



Cisco Intersight Workload Optimizer Target Configuration and User Guide

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS REFERENCED IN THIS DOCUMENTATION ARE SUBJECT TO CHANGE WITHOUT NOTICE. EXCEPT AS MAY OTHERWISE BE AGREED BY CISCO IN WRITING, ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS DOCUMENTATION ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED.

The Cisco End User License Agreement and any product specific license terms govern your use of any Cisco software, including this product documentation, and are located at: <http://www.cisco.com/go/eula>. Cisco product warranty information is available at <http://www.cisco.com/go/warranty>. US Federal Communications Commission Notices are found here <http://www.cisco.com/c/en/us/products/us-fcc-notice.html>.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Any products and features described in this document as in development or available at a future date remain in varying stages of development and will be offered on a when-and if-available basis. Any such product or feature roadmaps are subject to change at the sole discretion of Cisco, and Cisco will have no liability for delay in the delivery or failure to deliver any products or feature roadmap items that may be set forth in this document.

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

The documentation set for this product strives to use bias-free language. For the purposes of this documentation set, bias-free is defined as language that does not imply discrimination based on age, disability, gender, racial identity, ethnic identity, sexual orientation, socioeconomic status, and intersectionality. Exceptions may be present in the documentation due to language that is hardcoded in the user interfaces of the product software, language used based on RFP documentation, or language that is used by a referenced third-party product.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: [www.cisco.com go trademarks](http://www.cisco.com/go/trademarks). Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1721R)

© 2024 Cisco Systems, Inc. All rights reserved.

Contents

| | |
|--|-----|
| Documentation Overview..... | 7 |
| What’s New..... | 8 |
| Feature Updates and Notices..... | 11 |
| Product Overview..... | 13 |
| How Intersight Workload Optimizer Works..... | 13 |
| The Desired State..... | 14 |
| The Market and Virtual Currency..... | 14 |
| Risk Index..... | 15 |
| The Intersight Workload Optimizer Supply Chain..... | 16 |
| Intersight Workload Optimizer Targets..... | 16 |
| Sustainability Features..... | 19 |
| Getting Started..... | 20 |
| The Overview..... | 20 |
| APPLICATION View..... | 20 |
| ON-PREM View..... | 21 |
| CLOUD View..... | 22 |
| Configuring Targets..... | 26 |
| Supply Chain of Entities..... | 30 |
| Working With a Scoped View..... | 31 |
| Scoping the Intersight Workload Optimizer Session..... | 31 |
| Overview Charts..... | 34 |
| Details View..... | 35 |
| Scope Policies..... | 37 |
| List of Entities..... | 39 |
| Navigating With the Supply Chain..... | 40 |
| Viewing Cluster Headroom..... | 41 |
| Target Configuration..... | 42 |
| Cloud Targets..... | 45 |
| Amazon Web Services..... | 45 |
| Google Cloud..... | 58 |
| Microsoft Azure..... | 75 |
| Cloud Native Targets..... | 112 |
| Installing the Intersight Workload Optimizer Kubernetes Collector..... | 113 |
| Container Platform Monitored Resources..... | 118 |
| Container Platform Actions..... | 121 |

| | |
|---|-----|
| Applications and Databases Targets..... | 122 |
| Apache Tomcat..... | 122 |
| IBM WebSphere..... | 125 |
| JVM Application..... | 128 |
| MySQL..... | 130 |
| Oracle..... | 134 |
| SQL Server..... | 137 |
| Compute / Fabric Targets..... | 141 |
| Cisco UCS Manager..... | 142 |
| HPE OneView..... | 144 |
| Application Performance Management (APM)..... | 147 |
| Cisco AppDynamics..... | 148 |
| Dynatrace..... | 153 |
| New Relic..... | 159 |
| Hyperconverged Targets..... | 162 |
| Cisco HyperFlex..... | 163 |
| Nutanix Acropolis..... | 165 |
| Hypervisor Targets..... | 169 |
| Microsoft Hyper-V..... | 169 |
| vCenter Server..... | 176 |
| Orchestrator Targets..... | 182 |
| ServiceNow..... | 182 |
| Storage Targets..... | 183 |
| Dell EMC SC Series..... | 184 |
| EMC VMAX..... | 187 |
| EMC XtremIO..... | 189 |
| EMC ScaleIO..... | 191 |
| EMC VPLEX..... | 193 |
| HPE 3PAR..... | 194 |
| NetApp..... | 198 |
| Pure Storage FlashArray..... | 204 |
| User Interface Reference..... | 207 |
| Entity Types - Applications..... | 207 |
| Business Application..... | 208 |
| Business Transaction..... | 210 |
| Service..... | 213 |
| Application Component..... | 215 |
| Application Topology..... | 219 |
| Entity Types - Container Platform..... | 221 |
| Container Platform Service..... | 222 |

| | |
|--|-----|
| Container..... | 226 |
| Container Spec..... | 228 |
| Workload Controller..... | 235 |
| Container Pod..... | 241 |
| Namespace..... | 246 |
| Container Platform Cluster..... | 250 |
| Virtual Machine (Container Platform Node)..... | 257 |
| Container Platform CPU Metrics..... | 261 |
| Entity Types - Cloud Infrastructure..... | 263 |
| Virtual Machine (Cloud)..... | 264 |
| App Component Spec..... | 286 |
| Virtual Machine Spec..... | 287 |
| Database Server (Cloud)..... | 296 |
| Database (Cloud)..... | 308 |
| Document Collection..... | 322 |
| Volume (Cloud)..... | 326 |
| Zone..... | 332 |
| Region..... | 334 |
| Entity Types - On-prem Infrastructure..... | 335 |
| Virtual Machine (On-prem)..... | 336 |
| Database Server (On-prem)..... | 358 |
| Volume (On-prem)..... | 363 |
| Virtual Data Center (Private Cloud)..... | 364 |
| Host..... | 367 |
| Chassis..... | 374 |
| Data Center..... | 375 |
| Storage..... | 377 |
| Logical Pool..... | 385 |
| Disk Array..... | 387 |
| Storage Controller..... | 391 |
| IO Module..... | 393 |
| Switch..... | 393 |
| Intersight Workload Optimizer Actions..... | 395 |
| Working With Action Center..... | 396 |
| Action Details..... | 401 |
| Actions by Entity Type..... | 405 |
| Action Categories..... | 412 |
| Action Types..... | 413 |
| Action Acceptance Modes..... | 415 |
| Plans: Looking to the Future..... | 417 |

| | |
|---|-----|
| Plan Management..... | 418 |
| Setting Up Plan Scenarios..... | 418 |
| Plan Scenarios and Types..... | 424 |
| Configuring Nightly Plans..... | 492 |
| Placement: Reserve Workload Resources..... | 493 |
| Creating a Reservation..... | 495 |
| Managing Reservations..... | 497 |
| Dashboards: Focused Views..... | 498 |
| Built-in Dashboards..... | 498 |
| Creating and Editing Custom Dashboards..... | 502 |
| Creating and Editing Chart Widgets..... | 504 |
| Chart Types..... | 507 |
| Creating Groups..... | 566 |
| Working With Policies..... | 568 |
| Placement Policies..... | 569 |
| Automation Policies..... | 574 |
| Working With Schedules..... | 584 |
| Managing Calendar Schedules..... | 584 |
| Templates: Resource Allocations for New Entities..... | 587 |
| Creating Templates..... | 587 |
| VM Template Settings..... | 588 |
| Host Template Settings..... | 589 |
| HCI Host Template Settings..... | 591 |
| Storage Template Settings..... | 593 |
| Billing and Costs..... | 594 |
| Reserved Instance Settings..... | 594 |
| Price Adjustments..... | 595 |
| Currency Settings..... | 599 |
| Maintenance Options..... | 600 |
| Intersight Workload Optimizer Data Exports..... | 601 |
| Setting Up Redshift with Kafka..... | 601 |
| Setting Up a VPC, Subnet, and NAT Gateway..... | 602 |
| Provisioning Common Resources..... | 607 |
| Provisioning a New Tenant..... | 608 |
| Managing Data Exports Using Intersight API Docs..... | 610 |
| Viewing Exported Data..... | 611 |



Documentation Overview

Documentation for this product release includes the following general topics.

- [What's New \(on page 8\)](#) - Describes new features and improvements, as well as features that have been, or will be, deprecated/removed
- [Product Overview \(on page 13\)](#) - Provides an overview of the platform and its underlying architecture
- [Getting Started \(on page 20\)](#) - Describes the Overview page, actions, and policies
- [Target Configuration \(on page 42\)](#) - Provides a list of targets that the product can monitor, and describes how to configure each target properly
- [User Interface Reference \(on page 207\)](#) - Provides a list entities discovered from targets, and describes how to configure plans, charts, and administrative settings



What's New

Published on August 1, 2024

This release of Intersight Workload Optimizer includes improvements to existing features, and new features that improve your experience with the platform. We invite you to try them and see how they improve the capabilities of the platform.

■ Cloud Resource Management

– Support for AWS GPU Metrics

Intersight Workload Optimizer now discovers NVIDIA GPU metrics for supported AWS EC2 instance types and uses these metrics to generate VM scale actions.

Metrics include the number of utilized GPU cards and the amount of GPU memory in use. You can view these metrics in the *Capacity and Usage* and *Multiple Resources* charts. To optimize performance and costs, Intersight Workload Optimizer can recommend actions that scale down the number of GPU cards, or scale GPU memory up or down. To collect GPU metrics, be sure to configure CloudWatch as described in this [topic \(on page 55\)](#).

Currently, Intersight Workload Optimizer supports P2, P3, P3dn, G3, G4dn, G5, and G5g instance types with Linux AMIs. To discover these instance types, Intersight Workload Optimizer requires the `ec2:DescribeInstanceTypes` permission.

– Scaling of Standard VM Resources to AWS GPU and Accelerator Instance Types

Intersight Workload Optimizer can now recommend actions to scale standard VM resources (such as vCPU and vMem) to the following AWS EC2 instance types:

- GPU instance types
 - G3 instance family (based on NVIDIA Tesla M60 GPUs)
 - G5 instance family (based on NVIDIA A10G Tensor Core GPUs)
 - G5g instance family (based on NVIDIA T4G Tensor Core GPUs)
- Accelerator instance types
 - Inf1 instance family (based on AWS Inferentia chips)
 - Inf2 instance family (based on AWS Inferentia2 chips)

Intersight Workload Optimizer also creates the appropriate tier exclusion policies.

- Cross-target policies, such as:
 - `AWS GPU - 1 NVIDIA M60, 8 GiB Memory - Cloud Compute Tier Exclusion Policy`
This policy ensures that AWS VMs with certain GPU types only scale to an instance type with the same GPU configuration (card count and memory per card).
 - `AWS ML_ACCELERATOR - 1 Inferential - Cloud Compute Tier Exclusion Policy`
This policy ensures that AWS VMs with certain Accelerator types only scale to an instance type with the same Accelerator configuration (card count and memory per card).

- Per-target policies, such as:
 - `Cloud Compute Tier AWS:gpu - Cloud Compute Tier Exclusion Policy`
This policy ensures that any VMs in GPU supported instance families (G4dn, G4ad, G3, G3s, G5, G5g currently) do not scale out to instance families that do not support GPUs.
 - `Cloud Compute Tier AWS:infl - Cloud Compute Tier Exclusion Policy`
This policy ensures that any VMs in Inferentia1 instance families do not scale out to other instance families.

– **Support for AWS Standard Data Exports (CUR 2.0)**

The AWS Billing target can now retrieve billing data from a standard data export (CUR 2.0) that you set up in the AWS Billing and Cost Management console.

If you previously set up a legacy CUR export for use with the AWS Billing target and now want to switch to the standard data export:

1. Set up a standard data export. For more information, see [Setting Up a Standard Data Export \(CUR 2.0\) \(on page 47\)](#).
2. In the Intersight Workload Optimizer user interface, remove your existing AWS Billing target and then add a new one. In the new target, specify the S3 bucket name, S3 path prefix, and S3 bucket region of the standard data export.

NOTE:

You can continue to use a legacy CUR export until further notice.

– **AWS RDS Data in Discount Charts**

For AWS RDS Reserved Instances (RIs), utilization and coverage data is now available in the Discount Utilization and Discount Coverage charts.

NOTE:

Currently, utilization data is not considered in database server scaling recommendations. In addition, Intersight Workload Optimizer does not generate actions to purchase RDS RIs.

– **Delete Actions for Azure Cosmos Databases**

To help reduce your cloud expenses, Intersight Workload Optimizer can now recommend deleting an Azure Cosmos database with provisioned throughput but without any underlying document collection (container).

This action can be executed in Intersight Workload Optimizer manually or automatically. To execute actions, update the service principal for Intersight Workload Optimizer with the following permissions.

- `Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/delete`
- `Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/delete`
- `Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/delete`
- `Microsoft.DocumentDB/databaseAccounts/sqlDatabases/delete`

Intersight Workload Optimizer tracks savings associated with delete actions in the cloud savings charts.

– **Reconfigure Actions for Azure Cosmos Databases**

Intersight Workload Optimizer can now recommend removing unused provisioned throughput that is assigned to an Azure Cosmos database to help reduce your cloud expenses. To remove this resource, reconfigure the database from Azure.

For more information, see [Reconfigure Actions for Cosmos Databases \(on page 318\)](#).

– **Azure Volume Delete Action Enhancement**

Intersight Workload Optimizer now checks the `DiskState` property of Azure volumes to accurately determine their attachment state. If the `DiskState` is `Unattached`, Intersight Workload Optimizer generates an action to delete the volume.

– **Execution of Scale Actions for Google Cloud Volumes**

Scale actions for Google Cloud volumes can now be executed manually or automatically in Intersight Workload Optimizer. Before this release, scale actions can only be executed in Google Cloud.

To take advantage of this feature, update the role for the Intersight Workload Optimizer service account with the required permissions for executing scale actions. There are 24 new permissions to add. For the full list of permissions, see [Google Cloud Permissions \(on page 69\)](#).

– Enhancements to Cloud Savings and Investments Charts

- The cloud savings and investments charts can now break down data by tag values. These are the tag values that you set up in your cloud environment to categorize resources. With this feature, you can visualize the distribution of savings or investments across tag values for a given tag key. For example, you can visualize savings by teams in your organization.

NOTE:

For AWS, tag keys are limited to cost allocation tags.

- When you click **Show All** in the charts, you can now view and download a table that breaks down savings or investments by resource names.
- The charts now track savings and investments associated with Cosmos DB scale actions (for databases and document collections) and savings associated with delete actions (for databases).

■ On-prem Resource Management

– On-prem Database Server Actions and Policy Settings Enhancements

With this release, on-prem Database server entities and their related commodities (such as DB Memory and DB Cache Hit Rate) use percentile calculations to generate actions. Additionally, the on-prem Database Server policy settings now include Aggressiveness and Observation Periods as resizing sensitivity.

For more information, see [Action Details \(on page 402\)](#) and [On-prem Database Server Policies \(on page 360\)](#).

– Network Merge Placement Policy Enhancements

With this release, you can create groups of networks (static or dynamic) and associate them with a network merge placement policy, thus simplifying policy management.

For more information, see [Creating Placement Policies \(on page 569\)](#).

■ Application Performance Management

– Application Component Actions and Policy Settings Enhancements

With this release, Application Component entities and their related commodities (such as Heap and Garbage Collection) use percentile calculations to generate actions. Additionally, the Application Component policy settings now include Aggressiveness and Observation Periods as scaling constraints.

For more information, see [Action Details \(on page 402\)](#) and [Application Component Policies \(on page 216\)](#).

– New Relic Enhancements

This release introduces the ability to add a target name in the New Relic target page. You can now identify a New Relic target based on a target name instead of the Account ID in the target configuration page. Since this name is for display purposes only, you can customize the target name based on your needs and the name does not need to match any name in New Relic. Give each target a unique name when multiple New Relic targets are added to Intersight Workload Optimizer.

NOTE:

The target name field can only contain alphanumeric, space, or hyphen characters.

For more information, see [New Relic \(on page 159\)](#).

■ Container Platform Management

– Granular Resize Actions for Workload Controllers

You can now control the Workload Controller resize actions that Intersight Workload Optimizer generates and automates based on the specific resources that you want to resize and the resize direction. For example, for CPU limit resizes, you may want to automate resize down actions but require reviews of resize up actions. To enforce these rules, create a Workload Controller policy and then set the action mode for Resize CPU Limit Down to *automated*, and Resize CPU Limit Up to *manual*.

For more information, see [Workload Controller Policies \(on page 241\)](#).

– Container Spec Policy Enhancements

This release introduces the ability to configure container spec policies that specify tolerance levels for vCPU limits and requests, vMem limits and requests, and CPU throttling. With this feature, Intersight Workload Optimizer analysis can now generate more accurate resize actions that respect the tolerance levels that you configured.

To help you configure tolerance levels with ease, the enhanced policy page for container specs now includes individual tabs for vCPU limits and requests, vMem limits and requests, and CPU throttling.

For more information, see [Container Spec Policies \(on page 230\)](#).

– **Kubernetes Collector Updates**

Intersight Workload Optimizer continues to improve the management of resources in a Kubernetes cluster. To take advantage of these improvements, update the collector so it can pass new types of data to Intersight Workload Optimizer and execute any newly added commands.

For more information, see [Updating the Kubernetes Collector \(on page 117\)](#).

■ **User Interface Management**

– **Automation Policy Enhancements**

With this release, Automation and Orchestration settings in automation policies have been simplified and renamed to "Automation Workflows." Additionally, Action Generation and Action Acceptance settings are combined into a single Action Generation setting, with each action stage more clearly represented. You can now select multiple workflows per action stage and specify whether the workflow is critical (action will fail if the workflow fails) or non-critical (action will continue if the workflow fails).

For more information, see "Automation Workflows" in the *User Guide* and [Creating Automation Policies \(on page 577\)](#).

Feature Updates and Notices

Frequent changes to the product or third-party targets require that some features are updated, deprecated, removed, or no longer supported. See the following sections for more information about these features.

Updates

Consider the details and recommended actions that are provided.

| Feature | Status | Details and Recommended Action |
|--|------------------|--|
| Pure Storage targets running Purity 6.4.4 (Pure API 1.6) | Added to version | Intersight Workload Optimizer now supports Pure Storage targets running Purity 6.4.4 (Pure API 1.6). |
| Memory setting in host templates | Updated | Intersight Workload Optimizer now uses GB as the primary unit for setting the Memory capacity for host templates. Also, the Memory default value is now 1024 GB. |
| Azure permissions | Updated | Permissions were updated to more accurately reflect the minimum permissions for discovering workloads and executing actions. For example, permissions with wildcards (*) were replaced with the exact permissions. For details, see Azure Service Principal and Subscription Permissions (on page 86) . |
| Rate of Resize setting in on-prem VM policies | Updated | The Rate of Resize default value has changed from 2 to 3. If you have changed your default setting to 1 or want to keep the current default setting of 2, create a new policy that is scoped to all on-prem VMs and configure the Rate of Resize to your wanted setting. |

Notices

Features are labeled based on the following definitions:

- **Deprecated** – The feature is still supported but no longer developed or enhanced. The feature is not recommended for use and might become obsolete. Cisco might remove it in a subsequent release of the product.
- **Removed** – The feature is no longer available in the product.
- **Unsupported** – The feature is no longer supported in the product.

NOTE:

When a specific release or version of an integration partner technology reaches end-of-life (EOL) or its end of support date, Intersight Workload Optimizer no longer provides support for that version. Intersight Workload Optimizer follows integration partners' official EOL timeline for version support. Targeting a unsupported version, or one that is no longer supported by the vendor, is at your own risk.

Consider the details and recommended actions that are provided.

| Feature | Status | Details and Recommended Action |
|---|--|--|
| Microsoft Enterprise Agreement target | Reached end of support | Remove the target from the user interface and then add the Azure Billing target. For details, see Notice: Microsoft Enterprise Agreement (on page 100) . |
| New Relic REST API key and GraphQL API key | To be removed in a future release | The REST API and GraphQL API keys have been deprecated by New Relic. Intersight Workload Optimizer will remove the two fields by version , and you will must use the User Key field when adding New Relic as a target. |
| Azure unmanaged volumes | No longer discovered or monitored, if unattached | Microsoft recently started deprecating Azure unmanaged volumes (disks). In response, Intersight Workload Optimizer no longer discovers or monitors unmanaged volumes that are not attached to any VM. Intersight Workload Optimizer continues to discover and monitor unmanaged volumes that are attached to VMs, to establish their relationship with VMs in the supply chain. Intersight Workload Optimizer does not recommend actions for these volumes. |
| "Use hypervisor VMEM for Resize" setting in on-prem VM policies | Removed | This setting is replaced by the Collect Virtual Machine Metrics option in the target configuration pages for APM targets. |
| Billing Breakdown and Estimated Cost Breakdown charts | Removed | The charts have been removed from the product and can no longer be added to dashboards. If you have added these charts to your dashboards, delete them immediately and start using the replacement chart, Workload Cost Breakdown. |



Product Overview

Intersight Workload Optimizer is the premier solution for Application Resource Management (ARM), a hierarchical, application-driven approach that continuously analyzes applications' resource needs and generates fully automatable actions to ensure applications always get what they need to perform. It runs 24/7/365 and scales with the largest, most complex environments.

To perform Application Resource Management, Intersight Workload Optimizer represents your environment holistically as a *supply chain* of resource *buyers* and *sellers*, all working together to meet application demand. By empowering buyers (such as VMs) with a budget to seek the resources that applications need to perform, and sellers (such as hosts) to price their available CPU, memory, storage, and other resources based on utilization in real time, Intersight Workload Optimizer keeps your applications in an optimal state.

Intersight Workload Optimizer is a microservices architected platform that runs in your network or in the public cloud. It discovers and monitors your application environment through targets. It then performs analysis, anticipates risks to performance or efficiency, and recommends actions to avoid problems before they occur.

How Intersight Workload Optimizer Works

To keep your infrastructure in the desired state, Intersight Workload Optimizer performs Application Resource Management. This is an ongoing process that solves the problem of assuring application performance while simultaneously achieving the most efficient use of resources and respecting environment constraints to comply to business rules.

This is not a simple problem to solve. Application Resource Management has to consider many different resources and how they are used in relation to each other, and numerous control points for each resource. As you grow your infrastructure, the factors for each decision increase exponentially. Moreover, the environment is constantly changing – to stay in the desired state, you are constantly trying to hit a moving target.

To perform Application Resource Management, Intersight Workload Optimizer models the environment as a *market* made up of *buyers* and *sellers*. These buyers and sellers make up a *supply chain* that represents tiers of entities in your inventory. This supply chain represents the flow of resources from the datacenter, through the physical tiers of your environment, into the virtual tier and out to the cloud. By managing relationships between these buyers and sellers, Intersight Workload Optimizer provides closed-loop management of resources, from the datacenter, through to the application.

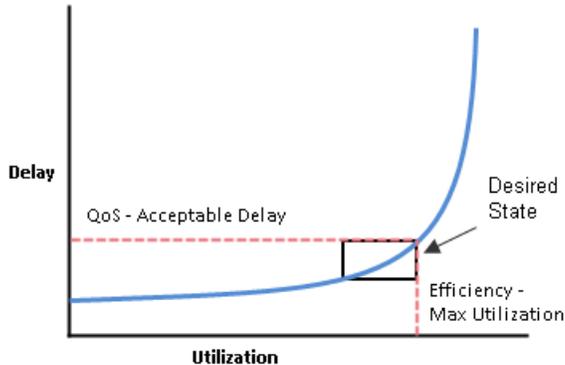
See [Supply Chain of Entities \(on page 30\)](#) for a visual layout of the buyer and seller relationships.

Intersight Workload Optimizer uses *Virtual Currency* to give a budget to buyers and assign cost to resources. This virtual currency assigns value across all tiers of your environment, making it possible to compare the cost of application transactions with the cost of space on a disk or physical space in a data center.

The price that a seller charges for a resource changes according to the seller's supply. As demand increases, prices increase. As prices change, buyers and sellers react. Buyers are free to look for other sellers that offer a better price, and sellers can duplicate themselves (open new storefronts) to meet increasing demand. Intersight Workload Optimizer uses its *Economic Scheduling Engine* to analyze the market and make these decisions. The effect is an invisible hand that dynamically guides your IT infrastructure to the optimal use of resources.

To get the most out of Intersight Workload Optimizer, you should understand how it models your environment, the kind of analysis it performs, and the desired state it works to achieve.

The Desired State



The goal of Application Resource Management is to assure performance while maintaining efficient use of resources. When performance and efficiency are both maintained, the environment is in the desired state. You can measure performance as a function of delay, where zero delay gives the ideal QoS for a given service. Efficient use of resources is a function of utilization where 100% utilization of a resource is the ideal for the most efficient utilization.

If you plot delay and utilization, the result is a curve that shows a correlation between utilization and delay. Up to a point, as you increase utilization, the increase in delay is slight. There comes a point on the curve where a slight increase in utilization results in an unacceptable increase in delay. On the other hand, there is a point in the curve where a reduction in utilization doesn't yield a meaningful increase in QoS. The desired state lies within these points on the curve.

You could set a threshold to post an alert whenever the threshold limit is crossed. In that case, you would never react to a problem until delay has already become unacceptable. To avoid that late reaction you could set the threshold to post an alert before the threshold limit is crossed. In that case, you guarantee QoS at the cost of over-provisioning – you increase operating costs and never achieve efficient utilization.

Instead of responding *after* a threshold is crossed, Intersight Workload Optimizer analyzes the operating conditions and constantly recommends actions to keep the entire environment within the desired state. If you execute these actions (or let Intersight Workload Optimizer execute them for you), the environment will maintain operating conditions that assure performance for your customers, while ensuring the lowest possible cost thanks to efficient utilization of your resources.

The Market and Virtual Currency

To perform Application Resource Management, Intersight Workload Optimizer models the environment as a market, and uses market analysis to manage resource supply and demand. For example, bottlenecks form when local workload demand exceeds the local capacity – in other words, when demand exceeds supply. By modeling the environment as a market, Intersight Workload Optimizer can use economic solutions to efficiently redistribute the demand or increase the supply.

Intersight Workload Optimizer uses two sets of abstraction to model the environment:

- Modeling the physical and virtual IT stack as a service supply chain

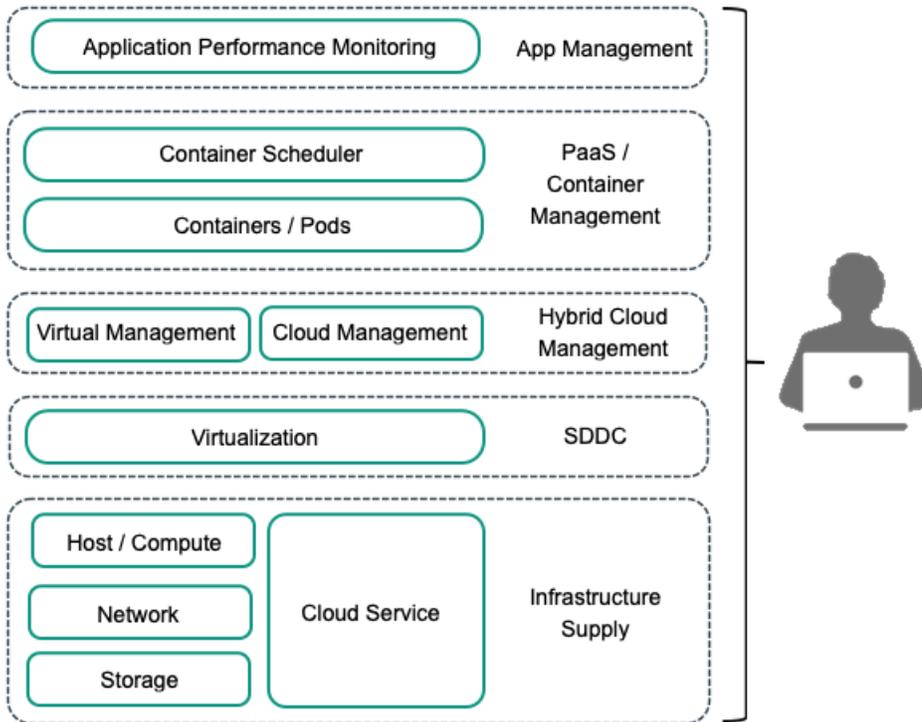
The supply chain models your environment as a set of managed entities. These include applications, VMs, hosts, storage, containers, availability zones (cloud), and data centers. Every entity is a buyer, a seller, or both. A host machine buys physical space, power, and cooling from a data center. The host sells resources such as CPU cycles and memory to VMs. In turn, VMs buy host services, and then sell their resources (VMem and VCPU) to containers, which then sell resources to applications.

See [Supply Chain of Entities \(on page 30\)](#) for a visual layout of the buyer and seller relationships.

- Using virtual currency to represent delay or QoS degradation, and to manage the supply and demand of services along the modeled supply chain

The system uses virtual currency to value these buy/sell transactions. Each managed entity has a running budget – the entity adds to its budget by providing resources to consumers, and the entity draws from its budget to pay for the

resources it consumes. The price of a resource is driven by its utilization – the more demand for a resource, the higher its price.



These abstractions open the whole spectrum of the environment to a single mode of analysis – market analysis. Resources and services can be priced to reflect changes in supply and demand, and pricing can drive resource allocation decisions. For example, a bottleneck (excess demand over supply) results in rising prices for the given resource. Applications competing for the same resource can lower their costs by shifting their workloads to other resource suppliers. As a result, utilization for that resource evens out across the environment and the bottleneck is resolved.

Risk Index

Intersight Workload Optimizer tracks prices for resources in terms of the *Risk Index*. The higher this index for a resource, the more heavily the resource is utilized, the greater the delay for consumers of that resource, and the greater the risk to your QoS. Intersight Workload Optimizer constantly works to keep the Risk Index within acceptable bounds.

You can think of Risk Index as the cost for a resource – Intersight Workload Optimizer works to keep the cost at a competitive level. This is not simply a matter of responding to threshold conditions. Intersight Workload Optimizer analyzes the full range of buyer/seller relationships, and each buyer constantly seeks out the most economical transaction that is available.

This last point is crucial to understanding Intersight Workload Optimizer. The virtual environment is dynamic, with constant changes to workload that correspond with the varying requests your customers make of your applications and services. By examining each buyer/seller relationship, Intersight Workload Optimizer arrives at the optimal workload distribution for the current state of the environment. In this way, it constantly drives your environment toward the desired state.

NOTE:

The default Intersight Workload Optimizer configuration is ready to use in many environments. However, you can fine-tune the configuration to address special services and resources in your environment. Intersight Workload Optimizer provides a full range of policies that you can set to control how the software manages specific groups of entities. Before you make such policy changes, you should understand default Intersight Workload Optimizer operation. For more information about policies, see [Working With Policies \(on page 568\)](#).

The Intersight Workload Optimizer Supply Chain

Intersight Workload Optimizer models your environment as a market of buyers and sellers. It discovers different types of entities in your environment via the targets you have added, and then maps these entities to the supply chain to manage the workloads they support. For example, for a hypervisor target, Intersight Workload Optimizer discovers VMs, the hosts and datastores that provide resources to the VMs, and the applications that use VM resources. The entities in your environment form a chain of supply and demand where some entities provide resources while others consume the supplied resources. Intersight Workload Optimizer *stitches* these entities together.

For information about specific members of the supply chain, see [Supply Chain of Entities \(on page 30\)](#).

Supply Chain Terminology

Cisco introduces specific terms to express IT resources and utilization in terms of supply and demand. These terms are largely intuitive, but you should understand how they relate to the issues and activities that are common for IT management.

| Term | Definition |
|-------------|---|
| Commodity | <p>The basic building block of Intersight Workload Optimizer supply and demand. All the resources that Intersight Workload Optimizer monitors are commodities. For example, the CPU capacity or memory that a host can provide are commodities. Intersight Workload Optimizer can also represent clusters and segments as commodities.</p> <p>When the user interface shows <i>commodities</i>, it's showing the resources a service provides. When the interface shows <i>commodities bought</i>, it's showing what that service consumes.</p> |
| Composed Of | <p>The resources or commodities that make up the given service. For example, in the user interface you might see that a certain VM is <i>composed of</i> commodities such as one or more physical CPUs, an Ethernet interface, and physical memory.</p> <p>Contrast <i>Composed Of</i> with <i>Consumes</i>, where consumption refers to the commodities the VM has bought. Also contrast <i>Composed Of</i> with the commodities a service offers for sale. A host might include four CPUs in its composition, but it offers CPU Cycles as a single commodity.</p> |
| Consumes | <p>The services and commodities a service has bought. A service <i>consumes</i> other commodities. For example, a VM consumes the commodities offered by a host, and an application consumes commodities from one or more VMs. In the user interface you can explore the services that provide the commodities the current service consumes.</p> |
| Entity | <p>A buyer or seller in the market. For example, a VM or a datastore is an entity.</p> |
| Environment | <p>The totality of data center, network, host, storage, VM, and application resources that you are monitoring.</p> |
| Inventory | <p>The list of all entities in your environment.</p> |
| Risk Index | <p>A measure of the risk to Quality of Service (QoS) that a consumer will experience. The higher the Risk Index on a provider, the more risk to QoS for any consumer of that provider's services.</p> <p>For example, a host provides resources to one or more VMs. The higher the Risk Index on the provider, the more likely that the VMs will experience QoS degradation.</p> <p>In most cases, for optimal operation the Risk Index on a provider should not go into double digits.</p> |

Intersight Workload Optimizer Targets

| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|--------------|---------------------|---|----------------------------|
| Cloud | Amazon Web Services | IWO Essentials | No |

| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|-----------------------------------|---|---|----------------------------|
| | Amazon Web Services Billing | IWO Essentials | No |
| | Google Cloud | IWO Essentials | No |
| | Google Cloud Billing | IWO Essentials | No |
| | Microsoft Azure Service Principal | IWO Essentials | No |
| | Microsoft Azure Billing | IWO Essentials | No |
| | Microsoft Azure Enterprise Agreement (deprecated) | IWO Essentials | No |
| Cloud Native | Kubernetes 1.8 or higher (IKS, CCP, Red Hat OpenShift, EKS, AKS, GKE) Deployed on-premises | IWO Advantage | Yes |
| | Kubernetes 1.8 or higher (Red Hat OpenShift, EKS, AKS, GKE) SaaS or deployed on public cloud | IWO Advantage | No |
| Applications and Databases | Apache Tomcat 7.x, 8.x, and 8.5.x Deployed on-premises | IWO Advantage | Yes |
| | IBM WebSphere Application Server 8.5+ Deployed on-premises | IWO Advantage | Yes |
| | JVM 6.0+ Deployed on-premises | IWO Advantage | Yes |
| | SQL Server 2012, 2014, 2016, 2017, and 2019 Deployed on-premises | IWO Advantage | Yes |
| | MySQL Server 8.0 Deployed on-premises | IWO Advantage | Yes |
| | Oracle 19c and 21c Deployed on-premises | IWO Advantage | Yes |
| Compute / Fabric | Cisco UCS Server (Standalone) | IWO Essentials | No |
| | Cisco UCS Domain (UCSM Managed) | IWO Essentials | No |
| | Cisco UCS Domain (Intersight Managed) | IWO Essentials | No |
| | HPE OneView 3.00.04 | IWO Essentials | No |

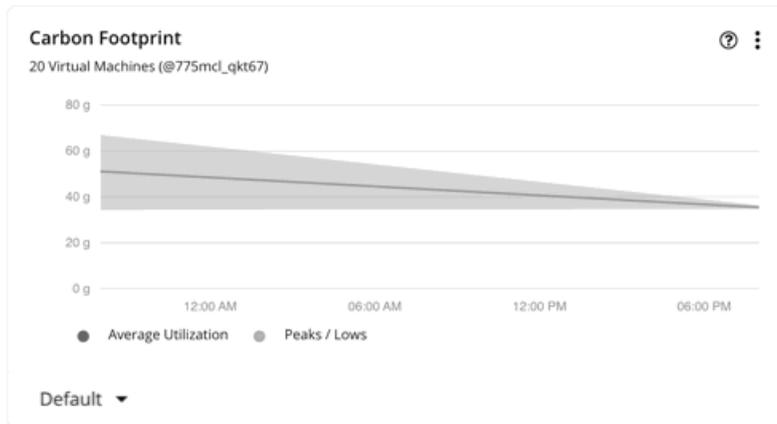
| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|--|---|---|----------------------------|
| Guest OS Process / APM (Application Performance Management) | New Relic <i>SaaS or deployed on public cloud</i> | IWO Premier | No |
| | Cisco AppDynamics 4.1+ <i>Deployed on-premises</i> | IWO Advantage | Yes |
| | Cisco AppDynamics <i>SaaS or deployed on public cloud</i> | IWO Advantage | No |
| | Dynatrace 1.1+ <i>Deployed on-premises</i> | IWO Premier | Yes |
| | Dynatrace <i>SaaS or deployed on public cloud</i> | IWO Premier | No |
| Hyperconverged | Cisco Hyperflex 3.5 | IWO Essentials | No |
| | Nutanix Acropolis | IWO Essentials | Yes |
| Hypervisor | Microsoft Hyper-V 2012 R2, 2016, 2019, 2022 | IWO Essentials | Yes |
| | VMware vCenter 6.0, 6.5, 6.7, and 7.0+ | IWO Essentials | Yes |
| Change Management | ServiceNow | IWO Advantage | No |
| Storage | HPE 3PAR InForm OS 3.2.2+, 3PAR SMI-S, 3PAR WSAPI | IWO Essentials | Yes |
| | Dell EMC SC Series | IWO Essentials | Yes |
| | EMC VMAX using SMI-S 8.1+ | IWO Essentials | Yes |
| | EMC ScaleIO 2.x and 3.x | IWO Essentials | Yes |
| | EMC VPLEX Local Architecture with 1:1 mapping of virtual volumes and LUNs | IWO Essentials | Yes |
| | NetApp ONTAP 8.0+ | IWO Essentials | Yes |
| | Pure Storage FlashArray running Purity 5.3.6 and 6.4.4 (Pure API 1.6) | IWO Essentials | Yes |
| | EMC XtremIO XMS 4.0+ | IWO Essentials | Yes |

Sustainability Features

Intersight Workload Optimizer helps companies reduce their energy consumption and carbon footprint by optimizing IT infrastructure while ensuring that applications get the resources that they need to perform optimally. This core functionality, along with visibility into sustainability data, is crucial to setting and achieving sustainability goals. To visualize this data, Intersight Workload Optimizer provides the following features.

Sustainability Charts

Intersight Workload Optimizer monitors the energy consumption and carbon footprint of your IT infrastructure, and then displays relevant data in charts. Currently, these capabilities are supported for hosts and VMs discovered via vCenter targets. When you set the scope to any of these entities and then click the **Details** tab, you can view sustainability data in the **Energy** and **Carbon Footprint** charts.



For details, see [Energy Chart \(on page 526\)](#) and [Carbon Footprint Chart \(on page 518\)](#).

Data Center Policies

Intersight Workload Optimizer [calculates \(on page 519\)](#) carbon footprint using industry standards that take into account energy consumption, datacenter efficiency, and carbon intensity data. You can create Data Center policies to adjust the calculations according to the requirements of your data centers. For example, a data center in a particular location might have different requirements than data centers in other locations. After you adjust the calculations via policies, Intersight Workload Optimizer can accurately report your organization's carbon footprint.

The screenshot shows the "Configure Data Centers Policy" interface. At the top, there is a back arrow and the title "Configure Data Centers Policy". Below the title, there is a "NAME" field with the value "Datacenter Defaults". Underneath, there is a section titled "OPERATIONAL CONSTRAINTS" with a minus sign icon. This section contains two configuration items: "Carbon Intensity (g/Wh)" with a dropdown menu and a value of "0.25", and "Power Usage Effectiveness" with a dropdown menu and a value of "1.5". Each item has an information icon (i) to its right.

For details, see [Data Center Policies \(on page 377\)](#).



Getting Started

To get started with the platform, open a web browser to your Intersight Workload Optimizer installation. The Intersight Workload Optimizer platform serves the user interface to your browser, where you can log in and get started managing your environment. In this way, you can access the unique capabilities of Intersight Workload Optimizer from any internet connection.

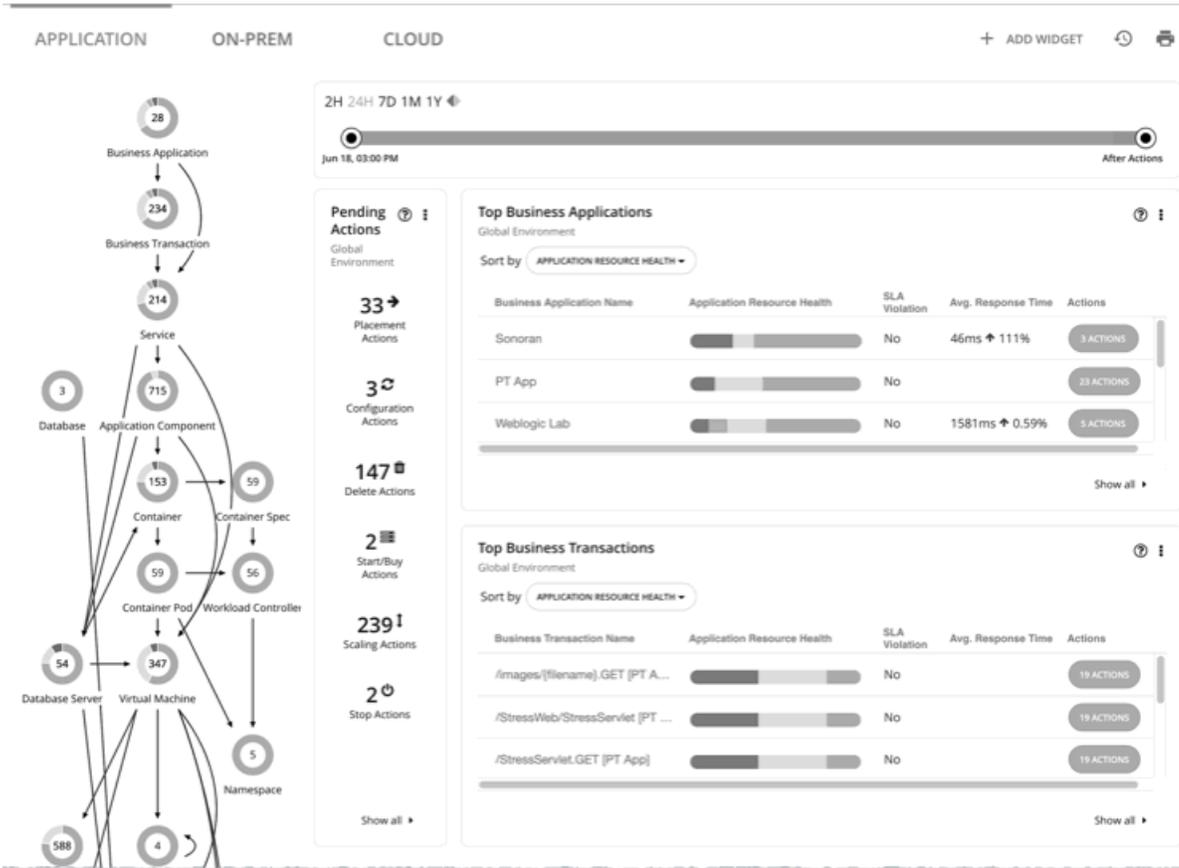
The Overview

In Intersight Workload Optimizer, navigate to **Workload Optimization > Overview**. From there you can:

- Choose a View to see overviews of your environment:
 - APPLICATION – See your environment in the context of your [Business Applications \(on page 208\)](#).
 - ON-PREM – See details for the on-prem environment. Notice that the Supply Chain excludes cloud entities and only shows the entities that are on-prem.
 - CLOUD – See details for the cloud environment, including pending actions, a listing of your cloud accounts by cost, the locations of cloud datacenters that you are using, estimated costs, and other cost-related information.
- Use the Supply Chain Navigator to inspect lists of entities
Click an entity tier in the Supply Chain to see a list of those entities. For example, click Virtual Machine to see a list of all the VMs in your environment.
- Navigate to other Intersight Workload Optimizer pages, including:
 - Search – Set the session scope to drill down to details about your environment
 - Plan – Run what-if scenarios
 - Place – Use Intersight Workload Optimizer to calculate the best placement for workloads, and execute the placement at the time you specify
 - Dashboards – Set up custom views with charts that focus on specifics in your environment
 - Settings – Configure Intersight Workload Optimizer to set up business rules and policies, define groups, and perform other administrative tasks

APPLICATION View

The **APPLICATION** view presents your environment in the context of your [Business Applications \(on page 208\)](#). See the overall health of your applications, examine any performance and compliance risks, and execute the actions that Intersight Workload Optimizer recommends to address these risks.



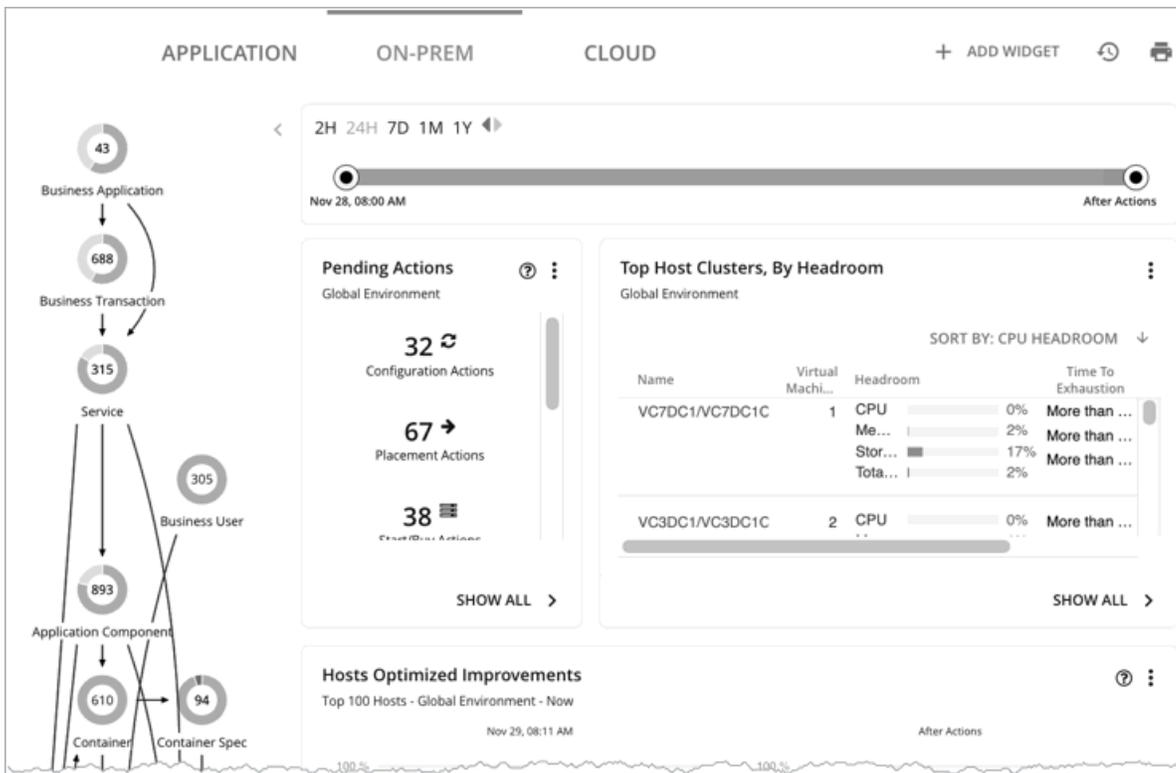
This view also shows the [Business Transactions \(on page 210\)](#) and [Services \(on page 213\)](#) that make up your Business Applications. You can see finer details and set SLOs at these levels of the application model.

NOTE:

If certain application entities do not stitch into the supply chain infrastructure for some reason, Intersight Workload Optimizer displays them in both the ON-PREM and the CLOUD views. Once Intersight Workload Optimizer can stitch them into the infrastructure, it classifies them according to the class of the infrastructure and displays them in the correct views.

ON-PREM View

When you set your session to the Global Scope, you can then select the **ON-PREM** view. This shows an overview of your on-prem environment. If you don't have any workload on the public cloud, then you should use this as your starting point for a Intersight Workload Optimizer session. If you have a hybrid environment (on-prem and on the public cloud), then you can refer to this view to see a detailed on-prem overview.

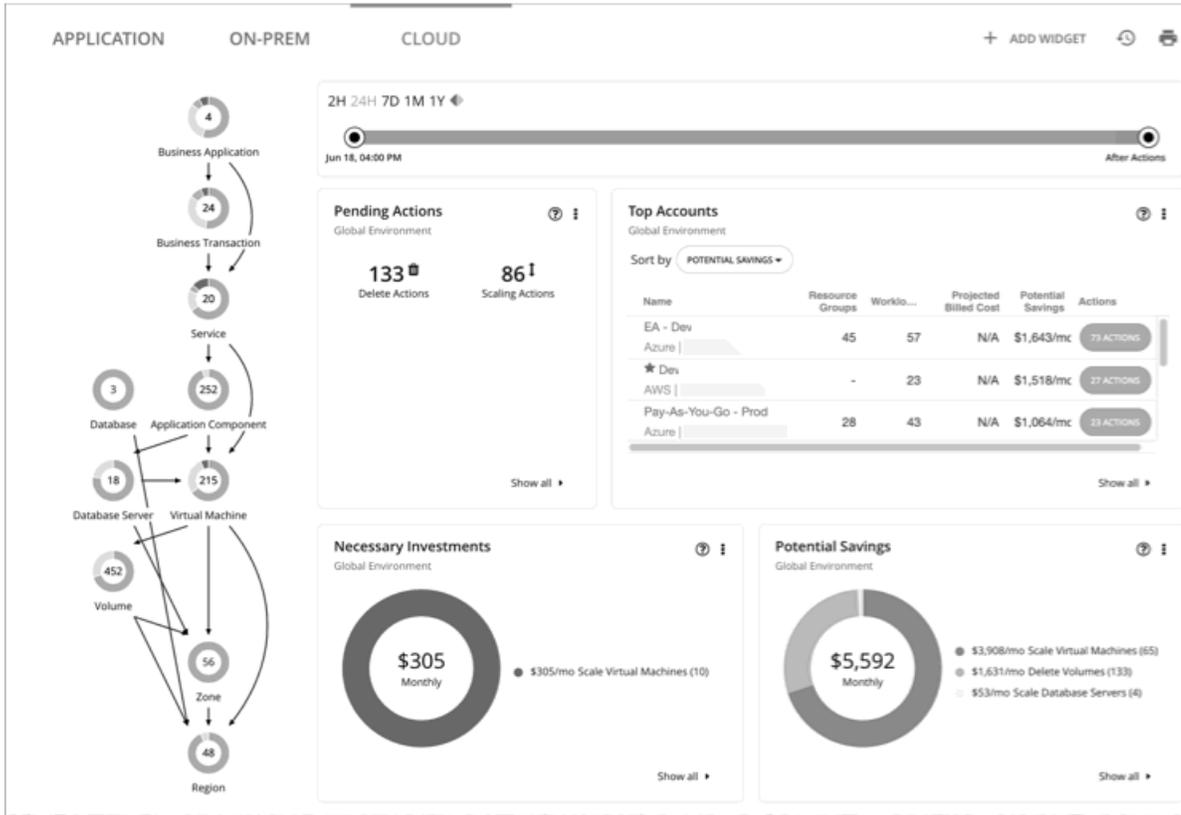


The Supply Chain shows all the on-prem entities in your environment. The charts show details about your environment, including:

- **Overviews of pending actions**
When appropriate, the overview includes estimated one-time savings or costs associated with the actions.
- **Top Host Cluster utilization**
See a list of the most utilized clusters. The chart shows these clusters, along with a count of actions for each. To drill down into the cluster details, click the cluster name. To see and execute the specific actions, click the **ACTIONS** button for that cluster. To see all the clusters in your environment, click **SHOW ALL**.
- **Optimized Improvements**
Compare current resource utilization with the utilization you would see if you choose to execute all the pending actions.
- **Action history**
You can see a history of all actions that have been recommended and executed, or of just the actions that have been accepted and executed.

CLOUD View

When you set your session to the Global Scope, you can then select the **CLOUD** view. This shows an overview of your cloud environment. If all your workload is on the public cloud, then you should use this as your starting point for a Intersight Workload Optimizer session. If you have a hybrid environment (on-prem and on the public cloud), then you can refer to this view to see a detailed cloud overview.



To view cloud cost information, you must have one or more public cloud targets set up in your Intersight Workload Optimizer installation. For information about setting up public cloud targets, see [Cloud Targets \(on page 45\)](#).

In this view, the Supply Chain shows all the cloud entities in your environment. The charts show details about your cloud environment, including:

- **Overviews of pending actions**
The overview includes the estimated monthly savings or cost associated with those actions.
- **Top Accounts utilization**
See a list of the most utilized public cloud accounts. The chart shows these accounts, along with an estimate of the monthly cost for each. To see all the cloud accounts in your environment, click **SHOW ALL**.
- **Necessary Investments and Potential Savings**
For the current set of pending actions, these charts show the impact in dollar value. Necessary Investments are from actions to provision more workloads or to resize workloads up. Potential Savings are from actions to resize down, or to purchase discounts and put them into active use.
- **Charts that show your current discounts.** For details, see [Discounts \(on page 25\)](#).
- **Billed Cost by Service**
This chart shows costs over time for each cloud service that you use in your cloud accounts. For example, you can see the cost for AWS CloudWatch, compared to the cost for AWS S3 storage.

Tracking Cloud Cost

Intersight Workload Optimizer tracks your cloud costs based on the cost information it discovers from targets (for example, accounts, billing reports, and on-demand or discount costs) and [price adjustments \(on page 595\)](#).

NOTE:

It is possible for Intersight Workload Optimizer to report negative amounts. For example, when discounts are larger than costs, the result is a negative amount. Currently, these amounts are not shown directly in cost-focused charts (such as the Expenses charts). To check for any negative amounts, hover on a data point in a chart and then review the data in the tooltip.

Cost for Services

Intersight Workload Optimizer uses the billing reports from your cloud service providers, as they are associated with your cloud targets. Intersight Workload Optimizer parses these reports to get cost breakdowns by service, service provider, Azure Resource Group, and cloud account. You can see cost data in the Expenses charts and Cost Breakdown by Tag charts.

Workload Expenses

Workloads are the VMs running in your environment, or other hosted processes such as database servers and containers. Intersight Workload Optimizer tracks the following expenses for your workloads:

- Compute

For compute expenses Intersight Workload Optimizer uses hourly expense per template as specified in the associated public cloud account.

- Storage

Intersight Workload Optimizer discovers the storage tier that supports a given workload, and uses the tier pricing to calculate storage cost.

- License

For AWS environments, Intersight Workload Optimizer can calculate OS costs. To calculate the OS cost for a VM, Intersight Workload Optimizer subtracts the template cost from the published workload cost. It assumes the difference is the license cost for that workload. If the OS is open source, then there will be no difference, and license cost is zero. Analysis does not consider AWS Marketplace costs.

For Azure environments, Intersight Workload Optimizer can track OS costs for existing VMs. For actions to purchase reservations, Intersight Workload Optimizer does not include the OS cost. Analysis considers the base OS cost, but does not consider additional costs for support or other add-on features that are bundled with the OS. The affected OS types are Ubuntu PRO, SUSE 24/7, and RHEL with HA.

Intersight Workload Optimizer uses this cost information when making scaling decisions, both in real time and in plans. You can see this information in Expenses charts and in the results of Migrate to Cloud plans.

Costs for Dedicated Tenancy on AWS

When you create VMs on AWS, you can specify their tenancy. When you specify Dedicated Tenancy (DT), the VMs you create are Amazon EC2 instances running on hardware that is dedicated to a single customer. To understand DT in the context of Intersight Workload Optimizer, you should consider:

- For AWS, the Intersight Workload Optimizer supply chain shows an Availability Zone as a Host. The supply chain does not indicate whether certain VMs have tenancy dedicated to specific resources in the given availability zone. Also, Intersight Workload Optimizer does not discover or show the costs for dedicated hosting of your workloads.
- Pricing for DT workloads is different than pricing for Shared Tenancy. Intersight Workload Optimizer does not discover that difference, and uses Shared Tenancy cost for the DT workloads. In action descriptions, the listed savings or investments will be based on Shared Tenancy costs.
- Intersight Workload Optimizer discovers the true costs of RIs for DT workloads. However, because the on-demand VM costs are based on Shared Tenancy, Intersight Workload Optimizer can overstate the savings you would get for purchasing and using RI capacity. In most cases, recommendations to purchase RIs will be correct. However, the time to achieve ROI could take longer than action descriptions and charts indicate.
- Some instance types that are valid for Shared Tenancy are not valid for DT. To see which instance types are valid for your DT VMs, consult the AWS documentation or your AWS representative.
- Under some circumstances Intersight Workload Optimizer can recommend changing a workload to a valid instance type for the tenant, even though the current type is already valid. This can happen when the instance type is not included in the Offer File for the tenancy. For example, assume the t3a template family does not support dedicated tenancy. However, assume that the user created a t3a instance with dedicated tenancy in the EC2 console. In that case, Intersight Workload Optimizer will see this as a misconfiguration and recommend changing to a different instance type.

To address these issues, you can create groups that set a scope to your DT workloads. For example, you can use naming conventions, tagging, or other means to identify your DT workloads. Then you can create dynamic groups based on those indicators. With those groups, you can create policies and dashboards that correspond to the differences you see in your DT environment. Use this approach to address issues for:

- Available Instance Types

To resize a workload, Intersight Workload Optimizer generates an action to change that workload to a different instance type. Because Intersight Workload Optimizer does not discover the difference between instance types that are valid for DT and for Shared Tenancy, it can recommend scaling a DT workload to an unavailable instance type. To avoid this, create a policy for the DT group, and exclude the unavailable instance types.

- Displaying Costs

Intersight Workload Optimizer charts show the costs for your environment. If the scope includes Dedicated Tenancy workloads, then the calculated cost will be incomplete. For example, since AWS does not return pricing data for converted RIs (that is, RIs that have been exchanged at least once) that are on *All Upfront* payment plans, Intersight Workload Optimizer does not include such RIs in its calculations of RI utilization or cost.

Use scope to minimize this effect. You can create separate dashboards for your DT and Shared Tenancy workloads.

Discounts

Intersight Workload Optimizer analysis takes advantage of cloud provider discounts to calculate optimal workload placement and to arrive at the best possible costs for your deployments on the cloud. Intersight Workload Optimizer discovers the following discounts:

- AWS EC2 Reserved Instances (RIs)
- AWS Compute Savings Plans
- AWS RDS Reserved DB Instances
- Azure reservations
- Google Cloud committed use discounts

The Cloud View in the Homepage includes the following charts that show discount data:

- [Discount Inventory \(on page 560\)](#)

This chart lists the cloud provider discounts discovered in your environment.

- [Discount Utilization \(on page 562\)](#)

This chart shows how well you have utilized your current discount [inventory \(on page 560\)](#). The desired goal is to maximize the utilization of your inventory and thus take full advantage of the discounted pricing offered by your cloud provider.

- [Discount Coverage \(on page 557\)](#)

This chart shows the percentage of cloud workloads (VMs and RDS database servers) covered by discounts. For VMs covered by discounts, you can reduce your costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

- [Recommended RI Purchases \(on page 555\)](#)

Intersight Workload Optimizer can recommend purchasing instance types at a discounted rate to help you increase the percentage of VMs covered by discounted pricing and reduce on-demand costs. This chart shows your pending purchases. Download the list of purchases and then send it your cloud provider or representative to initiate the purchase process.

NOTE:

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Currently, Intersight Workload Optimizer can recommend purchasing AWS EC2 RIs and Azure reservations.

Configuring Targets

A target is a service that performs management in your virtual environment. Intersight Workload Optimizer uses targets to monitor workload and to execute actions in your environment. When you configure a target, you specify the address of the service, and the credentials to connect as a client to it.

For each target, Intersight Workload Optimizer communicates with the service via the management protocol that it exposes – The REST API, SMI-S, XML, or some other management transport. Intersight Workload Optimizer uses this communication to discover the managed entities, monitor resource utilization, and execute actions.

To configure a target, you will choose the target type, specify the target's address or key, and then provide credentials to access the target. Intersight Workload Optimizer then discovers and validates the target, and then updates the supply chain with the entities that the target manages.

NOTE:

Intersight Workload Optimizer regularly checks the status of your targets. If target discovery or validation fails, the Target Configuration page updates the status. Under some circumstances, the target can become discoverable or valid again, but the status does not update. In this case, select the target and then click **Rediscover** or **Validate**.

For a list of supported targets and configuration requirements, see [Target Configuration \(on page 42\)](#).

Configuring a Target

1. Navigate to the Settings Page.



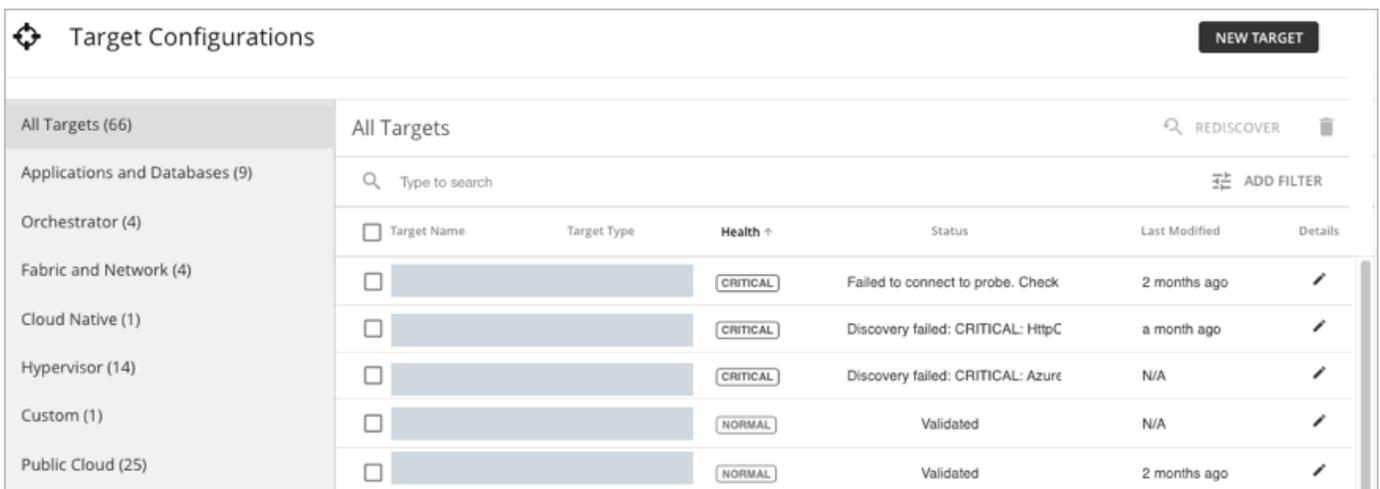
Click to navigate to the Settings Page. From there, you can perform a variety of Intersight Workload Optimizer configuration tasks.

2. Choose Target Configuration.



Click to navigate to the Target Configuration Page.

3. Review the list of targets.



| Target Name | Target Type | Health | Status | Last Modified | Details |
|-------------|-------------|----------|-----------------------------------|---------------|---------|
| [Redacted] | [Redacted] | CRITICAL | Failed to connect to probe. Check | 2 months ago | [Edit] |
| [Redacted] | [Redacted] | CRITICAL | Discovery failed: CRITICAL: HttpC | a month ago | [Edit] |
| [Redacted] | [Redacted] | CRITICAL | Discovery failed: CRITICAL: Azure | N/A | [Edit] |
| [Redacted] | [Redacted] | NORMAL | Validated | N/A | [Edit] |
| [Redacted] | [Redacted] | NORMAL | Validated | 2 months ago | [Edit] |

This page lists all the targets that you currently have configured for Intersight Workload Optimizer in a sortable table. The table is sorted by target health by default. You can inspect or edit these targets, or add a new target.

4. Filter the list of targets.

The screenshot shows the 'Target Configurations' page. On the left is a sidebar with target categories: All Targets (66), Applications and Databases (9), Orchestrator (4), Fabric and Network (4), Cloud Native (1), Hypervisor (14), Custom (1), and Public Cloud (25). The main area is titled 'All Targets' and contains a search bar, a 'REDISCOVER' button, and an 'ADD FILTER' button. Below these is a table with columns: Target Name, Target Type, Health, Status, Last Modified, and Details. The table lists five targets with their respective health and status.

| Target Name | Target Type | Health | Status | Last Modified | Details |
|-------------|-------------|----------|-----------------------------------|---------------|---------|
| [Redacted] | [Redacted] | CRITICAL | Failed to connect to probe. Check | 2 months ago | [Edit] |
| [Redacted] | [Redacted] | CRITICAL | Discovery failed: CRITICAL: HttpC | a month ago | [Edit] |
| [Redacted] | [Redacted] | CRITICAL | Discovery failed: CRITICAL: Azure | N/A | [Edit] |
| [Redacted] | [Redacted] | NORMAL | Validated | N/A | [Edit] |
| [Redacted] | [Redacted] | NORMAL | Validated | 2 months ago | [Edit] |

For a long list of targets, you can:

- Filter targets by target type.
- Use Search to filter targets by Target Name using a text string (partial matching is supported).
- Use Filter to filter targets by status (for example, only show validated targets). You can also use Filter to filter targets by target type or health.

5. Select one or more targets to work with.

This screenshot is identical to the one above, showing the 'Target Configurations' page with the same sidebar, search, filter, and table elements.

When you select a target you can:

- Rediscover
 - Direct Intersight Workload Optimizer to fully discover the entities that this target manages. This will rebuild the topology that is associated with this target.
- Delete
 - When you delete a target, Intersight Workload Optimizer removes all the associated entities from the supply chain.

6. View or edit the target details by clicking the icon under the **Details** column.

Target Configurations NEW TARGET

All Targets (66) REDISCOVER

Applications and Databases (9) Type to search ADD FILTER

Orchestrator (4)

Fabric and Network (4)

Cloud Native (1)

Hypervisor (14)

Custom (1)

Public Cloud (25)

| Target Name | Target Type | Health ↑ | Status | Last Modified | Details |
|--------------------------|-------------|----------|-----------------------------------|---------------|---------|
| <input type="checkbox"/> | [REDACTED] | CRITICAL | Failed to connect to probe. Check | 2 months ago | |
| <input type="checkbox"/> | [REDACTED] | CRITICAL | Discovery failed: CRITICAL: HttpC | a month ago | |
| <input type="checkbox"/> | [REDACTED] | CRITICAL | Discovery failed: CRITICAL: Azure | N/A | |
| <input type="checkbox"/> | [REDACTED] | NORMAL | Validated | N/A | |
| <input type="checkbox"/> | [REDACTED] | NORMAL | Validated | 2 months ago | |

Use the target details page to view or edit target details. In addition to target details, this page shows the last discovered and last modified dates, target health, state, and any related targets.

NOTE:

A value of "N/A" in the Last Modified column indicates that the target is a derived target, and therefore the details cannot be edited.

AWS Billing [?] [X]

NORMAL

Last Discovered: 5/12/2023, 09:48 | Last Modified By: administrator, 2 months ago

CONFIGURATION

CUSTOM TARGET NAME *

IAM Role

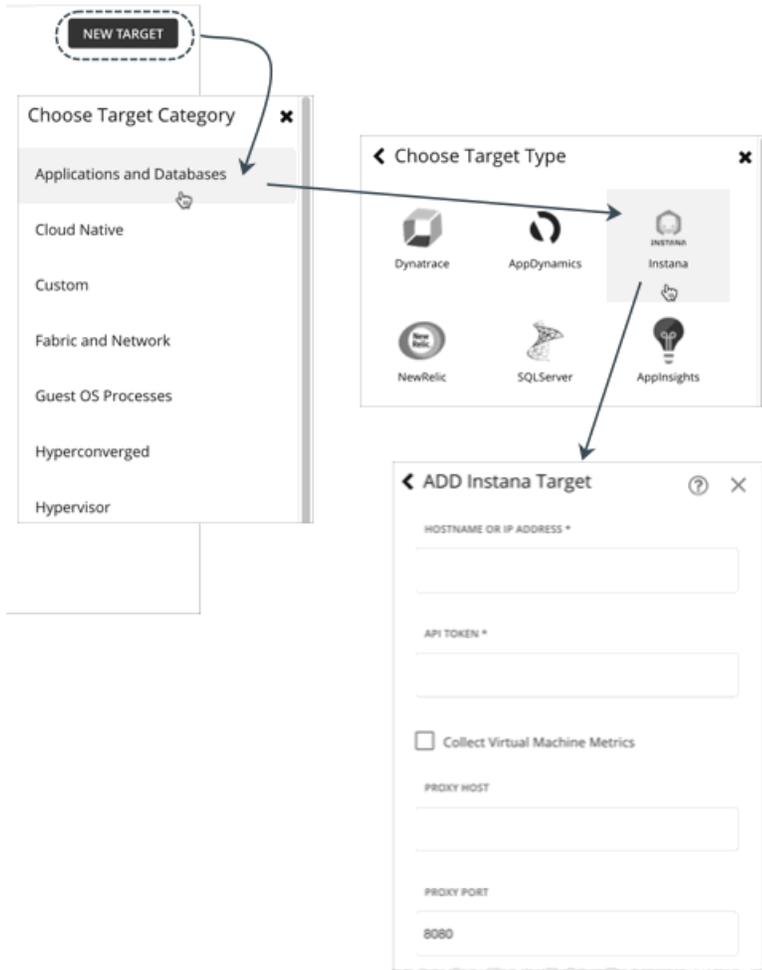
ACCESS KEY *

CURRENT STATE | RELATED TARGETS

Status Description
Validated

Stages
This target does not implement stages yet.

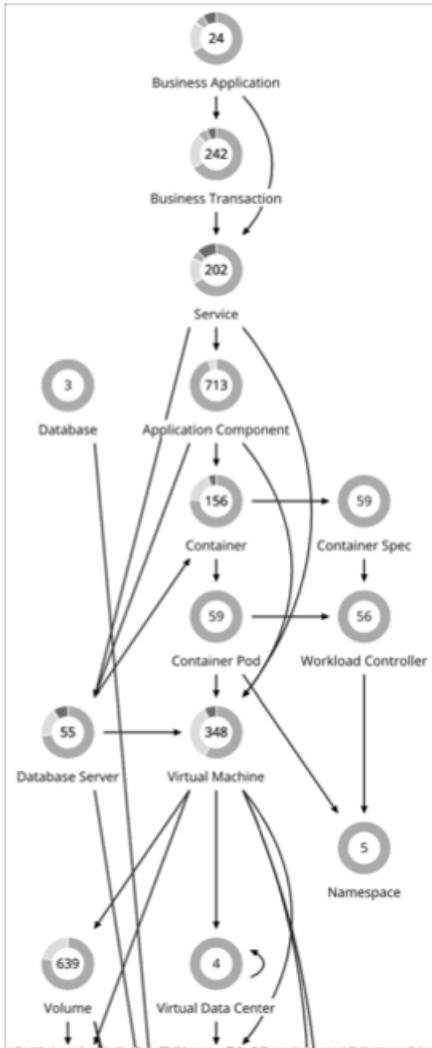
7. Create a new target and add it to Intersight Workload Optimizer.



Click **New Target**, select the target category and type, and then provide the address and credentials for that target. After you add the target, the Target Configuration page refreshes to show the current validation status.

- **Validating**
Validation is in progress.
- **Validated**
Validation was successful. Intersight Workload Optimizer can now monitor the target and will start discovering the entities that the target manages.
- **Validation Failed**
Validation was unsuccessful. Expand the target to see additional information.

Supply Chain of Entities



To perform Application Resource Management, Intersight Workload Optimizer models your environment as a market of buyers and sellers linked together in a supply chain. This supply chain represents the flow of resources from the datacenter, through the physical tiers of your environment, into the virtual tier and out to the cloud. By managing relationships between these buyers and sellers, Intersight Workload Optimizer provides closed-loop management of resources, from the datacenter, through to the application.

Reading the Supply Chain

By looking at the Supply Chain, you can see:

- How many entities you have on each tier
Each entry in the supply chain gives a count of entities for the given type.
- The overall health of entities in each tier
The ring for each entry indicates the percentage of pending actions for that tier in the datacenter. Ring colors indicate how critical the actions are - Green shows the percentage of entities that have no actions pending. To get actual counts of pending actions, hover on a ring to more details.

- The flow of resources between tiers

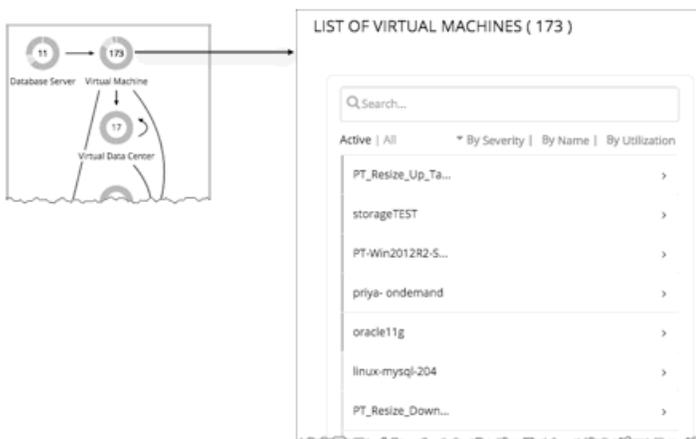
The arrow from one entry to another indicates the flow of resources. For example, the Virtual Machine entry has arrows to Hosts and to Storage. If the VMs are running in a Virtual Data Center, it will have another arrow to that as well. This means that your VMs consume resources from hosts, storage, and possibly from VDCs.

Listing Entities From the Overview

The Supply Chain shows the relationships of entities in your environment. When you're on the **Overview** with a global scope, the supply chain filters its display according to the view you have chosen:

- APPLICATIONS – All your [Business Applications \(on page 208\)](#)
- ON-PREM – All your on-prem entities
- CLOUD – All your entities on the public cloud

To see a list of entities, click an entity tier in the Supply Chain.



Working With a Scoped View

By default, the **Overview** shows a Global view of your environment. To drill down into specifics of your environment, you can set a scope to your Intersight Workload Optimizer session. A scoped view shows details about the specific entities in that scope.

Once you have set a scope, you can use the Supply Chain to zoom in on a related tier to see details about the entities on that tier.

If you find the current scope to be useful, you can save it as a named group. Using named groups is an easy way to return to different scopes that you have saved.

Scoping the Intersight Workload Optimizer Session

NOTE:

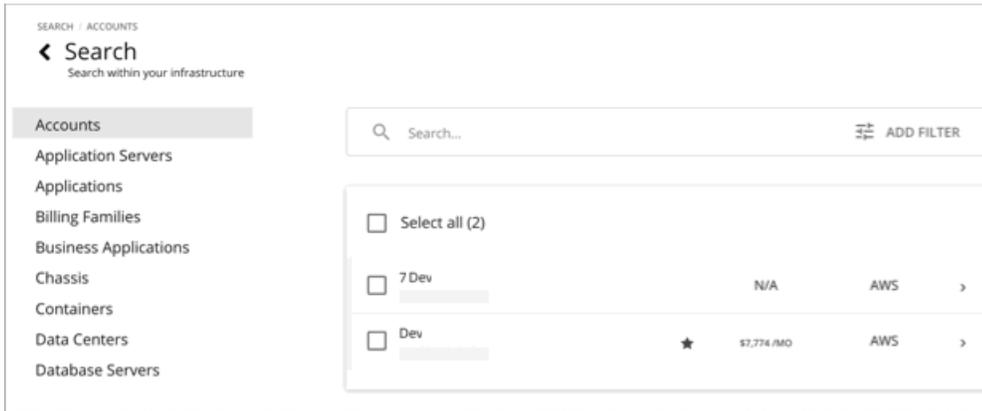
This page includes enhancements and a more modern look-and-feel that are only available when you enable the new design framework. To switch to the new framework, click the React icon  in the navigation bar of the user interface and then Turn ON the toggle. For more information, see "Design Framework for the User Interface" in the *User Guide*.

The default scope for the **Overview** shows an overview of the global environment. What if you want to focus on less than the global environment? Assume you are responsible for a subset of workloads in your environment. This could be:

- Workloads managed on a single host cluster
- The workloads in a single datacenter
- A custom group of workloads you have created in Intersight Workload Optimizer

It's easy to set the session scope so that Intersight Workload Optimizer zooms in on the part of the environment that you want to inspect. Once you set the scope, you can get a quick picture of system health for that scope. If you find a certain scope to be useful, you can save it as a named group that you can return to later.

1. Navigate to the Search Page.
Click **More**, then display the Search Page. This is where you can choose the scope you want.
2. Choose the type of entities to search.



In the Search Page, choose a category that you want to search through. You can focus on entities by type, by groups, or by clusters. When you select an entity type, the page updates to show all entities of that type.

3. Use **Search** to filter the listing.

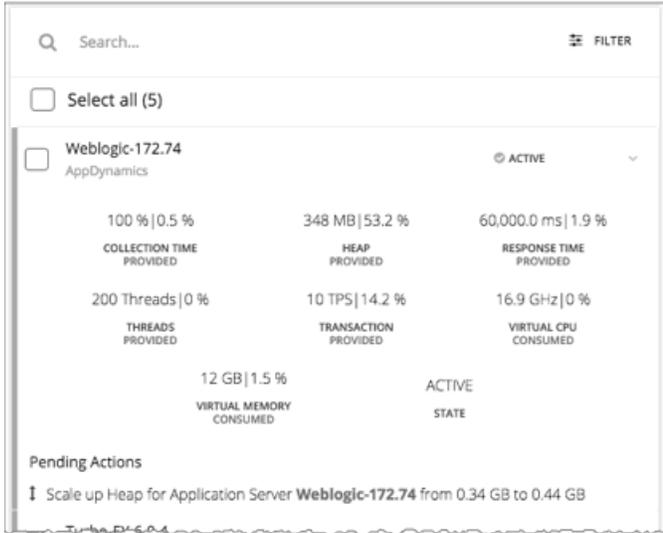
For example, if you're showing **Business Applications** and you search for "Dev", then you will see all Business Applications with "Dev" in their names.



4. Expand an entry to see details.
For example, expand a group or an entity to see utilization details and pending actions.

NOTE:

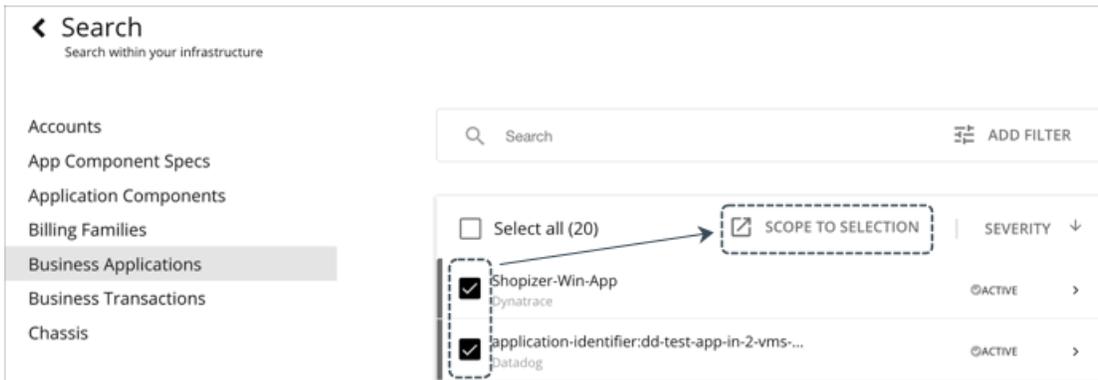
For hosts in the public cloud, utilization and capacity for host and datacenter resources don't affect Intersight Workload Optimizer calculations. When you expand an entry for a public cloud host, the details do not include information for these resources.



5. Select one or more entries to set the focus of the **Overview**.

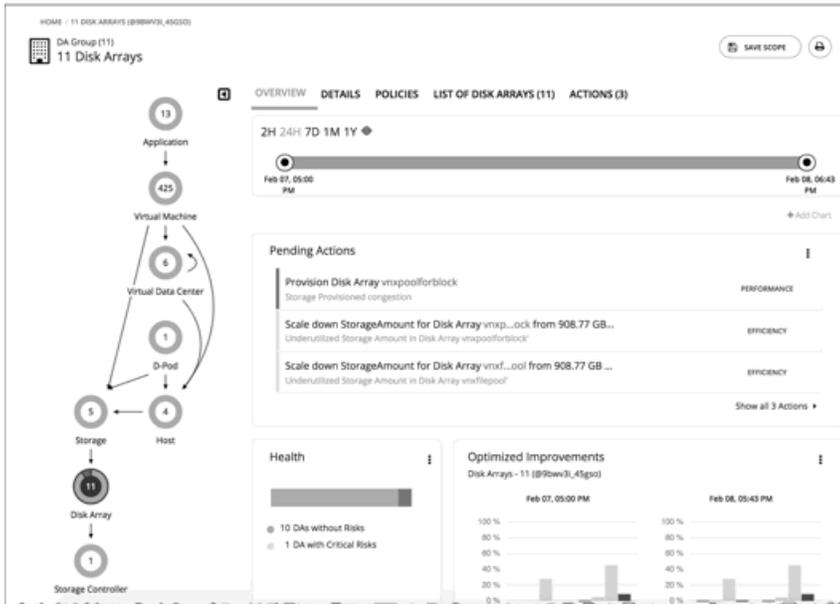
- Choose an entity type, and set the scope to one or more of those entities.
- For different types of groups, click to set a single group as your scope.

If you choose a category of entities to limit the list, then you can select one or more of the entities for your session scope. After you select the entities you want to include in your scope, click **SCOPE TO SELECTION** to set the session scope to those entities.



If you choose groups or clusters, then you can select a single entry to set the scope for your session. When you select an entry in the list, that sets the focus of the **Overview**. For example, if you select a host cluster in the **Search** listing, you set the **Overview** focus to that cluster. Use the **Overview** bread crumbs to set a different scope, or you can return to **Search** and set a different scope from there.

Overview Charts



The Overview Charts show your environment's overall operating health for the current session scope. A glance at the Overview gives you insights into service performance health, overall efficiency of your workload distribution, projections into the future, and trends over time.

The charts in this view show data for the current scope that you have set for the Intersight Workload Optimizer session. For the global scope, the charts roll up average, minimum, and peak values for the whole environment. When you reduce the scope (for example, set the scope to a cluster), the charts show values for the entities in that scope.

Some charts included in this view are:

- **Pending Actions**
See all the actions that are pending for the current scope.
- **Health**
Quickly see the health of the entities in this scope- How many entities have risks, and how critical the risks are.
- **Optimized Improvements**
A comparison of utilization in your environment before executing the pending actions, and then after.
- **Capacity and Usage**
This chart lists resources that are used by the current scope of entities, showing utilization as a percentage of the capacity that is currently in use.
- **Multiple Resources**
See the utilization over time of various resources that are used by the current scope of entities.
- **Top Entities**
For example, Top Virtual Machines. These charts list the top consumer entities in the current scope.
- **Risks Avoided**
Each action addresses one or more identified risks or opportunities in your environment. This chart shows how many risks have been addressed by the executed actions.
- **Accepted Actions**
This chart shows how many actions have been executed or ignored, and whether they have been executed manually or automatically.

What You Can Do:

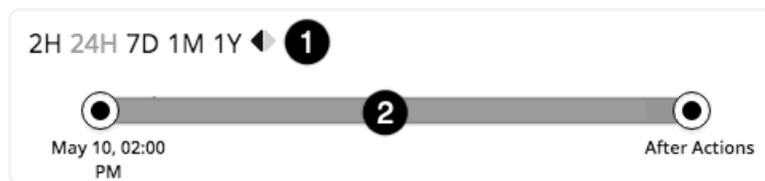
- Set scope: See [Scoping the Intersight Workload Optimizer Session \(on page 31\)](#)
- Create new charts: See [Creating and Editing Chart Widgets \(on page 504\)](#)

Setting Chart Focus

The charts update to reflect the focus that you have set for your viewing session. While viewing the Overview Charts, you can set the focus in different ways:

- Set Supply Chain Focus
Choose a tier in the supply chain to set the view focus - see [Navigating With the Supply Chain \(on page 40\)](#)
- Set Scope
Use **Search** to set the scope of the viewing session - see [Scoping the Intersight Workload Optimizer Session \(on page 31\)](#)

Chart Time Frame



You can set a time frame from recent hours to the past year (1), and set that to the charts in the view. Use the Time Slider to set specific start and end times within that range (2). The green section in the slider shows that you can set the time range to include a projection into the future. For this part of the time range, charts show the results you would see after you execute the current set of pending actions.

For most charts, you can also configure the chart to hard-code the time range. In that case, the chart always shows the same time scale, no matter what scale and range you set for the given view.

Note that Intersight Workload Optimizer stores historical data in its database. As you run Intersight Workload Optimizer in your environment for more time, then you can set a time range to show more history.

Details View

The Details View shows more details about the entities in your session scope. These charts focus on the utilization of resources by these entities, so you can get a sense of activity in that scope over time.



What You Can Do:

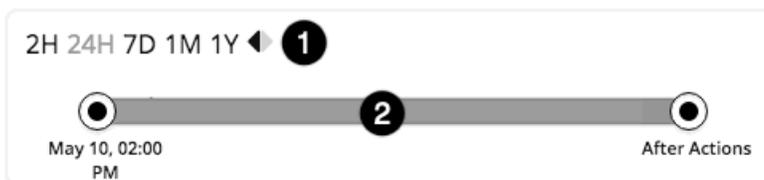
- Set scope: See [Scoping the Intersight Workload Optimizer Session \(on page 31\)](#)
- Create new charts: See [Creating and Editing Chart Widgets \(on page 504\)](#)

Setting Chart Focus

The charts update to reflect the focus that you have set for your viewing session. While viewing the Overview Charts, you can set the focus in different ways:

- Set Supply Chain Focus
Choose a tier in the supply chain to set the view focus - see [Navigating With the Supply Chain \(on page 40\)](#)
- Set Scope
Use **Search** to set the scope of the viewing session - see [Scoping the Intersight Workload Optimizer Session \(on page 31\)](#)

Chart Time Frame

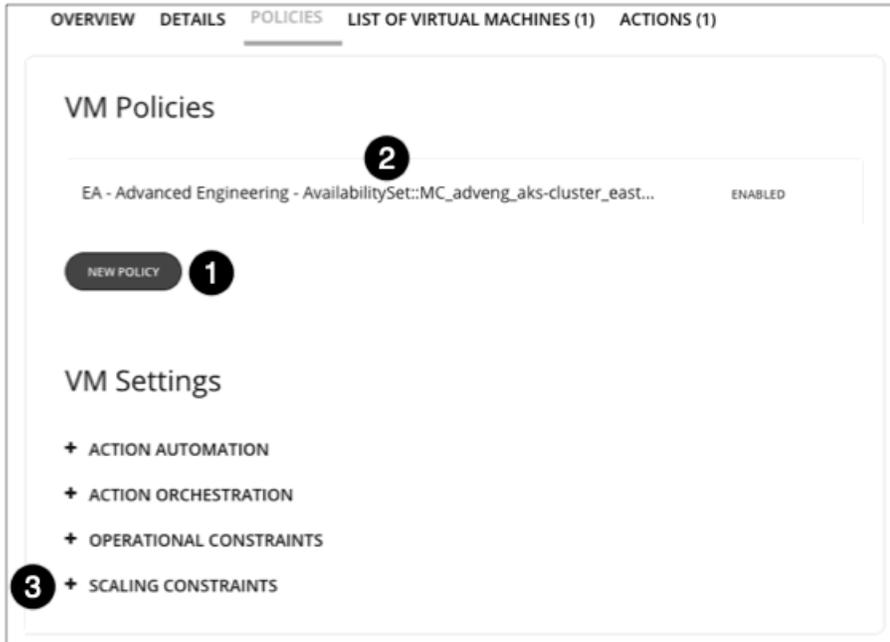


You can set a time frame from recent hours to the past year (1), and set that to the charts in the view. Use the Time Slider to set specific start and end times within that range (2). The green section in the slider shows that you can set the time range to include a projection into the future. For this part of the time range, charts show the results you would see after you execute the current set of pending actions.

For most charts, you can also configure the chart to hard-code the time range. In that case, the chart always shows the same time scale, no matter what scale and range you set for the given view.

Note that Intersight Workload Optimizer stores historical data in its database. As you run Intersight Workload Optimizer in your environment for more time, then you can set a time range to show more history.

Scope Policies



The Policy View gives you a look at the Automation Policies that are set for the entities in the current scope. For each policy, you can see whether it is enabled. In addition,

1. To create a policy, click **New Policy**. When you create a new policy, it automatically includes the current scope. You can add other groups to the policy scope.

NOTE:

You can enable more than one policy for the same scope. If two policies apply different values for the same setting, then the most conservative value takes effect.

2. To edit a policy, click the policy name. You can then change the policy settings and whether the policy is enabled.
3. To see the current policy settings, expand a settings category. For each setting, you can see which policy determines the value - either the default policy or a custom policy that is applied to this scope.

For more information, see [Automation Policies \(on page 574\)](#).

Entity Placement Constraints



When you drill down to a single entity, you can see details about the entity's relationships in the supply chain. This shows you which entities provide resources to this entity. When considering providers for this entity, you can see the name of each current provider, and how many alternative providers Intersight Workload Optimizer can choose from if the current one becomes overutilized.

Reviewing the constraints on an entity helps you understand the actions that Intersight Workload Optimizer recommends. If an action seems questionable to you, then you should look at the constraints on the affected entities. It's possible that some policy or constraint is in effect, and it keeps Intersight Workload Optimizer from recommending a more obvious action.

Experimenting With Placement Constraints

For each provider or consumer in the list, you can open a **Constraints** fly-out that gives more details about limits on the current element's supply chain relationships.

For example, assume the **PROVIDERS** list shows your VM's **CURRENT PLACEMENT** is on Host A, and for **OTHER POTENTIAL PLACEMENT** you see that Intersight Workload Optimizer can choose from 4 hosts. When you click **Constraints**, the flyout displays a list of host constraints that currently result in the four potential hosts (1) for this VM.

Host Constraints For "Oracle11g-Win-172.32" ✕

When you add constraints, you limit the placement decisions Turbonomic can make for your VM. Remove unnecessary constraints so Turbonomic can discover more placement options.

| <input type="checkbox"/> | CONSTRAINT TYPE | SCOPE NAME | SOURCE | POTENTIAL HOSTS |
|--------------------------|--------------------------|---------------------------------------|------------|-----------------|
| <input type="checkbox"/> | Cluster boundaries ⓘ | ACM/VACM Cluster | vCenter | 4 Hosts |
| <input type="checkbox"/> | Datacenter boundaries ⓘ | ACM | vCenter | 4 Hosts |
| <input type="checkbox"/> | Datastore Commodity ⓘ | QS4/ACM | vCenter | 4 Hosts |
| <input type="checkbox"/> | Network boundaries ⓘ | NetworkCommodity/Oracle11g-Win-172.32 | Turbonomic | 12 Hosts |
| <input type="checkbox"/> | Segmentation Commodity ⓘ | My Placement Policy | Turbonomic | 16 Hosts |
| <input type="checkbox"/> | LicenseAccessCommodity ⓘ | Linux | Turbonomic | 70 Hosts |

POTENTIAL HOSTS: 4 **1**
2 FIND MORE PLACEMENT OPTIONS

The list information includes:

- **CONSTRAINT TYPE**
Most constraints are boundaries that are inherent in your environment such as a cluster boundaries or a networks, or the can be constraint rules such as discovered HA or DRS rules authored Intersight Workload Optimizer placement policies (sometimes called *segments*)
- **SCOPE NAME**
For a given rule or constraint, the scope to which it was applied.
- **SOURCE**
If this is a discovered constraint, the source shows the type of target that imposes this constraint. For example, for a DRS rule the source will be vCenter.
- **POTENTIAL HOSTS**
For the given constraint, how many hosts that constraint allows. To see a list of the potential hosts, click the **POTENTIAL HOSTS** value.

To dig deeper into how these constraints affect your entity, click **FIND MORE PLACEMENT OPTIONS** (2). This puts you into a simulation mode that you can use to experiment with changing the effective constraints. For example, you might see that a cluster boundary is limiting your placement possibilities, and you would like the option to place the current VM on other clusters. Armed with this information, you could navigate to Policies and create a Merge Cluster policy.

The screenshot shows a configuration interface for constraints. On the left, a table lists various constraint types with toggle switches and their corresponding potential host counts. A 'POTENTIAL HOSTS: 12' label is highlighted with a '2'. On the right, a 'Related Entities' window displays a list of VMs with their status and host information.

| CONSTRAINT TYPE | SCOPE NAME | SOURCE | POTENTIAL HOSTS |
|------------------------|---------------------------------------|------------|-----------------|
| Cluster boundaries | ACM/ACM Cluster | vCenter | 4 Hosts |
| Datacenter boundaries | ACM | vCenter | 4 Hosts |
| Datastore Commodity | Q54-ACM | vCenter | 4 Hosts |
| Network boundaries | NetworkCommodity/Oracle11g-Win-172.32 | Turbonomic | 12 Hosts |
| Segmentation Commodity | My Placement Policy | Turbonomic | 16 Hosts |
| LicenseAccessCommodity | Linux | Turbonomic | 70 Hosts |

| Entity Name | Status |
|------------------------------|----------|
| dc17-host-03.eng.vmturbo.com | ACTIVE |
| hp-esx4.eng.vmturbo.com | FALLOVER |
| hp-esx7.eng.vmturbo.com | ACTIVE |
| hp-esx8.eng.vmturbo.com | ACTIVE |
| dc17-host-01.eng.vmturbo.com | ACTIVE |

In this mode you can enable and disable different combinations of constraints (1). As you do, the **POTENTIAL HOSTS** (2) label updates to show how many hosts are available to your entity. For example, by turning off the 4-Host constraints, you have 12 potential hosts for this VM. To see the resulting list of hosts, click the **POTENTIAL HOSTS** label (3).

List of Entities

The screenshot shows a list of 44 virtual machines. At the top, there is a search bar and sorting options. The first VM, 'i-2-32-VM', is expanded to show detailed resource usage and state. A '2' is placed over the expand/collapse icon for this VM.

Sort options: **By Virtual CPU** | By Severity | By Name | By Utilization

| VM Name | CPU | Memory | State |
|-------------|---------------------------------------|--|------------------------------------|
| i-2-32-VM | 52.06 GHz 5.00 % CPU Provisioned | 39.99 GB 12.50 % Memory Provisioned | IDLE State |
| iometer VM | 2.68 TB 0.02 % Storage Amount | 2.60 GHz 0.00 % Virtual CPU | 5.00 GB 0.00 % Virtual Memory |
| i-25-39-VM | | | |
| shai-test-4 | | | |

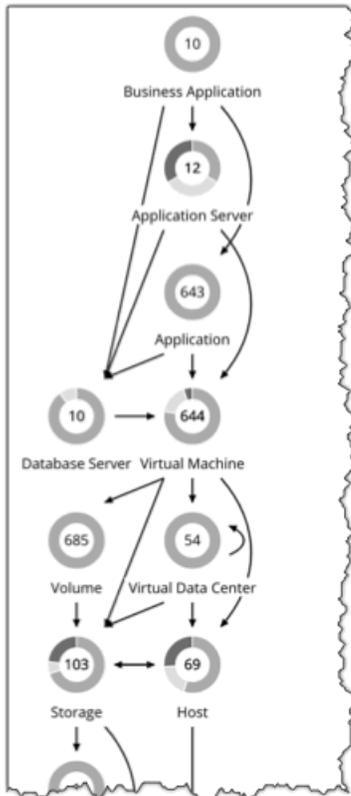
The list of entities is a quick way to drill down to details about your environment.

1. You can sort the list of entities by various categories.

- You can see specifics about resource consumption or state. For example, you can see the amount of capacity that has been assigned to a VM that is currently idle.

This list always updates to reflect the focus you have selected in the Supply Chain Navigator. When you select an entity type in the supply chain, the entities list updates to show the entities of that type for your current scope. For example, select Host to see a list of hosts in the current scope. For more information, see [Navigating with the Supply Chain \(on page 40\)](#).

Navigating With the Supply Chain



After you have set the scope of your Intersight Workload Optimizer session, you can use the Supply Chain to change the focus of the main view, and see details about different types of entities within the current scope.

Drilling Down in a Scoped Session

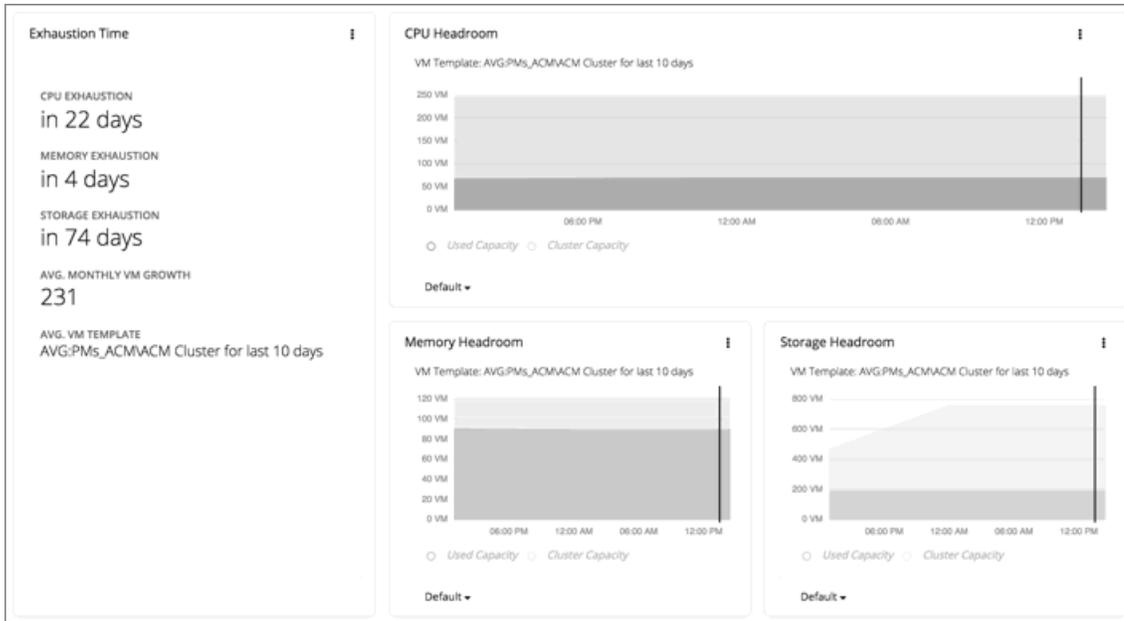
When you set a scope to your Intersight Workload Optimizer session, the **Overview** shows information about your environment, including:

- **Overview**
Charts and lists to give you an overview of your environment for the current scope. This overview corresponds to all the entities in scope.
- **Details** - Charts that give you a more detailed look at your environment for the given scope
- **Policies** - Any policies that are defined for the entities in the current scope
- **Entity Lists** - Details about the entities in the current scope
- **Pending Actions** - Actions that are pending for any entities in the current scope

The Supply Chain shows the currently selected tier of entities. To change the focus of the scoped view, select different tiers in the Supply Chain. The Policies, Entities List, and Pending Actions tabs update to focus on the tier you selected. These tabs show information for all the entities of that type that are in the current scope. For example, if you click the Host tier, these tabs update to show information about the hosts in your current scope.

To zoom in on a specific entity, you can click its name in the Entities List. This sets the scope to that specific entity. To return to the previous scope, use the browser's **Back** button.

Viewing Cluster Headroom



Cluster headroom shows you how much extra capacity your clusters have to host workloads. When you set the scope to a cluster, the **Overview** then includes charts that show headroom for that cluster, as well as time to exhaustion of the cluster resources.

To view host cluster headroom:

1. Navigate to the Search page.
2. Choose the **Host Clusters** category.
3. Select the cluster you want to view.
4. When the **Overview** displays, scroll down to show the headroom charts.

Make sure you have selected the Host tier in the supply chain navigator.

To calculate cluster capacity and headroom, Intersight Workload Optimizer runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current VMs to other hosts within the given cluster, if such moves would result in a more desirable workload distribution. The result of the plan is a calculation of how many more VMs the cluster can support.

To calculate VM headroom, the plan simulates adding VMs to your cluster. The plan assumes a certain capacity for these VMs, based on a specific VM template. For this reason, the count of VMs given for the headroom is an approximation based on that VM template.

To specify the templates these plans use, you can configure the nightly plans for each cluster. For more information, see [Configuring Nightly Plans \(on page 492\)](#)



Target Configuration

A target is an integration partner technology that provides resources or workload management services in your virtual environment. For example, Amazon Web Services (AWS) and VMware vCenter Server are supported targets.

Target configuration specifies the credentials that Intersight Workload Optimizer uses to connect to targets. Intersight Workload Optimizer connects through the management protocol that it exposes, such as REST API, SMI-S, XML, or some other management transport. It uses this connection to discover resources, monitor resource utilization, and recommend actions.

Points to consider:

- When a specific release or version of an integration partner technology reaches end-of-life (EOL) or its end of support date, Intersight Workload Optimizer no longer provides support for that version. Intersight Workload Optimizer follows integration partners' official EOL timeline for version support. Targeting a non-supported version, or one that is no longer supported by the vendor, is at your own risk.
- Intersight Workload Optimizer does not support duplicate instances of the same target. When configuring targets, you must not configure two or more targets to the same address in your environment. For example, you must not configure two different targets to the same AWS account, nor two targets to the same vCenter Server instance.

If you do configure duplicate targets, then actions can fail to execute with an error that begins: `Analysis Exception occurred...`

To resolve this issue, identify the duplicate targets, and delete them until you have only one target for each address.

Intersight Workload Optimizer supports the following targets.

| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|--------------|---|---|----------------------------|
| Cloud | Amazon Web Services | IWO Essentials | No |
| | Amazon Web Services Billing | IWO Essentials | No |
| | Google Cloud | IWO Essentials | No |
| | Google Cloud Billing | IWO Essentials | No |
| | Microsoft Azure Service Principal | IWO Essentials | No |
| | Microsoft Azure Billing | IWO Essentials | No |
| | Microsoft Azure Enterprise Agreement (deprecated) | IWO Essentials | No |

| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|--|--|---|----------------------------|
| Cloud Native | Kubernetes 1.8 or higher (IKS, CCP, Red Hat OpenShift, EKS, AKS, GKE) Deployed on-premises | IWO Advantage | Yes |
| | Kubernetes 1.8 or higher (Red Hat OpenShift, EKS, AKS, GKE) SaaS or deployed on public cloud | IWO Advantage | No |
| Applications and Databases | Apache Tomcat 7.x, 8.x, and 8.5.x Deployed on-premises | IWO Advantage | Yes |
| | IBM WebSphere Application Server 8.5+ Deployed on-premises | IWO Advantage | Yes |
| | JVM 6.0+ Deployed on-premises | IWO Advantage | Yes |
| | SQL Server 2012, 2014, 2016, 2017, and 2019 Deployed on-premises | IWO Advantage | Yes |
| | MySQL Server 8.0 Deployed on-premises | IWO Advantage | Yes |
| | Oracle 19c and 21c Deployed on-premises | IWO Advantage | Yes |
| Compute / Fabric | Cisco UCS Server (Standalone) | IWO Essentials | No |
| | Cisco UCS Domain (UCSM Managed) | IWO Essentials | No |
| | Cisco UCS Domain (Intersight Managed) | IWO Essentials | No |
| | HPE OneView 3.00.04 | IWO Essentials | No |
| Guest OS Process / APM (Application Performance Management) | New Relic SaaS or deployed on public cloud | IWO Premier | No |
| | Cisco AppDynamics 4.1+ Deployed on-premises | IWO Advantage | Yes |
| | Cisco AppDynamics SaaS or deployed on public cloud | IWO Advantage | No |
| | Dynatrace 1.1+ Deployed on-premises | IWO Premier | Yes |

| Category | Target Name | Minimum License Tier Required for Intersight Workload Optimizer | Intersight Assist Required |
|--------------------------|---|---|----------------------------|
| | Dynatrace SaaS or deployed on public cloud | IWO Premier | No |
| Hyperconverged | Cisco Hyperflex 3.5 | IWO Essentials | No |
| | Nutanix Acropolis | IWO Essentials | Yes |
| Hypervisor | Microsoft Hyper-V 2012 R2, 2016, 2019, 2022 | IWO Essentials | Yes |
| | VMware vCenter 6.0, 6.5, 6.7, and 7.0+ | IWO Essentials | Yes |
| Change Management | ServiceNow | IWO Advantage | No |
| Storage | HPE 3PAR InForm OS 3.2.2+, 3PAR SMI-S, 3PAR WSAPI | IWO Essentials | Yes |
| | Dell EMC SC Series | IWO Essentials | Yes |
| | EMC VMAX using SMI-S 8.1+ | IWO Essentials | Yes |
| | EMC ScaleIO 2.x and 3.x | IWO Essentials | Yes |
| | EMC VPLEX Local Architecture with 1:1 mapping of virtual volumes and LUNs | IWO Essentials | Yes |
| | NetApp ONTAP 8.0+ | IWO Essentials | Yes |
| | Pure Storage FlashArray running Purity 5.3.6 and 6.4.4 (Pure API 1.6) | IWO Essentials | Yes |
| | EMC XtremIO XMS 4.0+ | IWO Essentials | Yes |

Transport Layer Security Requirements

Intersight Workload Optimizer requires Transport Layer Security (TLS) version 1.2 to establish secure communications with targets. Most targets should have TLS 1.2 enabled. However, some targets might not have TLS enabled, or they might have enabled an earlier version. In that case, you will see handshake errors when Intersight Workload Optimizer tries to connect with the target service. When you go to the Target Configuration view, you will see a Validation Failed status for such targets.

If target validation fails because of TLS support, you might see validation errors with the following strings:

- `No appropriate protocol`
To correct this error, ensure that you have enabled the latest version of TLS that your target technology supports. If this does not resolve the issue, contact Cisco Technical Support.
- `Certificates do not conform to algorithm constraints`
To correct this error, refer to the documentation for your target technology for instructions to generate a certification key with a length of 2048 or greater on your target server. If this does not resolve the issue, contact Cisco Technical Support.

Cloud Targets

The public cloud provides compute, storage, and other resources on demand. Intersight Workload Optimizer can analyze the performance of workloads running on the public cloud, and scale workloads as demand requires.

With public cloud targets, you can use Intersight Workload Optimizer to scale workloads at the lowest possible cost, or reduce costs by purchasing discounts, deleting unattached volumes, or stopping workloads temporarily.

Amazon Web Services

Amazon Web Services (AWS) is Amazon's cloud computing platform. Intersight Workload Optimizer discovers your AWS resources through an IAM user, and then optimizes these resources to assure performance at the lowest possible cost.

To connect to AWS, follow the steps and guidelines outlined in the following topics:

- [Connecting to AWS \(on page 45\)](#)
- [AWS permissions \(on page 52\)](#)
- [AWS Billing permissions \(on page 52\)](#)

After connecting to AWS, Intersight Workload Optimizer monitors and optimizes the resources that it discovered. See the following topics for more information:

- [AWS Monitored Resources \(on page 52\)](#)
- [AWS Actions \(on page 57\)](#)

Connecting to AWS

To connect Intersight Workload Optimizer to your AWS environment, perform the following tasks:

1. [Set up an IAM user or IAM role in AWS \(on page 45\)](#).
Intersight Workload Optimizer discovers and monitors your AWS workloads through an IAM user that you set up in AWS.
2. [Set up a data export in AWS. \(on page 47\)](#)
Intersight Workload Optimizer uses a data export stored in an S3 bucket to visualize historical cloud expenses, and discover discounts and billing family relationships.
3. [Claim an AWS Billing target in Intersight Workload Optimizer. \(on page 49\)](#)
Authorize a secure connection to your billing data.
To authorize the connection, claim an AWS Billing target in the Intersight Workload Optimizer user interface.
4. [Claim an AWS target in Intersight Workload Optimizer. \(on page 50\)](#)
Authorize a secure connection to your workloads.
To authorize the connection, claim an AWS target in the Intersight Workload Optimizer user interface.

Setting Up an AWS IAM User

Intersight Workload Optimizer connects to your AWS environment through an IAM user.

For best practices on managing IAM identities, see the [AWS documentation](#).

Next Step

Set up an IAM user in AWS. For details, see this [topic \(on page 45\)](#).

Setting Up an AWS IAM User

Perform the following tasks to set up an IAM user for use with Intersight Workload Optimizer.

NOTE:

Intersight Workload Optimizer also supports IAM roles.

For information on supported IAM identities, see this [topic \(on page 45\)](#).

Guidelines

- Intersight Workload Optimizer recommends setting up an IAM user group that has the necessary permissions and then adding the IAM user to that group.
- If the IAM user that you are setting up also grants Intersight Workload Optimizer access to your billing data, the IAM user requires access to the S3 bucket that contains your data export. Billing access is not required.

Intersight Workload Optimizer uses a data export stored in an S3 bucket to visualize historical cloud expenses, and discover discounts and billing family relationships.

NOTE:

You will set up a data export in a later task.

Task Overview

To set up a IAM user, perform the following tasks in the AWS Management Console:

1. Create IAM policies that specify the permissions that Intersight Workload Optimizer needs to connect to AWS.
2. Create an IAM user and then assign the policies that you created to that user.
3. Generate an access key for the IAM user.

Creating IAM Policies

1. Sign in to the AWS Management Console and open the IAM console.
<https://console.aws.amazon.com/iam/>
2. In the navigation pane on the left, choose **Policies**.
3. Choose **Create policy**.
4. In the **Policy editor** section, choose **JSON**.
5. Paste the [minimum permissions](#) that Intersight Workload Optimizer requires to monitor workloads. To monitor workloads and execute actions within Intersight Workload Optimizer, use these [minimum permissions](#).
To retrieve billing data in your data export, use the [minimum permissions](#) for billing data monitoring.
6. Resolve any security warnings, errors, or general warnings generated during policy validation, and then choose **Next**.
7. In the **Review and create** page, type a **Policy Name** and a **Description** (optional) for the policy that you are creating.
8. Choose **Create policy**.

Creating an IAM User

1. In the navigation pane of the IAM console, select **Users** and then choose **Create user**.
2. Specify your preferred user name and then choose **Next**.
3. Select **Attach policies directly**, select the policy or policies that you created in the previous task, and then choose **Next**.
4. Review the user details and then choose **Create user**.

Generating an Access Key for the IAM User

1. In the navigation pane of the IAM console, select **Users** and then choose the user that you created in the previous task.
2. Choose **Security credentials**, scroll to the **Access keys** section, and then choose **Create access key**.
3. Choose **Third-party service** and then choose **Next**.
4. (Optional) Set a description tag value to describe the purpose of the access key.
5. Choose **Create access key**.
6. Record the access key ID and secret access key. You will need this information later when you claim an AWS target in the Intersight Workload Optimizer user interface.

Next Step

In AWS, set up a data export for use with the AWS Billing target. For details, see this [topic \(on page 47\)](#).

Setting Up an AWS Data Export

Intersight Workload Optimizer uses a data export stored in an S3 bucket to visualize historical cloud expenses, and discover discounts and billing family relationships.

Guidelines

- Intersight Workload Optimizer supports only the following data exports:
 - Standard data export (CUR 2.0)
 - Legacy CUR export
 All other data exports available in AWS are not supported.
- The IAM user or role that you set up for Intersight Workload Optimizer requires access to the S3 bucket that contains your data export. Billing access is not required.
- Intersight Workload Optimizer supports a data export created at the management account, but not member accounts.
- AWS publishes the daily data export twice a day, but it may take up to 24 hours for AWS to deliver the data export to the S3 bucket. Charts in Intersight Workload Optimizer, such as the Workload Cost Breakdown chart, will start to show data 1 to 2 hours after data export delivery.

Setting Up a Standard Data Export (CUR 2.0)

1. Sign in to the Billing and Cost Management console.
<https://console.aws.amazon.com/billing>
2. At the top-right section of the page, select the region for the S3 bucket.
3. In the navigation pane, find the **Cost Analysis** section and choose **Data Exports**.
4. Choose **Create**.
5. Choose **Standard data export** and configure the following settings.

- Export name

| Setting | Instructions |
|-------------|------------------------------|
| Export name | Specify your preferred name. |

- Data table content settings

| Setting | Instructions |
|---------------------------|--------------------------------------|
| Additional export content | Select Include resource IDs . |
| Time granularity | Choose Daily . |

- Column selection

By default, all columns are selected. To include only the columns that Intersight Workload Optimizer currently requires, select the following items:

- `bill_payer_account_id`
- `identity_line_item_id`
- `line_item_currency_code`
- `line_item_line_item_type`
- `line_item_normalization_factor`
- `line_item_normalized_usage_amount`
- `line_item_operation`
- `line_item_product_code`
- `line_item_resource_id`
- `line_item_unblended_cost`

- line_item_usage_account_id
- line_item_usage_amount
- line_item_usage_end_date
- line_item_usage_start_date
- line_item_usage_type
- pricing_public_on_demand_cost
- pricing_term
- pricing_unit
- product
- product_instance_type
- product_product_family
- product_sku
- reservation_reservation_a_r_n
- savings_plan_savings_plan_a_r_n
- savings_plan_savings_plan_effective_cost

■ Data export delivery options

| Setting | Instructions |
|----------------------------------|---|
| Compression type and file format | Choose gzip - text/csv . |
| File versioning | Choose Create new data export file . |

■ Data export storage settings

| Setting | Instructions |
|----------------|---|
| S3 bucket | Click Configure and then select an existing bucket or create a new one. If you created a new one, the region that you selected in a previous step is automatically specified. Note that the <code>AmazonS3FullAccess</code> and <code>Billing</code> permissions are required to set up the bucket. For detailed setup instructions, see the AWS documentation . Be sure to verify the bucket policy that AWS applies. |
| S3 path prefix | Specify your preferred prefix, such as <code>daily</code> . |

6. Review your settings and record the following information. You will need this information later when you add the AWS Billing target in the Intersight Workload Optimizer user interface.
 - S3 bucket name, such as `turbodataexport`
 - S3 path prefix, such as `daily`
 - S3 bucket region, such as `us-east-1`
7. Click **Create**.
AWS can now deliver the data export to the S3 bucket. It could take up to 24 hours for AWS to deliver the first data export.

Setting Up a Legacy CUR Export

1. Sign in to the Billing and Cost Management console.
<https://console.aws.amazon.com/billing>
2. At the top-right section of the page, select the region for the S3 bucket.
3. In the navigation pane, find the **Cost Analysis** section and choose **Data Exports**.
4. Choose **Create**.
5. Choose **Legacy CUR export** and configure the following settings.
 - Export name

| Setting | Instructions |
|-------------|------------------------------|
| Export name | Specify your preferred name. |

- Export content

| Setting | Instructions |
|---------------------------|--------------------------------------|
| Additional export content | Select Include resource IDs . |

- Data export delivery options

| Setting | Instructions |
|------------------------------|---|
| Report data time granularity | Choose Daily . |
| Report versioning | Choose Create new report version . |

- Data export storage settings

| Setting | Instructions |
|----------------|---|
| S3 bucket | <p>Click Configure and then select an existing bucket or create a new one.</p> <p>If you created a new one, the region that you selected in a previous step is automatically specified. Note that the <code>AmazonS3FullAccess</code> and <code>Billing</code> permissions are required to set up the bucket. For detailed setup instructions, see the AWS documentation.</p> <p>Be sure to verify the bucket policy that AWS applies.</p> |
| S3 path prefix | Specify your preferred prefix, such as <code>daily</code> . |

- Review your settings and record the following information. You will need this information later when you add the AWS Billing target in the Intersight Workload Optimizer user interface.
 - S3 bucket name, such as `turbodataexport`
 - S3 path prefix, such as `daily`
 - S3 bucket region, such as `us-east-1`
- Click **Create report**.

AWS can now deliver the data export to the S3 bucket. It could take up to 24 hours for AWS to deliver the first data export.

Next Step

In the Intersight Workload Optimizer user interface, claim an AWS Billing target using the data export settings and the IAM user or role that you set up. For details, see this [topic \(on page 49\)](#).

Claiming an AWS Billing Target

The AWS Billing target grants Intersight Workload Optimizer access to billing data from a data export stored in an S3 bucket. Intersight Workload Optimizer uses this data to visualize historical cloud expenses, and discover discounts and billing family relationships.

AWS member accounts and standalone accounts are not supported.

You can add multiple billing targets. Data for these targets will be aggregated and shown when you set the scope to your global environment.

NOTE:

The AWS Billing target only supports adding a data export associated with a management account. Member accounts or standalone accounts are not supported.

Before performing this task, be sure to [set up a data export \(on page 47\)](#) in AWS.

Adding an AWS Billing Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **AWS Billing**.
4. Configure the following settings:
 - **Custom Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in AWS.
 - **Access Key**
Specify the access key ID associated with the IAM user.
Access keys are long-term credentials for an IAM user or the AWS account root user.
[View security best practices in IAM](#)
 - **Secret Access Key**
Specify the secret access key associated with the IAM user.
 - **S3 Bucket Name**
Specify the full name of the S3 bucket that contains the data export.
 - **S3 Path Prefix**
Specify the S3 path prefix for the data export.
 - **S3 Bucket Region**
Specify the region of the S3 bucket that contains the data export.

Next Step

In the Intersight Workload Optimizer user interface, claim an AWS target using the IAM user or role that you set up. For details, see this [topic \(on page 50\)](#).

Claiming an AWS Target

Claim an AWS target in the Intersight Workload Optimizer user interface to monitor and optimize workloads in your AWS environment. This target specifies the IAM user that Intersight Workload Optimizer will use to connect to AWS.

Before performing this task, be sure to set up an IAM user in AWS.

Claiming an AWS Target Using an IAM User

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **AWS**.
4. Configure the following settings:
 - **Custom Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in AWS.
 - **Access Key**
Specify the access key ID associated with the IAM user.
Access keys are long-term credentials for an IAM user or the AWS account root user.
[View security best practices in IAM](#)
 - **Secret Access Key**
Specify the secret access key associated with the IAM user.

After Claiming an AWS Target

You have completed the required tasks for connecting to AWS.

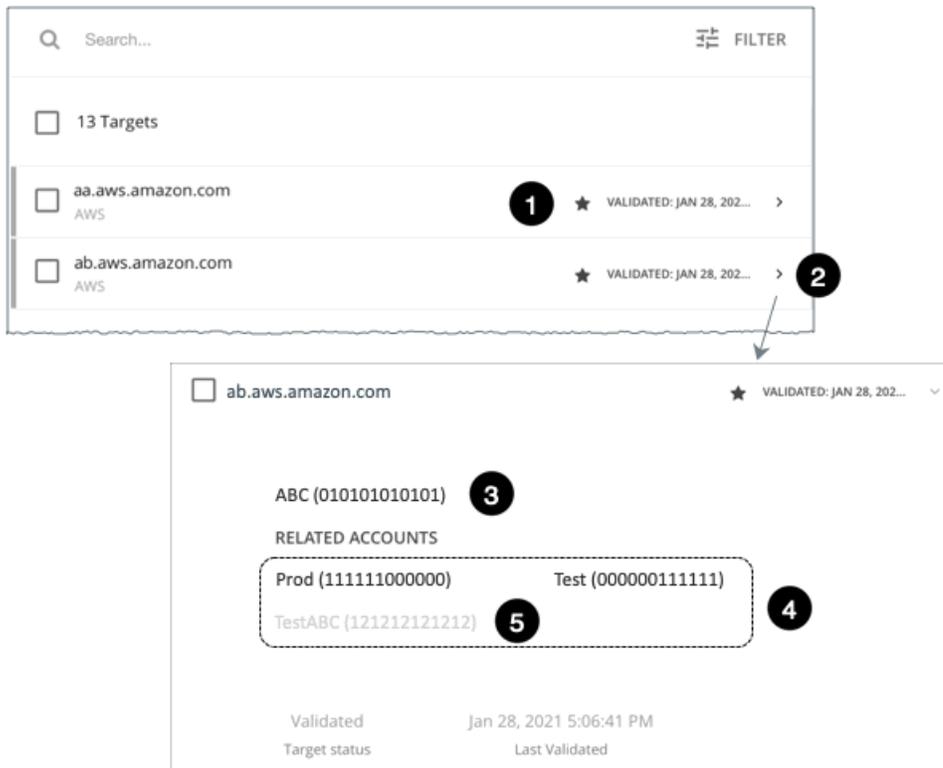
- Intersight Workload Optimizer can now monitor your AWS workloads and billing data, and recommend actions to optimize these workloads at the lowest possible cost.

Review the following topics for more information:

- [AWS Monitored Resources \(on page 52\)](#)
- [AWS Actions \(on page 57\)](#)
- If you have price adjustments for workloads based on your agreement with AWS, manually configure price adjustments in the Intersight Workload Optimizer user interface, in **Settings > Billing and Costs > Price Adjustments**. If price adjustments are not set, Intersight Workload Optimizer will use on-demand pricing, which could result in incorrect cost information in Intersight Workload Optimizer. For details, see this [topic \(on page 596\)](#).

Viewing AWS Accounts in the User Interface

The user interface displays the accounts discovered from your AWS targets. To identify these accounts with ease, refer to the following guidelines.



1. Management accounts appear in bold, with a star next to them.
2. You can expand the account entry to see the related member accounts.
3. If you expand the entry for a member account, the related accounts includes the management account, indicated by a star.
4. Related accounts includes the member accounts

NOTE:

If you expand the entry for a member account, then the related accounts includes the management account, indicated by a star.

5. A gray name indicates a member account that is not configured as a target.

Reference: AWS Permissions

The IAM user that you set up in AWS must specify the permissions that Intersight Workload Optimizer needs to discover and monitor your AWS workloads and billing data. Permissions to execute actions from Intersight Workload Optimizer are optional.

Click the links in the following table to view the minimum permissions for both the AWS and AWS Billing targets.

| Target | IAM Policy (JSON) |
|-------------|---|
| AWS Billing | Choose one: <ul style="list-style-type: none"> ■ Data export retrieval and workload monitoring ■ Data export retrieval, workload monitoring, and action execution |
| AWS | Choose one: <ul style="list-style-type: none"> ■ Workload monitoring ■ Workload monitoring and action execution |

AWS Monitored Resources

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|-----------------------------------|-------------------------------|
| AWS | Intersight Workload Optimizer |
| Elastic Compute Cloud (EC2) | Virtual Machine (VM) |
| Relational Database Service (RDS) | Database Server |
| Elastic Block Store (EBS) | Volume |
| Availability Zone | Zone |
| Region | Region |

Points to consider:

- Intersight Workload Optimizer supports discovery and management of entities in certain AWS regions. For details, see [Supported AWS Regions \(on page 53\)](#).

Monitored Resources for Virtual Machines

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

It is highly recommended that you enable collection of metrics in your environment. Enabling metrics allows Intersight Workload Optimizer to generate scale actions to optimize VM resource usage. For Intersight Workload Optimizer to collect metrics, you must enable the collection of these metrics on the VMs in your environment.

For details, see [AWS Memory Metrics Collection \(on page 54\)](#).
- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.
- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.
- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- I/O Throughput

I/O Throughput is the measurement of an entity's throughput to the underlying storage.

- Net Throughput

Net Throughput is the rate of message delivery over a port.

Monitored Resources for Database Servers

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- DB Cache Hit Rate

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

- Connection

Connection is the measurement of database connections utilized by applications.

Monitored Resources for Volumes

Intersight Workload Optimizer monitors the following resources:

- Storage Amount

Storage Amount is the storage capacity (disk size) of a volume.

Intersight Workload Optimizer discovers Storage Amount, but does not monitor utilization.

For a Kubeturbo (container) deployment that includes volumes, Kubeturbo monitors Storage Amount utilization for the volumes. You can view utilization information in the Capacity and Usage chart.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the measurement of IOPS capacity that is in use.

- I/O Throughput

I/O Throughput is the measurement of I/O throughput capacity that is in use.

Supported AWS Regions

Intersight Workload Optimizer supports discovery and management of entities in the following AWS regions:

| Region Code | Region Name | Notes |
|----------------|--------------------------|------------------------------------|
| af-south-1 | Africa (Cape Town) | Requires opt-in within AWS console |
| ap-south-1 | Asia Pacific (Mumbai) | |
| ap-northeast-1 | Asia Pacific (Tokyo) | |
| ap-northeast-2 | Asia Pacific (Seoul) | |
| ap-northeast-3 | Asia Pacific (Osaka) | |
| ap-southeast-1 | Asia Pacific (Singapore) | |
| ap-southeast-2 | Asia Pacific (Sydney) | |

| Region Code | Region Name | Notes |
|----------------|---------------------------|------------------------------------|
| ap-southeast-3 | Asia Pacific (Jakarta) | |
| ap-east-1 | Asia Pacific (Hong Kong) | Requires opt-in within AWS console |
| ca-central-1 | Canada (Central) | |
| eu-central-1 | Europe (Frankfurt) | |
| eu-south-1 | Europe (Milan) | Requires opt-in within AWS console |
| eu-west-1 | Europe (Ireland) | |
| eu-west-2 | Europe (London) | |
| eu-west-3 | Europe (Paris) | |
| eu-north-1 | Europe (Stockholm) | |
| me-south-1 | Middle East (Bahrain) | Requires opt-in within AWS console |
| sa-east-1 | South America (São Paulo) | |
| us-east-1 | US East (N. Virginia) | |
| us-east-2 | US East (Ohio) | |
| us-west-1 | US West (N. California) | |
| us-west-2 | US West (Oregon) | |

Unreachable AWS Regions

When Intersight Workload Optimizer fails to reach one or more AWS regions as part of discovering your AWS environment, then AWS discovery will fail for that target.

There may be policy decisions that prevent Intersight Workload Optimizer from reaching all AWS regions. For example, if you operate Intersight Workload Optimizer behind a firewall, you might not be able to reach all the regions that are available to your AWS account. In that case, you need to specify which regions you want Intersight Workload Optimizer to discover.

For information about how to specify the regions that you want Intersight Workload Optimizer to discover, contact your support representative.

AWS Metrics Collection

It is highly recommended that you enable collection of metrics in your environment. Enabling metrics allows Intersight Workload Optimizer to generate scale actions to optimize VM resource usage. For Intersight Workload Optimizer to collect metrics, you must enable the collection of these metrics on the VMs in your environment.

This topic describes the collection of the following metrics:

- Standard memory for AWS VMs
- Utilization of GPU cards and GPU memory for AWS VMs with Linux AMIs

Some of the steps to do this are different depending on whether your VM is running a Linux or Windows OS. To enable metrics collection, you must meet the following requirements:

- The VM image must have an SSM agent installed
 - Linux VMs:
 - By default, Linux AMIs dated 2017.09 and later include an installed SSM Agent.
 - Windows VMs:
 - You must install the SSM agent on the VMs. For more information, see [Working with SSM Agent](#).
- Access to the CloudWatch service

Your AWS Instance must have internet access or direct access to CloudWatch so it can push data to CloudWatch.

- **Access from Intersight Workload Optimizer**

For Intersight Workload Optimizer to access metrics, the account that it uses to connect to the AWS target must include the correct permissions. If you configured the AWS target via an AWS key (not an IAM role), then you must include the permissions as specified in the section for claiming an AWS target.

To set up the collection of metrics for your VMs:

1. Attach an IAM role to each VM instance.

Each EC2 instance must have an attached IAM role that grants CloudWatch access. To grant that access, include the `AmazonSSMFullAccess` policy in the role.

Use the AWS System Manager to attach the necessary roles to your VMs.

NOTE:

If you want to grant the role lesser access, you can use the `AmazonEC2RoleforSSM` policy. This is a custom policy that allows the action `ssm:GetParameter` to access the resource, `arn:aws:ssm:*:*:parameter/*`.

2. Install the CloudWatch agent on your Linux VMs.

Navigate to the AWS System Manager service for the account and region that you want to configure. In the service, navigate to the **Run Command** screen and set up the **AWS-ConfigureAWSPackage** command to install **AmazonCloudWatchAgent** on your VMs. For more information, see the AWS documentation.

3. Create configuration data for the CloudWatch agent.

The configuration data is a JSON object that you will add to as a parameter to the Parameter Store. The object must include the following, depending on whether it's for a Linux or a Windows VM instance.

- **Linux Configuration for Standard Memory**

```
{
  "agent": {
    "metrics_collection_interval": 60,
    "logfile": "/opt/aws/amazon-cloudwatch-agent/logs/amazon-cloudwatch-agent.log"
  },
  "metrics": {
    "namespace": "custom",
    "metrics_collected": {
      "mem": {
        "measurement": [
          {
            "name": "mem_available", "rename": "MemoryAvailable", "unit": "Bytes"
          }
        ]
      }
    }
  },
  "append_dimensions": {
    "AutoScalingGroupName": "${aws:AutoScalingGroupName}",
    "ImageId": "${aws:ImageId}",
    "InstanceId": "${aws:InstanceId}",
    "InstanceType": "${aws:InstanceType}"
  }
}
```

- **Linux Configuration for Standard Memory and GPU Card/Memory Utilization**

```
{
  "agent": {
    "metrics_collection_interval": 60,
```

```

"logfile":"/opt/aws/amazon-cloudwatch-agent/logs/amazon-cloudwatch-agent.log"
},
"metrics":{
  "namespace": "CWAgent",
  "metrics_collected":{
    "nvidia_gpu": {
      "measurement": [
        "utilization_gpu",
        "memory_used"
      ]
    },
    "mem":{
      "measurement":[
        {
          "name":"mem_available", "rename":"MemoryAvailable", "unit": "Bytes"
        }
      ]
    }
  },
  "append_dimensions":{
    "AutoScalingGroupName": "${aws:AutoScalingGroupName}",
    "ImageId": "${aws:ImageId}",
    "InstanceId": "${aws:InstanceId}",
    "InstanceType": "${aws:InstanceType}"
  }
}
}

```

■ Windows Configuration for Standard Memory

```

{
  "metrics": {
    "namespace": "Windows System",
    "append_dimensions": {
      "InstanceId": "${aws:InstanceId}"
    },
    "aggregation_dimensions" : [ ["InstanceId"] ],
    "metrics_collected": {
      "Memory": {
        "measurement": [
          {"name": "Available Bytes", "rename": "MemoryAvailable", "unit": "Bytes"}
        ],
        "metrics_collection_interval": 60
      },
      "Paging File": {
        "measurement": [
          {"name": "% Usage", "rename": "paging_used"}
        ],
        "metrics_collection_interval": 60,
        "resources": [
          "*"
        ]
      }
    }
  }
}

```

```
}

```

Note that you can configure optional parameters for the CW Namespace and region. However, if you configure more metrics for CloudWatch to collect, these metrics do not affect Intersight Workload Optimizer analysis and they do not show up in the user interface.

