Self-managed Red Hat OpenShift subscription guide

Table of contents

Introduction

This document will help you understand the subscription model for self-managed Red Hat® OpenShift® offerings and provide step-by-step instructions for how to approximate the number of entitlements needed for your Red Hat OpenShift environment. More accurate sizing information is available on request.

Red Hat OpenShift subscription offerings

Red Hat OpenShift provides a consistent application development and management platform across an open hybrid cloud environment, and supports on-premises, virtual, and physical infrastructure, and private cloud, public cloud, and edge deployments. There are 2 ways to operate and use Red Hat OpenShift: self-managed OpenShift and fully managed OpenShift cloud services.

Self-managed OpenShift allows you to install, operate, and manage Red Hat OpenShift environments with maximum control, flexibility, and customization, so you can operate your own environment starting with the infrastructure. Self-managed OpenShift is supported on-premises— using physical servers, virtualization, and in a private cloud environment—and in supported public cloud environments. You control upgrades, manage the lower-level infrastructure, and maintain service-level agreements (SLA).

OpenShift cloud services are fully managed and operated by Red Hat and its public cloud partners in major public clouds. A dedicated site reliability engineering (SRE) team manages and maintains Red Hat OpenShift core services and infrastructure, allowing your DevSecOps teams to concentrate on developing and deploying new applications and modernizing existing ones.

All editions of OpenShift offer a consistent user experience for developers and operations across every environment, allowing you to transfer your skills and applications to the cloud environments where your applications run best. Refer to the OpenShift documentation for your version to determine which components, features, and add-ons are included with each OpenShift entitlement level.

Self-managed OpenShift software offerings:

▸ **Red Hat OpenShift Kubernetes Engine:** A hybrid cloud, enterprise Kubernetes runtime engine that provides core Red Hat OpenShift functionality to deploy and run applications, which you install and manage in a datacenter, public cloud, or edge environment.

▸ **Red Hat OpenShift Container Platform:** A hybrid cloud, enterprise Kubernetes platform used to build, deploy, and run applications, which you install and manage in a datacenter, public cloud, and edge environment.

▸ **Red Hat OpenShift Platform Plus:** A single hybrid cloud platform that allows enterprises to build, deploy, run, and manage intelligent applications at scale across multiple clusters and cloud environments. Multiple layers of security, manageability, and automation provide consistency throughout the software supply chain. OpenShift Platform Plus subscriptions are available for x86-based clusters only.

OpenShift cloud services offerings:

▸ **Red Hat OpenShift Dedicated:** Fully managed Red Hat OpenShift service on Amazon Web Services (AWS) and Google Cloud. Read more.

▸ **Microsoft Azure Red Hat OpenShift:** Fully managed Red Hat OpenShift service on Microsoft Azure, jointly managed by Red Hat and Microsoft. Read more.

▸ **Red Hat OpenShift Service on AWS:** Fully managed Red Hat OpenShift service on Amazon Web Services, jointly managed by Red Hat and AWS. Read more.

▸ **Red Hat OpenShift Kubernetes Service on IBM Cloud:** Fully managed Red Hat OpenShift service on IBM Cloud, jointly managed by Red Hat and IBM. Read more.

Self-managed OpenShift environments

Self-managed OpenShift (Red Hat OpenShift Platform Plus, Red Hat OpenShift Container Platform, and Red Hat OpenShift Kubernetes Engine) can be used anywhere 64-bit Red Hat Enterprise Linux® is certified and supported. Refer to the documentation to learn more about OpenShift deployment methods and supported infrastructure types.

Subscription types

There are 2 types of subscriptions for Red Hat OpenShift Platform Plus, Red Hat OpenShift Container Platform, and Red Hat OpenShift Kubernetes Engine. There are 2 support levels available for each subscription.

1. x86 core or artificial intelligence (AI) accelerator card-based, 2 cores or 4 virtual central processing units (vCPUs), or 2 physical AI accelerators.

   ▸ When entitling CPU cores, this is based on the aggregate number of physical cores or vCPUs across all OpenShift compute nodes running across all OpenShift clusters.

   ▸ When entitling accelerators, this is based on any accelerator devices installed on OpenShift compute nodes that are being used by workloads launched from OpenShift. Each subscription covers 2 physical AI accelerator devices, regardless of the number of cores on the cards or whether virtual graphics processing units (GPUs) have been created.

   ▸ Available with Standard 8x5 or Premium 24x7 support SLA.

2. Bare-metal socket pair, up to 64 cores across 1-2 sockets. This subscription is available only for x86 and ARM bare-metal physical nodes where OpenShift is installed directly to the hardware. Accelerators must be entitled via sufficient additional AI accelerator subscriptions. Available with Standard 8x5 or Premium 24x7 support SLA.

