⚠ A translation for your language preference does not exist.

Unable to create more than 1024 Threads in OpenShift 4

SOLUTION 已验证 - 已更新 2022年十二月6日20:14 - English

环境

- Red Hat OpenShift Container Platform (RHOCP)
  - 4
- Red Hat OpenShift Service on AWS (ROSA)
  - 4.10
  - 4.9
  - 4.8
  - 4.7
  - For 4.11 ROSA clusters, the process is different. See this article for the process.

问题

- How to change the value of `pids_limit` in OpenShift 4?
- Getting the following exception when migrating from OCP3 to OCP4:

```
java.lang.OutOfMemoryError: unable to create new native thread
```

决议

A workaround is to increase the `pids_limit` in both the `KubeletConfig` and `ContainerRuntimeConfig` .

Note: For OSD and ROSA, refer to How to change the pids_limit in OSD or ROSA.

1. The steps below assume the cluster has no current `kubeletconfigs`. Verify this is true before proceeding:

```
$ oc get kubeletconfigs
No resources found in openshift-etcd namespace
```

2. Label the worker pool:

```
$ oc label machineconfigpool worker custom-crio=high-pid-limit custom-
kubelet=small-pods
```

3. Create a custom `KubeletConfig` with the following command, setting the **POD_PIDS_LIMIT** value as needed:

```
$ POD_PIDS_LIMIT=4096
$ oc apply -f - <<EOF
apiVersion: machineconfiguration.openshift.io/v1
kind: KubeletConfig
metadata:
  name: worker-kubeconfig-fix
spec:
  machineConfigPoolSelector:
    matchLabels:
      custom-kubelet: small-pods
  kubeletConfig:
      podPidsLimit: $POD_PIDS_LIMIT
EOF
```

4. Create a custom `ContainerRuntimeConfig` with the following command, setting the **PIDS_LIMIT** value according to the OpenShift version:

```
$ PIDS_LIMIT=4096
$ oc apply -f - <<EOF
apiVersion: machineconfiguration.openshift.io/v1
kind: ContainerRuntimeConfig
metadata:
 name: set-pids-limit
spec:
 machineConfigPoolSelector:
   matchLabels:
     custom-crio: high-pid-limit
 containerRuntimeConfig:
   pidsLimit: $PIDS_LIMIT
EOF
```

Note:

- Use `pidsLimit: -1` value for OpenShift versions ***4.5 to 4.7***. `pidsLimit: -1` means no limit will be enforced by CRI-O, and limits set by the kubelet will be honored.

- In OpenShift ***4.4 and older versions***, `pidsLimit: -1` has no effect and it is needed to specify a matching value to `podPidsLimit` in the `KubeletConfig` .

- In OpenShift ***4.8 and newer versions***, `pidsLimit: -1` is not valid. Per Red Hat BZ 2039187, it must be set to a value greater than 20. It is needed to specify a matching value to `podPidsLimit` in the `KubeletConfig` .
  The initial creation of a cluster may interfere with this configuration taking hold and sometimes may need more time before applying. Red Hat investigated this issue in bug report BZ 2100894 and delivered a fix in OpenShift 4.10.25 through errata RHSA-2022:5730. If this issue still occurs in your environment after updating, open a support case in the Red Hat Customer Portal referring to this solution.

IMPORTANT: Running the `oc apply` commands will result in a new Machine Config (MC) being created. This MC will be rolled out to each worker node. **Each node will cordon, drain, apply config, reboot, and uncordon**; so it is expected for the nodes to **reboot**.

根源

- It's needed to configure the `KubeletConfig` and `ContainerRuntimeConfig` resources.
- The `KubeletConfig` is the configuration for pods, while `ContainerRuntimeConfig` is for configuration for containers.

诊断步骤

1. Monitor `/sys/fs/cgroup/pids/pids.current` when the application is running to verify `java.lang.OutOfMemoryError: unable to create new native thread` or similar errors happen when it hits 1024.

2. Check if the CRI-O `pids_limit` is being set on the node where the application container is running:

```
$ crio config | grep pids_limit
INFO[2022-01-31 12:14:27.407346183Z] Starting CRI-O, version: 1.21.4-
4.rhaos4.8.git84fa55d.el8, git: ()
INFO Using default capabilities: CAP_CHOWN, CAP_DAC_OVERRIDE, CAP_FSETID,
CAP_FOWNER, CAP_SETGID, CAP_SETUID, CAP_SETPCAP, CAP_NET_BIND_SERVICE,
CAP_KILL
pids_limit = 4096
```

If not, the default (1024) applies.

