September 2024

Red Hat Modernization

Modernizing & Advancing the Data Center

Partner Enablement Package

How our valued partners can build modernization solutions and advance the data center with AI, based on latest Intel® technology

intel

Red Hat

Contents

Benefits & Value Proposition

For SIs / SPs

• Connect with your customers, show them how to save costs, deliver improved performance and modernize to a cloud native offering
• Server & Software upgrades unleash new AI & Security capabilities
• Identify multiple revenue streams by articulating the value of modernizing the data center

For End Customers

• Achieve lower TCO through server consolidation, resulting in energy savings
• Reduce the number of servers -> lower SW costs
• Cloudify the data center with OpenShift
    • Single platform to manage VMs and Containers
    • Flexibility to run On-Prem and in the Cloud

Why Faster Refresh Cycles And Modern Infrastructure Management Are Critical To Business Success

AI Ecosystem Support

"The AI landscape is evolving at breakneck speed, and organizations need solutions that are both agile and powerful to keep pace. Red Hat OpenShift AI, combined with the breadth and performance of Intel's AI portfolio, offers a compelling solution for businesses to accelerate their AI journey and unlock new possibilities."

Chris Wright, chief technology officer and senior vice president, Global Engineering, Red Hat

"Cisco is actively working to advance generative AI technology to bring new and exciting opportunities to customers, such as faster innovation, enhanced decision-making and mitigated risks. We strongly support the collaboration between Red Hat and Intel as we continue to work closely with them to offer new and innovative solutions for our joint customers."

Daniel McGinnis, vice president, Product Management, Cisco

"Dell Technologies helps customers take full advantage of artificial intelligence by enabling the deployment of AI models and applications, especially generative AI, across enterprise environments. Dell collaborates with Red Hat and Intel to make it easy for customers to deploy and operate Red Hat OpenShift AI on Dell APEX Cloud Platform and Dell PowerEdge with Intel accelerators."

Greg Findlen, senior vice president, Product Management, Dell Technologies

"Together, Red Hat and Intel are committed to helping organizations develop and unlock the full potential of scalable and more secure AI systems based on open ecosystems and standards. Intel's AI technologies help minimize the complexities associated with AI workloads and unlock improved performance, empowering businesses to innovate faster and achieve their AI goals."

Justin Hotard, executive vice president and general manager, Data Center and AI Group, Intel

"Lenovo is committed to shaping a future where AI enhances every aspect of computing environments. This commitment is embodied in our collaboration with Red Hat and Intel, aiming to harness the full potential of AI through our Lenovo Edge-to-Cloud Solutions. By using Red Hat OpenShift AI, powered by the latest 5th Gen Intel Xeon Scalable Processors with Intel AMX, and soon-to-include Intel Gaudi 3 AI Accelerators, we are optimizing and speeding up AI deployments. This synergy allows us to deliver sophisticated AI solutions that operate seamlessly across the hybrid cloud, meeting the evolving needs of our diverse clientele."

Kamran Amini, vice president and general manager, Server, Storage & Software Defined Solutions, Lenovo

"At Supermicro, we envision a world where AI is seamlessly integrated into every facet of our computing solutions. To bring this vision to life, Supermicro is working closely with industry leaders like Red Hat and Intel. By leveraging Red Hat OpenShift AI and the powerful capabilities of 5th Gen Intel Xeon Scalable Processors with Intel AMX and Intel's Gaud 2 and Gaudi 3 AI Accelerators—we are enhancing and accelerating AI operations across hybrid cloud environments, ensuring our customers have access to the most advanced AI tools on Supermicro servers."

Ray Pang, vice president, Technology & Business Enablement, Supermicro

Intel and Red Hat: Leaders in Open Technology

Together, Red Hat and Intel provide agile, cloud-ready network architectures based on high-performance, industry-standard platforms and open, software-defined infrastructure.

Partner Showcase:
Intel and Red Hat - Leaders in Open Technology

25 years in **partnership**

Top contributor **to Linux kernel and open source**

Solutions **optimized through joint roadmaps**

Why Intel + Red Hat

The Latest Technology

Intel collaborates closely with Red Hat to implement the newest technologies and software with a focus on optimization. Customers gain leading architecture, workload-optimized containers, and integrated solutions.

Advanced Security

Intel® Software Guard Extensions (Intel® SGX) and Intel® Trust Domain Extensions (Intel® TDX) provides advanced data protection.
(** Powered by Intel® Device Plugins Operator for Red Hat OpenShift Container Platform)

Check out the Intel® Device Plugins Operator here-->

Advanced AI Tools

Developers at industry-leading independent software vendors (ISVs) and enterprises use Intel® tools and framework optimizations to build their AI platforms, systems, and applications. Developers can learn, test, and deploy AI tools such as Intel® AI Analytics Toolkit (AI Kit), Intel® Distribution of OpenVINO™ toolkit, Intel® Tiber™ AI Studio, and Habana, using the Red Hat OpenShift AI sandbox.

Check out the Red Hat OpenShift AI Sandbox here-->

Featured Solutions

Streamline Enterprise AI Adoption and Deployment
Combining Red- Hat OpenShift AI and Intel Enterprise AI provides a unified, consistent, validated, and integrated hardware and software solution for enterprise AI, right out of the box.

GenAI Inferencing on UCS X-Series with 5th Gen Intel® Xeon® Scalable Processors on Red Hat OpenShift AI

Intel® Xeon® Scalable processors for your data center

Your business can count on:

- Shared architectural platform for 4th Gen Intel® Xeon® processors and 5th Gen Intel® Xeon® processors

- Designed to accelerate performance across the fastest-growing workloads

- Most built-in accelerators of any CPU on the market to help maximize performance and power efficiency for emerging workloads

- Advanced security technologies to help protect data

- Scale with the most choice, no matter the deployment path

- Intel's most sustainable data center processors
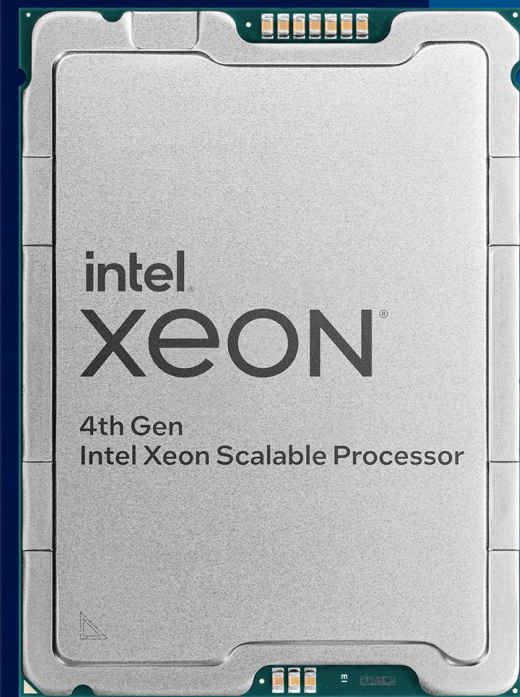
Upgrade Opportunity: CentOS Linux

Support ended on June 30th 2024 Current CentOS Linux users now need to find an alternative Linux offering to maintain support

Migration to **Red Hat Enterprise Linux (RHEL)** gives you an opportunity to gain more value for your organization

Security resources
Industry-leading security response and data

Partner ecosystem
4,900 ISVs, 5,500 IHVs

Life cycle support and flexibility
Reliable cadence, 10+ years major, 2+ year minors

Product leadership
Lead the Enterprise Linux ecosystem and roadmap

Proactive analytics
Vulnerability reporting, cost management, etc.

