Deployment Guide

Deploying Intel® Enterprise AI with Red Hat® OpenShift® AI Made Simple

Created July 8, 2024

Authors (alphabetically by last name): Veenadhari Bedida, Julie Fleischer, Sridhar Kayathi, Chaitanya Kulkarni, Raghu Moorthy, Hersh Pathak, Sharvil Shah, Filip Skirtun, Martin Xu

Background and Prerequisites

This document links to the resources needed to deploy Intel enterprise AI with Red Hat® OpenShift® AI. A cluster containing servers with Intel® Xeon® processors and, optionally, Intel® Gaudi® AI accelerators or a node or cluster containing servers with Intel® Flex GPUs, is a prerequisite. The steps should be completed in the order they are presented in this guide.

This guide was created for the following software versions:

- Red Hat OpenShift AI 2.10 / Red Hat OpenShift 4.14
- Intel Gaudi software and Operators 1.16

Install OpenShift

OpenShift can be installed either in a single node configuration or multi node configuration. The steps in this guide assume you have followed one of the two options below to install and run Red Hat OpenShift on a node or cluster.

SNO configuration is recommended when installing OpenShift on a single server where high availability is not needed (e.g., a single 8x Gaudi server, single Xeon server, or single Xeon server with one or more Intel Flex GPU add-in cards.

Install Single Node OpenShift (SNO)

Installing OpenShift in a cluster is recommended for multi-node configurations with more than one worker node, such as rack and multi-rack configurations.

Install OpenShift in a Cluster

Section 1: AI accelerator provisioning

Intel AI accelerators, including Intel Data Center GPUs and Intel Gaudi accelerators, are provisioned and managed using OpenShift Operators. The following procedures install and configure the Operators used to provision Intel AI accelerators on an OpenShift platform.

Note: These steps are not required for the Intel Advanced Matrix Extensions (Intel AMX) accelerators built into Intel Xeon processors (where available), as this support is already available in the Red Hat OpenShift installation described in the prerequisites.

Install and configure the Node Feature Discovery (NFD) Operator

The NFD Operator manages the NFD add-on deployment and lifecycle, by managing the detection of features and hardware configurations such as PCI cards, kernel, and OS version. NFD automatically detects and labels Intel AI accelerators including Intel Data Center GPUs and Intel Gaudi processors.

More Information: Node Feature Discovery Operator

Intel Data Center GPU

Install and Configure NFD Operator for Intel Data Center GPU

Verify NFD Operator for Intel Data Center GPU

Intel Gaudi Accelerator

Install and Configure NFD Operator for Intel Gaudi Accelerator

Use Machine Config Operator (MCO) to configure Kernel firmware search path for Intel Gaudi card

Note: This operation will reboot the node.

Provision Intel AI accelerators Out-of-Tree (OOT) kernel drivers with the Kernel Module Management (KMM) Operator

The KMM Operator is used to install and manage the OOT kernel drivers used to provision Intel Data Center GPUs and Intel Gaudi Accelerators.

More Information: Kernel Module Management Operator

Intel Data Center GPU

Install and Configure KMM Operator for Intel Data Center GPU

Verify KMM Operator for Intel Data Center GPU

Intel Gaudi Accelerator

Install and Configure KMM Operator for Intel Gaudi Accelerator

Install hardware-specific Operators

Intel Data Center GPU

Intel Device Plugins advertise Intel hardware features (resources) to OpenShift clusters, so that workloads running on the clusters can utilize those resources.

More Information: Intel Device Plugins Operator

Install Intel Device Plugins Operator

Configure and Verify Intel Device Plugins Operator for Intel Data Center GPU

Intel Gaudi Accelerator

The Habana AI Operator works with the KMM Operator to automate management of Habana AI software used to provision AI accelerators within the OpenShift cluster.

More Information: Habana AI Operator

Deploy Habana AI Operator

Install Red Hat OpenShift AI Operator

Red Hat OpenShift AI is a flexible, scalable artificial intelligence (AI) and machine learning (ML) platform that enables enterprises to create and deliver AI-enabled applications at scale across hybrid cloud environments.

Red Hat OpenShift AI is deployed as a regular Operator, installed on top of OpenShift.

Install Red Hat OpenShift AI Operator

Install accelerator profiles

Accelerator Profiles are required for using any Accelerator on OpenShift AI.

Intel Data Center GPU

Install Accelerator Profile: Intel Data Center GPU

Intel Gaudi Accelerator

Install Accelerator Profile: Intel Gaudi Accelerator

Install storage support

The following are additional storage options for an enterprise AI node or cluster.

Install OpenShift Data Foundation Support

Install Dell EMC Storage Support

Install Portworx Storage Support

Install Weka.io Storage Support

Section 2: Enterprise AI software installation and configuration

After the accelerator hardware has been provisioned, the enterprise AI stack(s) to provide the AI platform for creating, tuning and deploying AI applications can be installed.

Note: In the instructions below, the steps for enabling Intel Xeon processors or Intel AMX accelerators are the same as for enabling Intel Flex Series GPUs.

Install enterprise software

Intel Data Center CPU and GPU software

The AI Tools provide AI functionality, such as Pytorch, TensorFlow, and classical machine learning for data center CPUs and GPUs.

Jump to the "Install Intel AI Tools" Section of this Guide

Intel Gaudi Accelerator software

Intel Gaudi software provides AI functionality such as PyTorch, TensorFlow, and other GenAI apps for Intel Gaudi accelerators.

Jump to the "Install Intel Gaudi Software" Section of this Guide

Install distributed workload software

KubeRay and CodeFlare are used to provide distributed training and tuning for Intel Data Center CPUs and Intel Gaudi accelerators. At this time, KubeRay and CodeFlare are not supported with Intel Data Center GPUs.

