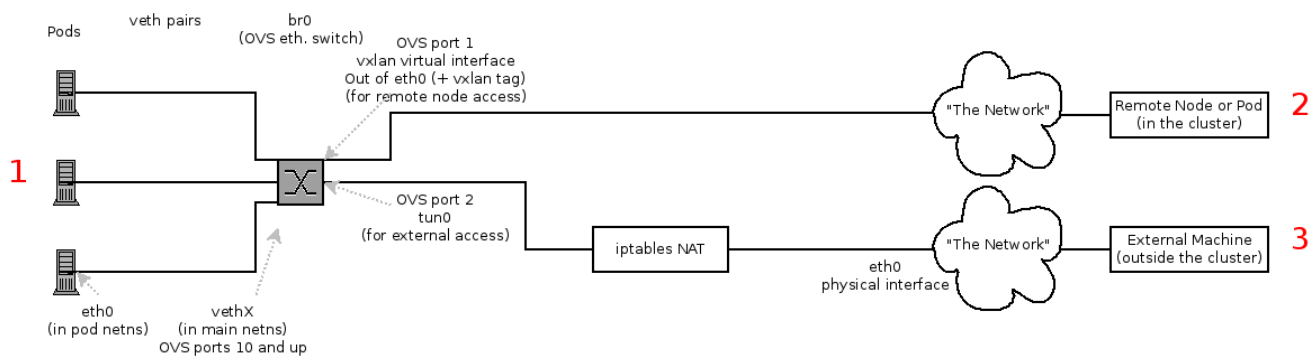
At this point we should have a list of one or more endpoints that you can't communicate with, but that have node to node connectivity. For each one, we need to work out what is wrong, but first you need to understand how the SDN sets up the networking on a node for the different pods.

### 35.7.1. The Interfaces on a Node

These are the interfaces that the OpenShift SDN creates:

- **br0**: The OVS bridge device that 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.
- **vethX** (in the main netns): A Linux virtual ethernet peer of **eth0** in the Docker netns. It will be attached to the OVS bridge on one of the other ports.

### 35.7.2. SDN Flows Inside a Node



Depending on what you are trying to access (or be accessed from) the path will vary. There are four different places the SDN connects (inside a node). They are labeled in red on the diagram above.

- **Pod**: Traffic is going from one pod to another on the same machine (1 to a different 1)
- **Remote Node (or Pod)**: Traffic is going from a local pod to a remote node or pod in the same cluster (1 to 2)
- **External Machine**: Traffic is going from a local pod outside the cluster (1 to 3)

Of course the opposite traffic flows are also possible.

### 35.7.3. Debugging Steps

#### 35.7.3.1. Is IP Forwarding Enabled?

Check that `sysctl net.ipv4.ip_forward` is set to 1 (and check the host if this is a VM)

#### 35.7.3.2. Are your routes correct?

Check the route tables with `ip route`:

```
# ip route
default via 192.168.122.1 dev ens3
10.128.0.0/14 dev tun0 proto kernel scope link      #
This sends all pod traffic into OVS
10.128.2.0/23 dev tun0 proto kernel scope link src 10.128.2.1 #
This is traffic going to local pods, overriding the above
169.254.0.0/16 dev ens3 scope link metric 1002      #
This is for Zeroconf (may not be present)
172.17.0.0/16 dev docker0 proto kernel scope link src 172.17.42.1 #
Docker's private IPs... used only by things directly configured by docker;
not OpenShift
192.168.122.0/24 dev ens3 proto kernel scope link src 192.168.122.46 #
The physical interface on the local subnet
```

You should see the 10.128.x.x lines (assuming you have your pod network set to the default range in your configuration). If you do not, check the OpenShift Container Platform logs (see the [Section 35.10](#), “Reading the Logs” section)

### 35.7.4. Is the Open vSwitch configured correctly?

Check the Open vSwitch bridges on both sides:

```
# ovs-vsctl list-br
br0
```

This should be **br0**.

You can list all of the ports that ovs knows about:

```
# ovs-ofctl -O OpenFlow13 dump-ports-desc br0
OFPST_PORT_DESC reply (OF1.3) (xid=0x2):
  1(vxlan0): addr:9e:f1:7d:4d:19:4f
    config:      0
    state:       0
    speed: 0 Mbps now, 0 Mbps max
  2(tun0): addr:6a:ef:90:24:a3:11
    config:      0
    state:       0
    speed: 0 Mbps now, 0 Mbps max
  8(vethe19c6ea): addr:1e:79:f3:a0:e8:8c
    config:      0
    state:       0
    current:     10GB-FD COPPER
    speed: 10000 Mbps now, 0 Mbps max
  LOCAL(br0): addr:0a:7f:b4:33:c2:43
    config:      PORT_DOWN
    state:       LINK_DOWN
    speed: 0 Mbps now, 0 Mbps max
```

In particular, the **vethX** devices for all of the active pods should be listed as ports.

Next, list the flows that are configured on that bridge:

```
# ovs-ofctl -O OpenFlow13 dump-flows br0
```

The results will vary slightly depending on whether you are using the **ovs-subnet** or **ovs-multitenant** plug-in, but there are certain general things you can look for:

1. Every remote node should have a flow matching **tun\_src=<node\_IP\_address>** (for incoming VXLAN traffic from that node) and another flow including the action **set\_field: <node\_IP\_address>->tun\_dst** (for outgoing VXLAN traffic to that node).
2. Every local pod should have flows matching **arp\_spa=<pod\_IP\_address>** and **arp\_tpa=<pod\_IP\_address>** (for incoming and outgoing ARP traffic for that pod), and flows matching **nw\_src=<pod\_IP\_address>** and **nw\_dst=<pod\_IP\_address>** (for incoming and outgoing IP traffic for that pod).