Support and expertise
Industry-leading team, TAMs

More than

90%

of companies in the Fortune 500 rely on Red Hat technologies[1]

How to Migrate

Watch Video

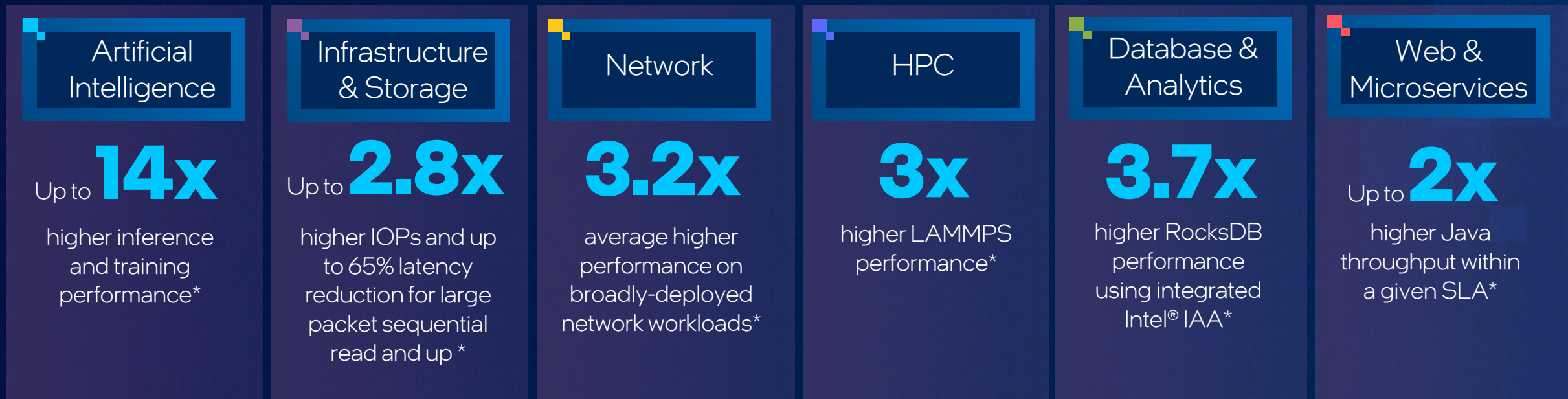Streamline your migration with an automated conversion tool

Convert2RHEL

[1]https://www.redhat.com/en/about/company

Deploy RHEL on Intel® Xeon®

Upgrade to 5th Gen Intel® Xeon® Scalable processors for significant performance gains

Artificial Intelligence
Up to **14x**

higher inference and training performance*

Infrastructure & Storage
Up to **2.8x**

higher IOPs and up to 65% latency reduction for large packet sequential read and up *

Network
3.2x

average higher performance on broadly-deployed network workloads*

HPC
3x

higher LAMMPS performance*

Database & Analytics
3.7x

higher RocksDB performance using integrated Intel® IAA*

Web & Microservices
Up to **2x**

higher Java throughput within a given SLA*

5th Gen Intel® Xeon® Scalable processors vs. 3rd Gen Intel® Xeon® processors

Migrate to RHEL on Intel® Xeon® Scalable processors to access significant performance gains (as above) + reduced power & TCO

READ MORE >

RHEL on 5th Gen Intel® Xeon®

Access the Intel® Xeon® Processor Advisor Suite to calculate the best route to lower TCO and your ROI

*See [A17, N16, N11, H10, D1, W3] at intel.com/processor claims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

Why Choose 5th Gen Intel® Xeon® processors for Server Refresh?

intel.
XEON®

Lower Total Cost of Ownership (TCO)

Intel's portfolio of hardware, software, systems, and tools can help advance your data center's overall efficiency, creating energy savings and reducing your carbon footprint, without sacrificing performance, while giving you the TCO and flexibility you need.

Up to
77%
reduction
in TCO[1]

Increase Efficiency

Intel® Accelerator Engines boost CPU utilization, reduce electricity consumption resulting in lower impact on the environment.

10x
better efficiency
(perf/watt)
with built-in
accelerators[3]

Gain Better Performance

By delivering more performance per core with built-in accelerators, 5th Gen Intel® Xeon® processors help you meet requirements for even the most demanding workloads.

84%
Performance
gain[2]

Optimize the Data Center

Intel® Xeon® processors deliver the low-latency, high-bandwidth capabilities required by modern and AI-infused workloads. Replacing aging infrastructure with these speedy and energy efficient processors will help you keep pace with rapidly evolving market needs.

Up to
16:1
server
consolidation[1]

Stay Secure with Confidential Computing

With Intel, you can choose from the most deployed confidential computing options in data centers on the market today—now including application or VM-level isolation.

300+
Organizations have engaged with Intel to develop and deploy Confidential Computing services[4]

LEARN MORE

What's the right transition for your customer?

Server Virtualization Opportunity

Global server virtualization software market size was

$7.88 billion in 2022[1]

Market is projected to reach

$12.32 billion by 2031[1]

Challenges of traditional Virtualization environments

$ $ $ Increasing cost

Slow Evolution

Supporting Growth

Risk

Developer Productivity

Benefits of Virtualized Environments

Cost reduction for operating infrastructure

Innovate at speed

Scalability

Security focused

Integrated development tools

[1] Business Research Insights

OpenShift Virtualization

Red Hat OpenShift Virtualization supports virtual machines alongside containers and serverless workloads, enabling a single platform and common set of enterprise tools to support rapidly evolving infrastructure.

