Red Hat OpenShift AI Self-Managed 2-latest

Working with accelerators

Working with accelerators from Red Hat OpenShift AI Self-Managed

Last Updated: 2024-08-27

Red Hat OpenShift AI Self-Managed 2-latest Working with accelerators

Working with accelerators from Red Hat OpenShift AI Self-Managed

Legal Notice

Copyright © 2024 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at
http://creativecommons.org/licenses/by-sa/3.0/
. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

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

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

Linux ® 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 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

Use accelerators to optimize the performance of your end-to-end data science workflows.

Table of Contents

PREFACE

Use accelerators, such as NVIDIA GPUs and Intel Gaudi AI accelerators, to optimize the performance of your end-to-end data science workflows.

CHAPTER 1. OVERVIEW OF ACCELERATORS

If you work with large data sets, you can use accelerators to optimize the performance of your data science models in OpenShift AI. With accelerators, you can scale your work, reduce latency, and increase productivity. You can use accelerators in OpenShift AI to assist your data scientists in the following tasks:

- Natural language processing (NLP)
- Inference
- Training deep neural networks
- Data cleansing and data processing

OpenShift AI supports the following accelerators:

- NVIDIA graphics processing units (GPUs)
  - To use compute-heavy workloads in your models, you can enable NVIDIA graphics processing units (GPUs) in OpenShift AI.
  - To enable GPUs on OpenShift, you must install the NVIDIA GPU Operator.
- Intel Gaudi AI accelerators
  - Intel provides hardware accelerators intended for deep learning workloads. You can use the Habana libraries and software associated with Intel Gaudi AI accelerators available from your notebook.
  - Before you can enable Intel Gaudi AI accelerators in OpenShift AI, you must install the necessary dependencies and the version of the HabanaAI Operator that matches the Habana version of the HabanaAI workbench image in your deployment. For more information about how to enable your OpenShift environment for Intel Gaudi AI accelerators, see HabanaAI Operator v1.10 for OpenShift and HabanaAI Operator v1.13 for OpenShift.
  - You can enable Intel Gaudi AI accelerators on-premises or with AWS DL1 compute nodes on an AWS instance.

Before you can use an accelerator in OpenShift AI, your OpenShift instance must contain an associated accelerator profile. For accelerators that are new to your deployment, you must configure an accelerator profile for the accelerator in context. You can create an accelerator profile from the **Settings → Accelerator profiles** page on the OpenShift AI dashboard. If your deployment contains existing accelerators that had associated accelerator profiles already configured, an accelerator profile is automatically created after you upgrade to the latest version of OpenShift AI.

Additional resources

- HabanaAI Operator v1.10 for OpenShift
- HabanaAI Operator v1.13 for OpenShift
- Habana, an Intel Company
- Amazon EC2 DL1 Instances

- lspci(8) - Linux man page

CHAPTER 2. ENABLING NVIDIA GPUS

Before you can use NVIDIA GPUs in OpenShift AI, you must install the NVIDIA GPU Operator.

Prerequisites

- You have logged in to your OpenShift cluster.
- You have the **cluster-admin** role in your OpenShift cluster.

Procedure

1. To enable GPU support on an OpenShift cluster in a disconnected or airgapped environment, follow the instructions here: Deploy GPU Operators in a disconnected or airgapped environment in the NVIDIA documentation.
2. Delete the **migration-gpu-status** ConfigMap.
   a. In the OpenShift web console, switch to the **Administrator** perspective.
   b. Set the **Project** to **All Projects** or **redhat-ods-applications** to ensure you can see the appropriate ConfigMap.
   c. Search for the **migration-gpu-status** ConfigMap.
   d. Click the action menu (⋮) and select **Delete ConfigMap** from the list.
   The **Delete ConfigMap** dialog appears.
   e. Inspect the dialog and confirm that you are deleting the correct ConfigMap.
   f. Click **Delete**.
3. Restart the dashboard replicaset.
   a. In the OpenShift web console, switch to the **Administrator** perspective.
   b. Click **Workloads** → **Deployments**.
   c. Set the **Project** to **All Projects** or **redhat-ods-applications** to ensure you can see the appropriate deployment.
   d. Search for the **rhods-dashboard** deployment.
   e. Click the action menu (⋮) and select **Restart Rollout** from the list.
   f. Wait until the **Status** column indicates that all pods in the rollout have fully restarted.