4. Create a parameter store.

a. Create a parameter.

In the AWS System Manager, navigate to **Parameter Store** and create a parameter. Copy and paste the JSON agent configuration (created in preceding steps) into the parameter **Value** field.

b. Name the parameter.

For example, `AmazonCloudWatch-MyMemoryParam`. You can use a different name, but per the Amazon documentation, the name *must* begin with `AmazonCloudWatch`. For more information, see [Store the CloudwatchConfig File in Parameter Store](#).

You must remember this parameter name.

c. Click to create the parameter.

5. Deploy the CloudWatch parameter to your VMs.

In the AWS System Manager, navigate to the **Run Command** screen to configure and run the **AmazonCloudWatch-ManagedAgent** command. The configuration should include:

- **Action:** `configure`
- **Mode:** `ec2`
- **Optional Configuration Source:** `ssm`
- **Optional Configuration Location:** Give the name of the parameter that you created earlier.
- **Optional Restart:** `yes` (this restarts the CloudWatch Agent, not the VM instance)
- **Targets:** The VMs that you will deploy the CloudWatch configuration to

When the command is configured, run it. This configures collection of metrics for your instances.

6. Verify that you are collecting metrics for your instances.

Navigate to the CloudWatch page, and display **Metrics** in the **CWAgent** namespace. Then inspect the instances by ID to verify that you can see `MemoryAvailable` or `utilization_gpu` and `memory_used` metrics if you are collecting GPU metrics.

AWS Actions

Intersight Workload Optimizer monitors the state and performance of your workloads and then recommends actions to optimize these workloads at the lowest possible cost.

NOTE:

Use the Potential Savings and Necessary Investments charts to view pending actions and evaluate their impact on your cloud expenditure.

Actions for Virtual Machines

Intersight Workload Optimizer supports the following actions:

- **Scale**

Change the VM instance to use a different instance type or tier to optimize performance and costs.

- **Discount-related actions**

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing RI coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

If you need more capacity, then Intersight Workload Optimizer will recommend actions to purchase additional RIs.

For details, see [Actions for AWS VMs \(on page 265\)](#).

Actions for Database Servers

Intersight Workload Optimizer supports the following actions:

- **Scale**

Scale compute and storage resources to optimize performance and costs.

For details, see [Cloud Database Server Actions \(on page 297\)](#).

Actions for Volumes

Intersight Workload Optimizer supports the following actions:

- **Scale**

Scale attached volumes to optimize performance and costs.

- **Delete**

Delete unattached volumes as a cost-saving measure. Intersight Workload Optimizer generates an action immediately after discovering an unattached volume.

For details, see [Cloud Volume Actions \(on page 327\)](#).

Google Cloud

Google Cloud is Google's cloud computing platform. Intersight Workload Optimizer discovers your Google Cloud resources through a service account, and then optimizes these resources to assure performance at the lowest possible cost.

To connect to Google Cloud, follow the steps and guidelines outlined in the following topics:

- [Connecting to Google Cloud \(on page 58\)](#)
- [Google Cloud permissions \(on page 69\)](#)
- [Google Cloud Billing permissions \(on page 72\)](#)

After connecting to Google Cloud, Intersight Workload Optimizer monitors and optimizes the resources that it discovered. See the following topics for more information:

- [Google Cloud Monitored Resources \(on page 73\)](#)
- [Google Cloud Actions \(on page 74\)](#)

Connecting to Google Cloud

To connect Intersight Workload Optimizer to your Google Cloud environment, perform the following tasks:

1. [Enable the required APIs in Google Cloud. \(on page 59\)](#)

When enabled, these APIs allow Intersight Workload Optimizer to discover your Google Cloud resources and billing data.

2. [Set up a service account for workload monitoring in Google Cloud. \(on page 59\)](#)

Intersight Workload Optimizer discovers and monitors your Google Cloud workloads through a service account that you set up in Google Cloud. You can set up a service account that accesses your entire organization, individual projects, or individual folders.

3. [Claim a Google Cloud target in Intersight Workload Optimizer. \(on page 64\)](#)

Authorize a secure connection through your service account.

To authorize the connection, claim a Google Cloud target in the Intersight Workload Optimizer user interface.

4. [Set up a billing export in Google Cloud. \(on page 65\)](#)

Intersight Workload Optimizer uses billing data from a billing export to BigQuery to discover Committed Use Discounts and visualize historical cloud expenses.

5. [Set up a service account for billing data monitoring in Google Cloud. \(on page 66\)](#)

Intersight Workload Optimizer retrieves billing data through a service account that you set up in Google Cloud. You can use the service account that you previously set up for workload monitoring, or create a new one specifically for billing data monitoring.

6. [Claim a Google Cloud Billing target in Intersight Workload Optimizer. \(on page 68\)](#)

Authorize a secure connection to BigQuery.

To authorize the connection, claim a Google Cloud Billing target in the Intersight Workload Optimizer user interface.

Enabling Required Google Cloud APIs

For Intersight Workload Optimizer to discover your Google Cloud environment and billing data, you must enable the following APIs:

- Cloud Resource Manager API
Creates, reads, and updates metadata for resource containers.
- Compute Engine API
Creates VMs and volumes.
- Cloud Billing API
Enables developers to manage billing for their projects programmatically.
- BigQuery API
A data platform for customers to create, manage, share, and query data.

To enable these APIs:

1. Navigate the Google Cloud Console to the library of APIs.
On the Google Cloud Console home page, navigate to **APIs & Services > Library**.
2. Search for the API you want to enable.
In the API Library **Search** box, enter the name of the API you want to enable. Then press **Enter** to execute the search. Repeat these steps for each of:
 - Cloud Resource Manager API
 - Compute Engine API
 - Cloud Billing API
 - BigQuery API
3. Enable the given API.
In the list that appears, click the API name to navigate to that API page. If the API is not already enabled, click **Enable**.
After you enable the given API, the console displays a details page for that API.
4. Navigate to the console Home page.
For each API you want to enable, navigate back to the home page and repeat these steps.

Setting Up a Google Cloud Service Account for Workload Monitoring

This topic describes the steps to set up a valid service account that Intersight Workload Optimizer will use to connect to your Google Cloud environment. The access that you grant this service account determines the Google Cloud workloads that Intersight Workload Optimizer discovers, monitors, and optimizes.

Task Overview

To set up a service account, perform the following tasks in Google Cloud:

1. Create a service account for workload monitoring.
2. Create a custom role and then assign the role to the service account.

Creating a Service Account for Workload Monitoring

Create a service account and generate a key file for the account. The key file is required when adding a Google Cloud target in the Intersight Workload Optimizer user interface.

For seamless monitoring of your Google Cloud workloads, create the service account in a project that does not typically hit the rate limits enforced by Google Cloud, such as a non-production project.

1. In the project that will host the new service account, open a `gcloud` CLI session.

2. Create a service account.

```
gcloud iam service-accounts create <SERVICE_ACCOUNT_NAME>
```

Where:

<SERVICE_ACCOUNT_NAME> is the internal name of the new service account. The name must be between 6 and 30 characters in length.

3. Record the following information for later use.

- <SERVICE_ACCOUNT_NAME>
- <PROJECT_ID>

NOTE:

<PROJECT_ID> identifies the project that hosts the service account. This information is needed if you need to review or edit the service account later.

4. Generate a key file for the service account.

```
gcloud iam service-accounts keys create <KEY_FILE_NAME> \
  --iam-account=<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com
```

Where:

- <KEY_FILE_NAME> is your preferred name for the key file.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.

5. Download the key file to your local machine. You will use the key file later when you add a Google Cloud target in the Intersight Workload Optimizer user interface.

```
cloudshell download <KEY_FILE_NAME>
```

Where:

<KEY_FILE_NAME> is the key file name that you specified in the previous step.

Overview of Custom Roles

You can assign the following custom roles to the service account that you created.

- (Required) Custom role for workload monitoring

This custom role specifies the permissions that Intersight Workload Optimizer needs to discover and monitor workloads in your entire organization, or in individual folders or projects.

| Google Cloud Resources to Monitor | Task |
|-----------------------------------|--|
| Organization | Create a custom role at the organization level. |
| Individual folders | Create a custom role at the organization level. It is not possible to create custom roles at the folder level. |
| Individual projects | Create a custom role at the project level. |

- (Optional) Custom role for action execution

To execute actions from Intersight Workload Optimizer, create a role that specifies the required permissions for executing actions.

For details, see one of the following topics:

- [Creating and Assigning Custom Roles \(Organization Level\) \(on page 61\)](#)
- [Creating and Assigning Custom Roles \(Project Level\) \(on page 62\)](#)

Creating and Assigning Custom Roles (Organization Level)

NOTE:

Skip to the [next section \(on page 62\)](#) if you want Intersight Workload Optimizer to monitor individual projects.

1. Create a custom role for workload monitoring.

```
gcloud iam roles create <ROLE_ID> --organization=<ORGANIZATION_ID> \
  --title='IWO Role: Access - Organization' \
  --description='Minimal Required Permissions for \
  IWO to manage the GCP Organization' \
  --permissions="compute.commitments.list,\
  compute.disks.get,compute.disks.list,compute.diskTypes.list,\
  compute.instances.get,compute.instances.list,\
  compute.instanceGroupManagers.get,\
  compute.instanceGroupManagers.list,compute.instanceGroups.list,\
  compute.machineTypes.get,compute.machineTypes.list,\
  compute.regions.list,compute.zones.list,\
  logging.views.get,logging.views.list,\
  monitoring.services.get,monitoring.services.list,\
  monitoring.timeSeries.list,resourceManager.projects.get,\
  serviceusage.services.get,billing.resourceAssociations.list,\
  resourceManager.folders.get,resourceManager.folders.list,\
  resourceManager.organizations.get,resourceManager.projects.get,\
  resourceManager.projects.list" --stage=ALPHA
```

Where:

- <ROLE_ID> is your preferred ID for the custom role.
- <ORGANIZATION_ID> is the organization that Intersight Workload Optimizer will monitor.

2. Assign the custom role to the service account.

```
gcloud organizations add-iam-policy-binding <ORGANIZATION_ID> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com --role=<ROLE_NAME>
```

Where:

- <ORGANIZATION_ID> is the organization that Intersight Workload Optimizer will monitor.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.
- <ROLE_NAME> is the complete path for the role name, expressed as follows:
organizations/<ORGANIZATION_ID>/roles/<ROLE_ID>

NOTE:

<ROLE_ID> was created in a previous step.

3. Add the predefined **Billing Account Viewer** role to the service account.

```
gcloud organizations add-iam-policy-binding <ORGANIZATION_ID> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com \
  --role=roles/billing.viewer
```

Where:

- <ORGANIZATION_ID> is your Google Cloud organization.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.

4. (Optional) Create a custom role to execute actions from Intersight Workload Optimizer.

```
gcloud iam roles create <ROLE_ID_ACTION> --organization=<ORGANIZATION_ID> \
  --title='IWO Role: Org Action Execution' \
  --description='Grant IWO permissions to \
  execute actions in the GCP Organization' \
  --permissions="compute.disks.create,compute.disks.createSnapshot,\
compute.disks.delete,compute.disks.resize,\
compute.disks.setLabels,compute.disks.update,\
compute.disks.use,compute.disks.useReadOnly,\
compute.globalOperations.get,\
compute.instanceGroups.get,compute.instanceGroups.list,\
compute.instanceGroups.use,compute.instances.attachDisk,\
compute.instances.detachDisk,compute.instances.setLabels,\
compute.instances.setMachineType,compute.instances.start,\
compute.instances.stop,compute.instances.useReadOnly,\
compute.instanceTemplates.list,compute.instantSnapshots.list,\
compute.regionOperations.get,compute.reservations.list,\
compute.resourcePolicies.use,compute.snapshots.create,\
compute.snapshots.delete,compute.snapshots.get,\
compute.snapshots.list,compute.snapshots.useReadOnly,\
compute.zoneOperations.get,\
iam.serviceAccounts.actAs" --stage=ALPHA
```

Where:

- <ROLE_ID_ACTION> is your preferred ID for the custom role for action execution.
- <ORGANIZATION_ID> is the organization that Intersight Workload Optimizer will monitor.

5. (Optional) Assign the custom role for action execution to the service account.

```
gcloud projects add-iam-policy-binding <ORGANIZATION_ID> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com --role=<ROLE_NAME_ACTION>
```

Where:

- <ORGANIZATION_ID> is the organization that Intersight Workload Optimizer will monitor.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.
- <ROLE_NAME_ACTION> is the complete path for the role name, expressed as follows:

```
organizations/<ORGANIZATION_ID>/roles/<ROLE_ID_ACTION>
```

NOTE:

<ROLE_ID_ACTION> was created in a previous step.

Creating and Assigning Custom Roles (Project Level)

NOTE: See the [previous section \(on page 61\)](#) if you want Intersight Workload Optimizer to monitor individual folders or your entire organization.

1. Create a custom role for workload monitoring.

```
gcloud iam roles create <ROLE_ID> --project=<PROJECT_ID_MONITOR> \
  --title='IWO Role: Min Access - Project' \
  --description='Minimal Required Permissions for \
  IWO to manage the GCP Project' \
  --permissions="resourceManager.projects.get,compute.regions.list,\
compute.zones.list,compute.machineTypes.list,\
compute.machineTypes.get,compute.disks.list,\
compute.disks.get,compute.diskTypes.list,\
compute.instances.list,compute.instances.get,\
compute.instanceGroupManagers.list,\
compute.instanceGroupManagers.get,\
compute.instanceGroups.list,\
compute.commitments.list,logging.views.list,\
logging.views.get,monitoring.services.get,\
monitoring.services.list,monitoring.timeSeries.list,\
serviceusage.services.get" --stage=ALPHA
```

Where:

- <ROLE_ID> is your preferred ID for the custom role.
- <PROJECT_ID_MONITOR> is the project that Intersight Workload Optimizer will monitor.

2. Assign the custom role to the service account.

```
gcloud projects add-iam-policy-binding <PROJECT_ID_MONITOR> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com --role=<ROLE_NAME>
```

Where:

- <PROJECT_ID_MONITOR> is the project that Intersight Workload Optimizer will monitor.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.
- <ROLE_NAME> is the complete path for the role name, expressed as follows:
 projects/<PROJECT_ID_MONITOR>/roles/<ROLE_ID>

NOTE:

<ROLE_ID> was created in a previous step.

3. Add the predefined **Billing Account Viewer** role to the service account.

```
gcloud organizations add-iam-policy-binding <ORGANIZATION_ID> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com \
  --role=roles/billing.viewer
```

Where:

- <ORGANIZATION_ID> is your Google Cloud organization.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.

4. (Optional) Create a custom role to execute actions from Intersight Workload Optimizer.

```
gcloud iam roles create <ROLE_ID_ACTION> --project=<PROJECT_ID_MONITOR> \
  --title='IWO Role: Project Action Execution' \
  --description='Grant IWO permissions to \
  execute actions in the GCP Project' \
  --permissions="compute.disks.create,compute.disks.createSnapshot,\
compute.disks.delete,compute.disks.resize,\
compute.disks.setLabels,compute.disks.update,\
compute.disks.use,compute.disks.useReadOnly,\
compute.globalOperations.get,\
compute.instanceGroups.get,compute.instanceGroups.list,\
compute.instanceGroups.use,compute.instances.attachDisk,\
compute.instances.detachDisk,compute.instances.setLabels,\
compute.instances.setMachineType,compute.instances.start,\
compute.instances.stop,compute.instances.useReadOnly,\
compute.instanceTemplates.list,compute.instantSnapshots.list,\
compute.regionOperations.get,compute.reservations.list,\
compute.resourcePolicies.use,compute.snapshots.create,\
compute.snapshots.delete,compute.snapshots.get,\
compute.snapshots.list,compute.snapshots.useReadOnly,\
compute.zoneOperations.get,\
iam.serviceAccounts.actAs" --stage=ALPHA
```

Where:

- <ROLE_ID_ACTION> is your preferred ID for the custom role for action execution.
- <PROJECT_ID_MONITOR> is the project that Intersight Workload Optimizer will monitor.

5. (Optional) Assign the custom role for action execution to the service account.

```
gcloud projects add-iam-policy-binding <PROJECT_ID_MONITOR> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com --role=<ROLE_NAME_ACTION>
```

Where:

- <PROJECT_ID_MONITOR> is the project that Intersight Workload Optimizer will monitor.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.
- <ROLE_NAME_ACTION> is the complete path for the role name, expressed as follows:

```
projects/<PROJECT_ID_MONITOR>/roles/<ROLE_ID_ACTION>
```

NOTE:

<ROLE_ID_ACTION> was created in a previous step.

Next Step

In the Intersight Workload Optimizer user interface, claim a Google Cloud target. For details, see this [topic \(on page 64\)](#).

Claiming a Google Cloud Target

Claim an Google Cloud target in the Intersight Workload Optimizer user interface to monitor and optimize workloads in your Google Cloud environment. This target specifies the service account that Intersight Workload Optimizer will use to connect to Google Cloud.

Before performing this task, be sure you have the key file for the service account that Intersight Workload Optimizer will use to connect to Google Cloud. If you do not have this key, follow the steps outlined in this [topic \(on page 59\)](#).

Claiming the Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **GCP**.
4. Configure the following settings:
 - **Name**
 Authorize a secure connection to BigQuery.
 Specify a name that uniquely identifies this connection.
 This name is for display purposes only and does not need to match any name in Google Cloud.
 - **Service Account Key (JSON)**
 Specify the service account key (JSON).
 This is the JSON for the key file that you generated in a previous task.

After Claiming a Google Cloud Target

Intersight Workload Optimizer starts to discover the projects that define compute, storage, and networking resources for your workloads. It then creates a derived target for each discovered project. Derived targets are not directly modifiable within Intersight Workload Optimizer but can be validated like any other target.

Intersight Workload Optimizer discovers a broader resource hierarchy if you claimed a target with permissions to access folders or your entire organization.

Next Step

In Google Cloud, set up a billing export. For details, see this [topic \(on page 65\)](#).

Setting Up a Google Cloud Billing Export

This topic describes the billing export that Intersight Workload Optimizer will use to monitor billing data exported to BigQuery. Without this data, Intersight Workload Optimizer cannot discover any cost data used for analysis.

When you export Google Cloud billing data from BigQuery, you can choose to export standard or detailed usage cost data. Intersight Workload Optimizer recommends detailed usage cost data because it includes granular information for VMs, which the platform uses to generate accurate VM scaling actions. This information is also reflected in cloud charts (such as the Workload Cost Breakdown chart) when you set the scope to individual VMs. Note that support for standard usage cost data will be discontinued in a future release.

Setup guidelines:

- For an overview of billing export, see this [Google Cloud page](#).
- To set up a billing export, follow the steps in this [Google Cloud page](#).
- Shortly after you set up a billing export, Google Cloud automatically creates the following billing data tables in the BigQuery dataset. Be sure to record these table names. You will need them later when you add a Google Cloud Billing target in the Intersight Workload Optimizer user interface.
 - [Standard usage cost data table](#)
 In your BigQuery dataset, this table is named `gcp_billing_export_v1_<Billing_Account_ID>`.
 - [Detailed usage cost data table](#)
 This table includes all the data fields from the standard usage cost data table, along with additional fields that provide resource-level cost data, such as a virtual machine or SSD that generates service usage. In your BigQuery dataset, this table is named `gcp_billing_export_resource_v1_<Billing_Account_ID>`.
 - [Pricing data export table](#)
 In your BigQuery dataset, this table is named `cloud_pricing_export`.

Next Step

In Google Cloud, set up a service account for billing data monitoring. For details, see this [topic \(on page 66\)](#).

Setting Up a Google Cloud Service Account for Billing Data Monitoring

This topic describes the steps to set up a valid service account that Intersight Workload Optimizer will use to connect to your Google Cloud environment and monitor billing data. Without this data, Intersight Workload Optimizer cannot discover any cost data used for analysis.

Task Overview

To set up a service account, perform the following tasks in Google Cloud:

1. Create a service account for billing data monitoring.
2. Create a custom role and then assign the role to the service account.
3. Add the Billing Account Viewer role to the service account.

Creating a Service Account for Billing Data Monitoring

Create a service account and generate a key file for the account. The key file is required when adding a Google Cloud Billing target in the Intersight Workload Optimizer user interface.

For seamless monitoring of your Google Cloud resources, create the service account in a project that does not typically hit the rate limits enforced by Google Cloud, such as a non-production project.

NOTE:

You can use the existing service account that you set up for workload monitoring, or create a new one specifically for billing data monitoring. Skip to the [next section \(on page 67\)](#) if you plan to use your existing service account.

1. In the project that will host the new service account, open a `gcloud` CLI session.
2. Create a service account.

```
gcloud iam service-accounts create <SERVICE_ACCOUNT_NAME>
```

Where:

<SERVICE_ACCOUNT_NAME> is the internal name of the new service account. The name must be between 6 and 30 characters in length.

3. Record the following information for later use.

- <SERVICE_ACCOUNT_NAME>
- <PROJECT_ID>

NOTE:

<PROJECT_ID> identifies the project that hosts the service account. This information is needed if you need to review or edit the service account later.

4. Generate a key file for the service account.

```
gcloud iam service-accounts keys create <KEY_FILE_NAME> \
  --iam-account=<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com
```

Where:

- <KEY_FILE_NAME> is your preferred name for the key file.
- <SERVICE_ACCOUNT_NAME> is the name of the service account that you created.
- <PROJECT_ID> is the project that hosts the service account.

5. Download the key file to your local machine. You will use the key file later when you add a Google Cloud target in the Intersight Workload Optimizer user interface.

```
cloudshell download <KEY_FILE_NAME>
```

Where:

<KEY_FILE_NAME> is the key file name that you specified in the previous step.

Creating and Assigning a Custom Role for Billing Data Monitoring

This custom role is **required** and must be created in the project that stores billing data. The role specifies the permissions that Intersight Workload Optimizer needs to discover and monitor billing data.

1. Create a custom role.

```
gcloud iam roles create <ROLE_ID_BILL> --project=<PROJECT_ID_BILL> \
  --title='IWO Billing Data Viewer Role' \
  --description='Minimal Required Permissions for \
  IWO to view \
  billed cost and pricing stored in the GCP Project' \
  --permissions="bigquery.jobs.create,bigquery.tables.get,\
  bigquery.tables.getData,bigquery.tables.list,\
  compute.commitments.list,compute.diskTypes.list,\
  compute.machineTypes.list,compute.regions.list,\
  compute.zones.list" --stage=ALPHA
```

Where:

- `<ROLE_ID_BILL>` is your preferred ID for the custom role.
- `<PROJECT_ID_BILL>` is the project that stores billing data.

2. Assign the custom role to the service account.

```
gcloud projects add-iam-policy-binding <PROJECT_ID_BILL> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com --role=<ROLE_NAME_BILL>
```

Where:

- `<PROJECT_ID_BILL>` is the project that stores billing data.
- `<SERVICE_ACCOUNT_NAME>` is the name of the service account that you created.
- `<PROJECT_ID>` is the project that hosts the service account.
- `<ROLE_NAME_BILL>` is the complete path for the role name, expressed as follows:

```
projects/<PROJECT_ID_BILL>/roles/<ROLE_ID_BILL>
```

NOTE:

`<ROLE_ID_BILL>` was created in a previous step.

Adding the Billing Account Viewer Role to the Service Account

1. Add the predefined **Billing Account Viewer** role to the service account.

```
gcloud organizations add-iam-policy-binding <ORGANIZATION_ID> \
  --member=serviceAccount:<SERVICE_ACCOUNT_NAME>@<PROJECT_ID>.iam.gserviceaccount.com \
  --role=roles/billing.viewer
```

Where:

- `<ORGANIZATION_ID>` is your Google Cloud organization.
- `<SERVICE_ACCOUNT_NAME>` is the name of the service account that you created.
- `<PROJECT_ID>` is the project that hosts the service account.

Next Step

In the Intersight Workload Optimizer user interface, claim a Google Cloud Billing target. For details, see this [topic \(on page 68\)](#).

Claiming a Google Cloud Billing Target

The Google Cloud Billing target grants Intersight Workload Optimizer access to billing data from a billing export to BigQuery. Intersight Workload Optimizer uses this data to visualize historical cloud expenses and discover Committed Use Discounts.

Points to consider:

- The Google Cloud Billing target can retrieve billing data from BigQuery data sets across all applicable regions, both inside and outside the US.
- You can add multiple billing targets. Data for these targets will be aggregated and shown when you set the scope to your global environment.

Before claiming a target, be sure you have the key file for the service account that Intersight Workload Optimizer will use to connect to Google Cloud and discover billing data. If you do not have this key, follow the steps outlined in this [topic \(on page 66\)](#).

Claiming the Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **GCP Billing**.
4. Configure the following settings:
 - **Name**
 Authorize a secure connection to BigQuery.
 Specify a name that uniquely identifies this connection.
 This name is for display purposes only and does not need to match any name in Google Cloud.
 - **Service Account Key**
 Specify the service account key (JSON).
 This is the JSON for the key file that you generated in a previous task.
 - **GCP Project ID**
 Specify the ID for the project that stores billing data.
 This is the unique ID assigned to the project associated with the billing account. Costs accrued to this project are charged to the billing account you are adding.
 - **BigQuery Settings**
 BigQuery is a data warehouse that helps you manage Google Cloud data. Intersight Workload Optimizer uses BigQuery resources to discover cost data for your environment. If you do not configure any of these fields, this target will not discover any cost data for Intersight Workload Optimizer analysis.

Configure the following settings.

- **BigQuery Cost Export Data Set Name**
 Specify the dataset name associated with your BigQuery billing export.
 This is the data set for billed costs. After you specify a data set, you must also specify the corresponding BigQuery Cost Export Table Name.
 You can find the data set name in the Google Cloud Billing dashboard under **Billing export / BIGQUERY EXPORT**.
- **BigQuery Cost Export Table Name**
 Specify the cost table name associated with the BigQuery billing export.
 The cost table names in your BigQuery dataset are as follows:
 - Standard usage cost data
`gcp_billing_export_v1_<Billing_Account_ID>`
 - Detailed usage cost data
`gcp_billing_export_resource_v1_<Billing_Account_ID>`
- **Enable Resource Level Detail From Cost Export Table**

When you export Google Cloud billing data from BigQuery, you can choose to export standard or detailed usage cost data. Intersight Workload Optimizer recommends detailed usage cost data because it includes granular information for VMs, which the platform uses to generate accurate VM scaling actions. This information is also reflected in cloud charts (such as the Workload Cost Breakdown chart) when you set the scope to individual VMs. Note that support for standard usage cost data will be discontinued in a future release.

For either option, specify the dataset name and cost table name associated with the BigQuery billing export.

- Detailed usage cost data

Select **Enable Resource Level Detail From Cost Export Table** and then specify the table name of the detailed usage cost data in the **BigQuery Cost Export Table Name** field.

- Standard usage cost data

Clear **Enable Resource Level Detail From Cost Export Table** and then specify the table name of the standard usage cost data in the **BigQuery Cost Export Table Name** field.

- BigQuery Pricing Export Table Name

Specify the table name associated with your BigQuery pricing export.

In your BigQuery dataset, the pricing export table is named *cloud_pricing_export*.

- BigQuery Pricing Export Data Set Name

Specify the dataset name associated with your BigQuery pricing export.

You can find the data set name in the Google Cloud Billing dashboard under **Billing export / BIGQUERY EXPORT**.

- Billing Account ID

Specify the billing account ID associated with your project. For help finding the ID, see this Google Cloud [page](#).

This field is required if you configure **BigQuery Pricing Export Data Set Name** and **BigQuery Pricing Export Table Name**.

After Claiming a Google Cloud Billing Target

You have completed the required tasks for connecting to Google Cloud. Intersight Workload Optimizer can now monitor your Google Cloud workloads and billing data, and recommend actions to optimize these workloads at the lowest possible cost. Review the following topics for more information:

- [Google Cloud Monitored Resources \(on page 73\)](#)
- [Google Cloud Actions \(on page 74\)](#)

Reference: Google Cloud Permissions

Intersight Workload Optimizer requires specific permissions to monitor your Google Cloud workloads and billing data.

Permissions for the Google Cloud Target

The Google Cloud target that you add to the Intersight Workload Optimizer user interface monitors your Google Cloud workloads. You can grant the target permissions to monitor workloads in individual projects, individual folders, or your entire organization.

Optionally, you can grant the target permissions to execute actions for your workloads automatically.

- **Project-level permissions**

NOTE:

Be sure to [set up a service account \(on page 59\)](#) in Google Cloud, [create a custom role \(on page 62\)](#), and then assign the custom role to the service account. You need to specify these permissions when you create the custom role.

When a service account has been properly configured for use with Intersight Workload Optimizer, [add a Google Cloud target \(on page 64\)](#) in the user interface.

- (Required) Monitoring permissions
 - `compute.commitments.list`
 - `compute.disks.get`

- `compute.disks.list`
- `compute.diskTypes.list`
- `compute.instances.get`
- `compute.instances.list`
- `compute.instanceGroupManagers.get`
- `compute.instanceGroupManagers.list`
- `compute.instanceGroups.list`
- `compute.machineTypes.get`
- `compute.machineTypes.list`
- `compute.regions.list`
- `compute.zones.list`
- `logging.views.get`
- `logging.views.list`
- `monitoring.services.get`
- `monitoring.services.list`
- `monitoring.timeSeries.list`
- `resourcemanager.projects.get`
- `serviceusage.services.get`
- (Optional) Action execution permissions
 - `compute.disks.create`
 - `compute.disks.createSnapshot`
 - `compute.disks.delete`
 - `compute.disks.resize`
 - `compute.disks.setLabels`
 - `compute.disks.update`
 - `compute.disks.use`
 - `compute.disks.useReadOnly`
 - `compute.globalOperations.get`
 - `compute.instanceGroups.get`
 - `compute.instanceGroups.list`
 - `compute.instanceGroups.use`
 - `compute.instances.attachDisk`
 - `compute.instances.detachDisk`
 - `compute.instances.setLabels`
 - `compute.instances.setMachineType`
 - `compute.instances.start`
 - `compute.instances.stop`
 - `compute.instances.useReadOnly`
 - `compute.instanceTemplates.list`
 - `compute.instantSnapshots.list`
 - `compute.regionOperations.get`
 - `compute.reservations.list`
 - `compute.resourcePolicies.use`
 - `compute.snapshots.create`
 - `compute.snapshots.delete`
 - `compute.snapshots.get`
 - `compute.snapshots.list`
 - `compute.snapshots.useReadOnly`
 - `compute.zoneOperations.get`

- iam.serviceAccounts.actAs

■ Folder-level permissions

To monitor workloads at the folder level, organization-level permissions are required (see the next item).

■ Organization-level permissions

NOTE:

Be sure to [set up a service account \(on page 59\)](#) in Google Cloud, [create a custom role \(on page 61\)](#), and then assign the custom role to the service account. You need to specify these permissions when you create the custom role.

When a service account has been properly configured for use with Intersight Workload Optimizer, [add a Google Cloud target \(on page 64\)](#) in the user interface.

- (Required) Monitoring permissions
 - Workload monitoring
 - compute.commitments.list
 - compute.disks.get
 - compute.disks.list
 - compute.diskTypes.list
 - compute.instances.get
 - compute.instances.list
 - compute.instanceGroupManagers.get
 - compute.instanceGroupManagers.list
 - compute.instanceGroups.list
 - compute.machineTypes.get
 - compute.machineTypes.list
 - compute.regions.list
 - compute.zones.list
 - logging.views.get
 - logging.views.list
 - monitoring.services.get
 - monitoring.services.list
 - monitoring.timeSeries.list
 - resourcemanager.projects.get
 - serviceusage.services.get
 - Resource hierarchy monitoring
 - billing.resourceAssociations.list
 - resourcemanager.folders.get
 - resourcemanager.folders.list
 - resourcemanager.organizations.get
 - resourcemanager.projects.get
 - resourcemanager.projects.list
- (Optional) Action execution permissions
 - compute.disks.create
 - compute.disks.createSnapshot
 - compute.disks.delete
 - compute.disks.resize
 - compute.disks.setLabels
 - compute.disks.update
 - compute.disks.use
 - compute.disks.useReadOnly
 - compute.globalOperations.get

- `compute.instanceGroups.get`
- `compute.instanceGroups.list`
- `compute.instanceGroups.use`
- `compute.instances.attachDisk`
- `compute.instances.detachDisk`
- `compute.instances.setLabels`
- `compute.instances.setMachineType`
- `compute.instances.start`
- `compute.instances.stop`
- `compute.instances.useReadOnly`
- `compute.instanceTemplates.list`
- `compute.instantSnapshots.list`
- `compute.regionOperations.get`
- `compute.reservations.list`
- `compute.resourcePolicies.use`
- `compute.snapshots.create`
- `compute.snapshots.delete`
- `compute.snapshots.get`
- `compute.snapshots.list`
- `compute.snapshots.useReadOnly`
- `compute.zoneOperations.get`
- `iam.serviceAccounts.actAs`

Permissions for the Google Cloud Billing Target

The Google Cloud Billing target grants Intersight Workload Optimizer access to billing data from a billing export to BigQuery. Intersight Workload Optimizer uses this data to visualize historical cloud expenses and discover Committed Use Discounts.

NOTE:

Be sure to [set up a service account \(on page 66\)](#) in Google Cloud, [create a custom role \(on page 67\)](#), and then assign the custom role to the service account. You need to specify these permissions when you create the custom role.

When a service account has been properly configured for use with Intersight Workload Optimizer, [add a Google Cloud Billing target \(on page 68\)](#) in the user interface.

- `bigquery.jobs.create`
- `bigquery.tables.get`
- `bigquery.tables.getData`
- `bigquery.tables.list`
- `compute.commitments.list`
- `compute.diskTypes.list`
- `compute.machineTypes.list`
- `compute.regions.list`
- `compute.zones.list`

Billing Data Exports for Google Cloud

When you export Google Cloud billing data from BigQuery, you can choose to export standard or detailed usage cost data. Intersight Workload Optimizer recommends detailed usage cost data because it includes granular information for VMs, which the platform uses to generate accurate VM scaling actions. This information is also reflected in cloud charts (such as the Workload Cost Breakdown chart) when you set the scope to individual VMs. Note that support for standard usage cost data will be discontinued in a future release.

If you configured your [Google Cloud Billing target \(on page 68\)](#) to retrieve standard usage cost data from BigQuery, switch to the detailed usage cost data by following these steps:

1. Export detailed usage cost data from BigQuery.
2. In the Intersight Workload Optimizer user interface, go to **Settings > Target Configuration** and then open your Google Cloud Billing target for editing.
3. Select **Enable Resource Level Detail From Cost Export Table**.
4. In the **BigQuery Cost Export Table Name** field, specify the name of the detailed usage cost data table.

Google Cloud Monitored Resources

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| Google Cloud | Intersight Workload Optimizer |
|-------------------------------|-------------------------------|
| Virtual Machine (VM) Instance | Virtual Machine (VM) |
| Disk | Volume |
| Zone | Zone |
| Region | Region |

Points to consider:

- Google Cloud *projects, folders, and billing accounts* do not appear as entities in the supply chain. Use Search to scope to these resources. In Search, projects are grouped under Accounts, folders under Folders, and billing accounts under Billing Families.
- Intersight Workload Optimizer supports discovery and management of workloads in all currently available Google Cloud [regions and zones](#).

Monitored Resources for Virtual Machines

NOTE:

Intersight Workload Optimizer discovers Google Cloud labels attached to VMs as tags. You can filter VMs by tags when you use Search or create groups. The Action Details page for a pending VM action also lists all the discovered tags.

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

Google Cloud collects memory metrics via [Ops Agent](#). In order for Intersight Workload Optimizer to retrieve these metrics, install and configure Ops Agent on each VM that it monitors. See Ops Agent installation instructions [here](#), and configuration details [here](#).

NOTE:

Google Cloud recommends using Ops Agent instead of its [legacy monitoring agent](#).

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

Intersight Workload Optimizer calculates CPU based on the normalized CPU frequency and the number of vCPUs for a given VM. Normalized CPU frequency takes into account performance variations seen in different models of a given CPU platform. Because frequency is normalized, charts might show utilization values that are slightly higher than 100% (for example, 100.03%) when capacity is fully utilized.

- Net Throughput

Net Throughput is the rate of message delivery over a port.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

■ Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

■ I/O Throughput

I/O Throughput is the measurement of an entity's throughput to the underlying storage.

For both Storage Access (IOPS) and I/O Throughput, Intersight Workload Optimizer calculates capacity or uses capacity data published by Google Cloud, depending on the VM's machine type and disk.

- Shared-core machine types share a physical core and are used for running small, non-resource intensive apps.

For shared-core machine types with *standard disks* or *SSDs*, Intersight Workload Optimizer calculates capacity using internal benchmark data, which takes into consideration the observed maximum limit that can be achieved for IOPS and I/O throughput. It then uses the calculated capacity to analyze utilization more accurately.

The use of internal benchmark data could result in differences in the capacity data shown in Intersight Workload Optimizer and Google Cloud.

- For machine types that are *not* shared-core:
 - Intersight Workload Optimizer uses [published](#) capacity data and assumes that I/O block size is 16KB per I/O.
 - For machine types with [persistent disks](#), Intersight Workload Optimizer assumes that the published capacity for the *SSD* disk type also applies to the *balanced* and *extreme* disk types. When a VM is attached to at least one of these disk types, capacity is assumed to be the per-VM limit for the *SSD* disk type. When a VM is attached only to the *standard* disk type, capacity is the per-VM limit for the standard disk type.

Monitored Resources for Volumes

Intersight Workload Optimizer monitors the following resources:

■ Storage Amount

Storage Amount is the storage capacity (disk size) of a volume.

Intersight Workload Optimizer discovers Storage Amount, but does not monitor utilization.

For a Kubeturbo (container) deployment that includes volumes, Kubeturbo monitors Storage Amount utilization for the volumes. You can view utilization information in the Capacity and Usage chart.

■ Storage Access (IOPS)

Storage Access, also known as IOPS, is the measurement of IOPS capacity that is in use.

■ I/O Throughput

I/O Throughput is the measurement of I/O throughput capacity that is in use.

NOTE:

Intersight Workload Optimizer also monitors the attachment state of volumes and then generates delete actions for unattached volumes.

Google Cloud Actions

Intersight Workload Optimizer monitors the state and performance of your workloads and then recommends actions to optimize these workloads at the lowest possible cost.

NOTE:

Use the Potential Savings and Necessary Investments charts to view pending actions and evaluate their impact on your cloud expenditure.

Actions for Virtual Machines

Intersight Workload Optimizer supports the following actions:

■ Scale

Change the VM instance to use a different instance type or tier to optimize performance and costs.

■ Discount-related actions

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing Committed Use Discount (CUD) coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

Actions to purchase CUDs will be introduced in a future release.

- **Reconfigure**

Google Cloud provides a specific set of machine types for each zone in a region. If you create a policy that restricts a VM to certain machine types and the zone it is currently on does not support all of those machine types, Intersight Workload Optimizer will recommend a reconfigure action as a way to notify you of the non-compliant VM.

For details, see [Actions for Google Cloud VMs \(on page 269\)](#).

Actions for Volumes

Intersight Workload Optimizer supports the following actions:

- **Scale**

Scale attached volumes to optimize performance and costs.

For scale actions that require a tier change or actions that scale up IOPS for extreme persistent disks, Google Cloud requires a snapshot of the volume and the creation of a new volume based on the snapshot. When a new volume is created from a snapshot, it effectively has a new identity, thus resulting in the loss of historical utilization and cost information for that volume.

- **Delete**

Delete unattached volumes as a cost-saving measure. Intersight Workload Optimizer generates an action immediately after discovering an unattached volume.

For additional information, see [Cloud Volume Actions \(on page 327\)](#).

Microsoft Azure

Microsoft Azure is Microsoft's cloud computing platform. Intersight Workload Optimizer discovers your Azure resources through a service principal, and then optimizes these resources to assure performance at the lowest possible cost.

To connect to Azure, follow the steps and guidelines outlined in the following topics:

- [Connecting to Azure \(on page 75\)](#)
- [Azure permissions \(on page 85\)](#)

After connecting to Azure, Intersight Workload Optimizer monitors and optimizes the resources that it discovered. See the following topics for more information:

- [Azure Monitored Resources \(on page 101\)](#)
- [Azure Actions \(on page 110\)](#)

Connecting to Azure

To connect Intersight Workload Optimizer to your Azure environment, perform the tasks outlined in this topic.

NOTE:

Intersight Workload Optimizer requires Resource Manager to monitor your Azure resources. If you are currently managing Azure resources using the classic deployment model, [migrate to Resource Manager](#) before connecting Intersight Workload Optimizer to Azure.

1. [Register resource providers. \(on page 76\)](#)
Register resource providers for the Azure subscriptions that Intersight Workload Optimizer will manage.
2. [Set up a service principal for workload monitoring in Azure. \(on page 76\)](#)
Intersight Workload Optimizer discovers and monitors your Azure workloads through a service principal that you set up in Azure.
3. [Claim an Azure Service Principal target in Intersight Workload Optimizer \(on page 79\)](#).
Authorize a secure connection through your service principal.

- To authorize the connection, claim an Azure Service Principal target in the Intersight Workload Optimizer user interface.
4. [Decide whether to use the Cost Details API or a cost export. \(on page 80\)](#)
Intersight Workload Optimizer can collect billing data through the Cost Details API or a cost export that you set up in Azure. Microsoft recommends the Cost Details API for billing data that is less than 2 GB in size, and a cost export for sizes that exceed 2 GB.
 5. [Set up a service principal for billing data monitoring in Azure. \(on page 81\)](#)
Intersight Workload Optimizer collects billing data through a service principal that you set up in Azure. You can use the service principal that you previously set up for workload monitoring, or create a new one specifically for billing data monitoring.
 6. [Claim an Azure Billing target in Intersight Workload Optimizer. \(on page 84\)](#)
Authorize a secure connection to your billing data.
To authorize the connection, claim an Azure Billing target in the Intersight Workload Optimizer user interface.

Registering Azure Resource Providers

A resource provider is a set of Azure REST operations that enable functionality for a specific Azure service. Registration configures your Azure subscriptions to work with resource providers.

Most resource providers are auto-registered and some need to be manually registered. In some cases, resource providers must be re-registered to support new locations that you may need to use.

Register the following resource providers for the Azure subscriptions that Intersight Workload Optimizer will manage:

- Microsoft.Compute
- Microsoft.Capacity
- Microsoft.Sql

To register resource providers, follow the registration steps in the [Azure documentation](#).

NOTE:

If you do not register these resource providers, you might notice certain errors in the Intersight Workload Optimizer logs, as documented in this [Azure page](#).

Next Step

Set up a service principal in Azure. For details, see this [topic \(on page 76\)](#).

Setting Up an Azure Service Principal for Workload Monitoring

This topic describes the steps to set up a valid service principal that Intersight Workload Optimizer will use to connect to your Azure environment and discover subscriptions. Intersight Workload Optimizer discovers, monitors, and optimizes the workloads in these subscriptions.

Task Overview

To set up a service principal for workload monitoring, perform the following tasks in the Azure portal:

1. Register Intersight Workload Optimizer with Microsoft Entra ID (formerly known as Azure Active Directory).
2. Create a client secret key.
3. Configure API permissions.
4. Enable access to the subscriptions that Intersight Workload Optimizer will manage.

Registering Intersight Workload Optimizer with Microsoft Entra ID (Formerly Known as Azure Active Directory)

Registration automatically creates a service principal that serves as Intersight Workload Optimizer's identity in the Microsoft Entra tenant.

1. Sign in to the Azure portal with an administrator or co-administrator account. This level of access is only required for setting up the service principal, and not for regular Intersight Workload Optimizer operations.

<https://portal.azure.com>

2. Browse to **Microsoft Entra ID > App registrations** and select **New registration**.
3. Configure the following settings:
 - **Name**
Specify your preferred name, such as Intersight Workload Optimizer.
 - **Supported Account Types**
Select **Accounts in this organizational directory only (Default Directory)**.
4. Select **Register**.
Microsoft Entra ID creates the app registration for the service principal.
5. Record the Application (client) ID and Directory (tenant) ID. You will need this information later when you claim an Azure target in the Intersight Workload Optimizer user interface.

Creating a Client Secret

1. Browse to **Microsoft Entra ID > App registrations** and then select the app registration for the service principal that you created for Intersight Workload Optimizer.
2. In the navigation menu, select **Certificates & secrets**.
3. Select **Client secrets** and then select **New client secret**.
4. Configure the following settings:
 - **Description**
Specify a meaningful description.
 - **Expires**
Choose *Never*.
5. Select **Add**.
6. Record the client secret value. You will need this information later when you claim an Azure service principal target in the Intersight Workload Optimizer user interface.

IMPORTANT:

The client secret value only displays once. It will no longer be available after you leave the page.

Configuring API Permissions

1. Browse to **Microsoft Entra ID > App registrations** and then select the app registration for the service principal that you created for Intersight Workload Optimizer.
2. In the navigation menu, select **API permissions**.
3. Click **Add a permission** and select **Azure Service Management**.
4. Select **Delegated permissions** and **user_impersonation**, and then click **Add permissions**.

Enabling Access to Subscriptions Using Azure Roles

Assign roles to the service principal to enable access to subscriptions.

To speed up the process of connecting Intersight Workload Optimizer to Azure, enable access at a high [scope level](#) (such as at the management group level). This automatically propagates permissions to lower levels of scope, such as subscriptions and resource groups.

You can assign the service principal either a custom role or built-in roles. If you have strict requirements for the service principal, assign a custom role.

- **Custom role**
When you create a custom role, you specify the permissions that Intersight Workload Optimizer needs to monitor workloads in the subscription. Optionally, specify permissions to execute actions from Intersight Workload Optimizer.
- **Built-in roles**

Built-in roles include default permissions that are sufficient for Intersight Workload Optimizer operations, but are more permissive than a custom role.

- (Required) The **Reader** role is required to monitor workloads.

You can also use a combination of **Reader** on the subscription, and **Storage Account - List Keys** on the storage account where memory metrics are stored. To set this up, you must create the **Storage Account - List Keys** role via the Azure CLI or APIs.

For example, you can edit the following to add your keys:

```
{
  "Name": "IWO Storage Key Access",
  "IsCustom": true,
  "Description": "Can list storageAccount keys.",
  "Actions": [
    "Microsoft.Storage/storageAccounts/listkeys/action"
  ],
  "NotActions": [
  ],
  "DataActions": [
  ],
  "NotDataActions": [
  ],
  "AssignableScopes": [
    "/subscriptions/<INSERT_SUBSCRIPTION_ID>"
  ]
}
```

If you save this JSON to a file named `listkeys.json`, you can execute the following command to create the permission:

```
az role definition create --role-definition listkeys.json
```

For other approaches to specify enhanced security, contact your support representative.

- (Optional) The **Owner** or **Contributor** role is needed to execute actions from Intersight Workload Optimizer. The Contributor role is the least privileged role for action execution.

To assign a custom role, see the next section. To assign built-in roles, skip to [Assigning Built-in Roles \(on page 79\)](#) below.

Assigning a Custom Role

This task assumes that when you assign a custom role to the service principal, you will upload a JSON file that specifies the required permissions. You can manually configure the permissions by following the wizard for creating custom roles, but complete instructions for running that wizard are not described in this task.

1. Create a JSON file that specifies the required permissions.
 - For permissions to monitor workloads, copy the JSON content found [here \(on page 94\)](#).
 - For permissions to monitor workloads and execute actions, copy the JSON content found [here \(on page 96\)](#).

NOTE:

Be sure to update the following information in the JSON file:

- `<RoleName>` - Specify your preferred name for the custom role.
 - `<Subscription_ID>` - Specify the ID of the subscription that Intersight Workload Optimizer will manage.
2. Browse to **Subscriptions**.
 3. Select a subscription that Intersight Workload Optimizer will manage.
 4. In the navigation bar, select **Access control (IAM)**.
 5. Click **Add > Add custom role**.

6. In the **Basics** tab:
 - a. In the **Baseline permissions** field, select **Start from JSON**.
 - b. In the **File** field, upload the JSON file that you created in a previous step. Azure notifies you if the JSON is valid.
7. Click **Review + create**.

Assigning Built-in Roles

1. Browse to **Subscriptions**.
2. Select a subscription that Intersight Workload Optimizer will manage.
3. In the navigation bar, select **Access control (IAM)**.
4. Click **Add > Add role assignment**.
5. In the **Role** tab, search for and select **Reader**.
(Optional) To execute actions from Intersight Workload Optimizer, select **Owner** or **Contributor**.
6. Click the **Members** tab. In that tab:
 - a. Click **Select members**.
 - b. Search for and select the app registration for the service principal that you created for Intersight Workload Optimizer.
 - c. Click **Select**.
7. Click the **Review + assign** tab to review your settings and then click **Review + assign**.

Next Step

Claim an Azure Service Principal target in Intersight Workload Optimizer. For details, see this [topic \(on page 79\)](#).

Claiming an Azure Service Principal Target

Claim an Azure Service Principal target in the Intersight Workload Optimizer user interface to monitor and optimize workloads in your Azure environment. This target specifies the service principal credentials that Intersight Workload Optimizer will use to connect to Azure.

Before performing this task, be sure you have the Application (client) ID, Directory (tenant) ID, and client secret value for the service principal. If you do not have these credentials, follow the steps outlined in this [topic \(on page 76\)](#).

Claiming the Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **Azure Service Principal**.
4. Configure the following settings:
 - **Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in Azure.
 - **Directory (Tenant) ID**
Specify the Directory ID associated with your app registration.
Format: 32 alphanumeric characters separated by hyphens
 - **Application (Client) ID**
Specify the application ID associated with your app registration.
Format: 32 alphanumeric characters separated by hyphens
 - **Client Secret Value**
Specify the client secret value associated with your app registration.

After Claiming an Azure Service Principal Target

Intersight Workload Optimizer starts to discover the resources in your Azure environment.

In the Target Configuration page, the service principal displays as the primary target that you can edit, while the subscriptions display as derived targets that cannot be edited but can be validated like any other target.

Next Step

Decide whether to use the Cost Details API or a cost export to enable access to your Azure billing data. For details, see this [topic \(on page 80\)](#).

Using the Cost Details API or a Cost Export

The Azure Billing target can collect billing data for your Microsoft Customer Agreement (MCA) or Enterprise Agreement (EA) accounts through the Azure [Cost Details API](#) or a cost export that you set up in your environment.

■ Cost Details API

Microsoft recommends the Cost Details API for billing data that is less than 2 GB in size. Larger sizes will likely result in timeouts.

■ Cost Export

Microsoft recommends a cost export for billing data that is at least 2 GB in size. The data export is in CSV format, and contains all the cost and usage data that Azure Cost Management collects.

NOTE:

Support for the Cost Details API started in the May 2023 release. New deployments of the product will use this API by default.

If you previously set up a cost export, the Azure Billing target will continue to use that cost export after the platform updates to the May 2023 (or later) release. However, the target might become disconnected due to a discovery failure error. If you encounter this issue, open the target for editing and then turn on the **Use a cost export** setting. After turning on the setting, you can turn it off to switch to the Cost Details API.

Cost Details API

If you choose the Cost Details API:

- There is nothing that you need to set up in Azure.
- If you run Intersight Workload Optimizer behind a firewall, be sure to allow unrestricted access to `*.blob.core.windows.net`.
- When you add an Azure Billing target in the Intersight Workload Optimizer user interface in a later task, be sure to *turn off* the **Use a cost export** option.

Next Step

Set up an Azure service principal for billing data monitoring. For details, see this [topic \(on page 81\)](#).

Cost Export

If you choose a cost export, set up a daily cost export of month-to-date costs in the Azure portal.

Cost Export Guidelines

- The cost export must be created as follows:
 - MCA accounts – Create the cost export at the Billing Profile scope.

NOTE:

For MCA accounts, if you previously set up a cost export at the Billing Account scope, delete the cost export and then create a new one at the Billing Profile scope. Create this cost export for each active Billing Profile that falls under your MCA Billing Account. All cost exports you create for your active Billing Profiles are required to have the same name.

- EA accounts – Create the cost export at the Billing Account scope.
- For both MCA and EA accounts, Subscription, Management Group, and Resource Group scopes are not supported.

- Provide the following permissions to the storage account and container associated with the cost export.
 - Storage account – Reader and Data Access
 - Container – Storage Blob Data Reader
- Intersight Workload Optimizer recommends creating a new cost export, even if you have an existing export that matches the setup noted here. Below is an example of a cost export setup.
 - Export details
 - Metric
Actual cost (Usage and Purchases)
 - Export type
Daily export of month-to-date costs
 - Storage
 - Use existing
 - Subscription
EA-Development
 - Storage account
turbocostexport
 - Container
cost-export-container
 - Directory
costExportDir

Additional Guidelines

- If the storage account for the cost export is behind a firewall, add the cluster IP range for Intersight Workload Optimizer to the firewall's allowlist.
- When you add an Azure Billing target in the Intersight Workload Optimizer user interface in a later task, be sure to *turn on* the **Use a cost export** option.

Next Step

Set up an Azure service principal for billing data monitoring. For details, see this [topic \(on page 81\)](#).

Setting Up an Azure Service Principal for Billing Data Monitoring

This topic describes the steps to set up a valid service principal that Intersight Workload Optimizer will use to connect to your Microsoft Customer Agreement (MCA) or Enterprise Agreement (EA) accounts and collect billing data. Intersight Workload Optimizer uses this data to visualize historical expenses and discover reservations.

NOTE:

For information on the current level of support for MCA and EA accounts, see this [topic \(on page 100\)](#).

Task Overview

To set up a service principal for billing data monitoring, perform the following tasks in the Azure portal:

- Scenario 1: Use the existing service principal that you set up for workload monitoring.
 1. Enable access to billing data. For more information, see the 'Enabling Access to Billing Data' section in this topic.
- Scenario 2: Create a new service principal specifically for billing data monitoring.
 1. Register Intersight Workload Optimizer with Microsoft Entra ID (formerly known as Azure Active Directory).
 2. Create a client secret key.
 3. Configure API permissions.
 4. Enable access to billing data.

See the next sections for detailed instructions.

Registering Intersight Workload Optimizer with Microsoft Entra ID (Formerly Known as Azure Active Directory)

NOTE:

Skip this task if you want to use the existing service principal that you set up for workload monitoring.

Registration automatically creates a service principal that serves as Intersight Workload Optimizer's identity in the Microsoft Entra tenant.

1. Sign in to the Azure portal with an administrator or co-administrator account. This level of access is only required for setting up the service principal, and not for regular Intersight Workload Optimizer operations.
<https://portal.azure.com>
2. Browse to **Microsoft Entra ID > App registrations** and select **New registration**.
3. Configure the following settings:
 - **Name**
Specify your preferred name, such as Intersight Workload Optimizer.
 - **Supported Account Types**
Select **Accounts in this organizational directory only (Default Directory)**.
4. Select **Register**.
Microsoft Entra ID creates the app registration for the service principal.
5. Record the Application (client) ID and Directory (tenant) ID. You will need this information later when you claim an Azure target in the Intersight Workload Optimizer user interface.

Creating a Client Secret

NOTE:

Skip this task if you want to use the existing service principal that you set up for workload monitoring.

1. Browse to **Microsoft Entra ID > App registrations** and then select the app registration for the service principal that you created for Intersight Workload Optimizer.
2. In the navigation menu, select **Certificates & secrets**.
3. Select **Client secrets** and then select **New client secret**.
4. Configure the following settings:
 - **Description**
Specify a meaningful description.
 - **Expires**
Choose *Never*.
5. Select **Add**.
6. Record the client secret value. You will need this information later when you claim an Azure service principal target in the Intersight Workload Optimizer user interface.

IMPORTANT:

The client secret value only displays once. It will no longer be available after you leave the page.

Configuring API Permissions

NOTE:

Skip this task if you want to use the existing service principal that you set up for workload monitoring.

1. Browse to **Microsoft Entra ID > App registrations** and then select the app registration for the service principal that you created for Intersight Workload Optimizer.
2. In the navigation menu, select **API permissions**.
3. Click **Add a permission** and select **Azure Service Management**.
4. Select **Delegated permissions** and **user_impersonation**, and then click **Add permissions**.

Enabling Access to Billing Data

To enable access to billing data, assign the required role to the service principal. Note that only users with the Global Administrator role and with elevated access can assign the role to the service principal. Elevated access grants permissions to assign roles in Azure subscriptions and management groups associated with Microsoft Entra ID (formerly known as Azure Active Directory). For instructions on elevating access for users, see the [Azure documentation](#).

The following roles are required:

- EA accounts – Enrollment Reader role
- MCA accounts – Billing Account Reader role

NOTE:

For information about the current level of support for EA and MCA accounts, see this [topic \(on page 100\)](#).

Assigning the Enrollment Reader Role

If you have EA accounts, assign the Enrollment Reader role to enable access to billing data.

1. Record the values for the following parameters. These values are required when assigning the role:

| Parameter | Value |
|------------------------------|---|
| billingAccountName | Billing account ID In the Azure portal, navigate to Cost management + Billing , open your EA account, and then select Overview . The screen that displays includes a field for the Billing Account ID. |
| billingRoleAssignmentName | A unique GUID To generate a GUID, use the New-Guid PowerShell command in Microsoft Learn or the Online GUID / UUID Generator website. |
| api-version | 2019-10-01-preview This is a static value. |
| properties.principalId | Object ID specified in Microsoft Entra ID For help locating the Object ID, see the Azure documentation . |
| properties.principalTenantId | Tenant ID specified in Microsoft Entra ID For help locating the Tenant ID, see the Azure documentation . |

2. Execute an API request in the Microsoft Learn portal to assign the Enrollment Reader role.
 - a. Open the [Microsoft Learn portal](#). Sign in when prompted.
 - b. In the following fields, specify the values that you recorded in the previous step.

| Field | Value |
|---------------------------|---------------------------|
| billingAccountName | {your_billing_account_ID} |
| billingRoleAssignmentName | {your_unique_GUID} |
| api-version | 2019-10-01-preview |

- c. In the **Body** field, paste the following content.

```
{
  "properties": {
    "principalId": "{your_object_ID}",
    "principalTenantId": "{your_tenant_ID}"
  }
}
```

```

    "roleDefinitionId": "/providers/Microsoft.Billing/billingAccounts/{your_billing_account_ID}/b
illingRoleAssignments/24f8edb6-1668-4659-b5e2-40bb5f3a7d7e"
  }
}

```

Be sure to replace the following variables with the values that you recorded in the previous step.

- `{your_object_ID}`
- `{your_tenant_ID}`
- `{your_billing_account_ID}`

NOTE:

24f8edb6-1668-4659-b5e2-40bb5f3a7d7e is the ID that Microsoft assigned to the Enrollment Reader role.

- d. Review the request preview and then click **Run**.

The Enrollment Reader role is now assigned.

Assigning the Billing Account Reader Role

If you have MCA accounts, assign the Billing Account Reader role to enable access to billing data.

1. Browse to **Cost Management + Billing > Access control (IAM)**.
2. Select **Add**.
3. Select **Billing account reader** and then select the app registration for the service principal that you created for Intersight Workload Optimizer.

For Azure reservations, the app registration for the service principal must have permissions to manage reservations. In most cases, permissions for the **Reader** role are sufficient.

NOTE:

Intersight Workload Optimizer also discovers reservations scoped to a *resource group*, but treats them as shared (in the Discount Inventory chart, the scope for these reservations is shown as *Shared**). This could result in unreliable actions, such as scaling VMs within the resource group to other reservations, which could potentially increase costs. If you have existing reservations scoped to a resource group, be sure to change their scope in Azure before executing VM scale actions. For best results, change their scope to *shared*.

Next Step

Claim an Azure Billing target in Intersight Workload Optimizer. For details, see this [topic \(on page 84\)](#).

Claiming an Azure Billing Target

The Azure Billing target grants Intersight Workload Optimizer access to billing data from the Azure Cost Details API or a cost export that you set up in Azure. Intersight Workload Optimizer uses this data to visualize historical cloud expenses and discover reservations. The Azure Billing target supports only Enterprise Agreements or Microsoft Customer Agreements bought directly from azure.com.

Before performing this task, be sure you have the required credentials for the service principal. Credentials include the billing account ID, application (client) ID, directory (tenant) ID, and client secret value. You also need to set up the appropriate role to enable access to billing data. If you do not have these credentials or if you have not set up a role, follow the steps outlined in this [topic \(on page 81\)](#).

IMPORTANT:

The Microsoft Enterprise Agreement target has reached end of support and will be removed from the product in a future release.

If you previously added a Microsoft Enterprise Agreement target in the Intersight Workload Optimizer user interface, remove the target before adding the Azure Billing target. It is not possible to have both the Microsoft Enterprise Agreement and Azure Billing targets managing the same subscriptions.

For more information about end of support for the Microsoft Enterprise Agreement target, see this [topic \(on page 100\)](#).

Claiming an Azure Billing Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Public Cloud**.
3. Select **Azure Billing**.
4. Configure the following settings:
 - **Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in Azure.
 - **Billing Account ID**
Specify the billing account ID.
You can find the billing ID in the **Cost Management + Billing** section of the Azure portal. The ID is different for MCA and EA accounts.
 - **MCA Billing Account ID**
In the Azure portal, navigate to **Cost management + Billing**, open your MCA billing account, and then select **Properties**. The screen that displays includes a field for the Billing Account ID.
 - **EA Billing Account ID (Enrollment Number)**
In the Azure portal, navigate to **Cost management + Billing**, open your EA account, and then select **Overview**. The screen that displays includes a field for the Billing Account ID.
 - **Use a cost export**
Specify the cost export name.
 - **Turned off:** Intersight Workload Optimizer will collect data via the Cost Details API.
 - **Turned on:** Intersight Workload Optimizer will collect data via a cost export.
To find the cost export name in the Azure portal, navigate to **Cost management + Billing** and then select **Exports**. A list of the cost exports you have created displays. For more information about setting up cost exports, see [Cost Export Setup \(on page 80\)](#).
 - **Directory (Tenant) ID**
Specify the Directory ID associated with your app registration.
Format: 32 alphanumeric characters separated by hyphens
 - **Application (Client) ID**
Specify the application ID associated with your app registration.
Format: 32 alphanumeric characters separated by hyphens
 - **Client Secret Value**
Specify the client secret value associated with your app registration.

NOTE:

After successfully adding a target, it can take up to 48 hours for any Azure reservation and billing data to display in the Intersight Workload Optimizer user interface.

After Claiming an Azure Billing Target

You have completed the required tasks for connecting to Azure. Intersight Workload Optimizer can now monitor your Azure workloads and billing data, and recommend actions to optimize these workloads at the lowest possible cost. Review the following topics for more information:

- [Azure Monitored Resources \(on page 101\)](#)
- [Azure Actions \(on page 110\)](#)

Reference: Azure Permissions

The service principal that you set up in Azure specifies the permissions that Intersight Workload Optimizer needs to discover and monitor your Azure workloads. Permissions to execute actions from Intersight Workload Optimizer are optional.

Minimum Permissions - Workload Monitoring

The following minimum permissions are required to monitor Azure workloads.

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| Role validation | <ul style="list-style-type: none"> ■ <code>Microsoft.Authorization/roleAssignments/read</code> Validates the role assigned to the Service Principal by checking if it has the minimum required permissions ■ <code>Microsoft.Authorization/roleDefinitions/read</code> Queries the permissions list from the assigned custom role |
| Discovery of subscriptions | <ul style="list-style-type: none"> ■ <code>Microsoft.Resources/subscriptions/read</code> Gets a list of accessible subscriptions for a tenant |
| Discovery of resource groups, locations, and SKUs | <ul style="list-style-type: none"> ■ <code>Microsoft.Resources/subscriptions/locations/read</code> List all locations available for the subscriptions ■ <code>Microsoft.Resources/subscriptions/resourceGroups/read</code> Discovers all resource groups for the subscriptions ■ <code>Microsoft.Compute/skus/read</code> Gets a list of <code>Microsoft.Compute</code> SKUs available for your subscription |
| Discovery of storage accounts | <ul style="list-style-type: none"> ■ <code>Microsoft.Storage/storageAccounts/read</code> Gets a list of storage accounts or gets the properties of a specific storage account |
| Discovery of metrics for various entities | <ul style="list-style-type: none"> ■ <code>Microsoft.Insights/Metrics/Read</code> Reads metrics for various resources from Azure Monitor ■ <code>Microsoft.OperationalInsights/workspaces/read</code> Queries a list of Log Analytics workspaces. Certain metrics (such as VM memory) are fetched from these workspaces. ■ <code>Microsoft.OperationalInsights/workspaces/query/read</code> Allows queries to data (such as metrics) stored in Log Analytics workspaces ■ <code>Microsoft.OperationalInsights/workspaces/query/InsightsMetrics/read</code> Queries the <code>InsightsMetrics</code> Log Analytics table for VM memory metrics. Azure Monitor Agent must be configured to send memory metrics to the table. ■ <code>Microsoft.OperationalInsights/workspaces/query/Perf/read</code> (Only required if the <code>Perf</code> Log Analytics table is configured for VM memory metrics, instead of the <code>InsightsMetrics</code> Log Analytics table) Queries the <code>Perf</code> Log Analytics table for VM memory metrics |
| Discovery of VMs | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/virtualMachines/instanceView/read</code> Gets the detailed runtime status of a VM and its resources ■ <code>Microsoft.Compute/virtualMachines/read</code> Gets the properties of a VM ■ <code>Microsoft.Compute/virtualMachines/extensions/read</code> (Only required if using storage account for VM memory configuration) |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|--|
| | <p>Gets the properties of a VM extension, to detect diagnostics extension before fetching memory metrics from storage account</p> |
| <p>Discovery of VM scale sets and availability sets</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/virtualMachineScaleSets/read</code> Gets the properties of a VM scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/networkInterfaces/read</code> Lists all the network interfaces of a VM scale set and gets their properties ■ <code>Microsoft.Compute/virtualMachineScaleSets/virtualMachines/instanceView/read</code> Retrieves the instance view of a VM in a scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/virtualMachines/read</code> Retrieves the properties of a VM in a scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/virtualMachines/extensions/read</code> (Only required if VM memory is configured onto a storage account table via a diagnostics agent extension) Gets the properties of an extension for a VM in a scale set ■ <code>Microsoft.Compute/availabilitySets/read</code> Lists all availability sets and gets their properties ■ <code>Microsoft.Compute/availabilitySets/vmSizes/read</code> Lists available sizes for creating or updating a VM in an availability set |
| <p>Discovery of reservations</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Capacity/reservationorders/reservations/read</code> Monitors reservations data and reads all reservations ■ <code>Microsoft.Capacity/catalogs/read</code> Reads the catalog of reservations |
| <p>Discovery of volumes</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/disks/read</code> Gets the properties of volumes ■ <code>Microsoft.Storage/storageAccounts/listkeys/action</code> (Only required if there are unmanaged disks that are attached to VMs) Discovers or queries unmanaged attached disks (volumes) in the storage account. Unmanaged disks that are not attached to VMs are not discovered. |
| <p>Discovery of SQL databases (vCore/DTU) and metrics</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Sql/servers/read</code> Lists all SQL servers in this subscription and gets their details ■ <code>Microsoft.Sql/servers/databases/read</code> Lists and gets details about all SQL databases for all SQL servers in the subscription ■ <code>Microsoft.Sql/servers/databases/metrics/read</code> Queries metrics for SQL databases |
| <p>Discovery of resources used in Azure Synapse Analytics</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Synapse/SKUs/read</code> Reads SKU details for a Synapse Analytics Service resource, such as SQL pools ■ <code>Microsoft.Synapse/workspaces/read</code> Reads details about Synapse workspaces |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| | <ul style="list-style-type: none"> ■ <code>Microsoft.Synapse/workspaces/keys/read</code> Gets details of Synapse workspace key definitions ■ <code>Microsoft.Synapse/workspaces/sqlDatabases/read</code> Reads a list of Synapse SQL Analytics databases ■ <code>Microsoft.Synapse/workspaces/sqlPools/read</code> Reads a list of Synapse SQL Analytics pools ■ <code>Microsoft.Synapse/workspaces/sqlPools/dataWarehouseUserActivities/read</code> Reads user activities on Synapse SQL Analytics pools ■ <code>Microsoft.Synapse/workspaces/sqlPools/extensions/read</code> Gets extensions for Synapse SQL Analytics pools ■ <code>Microsoft.Synapse/workspaces/sqlPools/operationStatuses/read</code> Reads the results of asynchronous operations on Synapse SQL Analytics pools ■ <code>Microsoft.Synapse/workspaces/sqlPools/usages/read</code> Reads usage metrics for Synapse SQL Analytics pools ■ <code>Microsoft.Synapse/workspaces/sqlUsages/read</code> Gets usage limits available for Synapse SQL Analytics pools |
| Discovery of App Services (plans/app instances) and metrics | <ul style="list-style-type: none"> ■ <code>Microsoft.Relay/namespaces/HybridConnections/read</code> Lists all Service Bus Hybrid Connections used by web apps ■ <code>Microsoft.Web/geoRegions/Read</code> Gets a list of available geographical regions for App Services ■ <code>Microsoft.Web/serverfarms/Read</code> Lists and gets the properties of App Service plans ■ <code>Microsoft.Web/serverfarms/sites/read</code> Gets a list of web apps that are part of App Service plans ■ <code>Microsoft.Web/serverfarms/skus/read</code> Gets SKUs for App Service plans ■ <code>Microsoft.Web/sites/read</code> Gets the properties of web apps that are part of App Service plans ■ <code>Microsoft.Web/sites/slots/Read</code> Gets the properties of a web app deployment slot ■ <code>Microsoft.Web/serverfarms/metrics/read</code> Queries metrics for App Services plans ■ <code>Microsoft.Web/serverfarms/usages/read</code> Gets usage information for App Service plans ■ <code>Microsoft.Web/sites/metrics/read</code> Gets metrics for web apps that are part of App Service plans ■ <code>Microsoft.Web/sites/usages/read</code> Gets usage information for web apps that are part of App Service plans |
| Discovery of Cosmos DB resources | <ul style="list-style-type: none"> ■ <code>Microsoft.DocumentDB/databaseAccounts/read</code> Gets the properties of a database account |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|--|
| | <ul style="list-style-type: none"> ■ <code>Microsoft.DocumentDB/databaseAccounts/databases/metrics/read</code> Queries metrics for a database account ■ <code>Microsoft.DocumentDB/databaseAccounts/databases/collections/metrics/read</code> Queries metrics for a container ■ <code>Microsoft.DocumentDB/databaseAccounts/metrics/read</code> Queries metrics for a database ■ <code>Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/read</code> Gets the properties of an Apache Cassandra keyspace ■ <code>Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/throughputSettings/read</code> Gets the throughput of an Apache Cassandra keyspace ■ <code>Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/read</code> Gets the properties of an Apache Cassandra table ■ <code>Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/throughputSettings/read</code> Gets the throughput of an Apache Cassandra table ■ <code>Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/read</code> Gets the properties of a MongoDB database ■ <code>Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/throughputSettings/read</code> Gets the throughput of a MongoDB database ■ <code>Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/read</code> Gets the properties of a MongoDB collection ■ <code>Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/throughputSettings/read</code> Gets the throughput of a MongoDB collection ■ <code>Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/read</code> Gets the properties of an Apache Gremlin database ■ <code>Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/throughputSettings/read</code> Gets the throughput of an Apache Gremlin database ■ <code>Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/read</code> Gets the properties of an Apache Gremlin graph ■ <code>Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/throughputSettings/read</code> Gets the throughput of an Apache Gremlin graph ■ <code>Microsoft.DocumentDB/databaseAccounts/tables/read</code> Gets the properties of an Azure table ■ <code>Microsoft.DocumentDB/databaseAccounts/tables/throughputSettings/read</code> |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|--|
| | <p>Gets the throughput of an Azure table</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/read <p>Gets the properties of a NoSQL account database</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/throughputSettings/read <p>Gets the throughput of a NoSQL account database</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/read <p>Get the properties of a NoSQL account container</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/throughputSettings/read <p>Gets the throughput of a NoSQL account container</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/usages/read <p>Gets the storage usage for a database account</p> |
| Discovery of clusters managed by Azure Kubernetes Service (AKS) | <ul style="list-style-type: none"> ■ Microsoft.ContainerService/managedClusters/read Discovers and lists managed container platform clusters ■ Microsoft.ContainerService/managedClusters/agentPools/read Discovers agent pools on managed container platform clusters |
| Discovery of desktop virtualization (VDI) | <ul style="list-style-type: none"> ■ Microsoft.DesktopVirtualization/hostpools/read Discovers Azure Desktop Virtualization host pools ■ Microsoft.DesktopVirtualization/hostpools/sessionhosts/read Discovers session hosts in Azure Desktop Virtualization host pools |
| Discovery of network resources | <ul style="list-style-type: none"> ■ Microsoft.Network/networkInterfaces/read Lists and gets details about network interfaces for resources ■ Microsoft.Network/publicIPAddresses/read Lists and gets details about public IP addresses for resources |
| Discovery of pricing information | <ul style="list-style-type: none"> ■ Microsoft.Commerce/RateCard/read Discovers pricing information from the pay-as-you-go rate card; also returns offer data, resource/meter metadata, and rates for the given subscription ■ Microsoft.Consumption/pricesheets/read Discovers pricing information from an Enterprise Agreement price sheet, and lists the price sheets data for a subscription or a management group |

Minimum Permissions - Action Execution

The following permissions are required only if you want to execute actions for Azure workloads from Intersight Workload Optimizer.

| Intersight Workload Optimizer Functionality | Required Permissions |
|--|--|
| Discovery of locks that could prevent action execution | <ul style="list-style-type: none"> ■ Microsoft.Authorization/locks/read |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| | <p>Lists all locks for a subscription, and creates action prerequisites that may prevent action execution if locks prevent write operations</p> |
| <p>Execution of actions for VMs</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/virtualMachines/deallocate/action</code> Stops a VM to execute a disruptive action; powers off the VM and releases the allocated compute resources ■ <code>Microsoft.Compute/virtualMachines/powerOff/action</code> Suspends a VM by powering it off. The VM will continue to be billed while suspended. ■ <code>Microsoft.Compute/virtualMachines/start/action</code> Restarts a VM that was stopped to execute a disruptive action ■ <code>Microsoft.Compute/virtualMachines/vmSizes/read</code> Lists available sizes that a VM can update to ■ <code>Microsoft.Compute/virtualMachines/write</code> Updates a VM (for example, its size) as part of executing scale actions ■ <code>Microsoft.Network/networkInterfaces/join/action</code> Allows a VM or VM scale set to rejoin its network after a scale action executes, by attaching the network interface to the VM ■ <code>Microsoft.KeyVault/vaults/deploy/action</code> (Only required if a VM to be scaled is using Azure Key Vault) Enables access to secrets in a key vault when deploying Azure resources to the VM ■ Azure Compute Gallery images (Only required during VM scaling execution if VM images are located in a separate Azure Compute Gallery image, such as in a different subscription) <ul style="list-style-type: none"> - <code>Microsoft.Compute/galleries/images/read</code> Gets the properties of Azure Compute Gallery images - <code>Microsoft.Compute/galleries/images/versions/read</code> Gets the versions for Azure Compute Gallery images - <code>Microsoft.Compute/galleries/read</code> Gets Azure Compute Gallery images - <code>Microsoft.Compute/images/read</code> Gets the properties of the image |
| <p>Execution of actions for VM scale sets and availability sets</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/virtualMachineScaleSets/deallocate/action</code> Stops a VM scale set to execute a disruptive action; powers off and releases the compute resources of the instances used by the VM scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/start/action</code> Restarts the VMs that were stopped to execute a disruptive action on VM scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/vmSizes/read</code> Lists available sizes for creating or updating a VM in a scale set ■ <code>Microsoft.Compute/virtualMachineScaleSets/write</code> Updates a VM scale set (for example, its size) as part of executing scale actions for VM scale sets ■ <code>Microsoft.Insights/AutoscaleSettings/Write</code> |

| Intersight Workload Optimizer Functionality | Required Permissions |
|--|---|
| | <p>Updates an autoscale setting as part of scale action execution</p> |
| <p>Execution of actions for Azure Kubernetes Service (AKS) nodes (VMs)</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.ContainerService/managedClusters/listClusterAdminCredential/action</code> Lists the <code>clusterAdmin</code> credentials of a managed cluster ■ <code>Microsoft.Compute/virtualMachineScaleSets/virtualMachines/delete</code> Deletes a specific VM in a VM scale set ■ <code>Microsoft.Compute/virtualMachines/delete</code> Deletes a specific VM ■ <code>Microsoft.ContainerService/managedClusters/write</code> Creates a new managed cluster or updates an existing one ■ <code>Microsoft.ContainerService/managedClusters/agentPools/write</code> Creates or updates an agent pool in the specified managed cluster ■ <code>Microsoft.OperationalInsights/workspaces/sharedkeys/read</code> Gets the shared keys for the Log Analytics workspace. These keys are used to connect Microsoft Operational Insights agents to the workspace. ■ <code>Microsoft.OperationsManagement/solutions/write</code> Create a new OMS solution ■ <code>Microsoft.OperationsManagement/solutions/read</code> Gets an exiting OMS solution |
| <p>Execution of actions for volumes</p> | <ul style="list-style-type: none"> ■ <code>Microsoft.Compute/disks/write</code> Resizes or changes the storage tier of volumes Executes scale actions for volumes and re-attaches volumes to VMs after scaling ■ <code>Microsoft.Compute/disks/delete</code> Deletes unattached volumes for managed disks ■ <code>Microsoft.Storage/storageAccounts/blobServices/containers/blobs/delete</code> Deletes unmanaged disks that became unattached after the deletion of unattached volumes ■ <code>Microsoft.Storage/storageAccounts/blobServices/containers/blobs/read</code> Checks if a page blob exists before deleting unmanaged disks that became unattached, and returns the properties of an existing page blob ■ <code>Microsoft.Storage/storageAccounts/blobServices/containers/read</code> Lists blob containers Finds containers hosting unmanaged disks that became unattached and are to be deleted ■ Recovery Services Vault (Only required if volumes are used for disaster recovery) Prevents the deletion of volumes used for disaster recovery, even if they become unattached <ul style="list-style-type: none"> - <code>Microsoft.RecoveryServices/Vaults/read</code> |

| Intersight Workload Optimizer Functionality | Required Permissions |
|--|--|
| | <p>Gets a list of Recovery Services Vaults containing replicated disks</p> <ul style="list-style-type: none"> - Microsoft.RecoveryServices/vaults/replicationProtectedItems/read <p>Reads the Disk IDs of any protected items in the vaults</p> |
| Execution of actions for SQL databases (vCore and DTU) | <ul style="list-style-type: none"> ■ Microsoft.Sql/servers/databases/write Executes scale actions for DTU and vCore databases, and updates database properties ■ Microsoft.Sql/servers/databases/pause/action Pauses a database as part of executing a scale or suspend action ■ Microsoft.Sql/servers/databases/resume/action Resumes a paused database as part of executing a scale or suspend action |
| Execution of actions for dedicated SQL pools for Azure Synapse Analytics | <ul style="list-style-type: none"> ■ Microsoft.Synapse/workspaces/sqlPools/pause/action Suspends or stops a Synapse SQL Analytics pool ■ Microsoft.Synapse/workspaces/sqlPools/resume/action Resumes a suspended or stopped Synapse SQL Analytics pool |
| Execution of actions for App Services (plans) | <ul style="list-style-type: none"> ■ Microsoft.Web/serverfarms/Delete Deletes an empty App Service plan (one that is not hosting any running apps) ■ Microsoft.Web/serverfarms/Write Updates an App Service plan as part of a scale action |
| Execution of actions for Cosmos DB databases and document collections | <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/delete Deletes an Apache Cassandra keyspace ■ Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/throughputSettings/write Updates the throughput of an Apache Cassandra keyspace ■ Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/throughputSettings/write Updates the throughput of an Apache Cassandra table ■ Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/throughputSettings/write Updates the throughput of a MongoDB collection ■ Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/delete Deletes a MongoDB database ■ Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/throughputSettings/write Updates the throughput of a MongoDB database ■ Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/delete Deletes an Apache Gremlin database ■ Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/throughputSettings/write Updates the throughput of an Apache Gremlin database ■ Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/throughputSettings/write |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|--|
| | <p>Updates the throughput of an Apache Gremlin graph</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/tables/throughputSettings/write <p>Updates the throughput of an Azure table</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/delete <p>Deletes a NoSQL account database</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/throughputSettings/write <p>Updates the throughput of a NoSQL account database</p> <ul style="list-style-type: none"> ■ Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/throughputSettings/write <p>Updates the throughput of NoSQL account container</p> |

Sample JSON - Minimum Permissions for Workload Monitoring

In Azure, you can create a custom role that specifies the permissions that Intersight Workload Optimizer needs to monitor workloads in your subscriptions.

When you create the role, you have the option of uploading a JSON file that specifies the permissions and settings for the role. You can copy the content in this section to the JSON file.

```
{
  "properties": {
    "roleName": "<RoleName>",
    "description": "",
    "assignableScopes": [
      "/subscriptions/<Subscription_ID>"
    ],
    "permissions": [
      {
        "actions": [
          "Microsoft.Authorization/roleAssignments/read",
          "Microsoft.Authorization/roleDefinitions/read",
          "Microsoft.Capacity/catalogs/read",
          "Microsoft.Capacity/reservationorders/reservations/read",
          "Microsoft.Commerce/RateCard/read",
          "Microsoft.Compute/availabilitySets/read",
          "Microsoft.Compute/availabilitySets/vmSizes/read",
          "Microsoft.Compute/disks/read",
          "Microsoft.Compute/skus/read",
          "Microsoft.Compute/virtualMachines/extensions/read",
          "Microsoft.Compute/virtualMachines/instanceView/read",
          "Microsoft.Compute/virtualMachines/read",
          "Microsoft.Compute/virtualMachineScaleSets/networkInterfaces/read",
          "Microsoft.Compute/virtualMachineScaleSets/read",
          "Microsoft.Compute/virtualMachineScaleSets/virtualMachines/extensions/read",
          "Microsoft.Compute/virtualMachineScaleSets/virtualMachines/instanceView/read",
          "Microsoft.Compute/virtualMachineScaleSets/virtualMachines/read",
          "Microsoft.Consumption/pricesheets/read",
          "Microsoft.ContainerService/managedClusters/agentPools/read",
          "Microsoft.ContainerService/managedClusters/read",
```

```

"Microsoft.DesktopVirtualization/hostpools/read",
"Microsoft.DesktopVirtualization/hostpools/sessionhosts/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/throughputSettings/r
ead",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/databases/collections/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/databases/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/throughputSettings/rea
d",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/read",
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/throughputSetting
s/read",
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/read",
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/read",
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/read",
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/throughputSettings/rea
d",
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/read",
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/tables/read",
"Microsoft.DocumentDB/databaseAccounts/tables/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/usages/read",
"Microsoft.Insights/Metrics/Read",
"Microsoft.Network/networkInterfaces/read",
"Microsoft.Network/publicIPAddresses/read",
"Microsoft.OperationalInsights/workspaces/query/InsightsMetrics/read",
"Microsoft.OperationalInsights/workspaces/query/Perf/read",
"Microsoft.OperationalInsights/workspaces/query/read",
"Microsoft.OperationalInsights/workspaces/read",
"Microsoft.Relay/namespaces/HybridConnections/read",
"Microsoft.Resources/subscriptions/locations/read",
"Microsoft.Resources/subscriptions/read",
"Microsoft.Resources/subscriptions/resourceGroups/read",
"Microsoft.Sql/servers/databases/metrics/read",
"Microsoft.Sql/servers/databases/read",
"Microsoft.Sql/servers/read",
"Microsoft.Storage/storageAccounts/listkeys/action",
"Microsoft.Storage/storageAccounts/read",
"Microsoft.Synapse/SKUs/read",
"Microsoft.Synapse/workspaces/keys/read",
"Microsoft.Synapse/workspaces/read",
"Microsoft.Synapse/workspaces/sqlDatabases/read",
"Microsoft.Synapse/workspaces/sqlPools/dataWarehouseUserActivities/read",
"Microsoft.Synapse/workspaces/sqlPools/extensions/read",
"Microsoft.Synapse/workspaces/sqlPools/operationStatuses/read",
"Microsoft.Synapse/workspaces/sqlPools/read",
"Microsoft.Synapse/workspaces/sqlPools/usages/read",
    
```

```

        "Microsoft.Synapse/workspaces/sqlUsages/read" ,
        "Microsoft.Web/geoRegions/Read" ,
        "Microsoft.Web/serverfarms/metrics/read" ,
        "Microsoft.Web/serverfarms/Read" ,
        "Microsoft.Web/serverfarms/sites/read" ,
        "Microsoft.Web/serverfarms/skus/read" ,
        "Microsoft.Web/serverfarms/usages/read" ,
        "Microsoft.Web/sites/metrics/read" ,
        "Microsoft.Web/sites/read" ,
        "Microsoft.Web/sites/slots/Read" ,
        "Microsoft.Web/sites/usages/read"
    ],
    "notActions": [],
    "dataActions": [],
    "notDataActions": []
}
]
}
}
}

```

NOTE:

Be sure to update the following information in the JSON file:

- *<RoleName>* - Specify your preferred name for the custom role.
- *<Subscription_ID>* - Specify the ID of the subscription that Intersight Workload Optimizer will manage.

Sample JSON - Minimum Permissions for Workload Monitoring and Action Execution

In Azure, you can create a custom role that specifies the permissions that Intersight Workload Optimizer needs to monitor workloads in your subscriptions and execute actions for these workloads.

When you create the role, you have the option of uploading a JSON file that specifies the permissions and settings for the role. You can copy the content in this section to the JSON file.

```

{
  "properties": {
    "roleName": "<RoleName>",
    "description": "",
    "assignableScopes": [
      "/subscriptions/<Subscription_ID>"
    ],
    "permissions": [
      {
        "actions": [
          "Microsoft.Authorization/locks/read",
          "Microsoft.Authorization/roleAssignments/read",
          "Microsoft.Authorization/roleDefinitions/read",
          "Microsoft.Capacity/catalogs/read",
          "Microsoft.Capacity/reservationorders/reservations/read",
          "Microsoft.Commerce/RateCard/read",
          "Microsoft.Compute/availabilitySets/read",
          "Microsoft.Compute/availabilitySets/vmSizes/read",
          "Microsoft.Compute/disks/delete",
          "Microsoft.Compute/disks/read",
          "Microsoft.Compute/disks/write",
          "Microsoft.Compute/galleries/images/read",

```

```

"Microsoft.Compute/galleries/images/versions/read",
"Microsoft.Compute/galleries/read",
"Microsoft.Compute/images/read",
"Microsoft.Compute/skus/read",
"Microsoft.Compute/virtualMachineScaleSets/deallocate/action",
"Microsoft.Compute/virtualMachineScaleSets/networkInterfaces/read",
"Microsoft.Compute/virtualMachineScaleSets/read",
"Microsoft.Compute/virtualMachineScaleSets/start/action",
"Microsoft.Compute/virtualMachineScaleSets/virtualMachines/delete",
"Microsoft.Compute/virtualMachineScaleSets/virtualMachines/extensions/read",
"Microsoft.Compute/virtualMachineScaleSets/virtualMachines/instanceView/read",
"Microsoft.Compute/virtualMachineScaleSets/virtualMachines/read",
"Microsoft.Compute/virtualMachineScaleSets/vmSizes/read",
"Microsoft.Compute/virtualMachineScaleSets/write",
"Microsoft.Compute/virtualMachines/deallocate/action",
"Microsoft.Compute/virtualMachines/delete",
"Microsoft.Compute/virtualMachines/extensions/read",
"Microsoft.Compute/virtualMachines/instanceView/read",
"Microsoft.Compute/virtualMachines/powerOff/action",
"Microsoft.Compute/virtualMachines/read",
"Microsoft.Compute/virtualMachines/start/action",
"Microsoft.Compute/virtualMachines/vmSizes/read",
"Microsoft.Compute/virtualMachines/write",
"Microsoft.Consumption/pricesheets/read",
"Microsoft.ContainerService/managedClusters/agentPools/read",
"Microsoft.ContainerService/managedClusters/agentPools/write",
"Microsoft.ContainerService/managedClusters/listClusterAdminCredential/action",
"Microsoft.ContainerService/managedClusters/read",
"Microsoft.ContainerService/managedClusters/write",
"Microsoft.DesktopVirtualization/hostpools/read",
"Microsoft.DesktopVirtualization/hostpools/sessionhosts/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/delete",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/throughputSettings/r
ead",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/tables/throughputSettings/w
rite",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/cassandraKeyspaces/throughputSettings/write",
"Microsoft.DocumentDB/databaseAccounts/databases/collections/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/databases/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/delete",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/throughputSettings/rea
d",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/graphs/throughputSettings/wri
te",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/throughputSettings/read",
"Microsoft.DocumentDB/databaseAccounts/gremlinDatabases/throughputSettings/write",
"Microsoft.DocumentDB/databaseAccounts/metrics/read",
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/read",

```

```

s/read" ,
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/collections/throughputSetting
s/write" ,
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/delete" ,
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/read" ,
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/throughputSettings/read" ,
"Microsoft.DocumentDB/databaseAccounts/mongodbDatabases/throughputSettings/write" ,
"Microsoft.DocumentDB/databaseAccounts/read" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/read" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/throughputSettings/rea
d" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/containers/throughputSettings/wri
te" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/delete" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/read" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/throughputSettings/read" ,
"Microsoft.DocumentDB/databaseAccounts/sqlDatabases/throughputSettings/write" ,
"Microsoft.DocumentDB/databaseAccounts/tables/read" ,
"Microsoft.DocumentDB/databaseAccounts/tables/throughputSettings/read" ,
"Microsoft.DocumentDB/databaseAccounts/tables/throughputSettings/write" ,
"Microsoft.DocumentDB/databaseAccounts/usages/read" ,
"Microsoft.Insights/AutoscaleSettings/Write" ,
"Microsoft.Insights/Metrics/Read" ,
"Microsoft.KeyVault/vaults/deploy/action" ,
"Microsoft.Migrate/migrateprojects/read" ,
"Microsoft.Migrate/migrateprojects/solutions/getconfig/action" ,
"Microsoft.Migrate/migrateprojects/solutions/read" ,
"Microsoft.Network/networkInterfaces/join/action" ,
"Microsoft.Network/networkInterfaces/read" ,
"Microsoft.Network/publicIPAddresses/read" ,
"Microsoft.OperationalInsights/workspaces/query/InsightsMetrics/read" ,
"Microsoft.OperationalInsights/workspaces/query/Perf/read" ,
"Microsoft.OperationalInsights/workspaces/query/read" ,
"Microsoft.OperationalInsights/workspaces/read" ,
"Microsoft.OperationalInsights/workspaces/sharedkeys/read" ,
"Microsoft.OperationsManagement/solutions/read" ,
"Microsoft.OperationsManagement/solutions/write" ,
"Microsoft.RecoveryServices/Vaults/read" ,
"Microsoft.RecoveryServices/vaults/replicationProtectedItems/read" ,
"Microsoft.Relay/namespaces/HybridConnections/read" ,
"Microsoft.Resources/subscriptions/locations/read" ,
"Microsoft.Resources/subscriptions/read" ,
"Microsoft.Resources/subscriptions/resourceGroups/read" ,
"Microsoft.Sql/servers/databases/metrics/read" ,
"Microsoft.Sql/servers/databases/pause/action" ,
"Microsoft.Sql/servers/databases/read" ,
"Microsoft.Sql/servers/databases/resume/action" ,
"Microsoft.Sql/servers/databases/write" ,
"Microsoft.Sql/servers/read" ,
"Microsoft.Storage/storageAccounts/blobServices/containers/read" ,
"Microsoft.Storage/storageAccounts/listkeys/action" ,
"Microsoft.Storage/storageAccounts/read" ,
"Microsoft.Synapse/SKUs/read" ,

```


| | |
|---|--|
| Intersight Workload Optimizer Functionality | Required Permissions |
| | Gets the configuration of a Migrate project solution |

Reference: Level of Support for Azure MCA and EA Accounts

Intersight Workload Optimizer uses a service principal to connect to your Microsoft Customer Agreement (MCA) or Enterprise Agreement (EA) accounts and discover billing data.

Currently, Intersight Workload Optimizer support for MCA and EA accounts is *identical* in the following areas:

- Discovery of pricing data
 - Pricing data is available for VMs, volumes, SQL databases (DTU and vCore), dedicated SQL pools (for Azure Synapse Analytics), and App Service plans (web apps, logic apps, and function apps).
- Usage of pricing data when recommending actions
 - All actions for Azure workloads are supported, including actions to optimize/buy reservations.
- Support for the Buy VM Reservations plan
 - The plan is fully supported.
- Scope for [cost exports \(on page 80\)](#) (if used to collect billing data)
 - The following scopes are *not* supported:
 - Subscription
 - Management Group
 - Resource Group

The level of support *differs* in the following areas.

| Item | MCA | EA |
|---|--|--|
| Required role | Billing Account Reader role | Enrollment Reader role |
| Discovery of Azure reservations | All reservations charged under the billing account are discovered. However, the percentage of workloads covered by reservations is currently not discovered. | All reservations charged under the billing account are discovered. |
| Discovery of Azure reservations costs in non-US dollar currencies | Enabled by default | Not enabled by default. To enable discovery, follow the instructions in this topic (on page 108) . |
| Scope for cost exports (on page 80) (if used to collect billing data) | Cost exports must be created at the <i>Billing Profile</i> scope. If you previously set up a cost export at the <i>Billing Account</i> scope, you must delete the cost export and then create a new one at the Billing Profile scope. Create this cost export for each active Billing Profile that falls under your MCA Billing Account. All cost exports you create for your active Billing Profiles are required to have the same name. | Cost exports must be created at the <i>Billing Account</i> scope. |

Notice: Microsoft Enterprise Agreement Target

The Microsoft Enterprise Agreement target has reached end of support and will be removed from the product in a future release.

Who is impacted by end of support?

End of support impacts customers who added, or plan to add, a Microsoft Enterprise Agreement target.

- Customers who added a Microsoft Enterprise Agreement target are advised to immediately remove the target from the user interface, and then set up an Azure Billing target for use with Intersight Workload Optimizer.
- Customers planning to add a Microsoft Enterprise Agreement target should set up an Azure Billing target instead.

The Azure Billing target leverages the new cost management APIs and cost exports recommended by [Microsoft](#).

What happens now that the Microsoft Enterprise Agreement target has reached end of support?

Any existing Microsoft Enterprise Agreement target remains in the user interface. If you encounter issues with the target, your Intersight Workload Optimizer representative will advise that you first remove the target and then set up an Azure Billing target. It is not possible to have both the Microsoft Enterprise Agreement and Azure Billing targets managing the same subscriptions.

What happens after the target is removed from the product?

- The Microsoft Enterprise Agreement target no longer appears in the list of targets in the Intersight Workload Optimizer user interface.
- Any existing Microsoft Enterprise Agreement target remains in the user interface but will no longer be discovered or validated.

How do I set up an Azure Billing target for use with Intersight Workload Optimizer?

Perform these tasks:

1. [Decide whether to use the Cost Details API or a cost export. \(on page 80\)](#)
2. [Set up a service principal for billing data monitoring in Azure. \(on page 81\)](#)
3. [Claim an Azure Billing target in Intersight Workload Optimizer. \(on page 84\)](#)

Azure Monitored Resources

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| Azure | Intersight Workload Optimizer |
|--|-------------------------------|
| Virtual Machine (VM) | Virtual Machine (VM) |
| Disk (Managed) | Volume |
| App Service - Plan | Virtual Machine Spec |
| App Service - Web App | App Component Spec |
| SQL Database (vCore or DTU) | Database |
| Synapse Analytics (Dedicated SQL Pool) | Database |
| Cosmos DB - Account | Database Server |
| Cosmos DB - Database | Database |
| Cosmos DB - Container | Document Collection |
| Region | Region |

Points to consider:

- Intersight Workload Optimizer supports discovery and management of entities in certain Azure regions. For details, see [Supported Azure Regions \(on page 104\)](#).

- When you first configure an Azure target, under some circumstances the target might show a `No Quotas Available` status. This means that Intersight Workload Optimizer cannot discover the available templates. This can happen when you initially set up the Azure account and you have not enabled any providers. If this occurs, you can install a single VM in your cloud subscription to make quotas available.
- An Azure subscription can use locked storage or locked resource groups. For such subscriptions, Intersight Workload Optimizer discovers incomplete data. Locked resources affect Intersight Workload Optimizer discovery in either of these scenarios:
 - Locked resource group

Intersight Workload Optimizer discovers all the entities in the resource group, but does not discover the resource group itself. For example, in the Top Accounts chart, the Resource Groups field will show no resource groups for a subscription that has a locked resource group.
 - Locked storage

Intersight Workload Optimizer discovers all the entities in the resource group except the locked storage. It also discovers the resource group.

Monitored Resources for Virtual Machines

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

It is highly recommended that you enable collection of metrics in your environment. Enabling metrics allows Intersight Workload Optimizer to generate scale actions to optimize VM resource usage. For Intersight Workload Optimizer to collect metrics, you must enable the collection of these metrics on the VMs in your environment.

For details, see [Azure Memory Metrics Collection \(on page 106\)](#).
- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.
- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.
- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- I/O Throughput

I/O Throughput is the measurement of an entity's throughput to the underlying storage.

Monitored Resources for Volumes

Intersight Workload Optimizer monitors the following resources:

NOTE:

Intersight Workload Optimizer discovers and optimizes Azure *managed* volumes.

According to this Microsoft [article](#), *unmanaged* volumes are deprecated and will be fully retired in 2025. In response, Intersight Workload Optimizer no longer discovers or monitors unmanaged volumes that are not attached to any VM. Unmanaged volumes that are attached to VMs will continue to be discovered and displayed in the user interface for your reference, but no action will be generated for these volumes.

- Storage Amount

Storage Amount is the storage capacity (disk size) of a volume.

Intersight Workload Optimizer discovers Storage Amount, but does not monitor utilization.

For a Kubeturbo (container) deployment that includes volumes, Kubeturbo monitors Storage Amount utilization for the volumes. You can view utilization information in the Capacity and Usage chart.
- Storage Access (IOPS)

Storage Access, also known as IOPS, is the measurement of IOPS capacity that is in use.

- I/O Throughput

I/O Throughput is the measurement of I/O throughput capacity that is in use.

Monitored Resources for Virtual Machine Specs (App Service Plans)

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Number of Replicas

Number of Replicas is the total number of VM instances underlying an App Service plan.

Monitored Resources for App Component Specs (App Service Instances)

Intersight Workload Optimizer monitors the following resources:

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

Monitored Resources for Database Servers (Cosmos DB Accounts)

Intersight Workload Optimizer monitors the following resources:

- Request Unit (RU)

Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

Monitored Resources for Databases

The resources that Intersight Workload Optimizer can monitor depend on the pricing model in place for the given database entity.

- SQL Database - DTU Pricing Model

- DTU

DTU is the measurement of compute capacity for the database. DTU represents CPU, memory, and IOPS/IO Throughput bundled as a single commodity.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- SQL Database - vCore Pricing Model

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- I/O Throughput
I/O Throughput is the measurement of an entity's throughput to the underlying storage.
- Dedicated SQL Pool (for Azure Synapse Analytics)
 - DWU
DWU (Data Warehousing Unit) is the measurement of compute capacity for the dedicated SQL pool. DWU represents CPU, memory, and IO Throughput bundled as a single commodity.
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Connection
Connection is the measurement of database connections utilized by applications.
- Cosmos DB Database
 - Request Unit (RU)
Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.

Monitored Resources for Document Collections (Cosmos DB Containers)

Intersight Workload Optimizer monitors the following resources:

- Request Unit (RU)
Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.

Supported Azure Regions

Intersight Workload Optimizer supports discovery and management of entities in the following Azure regions:

| Region Code | Region Name | Notes |
|----------------|------------------|-------|
| eastus | East US | |
| eastus2 | East US 2 | |
| centralus | Central US | |
| northcentralus | North Central US | |
| southcentralus | South Central US | |
| westcentralus | West Central US | |
| westus | West US | |
| westus2 | West US 2 | |
| westus3 | West US 3 | |
| canadaeast | Canada East | |
| canadacentral | Canada Central | |
| brazilsouth | Brazil South | |
| northeurope | North Europe | |

| Region Code | Region Name | Notes |
|--------------------|----------------------|-------|
| westeurope | West Europe | |
| francecentral | France Central | |
| ukwest | UK West | |
| uksouth | UK South | |
| germanywestcentral | Germany West Central | |
| norwayeast | Norway East | |
| switzerlandnorth | Switzerland North | |
| eastasia | East Asia | |
| southeastasia | Southeast Asia | |
| australiaeast | Australia East | |
| australiasoutheast | Australia Southeast | |
| australiacentral | Australia Central | |
| centralindia | Central India | |
| southindia | South India | |
| westindia | West India | |
| japaneast | Japan East | |
| japanwest | Japan West | |
| koreacentral | Korea Central | |
| koreasouth | Korea South | |
| uaenorth | UAE North | |
| southafricanorth | South Africa North | |

Support for Azure App Service

Azure App Service is an HTTP-based service for hosting apps. With Azure App Service, app developers can easily create enterprise-ready apps and deploy them on a scalable and reliable cloud infrastructure.

Azure App Service offers several types of apps, including web apps, mobile apps, API apps, and logic apps. Each app runs as a set of *app instances* and is associated with a *plan* that defines compute resources (CPU, memory, and storage) available to the app.

When you add an Azure account as a target:

- Intersight Workload Optimizer discovers all the plans in that account, except App Service Environment v3 I4, I5, and I6. Plans appear as 'Virtual Machine Spec' entities in the supply chain.
- For plans associated with *web apps*, Intersight Workload Optimizer discovers the related app instances. In the supply chain, app instances appear as 'App Component Spec' entities. Intersight Workload Optimizer generates actions to scale these plans to optimize app performance.
- For plans associated with the other types of apps, Intersight Workload Optimizer does not generate scale actions or discover the related app instances.
- For plans that are not associated with any type of app, Intersight Workload Optimizer generates delete actions as a cost-saving measure.

For details about scale and delete actions, see [Virtual Machine Spec \(on page 287\)](#).

To discover plans and app instances, you must provide permissions to support all the actions you want to perform. For a list of permissions, see [Azure Permissions \(on page 85\)](#).

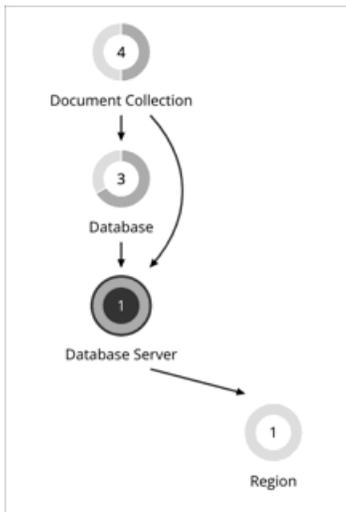
Support for Azure Cosmos DB

Azure Cosmos DB is a fully managed platform-as-a-service (PaaS) for modern app development. To deploy Cosmos DB, you create a Cosmos DB account in an Azure resource group in your subscription, and then create databases and containers within that account.

When you add Azure targets, Intersight Workload Optimizer discovers the accounts, databases and containers that make up your Cosmos DB deployment, and then maps them to the corresponding entities in the supply chain.

| | |
|-----------|---|
| Cosmos DB | Intersight Workload Optimizer |
| Account | Database Server (on page 307) |
| Database | Database (on page 317) |
| Container | Document Collection (on page 322) |

The following image shows how your Cosmos DB environment is represented in the supply chain.



An account has a total throughput limit that represents capacity, but actual throughput (measured in Request Units or RUs) is provisioned at the database or container (document collection) level.

Intersight Workload Optimizer optimizes databases and containers. It uses percentile calculations to measure Request Unit (RU) utilization for these entities and then recommends actions to scale RUs to optimize performance and costs.

Intersight Workload Optimizer also recommends cost-saving actions that reconfigure databases to remove unused provisioned throughput or delete databases that are not hosting document collections.

Intersight Workload Optimizer requires specific permissions to discover your Cosmos DB deployment. For details, see [Azure Permissions \(on page 85\)](#).

Azure VM Memory Metrics Collection

It is highly recommended that you enable collection of metrics in your environment. Enabling metrics allows Intersight Workload Optimizer to generate scale actions to optimize VM resource usage. For Intersight Workload Optimizer to collect metrics, you must enable the collection of these metrics on the VMs in your environment.

Intersight Workload Optimizer collects memory metrics for Azure VMs using the following mechanisms, in order:

1. Log Analytics Workspace - InsightsMetrics table

This is the preferred collection method. This mechanism is used when a VM is configured with the Azure Monitor Agent (AMA). This agent collects guest OS memory metrics and sends it to the Log Analytics Workspace `InsightsMetrics` table. Intersight Workload Optimizer then collects available memory metrics by querying this table.

For more information on the Azure Monitor Agent, see the [Azure documentation](#).

2. **Log Analytics Workspace - Perf table**

This mechanism is used when a VM is configured with the Log Analytics agent (also known as MMA and OMS) and is connected to a Log Analytics Workspace. This agent sends guest OS memory metrics to the Log Analytics Workspace Perf table. Intersight Workload Optimizer then collects available memory metrics by querying this table.

3. **Storage Account SDK Based Diagnostics Monitoring**

This legacy mechanism installs an Azure Diagnostics extension on the VM, which sends guest OS level memory metrics to a designated Storage Account table. During the discovery process, Intersight Workload Optimizer collects VM memory from these tables. If this mechanism is being used, Intersight Workload Optimizer needs the Storage Account Contributor to access the Storage Account keys and establish a connection that can retrieve VM memory statistics.

4. **Metrics REST API**

If memory cannot be collected using the previously-mentioned mechanisms, Intersight Workload Optimizer tries to collect host OS based memory for the VM via REST API calls, which is the same mechanism used for collecting other VM metrics, such as Percentage CPU usage. Intersight Workload Optimizer collects VM memory and generates memory-driven scale actions where appropriate.

NOTE:

Azure provides available memory (in bytes) via REST API in preview mode, or through the Azure Portal UI under the VM's **Monitoring > Metrics** section.

Azure Memory Source Groups in Intersight Workload Optimizer

In Intersight Workload Optimizer, Azure Memory Source groups reflect the source of VM memory metrics. To search for a specific Azure Memory Source group in the Intersight Workload Optimizer user interface, open the Search page, choose **Groups**, and enter `memory` in the search field. Each group contains the list of VMs for which memory was collected through a specific mechanism. These are dynamic groups that are continuously updated based on each discovery.

The Azure Memory Monitoring Unavailable group contains VMs for which memory could not be collected using any of the collection mechanisms. This group may contain VMs that are not in a RUNNING state. For these VMs, metrics data is often not available and memory-driven scale actions are not generated. However, other scale actions may be generated.

The screenshot shows the search interface with a search bar containing 'memory'. The left sidebar lists navigation options, with 'Groups' selected. The main area displays a table of search results:

| Group Name | Count | Type | Actions |
|---|-------|--------|---------|
| Azure Memory Monitoring Available (Log Analytics InsightsMetrics) | 78 | Static | > |
| Azure Memory Monitoring Available (Log Analytics Perf) | 2 | Static | > |
| Azure Memory Monitoring Available (Metrics API) | 55 | Static | > |
| Azure Memory Monitoring Available (Storage SDK Monitoring) | 6 | Static | > |
| Azure Memory Monitoring Unavailable | 52 | Static | > |

Azure VM Metrics Collection in Batch Mode

Intersight Workload Optimizer can collect memory, CPU, and other metrics for a large number of VMs when you enable VM metric collection in batch mode.

VM metric collection in batch mode is *optional* and is *disabled* by default. When enabled, Intersight Workload Optimizer makes a single API call to collect metrics for 50 VMs (this number is configurable), instead of a single API call for each VM. This can

reduce the likelihood of Azure throttling API requests from Intersight Workload Optimizer, which happens when the number of API calls exceeds a certain limit. Throttling can sometimes result in missing metrics in Intersight Workload Optimizer.

To enable VM metric collection in batch mode:

1. Enable the following property for the `mediation-azure` probe.

```
metrics.batch.api.enabled
```

In the Intersight Workload Optimizer Swagger API, set the property to `true` on the probe of type `Azure Subscription`.

Contact your Intersight Workload Optimizer representative if you need assistance setting the probe property.

2. Assign the `Monitoring Reader` role to the subscription containing the VMs for which metrics will be collected in batches.

Azure requires this role to enable the collection of metrics through the Azure Batch APIs. This role can view all monitoring data in a subscription but cannot modify any resources or edit any settings related to monitoring resources. For details about roles, see the [Azure documentation](#).

Discovery of Azure Reservations Costs in Non-US Dollar Currencies

Intersight Workload Optimizer discover costs for Azure reservations through the [Azure Retail API](#) and your billing accounts. By default, the API returns cost information in US dollars (USD), while your billing accounts return cost information in your local currency.

If you use Azure Enterprise Agreement and your local currency is one of the non-US dollar currencies listed in the [Azure documentation](#), Intersight Workload Optimizer might sometimes calculate unrealistic savings values.

NOTE:

By default, Intersight Workload Optimizer uses the dollar symbol (\$) when displaying the costs and savings that it discovers or calculates for your cloud workloads. You can set a different symbol to match your preferred currency. For example, if your cloud provider bills you in euros, change the currency symbol to €.

Go to **Settings > Billing and Costs > Currency** to change the symbol. Note that currency symbols are for display purposes only. Intersight Workload Optimizer does not convert monetary amounts when you switch symbols.

To enable the discovery of Azure reservations costs in your local currency, perform the following steps.

1. Get the probe UUID for Azure Pricing.
 - a. Log in to the Intersight Workload Optimizer user interface and then navigate to `https://<your_instance_address>/apidoc`.
 - b. Expand [INTERNAL USE - NOT SUPPORTED].
 - c. Expand `GET /probes`, click **Try it out**, and then click **Execute**.
 - d. Find the Azure Pricing probe.

GET /probes Get a list of all probes.

Parameters Cancel

No parameters

Execute Clear

Responses Response content type: application/json

Curl

```
curl -X 'GET' \
  'https://support.turbonomic.io/vmturbo/rest/probes' \
  -H 'accept: application/json'
```

Request URL

```
https://support.turbonomic.io/vmturbo/rest/probes
```

Server response

Code Details

200

Response body

```
{
  "isMultiline": false,
  "isTargetDisplayName": false,
  "valueType": "STRING",
  "description": "Password to use to connect to a proxy",
  "verificationRegex": ".*"
},
{
  "displayName": "Secure Proxy Connection",
  "name": "secureProxy",
  "defaultValue": "false",
  "isMandatory": false,
  "isSecret": false,
  "isMultiline": false,
  "isTargetDisplayName": false,
  "valueType": "BOOLEAN",
  "description": "Use SSL to connect to the proxy host",
  "verificationRegex": "(true|false)"
},
{
  "type": "Azure Pricing"
}
```

- e. Copy the UUID for the Azure Pricing probe.

GET /probes Get a list of all probes.

Parameters Cancel

No parameters

Execute Clear

Responses Response content type: application/json

Curl

```
curl -X 'GET' \
  'https://support.turbonomic.io/vmturbo/rest/probes' \
  -H 'accept: application/json'
```

Request URL

```
https://support.turbonomic.io/vmturbo/rest/probes
```

Server response

Code Details

200

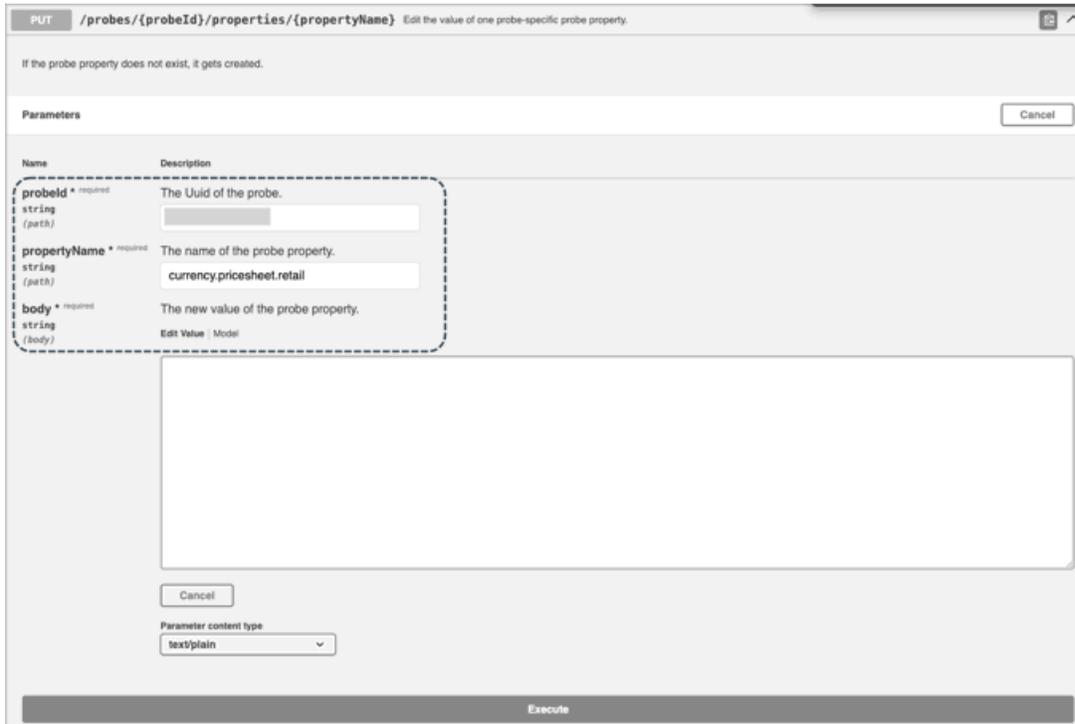
Response body

```
{
  "readOnly": false
},
{
  "uid": "XXXXXXXX-XXXX-XXXX-XXXX-XXXXXXXXXXXX",
  "category": "Public-Cloud",
  "isProbeRegistered": true,
  "identifyingFields": [
    "mcaBillingAccountId",
    "mcaBillingProfileId",
    "planId"
  ],
  "inputFields": [
    {

```

2. Set your local currency.
 - a. Expand PUT /probes/{probeId}/properties/{propertyName} and click **Try it out**.

- b. Configure the following settings:



PUT /probes/{probeId}/properties/{propertyName} Edit the value of one probe-specific probe property.

If the probe property does not exist, it gets created.

Parameters Cancel

| Name | Description |
|---|---|
| probeid * required string (path) | The Uuid of the probe. <input type="text"/> |
| propertyName * required string (path) | The name of the probe property. <input type="text" value="currency.priceshet.retail"/> |
| body * required string (body) | The new value of the probe property. <input type="text" value="Edit Value Model"/> |

Parameter content type

- probeid - Specify the UUID for the Azure Pricing probe.
- propertyName - Specify currency.priceshet.retail.
- body - Specify the 3 digit currency code. Codes are available in the [Azure documentation](#).

- c. Click **Execute** to apply your changes. A 200 response displays.

3. Restart the mediation-azurepricing pod from the backend.

```
kubectl delete pod mediation-azurepricing<tab>
```

NOTE:

After the restart, it may take Intersight Workload Optimizer a few hours to discover Azure reservations costs in the new currency.

Azure Actions

Intersight Workload Optimizer monitors the state and performance of your workloads and then recommends actions to optimize these workloads at the lowest possible cost.

NOTE:

Use the Potential Savings and Necessary Investments charts to view pending actions and evaluate their impact on your cloud expenditure.

Actions for Virtual Machines

Intersight Workload Optimizer supports the following actions:

- **Scale**

Change the VM instance to use a different instance type or tier to optimize performance and costs.

- **Discount-related actions**

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing Azure reservations coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

If you need more capacity, then Intersight Workload Optimizer will recommend actions to purchase additional reservations. For details, see [Actions for Azure VMs \(on page 268\)](#).

Actions for Volumes

Intersight Workload Optimizer discovers and optimizes Azure *managed* volumes.

According to this Microsoft [article](#), *unmanaged* volumes are deprecated and will be fully retired in 2025. In response, Intersight Workload Optimizer no longer discovers or monitors unmanaged volumes that are not attached to any VM. Unmanaged volumes that are attached to VMs will continue to be discovered and displayed in the user interface for your reference, but no action will be generated for these volumes.

Intersight Workload Optimizer supports the following actions:

- **Scale**
Scale attached volumes to optimize performance and costs.
- **Delete**
Delete unattached volumes as a cost-saving measure. Intersight Workload Optimizer generates an action immediately after discovering an unattached volume.

For details, see [Cloud Volume Actions \(on page 327\)](#).

Actions for Virtual Machine Specs (App Service Plans)

Intersight Workload Optimizer supports the following actions:

- **Scale**
Scale Azure App Service plans to optimize app performance or reduce costs, while complying with business policies.
- **Delete**
Delete empty Azure App Service plans as a cost-saving measure. A plan is considered empty if it is not hosting any running apps.

For details, see [Virtual Machine Spec Actions \(on page 288\)](#).

Actions for App Component Specs (App Service Instances)

None

Intersight Workload Optimizer does not recommend actions for App Component Specs, but it does recommend actions for the underlying Virtual Machine Specs. For details, see [Virtual Machine Spec Actions \(on page 288\)](#).

Actions for Database Servers (Cosmos DB Accounts)

None

Intersight Workload Optimizer does not recommend actions for a Cosmos DB account but it does recommend actions for the [databases \(on page 317\)](#) and [document collections \(on page 322\)](#) in the account.

Actions for Databases

Intersight Workload Optimizer supports the following actions:

- **Scale SQL Database**
 - DTU Pricing Model
Scale DTU and storage resources to optimize performance and costs.
 - vCore Pricing Model
Scale vCPU, vMem, IOPS, throughput and storage resources to optimize performance and costs.

For details, see [Scale Actions for SQL Databases \(on page 310\)](#).

- **Scale Cosmos DB Database**
Scale Request Units (RUs) to optimize performance and costs.
For details, see [Scale Actions for Cosmos DB Databases \(on page 317\)](#).
- **Reconfigure Cosmos DB Database**

Remove unused provisioned throughput to reduce costs.

For details, see [Reconfigure Actions for Cosmos DB Databases \(on page 318\)](#).

- **Delete Cosmos DB Database**

Delete a database with provisioned throughput but without any underlying document collection (container) to reduce costs.

For details, see [Delete Actions for Cosmos DB Databases \(on page 318\)](#).

- **Suspend/Stop Dedicated SQL Pool**

Suspend or stop a dedicated SQL pool (used in Azure Synapse Analytics) to reduce compute costs.

- Intersight Workload Optimizer analysis generates suspend actions for *idle* pools.

NOTE:

Currently, Intersight Workload Optimizer analysis does not generate actions to start a suspended pool. You can start a suspended pool from Azure.

For details, see [Suspend Actions for Dedicated SQL Pools \(on page 316\)](#).

Actions for Document Collections (Cosmos DB Containers)

Intersight Workload Optimizer supports the following actions:

- **Scale**

Scale Request Units (RUs) to optimize performance and costs.

Cloud Native Targets

To support cloud native environments, Intersight Workload Optimizer targets Kubernetes clusters. Intersight Workload Optimizer supports target clusters managed on Kubernetes 1.8 or higher, whether the clusters are managed directly through `kubeadm` or other platforms, including:

- Cisco Container Platform (CCP)
- Red Hat OpenShift
- Pivotal Kubernetes Service
- Amazon Elastic Kubernetes Service (EKS)
- Azure Kubernetes Service (AKS)
- Google Kubernetes Engine (GKE)

With cloud native targets, Intersight Workload Optimizer discovers entities related to container platforms in your environment. Discovery can also stitch the container cluster entities together with managed applications. For example, discovery can show the full application stack if your container environment includes applications managed by the following technologies, and you have added them as targets to Intersight Workload Optimizer:

- [Cisco AppDynamics \(on page 148\)](#)
- [Dynatrace \(on page 153\)](#)
- [New Relic \(on page 159\)](#)

Claiming a Cloud Native Target

To claim this target for a Kubernetes cluster, you first install the Intersight Workload Optimizer Kubernetes Collector on your Kubernetes cluster. The installation process generates a Device ID and a Claim Code that you can then use to add the collector as a target to manage the Kubernetes cluster.

The Intersight Workload Optimizer platform gathers information from your Kubernetes or Red Hat OpenShift environment via the collector that you install on your Kubernetes cluster. The collector collects information from your environment and passes it to Intersight Workload Optimizer. As it generates actions, Intersight Workload Optimizer then uses the collector to execute those actions in your cluster. In this way, Intersight Workload Optimizer users can execute actions from the user interface, policies can set up actions to execute automatically, and Intersight Workload Optimizer can automatically execute groups of related actions on the workloads in a container spec.

NOTE:

You must install a different collector on each Kubernetes cluster you want Intersight Workload Optimizer to manage.

You install the collector via a Helm chart. For installation instructions, see [Installing the Intersight Workload Optimizer Kubernetes Collector \(on page 113\)](#). The last installation step is to register the collector. This generates a Device ID and a Claim Code that you can then use to add the collector as a target. As you install the collector, you should record these values.

To claim a Kubernetes target, select **Cloud Native > Kubernetes** on the Target Configuration page and provide the following information:

- Device ID

This identifies the Kubernetes Collector you have installed for the given cluster. When you register the collector, this is returned as the `SerialNumber` token.
- Claim Code

This authorizes the connection between your Intersight Workload Optimizer account and the collector. When you register the collector, this is returned as the `SecurityToken`.

Installing the Intersight Workload Optimizer Kubernetes Collector

To install the Intersight Workload Optimizer Kubernetes Collector, you deploy it on a node in your Kubernetes cluster. From that node, the collector uses `kubelet` to reach all the other pods in the cluster. To download the latest version of the Intersight Kubernetes Collector, go to [Cisco Software Download](#).

The collector installs as two identical pods. This supports High Availability (HA), where one pod can take over if the currently active collector pod crashes.

You deploy the collector on one node per cluster, or one collector per control plane when using stretch clusters. The collector runs with a service account that has the `cluster-admin` role. This role enables the collector to execute Intersight Workload Optimizer actions within your Kubernetes cluster.

To communicate with Intersight Workload Optimizer, the collector installs with its own Device Connector. The device connector provides a secure way for the collector to send information and receive control instructions from the Cisco Intersight portal, using a secure Internet connection.

Installation Requirements

To use the Intersight Workload Optimizer Kubernetes Collector, your environment must meet the following requirements:

- Kubernetes 1.8 or higher
- Helm v2 or v3 installed:

To deploy the collector, you will use Helm to install a chart in the Kubernetes cluster.

The installation instructions assume you have Helm v2 or v3 installed and configured to install the chart on the node where you want the collector to run. For more information about the Helm client, see [HELM](#).

For Helm v2, you must also have Tiller installed. Tiller requires the `cluster-admin` role to install and run collector charts, and needs to run with a service account with `ClusterRole` access. For details about role-based access, see:

 - <https://helm.sh/docs/topics/rbac/>
 - <https://github.com/fnproject/fn-helm/issues/21>
- Network requirements:

The collector pods must have access to the kubelet on every node in the cluster. This access can be via one of:

 - `https + port=10250` (default)
 - `http + port=10255`
- Device Connector port requirements:

The device connector provides a secure way for the collector to send information and receive control instructions from the Cisco Intersight portal. The following table lists the port numbers that must be open for device connector communication:

| Port | Protocol | Description |
|------|----------|--|
| 443 | TCP/UDP | Required for communication between: <ul style="list-style-type: none"> – The device connector and the user's Web browser. – The device connector and the Kubernetes endpoints. |
| 80 | TCP | This port is optional for normal operation, but is required for initial monitoring of the device connector setup and when using the one-time device connector upgrade. This port is not used if the device connector is at the minimum supported version. |

■ Compute and storage requirements:

The collector pod typically runs well with no more than:

- 512 Mg Memory
- One core or 1 GHz CPU
- 10 GB of volume space

Deploying the Intersight Workload Optimizer Kubernetes Collector

To deploy the collector:

1. Download the Helm chart to your node cluster.

To get the Helm Chart files, download them from [Cisco Software Download](#).

2. Create a namespace for the collector.

If you are installing many instances of the collector on many node clusters, it can be convenient to use the same namespace for each deployment. Execute the following command, where `iwo-collector` is the namespace name for this example (you can use any valid namespace name).

```
kubect1 create namespace iwo-collector
```

3. Execute the install command.

When you execute the command, you will specify:

- `name`: The release name of the installed collector. The name must not exceed 20 characters in length.
- `namespace`: The namespace the collector is installed under
- (Optional) `targetName`: A name that identifies the cluster you are installing onto. This can be any name. The use interface uses this name as it displays each managed cluster in the list of targets. It also uses this name in other places where it displays data about the cluster.
- (Optional) Any other parameters specified in the table below that you need to override.

The following sample commands assume these parameter values:

- `my-iwo-k8s-collector` for the release name
- `iwo-collector` for the namespace
- `my-k8s-cluster` for the `targetName`

Before you actually install the collector, use the dry-run feature to test your deployment. Execute the command:

- Helm v2:

```
helm install --dry-run --debug <Chart_Location> --name my-iwo-k8s-collector --namespace iwo-collector --set targetName=my-k8s-cluster
```

- Helm v3:

```
helm install --dry-run --debug my-iwo-k8s-collector <Chart_Location> --namespace iwo-collector --set targetName=my-k8s-cluster
```

Inspect the output to make sure the results are as you expect. If the output is correct, execute the installation:

- Helm v2:

```
helm install <Chart_Location> --name my-iwo-k8s-collector --namespace iwo-collector --set targetName=my-k8s-cluster
```

- Helm v3:

```
helm install my-iwo-k8s-collector <Chart_Location> --namespace iwo-collector --set targetName=my-k8s-cluster
```

Note that you can provide custom values for the installation. The parameters you can access and optionally change are:

| Parameter | Default Value | Changes to Default Value | Parameter Type |
|---------------------------|--------------------|--|--|
| targetName | "Your_k8s_cluster" | Optional Note that this is required for multiple clusters. | String The name you want to use to identify your cluster |
| args.failVolumePodMoves | true | Optional Change to <i>false</i> if you want to move pods that have volumes attached. The pod(s) will be down during the move. | Boolean |
| args.kubelethttps | true | Optional Change to <i>false</i> if using Kubernetes 1.10 or older. | Boolean |
| args.kubeletport | 10250 | Optional Change to 10255 if using Kubernetes 1.10 or older. | Number |
| args.logginglevel | 2 | Optional | Number |
| args.stitchuuid | true | Optional Change to <i>false</i> if IaaS is VMM or Hyper-V. | Boolean |
| HANodeDetectors.nodeRoles | "\ "master\ " | Optional Use to automate policies to keep nodes of the same role limited to 1 instance per ESX host or availability zone. | String Uses regular expressions. Values are in quotes and comma separated. For example: "master" (default), "worker", "app" |

4. Verify the installation.

After you execute the installation, give Helm enough time to install the collector. It installs two collector pods to support HA for the collector. Each pod contains two containers – one for the device connector and one for the collector.

After enough time has passed to start up the pods, verify that they are running. Execute the following command, where `iwo-collector` is the namespace for this collector:

```
kubectl get pods -n iwo-collector
```

The entry for the pod should have the following values:

- NAME: `iwok8scollector-my-iwo-k8s-collector<Pod_ID>`, where `my-iwo-k8s-collector` is the name you provided when you installed the collector, and `<Pod_ID>` is a generated ID value.
- READY: 2/2
- STATUS: Running

For example, the output should be similar to:

| NAME | READY | STATUS | RESTARTS | AGE |
|--|-------|---------|----------|-----|
| <code>iwok8scollector-my-iwo-k8s-collector-57fcb8b874-s5ch8</code> | 2/2 | Running | 0 | 12s |
| <code>iwok8scollector-my-iwo-k8s-collector-57fcb8b874-c7dd2</code> | 2/2 | Running | 0 | 12s |

Record one of the full pod names for the next step. Either pod will do.

5. Register the collector to get its Device ID and Claim Code.

When you register the collector, you will forward its port 9110 and then connect to your Intersight instance to get the registration tokens. If you need a proxy for a connection that is external to your cluster or network, run a command to enable a proxy in the collector.

To register the collector:

- Forward the pod's port 9110:

Execute the command, where `iwo-collector` is the namespace you defined for this collector and `my-iwo-k8s-collector-57fcb8b874-s5ch8` is the full pod name:

```
kubectl -n iwo-collector port-forward my-iwo-k8s-collector-57fcb8b874-s5ch8 9110
```

- (Optional) Configure the proxy connection from the collector to `intersight.com`.

If your proxy does not require authentication, execute the following command:

```
curl -XPUT http://localhost:9110/HttpProxies -d '{"ProxyType":"Manual", "ProxyHost":"<My_Proxy_Server>", "ProxyPort":<My_Proxy_Port>}'
```

If your proxy requires authentication using username/password credentials, execute the following command:

```
curl -XPUT http://localhost:9110/HttpProxies -d '{"ProxyType":"Manual", "ProxyUsername":"<username>", "ProxyPassword":"<password>", "ProxyHost":"<My_Proxy_Server>", "ProxyPort":<My_Proxy_Port>}'
```

Where:

- `<My_Proxy_Server>` is the address of your proxy server

Note that you must *not* include the HTTP protocol in the proxy address. For example, if your proxy is located at `https://proxy-was.esl.cisco.com`, specify the following address:

```
proxy-was.esl.cisco.com
```

- `<My_Proxy_Port>` is the port your proxy server uses

- Get the Device ID and Claim Code:

Execute the following commands:

- Get the Device ID:

```
curl -s http://localhost:9110/DeviceIdentifiers
```

The command output should be similar to the following, where "ID" : is the Device ID value:

```
[
  {
    "Id": "22284c13-xxxx-yyyy-zzzz-93a14e4de07f"
  }
]* Closing connection 0
```

- Get the Claim Code:

```
curl -s http://localhost:9110/SecurityTokens
```

The command output should be similar to the following, where "Token" : is the Claim Code value:

```
[
  {
    "Token": "26AEAECDD67",
    "Duration": 599
  }
]* Closing connection 0
```

Record these values. You will provide them as credentials when you claim the collector as a Kubernetes target.