VMs   Containers

OpenShift Virtualization

Red Hat OpenShift Container Platform

Red Hat Enterprise Linux CoreOS

Physical machine
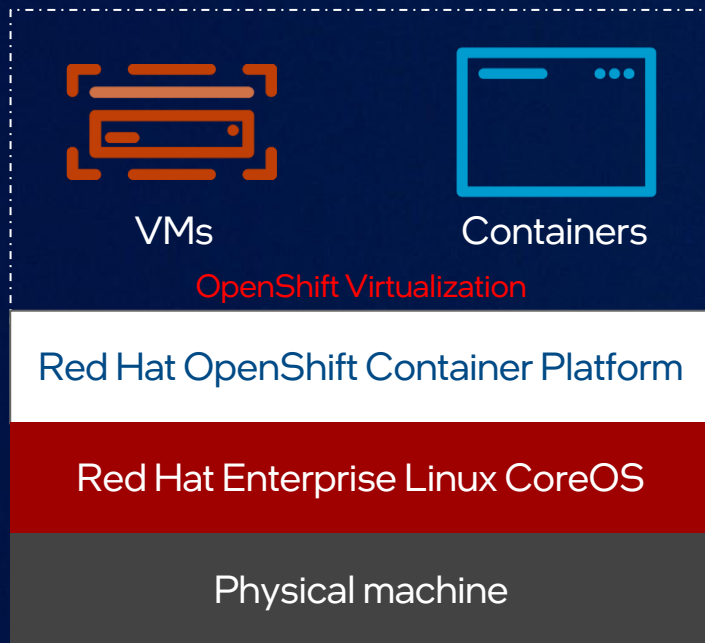
Features:

- Run VMs in OpenShift

- Performance, stability, scalability, and reliability of KVM, the Linux kernel-based hypervisor

- RHEL guest entitlements are included

- Supports Microsoft Windows guests - Microsoft Server Virtualization Validation Program (SVVP)

- Manageability and ecosystem of OpenShift

- Unified platform for running VMs and Containers

- Included feature of the OpenShift application platform

OpenShift Virtualization Business Value

Business Value:

- Trusted by FSI, healthcare, and government for enterprise software
- Hybrid cloud ready — Built for DevOps and DevSecOps practices for greater agility and flexibility

- Migration stability with similar CPU architecture
- Support for multiple Confidential Computing options for virtual machines and containers

Benefits to upgrading infrastructure

From 1st to 5th Gen Intel® Xeon® processor	Reduce servers	Reduce energy and CO2	Reduce TCO	Recover costs (months)
8168 → 8562Y+	60%	44%	41%	19
6138 → 8558	74%	39%	59%	10
5118 → 5520+	74%	55%	66%	5
4110 → 5520+	82%	63%	76%	3

*EMEA energy costs greater than US. See backup for configurations. Results may vary.

Save Power and Money on New Server Purchases

Deploy fewer 5th Gen Intel® Xeon® processor-based servers to meet performance and TCO goals

Comparisons of replacing **50** servers based on 3rd Gen Intel® Xeon® with new 5th Gen Intel® Xeon® processors

	Web (NGINX TLS)	Data Services (RocksDB)	Artificial Intelligence (NLP w/ Bert-Large)	Artificial Intelligence (Recommender w/ DLRM)
Number of 5th Gen Intel® Xeon® processor-based servers	29 servers	14 servers	13 servers	10 servers
Lower fleet energy	870 MWh	1,482 MWh	1,644 MWh	1,705 MWh
Reduced CO2 emissions*	368,757 kg	628,372 kg	696,768 kg	722,607 kg
TCO savings*	$442K	$1,198K	$1,273K	$1,425K

* Estimated over 4 years.

Replace Aging Servers to Save Energy and Costs

Significantly reduce data center infrastructure space, power and costs

Comparisons of replacing **50** servers based on 1st Gen Intel® Xeon® with new 5th Gen Intel® Xeon® processors

	Web (NGINX TLS)	Data Services (RocksDB)	Artificial Intelligence (NLP w/ Bert-Large)	Artificial Intelligence (Recommender w/ DLRM)
Number of 5th Gen Intel® Xeon® processor-based servers	7 servers	9 servers	3 servers	4 servers
Lower fleet energy	843.5 MWh	1191.5 MWh	1697.1 MWh	1259.2 MWh
Reduced CO2 emissions*	357,586 kg	505,136 kg	719,458 kg	533,822 kg
TCO savings*	$254K	$192K	$541K	$449K

Access the Intel® Xeon® Processor Advisor Suite to calculate the best route to lower TCO and ROI

*Estimated over 4 years.

*See [T6, T5, T7, T8] at intel.com/processor claims: 5th Gen Intel® Xeon® Scalable processors. Results may vary.

Live Migration Stability

Bring cloud-native functionality to virtual machines with Red Hat in a single open platform

Why Migrate?

| Empower operations teams to innovate | Take advantage of a cloud-like experience, everywhere | Benefit from a trusted enterprise open source platform |

Experience a complete application platform that integrates with existing infrastructure, tools, and services

Intel recommends migration from **Xeon® to Xeon®** whenever possible:
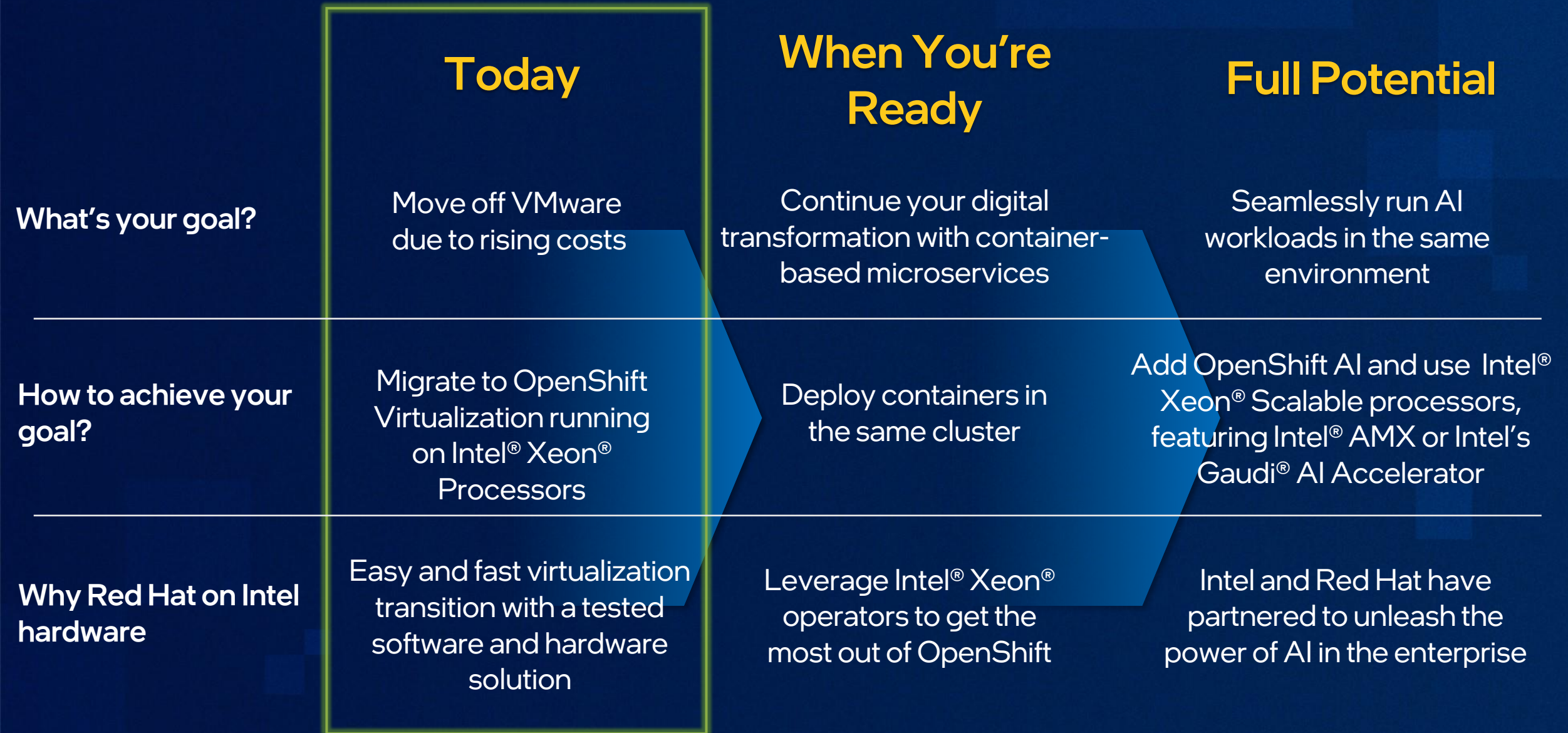
intel XEON® → intel XEON®