As with Red Hat Enterprise Linux:

▸ OpenShift subscriptions (Red Hat OpenShift Platform Plus, Red Hat OpenShift Container Platform, and Red Hat OpenShift Kubernetes Engine) are stackable to cover larger hosts.

▸ Core-based subscriptions can be distributed to cover all OpenShift compute nodes across all OpenShift clusters. For example, 100 2-core Red Hat OpenShift Platform Plus subscriptions will provide 200 cores (400 vCPUs) that can be used across any number of compute nodes, across all running OpenShift clusters.

Disaster recovery

Red Hat defines 3 types of disaster recovery (DR) environments—hot, warm, and cold. Paid Red Hat OpenShift subscriptions are needed for hot DR only.

▸ Hot DR systems are defined as fully functional and running concurrently with the production systems. They are ready to immediately receive traffic and take over in the event of a disaster within the primary environment. When data volumes are actively being replicated either synchronously or asynchronously between clusters they are considered to be "hot" DR systems.

▸ Warm DR systems are defined as already prepared to deploy and host containerized workload representing a reasonable facsimile of that found in the primary site, but containing no customer workload from the source cluster(s). Warm DR systems should not be participating in active data volume replication either synchronously or asynchronously between clusters. Warm DR recovery requires customer's data having to be restored onto the existing cluster hardware from a last backup located outside of the cluster.

▸ Cold DR systems are defined as having the infrastructure in place, but not the full technology (hardware, software, data) needed to restore service.

Hibernating clusters that are not expressly configured and designed for warm or cold DR—such as clusters running on cloud services that are temporarily hibernating as a result of lower demand—require subscriptions. When warm or cold DR clusters are brought out of hibernation for running workloads, subscriptions are required. Bringing a cluster out of hibernation temporarily for routine maintenance or tests does not require an additional subscription for any of the components in OpenShift software offerings.

For both warm DR and cold DR, Red Hat OpenShift subscriptions can be transferred from the primary environment to the DR environment when the disaster occurs to restore service and maintain compliance with Red Hat's subscription terms.

Migration and swing upgrades

Red Hat OpenShift 4 provides in-place upgrades between minor versions. Nevertheless, if you are upgrading from Red Hat OpenShift 3, or need to perform a swing upgrade between OpenShift 4 minor versions as a result of maintenance windows or other considerations, your Red Hat OpenShift subscription will cover both the original and destination infrastructure of a 1-way migration until such migration is complete. During the migration, Red Hat's subscription management tools will show your environment as being out-of-compliance on the basis of the number of Red Hat OpenShift subscriptions you purchased. Red Hat allows this for major version upgrades and will not require the purchase of additional subscriptions to get back into compliance during the migration. Finally, OpenShift provides tooling to assist in these migrations, along with Red Hat consulting services, if desired. See documentation on the migration toolkit for containers.

Entitlements for cores with hyperthreading

Making a determination about whether or not a particular OpenShift node uses 1 or more physical cores is determined by whether or not that system has multiple threads per core enabled. Discover how to determine whether a particular system supports hyperthreading.

Virtualized OpenShift nodes using logical CPU threads, also known as simultaneous multithreading (SMT) for AMD EPYC CPUs or hyperthreading with Intel CPUs, calculate their core utilization for OpenShift subscriptions based on the number of cores/CPUs assigned to the node, however each subscription covers 4 vCPUs/cores when logical CPU threads are used. Red Hat's subscription management tools assume logical CPU threads are enabled by default on all systems.

Bare metal subscriptions only count physical cores, logical CPU threads are not counted when calculating the number of subscriptions needed for bare metal.

Entitlements for virtualized compute nodes

When deploying OpenShift compute nodes to a hypervisor, such as VMware vSphere, Red Hat OpenShift Virtualization, or Red Hat OpenStack® Platform, the number of entitlements required is the lesser of the cores/threads assigned to the virtual compute nodes or the sum of cores of the physical servers.

For example, an OpenShift cluster composed of virtual compute nodes totaling 200 cores would require one hundred 2-core/4-thread entitlements when the underlying physical hypervisor nodes have more than 200 cores. The same 200 core OpenShift cluster, when deployed to a physical hypervisor cluster totaling just 120 cores—resulting in overcommitment of OpenShift CPU resources at the hypervisor—would need only 120 cores of entitlements (sixty 2-core/4-thread).