If either command returns an error similar to the following, confirm that the cluster and the collector can connect with `intersight.com`. This could indicate that you need to configure a proxy connection (see above).

```
{
  "code": "InternalServerError",
  "message": "Internal error while fetching Claim Code",
  "messageId": "",
  "messageParams": null,
  "traceId": "DCxxxxxxxxxxxxxxxxxxxxfc9e4de952584049"
}
```

Updating the Intersight Workload Optimizer Kubernetes Collector

Intersight Workload Optimizer continues to improve the management of resources in a Kubernetes cluster. To take advantage of these improvements, update the collector so it can pass new types of data to Intersight Workload Optimizer and execute any newly added commands.

When you update the collector, you will specify:

- `name`: The release name that you specified when you originally installed the collector
- `namespace`: The namespace where the collector was originally installed
- `Chart_Location`: Folder where the chart that was downloaded from Cisco Software Download is present
- `Overridden parameters`: If you overrode any of the parameters listed in the table above (such as `targetName`), specify them during the update.

The following sample commands assume these parameter values:

- `my-iwo-k8s-collector` for the release name (for Helm v3)
- `iwo-collector` for the namespace
- `targetName=my-k8s-collector` is an overridden parameter that was specified during the installation.

To update the collector:

1. Download the Helm chart to your node cluster.

To get the Helm Chart files, download them from [Cisco Software Download](#).

2. Execute the Helm command:

- Helm v2:

```
helm upgrade <Chart_Location> --namespace iwo-collector --set targetName=my-k8s-collector
```

- Helm v3:

```
helm upgrade my-iwo-k8s-collector <Chart_Location> --namespace iwo-collector --set targetName=my-k8s-collector
```

Removing the Intersight Workload Optimizer Kubernetes Collector

To remove the collector from your cluster, execute one of the following Helm commands, where `iwo-collector` is the namespace you used for the release, and `my-iwo-k8s-collector` is the release name:

- Helm v2:

```
helm uninstall -n iwo-collector my-iwo-k8s-collector
```

- Helm v3:

```
helm delete -n iwo-collector my-iwo-k8s-collector
```

Container Platform Monitored Resources

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|------------------------|---|
| Container Platform | Intersight Workload Optimizer |
| Service | Service (on page 213) |
| Container | Container (on page 226) |
| A container's spec | Container spec (on page 228) |
| Pod | Container Pod (on page 241) |
| Controller | Workload Controller (on page 235) |
| Namespace | Namespace (on page 246) |
| Cluster | Container platform cluster (on page 250) |
| Node | Virtual Machine (on page 257) |
| Persistent Volume (PV) | <p>Volume</p> <p>NOTE: If a container pod is attached to a volume (on page 326), Intersight Workload Optimizer discovers it as a Persistent Volume (PV), and shows which pods are connected to the PV.</p> |

Monitored Resources for Services

Intersight Workload Optimizer monitors the following resources:

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

Monitored Resources for Containers

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

VMem is the virtual memory utilized by a container against the memory limit. If no limit is set, node capacity is used.
- VMem Request

If applicable, VMem Request is the virtual memory utilized by a container against the memory request.
- VCPU

VCPU is the virtual CPU (in mCores) utilized by a container against the CPU limit. If no limit is set, node capacity is used).
- VCPU Request

If applicable, VCPU Request is the virtual CPU (in mCores) utilized by a container against the CPU request.
- VCPU Throttling

VCPU Throttling is the throttling of container virtual CPU that could impact response time, expressed as the percentage of throttling for all containers associated with a Container Spec. In the Capacity and Usage chart for containers, *used* and *utilization* values reflect the actual throttling percentage, while *capacity* value is always 100%.

Monitored Resources for Container Specs

Intersight Workload Optimizer monitors the historical usage of any instance of a container running for the workload (assuming the workload name stays the same). Charts show the trend of usage even with restarts or redeployments.

Monitored Resources for Container Pods

Intersight Workload Optimizer monitors the following resources:

- VMem

VMem is the virtual memory utilized by a pod against the node physical capacity.
- VMem Request

VMem Request is the virtual memory request allocated by a pod against the node allocatable capacity.
- VCPU

VCPU is the virtual CPU (in mCores) utilized by a pod against the node physical capacity.
- VCPU Request

VCPU Request is the virtual CPU request (in mCores) allocated by a pod against the node allocatable capacity.
- VMem Request Quota

If applicable, VMem Request Quota is the amount of virtual memory request a pod has allocated against the namespace quota.
- VCPU Request Quota

If applicable, VCPU Request Quota is the amount of virtual CPU request (in mCores) a pod has allocated against the namespace quota.
- VMem Limit Quota

If applicable, VMem Limit Quota is the amount of virtual memory limit a pod has allocated against the namespace quota.
- VCPU Limit Quota

If applicable, VCPU Limit Quota is the amount of virtual CPU limit (in mCores) a pod has allocated against the namespace quota.

Monitored Resources for Workload Controllers

Intersight Workload Optimizer monitors quotas (limits and requests) for VCPU and VMem, and associates how much each Workload Controller is contributing to a quota based on all replicas. This allows Intersight Workload Optimizer to generate rightsizing decisions, and manage the quota as a constraint to rightsizing. Metrics on resource consumption are shown in the Container Spec, Container, and Container Pod views.

Monitored Resources for Namespaces

Intersight Workload Optimizer monitors the following resources:

- VMem Request Quota
VMem Request Quota is the total amount of virtual memory request for all pods allocated to the namespace against the namespace quota.
- VCPU Request Quota
VCPU Request Quota is the total amount of virtual CPU request (in mCores) for all pods allocated to the namespace against the namespace quota.
- VMem Limit Quota
VMem Limit Quota is the total amount of virtual memory limit for all pods allocated to the namespace against the namespace quota.
- VCPU Limit Quota
VCPU Limit Quota is the total amount of virtual CPU limit (in mCores) for all pods allocated to the namespace against the namespace quota.

Monitored Resources for Container Platform Clusters

Intersight Workload Optimizer monitors resources for the containers, pods, nodes (VMs), and volumes in a cluster.

Monitored Resources for Nodes (VMs)

Intersight Workload Optimizer monitors the following resources for nodes that host pods. These resources are monitored along with the resources from the infrastructure probes, such as vCenter or a public cloud mediation probe.

- VMem
VMem is the virtual memory currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- VCPU
VCPU is the virtual CPU currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- Memory Request Allocation
Memory Request Allocation is the memory available to the node to support the ResourceQuota request parameter for a given Kubernetes namespace or Red Hat OpenShift project.
- CPU Request Allocation
CPU Request Allocation is the CPU available to the node to support the ResourceQuota request parameter for a given Kubernetes namespace or Red Hat OpenShift project.
- Virtual Memory Request
Virtual Memory Request is the memory currently guaranteed by all containers on the node with a memory request. The capacity for this resource is the Node Allocatable capacity, which is the amount of resources available for pods and can be less than the physical capacity.
- Virtual CPU Request
Virtual CPU Request is the CPU currently guaranteed by all containers on the node with a CPU request. The capacity for this resource is the Node Allocatable capacity, which is the amount of resources available for pods and can be less than the physical capacity.

- **Memory Allocation**

Memory Allocation is the memory ResourceQuota limit parameter for a given Kubernetes namespace or Red Hat OpenShift project.

- **CPU Allocation**

CPU Allocation is the CPU ResourceQuota limit parameter for a given Kubernetes namespace or Red Hat OpenShift project.

Container Platform Actions

Intersight Workload Optimizer monitors the state and performance of your containerized workloads and then recommends actions to optimize these workloads.

Actions for Services

None

Intersight Workload Optimizer does not recommend actions for services in container platform environments, but it does recommend actions for the replicas that back those services.

For details, see [Workload Controller Scale Actions \(on page 238\)](#).

Actions for Containers

None

Intersight Workload Optimizer does not recommend actions on containers.

Actions for Container Specs

Resize (via workload controllers)

A container spec retains the historical utilization data of ephemeral containers. Intersight Workload Optimizer uses this data to make resize decisions that assure optimal utilization of resources. By default, all replicas of the same container for the same workload type resize consistently.

For details, see [Workload Controller Resize Actions \(on page 236\)](#).

Actions for Container Pods

- **Move**

Move a pod between nodes (VMs) to address performance issues or improve infrastructure efficiency. For example, if a particular node is congested for CPU, you can move pods to a node with sufficient capacity. If a node is underutilized and is a candidate for suspension, you must first move the pods before you can safely suspend the node.

- **Provision/Suspend**

When recommending node provision or suspend actions, Intersight Workload Optimizer will also recommend provisioning pods (based on demand from DaemonSets) or suspending the related pods.

For details, see [Container Pod Actions \(on page 243\)](#).

Actions for Workload Controllers

Resize or Scale

Actions associated with a workload controller resize container specs vertically or scale replicas horizontally. This is a natural representation of these actions because the parent controller's container specs and number of replicas are modified. The workload controller then rolls out the changes in the running environment.

For details, see [Workload Controller Resize Actions \(on page 236\)](#) and [Workload Controller Scale Actions \(on page 238\)](#).

Actions for Namespaces

Resize Quota

Intersight Workload Optimizer treats quotas defined in a namespace as constraints when making resize decisions. If existing actions would exceed the namespace quotas, Intersight Workload Optimizer recommends actions to resize up the affected namespace quota.

Note that Intersight Workload Optimizer does not recommend actions to resize *down* a namespace quota. Such an action reduces the capacity that is already allocated to an application. The decision to resize down a namespace quota should include the application owner.

For details, see [Namespace Actions \(on page 248\)](#).

Actions for Container Platform Clusters

None

Intersight Workload Optimizer does not recommend actions for a container platform cluster. Instead, it recommends actions for the containers, pods, nodes (VMs), and volumes in the cluster. Intersight Workload Optimizer shows all of these actions when you scope to a container platform cluster and view the Pending Actions chart.

Actions for Nodes (VMs)

A node (cloud or on-prem) is represented as a Virtual Machine entity in the supply chain.

- **Provision**

Provision nodes to address workload congestion or meet application demand.

- **Suspend**

Suspend nodes after you have consolidated pods or defragmented node resources to improve infrastructure efficiency.

- **Reconfigure**

Reconfigure nodes that are currently in the `NotReady` state.

NOTE:

For nodes in the public cloud, Intersight Workload Optimizer reports the cost savings or investments attached to these actions.

For details, see [Node Actions \(on page 258\)](#).

Applications and Databases Targets

Applications and Databases targets support domains of particular application servers that are controlled by management servers. For such managed domains you will add the management server as a target, and Intersight Workload Optimizer will discover the managed application servers.

NOTE:

As it manages your applications environment, Intersight Workload Optimizer discovers connected application components to stitch them into a supply chain of entities. For connections that are made by name and not IP address, Intersight Workload Optimizer makes DNS calls to resolve these names to IP addresses. This can happen during repeated discovery cycles.

Supply Chain

Applications and Databases targets add Business Application, Business Transaction, Service, Application Component, Application Server, and Database Server entities to the supply chain. You can navigate to the associated target page to see how these entities map to the target nomenclature.

Apache Tomcat

Intersight Workload Optimizer supports connecting to individual Tomcat targets. Intersight Workload Optimizer connects to the Tomcat process as a remote client via remote JMX access. Target configuration includes the port used by the JMX/RMI registry.

Prerequisites

- A valid JMX user account for the Tomcat server.
 - If Tomcat security is enabled, this must be a Tomcat JMX user with a `readonly` role.
- Tomcat should run on JVM version 7 or 8.

- For VMware environments, VMware Tools must be installed on the VM that hosts the Tomcat server. For Hyper-V environments, Hyper-V Integration Services must be installed.

This ensures that the VM hosting the Tomcat server can get its IP address.

- Remote JMX access is enabled through a port that is opened to the firewall.
- Discovered infrastructure.

Intersight Workload Optimizer discovers Tomcat servers that are running on VMs or containers. The hosting VM or container must already be in your Intersight Workload Optimizer inventory.

To set the target for a server running on a VM, you must have first discovered the hosting VM through a hypervisor target. To set the target for a server running in a container, you must have configured container discovery for Tomcat applications.

Configuring JMX Remote Access

Intersight Workload Optimizer monitors and controls the Tomcat server via JMX Remote access. You must configure a JMX Remote port.

Note that to work with a firewall you should also set the RMI Server port – If you don't set an RMI port, then JMX sets an arbitrary *ephemeral port*, and you can't guarantee that the port will be open to your firewall.

There are two ways to set JMX Remote port on Linux platforms:

- Ports specified as system properties

You can set the port via the system property, `com.sun.management.jmxremote.port`. For example:

```
com.sun.management.jmxremote.port=8050
```

A common way to set this property is to declare it in the `CATALINA_OPTS` system variable – You can set this in the `setenv.sh` script. For example:

```
CATALINA_OPTS="$CATALINA_OPTS
-Dcom.sun.management.jmxremote
-Dcom.sun.management.jmxremote.port=8050"
export CATALINA_OPTS
```

Note that this sets the JMX Remote port, but it does not set the RMI Server port – Tomcat startup will specify an ephemeral port for the RMI server.

- Ports specified in a JMX Remote Lifecycle Listener

This listener component fixes the ports used by the JMX/RMI Server. When you configure the listener, you specify both the JMX Remote port and the RMI Server port. This is the preferred method when working with a firewall. For more information, see the Apache Tomcat documentation.

On Windows, the typical installation is with Tomcat as a service. There are two ways to set the JMX Remote port:

- Via `setenv.bat`

Add the property to the `CATALINA_OPTS` environment variable:

```
set "CATALINA_OPTS=%CATALINA_OPTS% -Dcom.sun.management.jmxremote.port=8050"
```

- Use the Tomcat configuration utility (`tomcat7w` or `tomcat8w`)

Set the port with the following command:

```
-Dcom.sun.management.jmxremote.port=8050"
```

To discover the JMX port that is set to an already running Tomcat, you can look in the following locations:

- For Linux platforms, look in the configuration files – Either:
 - `setenv.sh` – Assuming you configured the port by adding it to the `CATALINA_OPTS` environment variable

- \$CATALINA_HOME/conf/server.xml - Assuming you configured a JMX Remote Lifecycle Listener in this file
- For Windows platforms, look in:
 - setenv.bat - Assuming you configured the port by adding it to the CATALINA_OPTS environment variable
 - The Windows registry - Assuming you installed Tomcat as a Windows service using the Tomcat Configuration utility

Adding a Tomcat Target

You can add an individual Tomcat server as a target, or you can add all matching servers within a given scope.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **Tomcat**.
4. Configure the following settings:
 - **Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in Tomcat.
 - **Username**
Specify the username of an account with the Admin role.
 - **Password**
Specify the password of an account with the Admin role.
 - **Scope**
Specify the scope Intersight Workload Optimizer uses for application discovery.
The scope is a group of applications that are stitched to the underlying VMs when the VMs are discovered as part of a separate Intersight Workload Optimizer target.
If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.
 - **JMX Remote Port**
Specify a JMX port that is set to an already running Tomcat process.
 - **Full Validation**
If you select this option, Intersight Workload Optimizer attempts to authenticate all database servers in the selected scope. If Intersight Workload Optimizer is unable to authenticate a database server, the target is not added and no data is collected.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Application Component**
 - Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
 - Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
 - Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
 - Heap
Heap is the portion of a VM or container's memory allocated to individual applications.
 - Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
 - Threads

Threads is the measurement of thread capacity utilized by applications.

- Connection

Connection is the measurement of database connections utilized by applications.

- Remaining GC Capacity

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

■ Virtual Machine

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

Actions

Intersight Workload Optimizer supports the following actions:

■ Application Component (Tomcat Application)

- Resize Heap

This action can only be executed outside Intersight Workload Optimizer.

- Resize Thread Pool

This action can only be executed outside Intersight Workload Optimizer.

- Resize Connection Capacity

This action can only be executed outside Intersight Workload Optimizer.

■ Virtual Machine

- Provision additional resources (VMem, VCPU)

- Move Virtual Machine

- Move Virtual Machine Storage

- Reconfigure Storage

- Reconfigure Virtual Machine

- Suspend VM

- Provision VM

IBM WebSphere

The typical WebSphere deployment is a cell of WebSphere servers, controlled by a Deployment Manager. A cell makes up a managed domain that incorporates multiple VMS that host managed application servers. The Deployment Manager is a WebSphere instance that provides a single point of entry for the managed domain.

NOTE:

When adding a WebSphere Deployment Manager as a target, you must ensure that the name of each WebSphere node is resolvable to an IP address by the Intersight Workload Optimizer instance.

You may need to make changes to your DNS or the file `/etc/resolv.conf` on the Intersight Workload Optimizer instance to make it aware of the domain names in use in your environment.

To configure the WebSphere installation, you can use the WebSphere Integrated Solutions Console. This is a client that exposes configuration settings including the SOAP port and the PMI settings.

To manage the servers in an installation, WebSphere uses the Performance Monitoring Infrastructure (PMI). Each WebSphere server runs a PMI service that collects performance data from the various application server components. Intersight Workload Optimizer uses PMI for monitoring and control of the WebSphere installation.

Prerequisites

- A service user account

To execute actions the service account must have an Administrator role. For read-only monitoring and analysis, you can set the target with a more restricted role (Monitor), but then you will have to execute all recommended actions manually, through the WebSphere interface.

NOTE:

You cannot use Active Directory (AD) accounts to target WebSphere.

- The PMI service set to monitor at the Basic level or greater
- Discovered infrastructure

Intersight Workload Optimizer discovers WebSphere servers that are running on VMs or containers. The hosting VM or container must already be in your Intersight Workload Optimizer inventory.

To set the target for a server running on a VM, you must have first discovered the hosting VM through a Hypervisor target. To set the target for a server running in a container, you must have configured container discovery for WebSphere applications.

Finding the SOAP Connector Address

To configure a WebSphere target, you need to know the port that the server listens on for administrative communications. Launch the WebSphere Administration Console:

- Navigate to System **Administration > Deployment Manager**
- Under **Additional Properties**, click **Ports**

The entry for `SOAP_CONNECTOR_ADDRESS` gives the currently set port number.

Adding a WebSphere Target

You can add an individual WebSphere server as a target, or you can add all matching targets within a given scope.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **WebSphere**.
4. Configure the following settings:
 - **Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in WebSphere.
 - **Username**
Specify the username of an account with the Admin role.
 - **Password**
Specify the password of an account with the Admin role.
 - **Scope**
The scope is a group of applications that are stitched to the underlying VMs when the VMs are discovered as part of a separate Intersight Workload Optimizer target.
If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.
 - **Port Number**
Specify the WebSphere remote port number.
 - **Full Validation**
If you select this option, Intersight Workload Optimizer attempts to authenticate all database servers in the selected scope. If Intersight Workload Optimizer is unable to authenticate a database server, the target is not added and no data is collected.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Service

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

■ Application Component

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- Heap

Heap is the portion of a VM or container's memory allocated to individual applications.

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Threads

Threads is the measurement of thread capacity utilized by applications.

- Connection

Connection is the measurement of database connections utilized by applications.

- Remaining GC Capacity

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

Actions

Intersight Workload Optimizer supports the following actions:

■ Service

Intersight Workload Optimizer does not recommend actions to perform on the Service itself, but it does recommend actions to perform on the Application Components and hosting VMs. For example, assume a Service that manages three SQL databases. If a surge in requests degrades performance across all three databases, then Intersight Workload Optimizer can start a new application component to run another instance of the database application, and bind it to the service. On the other hand, if SQL requests drop off so that the load balancer only forwards requests to two of the databases, Intersight Workload Optimizer can suspend the dormant database and unbind it.

■ Application Component

- Resize Heap

This action can only be executed by Intersight Workload Optimizer if the entity is running in a domain controller. Actions for standalone entities can only be executed outside Intersight Workload Optimizer.

- Resize Connection Capacity

This action can only be executed outside Intersight Workload Optimizer.

■ Virtual Machine

- Provision additional resources (VMem, VCPU)
- Move Virtual Machine
- Reconfigure Storage
- Suspend VM

JVM Application

Intersight Workload Optimizer supports connecting to individual JVM Applications as targets. Intersight Workload Optimizer connects to the JVM process as a remote client via remote JMX access. Target configuration includes the port used by the JMX/RMI registry.

Prerequisites

- A valid JMX user account for the JVM application.
 - If JMX security is enabled this must be a JMX user with a `readonly` role
- The application should run on JVM version 6.0 or higher.
- For VMware environments, VMware Tools must be installed on the VM that hosts the application.

This ensures that the VM hosting the application can get the application's IP address

- Remote JMX access is enabled through a port that is opened to the firewall.
- Discovered infrastructure

Intersight Workload Optimizer discovers JVM applications that are running on VMs or containers. The hosting VM or container must already be in your Intersight Workload Optimizer inventory.

To set the target for a server running on a VM, you must have first discovered the hosting VM through a hypervisor target. To set the target for a server running in a container, you must have configured container discovery for JVM applications.

Adding JVM Application Targets

When you configure JVM targets, you declare a given scope and add all matching applications within that given scope. To do this, specify:

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **JVM**.
4. Configure the following settings:

- **Scope**

Specify the scope Intersight Workload Optimizer uses for application discovery.

The scope is a group of applications that are stitched to the underlying VMs when the VMs are discovered as part of a separate Intersight Workload Optimizer target.

If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.

- **Port Number**

Specify the JMX remote port number.

- **Username/Password**

Specify the credentials for a user account with an Admin role.

Credentials must match the credentials that you specify for the JMX login configuration when you start up the application. If you disable authentication on the application, you must still provide arbitrary values for Username and Password.

To disable JMX authentication, use the following flags in the command line as you start the application:

```
-Dcom.sun.management.jmxremote.authenticate=false
```

```
-Dcom.sun.management.jmxremote.ssl=false
```

Configuring JMX Remote Access

Intersight Workload Optimizer monitors and controls JVM applications via JMX Remote access. You must configure a JMX Remote port.

Note that to work with a firewall you should also set the RMI Server port – If you don't set an RMI port, then JMX sets an arbitrary *ephemeral port*, and you can't guarantee that the port will be open to your firewall.

To set the JMX Remote port, pass in the port at the command line when you start your application. For example, to set the port to 8090, start your application with the following options:

```
-Dcom.sun.management.jmxremote -Dcom.sun.management.jmxremote.port=8090
```

Multiple JVM Targets On Single VM

Note that you can specify targets with different ports, but that run on the same VM (use the same IP address). You can also specify targets via the same scope, but with different ports – This is another way to assign applications running on the same VM to different ports. To do this:

To do this, add the targets in two separate steps. For example, assume you want to add two JVM application targets, and they both run on the VM at 10.10.123.45. One application is on port 123, and the other application is on port 456. To specify these two targets:

- Specify the first target with the following parameters:

- Scope: VMs_myCluster.mycorp.com
- Port number: 123
- Username: AppUser
- Password: *****

Click **ADD**.

- Specify the second target with the following parameters:

- Scope: VMs_myCluster.mycorp.com
- Port number: 456
- Username: OtherAppUser
- Password: *****

Click **ADD**.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Application Component** (JVM Application)

- Heap
 - Heap is the portion of a VM or container's memory allocated to individual applications.
- Remaining GC Capacity
 - Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).
 - Data is collected if JVM profiler is enabled.

- **Virtual Machine**

- Virtual Memory (VMem)
 - Virtual Memory is the measurement of memory that is in use.
- Virtual CPU (VCPU)
 - Virtual CPU is the measurement of CPU that is in use.

Actions

Intersight Workload Optimizer supports the following actions:

- **Application Component** (JVM Application)

- Resize Heap
 - This action can only be executed outside Intersight Workload Optimizer.
- Resize Thread Pool

This action can only be executed outside Intersight Workload Optimizer.

- Resize Connection Capacity

This action can only be executed outside Intersight Workload Optimizer.

- Suspend VM

This action can only be executed by Intersight Workload Optimizer if a VM is hosted in a vCenter environment. Actions for applications running on other hypervisors can only be executed outside Intersight Workload Optimizer.

- Provision VM

This action can only be executed outside Intersight Workload Optimizer.

■ Virtual Machine

- Provision additional resources (VMem, VCPU)
- Move Virtual Machine
- Move Virtual Machine Storage
- Reconfigure Storage
- Reconfigure Virtual Machine
- Suspend VM
- Provision VM

MySQL

NOTE:

This type of target can run as SaaS or in on-prem data centers. When you claim the target, you can choose to turn ON or turn OFF **Connect through an Intersight Assist** as follows:

- If the target runs as SaaS:

Turn OFF **Connect through an Intersight Assist**.

You should be aware that for earlier versions of Intersight Workload Optimizer, to claim an AppDynamics target running as SaaS you were required to specify an Intersight Assist. If you claimed your target through an Assist, you can reclaim that target *without* using the Intersight Assist. To do that you must first delete the claimed target, and then claim the target anew with **Claim through an Intersight Assist** in the OFF position.

- If the target runs in an on-prem data center:

Turn ON **Connect through an Intersight Assist**.

To establish communication between this target on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The AppDynamics target must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.
- Claim the AppDynamics target with **Connect through an Intersight Assist** in the ON position.

Intersight Assist provides a secure way for on-prem targets to send information to and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

To manage MySQL databases, Intersight Workload Optimizer can connect to one or more database servers within a defined scope.

Prerequisites

- User permissions are enabled on the MySQL Server.

For more information, see [Enabling User Permissions on MySQL \(on page 132\)](#).

Adding a MySQL Database Target

You can add all matching targets within a given scope.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **MySQL**.
4. Configure the following settings:
 - **Target ID**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in MySQL.
 - **Username**
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
 - **Password**
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.
 - **Scope**
Specify the scope Intersight Workload Optimizer uses for application discovery.
The scope is a group of applications that are stitched to the underlying VMs when the VMs are discovered as part of a separate Intersight Workload Optimizer target.
If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.
 - **Port Number**
Specify the MySQL remote port number. If left blank, Intersight Workload Optimizer uses the MySQL default port of 3306.
 - **Full Validation**
If you select this option, Intersight Workload Optimizer attempts to authenticate all database servers in the selected scope. If Intersight Workload Optimizer is unable to authenticate a database server, the target is not added and no data is collected.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Database Server**
 - Database Memory (DBMem)
Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.
 - Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
 - Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
 - Connection
Connection is the measurement of database connections utilized by applications.
A database connection is a physical communication pathway that holds database sessions, which are logical entities in the database instance memory that represent the state of a current user login to a database. Connections should be managed properly.
 - DB Cache Hit Rate
DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.
- **Virtual Machine**
 - Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Virtual Storage

Virtual storage is the measurement of virtual storage capacity that is in use.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

Actions

Intersight Workload Optimizer supports the following actions:

■ Database Server

Resize

- Database memory (DBMem)

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

■ Virtual Machine

Resize

Resize resource capacity, reservation, or limit to improve performance.

Enabling User Permissions on MySQL Server

Follow the following steps to enable appropriate user permissions on a MySQL Server.

1. Edit the MySQL server's configuration file to grant user permissions.
 - a. Open a Secure Shell session on the server and open the `.conf` file on the MySQL server in an editor. Depending on the platform your MySQL is running on, you find the file at different locations.
 - Debian Linux
`/etc/mysql/my.cnf`
 - Red Hat Linux (Fedora or Rocky)
`/etc/my.cnf`
 - FreeBSD Linux
Create the file at `/var/db/mysql/my.cnf`
 - b. Make the following changes in the `[mysqld]` section:
 - Comment out the following line to enable remote connections over TCP/Is:
`skip-networking`
 - Add the following line to bind your MySQL server address:
`bind-address=<MySQL_IP_Address>`
 - Add the following line to enable the collection of Transaction metrics:

```
innodb_monitor_enable = trx_rw_commits, trx_nl_ro_commits, trx_ro_commits,
trx_rollbacks
```

For example, if your MySQL server has the address, 123.45.66.77, after you bound the IP address and enabled Transaction metrics, the section of the `.conf` file should look like the following example:

```
[mysqld]
user          = mysql
pid-file      = /var/run/mysqld/mysqld.pid
socket        = /var/run/mysqld/mysqld.sock
port          = 3306
basedir       = /usr
datadir       = /var/lib/mysql
tmpdir        = /tmp
language      = /usr/share/mysql/English
bind-address  = 123.45.66.77
# skip-networking
# Uncomment the following line for MySQL versions 5.6+
innodb_monitor_enable = trx_rw_commits, trx_nl_ro_commits, trx_ro_commits, trx_rollbacks
....
```

- c. Save the `.conf` file.

NOTE:

Some MySQL installations use multiple configuration files. If a setting you made does not have the wanted effect, make sure that a different configuration file is not overwriting the value.

2. Enable collection of Response Time metrics.

Run the following command to log in to to the MySQL server:

```
$mysql -u root -p mysql
```

Then run the following SQL commands:

```
UPDATE performance_schema.setup_instruments SET ENABLED = 'YES' WHERE NAME LIKE 'statement/sql%';
```

```
UPDATE performance_schema.setup_instruments SET TIMED = 'YES' WHERE NAME LIKE 'statement/sql%';
```

NOTE:

If you want these changes to take effect each time you restart the MySQL server, add these statements to a file, and start the server with the `--init-file` option. For example, if you name the file `MyInit.txt`, then start the MySQL server with the following option:

```
--init-file=MyInit.txt
```

3. Give your Intersight Workload Optimizer server remote access to the database.

If you are not already logged in to the MySQL server, run the following command:

```
$mysql -u root -p mysql
```

Then run the following commands.

Assume a user named `USER_NAME` with a password `PWD_STRING`. Then assume that your Intersight Workload Optimizer has an IP address of 10.10.123.45. The following command grants privileges to that Intersight Workload Optimizer, if it connects with the specified user account:

```
GRANT SELECT ON performance_schema.* TO 'USER_NAME'@'10.10.123.45' IDENTIFIED BY 'PWD_STRING';
GRANT PROCESS ON *.* TO 'USER_NAME'@'10.10.123.45' IDENTIFIED BY 'PWD_STRING';
FLUSH PRIVILEGES;
```

The `FLUSH PRIVILEGES` command causes MySQL to retain these settings upon restart. When you are finished running these SQL commands, log out of MySQL.

Oracle

To manage Oracle databases, Intersight Workload Optimizer can connect to one or more database servers within a defined scope.

To connect to an Oracle database, you will:

- Add a Dynamic Performance view to the Oracle database
- Configure a service account on the database that Intersight Workload Optimizer can use to log on
- Find the Service Name and port for the database

Version Support

Intersight Workload Optimizer officially supports all versions supported by Oracle, which are currently: Oracle 19c and 21c

Prerequisites

- User permissions that grant access to Intersight Workload Optimizer through a specific user account. See [Creating a Service User Account in Oracle \(on page 136\)](#).
- Dynamic Performance View (V\$) must be enabled. For more information, see [Adding a Dynamic Performance View \(on page 134\)](#).
- Access through the firewall to the Oracle database port that you specify for the Intersight Workload Optimizer target connection

Adding a Dynamic Performance View

In order to collect data from the Oracle database, Intersight Workload Optimizer uses the Dynamic Performance View (referred to as V\$). V\$ is not enabled by default. You must run a script to build the tables and views that are necessary to enable V\$. In some environments only the DBA has privileges to run this script.

To enable V\$:

- Open a secure shell session (ssh) on the database host as a system user or a user with the `sysdba` role
- In the shell session enter the following commands:

```
sqlplus /nolog
connect /as sysdba
CREATE USER My_Username IDENTIFIED BY My_Password container=all;
GRANT CONNECT TO My_Username container=all;
GRANT sysdba TO My_Username container=all;
```

NOTE:

If security or other practices prohibit assigning SYSDBA to this user, you can use the following command to provide access to all V\$ views:

```
GRANT select any dictionary TO My_Username;
```

This creates a user account named `My_Username` with full privileges to access the V\$ Dynamic Performance view.

Adding an Oracle Database to Intersight Workload Optimizer

You can add an individual database server as a target, or you can add all matching targets within a given scope.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.

3. Select **Oracle**.
4. Configure the following settings:
 - **Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in Oracle.
 - **Username**
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
For Intersight Workload Optimizer to execute actions, the account must have administrator privileges. You must have enabled user permissions to this user account, including remote access from the Intersight Workload Optimizer server.
 - **Password**
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.
 - **Scope**
Specify the scope Intersight Workload Optimizer uses for application discovery.
The scope is a group of virtual machines that contain the databases that are discovered as part of a separate Intersight Workload Optimizer target.
If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.
NOTE:
All database servers in the scope must share the same service name, credentials, and port. For databases that have a different value for any of these, you must create a separate target using those values.
 - **Oracle Port**
Specify the port that connects to the database. You must open the firewall on the database server to allow access through this port.
Any firewall on the database must allow access through this port. To find the port, open an SSH session (as a system or sysdba user) on the database's host, run `lsnrctl status`, and then check `PROTOCOL=tcp`.
 - **Oracle Service Name**
Specify the service name for the database that you are connecting to Intersight Workload Optimizer.
To find the service name, open an SSH session on the database's host and run the following commands:

```
sqlplus /no log
connect /as sysdba
SELECT SYS_CONTEXT ('userenv', 'db_name') FROM dual;
```
 - **Full Validation**
If you select this option, Intersight Workload Optimizer attempts to authenticate all database servers in the selected scope. If Intersight Workload Optimizer is unable to authenticate a database server, the target is not added and no data is collected.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Database Server**
 - Database Memory (DBMem)
Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.
Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM.
 - Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
 - Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Connection

Connection is the measurement of database connections utilized by applications.

A database connection is a physical communication pathway that holds database sessions, which are logical entities in the database instance memory that represent the state of a current user login to a database. Connections should be managed properly.

- Transaction Log

Transaction log is the measurement of storage capacity utilized by a Database Server for transaction logging.

- DB Cache Hit Rate

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

- **Virtual Machine**

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Virtual Storage

Virtual storage is the measurement of virtual storage capacity that is in use.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

Actions

Intersight Workload Optimizer supports the following actions:

- **Database Server**

- **Resize**

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

- **Virtual Machine**

- **Resize**

Resize resource capacity, reservation, or limit to improve performance.

Creating a Service User Account in Oracle

To collect data from the Oracle database, Intersight Workload Optimizer requires a service account that has privileges to access the `V$` Dynamic Performance view. To create this account:

- Open a secure shell session (ssh) on the database host as a system user or a user with the `sysdba` role
- In the shell session enter the following commands:

```

sqlplus /nolog
connect /as sysdba
CREATE USER My_Username IDENTIFIED BY My_Password container=all;
GRANT CONNECT TO My_Username container=all;
GRANT sysdba TO My_Username container=all;

```

This creates a user account named My_Username with full privileges to access the V\$ Dynamic Performance view.

NOTE:

The preceding example uses a fictitious username. To comply with Oracle 12C norms, the username should include a prefix of c##.

Some enterprises don't allow accounts with sysdba access. Cisco recommends using sysdba, according to the Oracle documentation. However, you can work with your Oracle DBA staff to provide read access to the following views, which are the ones that Intersight Workload Optimizer needs:

- V\$INSTANCE
- V\$LOG
- V\$LOGFILE
- V\$PARAMETER
- V\$PGASTAT
- V\$RESOURCE_LIMIT
- V\$SGASTAT
- V\$SYS_TIME_MODEL
- V\$SYSMETRIC
- V\$SYSSTAT

SQL Server

NOTE:

This type of target can run as SaaS or in on-prem data centers. When you claim the target, you can choose to turn ON or turn OFF **Connect through an Intersight Assist** as follows:

- If the target runs as SaaS:

Turn OFF **Connect through an Intersight Assist**.

You should be aware that for earlier versions of Intersight Workload Optimizer, to claim an AppDynamics target running as SaaS you were required to specify an Intersight Assist. If you claimed your target through an Assist, you can reclaim that target *without* using the Intersight Assist. To do that you must first delete the claimed target, and then claim the target anew with **Claim through an Intersight Assist** in the OFF position.

- If the target runs in an on-prem data center:

Turn ON **Connect through an Intersight Assist**.

To establish communication between this target on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The AppDynamics target must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.
- Claim the AppDynamics target with **Connect through an Intersight Assist** in the ON position.

Intersight Assist provides a secure way for on-prem targets to send information to and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports the following versions of this target:

SQL Server 2012, 2014, 2016, 2017, 2019, and 2022

Intersight Workload Optimizer discovers both standalone and clustered SQL Servers, and represents them as Database Server entities in the supply chain.

Prerequisites

- A user account with SQL permissions including `Connect SQL` and `View Server State` on the database
- The `Net.Tcp Port Sharing Service` and `Net.Tcp Listener Adapter` services must be running, and set to enabled.
- TCP/IP is enabled on the port used for Intersight Workload Optimizer discovery.

Creating a Service User Account

The user account that Intersight Workload Optimizer uses for its service login must include the following:

- The account must exist in the Security folder within the SQL Server Object Explorer, with the following properties:
 - Enable **SQL Server Authentication**
 - Disable **Enforce password policy**
- The account's security properties must include:
 - Permission to connect to the database through SQL
 - Permission to view the server state

Adding a SQL Server Target

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **SQLServer**.
4. Configure the following settings:
 - **Target Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in SQLServer.
 - **Username**
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
Username must not include the Active Directory domain.
 - **AD Domain**
Specify the Active Directory domain used by Intersight Workload Optimizer in conjunction with the username for authentication. Leave blank for local accounts.
 - **Password**
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.
Password must not include the Active Directory domain.
 - **Discovery Path**
Specify the hostname or IP address or scope when performing the discovery process.
Intersight Workload Optimizer discovers SQL instances through the SQL Server's hostname/IP address or your selected scope.
If you change your selection, the value associated with the deselected option is automatically removed.
 - **Hostname or IP address**
Specify the hostname or IP address of your MSSQL environment configuration. Intersight Workload Optimizer scans the targeted hostname or IP address and tries to connect to the target using the specified port. Intersight Workload Optimizer adds any instances of the target it finds as entities from which metrics are retrieved.
 - **Scope**
Specify the scope Intersight Workload Optimizer uses for application discovery.
The scope is a group of applications that are stitched to the underlying VMs when the VMs are discovered as part of a separate Intersight Workload Optimizer target.

If you set a scope, Intersight Workload Optimizer searches for virtual machines in the selected group. Intersight Workload Optimizer can monitor up to 500 virtual machines in a group. If you have more than 500 virtual machines in your environment, split them across smaller groups and then add those groups as individual targets.

- **Browsing Service Port**

Specify the UDP port for the browsing service that listens for incoming connections to the SQL instances running on the SQL Server. The default UDP port is 1434.

If the browsing service is reachable via the specified port, Intersight Workload Optimizer discovers the SQL instances used by the VM group that you defined as your scope, as well as the listening ports on those SQL instances.

If the service is unreachable, or if you did not specify a UDP port, Intersight Workload Optimizer uses the TCP port that you specified in the SQLServer Port field to discover SQL instances.

- **SQLServer Port**

Specify the TCP port for the SQL Server. The default TCP port is 1433.

Intersight Workload Optimizer uses this port if the browsing service is unreachable, or if you did not specify a browsing service port.

- **Full Validation**

If you select this option, Intersight Workload Optimizer attempts to authenticate all database servers in the selected scope. If Intersight Workload Optimizer is unable to authenticate a database server, the target is not added and no data is collected.

Standalone and Clustered SQL Servers

Intersight Workload Optimizer discovers both standalone and clustered SQL Servers, and represents them as Database Server entities in the supply chain.

- When you set the scope to a SQL Server entity and view the Entity Information chart, the **Server Configuration** field indicates whether that entity is a standalone server or is part of a cluster.
- When you search for Database Servers or create groups of Database Servers, use the `Server Configuration` filter to get a list of standalone or clustered servers.

For clustered environments:

- Intersight Workload Optimizer represents each SQL Server instance in the cluster as a Database Server entity, and automatically creates a group for these instances.
- To see all the auto-created groups, go to **Settings > Groups** and then search for group names starting with `MSSQL:Cluster:`. Click a group name to set the scope to that group. In the resulting supply chain, the Database Server entity shows the number of SQL Server instances in the cluster. This entity is stitched to the Virtual Machine entity, which represents the corresponding SQL Server nodes. Click **Database Server** to see a list of instances, and identify which instance is currently active and which ones are idle (redundant). Intersight Workload Optimizer only monitors resources for the active instance, and shows resource metrics when you set the scope to that instance.
- If you have several clusters, you can use filters to identify all the active/idle instances in those clusters. In Search, select **Database Servers**, and then set the filter as follows:
 - Active instances
`Server Configuration=Clustered and State=ACTIVE`
 - Idle (redundant) instances
`Server Configuration=Clustered and State=IDLE`

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Database Server**

NOTE:

For clustered environments, Intersight Workload Optimizer only monitors resources for the currently active SQL Server instance, and shows resource metrics when you set the scope to that instance.

- Connection

Connection is the measurement of database connections utilized by applications.

A database connection is a physical communication pathway that holds database sessions, which are logical entities in the database instance memory that represent the state of a current user login to a database. Connections should be managed properly.

- Database Memory (DBMem)

Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM.

- DB Cache Hit Rate

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Transaction Log

Transaction log is the measurement of storage capacity utilized by a Database Server for transaction logging.

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- Virtual Storage

Virtual storage is the measurement of virtual storage capacity that is in use.

- **Virtual Machine**

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Virtual Storage

Virtual storage is the measurement of virtual storage capacity that is in use.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

Actions

Intersight Workload Optimizer supports the following actions:

- **Database Server**

- **Resize**

- Connections

Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers.

- Database memory (DBMem)

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

- Transaction log

Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology.

Currently, Intersight Workload Optimizer does not support resize actions for Oracle and Database Servers on the Hyper-V platform (due to the lack of API support for virtual storage).

■ Virtual Machine

- Provision additional resources (VMem, VCPU)
- Move Virtual Machine
- Move Virtual Machine Storage
- Reconfigure Storage
- Reconfigure Virtual Machine
- Suspend VM
- Provision VM

NOTE:

Without separate targets to discover Guest OS Processes or Application Servers, Intersight Workload Optimizer does not generate actions on applications. Instead, it generates resize actions on the host VMs. For on-prem environments, if host utilization is high enough on the host running the application VM, Intersight Workload Optimizer can also recommend provisioning a new host.

To retrieve the IP address and DNS name of the node inside the SQL Server cluster, Intersight Workload Optimizer might enable advanced tools and `xp_cmdshell` options and run commands against `nslookup`.

Compute / Fabric Targets

A fabric target is a service that unites compute, network and storage access into a cohesive system. When you connect Intersight Workload Optimizer to fabric targets, it monitors the performance and resource consumption of your fabric interconnects, IO modules, chassis, and physical machines to assure application performance and utilize resources as efficiently as possible.

Once connected, Intersight Workload Optimizer discovers the blade servers that host the VMs, the chassis and datastores that provide resources to the blade servers, the IO modules and fabric interconnects that provide network resources, and the virtual datastores that provide storage resources to the VMs.

As part of this process, Intersight Workload Optimizer will stitch information from the fabric target and connected hypervisor targets to provide more granular data and information related to the applications and VMs running on the hypervisor-stitched blade servers. Combined with other targets, this information will support a hierarchical, application-driven approach to managing your environment.

For example:

When Intersight Workload Optimizer discovers that blade servers housed in a particular chassis have been designated as vCenter hosts, the supply chain stitches the blade servers and chassis to the corresponding vCenter data center to establish their relationship. When you set the scope to that data center and view the Health chart, you will see the blade servers in the list of hosts. In addition, when the data center is included in a merge policy (a policy that merges data centers for the purpose of VM placement), the VMs in the blade servers apply the policy, allowing them to move between data centers as necessary.

When you add application server targets, your applications and their individual components and services are discovered, enabling a view of your infrastructure from an individual application service to the physical hardware. Adding public cloud targets also allow for workloads to potentially migrate from your UCS infrastructure to the cloud, based on cost or available resources.

Supply Chain

Fabric targets add IO Module, Fabric Interconnect, Domain, and Chassis entities to the supply chain. The Chassis entities host physical machines (blade servers) – The physical machines also consume network connection commodities from IO Modules.

The Fabric Interconnect supplies connectivity to the overall network, and also hosts the UCS Manager for UCS Targets. The Domain serves as the bottom-level pool of network resource, supplying the Fabric Interconnect.

Cisco UCS Manager

The Cisco Unified Computing System (UCS) Manager is a management solution that participates in server, fabric, and storage provisioning, device discovery, inventory, configuration, diagnostics, monitoring, fault detection, auditing, and statistics collection. Intersight Workload Optimizer discovers these targets automatically.

UCS integrates all of these resources in a scalable multi-chassis platform to converge administration onto a single point. Managing these various entities on a network fabric with Intersight Workload Optimizer enables automation at the hardware level, including automated provisioning of hosts.

Intersight Managed UCS

Intersight Workload Optimizer supports Cisco UCS with Intersight Managed Mode.

Intersight Managed Mode (IMM) is a new architecture that manages the UCS Fabric Interconnected systems through a Redfish-based standard model. Intersight Managed Mode unifies the capabilities of the UCS Systems and the cloud-based flexibility of Intersight, thus unifying the management experience for the stand-alone and Fabric Interconnect attached systems. Intersight Management Model standardizes policy and operation management for UCS-FI-6454, UCS-FI-64108, UCS-FI-6536, and Cisco UCS M5, M6, and X-Series servers.

Intersight Workload Optimizer can discover these targets and various entities that belong to UCS. Intersight Workload Optimizer can also discover and list down Chassis, Redfish server hosts, IO Modules, Fabric Interconnects and Networks attached to the target connected. Support for usage statistics and automation will be added in the future.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| UCS | Intersight Workload Optimizer |
|----------------------------|-------------------------------|
| Server / Blade / Rack Unit | Host |
| Chassis | Chassis |
| Datacenter | Datacenter |
| IO Module | IO Module |
| Fabric Interconnect | Switch |
| Network | Network |

Fabric targets add IO Module, Fabric Interconnect (Switch), and Chassis entities to the supply chain. Hosts consume resources from Chassis entities, and network connection commodities from IO Modules. The Fabric Interconnect supplies connectivity to the overall network, and also hosts the UCS Manager. The Domain serves as the bottom-level pool of network resource, supplying the Fabric Interconnect. Be sure that all the FC, Ether, and Physical Ports are properly configured with suitable roles in UCS so that the supply chain is populated accurately without disjoints.

Claiming UCS Targets

If your installation of Cisco Intersight already claims your UCS device, then Intersight Workload Optimizer discovers the UCS environment automatically.

To claim a new UCS device, select the **Compute / Fabric** category and choose the type of device you want for a target. Then provide the following:

- Device ID

Enter the applicable Device ID. Endpoint devices connect to the Cisco Intersight portal through a Device Connector that is embedded in the management controller (Management VM for Cisco UCS Director) of each system. The Device Connector

provides a secure way for connected devices to send information and receive control instructions from the Cisco Intersight portal, by using a secure internet connection.

- **Claim Code**

The Claim Code authorizes your access. You can find this code in the Device Connector.

- **Click Claim.**

After you provide the information, click **Claim**. You can see the status of your claimed target in the **Targets** tab.

The following table provides the format of the device ID and the device connector location:

| Targets | Device ID Format and Example | Device Connector Location |
|---------------------------|---|--|
| Stand-alone UCS Server | Serial Number Example: NGTR12345 | From Admin > Device Connector in Cisco IMC |
| Intersight Managed Domain | Serial ID of the primary or subordinate FIs in this format: Serial number of FI-A or Serial number of FI-B Example: [SAL1924GKV6&SAL1913CJ7V] | Go to the Device Connector tab in the Device Console. |
| Cisco UCS Manager | Serial ID of the primary and subordinate FIs in this format: Serial number of FI-A & Serial number of FI-B Example: [SAL1924GKV6&SAL1913CJ7V] | From Admin > Device Connector in Cisco IMC |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

In IMM mode, the resources in the following table are not supported. Support will be added in the future.

- **Host**

- Power
Power is the measurement of electricity consumed by a given entity, expressed in watts.
- Memory (Mem)
Memory is the measurement of memory that is reserved or in use.
- CPU
CPU is the measurement of CPU that is reserved or in use.
- IO
IO is the utilization of a host's IO adapters.
- Net
Net is the utilization of data through the host's network adapters.
- Swap
Swap is the measurement of a host's swap space that is in use.
- Balloon
Balloon is the measurement of memory that is shared by VMs running on a host.
This commodity applies to ESX only.
- CPU Ready
CPU Ready is the measurement of a host's ready queue capacity that is in use.
This commodity applies to ESX only.

■ Chassis

- Power

Power is the measurement of electricity consumed by a given entity, expressed in watts.

- Cooling

Cooling is the percentage of the acceptable temperature range that is utilized by the entity. As the temperature nears the high or low running temperature limits, this percentage increases.

■ I/O Module

- Net Throughput

Net Throughput is the rate of message delivery over a port.

■ Switch

- Net Throughput

Net Throughput is the rate of message delivery over a port.

- PortChannel

PortChannel is the amalgamation of ports with a shared net throughput and utilization.

Actions

Intersight Workload Optimizer supports the following actions:

■ Host

- Start
- Provision
- Suspend

■ Chassis

- Provision

■ Switch

- Add Port to Port Channel
- Remove Port from Port Channel
- Add Port

HPE OneView

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

HPE OneView is a management solution that streamlines provisioning and lifecycle management across compute, storage, and fabric. Through a unified API, infrastructure can be configured, monitored, updated, and re-purposed.

HPE OneView integrates all of these resources in a scalable multi-enclosure platform to converge administration onto a single point. Managing these various entities on a network fabric with Intersight Workload Optimizer enables automation at the hardware level, including automated provisioning of hosts.

Prerequisites

- A service account Intersight Workload Optimizer can use to connect to HPE OneView.
- HPE OneView 2.0 and compatible hardware.
- The **Banner Page** option for the user account should be disabled in the HPE OneView user interface.
- You should disable **Require Acknowledgment** for the user account in the HPE OneView user interface.

Adding HPE OneView Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Fabric**.
3. Select **HPE OneView**.
4. Configure the following settings:
 - **Address**
Specify the IP address of the HPE OneView target. Intersight Workload Optimizer uses the HTTPS protocol by default. To force the HTTP protocol, specify the address as http://8.8.8.8 or 8.8.8.8:80.
This gives access to the Fabric Manager that resides on the VM.
 - **Username**
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
Specify the IP address and credentials for HPE OneView. Intersight Workload Optimizer discovers the fabric interfaces associated with that instance.

NOTE:
If the account is managed in Active Directory, include the case-sensitive domain name as part of the username. For example, MyDomain@john is not the same as mydomain@john. For local user accounts, just provide the username.
 - **Password**
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| HPE OneView | Intersight Workload Optimizer |
|---------------------|-------------------------------|
| IO Module | IO Module |
| Fabric Interconnect | Switch |
| Domain | Domain |
| Chassis | Physical Machines |

Fabric targets add IO Module, Fabric Interconnect (Switch), Domain, and Chassis entities to the supply chain. The Chassis entities host physical machines – The physical machines also consume network connection commodities from IO Modules. The Fabric Interconnect supplies connectivity to the overall network. The Domain serves as the bottom-level pool of network resource, supplying the Fabric Interconnect.

NOTE:

For HPE OneView targets, the "Fabric Interconnect" entity exists as a false "Switch", and only as a pass-through for network resources. Unlike other fabric targets, such as UCS, there is no physical hardware that serves this function.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Virtual Machine**

NOTE:

Intersight Workload Optimizer only monitors VM resources if HPE OneView is stitched to a hypervisor in the supply chain.

- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Storage
Virtual storage is the measurement of virtual storage capacity that is in use.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- Latency
Latency is the measurement of storage latency.

■ Host

- Power
Power is the measurement of electricity consumed by a given entity, expressed in watts.
- Memory (Mem)
Memory is the measurement of memory that is reserved or in use.
- CPU
CPU is the measurement of CPU that is reserved or in use.
- IO
IO is the utilization of a host's IO adapters.
- Net
Net is the utilization of data through the host's network adapters.
- Swap
Swap is the measurement of a host's swap space that is in use.
- Balloon
Balloon is the measurement of memory that is shared by VMs running on a host.
- CPU Ready
CPU Ready is the measurement of a host's ready queue capacity that is in use.

■ Chassis

- Power
Power is the measurement of electricity consumed by a given entity, expressed in watts.
- Cooling
Cooling is the percentage of the acceptable temperature range that is utilized by the entity. As the temperature nears the high or low running temperature limits, this percentage increases.

■ Storage

- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
- Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

■ I/O Module

- Net Throughput

Net Throughput is the rate of message delivery over a port.

■ Switch

- Net Throughput

Net Throughput is the rate of message delivery over a port.

- PortChannel

PortChannel is the amalgamation of ports with a shared net throughput and utilization.

Actions

Intersight Workload Optimizer supports the following actions:

■ Virtual Machine

- Provision additional resources (VMem, VCPU)
- Move Virtual Machine
- Move Virtual Machine Storage
- Reconfigure Storage
- Reconfigure Virtual Machine
- Suspend VM
- Provision VM

■ Host

- Start
- Provision
- Suspend

■ Switch

- Add Port to Port Channel
- Remove Port from Port Channel
- Add Port

Application Performance Management (APM)

For APM, Intersight Workload Optimizer supports Cisco AppDynamics targets. These targets add Business Application, Business Transaction, Service, Application Component, and Database entities to the supply chain. To see how these entities map to the AppDynamics nomenclature, see [Entity Mapping \(on page 150\)](#).

Cisco AppDynamics

NOTE:

This type of target can run as SaaS or in on-prem data centers. When you claim the target, you can choose to turn on or off **Connect through an Intersight Assist** as follows:

- If the target runs as SaaS:

Turn off **Connect through an Intersight Assist**.

In earlier versions of Intersight Workload Optimizer you were required to specify an Intersight Assist to claim an AppDynamics target running as a SaaS. If you claimed your target through an Assist, you can reclaim that target *without* using the Intersight Assist. To do that you must first delete the claimed target, and then claim the target anew with **Claim through an Intersight Assist** in the off position.

- If the target runs in an on-prem data center:

Turn on **Connect through an Intersight Assist**.

To establish communication between this target on the data center and Intersight Workload Optimizer, you must:

- Install an **Intersight Assist appliance** in the on-prem data center. The AppDynamics target must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.
- Claim the AppDynamics target with **Connect through an Intersight Assist** in the ON position.

Intersight Assist provides a secure way for on-prem targets to send information to and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports workload management of the application infrastructure that is monitored by AppDynamics, via adding the AppDynamics instance to Intersight Workload Optimizer as a target.

The Intersight Workload Optimizer integration with AppDynamics provides a full-stack view of your environment, from application to physical hardware. With information obtained from AppDynamics, Intersight Workload Optimizer is able to make recommendations and take actions to both assure performance and drive efficiency with the full knowledge of the demands of each individual application.

In its default configuration, the AppDynamics target collects up to 5000 AppDynamics nodes within the default collection period. Larger AppDynamics environments can take longer than one cycle to collect complete data.

Prerequisites

- A valid AppDynamics user account

For all types of application instances, the service account must have the `Read Only User` role. For monitoring database instances, this user must also have the `DB Monitoring User` role.

NOTE:

In newer versions of AppDynamics where these roles are available, they should be used instead:

- Applications and Dashboards Viewer
- DB Monitoring User
- Server Monitoring

To use a custom role, ensure that the role has the `View Server Visibility` permission for both applications and databases.

AppDynamics Database Servers

AppDynamics also monitors database servers. For your database servers to be correctly stitched to the rest of your environment, you must:

- Enable enhanced metric collection.

For Hyper-V hosts, you must install Hyper-V Integration Services on the target VM hosting the database. For more information, refer to the following integration services TechNet article:

<https://technet.microsoft.com/en-us/library/dn798297%28v=ws.11%29.aspx>

For VMware hosts, you must install VMware Tools on the target VMs.

- Ensure that the database name in AppDynamics is resolvable to an IP address by the Intersight Workload Optimizer instance.

You may need to make changes to your DNS or the file `/etc/resolv.conf` on the Intersight Workload Optimizer instance.

Claiming an AppDynamics Target

NOTE:

It is possible to monitor certain applications or database servers with both AppDynamics and Intersight Workload Optimizer, but this must be avoided as it causes the entities to appear duplicated in the market.

If an application is monitored by AppDynamics, do not add it as a separate Intersight Workload Optimizer application target.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **AppDynamics**.
4. Configure the following settings:
 - **Connect through an Intersight Assist**
 If you select this option, Intersight Workload Optimizer claims the target through an Intersight Assist instance.
 If your AppDynamics is deployed in your data center, then you must turn this ON and use an Intersight Assist to establish the connection with that target.
 If the target is a SaaS-based AppDynamics instance, then you should turn this option OFF.
 - **Intersight Assist**
 Specify the Intersight Assist instance that you use to claim this AppDynamics target.
 To provide this setting, you must turn on **Connect through an Intersight Assist**. You must also have already claimed at least one Intersight Assist instance.
 - **Hostname or IP Address**
 Specify the hostname or IP address of the AppDynamics controller instance.
 - **Port**
 Specify the port that Intersight Workload Optimizer uses to connect to the AppDynamics controller. By default, the HTTP port is 80 and the HTTPS port is 443. For SaaS-based AppDynamics instances, use port 443.
 - **Username**
 Specify the username or account ID with the "Read Only User" and "DB Monitoring User" permissions.
 The username can be found on the **License > Account** page in AppDynamics.
 The format must be *Central ID@tenant*, where *Central ID* can be either a username or an email ID and *tenant* is not a domain. If the Central ID is an email ID (such as *user@domain*), you must type `%40` instead of "@" in front of the domain since Central IDs must be URL encoded. The "@" symbol is accepted as a delimiter between the Central ID and tenant only.
 Examples: `<username>@tenant` or `<user%40domain>@tenant`.

NOTE:

The Central ID cannot contain any of the following special characters:

`\ / " [] : | < > + = ; , ? * , ' tab space @`

For Central IDs containing the "@" symbol, URL encode the "@" character as `%40`.

- **Password**
 Specify the password for the account used to connect to the AppDynamics instance.

NOTE:

The password cannot contain any of the following special characters:

\ / " [] : | < > + = ; , ? * , ' tab space @

For passwords containing the "@" symbol, URL encode the "@" character as %40.

- **Secure Connection**

If you select this option, Intersight Workload Optimizer connects to the target servers using HTTPS. Make sure that the required certificate is configured for use on the host.

- **Validate Server Certificates**

If you select this option, Intersight Workload Optimizer verifies the target certificate and proxy, if in use.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|--|-------------------------------|
| AppDynamics | Intersight Workload Optimizer |
| Business Application | Business Application |
| Business Transaction | Business Transaction |
| Tier | Service |
| Node | Application Component |
| Database | Database Server |
| Machine (when the machine type is Container) | Container |
| Server | Virtual Machine |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

The exact resources that are monitored will differ based on application type. This list includes all of the resources that you may see.

- **Application Component**

- Connection

Connection is the measurement of database connections utilized by applications.

- Heap

Heap is the portion of a VM or container's memory allocated to individual applications.

This commodity applies to Java, .NET, and Node.js only.

- Remaining GC Capacity

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Threads

Threads is the measurement of thread capacity utilized by applications.

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

This commodity applies to Java, .NET, and Node.js only.

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

This commodity applies to Java, .NET, and Node.js only.

- **Business Application**

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- **Business Transaction**

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- **Database Server**

- Connection

Connection is the measurement of database connections utilized by applications.

This commodity applies to MongoDB only.

- DB Cache Hit Rate

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

This commodity applies to SQL and Oracle only.

- Transaction Log

Transaction log is the measurement of storage capacity utilized by a Database Server for transaction logging.

This commodity applies to SQL only.

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

This commodity applies to SQL, MySQL, and Oracle only.

- **Service**

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

For container platform environments, this is the desired *weighted average* response time of all Application Component replicas associated with a Service.

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

For container platform environments, this is the maximum number of transactions per second that each Application Component replica can handle.

- **Virtual Machine**

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

NOTE:

To collect data, a machine agent must be present and database hardware monitoring must be enabled.

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

NOTE:

To collect data, a machine agent must be present and database hardware monitoring must be enabled.

For a VM, the resources you see depend on how the VM is discovered, and whether the VM provides resources for an application that is discovered by this target:

- If the VM hosts an application that is discovered through this target, then you see VM metrics that are discovered through this target.
- If the VM is discovered through a different target, and it does not host any application discovered through this target, you see VM metrics that are discovered through that different target.
- If the VM is discovered through this target, but it does not host any application that is discovered through this target, then Intersight Workload Optimizer does not display metrics for the VM.

Actions

NOTE:

The specific actions that Intersight Workload Optimizer recommends can differ, depending on the processes that Intersight Workload Optimizer discovers.

For other application components, Intersight Workload Optimizer can recommend actions based on the resources it can discover for the application. For example, Node.js® applications report CPU usage, so Intersight Workload Optimizer can generate vCPU resize actions and display them in the user interface.

Intersight Workload Optimizer supports the following actions:

- **Application Component**

- Resize Heap

This action can only be executed outside Intersight Workload Optimizer.

- **Database Server**

- Resize Connections

This action can only be executed outside Intersight Workload Optimizer.

- Resize Database Memory (DBMem)

This action can only be executed outside Intersight Workload Optimizer.

NOTE:

For different types of Database Servers, the AppDynamics target returns different metrics. This affects Intersight Workload Optimizer actions as follows:

- **MySQL:**

For MySQL database servers, analysis does not generate resize actions for DB Memory, Connections, or Transaction Log. The target does not discover DB Cache Hit Rate, DB Memory, Connections, or Transaction Log.

- **SQL Server:**

For SQL database servers, analysis does not generate resize actions for DB Memory, Connections, or Transaction Log. The target does not discover DB Memory or Connections.

- **MongoDB:**

For MongoDB database servers, analysis does not generate resize actions for DB Memory, Connections, or Transaction Log. The target does not discover DB Cache Hit Rate, DB Memory, Transactions, or Transaction Log.

- **Oracle:**

For Oracle database servers, analysis does not generate resize actions for DB Memory, Connections, or Transaction Log. The target does not discover DB Memory, Connections, or Transaction Log.

Dynatrace

NOTE:

This type of target can run as SaaS or in on-prem data centers. When you claim the target, you can choose to turn ON or turn OFF **Connect through an Intersight Assist** as follows:

- If the target runs as SaaS:
Turn OFF **Connect through an Intersight Assist**.
- If the target runs in an on-prem data center:
Turn ON **Connect through an Intersight Assist**.

To establish communication between this target on the data center and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem data center. The target that you are configuring must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.
- Claim the target that you are configuring with **Connect through an Intersight Assist** in the ON position.

Intersight Assist provides a secure way for on-prem targets to send information to and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports discovery of applications that are managed by the Dynatrace platform. Intersight Workload Optimizer includes the discovered information about these applications in its calculations for VM actions.

Prerequisites

- A Dynatrace Server instance
This instance must be configured to monitor applications that are running in your environment. Intersight Workload Optimizer supports both SaaS and on-prem Dynatrace server installations.
- Managed VMs that host applications managed by Dynatrace
For Intersight Workload Optimizer to discover applications through Dynatrace, the applications must be running on VMs in your environment. Also, Intersight Workload Optimizer targets such as hypervisors or public cloud targets must manage these VMs.
- An API access token with the proper scopes
Intersight Workload Optimizer uses the API token to authenticate its calls to the Dynatrace API. This token must have permission to run GET methods using the Dynatrace API, both Version 1 and Version 2. Generate a new generic access token with these scopes:

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| Monitoring | <ul style="list-style-type: none"> - API V1 scopes: <ul style="list-style-type: none"> • Access problem and event feed, metrics, and topology - API V2 scopes: <ul style="list-style-type: none"> • Read entities • Read metrics |

NOTE:

If you are updating to Intersight Workload Optimizer version or later, from a version that is earlier than , you must generate a new API token for each existing Dynatrace target. Then you must enter that token in the target configuration, and validate the target.

If the target still fails to validate after you update the access token, take note of your configuration settings, delete the target, and configure the target again. Be sure to use the new API token that you have generated.

- Custom calculated service metrics

For Intersight Workload Optimizer to discover Response Time and Transaction metrics for Dynatrace Application Component, you must configure custom calculated service metrics in Dynatrace. For details, see [Dynatrace Custom Calculated Service Metrics \(on page 157\)](#).

- The Environment ID

To claim a Dynatrace target, you must know the Environment ID for the Dynatrace installation. According to the Dynatrace documentation, you can identify the Environment ID in these ways:

- SaaS-based Dynatrace Server:

The Environment ID is the first part of the Dynatrace environment URL. For example, for the environment `https://abc123a.live.dynatrace.com`, the Environment ID is `abc123a`

- On-prem Dynatrace Server:

The Environment ID is the string after `/e/` in the Dynatrace environment URL. For example, for the environment address `https://managed-cluster/e/abc123a`, the Environment ID is `abc123a`

For more information, see the Dynatrace documentation at <https://www.dynatrace.com/support/help/get-started/monitoring-environment/environment-id/>

Claiming a Dynatrace Target

NOTE:

You can manage certain applications or database servers with both Dynatrace and Intersight Workload Optimizer. Avoid such a configuration because it can cause Intersight Workload Optimizer to generate duplicate entities in the market.

If you manage an application by using a Dynatrace server, and you configure that Dynatrace server as a Intersight Workload Optimizer target, ensure you have not added that application as a separate application target in Intersight Workload Optimizer.

On-Prem Dynatrace Target:

To claim an on-prem Dynatrace server instance as a target, specify:

- Connect through an Intersight Assist

Whether to claim the target by using an Intersight Assist instance.

If your Dynatrace server is deployed in your data center, then you must turn this ON and use an Intersight Assist to establish the connection with that target.

- Intersight Assist

The Intersight Assist instance that you use to claim this Dynatrace target.

To provide this setting, you must turn on **Connect through an Intersight Assist**. You must also have already claimed at least one Intersight Assist instance.

- Hostname or IP Address

For an on-prem installation of Dynatrace, give the hostname or IP for your Dynatrace server. For example, `10.10.12.34`.

- Environment ID

The unique string that identifies the environment this Dynatrace target manages.

- API Key

The token that Intersight Workload Optimizer can use to authenticate its calls to the Dynatrace API. This token must have permission to run GET methods by using the Dynatrace API V1 and V2. For more information, see the Prerequisites section.

SaaS-Based Dynatrace Target:

To claim a SaaS-based Dynatrace server instance as a target, specify:

- Connect through an Intersight Assist

Whether to claim the target via an Intersight Assist instance.

If the target is a SaaS-based Dynatrace server, turn this option OFF.

- Environment ID

The unique string that identifies the environment this Dynatrace target manages.

- API Key

The token that Intersight Workload Optimizer can use to authenticate its calls to the Dynatrace API. This token must have permission to run GET methods by using the Dynatrace API V1 and V2. For more information, see the Prerequisites section.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| Dynatrace | Intersight Workload Optimizer |
|-------------|---|
| Application | Business Application NOTE: For Dynatrace Applications, Intersight Workload Optimizer displays Business Application entities in the supply chain when they are active for at least the past 10 minutes. |
| Service | Service |
| Process | Application Component, Database Server |
| NA | Container |
| Host | Virtual Machine |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

The exact resources that are monitored will differ based on application type. This list includes all of the resources that you may see.

■ Application Component

- Heap
Heap is the portion of a VM or container's memory allocated to individual applications.
This commodity applies only to Java applications.
- Remaining GC Capacity
Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).
This commodity applies only to Java applications.
- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.

■ Business Application

- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

■ Container

- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.

■ Database Server

For Database Server applications, Intersight Workload Optimizer discovers metrics for MySQL and SQL Server databases only.

- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Database Memory (DBMem)
Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.
This commodity applies only to SQL and MySQL databases.
Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM.
- DB Cache Hit Rate
DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.
This commodity applies only to SQL databases.
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
This commodity applies only to SQL databases.

■ Service

- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

■ Virtual Machine

- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.

For a VM, the resources you see depend on how the VM is discovered, and whether the VM provides resources for an application that is discovered by this target:

- If the VM hosts an application that is discovered through this target, then you see VM metrics that are discovered through this target.
- If the VM is discovered through a different target, and it does not host any application discovered through this target, you see VM metrics that are discovered through that different target.
- If the VM is discovered through this target, but it does not host any application that is discovered through this target, then Intersight Workload Optimizer does not display metrics for the VM.

Actions

Intersight Workload Optimizer supports the following actions:

- **Application Component**

- Resize Heap

This action can only be executed outside Intersight Workload Optimizer.

- **Database Server**

- Resize Database Memory (DBMem)

This action can only be executed outside Intersight Workload Optimizer.

This commodity applies to MySQL only.

- **Workload Controller**

- **Scale**

Actions associated with a workload controller scale replicas horizontally. This is a natural representation of these actions because the parent controller's container specs and number of replicas are modified. The workload controller then rolls out the changes in the running environment.

For details, see [Workload Controller Scale Actions \(on page 238\)](#).

Dynatrace Custom Calculated Service Metrics

Dynatrace collects important metrics for services with no additional configurations. However, you might need more business or technical metrics that are specific to your application. These metrics can be calculated and derived based on the captured metric data with Dynatrace.

For Intersight Workload Optimizer to discover Dynatrace application component level metrics (such as Response Time and Transaction), you must configure custom calculated service metrics in Dynatrace.

NOTE:

Custom calculated service metrics use Dynatrace DDU (Davis data units).

Creating an Auto Tag for Service Entity Types

Before you can create custom calculated metrics, you must first create an auto tag for the service entity types.

Dynatrace Service entities are tagged with `service-autotag:service.custom.tag` after you create the auto tag.

1. Log in to your Dynatrace account.
2. Select **Settings > Tags > Automatically applied tags**.
3. Click **Create tag**.

Specify the following properties for the tag:

- **Tag name** - `service-autotag`

4. Click **Add a new rule**.

Specify the following properties for the rule:

- **Optional tag value** - `service.custom.tag`
- **Value Normalization** - Leave text as-is
- **Rule type** - Entity selector
- **Entity selector** - `type(SERVICE)`

5. Click **Save Changes**.

The tag now appears in the **Settings > Tags > Automatically applied tags** view.

6. Verify that the tag was created.
 - Select any Dynatrace Service.
 - In the **Properties and tags** section, verify that the new tag `service-autotag:service.custom.tag` is listed under **Tags**.

Creating Calculated Service Metrics for Response Time

NOTE:

Custom calculated service metrics use Dynatrace DDU (Davis data units).

Follow the steps to create the custom service metric `app.responsetime` to track Response time. The dimension `{Service:Instance}` named `service-instance` splits the values of this metric.

1. Log in to your Dynatrace account.
2. Select **Settings > Server-side service monitoring > Calculated service metrics**.
3. Click **Create new metric**.

Specify the following properties for the metric:

- **Metric name** - `app.responsetime`
 - **Metric source** - Response time
4. In the **Conditions** section, click **Add Condition**.
Specify the following properties for the condition:

```
Service tag equals service-autotag service.custom.tag
```

5. Click **Add**.
The selected services display in the **Preview** section.
6. Turn on the **Split by dimension** toggle to add dimensions for the metric.

Specify the following properties:

- **Dimension value pattern** - `{Service:Instance}`
 - **Dimension name** - `service-instance`
7. Click **Save Metric**.

The metric now appears in the **Settings > Server-side service monitoring > Calculated service metrics** view.

Creating Calculated Service Metrics for Request Count

NOTE:

Custom calculated service metrics use Dynatrace DDU (Davis data units).

Follow the steps to create the custom service metric `app.requestcount` to track Request count. The dimension `{Service:Instance}` named `service-instance` splits the values of this metric.

1. Log in to your Dynatrace account.
2. Select **Settings > Server-side service monitoring > Calculated service metrics**.
3. Click **Create new metric**.

Specify the following properties for the metric:

- **Metric name** - `app.requestcount`
 - **Metric source** - Request count
4. In the **Conditions** section, click **Add Condition**.
Specify the following properties for the condition:

```
Service tag equals service-autotag service.custom.tag
```

5. Click **Add**.
The selected services display in the **Preview** section.
6. Turn on the **Split by dimension** toggle to add dimensions for the new metric.

Specify the following properties:

- **Dimension value pattern** - `{Service:Instance}`
 - **Dimension name** - `service-instance`
7. Click **Save Metric**.

The metric now appears in the **Settings > Server-side service monitoring > Calculated service metrics** view.

New Relic

Intersight Workload Optimizer supports workload management of the application infrastructure monitored by New Relic, from application instance to host. With information obtained from New Relic, Intersight Workload Optimizer can make recommendations and take actions to both assure performance and drive efficiency to address the demands of each individual application.

For container platform environments, Intersight Workload Optimizer stitches containerized application components into the supply chain to provide a unified view of your applications.

Prerequisites

- A valid New Relic user account that includes both APM and infrastructure monitoring.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| New Relic | Intersight Workload Optimizer |
|--|-------------------------------|
| APM: Key Transactions | Business Transaction |
| APM: Application / Service (New Relic One) | Service |
| APM: Application Instance | Application Component |
| Infra: Database | Database Server |
| Infra: Host | Virtual Machine |

Supported Applications

Intersight Workload Optimizer discovers the following application types (and associated commodities) via the New Relic target:

| Application Type | Commodities |
|------------------|--|
| .NET | Virtual CPU, Virtual Memory, Response Time, Transactions |
| GO | Virtual CPU, Virtual Memory, Response Time, Transactions |
| Java | Virtual CPU, Virtual Memory, Response Time, Transactions, Heap, Collection Time, Threads |
| Node.js | Virtual CPU, Virtual Memory, Response Time, Transactions, Heap, Collection Time |
| PHP | Virtual CPU, Virtual Memory, Response Time, Transactions |
| Python | Virtual CPU, Virtual Memory, Response Time, Transactions |

Supported Databases

Intersight Workload Optimizer supports the following Database types and commodities:

NOTE: Database commodities are exposed only if the New Relic account used to connect to Intersight Workload Optimizer has a `New Relic Infrastructure Pro` subscription.

| Database | Commodities |
|----------|--|
| SQL | Cache Hit Rate, Virtual Memory, Transactions |

| Database | Commodities |
|----------|---|
| MySQL | Cache Hit Rate, Transactions NOTE: Intersight Workload Optimizer does not support MySQL version 8.0 or higher. For more information, see the New Relic documentation . |
| OracleDB | Cache Hit Rate, Transactions, Response Time |
| MongoDB | Virtual Memory, Connections |

Claiming a New Relic Target

NOTE:

If an application is monitored by New Relic, do not add it as a separate Intersight Workload Optimizer application target.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Applications and Databases**.
3. Select **New Relic**.
4. Configure the following settings:
 - **Name**
Specify a name that uniquely identifies this connection.
This name is for display purposes only and does not need to match any name in New Relic. The target name can contain alphanumeric, space, or hyphen characters.
 - **Account ID**
Specify the New Relic Account ID. It is recommended to use an Account ID to enable the New Relic APM service for Intersight Workload Optimizer to monitor Business Applications, Business Transactions, Services and Application Components.
 - **User Key**
Specify the User Key that is provided by the New Relic platform. It is recommended to use User Key instead of REST API Key and GraphQL API Key.
 - **Use REST API Key and GraphQL API Key toggle**

NOTE:

The REST API and GraphQL API keys have been deprecated by New Relic. Intersight Workload Optimizer will remove the two fields by version , and you will have to use the User Key field when adding New Relic as a target.

When enabled, Intersight Workload Optimizer uses the REST API Key and GraphQL API Key instead of the User Key.

- **REST API Key**
Specify the REST API Key that is provided by the New Relic platform.
In the New Relic environment, this key is called the User Key. For more information, see the [New Relic documentation](#).
- **GraphQL API Key**
Specify the GraphQL API Key that is provided by the New Relic platform.
In the New Relic environment, this key is called the User Key. For more information, see the [New Relic documentation](#).

For more information, see the [New Relic documentation](#).

- **Region**
If you select this option, Intersight Workload Optimizer uses the EU API endpoints.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

The exact resources that are monitored will differ based on application type. This list includes all of the resources that you may see.

■ Application Component

- Heap
Heap is the portion of a VM or container's memory allocated to individual applications.
- Remaining GC Capacity
Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).
- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Threads
Threads is the measurement of thread capacity utilized by applications.
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.

■ Database
NOTE:

Refer to the supported [Database types and commodities \(on page 159\)](#).

- Connection
Connection is the measurement of database connections utilized by applications.
- Database Memory (DBMem)
Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.
- DB Cache Hit Rate
DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.
- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.

■ Business Transaction

- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

■ Service

- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

- **Container**

- VCPU

VCPU is the virtual CPU (in mCores) utilized by a container against the CPU limit. If no limit is set, node capacity is used).

- Virtual Memory (VMem)

VMem is the virtual memory utilized by a container against the memory limit. If no limit is set, node capacity is used.

- **Virtual Machine**

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

For a VM, the resources you see depend on how the VM is discovered, and whether the VM provides resources for an application that is discovered by this target:

- If the VM hosts an application that is discovered through this target, then you see VM metrics that are discovered through this target.
- If the VM is discovered through a different target, and it does not host any application discovered through this target, you see VM metrics that are discovered through that different target.
- If the VM is discovered through this target, but it does not host any application that is discovered through this target, then Intersight Workload Optimizer does not display metrics for the VM.

Actions

NOTE:

The specific actions that Intersight Workload Optimizer recommends can differ, depending on the processes that Intersight Workload Optimizer discovers.

For other application components, Intersight Workload Optimizer can recommend actions based on the resources it can discover for the application. For example, Node.js® applications report CPU usage, so Intersight Workload Optimizer can generate vCPU resize actions and display them in the user interface.

Intersight Workload Optimizer supports the following actions:

- **Application Component**

- Resize Heap

This action can only be executed outside Intersight Workload Optimizer.

- Resize Threads

This action can only be executed outside Intersight Workload Optimizer.

- **Virtual Machine**

- Provision VM

This action can only be executed outside Intersight Workload Optimizer.

- Suspend VM

This action can only be executed outside Intersight Workload Optimizer.

Hyperconverged Targets

A hyper converged target is a service that unites compute, network and storage access into a cohesive system. When you connect Intersight Workload Optimizer to hyper converged targets, it will monitor the performance and resource consumption of your hyper converged infrastructure to maintain application performance while utilizing resources as efficiently as possible.

As part of this process, Intersight Workload Optimizer will stitch information from the hyper converged target to the associated hypervisor and fabric targets, supporting Application Resource Management (ARM) and providing deeper insight into the state of the hardware and information related to the entities in the supply chain. Combined with application server targets, this information will support a hierarchical, application-driven approach to managing your environment.

Cisco HyperFlex

Cisco HyperFlex provides a hyperconverged platform that combines the networking and compute power of UCS with the storage capabilities of the HyperFlex HX Data Platform. Intersight Workload Optimizer discovers these targets automatically.

With the additional and refined storage information provided by HyperFlex, Intersight Workload Optimizer narrows the Desired State and recommends actions using the joint compute and storage information, gaining valuable insight into the interconnected nature of your environment.

HyperFlex environments typically include:

- **Converged (HX) Nodes**
A combination of the cluster's storage devices into a single multi-tiered, object-based datastore.
- **Compute Nodes**
Cisco B or C series servers that make up the compute resources of the cluster, and are typically managed by a hypervisor.
- **Controller VMs**
Each HyperFlex node includes a Controller VM that intercepts and handles all the I/O from associated virtual machines. Intersight Workload Optimizer will not recommend actions for these VMs.

Claiming HyperFlex Targets

If your installation of Cisco Intersight has already claimed your HyperFlex device, then Intersight Workload Optimizer discovers the HyperFlex environment automatically.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Hyperconverged**.
3. Select **Cisco HyperFlex**.
4. Configure the following settings:
 - **Device ID:**
Enter the applicable Device ID. Endpoint devices connect to the Cisco Intersight portal through a Device Connector that is embedded in the management controller (Management VM for Cisco UCS Director) of each system. The Device Connector provides a secure way for connected devices to send information and receive control instructions from the Cisco Intersight portal, using a secure Internet connection.
 - The device ID format is the Cluster UUID. For example, xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx
 - Locate the device connector location via **HyperFlex Connect UI > Settings > Device Connector** in Cisco HyperFlex.
 - **Claim Code:**
The Claim Code authorizes your access. You can find this code in the Device Connector.
 - **Click Claim:**
After you provide the information, click **Claim**. You can see the status of your claimed target in the **Targets** tab.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|------------|-------------------------------|
| HyperFlex | Intersight Workload Optimizer |
| Volume | Storage |
| HX Cluster | Disk Array |

HyperFlex targets add Disk Array entities to the supply chain, and receive more granular information from the compute resources in your environment.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Storage

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

■ Disk Array

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

Actions

Intersight Workload Optimizer supports the following actions:

■ Storage

- Move

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

■ Disk Array

- Provision

This action can only be executed outside Intersight Workload Optimizer.

- Suspend

This action can only be executed outside Intersight Workload Optimizer.

- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

NOTE:

For this target, Intersight Workload Optimizer discovers the HX Cluster as a Disk Array. When you see a provision action on this entity, you should determine which of the following is most relevant, based on your environment:

- Add disks to converged nodes
- Add a new converged node
- Add a new HX Cluster

Nutanix Acropolis

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Nutanix products provide hyperconverged platforms that include VM hosting and a distributed storage fabric. The platform presents storage in two tiers – Local HDD storage and server-attached flash (hot storage).

Nutanix environments may include:

- One or more Nutanix appliances
 - An appliance contains up to four server nodes.
- Nutanix nodes
 - Servers that expose compute and storage resources. Each node provides local HDD and hot storage. Nodes combine to form a unified cluster that pools resources.
- Controller VMs
 - Each node includes a Controller VM that manages the node's resources within the cluster pool. To minimize storage latency, the Controller VM keeps the most frequently accessed data in the hot storage.

Intersight Workload Optimizer supports management of Nutanix fabrics, where the supply chain treats a Nutanix Storage Pool as a disk array. Intersight Workload Optimizer recognizes Nutanix storage tiers when calculating placement of VMs and VStorage. In addition, Intersight Workload Optimizer can recommend actions to scale flash capacity up or down by adding more hosts to the cluster, or more flash drives to the hosts.

To specify a Nutanix Acropolis target, provide the Cluster External IP address. This is a logical IP address that always connects to one of the active Controller VMs in the cluster. In this way, you can specify a Nutanix target without having to specify an explicit Controller VM.

NOTE:

The Controller VM must remain *pinned* to its host machine – You must not move the Controller VM to a different host. If the Nutanix cluster uses the Nutanix Acropolis OS to manage VMs, Intersight Workload Optimizer automatically pins the Controller VMs. However, if you use vCenter Server or Hyper-V to manage VMs on the hosts, you must configure a group to pin the Controller VMs. For more information, see [Pinning Nutanix Controller VMs \(on page 168\)](#).

Prerequisites

- A service account with cluster administrator rights on the Nutanix clusters for action execution. For entity discovery, a minimum of READ access is required.

Finding the Cluster External IP Address

To configure a Nutanix Acropolis target, provide the Cluster External IP address for the given Nutanix cluster.

The Cluster External IP address is a logical IP that resolves to the cluster's Prism Element Leader. If the Prism Element Leader fails, then the Cluster External IP address will resolve to the newly elected Prism Element Leader.

To find this IP address, open the Web Console (the Prism Element) on the cluster and navigate to the **Cluster Details** view. In this view you can see the **Cluster External IP** address. If there is no IP address specified, you can specify the address at this time. For more information, see the Nutanix documentation.

Operating Modes

A Nutanix node is a server that hosts VMs – In this sense the node functions as a hypervisor. A cluster of nodes can host VMs using the following Hypervisor technologies:

- Nutanix Acropolis
The native Nutanix host platform, which combines software-defined storage with built-in virtualization.
- VMware ESXi
- Microsoft Hyper-V

Controller VM Pinning

Each Nutanix node hosts a Controller VM that runs the Nutanix software and manages I/O for the hypervisor and all VMs running on the host. Each Controller VM must remain on its host node –The Controller VM must be *pinned* to that host, and must not be moved to any other host.

For more information about how to pin the Controller VM, see [Pinning Nutanix Controller VMs \(on page 168\)](#).

Adding Nutanix Targets

NOTE:

This describes how to add a Nutanix cluster to Intersight Workload Optimizer as a target. If Nutanix is not managing hosts running Acropolis as the hypervisor, you will have to add the vCenter or Hyper-V hypervisors as targets after you have added the Nutanix cluster as a target. For more information, see [Hypervisor Targets \(on page 169\)](#).

1. Click **Settings > Target Configuration**.
2. Click **New Target > Hyperconverged**.
3. Select **Nutanix**.
4. Configure the following settings:
 - Address
Specify the Cluster External IP address for the Nutanix cluster.
 - Port Number
Specify the listening port of the cluster.
 - Secure Connection
If you select this option, Intersight Workload Optimizer connects to the target servers using HTTPS. Make sure that the required certificate is configured for use on the host.
 - Username
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
 - Password
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|--------------|-------------------------------|
| Nutanix | Intersight Workload Optimizer |
| Container | Storage |
| Storage Pool | Disk Array |

| | |
|-----------------|-------------------------------|
| Nutanix | Intersight Workload Optimizer |
| Nutanix Cluster | Storage Controller |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Storage

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

■ Disk Array

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ Storage Controller

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

■ Virtual Machine (Nutanix VM)

- Move

VMotion to hosts can be automated, but storage moves on Nutanix can only be executed outside Intersight Workload Optimizer.

- Resize

Resize actions require the VM to power down, and power back on again.

- **Datastore (Storage)**

- Provision
- Resize Up/Down
- Suspend
- Move

This action can only be executed outside Intersight Workload Optimizer.

- **Storage Controller**

- Provision

This action can only be executed outside Intersight Workload Optimizer.

Pinning Nutanix Controller VMs

Each Nutanix node hosts a Controller VM that runs the Nutanix software and manages I/O for the hypervisor and all VMs running on the host. Each Controller VM must remain on its host node –The Controller VM must be *pinned* to that host, and must not be moved to any other host.

For a cluster using vCenter or Hyper-V hypervisors, you must use Intersight Workload Optimizer policies to pin the Controller VMs to their respective nodes. To do this, you will create a dynamic group of Nutanix Controller VMs, and then disable move actions for all members of this group.

To pin the Controller VMs:

1. Create a group of Controller VMs.

In Intersight Workload Optimizer you can create dynamic groups based on VM name – All VMs with matching names automatically belong to the group. Nutanix uses the following naming convention for Control VMs:

NTNX-*<SerialNumber>*-A-CVM, where *<SerialNumber>* is the serial number of the Controller VM.

You can create a dynamic group that automatically includes these Nutanix controller VMs. (For complete instructions on creating groups, see [Creating Groups \(on page 566\)](#).)

- Create a new group.

In Intersight Workload Optimizer navigate to **IWO > More > Settings > Groups** and create a new group.

- Set the group type to **Dynamic**.
- Add a filter to match VMs by their names.

Add a filter that uses the regular expression, `NTNX.*CVM`. This regular expression will match the Nutanix Controller VMs.

Be sure to save the group. All the Nutanix Controller VMs will automatically become members of this group.

2. Disable moves for all VMs in this group.

To do this, create an automation policy for the group and disable actions. (For complete instructions to create these policies, see [Creating Automation Policies \(on page 577\)](#).)

- In Intersight Workload Optimizer, navigate to **IWO > More > Settings > Policy**, and create an automation policy for VMs.
- Set the scope to the group you made.
- Disable moves for this group.
- Save the action acceptance mode.

Be sure to click **Apply Settings Change**.

Hypervisor Targets

A hypervisor is a service that creates and runs virtual machines (VMs), containers, or both, and provides these entities compute and storage resources. When you connect Intersight Workload Optimizer to hypervisor targets in your environment, Intersight Workload Optimizer assures application performance by using these resources as efficiently as possible.

After connected to a hypervisor target, Intersight Workload Optimizer discovers the VMs, containers, or physical machines that host the VMs or containers, data stores that provide storage resources to the physical machines, and virtual data stores that provide storage resources.

As more targets are added, Intersight Workload Optimizer discovers the resources belonging to your physical and virtual infrastructure. For example, adding the underlying hardware as part of a UCS, storage target, or both provides more visibility into the physical infrastructure of your environment. To extend the virtual infrastructure, application server or guest operating process targets can be added.

Intersight Workload Optimizer represents your environment holistically as a supply chain of resource buyers and sellers, all working together to meet application demand. Intersight Workload Optimizer maintains your environment within the desired state by empowering buyers and sellers:

- Buyers (VMs, instances, containers, and services) are given a budget to seek the resources that applications need to perform.
- Sellers can price their available resources (CPU, memory, storage, network) based on utilization in real-time.

Intersight Workload Optimizer does not support moves across stand-alone hosts or the merging of stand-alone hosts in any hypervisor. Intersight Workload Optimizer supports only VM moves across host cluster and merging host clusters.

For more information, see [Application Resource Management \(on page 13\)](#).

Supply Chain

Each hypervisor requires a physical machine (host) and one or more data stores to provide compute and storage resources. Virtual machines (VMs) or containers run on those physical resources, and the VMs in turn provide resources to applications.

At the end of the supply chain, physical machines consume resources from data centers.

If your environment includes SAN technologies such as disk arrays, then the storage consumes resources from that underlying technology. If you add these storage targets, then Intersight Workload Optimizer extends the supply chain analysis into the components that make up the disk array. For more information, see [Storage Manager Targets \(on page 183\)](#).

Microsoft Hyper-V

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

If you have a few Hyper-V hosts in your environment, you can add them individually as Intersight Workload Optimizer targets. Also, if you have deployed the Hyper-V hosts in a clustered domain (for example as a failover cluster), you can specify one Hyper-V host as a target and Intersight Workload Optimizer automatically adds the other members of that cluster.

For accurate Server Message Block (SMB) storage calculations, Intersight Workload Optimizer requires a VMM target.

Prerequisites

- [Create a service user account in Hyper-V. \(on page 172\)](#)
- [Configure remote management on each Hyper-V server. \(on page 173\)](#)

- Sync the time on each Hyper-V host with the rest of the managed Hyper-V environment.
- Ensure that your Hyper-V environment does not use Server Message Block (SMB) storage.
To manage SMB storage, Intersight Workload Optimizer requires a VMM target, and that VMM instance must manage the Hyper-V hypervisors and the SMB storage that they use.
Managing a Hyper-V plus SMB environment via Hyper-V targets will result in incorrect data collection for SMB storage.

Adding Hyper-V Targets

Once you've enabled remote management, you can add your Hyper-V hosts as targets.

1. Click **Settings > Target Configuration**.
2. Click **New Target > Hypervisors**.
3. Select **Hyper-V**.
4. Configure the following settings:
 - **Address**
Specify the fully qualified domain name of the Hyper-V host. If you are using the "Discover Host Cluster" option to add an entire cluster, enter the name of any one of the Hyper-V hosts in the cluster.

NOTE:
You can enter an IP address for the host, but you must first configure an SPN on the host. Cisco recommends that you use the FQDN in this field.
 - **Username**
Specify the username of the account Intersight Workload Optimizer uses to connect to the target.
If you select "Discover Host Cluster", use an account that is valid for all Hyper-V hosts in that cluster.
 - **Fully qualified domain name**
Specify the fully qualified domain name of the cluster to which the host belongs.
 - **Password**
Specify the password of the account Intersight Workload Optimizer uses to connect to the target.
 - **Port number**
Specify the port number that Intersight Workload Optimizer uses to connect to remote management. By default, this is port 5985 for HTTP and port 5986 for HTTPS.
 - **Discover Host Cluster**
If you select this option, Intersight Workload Optimizer discovers and adds all Hyper-V hosts in the named cluster. Each server must be configured to allow remote management.

You may find it helpful to configure WinRM using a GPO so new servers are configured automatically (see [Enabling WinRM Via a GPO \(on page 174\)](#)).
 - **Secure Connection**
If you select this option, Intersight Workload Optimizer connects to the target servers using HTTPS. Make sure that the required certificate is configured for use on the host.

Exporting Hyper-V Virtual Machines

In Hyper-V environments, you must be sure that all VMs have unique IDs.

Hyper-V supports the export of a VM, so that you can create exact copies of it by importing those exported files. The `Copy` import type creates a new unique ID for the imported VM. When importing VMs in your environment, you should always use the `Copy` import type.

Intersight Workload Optimizer uses the unique ID to discover and track a VM. If your environment includes multiple VMs with the same ID, then discovery assumes they are the same VM. As a result, the counts for VMs are incorrect.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Virtual Machine

- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Storage
Virtual storage is the measurement of virtual storage capacity that is in use.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- Latency
Latency is the measurement of storage latency.

■ Host

- Memory (Mem)
Memory is the measurement of memory that is reserved or in use.
- CPU
CPU is the measurement of CPU that is reserved or in use.
- IO
IO is the utilization of a host's IO adapters.
- Net
Net is the utilization of data through the host's network adapters.
- Swap
Swap is the measurement of a host's swap space that is in use.

■ Storage

- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
- Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS that it discovers on Logical Pool or Disk Array entities.

- Latency
Latency is the measurement of storage latency.

■ Datacenter

NOTE:

For datacenter entities, Intersight Workload Optimizer does not monitor resources directly from the datacenter, but from the hosts in the datacenter. See host monitored resources for details.

Actions

Intersight Workload Optimizer supports the following actions:

■ Virtual Machine

- Start
- Move
- Suspend
- Resize Up/Down
- Terminate

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

- Reconfigure

This action can only be executed outside Intersight Workload Optimizer.

■ Host

- Start
- Suspend
- Terminate

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

■ Storage

- Provision

This action can only be executed outside Intersight Workload Optimizer.

Creating A Service User Account

The service account Intersight Workload Optimizer uses to connect to a Hyper-V host must be an Active Directory domain account. The account must have full access to the cluster. To create such an account, execute the following command at a PowerShell prompt:

```
Grant-ClusterAccess <domain>\<service_account> -Full
```

Additionally, the service account must have specific local access rights on each host. The easiest way to grant Intersight Workload Optimizer the access it requires is to add the domain account to the Local Administrators group on each Hyper-V server.

Some enterprises require that the service account does not grant full administrator rights. In that case, you can create a restricted service account on every Hyper-V host.

NOTE:

Intersight Workload Optimizer does not support Restricted User Accounts on Windows 2012 Hyper-V nodes.

To create a restricted service account on your Hyper-V hosts:

1. Add the service account to each of the following local groups:
 - WinRMRemoteWMIUsers__ or Remote Management Users
 - Hyper-V Administrators
 - Performance Monitor Users

NOTE:

These groups are examples only. If your version of Windows Server does not include these groups, contact Technical Support for assistance.

2. Grant permissions to the service account.

In the WMI Management console, grant the following permissions to the service account:

- Enable Account
- Remote Enable
- Act as Operating System (for Windows 2016)

3. Configure the WinRM security descriptor to allow access by the service account:

- At a PowerShell prompt, run the command:

```
winrm configSDDL default
```

- In the "Permissions for Default" dialog box, grant the service account Read and Execute access.

Enabling Windows Remote Management

Intersight Workload Optimizer communicates with your Hyper-V servers using Web Services Management (WS-Management), which is implemented on Microsoft platforms using Windows Remote Management (WinRM). The following steps show how to enable WinRM on a single host, using the command line.

1. Ensure Windows Firewall is running on the host.

For you to configure WinRM successfully, Windows Firewall must be running on the host. For more information, see the Microsoft Knowledge Base article #2004640 (<http://support.microsoft.com/kb/2004640>).

2. Set up a Service Principal Name (SPN) for the host machine.

The machine must have a SPN of the form, `protocol/host_address`. For example, `WSMAN/10.99.9.2`.

To get a list of SPNs for the machine, run the following in the command window:

```
setspn -l <vmm-server-name>
```

If there is no valid SPN in the list, create one by running the command:

```
setspn -A protocol/host-address:port
```

where `port` is optional. For example:

```
setspn -A WSMAN/10.99.9.2:VMM-02
```

3. Set up the Windows Remote Management (WinRM) service to run on startup.

Run the `quickconfig` utility to set up the WinRM service. The `quickconfig` utility:

- Configures the WinRM service to auto-start
- Configures basic authentication and disables unencrypted traffic
- Creates a firewall exception for the current user profile
- Configures a listener for HTTP and HTTPS on any IP address
- Enables remote shell access

To run `quickconfig`, log into a command window as Administrator on the host machine. Then execute the following commands:

```
winrm quickconfig
```

Enter `y` to accept the `quickconfig` changes

4. Set permissions on the host machine.

Execute the following commands in the command window to modify the settings made by `quickconfig`:

- To set the memory capacity for remote shells:

```
winrm set winrm/config/winrs @{MaxMemoryPerShellMB="1024"}
```

- To set up an unsecured HTTP connection:

```
winrm set winrm/config/service @{AllowUnencrypted="true"}
```

```
winrm set winrm/config/service/Auth @{Basic="true"}
```

These steps showed you how to enable WinRM for a single host. Some users find the following methods useful for enabling WinRM on multiple hosts:

- [EnablingWinRmViaGlobal Policy Objects \(on page 174\)](#)
- [EnablingWinRMViaPowerShell \(on page 175\)](#)

Enabling WinRM Via Global Policy Objects

You can configure WinRM for all of your Hyper-V targets by creating and linking a Global Policy Object (GPO) within the Hyper-V domain and applying the GPO to all servers.

Follow the steps to enable Windows Remote Management (WinRM) for your Hyper-V targets.

1. On the AD domain controller, open the Group Policy Management Console (GPMC). If the GPMC is not installed, see <https://technet.microsoft.com/en-us/library/cc725932.aspx>.
2. Create a new Global Policy Object:
 - a. In the GPMC tree, right-click **Group Policy Objects** within the domain containing your Hyper-V servers.
 - b. Choose **Create a GPO in this domain**, and link it here.
 - c. Enter a name for the new GPO and click **OK**.
3. Specify the computers that need access:
 - a. Select the new GPO from the tree.
 - b. On the **Scope** tab, under **Security Filtering**, specify the computer or group of computers you want to grant access. Make sure you include all of your Hyper-V targets.
4. Right-click the new GPO and choose **Edit** to open the Group Policy Management Editor.
5. Configure the WinRM Service:
 - a. In the Group Policy Management Editor, select **Computer Configuration > Policies > Administrative Templates > Windows Components > Windows Remote Management (WinRM) > WinRM Service**.
 - b. Double-click each of following settings and configure as specified:

| Setting | Value |
|---|---------------------------|
| Allow automatic configuration of listeners ("Allow remote server management through WinRM" on older versions of Windows Server) | Enabled IPv4 filter: * |
| Allow Basic authentication | Enabled |

6. Configure the WinRM service to run automatically:
 - a. In the Group Policy Management Editor, expand **Computer Configuration > Preferences > Control Panel Settings**.
 - b. Under Control Panel Settings, right-click Services and choose **New > Service**.
 - c. In the New Service Properties window, configure the following settings:

| Setting | Value |
|--------------|-----------|
| Startup | Automatic |
| Service name | WinRM |

7. Enable Windows Remote Shell:
 - a. In the Group Policy Management Editor, select **Computer Configuration > Policies > Administrative Templates > Windows Components > Windows Remote Shell**.
 - b. Double-click the following setting and configure as specified:

| Setting | Value |
|----------------------------|---------|
| Allow Remote Shell Access: | Enabled |

8. Add a Windows Firewall exception:
 - a. In the Group Policy Management Editor, open **Computer Configuration > Windows Settings > Security Settings > Windows Firewall > Windows Firewall**.
 - b. Under Windows Firewall, right-click **Inbound Rules** and choose **New > Rule**.
 - c. In the New Inbound Rule Wizard, select **Predefined: Windows Remote Management and Allow the connection**.

The new group policy will be applied during the next policy process update. To apply the new policy immediately, run the following command at a Powershell prompt:

```
gpupdate /force
```

Enabling WinRM Via PowerShell

Using PsExec, you can run quickconfig on all your Hyper-V servers and change the default settings remotely. PsExec is a component of PsTools, which you can download from <https://technet.microsoft.com/en-us/sysinternals/bb897553.aspx>.

1. Create a text file containing the Hyper-V hostnames, for example:

```
hp-vx485
hp-vx486
```

2. Since Cisco requires changes to the default quickconfig settings, create a batch file containing the following command:

```
@echo off Powershell.exe Set-WSManQuickConfig -Force Powershell.exe Set-Item WSMan:\localhost\Shell\MaxMemoryPerShellMB 1024
```

NOTE:

If you are connecting via HTTP, you must include the following command:

```
Powershell.exe Set-Item WSMan:\localhost\Service\AllowUnencrypted -Value $True
```

3. Use PsExec to enable WinRM on the remote servers:

```
.\PsExec.exe @<hosts_file_path> -u <username> -p <password> -c <batch_file_path>
```

NOTE:

If you get an error message when executing this command, add the `-h` option (`.\PsExec.exe -h`).

Secure Setup of WSMan

Intersight Workload Optimizer provides a secure option for Hyper-V/VMM Targets which requires that WSMan be set up securely. Use PowerShell to generate a self-signed certificate, and create an HTTPS WinRM listener.

NOTE:

For clustered Hyper-V targets, you do not need to create a listener on each host. Only create a listener on the host that is being added to the "Address" field in the Target Configuration.

To enable secure WSMan on your Hyper-V host:

1. Generate a self-signed certificate using the following command:

```
New-SelfSignedCertificate -CertstoreLocation Cert:\LocalMachine\My -DnsName "myhost.example.org"
```

2. Find the thumbprint for the certificate for the host:

```
Get-childItem cert:\LocalMachine\My
```

3. Create an HTTPS WinRM listener for the host with the thumbprint you've found:

```
winrm create winrm/config/Listener?Address=*&Transport=HTTPS '@{Hostname="myhost.example.org"; CertificateThumbprint="THUMBPRINT_YOU_FOUND"}'
```

4. Verify the presence of configured listeners:

```
Get-WSManInstance -ResourceURI winrm/config/listener -Enumerate
```

vCenter Server

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

VMware vCenter Server provides a centralized management platform for VMware hypervisors. To manage your VMware environment with Intersight Workload Optimizer, you specify a vCenter Server instance as a target. Intersight Workload Optimizer discovers the infrastructure that target manages, and links it into a supply chain to deliver application performance management.

Prerequisites

- [Create a user account in vCenter. \(on page 180\)](#)

Intersight Workload Optimizer uses this user account to connect to your vCenter Server and execute actions.

General Considerations

Before you configure a vCenter Server target, you should consider the following:

- **Linked vCenters**

For linked vCenters, you must add each vCenter Server separately so Intersight Workload Optimizer can communicate with each vCenter Server through a separate API endpoint.

- **Restricting Intersight Workload Optimizer Access to Specific Clusters**

When you add a vCenter Server target, Intersight Workload Optimizer discovers all of the connected entities that are visible, based on the target account that it uses to connect to the vCenter Server target. If you have clusters or other entities you want to exclude from discovery, you can use the vSphere management client to the role of the Intersight Workload Optimizer account to `No access` for the given entities.

- **Shared Datastores**

If you add more than one vCenter Server target that manages the same datastore, you can enable or disable datastore browsing to discover wasted files on the shared datastore:

- Enable datastore browsing:

To properly enable browsing, you must turn on the **Enable Datastore Browsing** option in the target configuration for each vCenter Server target that manages the shared datastore.

- Disable datastore browsing:

If you don't want datastore browsing over shared datastores, you must turn *off* the **Enable Datastore Browsing** option in the target configuration for each vCenter Server target that manages the shared datastore.

If set **Enable Datastore Browsing** differently for separate targets that manage the same datastore, datastore browsing can give inconsistent results for active and wasted files.

■ **VSAN Permissions**

In order to enable VSAN support and discover groups based on storage profiles, you must ensure that the user role Intersight Workload Optimizer is assigned has the `Profile-driven storage view` permission enabled. This permission is *disabled* in the built-in `readonly` role.

Claiming vCenter Server Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Hypervisors**.
3. Select **vCenter**.
4. Configure the following settings:

■ Address

Specify the fully qualified domain name or IP address of the vCenter Server.

■ Username

Specify the username of the account Intersight Workload Optimizer uses to connect to the target.

Include the domain if required (`<domain>\<username>`).

■ Password

Specify the password of the account Intersight Workload Optimizer uses to connect to the target.

NOTE:

The password cannot contain any of the following special characters: \ / " [] : | < > + = ; , ? * , ' tab space @

■ Enable Datastore Browsing

If this option is selected, Intersight Workload Optimizer discovers unused storage.

■ Enable Guest Metrics

If this option is selected, Intersight Workload Optimizer collects advanced guest memory metrics that increase the accuracy of the VMEM data that Intersight Workload Optimizer uses for analysis of virtual machines. Guest metrics are only supported on vCenter Server version 6.5U3 or later. Guest VMs must run VMware Tools 10.3.2 Build 10338 or later.

To enable guest metrics, ensure the following:

- VMware Tools is installed and running on the target VMs.
- The user account has the `Performance.Modify Intervals` performance privilege.

For more information, see [vCenter Performance Privileges](#).

vCenter Server Imported Settings

In addition to discovering entities managed by the hypervisor, Intersight Workload Optimizer also imports a wide range of vSphere settings, such as Host DRS rules, annotations, Resource Pools, and DRS HA settings (See [Other Information Imported From vCenter \(on page 182\)](#)).

NOTE:

Intersight Workload Optimizer does not import Storage DRS rules at this time.

VMware vSphere 6.0 introduced the ability to move VMs between vCenters. If you enabled this feature in your VMware environment, you can configure Intersight Workload Optimizer to include cross vCenter Server vMotions in its recommendations. You must create a Workload Placement Policy that merges the data centers on the different vCenters, and then another policy to merge the given clusters.

For vCenter environments, you can create placement policies that merge data centers to support cross-vCenter moves. In this case, where a data center corresponds to a vCenter target, the merged clusters can be in different data centers. In this case, you must create two merge policies; one to merge the affected data centers, and another to merge the specific clusters.

To create a Merge policy:

1. In the Policy Management tab, select **Placement Policy**.
2. For **Type**, select **Merge**.
3. For **Merge**, choose the merge type, and click **Select**.
4. Choose the specific data centers or clusters to merge in this policy, then click **Select**.
5. Click **Save Policy**.

NOTE:

Since Intersight Workload Optimizer can only execute vMotions between clusters that use the same switch type (VSS or VDS), make sure any clusters you merge use the same switch type. Although Intersight Workload Optimizer will not initiate VSS → VDS vMotions, vSphere may do so. If this happens, Intersight Workload Optimizer displays a compliance violation notification.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Virtual Machine

- Energy
Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).
- Carbon Footprint
Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Virtual Storage
Virtual storage is the measurement of virtual storage capacity that is in use.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- Latency
Latency is the measurement of storage latency.

■ Host

- Energy
Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).
- Power
Power is the measurement of electricity consumed by a given entity, expressed in watts.
- Carbon Footprint
Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.
- Memory (Mem)
Memory is the measurement of memory that is reserved or in use.
- CPU
CPU is the measurement of CPU that is reserved or in use.
- IO
IO is the utilization of a host's IO adapters.
- Net

Net is the utilization of data through the host's network adapters.

- Swap

Swap is the measurement of a host's swap space that is in use.

- Balloon

Balloon is the measurement of memory that is shared by VMs running on a host.

This commodity applies to ESX only.

- CPU Ready

CPU Ready is the measurement of a host's ready queue capacity that is in use.

This commodity applies to ESX only.

- **Storage**

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

- Unattached Files

Unattached files are files that are not currently connected to a virtual machine.

- **Datacenter**

NOTE:

For datacenter entities, Intersight Workload Optimizer does not monitor resources directly from the datacenter, but from the hosts in the datacenter. See host monitored resources for details.

- **Provider Virtual Datacenter**

- Memory (Mem)

Memory is the measurement of memory that is reserved or in use.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage

Storage is the utilization of the storage attached to the entity.

- **Consumer Virtual Datacenter**

- Memory (Mem)

Memory is the measurement of memory that is reserved or in use.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage

Storage is the utilization of the storage attached to the entity.

Actions

In order to execute cross-vCenter migrations as a non-admin user, you must have the following permissions enabled for the user account in both the current and destination vCenter servers:

| Entity | Permissions |
|-----------------|---|
| Virtual Machine | Edit Inventory, Create From Existing (Move, Register, Remove, Unregister sub-options), Create New |
| Datacenter | Reconfigure Datacenter |
| Network | Assign Network |

Intersight Workload Optimizer supports the following actions:

■ **Virtual Machine**

- Start
- Move
- Move VM Storage
- Suspend
- Resize Up/Down
- Terminate

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

- Reconfigure

This action can only be executed outside Intersight Workload Optimizer.

■ **Host**

- Start
- Suspend
- Terminate

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

■ **Storage**

- Delete unattached file

Any file that has not been modified in the period defined in the storage policy. The default behavior is to generate an action after 15 consecutive days of non-use.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

Creating A Service User Account In vCenter

The service account you use must have specific permissions on the vCenter. The easiest way to grant Intersight Workload Optimizer the access it requires is to grant full administrator rights.

Some enterprises require that the service account does not grant full administrator rights. In that case, you can create a restricted service account that grants the following permissions to enable the required Intersight Workload Optimizer activities. For any environment implementing merge policies, these permissions need to be enabled at the vCenter level. The list of permissions is defined below.

vCenter Permissions

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| Monitoring | <ul style="list-style-type: none"> ■ Read-only role for all entity types |

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|---|
| | <p>Assign either Global permissions or permissions for the given vCenter Server instance to the target user or user group.</p> <ul style="list-style-type: none"> ■ Requirement to monitor VSAN and storage profiles <p>In order to enable VSAN support and discover groups based on storage profiles, you must enable the Profile-driven storage view permission. This permission is <i>disabled</i> in the built-in readonly role.</p> |
| Recommend Actions | <ul style="list-style-type: none"> ■ Read-only role for all entity types <p>Assign either Global permissions or permissions for the given vCenter Server instance to the target user or user group.</p> |
| Wasted Storage Reporting | <ul style="list-style-type: none"> ■ Datastore > Browse Datastore |
| Execute Delete Files | <ul style="list-style-type: none"> ■ Datastore > Low level file operations |
| Execute VM Move | <ul style="list-style-type: none"> ■ Resources > Assign VM to Resource Pool ■ Resources > Migrate Powered Off VMs ■ Resources > Migrate Powered On VMs ■ Resources > Modify Resource Pool ■ Resources > Query Vmotion |
| Execute VM Storage Move | <ul style="list-style-type: none"> ■ Datastore > Allocate Space ■ Datastore > Browse Datastore ■ Resources > Assign VM to Resource Pool ■ Resources > Migrate ■ Resources > Modify Resource Pool ■ Resources > Move Resource Pool ■ Resources > Query VMotion ■ Virtual Machine > Change Configuration > Change resource ■ Virtual Machine > Change Configuration > Change Swapfile placement |
| Execute VM Resize | <ul style="list-style-type: none"> ■ Virtual Machine > Change Configuration > Change CPU count ■ Virtual Machine > Change Configuration > Change Memory ■ Virtual Machine > Change Configuration > Change resource ■ Virtual Machine > Interaction > Reset ■ Virtual Machine > Interaction > Power Off ■ Virtual Machine > Interaction > Power On |
| Discover Tags | <ul style="list-style-type: none"> ■ Global > Global tag <p>You must also open ports 10443 and 7443 on the target server</p> |
| Assign Network | <ul style="list-style-type: none"> ■ Network > Assign |

Other Information Imported from vCenter

In addition to discovering entities managed by the vSphere hypervisors and their resources, Intersight Workload Optimizer:

- Imports any vSphere Host DRS rules when DRS is enabled, and displays them on the **Policy > Workload Placement** view under **Imported Placement Policies**. Imported rules are enabled by default, but you can disable them in Intersight Workload Optimizer.

NOTE:

In vCenter environments, Intersight Workload Optimizer does not import DRS rules if DRS is disabled on the hypervisor. Further, if Intersight Workload Optimizer did import an enabled DRS rule and somebody subsequently disables that DRS rule, then Intersight Workload Optimizer will discover that the rule was disabled and will remove the imported placement policy.

- Imports any custom annotations and displays related groupings in the **Inventory > Groups** tree view, under **VC Annotations**. The service account must enable the **Global > Global tag** privilege.
- For vCenter Server versions 5.5 and later, discovers Virtual Machine Storage Profiles and displays them as groups anywhere that you can set scope. The groups appear under **VC Storage Profiles**. You can use these discovered storage profiles the same as any other groups – For example, to scope dashboards, or to set the scope for specific action policies.
- Discovers resource pools and displays them as folders in the Inventory tree and as components in the Supply Chain Navigator. If you have the Cloud Control Module license, Intersight Workload Optimizer manages resource pools as Virtual Datacenters (VDCs) and can recommend resize actions. Root resource pools appear as Provider VDCs in the supply chain, whereas child resource pools appear as Consumer VDCs.
- Imports vSphere HA cluster settings and translates them into CPU and memory utilization constraints. These are displayed as cluster-level overrides under **Folders** on the **Policy > Analysis > Host** view.

Orchestrator Targets

Intersight Workload Optimizer supports the ServiceNow orchestrator target.

With orchestrator targets you can integrate Intersight Workload Optimizer actions with the orchestrator's application management process. For example, you can pass Intersight Workload Optimizer to a Change Request system for approval, and the system can pass the action back to Intersight Workload Optimizer for execution.

About Orchestrators

Orchestration targets assign workflows that execute multiple actions to make changes in your environment. Intersight Workload Optimizer discovers workflows that you have defined on the orchestrator. You can then set up an automation policy that maps workflows to actions. If the action acceptance mode is *Manual* or *Automated*, then when Intersight Workload Optimizer recommends the action, it will direct the orchestrator to use the mapped workflow to execute it. For example, you can configure policies that log Intersight Workload Optimizer actions in your ServiceNow instance, and that submit actions for approval in ServiceNow workflows.

ServiceNow

You can configure Intersight Workload Optimizer policies that log Intersight Workload Optimizer actions in your ServiceNow instance, and that submit actions for approval in ServiceNow workflows.

NOTE:

When creating the action orchestration policy as explained in the section above, the scope of the policy must match the scope of the ServiceNow target.

Prerequisites

- A ServiceNow user with the `web_service_admin` role and the custom role `x_turbo_turbonomic.user` that is created during installation that can communicate with Intersight Workload Optimizer via the REST API.

Adding ServiceNow Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Orchestrator**.
3. Select **ServiceNow**.
4. Configure the following settings:
 - **Address**
Specify the hostname of the ServiceNow instance without the `http` or `https` protocols. For example, `dev-env-266.service-now.com`.
 - **Username**
Specify the username for the account Intersight Workload Optimizer uses to connect to the ServiceNow instance.
 - **Password**
Specify the password for the account Intersight Workload Optimizer uses to connect to the ServiceNow instance.
 - **Client ID**
Specify the client ID Intersight Workload Optimizer uses if `Use OAuth` is checked.
 - **Client Secret**
Specify the password Intersight Workload Optimizer uses if `Use OAuth` is checked.
 - **Port**
Specify the port used to access the ServiceNow instance.
 - **Use OAuth**
Use OAuth to communicate with the ServiceNow target.
Currently, Intersight Workload Optimizer supports OAuth 2.0 when this protocol is enabled in the ServiceNow target.

ServiceNow Integration

In order to complete target addition, see the [Intersight Workload Optimizer Actions for ServiceNow](#) documentation.

Storage Targets

Adding a storage Target enables Intersight Workload Optimizer to connect to your storage subsystem through a native or SMI-S provider API. Intersight Workload Optimizer uses the target's API to access and collect information from each of the underlying disk arrays. The information is used to set disk performance characteristics according to the type and capacity of storage, leading to improved workload placement.

Similarly, Intersight Workload Optimizer determines the relationships between storage controllers and disk arrays, and the location of datastores within those arrays. This information also helps optimize workload placement at a more granular level.

For on-prem applications, this optimization will enable Intersight Workload Optimizer to make more informed decisions about which storage devices the workloads hosting your applications run on, and assist in assuring application SLO. In the cloud, storage data is handled as part of the public cloud target.

Both virtual machines and containers benefit from this level of optimization. In the case of short-lived containers, Intersight Workload Optimizer will suggest the best datastore to hold persistent data, and paired with a container or hypervisor target, will select the optimal match of compute and storage resources. For longer-lived containers and virtual machines, each workload will be continually assessed for SLA/SLO, and recommendations to move or resize storages will ensure the continued efficiency of your environment.

The following section describes the storage supply chain. For information on how to add specific storage targets, the resources Intersight Workload Optimizer can monitor for the various supply chain entities, and the actions it can take to optimize the environment, refer to the target configuration instructions for your specific storage type.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| EMC VMAX | EMC XtremIO | HPE 3Par | NetApp | Nutanix | Pure | Intersight Workload Optimizer |
|------------------------------|-----------------|----------------|--------------------|---------------|-------------|-------------------------------|
| Volume (Regular, Thin, Meta) | Volume | Virtual Volume | Volume | Container | Volume | Storage |
| Disk Group or Thin Pool | XTremIO Cluster | CPG | Aggregate | Storage Pool | Shelf Array | Disk Array |
| VMAX Array | XTremIO Cluster | Controller | Controller / Filer | Controller VM | Controller | Storage Controller |

Storage targets (storage controllers) add Storage Controller and Disk Array entities to the supply chain. Disk Array entities in turn host Storage entities (datastores).

Dell EMC SC Series

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports the management of Dell SC Series (Compellent) disk arrays and storage controllers. Intersight Workload Optimizer connects through the Dell Enterprise Manager and performs management as a client of the Enterprise Manager Data Collector.

The Dell Enterprise Manager is a management service that provides administration, management, and monitoring of multiple Storage Centers – Typically installed on a Windows VM.

When you specify a Dell Compellent target, you provide the IP address of the Dell Enterprise Manager. Intersight Workload Optimizer discovers the Compellent infrastructure through the SMI-S component which is typically installed as part of the Enterprise Manager.

NOTE: Before adding the Dell Compellent target to Intersight Workload Optimizer, confirm that the Storage Centers you want to manage show up in Dell Enterprise Manager (see “Storage Center Administration” in the *Dell Compellent Enterprise Manager Administrator’s Guide*). The SMI-S user account must be able to access all of the Storage Centers. If you add or remove Storage Centers later, Intersight Workload Optimizer will detect the changes during its next discovery cycle.

Prerequisites

- Dell Enterprise Manager Data Collector Service 6.2 or higher
- Dell Compellent SMI-S Provider
- Storage Centers added to Dell Enterprise Manager

Setting Up the Dell Compellent SMI-S Provider

Your Dell Compellent storage environment must include an enabled Dell Compellent SMI-S Provider. Configure the SMI-S Provider as described in the “SMI-S” section of the *Dell Storage Manager Administrator’s Guide*. The guide provides detailed steps to:

- Open the required ports on the server hosting the Enterprise Manager Data Collector.
- Enable SMI-S for the Data Collector.

- Add a user for SMI-S.
- If using HTTPS, associate the SSL certificate with the SMI-S Provider.

Adding Dell Compellent Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **Dell Compellent**.
4. Configure the following settings:
 - **Address**
The name or IP address of the Dell Enterprise Manager.
By default, Enterprise Manager provides SMI-S data over port 5988 (HTTP). If your installation uses a different HTTP port for SMI-S, include the port number in the Address field. For HTTPS, do not include the port number. Instead, select the `Use Secure Connection` check box.
 - **Username/Password**
Credentials for the SMI-S user you added when setting up the SMI-S provider.
 - **Use Secure connection**
Select this option to connect to the target using a secure connection (HTTPS). Do not enter a port in the Address field if this option is selected.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|----------------|-------------------------------|
| Dell | Intersight Workload Optimizer |
| Storage Center | Storage Controller |
| Storage Type | Disk Array |
| Volume | Storage |

Storage targets (storage controllers) add Storage Controller and Disk Array entities to the supply chain. Disk Array entities then host Storage entities (datastores).

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
 - Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
NOTE:
When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.
 - Latency
Latency is the measurement of storage latency.
- **Disk Array**
 - Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ Storage Controller

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

■ Storage

- Provision
- Resize Up
- Move

This action can only be executed outside Intersight Workload Optimizer.

■ Disk Array

- Provision
- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

This action can only be executed outside Intersight Workload Optimizer.

■ Storage Controller

- Provision

This action can only be executed outside Intersight Workload Optimizer.

EMC VMAX

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports management of VMAX2 and 3 Series storage arrays. The VMAX series is a family of enterprise storage arrays designed for SAN environments. Intersight Workload Optimizer connects to VMAX storage systems via an EMC SMI-S provider that has the disk arrays added to it. A single SMI-S provider can communicate with one or more disk arrays. When you specify an SMI-S provider as a target, Intersight Workload Optimizer discovers all the added disk arrays.

NOTE:

Intersight Workload Optimizer does not utilize Unisphere. Data is collected exclusively from the SMI-S provider.

Intersight Workload Optimizer will create Storage Groups based on the SLO levels defined in VMAX3 Targets. By default, Storage vMotion actions will respect these SLO levels based on the configured response time.

Prerequisites

- EMC SMI-S Provider V8.x
- A service account that Intersight Workload Optimizer can use to connect to the EMC SMI-S Provider (typically the default `admin` account)

Claiming VMAX Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **VMAX**.
4. Configure the following settings:
 - Address
The IP or hostname of the SMI-S provider. If the provider address begins with `https`, you must follow the IP with the port used to connect.
 - Use Secure Connection
If checked, port 5989 will be used to connect. If unchecked, port 5988 will be used.
 - Username
The Username for the SMI-S provider.
 - Password
The Password for the SMI-S provider.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|------------------------------|-------------------------------|
| EMC VMAX | Intersight Workload Optimizer |
| Volume (Regular, Thin, Meta) | Storage |

| | |
|--|-------------------------------|
| EMC VMAX | Intersight Workload Optimizer |
| Storage Resource Pool (VMAX3) / Thick Provisioned Pool (earlier) | Disk Array |
| Storage Group (VMAX3) / Thin Provisioned Pool (earlier) | Logical Pool |
| VMAX Array | Storage Controller |

Monitored Resources

When calculating available storage, Intersight Workload Optimizer excludes disks devoted to the VMAX operating system by default. If these disks are assigned to new RAID groups or storage pools, the capacity of those disks will then be considered when calculating the capacity of the Storage Controller.

Intersight Workload Optimizer monitors the following resources:

■ Storage

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

■ Logical Pool

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ Disk Array

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ Storage Controller

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

Actions

Intersight Workload Optimizer supports the following actions:

- **Storage**

- Provision (Clone)
- Delete
- Move
- Resize (V-Volumes only)

This action can only be executed outside Intersight Workload Optimizer.

- **Logical Pool**

- Resize

This action can only be executed outside Intersight Workload Optimizer.

EMC XtremIO

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

EMC® XtremIO® is a flash-based (SSD) storage solution, designed to push data to applications at higher speeds. The system building blocks are SAN appliances called X-Bricks. A deployment is organized into clusters of X-Bricks, and the clusters are managed by the XtremIO Management Server (XMS).

Intersight Workload Optimizer connects to X-Bricks through the XMS. The XMS presents a unified view of each connected X-Brick cluster, rather than exposing the individual X-Bricks within each cluster. Within Intersight Workload Optimizer, each X-Brick cluster displays as a single storage controller with an associated disk array.

The relationship between Storage entities and individual X-Bricks within the cluster is not exposed through the XMS – Intersight Workload Optimizer cannot make recommendations to move datastores from one X-Brick to another. Additionally, the X-Brick has a fixed form factor – Intersight Workload Optimizer does not recommend resize actions for disk array or storage controller resources.

Intersight Workload Optimizer recognizes XtremIO arrays as flash storage and sets the IOPS capacity on discovered arrays accordingly.

Prerequisites

- A service user account on XMS 4.0 or higher – typically the default `xmsadmin` account
- Intersight Workload Optimizer uses this account to connect to the XMS and execute commands through the XtremIO API.

Claiming XtremIO Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **EMC XtremIO**.

4. Configure the following settings:
 - Address
The name or IP address of the XtremIO Management Server (XMS).
 - Username/Password
Credentials for a user account on the XMS.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|-----------------|-------------------------------|
| XtremIO | Intersight Workload Optimizer |
| Volume | Storage |
| XtremIO Cluster | Disk Array |
| XtremIO Cluster | Storage Controller |

Storage targets (storage controllers) add Storage Controller and Disk Array entities to the supply chain. Disk Array entities then host Storage entities (datastores).

Monitored Resources

When calculating available storage, Intersight Workload Optimizer excludes disks devoted to the VNX operating system.

Intersight Workload Optimizer monitors the following resources:

- **Storage**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
 - Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:
When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

 - Latency
Latency is the measurement of storage latency.
- **Disk Array**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
 - Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
 - Latency
Latency is the measurement of storage latency.
- **Storage Controller**

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU
CPU is the measurement of CPU that is reserved or in use.
- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

- **Storage**
 - Provision
This action can only be executed outside Intersight Workload Optimizer.
 - Resize Up
This action can only be executed outside Intersight Workload Optimizer.
- **Storage Controller**
 - Provision
This action can only be executed outside Intersight Workload Optimizer.

EMC ScaleIO

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

EMC ScaleIO is an example of Software-Defined Storage for the datacenter. It creates a Virtual SAN overlaying commodity infrastructure that consists of multiple LAN-connected Servers with locally attached commodity Storage. It presents a standard Block Storage interface to Applications accessing the Virtual SAN.

Intersight Workload Optimizer communicates with the EMC ScaleIO system via the REST API Gateway.

Prerequisites

- EMC ScaleIO 2.x or 3.x
- A service account that Intersight Workload Optimizer can use to connect to the ScaleIO Gateway.

Claiming EMC ScaleIO Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **EMC ScaleIO**.

4. Configure the following settings:
 - Address
The IP or hostname of the Gateway.
 - Username
The Username for the Gateway service account.
 - Password
The Password for the Gateway service account.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|-------------------|-------------------------------|
| EMC ScaleIO | Intersight Workload Optimizer |
| Volume | Storage |
| Storage Pool | Disk Array |
| Protection Domain | Storage Controller |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

■ Storage

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
- Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

■ Disk Array

- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
- Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency
Latency is the measurement of storage latency.

■ Storage Controller

- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.

Actions

Intersight Workload Optimizer supports the following actions:

- **Storage**
 - Provision (Clone)
 - Resize (disabled by default)

This action can only be executed outside Intersight Workload Optimizer.
- **Disk Array**
 - Resize Disk Array

This action can only be executed outside Intersight Workload Optimizer.
- **Protection Domain**
 - Provision (Clone)

This action can only be executed outside Intersight Workload Optimizer.

EMC VPLEX

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports management of EMC VPLEX virtual storage systems in a local configuration, via the VPLEX API. Currently, Intersight Workload Optimizer does not support Metro or Geo configurations.

VPLEX is used to aggregate and refine data collected between connected Storage and Hypervisor targets. VPLEX supports one-to-one, one-to-many, and many-to-one relationships between virtual volumes and LUNs. Only one-to-one mapping between virtual volume and LUNs is supported by Intersight Workload Optimizer.

Prerequisites

- VPLEX Management Server
- Hypervisor target supported by Intersight Workload Optimizer
- Storage target supported by Intersight Workload Optimizer

NOTE:

In order for Intersight Workload Optimizer to make use of the information provided by VPLEX, you must also add the hypervisor and storage layered under it as targets.

VPLEX Permissions

| Intersight Workload Optimizer Functionality | Required Permissions |
|---|----------------------|
| Monitoring | Service Account |
| Action Execution | Admin account |

Claiming EMC VPLEX Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **EMC VPLEX**.
4. Configure the following settings:
 - **Address:**
The IP or Hostname of the VPLEX Management Server
 - **Username:**
The Username for the VPLEX Management Server
 - **Password:**
The Password for the VPLEX Management Server
 - **Port Number:**
The port number for the remote management connection. The default port number for the VPLEX Management server is 443
 - **Secure Connection:**
Select this option to use a secure connection (HTTPS)

NOTE:

The default port (443) uses a secure connection.

Actions

For this target, actions are generated and executed via the underlying storage targets. Intersight Workload Optimizer will use the enhanced visibility provided by VPLEX to make more accurate storage decisions, such as recommending storage vMotion between pools.

HPE 3PAR

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

HPE 3PAR StoreServ systems use controller nodes to manage pools of storage resources and present a single storage system to consumers. Intersight Workload Optimizer communicates with the HPE 3PAR system via both the WSAPI and SMI-S providers that are installed on the 3PAR controller node.

Prerequisites

- SMI-S Provider enabled and configured on the controller node.
- WSAPI Provider enabled and configured on the controller node.
- A service account on the controller node that Intersight Workload Optimizer can use to connect to the SMI-S and WSPAI providers.

NOTE:

For discovery and monitoring, the Intersight Workload Optimizer service account must have the `Browse` permission on all monitored domains. To exclude domains from monitoring, the Intersight Workload Optimizer service account must have no permissions on those domains. For action execution, Intersight Workload Optimizer requires the `Edit` permission.

Adding HPE 3PAR Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **HPE 3PAR**.
4. Configure the following settings:
 - **Address**
Specify the name or IP address of the 3PAR controller node.
By default, the controller provides SMI-S data over port 5988 for HTTP or port 5989 for HTTPS. If your installation uses a different port for SMI-S, include the Port in the Host name or IP address field.
 - **Username/Password**
Specify the credentials for a user account on the controller node.
 - **Use Secure Connection**
Select this option to connect to the target by using SSL.
 - **Web services API port**
For non-default configurations, specify the WSAPI port as defined in the HPE 3PAR Management Console.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| | |
|------------------|-------------------------------|
| HPE 3PAR | Intersight Workload Optimizer |
| Virtual Volume | Storage |
| CPG | Disk Array |
| AO Configuration | Logical Pool |
| Controller | Storage Controller |

Storage targets (storage controllers) add Storage Controller, Logical Pool and Disk Array entities to the supply chain. Logical Pool and Disk Array entities then host Storage entities (datastores).

Setting Up the SMI-S Provider

The HPE 3PAR SMI-S Provider should be installed on the controller node. It is disabled by default – you must ensure that it is installed properly and running on the controller node.

To enable the SMI-S provider:

1. Log into the HPE 3PAR Command Line Interface (CLI).
Open a secure shell session (ssh) on the controller node. Default credentials are `3paradm/3pardata`.
2. Check the current status of the SMI-S provider.
In the shell session, run the command:

`showcim`
3. If the CIM service is not running, start it.
To enable the CIM service and the SMI-S provider, run the command:

```
startcim
```

To stop the SMI-S provider, execute the command `stopcim -f -x`.

Setting Up the WSAPI Provider

The HPE 3PAR WSAPI Provider should be installed on the controller node.

To enable the WSAPI provider:

1. Log into the HPE 3PAR Command Line Interface (CLI).
Open a secure shell session (ssh) on the controller node. Default credentials are `3paradm/3pardata`.
2. Check the current status of the WSAPI provider.