- Live migration requires a similar CPU architecture to reduce risk in the migration process

- Red Hat Migration Tool Recommendations:
  - *"In all cases, the CPU features of the destination node must be a superset of the CPU features on the source node."*
  Source: Red Hat Migration <u>Webpage</u>

Get Started with the <u>Migration Toolkit for Virtualization</u>

Why Partner with Intel for Digital Transformation

	Today	**When You're Ready**	**Full Potential**
What's your goal?	Move off VMware due to rising costs	Continue your digital transformation with container-based microservices	Seamlessly run AI workloads in the same environment
How to achieve your goal?	Migrate to OpenShift Virtualization running on Intel® Xeon® Processors	Deploy containers in the same cluster	Add OpenShift AI and use Intel® Xeon® Scalable processors, featuring Intel® AMX or Intel's Gaudi® AI Accelerator
Why Red Hat on Intel hardware	Easy and fast virtualization transition with a tested software and hardware solution	Leverage Intel® Xeon® operators to get the most out of OpenShift	Intel and Red Hat have partnered to unleash the power of AI in the enterprise

Summary

1. OpenShift Virtualization provides a mature, stable and scalable solution that's easy to get started

2. Full automation including; declarative, self-healing and secure

3. Integration with a broad ecosystem of partners

4. Part of an application platform that allows added features and capabilities

ACCESS NOW > Red Hat OpenShift Virtualization + Intel — Better Together Pitch Deck

Next Steps

Engage customers to understand if they are happy with their existing Virtualization offering or if familiar with OCP Virtualization

Recommend OpenShift Virtualization running on Intel® 4th or 5th Gen Xeon® processors & Connect with corresponding Red Hat/Intel AE

Begin Migration Journey

Every Company Will Be an AI Company

Virtually every company will become an **"AI company,"** AI will be used to create new services, improve internal business processes, and/or improve employee productivity

AI is now central to business strategy

79% say AI is critical to success in 2024[1]
of corporate strategists

↑ **37%** growth is underway in the global AI market[2]
CAGR, 2023-2330

GenAI is supercharging AI's potential

80% of enterprises
will use GenAI by 2026[4]

AI architects face too many choices

46% of experts
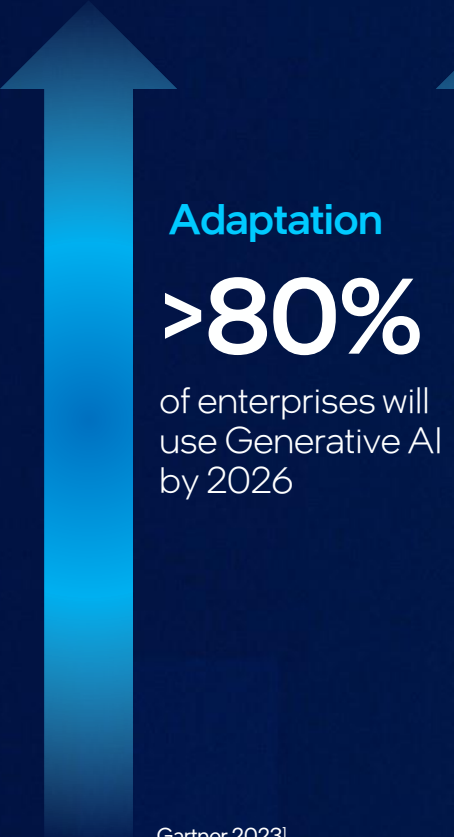say infrastructure is the biggest challenge in productizing LLMs[3]

[1] Skim AI, "10 Enterprise AI Statistics to Know in 2024." https://skimai.com/10-enterprise-ai-stats-to-know-in-2024/.
[2] Grandview Research, "Artificial Intelligence Market Size and Share Report, 2030." https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.
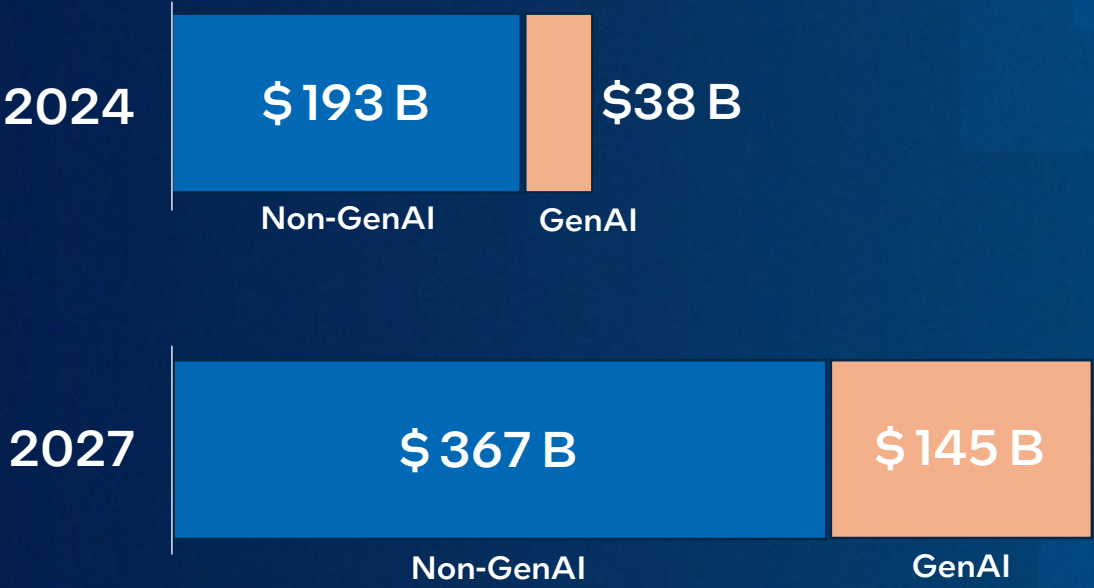[3] mlinsider. "The state of Generative AI and Machine Learning at the end of 2023." https://cnvrg.io/wp-content/uploads/2023/11/ML-Insider-Survey_2023_WEB.pdf.
[4] Gartner, October 11, 2023. "Gartner says More than 80% of Enterprises Will Have Used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026."
https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026.