Core bands

Red Hat OpenShift subscriptions use a system of measure called core bands. This means subscriptions (entitlements to deploy and use OpenShift) are applied and consumed at the OpenShift cluster level and apply to all eligible OpenShift compute nodes on that cluster. If you have multiple OpenShift clusters, you would aggregate the sum of cores consumed by the OpenShift compute nodes across all clusters to determine how many subscriptions are needed. For example, if you have 100 two-core Red Hat OpenShift Container Platform subscriptions, a total of 200 cores (400 vCPUs with hyperthreading) are available to be applied to the OpenShift compute nodes across all running OpenShift clusters.

Bare-metal server considerations

A physical server can be entitled using either core-based (2-core/4 logical CPU) or socket-based (up to 64 cores across 1-2 sockets) Red Hat OpenShift subscriptions. If core-based subscriptions are used, stack a sufficient number of them to cover the total number of physical cores in the server.

In addition to core-based subscriptions, Red Hat also offers OpenShift socket-based subscriptions. For certain deployment types, this is a more economical option. The socket-based subscriptions are limited to entitling an x86 server with up to 2 sockets with a total of 64 cores across them. The socket-based subscriptions are currently available for x86 and ARM servers only and not for the IBM Z and IBM Power architectures.

To entitle a physical server, stack 1 or more subscriptions to cover either the total number of sockets or physical cores in it (whichever is greater). For example, a server has 2 sockets and 48 cores. One subscription is needed because the server has 1-2 sockets and less than 64 cores, while a server with 2 sockets and 96 cores would need 2 subscriptions. Two subscriptions are needed to cover 96 cores because a single subscription covers a maximum of 64 cores. A single bare-metal subscription cannot be split across multiple physical hosts.

Bare-metal socket-pair subscriptions also come with infrastructure subscriptions for the control plane and infrastructure. Qualified control plane and infrastructure workloads (explained in detail below) may be deployed to either, or a mix of, virtual and physical servers when using socket-based subscriptions. A mixed cluster, composed of virtual and physical nodes, is a supported deployment architecture when deploying a platform agnostic, nonintegrated cluster without cloud provider or machine application programming interface (API) integrations.

Each physical, bare-metal server using socket-based entitlements can only be used as a single OpenShift node. Using a hypervisor, including OpenShift Virtualization, to create virtual OpenShift compute nodes will require entitling the virtual compute nodes using core-based subscriptions. This means that the bare-metal socket-pair model is best suited for resource intensive workloads like OpenShift Virtualization (where each workload is running a full virtual machine).

Finally, using the bare-metal socket-pair subscriptions does not change the limitation of the number of containers per node (currently 250-500).

Alternative architectures (ARM, IBM Z, IBM LinuxONE , IBM Power)

Note: While this document refers only to IBM Z from here on, all information that references IBM Z also applies to IBM LinuxONE.

Red Hat OpenShift Container Platform can also run on ARM, IBM Z, and IBM Power for customers using these platforms as the standard for building and deploying cloud-native applications and microservices. Only the core-based subscription model is supported for IBM Z and IBM Power platforms.

ARM clusters are entitled using the same rules as x86.

For IBM Z customers, Red Hat OpenShift does not require the entire physical node to be entitled, only the cores used by OpenShift. IBM Z customers know this as "subcapacity" entitlement. Customers using only a subset of the available cores (compute capacity) on their IBM Z environment for OpenShift Container Platform only require subscriptions for the subset that is used for the compute nodes. This applies regardless of how CPU partitioning is achieved, whether by CPU pooling, capping, separate logical partitions (LPARs), or other means.

For IBM Z, 1 Integrated Facility for Linux (IFL) requires 1 OpenShift core-based subscription. When no partitioning is used, up to 3 IFLs per cluster may be identified per OpenShift cluster for control plane or infrastructure services running on the host. These must be actively used for control plane and/or infra services to qualify and do not require OpenShift entitlements. Three-node compact cluster deployments require all IFLs to be entitled.

Red Hat OpenShift Platform Plus components beyond OpenShift Container Platform are not supported on IBM Z and IBM Power at this time, with the following exceptions:

▸ A standalone subscription of Red Hat Quay running on x86 architectures provides a global registry for multiple architectures, including IBM Z and IBM Power clusters. Red Hat Quay itself will not run on IBM Z or IBM Power.

▸ Red Hat Advanced Cluster Management for Kubernetes can be installed on IBM Z or IBM Power environments and manage Red Hat OpenShift nodes running on IBM Z or IBM Power environments.

▸ With Red Hat Advanced Cluster Security for Kubernetes, you can protect clusters running on Red Hat OpenShift on IBM Z or IBM Power by using the Red Hat Advanced Cluster Security Operator.

▸ Red Hat OpenShift Data Foundation fully supports installation on IBM Z or IBM Power.

Red Hat OpenShift Kubernetes Engine is not supported on IBM Z and IBM Power.