In the shell session, run the command:

```
showsapi
```

3. If the WSAPI service is not running, start it by running the command:

```
startwsapi
```

To allow only insecure connections run the command:

```
set wsapi -http enable
```

Or to allow only secure connections, run the command:

```
set wsapi -https enable
```

To stop the WSAPI provider, execute the command `stopwsapi -f`.

3Par Adaptive Optimization

Adaptive Optimization (AO) for HPE 3Par enables management of data storage across two or three tiers. AO places storage regions on the appropriate tier in response to periodic analysis that AO performs.

To work with the storage in an AO group, Intersight Workload Optimizer:

- Discovers each Common Provisioning Group (CPG) as a disk array
In the Intersight Workload Optimizer user interface, these disk arrays do not host storage – They appear empty. Intersight Workload Optimizer will not recommend storage moves between these disk arrays, because such moves would conflict with AO block-level placement.
- Creates a single logical pool that hosts all the datastores in an AO group
This logical pool represents the AO group, and it includes all the member CPGs. Intersight Workload Optimizer considers this single logical pool when it performs analysis – It can recommend moving storage into or out of the AO group. Also, Intersight Workload Optimizer aggregates resource capacity in this logical pool. For example, the IOPS capacity for the AO logical pool is a combination of IOPS capacity for the constituent CPGs.

You can see the AO logical pool in the Intersight Workload Optimizer user interface. The display name for this logical pool is the name of the AO Configuration.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

■ **Disk Array**

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ **Logical Pool**

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ **Storage Controller**

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

■ **Storage**

- Provision
- Resize Up/Down

■ **Disk Array**

- Provision

- Resize Up/Down
- **Logical Pool**
 - Provision

This action can only be executed outside Intersight Workload Optimizer.
 - Resize Up/Down

This action can only be executed outside Intersight Workload Optimizer.
- **Storage Controller**
 - Provision

This action can only be executed outside Intersight Workload Optimizer.

NetApp

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

The Storage Control Module adds support for NetApp filers running the Data ONTAP operating system. NetApp storage controllers are Storage Virtual Machines that manage storage arrays. Intersight Workload Optimizer connects to these storage controllers to support NetApp targets in Cluster-Mode (C-Mode).

Prerequisites

- Transport Layer Security (TLS) is enabled.
- A service account Intersight Workload Optimizer can use to connect to the NetApp target.

Enabling TLS

Starting with version 5.4, by default Intersight Workload Optimizer requires Transport Layer Security (TLS) version 1.2 to establish secure communications with targets. NetApp filers have TLS disabled by default, and the latest version they support is TLSv1. If your NetApp target fails to validate on Intersight Workload Optimizer 5.4 or later, this is probably the cause.

If target validation fails because of TLS support, you might see validation errors with the following strings:

- No appropriate protocol

To correct this error, ensure that you have enabled the latest version of TLS that your target technology supports. If this does not resolve the issue, contact Cisco Technical Support.
- Certificates does not conform to algorithm constraints

To correct this error, refer to your NetApp documentation for instructions to generate a certification key with a length of 2048 or greater on your target server. If this does not resolve the issue, please contact Cisco Technical Support.

For information about enabling TLS, see the Data ONTAP **System Administration Guide** for sections on the SSL protocol.

Service User Account – Administrator Role

To discover and fully manage NetApp disk arrays, Intersight Workload Optimizer must have a service account that grants privileges to execute commands through the NetApp filer's OnTap API (ontapi). In most cases, you can create the administrator account via the NetApp OnCommand System Manager, or from the NetApp command line – For example:

```
security login create -role admin -username Cisco -application ontapi -authmethod password
```

If you prefer not to grant full administrator rights, see [Creating Restricted Service Accounts In NetApp \(on page 201\)](#)

Claiming NetApp Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **NetApp**.
4. Configure the following settings:
 - **Address**
Specify the name or IP address of the NetApp cluster management server.
 - **Username/Password**
Specify the credentials for the NetApp service user account that you have configured for Intersight Workload Optimizer to use.
 - **Secure Connection**
Select this option to connect to the target by using SSL.

After validating the new target, Intersight Workload Optimizer discovers the connected storage entities. This table compares terms used in NetApp to those used in Intersight Workload Optimizer:

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| NetApp | Intersight Workload Optimizer |
|--------------------|-------------------------------|
| Volume | Storage |
| Aggregate | Disk Array |
| Controller / Filer | Storage Controller |

Storage targets (storage controllers) add Storage Controller and Disk Array entities to the supply chain. Disk Array entities then host Storage entities (datastores).

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

In NetApp environments, the storage controller shows 100% utilization when there are no more disks in a `SPARE` state that the storage controller can utilize in an aggregate. This does not indicate that the storage controller has no capacity.

■ Storage

- **Storage Amount**
Storage Amount is the measurement of storage capacity that is in use.
- **Storage Provisioned**
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:

When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.

- Latency

Latency is the measurement of storage latency.

- **Disk Array**

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

- **Storage Controller**

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

- **Storage**

- Move
- Provision

This action can only be executed outside Intersight Workload Optimizer.

- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

- **Disk Array**

- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

- Move

This action can only be executed outside Intersight Workload Optimizer.

- Provision

This action can only be executed outside Intersight Workload Optimizer.

- **Storage Controller**

- Provision

This action can only be executed outside Intersight Workload Optimizer.

Note that Intersight Workload Optimizer can automate moving a datastore to a disk array on the same storage controller, as well as moves to a disk array on a different storage controller.

Restricted Service Accounts In NetApp

While Intersight Workload Optimizer prefers a NetApp service account with administrator rights, it is possible to create an account that has limited access.

NetApp 9.x Restricted Service Account Setup

Complete the following steps to use a service account that does not have full administrator rights.

1. Log into the NetApp filer from a command shell.
2. Create a role and assign it permission to execute each of the following commands:

For example:

```
security login role create -role RoleName -cmddirname "storage aggregate show" -vserver Cluster-Name
```

The required capabilities are listed here:

- cluster identity modify
- cluster identity show
- lun create
- lun igroup create
- lun igroup modify
- lun igroup show
- lun mapping create
- lun mapping delete
- lun mapping show
- lun modify
- lun show
- network interface create
- network interface delete
- network interface modify
- network interface show
- statistics show
- storage aggregate create
- storage aggregate modify
- storage aggregate show

- `storage disk show`
 - `system controller flash-cache show`
 - `system node modify`
 - `system node show`
 - `version`
 - `volume create`
 - `volume modify`
 - `volume move modify`
 - `volume move show`
 - `volume move start`
 - `volume qtree create`
 - `volume qtree show`
 - `volume show`
 - `volume snapshot create`
 - `volume snapshot modify`
 - `volume snapshot show`
 - `vserver create`
 - `vserver fcp nodename`
 - `vserver iscsi nodename`
 - `vserver modify`
 - `vserver options`
 - `vserver show`
3. For execution privileges, run the following commands for the given role, where `Role-Name` is the name of the role you are creating, and `Cluster-Name` identifies the cluster you want the role to affect. You must execute these commands individually to set privileges that affect each individual cluster:
- `security login role create -role Role-Name -access all -cmddirname "volume offline" -vserver Cluster-Name`
 - `security login role create -role Role-Name -access all -cmddirname "volume unmount" -vserver Cluster-Name`

- `security login role create -role Role-Name -access all -cmddirname "volume move" -vserver Cluster-Name`
- `security login role create -role Role-Name -access all -cmddirname "volume delete" -vserver Cluster-Name`

4. Create a user that will use the newly-created role.

For example:

```
security login create -User-Name RoleUser -r Intersight Workload OptimizerRole
```

5. Enter a password for the new user when prompted.

6. Give the user access to the `ssh` and `ontapi` applications by using the following commands, replacing `Role-Name` and `RoleUser` with the role and user you created:

```
security login create -role Role-Name -username RoleUser -application ontapi -authmethod password
```

```
security login create -role Role-Name -username RoleUser -application ssh -authmethod password
```

NetApp C-Mode Restricted Service Account Setup

Complete the following steps to use a service account that does not have full administrator rights.

1. Log into the NetApp filer from a command shell.
2. Create a role and assign it permission to run each of the following commands:

- `aggr-get-iter`
- `igroup-get-iter`
- `cluster-identity-get`
- `lun-map-get-iter`
- `net-interface-get-iter`
- `storage-disk-get-iter`
- `system-get-node-info-iter`
- `volume-get-iter`
- `vserver-get-iter`
- `fcp-node-get-name`
- `flash-device-get-iter`
- `iscsi-node-get-name`
- `options-list-info`
- `qtree-list-iter`

- `system-get-version`
- `lun-get-iter`
- `snapshot-get-iter`
- `perf-object-get-instances`
- `volume-get-iter`
- `volume-move-get-iter`
- `volume-move-start`

For example, to enable volume offline, run the following command:

```
security login role create -role Role-Name -access all -cmddirname "volume offline" -vserver <cluster_name>
```

3. Create a user based on the role you create.

Give the user access to the `ssh` and `ontapi` applications. For example:

```
security login create -role Role-Name -username User-Name -application ontapi -authmethod password
```

Pure Storage FlashArray

NOTE:

This target runs in on-prem datacenters. To establish communication between targets on the datacenter and Intersight Workload Optimizer, you must:

- Install an Intersight Assist appliance in the on-prem datacenter. The target service must be accessible to the Intersight Assist appliance.
- Connect the Intersight Assist instance with Cisco Intersight.
- Log in to Cisco Intersight and claim the Intersight Assist instance as a target.

Intersight Assist provides a secure way for connected targets to send information and receive control instructions from Intersight Workload Optimizer, using a secure internet connection. For more information, see the [Cisco Intersight Assist Getting Started Guide](#).

Intersight Workload Optimizer supports management of the following Pure Storage technologies:

- FlashArray//C
- FlashArray//X

The following technologies are not supported:

- FlashBlade

Because of the improved performance of Pure Storage arrays, Intersight Workload Optimizer intelligently moves more demanding workloads to Flash-based data stores. Intersight Workload Optimizer analysis is also able to incorporate Pure Storage de-duplication and compression when recommending actions.

Prerequisites

- A service account Intersight Workload Optimizer can use to connect to the FlashArray.

This account needs privileges to execute commands through the Pure Storage API – Typically the default `pureuser` administrative account.

Claiming Pure Storage Targets

1. Click **Settings > Target Configuration**.
2. Click **New Target > Storage**.
3. Select **Pure**.
4. Configure the following settings:
 - **Address**
Specify the name or IP address of the Pure Storage FlashArray.
 - **Username/Password**
Specify the credentials for the service account that Intersight Workload Optimizer can use to connect to the FlashArray. The username must not contain the domain. For example, `Username=jjsmith` is correct, while `Username=myDomain\jjsmith` results in a failure to validate.
 - **Secure connection**
When checked, uses SSL to connect to the Pure target.

Most Pure installations do not accept insecure connections. If you receive an error when adding the target with secure connections disabled, try re-adding with this option enabled.

Entity Mapping

After validating your targets, Intersight Workload Optimizer updates the supply chain with the entities that it discovered. The following table describes the entity mapping between the target and Intersight Workload Optimizer.

| Pure | Intersight Workload Optimizer |
|-------------|-------------------------------|
| Volume | Storage |
| Shelf Array | Disk Array |
| Controller | Storage Controller |

Storage targets (storage controllers) add Storage Controller and Disk Array entities to the supply chain. Disk Array entities then host Storage entities (datastores).

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
 - Storage Access (IOPS)
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:
When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.
 - Latency
Latency is the measurement of storage latency.
- **Disk Array**
 - Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
 - Storage Provisioned

Storage provisioned is the utilization of the entity's capacity, including overprovisioning.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- Latency

Latency is the measurement of storage latency.

■ **Storage Controller**

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, its chart in the user interface will not display data.

- CPU

CPU is the measurement of CPU that is reserved or in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

Actions

Intersight Workload Optimizer supports the following actions:

■ **Storage**

- Resize Up

This action can only be executed outside Intersight Workload Optimizer.

■ **Storage Controller**

- Provision

This action can only be executed outside Intersight Workload Optimizer.

■ **Disk Array**

- Provision

This action can only be executed outside Intersight Workload Optimizer.

Pure Storage assigns all the disks managed by a storage controller to a single array, with a fixed form-factor. There are no actions to perform for an array. For example, there is no action to move a disk array from one storage controller to another. Likewise, there are no actions to move or provision volumes because of the fixed form-factor.



User Interface Reference

Cisco Intersight Workload Optimizer is a solution that assures application performance for any workload running in any virtualized or cloud environment.

After installing or setting up the product, you can see the results of Intersight Workload Optimizer analysis - actions to perform that directly improve your datacenter state - within 15 to 30 minutes.

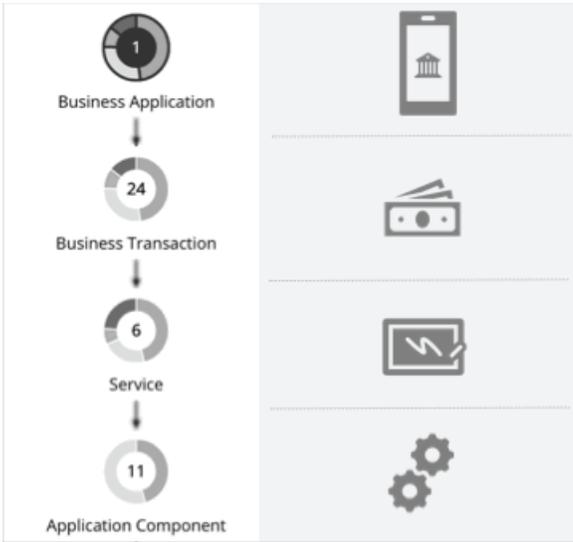
While it's true that setup is straightforward, the Intersight Workload Optimizer platform is rich in its coverage. Analysis takes your full stack into account, and Intersight Workload Optimizer recommends actions before alarms go off and the situation becomes critical. The user interface for such a product is necessarily rich with high-level information as well as fine details. You can use this interface and drill down to specific details that clarify the actions Intersight Workload Optimizer recommends.

Cisco also understands the need to define business rules in your environment. For example, you might need to ensure that certain applications have exclusive access to "golden" storage, while others can use less expensive resources. Or, you might want to ensure that certain workloads resize at scheduled times. Intersight Workload Optimizer automation supports this kind of business rule, as well as HA, affinity, discount purchase profiles, and many other rules that you expect to set up in a modern datacenter. The user interface includes tools to configure Intersight Workload Optimizer so that action recommendations respect the needs within your environment.

Entity Types - Applications

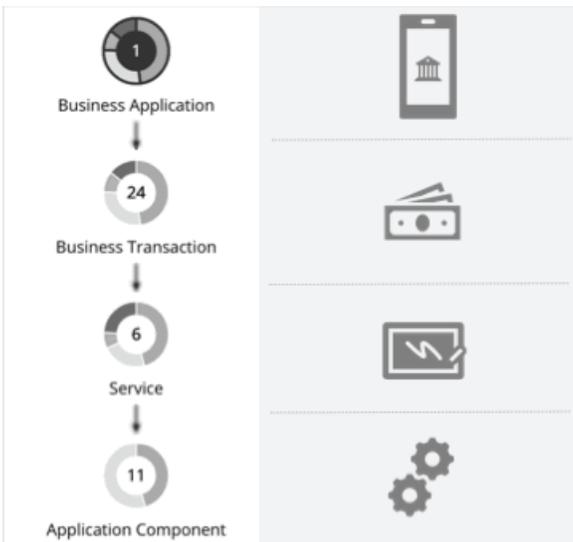
The supply chain strongly emphasizes our application-driven approach to managing your infrastructure. By showing the entity types that make up your applications at the top of the hierarchy, it is easier for you to see the health of your environment and evaluate actions from the perspective that matters - Application Performance.

For distributed applications the Intersight Workload Optimizer Supply Chain can include a Business Application entity, with underlying Business Transaction, Service, and Application Component entities. The application is ultimately hosted on one or more VMs or Containers.



Business Application

A Business Application is a logical grouping of Business Transactions, Services, Application Components, and other elements of the application model that work together to compose a complete application as end users would view it. For example, a mobile banking app is a Business Application with a *Business Transaction* that facilitates payments, a *Service* within the Business Transaction that records payment information, and underlying *Application Components* (such as JVMs) that enable the Service to perform its functions.



You can monitor overall performance, make resourcing decisions, and set policies in the context of your Business Applications.

Synopsis

| | |
|----------------|---|
| Synopsis | |
| Provides: | The complete application to end users |
| Consumes from: | Business Transactions, Services, Application Components, Database Servers, and the underlying nodes |

| Synopsis | |
|------------|---|
| Discovery: | <p>Intersight Workload Optimizer discovers the following:</p> <ul style="list-style-type: none"> ■ AppDynamics Business Applications ■ Dynatrace Applications <p>If you do not have these targets, you can create your own Business Applications using the Application Topology feature. For details, see Application Topology (on page 219).</p> |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

The **Response Time** and **Transaction** charts for a Business Application show average and peak/low values over time. You can gauge performance against the given SLOs. By default, Intersight Workload Optimizer estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Intersight Workload Optimizer does not recommend actions for a Business Application, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Application lists these actions, thus providing visibility into the risks that have a direct impact on the Business Application's performance.

Business Application Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

None

Intersight Workload Optimizer does not recommend actions for a Business Application, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Application lists these actions, thus providing visibility into the risks that have a direct impact on the Business Application's performance.

Transaction SLO

Enable this SLO if you are monitoring performance through your Business Applications.

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Response Time SLO

Enable this SLO if you are monitoring performance through your Business Applications.

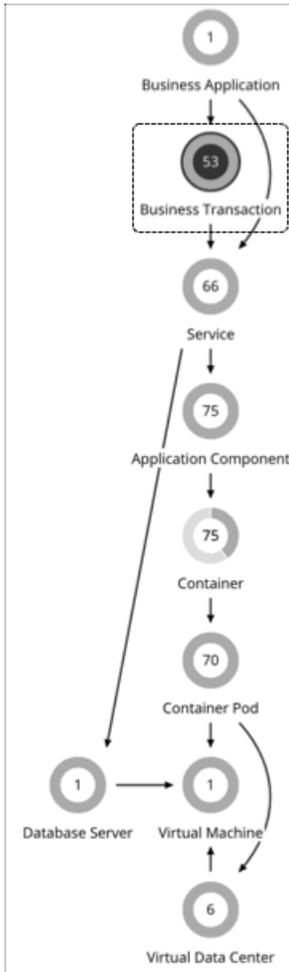
| Attribute | Default Setting/Value |
|--------------------------|---|
| Enable Response Time SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Response Time SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000 ms. You can change this to a different value. |

Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Business Transaction

A Business Transaction represents a capability within your Business Application that fulfills a response to a user-initiated request. Its performance directly impacts user experience. You can monitor performance as experienced by your end users in the context of Business Transactions.

Synopsis



| | |
|----------------|---|
| Synopsis | |
| Provides: | Response time and transactions to Business Applications |
| Consumes from: | Services (on page 213) , Application Components (on page 215) , Database Servers, and the underlying nodes |
| Discovery: | Intersight Workload Optimizer discovers the following: <ul style="list-style-type: none"> ■ AppDynamics Business Transactions ■ NewRelic Key Transactions If you do not have these targets, you can create your own Business Transactions using the Application Topology feature. For details, see Application Topology (on page 219) . |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

The **Response Time** and **Transaction** charts for a Business Transaction show average and peak/low values over time. You can gauge performance against the given SLOs. By default, Intersight Workload Optimizer estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Intersight Workload Optimizer does not recommend actions for a Business Transaction, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Transaction lists these actions, thus providing visibility into the risks that have a direct impact on the Business Transaction's performance.

Business Transaction Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

None

Intersight Workload Optimizer does not recommend actions for a Business Transaction, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Transaction lists these actions, thus providing visibility into the risks that have a direct impact on the Business Transaction's performance.

Transaction SLO

Enable this SLO if you are monitoring performance through your Business Transactions.

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Response Time SLO

Enable this SLO if you are monitoring performance through your Business Transactions.

| Attribute | Default Setting/Value |
|--------------------------|---|
| Enable Response Time SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Response Time SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000 ms. You can change this to a different value. |

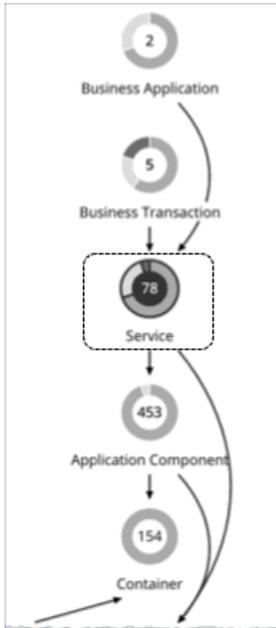
Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Service

A Service in the supply chain represents one or several Application Components that perform a defined, measurable function as part of an internal or user-initiated request. Its performance is key to understanding application performance, but only indirectly impacts user experience. You can measure performance as experienced internal to the Business Application in the context of Services.

This topic discusses Services discovered through APM targets and user-defined application topologies. For details about Services discovered through container platform targets, see this [topic \(on page 222\)](#).

Synopsis



| Synopsis | |
|----------------|---|
| Provides: | Response time and transactions to Business Transactions (on page 210) and Business Applications |
| Consumes from: | Application Components, Database Servers, and the underlying nodes |
| Discovery: | Intersight Workload Optimizer discovers the following: <ul style="list-style-type: none"> ■ AppDynamics Tiers ■ Dynatrace Services ■ Instana Services ■ NewRelic APM Applications / NewRelic Services (New Relic ONE) <p>NOTE: If you do not have an APM target, you can create your own Services using the Application Topology feature. For details, see Application Topology (on page 219).</p> |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Response Time

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Transaction

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity. The **Response Time** and **Transaction** charts for a Service show average and peak/low values over time. You can gauge performance against the given SLOs. By default, Intersight Workload Optimizer estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Intersight Workload Optimizer does not recommend actions for services discovered through APM targets, but it does recommend actions for the underlying Application Components and nodes. The Pending Actions chart for services list these actions, thus providing visibility into the risks that have a direct impact on their performance.

Service Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

This topic discusses policy settings for Services discovered through APM targets. For Services discovered through container platform targets, a different set of policy settings apply. For details, see this [topic \(on page 223\)](#).

Automation Workflow

Intersight Workload Optimizer does not recommend actions for services discovered through APM targets, but it does recommend actions for the underlying Application Components and nodes. The Pending Actions chart for services list these actions, thus providing visibility into the risks that have a direct impact on their performance.

Response Time SLO

Enable this SLO if you are monitoring performance through Services.

| Attribute | Default Setting/Value |
|--------------------------|---|
| Enable Response Time SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Response Time SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000 ms. You can change this to a different value. |

Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Transaction SLO

Enable this SLO if you are monitoring performance through Services.

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

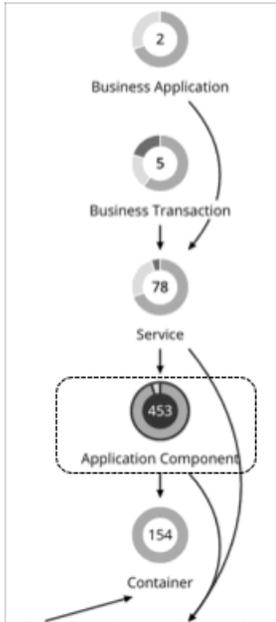
Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Application Component

An Application Component is a software component, application code, or a unit of processing within a [Service \(on page 213\)](#) that consumes resources to enable it to perform its function for the [Business Application \(on page 208\)](#). For example, Apache Tomcat is a Java Servlet container that hosts a range of Java applications on the web.

Intersight Workload Optimizer can recommend actions to adjust the amount of resources available to Application Components.

Synopsis



| Synopsis | |
|------------|--|
| Provides: | <ul style="list-style-type: none"> Response Time and Transactions to Services, Business Transactions (on page 210), and Business Applications Response Time, Transactions, Heap, Remaining GC Capacity, and Threads to end users |
| Consumes: | Compute resources from nodes |
| Discovery: | Intersight Workload Optimizer discovers the following: <ul style="list-style-type: none"> Apache Tomcat AppDynamics Nodes Dynatrace Processes NewRelic APM Application Instances SNMP WMI |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

The exact resources that are monitored will differ based on application type. This list includes all of the resources that you may see.

- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.

- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
- Heap
Heap is the portion of a VM or container's memory allocated to individual applications.
- Response Time
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Connection
Connection is the measurement of database connections utilized by applications.
- Remaining GC Capacity
Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).
- Threads
Threads is the measurement of thread capacity utilized by applications.

The charts for an Application Component show average and peak/low values over time. You can gauge performance against the given SLOs. By default, Intersight Workload Optimizer does not enable SLOs in the default policy for Application Components. It estimates SLOs based on monitored values, but does not use these values in its analysis.

NOTE:

In container platform environments, SLOs defined in a service policy override any SLOs set in the associated Application Components to prevent conflicts. In addition, the Response Time and Transaction charts for Application Components will show SLOs specified in the service policy. For more information, see this [topic \(on page 223\)](#).

Actions

Intersight Workload Optimizer supports the following actions:

Resize

Resize the following resources to maintain performance:

- Thread Pool
Intersight Workload Optimizer generates thread pool resize actions. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.
- Connections
Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers.
- Heap
Intersight Workload Optimizer generates Heap resize actions if an Application Component provides Heap and Remaining GC Capacity, and the underlying VM or container provides VMem. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.

NOTE:

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

The resources that Intersight Workload Optimizer can resize depend on the processes that it discovers from your Application Performance Management (APM) targets. Refer to the topic for a specific target to see a list of resources that can be resized.

Application Component Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about Application Component actions, see [Application Component Actions. \(on page 216\)](#)

| Action | Default Mode |
|---------------------------------|--------------|
| Resize connections (up or down) | Recommend |
| Resize heap (up or down) | Recommend |
| Resize thread pool (up or down) | Recommend |

Response Time SLO

Enable this SLO to monitor the performance of your Application Components.

NOTE:

In container platform environments, SLOs defined in a service policy override any SLOs set in the associated Application Components to prevent conflicts. In addition, the Response Time and Transaction charts for Application Components will show SLOs specified in the service policy. For more information, see this [topic \(on page 223\)](#).

| Attribute | Default Setting/Value |
|--------------------------|---|
| Enable Response Time SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Response Time SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000 ms. You can change this to a different value. |

Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Transaction SLO

Enable this SLO to monitor the performance of your Application Components.

NOTE:

In container platform environments, SLOs defined in a service policy override any SLOs set in the associated Application Components to prevent conflicts. In addition, the Response Time and Transaction charts for Application Components will show SLOs specified in the service policy. For more information, see this [topic \(on page 223\)](#).

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

Heap Utilization

The Heap utilization that you set here specifies the percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity. For example, a value of 80 means that Intersight Workload Optimizer considers 80% utilization to be 100% of capacity.

| Attribute | Default Value |
|----------------------|---------------|
| Heap Utilization (%) | 80 |

Intersight Workload Optimizer uses Heap utilization and Remaining GC Capacity (the percentage of CPU time *not* spent on garbage collection) when making scaling decisions. Assume Heap utilization is at 80%, which is 100% of capacity. However, if Remaining GC Capacity is at least 90% (in other words, CPU time spent on garbage collection is only 10% or less), an 80% Heap utilization does not indicate a shortage after all. As a result, Intersight Workload Optimizer will not recommend Heap scaling.

If Heap utilization is low and Remaining GC Capacity is high, Intersight Workload Optimizer will recommend resizing down Heap. If the opposite is true, then Intersight Workload Optimizer will recommend resizing up Heap.

Heap Scaling Increment

This increment specifies how many units to add or subtract when scaling Heap for an application component.

| Attribute | Default Value |
|------------------------|---------------|
| Heap Scaling Increment | 128 MB |

Do not set the increment value to be lower than what is necessary for the Application Component to operate. If the increment is too low, then it's possible there would be insufficient Heap for the Application Component to operate. When reducing allocation, Intersight Workload Optimizer will not leave an Application Component with less than the increment value. For example, if you use the default 128, then Intersight Workload Optimizer cannot reduce the Heap to less than 128 MB.

Aggressiveness and Observation Periods

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU, vMEM, Heap, and Garbage Collection. It then recommends actions to improve utilization based on the observed values for a given time period.

■ Aggressiveness

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce capacity. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce resources for that Application Component.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single burst to 100%, but for 95% of the samples, utilization never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce resource allocation.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 99th Percentile – More performance. Recommended for critical Application Components that need maximum guaranteed performance at all times, or those that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.

- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for Application Components that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 14 days. Use the **Max Observation Period** setting to adjust the number of days. To ensure that there are enough samples to analyze and drive resize actions, set the **Min Observation Period**.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 14 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the Application Component has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 7 Days or Last 3 Days

Intersight Workload Optimizer recommends an observation period of 14 days so it can recommend resize actions more often. Since Application Component resizing is minimally disruptive, resizing often should not introduce any noticeable performance risks.

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

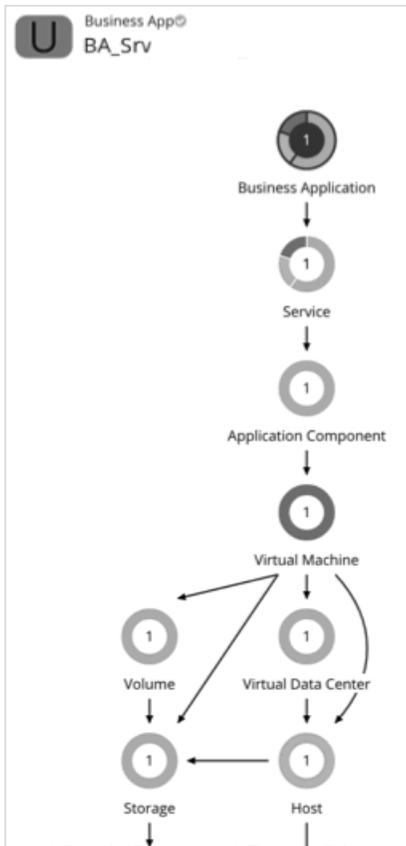
- More Elastic – None
- Less Elastic – 1, 3, or 7 Days

Application Topology

You can create your own [Business Applications \(on page 208\)](#), [Business Transactions \(on page 210\)](#), and [Services \(on page 213\)](#) without the need to load more application data into the platform. This is especially useful in environments with gaps in the application stack that is shown in the Intersight Workload Optimizer supply chain. For example, in the absence of an application monitoring target such as AppDynamics or Dynatrace, you do not see Business Applications in your supply chain. User-created application entities address those gaps.

When you create a new application entity, you identify interrelated application entities in the environment for which you want to measure performance.

Intersight Workload Optimizer then links them in a supply chain and represents them as a unified group. You can monitor overall performance for the group in the context of the new application entity. Drill down to the newly-created entity to monitor performance. You can also use Search to find the application entity and set it as your scope.



Intersight Workload Optimizer does not run analysis on any user-created application entity, but it aggregates the underlying risks the same way it does for auto-discovered entities. Intersight Workload Optimizer adds newly-created entity to the relevant charts. For example, if you created a new Service that has performance risks, it appears in the Top Services chart.

NOTE:

It might take up to 10 minutes to see newly created entities in the supply chain.

Creating Application Entities

1. Click **More**, then display the **Settings** Page.
2. Choose **Application Topology**.



3. Click **New Application Topology** and then choose **Automatic** or **Manual**.

- **Automatic**

Create a new application entity that is composed of tagged entities. Intersight Workload Optimizer automatically adds selected entities to the proper topology.

For example, create a new Business Application that is composed of VMs with the "Production" tag.

- a. Select the application entity type to create.
- b. Specify an entity name prefix to help you easily identify the application entities that Intersight Workload Optimizer creates for you.
- c. Specify the tags that identify the underlying entities.

- **Manual**

Create a new application entity that is composed of a specific set of application entities and nodes. Manually creating the topology allows for more flexibility.

- a. Select the application entity type that you want to create.
- b. Specify the application entity name.
- c. Select the underlying application entities and nodes.
- d. Enable or disable **Direct Link**.

- Disabled (default)

When **Direct Link** is disabled, Intersight Workload Optimizer creates a context-based definition of the application entity you are creating and automatically updates that definition as the entity evolves. You create flexible definitions with minimal effort.

The underlying application entities and nodes that you specified act as "seed entities" for creating the definition. Intersight Workload Optimizer uses these seed entities to identify the highest entity in the supply chain and any other related entities ("leaf entities"), and then creates a new context-based definition. The result is an application topology that closely matches your environment.

For example, your initial intent might be to create a new Business Application entity composed of several Services (seed entities), so you can monitor performance at the Service level. However, you might not be aware of other entities that might impact performance, making it more time-consuming to identify and resolve performance issues outside of the selected scope. With **Direct Link** disabled, Intersight Workload Optimizer might discover Application Components and VMs (leaf entities) that back the Services, and then show them in the supply chain. The result is a complete representation of the Business Application that shows performance risks at each level of the discovered application stack. As the composition of the Business Application changes, Intersight Workload Optimizer automatically updates the definition so your supply chain view remains current.

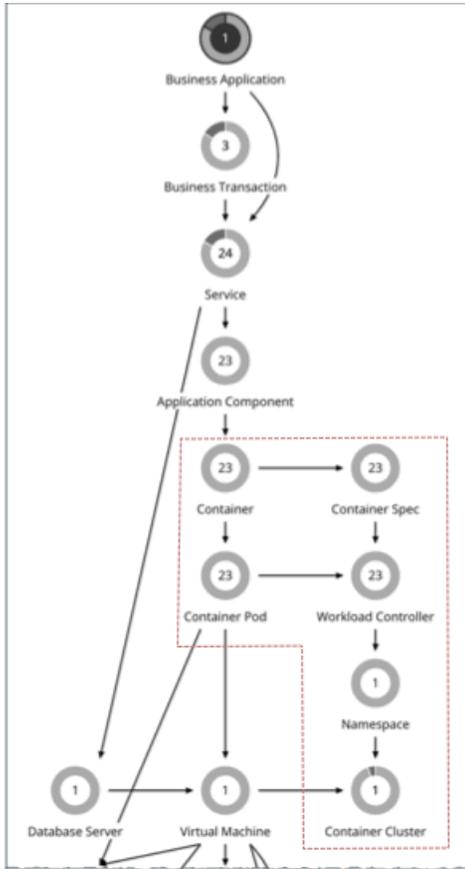
- Enabled

When **Direct Link** is enabled, Intersight Workload Optimizer creates a definition based solely on your selected entities. This option is ideal if you require full control of your definitions. For example, you might have a requirement to limit the scope of your performance monitoring to certain entities.

4. Click **Create Definition**.

Entity Types - Container Platform

Intersight Workload Optimizer discovers and monitors the entities that make up your container platform, and recommends actions to assure performance for the applications that consume resources from these entities.



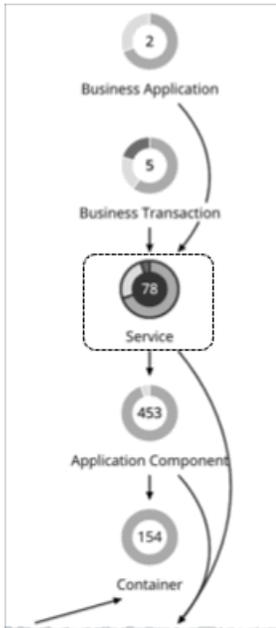
Container Platform Service

A service in container platform environments is a logical set of pods that represents a given application. The service exposes a single entry point for the application process. While the pods that comprise the service are ephemeral, the service is persistent. The service entity also gives historical tracking of the number of replicas that run to support the service.

NOTE:

For details about services discovered through APM targets, see this [topic \(on page 213\)](#).

Synopsis



| | |
|---------------------|---|
| Synopsis | |
| Provides: | Response time and transactions to Business Transactions (on page 210) and Business Applications |
| Consumes from: | Container pods and the underlying nodes |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Response Time**
Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- **Transaction**
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

The **Response Time** and **Transaction** charts for a Service show average and peak/low values over time. You can gauge performance against the given SLOs. By default, Intersight Workload Optimizer estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Intersight Workload Optimizer does not recommend actions for services in container platform environments, but it does recommend actions for the replicas that back those services.

For details, see [Workload Controller Scale Actions \(on page 238\)](#).

Container Platform Service Policies

Policies are required to generate [SLO-driven scale actions \(on page 238\)](#) for container platform services. You can create policies from the Intersight Workload Optimizer user interface or by using Custom Resources in your container platform clusters.

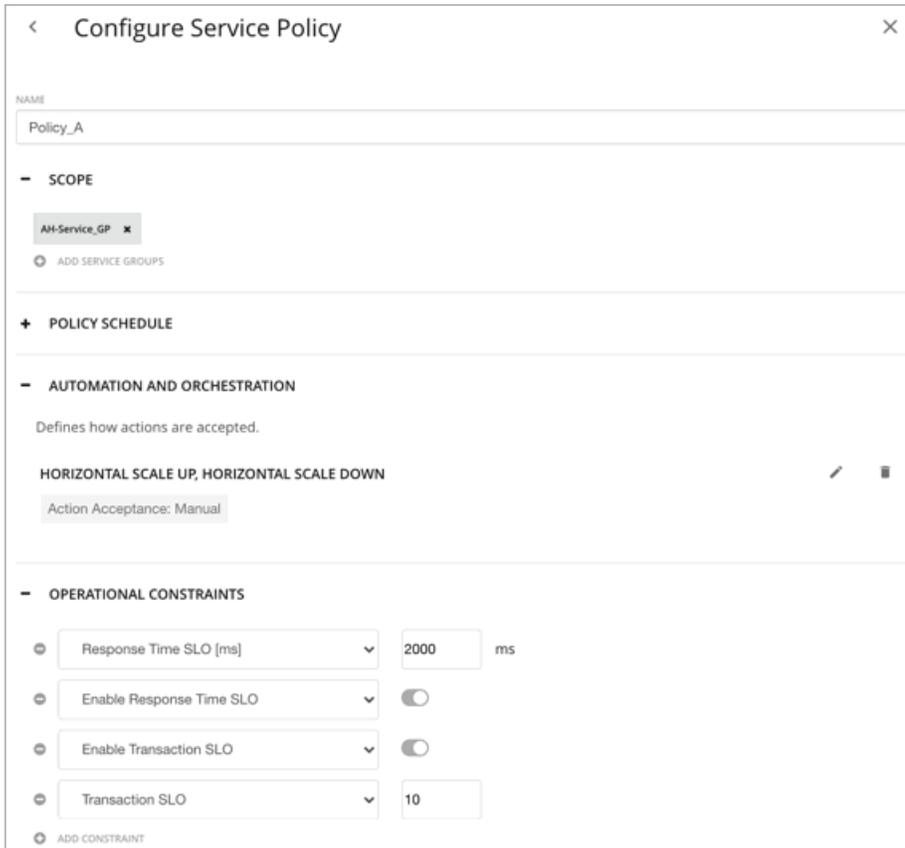
In addition to policies, certain requirements must be satisfied in order for Intersight Workload Optimizer to generate SLO-driven scale actions. For details, see this [topic \(on page 239\)](#).

NOTE:

When a group of services applies multiple conflicting policies, Intersight Workload Optimizer enforces the policy with the most conservative settings.

For services discovered through APM targets, a different set of policy settings apply. For details, see this [topic \(on page 214\)](#).

Creating Policies From the User Interface



The screenshot shows the 'Configure Service Policy' dialog box. It has a title bar with a back arrow and a close button. The main content is organized into sections:

- NAME:** A text input field containing 'Policy_A'.
- SCOPE:** A section with a minus sign. It contains a tag 'AH-Service_GP' with a close icon and an 'ADD SERVICE GROUPS' button.
- POLICY SCHEDULE:** A section with a plus sign, currently collapsed.
- AUTOMATION AND ORCHESTRATION:** A section with a minus sign. It contains the text 'Defines how actions are accepted.' and a dropdown menu for 'HORIZONTAL SCALE UP, HORIZONTAL SCALE DOWN' set to 'Action Acceptance: Manual'.
- OPERATIONAL CONSTRAINTS:** A section with a minus sign. It contains four rows of controls:
 - 'Response Time SLO [ms]' dropdown set to '2000' with a 'ms' unit label.
 - 'Enable Response Time SLO' dropdown with a toggle switch turned off.
 - 'Enable Transaction SLO' dropdown with a toggle switch turned off.
 - 'Transaction SLO' dropdown set to '10'.

At the bottom left of the section is an 'ADD CONSTRAINT' button.

■ Automation Workflow

Intersight Workload Optimizer does not recommend actions for services in container platform environments, but it does recommend actions for the replicas that back those services.

To generate these actions, you must turn on horizontal scaling (up and/or down) and specify your desired SLOs.

| Attribute | Default Setting/Value |
|-----------------------|-----------------------|
| Horizontal Scale Down | Off (Disabled) |
| Horizontal Scale Up | Off (Disabled) |

Action automation is recommended under the following circumstances:

- Your applications run as a set of services backed by a deployment.
- Services deploy via a namespace *without* quotas.
- Newly provisioned pods are placed on nodes with the same CPU speed.
- You do not have an upstream HPA (Horizontal Pod Autoscaling) enabled for the same workload.

It is recommended that you disable automation and manually execute actions under the following circumstances:

- Services deploy via a namespace *with* quotas.

- Newly created pods are scheduled on nodes with different CPU speeds.
- Services have non-resource constraints that could result in newly provisioned pods staying in the pending state.

■ Transaction SLO

Transaction SLO is the maximum number of transactions per second that each Application Component replica can handle.

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

■ Response Time SLO

Response Time SLO is the desired *weighted average* response time (in milliseconds) of all Application Component replicas associated with a Service.

| Attribute | Default Setting/Value |
|--------------------------|--|
| Enable Response Time SLO | Off |
| Response Time SLO [ms] | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000. You can change this to a different value. |

Minimum and Maximum Replicas

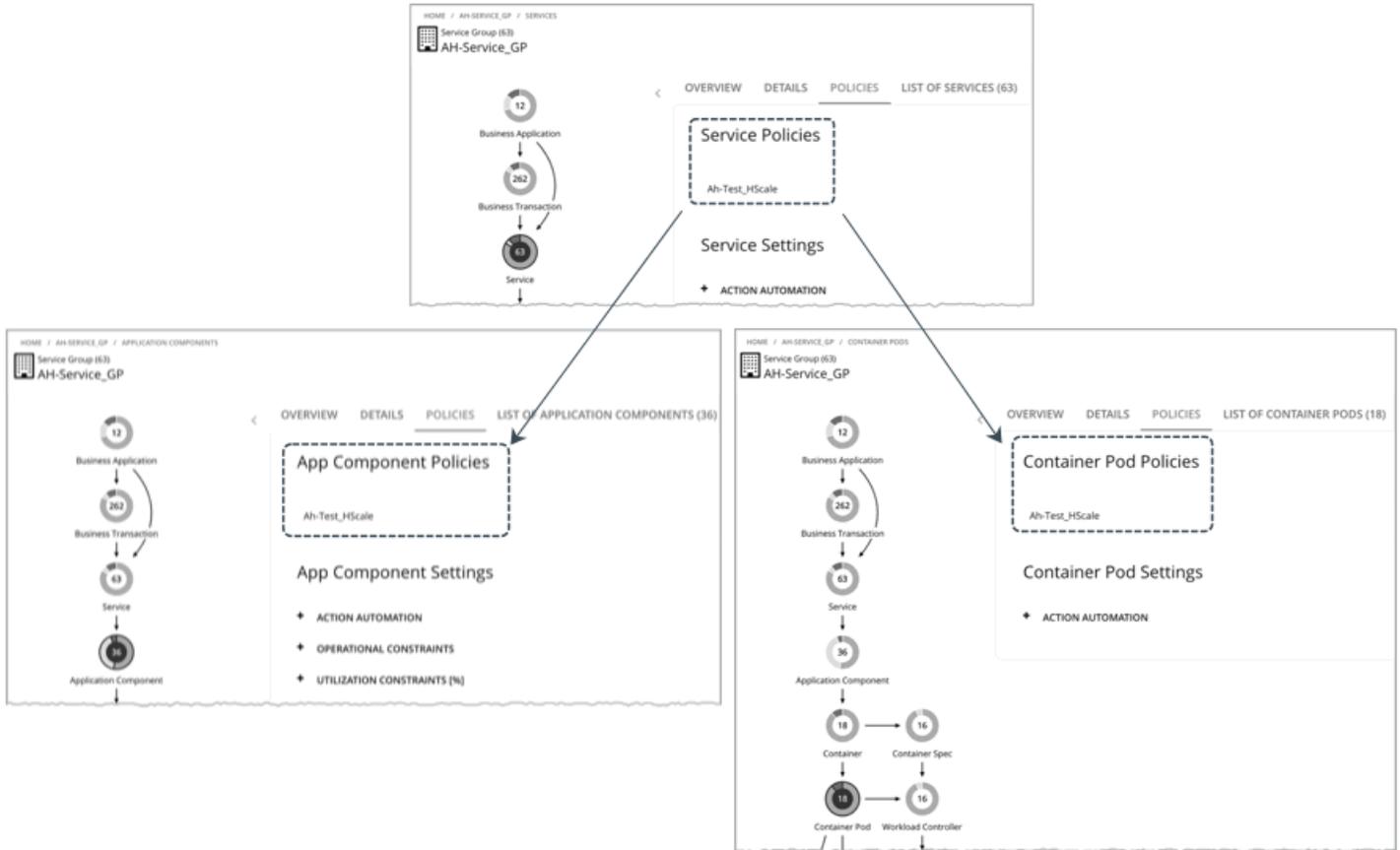
You can adjust the default values based on the characteristics of your applications or if you are planning for capacity. The maximum value also acts as a safeguard against overprovisioning of replicas.

| Attribute | Default Setting/Value |
|------------------|-----------------------|
| Minimum Replicas | 1 |
| Maximum Replicas | 10000 |

Propagation of Service Policy Settings

Settings in a service policy propagate to the associated pods and Application Components to establish their relationship and provide context.

For example, assume you created a group of services called `AH-Service_GP` and then applied the service policy `Ah-Test_HScale` to that group. When you set the scope to this group, `Ah-Test_HScale` displays as a policy in the entity views for services, Application Components, and Container Pods. You can click the policy name in any view to see or modify the policy settings.

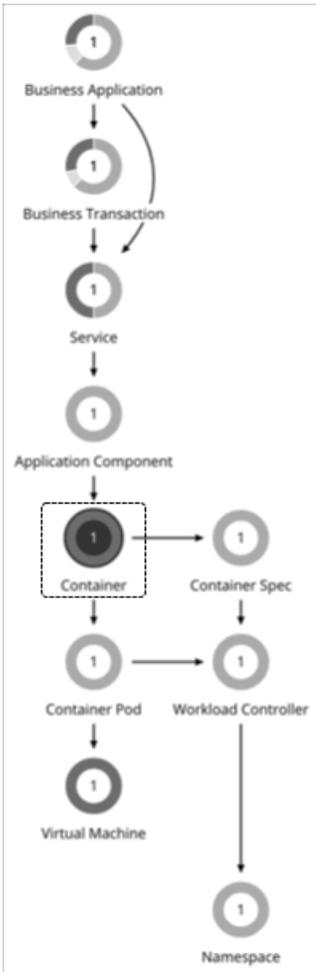


To prevent conflicts, SLO values in service policies override any SLOs set in Application Components. In addition, the Response Time and Transaction charts for Application Components show SLOs specified in the service policy.

Container

An application container is a standalone, executable image of software that includes components to host an application. Because the container instances that support an application can change at any time, containers are considered *ephemeral*.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Resources for the applications to use, including: <ul style="list-style-type: none"> Virtual CPU Virtual Memory |
| Consumes: | Resources from container pods, virtual machines, and virtual datacenters |
| Discovered through: | For container platforms, Intersight Workload Optimizer discovers containers through the Kuberturno agent that you deployed to your cluster. For Dynatrace and AppDynamics hosted on containers: <ul style="list-style-type: none"> Dynatrace: Intersight Workload Optimizer discovers containers through the metadata of processes. AppDynamics: Intersight Workload Optimizer discovers containers through container objects. |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)
 - VMem is the virtual memory utilized by a container against the memory limit. If no limit is set, node capacity is used.
- VMem Request

If applicable, VMem Request is the virtual memory utilized by a container against the memory request.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- VCPU Request

If applicable, VCPU Request is the virtual CPU (in mCores) utilized by a container against the CPU request.

- VCPU Throttling

VCPU Throttling is the throttling of container virtual CPU that could impact response time, expressed as the percentage of throttling for all containers associated with a Container Spec. In the Capacity and Usage chart for containers, *used* and *utilization* values reflect the actual throttling percentage, while *capacity* value is always 100%.

Actions

None

Intersight Workload Optimizer does not recommend actions on containers.

Container Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

None

Intersight Workload Optimizer does not recommend actions on containers.

Consistent Resizing

- *For groups in user-defined automation policies:*

| Attribute | Default Setting |
|---------------------|-----------------|
| Consistent Resizing | Off |

When you create a policy for a group of containers and turn on Consistent Resizing, Intersight Workload Optimizer resizes all the group members to the same size, such that they all support the top utilization of each resource commodity in the group. For example, assume container A shows top utilization of CPU, and container B shows top utilization of memory. Container resize actions would result in all the containers with CPU capacity to satisfy container A, and memory capacity to satisfy container B.

For an affected resize, the Actions List shows individual resize actions for each of the containers in the group. If you automate resizes, Intersight Workload Optimizer executes each resize individually in a way that avoids disruption to your workloads.

- *For auto-discovered groups:*

Intersight Workload Optimizer discovers container platform groups such as Deployments, ReplicationControllers, ReplicaSets, DaemonSets, and StatefulSets, and automatically enables Consistent Resizing in a read-only policy for each group. If you do not need to resize all the members consistently, create another policy for the group and turn off Consistent Resizing.

Container Spec

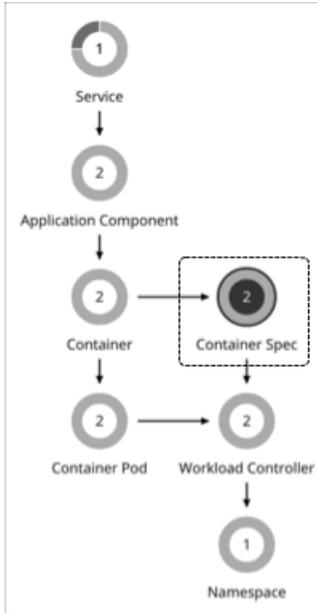
A container spec is a shared definition for all ephemeral container replicas with like properties. It is a persistent entity that maintains historical data for its ephemeral containers, and all the replicas that have run in the past. Intersight Workload Optimizer uses historical data to make container sizing decisions. Data includes:

- vCPU used by all container replicas
- vCPU request capacity (if applicable)
- vMem used by all container replicas

- vMem request capacity (if applicable)

In the Intersight Workload Optimizer supply chain, the count of replicas maps to the count of container entities that a container spec encompasses.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | N/A |
| Consumes: | N/A |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Monitored Resources

Intersight Workload Optimizer monitors the historical usage of any instance of a container running for the workload (assuming the workload name stays the same). Charts show the trend of usage even with restarts or redeployments.

Actions

Intersight Workload Optimizer supports the following actions:

Resize (via workload controllers)

A container spec retains the historical utilization data of ephemeral containers. Intersight Workload Optimizer uses this data to make resize decisions that assure optimal utilization of resources. By default, all replicas of the same container for the same workload type resize consistently.

For details, see [Workload Controller Resize Actions \(on page 236\)](#).

Constraint for Sidecar Container Specs

A container platform service might include [sidecar](#) container specs to provide additional services to a running pod, such as security or logging services. Sidecars injected at pod creation cannot be updated from the parent Workload Controller, causing a resize action to fail.

To prevent the execution of resize actions on injected sidecars, Intersight Workload Optimizer adds them to a group called "Injected Sidecars/All ContainerSpecs". This group applies a read-only policy that sets the action acceptance mode for resizes to *Recommend*. This means that you can only execute resizes outside of Intersight Workload Optimizer. The parent Workload Controller will continue to resize non-sidecar container specs as usual.

Container Spec Policies

This policy controls the analysis of resize actions generated on container specs. To define the acceptance mode of these actions, create a [workload controller policy \(on page 241\)](#).

Analysis Properties: Overview

Analysis properties specify the tolerance levels for vCPU limits and requests, vMem limits and requests, and CPU throttling that Intersight Workload Optimizer must consider when it generates resize actions. Resize actions outside these tolerance levels may or may not generate, depending on the action generation setting (on or off) that you specify.

For example, consider resizing vMem limits.

- As vMem demand increases, Intersight Workload Optimizer can generate vMem limit resize actions that fall within the capacity range that you specified for vMem limits. The action acceptance mode defined in the [policy \(on page 241\)](#) for the parent workload controller determines how resize actions are executed. For example, if the action acceptance mode is *automated*, Intersight Workload Optimizer executes resize actions automatically.
- If the container spec requests vMem outside the range and the action generation for out-of-range resize actions is turned on, Intersight Workload Optimizer generates resize actions even if they are outside the range. Note that out-of-range resize actions always generate in *recommend* mode, which means that you can view the actions in [Action Center \(on page 396\)](#), but you can only execute them in your cluster.

Analysis Properties: Resize Within Range

| Setting | Default State |
|-----------------------|---------------|
| Generate Actions | On |
| Set to Recommend mode | Off |

■ Generate Actions

If you turn on this setting, resize actions generate and take into consideration the analysis properties for vCPU limits and requests, vMem limits and requests, and CPU throttling.

If you turn off this setting, resize actions do not generate.

■ Set to Recommend mode

If you turn on this setting, resize actions generate in recommend mode, overriding the action acceptance mode configured in the policy for the parent workload controller. This means that you need to execute workload controller resize actions in your cluster. If this is not your intention, keep this setting turned off. You can turn it on for container specs that cannot be updated from their parent workload controllers, such as sidecars.

Analysis Properties: vCPU Limits

| Setting | Default Value/State |
|--|---------------------|
| Capacity range (in mCores) | 500 to 64000 |
| Resize below minimum – Generate action in recommend mode | Off |
| Resize above maximum – Generate action in recommend mode | On |
| Target utilization | Midpoint – 70% |

■ Capacity range

Intersight Workload Optimizer generates resize actions that fall within the specified capacity range.

■ Resize below minimum / Resize above maximum

If resize values fall outside the specified capacity range and this setting is:

- On – Intersight Workload Optimizer generates resize actions in recommend mode. You can execute these actions only in your cluster.
- Off – Intersight Workload Optimizer does not generate resize actions.

For example, if the capacity range is 500 to 64000 and the resize below minimum setting is turned off, Intersight Workload Optimizer does not generate resize actions that are lower than 500 mCores.

■ **Target utilization**

Midpoint determines the range of utilization that is considered optimal. The lower and upper limits for this range are always 5% below and above the midpoint, respectively. For example, A midpoint of 70% means that optimal utilization is between 65% and 75%. Intersight Workload Optimizer takes this optimal utilization range (along with other resize factors) into consideration when it generates resize actions. For example, it can generate a resize action to improve performance while keeping utilization within this range.

The slider shows 10% and 90% as the minimum and maximum midpoint values that you can specify, for your reference. These values are not configurable. This means that you cannot specify a midpoint value that is lower than 10% or higher than 90%. For example, if you set the midpoint to the minimum of 10%, the optimal utilization range is 5% to 15%.

Analysis Properties: vCPU Requests

| Setting | Default Value/State |
|--|---------------------|
| Minimum value (in mCores) | 10 |
| Resize below minimum – Generate action in recommend mode | On |
| Target utilization | Midpoint – 70% |

■ **Minimum value**

Intersight Workload Optimizer generates resize actions if resize values are equal to or higher than the minimum value.

■ **Resize below minimum**

If resize values are lower than the minimum specified value and this setting is:

- On – Intersight Workload Optimizer generates resize actions in recommend mode. You can execute these actions only in your cluster.
- Off – Intersight Workload Optimizer does not generate resize actions.

For example, if the minimum value is 10 and the resize below minimum setting is turned on, Intersight Workload Optimizer generates resize actions that are lower than 10 mCores. You can execute these actions only in your cluster.

■ **Target utilization**

Midpoint determines the range of utilization that is considered optimal. The lower and upper limits for this range are always 5% below and above the midpoint, respectively. For example, A midpoint of 70% means that optimal utilization is between 65% and 75%. Intersight Workload Optimizer takes this optimal utilization range (along with other resize factors) into consideration when it generates resize actions. For example, it can generate a resize action to improve performance while keeping utilization within this range.

The slider shows 10% and 90% as the minimum and maximum midpoint values that you can specify, for your reference. These values are not configurable. This means that you cannot specify a midpoint value that is lower than 10% or higher than 90%. For example, if you set the midpoint to the minimum of 10%, the optimal utilization range is 5% to 15%.

Analysis Properties: vMem Limits

| Setting | Default Value/State |
|--|---------------------|
| Capacity range (in MB) | 10 to 1048576 |
| Resize below minimum – Generate action in recommend mode | On |
| Resize above maximum – Generate action in recommend mode | On |
| Target utilization | Midpoint – 70% |

■ **Capacity range**

Intersight Workload Optimizer generates resize actions that fall within the specified capacity range.

■ **Resize below minimum / Resize above maximum**

If resize values fall outside the specified capacity range and this setting is:

- On – Intersight Workload Optimizer generates resize actions in recommend mode. You can execute these actions only in your cluster.
- Off – Intersight Workload Optimizer does not generate resize actions.

For example, if the capacity range is 10 to 104857 and the resize below minimum setting is turned on, Intersight Workload Optimizer generates resize actions that are lower than 10 MB. You can execute these actions only in your cluster.

■ Target utilization

Midpoint determines the range of utilization that is considered optimal. The lower and upper limits for this range are always 5% below and above the midpoint, respectively. For example, A midpoint of 70% means that optimal utilization is between 65% and 75%. Intersight Workload Optimizer takes this optimal utilization range (along with other resize factors) into consideration when it generates resize actions. For example, it can generate a resize action to improve performance while keeping utilization within this range.

The slider shows 10% and 90% as the minimum and maximum midpoint values that you can specify, for your reference. These values are not configurable. This means that you cannot specify a midpoint value that is lower than 10% or higher than 90%. For example, if you set the midpoint to the minimum of 10%, the optimal utilization range is 5% to 15%.

Analysis Properties: vMem Requests

| Setting | Default Value/State |
|--|---------------------|
| Minimum value (in MB) | 10 |
| Resize below minimum – Generate action in recommend mode | On |
| Target utilization | Midpoint – 70% |

■ Minimum value

Intersight Workload Optimizer generates resize actions if resize values are equal to or higher than the minimum value.

■ Resize below minimum

If resize values are lower than the minimum specified value and this setting is:

- On – Intersight Workload Optimizer generates resize actions in recommend mode. You can execute these actions only in your cluster.
- Off – Intersight Workload Optimizer does not generate resize actions.

For example, if the minimum value is 10 and the resize below minimum setting is turned on, Intersight Workload Optimizer generates resize actions that are lower than 10 MB. You can execute these actions only in your cluster.

■ Target utilization

Midpoint determines the range of utilization that is considered optimal. The lower and upper limits for this range are always 5% below and above the midpoint, respectively. For example, A midpoint of 70% means that optimal utilization is between 65% and 75%. Intersight Workload Optimizer takes this optimal utilization range (along with other resize factors) into consideration when it generates resize actions. For example, it can generate a resize action to improve performance while keeping utilization within this range.

The slider shows 10% and 90% as the minimum and maximum midpoint values that you can specify, for your reference. These values are not configurable. This means that you cannot specify a midpoint value that is lower than 10% or higher than 90%. For example, if you set the midpoint to the minimum of 10%, the optimal utilization range is 5% to 15%.

Analysis Properties: CPU Throttling

| Setting | Default Value |
|--------------------------|----------------|
| CPU throttling tolerance | Midpoint – 20% |

This value defines your acceptable level of throttling and directly impacts the resize actions generated on CPU limits.

A low percentage value indicates more sensitivity to throttling, while a high value indicates more tolerance for throttling and a higher risk of congestion.

Midpoint determines the range of throttling that will be tolerated. The slider shows 10% and 70% as the minimum and maximum midpoint values that you can specify, for your reference. These values are not configurable. This means that you cannot specify a midpoint value that is lower than 10% or higher than 70%.

Learn more about CPU throttling [here](#).

Scaling Constraints: Increment Constants

Intersight Workload Optimizer recommends changes in terms of the specified resize increments.

| Attribute | Default Value |
|---|---------------|
| Increment constant for vCPU Limit and vCPU Request (mCores) | 100 |
| Increment constant for vMem Limit and vMem Request (MB) | 128 |

For example, assume the vCPU request increment is 100 mCores and you have requested 800 mCores for a container. Intersight Workload Optimizer could recommend reducing the request by 100, down to 700 mCores.

For vMem, do not set the increment value to be lower than what is necessary for the container to operate. If the vMem increment is too low, Intersight Workload Optimizer might allocate insufficient vMem. For a container that is underutilized, Intersight Workload Optimizer will reduce vMem allocation by the increment amount, but it will not leave a container with zero vMem. For example, if you set this to 128, then Intersight Workload Optimizer cannot reduce the vMem to less than 128 MB.

Scaling Constraints: Rate of Resize

(For the *default* policy only)

| Attribute | Default Value |
|----------------|---------------|
| Rate of Resize | High |

When resizing resources, Intersight Workload Optimizer calculates the optimal values for vCPU and vMem, but it does not necessarily make a change to that value in one action. Intersight Workload Optimizer uses the rate of resize setting to determine how to make the change in a single action.

- **Low**

Change the value by one increment, only. For example, if the resize action calls for increasing vMem, and the increment is set at 128, Intersight Workload Optimizer increases vMem by 128 MB.

- **Medium**

Change the value by an increment that is 1/4 of the difference between the current value and the optimal value. For example, if the current vMem is 2 GB and the optimal vMem is 10 GB, then Intersight Workload Optimizer will raise vMem to 4 GB (or as close to that as the increment constant will allow).

- **High**

Change the value to be the optimal value. For example, if the current vMem is 2 GB and the optimal vMem is 8 GB, then Intersight Workload Optimizer will raise vMem to 8 GB (or as close to that as the increment constant will allow).

Scaling Constraints: Aggressiveness and Observation Periods

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU and vMem. It then recommends actions to improve utilization based on the observed values for a given time period.

- **Aggressiveness**

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 99th Percentile |

When evaluating vCPU and vMem performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 99th percentile. The percentile

utilization is the highest value that 99% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce the capacity for vCPU on a container. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. For most containers there are moments when peak vCPU reaches high levels. Assume utilization for a container peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce allocated vCPU for that container.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single vCPU burst to 100%, but for 99% of the samples vCPU never exceeded 50%. If you set **Aggressiveness** to 99th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce vCPU allocation for the container.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 100th Percentile
Least aggressive, recommended for critical workloads that need maximum guaranteed performance at all times
- 99th Percentile (Default)
Recommended setting to achieve maximum performance
- 90th Percentile
Most aggressive, recommended for non-production workloads that can stand higher resource utilization

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 30 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. (If the database has fewer days' data then it uses all of the stored historical data.)

A shorter period means there are fewer data points to account for when Intersight Workload Optimizer calculates utilization percentiles. This results in more dynamic, elastic resizing, while a longer period results in more stable or less elastic resizing. You can make the following settings:

- Less Elastic – Last 90 Days
- Recommended – Last 30 Days
- More Elastic – Last 7 Days

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | 1 Day |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

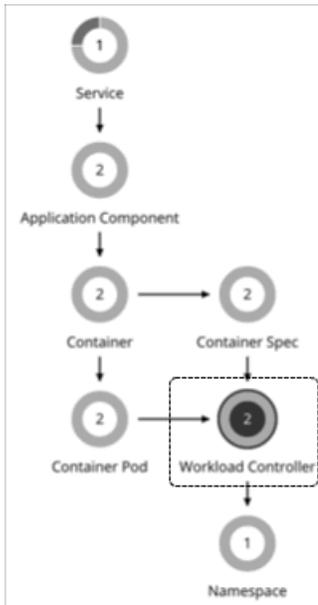
- More Elastic – None
- Recommended – 1 Day
- Less Elastic – 3 or 7 Days

Workload Controller

A workload controller is a container platform entity that watches the state of your pods and then requests changes where needed. Examples of workload controllers are `Deployments` and `StatefulSets`. A single workload controller can contain one or more container spec entities, and it can be related to one or more running replica pods. Like container specs, workload controllers are persistent.

You can execute actions to resize container specs or scale replicas when you set the scope to a workload controller.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | N/A |
| Consumes: | N/A |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Monitored Resources

Intersight Workload Optimizer monitors quotas (limits and requests) for VCPU and VMem, and associates how much each workload controller is contributing to a quota based on all replicas. This allows Intersight Workload Optimizer to generate rightsizing decisions, and manage the quota as a constraint to rightsizing. Metrics on resource consumption are shown in the Container Spec, Container, and Container Pod views.

Workload Controller Actions

Intersight Workload Optimizer supports the following actions:

Resize or Scale

Actions associated with a workload controller resize container specs vertically or scale replicas horizontally. This is a natural representation of these actions because the parent controller's container specs and number of replicas are modified. The workload controller then rolls out the changes in the running environment.

For details, see [Workload Controller Resize Actions \(on page 236\)](#) and [Workload Controller Scale Actions \(on page 238\)](#).

Workload Controller Resize Actions

Actions associated with a workload controller resize container specs vertically to assure optimal utilization of resources. The workload controller then rolls out the changes in the running environment.

Resize Up Actions In Response to CPU Throttling

For vCPU limit resizes, Intersight Workload Optimizer will recommend a resize up action, even if utilization percentile is low, to address slow response times associated with CPU throttling.

CPU throttling occurs when you configure a CPU limit on a container, which can inadvertently slow your applications' response time. Even if you have more than enough resources on your underlying node, your container workload will still be throttled because it was not configured properly. High response times are directly correlated to periods of high CPU throttling.

Learn more about CPU throttling [here](#).

Especially for sudden throttling spikes, Intersight Workload Optimizer will persist the related resize actions so you can evaluate these actions even after the spikes have gone away, and then execute them to prevent spikes from re-occurring. As throttling drops, Intersight Workload Optimizer will not recommend a resize down action right away, as this could result in subsequent back-and-forth upsize and downsize recommendations. Instead, it evaluates past throttling to decide when a resize down action is finally safe to execute. To ensure the timeliness of these actions and arrive at the optimal resize values to recommend, Intersight Workload Optimizer calculates fast and slow moving throttling averages, and then displays *smoothed* and *daily* averages in charts.

Smoothed average is an exponential moving average and moving variance method used on CPU throttling data. It allows analysis to generate a vCPU limit resize up action more quickly when throttling is detected, and be conservative on resize down to mitigate introducing throttling.

Action Propagation

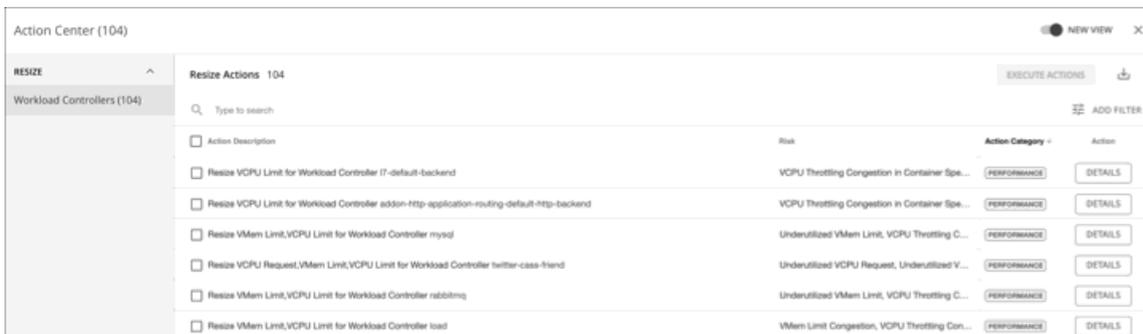
Resize actions propagate to application entities and the underlying container infrastructure to show the impact of these actions on the health of your applications and container environment.

Action Merging

Executing several resize actions via workload controllers can be very disruptive since pods need to restart with each resize. For replicas of the container scale group(s) related to a single workload controller, Intersight Workload Optimizer consolidates resize actions into one *merged action* to minimize disruptions. When a merged action has been executed (via the associated workload controller), all resizes for all related container specifications will be changed at the same time, and pods will restart once.

Action Execution

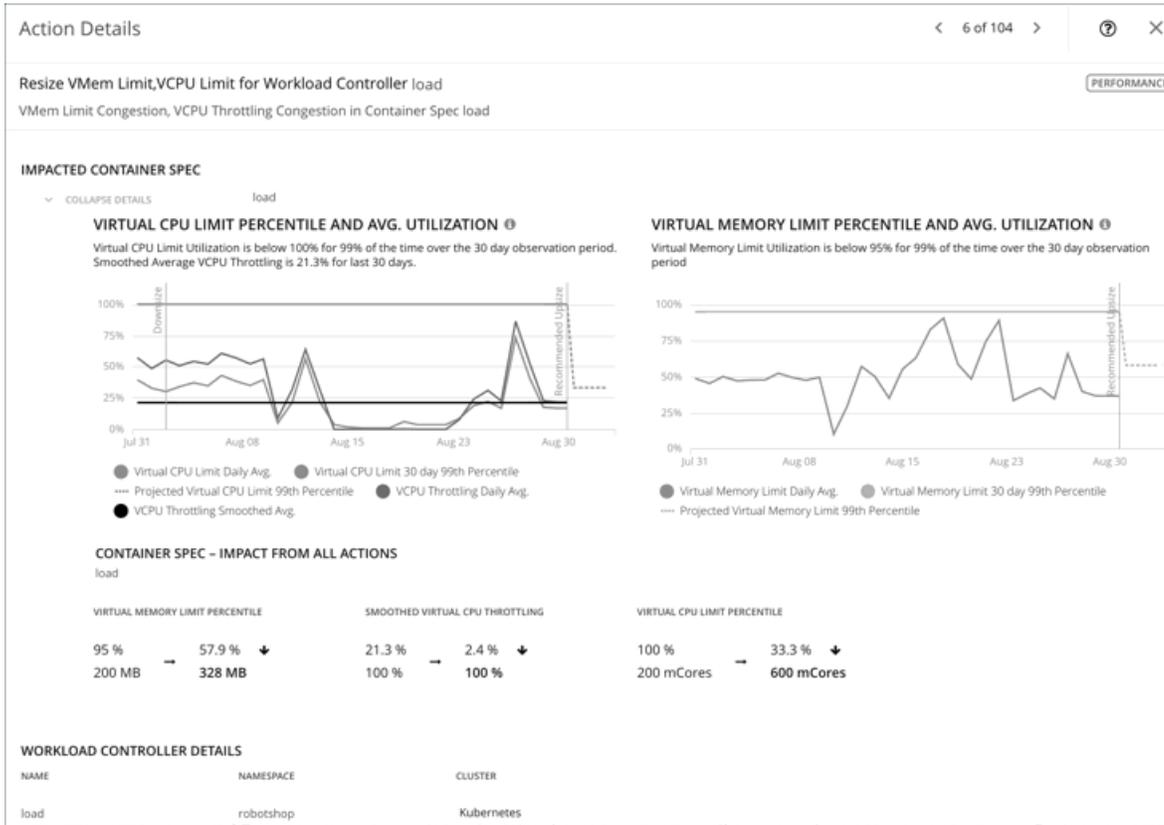
After you set the scope to workload controllers, go to Action Center to see the full list of resize actions that you can execute. This list includes individual and merged actions. You can filter the list to focus on specific actions, such as actions to address resource congestion or vCPU throttling.



| Action Description | Risk | Action Category | Action |
|--|--|-----------------|---------|
| Resize VCPU Limit for Workload Controller [?]-default-backend | VCPU Throttling Congestion in Container Spe... | PERFORMANCE | DETAILS |
| Resize VCPU Limit for Workload Controller add-on-http-application-routing-default-http-backend | VCPU Throttling Congestion in Container Spe... | PERFORMANCE | DETAILS |
| Resize VMem Limit,VCPU Limit for Workload Controller myid | Underutilized VMem Limit, VCPU Throttling C... | PERFORMANCE | DETAILS |
| Resize VCPU Request,VMem Limit,VCPU Limit for Workload Controller twitter-cass-frontend | Underutilized VCPU Request, Underutilized V... | PERFORMANCE | DETAILS |
| Resize VMem Limit,VCPU Limit for Workload Controller ratblmsj | Underutilized VMem Limit, VCPU Throttling C... | PERFORMANCE | DETAILS |
| Resize VMem Limit,VCPU Limit for Workload Controller load | VMem Limit Congestion, VCPU Throttling Con... | PERFORMANCE | DETAILS |

By default, resize actions are set in *Manual* mode at the workload controller level. This means that Intersight Workload Optimizer will not execute any action automatically, and you can manually select the actions that you want to execute. If you prefer to execute actions outside Intersight Workload Optimizer, create workload controller policies and set the resize action acceptance mode to *Recommend*. To automate actions, create workload controller policies and set the resize action acceptance mode to *Automatic*.

For each action, click DETAILS and expand the Details section to view time series charts that explain the reason for the action. These charts highlight *utilization percentiles* and *smoothed throttling averages* for a given observation period. Intersight Workload Optimizer uses percentile calculations to make accurate resize decisions.



These charts also:

- Plot daily average percentiles and throttling, for your reference.
- Show projected percentiles after you execute the action. If you have previously executed resize actions on the same workload controller, the charts show the resulting improvements in daily average utilization.

Put together, these charts allow you to easily recognize trends that drive Intersight Workload Optimizer's resize recommendations.

NOTE:

You can set scaling constraints in container spec policies to refine the percentile calculations. For details, see [Aggressiveness and Observation Periods \(on page 233\)](#).

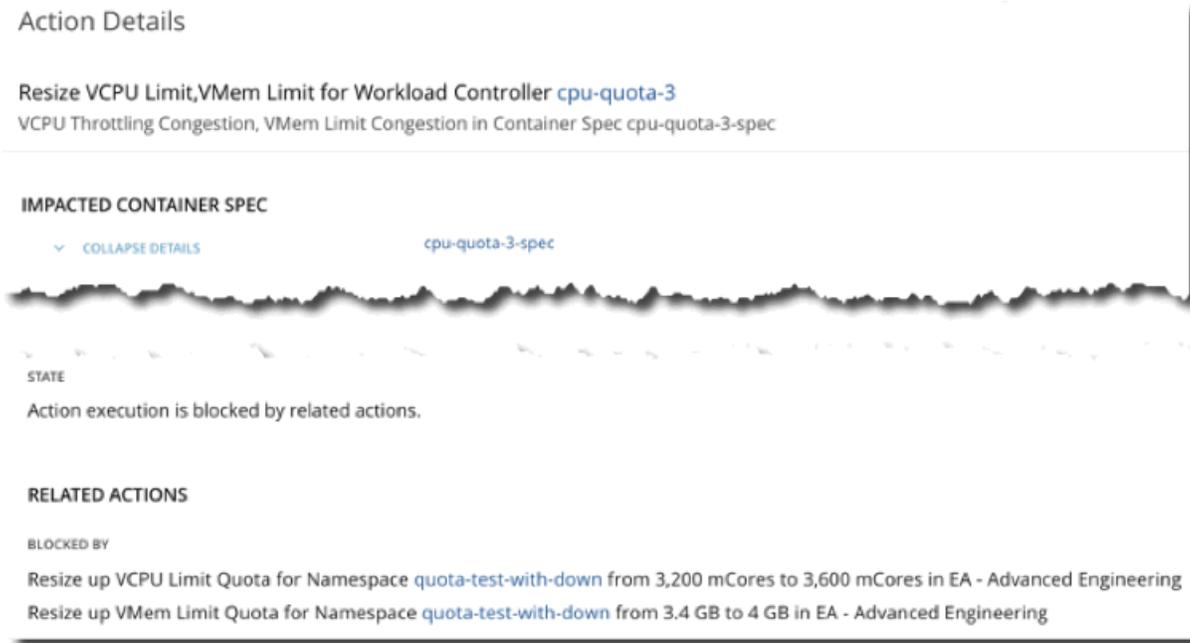
Blocking of Action Execution Due to Insufficient Quotas

Intersight Workload Optimizer treats quotas defined in a namespace as constraints when making resize decisions. If existing actions would exceed the namespace quotas, Intersight Workload Optimizer recommends actions to resize up the affected namespace quota.

NOTE:

For more information about namespace quotas, see [Namespace Monitored Resources \(on page 247\)](#).

The action details for a workload controller include descriptions of the affected container spec entities, and how the resources will change for each. If the resize exceeds current namespace quotas, then Intersight Workload Optimizer blocks the workload controller action. The action details list the Namespace actions that block execution of this resize in the **Related Actions** list.



Downloading Data for Executed Resize Actions

You can view and download data for executed resize actions to gain insight into workloads with performance risks due to resource limits, as well as opportunities to reclaim unused resource requests.

1. Add the Actions chart to your dashboard.
2. Configure the chart. Be sure to set the scope to workload controllers to narrow the results, select Tabular as the chart type, and then set the filter to either All Actions or Executed Actions.
3. In the chart, click **Show All**.
4. In the page that opens, click the download button at the top-right section of the page. If you set All Actions as the filter, select **Executed Actions**.
5. Open the downloaded file and go to the Resize Workload Controller sheet.

Pay attention to the following columns:

- Namespace, Container Spec, and Container Platform Cluster columns – Data in these columns is the most up-to-date at the time of download.
- Change column – This column highlights the impact of resize actions on the resources allocated to containers. Each value represents the difference between the Current Value and New Value columns.
- Container Spec column – If there is more than one container spec in a workload controller, you can sort this column to easily identify resize actions on individual container specs.
- Impacted Commodity column – This column shows if VMem or VCPU was resized. A resize action might show both commodities, which indicates that it is a *merged* action (an action that consolidates resizes related to a single workload controller, to minimize disruptions).

Workload Controller Scale Actions

Actions associated with a workload controller scale the replicas that back horizontally scalable [Container Platform Services \(on page 222\)](#) to maintain SLOs for your applications. The workload controller then rolls out the changes in the running environment.

For example, when current response time for an application is in direct violation of SLO, Intersight Workload Optimizer will recommend increasing the number of replicas to improve response time. If applications can meet SLOs using less resources, Intersight Workload Optimizer will recommend reducing the replica count to improve infrastructure efficiency.

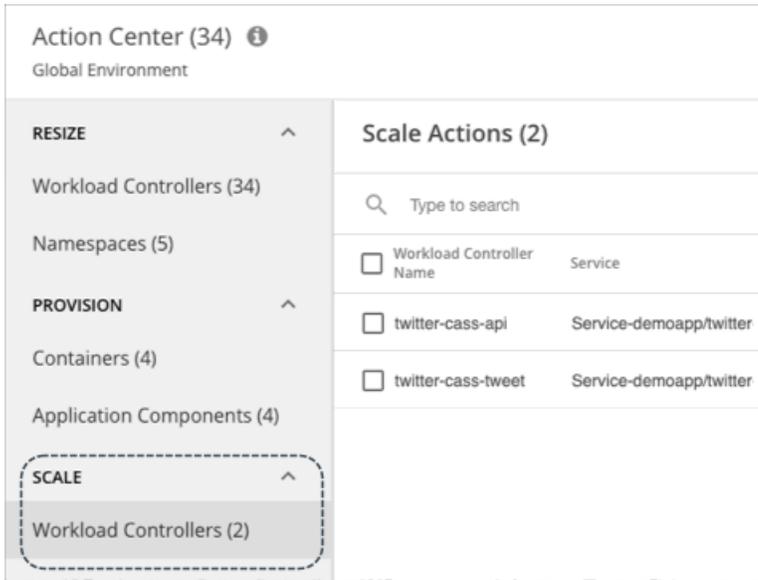
Action Generation Requirements

Intersight Workload Optimizer generates SLO-driven scale actions under the following conditions:

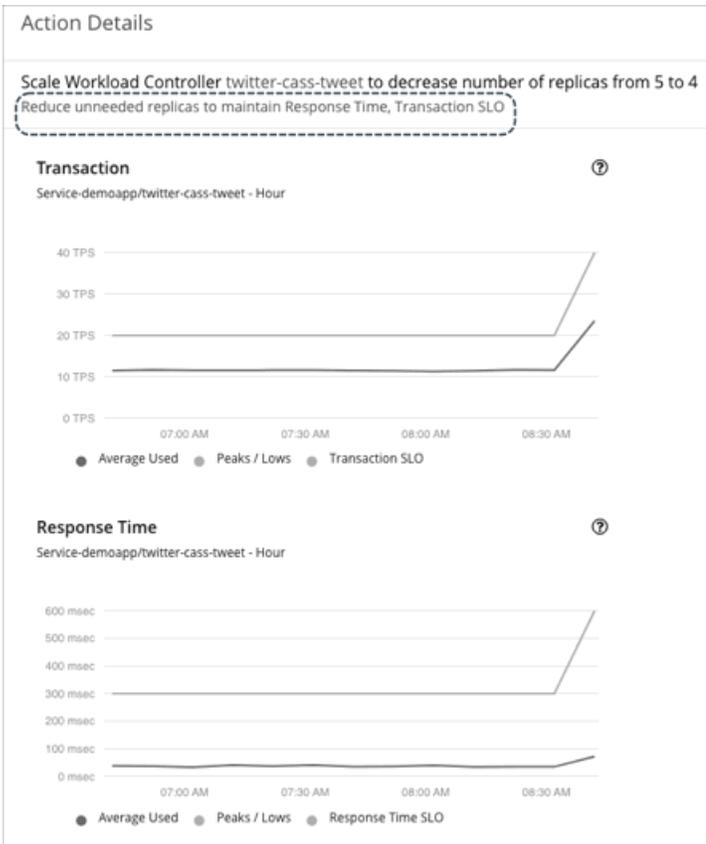
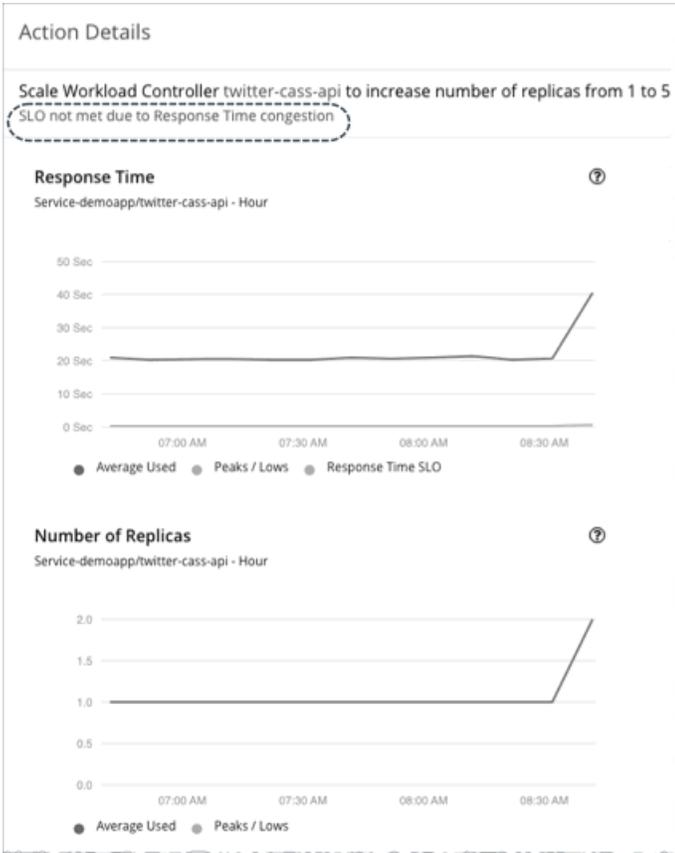
- You have created policies for the Services. You can create policies from the user interface or by using Custom Resources (CRs). For details, see this [topic \(on page 223\)](#).

Action Visibility

Intersight Workload Optimizer shows and executes SLO-driven scale actions via workload controllers. A single scale action represents the total number of replicas that you need to scale in or out to meet your SLOs.



When you examine an action, Response Time and/or Transaction SLO is indicated as the reason for the action.



Workload Controller Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

Actions associated with a workload controller resize container specs vertically or scale replicas horizontally. This is a natural representation of these actions because the parent controller's container specs and number of replicas are modified. The workload controller then rolls out the changes in the running environment.

Points to consider:

- Scale actions

Workload controller policies do not control scale actions. These actions are generated under the conditions described in this [topic \(on page 239\)](#).

- Resize actions

Container spec policies control the *generation* of resize actions, while workload controller policies control the *execution* of these actions.

| Action | Default Mode |
|--------------------------|----------------------|
| vCPU Limit Resize Down | Manual (automatable) |
| vCPU Limit Resize Up | Manual (automatable) |
| vCPU Request Resize Down | Manual (automatable) |
| vMem Limit Resize Down | Manual (automatable) |
| vMem Limit Resize Up | Manual (automatable) |
| vMem Request Resize Down | Manual (automatable) |

You can control action execution based on the specific resources that you want to resize and the resize direction. For example, for vCPU limit resizes, you may want to automate resize down actions, but require reviews of resize up actions and only execute the actions that receive an approval. To enforce these rules, create a workload controller policy and then set the action mode for vCPU Limit Resize Down to *automated*, and vCPU Limit Resize Up to *manual*.

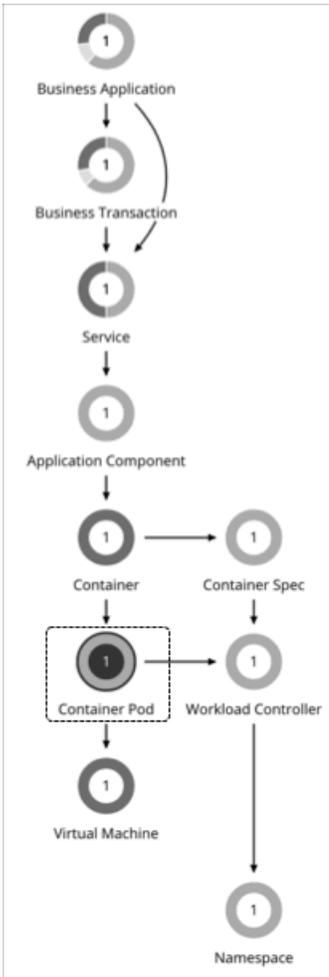
- Action orchestration is currently not supported.

Container Pod

A container pod is a group of one or more containers with shared storage or network resources. It is the smallest deployable unit of computing that you can create and manage in container platform environments.

A container pod specifies how to run containers together. Like containers, a container pod is *ephemeral*.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Resources for containers to use, including: <ul style="list-style-type: none"> ■ Virtual CPU ■ Virtual Memory |
| Consumes: | Resources from virtual machines and namespaces |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.
- VMem Request

VMem Request is the virtual memory request allocated by a pod against the node allocatable capacity.
- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.
- VCPU Request

VCPU Request is the virtual CPU request (in mCores) allocated by a pod against the node allocatable capacity.

Actions

Intersight Workload Optimizer supports the following actions:

- **Move**

Move a pod between nodes (VMs) to address performance issues or improve infrastructure efficiency. For example, if a particular node is congested for CPU, you can move pods to a node with sufficient capacity. If a node is underutilized and is a candidate for suspension, you must first move the pods before you can safely suspend the node.

- **Provision/Suspend**

When recommending node provision or suspend actions, Intersight Workload Optimizer will also recommend provisioning pods (based on demand from DaemonSets) or suspending the related pods.

Pod Move Actions

The following items impact the generation and execution of pod move actions:

- **Constraints**

Intersight Workload Optimizer respects the following constraints when making placement decisions for pods:

- Taints and tolerations are treated as constraints. For example, if a pod has a toleration attribute that restricts it from moving to a certain node, Intersight Workload Optimizer will not move that pod to the restricted node.
- Intersight Workload Optimizer imports node labels and treats them as constraints. For example, if a pod has a defined node label, Intersight Workload Optimizer will move that pod to a node with a matching label.
- Intersight Workload Optimizer recognizes pod affinity and anti-affinity policies.
- You can create placement policies to enforce constraints for pod move actions. For example, you can have a policy that allows pods to only move to certain nodes, or a policy that prevents pods from moving to certain nodes.

For more information, see [Creating Placement Policies \(on page 569\)](#).

- **Eviction Thresholds**

Intersight Workload Optimizer considers the memory/storage eviction thresholds of the destination node to ensure that the pod can be scheduled after it moves. Eviction thresholds for `imagefs` and `rootfs` are reflected as node effective capacity in the market analysis.

- **Temporary Quota Increases**

If a namespace quota is already fully utilized, Intersight Workload Optimizer temporarily increases the quota to allow a pod to move.

- **Security Context Constraints (SCCs)**

Red Hat OpenShift uses [SCCs](#) to control permissions for pods. This translates to permissions that users see within the containers of the pods, and the permissions for the processes running inside those pods.

When executing pod move actions, Kubeturbo normally runs with Red Hat OpenShift cluster administrator permissions to create a new pod and remove the old one. Because of this, the SCCs for the new pod are those that are available to a cluster administrator. It is therefore possible for the new pod to run with an SCC that has higher privileges than the old pod. For example, an old pod might have `restricted scc` access, while the new one might have `anyuid scc` access. This introduces a privilege escalation issue.

To prevent privilege escalation when moving pods, Kubeturbo enforces user impersonation, which carries the user-level SCCs of the old pod over to the new pod. To enforce user impersonation, Kubeturbo performs the following tasks:

- Create a [user impersonation account](#) for each SCC level.
- Create a [service account](#) and treat it as a user account for each SCC level currently running in a given cluster.
- Provide [role-based access](#) to SCCs used for impersonation via the service accounts. A service account is allowed to use only one SCC resource in the cluster.
- Create a `role binding` resource to allow service account access to a particular role.

All resources created to enforce user impersonation are removed when Kubeturbo shuts down.

Be aware that by default, an arbitrary pod running in a given cluster does not recognize the namespace it is configured to run in, which is a requirement for user impersonation enforcement. For Kubeturbo to recognize the namespaces for pods, it is recommended that you add an environment variable named `KUBETURBO_NAMESPACE` via the [downward API](#). Our standard installation methods add the following environment variable to the Kubeturbo deployment spec.

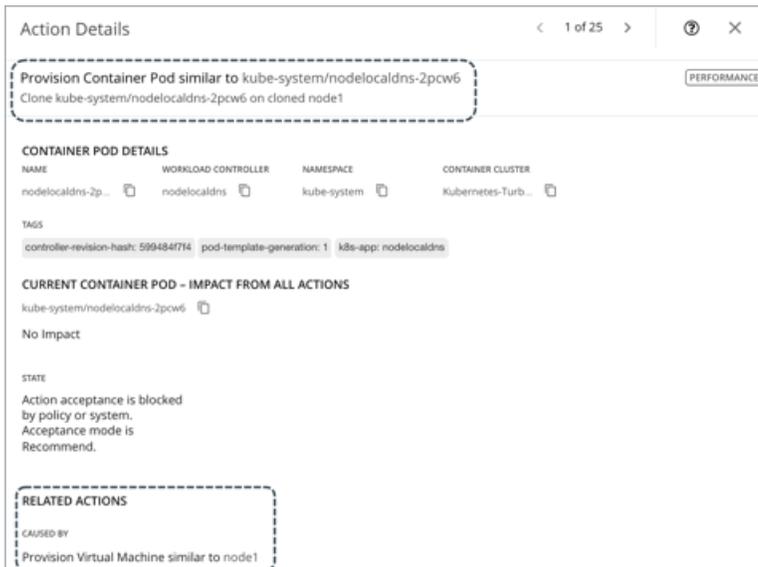
```
env:
  - name: KUBETURBO_NAMESPACE
    valueFrom:
      fieldRef:
        fieldPath: metadata.namespace
```

With this environment variable, Kubeturbo can successfully create the resources needed to enforce user impersonation. Without this variable, Kubeturbo creates the resources in the namespace called `default`. This might cause issues if you need to run multiple instances of Kubeturbo in the same cluster. For example, one instance might run as an observer, and another as an administrator. To ensure multiple Kubeturbo instances within the same cluster do not conflict when creating and removing user impersonation resources, run the instances in separate namespaces.

Pod Provision Action in Response to Node Provision

When recommending node provision actions, Intersight Workload Optimizer also recommends pod provision actions that reflect the projected demand from required DaemonSet pods, and respects the maximum number of pods allowed for a node. This ensures that any application workload can be placed on the new node and stay within the desired range of vMem/vCPU usage, vMem/vCPU request, and number of consumers.

The action details for a pod provision action shows the related node that you need to provision. Click the node name to set it at your scope.



The screenshot shows the 'Action Details' page for a pod provision action. The title is 'Provision Container Pod similar to kube-system/nodelocaldns-2pcw6' with a 'PERFORMANCE' tag. Below the title, it says 'Clone kube-system/nodelocaldns-2pcw6 on cloned node1'. The 'CONTAINER POD DETAILS' section includes a table with columns for NAME, WORKLOAD CONTROLLER, NAMESPACE, and CONTAINER CLUSTER. The table shows 'nodelocaldns-2p...' as the name, 'nodelocaldns' as the workload controller, 'kube-system' as the namespace, and 'Kubernetes-Turb...' as the container cluster. Below the table, there are tags: 'controller-revision-hash: 599484774', 'pod-template-generation: 1', and 'k8s-app: nodelocaldns'. The 'CURRENT CONTAINER POD - IMPACT FROM ALL ACTIONS' section shows 'kube-system/nodelocaldns-2pcw6' with 'No impact'. The 'STATE' section indicates 'Action acceptance is blocked by policy or system. Acceptance mode is Recommend.' The 'RELATED ACTIONS' section shows 'CAUSED BY' as 'Provision Virtual Machine similar to node1'.

Intersight Workload Optimizer treats [static pods](#) as DaemonSets for the purpose of provisioning nodes. Because a static pod provides a node with a specific capability, it is controlled by the node and is not accessible through the API server. If a node to be provisioned requires a static pod, Intersight Workload Optimizer generates actions to provision the node and the corresponding static pod.

Intersight Workload Optimizer creates an auto-generated group of static pods when it discovers a static pod on each node in a cluster. To view all the auto-generated groups, go to Search, select Groups, and then type `mirror pods` as your search keyword.

Search
Search within your infrastructure

Accounts

App Component Specs

Application Components

Billing Families

Business Applications

Business Transactions

Business Users

Chassis

Container Platform Clusters

Container Pods

Container Specs

Containers

Data Centers

Database Servers

Databases

Desktop Pools

Disk Arrays

Folders

Groups

Hosts

Search: mirror pods

| | | | NAME ↑ |
|-----------------------------------|-------------------------|----|--------|
| Mirror Pods | Kubernetes-ae-cluster-1 | 5 | Static |
| On-Prem Kubernetes-ae-cluster-1 | | | |
| Mirror Pods | Kubernetes-ae-cluster-2 | 5 | Static |
| On-Prem Kubernetes-ae-cluster-2 | | | |
| Mirror Pods | Kubernetes-DC11-PT-KBs | 12 | Static |
| On-Prem Kubernetes-DC11-PT-KBs | | | |
| Mirror Pods | Kubernetes-Hybrid | 4 | Static |
| Hybrid Kubernetes-Hybrid | | | |
| Mirror Pods | Kubernetes-OCP43-AWS | 8 | Static |
| Cloud Kubernetes-OCP43-AWS | | | |
| Mirror Pods | Kubernetes-OKD-311 | 9 | Static |
| On-Prem Kubernetes-OKD-311 | | | |
| Mirror Pods | Kubernetes-Turbonomic | 3 | Static |
| Hybrid Kubernetes-Turbonomic | | | |

Pod Suspension Action in Response to Node Suspension

When recommending node suspension actions, Intersight Workload Optimizer also recommends suspending the DaemonSet pods that are no longer required to run the suspended nodes.

The action details for a pod suspension action shows the related node that you need to suspend. Click the node name to set it at your scope.

Action Details

1 of 1

Suspend Container Pod openshift-sdn/sdn-rc94c

Suspend openshift-sdn/sdn-rc94c on suspended ocp47demo-2v5jc-worker-us-east-2a-ctc7p

CONTAINER POD DETAILS

| NAME | WORKLOAD CONTROLLER | NAMESPACE | CONTAINER CLUSTER |
|-----------|---------------------|---------------|-------------------|
| sdn-rc94c | sdn | openshift-sdn | Kubernetes-AWS... |

TAGS

app: sdn component: network controller-revision-hash: 55d797755b openshift.io/component: network pod-template-generation: 2 type: infra

CONTAINER POD STATS

openshift-sdn/sdn-rc94c

No Stats

STATE

Action acceptance is blocked by policy or system. Acceptance mode is Recommend.

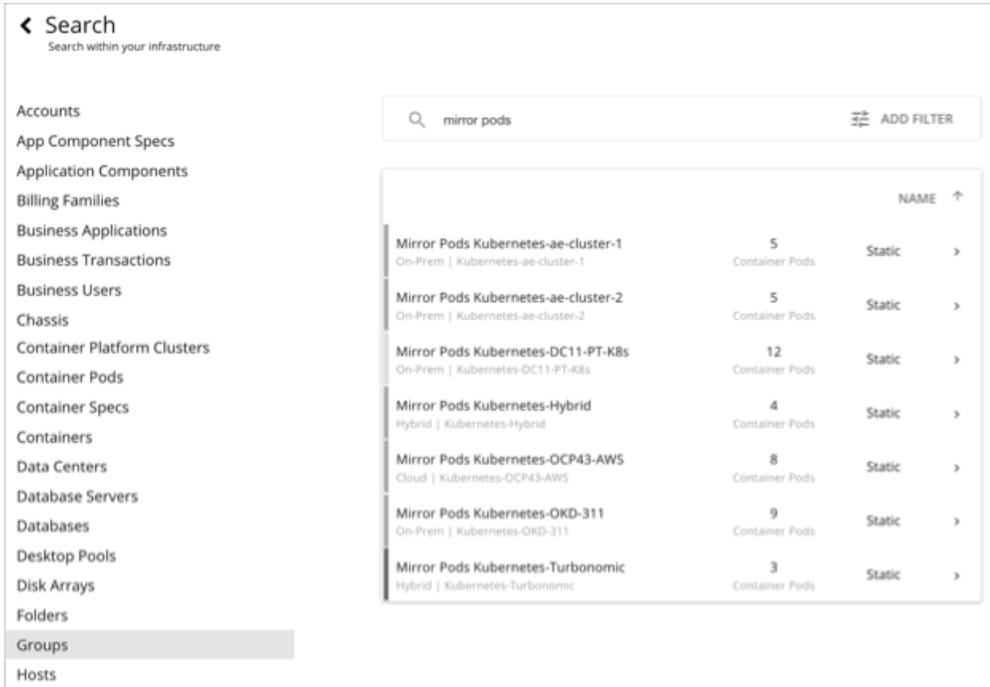
RELATED ACTIONS

CAUSED BY

Suspend Virtual Machine ocp47demo-2v5jc-worker-us-east-2a-ctc7p

Intersight Workload Optimizer treats [static pods](#) as DaemonSets for the purpose of suspending nodes. Because a static pod provides a node with a specific capability, it is controlled by the node and is not accessible through the API server. If the only workload type left on a node is a static pod, Intersight Workload Optimizer generates actions to suspend the node and the corresponding static pod.

Intersight Workload Optimizer creates an auto-generated group of static pods when it discovers a static pod on each node in a cluster. To view all the auto-generated groups, go to Search, select Groups, and then type `mirror pods` as your search keyword.



Container Pod Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about container pod actions, see [Container Pod Actions \(on page 243\)](#).

| Action | Default Mode |
|--------|--------------|
| Move | Manual |

Action orchestration is currently not supported.

Placement Policies

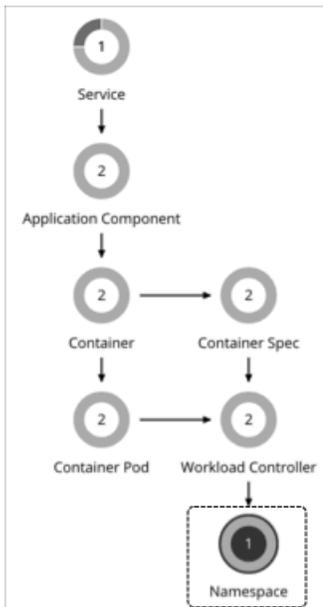
You can create placement policies to enforce constraints for pod move actions. For example, you can have a policy that allows pods to only move to certain nodes, or a policy that prevents pods from moving to certain nodes.

For more information, see [Creating Placement Policies \(on page 569\)](#).

Namespace

A namespace is a logical pool of resources in a container platform that manages workloads based on specific requirements or business needs. For example, administrators can pool resources for different organizations within the enterprise, and assign different policies to each pool.

Synopsis



| | |
|---------------------|---|
| Synopsis | |
| Provides: | N/A |
| Consumes: | N/A |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Labels and Annotations

Intersight Workload Optimizer discovers namespace labels and annotations as tag properties. You can filter namespaces by labels or annotations when you use Search or create Groups.

Monitored Resources

A namespace can include the following compute resource quotas:

- **VMem Request Quota**
VMem Request Quota is the total amount of virtual memory request for all pods allocated to the namespace against the namespace quota.
- **VCPU Request Quota**
VCPU Request Quota is the total amount of virtual CPU request (in mCores) for all pods allocated to the namespace against the namespace quota.
- **VMem Limit Quota**
VMem Limit Quota is the total amount of virtual memory limit for all pods allocated to the namespace against the namespace quota.
- **VCPU Limit Quota**
VCPU Limit Quota is the total amount of virtual CPU limit (in mCores) for all pods allocated to the namespace against the namespace quota.

When configured, these quotas define the capacity for the given namespace. Intersight Workload Optimizer monitors actual utilization of these quotas against cluster capacity.

Points to consider:

- When you scope to a namespace in the supply chain, you can see utilization data in the **Capacity and Usage** and **Namespace Multiple Resources** charts. With this data, you can understand how pods running in the namespace are consuming resources.
- The Capacity and Usage chart shows *Capacity* as the namespace quotas. *Used* values are the sum of resource limits and/or requests set for all pods in the namespace.

| Capacity and Usage | | | |
|------------------------|------------|------------|-------------|
| nsquota | | | |
| COMMODITY | CAPACITY | USED | UTILIZATION |
| Memory Request Quota ⓘ | 640 MB | 640 MB | 100% |
| CPU Limit Quota ⓘ | 500 mCores | 500 mCores | 100% |
| Memory Limit Quota ⓘ | 1.25 GB | 1.25 GB | 100% |
| CPU Request Quota ⓘ | 250 mCores | 100 mCores | 40% |
| Virtual Memory Request | 90.99 GB | 640 MB | 0.69% |

SHOW ALL >

For a namespace that does not have defined quotas, *Capacity* for the commodity is infinite (as shown in the image below). *Used* values are the sum of resource limits and/or requests set for all pods in the namespace. If these are not set, *Used* value is 0 (zero).

| Capacity and Usage | | | |
|------------------------|-----------|------------|-------------|
| openshift-sdn | | | |
| COMMODITY | CAPACITY | USED | UTILIZATION |
| Memory Request Quota ⓘ | ∞ | 4.99 GB | 0% |
| CPU Request Quota ⓘ | ∞ | 1.03 Cores | 0% |
| Memory Limit Quota ⓘ | ∞ | 0 KB | 0% |
| CPU Limit Quota ⓘ | ∞ | 0 mCores | 0% |
| Virtual Memory Request | 192.04 GB | 4.99 GB | 2.6% |

SHOW ALL >

NOTE:

If you download the data in the chart, the downloaded file shows infinite capacities as unusually large values (for example, 1,000,000,000 cores instead of the ∞ symbol).

- To see which namespaces use the most cluster resources, set the scope to a container platform cluster and see the **Top Namespaces** chart. You can use the data in the chart for showback analysis.
- When you run Optimize Container Cluster plans, Intersight Workload Optimizer can calculate increased namespace quotas in the plan results. For more information, see this [topic \(on page 427\)](#).

Namespace Actions

Intersight Workload Optimizer supports the following actions:

Resize Quota

Intersight Workload Optimizer treats quotas defined in a namespace as constraints when making resize decisions. If existing actions would exceed the namespace quotas, Intersight Workload Optimizer recommends actions to resize up the affected namespace quota.

Note that Intersight Workload Optimizer does not recommend actions to resize *down* a namespace quota. Such an action reduces the capacity that is already allocated to an application. The decision to resize down a namespace quota should include the application owner.

When you have a recommendation to resize namespace quotas, Intersight Workload Optimizer blocks execution of the resize actions for the affected Workload Containers. The action details show these blocked actions in the **Related Actions** list.

Action Details

Resize up VMem Limit Quota for Namespace `quota-test-with-down` from 3.4 GB to 4 GB
VMem Limit Congestion in Related Workload Controller

TAGS

`kubernetes.io/metadata.name: quota-test-with-down`

NAMESPACE - IMPACT FROM ALL ACTIONS

`quota-test-with-down`

| | |
|---------------------------|------------------------|
| MEMORY LIMIT QUOTA | CPU LIMIT QUOTA |
| 100 % → 84.8 % ↓ | 100 % → 88.9 % ↓ |
| 3.4 GB → 4 GB | 3.2 Cores → 3.6 Cores |

STATE

Action acceptance is blocked by policy or system.
Acceptance mode is Recommend.

RELATED ACTIONS

BLOCKING

Resize VCPU Limit,VMem Limit for Workload Controller `cpu-quota-3` in EA - Advanced Engineering
Resize VCPU Limit,VMem Limit for Workload Controller `cpu-quota-1` in EA - Advanced Engineering

NOTE:

For more information about execution of resize actions via workload controllers, see [Workload Controller Resize Actions \(on page 236\)](#).

Temporary Increases in Namespace Quotas

Intersight Workload Optimizer can increase namespace quotas temporarily under the following conditions:

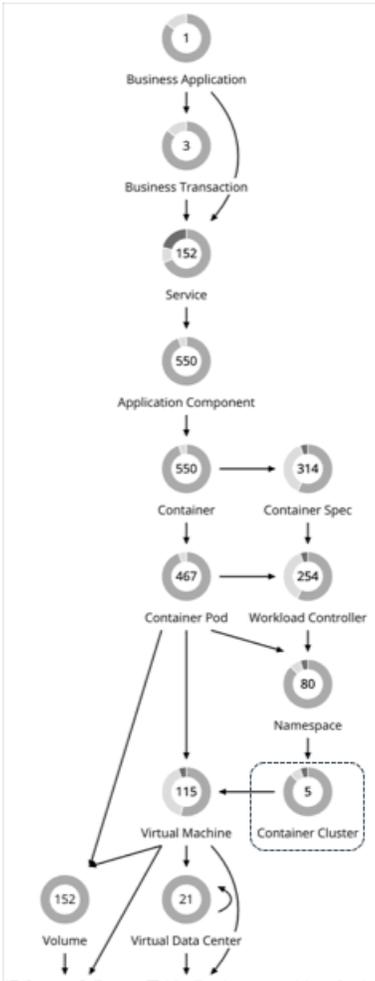
- Intersight Workload Optimizer temporarily increases the quota to allow the clone of the pod that is being moved to be scheduled.
- When executing actions to resize a workload controller in a namespace with a defined quota, Intersight Workload Optimizer can increase the namespace quota temporarily to accommodate new replicas.

When the execution of a resize action requires new replicas, a workload usually has a [rolling update](#) strategy defined. Even if the namespace quota is sufficient for the resource requirement of the new replicas, there is a chance that the quota is not enough to accommodate both the old and new replicas, as required by the rolling update. In this case, Kubeturbo calculates and then increases the quota based on the resources required by both the old and new replicas. Kubeturbo reverts the quota to its original size after the new replicas have been scheduled on a node.

Container Platform Cluster

A container platform cluster is a Kubernetes or Red Hat OpenShift cluster that Intersight Workload Optimizer discovers through KubeTurbo. With this entity type, Intersight Workload Optimizer can fully link the entire cluster with the underlying nodes, and then present all actions in a single view. This gives you full visibility into the actions that impact the health of your cluster.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | N/A |
| Consumes: | N/A |
| Discovered through: | KubeTurbo agent that you deployed to your cluster |

GitOps Integration

Intersight Workload Optimizer integrates with Git-based software through [Argo CD](#), a GitOps continuous delivery tool that manages ArgoCD Application resources using Custom Resources (CRs). With this integration, Intersight Workload Optimizer can identify the container platform cluster resources managed by Argo CD, and then execute resize actions against the Git-based version control software serving as Argo CD's 'source of truth'. Argo CD completes the action by synchronizing changes onto the running environment.

For detailed integration instructions, see this [topic \(on page 253\)](#).

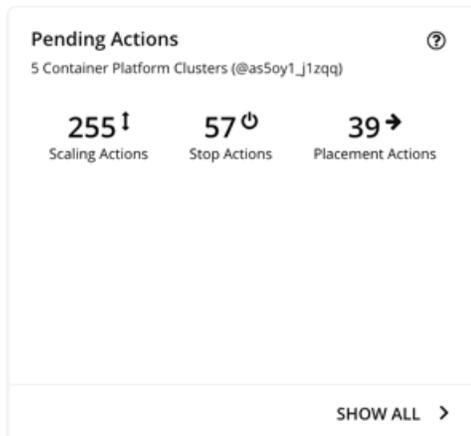
Monitored Resources

Intersight Workload Optimizer monitors resources for the containers, pods, nodes (VMs), and volumes in a cluster.

Actions

None

Intersight Workload Optimizer does not recommend actions for a container platform cluster. Instead, it recommends actions for the containers, pods, nodes (VMs), and volumes in the cluster. Intersight Workload Optimizer shows all of these actions when you scope to a container platform cluster and view the Pending Actions chart.



For actions on nodes (VMs):

- For actions to suspend or provision nodes in the public cloud, Intersight Workload Optimizer includes cost information (investments or savings) attached to those actions. Note that Intersight Workload Optimizer generates these actions *not* to optimize costs, but to assure performance and efficiency for your container infrastructure. Intersight Workload Optimizer reports costs to help you track your cloud spend.

To view cost information, set the scope to a cluster in the public cloud and view the Necessary Investments or Potential Savings charts. You can also set the scope to the global cloud environment to see total costs, or to individual container platform clusters or nodes.

- For nodes that make up an Amazon Elastic Kubernetes Service (EKS) or Azure Kubernetes Service (AKS) cluster, you can execute VM provision actions to increase the node count in a given node pool, and VM suspension actions to reduce the node count. You can manually or automatically execute these actions in Intersight Workload Optimizer if the cluster is also discovered through an Azure/AWS target in Intersight Workload Optimizer. This requires adding both a container platform target in your cluster and an Azure/AWS target where your cluster is deployed.
- For nodes that make up a Google Kubernetes Engine (GKE) cluster, you can manually execute VM provision and VM suspension actions directly in the Google Cloud console.
- Node pools and machine sets are ways to deploy and scale compute resources for container platform services hosted in the public cloud or Red Hat OpenShift on any infrastructure.

For the public cloud, Intersight Workload Optimizer uses default labels with the following patterns to discover the node pool types within each cluster:

- [AKS](#): agentpool
- [EKS](#):
//alpha.eksctl.io/nodegroup-name
eks.amazonaws.com/nodegroup
- [GKE](#): cloud.google.com/gke-nodepool

For [Red Hat OpenShift](#), Intersight Workload Optimizer creates node pools based on machine sets.

For both discovered and auto-created node pools, Intersight Workload Optimizer aggregates and visualizes actions for all the nodes in a pool to help you identify performance issues and optimization opportunities at the node pool level. Use the Top Node Pools chart to see actions and detailed information. By default, this chart displays when you set the scope to your global environment and then click the Container Platform Cluster entity in the supply chain.

Top Node Pools
Global Environment

SORT BY: POTENTIAL SAVINGS ↓

| Name | Node Count | Potential Savings ⓘ | Potential Investments ⓘ | Actions |
|---|------------|---------------------|-------------------------|------------|
| NodePool-machineset-ocp47demo-2v5jc-worker-us- | 6 | \$165.11/mo | \$165.11/mo | 26 ACTIONS |
| NodePool-eks-cluster-ng2-Kubernetes-EKS-withWin | 2 | \$8.69/mo | N/A | 1 ACTION |
| NodePool-agentpool-Kubernetes-AKS | 4 | \$0.00/mo | N/A | no actions |
| NodePool-ami-005fb2dc84caa293d-Kubernetes-EK! | 2 | \$0.00/mo | N/A | no actions |
| NodePool-ami-0a99721a12001ebd4-Kubernetes-EK | 2 | \$0.00/mo | N/A | no actions |
| NodePool-bar-Kubernetes-DC11-PT-K8s | 1 | \$0.00/mo | N/A | no actions |
| NodePool-eks-cluster-ng1-Kubernetes-EKS-withWin | 1 | \$0.00/mo | N/A | no actions |

SHOW ALL >

The chart shows the number of nodes and aggregated actions for each node pool. For node pools in the public cloud, the chart also shows the costs you would incur if you provision nodes and then scale their volumes, or the savings you would realize if you suspend nodes. To view individual actions, click the button under the Actions column. To see more details, including the full list of nodes for each pool, click the node pool name.

You can automate the execution of these actions through Intersight Workload Optimizer with Red Hat OpenShift Machine Operator. You can also manually execute node actions for AKS, EKS, or GKE through the cloud provider.

NOTE:

The following capabilities will be introduced in a future release:

- Policies for node pools
- Execution of node actions for GKE through Intersight Workload Optimizer

Cluster Health

To assess the health of each cluster, see the **Top Container Platform Clusters** chart in the predefined Container Platform Dashboard.

For each cluster, the chart shows the sum of resources used by containers and the underlying nodes. Click the **Actions** button to see a list of pending actions.

Top Container Platform Clusters
Global Environment

SORT BY: HEALTH ↓

| Container Cluster | Health | Virtual CPU Used | Virtual CPU Request | Virtual Memory Used | Virtual Memory Request | Actions |
|------------------------|--------|-------------------------|--------------------------|-----------------------|------------------------|-------------|
| Kubernetes-PT-AKS | | 8.2 Cores (41%) ↓ 0.31% | 13.01 Cores (68%) | 20.58 GB (28%) ↓ 0% | 21.66 GB (40%) | 60 ACTIONS |
| Kubernetes-DC11-PT-K8s | | 6.78 Cores (28%) ↓ 3% | 11.46 Cores (50%) | 24.67 GB (26%) ↑ 0.1% | 12.27 GB (13%) | 139 ACTIONS |
| Kubernetes-OCP47-AWS | | 13.36 Cores (26%) ↑ 2% | 12.87 Cores (27%) ↑ 3% | 90.84 GB (45%) ↓ 13% | 72.4 GB (38%) ↑ 0.69% | 217 ACTIONS |
| Kubernetes-OKD-311 | | 4.07 Cores (16%) ↓ 9% | 8.86 Cores (35%) ↑ 1% | 27.94 GB (30%) ↑ 23% | 21.35 GB (23%) ↑ 0.92% | 119 ACTIONS |
| Kubernetes-ocp-wdc02 | | 10.02 Cores (21%) ↓ 17% | 10.6 Cores (22%) ↓ 0.92% | 208.77 GB (55%) ↓ 24% | 103.17 GB (30%) ↑ 2% | 319 ACTIONS |
| Kubernetes-OCP43-AWS | | 3.32 Cores (21%) ↑ 17% | 7.41 Cores (51%) ↓ 0% | 23.72 GB (38%) ↑ 2% | 15.53 GB (26%) ↓ 0% | 16 ACTIONS |
| Kubernetes-Turbonomic | | 5.11 Cores (64%) ↑ 27% | 0.95 Cores (12%) | 61.83 GB (49%) ↑ 5% | 18.03 GB (14%) | 73 ACTIONS |

SHOW ALL >

The **Top Namespaces** chart shows the namespaces that use the most cluster resources. You can use the data in the chart for showback analysis.

Top Namespaces
Global Environment

SORT BY: HEALTH ↓

| Namespace | H. | Container Cluster | Virtual CPU Used | CPU Request Quota | CPU Limit Quota | Virtual Memory Used | Memory Request Quota | Memory Limit Quota | Actions |
|-----------------------|----|-------------------|-------------------------|-------------------|------------------|------------------------|----------------------|--------------------|------------|
| demoapp | | OKD-311 | 152 mCores (1%) ↓ 0.19% | 100 mCores (0%) | 200 mCores (0%) | 1.91 MB (0%) | 10 MB (0%) | 20 MB (0%) | 72 ACTIONS |
| action-merge-test | | OKD-311 | 16 mCores (0%) ↓ 5% | 100 mCores (0%) | 200 mCores (0%) | 153.23 MB (0%) ↑ 0.08% | 260 MB (0%) | 400 MB (0%) | 75 ACTIONS |
| action-merge-test2 | | OKD-311 | 202 mCores (1%) ↓ 0.07% | 90 mCores (0%) | 202 mCores (0%) | 2.2 MB (0%) ↑ 4% | 260 MB (0%) | 400 MB (0%) | 78 ACTIONS |
| turbo-operator-arsen | | OCP47-AWS | 92 mCores (0%) ↑ 279% | 200 mCores (0%) | 1 Cores (0%) | 202.93 MB (0%) ↑ 6% | 512 MB (0%) | 1 GB (0%) | 5 ACTIONS |
| provelt | | DC11-PT-K8s | 442 mCores (2%) ↓ 0.09% | 606 mCores (15%) | 5.78 Cores (72%) | 585.42 MB (1%) | 2.44 GB (24%) | 20.88 GB (42%) | 84 ACTIONS |
| instana-agent | | OCP47-AWS | 1.27 Cores (2%) ↑ 1% | 2.5 Cores (0%) | 7.5 Cores (0%) | 2.58 GB (1%) ↓ 4% | 2.5 GB (0%) | 2.5 GB (0%) | 17 ACTIONS |
| aqiqui-cpu-throttling | | DC11-PT-K8s | 757 mCores (3%) ↑ 0.18% | 600 mCores (0%) | 1.35 Cores (0%) | 23.71 MB (0%) | 50 MB (0%) | 100 MB (0%) | 78 ACTIONS |

SHOW ALL >

GitOps Integration

Intersight Workload Optimizer integrates with Git-based software through [Argo CD](#), a GitOps continuous delivery tool that manages ArgoCD Application resources using Custom Resources (CRs). With this integration, Intersight Workload Optimizer can identify the container platform cluster resources managed by Argo CD, and then execute resize actions against the Git-based version control software serving as Argo CD's 'source of truth'. Argo CD completes the action by synchronizing changes onto the running environment.

Intersight Workload Optimizer currently supports the following definitions:

| Workload Definition - Path to Version Control Software | Version Control Software (Git Source) | Application Packaging |
|--|---------------------------------------|---|
| ArgoCD Application CR in the cluster | GitHub or GitLab | None (straight YAMLS only). Operators and Helm are not supported. |
| ArgoCD ApplicationSet CR in the cluster | | |

Configurations

It is assumed that Argo CD is installed and active in the cluster managed by Intersight Workload Optimizer. Although Intersight Workload Optimizer will be able to discover most of the details from the Argo CD applications created to manage the resources, some details, like Git credentials will still need to be configured in Kubeturbo running in the same cluster.

- Create Developer Access Token in Git
 1. In GitHub under **Settings**, select **Developer Settings**.
 2. Select **Personal Access Tokens**.
 3. Generate a new token and grant the following permissions:
 - delete:packages
 - gist
 - notifications
 - repo
 - write:discussion
 - write:packages
 4. Copy the token, as you will need it in the next step when creating the secret.
- Create Secret to store Access Token
 1. Create a new secret in the same namespace that Kubeturbo is deployed in.
 2. Your key is `token`, value is access token created above in base64, and type is `Opaque`. See example:

```
kind: Secret
apiVersion: v1
metadata:
  name: github
  namespace: turbo
data:
  token: Z2hwX0xEdkFvcUZ0ZkabcDEfa053cGVwWXd2SDFkS2V1MQ==
type: Opaque
```

- Kubeturbo deployed with plane YAML

The options to be configured in the Kubeturbo deployment, if deployed using plane YAML, are listed below. These configurations will be globally used for all Argo CD apps deployed in that cluster.

| | |
|--|---|
| <code>--git-email</code> string | The email to be used to push changes to git. |
| <code>--git-secret-name</code> string | The name of the secret which holds the git credentials. |
| <code>--git-secret-namespace</code> string | The namespace of the secret which holds the git credentials. |
| <code>--git-username</code> string | The user name to be used to push changes to git. |
| <code>--git-commit-mode</code> | The commit mode that should be used for git action executions. One of <code>request direct</code> . Defaults to <code>direct</code> . |

The following example is a YAML file to elaborate the `args` section:

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: kubeturbo-test
  ....
  ....
spec:
  replicas: 1
  selector:
    matchLabels:
      app: kubeturbo-test
```

```

template:
  metadata:
    annotations:
      kubeturbo.io/monitored: "false"
    labels:
      app: kubeturbo-test
  spec:
    containers:
      - args:
          - --turboconfig=/etc/kubeturbo/turbo.config
          - --v=2
          - --kubelet-https=true
          - --kubelet-port=10237
          - --git-email=youremail@gmail.com
          - --git-secret-name=your-github-secret-name
          - --git-secret-namespace=turbo
          - --git-username=yourusername
          - --git-commit-mode=direct
        image: turbonomic/kubeturbo:latest
        imagePullPolicy: Always
        name: kubeturbo-test
        resources: {}
    ...
    ...
  
```

■ Kubeturbo deployed with Operator

Below options will need to be configured in the operator CR spec. These configurations will be globally used for all Argo CD apps deployed in that cluster.