AI Inflection Point

Adaptation

>80%

of enterprises will use Generative AI by 2026

Gartner 2023[1]

Accessibility

>80%

of world's data stored on-prem in the enterprise

Gartner 2023[1]

Expected Spend on Artificial Intelligence

2024

$193 B — Non-GenAI

$38 B — GenAI

2027

$367 B — Non-GenAI

$145 B — GenAI

IDC AI and Generative AI Spending Guide Forecast[2]

1 https://www.gartner.com/en/documents/4558899
2 https://www.idc.com/fdsfile/download/0cb3a164-e692-412b-a7ce-514e5f3bcee3

Transformative AI Joint Offerings

Red Hat OpenShift AI

Enterprise AI from **intel**®
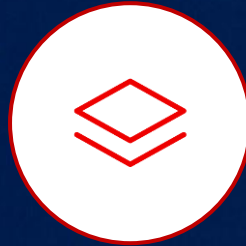
- Red Hat and Intel offer validated and integrated hardware and software elements for enterprise AI

- Together, they reduce the complexity of selecting and integrating solution components

- Customers attain faster time to value, with lower cost and less risk
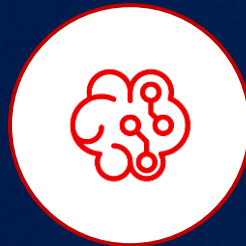
intel. + **Red Hat**

Red Hat® OpenShift® AI

Red Hat OpenShift AI

Provide data scientists an open ML platform

Built on top of OpenShift
A secure, supported and proven enterprise platform on hybrid cloud

Designed for machine learning
An easy-to-manage platform, with hardware acceleration and an ecosystem of ML tools

Empowered data science
A self-service infrastructure with foundational elements from open source

OpenShift is a Trusted Enterprise Kubernetes

Red Hat OpenShift

Red Hat OpenShift offers a comprehensive platform powered by Kubernetes to efficiently build, update, and scale applications. Enhance productivity and speed up the application deployment process by using a full suite of services, all adaptable to your preferred infrastructure.
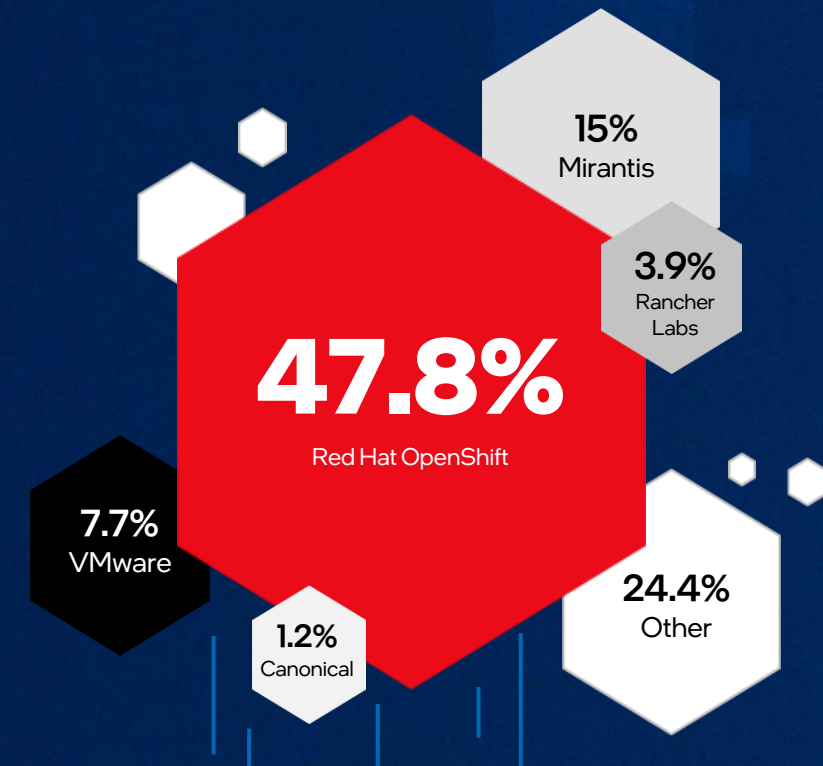
Red Hat OpenShift AI

Red Hat OpenShift AI serves as a versatile, scalable MLOps platform equipped with tools for constructing, deploying, and managing applications powered by AI. OpenShift AI was developed using open-source technologies: it offers reliable, operationally consistent features for teams to experiment, serve models, and launch innovative applications. OpenShift AI supports the complete lifecycle of AI/ML experiments and models, both on-premises and in the public cloud.

Learn more

Integrate and simplify production AI with Intel and Red Hat

The industry is moving towards a Hybrid Multi-Cloud operating model. Red Hat OpenShift is the most widely deployed multi-cloud container platform at **47% MSS**.

15%
Mirantis

3.9%
Rancher Labs

47.8%
Red Hat OpenShift

7.7%
VMware

24.4%
Other

1.2%
Canonical

Red Hat Supercharges AI Workloads with Intel® Technologies from the Datacenter to the Edge

• Red Hat OpenShift AI and Intel's AI technologies enable organizations to accelerate generative AI and machine learning innovation on a trusted, integrated platform

• Red Hat OpenShift AI now supports Intel® Gaudi® AI accelerators

• Get Started and accelerate AI development with Intel® Tiber™ Developer Cloud

PRESS RELEASE

Red Hat and Intel's open-source approach to AI innovation empowers organizations to deploy anywhere on their platform of choice, accelerating time to market and providing ubiquitous AI building blocks that are more cost-effective at scale.

