Solutions Reference Architecture
December 2023

Effortlessly Accelerate Your AI Applications and Workflows on Red Hat OpenShift Container Platform with 4th Gen Intel® Xeon® Scalable Processors

Intel's extensions of Red Hat OpenShift validated patterns take advantage of Intel® AMX to enhance AI training and inference workloads

Contents

Solution Benefits

- Easily deployable, fully functional, end-to-end workflows on the Red Hat® OpenShift® Container Platform can be created with no concern for configuration details.
- No-hassle enablement of the built-in AI accelerator on 4th Gen Intel® Xeon® Scalable processors.
- Fast inference and training for a variety of AI workloads at the edge.

Executive Summary

AI is being woven into nearly every field of endeavor. But sometimes, it can be overwhelming to know where to start to enable the full potential of AI. Red Hat and Intel are working together to simplify the deployment of hybrid cloud AI workloads.

Red Hat has developed validated patterns—all the code and Red Hat® OpenShift® Container Platform elements you need to deploy specific use cases—including recommender engines in retail and AI-powered image analytics for medical diagnosis. Validated patterns are continuously tested, and new ones are regularly added.

Intel has worked with Red Hat to show how easy it is to extend validated patterns to take advantage of specific hardware accelerators built into 4th Generation Intel® Xeon® Scalable processors. One such accelerator is Intel® Advanced Matrix Extensions (Intel® AMX), which is purpose-built to accelerate AI inference and training. With just a few extra lines of code and the use of Red Hat's Node Feature Discovery (NFD) operator, Intel extended the Multicloud GitOps validated pattern to take advantage of Intel AMX—potentially accelerating inference by as much as 10x compared to previous-generation Intel Xeon Scalable processors.[1]

In another example of validated pattern extension, Intel quantized a medical diagnosis machine-learning model to use a lower precision to accelerate image inference at the edge. Quantization is a technique used in AI to reduce the computational and memory costs of running inference.[2] Using the NFD operator to identify nodes equipped with Intel AMX and an open-source scaling tool, the extended validated pattern unleashed the power of AI to assess chest X-rays for the risk of pneumonia.

These extended validated patterns are available on GitHub to use as-is or as a foundation for further extensions. Red Hat's robust and flexible OpenShift platform, combined with continued innovation from Intel and the developer ecosystem, demystifies deploying and accelerating AI workloads.

Red Hat
OpenShift

Solution Brief

Business Challenge

Many industries, including retail and healthcare, are increasingly using AI in their workloads. Experts predict that 70% of businesses will use AI by 2030.[3] From personalized product recommendations to AI-powered disease diagnosis, the opportunities for AI to transform businesses and help save lives are enormous. Another industry trend is the move to cloud-native applications running on platforms like the latest Red Hat® OpenShift® Container Platform with Kubernetes. However, developers face several challenges when using AI and hybrid cloud deployments.

- Kubernetes environments are complex to set up from scratch, with many platform components and considerations.
- Code optimizations, like AI acceleration, are scattered across various sources and repositories, making them difficult to find and validate.

What if there was a way to quickly deploy a workload on the Red Hat OpenShift Container Platform, with minimal configuration and management overhead—plus easy access to optimizations that enable fast AI training and inference in the data center or at the edge?

Solution Value

The latest Red Hat OpenShift Container Platform provides DevOps teams and IT organizations with a hybrid cloud application platform for deploying new and existing edge applications on secure, scalable resources. But while the Red Hat OpenShift Container Platform is a powerful tool, it can be difficult to determine all the right components for a particular use case and workload.

Red Hat OpenShift validated patterns let you quickly create fully functional, end-to-end workflows on Red Hat OpenShift without worrying about configuration details and workflow management. Validated patterns are designed to connect multiple clouds and clusters, including edge clusters, can help you easily develop a solution that fulfills requirements and drives business success. Think of a validated pattern as a trusted recipe; you don't have to determine what ingredients are necessary or wonder if the recipe actually results in something edible—it just works.

Intel and Red Hat have collaborated to optimize Red Hat OpenShift and some of its validated patterns for 4th Generation Intel® Xeon® Scalable processors. Red Hat engineers maintain validated patterns to ensure they meet performance, scalability and reliability requirements and include all the necessary Red Hat OpenShift Container Platform components, as well as management software. They also automatically handle system configuration, security, networking, storage and application deployment.

Validated patterns can be extended to enable the Red Hat OpenShift Container Platform to use additional hardware features by using Red Hat's Node Feature Discovery (NFD) operator. For example, Intel® Advanced Matrix Extensions (Intel® AMX) is a built-in AI accelerator on 4th Gen Intel Xeon Scalable processors that can provide up to 10x faster training and inference compared to the previous generation of Intel Xeon Scalable processors.[4] Image and video recognition, scientific computing, financial modeling, healthcare, natural language processing and many more use cases can benefit from using Intel AMX to quickly uncover insights, increase customer satisfaction and potentially lower compute costs.

This reference architecture illustrates how easy it is to enable Intel AMX and gain its performance advantages.