3. Verify the kubelet `pids_limit` is being set in `/etc/kubernetes/kubelet.conf` and `SupportPodPidsLimit` is set running the following command:

```
$ oc debug node/ip-10-0-221-49.us-west-2.compute.internal -- cat
/host/etc/kubernetes/kubelet.conf | jq '.podPidsLimit, .featureGates'
Starting pod/ip-10-0-221-49us-west-2computeinternal-debug ...
To use host binaries, run `chroot /host`
Removing debug pod ...

2048
{
  "LegacyNodeRoleBehavior": false,
  "NodeDisruptionExclusion": true,
  "RotateKubeletServerCertificate": true,
  "SCTPSupport": true,
  "ServiceNodeExclusion": true,
  "SupportPodPidsLimit": true
}
```

If not, the default (1024) applies.

4. Verify the labels for the `ContainerRuntimeConfig` and `KubeletConfig` were created and applied:

```
$ oc get kubeletconfig,containerruntimeconfig
NAME                                AGE
kubeletconfig/worker-kubeconfig-fix  9d

NAME                                   AGE
containerruntimeconfig/set-pids-limit  15d

$ oc get mcp/worker -ojson | jq '.metadata.labels'
{
  "custom-crio": "high-pid-limit",
  "custom-kubelet": "small-pods",
  "machineconfiguration.openshift.io/mco-built-in": "",
  "pools.operator.machineconfiguration.openshift.io/worker": ""
}

$ oc get kubeletconfig/worker-kubeconfig-fix -ojson | jq
'.status.conditions[]'
{
  "lastTransitionTime": "2022-02-10T04:46:17Z",
  "message": "Success",
  "status": "True",
  "type": "Success"
}
```

产品（第） Red Hat OpenShift Container Platform Red Hat OpenShift Dedicated

元件 cri-o kubernetes openshift-node **类别** Customize or extend

标记 container_runtime crio kubernetes ocp_4 openshift shift_operator

This solution is part of Red Hat's fast-track publication program, providing a huge library of solutions that Red Hat engineers have created while supporting our customers. To give you the knowledge you need the instant it becomes available, these articles may be presented in a raw and unedited form.

People who viewed this solution also viewed

Apply KubeletConfig on OpenShift Infra pool

Solution - 2021年8月6日

Machine Config Operator is not merging different KubeletConfigs in Openshift 4

Solution - 2021年12月17日

Why doesn't ContainerRuntimeConfig get removed when the associated MachineConfigPool label is removed?

Solution - 2020年9月23日

6 评论

23 March 2021 5:10 PM

Benoit Hannebicq

COMMUNITY MEMBER

37 Points

The solution is not well updated,

回复

COMMUNITY MEMBER

26 Points

4 June 2022 7:20 AM

Lukman Hakim

works for me. I just verified the diagnostic steps given in the article and not through my application as it might take some time to do that.

回复

NEWBIE

5 Points

10 August 2021 8:07 PM

Kaleb Eckles

In the openshift mco github project, it gives the specific example of setting pidsLimit via a ContainerRuntimeConfig. (https://github.com/openshift/machine-config-operator/blob/master/docs/ContainerRuntimeConfigDesign.md). What is the need for the KubeletConfig and where is it documented that it is needed to set pidsLimit?

This doc does not work for me.

回复

NEWBIE

12 Points

23 November 2021 9:26 PM

Andrew Mellanby RHCSA

On ocp 4.8, I attempted to set the pids limit as described in the article: I see that it failed with this message: - lastTransitionTime: "2021-11-04T02:18:15Z" message: 'Error: invalid PidsLimit -1' status: "False" type: Failure

However it IS possible to set the limit to an arbitrary high number .. eg. 32000

回复

COMMUNITY MEMBER

26 Points

4 June 2022 7:19 AM

Lukman Hakim

Working for me after following all the steps. I applied it on one worker only by creating a custom mcp by following [Red Hat Customer Portal] (https://access.redhat.com/solutions/5688941).

When using custom label, need to use the below labels (replace worker with custom as in original solution to apply for particular node)

```
oc label machineconfigpool custom custom-crio=high-pid-limit
oc label machineconfigpool custom custom-kubelet=small-pods
```

回复

RED HAT

COMMUNITY MEMBER

55 Points

22 September 2022 2:02 AM

Anand R

By default KubeletConfig podPidsLimit is -1, So defining only CRC will be enough.

回复