If there are flows missing, look in the [Section 35.10, “Reading the Logs”](#) section.

#### 35.7.4.1. Is the iptables configuration correct?

Check the output from **iptables-save** to make sure you are not filtering traffic. However, OpenShift Container Platform sets up iptables rules during normal operation, so do not be surprised to see entries there.

#### 35.7.4.2. Is your external network correct?

Check external firewalls, if any, allow traffic to the target address (this is site-dependent, and beyond the purview of this guide).

## 35.8. DEBUGGING VIRTUAL NETWORKING

### 35.8.1. Builds on a Virtual Network are Failing

If you are installing OpenShift Container Platform using a virtual network (for example, OpenStack), and a build is failing, the maximum transmission unit (MTU) of the target node host might not be compatible with the MTU of the primary network interface (for example, **eth0**).

For a build to complete successfully, the MTU of an SDN must be less than the eth0 network MTU in order to pass data to between node hosts.

1. Check the MTU of your network by running the **ip addr** command:

```
# ip addr
---
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP qlen 1000
    link/ether fa:16:3e:56:4c:11 brd ff:ff:ff:ff:ff:ff
    inet 172.16.0.0/24 brd 172.16.0.0 scope global dynamic eth0
        valid_lft 168sec preferred_lft 168sec
    inet6 fe80::f816:3eff:fe56:4c11/64 scope link
        valid_lft forever preferred_lft forever
---
```

The MTU of the above network is 1500.

2. The MTU in your node configuration must be lower than the network value. Check the **mtu** in the node configuration of the targeted node host:

■

```
# cat /etc/origin/node/node-config.yaml
...
networkConfig:
  mtu: 1450
  networkPluginName: company/openshift-ovs-subnet
...
```

In the above node configuration file, the **mtu** value is lower than the network MTU, so no configuration is needed. If the **mtu** value was higher, edit the file and lower the value to at least 50 units fewer than the MTU of the primary network interface, then restart the node service. This would allow larger packets of data to pass between nodes.

## 35.9. DEBUGGING POD EGRESS

If you are trying to access an external service from a pod, e.g.:

```
curl -kv github.com
```

Make sure that the DNS is resolving correctly:

```
dig +search +noall +answer github.com
```

That should return the IP address for the github server, but check that you got back the correct address. If you get back no address, or the address of one of your machines, then you may be matching the wildcard entry in your local DNS server.

To fix that, you either need to make sure that DNS server that has the wildcard entry is not listed as a **nameserver** in your **/etc/resolv.conf** or you need to make sure that the wildcard domain is not listed in the **search** list.

If the correct IP address was returned, then try the debugging advice listed above in [Section 35.7, “Debugging Local Networking”](#). Your traffic should leave the Open vSwitch on port 2 to pass through the **iptables** rules, then out the route table normally.

## 35.10. READING THE LOGS

Run: **journalctl -u atomic-openshift-node.service --boot | less**

Look for the **Output of setup script:** line. Everything starting with '+' below that are the script steps. Look through that for obvious errors.

Following the script you should see lines with **Output of adding table=0**. Those are the OVS rules, and there should be no errors.

## 35.11. DEBUGGING KUBERNETES

Check **iptables -t nat -L** to make sure that the service is being NAT'd to the right port on the local machine for the **kubeproxy**.

**WARNING**

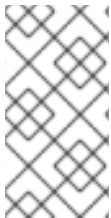
This is all changing soon... Kubeproxy is being eliminated and replaced with an **iptables**-only solution.

## 35.12. FINDING NETWORK ISSUES USING THE DIAGNOSTICS TOOL

As a cluster administrator, run the diagnostics tool to diagnose common network issues:

```
# oc adm diagnostics NetworkCheck
```

The diagnostics tool runs a series of checks for error conditions for the specified component. See the [Diagnostics Tool section](#) for more information.

**NOTE**

Currently, the diagnostics tool cannot diagnose IP failover issues. As a workaround, you can run the script at <https://raw.githubusercontent.com/openshift/openshift-sdn/master/hack/ipf-debug.sh> on the master (or from another machine with access to the master) to generate useful debugging information. However, this script is unsupported.

By default, **oc adm diagnostics NetworkCheck** logs errors into */tmp/openshift/*. This can be configured with the **--network-logdir** option:

```
# oc adm diagnostics NetworkCheck --network-logdir=<path/to/directory>
```

## 35.13. MISCELLANEOUS NOTES

### 35.13.1. Other clarifications on ingress

- Kube - declare a service as NodePort and it will claim that port on all machines in the cluster (on what interface?) and then route into kube-proxy and then to a backing pod. See <https://kubernetes.io/docs/concepts/services-networking/service/#type-nodeport> (some node must be accessible from outside)
- Kube - declare as a LoadBalancer and something *you* have to write does the rest
- OS/AE - Both use the router

### 35.13.2. TLS Handshake Timeout

When a pod fails to deploy, check its docker log for a TLS handshake timeout:

```
$ docker log <container_id>
...
[...] couldn't get deployment [...] TLS handshake timeout
...
```

This condition, and generally, errors in establishing a secure connection, may be caused by a large difference in the MTU values between tun0 and the primary interface (e.g., eth0), such as when tun0 MTU is 1500 and eth0 MTU is 9000 (jumbo frames).