The OpenShift AI UI is used to initialize the KubeRay and CodeFlare Operators. This process is the same regardless of accelerator type, but the `Ray-IPEX` image steps will need to be substituted with the corresponding Gaudi-specific steps.

Install KubeRay/CodeFlare

Install AI inference software

Intel Data Center CPU and GPU AI Inference

OpenVINO provides inference capabilities for Intel CPUs and Intel GPUs.

Jump to the "Install OpenVINO" Section of this Guide

Intel Gaudi Accelerator Inference

OpenVINO does not provide inference capabilities for Gaudi Accelerators. The section below covers the various model serving and inference options for Gaudi Accelerators.

Jump to the "Install Gaudi AI Inference Software" Section of this Guide

Note: OpenVINO Model Server (OVMS) does not support Intel Gaudi accelerators.

Install Intel AI Tools

Systems with Intel Data Center CPUs and GPUs require installation of Intel AI Tools, a oneAPI toolkit that helps speed up time to market and accelerate end-to-end machine learning and data science pipelines.

1. From **OperatorHub** on the web console, install the **oneAPI Analytics Toolkit Operator**.
2. After the operator is properly installed, create the **AIKitContainer** Custom Resource.
3. Use the following steps to launch the Intel-optimized PyTorch notebook image:
   a. Go to the OpenShift AI dashboard.
   b. Launch Jupyter application from the Enabled link in the left pane.
   c. Select the **Intel Optimized PyTorch**, **Intel Optimized TensorFlow**, or **Intel ML** radio button.
4. Select the container size from the **Container Size** dropdown menu.
5. Select the **Intel Data Center GPU Flex Series 140** accelerator if you prefer to use GPUs.
6. Click on the **Start Server** button to launch the Jupyter hub dashboard from the **Intel Optimized PyTorch** or **Intel Optimized TensorFlow** container.
7. Select the kernel from the top right bar to choose between CPU or GPU.

Install Intel Gaudi software

The Habana AI Notebook Image has already been integrated into Red Hat OpenShift AI version 2.10.

Start the Habana AI Notebook Image:

1. Search for `Networking -> Routes -> rhods-dashboard` on the web console and click on link.
2. Launch Jupyter application from the **Enabled** link in the left pane.
3. Click on the **Habana AI** image.
4. Choose the container size from the **Container Size** dropdown menu.
5. Choose the Gaudi accelerator.
6. Click on **Start server** and then **Open notebook** tab when ready.
7. Run the following command on the Gaudi-supported nodes to check for Gaudi resources:

```
oc describe <Gaudi_node_name> | grep habana.ai/gaudi
     habana.ai/gaudi:     8
     habana.ai/gaudi:     8
```

Install OpenVINO™ for AI inferencing with Intel Data Center GPUs

OpenVINO is a deep learning toolkit that enables developers to create a model once and deploy it across Intel hardware. It is available within OpenShift AI and can be installed from the OpenShift AI Operator Hub.

Install OpenVINO Operator

To install the OpenVINO Operator from within OpenShift AI, search for **OpenVINO Operator** from OperatorHub and click **Install**.

Verify OpenVINO installation

Verify the OpenVINO installation by running an OpenVINO notebook.

Collection of Ready-to-Run Jupyter Notebooks

Use OpenVINO Model Server (OVMS)

OpenVINO can be used as a server in deployment mode, referred to as OVMS, providing a high-performance system for serving models.

Perform Inference Using OVMS

Verify OVMS

OpenVINO Model Server Demos

Optimizing Large Language Models with OpenVINO

Install Gaudi AI inference software

Model servers make trained AI models available for inference in production environments. When a model is "served," it is available as a service, which can be accessed by an API.

About Model Serving in OpenShift AI

A model-serving runtime provides support for a given set of model frameworks and those frameworks' model formats. The box below provides links to pre-installed runtimes in OpenShift AI as well as details for how to create a custom runtime.

Creating Custom Runtimes in OpenShift AI

As one example, a custom TGI runtime can be created using the steps below.

Creating a Custom TGI Runtime

Section 3: Optional software and hardware installation

This section identifies additional software that can be installed on an Intel enterprise AI system with OpenShift AI. It includes references to software and packages that are in development and not yet available, denoted by white boxes which do not link to additional content. As those resources become available, the corresponding boxes and links will turn blue and become clickable.

Open Platform for Enterprise AI (OPEA) Validated Patterns for OpenShift

The following are popular OPEA Validated Patterns that are being ported to OpenShift AI.

- OPEA ChatQnA Validated Pattern
- OPEA Code Gen Validated Pattern
- OPEA SearchQnA Validated Pattern

Software and Tools to develop OPEA RAG applications

OPEA provides popular RAG software and tools that can be installed to develop RAG applications in OpenShift with Intel accelerators (e.g., Intel Gaudi accelerators, Intel AMX, Intel Flex GPU). The following link provides access to multiple tools utilized in GenAI examples. Specific tools needed will be dictated by the applications under development. All these tools can be installed in the Jupyter notebook environment within OpenShift AI.

- Install OPEA GenAI Tools

Edge AI software

The following popular software packages are useful when deploying AI applications at the Edge (typically devices based on Intel Xeon D, Intel Core and Intel Atom processors). These typically run on RHEL for Device Edge and leverage OpenVINO for inferencing.

- Guise AI
- Intel Tiber Edge Insights

Additional software

The following popular software options can be included in an OpenShift Cluster with Intel accelerators.

- IBM watsonx.ai
- Seekr.io
- Cloudera
- SAS Viya

Section 4: Helpful links

The following section will eventually provide helpful additional information for anyone interested in getting up and running using Intel enterprise AI with OpenShift AI.

Solution provided by:

0724/RM/MESH/356886-001US