Solutions Reference Architecture

Data Center | Artificial Intelligence

Boosting AI Performance with Red Hat OpenShift 4.12 on 4th Gen Intel® Xeon® Scalable Processors

Easily deploy and run all your data pipeline workloads on a validated open infrastructure featuring accelerators and optimized libraries and frameworks

Red Hat OpenShift Container Platform

Contents

Authors

Cloud & Enterprise Solution Group

Pawel Adamczyk
Cloud Solutions Engineer

Karol Brejna
Cloud Solutions Architect

Krzysztof Cieplucha
Cloud Solutions Architect

Izabela Irzynska
Cloud Solutions Engineer

Kamil Lipka
Cloud Solutions Engineer

Igor Marzynski
Cloud Solutions Engineer

Paulina Olszewska
Cloud Solutions Engineer

Malgorzata Rembas
Cloud Solutions Architect

Lukasz Sitkiewicz
Cloud Solutions Engineer

Filip Skirtun
Cloud Solutions Engineer

Lokendra Uppuluri
Cloud Software Architect

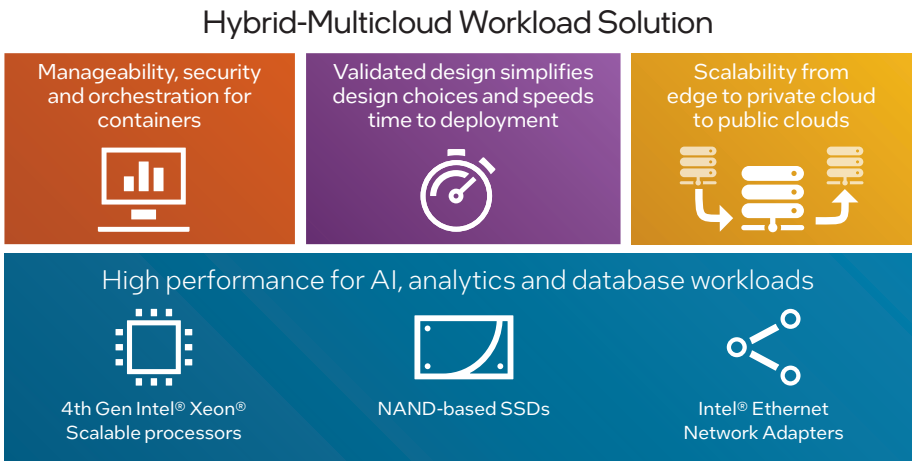
Solution Overview and Summary

Currently, hybrid-cloud-capable, cloud-native infrastructure is a major part of data center deployments, whether you are talking about databases, artificial intelligence (AI) or telecommunications workloads. Today's cloud-native applications use a distributed cloud approach. Some workloads or portions of the workload run in private clouds while others run in one or another public cloud.

Intel and Red Hat are working together to combine software and hardware components into a customized infrastructure tuned to the needs of achieving transformation and sustainability goals. The Intel® Solution for Red Hat® OpenShift® Container Platform delivers a cloud-native architecture that enables DevOps and IT to accelerate application deployment and easily scale across any cloud environment from on-premises, to hybrid, to the public cloud and edge.

You can run a variety of applications on your flexible OpenShift environment. AI is a growing workload across a wide spectrum of use cases, from information security and infrastructure management to business applications and automation. Specifically, this reference architecture describes the benefits of the 4th Generation Intel® Xeon® processor Scalable family and how to use the new Intel® Advanced Matrix Extensions (Intel® AMX), which is an AI accelerator, with Red Hat OpenShift 4.12. The result is an open, interoperable infrastructure that flexibly handles growing volumes of data and AI applications with ease to innovate faster for a competitive advantage.

The audience for this reference architecture includes enterprise infrastructure companies, network operators, communications service providers and cloud-first independent software vendors that offer their solutions through cloud service providers.



Solution Brief

Business Challenge

The amount of data that enterprises must store and analyze has been rising steadily for years, putting pressure on enterprises to modernize IT infrastructure that can handle growing workloads. To handle their burgeoning data, enterprises are focusing on distributed computing spread across hybrid and multicloud environments. As a result, enterprise IT needs more compute power to handle the proliferation of applications and data. Disparate or legacy technology can create integration challenges and impede progress.