### 35.13.3. Other debugging notes

- Peer interfaces (of a Linux virtual ethernet pair) can be determined with **ethtool -S *ifname***
- Driver type: **ethtool -i *ifname***



## CHAPTER 36. DIAGNOSTICS TOOL

### 36.1. OVERVIEW

The **oc adm diagnostics** command runs a series of checks for error conditions in the host or cluster. Specifically, it:

- Verifies that the default registry and router are running and correctly configured.
- Checks **ClusterRoleBindings** and **ClusterRoles** for consistency with base policy.
- Checks that all of the client configuration contexts are valid and can be connected to.
- Checks that SkyDNS is working properly and the pods have SDN connectivity.
- Validates master and node configuration on the host.
- Checks that nodes are running and available.
- Analyzes host logs for known errors.
- Checks that systemd units are configured as expected for the host.

### 36.2. USING THE DIAGNOSTICS TOOL

You can deploy OpenShift Container Platform in several ways. These include:

- Built from source
- Included within a VM image
- As a container image
- Using enterprise RPMs

Each method is suited for a different configuration and environment. To minimize environment assumptions, the diagnostics tool is included with the **openshift** binary to provide diagnostics within an OpenShift Container Platform server or client.

To use the diagnostics tool, preferably on a master host and as cluster administrator, run:

```
# oc adm diagnostics
```

This runs all available diagnostics and skips any that do not apply to the environment.

You can run a specific diagnostics by name or run specific diagnostics by name as you work to address issues. For example:

```
$ oc adm diagnostics
```

The options for the diagnostics tool require working configuration files. For example, the **NodeConfigCheck** does not run unless a node configuration is available.

The diagnostics tool uses the standard configuration file locations by default:

- Client:
  - As indicated by the **\$KUBECONFIG** environment variable
  - *~/.kube/config file*
- Master:
  - */etc/origin/master/master-config.yaml*
- Node:
  - */etc/origin/node/node-config.yaml*

You can specify non-standard locations with the **--config**, **--master-config**, and **--node-config** options. If a configuration file is not specified, related diagnostics are skipped.

Available diagnostics include:

| Diagnostic Name            | Purpose                                                                                                                                                                                                                                       |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>AggregatedLogging</b>   | Check the aggregated logging integration for proper configuration and operation.                                                                                                                                                              |
| <b>AnalyzeLogs</b>         | Check systemd service logs for problems. Does not require a configuration file to check against.                                                                                                                                              |
| <b>ClusterRegistry</b>     | Check that the cluster has a working Docker registry for builds and image streams.                                                                                                                                                            |
| <b>ClusterRoleBindings</b> | Check that the default cluster role bindings are present and contain the expected subjects according to base policy.                                                                                                                          |
| <b>ClusterRoles</b>        | Check that cluster roles are present and contain the expected permissions according to base policy.                                                                                                                                           |
| <b>ClusterRouter</b>       | Check for a working default router in the cluster.                                                                                                                                                                                            |
| <b>ConfigContexts</b>      | Check that each context in the client configuration is complete and has connectivity to its API server.                                                                                                                                       |
| <b>DiagnosticPod</b>       | Creates a pod that runs diagnostics from an application standpoint, which checks that DNS within the pod is working as expected and the credentials for the default service account authenticate correctly to the master API.                 |
| <b>EtcWriteVolume</b>      | Check the volume of writes against etcd for a time period and classify them by operation and key. This diagnostic only runs if specifically requested, because it does not run as quickly as other diagnostics and can increase load on etcd. |

| Diagnostic Name                   | Purpose                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>MasterConfigCheck</b>          | Check this host's master configuration file for problems.                                                                                                                                                                                                                                                                                                                                                                                                    |
| <b>MasterNode</b>                 | Check that the master running on this host is also running a node to verify that it is a member of the cluster SDN.                                                                                                                                                                                                                                                                                                                                          |
| <b>MetricsApiProxy</b>            | Check that the integrated Heapster metrics can be reached via the cluster API proxy.                                                                                                                                                                                                                                                                                                                                                                         |
| <b>NetworkCheck</b>               | <p>Create diagnostic pods on multiple nodes to diagnose common network issues from an application standpoint. For example, this checks that pods can connect to services, other pods, and the external network.</p> <p>If there are any errors, this diagnostic stores results and retrieved files in a local directory (<i>/tmp/openshift/</i>, by default) for further analysis. The directory can be specified with the <b>--network-logdir</b> flag.</p> |
| <b>NodeConfigCheck</b>            | Checks this host's node configuration file for problems.                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>NodeDefinitions</b>            | Check that the nodes defined in the master API are ready and can schedule pods.                                                                                                                                                                                                                                                                                                                                                                              |
| <b>RouteCertificateValidation</b> | Check all route certificates for those that might be rejected by extended validation.                                                                                                                                                                                                                                                                                                                                                                        |
| <b>ServiceExternalIPs</b>         | Check for existing services that specify external IPs, which are disallowed according to master configuration.                                                                                                                                                                                                                                                                                                                                               |
| <b>UnitStatus</b>                 | Check systemd status for units on this host related to OpenShift Container Platform. Does not require a configuration file to check against.                                                                                                                                                                                                                                                                                                                 |

### 36.3. RUNNING DIAGNOSTICS IN A SERVER ENVIRONMENT

An Ansible-deployed cluster provides additional diagnostic benefits for nodes within an OpenShift Container Platform cluster. These include:

- Master and node configuration is based on a configuration file in a standard location.
- Systemd units are configured to manage the server(s).
- Both master and node configuration files are in standard locations.
- Systemd units are created and configured for managing the nodes in a cluster.
- All components log to journald.