Intel AI Software

Intel AI Tools

Data Analytics
Distributed data ingestion and preprocessing

Intel® Distribution of Modin

Machine Learning
Training and prediction

Intel® Extension for Scikit-learn

Intel® Optimization for XGBoost

Deep Learning
Modeling, training, transfer learning, fine-tuning, inference

PyTorch Optimizations from Intel

TensorFlow Optimizations from Intel

Optimization & Inference
Perform model compression and deploy to the cloud, edge, or PC

Intel® Neural Compressor

OpenVINO

Intel® AI Reference Models — Pretrained, Intel-optimized models

Intel® Tiber™ Studio — Full-stack machine learning operating system

Intel® Gaudi® Software — Leverage dedicated AI accelerators

Open Platform for Enterprise AI

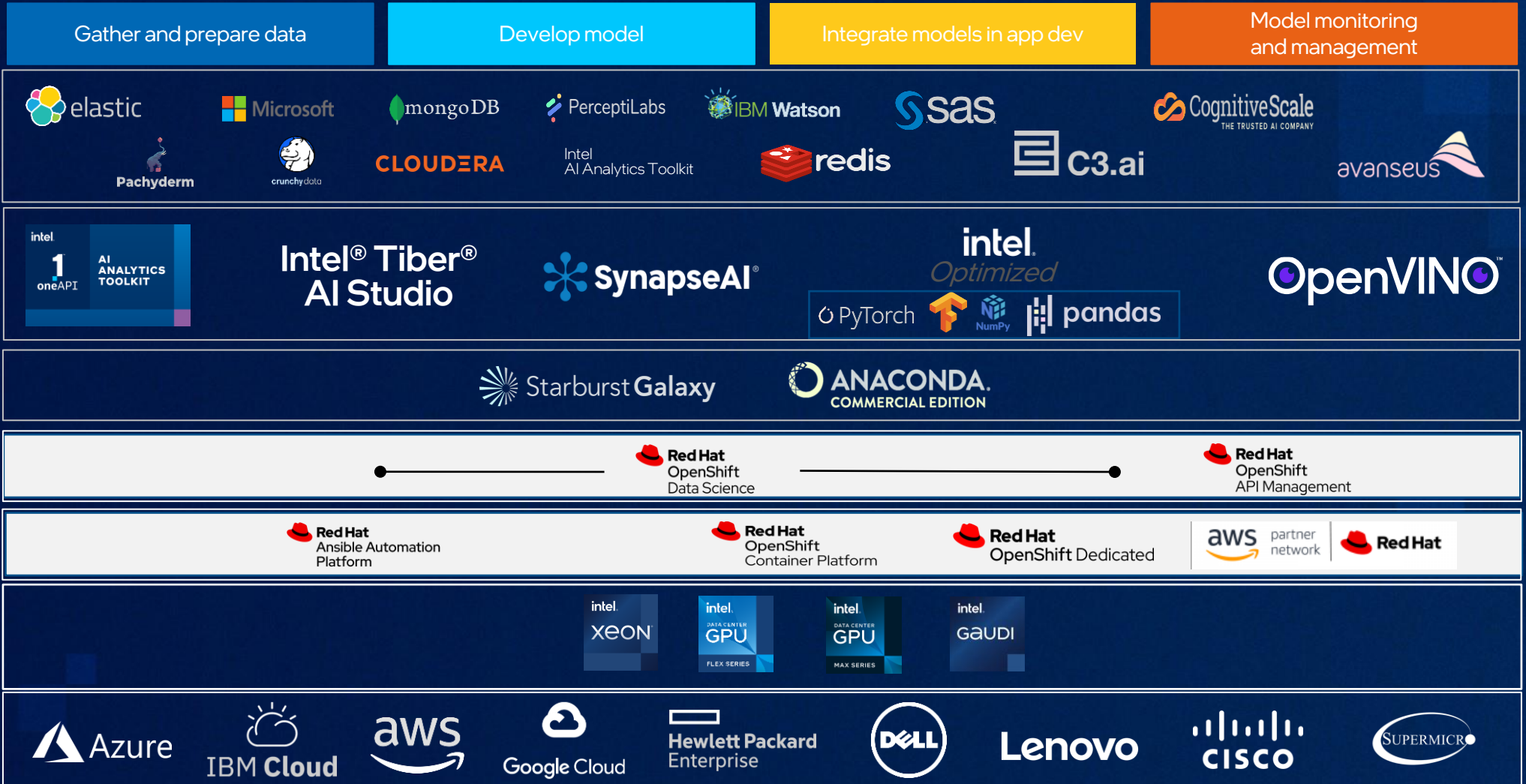**Intel® Tiber™ Edge Platform**

Intel® Tiber™ Developer Cloud

1 oneAPI   oneAPI Open, Cross-Architecture Programming Model

Red Hat® + Intel® AI

Integrating + Simplifying the AI Ecosystem

Retrain models

Gather and prepare data	Develop model	Integrate models in app dev	Model monitoring and management

Customer managed ISV software

elastic · Microsoft · mongoDB · PerceptiLabs · IBM Watson · SAS · CognitiveScale THE TRUSTED AI COMPANY · Pachyderm · crunchy data · CLOUDERA · Intel AI Analytics Toolkit · redis · C3.ai · avanseus

Intel® Software Components

intel 1 oneAPI AI ANALYTICS TOOLKIT · Intel® Tiber® AI Studio · SynapseAI® · intel Optimized (PyTorch · NumPy · pandas) · OpenVINO™

ISV managed cloud services

Starburst Galaxy · ANACONDA. COMMERCIAL EDITION

Red Hat managed cloud services

Red Hat OpenShift Data Science · Red Hat OpenShift API Management

RedHat Open hybrid cloud platform with self-service capabilities

Red Hat Ansible Automation Platform · Red Hat OpenShift Container Platform · Red Hat OpenShift Dedicated · aws partner network · Red Hat

Intel® Xeon® & Accelerators (GPU, Gaudi)

intel XEON · intel DATA CENTER GPU FLEX SERIES · intel DATA CENTER GPU MAX SERIES · intel GAUDI

Infrastructure

Azure · IBM Cloud · aws · Google Cloud · Hewlett Packard Enterprise · DELL · Lenovo · CISCO · SUPERMICRO
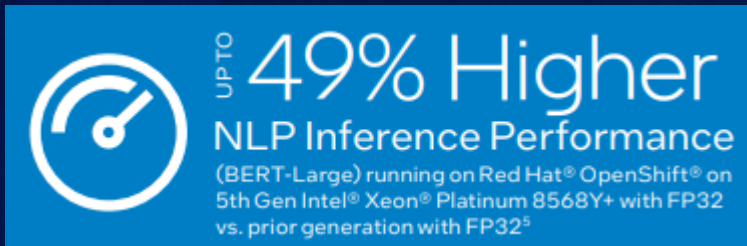
Accelerate Red Hat® OpenShift® AI Workflows:
Using Intel's Newest Processor Features with Red Hat® Validated Patterns