```

apiVersion: charts.helm.k8s.io/v1
kind: Kubeturbo
metadata:
  name: kubeturbo-sample
spec:
  HANodeConfig:
    nodeRoles: '"master"'
  args:
    kubelethttps: true
    kubeletport: 10250
    ...
    ...
  # [ArgoCD integration] The email to be used to push changes to git.
  gitEmail: "youremail@gmail.com"
  # [ArgoCD integration] The username to be used to push changes to git.
  gitUsername: "yourusername"
  # [ArgoCD integration] The name of the secret which holds the git credentials.
  gitSecretName: "your-github-secret-name"
  # [ArgoCD integration] The namespace of the secret which holds the git credentials.
  gitSecretNamespace: "turbo"
  # [ArgoCD integration] The commit mode that should be used for git action executions. One of {request|direct}. Defaults to direct.
  gitCommitMode: "direct"
  
```

(Optional) Fine Grained App Specific Configurations

It is possible that Argo CD apps syncing from different sources of truth (for example different GitHub repositories) exist in the same cluster. It is also possible that the credential information for different source repositories are different. The app specific fine grained configuration provides a mechanism to have different configuration information for different apps. It is achieved via a config custom resource. If using plane old YAML mechanism the user will need to first install the CRD from [here](#). The CRD can be deployed using the following command:

```
kubectl apply -f https://raw.githubusercontent.com/turbonomic/turbo-gitops/main/config/crd/bases/gitops.turbonomic.io_gitops.yaml
```

Once the type definition is available, a configuration could be used as provided in the sample below:

```
apiVersion: gitops.turbonomic.io/v1alpha1
kind: GitOps
metadata:
  name: gitops-sample
spec:
  config:
    - commitMode: direct
      credentials:
        email: test@turbonomic.com
        secretName: gitops-secret
        secretNamespace: gitops
        username: turbo
        selector: '^turbo.*$'
        whitelist:
          - app-name-1
          - app-name-2
```

This resource is name spaced and a single, or multiple non conflicting configurations could be supplied per namespace. If an app is selected as per the `whitelist` or the `selector` then the `credentials` and `commitMode` listed in the `GitOps` resource spec will be used instead of the global ones configured in `Kubeturbo`.

(Optional) Cluster Local Resource Update

It might be desirable to have Intersight Workload Optimizer update the resources locally and directly in the cluster, rather than pushing the changes back into the source of truth. There can be multiple reasons for choosing this.

- Security

Users want to minimize the credential and the source of truth access.

- Faster optimization

The turn around time for the resource to be updated would be bigger if the update has to be made back in the source of truth, because of the number of systems involved in the same, and possibility of failures and retries. Intersight Workload Optimizer should be able to push a commit or a PR back to the Git repository. If it is a PR, a user needs to approve the same.

Argo CD even if running in auto sync mode should be able to pull and sync the resource into the cluster. Instead, a resource could be updated locally in the cluster.

NOTE:

Argo CD has a default discovery interval of 3 minutes to discover changes in the Git source. Changes in Intersight Workload Optimizer can take up to 10 minutes or longer to reflect the changes.

- Multi cluster scenarios

It is possible that a unified app definition is used by multiple clusters and all of them syncing from a base definition. Optimizations from one cluster might not apply to another cluster. If the updates are made back to the source of truth, then the change applies to all the clusters.

- Newer or unsupported tooling in the pipeline

It is possible that Intersight Workload Optimizer does not have the capability to either understand the pipeline or directly interact with the GitOps system.

To use this pattern Argo CD app could be configured to skip sync of particular fields from the child resources.

`spec.ignoreDifferences` allows to not trigger an out of sync when the cluster local value changes, for details see the Argo CD documentation [here](#).

`RespectIgnoreDifferences` allows to ignore copying over the original values into the cluster local resources, in case the sync happens because of other field changes as described in Argo CD documentation [here](#).

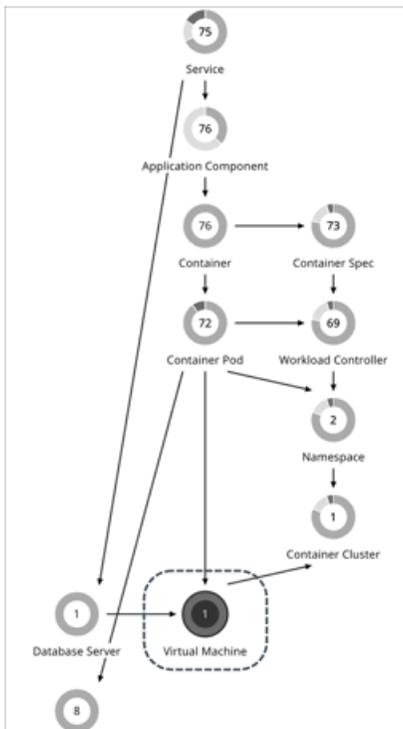
To utilize this pattern both options need to be used in the Argo CD applications.

Virtual Machine (Container Platform Node)

In container platform environments, a node is a virtual or physical machine that contains the services necessary to run pods. Intersight Workload Optimizer represents nodes as Virtual Machine entities in the supply chain.

Intersight Workload Optimizer can discover node roles and Master Nodes. It creates policies to keep nodes of the same role on unique host or Availability Zone providers, and policies to disable suspension of Master Nodes. Intersight Workload Optimizer also discovers and displays Node Pools and Red Hat OpenShift Machine Sets.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Resources to pods |
| Consumes: | Resources from container platform clusters |
| Discovered through: | Kubeturbo agent that you deployed to your cluster |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE: These are resources for nodes that host pods. They are monitored along with the resources from the infrastructure probes, such as vCenter or a public cloud mediation probe.

- **VMem**
VMem is the virtual memory currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- **VCPU**
VCPU is the virtual CPU currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- **Memory Request Allocation**
Memory Request Allocation is the memory available to the node to support the ResourceQuota request parameter for a given Kubernetes namespace or Red Hat OpenShift project.
- **CPU Request Allocation**
CPU Request Allocation is the CPU available to the node to support the ResourceQuota request parameter for a given Kubernetes namespace or Red Hat OpenShift project.
- **Virtual Memory Request**
Virtual Memory Request is the memory currently guaranteed by all containers on the node with a memory request. The capacity for this resource is the Node Allocatable capacity, which is the amount of resources available for pods and can be less than the physical capacity.
- **Virtual CPU Request**
Virtual CPU Request is the CPU currently guaranteed by all containers on the node with a CPU request. The capacity for this resource is the Node Allocatable capacity, which is the amount of resources available for pods and can be less than the physical capacity.
- **Memory Allocation**
Memory Allocation is the memory ResourceQuota limit parameter for a given Kubernetes namespace or Red Hat OpenShift project.
- **CPU Allocation**
CPU Allocation is the CPU ResourceQuota limit parameter for a given Kubernetes namespace or Red Hat OpenShift project.

Actions

Intersight Workload Optimizer supports the following actions:

- **Provision**
Provision nodes to address workload congestion or meet application demand.
- **Suspend**
Suspend nodes after you have consolidated pods or defragmented node resources to improve infrastructure efficiency.
- **Reconfigure**
Reconfigure nodes that are currently in the `NotReady` state.

NOTE:

For nodes in the public cloud, Intersight Workload Optimizer reports the cost savings or investments attached to node and provision actions. For example, you can see the additional costs you would incur if you provision nodes and then scale their volumes, or the savings you would realize if you suspend nodes. Note that performance and efficiency are the drivers of these actions, *not* cost. Cost information is included to help you track your cloud spend. For this reason, you will *not* see cost-optimization actions, including recommendations to re-allocate discounts or delete unattached volumes.

To view cost information, set the scope to a node and see the Necessary Investments and Potential Savings charts. You can also set the scope to a [container platform cluster \(on page 250\)](#) or the global cloud environment to view aggregated cost information.

Pre-execution Check for Node Provision and Suspension Actions

Before executing a node provision or suspension action for a MachineSet in a Red Hat OpenShift cluster, Kubeturbo checks the node count range specified in your `ConfigMap`. If the pre-execution check determines that the node count will fall outside the

range after action execution, the action executes but fails, and the failure is logged (for example, in the Executed Actions chart in the Intersight Workload Optimizer user interface). This mechanism ensures the overall stability and performance of the Red Hat OpenShift cluster.

NOTE:

The pre-execution check does not apply to AKS, EKS, and GKE node pools.

By default, the minimum node count is 1 and the maximum is 1000. You can customize these values by updating the `nodePoolSize` parameter in your Kubeturbo `ConfigMap`. This update does not require a restart of the Kubeturbo pod and takes effect after approximately one minute.

For a sample `ConfigMap` with the `nodePoolSize` parameter, see the Kubeturbo [GitHub repository](#).

Node Provision Actions

When recommending node provision actions, Intersight Workload Optimizer also recommends pod provision actions that reflect the projected demand from required `DaemonSet` pods, and respects the maximum number of pods allowed for a node. This ensures that any application workload can be placed on the new node and stay within the desired range of vMem/vCPU usage, vMem/vCPU request, and number of consumers.

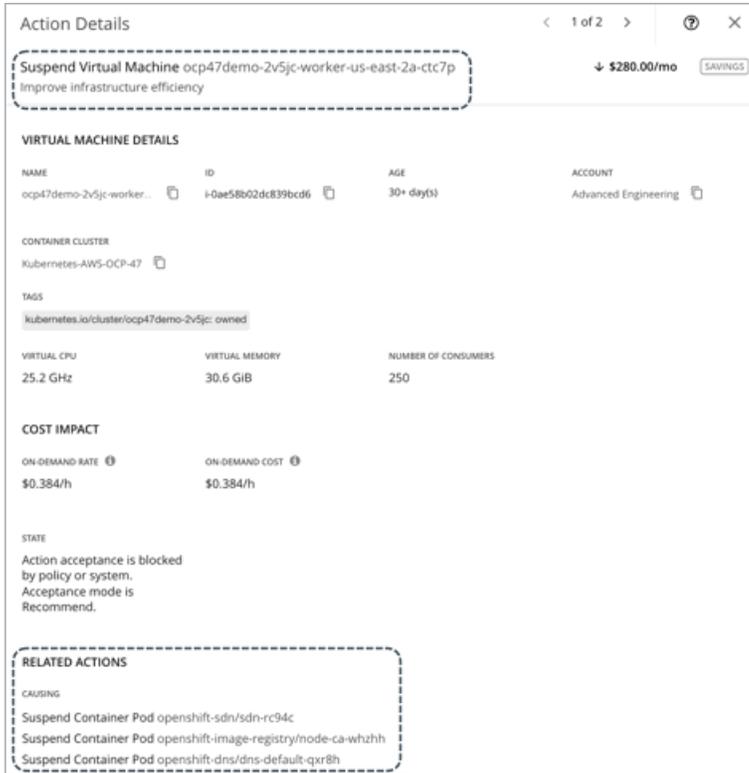
The action details for a node provision action show the related `DaemonSet` pods that are required for the node to run. Click a pod name to set it at your scope.

Intersight Workload Optimizer treats [static pods](#) as `DaemonSets` for the purpose of provisioning nodes. Because a static pod provides a node with a specific capability, it is controlled by the node and is not accessible through the API server. If a node to be provisioned requires a static pod, Intersight Workload Optimizer generates actions to provision the node and the corresponding static pod.

Node Suspension Actions

When recommending node suspension actions, Intersight Workload Optimizer also recommends suspending the `DaemonSet` pods that are no longer required to run the suspended nodes.

The action details for a node suspension action show the related `DaemonSet` pods that are no longer needed to run the suspended nodes. Click a pod name to set it at your scope.



Action Details < 1 of 2 > ? X

Suspend Virtual Machine ocp47demo-2v5jc-worker-us-east-2a-ctc7p
Improve infrastructure efficiency ↓ \$280.00/mo SAVINGS

VIRTUAL MACHINE DETAILS

| NAME | ID | AGE | ACCOUNT |
|---------------------------|--------------------|------------|----------------------|
| ocp47demo-2v5jc-worker... | f0ae58b02dc839bcd6 | 30+ day(s) | Advanced Engineering |

CONTAINER CLUSTER
Kubernetes-AWS-OCP-47

TAGS
kubernetes.io/cluster/ocp47demo-2v5jc: owned

| VIRTUAL CPU | VIRTUAL MEMORY | NUMBER OF CONSUMERS |
|-------------|----------------|---------------------|
| 25.2 GHz | 30.6 GiB | 250 |

COST IMPACT

| ON-DEMAND RATE | ON-DEMAND COST |
|----------------|----------------|
| \$0.384/h | \$0.384/h |

STATE
Action acceptance is blocked by policy or system. Acceptance mode is Recommend.

RELATED ACTIONS

CAUSING

- Suspend Container Pod openshift-sdn/sdn-rc94c
- Suspend Container Pod openshift-image-registry/node-ca-whzhh
- Suspend Container Pod openshift-dns/default-qxr8h

Intersight Workload Optimizer treats [static pods](#) as DaemonSets for the purpose of suspending nodes. Because a static pod provides a node with a specific capability, it is controlled by the node and is not accessible through the API server. If the only workload type left on a node is a static pod, Intersight Workload Optimizer generates actions to suspend the node and the corresponding static pod.

Node Reconfigure Actions

Intersight Workload Optimizer generates node reconfigure actions to notify you of nodes that are currently in the `NotReady` state.

A reconfigure action is read-only in Intersight Workload Optimizer and must be executed directly in the container platform cluster because the action might require restarting the node or the kubelet on the node. When Intersight Workload Optimizer discovers that a node's state is `Ready`, it removes the reconfigure action automatically and begins to monitor the health of the node and the associated container pods.

NOTE:

Intersight Workload Optimizer treats a node as a VM under certain circumstances. For example, it treats a node in vCenter as a VM that can move to a different host if the current host is congested. This means that for a `NotReady` node in vCenter, it is possible to see a VM move action along with the expected node reconfigure action. Both actions are valid and safe to execute since they achieve two different and non-conflicting results.

For each container platform cluster, Intersight Workload Optimizer creates an auto-generated group of `NotReady` nodes. To view all the auto-generated groups, go to Search, select Groups, and then type `notready` as your search keyword. Click a group to view the individual nodes and the pending reconfigure actions.

← Search
Search within your infrastructure

- Accounts
- App Component Specs
- Application Components
- Billing Families
- Business Applications
- Business Transactions
- Business Users
- Chassis
- Container Platform Clusters
- Container Pods
- Container Specs
- Containers
- Data Centers
- Database Servers
- Databases
- Desktop Pools
- Disk Arrays
- Folders
- Groups
- Hosts

ADD FILTER

| | | | NAME ↑ |
|--|---|------------------|----------|
| NotReady Nodes [Kubernetes-ae-cluster-1] <small>On-Prem Kubernetes-ae-cluster-1</small> | 2 | Virtual Machines | Static > |
| NotReady Nodes [Kubernetes-Hybrid] <small>Hybrid Kubernetes-Hybrid</small> | 2 | Virtual Machines | Static > |

When you examine a pending reconfigure action, you can click the link in the 'Entities Impacted by this Node' section to view a list of impacted pods.

Action Details

Reconfigure Virtual Machine ae-cluster1-node-group-afdd79db90
 The node is in a NotReady status

VIRTUAL MACHINE DETAILS

NAME
 ae-cluster1-node-group-afdd79db90 📄

| | |
|------------------------|------------------------|
| VMEM PERCENTILE | VCPU PERCENTILE |
| 31 % | 11 % |
| 8 GB | 4.8 GHz |

STATE
 Action acceptance is blocked by policy or system. Acceptance mode is Recommend.

ENTITIES IMPACTED BY THIS NODE

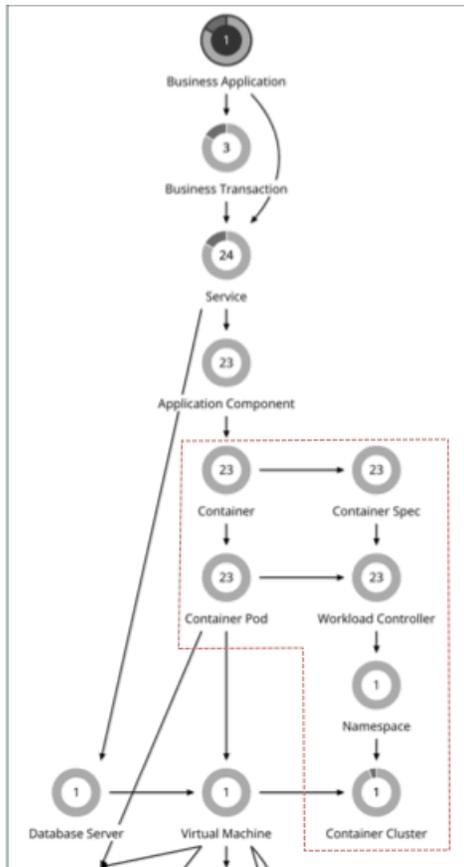
Reconfiguring this VM might activate the container pods that are currently in an unknown state.

[View List of Container Pods in Unknown State](#)

These pods are in the `Unknown` state and are not controllable. In the supply chain and in the list of container pods, these pods display with a gray color to help you differentiate them from other pods.

Container Platform CPU Metrics

To meet user requirements and align with container platform specifications, Intersight Workload Optimizer uses millicore (mCore) as the base unit for CPU metrics for your container platform.



These include metrics for the following CPU-related commodities:

- vCPU
- vCPU Request
- vCPU Limit Quota
- vCPU Request Quota
- vCPU Throttling

Intersight Workload Optimizer displays these commodities in charts, actions, policies, and plans. For example:

- In the Capacity and Usage chart for container platform entities, *capacity* and *used* values for CPU-related commodities are shown in mCores.
- In the supply chain, when you scope to a Workload Controller to view pending [resize actions \(on page 228\)](#) for a container, you will see utilization and resize values in mCores.
- When you create [container spec policies \(on page 230\)](#), resize thresholds and increment constants for CPU-related commodities are set in mCores.
- For an Optimize Container Cluster plan, the [plan results \(on page 430\)](#) for CPU-related commodities are shown in mCores.

For nodes (VMs) and Application Components:

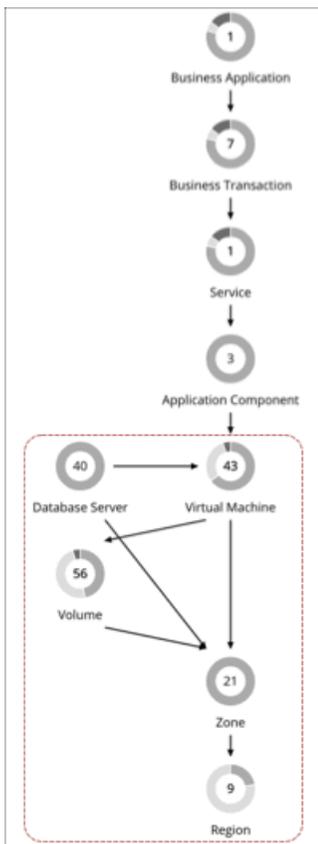
- For nodes stitched to your container platform, the base unit for *vCPU Request* is also mCore, since this commodity is provided only to the container platform.
- For both nodes and Application Components (standalone or stitched to your container platform), the base unit for *vCPU* is MHz, since this is a generic commodity. For example, when you view a pod move action, vCPU metrics for the current and destination nodes for the pod are expressed in MHz.

The following table summarizes the base units of CPU measurement that Intersight Workload Optimizer uses.

| Entity | CPU Commodity | | | | |
|----------------------------|---------------|--------------|------------------|--------------------|-----------------|
| | vCPU | vCPU Request | vCPU Limit Quota | vCPU Request Quota | vCPU Throttling |
| Container | mCore | mCore | mCore | mCore | mCore |
| Container spec | mCore | mCore | N/A | N/A | mCore |
| Workload Controller | N/A | N/A | mCore | mCore | N/A |
| Container Pod | mCore | mCore | mCore | mCore | N/A |
| Namespace | mCore | mCore | mCore | mCore | N/A |
| Container Platform Cluster | mCore | mCore | N/A | N/A | N/A |
| Node (VM) | MHz | mCore | N/A | N/A | N/A |
| Application Component | MHz | N/A | N/A | N/A | N/A |

Entity Types - Cloud Infrastructure

Intersight Workload Optimizer discovers and monitors the entities that make up your cloud infrastructure, and recommends actions to assure application performance at the lowest possible cost.



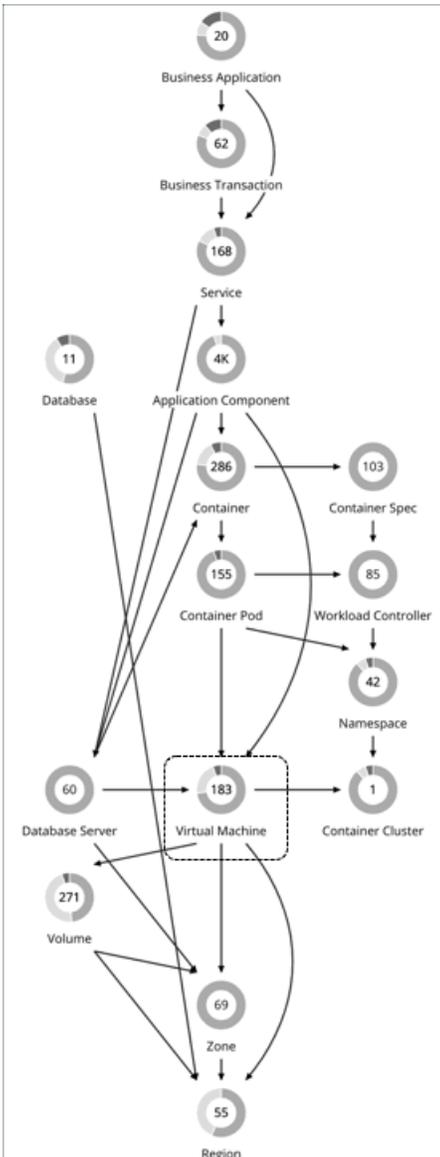
Virtual Machine (Cloud)

A virtual machine (VM) is a software emulation of a physical machine, including OS, virtual memory and CPUs, and network ports. VMs host applications, or they provide resources to container platforms.

NOTE:

Container platform nodes are also represented as Virtual Machines in the Intersight Workload Optimizer supply chain. For details, see this [topic \(on page 257\)](#).

Synopsis



| | |
|---------------------|---------------------------------------|
| Synopsis | |
| Provides: | Resources for applications to use |
| Consumes: | Resources from cloud zones or regions |
| Discovered through: | Public cloud targets |

Monitored Resources

For details about monitored resources for cloud VMs, see the following topics:

- [AWS Monitored Resources \(on page 52\)](#)
- [Azure Monitored Resources \(on page 101\)](#)
- [Google Cloud Monitored Resources \(on page 73\)](#)

Cloud VM Actions

For details about cloud VM actions, see the following topics:

- [Actions for AWS VMs \(on page 265\)](#)
- [Actions for Azure VMs \(on page 268\)](#)
- [Actions for Google Cloud VMs \(on page 269\)](#)

Actions for AWS VMs

Intersight Workload Optimizer supports the following actions:

- **Scale**

Change the VM instance to use a different instance type or tier to optimize performance and costs.

See additional information for scale actions below.
- **Discount-related actions**

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing RI coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

If you need more capacity, then Intersight Workload Optimizer will recommend actions to purchase additional RIs.

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Controlling Scale Actions for AWS VMs

For scale actions, you can create policies to control the scale actions that Intersight Workload Optimizer recommends. In those policies, choose from the following options:

- Cloud Scale All – execute all scaling actions
- Cloud Scale for Performance – only execute scaling actions that improve performance
- Cloud Scale for Savings – only execute scaling actions that reduce costs

The default action acceptance mode for these actions is *Manual*. When you examine the pending actions, only actions that satisfy the policy are allowed to execute. All other actions are read-only.

When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see [Default and User-defined Automation Policies \(on page 574\)](#).

Supported Instance Types for AWS VMs

Intersight Workload Optimizer considers all supported instance types when making scaling decisions for cloud VMs. If you want your VMs to *only scale to* or *avoid* certain instance types, create policies for those VMs.

You can view the instance types that Intersight Workload Optimizer supports from the user interface.

1. Navigate to **More > Settings > Policies**.
2. In the Policy Management page, search for and click **Virtual Machine Defaults**.
3. In the Configure Virtual Machine Policy page:
 - a. Scroll down to the bottom of the page.
 - b. Click **Add Scaling Constraint**.
 - c. Choose **Cloud Instance Types**.
 - d. Click **Edit**.

The policy page shows supported tiers for each cloud provider. A tier is a family of instance types, such as *M1* for Google Cloud, *a1* for AWS and *Basic_A1* for Azure. Expand a tier to see individual instance types and resource allocations.

Support for AWS EC2 GPU Instance Types

Intersight Workload Optimizer discovers NVIDIA [GPU metrics](#) for supported AWS EC2 instance types and uses these metrics to generate VM scale actions. Metrics include the number of utilized GPU cards and the amount of GPU memory in use. To collect GPU metrics from CloudWatch, be sure to configure CloudWatch as described in this [topic \(on page 55\)](#).

Currently, Intersight Workload Optimizer supports the following instance types with Linux AMIs.

- P2 instance family (based on NVIDIA Kepler K80 GPUs)
- P3/P3dn instance family (based on NVIDIA Volta V100 GPUs)
- G3 instance family (based on NVIDIA Tesla M60 GPUs)
- G4dn instance family (based on NVIDIA T4 GPUs)
- G5 instance family (based on NVIDIA A10G Tensor Core GPUs)
- G5g instance family (based on NVIDIA T4G Tensor Core GPUs)

To optimize performance and costs, Intersight Workload Optimizer can recommend actions that scale down the number of GPU cards, or scale GPU memory up or down.

Intersight Workload Optimizer can also recommend actions that scale standard VM resources (such as vCPU and vMem) to the supported GPU instance types and the G4ad instance family (based on AMD Radeon Pro V520 GPUs).

Intersight Workload Optimizer creates the appropriate read-only tier exclusion policies and displays them in the Policy Management page (**Settings > Policies**).

- Cross-target policies

Cross-target policies ensure that AWS VMs with certain GPU types only scale to an instance type with the same GPU configuration (card count and memory per card). Policies include:

- AWS GPU - NVIDIA M60 - Cloud Compute Tier Exclusion Policy
- AWS GPU - NVIDIA T4 - Cloud Compute Tier Exclusion Policy

- Per-target policies

Per-target policies ensure that any VMs in GPU supported instance families (G4dn, G4ad, G3, G3s, G5, G5g currently) do not scale out to instance families that do not support GPUs. The following policy is available.

Cloud Compute Tier AWS:gpu - Cloud Compute Tier Exclusion Policy

Support for AWS EC2 Accelerator Instance Types

Intersight Workload Optimizer can recommend actions to scale standard VM resources (such as vCPU and vMem) to the following AWS EC2 Accelerator instance types.

- Inf1 instance family (based on AWS Inferentia chips)
- Inf2 instance family (based on AWS Inferentia2 chips)

Intersight Workload Optimizer also creates the appropriate read-only tier exclusion policies and displays them in the Policy Management page (**Settings > Policies**).

- Cross-target policies

Cross-target policies ensure that AWS VMs with certain Accelerator types only scale to an instance type with the same Accelerator configuration (card count and memory per card). Policies include:

- AWS ML_ACCELERATOR - Inferentia1 - Cloud Compute Tier Exclusion Policy
- AWS ML_ACCELERATOR - Inferentia2 - Cloud Compute Tier Exclusion Policy

- Per-target policies

Per-target policies ensure that any VMs in Inferentia1 instance families do not scale out to other instance families. Policies include:

- Cloud Compute Tier AWS:inf1 - Cloud Compute Tier Exclusion Policy
- Cloud Compute Tier AWS:inf2 - Cloud Compute Tier Exclusion Policy

Scaling Prerequisites for AWS VMs

In AWS some instances require VMs to be configured in specific ways before they can scale to those instance types. If Intersight Workload Optimizer recommends scaling a VM that is not suitably configured onto one of these instances, then it sets

the action to *Recommend*, and describes the reason. Intersight Workload Optimizer will not automate the action, even if you have set the action acceptance mode for that scope to *Automatic*. You can execute the action manually, after you have properly configured the instance.

Note that if you have VMs that you cannot configure to support these requirements, then you can set up a policy to keep Intersight Workload Optimizer from making these recommendations. Create a group that contains these VMs, and then create policy for that scope. In the policy, exclude instance types by configuring the Cloud Instance Types scaling constraint. For information about excluding instance types, see [Cloud Instance Types \(on page 284\)](#).

The instance requirements that Intersight Workload Optimizer recognizes are:

- Enhanced Network Adapters

Some VMs can run on instances that support Enhanced Networking via the Elastic Network Adapter (ENA), while others can run on instances that do not offer this support. Intersight Workload Optimizer can recommend scaling a VM that does not support ENA onto an instance that does. However, you must enable ENA on the VM before executing the scaling action. If you scale a non-ENA VM to an instance that requires ENA, then AWS cannot start up the VM after the scaling action.

For information about ENA configuration, visit this [page](#).

- Linux AMI Virtualization Type

An Amazon Linux AMI can use ParaVirtual (PV) or Hardware Virtual Machine (HVM) virtualization. Intersight Workload Optimizer can recommend scaling a PV VM to an HVM instance that does not include the necessary PV drivers.

To check the virtualization type of an instance, open the Amazon EC2 console to the Details pane, and review the Virtualization field for that instance.

- 64-bit vs 32-bit

Not all AWS instance can support a 32-bit VMs. Intersight Workload Optimizer can recommend scaling a 32-bit VM to an instance that only supports a 64-bit platform.

- NVMe Block

Some instances expose EBS volumes as NVMe block devices, but not all VMs are configured with NVMe drivers. Intersight Workload Optimizer can recommend scaling such a VM to an instance that supports NVMe. Before executing the action, you must install the NVMe drivers on the VM.

In addition, Intersight Workload Optimizer recognizes processor types that you currently use for your VM. For scale actions, Intersight Workload Optimizer keeps your VMs on instance types with compatible processors. For example, if your VM is on an ARM-based instance, then Intersight Workload Optimizer will only recommend scaling to other compatible ARM-based instance types.

Scaling Storage for AWS VMs

When a VM needs more storage capacity Intersight Workload Optimizer recommends actions to scale its volume to an instance that provides more storage. Note that AWS supports both Elastic Block Store (EBS) and Instance storage. Intersight Workload Optimizer recognizes these storage types as it recommends volume actions.

If the root storage for your VM is Instance Storage, then Intersight Workload Optimizer will not recommend an action. This is because Instance Storage is ephemeral, and such an action would cause the VM to lose all the stored data.

If the root storage is EBS, then Intersight Workload Optimizer recommends volume actions. EBS is persistent, and the data will remain after the action. However, if the VM uses Instance Storage for extra storage, then Intersight Workload Optimizer does not include that storage in its calculations or actions.

Nodes in AWS EMR Clusters

Intersight Workload Optimizer treats nodes in AWS [EMR](#) clusters like regular VMs. As such, it could incorrectly generate scale actions for such nodes. After a node action executes, AWS detects the action as a defect, terminates the node, and replaces it with a new instance of the initial size. To avoid this issue, disable scale actions for nodes in EMR clusters.

AWS automatically assigns [system tags](#) to EMR clusters. To disable scale actions, create a VM group that uses these tags as a filter, and then create a VM policy that disables the 'Cloud Scale All' action type for the VM group.

Actions for Azure VMs

Intersight Workload Optimizer supports the following actions:

- **Scale**

Change the VM instance to use a different instance type or tier to optimize performance and costs.

See additional information for scale actions below.

- **Discount-related actions**

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing Azure reservations coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

If you need more capacity, then Intersight Workload Optimizer will recommend actions to purchase additional reservations.

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Controlling Scale Actions for Azure VMs

For scale actions, you can create policies to control the scale actions that Intersight Workload Optimizer recommends. In those policies, choose from the following options:

- Cloud Scale All – execute all scaling actions
- Cloud Scale for Performance – only execute scaling actions that improve performance
- Cloud Scale for Savings – only execute scaling actions that reduce costs

The default action acceptance mode for these actions is *Manual*. When you examine the pending actions, only actions that satisfy the policy are allowed to execute. All other actions are read-only.

When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see [Default and User-defined Automation Policies \(on page 574\)](#).

Supported Instance Types for Azure VMs

Intersight Workload Optimizer considers all supported instance types when making scaling decisions for cloud VMs. If you want your VMs to *only scale to* or *avoid* certain instance types, create policies for those VMs.

You can view the instance types that Intersight Workload Optimizer supports from the user interface.

1. Navigate to **More > Settings > Policies**.
2. In the Policy Management page, search for and click **Virtual Machine Defaults**.
3. In the Configure Virtual Machine Policy page:
 - a. Scroll down to the bottom of the page.
 - b. Click **Add Scaling Constraint**.
 - c. Choose **Cloud Instance Types**.
 - d. Click **Edit**.

The policy page shows supported tiers for each cloud provider. A tier is a family of instance types, such as *M1* for Google Cloud, *a1* for AWS and *Basic_A1* for Azure. Expand a tier to see individual instance types and resource allocations.

Azure Resource Group Discovery

For Azure environments that include Resource Groups, Intersight Workload Optimizer discovers the Azure Resource Groups and the tags that are used to identify these groups.

In the Intersight Workload Optimizer user interface, to search for a specific Azure Resource Group, choose **Resource Groups** in the Search Page.

You can set the scope of your Intersight Workload Optimizer session to an Azure Resource Group by choosing a group in the Search results and clicking **Scope To Selection**.

You can also use Azure tags as filter criteria when you create a custom Intersight Workload Optimizer resource group. You can choose the Azure Resource Groups that match the tag criteria to be members of the new custom group.

To find the available tags for a specific Azure Resource Group, add the Basic Info chart configured with Related Tag Information to your view or custom dashboard. See [Basic Info Charts \(on page 514\)](#).

NOTE:

When you inspect Resource Groups, Intersight Workload Optimizer does not currently show the billed costs for those Resource Groups.

Azure Instance Requirements

In Azure environments, some instance types require workloads to be configured in specific ways, and some workload configurations require instance types that support specific features. When Intersight Workload Optimizer generates resize actions in Azure, these actions consider the following features:

- **Accelerated Networking (AN)**

In an Azure environment, not all instance types support AN, and not all workloads on AN instances actually enable AN. Intersight Workload Optimizer maintains a dynamic group of workloads that have AN enabled, and it assigns a policy to that group to exclude any templates that do not support AN. In this way, if a workload is on an instance that supports AN, and that workload has enabled AN, then Intersight Workload Optimizer will not recommend an action that would move the workload to a non-AN instance.

- **Azure Premium Storage**

Intersight Workload Optimizer recognizes whether a workload uses Premium Storage, and will not recommend a resize to an instance that does not support Azure Premium Storage.

In addition, Intersight Workload Optimizer recognizes processor types that you currently use for your workloads. If your workload is on a GPU-based instance, then Intersight Workload Optimizer will only recommend moves to other compatible GPU-based instance types. For these workloads, Intersight Workload Optimizer does not recommend resize actions.

IOPS-aware Scaling for Azure VMs

Intersight Workload Optimizer considers IOPS utilization when making scaling decisions for Azure VMs. To measure utilization, Intersight Workload Optimizer takes into account a variety of attributes, such as per-disk IOPS utilization, whole VM IOPS utilization, cache settings, and IOPS capacity for the VMs. It also respects IOPS utilization and aggressiveness constraints that you set in VM policies. For details, see [Aggressiveness and Observation Periods \(on page 282\)](#).

Analysis impacts VM scaling decisions in different ways. For example:

- If your instance experiences IOPS bottlenecks, Intersight Workload Optimizer can recommend scaling up to a larger instance type to increase IOPS capacity, even if you do not fully use the current VCPU or VMEM resources.
- If your instance experiences underutilization of VMEM and VCPU, but high IOPS utilization, Intersight Workload Optimizer might not recommend scaling down. It might keep you on the larger instance to provide sufficient IOPS capacity.
- If the instance experiences underutilization of IOPS capacity along with normal utilization of other resources, you might see an action to resize to an instance that is very similar to the current one. If you inspect the action details, you should see that you are changing to a less expensive instance with less IOPS capacity.

Actions for Google Cloud VMs

Intersight Workload Optimizer supports the following actions:

- **Scale**

Change the VM instance to use a different instance type or tier to optimize performance and costs.

See additional information for scale actions below.

- **Discount-related actions**

If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing Committed Use Discount (CUD) coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

Actions to purchase CUDs will be introduced in a future release.

- **Reconfigure**

Google Cloud provides a specific set of machine types for each zone in a region. If you create a policy that restricts a VM to certain machine types and the zone it is currently on does not support all of those machine types, Intersight Workload Optimizer will recommend a reconfigure action as a way to notify you of the non-compliant VM.

For example, assume Zone A does not support machine types for the M1 family. When a VM in that zone applies a policy that restricts it to M1, Intersight Workload Optimizer will recommend that you reconfigure the VM.

Points to Consider for Google Cloud Scale Actions

- Intersight Workload Optimizer can generate scaling actions for the following VMs, but cannot execute the actions automatically:
 - VMs with local SSDs
Intersight Workload Optimizer can recommend scaling to a machine type that supports local SSDs and the number of disks required by the VM, but blocks action execution due to prerequisite steps that you can only perform from Google Cloud. You can view the prerequisite steps when you examine a pending action.
 - VMs configured with a [minimum CPU platform](#)
Google Cloud instance type families can support multiple CPU generations. A specific VM may be configured with a minimum CPU platform to prevent it from scaling to instance types with incompatible CPUs. When you examine a pending action for such a VM, verify that the recommended instance type runs a compatible CPU. Once verified, manually execute the action from Google Cloud.
- Intersight Workload Optimizer does not recognize prioritized attributions you may have set for CUDs. For example, if you have prioritized all your CUD allotments for a single project, Intersight Workload Optimizer can still recommend actions to apply CUD to other projects in your environment.
- Since all Google Cloud compute tiers have the same net throughput capacity, Intersight Workload Optimizer will not generate scaling actions in response to net throughput.
- Intersight Workload Optimizer does not recommend scaling actions for:
 - [Spot VMs](#)

NOTE:
Intersight Workload Optimizer discovers spot VMs, but does not recommend actions or monitor costs for these VMs.

 - VMs running on [sole-tenant nodes](#)
 - VMs with attached [GPUs](#)
 - VMs in [managed instance groups](#)
 - VMs running [custom machine types](#)

Controlling Scale Actions for Google Cloud VMs

For scale actions, you can create policies to control the scale actions that Intersight Workload Optimizer recommends. In those policies, choose from the following options:

- Cloud Scale All – execute all scaling actions
- Cloud Scale for Performance – only execute scaling actions that improve performance
- Cloud Scale for Savings – only execute scaling actions that reduce costs

The default action acceptance mode for these actions is *Manual*. When you examine the pending actions, only actions that satisfy the policy are allowed to execute. All other actions are read-only.

When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see [Default and User-defined Automation Policies \(on page 574\)](#).

Supported Instance Types for Google Cloud VMs

Intersight Workload Optimizer considers all supported instance types when making scaling decisions for cloud VMs. If you want your VMs to *only scale to* or *avoid* certain instance types, create policies for those VMs.

You can view the instance types that Intersight Workload Optimizer supports from the user interface.

1. Navigate to **More > Settings > Policies**.
2. In the Policy Management page, search for and click **Virtual Machine Defaults**.
3. In the Configure Virtual Machine Policy page:
 - a. Scroll down to the bottom of the page.
 - b. Click **Add Scaling Constraint**.
 - c. Choose **Cloud Instance Types**.
 - d. Click **Edit**.

The policy page shows supported tiers for each cloud provider. A tier is a family of instance types, such as *M1* for Google Cloud, *a1* for AWS and *Basic_A1* for Azure. Expand a tier to see individual instance types and resource allocations.

Cloud VM Uptime

For cloud VMs, Intersight Workload Optimizer includes *uptime* data in its cost calculations. This is especially important for VMs that do not run 24/7 and are charged on-demand rates. With uptime data, Intersight Workload Optimizer can calculate costs more accurately based on the amount of time a VM has been running.

The Action Details page shows uptime data for these VMs. Intersight Workload Optimizer calculates uptime based on the VM's age.



Key Concepts

■ Uptime

A percentage value that indicates how long a VM has been running over a period of time (age)

■ Age

The number of days that a VM has existed since first discovery. For VMs older than 30 days, Intersight Workload Optimizer displays a value of **30+ days**, but only calculates uptime over the last 30 days.

For newly discovered VMs, age is 0 (zero) on the day of discovery. If the VM is running at the time of discovery, uptime is 100%. Otherwise, uptime is 0% and remains unchanged until the VM is powered on. Intersight Workload Optimizer recalculates uptime every hour and then refreshes the data shown in the user interface.

Examples

- A VM that was first discovered 5 days (or 120 hours) ago and has been running for a total of 60 hours during that period has a current uptime value of 50%.
- A VM that was first discovered 2 months ago and has been running for a total of 180 hours over the last 30 days (or 720 hours) has a current uptime value of 25%.

Cost Calculations Using Uptime Data

Intersight Workload Optimizer uses uptime data to calculate estimated on-demand costs for your cloud VMs. For details about calculations, see [Estimated On-demand Monthly Costs for Cloud VMs \(on page 273\)](#).

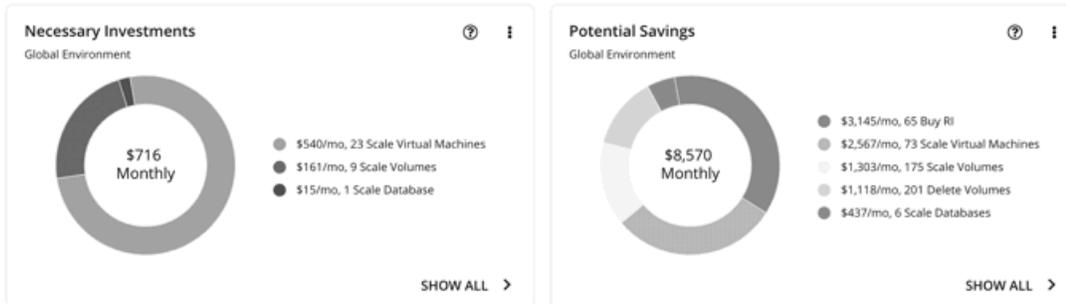
Uptime data impacts cost calculations, but not the actual scaling decisions that Intersight Workload Optimizer makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Uptime Data in Charts

Intersight Workload Optimizer recalculates uptime data every hour and then updates the values shown in charts. The following charts reflect the cost impact of uptime-based calculations:

- Potential Savings and Necessary Investment charts

The projected amounts in these charts include on-demand costs for cloud VMs.

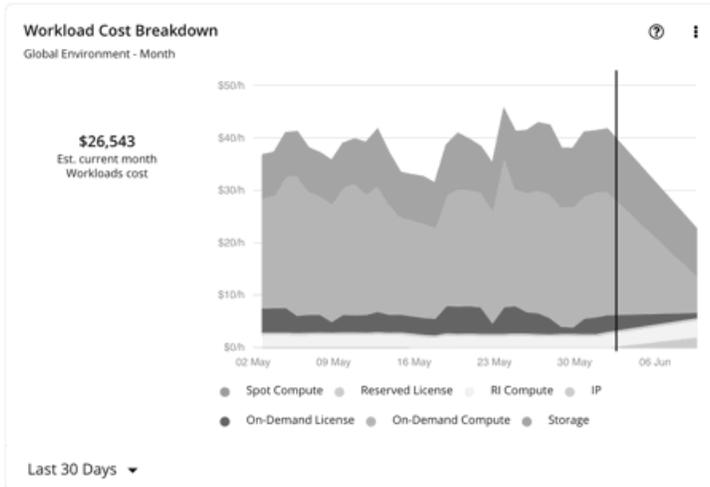


When you click **Show All** in these charts and view details for a pending VM action, the **Action Details** page shows on-demand costs before and after you execute the action, factoring in the VM's uptime value. The page also shows the VM's age.



■ Workload Cost Breakdown chart

This chart shows estimated costs over time, including on-demand costs for VMs.



■ The **Entity Information** chart shows the latest uptime and age data for a specific cloud VM.

| | |
|------------------------|-------------------|
| Number of VCPUs | 4 |
| Region | azure-Canada East |
| Account | [Redacted] |
| Resource Group | [Redacted] |
| Uptime ⓘ | 86.9% |
| Last Modification Time | N/A |
| Attachment State | Attached |
| Vendor ID [EA - PT2] | [Redacted] |
| Age ⓘ | 30+ days |

Estimated On-demand Costs for Cloud VMs

Intersight Workload Optimizer considers a variety of factors when calculating *Estimated On-demand Monthly Cost* for a cloud VM.

| VIRTUAL MACHINE DETAILS | | | | | | | |
|-------------------------|---------------|-----------------|---------|-------------------------------|--------------|--------------------|--------------|
| NAME | ID | AGE | ACCOUNT | REGION | | | |
| PT_Consistent_S... | | 30+ days | | aws-EU (Paris) | | | |
| VCPU PERCENTILE | | VMEM PERCENTILE | | NET THROUGHPUT | | IO THROUGHPUT | |
| 1 % | 0.6 % ↓ | 94 % | 47 % ↓ | 0 % | 0 % | 0 % | 0 % |
| 800 MHz | → 1.4 GHz | 1 GiB | → 2 GiB | 468.8 MB/s | → 468.8 MB/s | 260.6 MB/s | → 260.6 MB/s |
| ON-DEMAND RATE ⓘ | RI COVERAGE ⓘ | UPTIME ⓘ | | EST. ON-DEMAND MONTHLY COST ⓘ | | | |
| \$0.012/h | → \$0.021/h | 50% | → 0% | 95.3% | | \$4.1/mo → \$15/mo | |

AWS VMs and Azure VMs Without License Costs

Cost Calculation

For these VMs, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

$$\text{On-demand Rate} * \text{Usage Not Covered by Discounts} * \text{Uptime} * 730 = \text{Estimated On-demand Monthly Cost}$$

Where:

- **On-demand Rate** is the hourly cost for a VM's instance type *without* discount coverage (AWS RIs/Savings Plans or Azure reservations).
 - For AWS, this rate includes all license costs, but not storage or IP. You can obtain on-demand rates via [Amazon EC2 On-demand Pricing](#).
 - For Azure, the rate does *not* include license costs, storage, or IP. You can obtain on-demand rates via [Azure Pricing Calculator](#).

NOTE:

Azure VMs covered by Azure Hybrid Benefit do not have license costs.

- **Usage Not Covered by Discounts** is the percentage of hourly VM usage not covered by any discount. For example:
 - Discount Coverage = 20% (0.2)
 - Usage Not Covered by Discounts = 80% (0.8)
- **Uptime** is a percentage value that indicates how long a VM has been running since it was first discovered by Intersight Workload Optimizer. For VMs discovered more than 30 days ago, Intersight Workload Optimizer only calculates uptime over the last 30 days.

To estimate monthly on-demand costs, Intersight Workload Optimizer projects the current uptime value into the future. It assumes that future uptime will be similar to the current uptime.

- **730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.

The listed items above impact cost calculations, but not the actual scaling decisions that Intersight Workload Optimizer makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for an AWS VM:

VIRTUAL MACHINE DETAILS

| NAME | ID | AGE | ACCOUNT | REGION |
|--------------------|----|----------|---------|----------------|
| PT_Consistent_S... | | 30+ days | | aws-EU (Paris) |

| VCPU PERCENTILE | VMEM PERCENTILE | NET THROUGHPUT | IO THROUGHPUT |
|-------------------------------------|---------------------------------|---------------------------------------|---------------------------------------|
| 1 % 800 MHz → 0.6 % ↓ 1.4 GHz | 94 % 1 GiB → 47 % ↓ 2 GiB | 0 % 468.8 MB/s → 0 % 468.8 MB/s | 0 % 260.6 MB/s → 0 % 260.6 MB/s |

| | | | |
|---|---------------------------|-------------------|---|
| ON-DEMAND RATE ⓘ \$0.012/h → \$0.021/h | RI COVERAGE ⓘ 50% → 0% | UPTIME ⓘ 95.3% | EST. ON-DEMAND MONTHLY COST ⓘ \$4.1/mo → \$15/mo |
|---|---------------------------|-------------------|---|

| | Current Values | Values After Action Execution |
|--|----------------|-------------------------------|
| On-demand Rate | \$0.012/hr | \$0.021/hr |
| Discount Coverage | 50% (0.5) | 0% (0.0) |
| Usage Not Covered by Discounts <i>(calculated based on discount coverage)</i> | 50% (0.5) | 100% (1.0) |
| Uptime | 95.3% (.953) | |

Intersight Workload Optimizer calculates the following:

VIRTUAL MACHINE DETAILS

| NAME | ID | AGE | ACCOUNT | REGION |
|--------------------|----|----------|---------|----------------|
| PT_Consistent_S... | | 30+ days | | aws-EU (Paris) |

| VCPU PERCENTILE | VMEM PERCENTILE | NET THROUGHPUT | IO THROUGHPUT |
|-------------------------------------|---------------------------------|---------------------------------------|---------------------------------------|
| 1 % 800 MHz → 0.6 % ↓ 1.4 GHz | 94 % 1 GiB → 47 % ↓ 2 GiB | 0 % 468.8 MB/s → 0 % 468.8 MB/s | 0 % 260.6 MB/s → 0 % 260.6 MB/s |

| | | | |
|---|---------------------------|-------------------|---|
| ON-DEMAND RATE ⓘ \$0.012/h → \$0.021/h | RI COVERAGE ⓘ 50% → 0% | UPTIME ⓘ 95.3% | EST. ON-DEMAND MONTHLY COST ⓘ \$4.1/mo → \$15/mo |
|---|---------------------------|-------------------|---|

■ **Current Estimated On-demand Monthly Cost:**

$$0.012 * 0.5 * 0.953 * 730 = 4.1$$

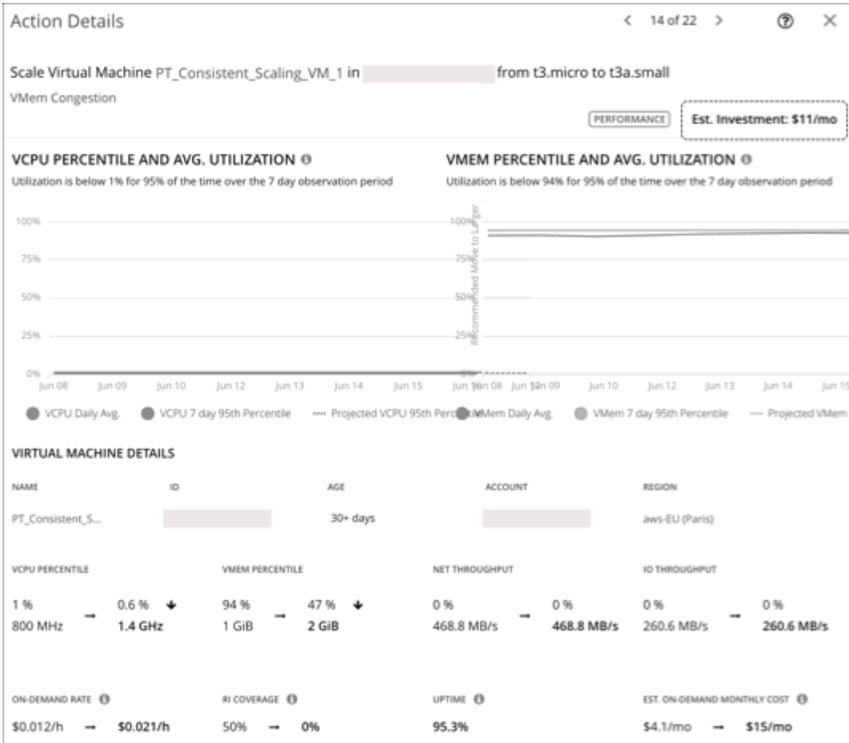
■ **Estimated On-demand Monthly Cost *after* executing the action:**

$$0.021 * 1.0 * 0.953 * 730 = 15$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

Since the Estimated On-demand Monthly Cost is projected to increase from \$4.1/month to \$15/month, Intersight Workload Optimizer treats the action as an investment and shows an estimated investment of \$11/month.



Azure VMs with License Costs

Cost Calculation

For VMs with license costs, Intersight Workload Optimizer first calculates the *On-demand Compute Rate*, which it then uses to calculate *Estimated On-demand Monthly Costs*.

1. On-demand Compute Rate Calculation

The calculation for On-demand Compute Rate can be expressed as follows:

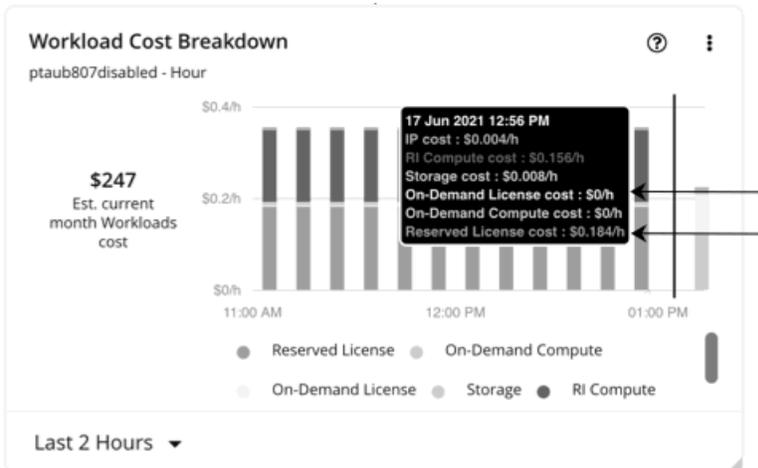
$$\text{On-demand Rate} - (\text{Reserved License Cost} + \text{On-demand License Cost}) = \text{On-demand Compute Rate}$$

Where:

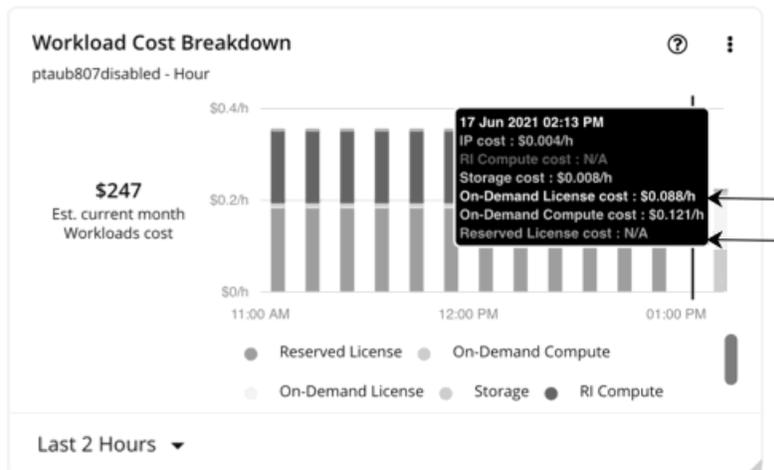
- **On-demand Rate** is the hourly cost for a VM's instance type *without* reservation coverage. This does *not* include license costs, storage, or IP. You can obtain on-demand rates via [Azure Pricing Calculator](#).
- **Reserved License Cost** and **On-demand License Cost** are the hourly costs for the VM's licenses. You can obtain license costs via Azure Pricing Calculator or the Intersight Workload Optimizer user interface.

From the user interface, set the scope to the Azure VM and then see the Workload Cost Breakdown chart. In the chart, set the time frame to Last 2 Hours, and then:

- Hover over the second to the last bar in the chart to obtain the *current* On-demand License Cost and Reserved License Cost.



- Hover over the last bar (after the vertical line) in the chart to obtain the On-demand License Cost and Reserved License Cost *after* you execute actions.



The *On-demand Compute Rate* and *License Cost (On-demand and Reserved)* are then used to calculate Estimated On-demand Monthly Costs.

2. Estimated On-demand Monthly Cost Calculation

The calculation can be expressed as follows:

$$(\text{On-demand Compute Rate} * \text{Usage Not Covered by Reservations}) + \text{License Cost} * \text{Uptime} * 730 = \text{Estimated On-demand Monthly Cost}$$

Where:

- **Usage Not Covered by Reservations** is the percentage of hourly VM usage not covered by any reservation. For example:
 - Reservation Coverage = 20% (0.2)
 - Usage Not Covered by Reservations = 80% (0.8)
 - **License Cost** is the sum of On-demand License Cost and Reserved License Cost.
 - **Uptime** is a percentage value that indicates how long a VM has been running since it was first discovered by Intersight Workload Optimizer. For VMs discovered more than 30 days ago, Intersight Workload Optimizer only calculates uptime over the last 30 days
- To estimate monthly on-demand costs, Intersight Workload Optimizer projects the current uptime value into the future. It assumes that future uptime will be similar to the current uptime.
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.

The listed items above impact cost calculations, but not the actual scaling decisions that Intersight Workload Optimizer makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

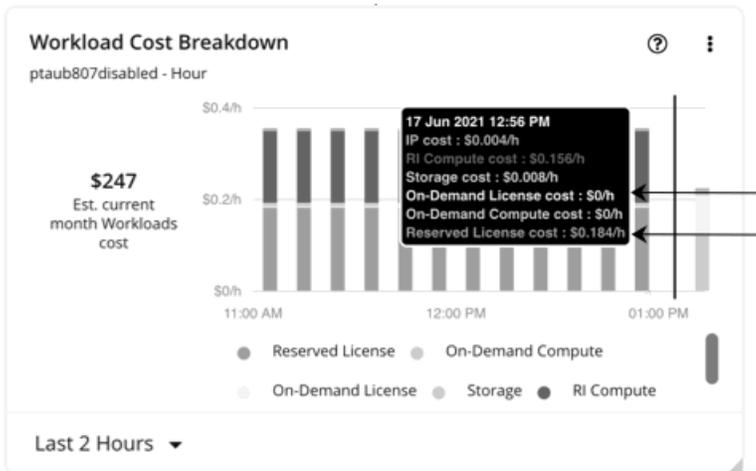
Example

Assume the following data for a pending scale action for an Azure VM with license costs:

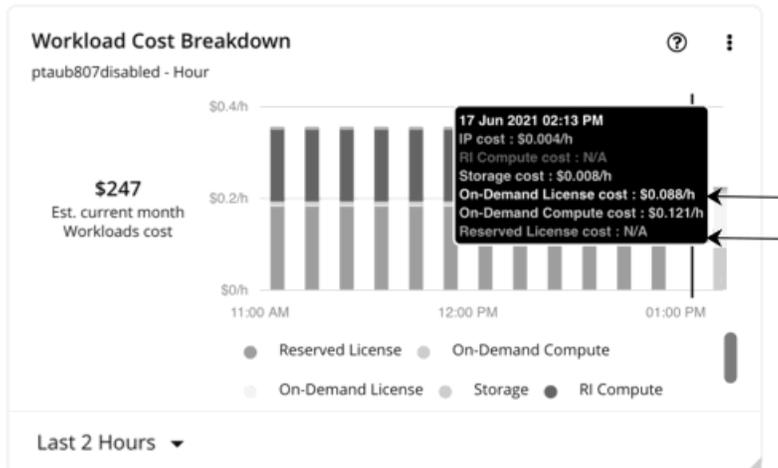
| VIRTUAL MACHINE DETAILS | | | | |
|----------------------------------|--------------------|-----------------------------|------------------|--------------------|
| NAME | ID | AGE | SUBSCRIPTION | LOCATION |
| ptaub807disable... | 25b9c0ce-0e08-4... | 30+ days | EA - Development | azure-East US 2 |
| VCPUs PERCENTILE | | VMEM PERCENTILE | | VM IOPS PERCENTILE |
| 2 % | 2.9 % ↑ | 9 % | 18 % ↑ | 0 % |
| 33.2 GHz | 22.7 GHz | 32 GiB | 16 GiB | 6,400 IOPS |
| VM STORAGE THROUGHPUT PERCENTILE | | EST. ON-DEMAND MONTHLY COST | | |
| 0 % | 0 % | 129/mo | → \$153/mo | |
| 96 MB/s | 48 MB/s | 100% → 0% | 96.1% | |
| ON-DEMAND RATE | | RI COVERAGE | UPTIME | |
| \$0.45/h | → \$0.218/h | 100% → 0% | 96.1% | |

| | Current Values | Values After Action Execution |
|----------------|----------------|-------------------------------|
| On-demand Rate | \$0.45/hr | \$0.218/hr |

Current license costs



License costs after action execution



| | Current Values | Values After Action Execution |
|------------------------|----------------|-------------------------------|
| On-demand License Cost | \$0/hr | \$0.088/hr |
| Reserved License Cost | \$0.184/hr | N/A |

1. Insight Workload Optimizer first calculates the following:

- Current On-demand Compute Rate:

$$0.45 - (0.184 + 0) = 0.266$$

- On-demand Compute Rate *after* executing the action:

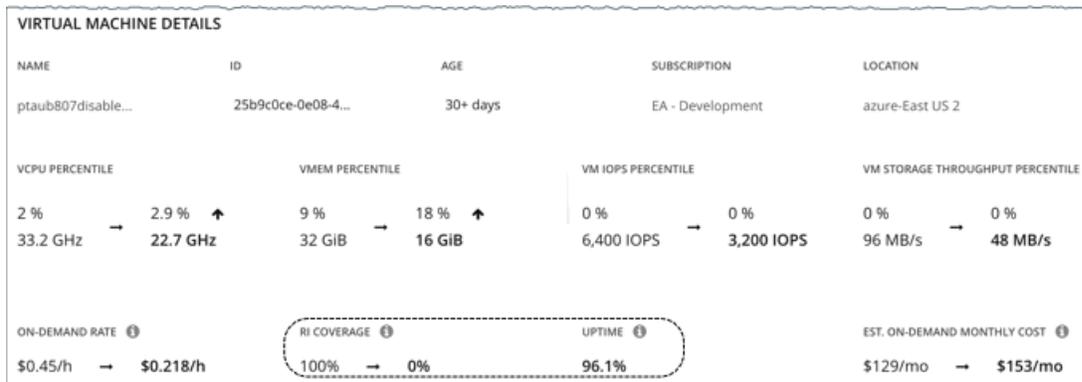
$$0.218 - (0 + 0.088) = 0.13$$

2. Insight Workload Optimizer can now calculate Estimated On-demand Monthly Cost based on:

- On-demand Compute Rate

| | Current Values | Values After Action Execution |
|------------------------|----------------|-------------------------------|
| On-demand Compute Rate | \$0.266/hr | \$0.13/hr |

- Usage Not Covered by Reservations and Uptime



| | Current Values | Values After Action Execution |
|----------------------|----------------|-------------------------------|
| Reservation Coverage | 100% (1.0) | 0% (0.0) |

| | Current Values | Values After Action Execution |
|--|----------------|-------------------------------|
| Usage Not Covered by Reservations <i>(calculated based on reservation coverage)</i> | 0% (0.0) | 100% (1.0) |
| Uptime | 96.1% (.961) | |

Intersight Workload Optimizer calculates the following:

VIRTUAL MACHINE DETAILS

| NAME | ID | AGE | SUBSCRIPTION | LOCATION |
|--------------------|--------------------|----------|------------------|-----------------|
| ptaub807disable... | 25b9c0ce-0e08-4... | 30+ days | EA - Development | azure-East US 2 |

| VCPU PERCENTILE | VMEM PERCENTILE | VM IOPS PERCENTILE | VM STORAGE THROUGHPUT PERCENTILE |
|------------------------------|--------------------------|--------------------------------|----------------------------------|
| 2 % 33.2 GHz → 22.7 GHz ↑ | 9 % 32 GiB → 16 GiB ↑ | 0 % 6,400 IOPS → 3,200 IOPS | 0 % 96 MB/s → 48 MB/s |

| ON-DEMAND RATE ⓘ | RI COVERAGE ⓘ | UPTIME ⓘ | EST. ON-DEMAND MONTHLY COST ⓘ |
|----------------------|---------------|----------|-------------------------------|
| \$0.45/h → \$0.218/h | 100% → 0% | 96.1% | \$129/mo → \$153/mo |

■ **Current** Estimated On-demand Monthly Cost:

$$(0.266 * 0.0) + 0.184 * 0.961 * 730 = 129$$

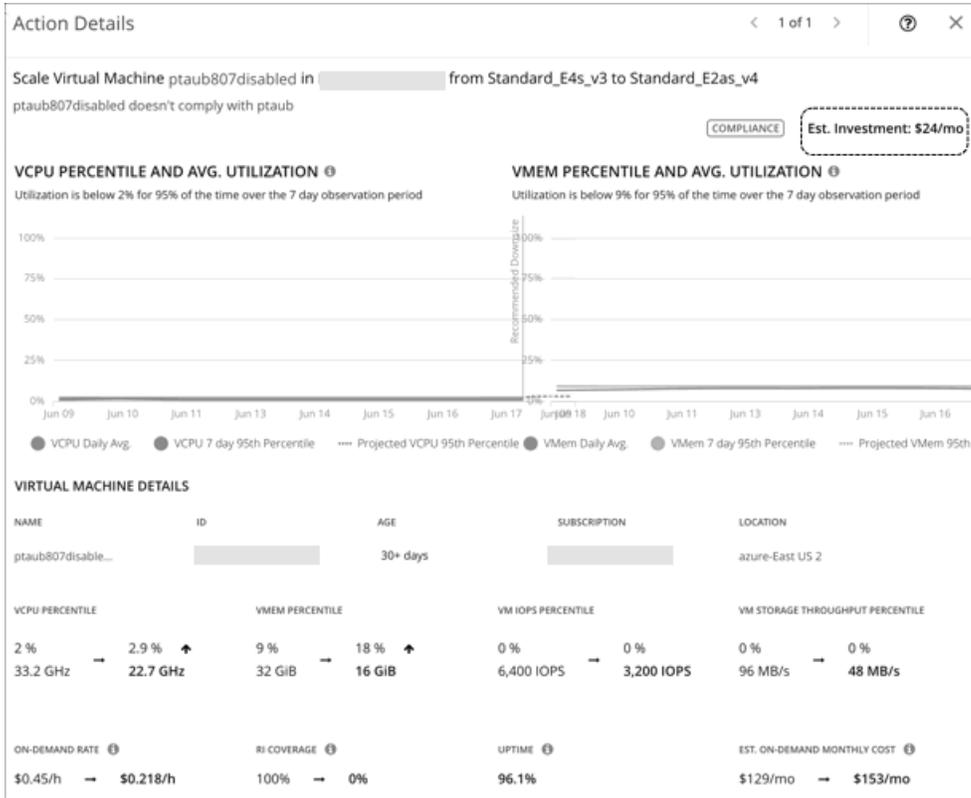
■ Estimated On-demand Monthly Cost *after* executing the action:

$$(0.13 * 1.0) + 0.088 * 0.961 * 730 = 153$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

Since the on-demand cost is projected to increase from \$129/month to \$153/month, Intersight Workload Optimizer treats the action as an investment and shows an estimated investment of \$24/month.



Cloud VM Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about cloud VM actions, see [Cloud VM Actions \(on page 265\)](#).

Cloud Scale

| Action | Default Mode | AWS, Azure, and Google Cloud |
|-----------------------------|--------------|---|
| Cloud Scale All | Manual | Auto |
| Cloud Scale for Performance | Manual | Auto |
| Cloud Scale for Savings | Manual | Auto |

Other Actions

| Action | Default Mode | AWS and Azure | Google Cloud |
|--|--------------|---|---|
| Buy discounts | Recommend | Rcmd | Not yet supported |
| Provision container platform node (VM) | Manual | Rcmd | Rcmd |
| Suspend container platform node (VM) | Manual | Auto | Auto |

Scaling Target Utilization

For VCPU, VMEM, and IO/Net Throughput Utilization:

These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Intersight Workload Optimizer calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Intersight Workload Optimizer recommends actions, in most cases you should never need to use them. If you want to control how Intersight Workload Optimizer recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

| Attribute | Default Value |
|---|---|
| Scaling Target VCPU Utilization | 70 The target utilization as a percentage of VCPU capacity. |
| Scaling Target VMEM Utilization | 90 The target utilization as a percentage of memory capacity. |
| Scaling Target IO Throughput Utilization | 70 The target utilization as a percentage of IO throughput (Read and Write) capacity. |
| Scaling Target Net Throughput Utilization | 70 The target utilization as a percentage of network throughput (Inbound and Outbound) capacity. |

For IOPS Utilization:

Intersight Workload Optimizer uses this setting in conjunction with aggressiveness constraints to control scaling actions for VMs. You can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

| Attribute | Default Value |
|---|--|
| Scaling Target IOPS Utilization (Azure VMs only) | 70 For Azure environments, the target percentile value Intersight Workload Optimizer will attempt to match. |

For details on how IOPS utilization affects scaling decisions, see [IOPS-aware Scaling for Azure VMs \(on page 269\)](#).

Aggressiveness and Observation Periods

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU, vMEM, and IOPS (Azure VMs only). It then recommends actions to improve utilization based on the observed values for a given time period.

■ Aggressiveness

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce the capacity for CPU on a VM. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. For most VMs, there are moments when peak CPU reaches

high levels, such as during reboots, patching, and other maintenance tasks. Assume utilization for a VM peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce allocated CPU for that VM.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single CPU burst to 100%, but for 95% of the samples CPU never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce CPU allocation for the VM.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 100th and 99th Percentile – More performance. Recommended for critical workloads that need maximum guaranteed performance at all times, or workloads that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.
- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures application performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for non-production workloads that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 30 days. Use the **Max Observation Period** setting to adjust the number of days. To ensure that there are enough samples to analyze and drive scaling actions, set the **Min Observation Period**.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 30 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the database has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 90 Days
- Recommended – Last 30 Days
- More Elastic – Last 7 Days

Intersight Workload Optimizer recommends an observation period of 30 days following the monthly workload maintenance cycle seen in many organizations. VMs typically peak during the maintenance window as patching and other maintenance tasks are carried out. A 30-day observation period means that Intersight Workload Optimizer can capture these peaks and increase the accuracy of its sizing recommendations.

You can set the value to 7 days if workloads need to resize more often in response to performance changes. For workloads that cannot handle changes very often or have longer usage periods, you can set the value to 90 days.

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic – None
- Less Elastic – 7 Days

Cloud Instance Types

| Attribute | Default Value |
|----------------------|---------------|
| Cloud Instance Types | None |

By default, Intersight Workload Optimizer considers all instance types currently available for scaling when making scaling decisions for VMs. However, you may have set up your cloud VMs to *only scale to* or *avoid* certain instance types to reduce complexity and cost, improve discount utilization, or meet application demand. Use this setting to identify the instance types that VMs can scale to.

NOTE:

Under most circumstances, when a cloud provider offers a new instance type that is meant to replace an older type, the provider offers it at a lower cost. However, a provider may provide a new instance type with identical costs as the older instance types. If this occurs, and capacity and cost are equal, Intersight Workload Optimizer cannot ensure that it chooses the newer instance type. To work around this issue, you can create an Action Automation policy that excludes the older instance type.

Click **Edit** to set your preferences. In the new page that displays, expand a **cloud tier** (a family of instance types, such as *a1* for AWS or *B-series* for Azure) to see individual instance types and the resources allocated to them. If you have several cloud providers, each provider will have its own tab.

Select your preferred instance types or cloud tiers, or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

If you selected a cloud tier and the service provider deploys new instance types to that tier later, then those instance types will automatically be included in your policy. Be sure to review your policies periodically to see if new instance types have been added to a tier. If you do not want to scale to those instance types, update the affected policies.

Consistent Resizing

| Attribute | Default Setting |
|---------------------|-----------------|
| Consistent Resizing | Off |

Consistent Resizing for User-defined Automation Policies

When you create a policy for a group of VMs and turn on Consistent Resizing, Intersight Workload Optimizer resizes all the group members to the same size, such that they all support the top utilization of each resource commodity in the group. For example, assume VM A shows top utilization of CPU, and VM B shows top utilization of memory. A resize action would result in all the VMs with CPU capacity to satisfy VM A, and memory capacity to satisfy VM B.

For an affected resize, the Actions List shows individual resize actions for each of the VMs in the group. If you automate resizes, Intersight Workload Optimizer executes each resize individually in a way that avoids disruption to your workloads.

Use this setting to enforce the same template across all VMs in a group when resizing VMs on the public cloud. In this way, Intersight Workload Optimizer can enforce a rule to size all the VMs in a group equally.

Consistent Resizing for Auto-discovered Groups

In public cloud environments, Intersight Workload Optimizer discovers groups that should keep all their VMs on the same template, and then creates read-only policies for them to implement Consistent Resizing. The details of this discovery and the associated policy vary depending on the cloud provider.

■ Azure

Intersight Workload Optimizer discovers Azure availability sets and scale sets.

- For availability sets, Intersight Workload Optimizer does *not* enable Consistent Resizing, but it can recommend scale actions for individual VMs in the availability set.

When a scale action for a VM in an availability set fails due to insufficient resources in the compute cluster, the action remains pending. When you hover over the pending action, you will see a message indicating that action execution has been temporarily disabled due to a previous execution error in the availability set. Intersight Workload Optimizer assumes that all other VMs in the availability set will fail to scale due to the same resource issue, so it creates a temporary policy that disables action execution for the availability set. Specifically, this policy sets the action

acceptance mode for scale actions to *Recommend* and stays in effect for 730 hours (one month). This means that for the duration of the policy, Intersight Workload Optimizer will continue to generate read-only, non-executable scale actions for individual VMs, so you can evaluate their resource requirements and plan accordingly. You can delete this policy if you need to re-enable action execution in the availability set.

- For scale sets, Intersight Workload Optimizer enables Consistent Resizing across all the VMs in the group. You execute those actions directly in Azure. If you do not need to resize all the members of a given scale set to a consistent template, create another policy for that scope and turn off Consistent Resizing.

■ AWS

Intersight Workload Optimizer discovers Auto Scaling Groups and automatically enables Consistent Resizing across all the VMs in each group. You can choose to execute all the actions for such a group, either manually or automatically. In that case, Intersight Workload Optimizer executes the resizes one VM at a time. If you do not need to resize all the members of a given Auto Scaling Group to a consistent template, create another policy for that scope and turn off Consistent Resizing.

If you select one or all actions for the group either manually or automatically, Intersight Workload Optimizer will change the Launch Configuration for the Auto Scaling Group but it will not terminate the EC2 instances.

Below are some use cases for employing Consistent Resizing for a group.

- If you have deployed load balancing for a group, then all the VMs in the group should experience similar utilization. In that case, if one VM needs to be resized, then it makes sense to resize them all consistently.
- A common HA configuration on the public cloud is to deploy mirror VMs to different availability zones, where the given application runs on only one of the VMs at a given time. The other VMs are on standby to recover in failover events. Without Consistent Resizing, Intersight Workload Optimizer would tend to size down or suspend the unused VMs, which would make them unready for the failover situation.

When working with Consistent Resizing, consider these points:

- You should not mix VMs in a group that has a Consistent Resizing policy, with other groups that enable Consistent Resizing. One VM can be a member of more than one group. If one VM (or more) in a group with Consistent Resizing is also in another group that has Consistent Resizing, then both groups enforce Consistent Resizing together, for all their group members.
- If one VM (or more) is in a group with Consistent Resizing turned on, and the same VMs are in a group with Consistent Resizing turned off, the affected VMs assume the ON setting. This is true if you created both groups, or if Intersight Workload Optimizer created one of the groups for Azure Scale Sets or AWS Auto Scaling Groups.
- For any group of VMs that enables Consistent Resizing, you should not mix the associated target technologies. For example, one group should not include VMs that are managed on both Azure and AWS platforms.
- Charts that show actions and risks assign the same risk statement to all the affected VMs. This can seem confusing. For example, assume one VM needs to resize to address vCPU risk, and 9 other VMs are set to resize consistently with it. Then charts will state that 10 VMs need to resize to address vCPU risks.

Ignore NVMe Constraints

For AWS, Intersight Workload Optimizer recognizes when a VM instance includes an NVMe driver. To respect NVMe constraints, it will not recommend scaling to an instance type that does not also include an NVMe driver. If you ignore NVMe constraints, then Intersight Workload Optimizer is free to scale the instance to a type that does not include an NVMe driver.

| Attribute | Default Setting |
|---------------------------------------|-----------------|
| Ignore NVMe Constraints (AWS only) | Off |

Instance Store Aware Scaling

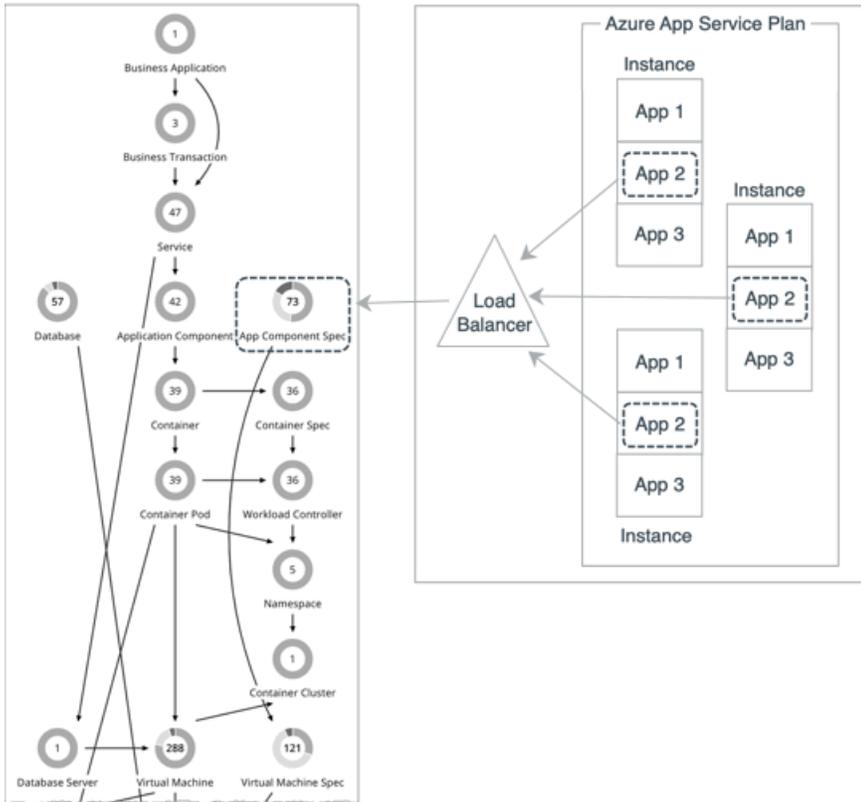
| Attribute | Default Setting |
|------------------------------|-----------------|
| Instance Store Aware Scaling | Off |

The template for your workload determines whether the workload can use an *instance store*, and it determines the instance store capacity. As Intersight Workload Optimizer calculates a resize or move action, it can recommend a new template that does not support instance stores, or that does not provide the same instance store capacity.

To ensure that resize actions respect the instance store requirements for your workloads, turn on **Instance Store Aware Scaling** for a given VM or for a group of VMs. When you turn this on for a given scope of VMs, then as it calculates move and resize actions, Intersight Workload Optimizer will only consider templates that support instance stores. In addition, it will not move a workload to a template that provides less instance store capacity.

App Component Spec

In [Azure App Service \(on page 105\)](#) deployments, an App Component Spec represents a set of app instances comprising a single web application. Intersight Workload Optimizer discovers App Component Specs when you add an Azure target with the necessary permissions.



NOTE:

For a list of permissions, see [Azure Permissions \(on page 85\)](#).

Intersight Workload Optimizer also discovers the *plans* that provide resources to app instances. The supply chain shows these plans as [Virtual Machine Specs \(on page 287\)](#) and links them with App Component Specs to establish their relationship.

Synopsis

| | |
|---------------------|----------------------------------|
| Synopsis | |
| Provides: | App services to end users |
| Consumes: | Resources from App Service plans |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Response Time
 - Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).
- Virtual CPU (VCPU)
 - Virtual CPU is the measurement of CPU that is in use.

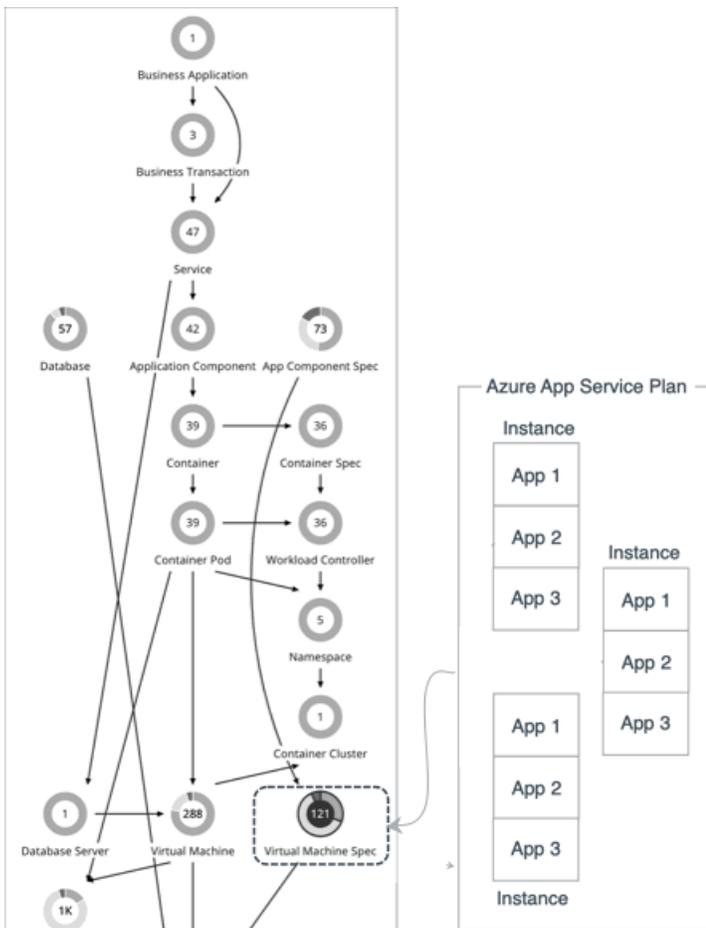
Actions

None

Intersight Workload Optimizer does not recommend actions for App Component Specs, but it does recommend actions for the underlying Virtual Machine Specs. For details, see [Virtual Machine Spec Actions \(on page 288\)](#).

Virtual Machine Spec

In [Azure App Service \(on page 105\)](#), plans define CPU, memory, and storage resources that are available to VM instances to run apps. When you add an Azure target with the necessary permissions, Intersight Workload Optimizer discovers the plans associated with apps, and shows them as Virtual Machine Specs in the supply chain. Currently, Intersight Workload Optimizer discovers all plans, except App Service Environment v3 I4, I5, and I6.



NOTE:

For a list of permissions, see [Azure Permissions \(on page 85\)](#).

Points to consider:

- Azure App Service offers several types of apps, including web apps, mobile apps, API apps, and logic apps. Intersight Workload Optimizer discovers the plans associated with these apps, but only recommends scale actions for plans associated with web apps. If a plan is no longer associated with any type of app, Intersight Workload Optimizer will recommend that you delete it.
- For web apps, Intersight Workload Optimizer also discovers the app instances that consume resources from a plan, and shows them as [App Component Specs \(on page 286\)](#) in the supply chain. The supply chain links App Component Specs with Virtual Machine Specs to establish their relationship.
- VM instances underlying a plan scale as a group. For this reason, Intersight Workload Optimizer represents these VM instances as a single Virtual Machine Spec entity and does *not* monitor them individually. The Entity Information chart for a Virtual Machine Spec shows the current number of VM instances, while resource charts (such as the Virtual CPU and Virtual Memory charts) show aggregated metrics for all VM instances.

Synopsis

| | |
|---------------------|---|
| Synopsis | |
| Provides: | Resources to apps (via App Component Specs) |
| Consumes: | Resources from Azure regions |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Virtual CPU (VCPU)
Virtual CPU is the measurement of CPU that is in use.
- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.
- Number of Replicas
Number of Replicas is the total number of VM instances underlying an App Service plan.

Actions

Intersight Workload Optimizer supports the following actions:

- **Scale**
Scale Azure App Service plans to optimize app performance or reduce costs, while complying with business policies.
- **Delete**
Delete empty Azure App Service plans as a cost-saving measure. A plan is considered empty if it is not hosting any running apps.

Scale Actions

Intersight Workload Optimizer supports vertical scaling actions for provisioned App Service plans. These actions change the *size* of all VM instances underlying a plan (for example, from small to large, or large to medium). Horizontal scaling actions, which change the *number* of VM instances underlying a plan, are currently not supported.

Vertical scaling recommendations rely on a variety of factors, including:

- [Resource utilization percentiles \(on page 290\)](#)
- [On-demand monthly costs \(on page 291\)](#)
- VM instance count

Intersight Workload Optimizer will only recommend vertical scaling actions on plans with six or less VM instances.

- Scaling eligibility
 - Eligible for scaling – Basic, Standard, Premium v2, Premium v3, Isolated, Isolated v2
 - Not eligible for scaling – Workflow Standard, Elastic Premium, Free, Shared, Dynamic/Serverless
- Azure-enforced constraints, including:
 - Region – Only recommend instance types in regions where they are available
 - Server rack – Only recommend instance types on server racks where they are available
 - Zone redundancy – If zone redundancy is enabled, only recommend instance types that support zone redundancy
 - Deployment slots – Only recommend instance types that support the currently configured number of deployment slots that can be added to apps
 - Hybrid connections – Only recommend scaling to instance types that support the currently configured number of hybrid connections for a plan

NOTE:

To see the number of deployment slots and hybrid connections configured for a plan, set the scope to the corresponding Virtual Machine Spec and then view the Entity Information chart.

- Scaling constraints that you set in Intersight Workload Optimizer [policies \(on page 295\)](#) for Virtual Machine Specs

For example, you can set a constraint if you want App Service plans to *only scale to* or *avoid* certain instance types. For scale actions, you can create policies to control the scale actions that Intersight Workload Optimizer recommends. In those policies, choose from the following options:

 - Cloud Scale All – execute all scaling actions
 - Cloud Scale for Performance – only execute scaling actions that improve performance
 - Cloud Scale for Savings – only execute scaling actions that reduce costs

The default action acceptance mode for these actions is *Manual*. When you examine the pending actions, only actions that satisfy the policy are allowed to execute. All other actions are read-only.

When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see [Default and User-defined Automation Policies \(on page 574\)](#).

Delete Actions

When Intersight Workload Optimizer discovers an empty plan (i.e., a plan that is not hosting any running apps), it will immediately recommend that you delete the plan as a cost-saving measure. Intersight Workload Optimizer can recommend deleting provisioned App Service plans, as well as Elastic Premium and Workflow Standard plans.

If a currently empty plan is not deleted and is subsequently discovered as used, Intersight Workload Optimizer removes the delete action attached to it.

Delete actions include the 'Days Empty' information that indicates how long a plan has been empty.

| RESOURCE IMPACT | | CURRENT | AFTER ACTIONS |
|-----------------|--|-----------|---------------|
| Virtual Memory | | 7 GB | - |
| Virtual CPU | | 18.36 GHz | - |

| APP SERVICE PLAN DETAILS | |
|--------------------------|-------------------------|
| Name | paas-asp-zone-redundant |
| Id | paas-asp-zone-redundant |
| Subscription | [redacted] |
| Location | azure-East US |
| Days Empty | 7 |

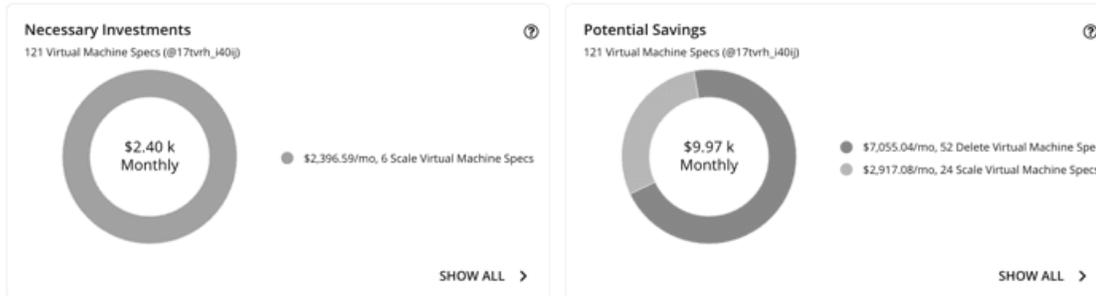
You can control the delete actions that Intersight Workload Optimizer recommends, based on the 'Days Empty' value that you set. For example, if you want Intersight Workload Optimizer to only generate delete actions for plans that have been empty for at least 5 days, perform these steps:

1. In the default policy for Virtual Machine Specs, *disable* delete actions.

2. Create a dynamic group of Virtual Machine Specs and set the 'Days Empty' filter to `Days Empty > = 5`.
3. Create a custom Virtual Machine Spec policy, set the scope to the group that you just created, and then *enable* delete actions in that policy.

Actions in Charts

Use the Necessary Investments and Potential Savings charts to view pending Virtual Machine Spec actions. These charts show total monthly investments and savings, assuming you execute all the actions.



Click **Show All** for each chart to review and execute the actions.

The table shows the following:

- Actions that are pending for each Virtual Machine Spec
- Savings or investments for each Virtual Machine Spec

Utilization Charts for Scale Actions

Intersight Workload Optimizer uses percentile calculations to measure resource utilization, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on an App Service plan, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.

Action Details
< 1 of 24 >
?
✕

Scale Virtual Machine Spec ASP-CloudPaaS-3e from I2v2 to I1v2 in ↓ \$845.34/mo SAVINGS

Underutilized vCPU, vMEM

VCPU PERCENTILE AND AVG. UTILIZATION

vCPU Utilization is below 2% for 95% of the time over the 30 day observation period

VMEM PERCENTILE AND AVG. UTILIZATION

vMEM Utilization is below 14% for 95% of the time over the 30 day observation period

ACTION ESSENTIALS

State: ✔ Action can be accepted and executed immediately.

Non-Disruptive: ✕ Downtime is required to execute.

Reversible: ✔ Action can be manually reverted.

APP SERVICE PLAN DETAILS

Name: ASP-CloudPaaS-3e

Id: ASP-CloudPaaS-3e

Subscription:

Location: azure-East US

| | CURRENT | AFTER ACTIONS | |
|---------------------------------|----------|---------------|-------------|
| Plan Tier | I2v2 | I1v2 | |
| VMem, Capacity | 16 GB | 8 GB | ↓ 8 GB |
| VMem, P95th Utilization | 14% | 28% | ↑ 14 % |
| VCPU, Capacity | 34.1 GHz | 17.05 GHz | ↓ 17.05 GHz |
| VCPU, P95th Utilization | 2% | 4% | ↑ 2 % |
| Storage, Capacity | 0.98 TB | 0.98 TB | - |
| Storage, Utilization | 0% | 0% | - |
| Number Of Replicas, Capacity | 100 | 100 | - |
| Number Of Replicas, Utilization | 3% | 3% | - |

| | CURRENT | AFTER ACTIONS | |
|----------------|---------------|---------------|---------------|
| Compute Cost | \$563.56/mo | \$281.78/mo | ↓ \$281.78/mo |
| Instance Count | 3 | 3 | |
| Total Cost | \$1,690.68/mo | \$845.34/mo | ↓ \$845.34/mo |
| Total Savings | | \$845.34/mo | |

POLICIES

STORAGE AMOUNT AVG. UTILIZATION

Storage Amount Utilization average is equal to 0%

The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the App Service plan, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's scaling recommendations.

NOTE:

You can set scaling constraints in Virtual Machine Spec policies to refine the percentile calculations. For details, see [Scaling Sensitivity \(on page 295\)](#).

Disruptiveness and Reversibility of Scale Actions

Intersight Workload Optimizer always recommends scaling to a different instance type, so all scaling actions are disruptive and require downtime. You can reverse an action by scaling an App Service plan back to its original instance type.

Estimated On-demand Monthly Costs for Azure App Service Plans

Intersight Workload Optimizer considers a variety of factors when calculating estimated on-demand monthly costs for Azure App Service plans.

NOTE:

Azure App Service plans appear as Virtual Machine Spec entities in the supply chain.

Cisco Intersight Workload Optimizer Target Configuration and User Guide

291

Action Details
< 1 of 24 >
?
✕

Scale Virtual Machine Spec ASP-CloudPaaS-3e from I2v2 to I1v1 in ↓ \$845.34/mo SAVINGS

Underutilized vCPU, vMEM

VCPU PERCENTILE AND AVG. UTILIZATION

vCPU Utilization is below 2% for 95% of the time over the 30 day observation period

● vCPU Daily Avg. - - - Projected vCPU 95th Percentile
● vCPU 30 day 95th Percentile

ACTION ESSENTIALS

State ✔ Action can be accepted and executed immediately.

Non-Disruptive ✕ Downtime is required to execute.

Reversible ✔ Action can be manually reverted.

VMEM PERCENTILE AND AVG. UTILIZATION

vMEM Utilization is below 14% for 95% of the time over the 30 day observation period

● vMEM Daily Avg. - - - Projected vMEM 95th Percentile
● vMEM 30 day 95th Percentile

APP SERVICE PLAN DETAILS

Name ASP-CloudPaaS-3e

Id ASP-CloudPaaS-3e

Subscription

Location azure-East US

STORAGE AMOUNT AVG. UTILIZATION

Storage Amount Utilization average is equal to 0%

● Storage Amount Daily Avg. - - - Projected Avg. Storage Amount

RESOURCE IMPACT

| | CURRENT | AFTER ACTIONS | |
|---------------------------------|----------|---------------|-------------|
| Plan Tier | I2v2 | I1v1 | |
| VMem, Capacity | 16 GB | 8 GB | ↓ 8 GB |
| VMem, P95th Utilization | 14% | 28% | ↑ 14 % |
| VCPU, Capacity | 34.1 GHz | 17.05 GHz | ↓ 17.05 GHz |
| VCPU, P95th Utilization | 2% | 4% | ↑ 2 % |
| Storage, Capacity | 0.98 TB | 0.98 TB | - |
| Storage, Utilization | 0% | 0% | - |
| Number Of Replicas, Capacity | 100 | 100 | - |
| Number Of Replicas, Utilization | 3% | 3% | - |

COST IMPACT

| | CURRENT | AFTER ACTIONS | |
|----------------------|----------------------|--------------------|----------------------|
| Compute Cost | \$563.56/mo | \$281.78/mo | ↓ \$281.78/mo |
| Instance Count | 3 | 3 | |
| Total Cost | \$1,690.68/mo | \$845.34/mo | ↓ \$845.34/mo |
| Total Savings | | \$845.34/mo | |

POLICIES

Cost Calculation

The calculation for estimated on-demand monthly cost can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) * \text{Number of Instances} = \text{Estimated On-demand Monthly Cost}$$

Where:

- **On-demand Compute Rate** is the **hourly** cost for an App Service plan's instance type. You can obtain on-demand rates via [App Service Pricing](#).
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to calculate monthly costs.
- **Number of Instances** is the total number of VM instances underlying the App Service plan.

The listed items above impact cost calculations and the scaling decisions that Intersight Workload Optimizer makes. These decisions also rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending action to scale an Azure Service plan from the I2V2 to the I1V1 instance type.

| | Current Values | Values After Action Execution |
|------------------------|----------------|-------------------------------|
| On-demand Compute Rate | \$0.772/hr | \$0.386/hr |
| Number of Instances | 3 | 3 |

Intersight Workload Optimizer calculates the following:

- *Current* estimated on-demand monthly cost:

$$(\$0.772 * 730) * 3 = \$1690.68/\text{Mo.}$$

- Estimated on-demand monthly cost *after* executing the action:

$$(\$0.386 * 730) * 3 = \$845.34/\text{Mo.}$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

The estimated on-demand monthly cost is projected to decrease from \$1690.68/month to \$845.34/month, as shown in the Details section of the pending action.

| COST IMPACT ⓘ | CURRENT | AFTER ACTIONS | |
|-------------------|----------------------|--------------------|----------------------|
| Compute Cost ⓘ | \$563.56/mo | \$281.78/mo | ↓ \$281.78/mo |
| Instance Count ⓘ | 3 | 3 | |
| Total Cost | \$1,690.68/mo | \$845.34/mo | ↓ \$845.34/mo |
| Total Savings | | \$845.34/mo | |
| POLICIES | | | |

Intersight Workload Optimizer treats the action as a cost-saving measure, and shows total savings of \$845.34/month.

| COST IMPACT ⓘ | CURRENT | AFTER ACTIONS | |
|----------------------|---------------|--------------------|---------------|
| Compute Cost ⓘ | \$563.56/mo | \$281.78/mo | ↓ \$281.78/mo |
| Instance Count ⓘ | 3 | 3 | |
| Total Cost | \$1,690.68/mo | \$845.34/mo | ↓ \$845.34/mo |
| Total Savings | | \$845.34/mo | |
| POLICIES | | | |

Estimated On-demand Monthly Savings for Empty Azure App Service Plans

Intersight Workload Optimizer considers an empty App Service plan's on-demand compute rate and VM instance count when calculating the estimated on-demand monthly savings that you would realize when you delete the plan. A plan is considered empty if it is not hosting any running apps.

NOTE:

Azure App Service plans appear as Virtual Machine Spec entities in the supply chain.

Action Details
< 10 of 52 >
?
✕

Delete Empty P2v2 App Service Plan paas-asp-zone-redundant from ↓ \$147.46/mo SAVINGS

Increase savings

| | |
|---|--|
| <p>ACTION ESSENTIALS</p> <p>State 📄 Action can be accepted and executed immediately.</p> <p>Non-Disruptive ✓ Downtime is not required to execute.</p> <p>Reversible ✕ Action cannot be manually reverted.</p> | <p>APP SERVICE PLAN DETAILS</p> <p>Name paas-asp-zone-redundant 📄</p> <p>Id paas-asp-zone-redundant 📄</p> <p>Subscription 📄</p> <p>Location azure-East US 📄</p> <p>Days Empty 7</p> |
|---|--|

| | | | |
|------------------------|----------------|----------------------|--|
| RESOURCE IMPACT | CURRENT | AFTER ACTIONS | |
| Virtual Memory | 7 GB | - | |
| Virtual CPU | 18.36 GHz | - | |

| | | | |
|----------------------|----------------|----------------------|---------------|
| COST IMPACT ? | CURRENT | AFTER ACTIONS | |
| Compute Cost ⓘ | \$147.46/mo | N/A | - |
| Instance Count ⓘ | 1 | N/A | |
| Total Cost | \$147.46/mo | \$0.00/mo | ↓ \$147.46/mo |
| Total Savings | | \$147.46/mo | |

Savings Calculation

The calculation for estimated on-demand monthly savings can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) * \text{Number of Instances} = \text{Estimated On-demand Monthly Savings}$$

Where:

- **On-demand Compute Rate** is the **hourly** cost for an App Service plan's instance type.
You can obtain on-demand rates via [App Service Pricing](#).
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to calculate monthly savings.
- **Number of Instances** is the total number of VM instances underlying the App Service plan.

Example

Assume the following data for a pending action to delete an empty Azure Service plan on the P2V2 instance type.

| | Current Values |
|------------------------|----------------|
| On-demand Compute Rate | \$0.202/hr |
| Number of Instances | 1 |

Intersight Workload Optimizer calculates savings as follows:

$$(\$0.202 * 730) * 1 = \$147.46/\text{Mo.}$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

Intersight Workload Optimizer shows total savings of \$147.46/month.

| COST IMPACT ? | CURRENT | AFTER ACTIONS |
|------------------|-------------|-------------------------|
| Compute Cost ⓘ | \$147.46/mo | N/A |
| Instance Count ⓘ | 1 | N/A |
| Total Cost | \$147.46/mo | \$0.00/mo ↓ \$147.46/mo |
| Total Savings | | \$147.46/mo |

Virtual Machine Spec Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about Virtual Machine Spec actions, see [Virtual Machine Spec Actions \(on page 288\)](#).

| Action | Default Mode |
|-----------------------------|--------------|
| Cloud Scale All | Manual |
| Cloud Scale for Performance | Manual |
| Cloud Scale for Savings | Manual |
| Delete Virtual Machine Spec | Manual |

Scaling Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU and vMem. It then recommends actions to improve utilization based on the observed values for a given time period.

Cloud Instance Types

| Attribute | Default Value |
|----------------------|---------------|
| Cloud Instance Types | None |

By default, Intersight Workload Optimizer considers all instance types currently available for scaling when making scaling decisions for Virtual Machine Specs. However, you may have set up your Virtual Machine Specs to *only scale to* or *avoid* certain instance types to reduce complexity and cost, improve discount utilization, or meet application demand. Use this setting to identify the instance types that Virtual Machine Specs can scale to.

Click **Edit** to set your preferences. In the new page that displays, expand a **cloud tier** (a family of instance types, such as *Basic*) to see individual instance types and the resources allocated to them.

Select your preferred instance types or cloud tiers, or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

If you selected a cloud tier and the service provider deploys new instance types to that tier later, then those instance types will automatically be included in your policy. Be sure to review your policies periodically to see if new instance types have been added to a tier. If you do not want to scale to those instance types, update the affected policies.

Scaling Target Utilization

The utilization that you set here specifies the percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|-----------|---------------|
| VCPU | 70 |
| VMEM | 90 |
| Storage | 90 |

These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Intersight Workload Optimizer calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Intersight Workload Optimizer recommends actions, in most cases you should never need to use them. If you want to control how Intersight Workload Optimizer recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

Database Server (Cloud)

Intersight Workload Optimizer represents the following public cloud resources as Database Server entities in the supply chain:

- [AWS RDS \(on page 296\)](#)
- [Azure Cosmos DB Accounts \(on page 307\)](#)

NOTE:

Azure SQL databases, dedicated SQL pools, and Cosmos DB databases appear as *Database* entities in the supply chain. For details, see [Database \(Cloud\) \(on page 308\)](#).

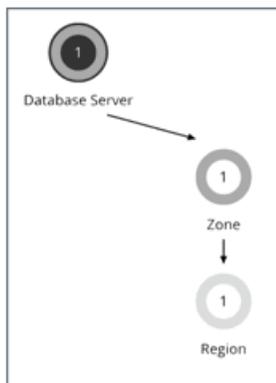
Create groups in Intersight Workload Optimizer so you can manage AWS RDS and Azure Cosmos DB accounts with ease. To create groups, go to **Settings > Groups**, and then select **New Group > Database Server**. In the page that displays, choose **Dynamic** (recommended) and then set the `Cloud Provider` filter as follows:

- `AWS` - use to create a group of AWS RDS Database Servers
- `Azure` - use to create a group of Azure Cosmos DB accounts

Database Server - AWS RDS

In AWS public cloud environments, a Database Server is a relational database that you have configured using AWS Relational Database Service (RDS). Intersight Workload Optimizer discovers RDS instances through your AWS targets, and then generates scaling actions as needed.

Synopsis



AWS RDS

| | |
|---------------------|--|
| Synopsis | |
| Provides: | Database services to cloud applications and end users |
| Consumes: | Compute and storage resources in the availability zone |
| Discovered through: | AWS targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Virtual Memory (VMem)**
Virtual Memory is the measurement of memory that is in use.
- **Virtual CPU (VCPU)**
Virtual CPU is the measurement of CPU that is in use.
- **Storage Amount**
Storage Amount is the measurement of storage capacity that is in use.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- **DB Cache Hit Rate**
DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.
- **Connection**
Connection is the measurement of database connections utilized by applications.

Actions

Intersight Workload Optimizer supports the following actions:

- **Scale**
Scale compute and storage resources to optimize performance and costs.

Scale Actions for Database Servers

To recommend accurate scaling actions, Intersight Workload Optimizer analyzes resource utilization percentiles and collects relevant metrics (such as connections utilization) from AWS. It also takes into consideration constraints defined in [policies \(on page 304\)](#).

Consider the following scenarios and actions:

- To address vCPU congestion, Intersight Workload Optimizer can recommend scaling a Database Server to the instance type that can adequately meet demand at the lowest possible cost. If vCPU is underutilized, it can recommend scaling to a smaller instance type.
- To address IOPS congestion, Intersight Workload Optimizer can recommend increasing provisioned IOPS or scaling the Database Server to a different storage type. For gp2 storage, it can recommend increasing disk size to increase provisioned IOPS. After executing these actions, Intersight Workload Optimizer will not recommend new actions for the next six hours, in compliance with AWS's "cooldown" period for EBS storage.
- Intersight Workload Optimizer analyzes DB cache hit rate before making vMem scaling decisions. To perform its analysis, it collects cache hit rate metrics for Database Servers with [Performance Insights](#) enabled.

For Database Servers with cache hit rate metrics, Intersight Workload Optimizer considers at least 90% cache hit rate to be optimal. This percentage value is not configurable.

- A cache hit rate value equal to or greater than 90% indicates efficiency. For this reason, Intersight Workload Optimizer will not recommend an action even if vMem utilization is high. If vMem utilization is low, it will recommend scaling to a smaller instance type.
- When the cache hit rate is below 90%, Intersight Workload Optimizer will also not recommend an action, provided that vMem utilization remains low. If vMem utilization is high, then it will recommend scaling to a larger instance type.

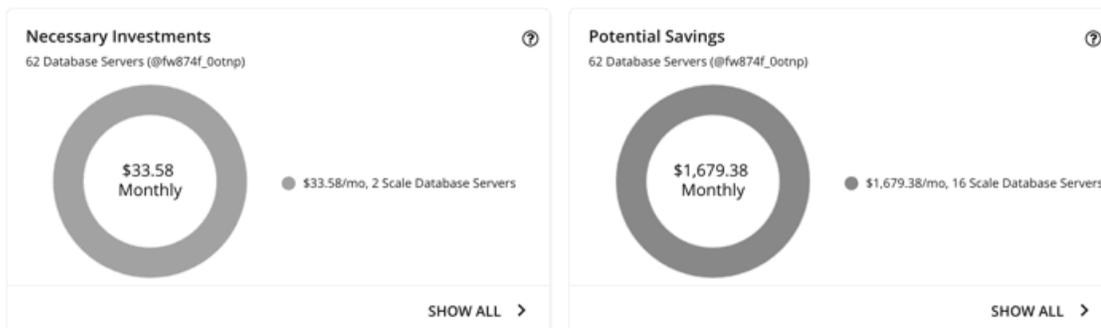
Notes on Performance Insights and cache hit rate metrics:

- Performance Insights is enabled by default on a majority of AWS Database Servers. In the Intersight Workload Optimizer user interface, you can use Search and then apply the Performance Insights filter to see which Database Servers have Performance Insights enabled.
- If Performance Insights is disabled or is not supported for your AWS Database Server engines or regions, Intersight Workload Optimizer will not have cache hit rate metrics to analyze and will therefore not generate actions in direct response to vMem utilization. For a list of supported engines and regions, see this [AWS page](#).
- An action to scale to a different instance type in response to vCPU utilization might also include vMem changes, but vMem utilization alone (without cache hit rate metrics) will not drive actions.

Intersight Workload Optimizer also considers Connections utilization and capacity when making scaling decisions. It collects utilization metrics from CloudWatch and calculates capacity based on the maximum number of simultaneous connections configured for the Database Server. The maximum number varies by Database Server engine type and memory allocation, and is set in the [parameter group](#) associated with a Database Server. Intersight Workload Optimizer currently supports Database Servers associated with parameter groups that use [default values](#). For example, consider a MySQL Database Server that is on a `db.t3.large` instance type with 8 GB (8589934592 bytes) of memory, and is associated with a parameter group that uses the default value `{DBInstanceClassMemory/12582880}`. In this case, Intersight Workload Optimizer calculates capacity as 682 connections (or `{8589934592/12582880}`). When Intersight Workload Optimizer generates an action due to vMem underutilization and sees that Connections utilization is only 15% of capacity (or roughly 100 connections), it picks a smaller instance type that is adequate for both the vMem and Connections requirements of the Database Server. For example, it could pick `db.t2.small`, which provides 2 GB of memory and a maximum of 170 connections.

Scale Actions in Charts

Use the Necessary Investments and Potential Savings charts to view pending Database Server actions. These charts show total monthly investments and savings, assuming you execute all the actions.



Click **Show All** for each chart to review and execute the actions.

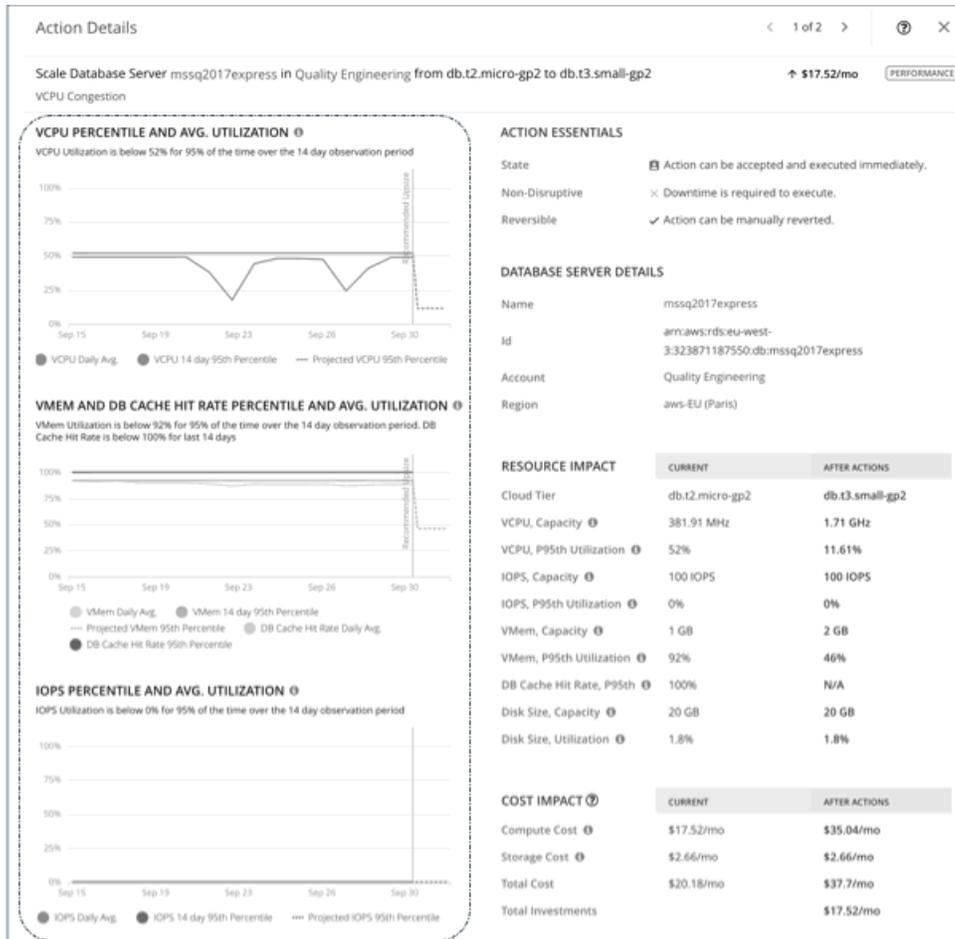
The table lists all the actions that are pending for Database Servers, and the savings or investments for each action.

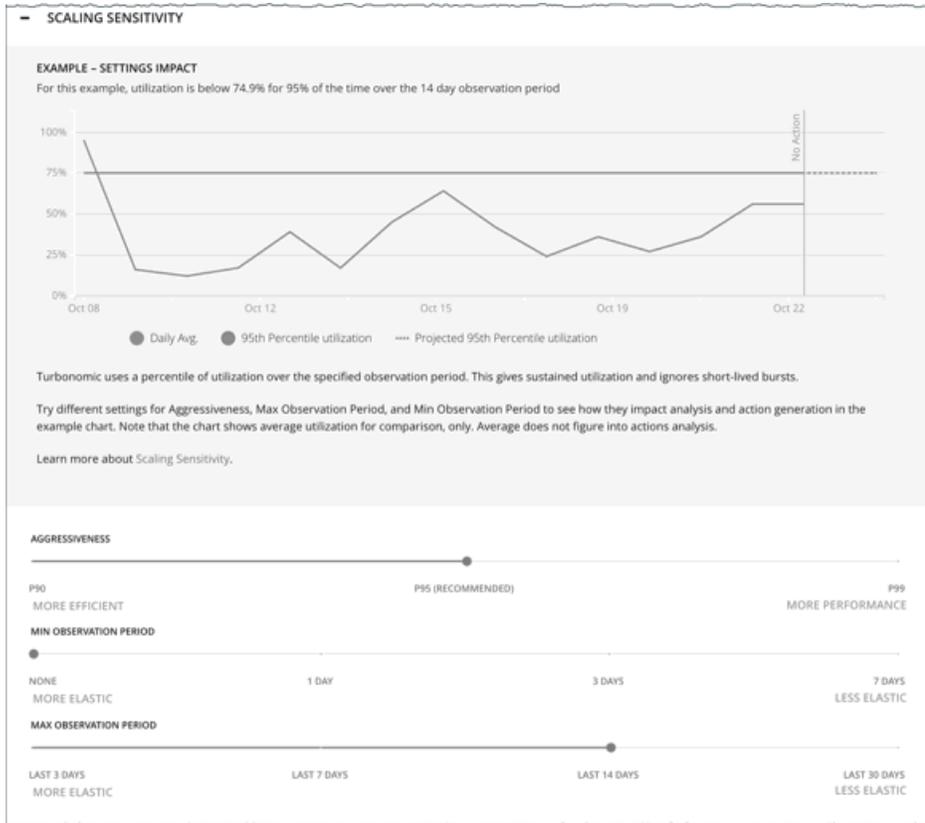
| Scale Actions 16 Savings \$1,679.38/mo | | | | | | | | | | EXECUTE ACTIONS | |
|--|---------|----------------|------------|------------------|-----------|----------------------|---------------|-----------------|---------------|-----------------|--|
| Database Server Name | Account | Non-Disrupt... | Reversible | Instance Type | On-Demand | New Instance Type | New On-Demand | Action Category | Savings | Action | |
| <input type="checkbox"/> rds-mariamultiatz | | × | ✓ | db.t3.small-io1 | \$0.924/h | db.t3.small-stand... | \$0.095/h | SAVINGS | ↓ \$604.80/rr | DETAILS | |
| <input type="checkbox"/> testioautoscaling | | × | ✓ | db.t3.micro-io1 | \$0.446/h | db.t3.micro-gp2 | \$0.033/h | SAVINGS | ↓ \$301.00/rr | DETAILS | |
| <input type="checkbox"/> rds-maria-io1 | | × | ✓ | db.t3.micro-io1 | \$0.418/h | db.t3.micro-stand... | \$0.031/h | SAVINGS | ↓ \$282.50/rr | DETAILS | |
| <input type="checkbox"/> btc-dbs-1 | | × | ✓ | db.m5.xlarge-io1 | \$0.514/h | db.r6g.large-stan... | \$0.219/h | SAVINGS | ↓ \$215.20/rr | DETAILS | |

For details on how Intersight Workload Optimizer calculates savings or investments, see [Estimated On-demand Costs for Cloud Database Servers \(on page 302\)](#).

Utilization Charts for Scale Actions

Intersight Workload Optimizer uses percentile calculations to measure resource utilization more accurately, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scale action on a Database Server, you will see charts that highlight *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.





For details, see [Scaling Sensitivity \(on page 305\)](#).

Non-disruptive and Reversible Scaling Actions

Intersight Workload Optimizer indicates whether a pending action is non-disruptive or reversible to help you decide how to handle the action.

| Scale Actions 16 Savings \$1,679.38/mo | | | | | |
|---|---------|----------------|------------|------------------|----------------|
| Type to search | | | | | |
| Database Server Name | Account | Non-Disruptive | Reversible | Instance Type | On-Demand Cost |
| <input type="checkbox"/> rds-mariamulti-az | | ✗ | ✓ | db.t3.small-io1 | \$0.924/h |
| <input type="checkbox"/> testioautoscalingenabled | | ✗ | ✓ | db.t3.micro-io1 | \$0.446/h |
| <input type="checkbox"/> rds-maria-io1 | | ✗ | ✓ | db.t3.micro-io1 | \$0.418/h |
| <input type="checkbox"/> btc-dbs-1 | | ✗ | ✓ | db.m5.xlarge-io1 | \$0.514/h |

The following table describes the disruptiveness and reversibility of the actions that Intersight Workload Optimizer recommends:

| Action | Disruptive | Reversible |
|--|------------|------------|
| Scaling to a different instance type | Yes | Yes |
| Scaling up storage amount | No | No |
| Scaling up storage access (provisioned IOPS) | No | Yes |

| Action | Disruptive | Reversible |
|--|------------|------------|
| Scaling to a different storage type + Scaling up storage amount | Yes | No |
| Scaling to a different storage type + Scaling up storage access (provisioned IOPS) | Yes | Yes |
| Scaling to a different storage type + Scaling up storage amount + Scaling up storage access (provisioned IOPS) | Yes | No |

You can set action acceptance modes in policies to specify the degree of automation for these actions.

Configure Database Server Policy

- + SCOPE ⓘ
- + POLICY SCHEDULE ⓘ
- AUTOMATION AND ORCHESTRATION
 - Database Server Scaling Actions
Generate cloud database server scaling actions
 - DISRUPTIVE IRREVERSIBLE SCALING**
AWS RDS storage amount scale and instance type or storage tier change
Action Acceptance: Recommend
 - DISRUPTIVE REVERSIBLE SCALING**
AWS RDS instance type scaling, storage tier changes
Action Acceptance: Recommend
 - NON-DISRUPTIVE IRREVERSIBLE SCALING**
AWS RDS storage amount scaling
Action Acceptance: Recommend
 - NON-DISRUPTIVE REVERSIBLE SCALING**
AWS RDS provisioned IOPS scaling
Action Acceptance: Recommend

Unavailable Instance Types

A scale action could fail if the target instance type is unavailable in the availability zone for some reason. Your AWS environment might show the instance type as available, but when the scaling action executes, the following error displays in AWS:

```
Cannot modify the instance class because there are no instances of the requested class available in the current instance's availability zone. Please try your request again at a later time.
```

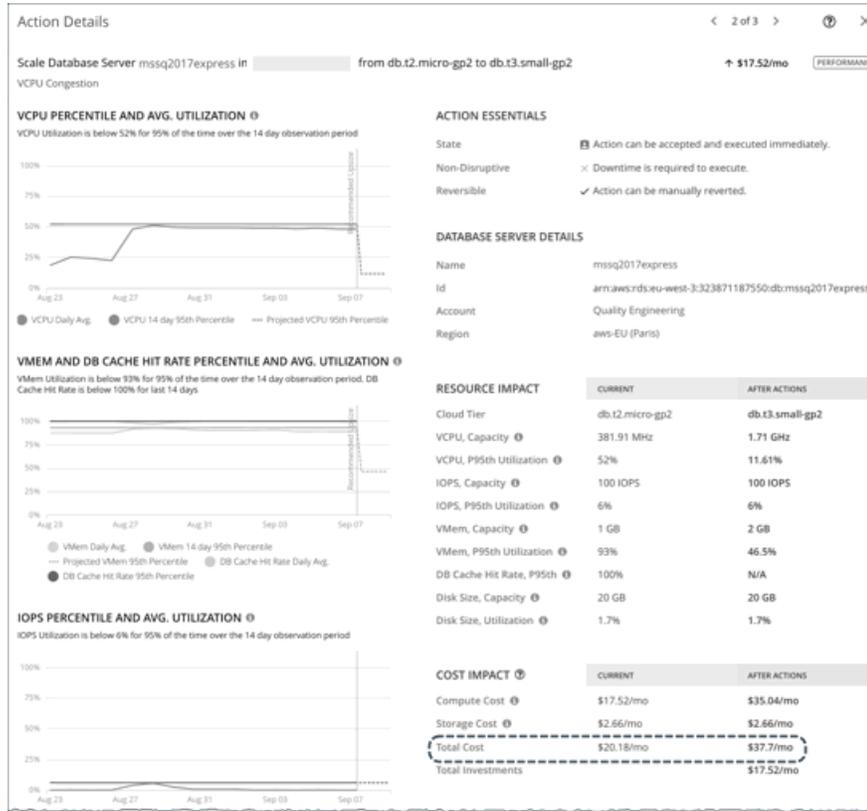
NOTE:

For details about this error, see this [AWS page](#).

When this error occurs, Intersight Workload Optimizer modifies the default Database Server policy to exclude the instance type from its scaling list. When the Database Server is available again, Intersight Workload Optimizer adds it back to the scaling list. For details about this list, see [Cloud Instance Types \(on page 306\)](#).

Estimated On-demand Costs for AWS RDS Database Servers

Intersight Workload Optimizer considers a variety of factors when calculating *Estimated On-demand Monthly Cost* for an AWS RDS Database Server.



Non-Aurora Database Servers

Cost Calculation

For non-Aurora Database Servers, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) + (\text{Provisioned Database Storage Rate} * \text{Provisioned Database Storage Amount}) + (\text{Provisioned IOPS Rate} * \text{Provisioned IOPS Amount}) = \text{Estimated On-demand Monthly Cost}$$

Where:

- **On-demand Compute Rate** is the hourly cost for a Database Server's instance type
 You can obtain on-demand rates via [Amazon RDS Pricing](#).
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.
- **Provisioned Database Storage Rate** is the hourly cost for a Database Server's provisioned database storage
 You can obtain provisioned database storage rates via [Amazon RDS Pricing](#).
- **Provisioned IOPS Rate** is the monthly cost for a Database Server's provisioned IOPS
 Provisioned IOPS apply only to Database Servers on Provisioned IOPS SSD (io1) storage. You can obtain information about Provisioned IOPS SSD storage via the [RDS User Guide](#).
 You can obtain provisioned IOPS rates via [Amazon RDS Pricing](#).

The listed items above impact cost calculations and the scaling decisions that Intersight Workload Optimizer makes. These decisions also rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for SQL Server Express Edition (Single A-Z deployment):

| | Current Values | Values After Action Execution |
|-------------------------------------|----------------|-------------------------------|
| On-demand Compute Rate | \$0.024/hr | \$0.048/hr |
| Provisioned Database Storage Rate | \$0.133/hr | \$0.133/hr |
| Provisioned Database Storage Amount | 20 GB | 20 GB |
| Provisioned IOPS Rate | \$0.00 | \$0.00 |
| Provisioned IOPS Amount | 0 | 0 |

Intersight Workload Optimizer calculates the following:

- **Current** Estimated On-demand Monthly Cost:

$$(0.024 * 730) + (0.133 * 20) + (0.00 * 0) = 20.18$$

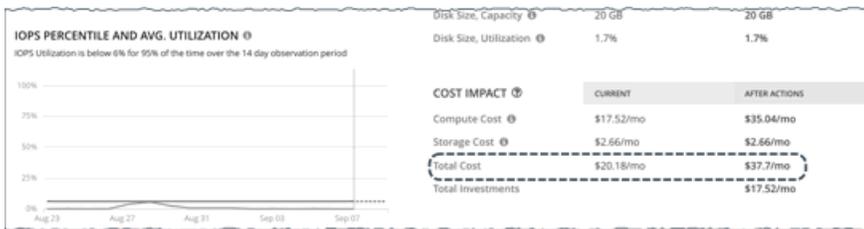
- Estimated On-demand Monthly Cost *after* executing the action:

$$(0.048 * 730) + (0.133 * 20) + (0.00 * 0) = 37.7$$

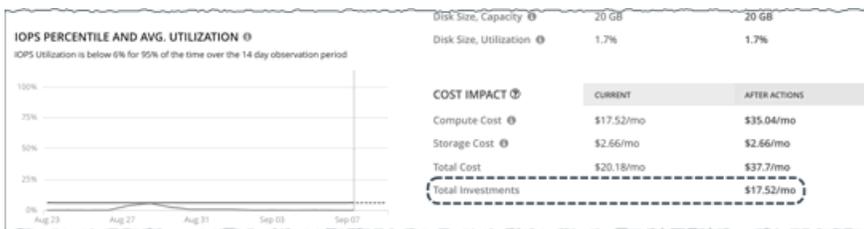
NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

The Estimated On-demand Monthly Cost is projected to increase from \$20.18/month to \$37.7/month, as shown in the Details section of the pending action.



Intersight Workload Optimizer treats the action as an investment and shows an estimated investment of \$17.52/month.



Aurora Database Servers

Cost Calculation

For Aurora Database Servers, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) + (\text{Provisioned Database Storage Rate} * \text{Provisioned Database Storage Amount}) + (\text{I/O Request Rate} * (\text{Hourly Billed I/O Operation Count} * 730)) = \text{Estimated On-demand Monthly Cost}$$

Where:

- **On-demand Compute Rate** is the hourly cost for a Database Server's instance type
You can obtain on-demand rates via [Amazon Aurora Pricing](#).
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.
- **Provisioned Database Storage Rate** is the hourly cost for a Database Server's provisioned database storage
You can obtain provisioned database storage rates via [Amazon Aurora Pricing](#).

- **I/O Request Rate** is the cost per one million read/write I/O requests

You can obtain I/O request rates via [Amazon Aurora Pricing](#).

- **Hourly Billed I/O Operation Count** is the average sum of read and write I/O operations per hour over the last month

The listed items above impact cost calculations. Except for I/O request rate, these items affect the actual scaling decisions that Intersight Workload Optimizer makes. These decisions also rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for Aurora MySQL-Compatible Edition:

| | Current Values | Values After Action Execution |
|-------------------------------------|-----------------------------|-------------------------------|
| On-demand Compute Rate | \$0.164/hr | \$0.041/hr |
| Provisioned Database Storage Rate | \$0.10/hr | \$0.10/hr |
| Provisioned Database Storage Amount | 100 | 100 |
| I/O Request Rate | \$0.20/one million requests | \$0.20/one million requests |
| Hourly Billed I/O Operation Count | 2000 | 2000 |

Intersight Workload Optimizer calculates the following:

- **Current** Estimated On-demand Monthly Cost:

$$(0.164 * 730) + (0.10 * 100) + ((0.20 / 1000000) * (2000 * 730)) = 130.01$$

- Estimated On-demand Monthly Cost *after* executing the action:

$$(0.041 * 730) + (0.10 * 100) + ((0.20 / 1000000) * (2000 * 730)) = 40.22$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

Since the Estimated On-demand Monthly Cost is projected to decrease from \$130.01/month to \$40.22/month, Intersight Workload Optimizer treats the action as a cost-saving measure and shows estimated savings of \$89.79/month.

AWS RDS Database Server Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about cloud Database Server actions, see [AWS RDS Database Server Actions \(on page 297\)](#) and [Non-disruptive and Reversible Scaling Actions \(on page 300\)](#).

| Action | Default Setting/Mode |
|-------------------------------------|----------------------|
| Database Server Scaling Actions | On |
| Disruptive irreversible scaling | Recommend |
| Disruptive reversible scaling | Recommend |
| Non-disruptive irreversible scaling | Recommend |
| Non-disruptive reversible scaling | Recommend |

Scaling Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU, vMem, DB Cache Hit Rate, and IOPS. It then recommends actions to improve utilization based on the observed values for a given time period.

■ Aggressiveness

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce capacity. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce resources for that Database Server.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single burst to 100%, but for 95% of the samples, utilization never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce resource allocation.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 99th Percentile – More performance. Recommended for critical Database Servers that need maximum guaranteed performance at all times, or those that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.
- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for Database Servers that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 14 days. Use the **Max Observation Period** setting to adjust the number of days. To ensure that there are enough samples to analyze and drive scaling actions, set the **Min Observation Period**.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 14 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the Database Server has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 7 Days or Last 3 Days

Intersight Workload Optimizer recommends an observation period of 14 days so it can recommend scaling actions more often. Since Database Server scaling is minimally disruptive, scaling often should not introduce any noticeable performance risks.

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic - None
- Less Elastic - 1, 3, or 7 Days

Cloud Instance Types

By default, Intersight Workload Optimizer considers all instance types currently available for scaling when making scaling decisions for Database Servers. However, you may have set up your Database Servers to *only scale to* or *avoid* certain instance types to reduce complexity and cost, or meet demand. Use this setting to identify the instance types that Database Servers can scale to.

NOTE:

Intersight Workload Optimizer automatically discovers and enforces Database Server tier exclusions configured in your AWS environment. You do not need to configure these tier exclusions in policies. To see a list of tier exclusions that are currently enforced, set the scope to one or several Database Servers and click the **Policies** tab.

| Attribute | Default Value |
|----------------------|---------------|
| Cloud Instance Types | None |

Click **Edit** to set your preferences. In the new page that displays, expand a **cloud tier** (a family of instance types, such as *db.m1*) to see individual instance types and the resources allocated to them.

Select your preferred instance types or cloud tiers, or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

NOTE:

This policy setting is not available in plans.

If you selected a cloud tier and the service provider deploys new instance types to that tier later, then those instance types will automatically be included in your policy. Be sure to review your policies periodically to see if new instance types have been added to a tier. If you do not want to scale to those instance types, update the affected policies.

Scaling Target Utilization

This is the target utilization as a percentage of capacity.

| Attribute | Default Value |
|----------------|---------------|
| VCPU | 70 |
| VMEM | 90 |
| IOPS | 70 |
| Storage Amount | 90 |

These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Intersight Workload Optimizer calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Intersight Workload Optimizer recommends actions, in most cases you should never need to use them. If you want to control how Intersight Workload Optimizer recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

Database Server - Azure Cosmos DB Account

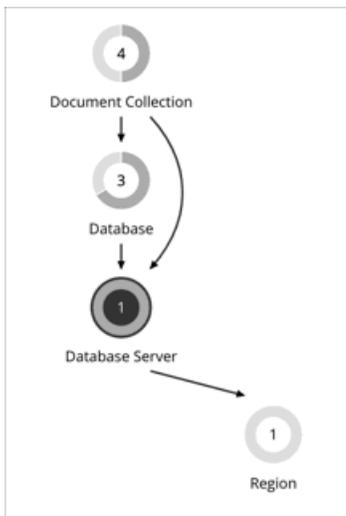
In [Azure Cosmos DB \(on page 106\)](#), an account is an entity that manages [databases \(on page 317\)](#) and containers ([document collections \(on page 322\)](#)). An account has a total throughput limit that represents capacity, but actual throughput (measured in Request Units or RUs) is provisioned at the database or container level.

Intersight Workload Optimizer discovers Azure Cosmos DB accounts through your Azure targets, and represents them as database server entities in the supply chain.

NOTE:

Intersight Workload Optimizer does not discover Cosmos DB serverless accounts.

Synopsis



| Synopsis | |
|---------------------|--|
| Provides: | RUs to Cosmos DB databases and containers (document collections) |
| Consumes: | Storage resources |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Request Unit (RU)

Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.
- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

Actions

None

Intersight Workload Optimizer does not recommend actions for a Cosmos DB account but it does recommend actions for the [databases \(on page 317\)](#) and [document collections \(on page 322\)](#) in the account.

Database (Cloud)

Intersight Workload Optimizer represents the following public cloud resources as Database entities in the supply chain:

- [Azure SQL databases \(vCore or DTU pricing model\) \(on page 308\)](#)
- [Azure dedicated SQL pool \(on page 314\)](#)
- [Azure Cosmos DB databases \(on page 317\)](#)

NOTE:

AWS RDS databases appear as *Database Server* entities in the supply chain. For details, see [Database Server - AWS RDS \(on page 296\)](#).

Create groups in Intersight Workload Optimizer so you can manage SQL databases, dedicated SQL pools, and Cosmos DB databases with ease. When you set the scope to a particular group, you can evaluate pending actions, view metrics, and assign policies as usual. You can also view the status of individual members. For example, you can see if a dedicated SQL pool is currently active or suspended.

To create groups, go to **Settings > Groups**, and then click **New Group > Database**. In the page that displays, choose **Dynamic** (recommended) and then set the **Pricing Model** filter as follows:

- DTU and/or vCore - use to create a group of SQL databases
- DWU - use to create a group of dedicated SQL pools
- RU - use to create a group of Cosmos DB databases

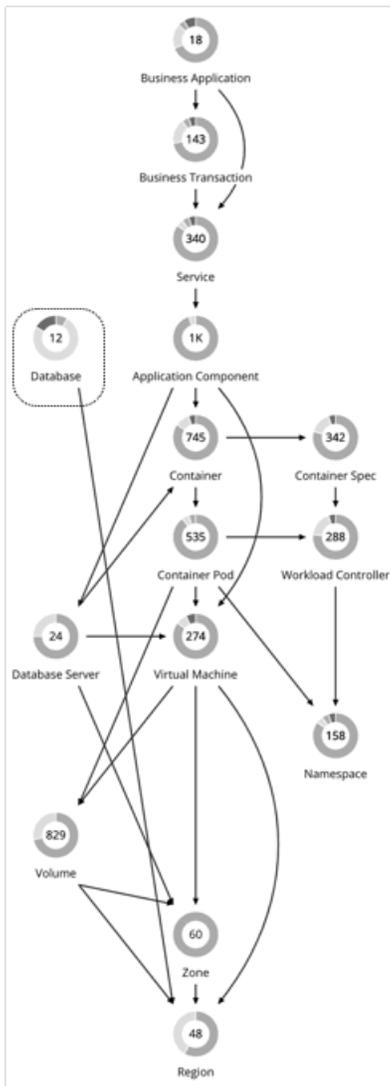
Database - Azure SQL

Azure SQL database is an individual database that is managed under the DTU (Database Transaction Unit) or vCore [pricing model](#).

- DTU Pricing Model
In the DTU model, Azure bundles CPU, memory, and IOPS as a single DTU metric. Intersight Workload Optimizer actions on these databases consider both DTU and storage utilization.
- vCore Pricing Model
In the vCore model, analysis can track CPU, memory, IOPS, and throughput metrics in isolation. Intersight Workload Optimizer actions on these databases are driven by CPU, memory, IOPS, throughput and storage utilization.

Intersight Workload Optimizer discovers SQL databases through your Azure targets, and represents them as Database entities in the supply chain.

Synopsis



Synopsis

| | |
|---------------------|--|
| Provides: | Transactions to end users |
| Consumes: | <ul style="list-style-type: none"> ■ DTU Pricing Model: DTU and storage resources in an Azure region ■ vCore Pricing Model: vCPU, vMem, IOPS, throughput, and storage resources in an Azure region |
| Discovered through: | Azure targets |

Monitored Resources

The resources that Intersight Workload Optimizer can monitor depend on the pricing model in place for the given database entity.

- DTU Pricing Model
 - DTU

DTU is the measurement of compute capacity for the database. DTU represents CPU, memory, and IOPS/IO Throughput bundled as a single commodity.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- vCore Pricing Model

- Virtual Memory (VMem)

Virtual Memory is the measurement of memory that is in use.

- Virtual CPU (VCPU)

Virtual CPU is the measurement of CPU that is in use.

- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.

- Storage Access (IOPS)

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

- I/O Throughput

I/O Throughput is the measurement of an entity's throughput to the underlying storage.

Scale Actions for SQL Databases

Intersight Workload Optimizer database scaling actions aim to increase resource utilization and reduce costs while complying with business policies.

Intersight Workload Optimizer drives scaling actions based on the utilization of resources, and treats the following limits as constraints when it makes scaling decisions:

- Maximum concurrent sessions

This is the maximum number of database connections at a time.

- Maximum concurrent workers

This is the maximum number of database processes that can handle queries at a time.

Points to consider:

- Intersight Workload Optimizer will *not* recommend:

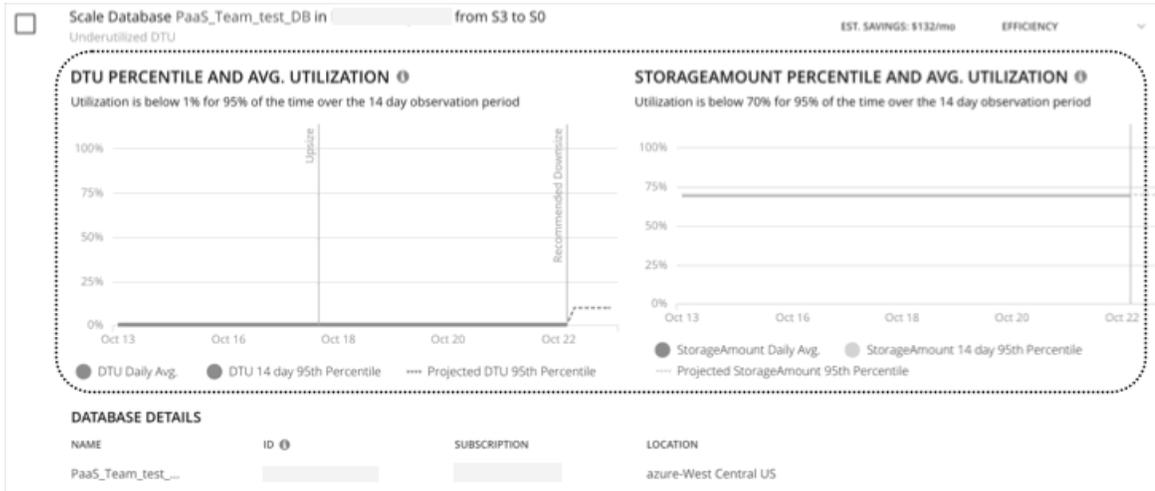
- Scaling from one pricing model to another
- Scaling vCore databases to instance types running Gen4 hardware. This hardware generation has been retired and pricing information can no longer be retrieved via the Azure API.
- Scaling vCore databases on the [serverless compute tier](#)
- Scaling provisioned memory for vCore databases on the Hyperscale service tier. VMem utilization data is currently unavailable for Hyperscale due to an issue in the Azure API.

- On DTU databases, a single action can scale both DTU and storage. On vCore databases, a single action can scale vCPU, vMem, IOPS, throughput, and storage.

- In some cases, Intersight Workload Optimizer might recommend scaling up storage, even if there is no storage pressure on the database, to take advantage of storage provided at no extra cost. For example, Intersight Workload Optimizer might recommend scaling from the S3 to the S0 tier because of low DTU and storage utilization. Since the S0 tier includes 250 GB of storage at no extra cost, Intersight Workload Optimizer will also recommend scaling up to this storage amount. If you want to scale DTU but keep the storage amount unchanged, adjust the values for aggressiveness (percentile) and observation period in your database policies.

Utilization Charts for Scale Actions

Intersight Workload Optimizer uses percentile calculations to measure resource utilization, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on a database, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.



The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the database, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's scaling recommendations.

NOTE:

You can set scaling constraints in database policies to refine the percentile calculations. For details, see [Aggressiveness and Observation Period \(on page 319\)](#).

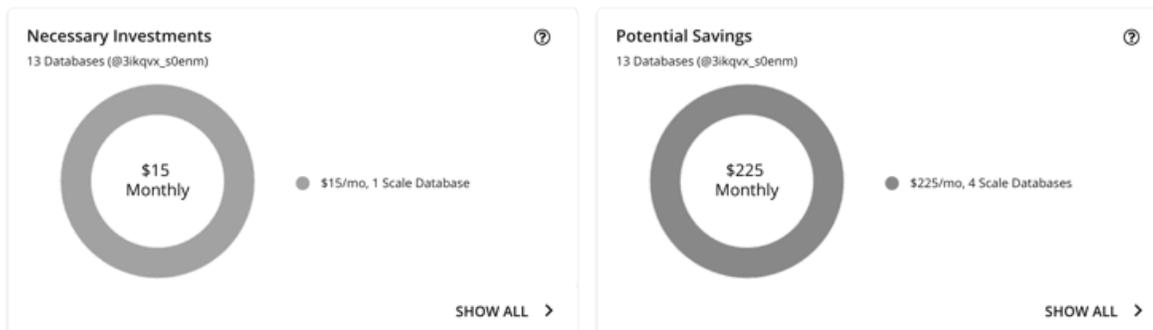
Non-disruptive and Reversible Scaling Actions

All scaling actions shown in the Action Center view and Action Details page are non-disruptive and reversible.

For actions to scale vCore databases from General Purpose or Business Critical to Hyperscale, there are certain caveats associated with reversing such actions. To learn more, see the [Azure documentation](#).

Actions in Charts

Use the Necessary Investments and Potential Savings charts to view pending database actions. These charts show total monthly investments and savings, assuming you execute all the actions.



Click **Show All** for each chart to review and execute the actions.

The table shows the following:

- Actions that are pending for each database
- Savings or investments for each database

Estimated On-demand Costs for Azure SQL Databases

Intersight Workload Optimizer considers a variety of factors when calculating *Estimated On-demand Monthly Cost* for an Azure SQL Database.

| RESOURCE IMPACT | CURRENT | AFTER ACTIONS |
|-----------------------------------|---------|---------------|
| Cloud Tier | S3 | S0 |
| DTU, Capacity | 100 | 10 |
| DTU, P95th Utilization | 1% | 10% |
| Storage Amount, Capacity | 300 GB | 250 GB |
| Storage Amount, P95th Utilization | 1% | 1.2% |

| COST IMPACT | CURRENT | AFTER ACTIONS |
|----------------------|--------------------|--------------------|
| Compute Cost | \$146.91/mo | \$14.69/mo |
| Storage Cost | \$11.05/mo | \$0.00/mo |
| Total Cost | \$157.96/mo | \$14.69/mo |
| Total Savings | | \$143.27/mo |

Azure SQL DTU Databases

Cost Calculation

For Azure SQL DTU Databases, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) + (\text{Provisioned Database Storage Rate} * (\text{Provisioned Database Storage Amount} - \text{Performance Level Included Storage})) = \text{Estimated On-demand Monthly Cost}$$

Where:

- On-demand Compute Rate** is the **hourly** cost for a Database's instance type
 You can obtain on-demand rates via [Azure SQL Database Pricing](#).
- 730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.
- Provisioned Database Storage Rate** is the cost for 1 GB / mo. of a Database's provisioned storage
 You can obtain provisioned database storage rates via [Azure SQL Database Pricing](#).
- Performance Level Included Storage** is the storage amount included in the price of the selected Performance Level of a DTU Database
 You can obtain information on DTU storage limits via [DTU Storage Limits](#).

The listed items above impact cost calculations and the scaling decisions that Intersight Workload Optimizer makes. These decisions also rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for an Azure SQL DTU Database:

| | Current Values | Values After Action Execution |
|-------------------------------------|----------------------|-------------------------------|
| On-demand Compute Rate | \$0.20125/hr | \$0.020125/hr |
| Provisioned Database Storage Rate | \$0.221 per 1 GB/Mo. | \$0.221 per 1 GB/Mo. |
| Performance Level Included Storage | 250 GB | 250 GB |
| Provisioned Database Storage Amount | 300 GB | 250 GB |

Intersight Workload Optimizer calculates the following:

- **Current** Estimated On-demand Monthly Cost:

$$(\$0.20125 * 730) + (\$0.221 * (300 - 250)) = \$157.96/\text{Mo.}$$

- Estimated On-demand Monthly Cost *after* executing the action:

$$(\$0.020125 * 730) + (\$0.221 * (250 - 250)) = \$14.69/\text{Mo.}$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

The Estimated On-demand Monthly Cost is projected to decrease from \$157.96/month to \$14.69/month, as shown in the Details section of the pending action.

| COST IMPACT ? | CURRENT | AFTER ACTIONS |
|----------------|-------------|---------------|
| Compute Cost ⓘ | \$146.91/mo | \$14.69/mo |
| Storage Cost ⓘ | \$11.05/mo | \$0.00/mo |
| Total Cost | \$157.96/mo | \$14.69/mo |
| Total Savings | | \$143.27/mo |

Intersight Workload Optimizer treats the action as a saving, and shows an estimated savings of \$143.27/month.

| COST IMPACT ? | CURRENT | AFTER ACTIONS |
|----------------|-------------|---------------|
| Compute Cost ⓘ | \$146.91/mo | \$14.69/mo |
| Storage Cost ⓘ | \$11.05/mo | \$0.00/mo |
| Total Cost | \$157.96/mo | \$14.69/mo |
| Total Savings | | \$143.27/mo |

Azure SQL vCore Databases

Cost Calculation

For Azure SQL vCore Databases, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

$$(\text{On-demand Compute Rate} * 730) + (\text{SQL License Rate} * 730) + (\text{Provisioned Database Storage Rate} * (\text{Provisioned Database Storage Amount} + \text{Log Space Allocated})) = \text{Estimated On-demand Monthly Cost}$$

Where:

- **On-demand Compute Rate** is the hourly cost for a Database's instance type
You can obtain on-demand rates via [Azure SQL Database Pricing](#).
- **730** represents the number of hours per month that Intersight Workload Optimizer uses to estimate monthly costs.
- **SQL License Rate** is the hourly cost for a Database's SQL license
You can obtain SQL license rates via [Azure SQL Database Pricing](#).
Note: "Pay as you go" prices in the link above represent the sum of compute and license costs, while "Azure Hybrid Benefit Price" values represent compute costs only.
- **Provisioned Database Storage Rate** is the cost for 1 GB / mo. of a Database's provisioned storage
You can obtain provisioned database storage rates via [Azure SQL Database Pricing](#).
- **Log Space Allocated** is the log storage space automatically allocated to single Database instance by Azure.

Note: Log storage space is considered in database cost calculations, but not reflected in Storage capacity.

You can obtain provisioned database storage rates via [Azure SQL Database Pricing](#).

The listed items above impact cost calculations and the scaling decisions that Intersight Workload Optimizer makes. These decisions also rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for an Azure SQL vCore Database:

| | Current Values | Values After Action Execution |
|-------------------------------------|----------------|-------------------------------|
| On-demand Compute Rate | \$1.068/hr | \$0.304/hr |
| SQL License Rate | \$0.799728/hr | \$0.199932/hr |
| Provisioned Database Storage Rate | \$0.115/hr | \$0.115/hr |
| Provisioned Database Storage Amount | 32 GB | 5 GB |

Intersight Workload Optimizer calculates the following:

- **Current** Estimated On-demand Monthly Cost:

$$(\$1.068 * 730) + (\$0.799728 * 730) + (\$0.115 * (32 + 9.6)) = \$1368.23/\text{Mo.}$$

- Estimated On-demand Monthly Cost *after* executing the action:

$$(\$0.304 * 730) + (\$0.199932 * 730) + (\$0.115 * (5 + 1.5)) = \$368.62/\text{Mo.}$$

NOTE:

Intersight Workload Optimizer rounds the calculated values that it displays in the user interface.

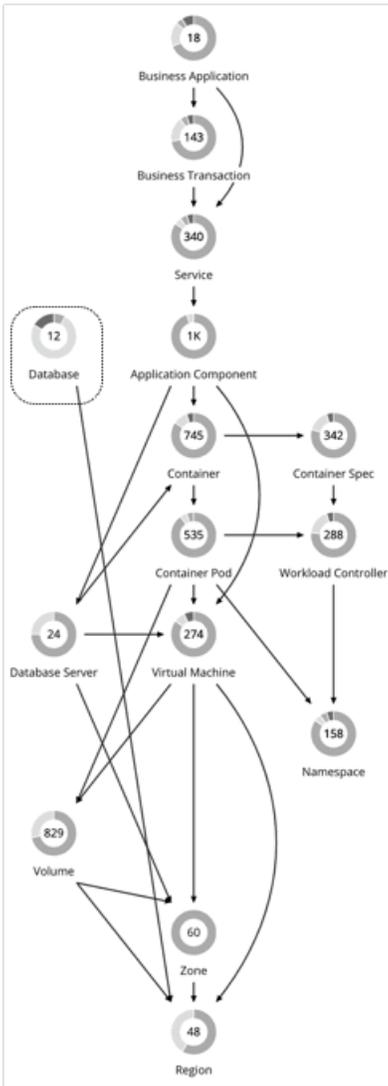
Since the Estimated On-demand Monthly Cost is projected to decrease from \$1368.23/month to \$368.62/month, Intersight Workload Optimizer treats the action as a cost-saving measure and shows estimated savings of \$999.61/month.

Database – Azure Dedicated SQL Pool

A [dedicated SQL pool](#) in Azure Synapse Analytics represents a collection of analytic resources that Azure provisions when you use Synapse SQL. Compute capacity for a dedicated SQL pool is expressed in Data Warehousing Units (DWU). DWU represents CPU, memory, and IO Throughput bundled as a single commodity.

Intersight Workload Optimizer discovers dedicated SQL pools through your Azure targets, and represents them as Database entities in the supply chain.

Synopsis



Synopsis

| | |
|---------------------|---|
| Provides: | Analytic resources to Azure Synapse Analytics |
| Consumes: | DWU and storage resources in an Azure region |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- DWU

DWU (Data Warehousing Unit) is the measurement of compute capacity for the dedicated SQL pool. DWU represents CPU, memory, and IO Throughput bundled as a single commodity.
- Storage Amount

Storage Amount is the measurement of storage capacity that is in use.
- Connection

Connection is the measurement of database connections utilized by applications.

Suspend Actions for Dedicated SQL Pools

When Intersight Workload Optimizer analysis discovers an idle dedicated SQL pool (i.e., a pool with no active connections for a specific period of time), it will immediately recommend that you suspend the pool as a cost-saving measure, and show the action in Action Center. This action stops the compute instance for the SQL pool, but keeps storage running. In the Action Details page for a specific action, the compute cost is 0 (zero) *after* suspension, while storage cost remains the same.

Action Details

Suspend Database paassqlpool in [redacted]
Improve infrastructure efficiency

| ACTION ESSENTIALS | | DATABASE DETAILS | |
|------------------------|--|----------------------|--------------------------------|
| State | 📌 Action can be accepted and executed immediately. | Name | paassqlpool 🗑️ |
| Non-Disruptive | ✓ Downtime is not required to execute. | Id | paassqlpool 🗑️ |
| Reversible | ✓ Action can be manually reverted. | Subscription | [redacted] 🗑️ |
| RESOURCE IMPACT | | Workspace | paas-synapse-workspace 🗑️ |
| | CURRENT | AFTER ACTIONS | Pool Type |
| DWU, Capacity 📉 | 100 | - | Dedicated SQL pool 🗑️ |
| DWU, Utilization 📉 | 0% | - | Pricing Model |
| Storage, Capacity 📉 | 1.34 GB | 1.34 GB | DWU 🗑️ |
| Storage, Utilization 📉 | 100% | 100% | Idle Time |
| | | | 14 days 12 hours 47 minutes 🗑️ |
| COST IMPACT | | | |
| | CURRENT | AFTER ACTIONS | |
| Compute Cost 📉 | \$876.00/mo | \$0.00/mo | ↓ \$876.00/mo |
| Storage Cost 📉 | \$23.00/mo | \$23.00/mo | - |
| Total Cost | \$899.00/mo | \$23.00/mo | ↓ \$876.00/mo |
| Total Savings | | \$876.00/mo | |

To calculate the current compute cost, Intersight Workload Optimizer retrieves the hourly [on-demand cost](#) from Azure and then multiplies the result by 730, which represents the number of hours per month that the product uses to estimate monthly costs.

If a currently idle pool is not suspended and is subsequently discovered as active, Intersight Workload Optimizer removes the suspend action attached to it.

Suspend actions include the 'Idle Time' information that indicates how long a dedicated SQL pool has been idle. By default, Intersight Workload Optimizer will generate suspend actions if dedicated SQL pools have been idle for at least one hour.

Action Details

Suspend Database paassqlpool in [redacted]
Improve infrastructure efficiency

| ACTION ESSENTIALS | | DATABASE DETAILS | |
|------------------------|--|----------------------|--------------------------------|
| State | 📌 Action can be accepted and executed immediately. | Name | paassqlpool 🗑️ |
| Non-Disruptive | ✓ Downtime is not required to execute. | Id | paassqlpool 🗑️ |
| Reversible | ✓ Action can be manually reverted. | Subscription | [redacted] 🗑️ |
| RESOURCE IMPACT | | Workspace | paas-synapse-workspace 🗑️ |
| | CURRENT | AFTER ACTIONS | Pool Type |
| DWU, Capacity 📉 | 100 | - | Dedicated SQL pool 🗑️ |
| DWU, Utilization 📉 | 0% | - | Pricing Model |
| Storage, Capacity 📉 | 1.34 GB | 1.34 GB | DWU 🗑️ |
| Storage, Utilization 📉 | 100% | 100% | Idle Time |
| | | | 14 days 12 hours 47 minutes 🗑️ |
| COST IMPACT | | | |
| | CURRENT | AFTER ACTIONS | |
| Compute Cost 📉 | \$876.00/mo | \$0.00/mo | ↓ \$876.00/mo |
| Storage Cost 📉 | \$23.00/mo | \$23.00/mo | - |
| Total Cost | \$899.00/mo | \$23.00/mo | ↓ \$876.00/mo |
| Total Savings | | \$876.00/mo | |

You can control the suspend actions that Intersight Workload Optimizer generates, based on your preferred idle time value.

For example, if you want Intersight Workload Optimizer to only generate suspend actions for dedicated SQL pools that have been idle for at least one day, perform these steps:

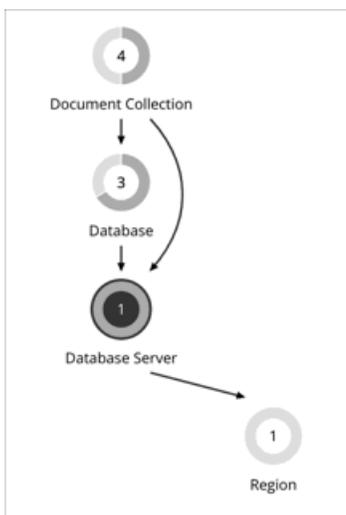
1. Create a dynamic group of Databases and set the 'Hours Idle' filter to `Hours Idle < = 24`.
2. Create a custom Database policy, set the scope to the group that you just created, and then *disable* suspend actions in that policy.

Database - Azure Cosmos DB

In [Azure Cosmos DB \(on page 106\)](#), a database is a group of containers ([document collections \(on page 322\)](#)) and is similar to a namespace. Cosmos DB databases are managed through [accounts \(on page 307\)](#).

Intersight Workload Optimizer discovers Cosmos DB databases through your Azure targets, and represents them as database entities in the supply chain.

Synopsis



| Synopsis | |
|---------------------|--|
| Provides: | Resources for database operations supported by Azure Cosmos DB |
| Consumes: | Request Units (RUs) from Azure Cosmos DB accounts |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Request Unit (RU)
 - Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.

Scale Actions for Cosmos DB Databases

Intersight Workload Optimizer can scale databases with provisioned throughput. It uses percentile calculations to measure RU utilization and then recommends actions to scale RUs up or down to optimize performance and costs.

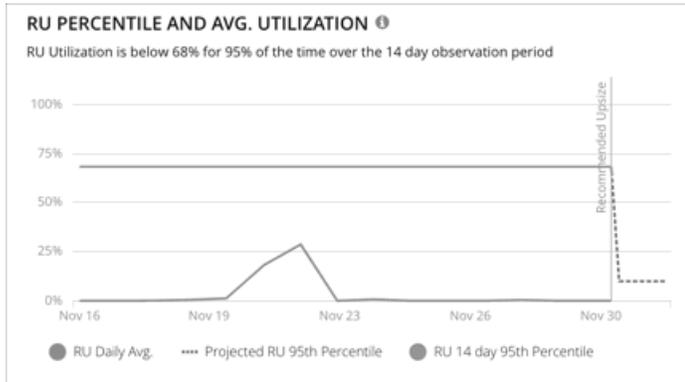
Points to consider:

- For scale down actions, Intersight Workload Optimizer will never recommend scaling below the minimum RU value calculated by Azure or configured in a database policy. For details, see [Minimum RU Capacity for Scale Down Actions \(on page 321\)](#).

- Scale out actions do not apply since there is only one tier for Cosmos DB.
- Intersight Workload Optimizer discovers but does not scale databases without provisioned throughput.

Utilization Charts for Scale Actions

Intersight Workload Optimizer uses percentile calculations to measure resource utilization, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on a database, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.



The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the database, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's scaling recommendations.

NOTE:

You can set scaling constraints in database policies to refine the percentile calculations. For details, see [Aggressiveness and Observation Period \(on page 319\)](#).

Reconfigure Actions for Cosmos DB Databases

Intersight Workload Optimizer recommends removing unused throughput that is assigned to a database to help reduce your cloud expenses. To remove this resource, reconfigure the database from Azure. It is not possible to execute reconfigure actions from Intersight Workload Optimizer.

Intersight Workload Optimizer generates a reconfigure action if a database and all the underlying document collections have their own provisioned throughput. In this scenario, provisioned throughput at the database level is not used by any document collection and is therefore safe to remove. This action translates into cost savings. When you review a database reconfigure action in Action Center, the Cost Impact column indicates the cost savings that you would realize if you execute the action.

Points to consider:

- Intersight Workload Optimizer does not generate a reconfigure action if a database manages at least one document collection without provisioned throughput. In this scenario, the document collection is assumed to be using provisioned throughput at the database level, so it is not safe to remove that resource.
- Reconfigure actions are generated in recommend mode since it is not possible to execute these actions from Intersight Workload Optimizer.

You can disable the generation of reconfigure actions in custom policies for select databases or the default database policy (for a global effect).

Delete Actions for Cosmos DB Databases

To help reduce your cloud expenses, Intersight Workload Optimizer recommends deleting a database with provisioned throughput but without any underlying document collection. When you review a database delete action in Action Center, the Cost Impact column indicates the cost savings that you would realize if you execute the action.

This action can be executed in Intersight Workload Optimizer manually or automatically. You can disable the generation of delete actions in custom policies for select databases or the default database policy (for a global effect).

Action Disruptiveness and Reversibility

The Action Center view and Action Details page indicate the disruptiveness and reversibility of actions for Cosmos DB databases.

- Scale actions
 - All scale actions are non-disruptive and reversible.
- Reconfigure and delete actions
 - All reconfigure and delete actions are disruptive and irreversible.

Cloud Database Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

| Action | Applies To | Default Setting |
|------------------------------|--------------------------|-----------------|
| Cloud DB scale | Azure SQL vCore and DTU | Manual |
| Database suspend actions | Azure dedicated SQL pool | On, Manual |
| Database reconfigure actions | Azure Cosmos DB | On, Recommend |
| Database delete actions | Azure Cosmos DB | On, Manual |

Scaling Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for DTU, RU, and storage. It then recommends actions to improve utilization based on the observed values for a given time period.

- **Aggressiveness**

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce capacity. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce resources for that database.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single burst to 100%, but for 95% of the samples, utilization never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce resource allocation.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 99th Percentile – More performance. Recommended for critical databases that need maximum guaranteed performance at all times, or those that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.
- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for databases that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 14 days. Use the **Max Observation Period** setting to adjust the number of days.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 14 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the database has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 7 Days or Last 3 Days

Intersight Workload Optimizer recommends an observation period of 14 days so it can recommend scaling actions more often. Since Azure SQL DB scaling is minimally disruptive, with near-zero downtime, scaling often should not introduce any noticeable performance risks.

NOTE:

For more information about Azure scaling downtimes, see the [Azure documentation](#).

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic – None
- Less Elastic – 7 Days

Cloud Instance Types

By default, Intersight Workload Optimizer considers all instance types currently available for scaling when making scaling decisions for SQL databases. However, you may have set up your SQL databases to *only scale to* or *avoid* certain instance types to reduce complexity and cost, or meet demand. Use this setting to identify the instance types that SQL databases can scale to.

| Attribute | Default Value |
|----------------------|---------------|
| Cloud Instance Types | None |

Click **Edit** to set your preferences. In the new page that displays, expand a **cloud tier** (a family of instance types, such as *Premium*) to see individual instance types and the resources allocated to them.

Select your preferred instance types or cloud tiers, or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

NOTE:

This policy setting is not available in plans.

If you selected a cloud tier and the service provider deploys new instance types to that tier later, then those instance types will automatically be included in your policy. Be sure to review your policies periodically to see if new instance types have been added to a tier. If you do not want to scale to those instance types, update the affected policies.

Scaling Target Utilization

The utilization that you set here specifies the percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

The settings you make depend on the pricing model in place for the workloads in the policy scope.

- vCore Pricing Model

To meet individual VCPU, VMEM, or IOPS/Throughput targets, the workloads must be charged according to the vCore pricing model.

| Attribute | Default Value |
|----------------------------|---------------|
| VCPU | 70 |
| VMEM | 90 |
| IOPS/Throughput | 70 |
| Storage Amount Utilization | 90 |

- DTU Pricing Model

To meet a target DTU utilization, the workloads must be charged according to the DTU pricing model.

| Attribute | Default Value |
|----------------------------|---------------|
| DTU Utilization | 70 |
| Storage Amount Utilization | 90 |

- RU Pricing Model

To meet a target RU utilization or individual RU targets, the workloads must be charged according to the RU pricing model.

| Attribute | Default Value |
|----------------|---------------|
| RU Utilization | 70 |

These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Intersight Workload Optimizer calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Intersight Workload Optimizer recommends actions, in most cases you should never need to use them. If you want to control how Intersight Workload Optimizer recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

Minimum RU Capacity for Scale Down Actions

This setting applies only to Cosmos DB databases. Intersight Workload Optimizer checks this scaling constraint when it recommends scale down actions to ensure that Cosmos DB databases never scale below the specified value.

| Attribute | Default Value |
|--|---------------|
| Minimum RU capacity for scale down actions | 400 |

Be sure to specify a value that is an increment of 100. By design, Azure increases or decreases the number of RUs at any time in increments or decrements of 100 RUs.

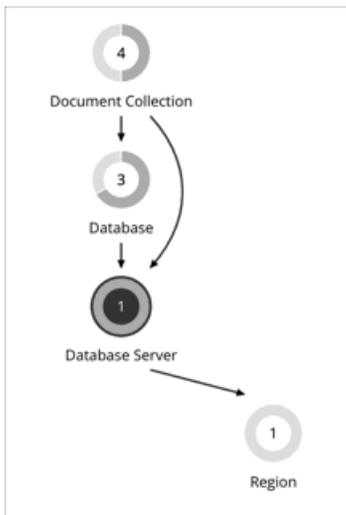
In addition to this scaling constraint, Intersight Workload Optimizer also checks the [minimum RU limit](#) that Azure calculated. This limit takes precedence over the value specified in the policy. For example, if the minimum RU limit calculated by Azure is 500 while the minimum RU capacity in the database policy is 400, Intersight Workload Optimizer will never recommend scaling below 500 RUs.

Document Collection

In [Azure Cosmos DB \(on page 106\)](#), containers are entities that store data on one or more partitions. Containers are grouped using [databases \(on page 317\)](#) and managed through [accounts \(on page 307\)](#).

Intersight Workload Optimizer discovers Azure Cosmos DB containers through your Azure targets, and represents them as document collection entities in the supply chain.

Synopsis



| Synopsis | |
|---------------------|--|
| Provides: | Resources for database operations supported by Azure Cosmos DB |
| Consumes: | Request Units (RUs) from Azure Cosmos DB accounts |
| Discovered through: | Azure targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Request Unit (RU)

Request Unit (RU) is a performance currency that abstracts CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB. Azure Cosmos DB normalizes the cost of all database operations using RUs.

Scale Actions for Document Collections

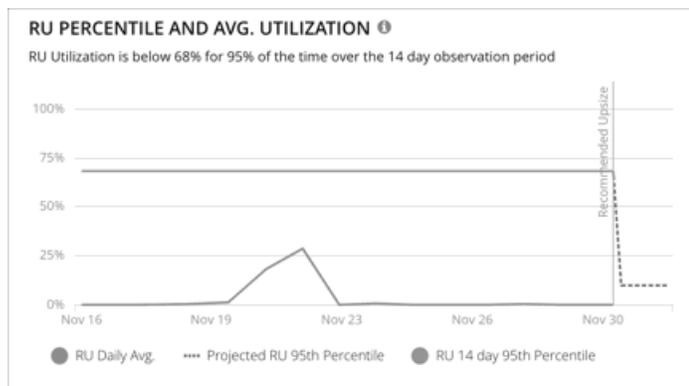
Intersight Workload Optimizer can scale document collections with standard (manually provisioned) dedicated throughput. It uses percentile calculations to measure RU utilization and then recommends actions to scale RUs up or down to optimize performance and costs.

Points to consider:

- For scale down actions, Intersight Workload Optimizer will never recommend scaling below the minimum RU value calculated by Azure or configured in a document collection policy. For details, see [Minimum RU Capacity for Scale Down Actions \(on page 325\)](#).
- Scale out actions do not apply since there is only one tier for Cosmos DB.
- Intersight Workload Optimizer discovers but does not scale the following entities:
 - Document collections with auto-scaled dedicated throughput
 - Document collections with shared throughput, where throughput is provisioned at the database level and then shared by document collections within the database

Utilization Charts for Scale Actions

Intersight Workload Optimizer uses percentile calculations to measure resource utilization, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on a document collection, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.



The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the document collection, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's scaling recommendations.

NOTE:

You can set scaling constraints in document collection policies to refine the percentile calculations. For details, see [Aggressiveness and Observation Period \(on page 324\)](#).

Non-disruptive and Reversible Scaling Actions

All scaling actions shown in the Action Center view and Action Details page are non-disruptive and reversible.

Document Collection Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Action Automation and Orchestration

For details about cloud document collection actions, see [Document Collection Actions \(on page 323\)](#).

| Action | Default Setting |
|---------------------------------|-----------------|
| Cloud Document Collection Scale | Manual |

Scaling Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for Request Units (RUs). It then recommends actions to improve utilization based on the observed values for a given time period.

■ Aggressiveness

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. This is important in the cloud, so that analysis can better exploit the elasticity of the cloud. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce capacity. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce resources for that document collection.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single burst to 100%, but for 95% of the samples, utilization never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce resource allocation.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 99th Percentile – More performance. Recommended for critical document collections that need maximum guaranteed performance at all times, or those that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.
- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for document collections that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 14 days. Use the **Max Observation Period** setting to adjust the number of days.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 14 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the document collection has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 7 Days or Last 3 Days

Intersight Workload Optimizer recommends an observation period of 14 days so it can recommend scaling actions more often. Since Azure SQL DB scaling is minimally disruptive, with near-zero downtime, scaling often should not introduce any noticeable performance risks.

NOTE:

For more information about Azure scaling downtimes, see the [Azure documentation](#).

■ **Min Observation Period**

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic - None
- Less Elastic - 7 Days

Scaling Target Utilization

The utilization that you set here specifies the percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

To meet a target RU utilization or individual RU targets, the workloads must be charged according to the RU pricing model.

| Attribute | Default Value |
|----------------|---------------|
| RU Utilization | 70 |

These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Intersight Workload Optimizer calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Intersight Workload Optimizer recommends actions, in most cases you should never need to use them. If you want to control how Intersight Workload Optimizer recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

Minimum RU Capacity for Scale Down Actions

Intersight Workload Optimizer checks this scaling constraint when it recommends scale down actions to ensure that document collections never scale below the specified value.

| Attribute | Default Value |
|--|---------------|
| Minimum RU capacity for scale down actions | 400 |

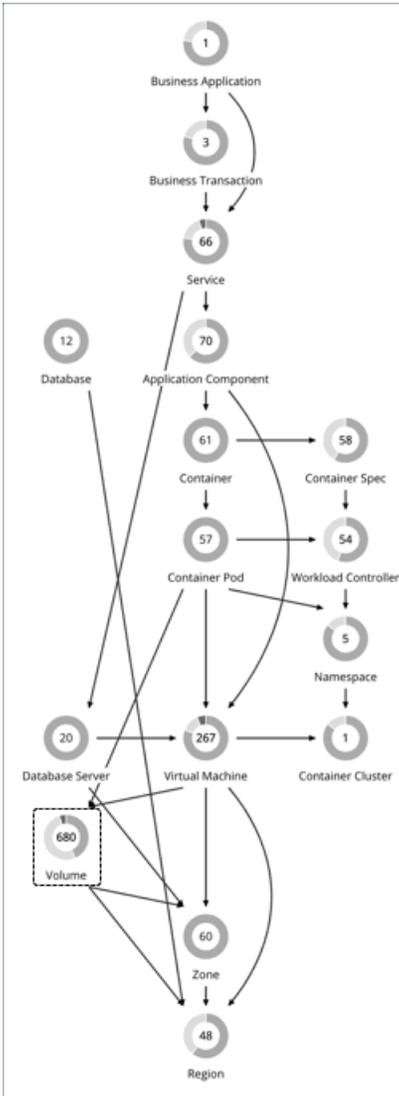
Be sure to specify a value that is an increment of 100. By design, Azure increases or decreases the number of RUs at any time in increments or decrements of 100 RUs.

In addition to this scaling constraint, Intersight Workload Optimizer also checks the [minimum RU limit](#) that Azure calculated. This limit takes precedence over the value specified in the policy. For example, if the minimum RU limit calculated by Azure is 500 while the minimum RU capacity in the document collection policy is 400, Intersight Workload Optimizer will never recommend scaling below 500 RUs.

Volume (Cloud)

A cloud volume is a storage device that you can attach to a VM. You can use an attached volume the same as a physical hard drive.

Synopsis



| Synopsis | |
|---------------------|--|
| Provides: | Storage resources for VMs to use, including: <ul style="list-style-type: none"> ■ Storage Access ■ Storage Amount ■ IO Throughput |
| Consumes: | Storage services provided by zones or regions |
| Discovered through: | Public cloud targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage Amount**
Storage Amount is the storage capacity (disk size) of a volume.
Intersight Workload Optimizer discovers Storage Amount, but does not monitor utilization.
For a Kubeturbo (container) deployment that includes volumes, Kubeturbo monitors Storage Amount utilization for the volumes. You can view utilization information in the Capacity and Usage chart.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the measurement of IOPS capacity that is in use.
- **I/O Throughput**
I/O Throughput is the measurement of I/O throughput capacity that is in use.

NOTE:

Intersight Workload Optimizer also monitors the attachment state of volumes and then generates delete actions for unattached volumes.

Actions

Intersight Workload Optimizer supports the following actions:

- **Scale**
Scale attached volumes to optimize performance and costs.
- **Delete**
Delete unattached volumes as a cost-saving measure. Intersight Workload Optimizer generates an action immediately after discovering an unattached volume.

Scale Actions

Scale attached volumes to optimize performance and costs.

Intersight Workload Optimizer can recommend:

- Scaling within the same tier (scale up or down), or from one tier to another
Examples:
 - An action to scale down IOPS for a high performance volume, such as Azure Managed Ultra
 - An action to scale a volume from the *io1* to the *gp2* tier

These actions can reduce costs significantly while continuing to assure performance. In addition, they are *non-disruptive* and *reversible*.

NOTE:

For details about action disruptiveness and reversibility, see [Non-disruptive and Reversible Scaling Actions \(on page 331\)](#).

- Scaling from one tier to another and then within the new tier, in a single action
For example, to achieve higher IOPS performance for VMs that are premium storage capable, you might see an action to scale the corresponding volume from *Standard* to *Premium*, and then scale up volume capacity from *32GB* to *256 GB*. This action increases costs and is *irreversible*, but is more cost effective than scaling up within the original tier.

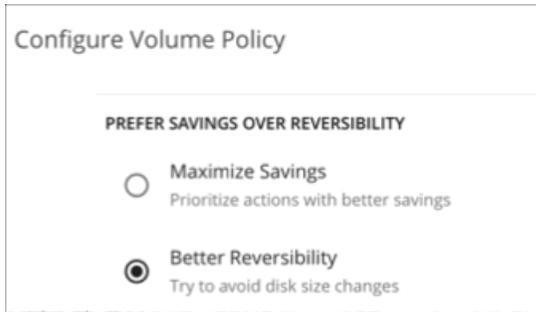
NOTE:

Not all VM types are premium storage capable. For details, see the [Azure Documentation](#).

Points to consider:

- When there are multiple disks attached to a volume, every volume scale action can potentially disrupt the same VM multiple times and some of the actions may fail due to concurrency. To mitigate these issues, Intersight Workload Optimizer identifies all volume actions associated with a particular VM and combines them into a single unit for execution, thus reducing disruptions and the chance of failures due to concurrency. This approach applies to scale actions in *Manual* or *Automated* mode.

- You can create policies to control the scaling actions that Intersight Workload Optimizer generates.
 - Intersight Workload Optimizer includes a policy setting that lets you choose between two outcomes – better savings (default) and better reversibility. If you choose reversibility, which can increase costs, Intersight Workload Optimizer will prioritize actions to change tiers whenever possible.



- Set scaling constraints if you want volumes to *only scale to* or *avoid* certain tiers. For details, see [Cloud Storage Tiers \(on page 332\)](#).

Delete Actions

Delete unattached volumes as a cost-saving measure.

NOTE:

If you delete an Azure volume and then later deploy a new one with an identical name, charts will include historical data from the volume that you deleted.

Exceptions for Azure Site Recovery and Azure Backup Volumes

Intersight Workload Optimizer discovers Azure Site Recovery and Azure Backup volumes when you add Azure targets. Even though these volumes are always unattached, Intersight Workload Optimizer will never recommend deleting them because they are critical to business continuity and disaster recovery.

To identify Azure Site Recovery volumes, Intersight Workload Optimizer checks an Azure resource called [Recovery Services vault](#), which includes information specific to the volumes. It also checks for the `ASR-ReplicaDisk` tag, which Azure automatically assigns to the volumes.

For Azure Backup volumes, Intersight Workload Optimizer checks for the `RSVaultBackup` tag.

It is important that you do not remove these tags. If these tags have been removed for some reason, create a volume policy for the affected volumes and disable the *Delete* action in that policy.

Action Execution for Locked Volumes

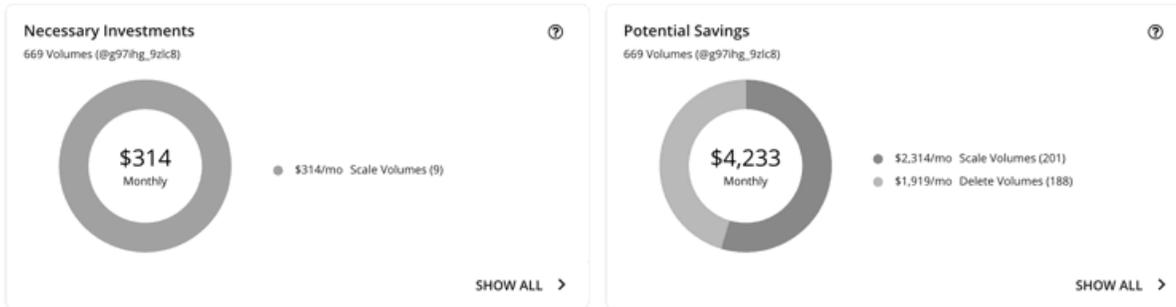
For Azure environments, Intersight Workload Optimizer can recommend scale and delete actions for [locked volumes](#), but the lock level configured for the volumes may prevent some actions from executing.

- For volumes with the `ReadOnly` lock, both scale and delete actions are *not* executable.
- For volumes with the `CanNotDelete` lock, delete actions are *not* executable, but scale actions are executable.

You must sign in to Azure and then remove the locks for the affected volumes before you can execute actions. To identify the specific locks that you need to remove, open the Action Details page for a pending volume action and see the **Execution Prerequisites** section.

Volume Actions in Charts

Use the Necessary Investments and Potential Savings charts to view pending volume actions. These charts show total monthly investments and savings, assuming you execute all the actions.



Click **Show All** for each chart to review and execute the actions.

The table shows the following:

- Actions that are pending for each volume
- Savings or investments for each volume
- For *Delete* actions in the Potential Savings chart:

| Potential Savings | | | | | | | | | | |
|---------------------------------------|-----------------|----------------------|-----------------|--------|------------|-----------------|--------------|-----------------|-------------|---------|
| DELETE ^ | | | | | | | | | | |
| AWS AZURE | | | | | | | | | | |
| Volumes (191) | | | | | | | | | | |
| Delete Actions 189 Savings \$2,033/mo | | | | | | | | | | |
| EXECUTE ACTIONS | | | | | | | | | | |
| SCALE ^ | | | | | | | | | | |
| Volumes (157) | | | | | | | | | | |
| <input type="checkbox"/> | Volume ID | Subscription | Tier Type | Size | State | Days Unattached | Image Disk | Action Category | Savings | Action |
| <input type="checkbox"/> | aks-agentpool-3 | Pay-As-You-Go - Prod | Managed Premium | 30 GiB | Unattached | 4 | /Subscrip... | SAVINGS | ↓ \$5.28/mo | DETAILS |
| <input type="checkbox"/> | aks-agentpool-4 | Pay-As-You-Go - Prod | Managed Premium | 30 GiB | Unattached | 4 | /Subscrip... | SAVINGS | ↓ \$5.28/mo | DETAILS |
| <input type="checkbox"/> | aks-agentpool-8 | Pay-As-You-Go - Prod | Managed Premium | 30 GiB | Unattached | 4 | /Subscrip... | SAVINGS | ↓ \$5.28/mo | DETAILS |
| <input type="checkbox"/> | aks-agentpool-7 | Pay-As-You-Go - Prod | Managed Premium | 30 GiB | Unattached | 4 | /Subscrip... | SAVINGS | ↓ \$5.28/mo | DETAILS |

- Number of days a volume has been unattached
- This information helps you decide whether to take the action.

Intersight Workload Optimizer polls your cloud volumes every 6 hours, and then records their state (attached or unattached) at the time of polling. It does not account for changes in state between polls.

For newly unattached volumes, Intersight Workload Optimizer shows a dash symbol (-) if a volume has been unattached within the last 6 hours. A value of 0 (zero) means that a volume has been unattached within the last 24 hours.

Once Intersight Workload Optimizer discovers an unattached volume, it immediately recommends that you delete it. If a currently unattached volume is not deleted and is subsequently discovered as attached, Intersight Workload Optimizer removes the *Delete* action attached to it, and then resets the unattached period.

NOTE:

For volumes that have been deleted from the cloud provider and are no longer discoverable, Intersight Workload Optimizer removes them from its records after 15 days.

To see the last known VM attached to the volume, click **DETAILS**.

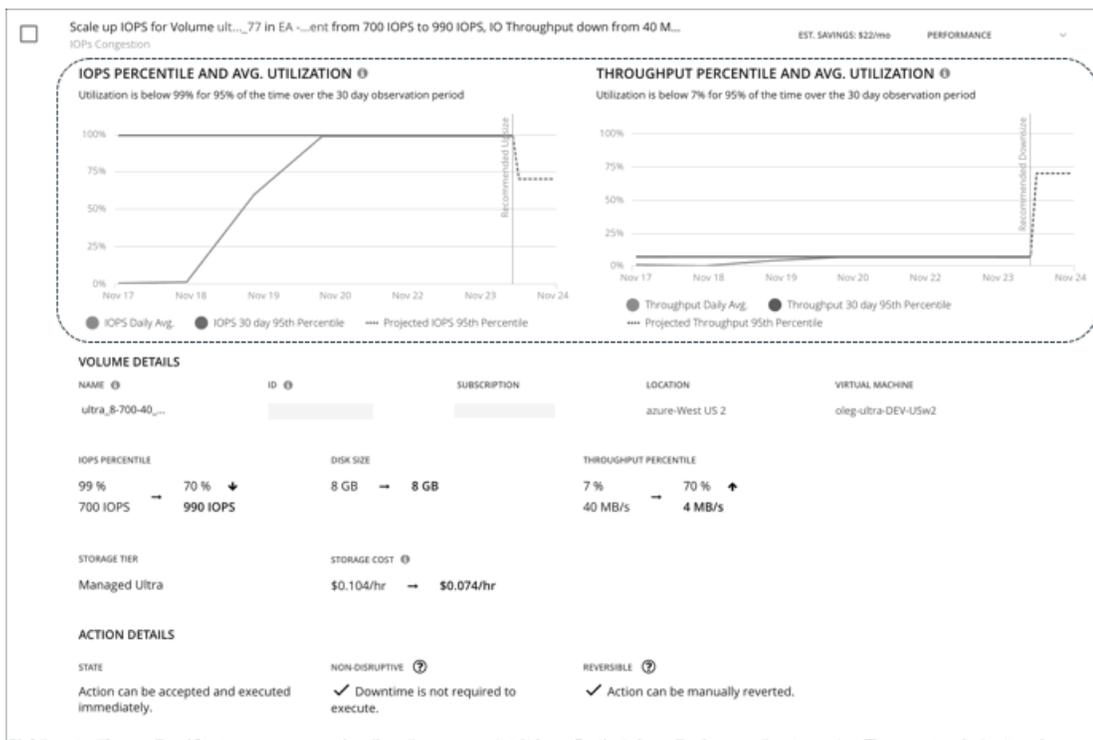
- For *Scale* actions in the Potential Savings or Necessary Investments chart:

| Potential Savings | | | | | | | | | | | | | | | |
|--------------------------|-----------------|--------------------------------------|----------------|------------|-------------|-----------|------|----------|-------------|-----------|----------|----------|-----------------|------------|---------|
| SCALE | | Scale Actions 201 Savings \$2,314/mo | | | | | | | | | | | EXECUTE ACTIONS | | |
| Volumes (201) | | | | | | | | | | | | | | | |
| DELETE | | | | | | | | | | | | | | | |
| Volumes (188) | | | | | | | | | | | | | | | |
| <input type="checkbox"/> | Volume Name | Account | Non-Disruptive | Reversible | Tier | Disk Size | IOPS | Cost | New Tier | Disk Size | New IOPS | New Cost | Action Category | Savings | Action |
| <input type="checkbox"/> | snap_1rml_dotn | Dev | ✗ | ✓ | Managed ... | 1 TB | 5000 | \$148/mo | Managed ... | 1 TB | 500 | \$41/mo | SAVINGS | ↓ \$108/mo | DETAILS |
| <input type="checkbox"/> | PTEricDisks2_Di | Prod | ✓ | ✓ | Managed ... | 256 GB | 1500 | \$284/mo | Managed ... | 256 GB | 2143 | \$180/mo | PERFORMANCE | ↓ \$104/mo | DETAILS |
| <input type="checkbox"/> | SQLServerDyna' | Dev | ✗ | ✓ | Managed ... | 1 TB | 5000 | \$135/mo | Managed ... | 1 TB | 500 | \$41/mo | SAVINGS | ↓ \$94/mo | DETAILS |
| <input type="checkbox"/> | SQLServerDyna' | Dev | ✗ | ✓ | Managed ... | 1 TB | 5000 | \$135/mo | Managed ... | 1 TB | 500 | \$41/mo | SAVINGS | ↓ \$94/mo | DETAILS |
| <input type="checkbox"/> | SQLServerTestV | Dev | ✗ | ✓ | Managed ... | 1 TB | 5000 | \$135/mo | Managed ... | 1 TB | 500 | \$41/mo | SAVINGS | ↓ \$94/mo | DETAILS |
| <input type="checkbox"/> | SQLServerTestV | Dev | ✗ | ✓ | Managed ... | 1 TB | 5000 | \$135/mo | Managed ... | 1 TB | 500 | \$41/mo | SAVINGS | ↓ \$94/mo | DETAILS |

- Whether actions are non-disruptive or reversible
- Changes the actions will effect (for example, changes in tiers and/or resource allocations)

When you click the **DETAILS** button for a scaling action, you will see utilization charts that help explain the reason for the action.

Utilization Charts for Volume Scaling Actions



Intersight Workload Optimizer uses percentile calculations to measure IOPS and throughput more accurately, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on a volume, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.

The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the volume, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's scaling recommendations.

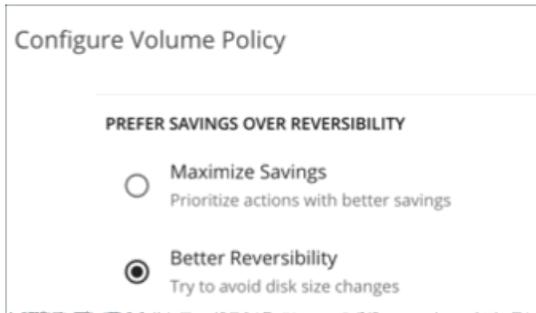
NOTE:

You can set scaling sensitivity in volume policies to refine the percentile calculations. For details, see [Scaling Sensitivity \(on page 332\)](#).

Disruptiveness and Reversibility of Volume Scaling Actions

Scaling actions can sometimes be disruptive and/or irreversible.

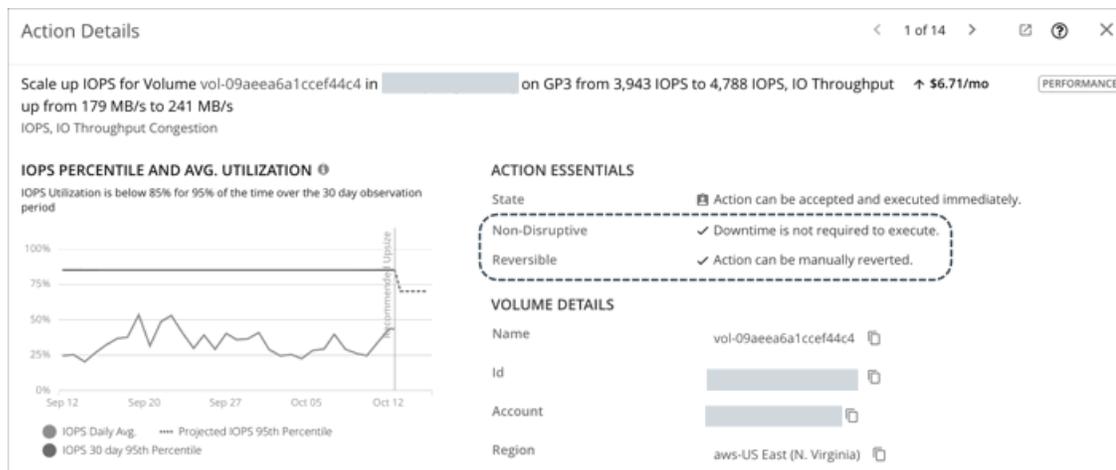
- Scaling actions are *disruptive* if the cloud provider must reboot a VM to execute a storage change. When a storage action is disruptive, expect a single reboot that usually results in 2 to 3 minutes of downtime.
- Scaling actions are irreversible if the storage must grow in size to subsequently increase IOPS or throughput capacity. In this case, shrinking the storage size later would not be possible. If you prefer reversible volume actions, create a volume policy and choose **Better Reversibility**.



The following table describes the disruptiveness and reversibility of the scaling actions that Intersight Workload Optimizer recommends.

| Nature of Scaling Action | Applicable Actions |
|---------------------------------|--|
| Non-disruptive and reversible | <ul style="list-style-type: none"> All scaling actions for AWS storage IOPS or throughput scaling for Azure ultra disks IOPS scaling for Google Cloud Hyperdisk Extreme Throughput scaling for Google Cloud Hyperdisk Throughput |
| Disruptive but reversible | <ul style="list-style-type: none"> All scaling actions for Azure storage, except IOPS or throughput scaling for ultra storage Storage tier changes for Google Cloud workloads IOPS scaling for Google Cloud extreme persistent disks |
| Non-disruptive but irreversible | Disk size scaling for AWS and Google Cloud storage |
| Disruptive and irreversible | Disk size scaling for Azure storage |

The action details for a pending volume scaling action indicates whether the action is non-disruptive or reversible.



Cloud Volume Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about cloud volume actions, see [Cloud Volume Actions \(on page 327\)](#).

| Action | Default Mode |
|---|--------------|
| Scale <ul style="list-style-type: none"> ■ Non-disruptive reversible scaling ■ Disruptive reversible scaling ■ Disruptive irreversible scaling For details about action disruptiveness and reversibility, see this topic (on page 331) . | Manual |
| Delete | Manual |

Prefer Savings Over Reversibility

Executing storage scaling actions can sometimes be irreversible if the volume must grow in size to subsequently increase IOPS or throughput capacity. In this case, shrinking that volume's size later would not be possible. If you prefer reversible volume actions, create a volume policy and choose **Better Reversibility**.

Scaling Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for IOPS and throughput. It then recommends actions to improve utilization based on the observed values for a given time period.

Scaling Target IOPS/Throughput Utilization

This is the target utilization as a percentage of capacity.

| Attribute | Default Value |
|--|---------------|
| Scaling Target IOPS/Throughput Utilization | 70 |

Cloud Storage Tiers

By default, Intersight Workload Optimizer considers all storage tiers currently available for scaling when making scaling decisions for volumes. However, you may have set up your cloud volumes to *only scale to* or *avoid* certain tiers to reduce complexity and cost, or meet demand. Use this setting to identify the tiers that volumes can scale to.

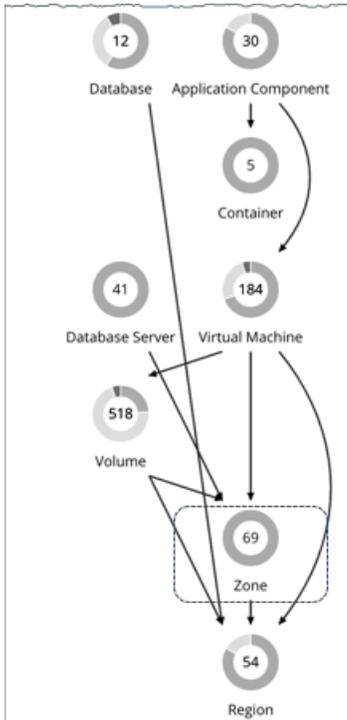
| Attribute | Default Value |
|---------------------|---------------|
| Cloud Storage Tiers | None |

Click **Edit** to set your preferences. In the new page that displays, select your preferred cloud tiers or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

Zone

A zone represents an Availability Zone in your public cloud account or subscription. A zone is a location within a given region that serves as a datacenter to host the workloads that you run in your environment.

Synopsis



| Synopsis | |
|---------------------|--------------------------------------|
| Provides: | Compute and storage resources to VMs |
| Consumes: | Region resources |
| Discovered through: | Public cloud targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Templates**
The templates and template families that each zone or region delivers. This includes template capacity and cost for workload resources.
- **Account Services**
These include storage modes, services the accounts offer for enhanced metrics, and services for different storage capabilities.
- **(AWS only) Relational Database Services (RDS)**
The RDS capabilities each cloud account provides.
- **Storage Tiers**
Intersight Workload Optimizer discovers the storage tier that supports your workloads, and uses the tier pricing to calculate storage cost.
- **Billing**
Intersight Workload Optimizer discovers the billing across the zones and regions to predict costs in the future, and to track ongoing costs. This includes comparing on-demand pricing to discount billing.
- **Virtual Memory**
The percentage utilized of memory capacity for all the workloads in the zone.
- **Virtual CPU**

- The percentage utilized of VCPU capacity for all the workloads in the zone.
- Storage Access

For environments that measure storage access, the percentage utilized of access capacity for the zone.
- Storage Amount

The percentage utilized of storage capacity for the zone.
- IO Throughput

For environments that measure IO throughput, the percentage utilized of throughput capacity for the zone.
- Net Throughput

For environments that measure Net throughput, the percentage utilized of throughput capacity for the zone.

Actions

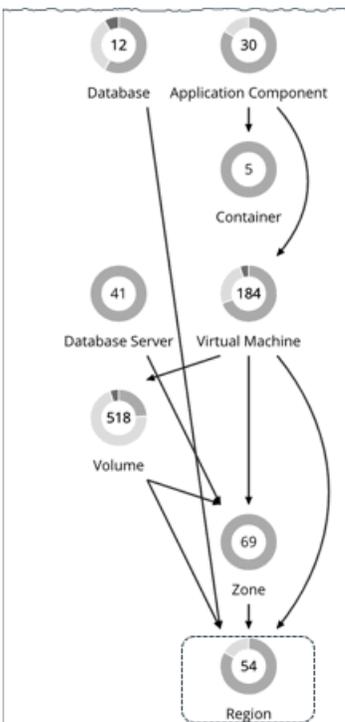
None

Intersight Workload Optimizer does not recommend actions for a cloud zone.

Region

A Region represents a geographical area that is home to one or more Availability Zones. Regions are often isolated from each other, and you can incur a cost for data transfer between them.

Synopsis



| | |
|---------------------|-------------------------------------|
| Synopsis | |
| Provides: | Host and storage resources to zones |
| Consumes: | N/A |
| Discovered through: | Public cloud targets |

Monitored Resources

Intersight Workload Optimizer does not monitor resources directly from the region, but it does monitor the following resources, aggregated for the zones in a region:

- **Virtual Memory**
The percentage utilized of memory capacity for workloads in the zones.
- **Virtual CPU**
The percentage utilized of VCPU capacity for workloads in the zones.
- **Storage Access**
For environments that measure storage access, the percentage utilized of access capacity for the zones.
- **Storage Amount**
The percentage utilized of storage capacity for the zones.
- **IO Throughput**
For environments that measure IO throughput, the percentage utilized of throughput capacity for the zones.
- **Net Throughput**
For environments that measure Net throughput, the percentage utilized of throughput capacity for the zones.

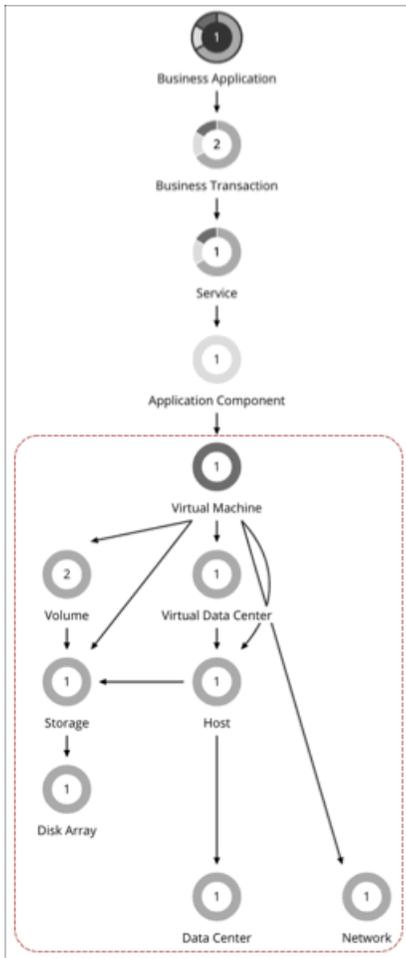
Actions

None

Intersight Workload Optimizer does not recommend actions for a cloud region.

Entity Types - On-prem Infrastructure

Intersight Workload Optimizer discovers and monitors the entities that make up your on-prem infrastructure, and recommends actions to assure performance for the applications that consume resources from these entities.



Virtual Machine (On-prem)

A virtual machine (VM) is a software emulation of a physical machine, including OS, virtual memory and CPUs, and network ports. VMs host applications, or they provide resources to container platforms.

NOTE:

Container platform nodes are also represented as Virtual Machines in the Intersight Workload Optimizer supply chain. For details, see this [topic \(on page 257\)](#).

| | |
|---------------------|--------------------|
| Synopsis | |
| Discovered through: | Hypervisor targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Virtual Memory (VMem)**
Virtual Memory is the measurement of memory that is in use.
- **Virtual CPU (VCPU)**
Virtual CPU is the measurement of CPU that is in use.
- **Processing Unit (PU)**
For PowerVM, processing unit is a unit of measurement for shared processing power across one or more virtual processors. One shared processing unit on one virtual processor accomplishes approximately the same work as one dedicated processor.
- **Virtual Processor (VP)**
For PowerVM, virtual processor (VP) is the guest OS representation of the capacity of the VM. VPs have a one-to-one correlation with PUs. Therefore 1 VP that is utilized in a VM equates to 1 PU being used. However, VPs do not have to equal to PUs on a system. VPs are a virtual representation. Therefore they can be overprovisioned.
- **Virtual Storage**
Virtual storage is the measurement of virtual storage capacity that is in use.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- **Latency**
Latency is the measurement of storage latency.

For VMs discovered via vCenter, the following resources are also monitored:

- **Energy**
Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).
- **Carbon Footprint**
Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.

Actions

Intersight Workload Optimizer supports the following actions:

- **Resize**
Resize resource capacity, reservation, or limit to improve performance.
 - **Resize resource capacity**
Change the capacity of a resource that is allocated for the VM. For example, a resize action might recommend increasing the VMem available to a VM. Before recommending this action, Intersight Workload Optimizer verifies that the VM's cluster can adequately support the new size. If the cluster is highly utilized, Intersight Workload Optimizer will recommend a move action, taking into consideration the capacity of the new cluster and compliance with existing placement policies.
For hypervisor targets, Intersight Workload Optimizer can resize vCPU by changing the VM's socket or cores per socket count. For details, see [VCPU Scaling Controls \(on page 348\)](#).
 - **Resize resource reservation**
Change the amount of a resource that is reserved for a VM. For example, a VM could have an excess amount of memory reserved. That can cause memory congestion on the host – A resize action might recommend reducing the amount reserved, freeing up that resource and reducing congestion
 - **Resize resource limit**

Change the limit that is set on the VM for a resource. For example, a VM could have a memory limit set on it. If the VM is experiencing memory shortage, an action that decreases or removes the limit could improve performance on that VM.

When multiple resize actions that relate to the same vCenter VM are accepted (manually or automated), they execute together to avoid multiple restarts for the same VM. For example, a VM has both a vCPU resize action and a vMEM resize action. When selecting both actions for execution, the actions combine into the fewest vCenter tasks to minimize the disruption on the VM instead of one set of tasks per action. Only the VMware vCenter integration supports this functionality at this time.

■ **Resize PU / Resize VP (IBM PowerVM)**

Resize PU (processing units) or VP (virtual processor) to optimize LPAR processing unit allocation and virtual processor capacity based on historical demand collected from the HMC.

Intersight Workload Optimizer does not support native execution of Resize PU (processing units) and Resize VP (virtual processor) actions. However, actions can be executed using external automation platforms through automation workflows such as Action Scripts or Webhooks.

The following policy settings affect the resize actions that Intersight Workload Optimizer generates:

- Min/Max Observation Period
Observation period settings only apply to Resize VP actions.
For details, see [Aggressiveness and Observation Periods \(on page 346\)](#).
- Increment Constant for Processing Unit
Change the increment constant value to use for VM PU resize actions. PUs can be a decimal value in a VM setting.
For details, see [Increment Constant for Processing Unit \(on page 344\)](#).
- Increment Constant for Virtual Processors
Change the increment constant value to use for VM VP resize actions. VPs must be an integer value in a VM setting.
For details, see [Increment Constant for Virtual Processors \(on page 344\)](#).

■ **Move**

Move a VM due to high resource utilization on VM or host, excess IOPS or latency in VStorage, workload placement violation, underutilized host (move VM before suspending host).

■ **Reconfigure**

Change a VM's configuration to comply with a policy.

For hypervisor targets, Intersight Workload Optimizer can reconfigure VMs that violate vCPU scaling policies. For details, see [VCPU Scaling Controls \(on page 348\)](#).

Tuned Scaling for On-prem VMs

For resizing VMs, Intersight Workload Optimizer includes tuned scaling action settings. These settings give you increased control over the action acceptance mode for various resize actions. With this feature, you can automate resize actions within a normal range (the tuned scaling range), and direct Intersight Workload Optimizer to take more conservative actions when resizes are outside the range.

For example, consider resizing VMs to add more memory. As memory demand increases on a VM, Intersight Workload Optimizer can automatically allocate more memory. If the hosted application is in a runaway state (always requesting more memory) and ultimately falls outside of the normal range, Intersight Workload Optimizer will not automate memory resize for the VM.

To configure tuned scaling:

1. Create a VM policy.
2. Under **Action Automation**, configure the action acceptance mode for the various resize actions.
 - VCPU Resize Up
 - VCPU Resize Down
 - VCPU Resize Above Max
 - VCPU Resize Below Min
 - VMEM Resize Up

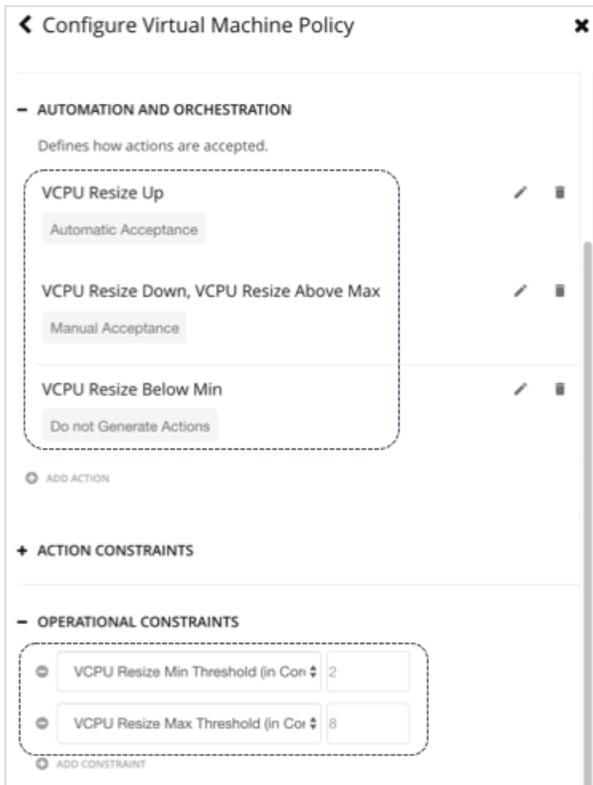
- VMEM Resize Down
- VMEM Resize Above Max
- VMEM Resize Below Min

NOTE:

Resize Up and **Resize Down** settings are for conditions within the tuned scaling range, while **Above Max** and **Below Min** settings are for outlying conditions.

3. Under **Operational Constraints**, specify the tuned scaling range.
 - VCPU Resize Max Threshold
 - VCPU Resize Min Threshold
 - VMEM Resize Max Threshold
 - VMEM Resize Min Threshold

For example, assume the following settings:



As VCPU utilization for a VM changes over time, Intersight Workload Optimizer handles resize actions as follows.

| Current | Resize Request | Action Acceptance Mode | Result |
|---------|-----------------------|---|---|
| 6 VCPUs | Resize up to 8 VCPUs | Automatic Since the VM will have 8 VCPUs after the requested resize, which is within the VCPU Resize Max threshold of 8, Intersight Workload Optimizer executes the VCPU Resize Up action automatically. | 8 VCPUs |
| 8 VCPUs | Resize up to 10 VCPUs | Manual Since the VM will have 10 VCPUs after the requested | 10 VCPUs (if you executed the pending action) |

| Current | Resize Request | Action Acceptance Mode | Result |
|----------|------------------------|---|--|
| | | resize, which is above the VCPU Resize Max threshold of 8, Intersight Workload Optimizer posts the VCPU Resize Up action (as a pending action) and provides the option to execute that action through the user interface. | |
| 10 VCPUs | Resize down to 2 VCPUs | Manual Since the VM will have 2 VCPUs after the requested resize, which is within the VCPU Resize Min threshold of 2, Intersight Workload Optimizer posts the VCPU Resize Down action (as a pending action) and provides the option to execute that action through the user interface. | 2 VCPUs (if you executed the pending action) |
| 2 VCPUs | Resize down to 1 VCPU | Not Generated Since the VM will have 1 VCPU after the requested resize, which is below the VCPU Resize Min threshold of 2, Intersight Workload Optimizer does not generate the VCPU Resize Down action to comply with the policy. | 2 VCPUs |

Action policies include scope to determine which entities will be affected by the given policy. It's possible for two or more policies to affect the same entities. As is true for other policy settings, tuned scaling uses the most conservative settings for the affected entities. The effective action acceptance mode will be the most conservative, and the effective tuned scaling range will be the narrowest range (the lowest MAX and highest MIN) out of the multiple policies that affect the given entities. For more information, see [Default and User-defined Policies \(on page 574\)](#).

You can schedule automation policies to take effect during a certain window of time. You can include tuned scaling settings in a scheduled window, the same as you can schedule other policy settings. For more information, see [Policy Schedule \(on page 583\)](#).

On-prem VM Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about on-prem VM actions, see [On-prem VM Actions \(on page 338\)](#).

| Action | Default Mode | vCenter | Hyper-V |
|--------|--------------|---|--|
| Move | Manual |  |  |

| Action | Default Mode | vCenter | Hyper-V |
|--|--------------|---|--|
| Provision (container platform nodes only) | Manual |  |  |
| Reconfigure | Recommend |  |  |
| Start | Manual |  |  |
| Storage Move | Recommend |  |  |
| Suspend (container platform nodes only) | Manual |  |  |
| vCPU Resize Above Max (uses tuned scaling) | Recommend |  |  |
| vCPU Resize Below Min (uses tuned scaling) | Recommend |  |  |
| vCPU Resize Down (uses tuned scaling) | Manual |  |  |
| vCPU Resize Up (uses tuned scaling) | Manual |  |  |
| vMem Resize Above Max (uses tuned scaling) | Recommend |  |  |
| vMem Resize Below Min (uses tuned scaling) | Recommend |  |  |
| vMem Resize Down (uses tuned scaling) | Manual |  |  |
| vMem Resize Up (uses tuned scaling) | Manual |  |  |

Non Disruptive Mode

VM actions include the modifier, **Enforce Non Disruptive Mode**. When you enable this modifier, Intersight Workload Optimizer ensures that a resize action in *Automatic* or *Manual* mode will not require a reboot or any other disruption to the affected VM. If the action will disrupt the VM, Intersight Workload Optimizer posts the action in *Recommend* mode.

| Attribute | Default Setting |
|-----------------------------|-----------------|
| Enforce Non Disruptive Mode | Off |

This setting has no effect on actions set to *Recommend* mode. Intersight Workload Optimizer will continue to post those actions for you to evaluate.

You can enforce non disruptive mode in the default VM policy, and then schedule action policies to automate resize actions during downtimes. Be aware that scheduled actions do not respect the enforced non disruptive mode – Scheduled resize actions will execute during the scheduled window even if they require a reboot. This is useful for setting up certain action behaviors, but you must be aware that enforced non disruptive mode has no effect on scheduled actions.

NOTE:

When you configure a schedule window for a VM resize action, to ensure Intersight Workload Optimizer will execute the action during the scheduled time, you must turn off the **Enforce Non Disruptive Mode** setting for that scheduled policy. Even if you turn the setting off for the global policy, you still must turn the setting off for your scheduled policy. Otherwise Intersight Workload Optimizer will not execute the resize action.

Shared-Nothing Migration

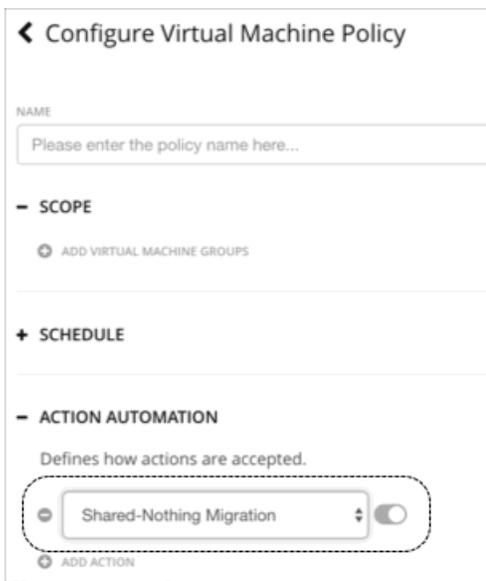
If you have enabled both storage and VM moves, Intersight Workload Optimizer can perform shared-nothing migrations, which move the VM and the stored VM files simultaneously. For example, assume a VM on a host also uses local storage on that host. In that case, Intersight Workload Optimizer can move that VM and move its data to a different datastore in a single action.

| Attribute | Default Setting |
|--------------------------|-----------------|
| Shared-Nothing Migration | Off |

Currently, the following targets support shared-nothing migrations:

- vSphere, versions 5.1 or greater
- VMM for Hyper-V 2012 or later

Because of this feature's potential impact on performance, it is turned off by default. Intersight Workload Optimizer recommends enabling it only on VMs that need it. To do this, you must first set the action acceptance mode for VM and storage moves to either *Manual* or *Automatic*, and then enable the feature in a VM policy.



If a policy that enables this feature conflicts with a more conservative policy, the latter policy wins. For example, if compute move is set to *Manual*, storage move is set to *Recommend*, and shared-nothing migration is turned on, shared-nothing migration is in effect but remains in *Recommended* state.

NOTE:

Intersight Workload Optimizer does not simulate shared-nothing migrations in plans.

Resize Thresholds

Intersight Workload Optimizer uses these settings to set up **tuned scaling** actions for on-prem VMs. Tuned scaling gives you increased control over the action mode for various resize actions. With this feature, you can automate resize actions within a normal range (the tuned scaling range), and direct Intersight Workload Optimizer to take more conservative actions when a resize is outside of the range.

For details, see [Tuned Scaling for On-prem VMs \(on page 339\)](#).

| Attribute | Default Value |
|--------------------------------------|---|
| vCPU Resize Max Threshold (in Cores) | 64 Cores Tuned Scaling Range Upper Limit |
| vCPU Resize Min Threshold (in Cores) | 1 Core Tuned Scaling Range Lower Limit |

| Attribute | Default Value |
|---------------------------|--|
| VMEM Resize Min Threshold | 512 MB Tuned Scaling Range Lower Limit |
| VMEM Resize Max Threshold | 131072 MB Tuned Scaling Range Upper Limit |

Move Enabled GPUs

Use this setting to construct a regex statement to identify which GPU models should support workload placement automation. By default, `.*_(?:a16|t4).*$` is the regex pattern used for this field.

On-prem virtual machines with vGPU types that match the specified pattern are enabled for moves. Leave this field blank to disable moves for all VMs in the selected scope.

To enable vMotion for vGPU Virtual Machines, you must change the `vgpu.hotmigrate.enabled` advanced setting. For more information, see [Configure Advanced Settings in the vCenter Server Configuration](#) in the VMware documentation.

Increment Constant for Processing Units

For PowerVM, processing unit is a unit of measurement for shared processing power across one or more virtual processors. One shared processing unit on one virtual processor accomplishes approximately the same work as one dedicated processor.

Intersight Workload Optimizer recommends VM PU resize actions in multiples of this increment constant value. For example, if Intersight Workload Optimizer determines that a VM with 1.0 PU needs 1.2 PU and the increment is set to 0.5, then a resize up action is not generated. Increasing this increment reduces the frequency of resize actions so that when they appear the change is more significant.

| Attribute | Default Value |
|---|---------------|
| Increment constant for Processing Units | 0.1 |

Increment Constant for Virtual Processors

For PowerVM, virtual processor (VP) is the guest OS representation of the capacity of the VM. VPs have a one-to-one correlation with PUs. Therefore 1 VP that is utilized in a VM equates to 1 PU being used. However, VPs do not have to equal to PUs on a system. VPs are a virtual representation. Therefore they can be overprovisioned.

Intersight Workload Optimizer recommends VM VP resize actions in multiples of this increment constant value. For example, if Intersight Workload Optimizer determines that a VM with 2 VPs needs 1 VP and the increment is set to 2, then a resize down action is not generated. Increasing this increment reduces the frequency of resize actions so that when they appear the change is more significant.

| Attribute | Default Value |
|---|---------------|
| Increment constant for Virtual Processors | 1 |

Resize VStorage

The default setting disables resize actions. This is usually preferred because VStorage resize requires that you reformat the storage. The increment constant takes effect if you enable resizing.

| Attribute | Default Setting/Value |
|---------------------------------|--|
| Resize VStorage | Disabled |
| Increment constant for VStorage | None If you enable resize, Intersight Workload Optimizer uses the default value of 1024 GB. You can change this to a different value. |

vCPU Scaling Controls

For details, see [VCPU Scaling Controls \(on page 348\)](#).

Resize Increments

These increments specify how many units to add or subtract when resizing the given resource allocation for a VM.

| Attribute | Default Value |
|---------------------------------|---------------|
| Increment constant for VMem | 1024 MB |
| Increment constant for VStorage | 1024 GB |

NOTE:

vCPU resize increments are configured in conjunction with vCPU scaling controls. For details, see [VCPU Scaling Controls \(on page 348\)](#).

For VMem, you should not set the increment value to be lower than what is necessary for the VM to operate. If the VMem increment is too low, then it's possible that Intersight Workload Optimizer would allocate insufficient VMem for the machine to operate. For a VM that is under utilized, Intersight Workload Optimizer will reduce VMem allocation by the increment amount, but it will not leave a VM with zero VMem. For example, if you set this to 1024, then Intersight Workload Optimizer cannot reduce the VMem to less than 1024 MB.

Rate of Resize

When resizing resources for a VM, Intersight Workload Optimizer calculates the optimal values for VMem, VCPU, and VStorage. But it does not necessarily make a change to that value in one action. Intersight Workload Optimizer uses the Rate of Resize setting to determine how to make the change in a single action.

NOTE:

In Intersight Workload Optimizer, the Rate of Resize default value changed from 2 to 3. If you have changed your default setting to 1 or want to keep the current default setting of 2, create a new policy scoped to all on-prem VMs and configure the Rate of Resize to your desired setting.

| Attribute | Default Value |
|----------------|---------------|
| Rate of Resize | High (3) |

■ Low

Change the value by one increment, only. For example, if the resize action calls for increasing VMem, and the increment is set at 1024, Intersight Workload Optimizer increases VMem by 1024 MB.

■ Medium

Change the value by an increment that is 1/4 of the difference between the current value and the optimal value. For example, if the current VMem is 2 GB and the optimal VMem is 10 GB, then Intersight Workload Optimizer will raise VMem to 4 GB (or as close to that as the increment constant will allow).

■ High

Change the value to be the optimal value. For example, if the current VMem is 2 GB and the optimal VMem is 8 GB, then Intersight Workload Optimizer will raise VMem to 8 GB (or as close to that as the increment constant will allow).

Consistent Resizing

| Attribute | Default Setting |
|---------------------|-----------------|
| Consistent Resizing | Off |

When you create a policy for a group of VMs and turn on Consistent Resizing, Intersight Workload Optimizer resizes all the group members to the same size, such that they all support the top utilization of each resource commodity in the group. For example, if you have deployed load balancing for a group, then all the VMs in the group should experience similar utilization. In that case, if one VM needs to be resized, then it makes sense to resize them all consistently.

Assume VM A shows top utilization of CPU, and VM B shows top utilization of memory. A resize action would result in all the VMs with CPU capacity to satisfy VM A, and memory capacity to satisfy VM B.

NOTE:

If the VMs in the group have different core speeds, then CPU scaling actions might not be consistent. For example, if you set the maximum target CPU size to 2, Intersight Workload Optimizer might recommend resizing to more than 2 CPUs to account for the VMs with slower cores.

To avoid this problem, be sure that the group only includes VMs with the same core speed.

For an affected resize, the Actions List shows individual resize actions for each of the VMs in the group. To avoid the possibility of resizing VMs disruptively at the same time, you must create automation policies with non-overlapping schedules. For example, if VMs A and B are in the same consistent resizing group, create two policies that resize the VMs at different times of the day.

- For Policy 1, set the scope to a group containing VM A and enable resize automation between, say, 01:00 and 01:45.
- For Policy 2 set the scope to a group containing VM B and enable resize automation between 02:00 and 02:45.

When working with Consistent Resizing, consider these points:

- You should not mix VMs in a group that has a Consistent Resizing policy, with other groups that enable Consistent Resizing. One VM can be a member of more than one group. If one VM (or more) in a group with Consistent Resizing is also in another group that has Consistent Resizing, then both groups enforce Consistent Resizing together, for all their group members.
- For any group of VMs that enables Consistent Resizing, you should not mix the associated target technologies. For example, one group should not include VMs that are on Hyper-V and vCenter platforms.
- Charts that show actions and risks assign the same risk statement to all the affected VMs. This can seem confusing. For example, assume one VM needs to resize to address vCPU risk, and 9 other VMs are set to resize consistently with it. Then charts will state that 10 VMs need to resize to address vCPU risks.

Fault Tolerance

Use this setting to set the number of VMs for which to tolerate failure.

NOTE:

If this number exceeds the number of VMs in the group, Intersight Workload Optimizer automatically uses N-1, where N is the number of VMs in the group.

Aggressiveness and Observation Periods

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for vCPU, vMEM, IOPS, and VPs. It then recommends actions to improve utilization based on the observed values for a given time period.

■ **Aggressiveness**

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 75th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to scale the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 75th percentile. The percentile utilization is the highest value that 75% of the observed samples fall below. Compare that to average utilization, which is the average of all the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce the capacity for CPU on a VM. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. For most VMs, there are moments when peak CPU reaches high levels, such as during reboots, patching, and other maintenance tasks. Assume utilization for a VM peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce allocated CPU for that VM.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single CPU burst to 100%, but for 75% of the samples CPU never exceeded

50%. If you set **Aggressiveness** to 75th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce CPU allocation for the VM.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples.

By default, Intersight Workload Optimizer uses samples from the last 30 days. Use the **Max Observation Period** setting to adjust the number of days. To ensure that there are enough samples to analyze and drive scaling actions, set the **Min Observation Period**.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 90 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the database has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 90 Days
- Recommended – Last 30 Days
- More Elastic – Last 7 Days

Intersight Workload Optimizer recommends an observation period of 30 days following the monthly workload maintenance cycle seen in many organizations. VMs typically peak during the maintenance window as patching and other maintenance tasks are carried out. A 30-day observation period means that Intersight Workload Optimizer can capture these peaks and increase the accuracy of its sizing recommendations.

You can set the value to 7 days if workloads need to resize more often in response to performance changes. For workloads that cannot handle changes very often or have longer usage periods, you can set the value to 90 days.

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic – None
- Less Elastic – 7 Days

Placement Policies

Intersight Workload Optimizer supports placement policies for on-prem VMs, as follows:

- You can create placement policies to enforce constraints for VM placements.

For example, the VMs in a consumer group can only run on a host that is in the provider group. You can limit the number of consumers that can run on a single provider - for hosts in the provider group, only 2 instances of VMs in the consumer group can run on the same host. Or no more than the specified number of VMs can use the same storage device.
- For VMs that require paid licenses, you can create placement policies that set up certain hosts to be the VMs' preferred license providers. Intersight Workload Optimizer can then recommend consolidating VMs or reconfiguring hosts in response to changing demand for licenses.

For more information, see [Creating Placement Policies \(on page 569\)](#).

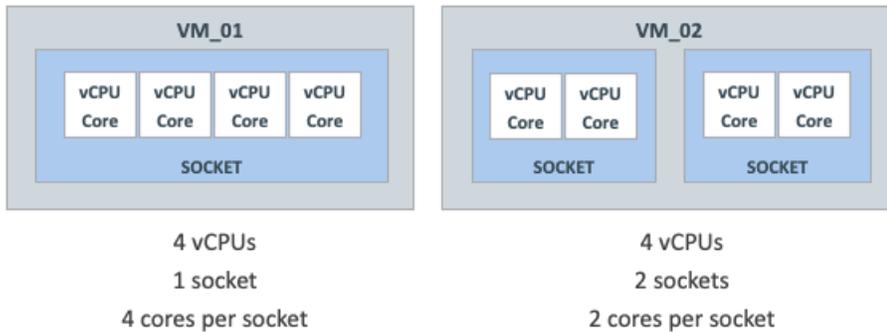
NOTE:

For VMM targets, Cisco automatically imports your Availability Sets, representing them as placement policies for the affected infrastructure. To see these availability sets, go to the **Settings > Policies** page and click **Imported Placement Policies**.

For more information, see [Importing Workload Placement Policies \(on page 569\)](#).

vCPU Scaling Controls

Intersight Workload Optimizer represents the compute capacity of a VM in MHz and vCPUs. The following diagram shows how a VM with four vCPUs can be configured differently in terms of sockets and cores.



Intersight Workload Optimizer can resize the compute capacity by changing the number of sockets or cores per socket, depending on:

- The policy assigned to the VM

[On-prem VM policies \(on page 341\)](#) include vCPU Scaling Controls that give you granular control over how VM compute resources are *resized* to maintain performance or *reconfigured* to comply with your operational policies. You can create policies for different VM groups based on their resource needs and characteristics, and decide whether to automate resize and reconfigure actions in those policies.
- The hypervisor that manages the VM

Hypervisor targets have varying degrees of support for vCPU Scaling Controls. VMware vSphere supports all scaling controls, while Hyper-V and Nutanix AHV provide limited support. For details, see the *Hypervisor Support* section below.

vCPU Scaling Control Modes and Options

Intersight Workload Optimizer provides **simple** and **advanced** controls to automate compute resource management actions in compliance with your policies. It also provides a **legacy** control based on units of MHz.

The controls you choose depend on your operational policies regarding the VM configuration of sockets and cores per socket, and your choice of hypervisor. For example, your operational policies may dictate a certain VM configuration that must be respected when resizing a VM's compute resources. Changing sockets is the least disruptive, but for some workloads, it may be preferable to change cores per socket due to socket licensing or operating system constraints. For larger VMs where Non-Uniform Memory Access (NUMA) must be considered for performance reasons, it may be preferable to balance vCPUs across host sockets.

The following tables explain the exact operation for each mode.

Simple Controls

Simple controls change compute resources based on units of vCPU.

| vCPU Scaling Option | Unit | Sockets | Cores Per Socket | Resize Action | Reconfigure Action |
|---------------------|-------|---------------------------------------|-----------------------------------|---|--------------------|
| Change virtual CPUs | vCPUs | Intersight Workload Optimizer decides | Reconfigured to 1 core per socket | <ul style="list-style-type: none"> Non-disruptive if hot-add is enabled and VM sockets are increasing Disruptive if VM cores per socket does not equal 1, even if | Disruptive |

| vCPU Scaling Option | Unit | Sockets | Cores Per Socket | Resize Action | Reconfigure Action |
|---------------------|------|---------|------------------|--------------------|--------------------|
| | | | | hot-add is enabled | |

Advanced Controls

Advanced controls allow you to change sockets or cores per socket, and configure additional options.

| vCPU Scaling Option | Unit | Sockets | Cores Per Socket | Resize Action | Reconfigure Action |
|-------------------------|-------------------|---------------------------------------|---------------------------------------|---|---|
| Change sockets | 1 socket | Intersight Workload Optimizer decides | Preserve VM cores per socket | Non-disruptive if hot-add is enabled and VM sockets are increasing | Not generated |
| Change sockets | 1 socket | Intersight Workload Optimizer decides | User-specified VM cores per socket | <ul style="list-style-type: none"> ■ Non-disruptive if hot-add is enabled ■ Disruptive if VM cores per socket does not match user-specified value | Disruptive if VM cores per socket does not match user-specified value |
| Change cores per socket | 1 core per socket | Preserve VM sockets | Intersight Workload Optimizer decides | Disruptive | Not generated |
| Change cores per socket | 1 core per socket | Match host sockets | Intersight Workload Optimizer decides | Disruptive | <ul style="list-style-type: none"> ■ Non-disruptive if hot-add is enabled and VM sockets are increasing ■ Disruptive if cores per socket is changed |
| Change cores per socket | 1 core per socket | User-specified VM sockets | Intersight Workload Optimizer decides | Disruptive | <ul style="list-style-type: none"> ■ Non-disruptive if hot-add is enabled and VM sockets are increasing ■ Disruptive if cores per socket is changed |

Legacy Controls

Legacy controls change compute resources based on units of MHz.

| vCPU Scaling Option | Unit | Sockets | Cores Per Socket | Resize Action | Reconfigure Action |
|---------------------|------|---------------------------------------|--|--|--------------------|
| MHz legacy behavior | MHz | Intersight Workload Optimizer decides | <ul style="list-style-type: none"> Assumes 1 core per socket Execution preserves actual cores per socket | Non-disruptive if hot-add is enabled and VM sockets are increasing | Not generated |

Points to consider:

- If [non-disruptive mode \(on page 342\)](#) is enabled, disruptive actions are not automated and must be executed manually.
- Older Guest OSes and applications may be sensitive to changes in the vCPU architecture that could result in power-on issues or kernel panics/BSODs. Some workloads require manual help with such changes, so always test certain classes of applications and guest operating systems before enabling any automation that changes the vCPU architecture. Use the Intersight Workload Optimizer knowledge of the application domain and Guest OS to scope them out of policies.

Scaling Option: Change Virtual CPUs

In this scaling option, Intersight Workload Optimizer adds or removes compute resources in increments of vCPUs. To achieve this, it changes the number of VM sockets and enforces 1 core per socket (if not already enforced).

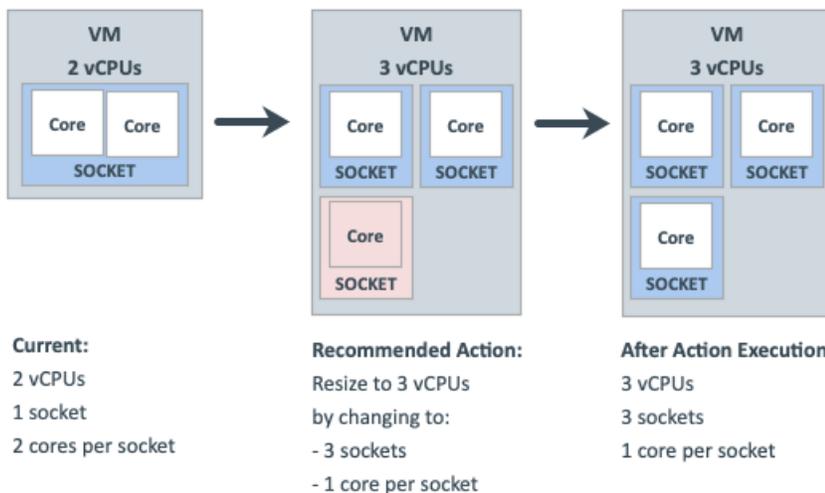
- If a VM requires a change to compute resources, Intersight Workload Optimizer generates a resize vCPU action that assumes 1 core per socket. If the VM currently does not have 1 core per socket, Intersight Workload Optimizer reconfigures it to 1 core per socket as part of action execution.
- If a VM is already optimally sized, but its current cores per socket is not 1, Intersight Workload Optimizer generates a reconfigure vCPU action to change cores per socket to 1, thereby bringing the VM into compliance with the policy.

This scaling option is ideal under the following scenarios:

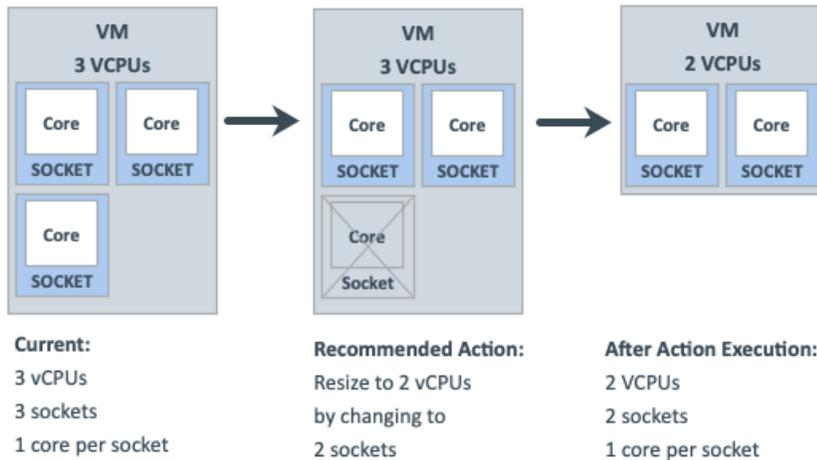
- Your environment has a large number of small VMs where precise vCPU scaling is the priority.
- You have VMs that already have 1 core per socket and require on-demand upsizes on these VMs to be non-disruptive.

For example, a VM currently has 1 socket and 2 cores per socket, and applies a policy that changes vCPU in increments of 1.

- If Intersight Workload Optimizer determines that the VM needs to increase compute capacity by 1 vCPU (i.e., from 2 to 3 vCPUs), a resize up action changes sockets from 2 to 3, and cores per socket from 2 to 1.



- When the same VM needs to reduce compute capacity by 1 vCPU (i.e., from 3 to 2 vCPUs), a resize down action changes sockets from 3 to 2.



Scaling Option: Change Sockets

In this scaling option, Intersight Workload Optimizer adds or removes compute resources by changing VM sockets.

- If a VM requires a change to compute resources, Intersight Workload Optimizer generates a resize vCPU action that considers the current cores per socket value (if the 'Preserve existing VM cores per socket' option is set) or uses the user-specified cores per socket value. If the VM's current cores per socket value violates a policy (i.e., does not match the user-specified value), Intersight Workload Optimizer reconfigures the VM's cores per socket value as part of action execution, thereby bringing the VM into compliance with the policy, while at the same time providing the required change to compute resources.
- If a VM is already optimally sized, but its current cores per socket value violates a policy (i.e., does not match the user-specified value, if set), Intersight Workload Optimizer generates a reconfigure vCPU action to change cores per socket to the user-specified value, thereby bringing the VM into compliance with the policy.

Change Sockets and Preserve VM Cores Per Socket

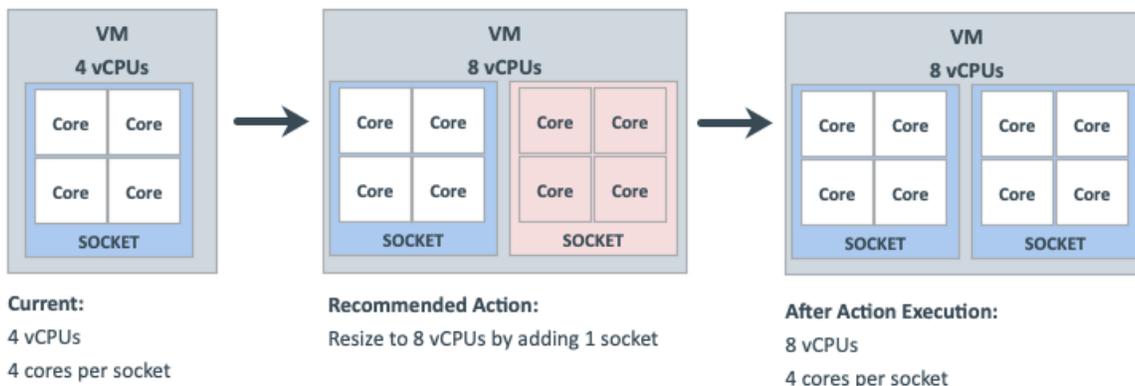
In this scaling option, Intersight Workload Optimizer adds or removes compute resources by changing VM sockets in increments of 1, and preserves VM cores per socket.

This scaling option is ideal under the following scenarios:

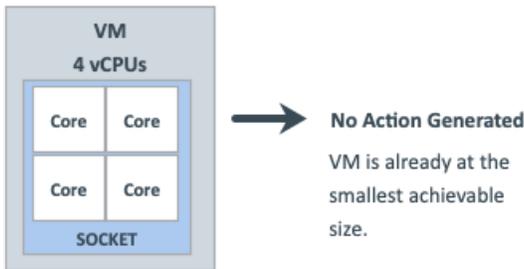
- You require Intersight Workload Optimizer to leave the VM cores per socket configuration unchanged for operational policy reasons (such as compliance with an application support contract policy).
- You have VMs that need to upsize non-disruptively to meet rising application demand.
- You have VMs with even numbers of cores per socket and are required to scale in even increments of vCPUs.

For example, a VM currently has 1 socket and 4 cores per socket, and applies a policy that changes sockets and preserves VM cores per socket. Intersight Workload Optimizer has determined that the VM requires a change in compute capacity of 1 vCPU.

- To increase compute capacity by 1 vCPU, a resize up action adds 1 socket. Because this new socket must have 4 cores to preserve VM cores per socket, the end result is 2 sockets with a total of 8 vCPUs.



- It is not possible to reduce compute capacity by 1 vCPU because the VM is already at the smallest achievable size. Therefore, no action generates.



Current:
 4 vCPUs
 4 cores per socket

Change Sockets and Specify Cores Per Socket

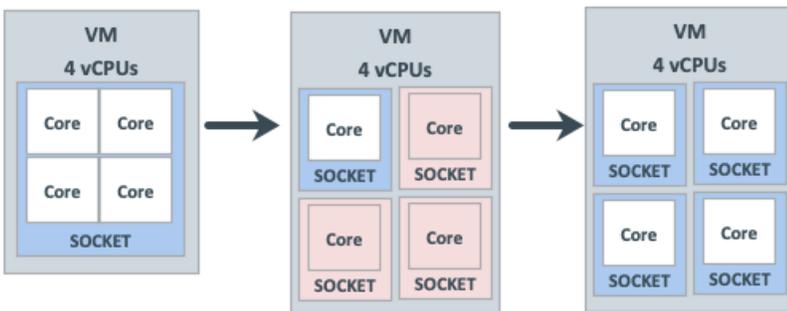
In this scaling option, Intersight Workload Optimizer adds or removes compute resources by changing VM sockets in increments of 1, and reconfigures VM cores per socket according to the value that you specify.

This scaling option is ideal under the following scenarios:

- You require any odd number vCPUs for a VM to be an even number, by setting an even number of cores per socket.
- You want a quick, script-less bulk disruptive conversion of VMs to a specific cores per socket without negatively impacting compute capacity (vCPUs).
- You have older Guest OSES and applications that are sensitive to vCPU architecture changes that could result in power-on issues or kernel panics/BSODs. Some workloads require manual help with such changes so always test certain classes of applications and OSES before enabling any automation that changes the vCPU architecture. Use the Intersight Workload Optimizer knowledge of the application domain and Guest OS to scope them out of policies.

For example, a VM currently has 1 socket and 4 cores per socket, and applies a policy that changes sockets and enforces the user-specified 1 core per socket. Intersight Workload Optimizer has determined that the VM is already optimally sized, so a resize action is not necessary.

- Since the VM is in violation of policy, Intersight Workload Optimizer changes sockets from 1 to 4, and cores per socket from 4 to 1.

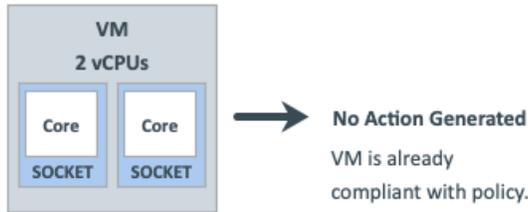


Current:
 4 vCPUs
 1 socket
 4 cores per socket (non-compliant)

Recommended Action:
 To comply with policy, change sockets to 4 and cores per socket to 1.

After Action Execution:
 4 VCPUs
 4 sockets
 1 core per socket (compliant)

- When the VM is compliant with policy, no action generates.



Current:
 2 vCPUs
 2 sockets
 1 core per socket (**compliant**)

Scaling Option: Change Cores Per Socket

In this scaling option, Intersight Workload Optimizer adds or removes compute resources by changing the VM cores per socket.

- If a VM requires a change to compute resources, Intersight Workload Optimizer generates a resize vCPU action that considers the current socket value (if the 'Preserve existing VM sockets' option is set), respects the user-specified socket value, or matches VM sockets to the host socket value. If the VM's current socket value violates a policy (i.e., does not match the user-specified or host socket value), Intersight Workload Optimizer reconfigures the VM's socket value as part of action execution, thereby bringing the VM into compliance with the policy while at the same time providing the required change to compute resources.
- If a VM is already optimally sized, but its current socket value violates a policy, Intersight Workload Optimizer generates a reconfigure vCPU action to change the sockets to the user-specified or host socket value, thereby bringing the VM into compliance with the policy.

Older Guest OSES and applications may be sensitive to vCPU architecture changes that could result in power-on issues or kernel panics/BSODs. Some workloads require manual help with such changes so always test certain classes of applications and OSES before enabling any automation that changes the vCPU architecture. Use the Intersight Workload Optimizer knowledge of the application domain and Guest OS to scope them out of policies.

Change Cores Per Socket and Preserve VM Sockets

In this scaling option, Intersight Workload Optimizer adds or removes compute resources by changing the VM cores per socket in increments of 1, and preserves VM sockets.

This scaling option is ideal under the following scenarios:

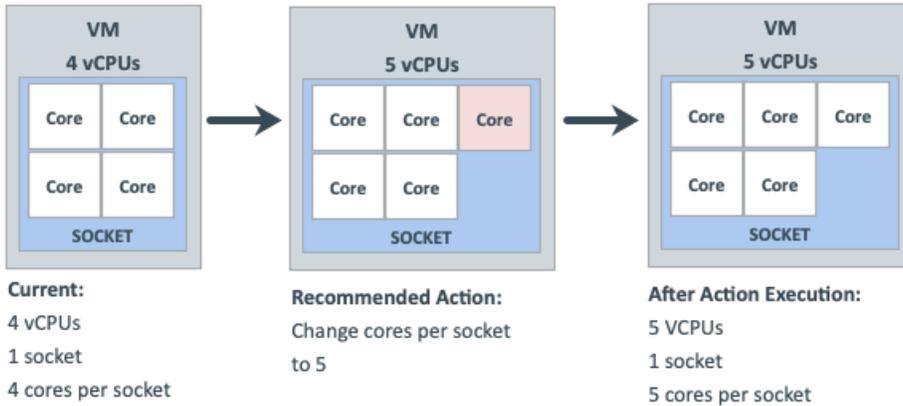
- You require Intersight Workload Optimizer to leave the VM sockets configuration unchanged for operational policy reasons (such as socket-based licensing or compliance with an application support contract policy).
- You have VDI VMs that are at their maximum Guest OS socket limitation, but require more compute resources.
- You have VMs that are configured with NUMA considerations.

NOTE:

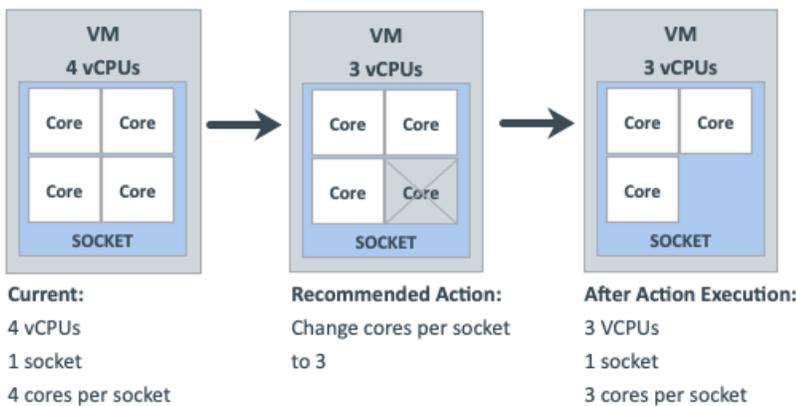
You can also use the 'Match Host Sockets' scaling option (discussed below) for NUMA sensitive VMs.

For example, a VM currently has 1 socket and 4 cores per socket, and applies a policy that changes cores per socket and preserves VM sockets. Intersight Workload Optimizer has determined that the VM requires a change in compute capacity of 1 vCPU.

- To increase compute capacity by 1 vCPU, a resize up action changes cores per socket from 4 to 5.



- To reduce compute capacity by 1 vCPU, a resize down action changes cores per socket from 4 to 3.



Change Cores Per Socket and Match Host Sockets

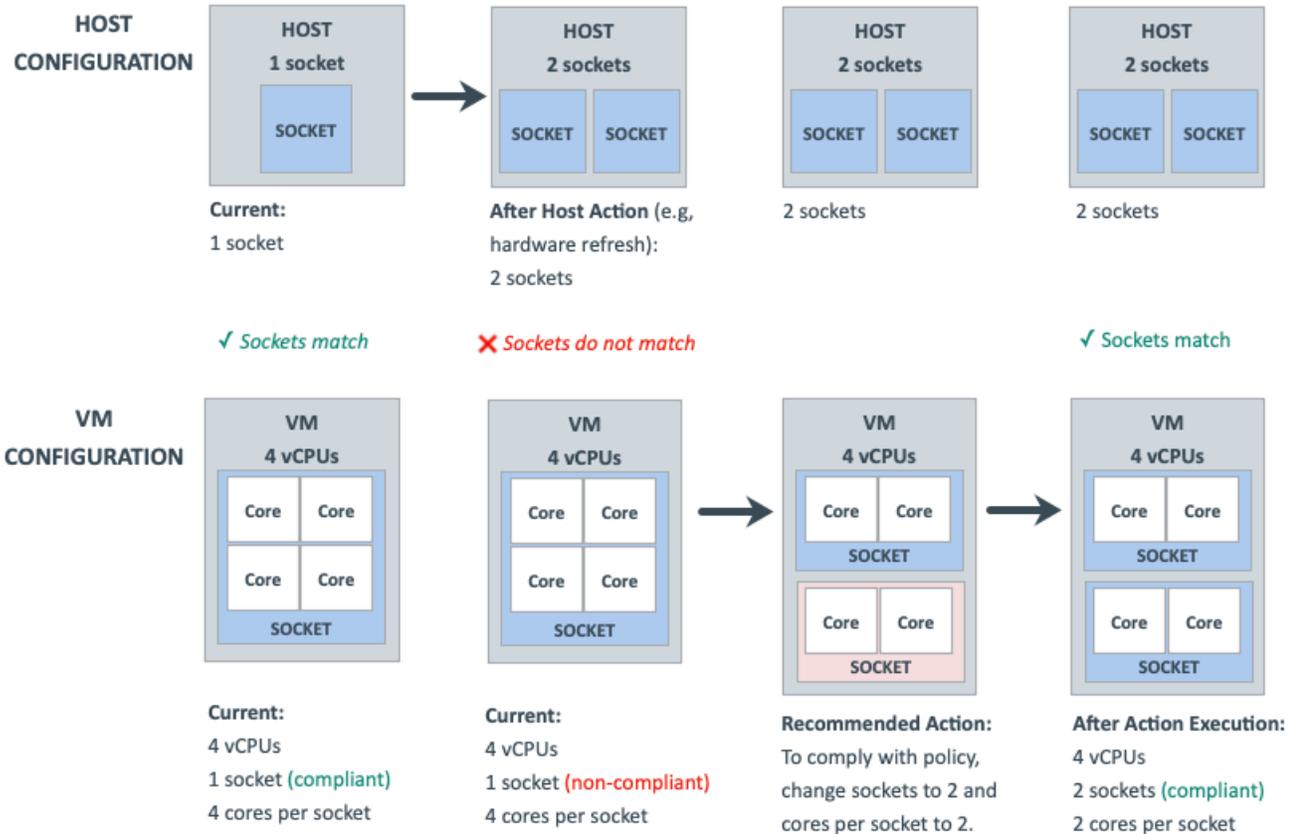
In this scaling option, Intersight Workload Optimizer reconfigures VM sockets to match the number of host sockets, thereby balancing vCPUs evenly across physical sockets. It also changes VM cores per socket to maintain the same compute capacity (vCPU).

This scaling option is ideal under the following scenarios:

- You have large VMs that may realize a performance benefit from reflecting the physical host CPU architecture within the Guest OS so that the application can optimize thread memory access to within a NUMA node.
- You have NUMA sensitive VMs that are migrating between hosts with different CPU architectures. Intersight Workload Optimizer can place the VMs on the best host and then generate an action to reconfigure the VMs to match the host sockets automatically. You can attach a schedule to the policy to automate disruptive reconfigure actions within a maintenance window.

For example, a VM currently has 1 socket and 4 cores per socket, and is on a host with 1 socket. The VM applies a policy that changes cores per socket and matches host sockets. Intersight Workload Optimizer has determined that the VM is already optimally sized, so a resize action is not necessary.

When the host socket value changes from 1 to 2, the VM is suddenly in violation of policy. To bring the VM into compliance while maintaining the same vCPU capacity (since the VM is already optimally sized), Intersight Workload Optimizer must distribute 4 cores between 2 sockets. The end result is 2 sockets and 2 cores per socket.



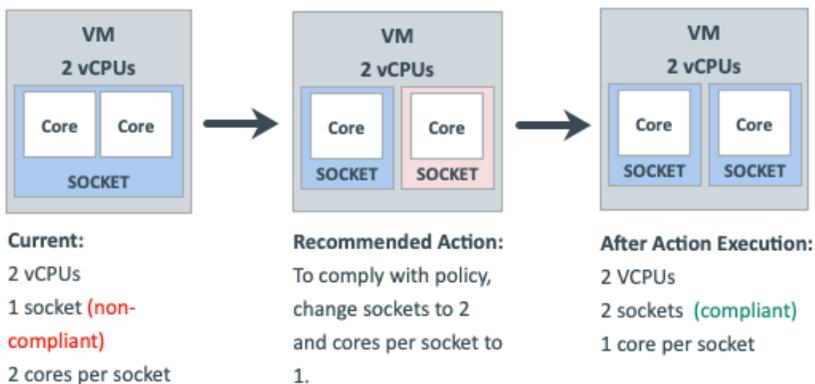
Change Cores Per Socket and Specify Sockets

In this scaling option, Intersight Workload Optimizer reconfigures VM sockets according to the value that you specify, and changes VM cores per socket to maintain the same compute capacity (vCPU).

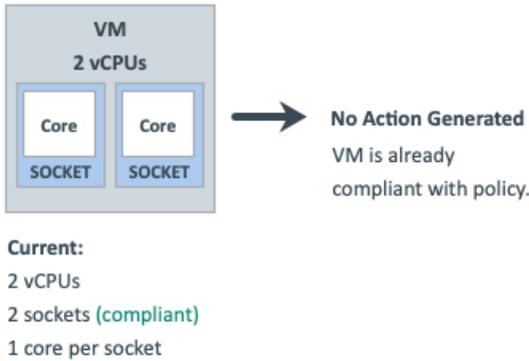
This scaling option is ideal if you have VMs that require a specific socket value for operational policy reasons (such as socket-based licensing or compliance with an application support contract policy).

For example, a VM currently has 1 socket and 2 cores per socket, and applies a policy that changes cores per socket and enforces the user-specified 2 sockets. Intersight Workload Optimizer has determined that the VM is already optimally sized, so a resize action is not necessary.

- Since the VM is in violation of policy, Intersight Workload Optimizer changes sockets from 1 to 2, and cores per socket from 2 to 1.



- When the VM is compliant with policy, no action generates.



Scaling Option: Change MHz Legacy Behavior

In this scaling option, Intersight Workload Optimizer adds or removes compute resources in increments of MHz (1800 MHz by default).

If a VM requires a change to compute resources, Intersight Workload Optimizer generates a resize vCPU action that assumes 1 core per socket, regardless of the VM's actual cores per socket.

If Intersight Workload Optimizer discovers the actual number of cores per socket as part of action execution, it adjusts the action accordingly.

For example, a VM currently has 4 vCPUs with 2 sockets and 2 cores per socket. Intersight Workload Optimizer may generate an action to resize from 4 to 5 vCPUs. However, as part of action execution, the VM socket count changes from 2 to 3, so the end result is 6 vCPUs. Conversely, the same VM may have an action to resize from 4 to 3 vCPUs, but nothing changes as part of action execution.

Hypervisor Support

For **VMware vSphere**, Intersight Workload Optimizer supports all vCPU scaling options, including changing a VM's number of sockets or cores per socket. Increasing the number of sockets is non-disruptive if CPU hot-add is enabled on a VM, while reducing the socket count always requires a restart and is therefore disruptive.

For **Hyper-V** and **Nutanix AHV**, cores per socket and hot-add features have varying degrees of support.

| vCPU Scaling Option | vSphere | Hyper-V | Nutanix AHV (Single Core) | Nutanix AHV (Multi Core) |
|--|-----------|---|--|--|
| Change virtual CPUs | Supported | Supported | Supported | Not supported by hypervisor |
| Change sockets – Preserve existing VM cores per socket | Supported | Supported | Supported | Supported |
| Change sockets – User specified cores per socket | Supported | Not supported by hypervisor | Not supported by hypervisor | Not supported by hypervisor |
| Change cores per socket – Preserve existing VM sockets | Supported | Not supported by hypervisor NOTE: Intersight Workload Optimizer assumes one core per socket and only changes sockets. | Not supported by Intersight Workload Optimizer | Not supported by Intersight Workload Optimizer |
| Change cores per socket – Match host sockets | | | | |
| Change cores per socket – User specified sockets | | | | |

Tie Breakers

When a single VM applies multiple conflicting policies, Intersight Workload Optimizer uses the following tie breakers that follow the principle of least disruptive and most conservative:

- vCPU Scaling Control
 - "Sockets" wins over "Cores per socket" wins over "Virtual CPU" wins over "MHz legacy behavior".

NOTE:

Policies created before the introduction of vCPU scaling controls (i.e., any policy before version) will continue to use the "MHz legacy behavior" option but will not be enforced when policy conflicts arise. You can remove these policies or update them to use the newer scaling controls.

- Sockets setting
 - "Preserve existing VM cores per socket" wins over "User-specified core per socket".
- Cores Per Socket setting
 - "Preserve existing VM sockets" wins over "User-specified socket" wins over "Match host sockets".
- User-specified value
 - The lowest value wins.
- Increment Size value
 - The lowest value wins.

For example, assume a VM belonging to two groups that apply different policies. Policy A changes cores per socket and matches host sockets, while Policy B changes sockets and preserves cores per socket. In this scenario, the VM applies Policy B. Changing sockets wins over changing cores per socket because it is less disruptive.