Keeping to the default location of the configuration files placed by an Ansible-deployed cluster ensures that running **oc adm diagnostics** works without any flags. If you are not using the default location for the configuration files, you must use the **--master-config** and **--node-config** options:

```
# oc adm diagnostics --master-config=<file_path> --node-config=<file_path>
```

Systemd units and logs entries in journald are necessary for the current log diagnostic logic. For other deployment types, logs can be stored in single files, stored in files that combine node and master logs, or printed to stdout. If log entries do not use journald, the log diagnostics cannot work and do not run.

## 36.4. RUNNING DIAGNOSTICS IN A CLIENT ENVIRONMENT

You can run the diagnostics tool as an ordinary user or a **cluster-admin**, and it runs using the level of permissions granted to the account from which you run it.

A client with ordinary access can diagnose its connection to the master and run a diagnostic pod. If multiple users or masters are configured, connections are tested for all, but the diagnostic pod only runs against the current user, server, or project.

A client with **cluster-admin** access can diagnose the status of infrastructure such as nodes, registry, and router. In each case, running **oc adm diagnostics** searches for the standard client configuration file in its standard location and uses it if available.

## 36.5. ANSIBLE-BASED HEALTH CHECKS

Additional diagnostic health checks are available through the [Ansible-based tooling](#) used to install and manage OpenShift Container Platform clusters. They can report common deployment problems for the current OpenShift Container Platform installation.

These checks can be run either using the **ansible-playbook** command (the same method used during [Advanced Installation](#)) or as a [containerized version](#) of **openshift-ansible**. For the **ansible-playbook** method, the checks are provided by the **atomic-openshift-utils** RPM package. For the containerized method, the **openshift3/ose-ansible** container image is distributed via the [Red Hat Container Registry](#). Example usage for each method are provided in subsequent sections.

The following health checks are a set of diagnostic tasks that are meant to be run against the Ansible inventory file for a deployed OpenShift Container Platform cluster using the provided **health.yml** playbook.



## WARNING

Due to potential changes the health check playbooks can make to the environment, you must run the playbooks against only Ansible-deployed clusters and using the same inventory file used for deployment. The changes consist of installing dependencies so that the checks can gather the required information. In some circumstances, additional system components, such as **docker** or networking configurations, can change if their current state differs from the configuration in the inventory file. You should run these health checks only if you do not expect the inventory file to make any changes to the existing cluster configuration.

**Table 36.1. Diagnostic Health Checks**

| Check Name                 | Purpose                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>etcd_imagedata_size</b> | <p>This check measures the total size of OpenShift Container Platform image data in an etcd cluster. The check fails if the calculated size exceeds a user-defined limit. If no limit is specified, this check fails if the size of image data amounts to 50% or more of the currently used space in the etcd cluster.</p> <p>A failure from this check indicates that a significant amount of space in etcd is being taken up by OpenShift Container Platform image data, which can eventually result in the etcd cluster crashing.</p> <p>A user-defined limit may be set by passing the <b>etcd_max_image_data_size_bytes</b> variable. For example, setting <b>etcd_max_image_data_size_bytes=40000000000</b> causes the check to fail if the total size of image data stored in etcd exceeds 40 GB.</p> |
| <b>etcd_traffic</b>        | <p>This check detects higher-than-normal traffic on an etcd host. It fails if a <b>journalctl</b> log entry with an etcd sync duration warning is found.</p> <p>For further information on improving etcd performance, see <a href="#">Recommended Practices for OpenShift Container Platform etcd Hosts</a> and the <a href="#">Red Hat Knowledgebase</a>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                              |

| Check Name                                     | Purpose                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>etcd_volume</b>                             | <p>This check ensures that the volume usage for an etcd cluster is below a maximum user-specified threshold. If no maximum threshold value is specified, it is defaulted to <b>90%</b> of the total volume size.</p> <p>A user-defined limit may be set by passing the <b>etcd_device_usage_threshold_percent</b> variable.</p>                                                                                                                                                                                                                                                                                                     |
| <b>docker_storage</b>                          | <p>Only runs on hosts that depend on the <b>docker</b> daemon (nodes and containerized installations). Checks that <b>docker</b>'s total usage does not exceed a user-defined limit. If no user-defined limit is set, <b>docker</b>'s maximum usage threshold defaults to 90% of the total size available.</p> <p>You can set the threshold limit for total percent usage with a variable in the inventory file, for example <b>max_thinpool_data_usage_percent=90</b>.</p> <p>This also checks that <b>docker</b>'s storage is using a <a href="#">supported configuration</a>.</p>                                                |
| <b>curator, elasticsearch, fluentd, kibana</b> | <p>This set of checks verifies that Curator, Kibana, Elasticsearch, and Fluentd pods have been deployed and are in a <b>running</b> state, and that a connection can be established between the control host and the exposed Kibana URL. These checks run only if the <b>openshift_logging_install_logging</b> inventory variable is set to <b>true</b> to ensure that they are executed in a deployment where <a href="#">cluster logging</a> is enabled.</p>                                                                                                                                                                      |
| <b>logging_index_time</b>                      | <p>This check detects higher than normal time delays between log creation and log aggregation by Elasticsearch in a logging stack deployment. It fails if a new log entry cannot be queried through Elasticsearch within a timeout (by default, 30 seconds). The check only runs if logging is enabled.</p> <p>A user-defined timeout may be set by passing the <b>openshift_check_logging_index_timeout_seconds</b> variable. For example, setting <b>openshift_check_logging_index_timeout_seconds=45</b> causes the check to fail if a newly-created log entry is not able to be queried via Elasticsearch after 45 seconds.</p> |



## NOTE

A similar set of checks meant to run as part of the installation process can be found in [Configuring Cluster Pre-install Checks](#). Another set of checks for checking certificate expiration can be found in [Redeploying Certificates](#).