Microsoft Windows Server containers support

Self-managed OpenShift supports a subset of installation infrastructures and OpenShift features using Microsoft Windows Server containers. Windows Server containers are only supported on Microsoft Windows Server-based compute nodes. The control and infrastructure planes of the OpenShift environment must be running on x86 infrastructure using Red Hat Enterprise Linux or Red Hat Enterprise Linux CoreOS. For this reason, Windows Server container support is sold as a standalone subscription priced by core.

Red Hat OpenShift Platform Plus and Red Hat OpenShift Container Platform infrastructure can be used to deploy and manage Windows Server compute nodes. Microsoft Windows Server container support for Red Hat OpenShift subscriptions must be purchased as a separate add-on.

Red Hat Advanced Cluster Management and Red Hat Advanced Cluster Security do not support managing Microsoft Windows nodes, but Red Hat Quay running on x86 architectures can manage container images for Microsoft Windows Server-based workloads.

Red Hat OpenShift Platform Plus component support

The components of the Red Hat OpenShift Platform Plus subscription have different levels of support for alternative (non-x86) architectures. See the article, Red Hat OpenShift Container Platform 4.16 Multi Architecture Component Availability Matrix. For details about interoperability between the OpenShift Platform Plus components and non-x86 architectures, please see the compatibility matrices for Red Hat OpenShift Container Platform, Red Hat Advanced Cluster Management, Red Hat Advanced Cluster Security, Red Hat Quay and Red Hat OpenShift Data Foundation.

Red Hat OpenShift Platform Plus includes additional software beyond the core OpenShift Container Platform to help you manage and secure your OpenShift environment at scale across multiple clusters and multiple clouds. Red Hat OpenShift Platform Plus is available both in the core-based and bare-metal socket-pair subscription models with the previously mentioned limitations.

The additional software included with Red Hat OpenShift Platform Plus is generally limited to managing the nodes entitled with OpenShift Platform Plus subscriptions. For example, the subscription for Red Hat Advanced Cluster Management included with OpenShift Platform Plus can only be used to manage OpenShift Platform Plus entitled nodes and clusters. Customers who also wish to manage non-OpenShift Platform Plus entitled nodes and clusters, for example Red Hat OpenShift Services on AWS clusters, would need to purchase additional Red Hat Advanced Cluster Management add-on subscriptions to cover those clusters.

The additional software subscriptions are also inseparable from the OpenShift Platform Plus subscription. For example, you cannot purchase 100 OpenShift Platform Plus subscriptions, install 200 cores of Red Hat OpenShift Container Platform subscriptions, and separately use Red Hat

Advanced Cluster Management to manage 200 cores of Azure Red Hat OpenShift with the same subscription. The additional software can only be used to manage the same 200 cores on which the core OpenShift Platform Plus software is installed.

Specific rules for each layered product are:

▸ **Red Hat Advanced Cluster Management for Kubernetes:** An OpenShift Platform Plus subscription allows you to install as many Red Hat Advanced Cluster Management central instances as needed to manage your environment, and covers the management of all nodes and clusters entitled with OpenShift Platform Plus, including control plane and infrastructure nodes. If you wish to manage nodes and clusters without OpenShift Platform Plus entitlements (for example, if you also have self-managed OpenShift Container Platform or Red Hat OpenShift Kubernetes Engine entitled clusters, clusters running in a fully managed OpenShift cloud, or third-party Kubernetes environments supported by Red Hat Advanced Cluster Management), then you need to purchase Red Hat Advanced Cluster Management add-on subscriptions to cover those environments. You can choose to manage them centrally from the Red Hat Advanced Cluster Management console installed on OpenShift Platform Plus, or from a separate central application if that meets your requirement. Learn more about Red Hat Advanced Cluster Management subscriptions, Red Hat Advanced Cluster Management supported environments, and Red Hat Advanced Cluster Management best practices.

▸ **Red Hat Advanced Cluster Security for Kubernetes:** The OpenShift Platform Plus subscription allows you to install as many Red Hat Advanced Cluster Security central applications as needed to manage your environment, and covers the management of all nodes and clusters entitled with OpenShift Platform Plus, including control plane and infrastructure nodes. If you want to manage nodes and clusters without OpenShift Platform Plus entitlements (for example, if you also have self-managed OpenShift Container Platform or OpenShift Kubernetes Engine entitled clusters, clusters running in a fully managed Red Hat OpenShift cloud, or third-party Kubernetes environments supported by Red Hat Advanced Cluster Security) you need to purchase Red Hat Advanced Cluster Security add-on subscriptions to cover those environments. Red Hat's suggested practice is to manage each environment with a separate Red Hat Advanced Cluster Security central application. Learn more about Red Hat Advanced Cluster Security supported environments.