Verification

- The NVIDIA GPU Operator appears on the **Operators** → **Installed Operators** page in the OpenShift web console.
- The reset **migration-gpu-status** instance is present on the **Instances** tab on the **AcceleratorProfile** custom resource definition (CRD) details page.

NOTE

In OpenShift AI 2-latest, Red Hat supports the use of accelerators within the same cluster only. Red Hat does not support remote direct memory access (RDMA) between accelerators, or the use of accelerators across a network, for example, by using technology such as NVIDIA GPUDirect or NVLink.

After installing the NVIDIA GPU Operator, create an accelerator profile as described in Working with accelerator profiles.

CHAPTER 3. INTEL GAUDI AI ACCELERATOR INTEGRATION

To accelerate your high-performance deep learning (DL) models, you can integrate Intel Gaudi AI accelerators in OpenShift AI. OpenShift AI also includes the HabanaAI workbench image, which is pre-built and ready for your data scientists to use after you install or upgrade OpenShift AI.

Before you can enable Intel Gaudi AI accelerators in OpenShift AI, you must install the necessary dependencies and the version of the HabanaAI Operator that matches the Habana version of the HabanaAI workbench image in your deployment. This allows your data scientists to use Habana libraries and software associated with Intel Gaudi AI accelerators from their workbench.

For more information about how to enable your OpenShift environment for Intel Gaudi AI accelerators, see HabanaAI Operator v1.10 for OpenShift and HabanaAI Operator v1.13 for OpenShift.

IMPORTANT

Currently, Intel Gaudi AI Accelerator integration is only supported in OpenShift 4.12.

You can use Intel Gaudi AI accelerators on OpenShift AI with versions 1.10.0 and 1.13.0 of the Habana AI Operator. The version of the HabanaAI Operator that you install must match the Habana version of the HabanaAI workbench image in your deployment. This means that only one version of HabanaAI workbench image will work for you at a time.

For information about the supported configurations for versions 1.10 and 1.13 of the Habana AI Operator, see Support Matrix v1.10.0 and Support Matrix v1.13.0.

You can use Intel Gaudi AI accelerators in an Amazon EC2 DL1 instance on OpenShift. Therefore, your OpenShift platform must support EC2 DL1 instances. Intel Gaudi AI accelerators are available to your data scientists when they create a workbench instance or serve a model.

To identify the Intel Gaudi AI accelerators present in your deployment, use the **lspci** utility. For more information, see lspci(8) - Linux man page .

IMPORTANT

If the **lspci** utility indicates that Intel Gaudi AI accelerators are present in your deployment, it does not necessarily mean that the devices are ready to use.

Before you can use your Intel Gaudi AI accelerators, you must enable them in your OpenShift environment and configure an accelerator profile for each device. For more information about how to enable your OpenShift environment for Intel Gaudi AI accelerators, see HabanaAI Operator for OpenShift.

Additional resources

- HabanaAI Operator v1.10 for OpenShift
- HabanaAI Operator v1.13 for OpenShift
- lspci(8) - Linux man page
- Amazon EC2 DL1 Instances
- Support Matrix v1.10.0

- Support Matrix v1.13.0
- What version of the Kubernetes API is included with each OpenShift 4.x release?

3.1. ENABLING INTEL GAUDI AI ACCELERATORS

Before you can use Intel Gaudi AI accelerators in OpenShift AI, you must install the necessary dependencies and deploy the HabanaAI Operator.

Prerequisites

- You have logged in to OpenShift.
- You have the **cluster-admin** role in OpenShift.

Procedure

1. To enable Intel Gaudi AI accelerators in OpenShift AI, follow the instructions at HabanaAI Operator for OpenShift.

Verification

- From the **Administrator** perspective, the following Operators appear on the **Operators → Installed Operators** page.
    - HabanaAI
    - Node Feature Discovery (NFD)
    - Kernel Module Management (KMM)

After installing the HabanaAI Operator, create an accelerator profile as described in Working with accelerator profiles.