For successful digital transformation, enterprises must invest in an infrastructure that can provide the foundation to meet these new demands and provide IT and developers with the ability to design and move applications consistently across different environments from the data center to the cloud and to the edge. In short, enterprises seek a cohesive collection of technologies that can propel their business into the digital future.

Solution Value

This solution integrates enterprise-level validated services and components with new performance-enhancing features of 4th Gen Intel® Xeon® Scalable processors.

- **Performance.** High throughput for AI workloads with Intel® Advanced Matrix Extensions (Intel® AMX) using the Intel® Optimization for TensorFlow and Intel® Distribution of OpenVINO™ toolkit.
- **Time to market.** Validated solution using containerized tools and sample workloads provides ease of deployment and helps reduce time to market.

Optimize Efficiency for AI Processing with Intel® Advanced Matrix Extensions

We introduced Intel AMX capabilities that provide massive speedup to the tensor processing that is at the heart of deep-learning algorithms. With Intel AMX, we can perform 2,048 INT8 operations per cycle (versus 256 without Intel AMX) and 1,024 BFLOAT16 operations per cycle (versus 64 without Intel AMX).[1]

4th Gen Intel Xeon Scalable processors can be fine-tuned to peak efficiency by using the Intel® oneAPI Deep Neural Network Library (oneDNN), which is part of the oneAPI toolkit and integrated into TensorFlow and PyTorch AI frameworks and with the Intel Distribution of OpenVINO toolkit. You can also use the oneAPI toolkit to write instructions that remove the administrative burden of manually assigning the right accelerator to an AI or non-AI workload. This enables the 4th Gen Intel Xeon Scalable processor to run all your data pipeline workloads and automatically select the accelerator for peak performance and optimal resource utilization.

Solution Benefits

- Accelerate time to deployment with a validated hardware and software recommendation based on popular enterprise use cases.
- Optimize efficiency for AI processing with Intel® Advanced Matrix Extensions (Intel® AMX).
- Speed innovation and ease application development using readily available containers of Intel® architecture-optimized AI libraries and models, such as Intel® Optimization for TensorFlow and Intel® Distribution of OpenVINO™ toolkit.

Solution Architecture

The Red Hat OpenShift Container Platform provides a consistent and security-enabled Kubernetes cloud-native, hybrid-multicloud experience (see Figure 1). It accommodates a large, scalable mix of microservices-oriented applications and their dependent components. The Red Hat OpenShift Container Platform uses the Container Runtime Interface–Open Container Initiative engine and Kubernetes-based orchestration. It provides container-as-a-service (CaaS) and platform-as-a-service (PaaS) workflows for developers and existing applications.

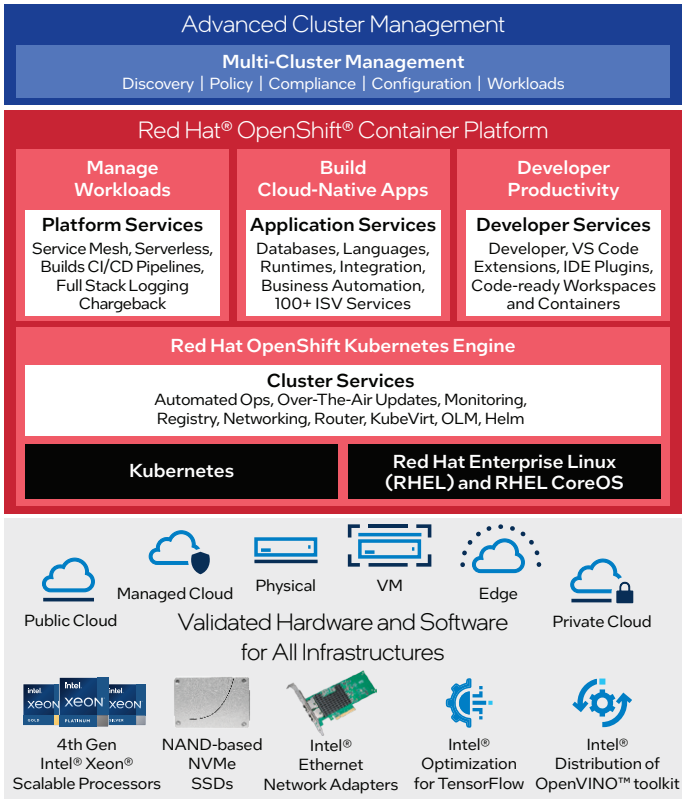


Figure 1. Red Hat® OpenShift® Container Platform helps enterprises develop, deploy and manage innovative applications at scale.