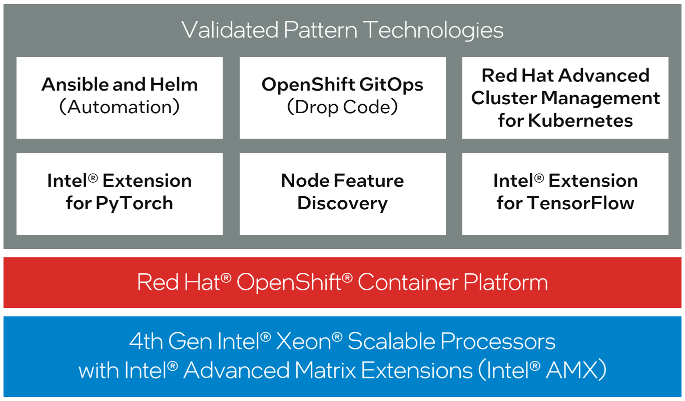
10x FASTER

training and inference

compared to the previous generation of Intel® Xeon® Scalable processors[1]

Solution Architecture Highlights

Red Hat's validated patterns contain all the code needed to help build your technology stack so that you can bring solutions to life more quickly. All the steps are fully automated through GitOps processes to automate deployments consistently and at scale. Moreover, unlike static reference architectures, Red Hat validated patterns are continuously updated and refined based on customer feedback, industry trends and technological advancements. As an example of how the IT community can contribute to continuous innovation, with just a few lines of code, Intel has customized two Red Hat validated patterns to enable the use of Intel AMX (see Figure 1).



Validated Pattern Technologies

Ansible and Helm (Automation)

OpenShift GitOps (Drop Code)

Red Hat Advanced Cluster Management for Kubernetes

Intel® Extension for PyTorch

Node Feature Discovery

Intel® Extension for TensorFlow

Red Hat® OpenShift® Container Platform

4th Gen Intel® Xeon® Scalable Processors with Intel® Advanced Matrix Extensions (Intel® AMX)

Figure 1. Extending Red Hat® validated patterns to include Intel® AMX can accelerate AI workloads.

A Closer Look at Intel's Extension of Red Hat Validated Patterns

The following sections briefly describe some of the most relevant technologies from Red Hat and Intel that power the validated pattern use cases illustrated in this reference architecture.

Red Hat OpenShift

Red Hat OpenShift offers a comprehensive platform for efficiently building, updating and scaling applications. You can enhance productivity and speed up your application deployment process by utilizing a full suite of services, all adaptable to your preferred infrastructure.

Red Hat OpenShift Operators

Red Hat OpenShift Container Platform relies heavily on operators, which are a group of software extensions and tools used to automate, manage and simplify the deployment and operation of complex applications and services. Software suppliers and developers often create operators with in-depth knowledge of a specific application or service.

The NFD operator controls the detection of hardware characteristics and configuration in a Red Hat OpenShift Container Platform cluster and labels the nodes with hardware-specific information. By marking nodes, it indicates the presence of a technology, accelerator or module on the host. By determining which nodes have a required hardware feature, DevOps teams can ensure that workflows requiring a specific feature (such as Intel AMX) land on an appropriate node.

Knative Serving

To declare and control how serverless applications behave within the cluster, Knative Serving creates a set of Kubernetes Custom Resource Definitions (CRDs). Its purpose is to automatically scale the number of containers up or down and assign them to nodes with required labels.

Argo CD

Argo CD is a declarative GitOps continuous delivery tool for Kubernetes that uses Git as the sole source of truth. It manages application definitions, configurations and environment versions in a declarative and automated manner while controlling component versions. Argo CD makes deploying and managing applications easier, faster and less problematic.

Argo CD is used in both of the extended validated patterns described in this reference architecture.

Intel AMX

Intel AMX is a built-in accelerator that enables 4th Gen Intel Xeon Scalable processors to optimize deep-learning (DL) training and inferencing workloads. With Intel AMX, 4th Gen Intel Xeon Scalable processors can quickly pivot between optimizing general computing and AI workloads. Imagine an automobile that could excel at city driving and quickly change to deliver Formula 1 racing performance. 4th Gen Intel Xeon Scalable processors deliver this type of flexibility. Developers can code AI functionality to take advantage of the Intel AMX instruction set, and they can code non-AI functionality to use the processor instruction set architecture. Intel has integrated the Intel® oneAPI Deep Neural Network Library (oneDNN) into popular open-source tools for AI applications, including TensorFlow, PyTorch, PaddlePaddle and ONNX.

Use Cases

In this document, we describe how Intel extended two validated patterns–Multicloud GitOps and Medical Diagnosis–to take advantage of Intel AMX performance gains.

Important note: The Linux kernel detects Intel AMX at run-time, so it is unnecessary to enable and configure it separately. For both patterns, the NFD operator is deployed to allow easy detection and consumption of Intel features and accelerators such as Intel AMX.

After pattern deployment, you can validate Intel AMX availability on the nodes. Log in to the Red Hat OpenShift Container Platform cluster and run the following command:

```
$ oc get nodes --show-labels
```

If the NFD operator was deployed successfully, nodes equipped with Intel AMX are labeled as shown below:

```
feature.node.kubernetes.io/cpu-cpuid.AMXBF16=true
feature.node.kubernetes.io/cpu-cpuid.AMXINT8=true
feature.node.kubernetes.io/cpu-cpuid.AMXTILE=true
```