Additional resources

- HabanaAI Operator v1.10 for OpenShift.
- HabanaAI Operator v1.13 for OpenShift.

CHAPTER 4. WORKING WITH ACCELERATOR PROFILES

To configure accelerators for your data scientists to use in OpenShift AI, you must create an associated accelerator profile. An accelerator profile is a custom resource definition (CRD) on OpenShift that has an AcceleratorProfile resource, and defines the specification of the accelerator. You can create and manage accelerator profiles by selecting **Settings → Accelerator profiles** on the OpenShift AI dashboard.

For accelerators that are new to your deployment, you must manually configure an accelerator profile for each accelerator. If your deployment contains an accelerator before you upgrade, the associated accelerator profile remains after the upgrade. You can manage the accelerators that appear to your data scientists by assigning specific accelerator profiles to your custom notebook images. This example shows the code for a Habana Gaudi 1 accelerator profile:

```
---
apiVersion: dashboard.opendatahub.io/v1alpha
kind: AcceleratorProfile
metadata:
  name: hpu-profile-first-gen-gaudi
spec:
  displayName: Habana HPU - 1st Gen Gaudi
  description: First Generation Habana Gaudi device
  enabled: true
  identifier: habana.ai/gaudi
  tolerations:
    - effect: NoSchedule
      key: habana.ai/gaudi
      operator: Exists
---
```

The accelerator profile code appears on the **Instances** tab on the details page for the **AcceleratorProfile** custom resource definition (CRD). For more information about accelerator profile attributes, see the following table:

Table 4.1. Accelerator profile attributes

Attribute	Type	Required	Description
displayName	String	Required	The display name of the accelerator profile.
description	String	Optional	Descriptive text defining the accelerator profile.
identifier	String	Required	A unique identifier defining the accelerator resource.
enabled	Boolean	Required	Determines if the accelerator is visible in OpenShift AI.
tolerations	Array	Optional	The tolerations that can apply to notebooks and serving runtimes that use the accelerator. For more information about the toleration attributes that OpenShift AI supports, see Toleration v1 core.

Additional resources

- Toleration v1 core
- Understanding taints and tolerations
- Managing resources from custom resource definitions

4.1. VIEWING ACCELERATOR PROFILES

If you have defined accelerator profiles for OpenShift AI, you can view, enable, and disable them from the **Accelerator profiles** page.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are assigned the **cluster-admin** role in OpenShift.
- Your deployment contains existing accelerator profiles.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Accelerator profiles**.
   The **Accelerator profiles** page appears, displaying existing accelerator profiles.
2. Inspect the list of accelerator profiles. To enable or disable an accelerator profile, on the row containing the accelerator profile, click the toggle in the **Enable** column.

Verification

- The **Accelerator profiles** page appears appears, displaying existing accelerator profiles.

4.2. CREATING AN ACCELERATOR PROFILE

To configure accelerators for your data scientists to use in OpenShift AI, you must create an associated accelerator profile.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are assigned the **cluster-admin** role in OpenShift.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Accelerator profiles**.
   The **Accelerator profiles** page appears, displaying existing accelerator profiles. To enable or disable an existing accelerator profile, on the row containing the relevant accelerator profile, click the toggle in the **Enable** column.
2. Click **Create accelerator profile**.
   The **Create accelerator profile** dialog appears.

3. In the **Name** field, enter a name for the accelerator profile.

4. In the **Identifier** field, enter a unique string that identifies the hardware accelerator associated with the accelerator profile.

5. Optional: In the **Description** field, enter a description for the accelerator profile.

6. To enable or disable the accelerator profile immediately after creation, click the toggle in the **Enable** column.

7. Optional: Add a toleration to schedule pods with matching taints.

    a. Click **Add toleration**.
    The **Add toleration** dialog opens.

    b. From the **Operator** list, select one of the following options:

    - **Equal** - The **key/value/effect** parameters must match. This is the default.