▸ **Red Hat Quay:** The OpenShift Platform Plus subscription allows you to install Red Hat Quay on any cluster that has an OpenShift Platform Plus entitlement. There is no limit on the number of Quay deployments you can install on your OpenShift Platform Plus clusters. Quay can then serve any supported Kubernetes environment you wish, including the OpenShift Platform Plus environment, other self-managed OpenShift clusters, managed OpenShift services, and supported third-party Kubernetes. If you wish to install Quay in a non-OpenShift Platform Plus environment, you will need to purchase a separate Red Hat Quay subscription. Red Hat Quay is also available as a fully managed SaaS offering.

▸ **Red Hat OpenShift Data Foundation.** The OpenShift Platform Plus subscription allows you to install Red Hat OpenShift Data Foundation Essentials on any cluster that has an OpenShift Platform Plus entitlement. The Red Hat Data Foundation entitlement is limited to the features available in Essentials, and to 256TB of data storage per OpenShift cluster. You can choose to extend functionality and capacity through additional subscriptions. See the OpenShift Data Foundation Subscription Guide (customer portal login required) or consult with a Red Hat sales representative for further guidance.

Subscription requirements for OpenShift node types and features

Red Hat OpenShift control plane and infrastructure nodes

Each self-managed Red Hat OpenShift subscription includes entitlements for Red Hat OpenShift, Red Hat Enterprise Linux, and other OpenShift-related components. These entitlements are included for running OpenShift control plane and infrastructure workloads and do not need to be accounted for when determining the number of subscriptions needed.

Red Hat OpenShift Platform Plus subscriptions include the management of these control plane and infrastructure nodes by Red Hat Advanced Cluster Security for Kubernetes and Red Hat Advanced Cluster Management for Kubernetes.

Infrastructure nodes

A default install deploys a highly available OpenShift control plane composed of three control plane nodes, in addition to OpenShift compute nodes, to run end-user applications. By default, Kubernetes control plane components (e.g. API server, etcd, scheduler) and supporting cluster services (e.g. monitoring, registry) will be deployed on the OpenShift control plane nodes. However, you may decide to move some of these supporting cluster services to dedicated infrastructure nodes.

To qualify as an infrastructure node and use the included entitlement, only components that are supporting the cluster, and not part of an end-user application, may be running on those instances. Examples include:

▸ OpenShift registry.

▸ OpenShift Ingress Router (local and global and multicluster ingress).

▸ OpenShift monitoring.

▸ OpenShift log management.

▸ HAProxy-based instances used for cluster ingress.

▸ Red Hat Quay.

▸ Red Hat OpenShift Data Foundation.

▸ Red Hat Advanced Cluster Management for Kubernetes.

▸ Red Hat Advanced Cluster Security for Kubernetes.

▸ Red Hat OpenShift GitOps.

▸ Red Hat OpenShift Pipelines.

▸ Hosted control planes for Red Hat OpenShift.

You may deploy and run custom and third-party agents and tools for monitoring, log data collection and forwarding, hardware drivers, infrastructure integration such as virtualization agents, etc., to infrastructure nodes without disqualifying the node for infrastructure licensing. However, this is limited only to agents and related components, including controller pods for operators and does not include the custom or third-party software suite. Examples of non-Red Hat software that qualify as infrastructure workload include:

▸ Custom and third-party monitoring agents.

- ▸ Container network interface (CNI) and container storage interface (CSI) drivers and controllers (plug-ins).

- ▸ Hardware or virtualization enablement accelerators.

- ▸ Controller pods used for Kubernetes CRD or Operators (custom or third-party software).

No other end-user application instances or types may be run on an infrastructure node using the included entitlement. To run other infrastructure workloads as application instances on Red Hat OpenShift, you must run those instances on regular application nodes. You can verify with Red Hat whether an app or service qualifies as an infrastructure workload.

Control plane nodes

OpenShift Kubernetes control plane nodes generally are not used as compute nodes, and by default, will not run application instances. However, you may choose to use a control plane node as a node for hosting end-user applications. Whether a control plane node requires a full OpenShift subscription depends on whether it runs supporting OpenShift cluster components only or end-user applications. See the infrastructure nodes section to identify qualified workloads that do not require a subscription.

In a compact 3-node cluster, end-user application workloads are run on the control plane nodes. There is no special pricing for this, and you would count the cores on the 3 nodes regardless of the role they play.

Single node OpenShift

A single node OpenShift instance deploys all OpenShift services and end-user applications to a single physical or virtual node, with optimizations to reduce the footprint and maximize resources available to applications. As with compact 3-node clusters above, there are no special accommodations for this deployment model, all cores on the node need to be entitled.