Solution Benefits

Extended Red Hat® validated patterns from Intel make it easy to enable AI and security features :

- **Effortless deployment** of comprehensive, fully operational AI workflows, eliminating the need for extensive manual configurations
- **Accelerated AI model training and inference** with Intel® Advanced Matrix Extensions
- **Enhanced security** using Intel® Software Guard Extensions, which supports confidential computing
- **Offloading of encryption and compression** to a dedicated accelerator to improve application performance, using Intel® Quick Assist Technology

PROOF POINT

UP TO **49% Higher**
NLP Inference Performance
(BERT-Large) running on Red Hat® OpenShift® on
5th Gen Intel® Xeon® Platinum 8568Y+ with FP32
vs. prior generation with FP32[5]

*configuration details in Solution Reference Architecture linked across, results may vary

READ MORE

Solution Reference Architecture

Solution Architecture Highlights

Figure 1 illustrates the solution stack for Intel's extension of Red Hat validated patterns. Through node feature discovery, the patterns can determine if a particular feature, such as Intel AMX or Intel SGX, is available. OpenShift AI, Kubernetes management software, and other pattern components all integrate with OpenShift.

Validated Pattern Technologies

Red Hat® Ansible® Automation Platform and Helm	OpenShift® GitOps (Drop Code)	Red Hat® Advanced Cluster Management for Kubernetes
Hashicorp Vault	Node Feature Discovery	Red Hat® OpenShift® AI

Red Hat® OpenShift®

5th Gen Intel® Xeon® Scalable Processors

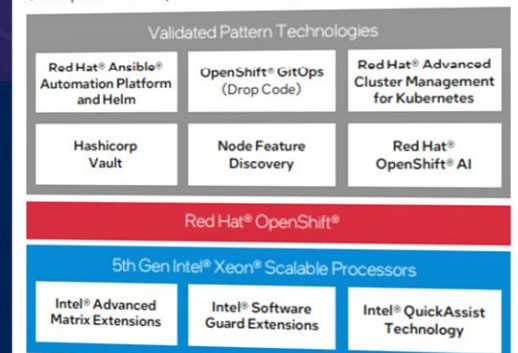| Intel® Advanced Matrix Extensions	Intel® Software Guard Extensions	Intel® QuickAssist Technology

Figure 1. Extending Red Hat® validated patterns to include Intel® technologies can accelerate AI workloads and enhance security.

Improve Time-to-Value with Validated Patterns

With **Red Hat Validated Patterns**, organizations can find tested and supported solutions for many AI use cases, from customized chatbots to fraud detection to medical diagnosis.

Benefits include:

- **Automatic deployment** of full application stack

- **Wide range of customization**, including Intel® accelerators like:
  - Intel ® Advanced Matrix Extensions (Intel® AMX)
  - Intel® Quick Assist Technology (Intel® QAT)
  - Intel® Software Guard Extensions (Intel® SGX)
  - Intel® Trust Domain Extensions (Intel® TDX)

- **Variety of use cases** including Medical Diagnosis and Fraud Detection

- List of validated patterns with maturity and support information is publicly available on https://validatedpatterns.io/patterns/

LEARN MORE >   Integrate and Simplify Production AI with Intel and Red Hat

Accelerating Generative AI and Large Language Models with Intel® Gaudi® AI Accelerator

intel.
GAUDI

Intel® Gaudi® 2 delivers leading performance and optimal cost savings for AI training

[Press Release](#)

The Gaudi® 2 deep learning accelerator performs competitively on deep learning training and inference, with up to **2.4x faster performance than Nvidia A100**[1]

[Newsroom](#)

[Tech Article](#)

Gaudi® 2 delivers compelling performance vs. Nvidia's H100[2,3] **for GPT-3 and GPT-J**

[Newsroom](#)

[MLCommons Announcement](#)

COMING SOON:
Intel® Gaudi® 3 AI accelerator
Bringing Choice to GenAI with Performance, Scalability and Efficiency

Intel® Gaudi® 3 will deliver a significant leap in AI training and inference for global enterprises looking to deploy GenAI at scale

READ MORE

[Intel® Gaudi® 2 Remains Only Benchmarked Alternative to NV H100 for GenAI Performance](#)

[Intel Breaks Down Proprietary Walls to Bring Choice to Enterprise GenAI Market](#)

[1] Performance varies by use, configuration, and other factors; workloads and configuration details available at: intel.com/performanceindex Results may vary.

[2,3] Performance varies by use, configuration, and other factors; workloads and configuration details available at: https://mlcommons.org/2023/09/mlperf-results-highlight-growing-importance-of-generative-ai-and-storage/ Results may vary.

Next Steps

Sign up for Intel® Tiber™ Developer Cloud

Try a 60-day trial of Red Hat OpenShift AI

Download pre-packaged cloud-native Intel® AI software

MORE INFORMATION

- Red Hat® and Intel® AI and Machine Learning: The Perfect Combination for Data Scientists
- Essential Tools for Jumpstarting AI Development Projects
- How to Use Intel®-Optimized AI Software in the Cloud
- Speed Up Machine Learning Training on CPUs with AI Tools
- Deploying Intel Enterprise AI with Red Hat OpenShift AI Made Simple
- Red Hat OpenShift AI + Intel — Better Together Pitch Deck

Developer Resources

- Create a Red Hat OpenShift AI Account
  - Developer Sandbox

- Intel Technology Enabling for OpenShift
  - GitHub

- Developer Resources from Intel and Red Hat

- Optimize AI with Powerful Drop-In Acceleration from Intel & Red Hat*

- How to Get the Most Out of Red Hat OpenShift* AI with Intel® AI Tools

- Streamline AI Adoption and Deployment Using Intel Enterprise AI with Red Hat® OpenShift® AI

AI application evaluation on Intel® hardware through Red Hat* OpenShift* AI platform using the Intel® Developer Cloud for the Edge
- devcloud.intel.com/edge

- Container Playground Entry point
  - https://www.intel.com/devcloud-containers

- Container Playground Developer Guide
  - https://www.intel.com/content/www/us/en/develop/documentation/devcloud-containers/top.html

Notices and Disclaimers

What's the right transition for your customer?