To see which policies are in effect after the tie-break decision, set the scope to a VM or group of VMs and then click the Policies tab.

Policy Cookbook

Tips:

- Use the following filters when searching for or creating VM groups:
 - Number of vCPUs
 - Number of Sockets
 - Cores per Socket
 - Target Type
 - Hot-Add Enabled
- For the least disruptive on-demand upsize of vCPU, enable hot-add on the VM and change sockets while preserving cores per socket.
- For the most precise compute resource management, change cores per socket.
- For NUMA considerations, change cores per socket and match host sockets.
- Check Guest OS application and license compatibility when changing vCPU architecture and before automating actions.

How to...

- Manage VM compute capacity by changing the number of vCPUs in increments of 2.

A VM will be reconfigured if required to use 1 core per socket, and resized by changing sockets. Actions are disruptive if the VM does not already have 1 core per socket or if hot-add is not enabled.

 1. Create a group of VMs that can have 1 core per socket and scale in sockets.
 2. Assign the group a policy with the following settings:
 - vCPU Scaling Controls
 - Change: Virtual CPU
 - Increment size: 2
 - (Optional) vCPU Resize Min/Max Threshold
- Reconfigure all odd-numbered vCPU VMs to be even-numbered, and then manage compute in even numbers of CPUs.

A VM will be reconfigured if required to use 2 cores per socket, and resized by changing sockets. Actions are disruptive if the VM does not already have 2 cores per socket or if hot-add is not enabled.

1. Create a group of VMs that can have 2 cores per socket and scale in sockets.
 2. Assign the group a policy with the following settings:
 - vCPU Scaling Controls
 - Change: Sockets
 - User specified cores per socket: 2
 - (Optional) vCPU Resize Min/Max Threshold
- Ensure that large VMs always balance their vCPU cores across all physical host sockets (for example, NUMA VMs and Database Server VMs).

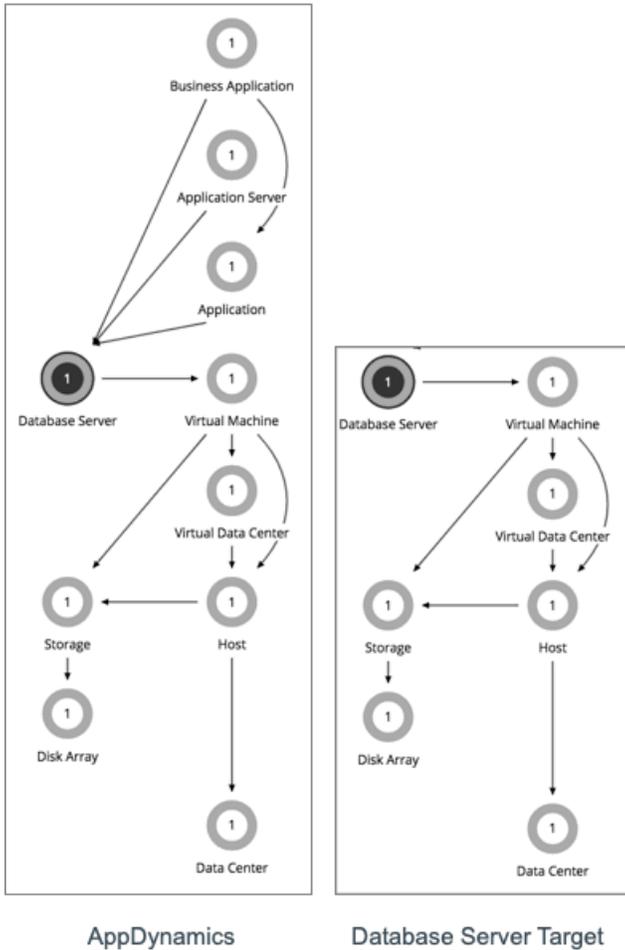
A VM will be reconfigured if its socket count does not match the host socket count. The cores per socket count may be adjusted to maintain the overall compute capacity (number of vCPUs). Resize actions are disruptive because cores per socket will change. Reconfigure actions are non-disruptive if VM sockets are increasing, hot-add is enabled, and there are no changes to cores per socket.

1. Create a group of VMs using the filters that you require to identify typically larger VMs.
 2. Assign the group a policy with the following settings:
 - vCPU Scaling Controls
 - Change: Cores per socket
 - Sockets: Match host sockets
 - (Optional) vCPU Resize Min/Max Threshold
- Keep VMs to 2 sockets only and manage compute by changing cores.
- VMs in the group will be reconfigured to 2 sockets if required, and resized by changing the cores per socket count while keeping the sockets fixed at 2, thus ensuring compliance with socket-based licensing. Resize actions are disruptive because cores per socket will change. Reconfigure actions are non-disruptive if VM sockets are increasing, hot-add is enabled, and there are no changes to cores per socket.
1. Create a VM group containing the socket-licensed VMs.
 2. Assign the group a policy with the following settings:
 - vCPU Scaling Controls
 - Change: Cores per socket
 - User specified sockets: 2
 - (Optional) vCPU Resize Min/Max Threshold

Database Server (On-prem)

For on-prem, a Database Server is a database discovered through one of the associated database application targets or through APM solutions.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | <ul style="list-style-type: none"> ■ Response Time, Transactions, DBmem, Cache Hit Rate, and TransactionLog to end users ■ Connections to Application Components |
| Consumes: | VM resources, including VCPU, VMem, and VStorage |
| Discovered through: | <ul style="list-style-type: none"> ■ AppDynamics targets ■ Database Server targets ■ Dynatrace MySQL and SQL Server processes ■ NewRelic Infrastructure Integration (NRI): MySql, SQL Server, MongoDB, OracleDB |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Virtual Memory (VMem)
Virtual Memory is the measurement of memory that is in use.
- Transaction
Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.
- Database Memory (DBMem)
Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.

- **Connection**

Connection is the measurement of database connections utilized by applications.

- **DB Cache Hit Rate**

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

Actions

Intersight Workload Optimizer supports the following actions:

Resize

Resize the following resources:

- **Connections**

Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers.

- **Database memory (DBMem)**

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

- **Transaction log**

Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology.

Currently, Intersight Workload Optimizer does not support resize actions for Oracle and Database Servers on the Hyper-V platform (due to the lack of API support for virtual storage).

On-prem Database Server Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

| Action | Default Mode |
|------------------------|--------------|
| Resize | Manual |
| Resize DBMem (Up/Down) | Manual |

Resizing Sensitivity

Intersight Workload Optimizer uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Intersight Workload Optimizer uses these settings to calculate utilization percentiles for DB Memory and DB Cache Hit Rate. It then recommends actions to improve utilization based on the observed values for a given time period.

- **Aggressiveness**

| Attribute | Default Value |
|----------------|-----------------|
| Aggressiveness | 95th Percentile |

When evaluating performance, Intersight Workload Optimizer considers resource utilization as a percentage of capacity. The utilization drives actions to resize the available capacity either up or down. To measure utilization, the analysis considers a given utilization percentile. For example, assume a 95th percentile. The percentile utilization is the highest value that 95% of the observed samples fall below. Compare that to average utilization, which is the average of *all* the observed samples.

Using a percentile, Intersight Workload Optimizer can recommend more relevant actions. For scheduled policies, the more relevant actions will tend to remain viable when their execution is put off to a later time.

For example, consider decisions to reduce capacity. Without using a percentile, Intersight Workload Optimizer never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Intersight Workload Optimizer will not reduce resources for that Application Component.

With **Aggressiveness**, instead of using the single highest utilization value, Intersight Workload Optimizer uses the percentile you set. For the above example, assume a single burst to 100%, but for 95% of the samples, utilization never exceeded 50%. If you set **Aggressiveness** to 95th Percentile, then Intersight Workload Optimizer can see this as an opportunity to reduce resource allocation.

In summary, a percentile evaluates the sustained resource utilization, and ignores bursts that occurred for a small portion of the samples. You can think of this as aggressiveness of resizing, as follows:

- 99th Percentile – More performance. Recommended for critical Application Components that need maximum guaranteed performance at all times, or those that need to tolerate sudden and previously unseen spikes in utilization, even though sustained utilization is low.
- 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings. This assures performance while avoiding reactive peak sizing due to transient spikes, thus allowing you to take advantage of the elastic ability of the cloud.
- 90th Percentile – More efficiency. Recommended for Application Components that can stand higher resource utilization.

By default, Intersight Workload Optimizer uses samples from the last 14 days. Use the **Max Observation Period** setting to adjust the number of days. To ensure that there are enough samples to analyze and drive resize actions, set the **Min Observation Period**.

■ Max Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Max Observation Period | Last 14 Days |

To refine the calculation of resource utilization percentiles, you can set the sample time to consider. Intersight Workload Optimizer uses historical data from up to the number of days that you specify as a sample period. If the Database Server has fewer days' data then it uses all of the stored historical data.

You can make the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 7 Days or Last 3 Days

Intersight Workload Optimizer recommends an observation period of 14 days so it can recommend resize actions more often. Since Database Server resizing is minimally disruptive, resizing often should not introduce any noticeable performance risks.

■ Min Observation Period

| Attribute | Default Value |
|------------------------|---------------|
| Min Observation Period | None |

This setting ensures historical data for a minimum number of days before Intersight Workload Optimizer will generate an action based on the percentile set in **Aggressiveness**. This ensures a minimum set of data points before it generates the action.

Especially for scheduled actions, it is important that resize calculations use enough historical data to generate actions that will remain viable even during a scheduled maintenance window. A maintenance window is usually set for "down" time,

when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic – None
- Less Elastic – 1, 3, or 7 Days

Transaction SLO

Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

| Attribute | Default Setting/Value |
|------------------------|--|
| Enable Transaction SLO | Off |
| Transaction SLO | None If you enable SLO, Intersight Workload Optimizer uses the default value of 10. You can change this to a different value. |

Response Time SLO

Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Intersight Workload Optimizer sets the risk index to 100%.

| Attribute | Default Setting/Value |
|--------------------------|--|
| Enable Response Time SLO | Off Intersight Workload Optimizer estimates SLO based on monitored values. |
| Response Time SLO [ms] | None If you enable SLO, Intersight Workload Optimizer uses the default value of 2000. You can change this to a different value. |

DBMem Scaling Increment

This increment specifies how many units to add or subtract when scaling DBMem.

| Attribute | Default Value |
|------------------------------|---------------|
| DBMem Scaling Increment (MB) | 128 |

Do not set the increment value to be lower than what is necessary for the database server to operate. If the increment is too low, then it's possible there would be insufficient DBMem. When reducing allocation, Intersight Workload Optimizer will not leave a database server with less than the increment value. For example, if you use the default 128, then Intersight Workload Optimizer cannot reduce DBMem to less than 128 MB.

DBMem Utilization

The utilization that you set here specifies the percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

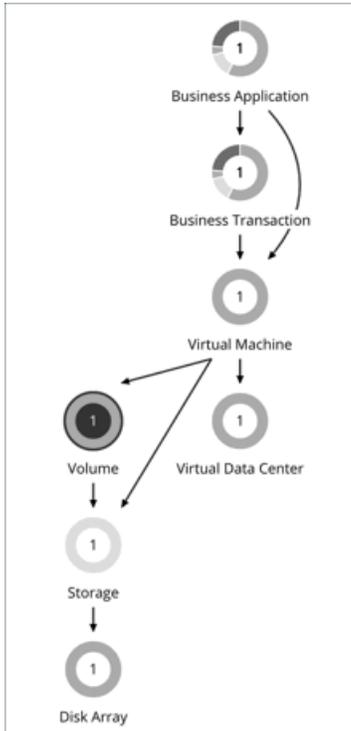
| Attribute | Default Value |
|-----------------------|---------------|
| DBMem Utilization (%) | 100 |

For example, a value of 80 means that Intersight Workload Optimizer considers 80% utilization to be 100% of capacity. Intersight Workload Optimizer recommends actions that avoid utilization beyond the given value.

Volume (On-prem)

On-prem volumes represent VM disks discovered by hypervisor targets. A VM will have one volume for each configured disk and another volume (representing the configuration) that always moves with Disk 1.

Synopsis



| Synopsis | |
|---------------------|--|
| Provides: | Storage resources for VMs to use Set the scope to a volume and view the Entity Information chart to see a list of VM-related files (such as VMDKs) contained in the volume. Set the scope to a VM to see a list of volumes attached to the VM. |
| Consumes: | Datacenter resources |
| Discovered through: | Hypervisor targets |

Actions

Intersight Workload Optimizer supports the following actions:

■ Move

Move a VM's volume (virtual storage) due to excess utilization of the current datastore, or for more efficient utilization of datastores in the environment.

Points to consider:

- The default global policy includes a setting that directs Intersight Workload Optimizer to use relevant metrics when analyzing and recommending actions for volumes. For details, see [Enable Analysis of On-prem Volumes \(on page 576\)](#).
- Intersight Workload Optimizer will not recommend moving a volume to a datastore that is currently in maintenance mode. Any volume in that datastore should move to an active datastore (for example, via vMotion).

■ Reconfigure

Reconfigure a VM's volume (virtual storage) to comply with placement policies.

On-prem Volume Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Placement Policies

By default, all on-prem volumes associated with a storage will move together rather than independently. You can create placement policies to place individual volumes on groups of storage. To ensure successful placement, be sure to also turn on the setting `Enable Analysis of On-prem Volumes` in the default global policy.

For more information, see [Creating Placement Policies \(on page 569\)](#) and [Enable Analysis of On-prem Volumes \(on page 576\)](#)

Automation Workflow

| Action | Default Mode |
|--------|--------------|
| Move | Manual |

Cloud Storage Tiers

This policy setting works with plans that simulate migration of on-prem volumes to the cloud. When you create the policy, be sure to set the scope to on-prem volumes and then select the cloud storage tiers that they can migrate to. Intersight Workload Optimizer treats these tiers as constraints when you run a Migrate to Cloud plan that includes the volumes defined in the policy.

| Attribute | Default Value |
|---------------------|---------------|
| Cloud Storage Tiers | None |

Click **Edit** to set your preferences. In the new page that displays, expand a **cloud tier** (a family of instance types, such as *Premium*) to see individual instance types.

Select your preferred instance types or cloud tiers, or clear the ones that you want to avoid. After you save your changes, the main page refreshes to reflect your selections.

Virtual Data Center (Private Cloud)

A virtual data center (vDC) is a collection or pool of resources that groups the resources around specific requirements or business needs. In private cloud environments, Intersight Workload Optimizer discovers the infrastructure that provides resources to the cloud, and the workloads that run on the cloud. To manage these resources, private clouds organize the infrastructure into Provider and Consumer virtual data centers.

NOTE:

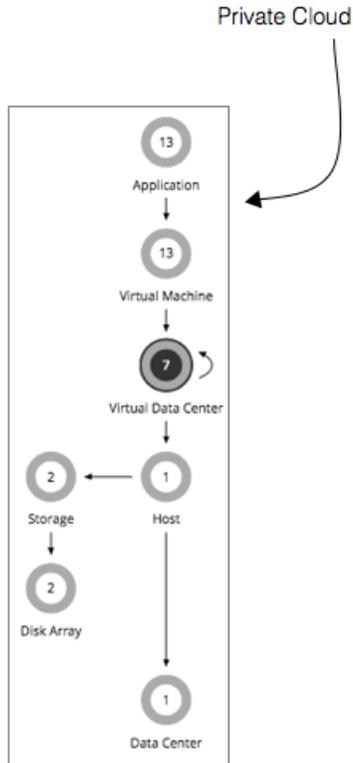
Different targets use different names to refer to virtual data centers. In the Intersight Workload Optimizer supply chain, these entities are all represented by Consumer and Provider VDCs, as follows:

| | | |
|-------------------------------|-----------------------|-----------------------|
| Intersight Workload Optimizer | vCenter Server | VMM |
| Consumer VDC | Resource Pool (Child) | Tenant or TenantQuota |
| Provider VDC | Resource Pool (Root) | Cloud |

Provider Virtual Data Centers

A provider virtual data center (vDC) is a collection of physical resources (hosts and data stores) within a cloud stack. The cloud administrator has access to these resources, and defines the data center members. A Provider vDC is created to manage resources that will be allocated to external customers through one or more Consumer vDCs.

Synopsis



A Provider vDC gains its budget by selling resources to the Consumer vDCs that it hosts. If utilization falls off, the data center loses budget. Ultimately, if the budget isn't enough to pay for the services it consumes, Intersight Workload Optimizer will recommend decommissioning the Provider vDC.

| Synopsis | |
|---------------------|--|
| Provides: | Physical resources (hosts and datastores) to Consumer vDCs |
| Consumes: | Hosts and datastores from the physical infrastructure |
| Discovered through: | Private Cloud Stack Managers |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Memory (Mem)
Memory is the measurement of memory that is reserved or in use.
- CPU
CPU is the measurement of CPU that is reserved or in use.
- Storage
Storage is the utilization of the storage attached to the entity.

Actions

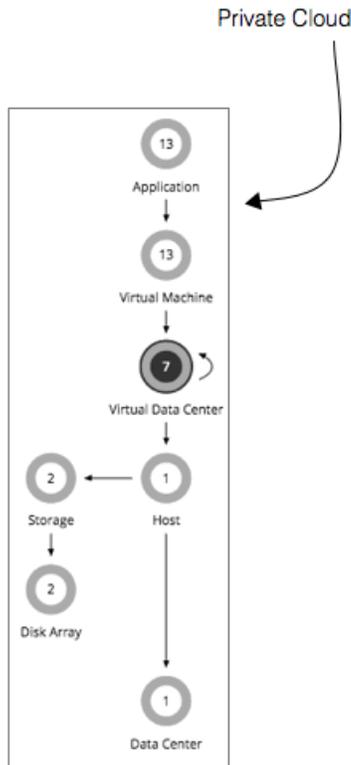
None

Intersight Workload Optimizer does not recommend actions for a Virtual Data Center. Instead, it recommends actions for the entities that provide resources to the Virtual Data Center.

Consumer Virtual Data Centers

A Consumer Virtual Data center (vDC) is a collection of resources that are available for external customers to manage workload through the private cloud. It is an environment customers can use to store, deploy, and operate virtual systems. Consumer Data centers use the resources supplied by a Provider Data center.

Synopsis



A Consumer vDC gains its budget as a function of its activity. The higher the utilization of the vDC, the more Intersight Workload Optimizer assumes the vDC is selling its services to a user. If utilization is high enough on a Consumer vDC, Intersight Workload Optimizer can increase resources for the vDC. If utilization falls off, Intersight Workload Optimizer can reduce resource capacity, or ultimately recommend terminating the vDC.

Intersight Workload Optimizer can also resize VMs through the Consumer vDC in response to changes in VM utilization.

| | |
|---------------------|-----------------------------------|
| Synopsis | |
| Provides: | Resources to host virtual systems |
| Consumes: | Provider vDC |
| Discovered through: | Cloud Stack Managers |

While users can see some of the physical resources that support the Consumer vDC, consumer-level users cannot modify these physical resources. Users of Consumer vDCs make changes to how the virtual devices are deployed in that environment, but they must ask the Provider vDC administrator to add more physical resources to be used by the Consumer vDC. Likewise,

Intersight Workload Optimizer can change resources on the VMs running in the vDC, but it does not make any changes to physical resources through this vDC.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Memory (Mem)
 - Memory is the measurement of memory that is reserved or in use.
- CPU
 - CPU is the measurement of CPU that is reserved or in use.
- Storage
 - Storage is the utilization of the storage attached to the entity.

Actions

Intersight Workload Optimizer does not recommend actions to perform on a Consumer vDC. Instead, it recommends actions to perform on the entities running in the Provider vDC.

Host

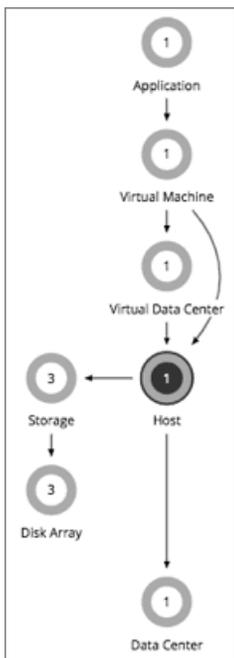
For on-prem environments, a host is a server that runs processes, including hypervisor processes to host virtual workloads. Note that a host is not necessarily a physical piece of hardware. A VM can be set up as a server that runs a hypervisor, and in turn it can host other VMs within its processing space. However, it is typical to use physical hardware as your hosts.

NOTE:

To support vSAN storage in your environment, you can deploy HCI Hosts. Intersight Workload Optimizer discovers the vSAN as a storage entity that consumes resources from the underlying hosts. For more information, see [vSAN Storage \(on page 379\)](#).

On the public cloud a host is an availability zone. This is where your cloud workloads run. For details, see [Zone \(on page 332\)](#).

Synopsis



A host gains its budget by selling resources to the workloads that run on it. The more workloads running on a host, the more budget the host has to purchase storage and datacenter resources. If utilization of a host is high enough, Intersight Workload

Optimizer can recommend that you provision a new one. If utilization falls off, the host loses budget. Ultimately, if the budget is not enough to pay for the services it consumes, Intersight Workload Optimizer will recommend to suspend or power off the host.

| Synopsis | |
|---------------------|--|
| Provides: | Host resources for VMs to use, including: <ul style="list-style-type: none"> ■ Memory ■ CPU ■ IO throughput ■ Net (network throughput) ■ Swap ■ Ballooning (sharing of memory among hosted VMs) ■ CPU Ready Queue |
| Consumes: | Datacenter resources (physical space, cooling, etc.) and storage |
| Discovered through: | Hypervisor targets For some hypervisor vendors, the host is the target, and for others the hosts are managed by the specified target. |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Memory (Mem)**
Memory is the measurement of memory that is reserved or in use.
- **CPU**
CPU is the measurement of CPU that is reserved or in use.
- **IO**
IO is the utilization of a host's IO adapters.
- **Net**
Net is the utilization of data through the host's network adapters.
- **Swap**
Swap is the measurement of a host's swap space that is in use.
- **Balloon**
Balloon is the measurement of memory that is shared by VMs running on a host.
- **CPU Ready**
CPU Ready is the measurement of a host's ready queue capacity that is in use.

For hosts discovered via vCenter targets, the following resources are monitored:

- **Energy**
Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).
- **Carbon Footprint**
Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.

Actions

Intersight Workload Optimizer supports the following actions:

- **Start**
Start a suspended host when there is increased demand for physical resources.
- **Provision**

Provision a new host in the environment when there is increased demand for physical resources. Intersight Workload Optimizer can then move workloads to that host.

- **Suspend**

When physical resources are underutilized on a host, move existing workloads to other hosts and then suspend the host.

NOTE:

Intersight Workload Optimizer discovers VMware HA configurations in clusters, and considers the reserved resources in its calculations. For tolerated host failures, or a reserved percentage of cluster resources, Intersight Workload Optimizer automatically sets utilization constraints for that cluster. If you configure a failover host, Intersight Workload Optimizer reserves that host for HA and will not move VMs to it.

DRS Automation Settings

Intersight Workload Optimizer automatically discovers DRS automation settings for vSphere hosts managed through vCenter. When you set the scope to a vSphere host and then view the Entity Information chart, the following information displays:

- Vendor Automation Mode

The chart shows the automation mode discovered from vCenter – Not Automated, Partially Automated, or Fully Automated.

- Vendor Migration Level

Intersight Workload Optimizer assigns a vendor migration level based on the migration level discovered from vCenter. The chart only shows the assigned migration level (i.e., the Intersight Workload Optimizer Vendor Migration Level).

| Intersight Workload Optimizer Vendor Migration Level | vCenter Migration Level |
|--|-------------------------|
| 1 (Conservative) | 5 |
| 2 (Less Conservative) | 4 |
| 3 (Moderate) | 3 |
| 4 (Less Aggressive) | 2 |
| 5 (Aggressive) | 1 |

Host Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For details about host actions, see [Host Actions \(on page 368\)](#).

| Action | Default Mode | vCenter | Hyper-V | UCS |
|-----------|--------------|---------|---------|-----|
| Start | Recommend | | | |
| Suspend | Recommend | | | |
| Provision | Recommend | | | |

For ServiceNow:

- Host provision actions will not generate a CR.
- For host suspend actions to succeed, it must be enabled in the given hypervisor, and there must be no VMs currently running on that host.

Maintenance Automation Avoidance

| Attribute | Default Setting |
|----------------------------------|-----------------|
| Maintenance Automation Avoidance | 30 minutes |

The Maintenance Automation Avoidance setting applies to vCenter environments with DRS clusters. Intersight Workload Optimizer uses this setting when:

- Intersight Workload Optimizer actions to move VMs from one host to another are automated.
- The DRS [automation level](#) is *Fully Automated*, regardless of [migration threshold](#).

NOTE:

Intersight Workload Optimizer automatically discovers DRS automation levels and migration thresholds and displays them in the Entity Information chart for hosts.

- Host maintenance is in effect.

This setting prevents action conflicts between Intersight Workload Optimizer and DRS.

When a host enters maintenance mode, DRS starts to move VMs on the host to other hosts to prepare for maintenance. In response, Intersight Workload Optimizer clears all pending actions to and from the host. For example, assume a cluster with Host_01, Host_02, and Host_03. When Host_01 enters maintenance mode, Intersight Workload Optimizer removes the following pending actions from the system:

- Move a VM on Host_01 to Host_02.
This prevents a potential conflict with a DRS action that moves the VM to Host_03.
- Move a VM on Host_02 to Host_01.
Since Host_02 and Host_03 are not in maintenance mode, Intersight Workload Optimizer might recommend moving the VM from Host_02 to Host_03 as an alternative action.

In addition, Intersight Workload Optimizer treats a host entering or in maintenance mode as uncontrollable and stops generating actions for the host. The host remains uncontrollable after it leaves maintenance mode, but is within the Maintenance Automation Avoidance period that you specified. During rolling host maintenance operations on DRS clusters, where hosts undergo maintenance on a staggered basis, this gives DRS a window (30 minutes by default) to move VMs from hosts entering maintenance to hosts that have recently left maintenance, thereby avoiding any potential conflict.

When the Maintenance Automation Avoidance period is over, Intersight Workload Optimizer treats the host as controllable and resumes action generation. At this stage, it is assumed that all critical DRS activities on the host have been completed, so Intersight Workload Optimizer actions should be safe to execute.

The following table summarizes Intersight Workload Optimizer's response at various stages of maintenance.

| Maintenance Status | DRS Activities | Host Status in Intersight Workload Optimizer | Intersight Workload Optimizer Pending Actions | Intersight Workload Optimizer New Actions |
|---|--|--|---|---|
| Host is entering maintenance mode. | Increased number of DRS activities moving VMs away from the host entering maintenance | X Uncontrollable (Maintenance) | # Removed from the system | X Not generated |
| Host is in maintenance mode. | Maintenance tasks on the host | X Uncontrollable (Maintenance) | N/A | X Not generated |
| Host has left maintenance mode but is within the Maintenance Automation Avoidance window. | Increased number of DRS activities moving VMs away from other hosts entering maintenance | X Uncontrollable (Maintenance) | N/A | X Not generated |

| Maintenance Status | DRS Activities | Host Status in Intersight Workload Optimizer | Intersight Workload Optimizer Pending Actions | Intersight Workload Optimizer New Actions |
|--|--|--|---|---|
| Host has left maintenance mode and is outside the Maintenance Automation Avoidance window. | Minimal number of DRS activities on the host | # Controllable | N/A | # Generated |

Points to consider:

- You can set a different Maintenance Automation Avoidance value that aligns with your host maintenance practices. For example, if moving VMs back to a host typically takes an hour, specify a value of 60.
- You can set a global value in the default policy for hosts, or specific values in automation policies that you create for your clusters.
- For rolling maintenance of hosts in a cluster, where hosts undergo maintenance on a staggered basis, there could be a point in the process where some or all hosts are uncontrollable. This means that Intersight Workload Optimizer cannot recommend actions to alleviate pressure on overburdened hosts. As such, these hosts could lose performance while they are uncontrollable.
- This setting has no effect on clusters where the DRS automation level is *Manual* or *Partially Automated*. As soon as a host enters maintenance mode, Intersight Workload Optimizer automates the first action to move a VM to another host, and then stops recommending actions. After the host leaves maintenance mode, Intersight Workload Optimizer automates actions to manage the performance of the cluster as normal.

Utilization Constraints

Utilization constraints affect the actions Intersight Workload Optimizer recommends as it manages your environment. Intersight Workload Optimizer recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|-------------------------|---------------|
| Net Throughput | 50 |
| Memory Utilization | 100 |
| IO Throughput | 50 |
| Swapping Utilization | 20 |
| CPU Utilization | 100 |
| Ready Queue Utilization | 50 |

For example:

- Setting 50 for Net Throughput means that Intersight Workload Optimizer considers 50% utilization of that throughput to be 100% of capacity and 25% utilization to be 50% of capacity.
- Setting 100 for Memory Utilization means that Intersight Workload Optimizer capacity reflects the physical capacity for this resource.

Desired State

The desired state for your environment is an n-dimensional sphere that encompasses the fittest conditions your environment can achieve.

| Attribute | Default Value |
|-----------|---------------|
| Center | 70 |
| Diameter | 10 |

The multiple dimensions of this sphere are defined by the resource metrics in your environment. Metric dimensions include VMem, storage, CPU, etc. While the metrics on the devices in your environment can be any value, the desired state, this n-dimensional sphere, is the subset of metric values that assures the best performance while achieving the most efficient utilization of resources that is possible.

The Desired State settings define the center of the sphere as well as its diameter. This is a way for you to customize what Intersight Workload Optimizer considers to be the desired state.

Setting the center of the sphere chooses the priority for Intersight Workload Optimizer analysis. If you set the balance in favor of efficiency, Intersight Workload Optimizer tends to place more VMs on fewer physical hosts, and to give them storage capacity from fewer data stores. As a result, high utilization can have more impact on QoS. With a balance in favor of performance, Intersight Workload Optimizer tends to spread virtual loads across more physical devices. This can result in the provisioning of excess resources.

The diameter setting determines the range of deviation from the center that can encompass the desired state. If you specify a large diameter, Intersight Workload Optimizer will have more variation in the way it distributes workload across hosting devices.

As you move each slider, a tooltip displays the numerical value of the setting. **Center** indicates the percentage of resource utilization you want, within the range you specify as **Diameter**. For example, if you want utilization of 75%, plus or minus 10%, then you would set **Center** = 75 and **Diameter** = 20. Intersight Workload Optimizer recommends actions that tend toward this desired state much as possible, given the dependencies within the current environment.

NOTE:

The setting for Target Utilization can have an effect on plans that you run. If you disable provisioning and suspension for hosts and datastores, then you should always set Center and Diameter to their default values.

Over Provisioning Constraints

These attributes allow capacity planners to overprovision CPU and Memory resources at the cluster level, while maintaining a limit on resource usage for the hosts for compliance. Overprovisioning at the cluster level increases density and efficiency by analyzing the cluster as a whole. There are two cluster level policy settings to control overprovisioning capacity percentage (CPU Overprovisioned Percentage and Memory Overprovisioned Percentage) and two host level settings to control the maximum-allowed utilization percentage (Host CPU Overprovisioned Max Util and Host Memory Overprovisioned Max Util).

| Attribute | Default Value |
|--------------------------------------|---------------|
| CPU Overprovisioned Percentage | 30000 |
| Memory Overprovisioned Percentage | 1000 |
| Host CPU Overprovisioned Max Util | 200 |
| Host Memory Overprovisioned Max Util | 200 |

■ CPU Overprovisioned Percentage

CPU Overprovisioned Percentage is a cluster level setting to control overprovisioning capacity percentage for CPU Provisioned. Setting 30000 for CPU Overprovisioned Percentage means that the CPU Provisioned capacity will be 300 times the actual CPU capacity. The calculation for CPU Provisioned can be expressed as follows:

$$\text{CPU} * (\text{CPU Overprovisioned Percentage} / 100)$$

Assume the following data for determining the CPU Provisioned capacity:

| Attribute | Value |
|--------------------------------|----------|
| CPU | 38.4 GHz |
| CPU Overprovisioned Percentage | 30000 |

In this example, Intersight Workload Optimizer calculates the CPU Provisioned capacity as follows:

$$38.4 * (30000 / 100) = 11.52 \text{ THz}$$

■ Memory Overprovisioned Percentage

Memory Overprovisioned Percentage is a cluster level setting to control overprovisioning capacity percentage for Memory Provisioned. Setting 1000 for **Memory Overprovisioned Percentage** means that the Memory Provisioned capacity will be 10 times the actual Memory capacity. The calculation for Memory Provisioned can be expressed as follows:

$$\text{Memory} * (\text{Memory Overprovisioned Percentage} / 100)$$

Assume the following data for determining the Memory Provisioned capacity:

| Attribute | Value |
|-----------------------------------|-----------|
| Memory | 287.91 GB |
| Memory Overprovisioned Percentage | 1000 |

In this example, Intersight Workload Optimizer calculates the CPU Provisioned capacity as follows:

$$287.91 * (1000 / 100) = 2.81 \text{ TB}$$

■ Host CPU Overprovisioned Max Util

Host CPU Overprovisioned Max Util is a host level setting to control max allowed utilization percentage for CPU Provisioned. Setting 200 for Host CPU Overprovisioned Max Util means that the maximum-allowed utilization percentage for CPU Provisioned capacity is 200. This means that once the host reaches the specified percentage of utilization, no other VM can be placed over the same host. The calculation for CPU Provisioned can be expressed as follows:

$$\text{CPU} * (\text{CPU Overprovisioned Percentage} / 100)$$

Then, the calculation for Max Allowed CPU Provisioned Used can be expressed as follows:

$$\text{CPU Provisioned} * (\text{Host CPU Overprovisioned Max Util} / 100)$$

Assume the following data for determining the CPU Provisioned capacity:

| Attribute | Value |
|-----------------------------------|----------|
| CPU | 38.4 GHz |
| CPU Overprovisioned Percentage | 30000 |
| Host CPU Overprovisioned Max Util | 105% |

In this example, Intersight Workload Optimizer calculates the CPU Provisioned capacity as follows:

$$38.4 * (30000 / 100) = 11.52 \text{ THz}$$

Then, Intersight Workload Optimizer calculates the Max Allowed CPU Provisioned Used as follows:

$$11.52 * (105 / 100) = 12.1 \text{ THz}$$

■ Host Memory Overprovisioned Max Util

Host Memory Overprovisioned Max Util is a host level setting to control max allowed utilization percentage for Memory Provisioned. Setting 200 for Host Memory Overprovisioned Max Util means that the maximum-allowed utilization for Memory Provisioned capacity is 200. This means that once the host reaches the specified percentage of utilization, no other VM can be placed over the same host. The calculation for Memory Provisioned can be expressed as follows:

$$\text{Memory} * (\text{Memory Overprovisioned Percentage} / 100)$$

Then, the calculation for Max Allowed Memory Provisioned Used can be expressed as follows:

$$\text{Memory Provisioned} * (\text{Host Memory Overprovisioned Max Util} / 100)$$

Assume the following data for determining the Memory Provisioned capacity:

| Attribute | Value |
|-----------------------------------|-----------|
| Memory | 287.91 GB |
| Memory Overprovisioned Percentage | 1000 |

| Attribute | Value |
|--------------------------------------|-------|
| Host Memory Overprovisioned Max Util | 108% |

In this example, Intersight Workload Optimizer calculates the Memory Provisioned capacity as follows:

$$287.91 * (1000 / 100) = 2.81 \text{ TB}$$

Then, Intersight Workload Optimizer calculates the Max Allowed Memory Provisioned Used as follows:

$$2.81 * (108 / 100) = 3.03 \text{ TB}$$

Points to consider:

Overprovisioning works on both the cluster and host levels. The following example uses the Memory commodity; however, the same considerations are applicable for CPU.

- **Cluster level:** Let's say you have a cluster with four hosts, each with 50 GB Memory capacity. The cluster's actual Memory capacity will be $4 * 50 = 200$ GB. If the cluster Memory Overprovisioned Percentage is set to 1000, then the overall cluster's Memory Provisioned will be $200 * (1000 / 100) = 2000$ GB.
- **Host level:** On host level, each host is allowed to be overprovisioned to twice the cluster's Memory Overprovisioned Percentage. In this example, each host in the cluster can be overprovisioned up to $50 * (1000 / 100) * 2 = 1000$ GB, as long as the overall sum for the Memory Provisioned capacity of all the hosts still adds up to the cluster's Memory Provisioned capacity (2000 GB in this example).
- In some financial institutions, customers have special groups of hosts called "Platinum Hosts" that have a policy stating that the overprovisioned commodity cannot breach the cluster's overprovisioned ratio. In other words, the overprovisioned commodity (in this example, Memory Provisioned) should not breach the 100% utilization mark on the "Platinum Hosts." In such cases, you can set the Host Memory Overprovisioned Max Util attribute to 100, meaning that the maximum-allowed Memory Provisioned Used for each platinum host can only go up to $50 * \text{Cluster's Memory Overprovisioned Percentage} * \text{Host Memory Overprovisioned Max Util}$.

In this example, Intersight Workload Optimizer calculates the Memory Provisioned Used as follows:

$$50 * (1000 / 100) * (100/100) = 500 \text{ GB}$$

Placement Policies

You can create placement policies that merge multiple clusters into a single logical group for workload placement.

For example, you can merge three host clusters in a single provider group. This enables Intersight Workload Optimizer to move workload from a host in one of the clusters to a host in any of the merged clusters to increase efficiency in your environment.

For more information, see [Creating Placement Policies \(on page 569\)](#).

NOTE:

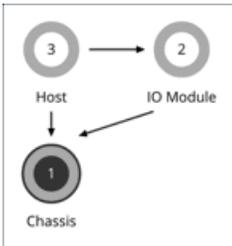
For vCenter, Cisco automatically imports any vSphere Host DRS rules when DRS is enabled, and displays them on the **Settings > Policies** page under **Imported Placement Policies**.

For more information, see [Importing Workload Placement Policies \(on page 569\)](#).

Chassis

A chassis houses the servers that are part of a computing fabric. It provides compute, memory, storage, and bandwidth resources.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Chassis resources (physical space, cooling, etc.) |
| Consumes: | N/A |
| Discovered through: | Fabric Manager targets |

NOTE:

When Intersight Workload Optimizer discovers that blade servers housed in a particular chassis have been designated as vCenter hosts, the supply chain stitches the blade servers and chassis to the corresponding vCenter data center to establish their relationship. When you set the scope to that data center and view the Health chart, you will see the blade servers in the list of hosts. In addition, when the data center is included in a merge policy (a policy that merges data centers for the purpose of VM placement), the VMs in the blade servers apply the policy, allowing them to move between data centers as necessary.

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Power
 - Power is the measurement of electricity consumed by a given entity, expressed in watts.
- Cooling
 - Cooling is the percentage of the acceptable temperature range that is utilized by the entity. As the temperature nears the high or low running temperature limits, this percentage increases.

Actions

None

Intersight Workload Optimizer does not recommend actions for a chassis.

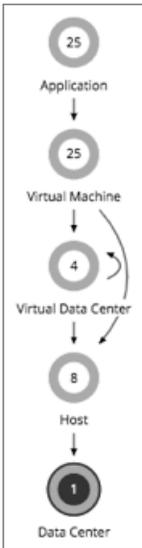
Data Center

A data center is the sum of VMs, hosts, datastores, and network devices that are managed by a given hypervisor target. A data center provides compute, memory, storage, and bandwidth resources.

NOTE:

For public cloud environments, a data center is the cloud region. The hosts that get resources from the data center are availability zones within that region. For details, see [Region \(on page 334\)](#) and [Zone \(on page 332\)](#).

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Compute, memory, storage, and bandwidth resources |
| Consumes: | N/A |
| Discovered through: | Hypervisor targets |

NOTE:

When Intersight Workload Optimizer discovers that blade servers housed in a particular chassis have been designated as vCenter hosts, the supply chain stitches the blade servers and chassis to the corresponding vCenter data center to establish their relationship. When you set the scope to that data center and view the Health chart, you will see the blade servers in the list of hosts. In addition, when the data center is included in a merge policy (a policy that merges data centers for the purpose of VM placement), the VMs in the blade servers apply the policy, allowing them to move between data centers as necessary.

Monitored Resources

Intersight Workload Optimizer does not monitor resources directly from the data center, but it does monitor the following resources, aggregated for the hosts in a data center:

- **Memory (Mem)**
Memory is the measurement of memory that is reserved or in use.
- **CPU**
CPU is the measurement of CPU that is reserved or in use.
- **IO**
IO is the utilization of a host's IO adapters.
- **Net**
Net is the utilization of data through the host's network adapters.
- **Swap**
Swap is the measurement of a host's swap space that is in use.
- **Balloon**
Balloon is the measurement of memory that is shared by VMs running on a host.
- **CPU Ready**
CPU Ready is the measurement of a host's ready queue capacity that is in use.

For vCenter targets, the following resources are also monitored:

- Energy
Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).
- Carbon Footprint
Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.

Actions

None

Intersight Workload Optimizer does not recommend actions for a data center. Instead, it recommends actions for the entities running in the data center.

Data Center Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Operational Constraints

Intersight Workload Optimizer [calculates \(on page 519\)](#) carbon footprint using industry standards that take into account energy consumption, datacenter efficiency, and carbon intensity data. You can create Data Center policies to adjust the calculations according to the requirements of your data centers. For example, a data center in a particular location might have different requirements than data centers in other locations. After you adjust the calculations via policies, Intersight Workload Optimizer can accurately report your organization's carbon footprint.

| Attribute | Default Value |
|---------------------------------|---------------|
| Power Usage Effectiveness (PUE) | 1.5 |
| Carbon Intensity (CI) | .25 (g/Wh) |

- Carbon Intensity (CI) is a measurement of how 'clean' electricity is. It refers to how many grams of carbon dioxide (CO₂) are released to produce 1 watt-hour (Wh) of electricity. Electricity that is generated using fossil fuels is more carbon intensive, as the process by which it is generated creates CO₂ emissions. Renewable energy sources, such as wind, hydro, or solar power produce next to no CO₂ emissions, so their carbon intensity value is much lower and often zero.
- Power Usage Effectiveness (PUE) is a ratio that describes how efficiently a computer data center uses energy; specifically, how much energy is used by the computing equipment. PUE is the ratio of the total amount of energy used by a computer data center facility to the energy delivered to computing equipment. The closer PUE is to 1, the more efficient the computer data center.

Placement Policies

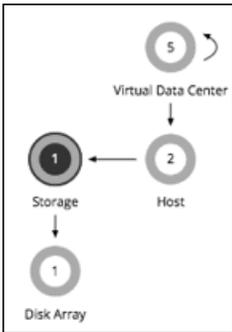
For vCenter environments, you can create placement policies that merge data centers to support cross-vCenter moves. In this case, where a data center corresponds to a vCenter target, the merged clusters can be in different data centers. In this case, you must create two merge policies; one to merge the affected data centers, and another to merge the specific clusters.

For more information, see [Creating Placement Policies \(on page 569\)](#).

Storage

Intersight Workload Optimizer represents storage as Datastores. A Datastore is a logical grouping of one or more physical storage devices that serve workload storage requirements.

Synopsis



A Datastore gains its budget by selling resources to the VMs it serves. If utilization of a Datastore is high enough, Intersight Workload Optimizer can recommend that you provision a new one.

| Synopsis | |
|---------------------|---|
| Provides: | Host resources for VMs to use, including: <ul style="list-style-type: none"> ■ Storage Amount ■ IOPS ■ Latency |
| Consumes: | Disk arrays (or aggregates) |
| Discovered through: | Hypervisor and Storage Controller targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage Amount**
Storage Amount is the measurement of storage capacity that is in use.
- **Storage Provisioned**
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.

NOTE:
When it generates actions, Intersight Workload Optimizer does not consider IOPS throttling that it discovers on storage entities. Analysis uses the IOPS it discovers on Logical Pool or Disk Array entities.
- **Latency**
Latency is the measurement of storage latency.

Storage Actions

Intersight Workload Optimizer supports the following actions:

- **Move**
For high utilization of physical storage, move datastore to a different disk array (aggregate).
- **Provision**
For high utilization of storage resources, provision a new datastore.
- **Resize**
Increase or decrease the datastore capacity.
- **Start**

For high utilization of storage resources, start a suspended datastore.

- **Suspend**

For low utilization of storage resources, move served VMs to other datastores and suspend this one.

- **Delete (datastore or volume)**

Delete a datastore or volume that has been suspended for a period of time.

- **Delete (unattached files)**

Delete a file on a datastore that has not been accessed or modified for a period of time.

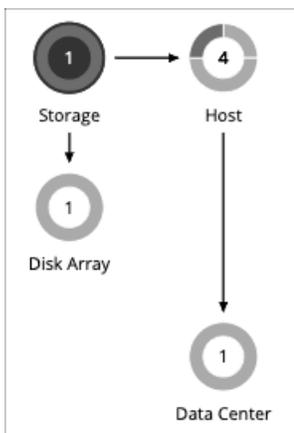
Storage resize actions use Intersight Workload Optimizer tuned scaling settings. This gives you increased control over the action acceptance mode Intersight Workload Optimizer will use for the affected actions. For an overview of tuned scaling, see [Tuned Scaling for On-prem VMs \(on page 339\)](#).

You can create placement policies to enforce constraints for storage move actions. For example, you can have a policy that allows storage to only move to certain disk arrays, or a policy that prevents storage from moving to certain disk arrays.

For more information, see [Creating Placement Policies \(on page 569\)](#).

vSAN Storage

Overview



For environments that use hyperconverged infrastructure to provide storage on a vSAN, Intersight Workload Optimizer can discover the storage provided by a host cluster as a single Storage entity. This Storage entity represents the full storage capacity that is provided by that host cluster.

Intersight Workload Optimizer supports VMware vSAN, but does not support stretched vSAN clusters. Adding stretched clusters can cause the generation of incorrect storage recommendations and actions.

vSAN Storage Capacity

When you consider vSAN capacity, you need to compare *Raw Capacity* with *Usable Capacity*.

- **Raw Capacity**

Intersight Workload Optimizer discovers Raw Capacity configured in vCenter and uses it to calculate Usable Capacity. Raw Capacity displays in the Entity Information chart.

- **Usable Capacity**

Intersight Workload Optimizer calculates Usable Capacity and then uses the calculated value to drive scaling actions. Intersight Workload Optimizer can recommend scaling the Storage Amount, Storage Provisioned, or Storage Access capacity. Usable Capacity displays in the Capacity and Usage chart.

Usable Capacity Calculation

To calculate Usable Capacity, Intersight Workload Optimizer considers a variety of attributes, including:

- Raw Capacity and Largest Host Capacity

Intersight Workload Optimizer compares the Raw Capacity for all the hosts in the cluster and then uses the largest value as Largest Host Capacity.

- RAID Factor

Intersight Workload Optimizer calculates RAID Factor based on the *Failures to Tolerate* (FTT) value and *Redundancy Method* that it discovers. FTT specifies how many failures a given cluster can tolerate, while Redundancy Method specifies the RAID level for the cluster.

| FTT | Redundancy Method | RAID Factor |
|-----|-------------------|-------------|
| 0 | RAID1 | 1 |
| 1 | RAID1 | 1/2 |
| 2 | RAID1 | 1/3 |
| 1 | RAID5/6 | 3/4 |
| 2 | RAID5/6 | 2/3 |

NOTE:

If discovery fails for some reason, Intersight Workload Optimizer uses a RAID Factor of 1.

- Host Capacity Reservation, Slack Space Percentage, and Compression Ratio

You can control the values for these attributes in storage policies. For details about these attributes and their effect on usable capacity calculations, see [Hyper-converged Infrastructure Settings \(on page 384\)](#).

The calculation for Usable Capacity can be expressed as:

$$\text{Usable Capacity} = (\text{Raw Capacity} - \text{Largest Host Capacity} * \text{Host Capacity Reservation}) * \text{Slack Space Percentage} * \text{RAID Factor} * \text{Compression Ratio}$$

If the result of the calculation is zero or a negative value, Intersight Workload Optimizer sets the Usable Capacity to 1 MB.

Capacity and Usage Chart for vSAN Storage

The **Capacity and Usage** chart for vSAN storage shows two Storage Amounts - *Consumed* (bought) and *Provided* (sold). This is because vSAN storage can buy and sell commodities to hosts.

For the *Provided* Storage Amount, the *Capacity* value corresponds to *Usable Capacity*, while the *Used* value indicates utilization.

Entity Information Chart for vSAN Storage

The **Entity Information** chart includes the following information:

- HCI Technology Type

The technology that supports this storage cluster. For this release, Intersight Workload Optimizer supports VMware vSAN technology.

- Capacity

Intersight Workload Optimizer displays rounded values for the following, which might be slightly different from the values it discovers from vCenter:

- Raw Capacity

The sum of the Raw Capacity that each storage capacity device provides.

- Raw Free Space

How much of the Raw Capacity is not currently in use.

- Raw Uncommitted Space

In terms of Raw Capacity, how much space is available according to your thin/thick provisioning.

- **Redundancy Method and Failures to Tolerate**

Redundancy Method specifies the RAID level employed for the cluster. RAID level impacts how much Usable Capacity you can see for a given Raw Capacity. You can use a RAID calculator to determine how the RAID level impacts your Usable Capacity.

Failures to Tolerate specifies how many capacity device failures a given cluster can tolerate. In practical terms, this means how many hosts can come down at the same time, without affecting storage. This value should match the RAID level.

Actions to Add vSAN Capacity

To scale up storage amount, you add additional hosts that are configured to include their storage in the vSAN array.

When you scope the session to the vSAN storage, you can see actions to scale:

- Storage Amount
- Storage Provisioned
- Storage Access

The action to scale up the storage indicates the amount of storage you need to add. It appears as a recommended action. In fact, to add storage you must add a new host.

When you scope the session to hosts that provide the capacity devices to the storage, you can see the following actions that are related to scaling up the storage capacity:

- Scale up StorageAmount for Storage [MyVsanStorageCluster]
- Provision Host [VSAN_HostName]

The action to provision a host includes details about the storage cluster. Because you need to manually add hosts to your on-prem environment, this appears as a recommended action.

Planning With vSAN Storage

For *Hardware Replace* and *Custom* plans, you can use HCI Host templates to add vSAN capacity. These represent the hosts that add storage capacity to a vSAN cluster. For more information, see [HCI Host Template Settings \(on page 591\)](#).

Under certain circumstances, *Add Virtual Machines* plans can fail to place workloads, or it can fail to generate actions to increase storage capacity by provisioning new hosts.

- If you scope the plan to a user-created group that only provides vSAN storage, or to a discovered storage cluster group, then the plan can fail to place VMs with multiple volumes. This can occur for VMs that use conventional storage (not vSAN) along with vSAN storage.
- If you scope the plan to a vSAN host group and add VMs, the plan can fail to increase storage capacity by provisioning new hosts. For example, assume you scope the plan to a vSAN host group and add 20 VMs to the environment. In that case, you need hosts to provide compute capacity for the VMs, and you also need hosts to provide storage capacity. The plan can represent the compute provisioning correctly, but it can incorrectly fail to add more storage capacity to the vSAN.
- If the vSAN RAID type is `Raid6/FTT=2`, if you scope the plan to any vSAN groups then the plan will fail to place any of the VMs.

Storage Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

The following are the storage actions and automation support for environments that do not include Disk Array Storage Controllers as targets. For details about these actions, see [Storage Actions \(on page 378\)](#).

| Action | Default Mode | vCenter | Hyper-V |
|--------------------|--------------|---|---|
| Delete (Datastore) | Manual |  |  |
| Delete (Volume) | Manual |  |  |

| Action | Default Mode | vCenter | Hyper-V |
|---|--------------|---|---|
| Delete Unattached Files Applicable for the on-prem environment only. | Manual |  |  |
| Move | Recommend |  |  |
| Provision | Recommend |  |  |
| Resize (Up, Down, Above Max, or Below Min - using tuned scaling) | Recommend |  |  |
| Start | Recommend |  |  |
| Suspend | Disabled |  |  |

For datastores on disk arrays:

| Action | Default Mode | NetApp ONTAP | VMAX | Pure Storage |
|--|--------------|---|---|---|
| Delete (Volume) | Recommend |  |  |  |
| Suspend | Manual |  |  |  |
| Delete (Datastore) | Disabled |  |  |  |
| Move | Recommend |  |  |  |
| Provision | Recommend |  |  |  |
| Start | Recommend |  |  |  |
| Resize (Up, Down, Above Max, or Below Min - using tuned scaling) | Recommend |  |  |  |

For ServiceNow:

- Storage suspend and vSAN storage resize actions will not generate a CR.
- Currently Intersight Workload Optimizer can only execute a CR for storage provision actions on Pure and Dell Compellent storage.

Utilization Constraints

Utilization constraints affect the actions Intersight Workload Optimizer recommends as it manages your environment. Intersight Workload Optimizer recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|---------------------------------|---------------|
| Storage Provisioned Utilization | 100 |
| IOPS Utilization | 100 |
| Storage Amount Utilization | 90 |
| Latency Utilization | 100 |

For example, setting 90 for Storage Amount Utilization means that Intersight Workload Optimizer considers 90% utilization of the physical storage to be 100% of capacity.

Storage Settings

| Attribute | Default Setting/Value |
|------------------------------------|--|
| Storage Latency Capacity | 277 ms |
| IOPS Capacity | 50000 |
| Storage Overprovisioned Percentage | 2000 |
| Minimum Unattached Files Size | 1000 KB |
| Directories to Ignore | \\.dvsData.* \\.snapshot.* \\.vSphere-HA.* \\.naa.* \\.etc.* lost +found.* |
| Files to Ignore | config\\.db stats\\.db.* |
| Generate Delete Action after | 15 Day(s) |
| Delete file after | 30 Day(s) |

■ Storage Overprovisioned Percentage

Storage Overprovisioned Percentage sets how much overprovisioning Intersight Workload Optimizer assumes when recommending actions for VM datastores. For example, if a datastore has a 30 GB capacity, and Storage Overprovisioned Percentage is set to 2000, Intersight Workload Optimizer will treat the datastore as though it has a capacity of 60 GB, or 200% of the actual datastore capacity.

■ IOPS Capacity

IOPS Capacity is the IOPS setting for individual datastores. To set a specific capacity for one group of datastores, select that group as the property scope and override the global setting for that scope.

Note that IOPS capacity for a disk array takes precedence – Datastores that are members of a disk array always have the IOPS capacity that is set to the disk array.

Intersight Workload Optimizer considers these settings when calculating utilization percentage. For example, assume IOPS Capacity of 500 for datastores. If utilization on a datastore is 250 IOPS, then the datastore is at 50% of capacity for that metric.

■ Storage Latency Capacity

Storage Latency Capacity sets the maximum storage latency to tolerate on a datastore, in ms. The default setting is 100 ms.

Intersight Workload Optimizer measures the latency experienced by all VMs and hosts that access the datastore. Assume a default setting of 100 ms. If a datastore exhibits latency of 50 ms, then the Intersight Workload Optimizer will show latency utilization of 50%.

For VMAX environments, Intersight Workload Optimizer discovers SLO for storage latency that you set in VMAX and uses it in analysis. However, if you set a higher storage latency value in a Intersight Workload Optimizer policy, analysis will use that value instead.

■ Minimum Unattached Files Size

You can make settings to control how Intersight Workload Optimizer tracks and reports on unattached storage in your environment. Unattached storage is any disk space devoted to files that are not required for operations of the devices or applications in your environment. Unattached storage may indicate opportunities for you to free up disk space, and provide more storage capacity to running VMs and applications.

If there are groups of datastores you do not want to track for unattached storage, set the given scope and disable datastore browsing there. If you prefer not to use Intersight Workload Optimizer resources to track unattached storage, leave the global setting checked.

The settings for **Directories to Ignore** and **Files to Ignore** specify directories and files that Intersight Workload Optimizer will not consider when looking for unattached data storage space. Separate items in these lists with the OR bar (“|”).

■ Generate Delete Action after

Generate Delete Action after specifies the duration of days a file is inactive before the "Delete Unattached Files" action is generated.

- **Delete file after**

Delete file after specifies the duration of days a file can be inactive before the delete action can be automatically executed. This setting must be larger than the duration of days the file is inactive.

Enable automatic deletion of unattached on-prem files by changing the Automation and Orchestration settings for "Delete Unattached Files" actions.

NOTE:

When the "Delete Unattached Files" action type is set to automated, Intersight Workload Optimizer deletes the unattached on-prem files after the specified number of days of inactivity. The action is not executed if the action type is left at the default (manual).

Scaling Constraints

| Attribute | Default Value |
|---------------------------------------|---------------|
| Increment Constant for Storage Amount | 100 GB |
| Rate of Resize | High (3) |

- **Increment Constant for Storage Amount**

This setting controls how many GB to add or subtract when resizing the allocation for a datastore.

- **Rate of Resize**

Intersight Workload Optimizer uses the Rate of Resize setting to determine how to make storage resize changes in a single action.

- **Low**

Change the value by one increment only.

- **Medium**

Change the value by an increment that is 1/4 of the difference between the current value and the optimal value.

- **High**

Change the value to be the optimal value.

This default value ensures that resizing to the desired state can be achieved in a single action. This is more efficient than smaller, incremental resizes.

Hyperconverged Infrastructure Settings

Intersight Workload Optimizer considers these settings when calculating capacity and utilization for hyperconverged environments.

| Attribute | Default Setting/Value |
|-----------------------------------|-----------------------|
| Compression Ratio | 1 |
| Host IOPS Capacity | 50000 |
| Slack Space Percentage | 25 |
| Host Capacity Reservation | 1 |
| Usable Space Includes Compression | Off |

NOTE:

Intersight Workload Optimizer uses Host Capacity Reservation, Slack Space Percentage, and Compression Ratio to calculate vSAN usable capacity and drive scaling actions. For more information about usable capacity and how it is calculated, see [vSAN Storage \(on page 379\)](#).

- **Host Capacity Reservation**

When a host must be taken out of service for maintenance, vSphere will evacuate the data from that host and move it to other hosts in the cluster to maintain the integrity of the replication demanded by the storage policy. For this to happen, there must be enough free raw capacity available to accept the data being evacuated.

Intersight Workload Optimizer uses this setting to determine how many hosts worth of capacity it should subtract from the raw capacity amount before calculating usable capacity. This is not the same as redundancy. It does not specify how the array distributes data to maintain integrity.

- **Host IOPS Capacity**

In addition to calculating usable capacity, Intersight Workload Optimizer needs an estimate of datastore IOPS capacity (storage access). Intersight Workload Optimizer uses the value that you set to provide an estimate of effective IOPS capacity for each host in the cluster. Total IOPS capacity is the number of hosts in the cluster multiplied by Host IOPS Capacity.

- **Slack Space Percentage**

It is recommended that a vSAN datastore never be filled to prevent vSphere from moving objects/files around the cluster to balance the datastore across all the hosts.

Intersight Workload Optimizer reduces usable capacity by the percentage that you set.

- **Compression Ratio**

vSAN supports both deduplication and compression, which may increase the amount of usable capacity on the datastore. Intersight Workload Optimizer does not try to predict the deduplication or compression ratio, but you can choose to include a compression ratio into the usable capacity calculation. This captures the ratio achieved both by compression and deduplication.

The compression ratio that you set acts as a multiplier on the raw capacity to calculate usable capacity. For example, a compression ratio of 2 would double the amount of usable capacity. The default value of 1 means no compression.

- **Usable Space Includes Compression**

Turn this on if you want Intersight Workload Optimizer to consider the compression ratio when calculating storage utilization and capacity. Whether this is on or off, Intersight Workload Optimizer always considers compression when calculating utilization of StorageProvisioned.

Placement Policies

Intersight Workload Optimizer supports placement policies for storage and storage clusters.

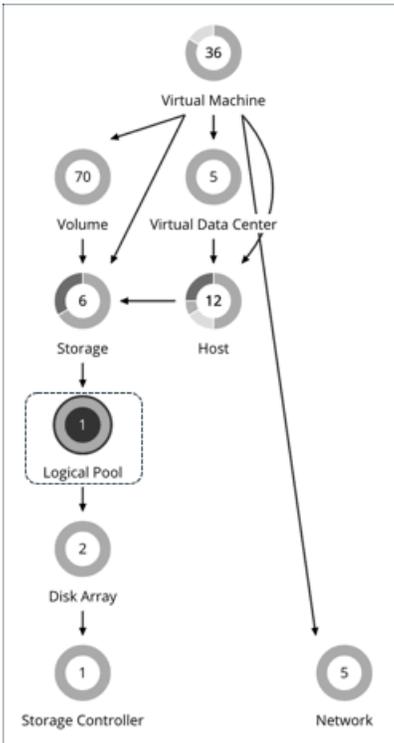
- You can create placement policies to enforce constraints for storage move actions. For example, you can have a policy that allows storage to only move to certain disk arrays, or a policy that prevents storage from moving to certain disk arrays.
- You can create placement policies that merge multiple clusters into a single logical group for workload placement.

For more information, see [Creating Placement Policies \(on page 569\)](#).

Logical Pool

A logical pool represents storage resources that are managed together and presented as a single storage system. Intersight Workload Optimizer analysis identifies performance and efficiency opportunities for a logical pool. For example, it can recommend moving resources into or out of a logical pool, or aggregating resource capacity within the pool.

Synopsis



| Synopsis | |
|---------------------|----------------------|
| Provides: | Storage resources |
| Consumes: | Disk array resources |
| Discovered through: | Storage targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- **Storage Amount**
Storage Amount is the measurement of storage capacity that is in use.
- **Storage Provisioned**
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- **Latency**
Latency is the measurement of storage latency.

Logical Pool Actions

Intersight Workload Optimizer supports the following actions:

- **Move**
For high utilization of physical storage, move the logical pool to a different disk array (aggregate).
- **Provision**
For high utilization of storage resources, provision a new logical pool.

- **Resize**
Increase or decrease the logical pool capacity.
- **Start**
For high utilization of storage resources, start a suspended logical pool.
- **Suspend**
For low utilization of storage resources, move served VMs to other logical pools and suspend this one.

Logical Pool Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

| Action | Default Mode |
|-----------|--------------|
| Move | Disabled |
| Provision | Disabled |
| Resize | Recommend |
| Start | Disabled |
| Suspend | Disabled |

Storage Settings

| Attribute | Default Value |
|------------------------------------|---------------|
| IOPS Capacity | 50000 |
| Storage Latency Capacity | 100 ms |
| Storage Overprovisioned Percentage | 200 |

- **Storage Latency Capacity**
Storage Latency Capacity sets the maximum storage latency to tolerate on a logical pool, in ms. The default setting is 100 ms.
- **Storage Overprovisioned Percentage**
Storage Overprovisioned Percentage sets how much overprovisioning Intersight Workload Optimizer assumes when recommending actions for logical pools.
- **IOPS Capacity**
IOPS Capacity is the IOPS setting for individual logical pools.

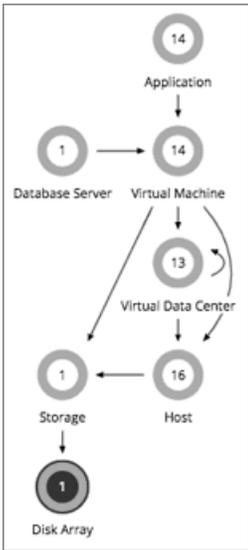
NOTE:

Intersight Workload Optimizer discovers storage latency and IOPS capacities that you set in your environment (for example VMAX) and uses them in its analysis. These capacities will be overridden by values that you set in Intersight Workload Optimizer policies.

Disk Array

A Disk Array (an aggregate) is a data storage system made up of multiple disk drives. For example, a RAID is an aggregate that implements redundancy and other data management features. A disk array provides storage volumes to serve the storage requirements of physical machines. It uses the resources of one storage controller, which manages the disk array operation.

Synopsis



| Synopsis | |
|---------------------|---|
| Provides: | Storage resources for datastores to use, including: <ul style="list-style-type: none"> ■ Storage Amount ■ Storage Provisioned ■ IOPS (storage access operations per second) ■ Latency (capacity for disk latency in ms) |
| Consumes from: | Storage Controllers |
| Discovered through: | Storage Controller targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

NOTE:

Not all targets of the same type provide all possible commodities. For example, some storage controllers do not expose CPU activity. When a metric is not collected, the corresponding chart in the user interface will not display data.

- **Storage Amount**
Storage Amount is the measurement of storage capacity that is in use.
- **Storage Provisioned**
Storage provisioned is the utilization of the entity's capacity, including overprovisioning.
- **Storage Access (IOPS)**
Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity.
- **Latency**
Latency is the measurement of storage latency.

Disk Array Actions

Intersight Workload Optimizer supports the following actions:

- **Provision**

For high utilization of the disk array's storage, provision a new disk array. This action can only be executed outside Intersight Workload Optimizer.

- **Start**

For high utilization of disk array, start a suspended disk array. This action can only be executed outside Intersight Workload Optimizer.

- **Suspend**

For low utilization of the disk array's storage, move VMs to other datastores and suspend volumes on the disk array. This action can only be executed outside Intersight Workload Optimizer.

- **Move**

(Only for NetApp Cluster-Mode) For high utilization of Storage Controller resources, Intersight Workload Optimizer can move an aggregate to another storage controller. The storage controllers must be running.

For high IOPS or latency, a move is always off of the current disk array. All the volumes on a given disk array show the same IOPS and Latency, so moving to a volume on the same array would not fix these issues.

- **Move VM**

For high utilization of Storage on a volume, Intersight Workload Optimizer can move a VM to another volume. The new volume can be on the current disk array, on some other disk array, or on any other datastore.

For high IOPS or latency, a move is always off of the current disk array. All the volumes on a given disk array show the same IOPS and Latency, so moving to a volume on the same array would not fix these issues.

- **Move Datastore**

To balance utilization of disk array resources, Intersight Workload Optimizer can move a datastore to another array.

Disk Array Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

The following table describes the default action acceptance modes for disk array actions and automation support for environments that have Disk Array Storage Controllers as targets.

| Action | Default Mode | NetApp ONTAP | VMAX | Pure Storage |
|-------------|--------------|--------------|------|--------------|
| Move | Disabled | | N/A | N/A |
| Provision | Recommend | | N/A | |
| Resize (up) | Recommend | | | |
| Start | Recommend | N/A | N/A | N/A |
| Suspend | Disabled | N/A | N/A | N/A |

Action Automation for NetApp Storage Systems

For NetApp storage systems, the actions Intersight Workload Optimizer can automatically perform depend on the NetApp version you are running, and whether the system is running in cluster mode:

| Automated Action | Cluster-Mode |
|---|--------------|
| Move VM between datastores, on the same disk array | Yes |
| Move VM between datastores on different disk arrays | Yes |
| | Yes |

| Automated Action | Cluster-Mode |
|---|----------------------|
| Move Datastore between disk arrays on the same storage controller | |
| Move Datastore between disk arrays on different storage controllers | Yes |
| Resize Storage | Yes |
| Resize Disk Array | No – Resize up, only |

In addition, for a system running in Cluster-Mode, Intersight Workload Optimizer can recommend moving an aggregate to another storage controller.

Utilization Constraints

Utilization constraints affect the actions Intersight Workload Optimizer recommends as it manages your environment. Intersight Workload Optimizer recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|----------------------------|---------------|
| Storage Amount Utilization | 90 |

Storage Settings

Set capacity for specific storage resources.

| Attribute | Default Value |
|--|---------------|
| SSD Disk IOPS Capacity | 50000 |
| 15k Disk IOPS Capacity | 1600 |
| VSeries LUN IOPS Capacity | 5000 |
| Storage Latency Capacity | 100 ms |
| 7.2k Disk IOPS Capacity | 800 |
| Storage Overprovisioned Percentage | 200 |
| IOPS Capacity A generic setting for disk array IOPS capacity (see Disk Array IOPS Capacity below). | 5000 |
| 10k Disk IOPS Capacity | 1200 |
| Disk Array IOPS Capacity | 10000 |

NOTE:

Intersight Workload Optimizer discovers storage latency and IOPS capacities that you set in your environment (for example VMAX) and uses them in its analysis. These capacities will be overridden by values that you set in Intersight Workload Optimizer policies.

■ IOPS Capacity

The capacity of IOPS (IO operations per second) that your storage devices can support. Intersight Workload Optimizer considers these settings when calculating utilization percentage. For example, assume IOPS Capacity of 5000 for a disk array. If utilization on the array is 2500 IOPS, then the disk array is at 50% of capacity for that metric.

Note that the IOPS setting for an array will determine IOPS calculations for all the storage on that array. If you made different IOPS settings for individual datastores hosted by the array, Intersight Workload Optimizer ignores the datastore settings and uses the disk array settings.

- Various Disk IOPS Capacity settings (**SSD Disk IOPS**, **7.2k Disk IOPS**, etc)

IOPS capacity settings for the different types of physical drives that are discovered on a disk array. If the storage controller exposes the types of disks in the array, Intersight Workload Optimizer uses multiples of these values to calculate the IOPS capacity of the disk array.

- **Disk Array IOPS Capacity**

Some disk arrays do not expose data for their individual disks – This is typical for flash arrays, or arrays that aggregate storage utilization across multiple tiers. Intersight Workload Optimizer uses this setting for the IOPS capacity of such disk arrays. Set it to the global scope to specify IOPS capacity for all disk arrays. To override this setting, set a disk array or group of disk arrays as the property scope, and then set the value you want for **IOPS Capacity**.

NOTE:

The user interface shows a disk array entity for any array that is discovered through a valid disk array or storage controller target. It also shows *placeholder* disk arrays for disk arrays that are not discovered through a configured target. For example, you might have disk arrays that Intersight Workload Optimizer does not natively support. Or you might have storage that is not hosted by any disk array. Such *placeholder* disk array entities appear with the string "DiskArray-" prefixed to their names. The user interface allows you to set IOPS Capacity to these placeholders, but those settings have no effect. To set IOPS Capacity for that storage, you must set it to the individual datastores.

■ **Storage Overprovisioned Percentage**

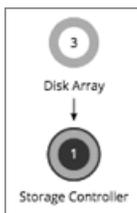
This setting indicates how much overprovisioning Intersight Workload Optimizer assumes when recommending actions for disk arrays. For example, if a disk array has a 30 TB capacity, and Storage Overprovisioned Percentage is set to 200, Intersight Workload Optimizer will treat the datastore as though it has a capacity of 60 TB, or 200% of the actual disk array capacity.

If the disk array is returning good Data Reduction Rates (DRR) in excess of the Storage Overprovisioned Percentage, actions to provision disk array may be generated. Consider increasing the percentage upwards closer to the real DRR. For example, if DRR is 12 to 1, the percentage could be closer to 1200.

Storage Controller

A Storage Controller is a device that manages one or more disk arrays. The storage controller provides CPU cycles to perform storage management tasks for each disk array it manages.

Synopsis



A storage controller gains its budget by selling resources to the disk arrays it manages. If utilization of the storage controller’s CPU resources is high enough, Intersight Workload Optimizer can recommend that you provision a new one and move disk arrays (aggregates) to it.

| Synopsis | |
|---------------------|---|
| Provides: | CPU resources to manage disk arrays. |
| Consumes: | NA |
| Discovered through: | Intersight Workload Optimizer directly accesses storage controller targets. |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- CPU
CPU is the measurement of CPU that is reserved or in use.
- Storage Amount
Storage Amount is the measurement of storage capacity that is in use.

NOTE:

In NetApp environments, the storage controller shows 100% utilization when there are no more disks in a `SPARE` state that the storage controller can utilize in an aggregate. This does not indicate that the storage controller has no capacity.

Actions

Intersight Workload Optimizer supports the following actions:

Provision

For high utilization of the storage controller's CPU, provision a new storage controller, and then move disk arrays to it.

Storage Controller Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

Actions for individual Disk Array Storage Controllers:

| Action | Default Mode | NetApp ONTAP | VMAX | Pure Storage |
|-----------|--------------|---|---|---|
| Provision | Disabled |  |  |  |

Utilization Constraints

Utilization constraints affect the actions Intersight Workload Optimizer recommends as it manages your environment. Intersight Workload Optimizer recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|----------------------------|---|
| CPU Utilization | 100 Maximum allowed utilization of Storage Controller CPU (from 20 to 100). |
| Storage Amount Utilization | 90 Maximum allowed utilization of storage that is managed by the Storage Controller. |

Storage Settings

Set capacity for specific storage resources.

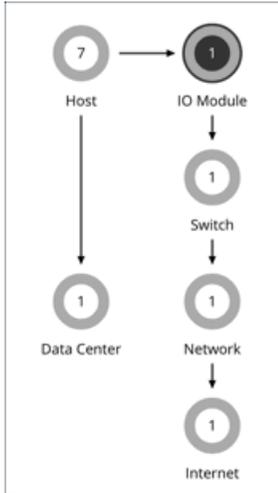
| Attribute | Default Value |
|--------------------------|---------------|
| IOPS Capacity | 5000 |
| Storage Latency Capacity | 100 ms |

IO Module

An IO Module connects the compute resources on a chassis to the fabric domain via the Fabric Interconnect. It provides the servers on the chassis with Net resources. Typical installations provide two IO Modules per chassis.

Intersight Workload Optimizer supports IO Modules when you have installed the Fabric Control Module license.

Synopsis



| Synopsis | |
|---------------------|---------------------------------|
| Provides: | Net resources |
| Consumes from: | Chassis and Fabric Interconnect |
| Discovered through: | Fabric Manager targets |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Net Throughput**
 Net Throughput is the rate of message delivery over a port.

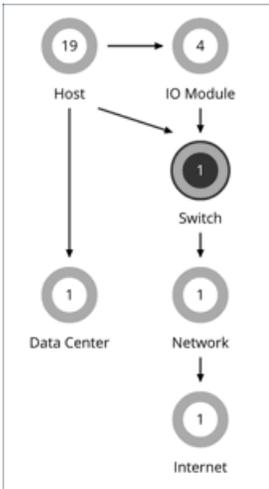
Actions

- Suspend**
 For low utilization of net resources, move VMs to another host that provides compatible network connectivity and suspend this IO Module.

Switch

A switch connects servers in a computing fabric to the fabric's network and storage resources. It provides network bandwidth to the servers in the platform.

Synopsis



| Synopsis | |
|---------------------|--------------------------------------|
| Provides: | Net resources |
| Consumes: | N/A |
| Discovered through: | Fabric Manager targets (such as UCS) |

Monitored Resources

Intersight Workload Optimizer monitors the following resources:

- Net Throughput
Net Throughput is the rate of message delivery over a port.
- PortChannel
PortChannel is the amalgamation of ports with a shared net throughput and utilization.

Actions

Intersight Workload Optimizer supports the following actions:

Resize

Resize PortChannel for a switch to increase bandwidth.

Switch Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

Automation Workflow

For environments that have Fabric Managers as targets:

| Action | Default Mode | Cisco UCS |
|--------|--------------|-------------------------------------|
| Resize | Recommend | <input type="button" value="Rcmd"/> |
| Start | Recommend | <input type="button" value="Rcmd"/> |

| Action | Default Mode | Cisco UCS |
|-----------|--------------|-----------|
| Provision | Recommend | |
| Suspend | Disabled | |
| Move | Disabled | |

Utilization Constraints

Utilization constraints affect the actions Intersight Workload Optimizer recommends as it manages your environment. Intersight Workload Optimizer recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Intersight Workload Optimizer will consider to be 100% of capacity.

| Attribute | Default Value |
|-----------------------|---------------|
| Switch Net Throughput | 70 |

Intersight Workload Optimizer Actions

After you deploy your targets, Intersight Workload Optimizer starts its analysis of your environment. This holistic analysis identifies problems and the actions you can take to *resolve* and *avoid* these problems. Intersight Workload Optimizer then generates a set of actions for that particular analysis and displays it in Action Center for you to investigate.

The screenshot shows the Action Center interface with the following details:

- Header:** Action Center (1112) VC7DC1/VC7DCC6
- Filters:** ALL (1,112), ON-PREM (78), CLOUD (967)
- Section:** RESIZE (58) actions. Sub-sections include VMEM Reclaim (46 GB) and VCPU Reclaim (20 vCPU). An EXECUTE ACTIONS button is present.
- Table:** A table listing actions for Virtual Machines (58). The table has columns for Virtual Machine Name, Risk, Resize Direction, Current Value, New Value, Resize Attribute, Action Category, and Action.

| Virtual Machine Name | Risk | Resize Direction | Current Value | New Value | Resize Attribute | Action Category | Action |
|------------------------------|----------------|------------------|---------------|-----------|------------------|-----------------|---------|
| System_Test | Virtual Memory | Upsize | 2 GB | 3 GB | Capacity | PERFORMANCE | DETAILS |
| qe-ocp412-n79zz-worker-v4bnp | Virtual CPU | Upsize | 4 vCPU | 8 vCPU | Capacity | PERFORMANCE | DETAILS |
| qe-ocp412-n79zz-worker-v4bnp | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |
| qe-ocp412-n79zz-worker-hzktq | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |
| ocp412.turbo.rtp | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |

To get the best results from Intersight Workload Optimizer, execute these actions promptly and consider automating as many of them as possible. You can execute these actions from the user interface or outside Intersight Workload Optimizer. To automate these actions, create an [automation policy \(on page 577\)](#) or change action acceptance to *Automatic* in the [default policies \(on page 575\)](#).

At first glance, individual actions might appear trivial and it is instinctively convenient to ignore them. It is important to keep in mind that executing a single action can impact other workloads in a meaningful way, helping move these other workloads closer to their desired state. However, if you find that a recommended action is not acceptable (for example, if it violates existing business rules), you can set up a policy with your preferred action.

In some cases, actions can introduce disruptions that you want to avoid at all costs. For example, during critical hours, Intersight Workload Optimizer might execute a resize action on a mission critical resource, which then requires that resource to restart. It is important to anticipate these disruptions and plan accordingly. For example, you can create a group for all critical resources and then schedule the execution of actions to off-peak hours or weekends.

Working With Action Center

When you start using Intersight Workload Optimizer, all the actions that the product generates appear as pending. Use Action Center to review and execute pending actions.

Action Center (1112) VC7DC1/VC7DCC6

ALL (1,112) ON-PREM (78) CLOUD (967)

RESIZE ^ **Resize Actions (58)** VMEM Reclaim 46 GB VCPU Reclaim 20 vCPU EXECUTE ACTIONS ⚙️ ⬇️ 🖨️

Virtual Machines (58) 🔍 Type to search ⚙️ ADD FILTER

| <input type="checkbox"/> | Virtual Machine Name | Risk | Resize Direction | Current Value | New Value | Resize Attribute | Action Category | Action |
|--------------------------|------------------------------|----------------|------------------|---------------|-----------|------------------|-----------------|---------|
| <input type="checkbox"/> | System_Test | Virtual Memory | Upsize | 2 GB | 3 GB | Capacity | PERFORMANCE | DETAILS |
| <input type="checkbox"/> | qe-ocp412-n79zz-worker-v4bkg | Virtual CPU | Upsize | 4 vCPU | 8 vCPU | Capacity | PERFORMANCE | DETAILS |
| <input type="checkbox"/> | qe-ocp412-n79zz-worker-v4bkg | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |
| <input type="checkbox"/> | qe-ocp412-n79zz-worker-hzktq | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |
| <input type="checkbox"/> | ocp412.turbo.rtp | Virtual Memory | Upsize | 16 GB | 18 GB | Capacity | PERFORMANCE | DETAILS |

Viewing Action Center

Action Center displays in areas of the product that reference actions, providing a consistent view and experience. Here are some ways you can view Action Center.

- To focus on actions for a specific entity type, click the entity type in the supply chain, and then click the **Actions** tab.

Application: Services (151)

OVERVIEW DETAILS POLICIES LIST OF SERVICES (151) **ACTIONS (69)**

RESIZE ^ **Resize Actions (46)**

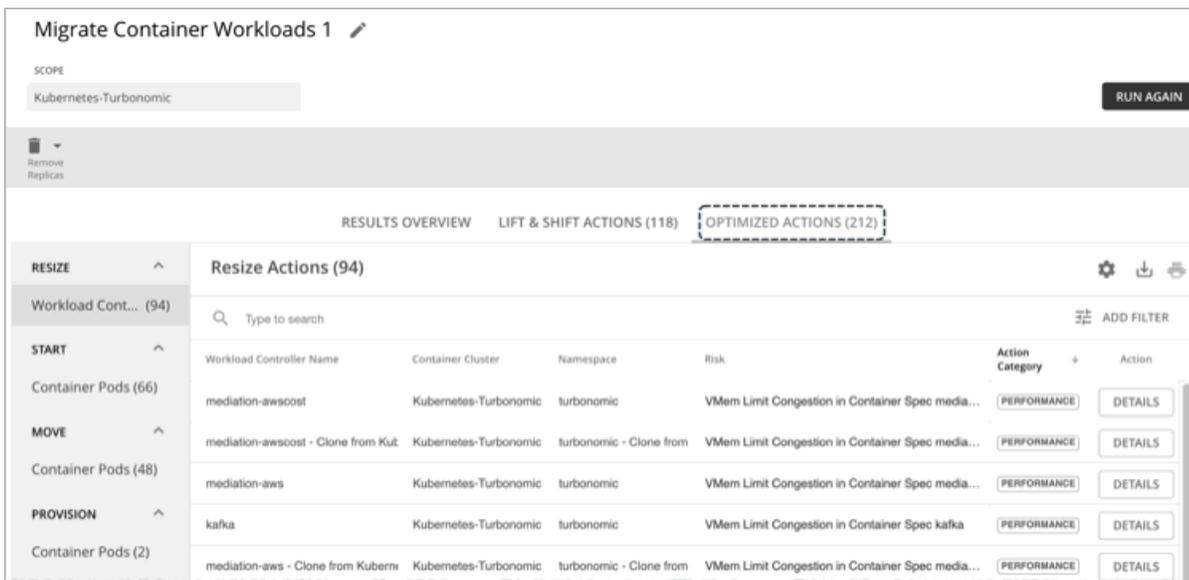
Workload Co... (46) 🔍 Type to search

| <input type="checkbox"/> | Workload Controller Name | Container Cluster | Namespace |
|--------------------------|--------------------------|-----------------------|------------|
| <input type="checkbox"/> | mediation-awscost | Kubernetes-Turbonomic | turbonomic |
| <input type="checkbox"/> | mediation-aws | Kubernetes-Turbonomic | turbonomic |
| <input type="checkbox"/> | kafka | Kubernetes-Turbonomic | turbonomic |
| <input type="checkbox"/> | mediation-azure | Kubernetes-Turbonomic | turbonomic |
| <input type="checkbox"/> | mediation-gcpcost | Kubernetes-Turbonomic | turbonomic |

- You can launch Action Center from action-focused charts, such as the Pending Actions, Top Utilized, and Potential Savings charts.

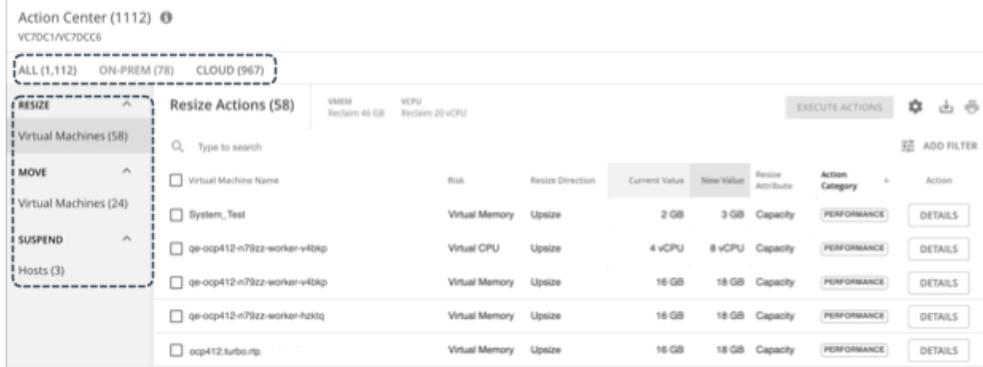
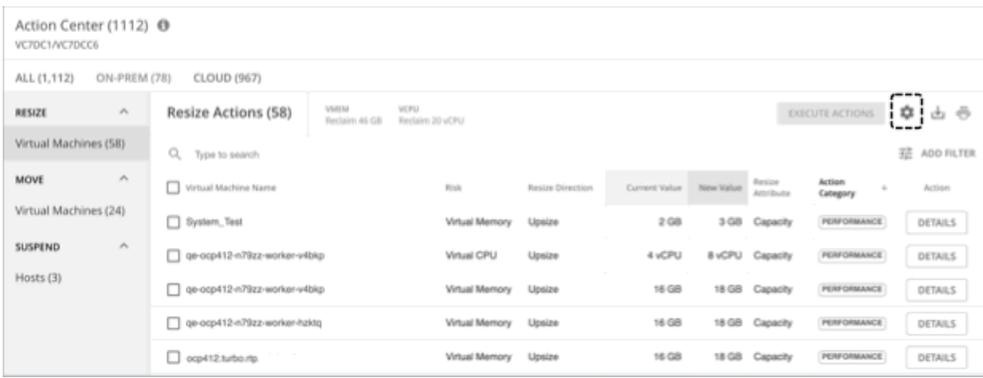


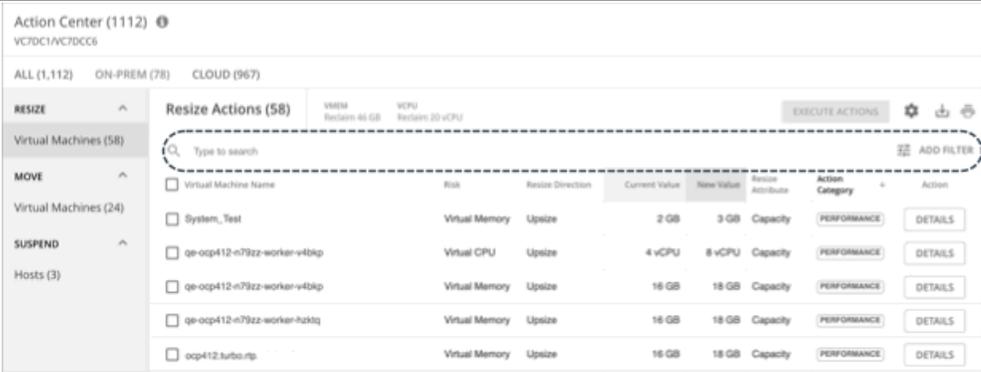
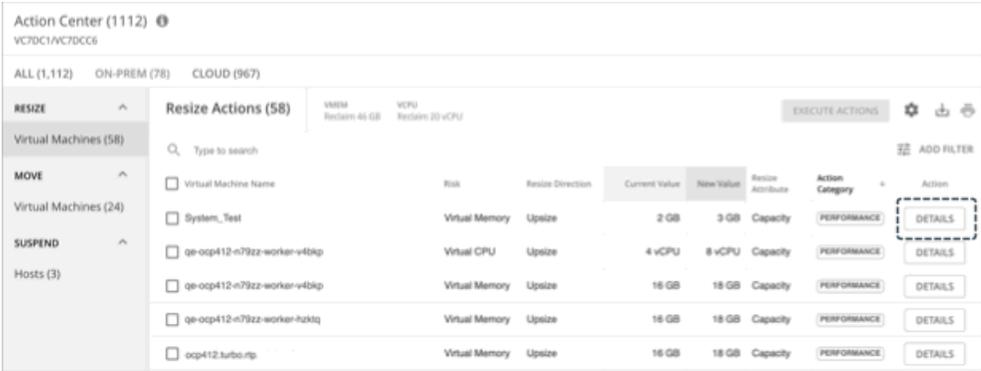
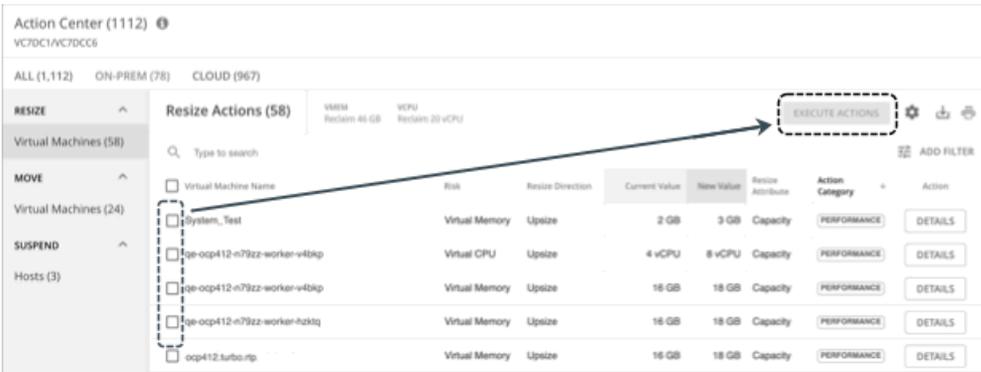
- After you run a [plan \(on page 418\)](#), click an Actions tab to open Action Center.

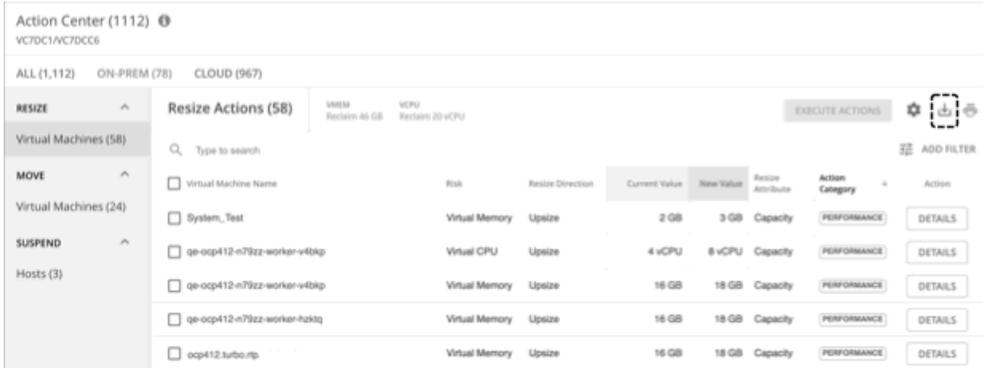
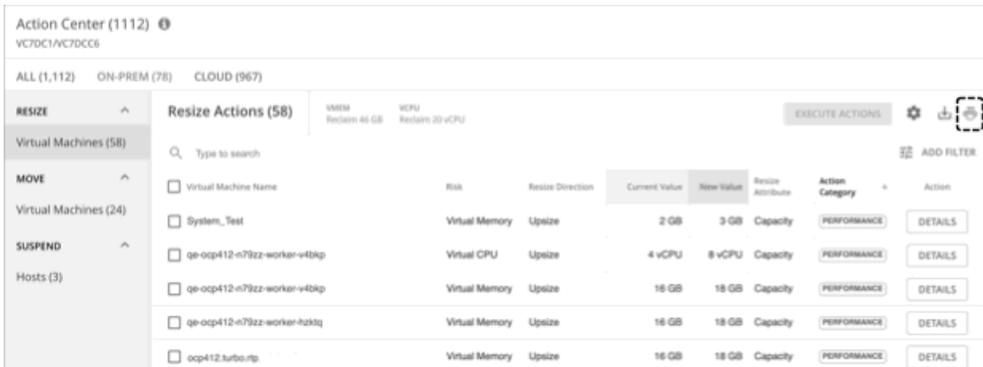


Action Center Features

Action Center provides a comprehensive view of actions. The following features are available to help you work with actions with ease.

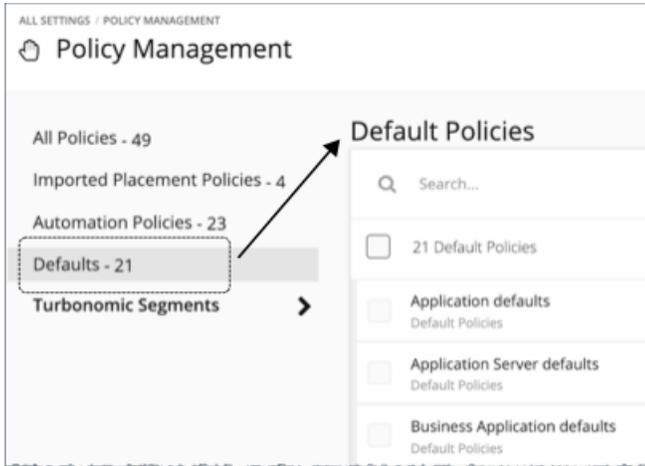
| Feature | Description |
|-------------------------------------|--|
| <p>Action groups</p> | <p>Action Center provides meaningful groups of actions.</p>  <ul style="list-style-type: none"> ■ Actions are grouped by environment type (on-prem or cloud). For a cloud environment, Action Center also groups actions by cloud provider. ■ Actions are grouped by action types (on page 413). These actions may apply to one or several entities. |
| <p>Table columns</p> | <p>Actions are presented in tabular format. Columns in the table provide information that is relevant to the action type and entity that you are currently viewing.</p>  |
| <p>Configure columns</p> | <p>The Configure columns button allows you to customize the display of columns. You can show or hide certain columns, or rearrange them as needed.</p>  |
| <p>Search and Add Filter button</p> | <p>For a long list of actions, use Search or filters to narrow the results. The filter options you see are those that are relevant to the action type and entity that you are currently viewing.</p> |

| Feature | Description |
|------------------|---|
| |  |
| Details button | <p>The Details button for each action opens the Action Details (on page 401) page, which provides additional information to help you understand why Intersight Workload Optimizer recommends an action and what you would gain if you execute it.</p>  |
| Action execution | <p>If an action is executable, you can execute it from Action Center. Select one or several actions that you want to execute, and then click Execute Actions.</p>  <ul style="list-style-type: none"> ■ Some actions may not be executable. For example, the action acceptance (on page 415) setting in a policy may have been set to <i>Recommend</i>, or the underlying technology for the entity does not support automation. <p>The check box for a non-executable action is grayed out. Review the action and then execute it outside Intersight Workload Optimizer.</p> ■ Some actions can only be executed after certain prerequisites are met. For example, in order to suspend Host A, VM_01 in the host must first move to Host B. However, Host B only has capacity for one VM and is currently hosting VM_02. In this case, Host A suspension is blocked by two prerequisite actions - VM_02 moving to another host and VM_01 moving to Host B. |

| Feature | Description |
|------------------------------------|---|
| | <p>The check box for an action with prerequisites is grayed out and includes the prohibition symbol (⊘). Click the Details button for the action to view prerequisite information. When all the prerequisites have been satisfied, the action becomes executable.</p> |
| <p>Download button</p> | <p>Download the actions you are currently viewing, or all the actions that are currently pending.</p>  |
| <p>Print Action Details button</p> | <p>Select one or several actions (maximum 75) that you want to print, and then click Print Action Details to open a printable preview of the selected Action Details (on page 401) pages. You can then print the action details for all of the selected actions at once.</p>  |

Action Handling

Intersight Workload Optimizer will never execute actions automatically, unless you tell it to. If you examine the default policies that ship with the product, you will notice that these policies do not enable automation on any action. Intersight Workload Optimizer gives you full control over all automation decisions.

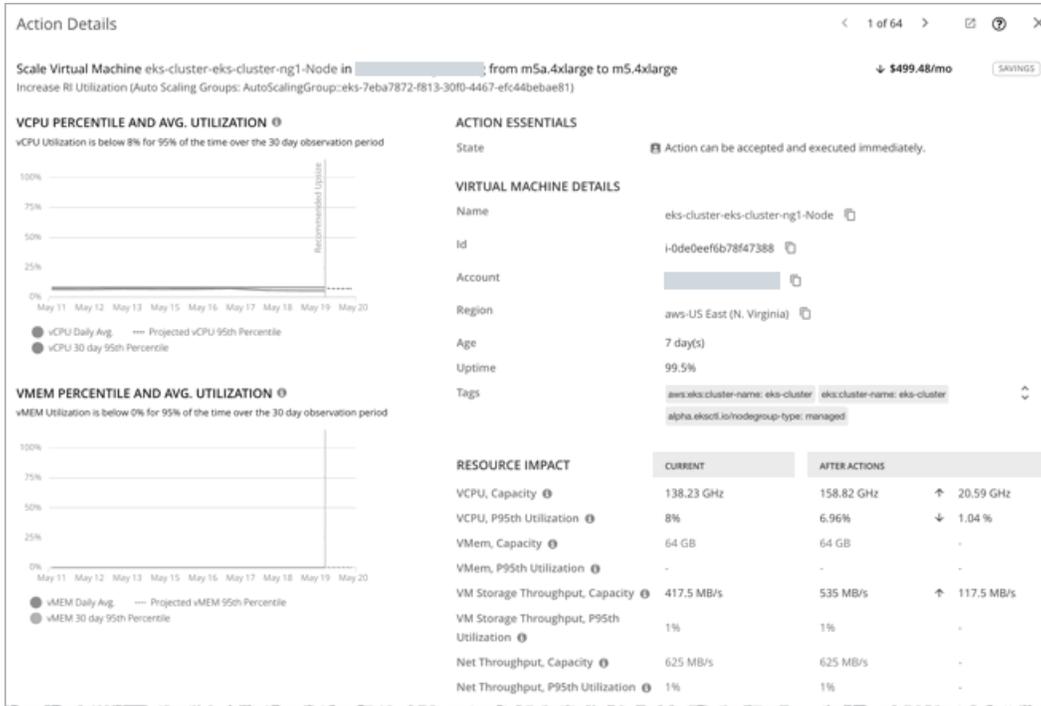


At first, your goal may be to evaluate all the actions in Action Center and then execute those that result in immediate improvements in performance and resource utilization. Over time, you develop and fine-tune your action-handling process to meet productivity goals and respond to changing business needs. This process could lead to the following key decisions:

- Disabling actions that should never execute, such as those that violate business rules
Intersight Workload Optimizer will not consider recommending disabled actions when it performs its analysis.
- Allowing certain actions to execute automatically, such as those that assure performance on mission-critical resources
Automation simplifies your task, while ensuring that workloads continue to have adequate resources to perform optimally. As such, it is important that you set the goal of automating as many actions as possible. This requires evaluating which actions are safe to automate, and on which entities.
- Continuing to let Intersight Workload Optimizer post certain actions so you can execute them on a case-by-case basis
For example, certain actions might require the approval of specific individuals. In this case, you would want Intersight Workload Optimizer to post those actions for review and only execute the actions that receive an approval.
These are the actions that you would look for in Action Center. These actions stop showing after you execute them, if you disable or automate them, or if the environment changes in the next market analysis such that the actions are no longer needed.

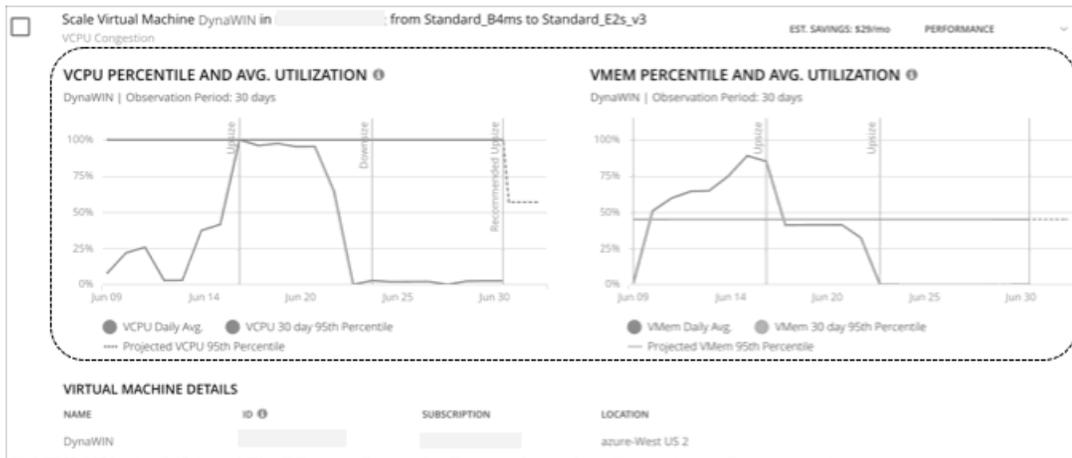
Action Details

The Action Details page provides a comprehensive set of information to help you understand why Intersight Workload Optimizer recommends an action and what you would gain if you execute it.



Utilization Charts

Intersight Workload Optimizer uses percentile calculations to measure resource utilization more accurately, and drive actions that improve overall utilization and reduce costs for cloud workloads. When you examine the details for an entity or pending action, you will see charts that highlight resource *utilization percentiles* for a given observation period, and the projected percentiles after you execute the action.



The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the entity, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Intersight Workload Optimizer's recommendations.

Notes:

- You can set constraints in policies to refine the percentile calculations.
- After you execute an action, it might take some time for the charts to reflect the resulting improvements.

Entities with Utilization Charts

Utilization charts display for actions on the following entity types:

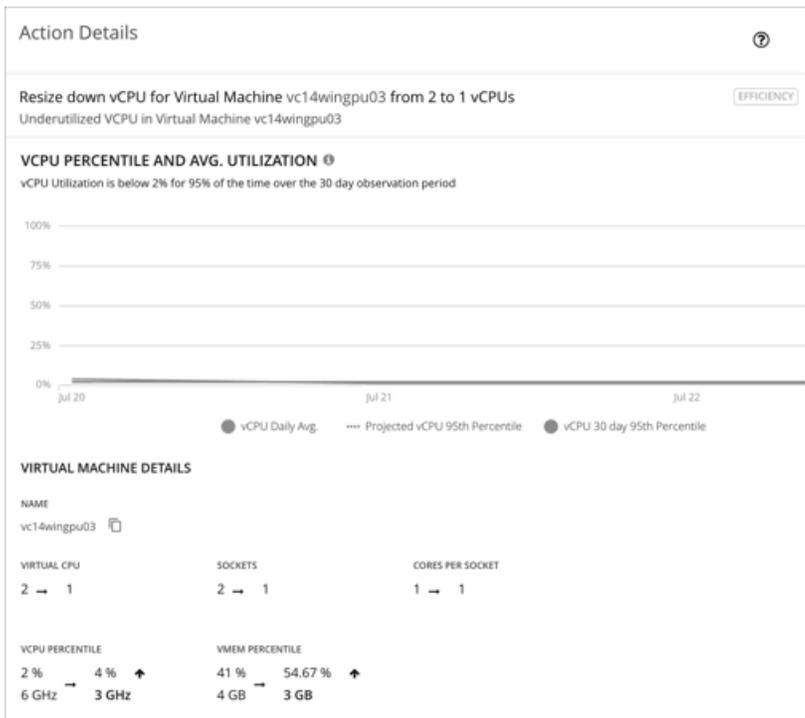
| Entity Type | Monitored Resources | | Notes |
|---------------------------|---|---|---|
| | Percentile Utilization | Average Utilization | |
| Application Component | <ul style="list-style-type: none"> ■ Heap ■ Garbage Collection | <ul style="list-style-type: none"> ■ Heap ■ Garbage Collection | <p>For Application Components, you will see either a Heap or Garbage Collection chart for Heap resize actions.</p> <p>See Application Component Actions (on page 216).</p> |
| Virtual Machine | <ul style="list-style-type: none"> ■ vCPU ■ vMem | <ul style="list-style-type: none"> ■ vCPU ■ vMem | <p>For on-prem VMs, you will see either a VCPU or VMem chart, depending on the commodity that needs to scale. For cloud VMs and VMs in Migrate to Cloud plans, both charts display.</p> <p>These charts also appear when you scope to a given VM (on-prem or cloud) and view the Details page. They also appear in Migrate to Cloud plan results.</p> |
| Virtual Machine Spec | <ul style="list-style-type: none"> ■ vCPU ■ vMem | <ul style="list-style-type: none"> ■ vCPU ■ vMem ■ Storage ■ Number of replicas | <p>See Virtual Machine Spec Actions (on page 288).</p> |
| Database (cloud) | <ul style="list-style-type: none"> ■ DTU Pricing Model <ul style="list-style-type: none"> – DTU ■ vCore Pricing Model <ul style="list-style-type: none"> – vCPU – vMem – IOPS – Throughput ■ RU Pricing Model <ul style="list-style-type: none"> – RU | <ul style="list-style-type: none"> ■ DTU Pricing Model <ul style="list-style-type: none"> – DTU – Storage ■ vCore Pricing Model <ul style="list-style-type: none"> – vCPU – vMem – IOPS – Throughput – Storage ■ RU Pricing Model <ul style="list-style-type: none"> – RU | <p>See Cloud Database (on page 308).</p> |
| Database Server (AWS RDS) | <ul style="list-style-type: none"> ■ vCPU ■ vMem ■ IOPS | <ul style="list-style-type: none"> ■ vCPU ■ vMem ■ IOPS | <p>See AWS RDS Actions (on page 297).</p> |
| Database Server (On-prem) | <ul style="list-style-type: none"> ■ DB Memory ■ DB Cache Hit Rate | <ul style="list-style-type: none"> ■ DB Memory ■ DB Cache Hit Rate | <p>For On-prem Database Servers, you will see either a DB Memory or DB Cache Hit Rate chart for DB Memory resize actions.</p> <p>See Database Server (On-prem) Actions (on page 360).</p> |

| Entity Type | Monitored Resources | | Notes |
|---------------------|--|---|--|
| | Percentile Utilization | Average Utilization | |
| Document Collection | RU | RU | See Document Collection Actions (on page 323) . |
| Volume (cloud) | <ul style="list-style-type: none"> IOPS Throughput | <ul style="list-style-type: none"> IOPS Throughput | These charts also appear when you scope to a given volume and view the Details page. See Cloud Volume Actions (on page 327) . |
| Workload Controller | <ul style="list-style-type: none"> vCPU limits and requests vMem limits and requests | <ul style="list-style-type: none"> vCPU limits, throttling, and requests vMem limits and requests | See Container Actions (on page 228) . |

Action Details - vCPU Resizes for On-prem VMs

The Action Details page provides a comprehensive set of information to help you understand why Intersight Workload Optimizer recommends a vCPU resize down action and the impact of this action on the VM and its host.

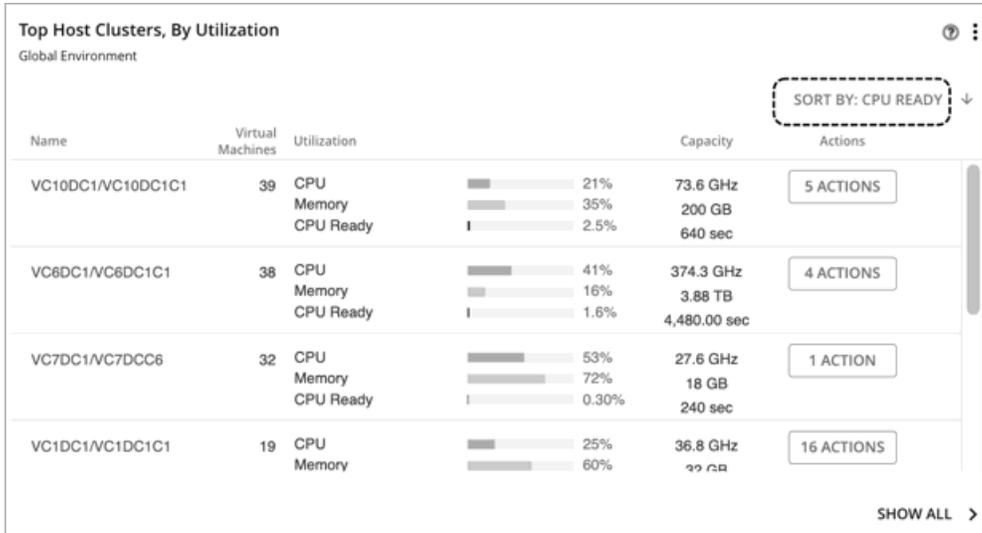
Actions to resize down vCPU can result in significant performance improvements in environments where CPU Ready on virtualization hosts is high.



Addressing CPU Ready Issues by Resizing Down vCPU

CPU Ready is a measurement of how long a VM is waiting for access to the host's physical CPU. In VMware environments, a CPU Ready value of 5% (or 1000 milliseconds) is considered a performance risk for general purpose VMs. As the VM and its application increase in priority/performance requirements, the acceptable CPU Ready value decreases.

To identify risks related to CPU Ready in your environment, add the Top Host Clusters by Utilization chart to your dashboard and then sort the chart data by CPU Ready.



As you examine the pending actions for these clusters, check for actions to resize down VM vCPU and then prioritize them for execution. These actions offer the greatest opportunity to resolve CPU Ready issues.

Actions by Entity Type

Intersight Workload Optimizer generates actions based on how entity types use or provide resources, and what each entity type supports.

The following tables show the actions that each entity type supports:

Application Entity Types

| Entity Type | Supported Actions |
|----------------------|--|
| Business Application | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a Business Application, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Application lists these actions, thus providing visibility into the risks that have a direct impact on the Business Application's performance.</p> |
| Business Transaction | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a Business Transaction, but it does recommend actions for the underlying Application Components and infrastructure. The Pending Actions chart for a Business Transaction lists these actions, thus providing visibility into the risks that have a direct impact on the Business Transaction's performance.</p> |
| Service | <p><i>For APM services:</i></p> <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for services discovered through APM targets, but it does recommend actions for the underlying Application Components and nodes. The Pending Actions chart for services list these actions, thus providing visibility into the risks that have a direct impact on their performance.</p> <p><i>For container platform services:</i></p> <p>None</p> |

| Entity Type | Supported Actions |
|-----------------------|--|
| | <p>Intersight Workload Optimizer does not recommend actions for services in container platform environments, but it does recommend actions for the replicas that back those services.</p> <p>For details, see Workload Controller Scale Actions (on page 238).</p> |
| Application Component | <p>Resize</p> <p>Resize the following resources to maintain performance:</p> <ul style="list-style-type: none"> ■ Thread Pool <p>Intersight Workload Optimizer generates thread pool resize actions. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.</p> ■ Connections <p>Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers.</p> ■ Heap <p>Intersight Workload Optimizer generates Heap resize actions if an Application Component provides Heap and Remaining GC Capacity, and the underlying VM or container provides VMem. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.</p> <p>NOTE: Remaining GC capacity is the measurement of Application Component uptime that is <i>not</i> spent on garbage collection (GC).</p> <p>The resources that Intersight Workload Optimizer can resize depend on the processes that it discovers from your Application Performance Management (APM) targets. Refer to the topic for a specific target to see a list of resources that can be resized.</p> |

Container Platform Entity Types

| Entity Type | Supported Actions |
|----------------|--|
| Service | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for services in container platform environments, but it does recommend actions for the replicas that back those services.</p> <p>For details, see Workload Controller Scale Actions (on page 238).</p> |
| Container | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions on containers.</p> |
| Container spec | <p>Resize (via workload controllers)</p> <p>A container spec retains the historical utilization data of ephemeral containers. Intersight Workload Optimizer uses this data to make resize decisions that assure optimal utilization of resources. By default, all replicas of the same container for the same workload type resize consistently.</p> <p>For details, see Workload Controller Resize Actions (on page 236).</p> |
| Namespace | <p>Resize Quota</p> <p>Intersight Workload Optimizer treats quotas defined in a namespace as constraints when making resize decisions. If existing actions would exceed the namespace quotas, Intersight Workload Optimizer recommends actions to resize up the affected namespace quota.</p> <p>Note that Intersight Workload Optimizer does not recommend actions to resize <i>down</i> a namespace quota. Such an action reduces the capacity that is already allocated to</p> |

| Entity Type | Supported Actions |
|------------------------------|---|
| | <p>an application. The decision to resize down a namespace quota should include the application owner.</p> <p>For details, see Namespace Actions (on page 248).</p> |
| Workload controller | <p>Resize or Scale</p> <p>Actions associated with a workload controller resize container specs vertically or scale replicas horizontally. This is a natural representation of these actions because the parent controller's container specs and number of replicas are modified. The workload controller then rolls out the changes in the running environment.</p> <p>For details, see Workload Controller Resize Actions (on page 236) and Workload Controller Scale Actions (on page 238).</p> |
| Container pod | <ul style="list-style-type: none"> ■ Move <p>Move a pod between nodes (VMs) to address performance issues or improve infrastructure efficiency. For example, if a particular node is congested for CPU, you can move pods to a node with sufficient capacity. If a node is underutilized and is a candidate for suspension, you must first move the pods before you can safely suspend the node.</p> ■ Provision/Suspend <p>When recommending node provision or suspend actions, Intersight Workload Optimizer will also recommend provisioning pods (based on demand from DaemonSets) or suspending the related pods.</p> <p>For details, see Container Pod Actions (on page 243).</p> |
| Container platform cluster | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a container platform cluster. Instead, it recommends actions for the containers, pods, nodes (VMs), and volumes in the cluster. Intersight Workload Optimizer shows all of these actions when you scope to a container platform cluster and view the Pending Actions chart.</p> |
| Container platform node (VM) | <p>A node (cloud or on-prem) is represented as a Virtual Machine entity in the supply chain.</p> <ul style="list-style-type: none"> ■ Provision <p>Provision nodes to address workload congestion or meet application demand.</p> ■ Suspend <p>Suspend nodes after you have consolidated pods or defragmented node resources to improve infrastructure efficiency.</p> ■ Reconfigure <p>Reconfigure nodes that are currently in the <code>NotReady</code> state.</p> <p>NOTE: For nodes in the public cloud, Intersight Workload Optimizer reports the cost savings or investments attached to these actions.</p> <p>For details, see Node Actions (on page 258).</p> |

Cloud Infrastructure Entity Types

| Entity Type | Supported Actions |
|-------------------------|---|
| Virtual Machine (Cloud) | <p>For details about cloud VM actions, see the following topics:</p> <ul style="list-style-type: none"> ■ Actions for AWS VMs (on page 265) ■ Actions for Azure VMs (on page 268) |

| Entity Type | Supported Actions |
|-------------------------|---|
| | <ul style="list-style-type: none"> ■ Actions for Google Cloud VMs (on page 269) |
| Virtual Machine Spec | <ul style="list-style-type: none"> ■ Scale Scale Azure App Service plans to optimize app performance or reduce costs, while complying with business policies. ■ Delete Delete empty Azure App Service plans as a cost-saving measure. A plan is considered empty if it is not hosting any running apps. For details, see Virtual Machine Spec Actions (on page 288). |
| App Component Spec | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for App Component Specs, but it does recommend actions for the underlying Virtual Machine Specs. For details, see Virtual Machine Spec Actions (on page 288).</p> |
| Database Server (Cloud) | <ul style="list-style-type: none"> ■ For AWS RDS: <ul style="list-style-type: none"> – Scale Scale compute and storage resources to optimize performance and costs. For details, see Cloud Database Server Actions (on page 297). ■ For Azure Cosmos DB Accounts: <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a Cosmos DB account but it does recommend actions for the databases (on page 317) and document collections (on page 322) in the account.</p> |
| Database (Cloud) | <ul style="list-style-type: none"> ■ Scale SQL Database <ul style="list-style-type: none"> – DTU Pricing Model Scale DTU and storage resources to optimize performance and costs. – vCore Pricing Model Scale vCPU, vMem, IOPS, throughput and storage resources to optimize performance and costs. For details, see Scale Actions for SQL Databases (on page 310). ■ Scale Cosmos DB Database Scale Request Units (RUs) to optimize performance and costs. For details, see Scale Actions for Cosmos DB Databases (on page 317). ■ Reconfigure Cosmos DB Database Remove unused provisioned throughput to reduce costs. For details, see Reconfigure Actions for Cosmos DB Databases (on page 318). ■ Delete Cosmos DB Database Delete a database with provisioned throughput but without any underlying document collection (container) to reduce costs. For details, see Delete Actions for Cosmos DB Databases (on page 318). ■ Suspend/Stop Dedicated SQL Pool Suspend or stop a dedicated SQL pool (used in Azure Synapse Analytics) to reduce compute costs. <ul style="list-style-type: none"> – Intersight Workload Optimizer analysis generates suspend actions for <i>idle</i> pools. |

| Entity Type | Supported Actions |
|---------------------|--|
| | <p>NOTE: Currently, Intersight Workload Optimizer analysis does not generate actions to start a suspended pool. You can start a suspended pool from Azure.</p> <p>For details, see Suspend Actions for Dedicated SQL Pools (on page 316).</p> |
| Document Collection | <p>Scale Scale Request Units (RUs) to optimize performance and costs. For details, see Document Collection Actions (on page 323).</p> |
| Volume (Cloud) | <ul style="list-style-type: none"> ■ Scale Scale attached volumes to optimize performance and costs. ■ Delete Delete unattached volumes as a cost-saving measure. Intersight Workload Optimizer generates an action immediately after discovering an unattached volume. For details, see Cloud Volume Actions (on page 327). |
| Zone | <p>None Intersight Workload Optimizer does not recommend actions for a cloud zone.</p> |
| Region | <p>None Intersight Workload Optimizer does not recommend actions for a cloud region.</p> |

On-prem Infrastructure Entity Types

| Entity Type | Supported Actions |
|---------------------------|---|
| Virtual Machine (On-prem) | <ul style="list-style-type: none"> ■ Resize Resize resource capacity, reservation, or limit to improve performance. ■ Resize PU / Resize VP (IBM PowerVM) Resize PU (processing units) or VP (virtual processor) to optimize LPAR processing unit allocation and virtual processor capacity based on historical demand collected from the HMC. ■ Move Move a VM due to high resource utilization on VM or host, excess IOPS or latency in VStorage, workload placement violation, underutilized host (move VM before suspending host). ■ Reconfigure Change a VM's configuration to comply with a policy. For hypervisor targets, Intersight Workload Optimizer can reconfigure VMs that violate vCPU scaling policies. For details, see vCPU Scaling Controls (on page 348). <p>For details, see On-prem VM Actions (on page 338).</p> |
| Volume (On-prem) | <ul style="list-style-type: none"> ■ Move Move a VM's volume (virtual storage) due to excess utilization of the current datastore, or for more efficient utilization of datastores in the environment. Points to consider: <ul style="list-style-type: none"> – The default global policy includes a setting that directs Intersight Workload Optimizer to use relevant metrics when analyzing and recommending actions |

| Entity Type | Supported Actions |
|---------------------------|---|
| | <p>for volumes. For details, see Enable Analysis of On-prem Volumes (on page 576).</p> <ul style="list-style-type: none"> – Intersight Workload Optimizer will not recommend moving a volume to a datastore that is currently in maintenance mode. Any volume in that datastore should move to an active datastore (for example, via vMotion). <ul style="list-style-type: none"> ■ Reconfigure Reconfigure a VM's volume (virtual storage) to comply with placement policies. |
| Database Server (On-prem) | <p>Resize</p> <p>Resize the following resources:</p> <ul style="list-style-type: none"> ■ Connections Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers. ■ Database memory (DBMem) Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary. A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates. When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates. ■ Transaction log Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology. Currently, Intersight Workload Optimizer does not support resize actions for Oracle and Database Servers on the Hyper-V platform (due to the lack of API support for virtual storage). |
| Virtual Data Center | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a Virtual Data Center. Instead, it recommends actions for the entities that provide resources to the Virtual Data Center.</p> |