Red Hat Enterprise Linux entitlements

Red Hat Enterprise Linux entitlements for OpenShift compute nodes are included with the OpenShift entitlement. This includes entitled Pods for applications and guest OS entitlements for OpenShift Virtualization virtual machines. However, OpenShift subscriptions do not include other entitlements for Red Hat Enterprise Linux nodes with the following exception:

- ▸ A Red Hat Enterprise Linux node used specifically for bare-metal IPI provisioning.

Red Hat Enterprise Linux entitlements are not included for external nodes hosting services that OpenShift depends on, such as internet proxies, load balancers, or the mirror registry. Red Hat Satellite is not included as part of the entitlement.

Bootstrap container registry for mirroring OpenShift container images

The mirror registry for OpenShift is a Quay entitlement for the single purpose of easing the process of mirroring content required for bootstrapping disconnected OpenShift clusters and is included as part of the OpenShift subscription. This is a limited support entitlement for a minimal Quay deployment created by a specific installer, which allows you to run a Quay registry on a preprovisioned and customer-managed Red Hat Enterprise Linux host.

Note: You are permitted to use Quay as a registry mirror limited to the purpose of mirroring the OpenShift release payload, OperatorHub content, sample Operator images, Cincinnati graph image, and images related to the Red Hat OpenStack Platform offerings including Red Hat OpenStack Services on OpenShift, Red Hat OpenShift Data Foundation, and Red Hat Ceph® Storage.

The mirror registry for OpenShift is not intended to be a general-purpose registry working at arbitrary scale. Nonetheless, a limited set of custom images is permitted to be stored there, which contains required auxiliary software, for example agents. These agents must be scoped only to the node level and not provide external-facing application services themselves, and end users may not interact with them directly. Examples of these may include:

▸ Monitoring agents.

▸ CNI and CSI providers.

▸ Hardware or virtualization enablement agents.

▸ Operators supporting independent software vendor (ISV) services.

▸ Custom operators as deployment controllers.

Accelerator devices

In recent years, specific hardware technologies have come onto the market that allow certain compute workloads to perform with increased speed. These types of hardware devices are c ollectively referred to as accelerators or AI accelerators in some Red Hat content. There are several types of hardware devices available for modern servers that can be classified as accelerators, including but not limited to GPUs, tensor processing units (TPUs), application-specific integrated circuit (ASICs), neural processing units (NPUs), and field-programmable gate arrays (FPGAs).

These accelerators are typically a card, board or other physical device occupying a peripheral component interconnect (PCI) slot in a server. This is almost always the unit quantity you have purchased from the accelerator vendor. For example, in a server whose vendor says it has "8 GPUs," this almost always means 8 physical accelerator units.

These accelerators require dedicated core-based subscriptions of the same type, i.e., OpenShift Kubernetes Engine, OpenShift Container Platform, or OpenShift Platform Plus, as the cores used for x86 compute workloads. For example, if the cluster compute nodes are subscribed using OpenShift Container Platform, then additional new AI accelerator card entitlements are needed. This is the same subscription as the core-based entitlements to make the counting of the computational workload more straightforward. As with normal Red Hat subscription practices, the same Red Hat OpenShift edition subscription and support level are required between the AI accelerator card and x86 core entitlements found in the same cluster.

Each AI accelerator subscription covers up to 2 physical accelerator devices. For example, focusing just on the AI accelerator subscriptions:

▸ A physical compute node with 4 physical GPU devices would require 2 x AI accelerator subscriptions in addition to the CPU core or socket subscriptions covering the compute node.

▸ A single virtual compute node with 1 physical GPU device presented to the virtual machines (VMs) as multiple vGPUs would require 1 AI accelerator subscription, as counting is based on physical accelerators, not virtual accelerators. Another physical GPU accelerator could be added at a later time or 1 AI accelerator from a totally separate node for a total of 2 physical accelerators without requiring an additional AI accelerator subscription.

When are AI accelerator subscriptions needed for accelerator devices?

Accelerators are only counted when they are used to execute a compute workload. A workload is considered a compute workload when the primary purpose of the workload is not actively drawing pixels on a user's screen in near real-time or moving data across a network.

This distinction is important because VFX (visual effects) and streaming applications may use GPUs and other accelerator hardware, but the primary purpose in these cases is drawing pixels on a screen. In some cases, the primary function is moving data across a network (such as data processing units dedicated to network functions), which also is not considered compute.

Examples of compute workloads include:

▸ Traditional software applications, such as Java, Python, and Perl.

▸ LLMs or other compute-intensive software.

▸ Data science model training and tuning.

▸ Scientific modeling and physical simulation like protein folding and fluid dynamics.