    - **Exists** - The **key/effect** parameters must match. You must leave a blank value parameter, which matches any.

    c. From the **Effect** list, select one of the following options:

    - **None**

    - **NoSchedule** - New pods that do not match the taint are not scheduled onto that node. Existing pods on the node remain.

    - **PreferNoSchedule** - New pods that do not match the taint might be scheduled onto that node, but the scheduler tries not to. Existing pods on the node remain.

    - **NoExecute** - New pods that do not match the taint cannot be scheduled onto that node. Existing pods on the node that do not have a matching toleration are removed.

    d. In the **Key** field, enter a toleration key. The key 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.

    e. In the **Value** field, enter a toleration value. The value 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.

    f. In the **Toleration Seconds** section, select one of the following options to specify how long a pod stays bound to a node that has a node condition.

    - **Forever** - Pods stays permanently bound to a node.

    - **Custom value** - Enter a value, in seconds, to define how long pods stay bound to a node that has a node condition.

    g. Click **Add**.

8. Click **Create accelerator profile**.

Verification

- The accelerator profile appears on the **Accelerator profiles** page.

- The **Accelerator** list appears on the **Start a notebook server** page. After you select an accelerator, the **Number of accelerators** field appears, which you can use to choose the number of accelerators for your notebook server.

- The accelerator profile appears on the **Instances** tab on the details page for the **AcceleratorProfile** custom resource definition (CRD).

Additional resources

- Toleration v1 core

- Understanding taints and tolerations

- Managing resources from custom resource definitions

4.3. UPDATING AN ACCELERATOR PROFILE

You can update the existing accelerator profiles in your deployment. You might want to change important identifying information, such as the display name, the identifier, or the description.

Prerequisites

- You have logged in to Red Hat OpenShift AI.

- You are assigned the **cluster-admin** role in OpenShift.

- The accelerator profile exists in your deployment.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Notebook images**.
   The **Notebook images** page appears. Previously imported notebook images are displayed. To enable or disable a previously imported notebook image, on the row containing the relevant notebook image, click the toggle in the **Enable** column.

2. Click the action menu (⋮) and select **Edit** from the list.
   The **Edit accelerator profile** dialog opens.

3. In the **Name** field, update the accelerator profile name.

4. In the **Identifier** field, update the unique string that identifies the hardware accelerator associated with the accelerator profile, if applicable.

5. Optional: In the **Description** field, update the accelerator profile.

6. To enable or disable the accelerator profile immediately after creation, click the toggle in the **Enable** column.

7. Optional: Add a toleration to schedule pods with matching taints.

   a. Click **Add toleration**.
      The **Add toleration** dialog opens.

   b. From the **Operator** list, select one of the following options:

      - **Equal** - The **key/value/effect** parameters must match. This is the default.

- **Exists** - The **key/effect** parameters must match. You must leave a blank value parameter, which matches any.

c. From the **Effect** list, select one of the following options:

- **None**
- **NoSchedule** - New pods that do not match the taint are not scheduled onto that node. Existing pods on the node remain.
- **PreferNoSchedule** - New pods that do not match the taint might be scheduled onto that node, but the scheduler tries not to. Existing pods on the node remain.
- **NoExecute** - New pods that do not match the taint cannot be scheduled onto that node. Existing pods on the node that do not have a matching toleration are removed.

d. In the **Key** field, enter a toleration key. The key 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.

e. In the **Value** field, enter a toleration value. The value 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.

f. In the **Toleration Seconds** section, select one of the following options to specify how long a pod stays bound to a node that has a node condition.

- **Forever** - Pods stays permanently bound to a node.
- **Custom value** - Enter a value, in seconds, to define how long pods stay bound to a node that has a node condition.

g. Click **Add**.

8. If your accelerator profile contains existing tolerations, you can edit them.

   a. Click the action menu (⋮) on the row containing the toleration that you want to edit and select **Edit** from the list.

   b. Complete the applicable fields to update the details of the toleration.

   c. Click **Update**.

9. Click **Update accelerator profile**.

Verification

- If your accelerator profile has new identifying information, this information appears in the **Accelerator** list on the **Start a notebook server** page.