### 36.5.1. Running Health Checks via ansible-playbook

To run the **openshift-ansible** health checks using the **ansible-playbook** command, specify your cluster's inventory file and run the **health.yml** playbook:

```
# ansible-playbook -i <inventory_file> \
    /usr/share/ansible/openshift-ansible/playbooks/byo/openshift-
    checks/health.yml
```

To set variables in the command line, include the **-e** flag with any desired variables in **key=value** format. For example:

```
# ansible-playbook -i <inventory_file> \
    /usr/share/ansible/openshift-ansible/playbooks/byo/openshift-
    checks/health.yml
    -e openshift_check_logging_index_timeout_seconds=45
    -e etcd_max_image_data_size_bytes=40000000000
```

To disable specific checks, include the variable **openshift\_disable\_check** with a comma-delimited list of check names in your inventory file before running the playbook. For example:

```
openshift_disable_check=etcd_traffic,etcd_volume
```

Alternatively, set any checks to disable as variables with **-e openshift\_disable\_check=<check1>,<check2>** when running the **ansible-playbook** command.

### 36.5.2. Running Health Checks via Docker CLI

You can run the **openshift-ansible** playbooks in a Docker container, avoiding the need for installing and configuring Ansible, on any host that can run the **ose-ansible** image via the Docker CLI.

Run the following as a non-root user that has privileges to run containers:

```
# docker run -u `id -u` \ 1
    -v $HOME/.ssh/id_rsa:/opt/app-root/src/.ssh/id_rsa:Z,ro \ 2
    -v /etc/ansible/hosts:/tmp/inventory:ro \ 3
    -e INVENTORY_FILE=/tmp/inventory \
    -e PLAYBOOK_FILE=playbooks/byo/openshift-checks/health.yml \ 4
    -e OPTS="-v -e openshift_check_logging_index_timeout_seconds=45 -e
    etcd_max_image_data_size_bytes=40000000000" \ 5
    openshift3/ose-ansible
```

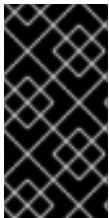
**1** These options make the container run with the same UID as the current user, which is required for permissions so that the SSH key can be read inside the container (SSH private keys are expected to be readable only by their owner).

**2**

Mount SSH keys as a volume under **`/opt/app-root/src/.ssh`** under normal usage when running the container as a non-root user.

- 3 Change **`/etc/ansible/hosts`** to the location of the cluster's inventory file, if different. This file is bind-mounted to **`/tmp/inventory`**, which is used according to the **`INVENTORY_FILE`** environment variable in the container.
- 4 The **`PLAYBOOK_FILE`** environment variable is set to the location of the **`health.yml`** playbook relative to **`/usr/share/ansible/openshift-ansible`** inside the container.
- 5 Set any variables desired for a single run with the **`-e key=value`** format.

In the previous command, the SSH key is mounted with the **`:Z`** option so that the container can read the SSH key from its restricted SELinux context. Adding this option means that your original SSH key file is relabeled similarly to **`system_u:object_r:container_file_t:s0:c113,c247`**. For more details about **`:Z`**, see the **`docker-run(1)`** man page.



### IMPORTANT

These volume mount specifications can have unexpected consequences. For example, if you mount, and therefore relabel, the **`$HOME/.ssh`** directory, **`sshd`** becomes unable to access the public keys to allow remote login. To avoid altering the original file labels, mount a copy of the SSH key or directory.

Mounting an entire **`.ssh`** directory can be helpful for:

- Allowing you to use an SSH configuration to match keys with hosts or modify other connection parameters.
- Allowing a user to provide a **`known_hosts`** file and have SSH validate host keys. This is disabled by the default configuration and can be re-enabled with an environment variable by adding **`-e ANSIBLE_HOST_KEY_CHECKING=True`** to the **`docker`** command line.



## CHAPTER 37. IDLING APPLICATIONS

### 37.1. OVERVIEW

As an OpenShift Container Platform administrator, you can idle applications to reduce resource consumption. This is useful when deployed on a public cloud where cost is related to resource consumption.

If any scalable resources are not in use, OpenShift Container Platform discovers, then idles them, by scaling them to 0 replicas. When network traffic is directed to the resources, they are unidled by scaling up the replicas, then operation continues.

Applications are made of services, as well as other scalable resources, such as deployment configurations. The action of idling an application involves idling all associated resources.

### 37.2. IDLING APPLICATIONS

Idling an application involves finding the scalable resources (deployment configurations, replication controllers, and others) associated with a service. Idling an application finds the service and marks it as idled, scaling down the resources to zero replicas.

You can use the **oc idle** command to [idle a single service](#), or use the **--resource-names-file** option to [idle multiple services](#).

#### 37.2.1. Idling Single Services

Idle a single service with the following command:

```
$ oc idle <service>
```

#### 37.2.2. Idling Multiple Services

Idle multiple services by creating a list of the desired services, then using the **--resource-names-file** option with the **oc idle** command.

This is helpful if an application spans across a set of services within a project, or when idling multiple services in conjunction with a script in order to idle multiple applications in bulk within the same project.

1. Create a text file containing a list of the services, each on their own line.
2. Idle the services using the **--resource-names-file** option:

```
$ oc idle --resource-names-file <filename>
```



#### NOTE

The idle command is limited to a single project. For idling applications across a cluster, run the idle command for each project individually.