Hosted control planes and OpenShift on OpenShift Virtualization

Only bare-metal OpenShift nodes hosting these 2 components qualify as infrastructure nodes that are included in self-managed OpenShift subscriptions.

1. Virtual OpenShift compute nodes

2. Hosted control plane(s) workloads

The hosted virtual OpenShift compute nodes require core-based subscriptions and cannot be joined back to the hosting bare-metal cluster.

Example entitlements for an initial self-managed Red Hat OpenShift deployment

To conduct a more thorough sizing exercise to determine how many self-managed OpenShift (Red Hat OpenShift Platform Plus, Red Hat OpenShift Container Platform, or Red Hat OpenShift Kubernetes Engine) or add-on subscriptions you need, use the following questions and examples.

A few basic OpenShift terms are used in these sizing exercises:

▸ **Pod:** The smallest deployable Kubernetes unit in OpenShift. A Kubernetes pod instance could have a single container or multiple containers running as sidecars.

▸ **Application instance:** An application may be a single-pod instance or may be deployed across multiple-pod instances that make up an application service. For example, a highly available Tomcat application service may consist of 2 or more Tomcat pods.

▸ **Compute node:** Instances (VMs or bare-metal hosts) of Red Hat Enterprise Linux or Red Hat Enterprise Linux CoreOS where end-user application pods run. OpenShift environments can have many compute nodes.

▸ **Control plane nodes:** Instances (VMs or bare-metal hosts) of Red Hat Enterprise Linux CoreOS that act as the Kubernetes orchestration or management layer for OpenShift. Control plane nodes are included in self-managed OpenShift subscriptions. See the Red Hat OpenShift control plane and infrastructure nodes section for more details.

▸ **Infrastructure nodes:** Instances (virtual or physical hosts) of Red Hat Enterprise Linux or Red Hat Enterprise Linux CoreOS that are running pods supporting an OpenShift cluster infrastructure or running the HAProxy-based load balancer for ingress traffic. Infrastructure nodes are included in self-managed OpenShift subscriptions. See Red Hat OpenShift control plane and infrastructure node sections below for more details.

▸ **Cluster:** An OpenShift Kubernetes cluster consisting of a control plane and 1 or more compute nodes.

In summary:

▸ Applications are packaged in container images.

▸ Containers are deployed as pods.

▸ Pods run on Kubernetes compute nodes, which are managed by the Kubernetes control plane nodes.

The following example bill of materials provides an extremely flexible, scalable Red Hat OpenShift environment designed to run as VMs and support hundreds of application containers:

▸ **16 x OpenShift Platform Plus, 2-Core Premium subscriptions, including:**

  ▸ Control plane nodes (3 VMs).

  ▸ Additional infrastructure nodes (3 VMs).

  ▸ Compute nodes (16 VMs sized at 2 cores or 4 vCPUs).

  ▸ Multicluster management, advanced observability, and policy compliance.

  ▸ Declarative security and active threat detection and response.

  ▸ Scalable global container registry.

  ▸ Persistent storage for applications and OpenShift infrastructure services.

Optional:

▸ **16 x Red Hat OpenShift Data Foundation Advanced:** Adds enhanced scalability, granular encryption, disaster recovery functionality, data security, and resilient block and file for file, block, and object storage services for workloads deployed inside Red Hat OpenShift as well as OpenShift infrastructure services. This is an optional add-on for customers running stateful applications that require persistent storage, or who want to build and operate a dedicated external storage cluster shared by multiple OpenShift clusters. Red Hat OpenShift Data Foundation Advanced is also available as part of a bundle called *Red Hat OpenShift Platform Plus with Red Hat OpenShift Data Foundation Advanced.*

Red Hat offers many additional application services and runtimes that have their own subscription and consumption models.

How to estimate your entitlement requirements

Red Hat OpenShift subscriptions do not limit application instances. You can run as many application instances in the Red Hat OpenShift environment as the underlying hardware and infrastructure will support. Larger-capacity hardware can run many application instances on a small number of hosts, while smaller-capacity hardware may require many hosts to run many application instances. The primary factor in determining the size of a Red Hat OpenShift environment is how many pods, or application instances, will be running at any given time.

Step 1: Determine standard VM or hardware cores and memory

You may have a standard VM size for application instances or, if you typically deploy on bare metal, a standard server configuration. The following questions will help you more accurately understand your VM and hardware needs. Remember that, in most cases, 2 vCPUs are equivalent to 1 core.

Table 1: VM and hardware sizing questions

Relevant questions	Example answers
What is the memory capacity of the VMs you will use for nodes?	Our VMs have 64GB of memory and 4 vCPUs and hyperthreading is used.
What is the number of vCPUs for the VMs you will use for nodes?	Our bare-metal systems have 128GB of RAM, 2 CPU sockets (64 cores), hyperthreading is used, and a single GPU is present.
Is hyperthreading in use?	
What number of AI accelerators are in use?	