Guide your customer to the best refresh option

Current install base	Good	Better	Best
Intel Xeon Silver 4000 Series	4514Y 16C/150W	4516+ 24C/185W	5520+ 28C/205W
Intel Xeon Gold 5000 Series	5515Y 8C/185W	5520+ 28C/205W	6530 32C/270W
Intel Xeon Gold 6000 Series	6526Y 16C/195W	6548Y 32C/250W	8558 48C/330W
Intel Xeon Platinum 8000 Series	8562Y+ 32C/300W	8568Y+ 48C/350W	8592+ 64C/350W

Refresh your infrastructure today, to be ready for tomorrow's demands

Refresh from 1st Gen Intel® Xeon® to 5th Gen Intel Xeon

AI
(BertLarge)

up to **42x**
higher performance[1]

up to **20x**
higher performance/ watt[1]

Refresh from 2nd Gen Intel® Xeon® to 5th Gen Intel Xeon

AI
(DLRM)

up to **10x**
higher performance[2]

up to **6x**
higher performance/ watt[2]

Intel vs. AMD

AI
(Recommender - DLRM)

up to **2.8x**
higher batched inference performance[3]

up to **2.6x**
higher performance/ watt[3]

[1,2,3] See [A37, A38, A208] at intel.com/processorclaims:5th Gen Intel Xeon Scalable processors. Results may vary
*Enabled in platform BIOS, visit intel.com/processorclaims: 5th Gen Intel Scalable processors for more information

intel XEON®

Exceptional Performance & Efficiency

5th Gen Intel® Xeon® processors deliver impressive performance-per-watt gains across all workloads, plus outsized performance and lower TCO

- Optimize AI, HPC, network, data analytics & storage workloads with Intel® Accelerator Engines

- Enhanced platform capabilities
  - 3x increase in shared last-level cache
  - PCIe 5 – Double I/O bandwidth
  - DDR5 – Increase memory bandwidth
  - CXL® 1.1 – Next Gen I/O for low latency and performance
  - Optimized Power Mode enables energy savings with minimal impact of performance*
  - Advanced security technologies to help protect data with Intel Software Guard Extensions (Intel SGX) and Intel® Trust Domain Extensions (Intel TDX)

- To learn more about SKU transitions and refresh opportunities:
  - https://xeonprocessoradvisor.intel.com

→ TCO Advisor Tool

What's the right transition for your customer?

Guide your customer to the best refresh option

Current install base	Good	Better	Best
Intel Xeon Silver 4000 Series	4514Y 16C/150W	4516+ 24C/185W	5520+ 28C/205W
Intel Xeon Gold 5000 Series	5515Y 8C/185W	5520+ 28C/205W	6530 32C/270W
Intel Xeon Gold 6000 Series	6526Y 16C/195W	6548Y 32C/250W	8558 48C/330W
Intel Xeon Platinum 8000 Series	8562Y+ 32C/300W	8568Y+ 48C/350W	8592+ 64C/350W

Intel® Xeon®, the processor designed for AI

Total Cost of Ownership Savings
3rd Gen Intel® Xeon® to 5th Gen Intel Xeon

AI
(Recommender -DLRM)

up to **5:1**
server consolidation[1]

up to **72%**
TCO savings[1]

Performance & Efficiency Gains
3rd Gen Intel® Xeon® to 5th Gen Intel Xeon

AI
(Real Time Inference)

up to **14x**
higher performance[2]

up to **9.5x**
higher performance/ watt[2]

Intel vs. AMD

AI
(Recommender - DLRM)

up to **2.8x**
higher batched inference performance[3]

up to **2.6x**
higher performance/ watt[3]

[1, 2, 3] See [T12, A16, A208] at intel.com/processorclaims:5th Gen Intel Xeon Scalable processors. Results may vary
*Enabled in platform BIOS

---

intel. **XEON**

Exceptional Performance & Efficiency

5th Gen Intel® Xeon® processors deliver impressive performance-per-watt gains across all workloads, plus outsized performance and lower TCO

- Optimize AI, HPC, network, data analytics & storage workloads with Intel® Accelerator Engines

- Enhanced platform capabilities
  - 3x increase in shared last-level cache
  - PCIe 5 – Double I/O bandwidth
  - DDR5 – Increase memory bandwidth
  - CXL® 1.1 – Next Gen I/O for low latency and performance
  - Optimized Power Mode* enables energy savings with minimal impact of performance
  - Advanced security technologies to help protect data with Intel Software Guard Extensions (Intel SGX) and Intel® Trust Domain Extensions (Intel TDX)

- To learn more about SKU transitions and refresh opportunities:
  - https://xeonprocessoradvisor.intel.com

TCO Advisor Tool

Benefits to upgrading infrastructure configurations

3:1 - 5th Gen Xeon Server Consolidation:
Up to 3:1 consolidation and 61% TCO savings with 5th Gen Intel Xeon processors: Calculations as of June 2024 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a 5-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4214 processors with new servers using new Intel Xeon 5520+ processors. Results may vary.

[4214: https://www.spec.org/cpu2017/results/res2020q2/cpu2017-20200427-22184.html
5520+: https://www.spec.org/cpu2017/results/res2024q2/cpu2017-20240325-42544.html]

1st Gen Intel Xeon to 5th Gen Intel Xeon:
Calculations as of June, 2024 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a 5-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4110 processors with new servers using new 4th Gen Intel Xeon processors. Results may vary. Performance measurements based on published SPECrate®2017_int_base on spec.org as of March 28, 2023

8168: https://www.spec.org/cpu2017/results/res2019q3/cpu2017-20190827-17261.html
8562Y+: https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240311-42071.html
6138: https://www.spec.org/cpu2017/results/res2021q4/cpu2017-20211122-30177.html
8558: https://www.spec.org/cpu2017/results/res2024q2/cpu2017-20240325-42453.html
5118: https://www.spec.org/cpu2017/results/res2019q3/cpu2017-20190805-16469.html
5520+: https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240311-41976.html
4110: https://www.spec.org/cpu2017/results/res2020q4/cpu2017-20201015-24218.html

2nd Gen Intel Xeon to 5th Gen Intel Xeon:
Calculations as of June 2024 based on the Intel® Node TCO & Power Calculator using default cost, power and TCO assumptions over a multi-year TCO horizon comparing replacing 50 older servers with Intel Xeon 4110 processors with new servers using new 4th Gen Intel Xeon processors. Results may vary. Performance measurements based on published SPECrate®2017_int_base on spec.org as of March 28, 2023

8260: https://www.spec.org/cpu2017/results/res2021q3/cpu2017-20210815-28685.html
8558: https://www.spec.org/cpu2017/results/res2024q2/cpu2017-20240325-42453.html
6230: https://www.spec.org/cpu2017/results/res2021q1/cpu2017-20210301-25037.html
5218: https://www.spec.org/cpu2017/results/res2020q3/cpu2017-20200731-23587.html
6538Y+: https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240115-40663.html
4214: https://www.spec.org/cpu2017/results/res2021q3/cpu2017-20210824-28887.html
5520+: https://www.spec.org/cpu2017/results/res2024q1/cpu2017-20240311-41976.html