### 37.3. UNIDLING APPLICATIONS

Application services become active again when they receive network traffic and will be scaled back up their previous state. This includes both traffic to the services and traffic passing through routes.

Applications may be manually unidled by scaling up the resources. For example, to scale up a deploymentconfig, run the command:

```
$ oc scale --replicas=1 dc <deploymentconfig>
```

**NOTE**

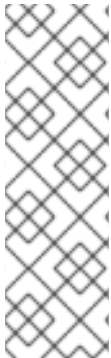
Automatic unidling by a router is currently only supported by the default HAProxy router.

## CHAPTER 38. ANALYZING CLUSTER CAPACITY

### 38.1. OVERVIEW

As a cluster administrator, you can use the cluster capacity tool to view the number of pods that can be scheduled to increase the current resources before they become exhausted, and to ensure any future pods can be scheduled. This capacity comes from an individual node host in a cluster, and includes CPU, memory, disk space, and others.

The cluster capacity tool simulates a sequence of scheduling decisions to determine how many instances of an input pod can be scheduled on the cluster before it is exhausted of resources to provide a more accurate estimation.



#### NOTE

The remaining allocatable capacity is a rough estimation, because it does not count all of the resources being distributed among nodes. It analyzes only the remaining resources and estimates the available capacity that is still consumable in terms of a number of instances of a pod with given requirements that can be scheduled in a cluster.

Also, pods might only have scheduling support on particular sets of nodes based on its selection and affinity criteria. As a result, the estimation of which remaining pods a cluster can schedule can be difficult.

You can run the cluster capacity analysis tool as a stand-alone utility from the command line, or [as a job](#) in a pod inside an OpenShift Container Platform cluster. Running it as job inside of a pod enables you to run it multiple times without intervention.

### 38.2. RUNNING CLUSTER CAPACITY ANALYSIS ON THE COMMAND LINE

To run the tool on the command line:

```
$ cluster-capacity --kubeconfig <path-to-kubeconfig> \
  --podspec <path-to-pod-spec>
```

The **--kubeconfig** option indicates your Kubernetes configuration file, and the **--podspec** option indicates a sample pod specification file, which the tool uses for estimating resource usage. The **podspec** specifies its resource requirements as **limits** or **requests**. The cluster capacity tool takes the pod's resource requirements into account for its estimation analysis.

An example of the pod specification input is:

```
apiVersion: v1
kind: Pod
metadata:
  name: small-pod
  labels:
    app: guestbook
    tier: frontend
spec:
  containers:
  - name: php-redis
```

```

image: gcr.io/google-samples/gb-frontend:v4
imagePullPolicy: Always
resources:
  limits:
    cpu: 150m
    memory: 100Mi
  requests:
    cpu: 150m
    memory: 100Mi

```

You can also add the `--verbose` option to output a detailed description of how many pods can be scheduled on each node in the cluster:

```

$ cluster-capacity --kubeconfig <path-to-kubeconfig> \
  --podspec <path-to-pod-spec> --verbose

```

The output will look similar to the following:

```

small-pod pod requirements:
- CPU: 150m
- Memory: 100Mi

```

The cluster can schedule 52 instance(s) of the pod small-pod.

```

Termination reason: Unschedulable: No nodes are available that match all
of the
following predicates:: Insufficient cpu (2).

```

```

Pod distribution among nodes:
small-pod
- 192.168.124.214: 26 instance(s)
- 192.168.124.120: 26 instance(s)

```

In the above example, the number of estimated pods that can be scheduled onto the cluster is 52.

### 38.3. RUNNING CLUSTER CAPACITY AS A JOB INSIDE OF A POD

Running the cluster capacity tool as a job inside of a pod has the advantage of being able to be run multiple times without needing user intervention. Running the cluster capacity tool as a job involves using a **ConfigMap**.

1. Create the cluster role:

```

$ cat << EOF | oc create -f -
kind: ClusterRole
apiVersion: v1
metadata:
  name: cluster-capacity-role
rules:
- apiGroups: [""]
  resources: ["pods", "nodes", "persistentvolumeclaims",
    "persistentvolumes", "services"]
  verbs: ["get", "watch", "list"]
EOF

```

2. Create the service account:

```
$ oc create sa cluster-capacity-sa
```

3. Add the role to the service account:

```
$ oc adm policy add-cluster-role-to-user cluster-capacity-role \
  system:serviceaccount:default:cluster-capacity-sa
```

4. Define and create the pod specification:

```
apiVersion: v1
kind: Pod
metadata:
  name: small-pod
  labels:
    app: guestbook
    tier: frontend
spec:
  containers:
  - name: php-redis
    image: gcr.io/google-samples/gb-frontend:v4
    imagePullPolicy: Always
    resources:
      limits:
        cpu: 150m
        memory: 100Mi
      requests:
        cpu: 150m
        memory: 100Mi
```

5. The cluster capacity analysis is mounted in a volume using a **ConfigMap** named **cluster-capacity-configmap** to mount input pod spec file **pod.yaml** into a volume **test-volume** at the path **/test-pod**.