Configuration Summary

Introduction

The following hardware and software were used for performance evaluation of Intel AMX acceleration with 4th Gen Intel Xeon Scalable processors.

Table 1. Hardware Bill of Materials

Component	Description	Quantity per Node
3x Control Plane Nodes		
Processor	Intel® Xeon® Gold 6330 processor (28 cores, 2.0 GHz)	2
Memory	256 GB	16x 16 GB DDR4 3200 MT/s
Boot Drive	Solidigm D3-S4510 240 GB	2
Network	Intel® Ethernet Controller E810-C for QSFP	1
3x-6x Compute/Worker Nodes		
Processor	Intel Xeon Gold 6438Y+ processor (32 cores, 2.0 GHz)	2
Memory	512 GB	16x 32 GB DDR5 4800 MT/s
Boot Drive	Solidigm D3-S4510 240 GB	2
Storage Drive	Solidigm D7-P5510 3.84 TB	2
Network	Intel Ethernet Network Adapter E810-CQDA2 for OCP 3.0	1

Table 2. Software Versions

Software	Description
Red Hat® OpenShift® Container Platform	4.12
Red Hat Enterprise Linux CoreOS	4.12
Intel® Optimization for TensorFlow	2.11
Intel® Distribution of OpenVINO™ toolkit	2022.3.0

Operators and Red Hat OpenShift Container Platform

Developers and Kubernetes administrators can use the Red Hat Marketplace to gain automation advantages while enabling the portability of the services across Kubernetes environments. Developers can choose operators for a wide variety of tasks, including AI and machine learning, databases, integration and delivery, logging and tracing, monitoring, networking, security, storage, and streaming and messaging.

Once installed on a cluster, operators are listed in the Red Hat OpenShift Container Platform Developer Catalog, which provides a self-service experience. Developers don't need to be an expert in applications deployment such as Ceph Object Storage, Kubeflow, Jupyterhub, Apache Spark, Seldon, Prometheus, Grafana, Argo, TensorFlow or Scikit-learn—they just install the operators they need to accomplish their application goals. The result is that teams can spend more time solving critical business needs and less on installing and maintaining infrastructure.

Implementation Guide

AI Workload Selection and Preparation

Two different AI models were selected to showcase AI inferencing acceleration with Intel AMX:

- DIEN is commonly used in recommendation systems.
- BERT-Large is used in natural language processing (NLP) systems such as chat bots.

We chose these models because of their relevance to popular use cases in recommendation systems and NLP, respectively. Both models, with the corresponding container images optimized for various precisions, are made publicly available by Intel and the corresponding datasets on which the models are trained on are publicly available as well.

- DIEN was run using the Intel Optimization for TensorFlow runtime.
- BERT-Large was run using the Intel Distribution of OpenVINO toolkit runtime.

To run inferencing with the DIEN model, you must obtain the DIEN container image from the Intel Docker Hub and download the Amazon book reviews dataset. To obtain the dataset, it is necessary to install the wget and bzip2 packages inside the running container. Detailed steps are described in the README document located in the IntelAI GitHub repository.

For the BERT-Large model, you must obtain the Intel Distribution of OpenVINO runtime container from the Red Hat Catalog. Download the BERT-Large model compatible with FP32&BF16 precision. For inference at INT8 precision, a different model must be obtained separately from the same source. Both BERT-Large models can be downloaded through the Model Downloader tool. The tool can be installed as specified in the instructions found in OpenVINO documentation. The final step before running the inference is building the Benchmark C++ tool. Instructions to build and operate the tool can be found in the "Benchmark C++ Tool" article.

By default, Red Hat OpenShift prevents containers from running as an arbitrary user in the system. Running the pods with our workload successfully requires applying "anyuid" Security Context Constraint to the default ServiceAccount in the namespace. Detailed procedure can be found in Red Hat OpenShift documentation.