Step 2: Calculate number of application instances needed

Next, determine how many application instances, or pods, you plan to deploy. When sizing the environment, any application component deployed on Red Hat OpenShift—such as a database, front-end static server, or message broker instance—is considered an application instance.

This figure can be an approximation to help you calculate a gross estimate of your Red Hat OpenShift environment size. CPU, memory oversubscription, quotas and limits, and other features can be used to further refine this estimate.

Table 2: Application and instance sizing questions

Relevant questions	Example answers
How many application instances do you anticipate deploying in each Red Hat OpenShift environment?	We have around 1,250 application instances in our development environment and around 250 application instances in production.
What type of applications are they (e.g., language, framework, database)?	We mainly deploy Java but have some Microsoft. NET Core and Ruby applications as well. We also use a lot of MySQL.

Step 3: Determine preferred maximum OpenShift node utilization

We recommend reserving some space in case of increased demand, especially if autoscaling is enabled for workloads. Your preferred utilization will vary on the basis of historical load for the applications that will run on OpenShift.

Table 3: Preferred maximum OpenShift node utilization questions

Relevant questions	Example answers
How much space do I want to reserve for increased demand?	We want to run nodes at a maximum average of 80% of total capacity (leaving 20% in reserve).

Step 4: Determine total memory footprint

Next, calculate the total memory footprint of the deployed applications. If you are considering a completely greenfield environment, memory use data may not be available, but you can use educated approximations—for example, 1GB of memory per Java application instance—to make an estimate.

Table 4: OpenShift memory footprint questions

Relevant questions	Example answers
What is the average memory footprint of applications?	Our application instances use 2GB of memory or less.
	OR
	We typically allocate 2GB for JVM heap.

Step 5: Calculate totals

Finally, determine the number of OpenShift subscriptions needed on the basis of the data gathered in steps 1-4.

▸ Effective per-node memory capacity (GB)
  = Preferred maximum OpenShift node utilization (%) * standard VM or hardware memory

▶ Total memory utilization
  = Application instances * average application memory footprint

▶ Number of nodes required to cover utilization
  = Total memory utilization / standard VM or hardware memory

▶ Total required cores
  = Number of nodes required to cover utilization * standard VM or hardware cores

▶ Effective virtual cores
  = Total required cores / 2

▶ Number of OpenShift Platform Plus subscriptions[1]
  = Total cores / 2 OR
  = Effective virtual cores / 2

Example calculation for virtualized environments

System sizing (from steps 1-5)

▶ Standard number of VM cores = 4 (hyperthreading used, 2 effective virtual cores)

▶ Standard VM memory = 64GB

▶ Preferred maximum node utilization = 80%

▶ Average application memory footprint = 2GB

▶ Number of application instances = 1500

Subscription calculations

▶ Effective node memory capacity
  = 80% preferred maximum node utilization * 64GB standard VM memory
  = 51GB

▶ Total memory utilization
  = 1500 application instances * 2GB average application memory footprint
  = 3000GB

▶ Nodes required to cover utilization
  = 3000GB total memory utilization / 51GB effective node memory capacity
  = 59 nodes

▶ Total cores
  = 59 nodes required * 2 cores per node
  = 118 total cores

▶ Total subscriptions
  = 118 total cores / 2 cores per subscription
  = 59 subscriptions

In this example, fifty-nine 2-core OpenShift Platform Plus 2-core subscriptions would be needed.

---

1 *If hyperthreading is in use, 2 virtual cores count only as 1 core of a subscription. See the section on Cores versus vCPUs and hyperthreading for details on whether to use effective or actual cores in this calculation.*

Note: Red Hat OpenShift supports many features and functions which affect scalability, Pod scheduling, idling, and resource quota/limiting features. The previous calculations are guidelines, and you may be able to tune your actual environment for better resource use or smaller total environment size. OpenShift Platform Plus customers should take into account the needs of the additional software applications (Red Hat Advanced Cluster Management, Red Hat Advanced Cluster Security, and Quay) including storage and compute resources, even though they may not require additional compute subscriptions.

If working with a 3rd-party reseller, refer to their specific terms and agreements for Red Hat products and services.

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

f facebook.com/redhatinc
𝕏 @RedHat
in linkedin.com/company/red-hat

North America	**Europe, Middle East, and Africa**	**Asia Pacific**	**Latin America**
1 888 REDHAT1	00800 7334 2835 europe@redhat.com	+65 6490 4200 apac@redhat.com	+54 11 4329 7300 info-latam@redhat.com

redhat.com
#1434155_1024