If you haven't created a **ConfigMap**, create one before creating the job:

```
$ oc create configmap cluster-capacity-configmap \
  --from-file=pod.yaml=pod.yaml
```

6. Create the job using the below example of a job specification file:

```
apiVersion: batch/v1
kind: Job
metadata:
  name: cluster-capacity-job
spec:
  parallelism: 1
  completions: 1
  template:
    metadata:
      name: cluster-capacity-pod
    spec:
      containers:
      - name: cluster-capacity
```

```

        image: openshift/origin-cluster-capacity
        imagePullPolicy: "Always"
        volumeMounts:
        - mountPath: /test-pod
          name: test-volume
        env:
        - name: CC_INCLUSTER 1
          value: "true"
        command:
        - "/bin/sh"
        - "-ec"
        - |
          /bin/cluster-capacity --podspec=/test-pod/pod.yaml --
verbose
        restartPolicy: "Never"
        serviceAccountName: cluster-capacity-sa
        volumes:
        - name: test-volume
          configMap:
            name: cluster-capacity-configmap

```

- 1** A required environment variable letting the cluster capacity tool know that it is running inside a cluster as a pod.
- The **pod.yaml** key of the **ConfigMap** is the same as the pod specification file name, though it is not required. By doing this, the input pod spec file can be accessed inside the pod as **/test-pod/pod.yaml**.

7. Run the cluster capacity image as a job in a pod:

```
$ oc create -f cluster-capacity-job.yaml
```

8. Check the job logs to find the number of pods that can be scheduled in the cluster:

```

$ oc logs jobs/cluster-capacity-job
small-pod pod requirements:
- CPU: 150m
- Memory: 100Mi

```

The cluster can schedule 52 instance(s) of the pod small-pod.

Termination reason: Unschedulable: No nodes are available that match all of the following predicates:: Insufficient cpu (2).

```

Pod distribution among nodes:
small-pod
- 192.168.124.214: 26 instance(s)
- 192.168.124.120: 26 instance(s)

```

## CHAPTER 39. REVISION HISTORY: CLUSTER ADMINISTRATION

### 39.1. TUES MAR 06 2018

| Affected Topic                                        | Description of Change                                                              |
|-------------------------------------------------------|------------------------------------------------------------------------------------|
| <a href="#">Managing Security Context Constraints</a> | Added a new <a href="#">Example Security Context Constraints Settings</a> section. |

### 39.2. FRI FEB 23 2018

| Affected Topic                                        | Description of Change                                                                                                                   |
|-------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Default Scheduling</a>                    | Reorganized topic and updated the list of predicates and policies in the <a href="#">Scheduler Policy</a> section.                      |
| <a href="#">Controlling Pod Placement</a>             | Created new topic from the Controlling Pod Placement section of the <a href="#">Default Scheduling</a> topic. No change to the content. |
| <a href="#">Managing Security Context Constraints</a> | Noted the importance of <code>-z</code> flag usage when <a href="#">granting access to service accounts</a> .                           |
| <a href="#">Configuring Service Accounts</a>          | Noted the importance of <code>-z</code> flag usage when <a href="#">granting access to service accounts</a> .                           |

### 39.3. TUES FEB 20 2018

| Affected Topic                 | Description of Change                                                                                        |
|--------------------------------|--------------------------------------------------------------------------------------------------------------|
| <a href="#">Managing Nodes</a> | Replaced outdated <code>oadm manage-node --evacuate</code> commands with <code>oc adm drain</code> commands. |

### 39.4. FRI FEB 16 2018

| Affected Topic                                  | Description of Change                                                      |
|-------------------------------------------------|----------------------------------------------------------------------------|
| <a href="#">Handling Out of Resource Errors</a> | Adjusted the math in the <a href="#">Example Scenario</a> .                |
| <a href="#">Garbage Collection</a>              | Added clarifying details about the default behavior of garbage collection. |

### 39.5. TUE FEB 06 2018

| Affected Topic                      | Description of Change                                                                                                                                                                                          |
|-------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Managing Networking</a> | Changed the <a href="#">Disabling Host Name Collision Prevention For Routes and Ingress Objects</a> section to mention the ability to give users the rights to edits host names on routes and ingress objects. |

## 39.6. TUE NOV 21 2017

| Affected Topic                 | Description of Change                                                                    |
|--------------------------------|------------------------------------------------------------------------------------------|
| <a href="#">Managing Nodes</a> | Added new section on <a href="#">Resetting Docker Storage</a> to free up space on nodes. |

## 39.7. FRI NOV 10 2017

| Affected Topic                   | Description of Change                                                                                         |
|----------------------------------|---------------------------------------------------------------------------------------------------------------|
| <a href="#">Service Accounts</a> | Changed <b>serviceaccounts</b> to <b>serviceaccount</b> in the <a href="#">User Names and Groups</a> section. |

## 39.8. FRI NOV 03 2017

| Affected Topic                                     | Description of Change                                                           |
|----------------------------------------------------|---------------------------------------------------------------------------------|
| <a href="#">Encrypting Data at Datastore Layer</a> | Added a note that etcd v3 or later is required in order to use data encryption. |

## 39.9. TUE OCT 24 2017

| Affected Topic                     | Description of Change                                               |
|------------------------------------|---------------------------------------------------------------------|
| <a href="#">Backup and Restore</a> | Added a new <a href="#">Containerized etcd Deployments</a> section. |

## 39.10. WED OCT 11 2017

| Affected Topic                     | Description of Change                                             |
|------------------------------------|-------------------------------------------------------------------|
| <a href="#">Pruning Objects</a>    | Added details on secure versus insecure image pruning.            |
| <a href="#">Backup and Restore</a> | Added a new <a href="#">Registry Certificates Backup</a> section. |

## 39.11. MON OCT 02 2017



| Affected Topic                                     | Description of Change                                                                                                      |
|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Managing Networking</a>                | Added more information to the <a href="#">VMWare vSphere</a> section.                                                      |
| <a href="#">Encrypting Data at Datastore Layer</a> | Removed the "experimental support" language for the data encryption feature, as the feature is fully supported as of v3.6. |