The benchmark parameters (batch size, precision) can be modified as detailed in the "Benchmark C++ Tool" article mentioned in the previous paragraph.

Results and Use Cases

The Intel® Solution for Red Hat® OpenShift® Container Platform can be used to accelerate performance and simplify deployment for a wide variety of AI workloads, such as object detection, image classification and NLP. The solution is also suitable for non-AI workloads like databases and online transactional processing (OLTP). The following sections provide a sampling of the performance benefits for NLP and recommendation system workloads on 4th Gen Intel Xeon Scalable processors. We expect the Intel Solution for Red Hat OpenShift Container Platform will provide similar results.

Natural Language Processing: Smoother Experiences with Faster Responses

Higher-performance NLP inference can help enable more responsive smart assistants, chatbots, predictive text, language translation and more. 4th Gen Intel Xeon Scalable processors feature Intel AMX, a built-in AI accelerator that speeds up deep-learning inference on the CPU, without the cost and complexity of a discrete accelerator. Intel AMX achieves:

- Up to 5.7x end-to-end real-time inference performance speedup with Intel AMX (BF16) compared with the prior generation (FP32) on Document Level Sentiment Analysis (DLSA) with Hugging Face (IMDB).[2]
- Up to 6.2x higher real-time NLP inference performance (BERT) with Intel AMX (BF16) versus the prior generation (FP32).[3]

Up To
5.7x Higher
End-to-End Real-Time Inference Performance Speedup[2]

Up To
6.2x Higher
Real-Time NLP Inference Performance[3]

Intel AMX also excels in transfer learning and retraining, so you can keep your models current without the need for additional hardware.

Recommendation Systems: Recommendations in Real Time

Deliver fast, personalized product or content recommendations that doesn't impede the user experience by using a deep-learning-based recommendation system that accounts for real-time user behavior signals and context features, such as time and location. 4th Gen Intel Xeon Scalable processors with Intel AMX helps speed up deep-learning inference and accelerate small model training on the CPU:

- Up to 6.3x higher batch recommendation system inference performance (DLRM) with Intel AMX (BF16) versus the prior generation with FP32.[4]
- Up to 4x higher recommendation system training performance (DLRM) with Intel AMX (BF16) versus the prior generation with FP32.[4]

Up To
6.3x Higher
Batch Recommendation System Inference Performance[4]

Up To
4x Higher
Recommendation System Training Performance[4]

Conclusion

This solutions reference architecture highlights the benefits of 4th Gen Intel Xeon Scalable processors with Intel AMX for AI workload acceleration. It also shows the gen-over-gen performance improvement for AI workloads on Intel® processors. Using the Intel Solution for Red Hat OpenShift Container Platform can improve the time to market for AI solutions using the 4th Gen Intel Xeon Scalable processors due to the combination of the following:

- Intel AMX works out of the box without any Red Hat OpenShift configuration changes.
- Many pretrained AI models and container images compatible with 4th Gen Intel Xeon Scalable processors and Intel AMX (with support for different precision levels) are made publicly available by Intel.
- A reference architecture consisting of recommended hardware and software, validated by benchmarking AI workloads using popular deep-learning models like DIEN and Bert-Large running on Red Hat OpenShift 4.12.

Learn More

You may also find the following resources useful:

- 4th Gen Intel Xeon Scalable processors
- "Benchmark C++ Tool" article
- IntelAI GitHub repository
- Intel Docker Hub containers for recommendation
- Intel Ethernet products
- OpenVINO Bert-Large model FP32&BF16
- OpenVINO Bert-Large model INT8
- OpenVINO documentation
- Red Hat Catalog
- Red Hat Marketplace
- Red Hat OpenShift Container Platform

[1] https://edc.intel.com/content/www/tw/zh/products/performance/benchmarks/architecture-day-2021/. Results may vary.

[2] See claim [A2] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/. 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[3] See claim [A19] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/. 4th Gen Intel® Xeon® Scalable processors. Results may vary.

[4] See claim [A21] at https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/. 4th Gen Intel® Xeon® Scalable processors. Results may vary.

Performance varies by use, configuration and other factors. Learn more at intel.com/PerformanceIndex.