Additional resources

- Toleration v1 core
- Understanding taints and tolerations
- Managing resources from custom resource definitions

4.4. DELETING AN ACCELERATOR PROFILE

To discard accelerator profiles that you no longer require, you can delete them so that they do not appear on the dashboard.

Prerequisites

- You have logged in to Red Hat OpenShift AI.
- You are assigned the **cluster-admin** role in OpenShift.
- The accelerator profile that you want to delete exists in your deployment.

Procedure

1. From the OpenShift AI dashboard, click **Settings** → **Accelerator profiles**.
   The **Accelerator profiles** page appears, displaying existing accelerator profiles.
2. Click the action menu (⋮) beside the accelerator profile that you want to delete and click **Delete**.
   The **Delete accelerator profile** dialog opens.
3. Enter the name of the accelerator profile in the text field to confirm that you intend to delete it.
4. Click **Delete**.

Verification

- The accelerator profile no longer appears on the **Accelerator profiles** page.

Additional resources

- Toleration v1 core
- Understanding taints and tolerations
- Managing resources from custom resource definitions

4.5. CONFIGURING A RECOMMENDED ACCELERATOR FOR NOTEBOOK IMAGES

To help you indicate the most suitable accelerators to your data scientists, you can configure a recommended tag to appear on the dashboard.

Prerequisites

- You have logged in to OpenShift.
- You have the **cluster-admin** role in OpenShift.
- You have existing notebook images in your deployment.

Procedure

1. From the OpenShift AI dashboard, click **Settings → Notebook images**.
   The **Notebook images** page appears. Previously imported notebook images are displayed.

2. Click the action menu (⋮) and select **Edit** from the list.
   The **Update notebook image** dialog opens.

3. From the **Accelerator identifier** list, select an identifier to set its accelerator as recommended with the notebook image. If the notebook image contains only one accelerator identifier, the identifier name displays by default.

4. Click **Update**.

   NOTE

   If you have already configured an accelerator identifier for a notebook image, you can specify a recommended accelerator for the notebook image by creating an associated accelerator profile. To do this, click **Create profile** on the row containing the notebook image and complete the relevant fields. If the notebook image does not contain an accelerator identifier, you must manually configure one before creating an associated accelerator profile.

Verification

- When your data scientists select an accelerator with a specific notebook image, a tag appears next to the corresponding accelerator indicating its compatibility.

4.6. CONFIGURING A RECOMMENDED ACCELERATOR FOR SERVING RUNTIMES

To help you indicate the most suitable accelerators to your data scientists, you can configure a recommended accelerator tag for your serving runtimes.

Prerequisites

- You have logged in to Red Hat OpenShift AI.

- If you use specialized OpenShift AI groups, you are part of the admin group (for example, **{oai-admin-group}**) in OpenShift.

Procedure

1. From the OpenShift AI dashboard, click **Settings** > **Serving runtimes**.
   The **Serving runtimes** page opens and shows the model-serving runtimes that are already installed and enabled in your OpenShift AI deployment. By default, the OpenVINO Model Server runtime is pre-installed and enabled in OpenShift AI.

2. Edit your custom runtime that you want to add the recommended accelerator tag to, click the action menu (⋮) and select **Edit**.
   A page with an embedded YAML editor opens.

NOTE

You cannot directly edit the OpenVINO Model Server runtime that is included in OpenShift AI by default. However, you can *clone* this runtime and edit the cloned version. You can then add the edited clone as a new, custom runtime. To do this, click the action menu beside the OpenVINO Model Server and select **Duplicate**.

3. In the editor, enter the YAML code to apply the annotation **opendatahub.io/recommended-accelerators**. The excerpt in this example shows the annotation to set a recommended tag for an NVIDIA GPU accelerator:

```
metadata:
 annotations:
  opendatahub.io/recommended-accelerators: '["nvidia.com/gpu"]'
```

4. Click **Update**.

Verification

- When your data scientists select an accelerator with a specific serving runtime, a tag appears next to the corresponding accelerator indicating its compatibility.