## 39.12. MON SEP 18 2017

| Affected Topic                           | Description of Change                                                                              |
|------------------------------------------|----------------------------------------------------------------------------------------------------|
| <a href="#">Diagnostics Tool</a>         | Added more information about tool usage to the <a href="#">Using the Diagnostics Tool</a> section. |
| <a href="#">Opaque Integer Resources</a> | Moved information on opaque integer resources to Administrator Guide                               |
| <a href="#">Setting Limit Ranges</a>     | Added link to information on how CPU and memory are calculated.                                    |

## 39.13. FRI SEP 08 2017

| Affected Topic                                     | Description of Change                                                                      |
|----------------------------------------------------|--------------------------------------------------------------------------------------------|
| <a href="#">Encrypting Data at Datastore Layer</a> | New topic on how to enable and configure encryption of secret data at the datastore layer. |

## 39.14. TUE AUG 29 2017

| Affected Topic                                | Description of Change                                                                                                                                                |
|-----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Image Policy</a>                  | Added note clarifying the need for the image prefix to set the default registry string in the <a href="#">Configuring the ImagePolicy Admission Plug-in</a> section. |
| <a href="#">Pruning Objects</a>               | Added valid units of measurement for <b>--keep-younger-than</b> .                                                                                                    |
| <a href="#">Troubleshooting OpenShift SDN</a> | Changed the Further Help section to <a href="#">Finding Network Issues Using the Diagnostics Tool</a> and added information about the Diagnostic Tool.               |
| <a href="#">Troubleshooting OpenShift SDN</a> | Corrected <b>vxlan0</b> to <b>vxlan_sys_4789</b> in the <a href="#">Debugging Local Networking</a> section.                                                          |

## 39.15. TUE AUG 22 2017

| Affected Topic                      | Description of Change                                                                                                                                                            |
|-------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Managing Networking</a> | Added admonition to the <a href="#">Using an Egress Router to Allow External Resources to Recognize Pod Traffic</a> section about Amazon AWS not working with the egress router. |
| <a href="#">Diagnostics Tool</a>    | Enhanced the <a href="#">Ansible-based Health Checks</a> section with information on running via ansible-playbook or Docker CLI.                                                 |

## 39.16. MON AUG 14 2017

| Affected Topic                     | Description of Change                                                                                                 |
|------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| <a href="#">Garbage Collection</a> | Changed the <b>image-gc-high-threshold</b> default value to 85 from 90.                                               |
| <a href="#">High Availability</a>  | Added verbiage clarifying the example outlined in the <a href="#">Configuring a Highly-available Service</a> section. |

## 39.17. WED AUG 09 2017

OpenShift Container Platform 3.6 Initial Release

| Affected Topic                      | Description of Change                                                                                                                                                                                                                                                                      |
|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Managing Pods</a>       | Added information about allowing domain names in <b>EgressNetworkPolicy</b> .                                                                                                                                                                                                              |
| <a href="#">Managing Networking</a> | Added an admonition about DNS and egress network policy to the <a href="#">Using an Egress Firewall to Limit Access to External Resources</a> section.                                                                                                                                     |
|                                     | Added procedure to the <a href="#">Enabling NetworkPolicy</a> section.                                                                                                                                                                                                                     |
|                                     | Removed the Technology Preview designation for SDN Multicast.                                                                                                                                                                                                                              |
|                                     | Added the <a href="#">Using iptables Rules to Limit Access to External Resources</a> section, and various edits.                                                                                                                                                                           |
|                                     | Added a note about limitations with the egress network policy.                                                                                                                                                                                                                             |
|                                     | Added the <a href="#">Egress Router Modes</a> , <a href="#">Redirecting to Multiple Destinations</a> , <a href="#">Using a ConfigMap to specify EGRESS_DESTINATION</a> , and <a href="#">Deploying an Egress Router Pod in Redirect Mode</a> sections, as well as various content changes. |
|                                     | Added the <a href="#">Disabling Host Name Collision Prevention For Ingress Objects</a> section.                                                                                                                                                                                            |
| <a href="#">Image Policy</a>        | Added details about using image streams in Kubernetes resources.                                                                                                                                                                                                                           |

| Affected Topic                                   | Description of Change                                                                                                                                                                                        |
|--------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Image Signatures</a>                 | Added the <a href="#">Verifying Image Signatures Using OpenShift CLI</a> section.                                                                                                                            |
| <a href="#">Scheduling</a>                       | Added subsections for new scheduling features. Moved the current Scheduling topic into section as <a href="#">Default Scheduling</a> .                                                                       |
| <a href="#">Setting Quotas</a>                   | Added list of storage resources that can be managed by quota to the <a href="#">Resources Managed by Quota</a> section and added gold and bronze storage classes to <b>storage-consumption.yaml</b> example. |
| <a href="#">Pruning Objects</a>                  | Added note to <a href="#">Pruning Builds</a> linking to the <a href="#">Build Pruning</a> section.                                                                                                           |
| <a href="#">Overcommitting</a>                   | Added the <a href="#">Tune Buffer Chunk Limit</a> section.                                                                                                                                                   |
|                                                  | Added new section on <a href="#">reserving resources for pods based on QOS level</a> .                                                                                                                       |
| <a href="#">Monitoring and Debugging Routers</a> | Described the <b>ROUTER_SYSLOG_FORMAT</b> environment variable.                                                                                                                                              |
| <a href="#">Diagnostics Tool</a>                 | Added <a href="#">Additional Diagnostic Checks via Ansible</a> section.                                                                                                                                      |
| <a href="#">Analyzing Cluster Capacity</a>       | Added the Analyzing Cluster Capacity file.                                                                                                                                                                   |