| Host | <ul style="list-style-type: none"> ■ Start Start a suspended host when there is increased demand for physical resources. ■ Provision Provision a new host in the environment when there is increased demand for physical resources. Intersight Workload Optimizer can then move workloads to that host. ■ Suspend When physical resources are underutilized on a host, move existing workloads to other hosts and then suspend the host. <p>For details, see Host Actions (on page 368).</p> |
| Chassis | <p>None</p> <p>Intersight Workload Optimizer does not recommend actions for a chassis.</p> |
| Data Center | <p>None</p> |

| Entity Type | Supported Actions |
|--------------|--|
| | <p>Intersight Workload Optimizer does not recommend actions for a data center. Instead, it recommends actions for the entities running in the data center.</p> |
| Storage | <ul style="list-style-type: none"> ■ Move For high utilization of physical storage, move datastore to a different disk array (aggregate). ■ Provision For high utilization of storage resources, provision a new datastore. ■ Resize Increase or decrease the datastore capacity. ■ Start For high utilization of storage resources, start a suspended datastore. ■ Suspend For low utilization of storage resources, move served VMs to other datastores and suspend this one. ■ Delete (datastore or volume) Delete a datastore or volume that has been suspended for a period of time. ■ Delete (unattached files) Delete a file on a datastore that has not been accessed or modified for a period of time. <p>For details, see Storage Actions (on page 378).</p> |
| Logical Pool | <ul style="list-style-type: none"> ■ Move For high utilization of physical storage, move the logical pool to a different disk array (aggregate). ■ Provision For high utilization of storage resources, provision a new logical pool. ■ Resize Increase or decrease the logical pool capacity. ■ Start For high utilization of storage resources, start a suspended logical pool. ■ Suspend For low utilization of storage resources, move served VMs to other logical pools and suspend this one. |
| Disk Array | <ul style="list-style-type: none"> ■ Provision For high utilization of the disk array's storage, provision a new disk array. This action can only be executed outside Intersight Workload Optimizer. ■ Start For high utilization of disk array, start a suspended disk array. This action can only be executed outside Intersight Workload Optimizer. ■ Suspend For low utilization of the disk array's storage, move VMs to other datastores and suspend volumes on the disk array. This action can only be executed outside Intersight Workload Optimizer. ■ Move (Only for NetApp Cluster-Mode) For high utilization of Storage Controller resources, Intersight Workload Optimizer can move an aggregate to another storage controller. The storage controllers must be running. |

| Entity Type | Supported Actions |
|--------------------|---|
| | <p>For high IOPS or latency, a move is always off of the current disk array. All the volumes on a given disk array show the same IOPS and Latency, so moving to a volume on the same array would not fix these issues.</p> <ul style="list-style-type: none"> ■ Move VM <p>For high utilization of Storage on a volume, Intersight Workload Optimizer can move a VM to another volume. The new volume can be on the current disk array, on some other disk array, or on any other datastore.</p> <p>For high IOPS or latency, a move is always off of the current disk array. All the volumes on a given disk array show the same IOPS and Latency, so moving to a volume on the same array would not fix these issues.</p> <ul style="list-style-type: none"> ■ Move Datastore <p>To balance utilization of disk array resources, Intersight Workload Optimizer can move a datastore to another array.</p> |
| Storage Controller | <p>Provision</p> <p>For high utilization of the storage controller's CPU, provision a new storage controller, and then move disk arrays to it.</p> |
| IO Module | <p>Suspend</p> <p>For low utilization of net resources, move VMs to another host that provides compatible network connectivity and suspend this IO Module.</p> |
| Switch | <p>Resize</p> <p>Resize PortChannel for a switch to increase bandwidth.</p> |

Action Categories

Intersight Workload Optimizer groups entries in the Actions List by different categories. These categories do not strictly define the severity of an issue, but they indicate the nature of the issue.

Compliance

A virtual environment can include policies that limit availability of resources. It's possible that the environment configuration violates these defined policies. In such cases, Intersight Workload Optimizer identifies the violation and recommends actions that bring the entity back into compliance.

Efficiency

Efficient utilization of resources is an important part of running in the desired state. Running efficiently maximizes your investment and reduces cost. When Intersight Workload Optimizer discovers underutilized resources, it recommends actions to consolidate your operations. For example, it can recommend that you move certain VMs onto a different host, and then suspend the original host.

Performance

Ultimately, the reason to manage workloads in your environment is to assure performance and meet productivity goals. Intersight Workload Optimizer can recommends actions when it detects conditions that directly put performance at risk. You can consider these critical conditions, and you should execute the recommended actions as soon as possible.

Prevention

Intersight Workload Optimizer constantly monitors conditions, and works to keep your environment running in a desired state. As it finds issues that risk moving the environment out of this state, it recommends preventive actions. You should attend to these issues, and perform the associated actions. If you do not, the environment may drift away from the desired state, and performance may be put at risk.

Action Types

Intersight Workload Optimizer performs the following general types of actions:

- Buy – Purchase cloud instances at a [discounted \(on page 25\)](#) rate to reduce on-demand costs
- Delete – Remove unused entities (for example, datastores on disk arrays or unattached volumes)
- Move – Move an entity from one provider to another to address performance issues or improve infrastructure efficiency
- Optimization – Increase discount coverage for cloud workloads to reduce on-demand costs
- Provision – Add resource providers to the environment to increase capacity
- Reconfigure – Correct a misconfiguration to bring an entity into compliance
- Scale/Resize – Change the resources allocated to workloads to meet demand
- Start – Start a stopped/suspended entity
- Stop/Suspend – Pause an entity for a period of time to increase efficient use of resources

For information on the actions that Intersight Workload Optimizer generates for specific entities, see [Actions by Entity Type \(on page 405\)](#).

Placement

Placement actions determine the best provider for a consumer. These include initial placement for a new entity, and move actions that change a consumer to use a different provider. For example, moving a VM assigns it to a different host. Moving a VM's storage means the VM will use a different datastore.

Placement Constraints

When making placement decisions, Intersight Workload Optimizer checks for placement constraints to limit the set of providers for a given consumer. It respects automatic placement constraints, including cluster boundaries and DRS rules. It also considers user-configured constraints defined in a placement policy to ensure compliance to specific business requirements.

Reviewing the constraints on an entity helps you understand the actions that Intersight Workload Optimizer recommends. If an action seems questionable to you, then you should look at the constraints on the affected entities. It's possible that some policy or constraint is in effect, and it keeps Intersight Workload Optimizer from recommending a more obvious action. For details, see [Entity Placement Constraints \(on page 37\)](#).

You can run plans to see what happens if you turn off constraints, or disable or enable certain placement policies.

Effective CPU Capacity

CPU processor speed is not necessarily an effective indicator of CPU capacity. For example, processor architecture can make a slower CPU have a greater effective capacity. Newer models of machines can often have fewer cores or less clock speed, but still have a higher effective capacity.

When placing VMs on hosts in the on-prem environment, Intersight Workload Optimizer discovers the effective CPU capacity of your hosts. This increases the accuracy of placement calculations so that newer, more efficient hosts will show a greater effective capacity than less efficient hosts that might have larger or faster processors.

To discover the effective capacity, Intersight Workload Optimizer uses benchmark data from spec.org. This benchmark data maps to effective capacity settings that Intersight Workload Optimizer uses to make placement calculations.

You can see a catalog of these benchmark data and choose from listed processors when you edit Host templates. For more information, see [Selecting CPUs from the Catalog \(on page 590\)](#).

Shared-Nothing Migration Actions

If you have enabled both storage and VM moves, Intersight Workload Optimizer can perform shared-nothing migrations, which move the VM and the stored VM files simultaneously. For details, see [Shared-Nothing Migration \(on page 343\)](#).

Cross-vCenter vMotion

VMware vSphere 6.0 introduces functionality that enables migration of virtual machines between different vCenter Server instances. Intersight Workload Optimizer supports this capability through *Merge* placement policies (see [Creating Placement](#)

[Policies \(on page 569\)](#)). It considers cross-vCenter locations when calculating placement, and can recommend or execute moves to different vCenter servers.

Moves on the Public Cloud

On the public cloud you do not place workloads on physical hosts. In the Intersight Workload Optimizer Supply Chain, the Host nodes represent availability zones. Intersight Workload Optimizer can recommend moving a workload to a different zone, if such a move can reduce your cloud cost. These moves recognize constraints, such as availability of instances types and [discounts \(on page 25\)](#) in the given zones.

In AWS environments, a VM can use Elastic Block Stores (EBS) or Instance Storage. If the VM's root storage is EBS, then Intersight Workload Optimizer can recommend a VM move. However, because Instance Storage is ephemeral and a move would lose the stored data, Intersight Workload Optimizer does not recommend moving a VM that has Instance Storage as its root storage.

If a VM is running within a billing family, then Intersight Workload Optimizer only recommends moving that VM to other regions within that billing family.

In AWS environments that use RIs, Intersight Workload Optimizer recognizes Availability Zones that you have specified for your RI purchases. For move and resize actions, Intersight Workload Optimizer gives precedence to these RIs in the given zone. All else being equal for a given zone, if you have Zone RIs with reserved capacity and RIs that do not reserve capacity, Intersight Workload Optimizer will use the Zone RI first.

Scaling

Scaling actions update capacity in your environment. For vertical scaling, Intersight Workload Optimizer increases or decreases the capacity of resources on existing entities. For horizontal scaling it provisions new providers. For example, provisioning a host adds more compute capacity that is available to run VMs. Provisioning a VM adds capacity to run applications.

Intersight Workload Optimizer can provision the following:

- Containers
- VMs
- Hosts
- Storage
- Storage Controllers (only for planning scenarios)
- Disk Arrays

Under certain circumstances, Intersight Workload Optimizer can also recommend that you provision a virtual datacenter.

Discount Optimization

To reduce your cloud costs, Intersight Workload Optimizer can recommend scaling VMs to instance types that are charged discounted rates.

NOTE:

Under most circumstances, when a cloud provider offers a new instance type that is meant to replace an older type, the provider offers it at a lower cost. However, a provider may provide a new instance type with identical costs as the older instance types. If this occurs, and capacity and cost are equal, Intersight Workload Optimizer cannot ensure that it chooses the newer instance type. To work around this issue, you can create an Action Automation policy that excludes the older instance type.

- [Discount Utilization \(on page 562\)](#)

This chart shows how well you have utilized your current discount [inventory \(on page 560\)](#). The desired goal is to maximize the utilization of your inventory and thus take full advantage of the discounted pricing offered by your cloud provider.

- [Discount Coverage \(on page 557\)](#)

This chart shows the percentage of cloud workloads (VMs and RDS database servers) covered by discounts. For VMs covered by discounts, you can reduce your costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

- [Discount Inventory \(on page 560\)](#)

This chart lists the cloud provider discounts discovered in your environment.

Discount optimization actions are not executed by Intersight Workload Optimizer users. They reflect capacity reassignments performed by your cloud provider.

Configuration

These are reconfigure and resize actions. Reconfigure actions can add necessary network access, or reconfigure storage. Resize actions allocate more or less resource capacity on an entity, which can include adding or reducing VCPUs or VMem on a VM, adding or reducing capacity on a datastore, and adding or reducing volumes in a disk array.

Intersight Workload Optimizer can reconfigure the following:

- VMs
- Containers
- Storage
- Disk Arrays
- Virtual Datacenters

Start/Buy

Intersight Workload Optimizer can recommend that you start a suspended entity to add capacity to the environment. It can also recommend purchasing cloud provider [discounts \(on page 25\)](#) to reduce costs for your current workload.

Stop

Stop actions suspend entities without removing them from the environment. Suspended capacity is still available to be brought back online, but is currently not available for use. Suspended resources are candidates for termination.

Intersight Workload Optimizer can suspend the following:

- Application Components
- Container Pods
- Disk Arrays
- Hosts
- Storage (on-prem)
- Virtual datacenter

Delete

Intersight Workload Optimizer can recommend deleting unnecessary resources to improve infrastructure efficiency. In public cloud environments, delete actions also introduce cost savings. For example, Intersight Workload Optimizer might recommend that you delete wasted files to free up storage space, or delete unattached storage in your cloud environment.

In on-prem environments, you can sort Delete file actions based on file size or filter based on the file directory or the number of days a file has been unattached to understand the benefits of deleting unattached files. You are able to configure automation policies to automatically delete unattached files after a defined period of non-use, and can manually execute delete actions if automation is not desired.

Action Acceptance Modes

Action acceptance modes specify the degree of automation for the generated actions. For example, in some environments you might not want to automate resize down of VMs because that is a disruptive action. You would use action acceptance modes in a policy to set that business rule.

Intersight Workload Optimizer supports the following action acceptance modes:

- **Recommend Only** – Recommend the action so that a user can execute it outside Intersight Workload Optimizer
- **Manual** – Recommend the action, and provide the option to execute that action through the Intersight Workload Optimizer user interface
- **Automated** – Execute the action automatically.

For automated resize or move actions on the same entity, Intersight Workload Optimizer waits five minutes between each action to avoid failures associated with trying to execute all actions at once. Any action awaiting execution stays in queue. For example, if a VM has both vCPU and vMem resize actions, Intersight Workload Optimizer could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.

- **Automated when approved** – Execute the action automatically only after they are externally approved. This option is available only after adding a Service Now target.

The Pending Actions charts only count actions in *Recommend* or *Manual* mode.

Automated actions appear in the following charts:

- **All Actions** chart on the **Overview** and the On-prem Executive Dashboard
- **Accepted Actions** chart on the **Overview**

Setting Action Acceptance Modes

There are two ways to configure action acceptance modes:

- Change the action acceptance mode in a [default policy \(on page 575\)](#).
- Create an [automation policy \(on page 577\)](#), scope the policy to specific entities or groups, and then select the action acceptance mode for each action.

Intersight Workload Optimizer allows you to create dynamic groups to ensure that entities discovered in the future automatically add to a group and apply the policy of that group. If a conflict arises as a result of an entity belonging to several groups, the entity applies the policy with the most conservative action.

Action Acceptance Overrides

Under some conditions, Intersight Workload Optimizer changes the action acceptance mode of an action from *Manual* to *Recommend*.

Intersight Workload Optimizer makes this change as a safeguard against executing actions that the underlying infrastructure cannot support. For example, assume you have VM move actions set to *Manual*. Then assume Intersight Workload Optimizer analysis wants to move a VM onto a host that is already utilized fully. In this case, there would be other actions to move workloads *off* of the given host to make room for this new VM. However, because moves are *Manual*, the host might not be properly cleared off yet. In that case, Intersight Workload Optimizer changes actions to move workloads *to* the host from *Manual* to *Recommend*.

For cloud environments, some instances require workloads to be configured in specific ways before they can move to those instance types. If Intersight Workload Optimizer recommends moving a workload that is not suitably configured onto one of these instances, then it changes the action acceptance mode from *Manual* to *Recommend*, and then describes the reason.

Plans: Looking to the Future

CONFIGURATION

- SCOPE: DC14\DC14-Cluster
- Buttons: Add, Replace, Remove, Actions, Ignore Constraints, Placement Policies, Utilization, Baseline, Desired State, Projection
- Buttons: RUN AGAIN

RESULTS OVERVIEW PLAN ACTIONS (170)

Plan has 38 unplaced workloads

Plan Summary

| | Current | After Plan | Difference | % |
|------------------|---------|------------|------------|-----------|
| Virtual Machines | 30 | 30 | 0 | 0 % |
| Hosts | 3 | 15 | 12 | ▲400 % |
| Storage | 6 | 14 | 8 | ▲133.3 % |
| CPU | 6 Cores | 32 Cores | 26 | ▲433.3 % |
| Memory | 12 GB | 246.5 GB | 234.5 GB | ▲1958.3 % |
| Storage Amount | 3.6 GB | 6.3 GB | 2.7 GB | ▲50 % |
| Host Density | 10:1 | 2:1 | 8:1 | ▼80 % |
| Storage Density | 5:1 | 2:1 | 3:1 | ▼60 % |

Show all ▶

Use the Plan Page to run simulations for what-if scenarios that explore possibilities such as:

- Reducing cost while assuring performance for your workloads
- Impact of scaling resources
- Changing hardware supply
- Projected infrastructure requirements
- Optimal workload distribution to meet historical peaks demands
- Optimal workload distribution across existing resources

How Plans Work

To run a plan scenario, Intersight Workload Optimizer creates a snapshot copy of your real-time market and modifies that snapshot according to the scenario. It then uses the Economic Scheduling Engine to perform analysis on that plan market. A scenario can modify the snapshot market by changing the workload, adding or removing hardware resources, or eliminating constraints such as cluster boundaries or placement policies.

As it runs a plan, Intersight Workload Optimizer continuously analyzes the plan market until it arrives at the optimal conditions that market can achieve. When it reaches that point, the Economic Scheduling Engine cannot find better prices for any of the resources demanded by the workload – the plan stops running, and it displays the results as the plan's desired state. The display includes the resulting workload distribution across hosts and datastores, as well as a list of actions the plan executed to achieve the desired result.

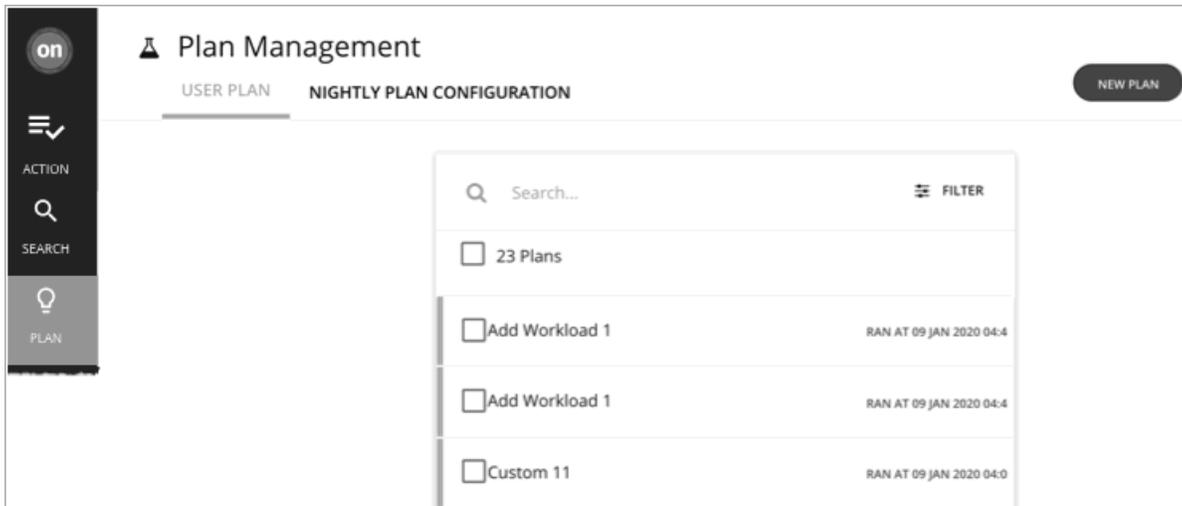
For example, assume a scenario that adds virtual machines to a cluster. To run the plan, Intersight Workload Optimizer takes a snapshot of the current market, and adds the VMs to the specified cluster. Intersight Workload Optimizer then runs analysis on the plan market, where each entity in the supply chain shops for the resources it needs, always looking for a better price – looking for those resources from less-utilized suppliers. This analysis continues until all the resources are provided at the best possible price.

The results might show that you can add more workload to your environment, even if you reduce compute resources by suspending physical machines. The recommended actions would then indicate which hosts you can take offline, and how to distribute your virtual machines among the remaining hosts.

Idle Workloads

Plans calculate optimal placement and optimal resource allocation for the given workload. However, plans do not include *idle* workloads. This is because an idle VM shows no utilization, so the plan cannot determine optimal placement or what percentage of allocated resources that workload will require when it restarts.

Plan Management



The Plan Management Page is your starting point for creating new plans, viewing saved plans, and deleting saved plans that you don't need anymore. To display this page, click **Plan** in the Intersight Workload Optimizer navigation bar.

- Create new plans

To create a new plan, click the **NEW PLAN** button. See [Setting Up Plan Scenarios \(on page 418\)](#).

- View saved plans

After you create and run a plan, Intersight Workload Optimizer saves it and then shows it in the Plan Management Page. You can open the saved plan to review the results, or you can change its configuration and run it again.

NOTE:

You can also view saved plans from the Search page, under the **Plans** category.

- Delete saved plans

To delete a saved plan, turn on the plan's checkbox and then click the **Delete** button.

- Configure nightly plans

Intersight Workload Optimizer runs nightly plans to calculate headroom for the clusters in your on-prem environment. For each cluster plan, you can set which VM template to use in these calculations. See [Configuring Nightly Plans \(on page 492\)](#).

Setting Up Plan Scenarios

A plan scenario specifies the overall configuration of a plan. Creating the plan scenario is how you set up a what-if scenario to see the results you would get if you changed your environment in some way.

This topic walks you through the general process of setting up a plan scenario.

1. Plan Entry Points

You can begin creating a plan scenario from different places in the user interface:

- From the Plan Page

Navigate to the Plan Page and click **NEW PLAN**. This plan has no scope. You will specify the scope after selecting the plan type.

- From the **Overview**

To start a plan scenario from the **Overview**, you must first go to the **Search** page to set the scope.

Set the scope to *a specific* Account, Billing Family, VM Group, or Region to start an Optimize Cloud plan.

- Cloud scope

If you set the scope to *a specific* Account, Billing Family, VM Group, or Region, you can start an Optimize Cloud or Buy VM Reservations plan.

- On-prem scope

If you set the scope to *a specific* Host Cluster, Datacenter, Group, Storage Cluster, or Virtual Datacenter, you can start any plan. You may need to go through additional steps, depending on your chosen plan type. For example, if you scope to a host cluster and choose the Add Virtual Machines plan type, the plan wizard prompts you to select the most suitable templates for the VMs you plan to add to the cluster.

- Container platform cluster scope

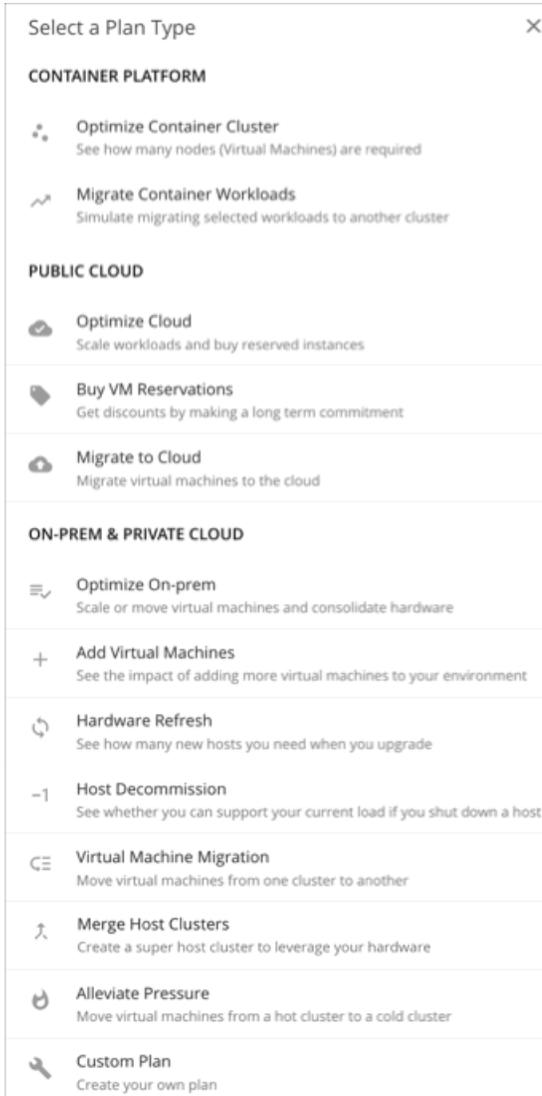
If you set the scope to *a specific* container platform cluster, you can start an Optimize Container Cluster or Migrate Container Workloads plan.

For details, see [Scoping the Intersight Workload Optimizer Session \(on page 31\)](#).

After setting the scope, the **Plan** button appears in the **Overview**.

2. Plan Types

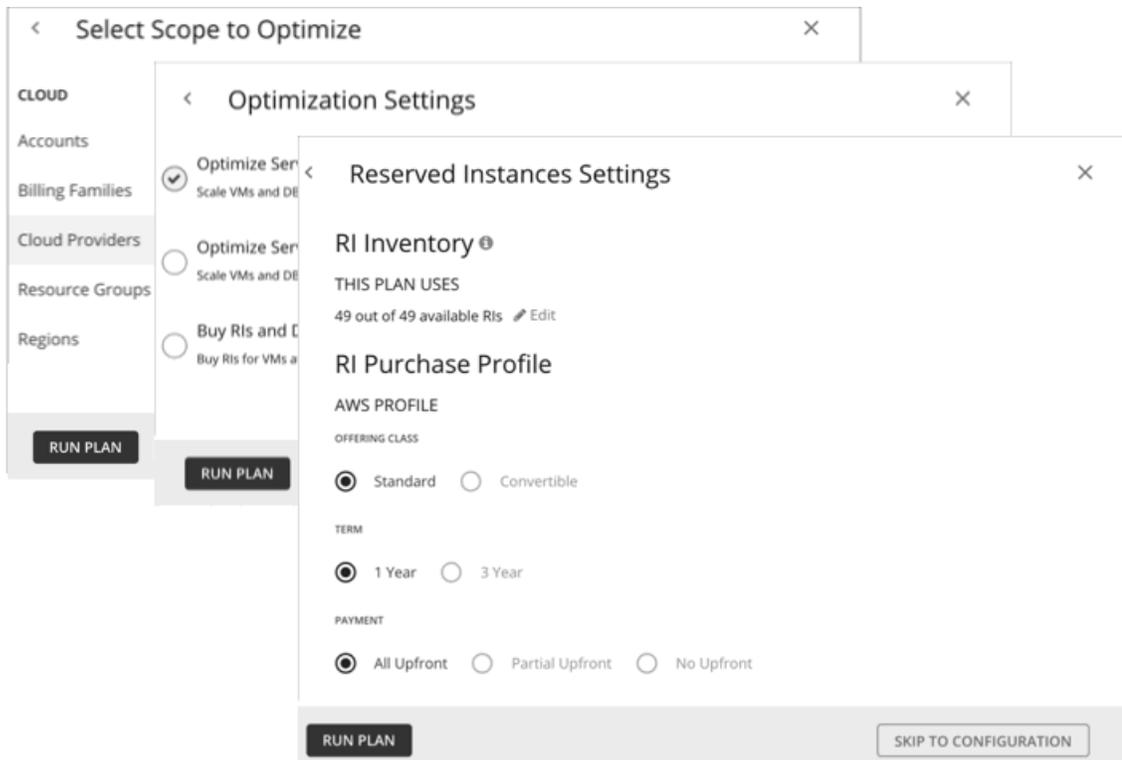
Select from the list of plan types. For more information, see [Plan Scenarios and Types \(on page 424\)](#).



Intersight Workload Optimizer opens the appropriate plan wizard.

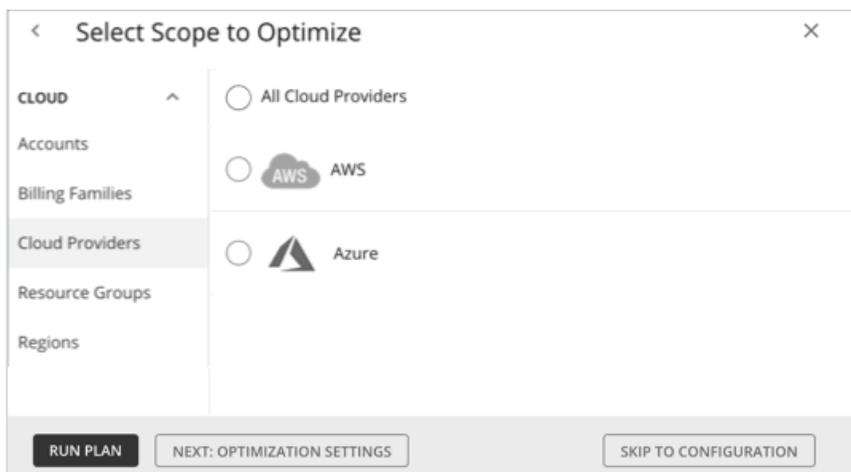
3. Plan Wizards

Each plan type includes a wizard to guide you through creating the scenario. The wizard leads you through the required configuration steps to create a plan that answers a specific question. After you make the required settings, you can skip ahead and run the plan, or continue through all the optional steps.



4. Plan Scope

All plans require a scope. For example, to configure an Optimize Cloud plan, you set the scope to all or specific cloud providers or accounts.



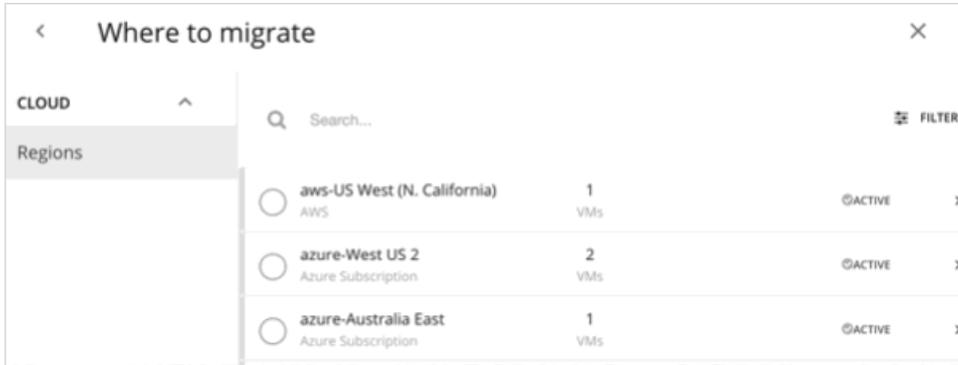
It usually helps to focus on a subset of your environment. For a very large environment, scoped plans run faster.

To narrow the scope, select a group from the list on the left side of the page. The page then refreshes to include only the entities belonging to that group.

Use **Search** or **Filter** to sort through a long list.

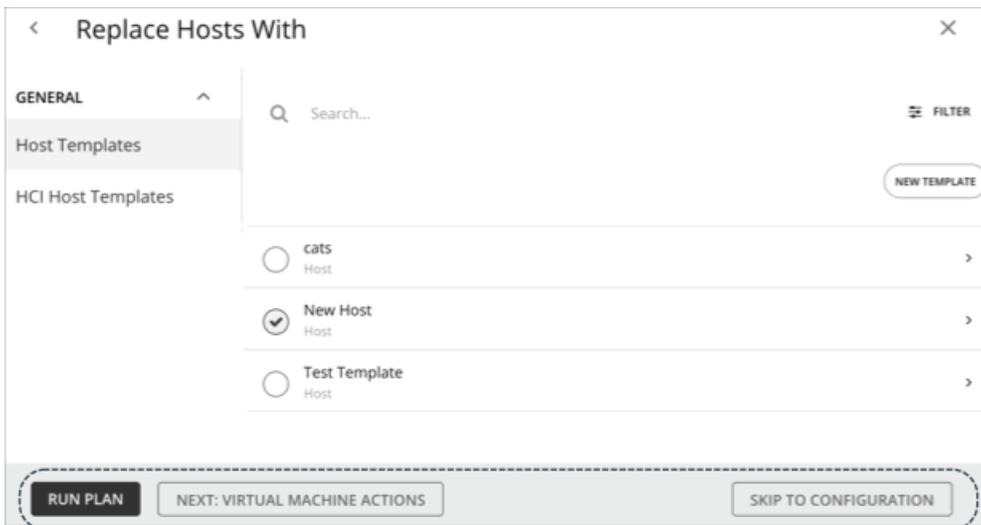
5. Additional Plan Information

The wizard prompts you for any additional information required to run the plan. For example, for a Hardware Refresh plan, you need to identify the hosts that will replace the scoped hosts. For a Migrate to Cloud plan, you need to identify the cloud service provider, region, or group you want the scoped workloads to migrate to.



6. Run the Plan

After you provide the minimum required information for running a plan, the wizard shows you the following options:



- **Run Plan:** Immediately run the plan.
- **Next: [Step]:** Continue with the rest of the wizard and then run the plan.
- **Skip to Configuration:** Skip the rest of the wizard and go to the Plan Page to:
 - Customize the plan settings.
 - See a preview of the plan scenario.
 - Run the plan.

NOTE: For a custom plan, the only option available is **Configure Plan**. Click this button to open the Plan Page, configure the plan settings, and then run the plan.

7. The Plan Page

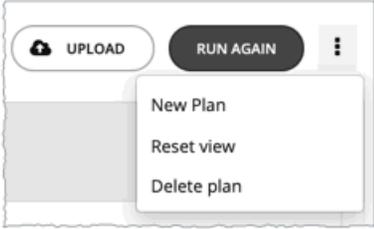
The Plan Page first displays if you skip the wizard or as soon as you run a plan.

For a plan with a large scope, it might take some time before you see the results. You can navigate away from the Plan Page and check the status in the Plan Management Page. You can also cancel a plan that is in progress.

The Plan Page shows the following sections:

| | Current | Optimized | Difference | % |
|---|----------------|----------------|----------------|------------------|
| Workloads with performance risks ⓘ | 0 Out of 1 | 0 Out of 1 | 0 | - |
| Workloads with efficiency opportunities ⓘ | 0 Out of 1 | 0 Out of 1 | 0 | - |
| Workloads out of compliance ⓘ | 1 Out of 1 | 0 Out of 1 | 1 | - |
| On-Demand Compute ⓘ | \$0/mo | \$0/mo | \$0/mo | 0 % |
| Reserved Compute ⓘ | \$0/mo | \$0/mo | \$0/mo | 0 % |
| On-Demand Database ⓘ | \$17/mo | \$83/mo | \$66/mo | ▲ 388.2 % |
| Storage ⓘ | \$0/mo | \$0/mo | \$0/mo | 0 % |
| Total ⓘ | \$17/mo | \$83/mo | \$66/mo | ▲ 388.2 % |

| Plan Page Sections | Description |
|--------------------------|--|
| A. Plan name | Intersight Workload Optimizer automatically generates a name when you create a new plan. Change the name to something that helps you recognize the purpose of this plan. |
| B. Plan scope | Review the scope that you set in a previous step. NOTE: It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (⋮), and then select New Plan . |
| C. Configuration toolbar | Configure additional settings for the plan. You can name the plan, change workload demand and the supply of resources, and specify other changes to the plan market. The toolbar items that display depend on the plan you are creating. |
| D. Configuration summary | Review the plan's configuration settings. You can remove any setting by clicking the x mark on the right. Use the toolbar on top to change the settings. As you make changes to the plan scenario, those changes immediately appear in the Configuration summary. |
| E. Additional options | See what else you can do with the plan. <ul style="list-style-type: none"> ■ Upload: (For Azure only) Upload the results of a Migrate to Cloud plan to the Azure Migrate portal. For details, see Uploading the Results to Azure (on page 467). ■ Run / Run Again: <ul style="list-style-type: none"> – If a plan has not run, click Run and then check the plan results. – If the plan has run and you want to run it again with a different set of configuration settings, click Run Again. This runs the plan scenario against the market in its current state. ■ ⋮ : Click to see more options. |

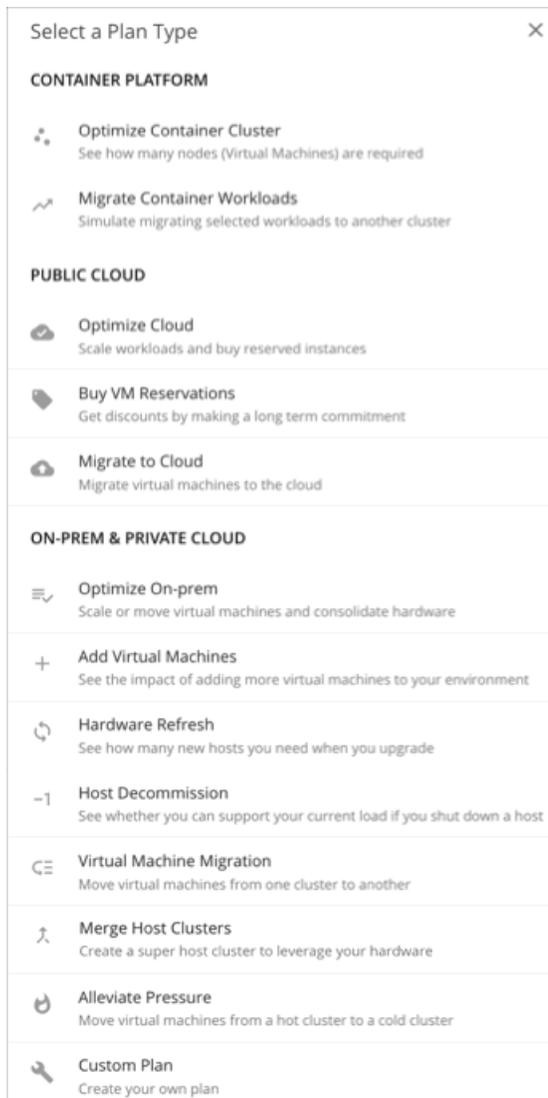
| Plan Page Sections | Description |
|--------------------|---|
| |  <ul style="list-style-type: none"> - New Plan: Configure a new plan. You can choose this option if you want to change the scope of the current plan, which requires that you start over and configure a new plan. - Reset view: Restore charts to their default views. For example, if you changed the commodities displayed in the Optimized Improvements or Comparison charts, you can discard those changes by choosing this option. - Delete plan: Choose if you no longer need the plan. |
| F. Plan results | <p>Review the results in the charts provided.</p> <p>For a plan that has not run, you will see a Scope Preview chart and a one-time message instructing you to run the plan.</p> |

8. Plan Management

All the plans you have created display in the [Plan Management Page \(on page 418\)](#).

Plan Scenarios and Types

To simulate different plan scenarios, Intersight Workload Optimizer provides the following general types of plans:



Optimize Container Cluster

Run an Optimize Container Cluster plan to identify performance and efficiency opportunities for a single cluster. The results show the optimal number of nodes you need to assure performance for your existing workloads, and the impact of actions on the health of your container workloads and infrastructure.

Migrate Container Workloads

Run a Migrate Container Workloads plan to simulate the migration of container workloads from one cluster to another. The plan compares results from a 'lift-and-shift only' scenario against a Intersight Workload Optimizer optimized plan. The results further highlight the actions you need to take to maintain and optimize workload performance in the new cluster.

Optimize Cloud

For the scope of your public cloud environment that you want to examine, run a plan to see all the opportunities you have to reduce cost while assuring performance for your workloads. This includes suggestions to buy [discounts \(on page 25\)](#), comparisons of template and storage usage, and a comparison of current to optimized cost.

Buy VM Reservations

Run the Buy VM Reservations plan to see the most cost-effective [discount \(on page 25\)](#) purchases that will continue to assure performance for your cloud VMs.

Migrate to Cloud

A Migrate to Cloud plan simulates migration of on-prem VMs to the cloud, or migration of VMs from one cloud provider to another.

NOTE:

For migrations within your on-prem environment, use the *Virtual Machine Migration* plan type.

Optimize On-prem

See the effects of executing certain actions, such as scaling virtual machines, suspending hosts, or provisioning storage, to your on-prem environment.

Add Virtual Machines

Adding virtual machines increases the demand that you place on your environment's infrastructure. You can set up a plan to add individual VMs or groups of VMs in your environment, or based on templates.

Hardware Refresh

Choose hosts that you want to replace with different hardware. For example, assume you are planning to upgrade the hosts in a cluster. How many do you need to deploy, and still assure performance of your applications? Create templates to represent the upgraded hosts and let the plan figure out how many hosts you really need.

To increase the accuracy of the plan results, Intersight Workload Optimizer analysis considers a cluster's overall resource utilization over the last ten days. The platform identifies the day within those ten days when percentile utilization for the cluster reached 90%, and then uses each VM's actual utilization data *on that day* to perform its analysis.

NOTE:

If you configure a Hardware Refresh plan to use a baseline snapshot, the plan will use that snapshot's data instead of the cluster's percentile data.

Host Decommission

If your environment includes underutilized hardware, you can use a plan to see whether you can decommission hosts without affecting the workloads that depend on them.

Virtual Machine Migration

Use this plan type to simulate workload migrations within your on-prem environment.

You can see whether you have enough resources to move your workload from its current provider group to another. For example, assume you want to decommission one datacenter and move all its workload to a different datacenter. Does the target datacenter have enough physical resources to support the workload move? Where should that workload be placed? How can you calculate the effect such a change would have on your overall infrastructure?

To calculate this information, create a plan that:

- Limits the plan scope to two data centers (or clusters) – the one you will decommission, and the one that will take on the extra workload
- Removes all the hardware from the decommissioned datacenter
- Calculates workload placement across datacenter (or cluster) boundaries
- Does not provision new hardware to support the workload

Merge Host Clusters

See the effects of merging two or more host clusters. For example, you can see if merging the host clusters would require provisioning additional storage to support current demand, or if ignoring cluster boundaries would improve performance and efficiency.

Alleviate Pressure

Choose a cluster that shows bottlenecks or other risks to performance, and check to see the minimal changes you can make by migrating some workloads to another cluster. The cluster that is showing risks is a *hot* cluster, and the cluster you will migrate to is a *cold* cluster.

Custom Plan

With a custom plan, you skip directly to the plan configuration after specifying the plan scope, and set up whatever type of scenario you want.

You would also choose **Custom Plan** if you need to run plans that include containers and container pods.

Optimize Container Cluster Plan

The screenshot displays the 'Optimize Container Cluster 1' interface. The scope is set to 'Kubernetes-PT-AKS'. There are buttons for 'RUN AGAIN', a download icon, and a settings icon. Below the scope are 'Add', 'Remove', and 'Actions' buttons. The left sidebar shows configuration settings for Container Spec, Container Pod, and Virtual Machines, all with 'ENABLED' status. The main area shows 'RESULTS OVERVIEW' and 'PLAN ACTIONS (258)'. A table titled 'Optimize Container Cluster Summary' provides a comparison of current and after-plan states for various metrics.

| | Current | After Plan | Difference | % |
|-------------------------------|------------|------------|------------|----------|
| Container Pods | 118 | 113 | 5 | ▼ 4.2 % |
| Virtual Machines | 8 | 7 | 1 | ▼ 12.5 % |
| Pod Density | 14.8 : 1 | 16.1 : 1 | 1.3 : 1 | ▲ 8.8 % |
| Cluster CPU Capacity | 20 Cores | 22 Cores | 2 Cores | ▲ 10 % |
| Cluster Memory Capacity | 74 GB | 82.9 GB | 8.9 GB | ▲ 12.1 % |
| Cluster Allocatable CPU | 19.2 Cores | 21.1 Cores | 2 Cores | ▲ 10.3 % |
| Cluster Allocatable Memory | 53.8 GB | 62.6 GB | 8.8 GB | ▲ 16.4 % |
| Cluster CPU Overcommitment | 261 % | 223.2 % | 37.8 % | ▼ 14.5 % |
| Cluster Memory Overcommitment | 83.7 % | 45.4 % | 38.3 % | ▼ 45.8 % |

At the bottom right of the table area, there is a 'SHOW ALL >' link.

Run an Optimize Container Cluster plan to identify performance and efficiency opportunities for a single container platform cluster. The results show the optimal number of nodes you need to assure performance for your existing workloads, and the impact of actions on the health of your container workloads and infrastructure. For example, you can see how container resize actions change the limits and requests allocated per namespace, or how node provision/suspend actions impact allocatable capacity for the cluster. For a cluster in the public cloud, the results also include the cost impact of node actions.

You can scope the plan to a:

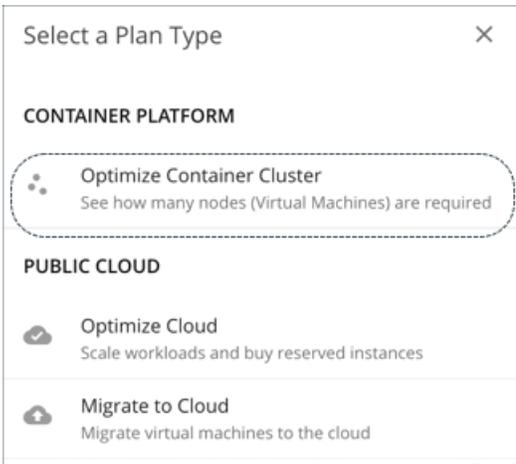
- Standalone cluster
- Cluster in an on-prem or public cloud environment

Scoping to a group within a cluster (such as a group of nodes) is currently not supported.

Configuring an Optimize Container Cluster Plan

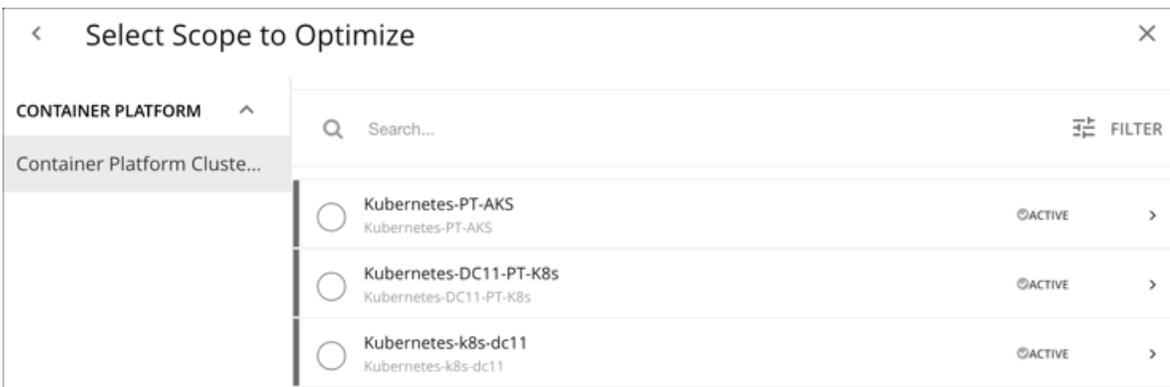
You can start an Optimize Container Cluster plan when you open the Plan page or set the scope to a container platform cluster.

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).



1. Scope

Select a container platform cluster to optimize.



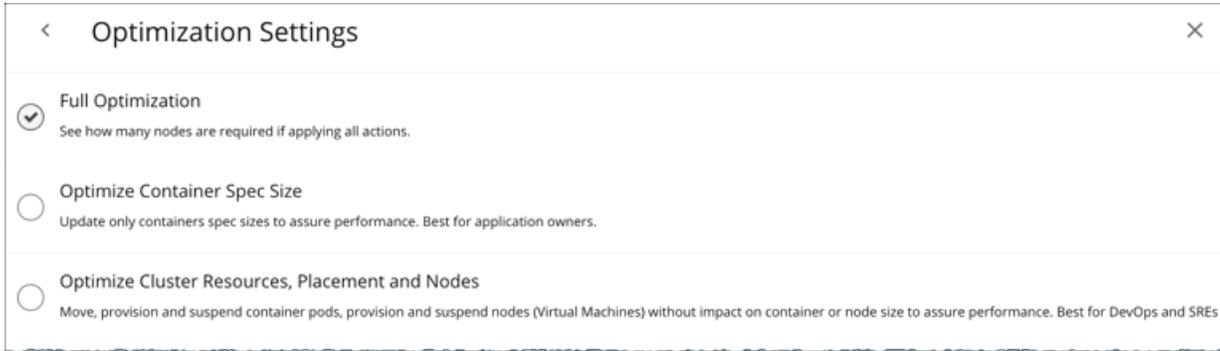
Scoping to a group within a cluster (such as a group of nodes) is currently not supported.

NOTE:

After selecting a cluster, you can skip the next step (**Optimization Settings**) and run the plan. Intersight Workload Optimizer runs the **Full Optimization** scenario in this case.

2. Optimization Settings

Choose from the given optimization scenarios.



■ Full Optimization

Intersight Workload Optimizer will recommend all relevant actions to optimize the cluster. For example, it can recommend provisioning nodes or resizing containers to meet application demand, or moving pods from one node to another to reduce congestion.

Intersight Workload Optimizer can recommend the following actions:

- Resize namespace compute resource quotas
- Resize container limits and requests
- Move pods
- Provision or suspend nodes
- Scale volumes

NOTE:

For a cluster in the public cloud, Intersight Workload Optimizer shows the cost impact of actions on nodes and volumes, to help you track your cloud spend. Intersight Workload Optimizer only reports the costs attached to these actions, and does *not* perform cost analysis on the cluster.

For a cluster in an on-prem environment, Intersight Workload Optimizer can also recommend the following actions:

- Move VMs
- Provision or suspend hosts
- Provision or suspend storage

■ Optimize Container Spec Size

Intersight Workload Optimizer will only recommend resizing container limits and requests. This is ideal for application owners who manage the containers that their applications run on, but not the underlying container infrastructure.

■ Optimize Cluster Resources, Placement, and Nodes

Intersight Workload Optimizer will recommend *all* relevant actions, *except* resizing container limits and requests. This is ideal for teams who oversee the health of your container infrastructure, and want to evaluate the impact of *not* rightsizing workloads.

After selecting an optimization scenario, you can:

- Run the plan.
- Or
- Choose **Skip to Configuration** to configure additional settings. See the next section for details.

(Optional) Additional Plan Settings

You can fine tune your selected optimization scenario or include additional scenarios before you run the plan.

- Enable or disable actions

Fine tune your optimization scenario by enabling or disabling actions for containers, pods, or nodes. For example, you may have selected **Full Optimization**, but only for containers, nodes, and pods that are allowed to move. In this case, you would disable move actions for the pods that should never move.

For clusters in on-prem environments, you can also enable or disable actions for hosts and storage.

IMPORTANT:

To avoid seeing inaccurate plan results, do *not* disable all actions.

- Add pods

See resource changes if you add more pods to the cluster. For example, you might need to provision nodes to accommodate the new pods.

Select an existing pod within or outside the selected cluster, and then specify how many copies to add. The plan simulates adding pods with the same resources as the selected pod.

- Remove pods or nodes

See the effect of removing pods or nodes from the cluster. For example, pod density could improve significantly if you remove pods that you no longer need, or certain pods might become unplaced if you remove nodes.

Working with Optimize Container Cluster Plan Results

After the plan runs, you can view the results to see how the plan settings you configured affect your environment.

The screenshot shows the 'Optimize Container Cluster 1' interface. On the left, there is a 'CONFIGURATION' sidebar with three sections: 'Container Spec Action Settings' (Scale for ContainerSpec: ENABLED), 'Container Pod Action Settings' (Move for Container Pods: ENABLED), and 'Virtual Machines Action Settings' (Provision for Virtual Machines: ENABLED, Suspend for Virtual Machines: ENABLED, Scale for Virtual Machines: ENABLED). The main area is titled 'RESULTS OVERVIEW' and 'PLAN ACTIONS (258)'. It features a table titled 'Optimize Container Cluster Summary' for 'Kubernetes-PT-AKS'.

| | Current | After Plan | Difference | % |
|-------------------------------|------------|------------|------------|----------|
| Container Pods | 118 | 113 | 5 | ▼ 4.2 % |
| Virtual Machines | 8 | 7 | 1 | ▼ 12.5 % |
| Pod Density | 14.8 : 1 | 16.1 : 1 | 1.3 : 1 | ▲ 8.8 % |
| Cluster CPU Capacity | 20 Cores | 22 Cores | 2 Cores | ▲ 10 % |
| Cluster Memory Capacity | 74 GB | 82.9 GB | 8.9 GB | ▲ 12.1 % |
| Cluster Allocatable CPU | 19.2 Cores | 21.1 Cores | 2 Cores | ▲ 10.3 % |
| Cluster Allocatable Memory | 53.8 GB | 62.6 GB | 8.8 GB | ▲ 16.4 % |
| Cluster CPU Overcommitment | 261 % | 223.2 % | 37.8 % | ▼ 14.5 % |
| Cluster Memory Overcommitment | 83.7 % | 45.4 % | 38.3 % | ▼ 45.8 % |

At the bottom right of the table area, there is a 'SHOW ALL >' link.

General Guidelines

Familiarize yourself with these common terms that appear in the plan results:

- A container pod represents the compute demand from a running pod.
- A node (virtualized or bare metal) is represented as a VM.
- *Used* (or *Usage*) values represent actual resource consumption. For example, a node that consumes 100 MB of memory has a used value of 100 MB.
- *Utilization* values represent used/usage values against capacity. For example, a node that consumes 100 MB of memory against a total capacity of 500 MB has a utilization value of 20%.

Optimize Container Cluster Summary

RESULTS OVERVIEW PLAN ACTIONS (258)

| Optimize Container Cluster Summary ? | | | | |
|---|------------|------------|------------|----------|
| Kubernetes-PT-AKS | | | | |
| | Current | After Plan | Difference | % |
| Container Pods | 118 | 113 | 5 | ▼ 4.2 % |
| Virtual Machines | 8 | 7 | 1 | ▼ 12.5 % |
| Pod Density | 14.8 : 1 | 16.1 : 1 | 1.3 : 1 | ▲ 8.8 % |
| Cluster CPU Capacity | 20 Cores | 22 Cores | 2 Cores | ▲ 10 % |
| Cluster Memory Capacity | 74 GB | 82.9 GB | 8.9 GB | ▲ 12.1 % |
| Cluster Allocatable CPU | 19.2 Cores | 21.1 Cores | 2 Cores | ▲ 10.3 % |
| Cluster Allocatable Memory | 53.8 GB | 62.6 GB | 8.8 GB | ▲ 16.4 % |
| Cluster CPU Overcommitment | 261 % | 223.2 % | 37.8 % | ▼ 14.5 % |
| Cluster Memory Overcommitment | 83.7 % | 45.4 % | 38.3 % | ▼ 45.8 % |

SHOW ALL >

This chart shows how your container environment and the underlying resources will change after you execute the actions that the plan recommends. The chart shows the following information:

- **Container Pods**

Count of active container pods in the plan.

- **Virtual Machines**

Count of active nodes in the plan. This chart does not count "non-participating" entities in the real-time market, such as suspended nodes.

- **Pod Density**

Average number of pods per node.

For the total number of pods against the node capacity (maximum pods per node), see the **Number of Consumers** data in the following charts:

- Nodes (VMs) Optimized Improvements
- Nodes (VMs) Comparison
- Container Cluster Optimized Improvements
- Container Cluster Comparison

- **Cluster CPU Capacity**

Total CPU capacity for the cluster. The 'After Plan' result indicates how much CPU capacity will result in the optimal number of nodes required to run workloads.

- **Cluster Memory Capacity**

Total memory capacity for the cluster. The 'After Plan' result indicates how much memory capacity will result in the optimal number of nodes required to run workloads.

- **Cluster Allocatable CPU**

Total amount of cluster CPU [available](#) for pod requests. The 'After Plan' result indicates how much of the allocatable CPU capacity will change if you provision or suspend nodes.

- **Cluster Allocatable Memory**

Total amount of cluster memory [available](#) for pod requests. The 'After Plan' result indicates how much of the allocatable memory capacity will change if you provision or suspend nodes.

■ Cluster CPU Overcommitment

(Only for containers with CPU limits) This indicates whether the CPU limits exceed the capacity of the underlying nodes. A value greater than 100% indicates overcommitment. Intersight Workload Optimizer manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.

Intersight Workload Optimizer only calculates overcommitment in plans. The calculation can be expressed as:

$$\text{Overcommitment} = \text{Sum of CPU limits for all containers} / \text{Sum of CPU capacity for all nodes}$$

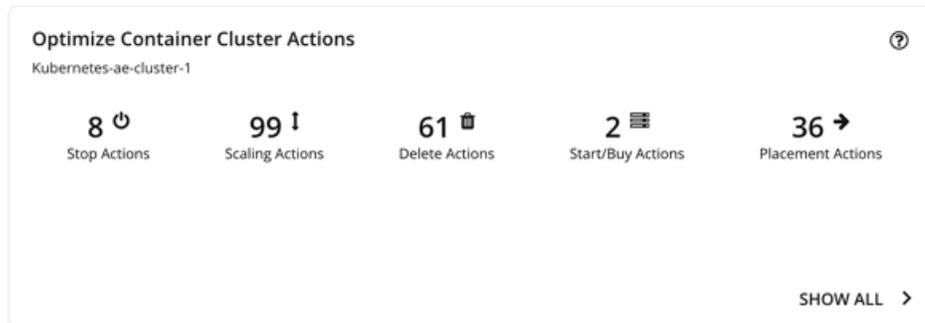
■ Cluster Memory Overcommitment

(Only for containers with memory limits) This indicates whether the memory limits exceed the capacity of the underlying nodes. A value greater than 100% indicates overcommitment. Intersight Workload Optimizer manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.

Intersight Workload Optimizer only calculates overcommitment in plans. The calculation can be expressed as:

$$\text{Overcommitment} = \text{Sum of memory limits for all containers} / \text{Sum of memory capacity for all nodes}$$

Optimize Container Cluster Actions



This chart summarizes the actions that you need to execute to achieve the plan results. For example, you might need to resize limits and requests for containers (via the associated Workload Controllers) to address performance issues. Or, you might need to move pods from one node to another to reduce congestion.

Smarter redistribution and workload rightsizing also drive cluster optimization, resulting in the need to provision node(s) based on application demand, or to defragment node resources to enable node suspension.

Intersight Workload Optimizer can recommend the following actions:

- Resize namespace compute resource quotas
- Resize container limits and requests

NOTE:

Executing several container resize actions can be very disruptive since pods need to restart with each resize. For replicas of the container scale group(s) related to a single Workload Controller, Intersight Workload Optimizer consolidates resize actions into one *merged action* to minimize disruptions. When a merged action has been executed (via the associated Workload Controller), all resizes for all related container specifications will be changed at the same time, and pods will restart once.

- Move pods
- Provision or suspend nodes
- Scale volumes

NOTE:

For a cluster in the public cloud, Intersight Workload Optimizer shows the cost impact of actions on nodes and volumes, to help you track your cloud spend. Intersight Workload Optimizer only reports the costs attached to these actions, and does *not* perform cost analysis on the cluster. See the Optimized Savings and Optimized Investments charts for more information.

For an on-prem cluster, Intersight Workload Optimizer can also recommend the following actions:

- Move VMs
- Provision or suspend hosts
- Provision or suspend storage

Optimized Savings



For a cluster in the public cloud, Intersight Workload Optimizer shows the savings you would realize if you execute the actions (such as node suspension) that the plan recommends to increase infrastructure efficiency. Note that efficiency is the driver of this action, *not* cost. Cost information is included to help you track your cloud spend.

The chart shows total monthly savings. Click **Show All** to view the actions with cost savings.

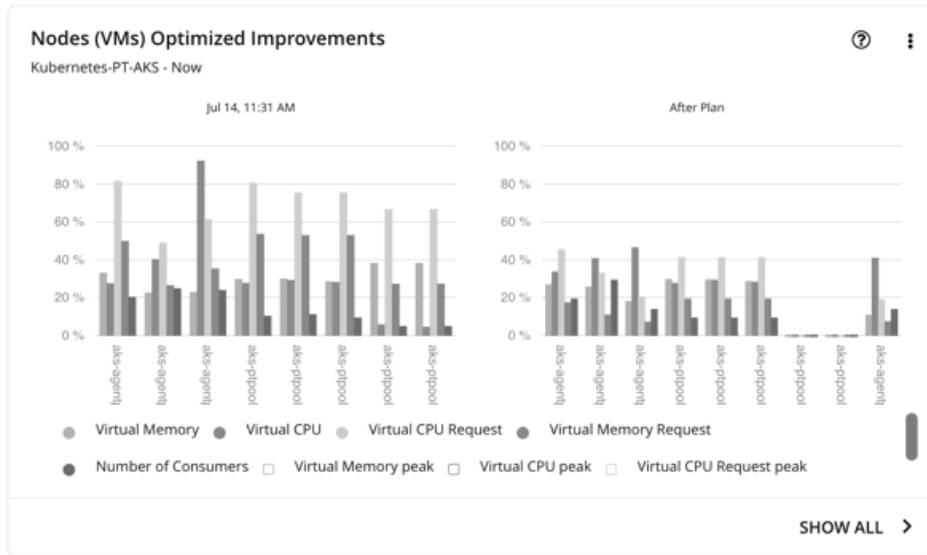
Optimized Investments



For a cluster in the public cloud, Intersight Workload Optimizer shows the costs you would incur if you execute the node and volume scaling actions that the plan recommends to address performance issues. For example, if some applications risk losing performance, Intersight Workload Optimizer can recommend provisioning nodes to increase capacity. This chart shows how these actions translate to an increase in expenditure. Note that performance and efficiency are the drivers of these actions, *not* cost. Cost information is included to help you plan for the increase in capacity.

The chart shows total monthly investments. Click **Show All** to view the actions that require investments.

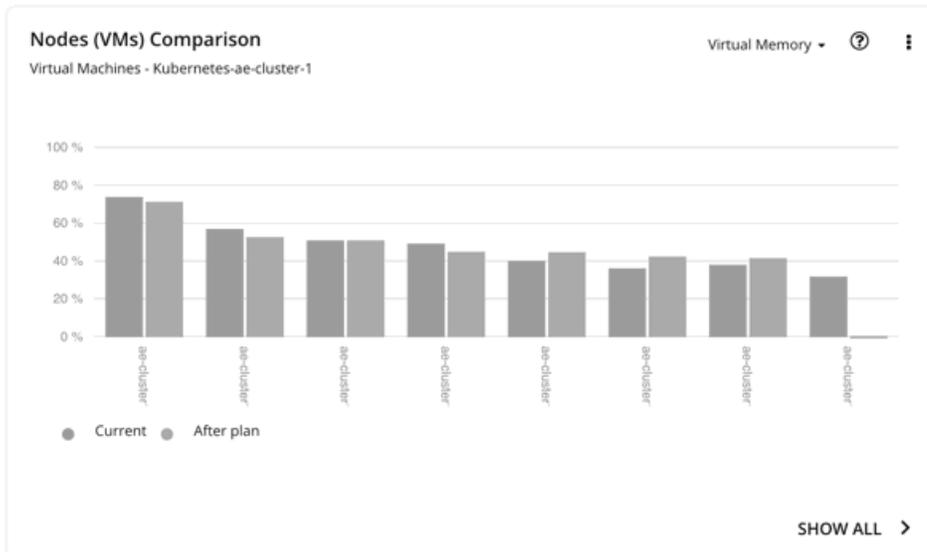
Nodes (VMs) Optimized Improvements



This chart compares the following before and after the plan:

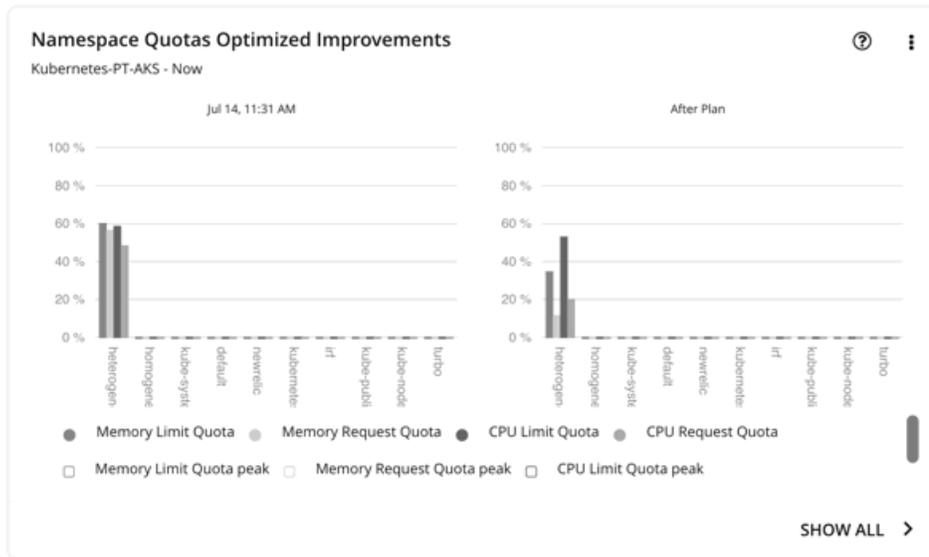
- Utilization of the following for all nodes:
 - vMem
 - vCPU
 - vMem Request
 - vCPU Request
- Number of pods consuming resources against the maximum pod capacity for all the nodes

Nodes (VMs) Comparison



This chart compares node resource utilization (one metric at a time) before and after the plan.

Namespace Quotas Optimized Improvements



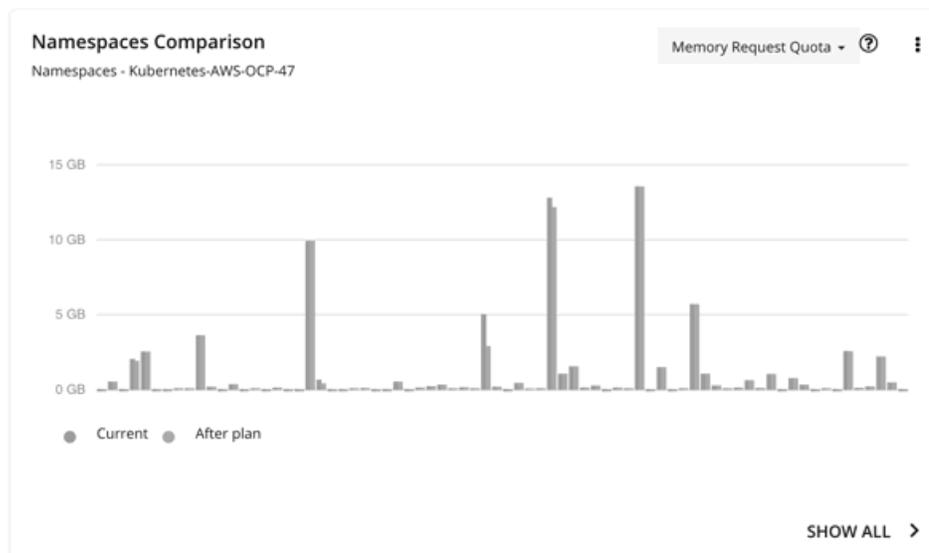
This chart shows pod utilization of resource quotas defined in namespaces. Resource quotas include:

- CPU Limit Quota
- Memory Limit Quota
- CPU Request Quota
- Memory Request Quota

For namespaces without defined quotas, utilization is 0 (zero).

With or without quotas, you can see the sum of pod limits and requests per namespace. Go to the top-right section of the Plan Results page, click the download button, and select **Namespace**. Utilization data in the downloaded file shows these limits and requests. You can also compare usage values in the Namespaces Comparison chart.

Namespaces Comparison



This chart compares namespace quota usage (one metric at a time) before and after the plan.

Use this chart to see how container resizing changes the limits and requests allocated per namespace, whether you leverage quotas or not.

To achieve the 'After Plan' results, click **Show All**. In the Details page that opens, go to the Name column and then click the namespace link. This opens another page with a list of pending actions for the namespace.

Namespaces Comparison

| Name | Current | | | After Plan | | |
|-----------------|--------------------|-------------------------|--------------------|--------------------|-------------------------|--------------------|
| | Memory Limit Quota | Memory Limit Quota Used | Memory Limit Quota | Memory Limit Quota | Memory Limit Quota Used | Memory Limit Quota |
| turbonomic | ∞ | 570.7 GB | 0 % | ∞ | 70.3 GB | 0 % |
| kube-node-lease | ∞ | 0 KB | 0 % | ∞ | 0 KB | 0 % |

Namespace **turbonomic**

OVERVIEW | DETAILS | POLICIES | **ACTIONS (47)**

RESIZE ^

Workloa... (46)

PROVISI... ^

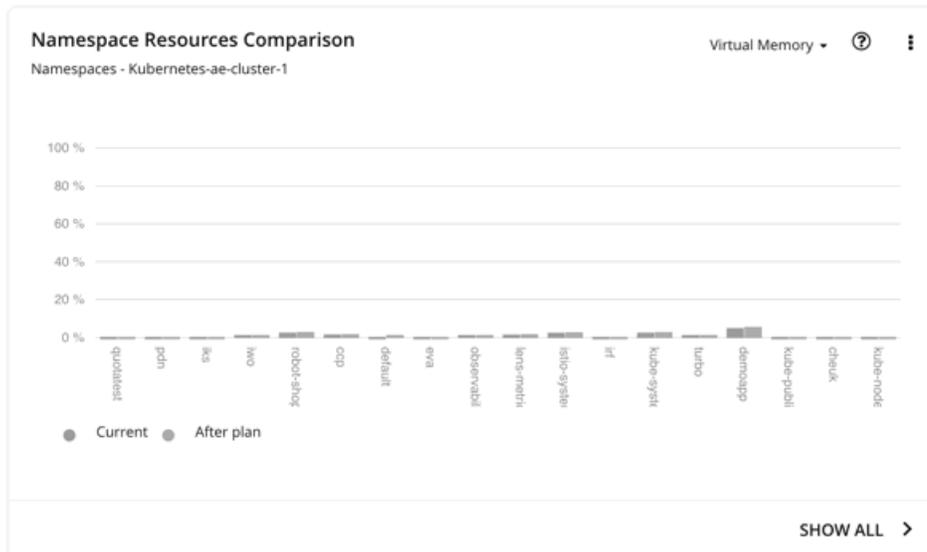
Virtual Ma... (1)

Resize Actions (46)

Type to search

- Workload Controller Name Container Cluster
- mediation-awscost Kubernetes-Turbonomic
- mediation-aws Kubernetes-Turbonomic

Namespace Resources Comparison



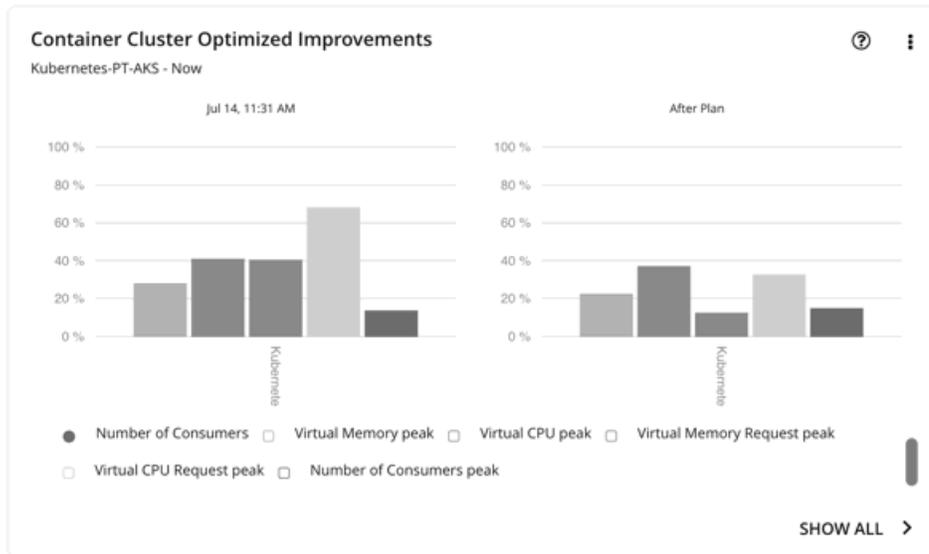
This chart shows how much cluster resources per namespace are utilized by pods. Utilization can be expressed as follows:

$$\text{Utilization} = \text{Sum of actual vMem/vCPU used by pods} / \text{vMem/vCPU capacity for the cluster}$$

This information helps you understand which namespaces use the most cluster resources. You can also use it for showback analysis. vMem and vCPU utilized by pods in the namespaces would change when the number of nodes changes as a result of executing the plan actions.

This chart is especially useful if you do not have resource quotas defined in your namespaces.

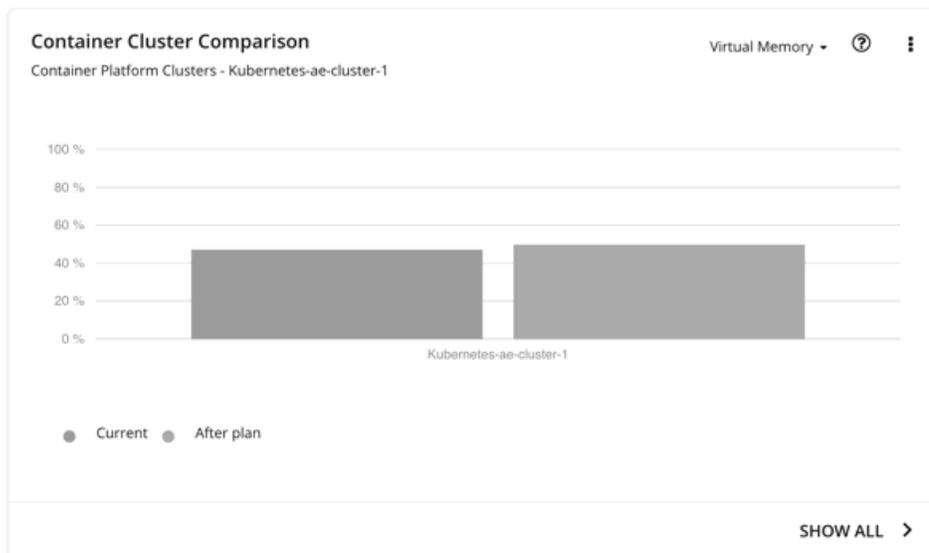
Container Cluster Optimized Improvements



This chart shows the following, assuming you execute all actions in the plan:

- Changes to the utilization of cluster resources
- Overcommitment values

Container Cluster Comparison



This chart compares the following before and after the plan:

- Utilization of cluster resources (one metric at a time)
- Overcommitment values

Optimized Improvements for Hosts, Storage, and Virtual Machines

Use these charts if you ran the plan in an on-prem cluster. These charts show how the utilization of resources would change if you accept all of the actions listed in the Plan Actions chart.

Hosts, Storage Devices, and Virtual Machines Comparison

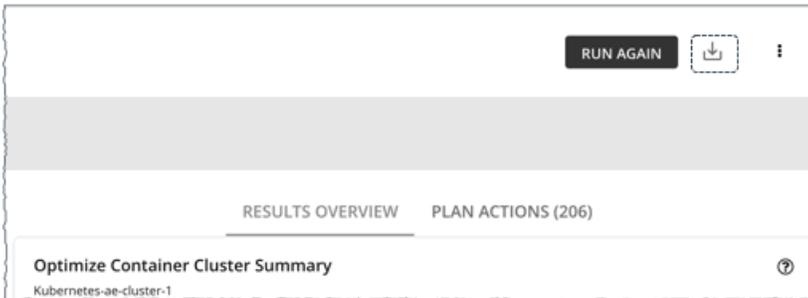
Use these charts if you ran the plan in an on-prem cluster. These charts show how the utilization of a particular commodity (such as memory or CPU) for each entity in the plan would change if you execute the recommended actions.

NOTE:

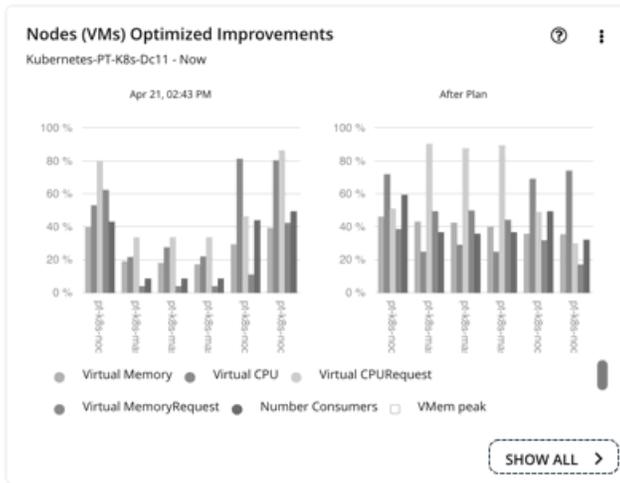
For the Storage Devices Comparison chart, if you set the view to **VM Per Storage** and click **Show all**, the total number of VMs sometimes does not match the number in the Summary chart. This happens if there are VMs in the plan that use multiple storage devices. The Storage Devices Comparison chart counts those VMs multiple times, depending on the number of storage devices they use, while the Plan Summary chart shows the actual number of VMs.

Downloading Plan Results

To download results for nodes, namespaces, or the cluster, click the download button at the top-right section of the Plan Results page.



You can also download the plan results shown in individual charts. Click the **Show All** button for a chart, and then the download button at the top-right section of the Details page.



Nodes (VMs) Optimized Improvements

| Virtual Machine Name | Container Cluster | Current | | | | | |
|----------------------|---------------------|-----------------|-----------------|-------------|-------------|------------------|------------------|
| | | Number Consu... | Number Consu... | Virtual CPU | Virtual CPU | Virtual CPURe... | Virtual CPURe... |
| pt-k8s-node-2 | Kubernetes-PT-K8... | 49.1 % | 49.1 % | 80 % | 82.1 % | 86.1 % | 86.1 % |
| pt-k8s-node-1 | Kubernetes-PT-K8... | 42.7 % | 42.7 % | 52.7 % | 59.5 % | 79.5 % | 79.5 % |

For charts that display infinite capacities (for example, the Namespaces Comparison chart), the downloaded file shows an unusually high value, such as 1,000,000,000 cores, instead of the ∞ symbol.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Migrate Container Workloads Plan

Run a Migrate Container Workloads plan to simulate the migration of workloads from one container platform cluster to another. For example, when you are ready to deploy workloads in your test cluster to production, you can run this plan to evaluate resource requirements in the production cluster. You might also have a need to decommission an existing cluster or consolidate workloads in different clusters into a single cluster, both of which require moving existing workloads to a new cluster. Run this plan to see if all workloads can be placed in the new cluster.

The destination cluster (i.e., the cluster that you choose for migration) can be a:

- Standalone cluster
- Cluster in an on-prem or public cloud environment. For cloud environments, cost information is included in the plan results.

The plan results show the impact of workload migration on the destination cluster, based on two migration scenarios:

| | BEFORE MIGRATION | LIFT & SHIFT | OPTIMIZED |
|-------------------------------|------------------|--------------|--------------|
| Workload Controllers | 69 | 71 | 71 |
| Container Pods | 141 | 145 | 145 |
| Virtual Machines | 7 | 7 | 5 |
| Pod Density | 20.1 : 1 | 20.7 : 1 | 29 : 1 |
| Cluster CPU Capacity | 17 Cores | 27 Cores | 17 Cores |
| Cluster CPU Allocatable | 17 Cores | 17 Cores | 17 Cores |
| Cluster CPU Request | 4.6 Cores | 4.6 Cores | 3.85 Cores |
| Cluster CPU Limit | 773.81 Cores | 269.81 Cores | 270.22 Cores |
| Cluster CPU Overcommitment | 4551.8 % | 999.3 % | 1589.5 % |
| Cluster Memory Capacity | 39.94 GB | 61.94 GB | 39.94 GB |
| Cluster Memory Allocatable | 39.45 GB | 39.45 GB | 39.45 GB |
| Cluster Memory Request | 5.82 GB | 5.83 GB | 5.83 GB |
| Cluster Memory Limit | 27.72 GB | 22.05 GB | 15.81 GB |
| Cluster Memory Overcommitment | 69.4 % | 35.6 % | 39.6 % |

SHOW ALL >

■ Lift & Shift

The Lift & Shift scenario migrates your container workloads based on the resources currently available in the destination cluster.

■ Optimized

The Optimized scenario identifies opportunities to optimize performance in the destination cluster. For example, after analyzing historical resource utilization, Intersight Workload Optimizer might recommend resizing limits and requests for containers (via the associated workload controllers) to maintain performance. If you were to migrate workloads based on the resources currently available in the destination cluster, then your applications could risk losing performance.

NOTE:

The plan results do not show data for the source cluster.

The results further highlight the actions you need to take to maintain and optimize workload performance in the destination cluster.

Configuring a Migrate Container Workloads Plan

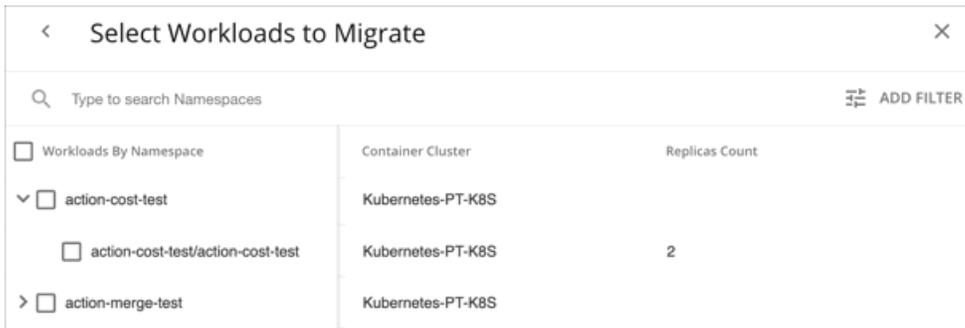
You can begin creating a plan scenario from two places in the user interface:

- From the Plan Page (Recommended)
 - On the left menu, click **Plan** to open the Plan Page. Click **New Plan**, and then select **Migrate Container Workloads**.
- From the supply chain
 - To start a plan scenario from the supply chain, set the scope to a container platform cluster and then click **Plan** at the top-right corner of the page.

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

1. Scope

Select the container workloads that you want to migrate.



| Select Workloads to Migrate | | |
|--|-------------------|----------------|
| Type to search Namespaces | | |
| | Container Cluster | Replicas Count |
| <input type="checkbox"/> Workloads By Namespace | | |
| <input type="checkbox"/> action-cost-test <ul style="list-style-type: none"> <input type="checkbox"/> action-cost-test/action-cost-test | Kubernetes-PT-K8S | |
| | Kubernetes-PT-K8S | 2 |
| <input type="checkbox"/> action-merge-test | Kubernetes-PT-K8S | |

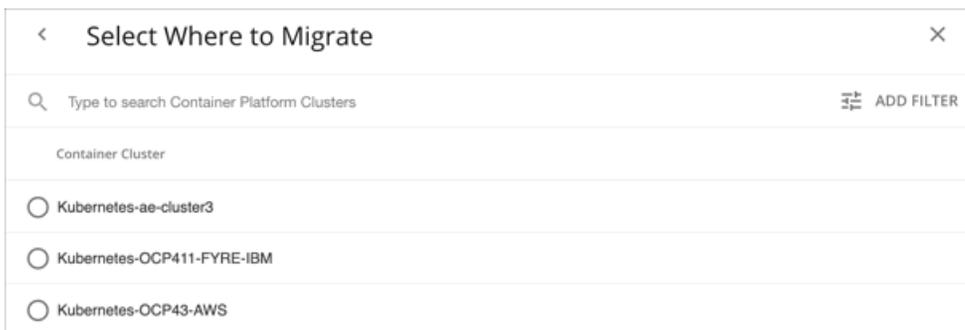
Container workloads are grouped by namespace. Expand a namespace to see individual Workload Controllers that manage workloads, and the number of container replicas for each Workload Controller. You can adjust the number of replicas in a later step. For example, you can add replicas to meet workload demand.

To define your scope, select individual namespaces and/or Workload Controllers, from one or several clusters. For example, if you are migrating workloads in two clusters, select the namespaces in those clusters.

To find specific workloads, type a keyword in the Search bar or click **Add Filter**. You can filter workloads by container platform cluster, namespace, tag, or Workload Controller.

2. Where to Migrate

Choose the destination cluster for the container workloads you selected.



| Select Where to Migrate | |
|--|----------------------------|
| Type to search Container Platform Clusters | |
| Container Cluster | |
| <input type="radio"/> | Kubernetes-ae-cluster3 |
| <input type="radio"/> | Kubernetes-OCP411-FYRE-IBM |
| <input type="radio"/> | Kubernetes-OCP43-AWS |

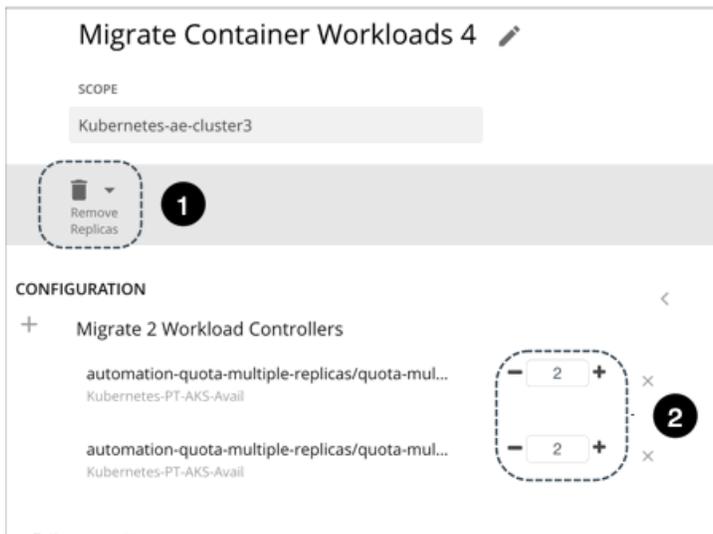
You can migrate to a:

- Standalone container platform cluster

- Container platform cluster in an on-prem or public cloud environment

After selecting a cluster, you can:

- Run the plan.
- Or
- Choose **Skip to Configuration** to configure additional settings, including:



1. Removing existing replicas from the destination cluster. You may need to do this to make room for the workloads you are migrating.
2. Adjusting the number of replicas for the Workload Controllers in scope. For example, you may need to add replicas to meet workload demand.

Working with Migrate Container Workloads Plan Results

After the plan runs, you can view the results to see the impact of workload migration on the *destination* cluster. The results do *not* show data for the *source* cluster (i.e., the cluster that currently hosts the workloads).

General Guidelines

Familiarize yourself with these common terms that appear in the plan results:

- A container pod represents the compute demand from a running pod.
- A node (virtualized or bare metal) is represented as a VM.
- *Used* (or *Usage*) values represent actual resource consumption. For example, a node that consumes 100 MB of memory has a used value of 100 MB.
- *Utilization* values represent used/usage values against capacity. For example, a node that consumes 100 MB of memory against a total capacity of 500 MB has a utilization value of 20%.

Migrate Container Workloads Summary

RESULTS OVERVIEW LIFT & SHIFT ACTIONS (27) OPTIMIZED ACTIONS (38)

Migrate Container Workloads Summary ?

Kubernetes-ae-cluster3

| | BEFORE MIGRATION | LIFT & SHIFT | OPTIMIZED |
|-------------------------------|------------------|--------------|--------------|
| Workload Controllers | 69 | 71 | 71 |
| Container Pods | 141 | 145 | 145 |
| Virtual Machines | 7 | 7 | 5 |
| Pod Density | 20.1 : 1 | 20.7 : 1 | 29 : 1 |
| Cluster CPU Capacity | 17 Cores | 27 Cores | 17 Cores |
| Cluster CPU Allocatable | 17 Cores | 17 Cores | 17 Cores |
| Cluster CPU Request | 4.6 Cores | 4.6 Cores | 3.85 Cores |
| Cluster CPU Limit | 773.81 Cores | 269.81 Cores | 270.22 Cores |
| Cluster CPU Overcommitment | 4551.8 % | 999.3 % | 1589.5 % |
| Cluster Memory Capacity | 39.94 GB | 61.94 GB | 39.94 GB |
| Cluster Memory Allocatable | 39.45 GB | 39.45 GB | 39.45 GB |
| Cluster Memory Request | 5.82 GB | 5.83 GB | 5.83 GB |
| Cluster Memory Limit | 27.72 GB | 22.05 GB | 15.81 GB |
| Cluster Memory Overcommitment | 69.4 % | 35.6 % | 39.6 % |

SHOW ALL >

Table Columns

Table columns shows the following information for the destination cluster:

- Before Migration**
 This column shows the state of the destination cluster before workload migration.
- Lift & Shift**
 The Lift & Shift scenario migrates your container workloads based on the resources currently available in the destination cluster.
- Optimized**
 The Optimized scenario identifies opportunities to optimize performance in the destination cluster. For example, after analyzing historical resource utilization, Intersight Workload Optimizer might recommend resizing limits and requests for containers (via the associated workload controllers) to maintain performance. If you were to migrate workloads based on the resources currently available in the destination cluster, then your applications could risk losing performance.

Table Rows

Table rows show the following information for the destination cluster:

| Row | Description |
|----------------------|--|
| Workload Controllers | Count of active Workload Controllers. This chart does not count "non-participating" entities in the real-time market, such as inactive Workload Controllers. |
| Container Pods | Count of active (running) container pods. |
| Virtual Machines | Count of active nodes (virtualized or bare metal). This chart does not count "non-participating" nodes in the real-time market, such as suspended nodes. |
| Pod Density | Average number of pods per node. The 'Optimized' result shows if you can improve density by increasing the number of pods per node. |

| Row | Description |
|-------------------------------|--|
| | <p>For the total number of pods against the node capacity (maximum pods per node), see the Number of Consumers data in the following charts:</p> <ul style="list-style-type: none"> ■ Nodes (VMs) Optimized Improvements ■ Container Platform Cluster Optimized Improvements |
| Cluster CPU Capacity | Total CPU capacity for the cluster. The 'Optimized' result indicates how much CPU capacity will result in the optimal number of nodes required to run workloads. |
| Cluster CPU Allocatable | Total amount of cluster CPU available for pod requests. The 'Optimized' result indicates how much of the allocatable CPU capacity will change if you provision or suspend nodes. |
| Cluster CPU Request | Total CPU Request capacity for the cluster. |
| Cluster CPU Limit | Total CPU Limit capacity for the cluster. |
| Cluster CPU Overcommitment | <p>(Only for containers with CPU limits) This indicates whether the CPU limits exceed the capacity of the underlying nodes. A value greater than 100% indicates overcommitment. Intersight Workload Optimizer manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.</p> <p>Intersight Workload Optimizer only calculates overcommitment in plans. The calculation can be expressed as:</p> $\text{Overcommitment} = \frac{\text{Sum of CPU limits for all containers}}{\text{Sum of CPU capacity for all nodes}}$ |
| Cluster Memory Capacity | Total memory capacity for the cluster. The 'Optimized' result indicates how much memory capacity will result in the optimal number of nodes required to run workloads. |
| Cluster Memory Allocatable | Total amount of cluster memory available for pod requests. The 'Optimized' result indicates how much of the allocatable memory capacity will change if you provision or suspend nodes. |
| Cluster Memory Request | Total Memory Request capacity for the cluster. |
| Cluster Memory Limit | Total Memory Limit capacity for the cluster. |
| Cluster Memory Overcommitment | <p>(Only for containers with memory limits) This indicates whether the memory limits exceed the capacity of the underlying nodes. A value greater than 100% indicates overcommitment. Intersight Workload Optimizer manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.</p> <p>Intersight Workload Optimizer only calculates overcommitment in plans. The calculation can be expressed as:</p> $\text{Overcommitment} = \frac{\text{Sum of memory limits for all containers}}{\text{Sum of memory capacity for all nodes}}$ |

Plan Actions

Intersight Workload Optimizer shows separate tabs for **Lift & Shift** and **Optimized** migration actions. You can download the list of actions as a CSV file.

| RESULTS OVERVIEW | | LIFT & SHIFT ACTIONS (51) | OPTIMIZED ACTIONS (97) |
|---------------------|---|---------------------------|---|
| RESIZE ^ | Resize Actions (46) | | |
| Workload Co... (46) | <input type="text" value="Type to search"/> | | |
| MOVE ^ | Workload Controller Name | Container Cluster | Risk |
| Container Pods (42) | mediation-awscost | Kubernetes-Turbonomic | VMem Limit Congestion in Container Spec media... |
| START ^ | kafka | Kubernetes-Turbonomic | VMem Limit Congestion in Container Spec kafka |
| Container Pods (6) | mediation-aws | Kubernetes-Turbonomic | VMem Limit Congestion in Container Spec media... |
| PROVISION ^ | cost | Kubernetes-Turbonomic | Underutilized VMem Limit in Container Spec cost |
| Container Pods (2) | mediation-gcpsa | Kubernetes-Turbonomic | Underutilized VMem Limit in Container Spec med... |

These tabs show the actions that you need to execute to achieve the plan results. For example, you might need to resize limits and requests for containers (via the associated Workload Controllers) to address performance issues. Or, you might need to move pods from one node to another to reduce congestion.

Smarter redistribution and workload rightsizing also drive actions, resulting in the need to provision node(s) based on application demand, or to defragment node resources to enable node suspension.

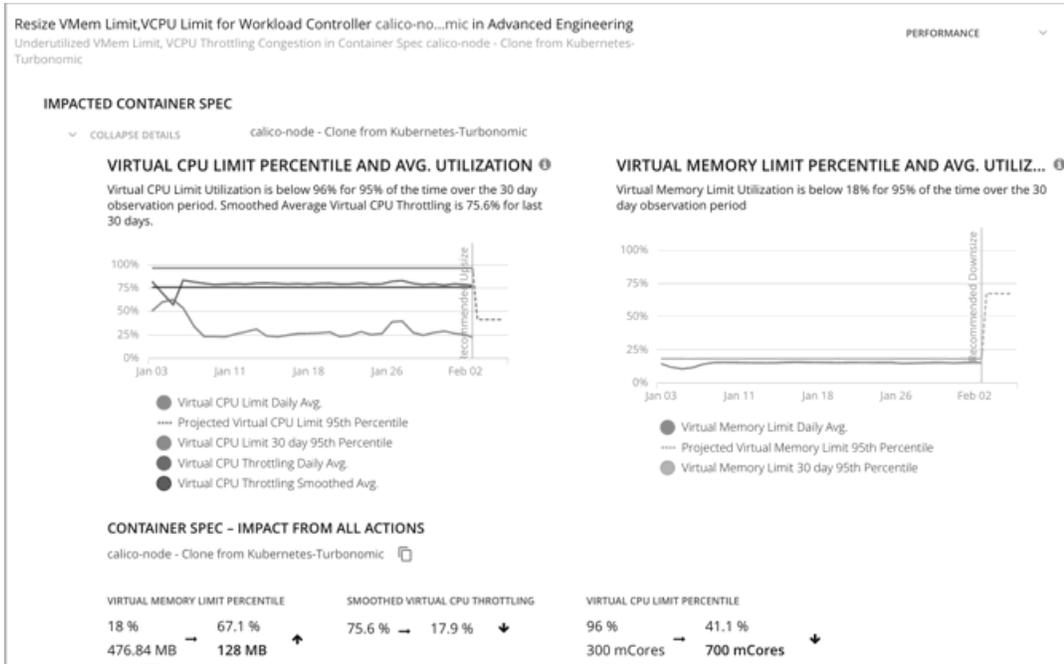
Intersight Workload Optimizer can recommend the following actions:

| Entity | Lift & Shift Actions | Optimized Actions |
|---------------------|--|--|
| Workload Controller | None | Resize container limits and requests See "Workload Controller Resize Actions" below for additional details. |
| Namespace | None | Resize quotas (if container resize actions would exceed the namespace quotas) |
| Pod | <ul style="list-style-type: none"> ■ Move ■ Provision (if node provision is required) ■ Reconfigure | <ul style="list-style-type: none"> ■ Move ■ Provision (if node provision is required) ■ Suspend (if node suspension is required) ■ Reconfigure |
| Node (VM) | <ul style="list-style-type: none"> ■ Provision ■ Move on-prem nodes | <ul style="list-style-type: none"> ■ Provision ■ Suspend ■ Move on-prem nodes |
| Cloud volume | Scale | Scale |
| On-prem storage | <ul style="list-style-type: none"> ■ Provision ■ Suspend | <ul style="list-style-type: none"> ■ Provision ■ Suspend |
| On-prem host | <ul style="list-style-type: none"> ■ Provision ■ Suspend | <ul style="list-style-type: none"> ■ Provision ■ Suspend |

Additional information:

- Workload Controller Resize Actions

When you expand an action on a Workload Controller, you will see charts that track VCPU and VMem utilization for the impacted container spec. With these charts, you can easily recognize the utilization trends that Intersight Workload Optimizer analyzed to make accurate resize decisions.



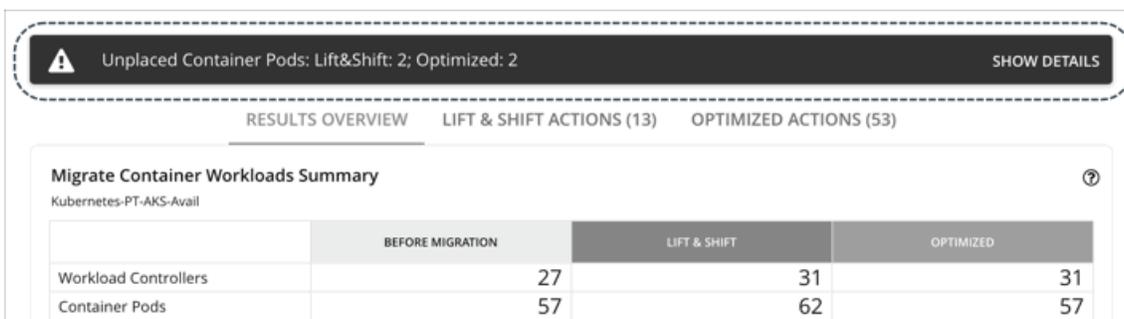
For more information about these charts, see [Utilization Charts \(on page 402\)](#).

Executing several container resize actions can be very disruptive since pods need to restart with each resize. For replicas of the container scale group(s) related to a single Workload Controller, Intersight Workload Optimizer consolidates resize actions into one *merged action* to minimize disruptions. When a merged action has been executed (via the associated Workload Controller), all resizes for all related container specifications will be changed at the same time, and pods will restart once.

■ Pod Constraints

A Migrate Container Workloads plan evaluates pod constraints on the source cluster when making placement decisions for pods. The plan enforces taints, tolerations, and `nodeSelector` specifications if there are matching constraints in the destination cluster. If these constraints cannot be achieved in the destination cluster, the plan ignores them to guarantee placement.

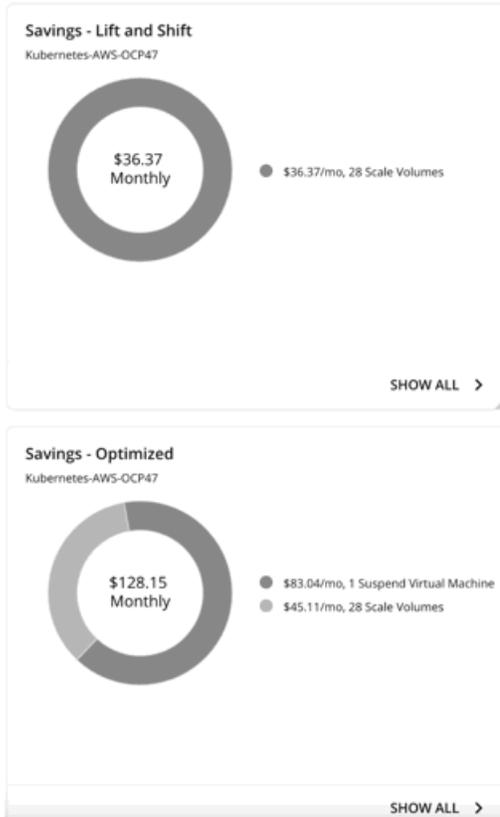
The plan always enforces affinity and anti-affinity constraints, which could result in unplaced pods in the destination cluster. If this happens, the plan generates reconfigure actions for the unplaced pods and shows a notification in the plan results.



Click **Show Details** to see the list of pods and the reasons for their non-placement. The charts in the plan results do not count these pods.

Savings

For a destination cluster in the public cloud, Intersight Workload Optimizer shows the savings you would realize if you execute the actions that the plan recommends. The results show separate charts for the **Lift and Shift** and **Optimized** scenarios, so you can compare the impact of actions on your cloud expenses.



For example, in both scenarios, the plan might recommend scaling volumes to new tiers to address performance issues. If these new tiers happen to be more cost-effective than the current tiers, then the actions are treated as cost-saving measures. For the optimized scenario, the plan might also recommend suspending certain nodes to increase infrastructure efficiency, which could introduce additional savings.

Note that application performance and infrastructure efficiency are the drivers of these actions, *not* cost optimization. Cost information is included to help you track your cloud expenses.

The charts show total monthly savings. Click **Show All** to view the actions with cost savings.

NOTE:

An empty chart indicates that no savings will be realized after you execute the recommended actions.

Investments

For a destination cluster in the public cloud, Intersight Workload Optimizer shows the costs you would incur if you execute the actions that the plan recommends. The results show separate charts for the **Optimized** and **Lift and Shift** scenarios, so you can compare the impact of actions on your cloud expenses.



For example, if some applications risk losing performance, the plan might recommend provisioning nodes to increase capacity. The charts show how node provision actions translate to an increase in expenditure. You can use cost information in the charts as you seek approval to execute these actions.

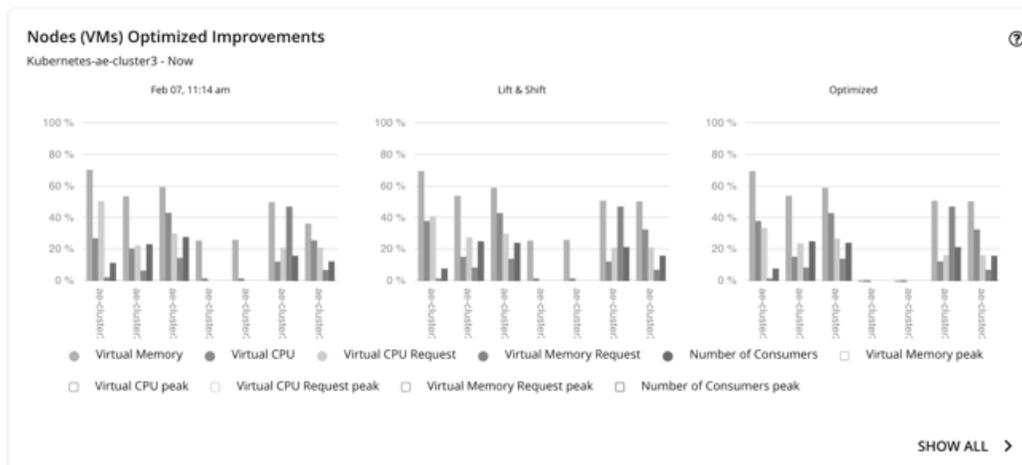
The charts show total monthly investments. Click **Show All** to view the actions that require investments.

NOTE:

An empty chart indicates that no investments are required to execute the recommended actions.

Nodes (VMs) Optimized Improvements

This chart compares the following before and after the plan:



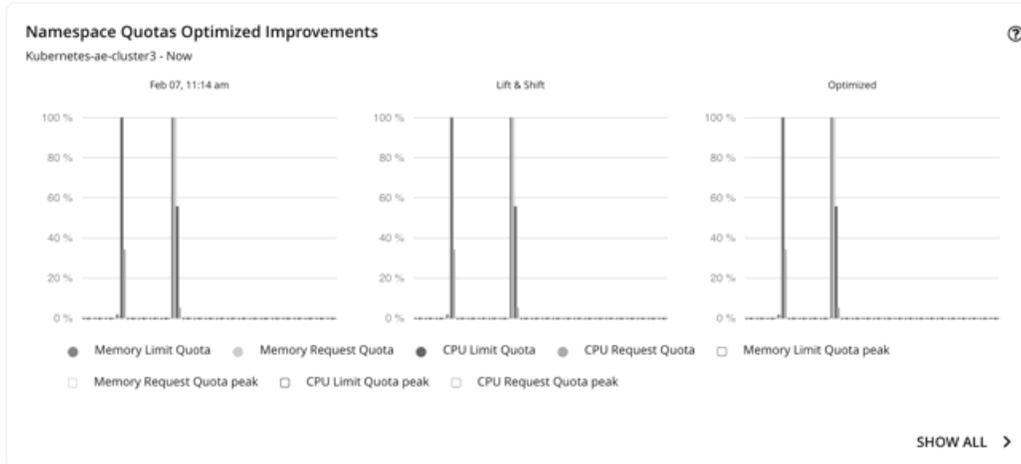
- Utilization of the following resources for all nodes:
 - vMem

- vCPU
- vMem Request
- vCPU Request
- Number of pods consuming resources against the maximum pod capacity for all the nodes

Use this chart to identify nodes that are candidates for suspension, or resources with unusual utilization. You can also use this chart to drill down to a specific resource. Click **Show All** for more details.

Namespace Quotas Optimized Improvements

This chart shows pod utilization of resource quotas defined in namespaces. Resource quotas include:



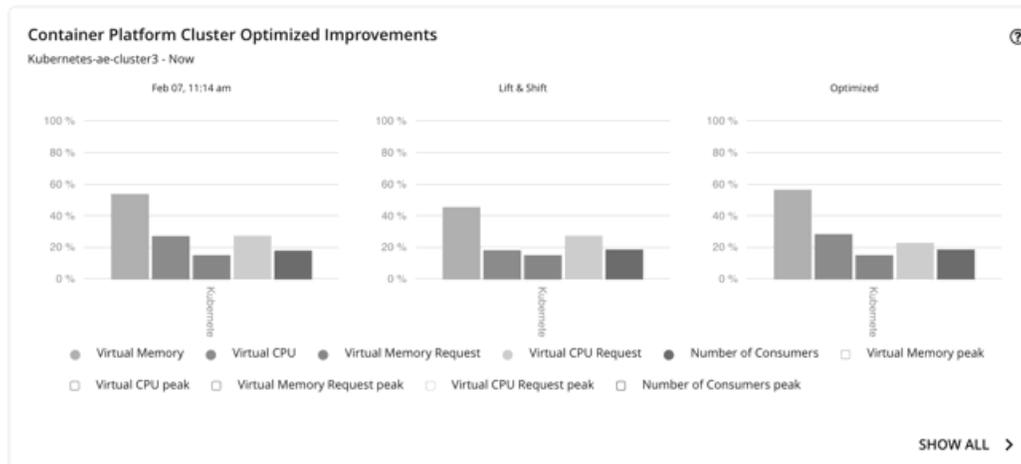
- CPU Limit Quota
- Memory Limit Quota
- CPU Request Quota
- Memory Request Quota

The chart highlights utilization improvements as a result of namespace resizing.

For namespaces without defined quotas, utilization is 0 (zero).

Container Platform Cluster Optimized Improvements

This chart shows the following, assuming you execute all actions in the plan:



- Changes to the utilization of cluster resources
- Overcommitment values

Optimized Improvements for Hosts and Storage

Use these charts if you ran the plan on an on-prem cluster. These charts show how the utilization of resources would change assuming you accept all of the actions that the plan recommends.

Downloading Plan Results

You can download the plan results shown in individual charts. Click the **Show All** button for a chart, and then the download button at the top-right section of the Details page.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Optimize Cloud Plan

Run the Optimize Cloud plan to see how you can maximize savings while still assuring performance for your applications and workloads. This plan identifies ways to optimize your costs by choosing the best templates (most adequate compute resources), regions, accounts, or resource groups to host your workloads. The plan also identifies workloads that can change over to discounted pricing, and it compares your current costs to the costs you would get after executing the plan recommendations.


Optimize Cloud 13

All Cloud Providers
RUN AGAIN

CONFIGURATION

Discount Settings

- Purchase Ris ENABLED ×
- AWS: Offering Class STANDARD ×
- AWS: Term 3 YEARS ×
- AWS: Payment ALL UPFRONT ×
- Azure: Term 1 YEAR ×
- Discount Inventory 28 OUT OF 28 ACTIVE ×

Virtual Machine Action Settings

- Scale for Virtual Machines ENABLED

RESULTS OVERVIEW PLAN ACTIONS (381)

Cloud Cost Comparison

All Cloud Providers

| | CURRENT | OPTIMIZED | DIFFERENCE | % |
|---|------------------------|------------------------|------------------------|-----------------|
| Workloads with performance risks | 29 Out Of 855 | 0 Out Of 855 | 29 | - |
| Workloads with efficiency opportunities | 125 Out Of 855 | 0 Out Of 855 | 125 | - |
| Workloads out of compliance | 1 Out Of 855 | 0 Out Of 855 | 1 | - |
| RI Coverage | 23 % | 53 % | | ▲ 130.4 % |
| RI Utilization | 40 % | 79 % | | ▲ 97.5 % |
| CUD vCPU Coverage | 8 % | 8 % | | 0 % |
| CUD vCPU Utilization | 45 % | 50 % | | ▲ 11.1 % |
| CUD Mem Coverage | 4 % | 5 % | | ▲ 25 % |
| CUD Mem Utilization | 36 % | 42 % | | ▲ 16.7 % |
| On-Demand Compute Cost | \$19,955.00 /mo | \$10,060.00 /mo | -\$9,895.00 /mo | ▼ 49.6 % |
| Reserved Compute Cost | \$2,349.00 /mo | \$5,945.00 /mo | \$3,596.00 /mo | ▲ 153.1 % |
| On-Demand Database Cost | \$11,641.00 /mo | \$11,621.00 /mo | -\$20.00 /mo | ▼ 0.2 % |
| Storage Cost | \$10,476.00 /mo | \$8,597.00 /mo | -\$1,879.00 /mo | ▼ 17.9 % |
| Total Cost | \$44,421.00 /mo | \$36,223.00 /mo | -\$8,198.00 /mo | ▼ 18.5 % |

Cisco Intersight Workload Optimizer Target Configuration and User Guide

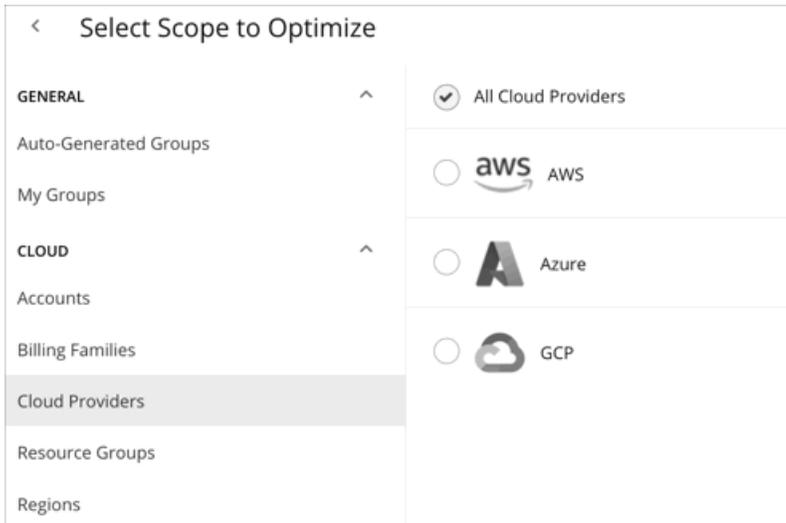
449

Configuring an Optimize Cloud Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

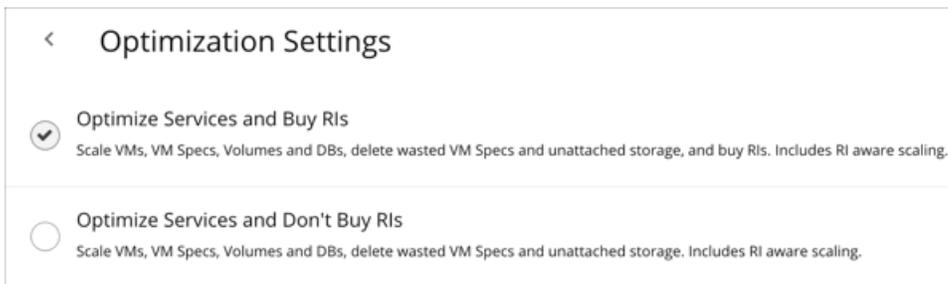
1. Scope

You can scope by:



- **Accounts**
Choose AWS accounts, Azure subscriptions, or Google Cloud projects as the plan's scope. Be aware that the plan will not recommend discount purchases if you scope to this level. To optimize discount purchases for a limited scope, choose a Billing Family.
- **Billing Families**
Include discount purchases in the planning for a scope that is limited to a single billing family. The plan calculates discount purchases through the billing family's management account. Discount purchases are currently supported for AWS and Azure, but not Google Cloud.
- **Cloud Providers**
See how you can optimize all your AWS, Azure, and Google Cloud workloads.
- **Resource Groups**
Intersight Workload Optimizer discovers Azure resource groups. You can select one or more resource groups for the plan scope.
- **Regions**
Focus the plan on a provider's region.

2. Optimization Settings



Choose from the given optimization options. Note that if you set a plan's scope to a resource group, Intersight Workload Optimizer will optimize services without recommending discount purchases.

If your goal is to purchase discounts for VMs at their current sizes, use the Buy VM Reservations plan type. For details, see [Buy VM Reservations Plan \(on page 455\)](#).

NOTE:

If you turn on the **Disable All Actions** setting in the global default policy and then run an Optimize Cloud plan with VM scaling and discount purchases enabled, the plan results show inaccurate discount recommendations.

Turn off **Disable All Actions** to resolve this issue. Be aware that after you turn off this setting, it will take Intersight Workload Optimizer a week to reflect accurate results in Optimize Cloud plans.

3. Discount Settings

<
Discount Settings

Purchase RIs

AWS PROFILE

OFFERING CLASS

Standard

Convertible

TERM

1 Year

3 Years

PAYMENT

All Upfront

Partial Upfront

No Upfront

AZURE PROFILE

TERM

1 Year

3 Years

Discount Inventory

This plan uses 29 out of 29 available discounts ✎ EDIT

- Currently, **Purchase RIs** only applies to AWS and Azure. For **AWS/Azure Profile**, the settings that you have set up for real-time analysis are selected by default. You can change the settings to see how they affect costs.
 - Offering Class

For AWS environments, choose the offering class that corresponds to the RI types that you typically use in your environment.
 - Term

For AWS and Azure environments, choose the payment terms you contract for your discounts. TERM can be one of **1 Year** or **3 Year**. Typically, longer term payment plans cost less per year.
 - Payment

The payment option that you prefer for your AWS RIs:

 - All Upfront – You make full payment at the start of the RI term.
 - Partial Upfront – You make a portion of the payment at the start of the term, with the remain cost paid at an hourly rate.

- No Upfront – You pay for the RIs at an hourly rate, for the duration of the term.
- For **Discount Inventory**, the discounts for the current scope are selected by default. Click **Edit** to make changes.

Working With Optimize Cloud Plan Results

After the Optimize Cloud plan runs, you can view the results to see how you can maximize savings or make other improvements to your cloud environment.

The plan results:

- Compare current to optimized costs, including on-demand compute, discounted compute, on-demand database, and storage costs
- Compare current and optimized breakdowns of workload tiers
- Compare breakdowns of storage tiers
- Project the discount coverage (how many workloads are covered by discounted pricing) and utilization (percentage of discounts that are active)
- Show the cost benefit of moving workloads from on-demand to discounted pricing

Viewing the Results

The screenshot shows the 'Optimize Cloud 13' interface. On the left is a 'CONFIGURATION' sidebar with settings for Discount Settings (Purchase RIs: ENABLED, AWS Offering Class: STANDARD, AWS Term: 3 YEARS, AWS Payment: ALL UPFRONT, Azure Term: 1 YEAR, Discount Inventory: 28 OUT OF 28 ACTIVE) and Virtual Machine Action Settings (Scale for Virtual Machines: ENABLED). The main area is titled 'RESULTS OVERVIEW' and 'PLAN ACTIONS (381)'. It features a 'Cloud Cost Comparison' table for 'All Cloud Providers'.

| | CURRENT | OPTIMIZED | DIFFERENCE | % |
|---|------------------------|------------------------|------------------------|-----------------|
| Workloads with performance risks | 29 Out Of 855 | 0 Out Of 855 | 29 | - |
| Workloads with efficiency opportunities | 125 Out Of 855 | 0 Out Of 855 | 125 | - |
| Workloads out of compliance | 1 Out Of 855 | 0 Out Of 855 | 1 | - |
| RI Coverage | 23 % | 53 % | | ▲ 130.4 % |
| RI Utilization | 40 % | 79 % | | ▲ 97.5 % |
| CUD vCPU Coverage | 8 % | 8 % | | 0 % |
| CUD vCPU Utilization | 45 % | 50 % | | ▲ 11.1 % |
| CUD Mem Coverage | 4 % | 5 % | | ▲ 25 % |
| CUD Mem Utilization | 36 % | 42 % | | ▲ 16.7 % |
| On-Demand Compute Cost | \$19,955.00 /mo | \$10,060.00 /mo | -\$9,895.00 /mo | ▼ 49.6 % |
| Reserved Compute Cost | \$2,349.00 /mo | \$5,945.00 /mo | \$3,596.00 /mo | ▲ 153.1 % |
| On-Demand Database Cost | \$11,641.00 /mo | \$11,621.00 /mo | -\$20.00 /mo | ▼ 0.2 % |
| Storage Cost | \$10,476.00 /mo | \$8,597.00 /mo | -\$1,879.00 /mo | ▼ 17.9 % |
| Total Cost | \$44,421.00 /mo | \$36,223.00 /mo | -\$8,198.00 /mo | ▼ 18.5 % |

The plan results include the following charts:

■ Cloud Cost Comparison

This chart highlights any difference in cost as a result of optimization. For example, undersized VMs risk losing performance and should therefore scale up. This could contribute to an increase in cost. On the other hand, oversized VMs can scale down to less expensive instances, so cost should go down. The values under the % column indicate the percentage of VMs that are affected by optimization cost calculations.

Intersight Workload Optimizer can also recommend that you purchase discounts to reduce costs. The analysis looks at workload history to identify workloads that can move from on-demand to discounted pricing. This considers the count of workloads in a family, plus their hours of active-state condition, to arrive at the discounts you should purchase. Since discounted costs are incurred at the account level, the Cloud Cost Comparison chart will present discounted costs or charges when you scope to an account or group of accounts (including a billing family).

For AWS clouds, Intersight Workload Optimizer can get the information it needs to display license costs for database instances. For Azure clouds, Intersight Workload Optimizer does not display database license costs because Azure does not make that information available.

■ Workload Mapping

This chart shows the types of tiers you currently use, compared to the tiers the plan recommends, including how many of each type, plus the costs for each.

To see a detailed breakdown of the template costs, click **SHOW CHANGES** at the bottom of the chart.

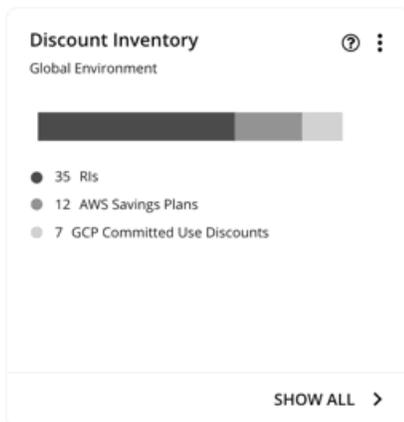
■ Volume Tier Summary

This chart shows the current distribution of volumes that support your workloads, and the optimized distribution if you execute the actions that the plan recommends.

The difference in the result reflects the number of unattached volumes. To see a list of unattached volumes, click **Show changes** at the bottom of the chart.

■ Discount Inventory

This chart lists the cloud provider discounts discovered in your environment. For a tabular listing, click **Show All** at the bottom of the chart. In the tabular listing, you can see if a discount expired before the specified purchase date.



■ Recommended RI Purchases

Intersight Workload Optimizer can recommend purchasing instance types at a discounted rate to help you increase the percentage of VMs covered by discounted pricing and reduce on-demand costs. This chart shows your pending purchases. Download the list of purchases and then send it your cloud provider or representative to initiate the purchase process.

NOTE:

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Currently, Intersight Workload Optimizer can recommend purchasing AWS EC2 RIs and Azure reservations.

Recommended RI Purchases ?

- 5 t3a.nano
- 3 t2.nano
- 3 m5.large
- 2 m5a.large
- 2 t3.nano
- 5 Other Instance Types

[SHOW ALL >](#)

Click **Show All** to see a table with details for each discount.

The table shows the properties, up-front cost, and break-even period for each discount. The break-even period is the time at which savings will exceed the up-front cost, rounded to the month. The Cost Impact column indicates the monthly savings you would realize when you buy a specific discount.

When you choose one or more check boxes, the total count, up-front cost, and savings appear at the top.

Viewing Plan Actions

Click the **Plan Actions** tab on top of the page to view a list of actions that you need to execute to achieve the plan results.

| | | RESULTS OVERVIEW | | PLAN ACTIONS (274) | | | | | | | |
|---|--|--|---------------|--------------------------------|----------------------------------|-------------------|-----------------------|--------------------|-----------------|---------------|---------|
| DELETE ^ Volumes (131) | | Scale Actions (28) | | TOTAL SAVINGS \$3,111.76/mo | TOTAL INVESTMENTS \$242.80/mo | | | | | | ⚙️ ⬇️ |
| SCALE ^ Volumes (79) | | <input type="text" value="Type to search"/> ⚙️ ADD FILTER | | | | | | | | | |
| Virtual Machines (28) | | Virtual Machine Name | Instance Type | Discount Coverage | On-Demand Cost | New Instance Type | New Discount Coverage | New On-Demand Cost | Action Category | Cost Impact | Action |
| Database Servers (7) | | eks-cluster-eks-cls | m5a.4xlarge | 0% | \$0.684/h | r5a.2xlarge | 75% | \$0.112/h | SAVINGS | ↓ \$417.49/mo | DETAILS |
| BUY ^ Reserved Instance (15) | | ocp47demo-2v5jc | m5a.4xlarge | 0% | \$0.684/h | r5a.2xlarge | 75% | \$0.112/h | SAVINGS | ↓ \$417.49/mo | DETAILS |
| | | cn-latest-lzdjn-wor | m6i.2xlarge | 0% | \$0.382/h | r5a.xlarge | 100% | \$0.00/h | SAVINGS | ↓ \$278.81/mo | DETAILS |
| | | cn-latest-lzdjn-wor | m6i.2xlarge | 0% | \$0.382/h | r5.xlarge | 50% | \$0.125/h | SAVINGS | ↓ \$187.33/mo | DETAILS |
| | | pt-ocp-apm-gr2qp | r5a.xlarge | 0% | \$0.247/h | r5.xlarge | 100% | \$0.00/h | PERFORMANCE | ↓ \$180.07/mo | DETAILS |

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

Optimize Cloud 56
✎

SCOPE

Prod

⬇️
⬆️

Actions
RI Profile

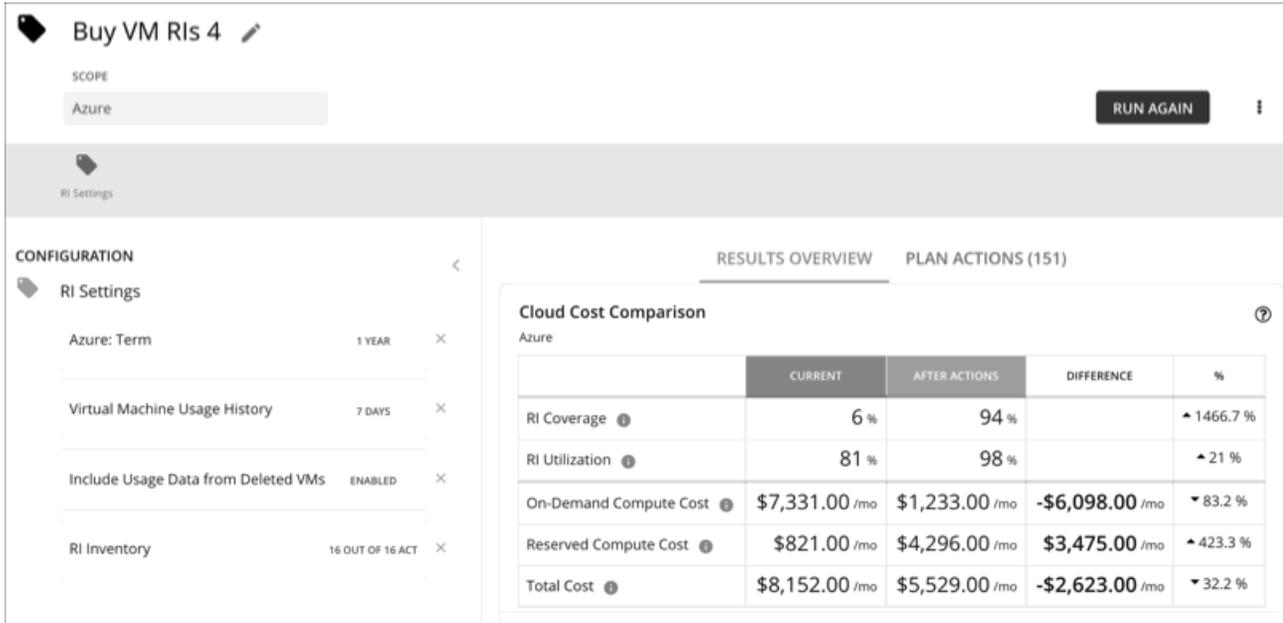
- **Actions**
Use this to enable or disable automatic Scale actions for the virtual machines in the plan.
- **Discount Settings**
See [Discount Settings \(on page 451\)](#).

NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Buy VM Reservations Plan



The screenshot shows the 'Buy VM RIs 4' plan configuration page. The scope is set to 'Azure'. A 'RUN AGAIN' button is visible in the top right. The left sidebar shows 'RI Settings' with options for 'Azure: Term' (1 YEAR), 'Virtual Machine Usage History' (7 DAYS), 'Include Usage Data from Deleted VMs' (ENABLED), and 'RI Inventory' (16 OUT OF 16 ACT). The main area displays a 'Cloud Cost Comparison' table for Azure.

| | CURRENT | AFTER ACTIONS | DIFFERENCE | % |
|------------------------|-----------------------|-----------------------|------------------------|-----------------|
| RI Coverage | 6 % | 94 % | | ▲ 1466.7 % |
| RI Utilization | 81 % | 98 % | | ▲ 21 % |
| On-Demand Compute Cost | \$7,331.00 /mo | \$1,233.00 /mo | -\$6,098.00 /mo | ▼ 83.2 % |
| Reserved Compute Cost | \$821.00 /mo | \$4,296.00 /mo | \$3,475.00 /mo | ▲ 423.3 % |
| Total Cost | \$8,152.00 /mo | \$5,529.00 /mo | -\$2,623.00 /mo | ▼ 32.2 % |

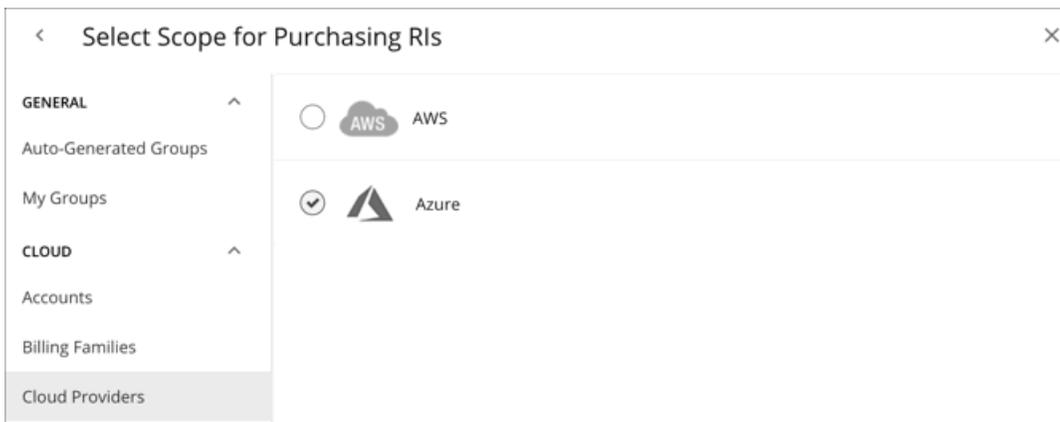
Run the Buy VM Reservations plan to see discount purchase opportunities that can significantly reduce on-demand costs for your cloud VMs. When calculating purchases, Intersight Workload Optimizer evaluates all purchasing options for your selected scope and usage data for the VMs in that scope. It then compares your current costs to the costs you would get after executing the plan recommendations.

Currently, Intersight Workload Optimizer can recommend purchasing AWS EC2 RIs and Azure reservations.

Configuring a Buy VM Reservations Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

1. Scope

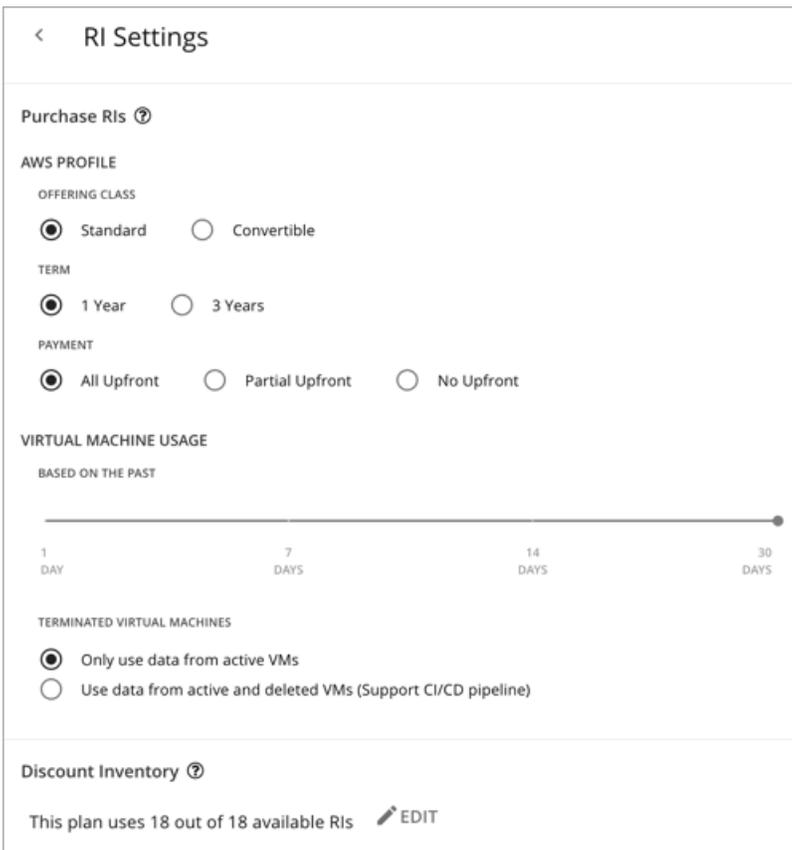


The screenshot shows the 'Select Scope for Purchasing RIs' dialog box. It has a left sidebar with categories: 'GENERAL', 'Auto-Generated Groups', 'My Groups', 'CLOUD', 'Accounts', 'Billing Families', and 'Cloud Providers'. The 'CLOUD' section is expanded, showing 'AWS' and 'Azure' options. 'Azure' is selected with a checkmark.

You can scope by:

- **Accounts**
Choose AWS accounts or Azure subscriptions for the plan's scope.
- **Billing Families**
Include discount purchases for a billing family. The plan calculates discount purchases through the billing family's master account.
- **Cloud Providers**
See purchase opportunities for your AWS or Azure environment.
- **Regions**
Focus the plan on a cloud provider's region.

2. RI Settings



Purchase RIs

Allow the plan to buy discounts based on the following configurations:

- **Profile**

The settings that you have set up for real-time analysis are selected by default. You can change the settings to see how they affect costs.

- **Offering Class**

For AWS environments, choose the offering class that corresponds to the RI types that you typically use in your environment.

- **Term**

For AWS and Azure environments, choose the payment terms you contract for your discounts. TERM can be one of **1 Year** or **3 Year**. Typically, longer term payment plans cost less per year.

- **Payment**

The payment option that you prefer for your AWS RIs:

- All Upfront – You make full payment at the start of the RI term.
- Partial Upfront – You make a portion of the payment at the start of the term, with the remain cost paid at an hourly rate.
- No Upfront – You pay for the RIs at an hourly rate, for the duration of the term.

■ Virtual Machine Usage

Specify the time frame you want the plan to use when it calculates your discount purchases.

■ Terminated Virtual Machines

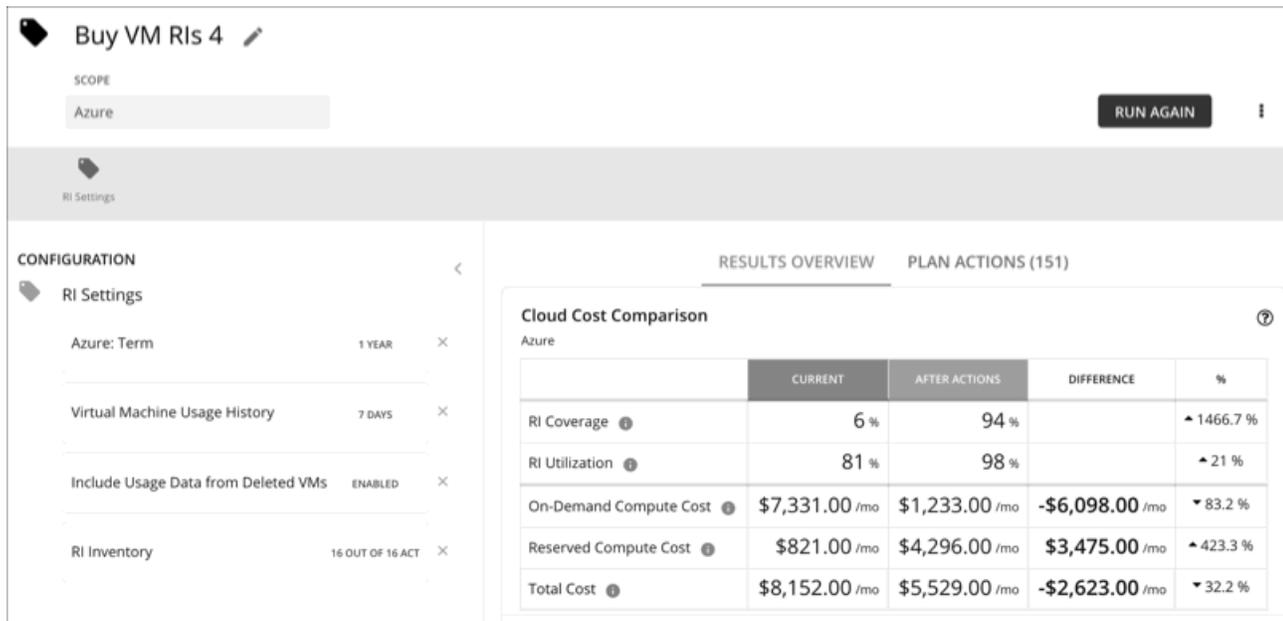
- **Only use data from active VMs** – Select this option if you terminate your VMs permanently.
- **Use data from active and deleted VMs (Support CI/CD pipeline)** – Select this option if you want to use data from a CI/CD pipeline that regularly deploys and terminates VMs.

Discount Inventory

Select your discount inventory for the plan. You can use the default selection or any of the available discounts for your scope.

Working With Buy VM Reservations Plan Results

After the Buy VM Reservations runs, you can view the results to see discount and optimization opportunities for your cloud environment.



Viewing the Results

The plan results include the following charts:

■ Cloud Cost Comparison

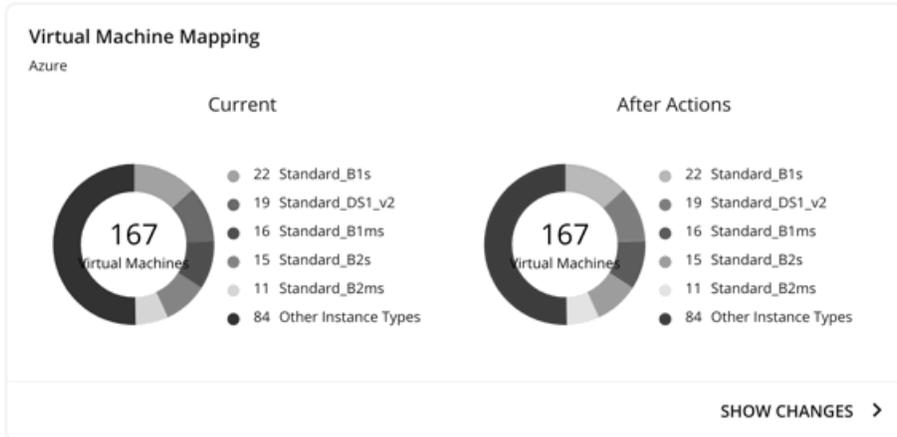
This chart highlights changes to your existing discount coverage and utilization if you execute all the actions that the plan recommends. Actions include increasing coverage or purchasing additional instance types at a discounted rate. Your cloud provider will adjust discount allocations when the actions have completed.

- Analysis evaluates ways to increase your current discount coverage so you can take full advantage of discounted pricing.
- The plan can recommend purchase actions to reduce your costs further. The analysis looks at historical VM usage and uptime to arrive at the number of instance types you should purchase.

You can compare current and after-action costs, including on-demand compute, discounted compute, and total costs. Purchase actions increase your discounted compute cost, but can lower your on-demand compute cost significantly as discount coverage increases. The end result is a reduction to your total cost.

■ **Virtual Machine Mapping**

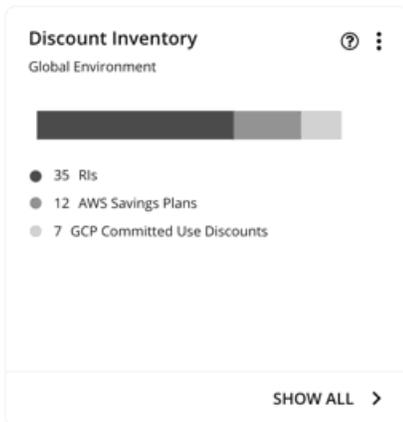
This chart shows the instance types for the VMs included in the plan.



Click **Show Changes** to see details for each VM with discount coverage changes. The table maps VMs to instance types, and shows how changes in discount coverage can reduce on-demand cost.

■ **Discount Inventory**

This chart lists the cloud provider discounts discovered in your environment. For a tabular listing, click **Show All** at the bottom of the chart. In the tabular listing, you can see if a discount expired before the specified purchase date.



■ **Recommended RI Purchases**

Intersight Workload Optimizer can recommend purchasing instance types at a discounted rate to help you increase the percentage of VMs covered by discounted pricing and reduce on-demand costs. This chart shows your pending purchases. Download the list of purchases and then send it your cloud provider or representative to initiate the purchase process.

Recommended RI Purchases ?

- 5 t3a.nano
- 3 t2.nano
- 3 m5.large
- 2 m5a.large
- 2 t3.nano
- 5 Other Instance Types

[SHOW ALL >](#)

Click **Show All** to see a table with details for each discount.

The table shows the properties, up-front cost, and break-even period for each discount. The break-even period is the time at which savings will exceed the up-front cost, rounded to the month. The Cost Impact column indicates the monthly savings you would realize when you buy a specific discount.

When you choose one or more check boxes, the total count, up-front cost, and savings appear at the top.

Viewing Plan Actions

Click the **Plan Actions** tab on top of the page to view a list of actions that you need to execute to achieve the plan results.

| | | RESULTS OVERVIEW | | PLAN ACTIONS (274) | | | | | | | |
|---------------|---|---------------------------|---------------|--------------------|-------------------|-------------------|-----------------------|--------------------|-----------------|---------------|---------|
| DELETE | ^ | Scale Actions (28) | | TOTAL SAVINGS | TOTAL INVESTMENTS | | | | | | |
| | | | | \$3,111.76/mo | \$242.80/mo | | | | | | |
| | | Type to search | | ADD FILTER | | | | | | | |
| SCALE | ^ | Virtual Machine Name | Instance Type | Discount Coverage | On-Demand Cost | New Instance Type | New Discount Coverage | New On-Demand Cost | Action Category | Cost Impact | Action |
| | | eks-cluster-eks-cl | m5a.4xlarge | 0% | \$0.684/h | r5a.2xlarge | 75% | \$0.112/h | SAVINGS | ↓ \$417.49/mo | DETAILS |
| | | ocp47demo-2v5jc | m5a.4xlarge | 0% | \$0.684/h | r5a.2xlarge | 75% | \$0.112/h | SAVINGS | ↓ \$417.49/mo | DETAILS |
| | | cn-latest-lzdjn-wor | m6i.2xlarge | 0% | \$0.382/h | r5a.xlarge | 100% | \$0.00/h | SAVINGS | ↓ \$278.81/mo | DETAILS |
| | | cn-latest-lzdjn-wor | m6i.2xlarge | 0% | \$0.382/h | r5.xlarge | 50% | \$0.125/h | SAVINGS | ↓ \$187.33/mo | DETAILS |
| | | pt-ocp-apm-gr2qp | r5a.xlarge | 0% | \$0.247/h | r5.xlarge | 100% | \$0.00/h | PERFORMANCE | ↓ \$180.07/mo | DETAILS |

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

- **RI Settings**

Update your purchase settings to see how they impact results. For example, you can configure a longer timeframe so that the plan can include additional VM usage data in its analysis. For details, see [Purchase RIs \(on page 456\)](#).

- **Discount Inventory**

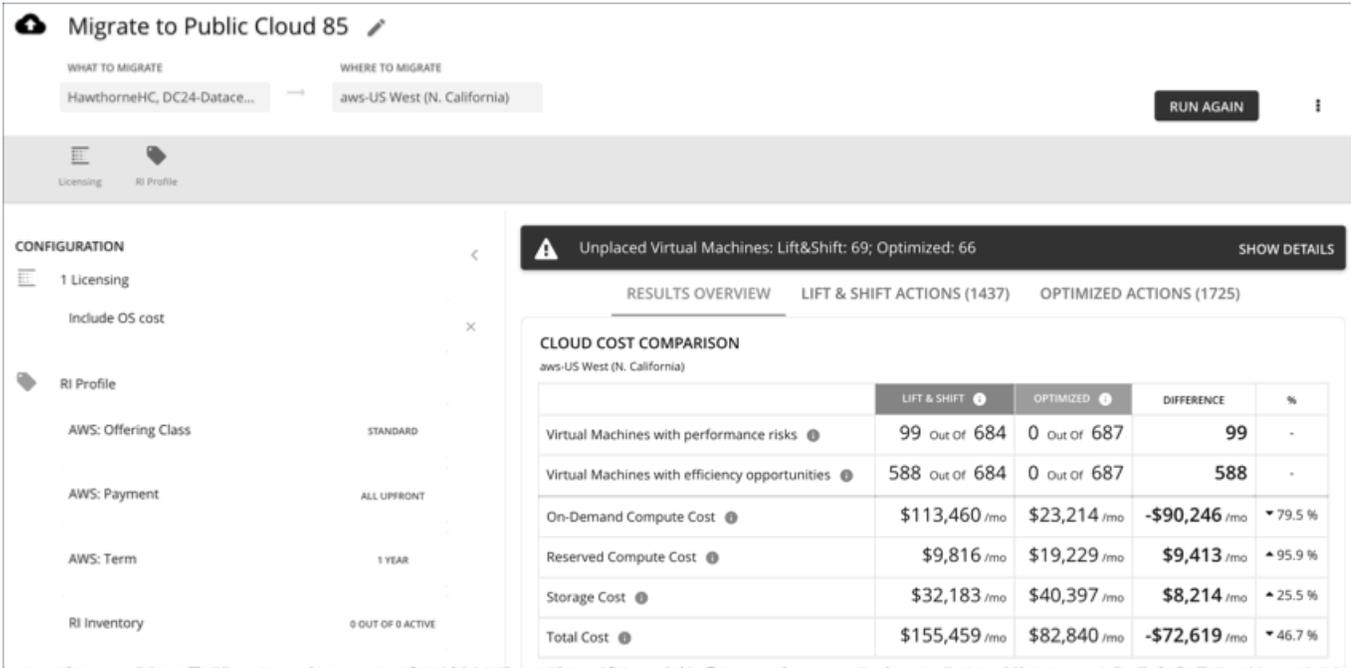
Use the default selection or any of the available discounts for your scope.

NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Migrate to Cloud Plan



The screenshot shows the 'Migrate to Public Cloud 85' interface. At the top, it displays 'WHAT TO MIGRATE' (HawthorneHC, DC24-Data...) and 'WHERE TO MIGRATE' (aws-US West (N. California)). A 'RUN AGAIN' button is visible. Below this is a 'CONFIGURATION' sidebar with options like '1 Licensing', 'Include OS cost', 'RI Profile', 'AWS: Offering Class' (STANDARD), 'AWS: Payment' (ALL UPFRONT), 'AWS: Term' (1 YEAR), and 'RI Inventory' (0 OUT OF 9 ACTIVE). A warning banner indicates 'Unplaced Virtual Machines: Lift&Shift: 69; Optimized: 66'. The main area shows 'RESULTS OVERVIEW' with tabs for 'LIFT & SHIFT ACTIONS (1437)' and 'OPTIMIZED ACTIONS (1725)'. A 'CLOUD COST COMPARISON' table is displayed for 'aws-US West (N. California)'.

| | LIFT & SHIFT | OPTIMIZED | DIFFERENCE | % |
|--|----------------------|---------------------|----------------------|-----------------|
| Virtual Machines with performance risks | 99 Out Of 684 | 0 Out Of 687 | 99 | - |
| Virtual Machines with efficiency opportunities | 588 Out Of 684 | 0 Out Of 687 | 588 | - |
| On-Demand Compute Cost | \$113,460 /mo | \$23,214 /mo | -\$90,246 /mo | ▼ 79.5 % |
| Reserved Compute Cost | \$9,816 /mo | \$19,229 /mo | \$9,413 /mo | ▲ 95.9 % |
| Storage Cost | \$32,183 /mo | \$40,397 /mo | \$8,214 /mo | ▲ 25.5 % |
| Total Cost | \$155,459 /mo | \$82,840 /mo | -\$72,619 /mo | ▼ 46.7 % |

A Migrate to Cloud plan simulates migration of on-prem VMs to the cloud, or migration of VMs from one cloud provider to another. This plan focuses on optimizing performance and costs by choosing the most suitable cloud resources for your VMs and the volumes they use. To further optimize your costs, the plan can recommend moving workloads from on-demand to discounted pricing, and purchasing more discounts.

The plan calculates costs according to the billing and price adjustments that you have negotiated with your cloud provider. Costs include compute, service (such as IP services), and license costs. The plan also calculates discount purchases for VMs that can benefit from discounted pricing.

NOTE:

If your instance of Intersight Workload Optimizer is inoperative for a period of time, that can affect the cost calculations. To calculate costs for a VM that it will migrate to the cloud, Intersight Workload Optimizer considers the VM's history. For example, if the VM has been stable for 16 of the last 21 days, then Intersight Workload Optimizer will plan for that VM to use a discount. In this way, the plan calculates the best cost for the migration. However, if Intersight Workload Optimizer is inoperative for any time, that can impact the historical data such that the plan will *not* recognize a VM as stable, even though it is.

Points to consider:

- AWS includes EC2 Spot Instances that offer steep discounts. A plan that migrates from AWS to Azure will not migrate VMs that run on Spot Instances.
- Do not use this plan type to migrate within the same cloud provider (for example, moving VMs from one Azure subscription to another) as a way to test the effect on pricing. The results from such a plan would not be reliable.
- For migrations within your on-prem environment, use the *Virtual Machine Migration* plan type.
- Before migrating, consider turning on a setting in the default global policy that enables metrics collection for on-prem volumes attached to VMs. This allows Intersight Workload Optimizer to make more accurate placement decisions for the VMs and volumes you are migrating. For details, see [Enable Analysis of On-prem Volumes \(on page 576\)](#).

The plan results show:

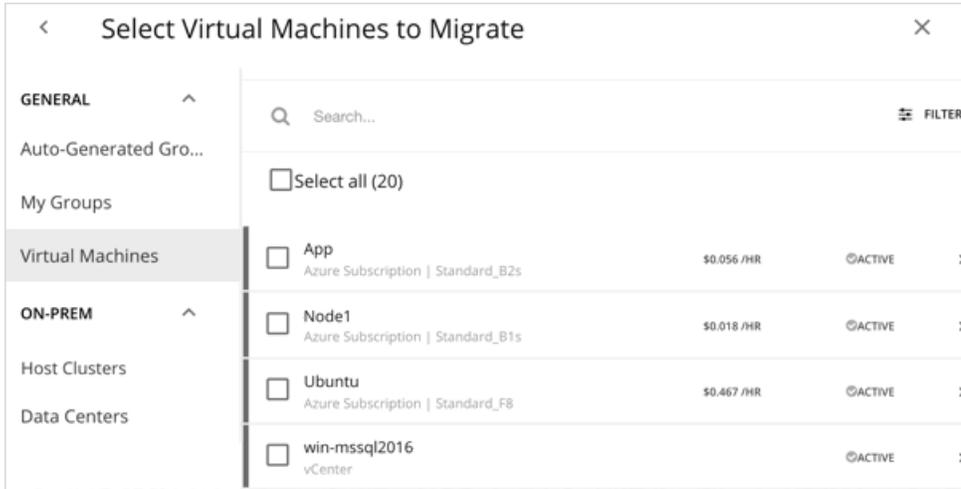
- Projected costs
- Actions to execute your migration and optimize costs and performance
- Optimal cloud instances to use, combining efficient purchase of resources with assured application performance
- The cost benefit of moving workloads from on-demand to discounted pricing
- Discounts you should purchase

Configuring a Migrate to Cloud Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

1. Scope

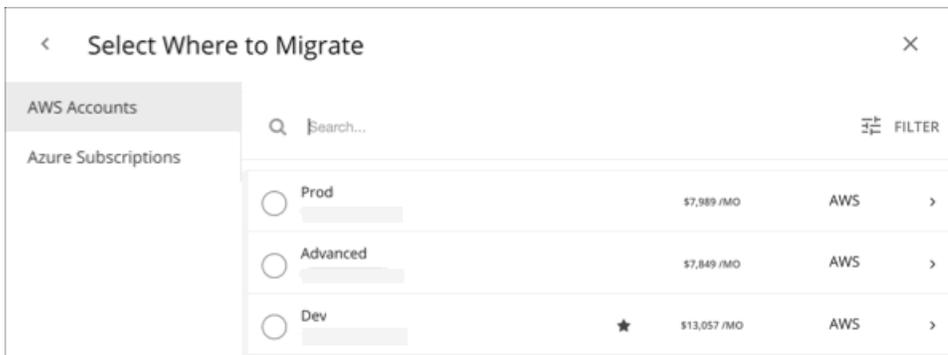
Select the VMs that you want to migrate. You can select VM groups and/or individual VMs.



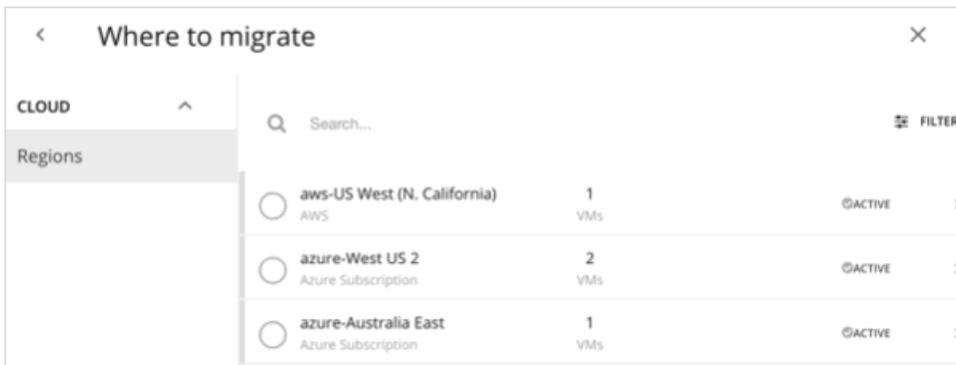
If you select an Auto Scaling Group, Intersight Workload Optimizer simulates migrating the VMs individually, and not as a group.

2. Where to Migrate

Choose a billing account (AWS account or Azure subscription).



Choose a region. Intersight Workload Optimizer shows all the regions that you can access from your target cloud accounts.

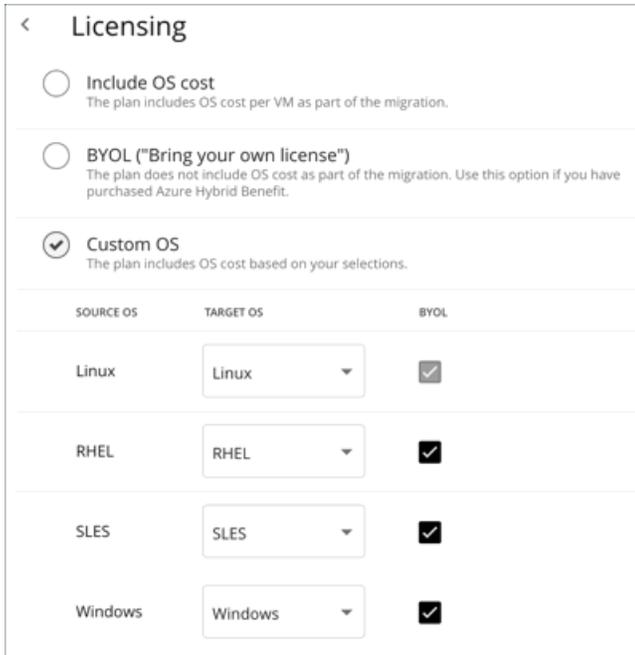


By default, Intersight Workload Optimizer considers all instance types in the selected region when making placement decisions for the scoped VMs and the volumes they use. However, you may have set up constraints in policies that limit migration to

certain instance types. If there are VMs and volumes in your scope that are affected by those policies, Intersight Workload Optimizer will only consider the instance types defined in the policies.

3. Licensing (OS Migration Profile)

Select an OS Profile for this migration.



| SOURCE OS | TARGET OS | BYOL |
|-----------|-----------|-------------------------------------|
| Linux | Linux | <input checked="" type="checkbox"/> |
| RHEL | RHEL | <input checked="" type="checkbox"/> |
| SLES | SLES | <input checked="" type="checkbox"/> |
| Windows | Windows | <input checked="" type="checkbox"/> |

On the cloud, instances usually include an OS platform to run processes on the VM. As you migrate VMs to the cloud, you can specify the OS you prefer to run. You can keep the same OS that the original VM has, or map it to a different OS.

- Include OS cost

As Intersight Workload Optimizer calculates placement for the migrated workloads, it will include costs for instances that provide the same OS that the VM already has.

- BYOL (Bring your own license)

This is the same as the **Include OS cost** option, except the plan does not include OS licensing costs in any of the cost calculations for on-cloud placement.

- Custom OS

For each of the listed OS types, map the migrated VM to the OS you choose. The OS types are:

- Linux – Any open source distribution of Linux. For the migration, Intersight Workload Optimizer will choose instances that provide the Linux platform that the cloud service provider delivers as a free platform. Note that this is always BYOL, because it assumes a free OS license.
- RHEL (Red Hat Enterprise Linux)
- SLES (SUSE Linux Enterprise Server)
- Windows

If you enable **BYOL** for RHEL, SLES, or Windows, Intersight Workload Optimizer assumes that you are paying for the OS license, and will not include the license cost in the plan results. If you do not enable **BYOL**, Intersight Workload Optimizer gets the license cost from the service provider and includes that cost in the plan results.

4. Reserved Instances Settings

<
Discount Settings

Purchase RIs

AWS PROFILE

OFFERING CLASS

Standard

Convertible

TERM

1 Year

3 Years

PAYMENT

All Upfront

Partial Upfront

No Upfront

AZURE PROFILE

TERM

1 Year

3 Years

Discount Inventory

This plan uses 29 out of 29 available discounts ✎ EDIT

- Currently, **Purchase RIs** only applies to AWS and Azure. For **AWS/Azure Profile**, the settings that you have set up for real-time analysis are selected by default. You can change the settings to see how they affect costs.
 - Offering Class

For AWS environments, choose the offering class that corresponds to the RI types that you typically use in your environment.
 - Term

For AWS and Azure environments, choose the payment terms you contract for your discounts. TERM can be one of **1 Year** or **3 Year**. Typically, longer term payment plans cost less per year.
 - Payment

The payment option that you prefer for your AWS RIs:

 - All Upfront – You make full payment at the start of the RI term.
 - Partial Upfront – You make a portion of the payment at the start of the term, with the remain cost paid at an hourly rate.
 - No Upfront – You pay for the RIs at an hourly rate, for the duration of the term.
- For **Discount Inventory**, the discounts for the current scope are selected by default. Click **Edit** to make changes.

Working With Migrate to Cloud Plan Results

The Migrate to Cloud plan results show the cloud resources and costs for the VMs you plan to migrate, and the actions required for migration.

The screenshot shows the 'Migrate to Public Cloud 85' interface. At the top, it displays 'WHAT TO MIGRATE' (HawthorneHC, DC24-Data...) and 'WHERE TO MIGRATE' (aws-US West (N. California)). A 'RUN AGAIN' button is visible. Below this is a 'CONFIGURATION' sidebar with options like '1 Licensing', 'Include OS cost', 'RI Profile', 'AWS: Offering Class' (STANDARD), 'AWS: Payment' (ALL UPFRONT), 'AWS: Term' (1 YEAR), and 'RI Inventory' (0 OUT OF 0 ACTIVE). The main area features a 'RESULTS OVERVIEW' section with a warning banner: 'Unplaced Virtual Machines: Lift&Shift: 69; Optimized: 66'. Below the banner are tabs for 'RESULTS OVERVIEW', 'LIFT & SHIFT ACTIONS (1437)', and 'OPTIMIZED ACTIONS (1725)'. A 'CLOUD COST COMPARISON' table is shown for 'aws-US West (N. California)'. The table compares 'LIFT & SHIFT' and 'OPTIMIZED' scenarios across various cost categories.

| | LIFT & SHIFT | OPTIMIZED | DIFFERENCE | % |
|--|----------------------|---------------------|----------------------|-----------------|
| Virtual Machines with performance risks | 99 Out Of 684 | 0 Out Of 687 | 99 | - |
| Virtual Machines with efficiency opportunities | 588 Out Of 684 | 0 Out Of 687 | 588 | - |
| On-Demand Compute Cost | \$113,460 /mo | \$23,214 /mo | -\$90,246 /mo | ▼ 79.5 % |
| Reserved Compute Cost | \$9,816 /mo | \$19,229 /mo | \$9,413 /mo | ▲ 95.9 % |
| Storage Cost | \$32,183 /mo | \$40,397 /mo | \$8,214 /mo | ▲ 25.5 % |
| Total Cost | \$155,459 /mo | \$82,840 /mo | -\$72,619 /mo | ▼ 46.7 % |

Intersight Workload Optimizer shows results for two migration scenarios:

■ **Lift & Shift**

Lift & Shift migrates your VMs to cloud instances that match their current resource allocations.

■ **Optimized**

As Intersight Workload Optimizer runs the plan, it looks for opportunities to optimize cost and performance. For example, it might discover overprovisioned VMs after analyzing the historical utilization of VM resources. If you were to migrate such VMs to instances that match their current allocations, then you would spend more than necessary. For an optimized migration, Intersight Workload Optimizer can recommend migrating to less expensive instances while still assuring performance, and then show the resulting savings. In addition, when you examine the actions for an optimized migration, you will see charts that plot the historical utilization data used in the analysis.

Results Overview

The Results Overview section shows the following:

■ **Unplaced VMs**

If the plan's scope includes VMs that cannot be migrated, the results include a notification indicating the number of VMs. Click **Show Details** to see the list of VMs and the reasons for their non-placement.

This close-up shows the notification banner: 'Unplaced Virtual Machines: Lift&Shift: 69; Optimized: 66' with a 'SHOW DETAILS' button. Below it are the navigation tabs: 'RESULTS OVERVIEW', 'LIFT & SHIFT ACTIONS (1437)', and 'OPTIMIZED ACTIONS (1725)'. The 'CLOUD COST COMPARISON' table header is also visible.

The charts in the plan results do not count these VMs.

Intersight Workload Optimizer displays adjusted CPU values for unplaced VMs. These values are the actual metrics used in analysis and are calculated using [benchmark data](#). CPU values shown in other places (such as the Capacity and Usage chart) are unadjusted values obtained from targets.

■ **Cloud Cost Comparison Chart**

This chart highlights any difference in cost as a result of optimization. For example, undersized VMs risk losing performance and should therefore scale up. This could contribute to an increase in cost. On the other hand, oversized VMs can scale down to less expensive instances, so cost should go down. The values under the % column indicate the percentage of VMs that are affected by optimization cost calculations.

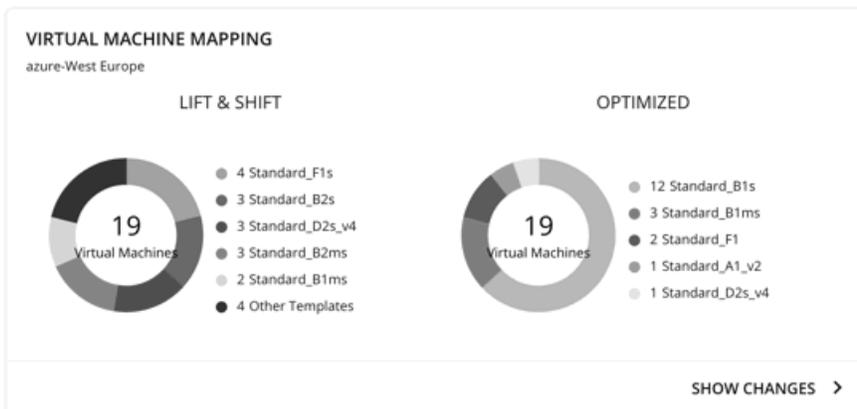
| CLOUD COST COMPARISON | | | | |
|--|----------------|--------------|---------------|----------|
| aws-US West (N. California) | | | | |
| | LIFT & SHIFT | OPTIMIZED | DIFFERENCE | % |
| Virtual Machines with performance risks | 99 Out Of 684 | 0 Out Of 687 | 99 | - |
| Virtual Machines with efficiency opportunities | 588 Out Of 684 | 0 Out Of 687 | 588 | - |
| On-Demand Compute Cost | \$113,460 /mo | \$23,214 /mo | -\$90,246 /mo | ▼ 79.5 % |
| Reserved Compute Cost | \$9,816 /mo | \$19,229 /mo | \$9,413 /mo | ▲ 95.9 % |
| Storage Cost | \$32,183 /mo | \$40,397 /mo | \$8,214 /mo | ▲ 25.5 % |
| Total Cost | \$155,459 /mo | \$82,840 /mo | -\$72,619 /mo | ▼ 46.7 % |

NOTE:

For Azure, the results do not include the license cost for the migrated VMs.

■ **Virtual Machine Mapping Chart**

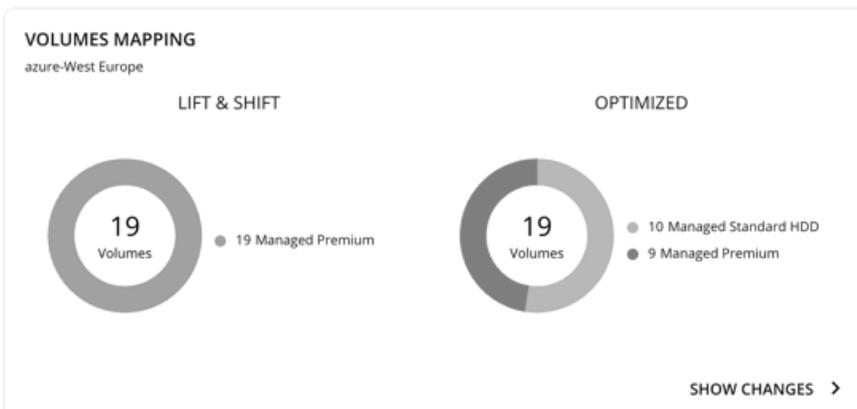
This chart gives a breakdown of the instance types that the plan recommends for the migration, including how many of each is needed.



Click **Show Changes** to see a table with details for each VM in the plan. The table maps VMs to instance types. It also shows the properties and monthly cost for each instance type, and indicates whether Intersight Workload Optimizer recommends buying discounts. Under the **Actions** column, click **Details** to compare Lift & Shift and Optimized actions.

■ **Volume Tier Summary Chart**

This chart gives a breakdown of the volume types that the plan recommends for the migration, including how many of each is needed.



Click **Show Changes** to see a table with details for each volume in the plan. The table maps the volumes you plan to migrate to the volume types that Intersight Workload Optimizer recommends. It also shows the properties and monthly cost for each volume type. Under the **Actions** column, click **Details** to compare Lift & Shift and Optimized actions.

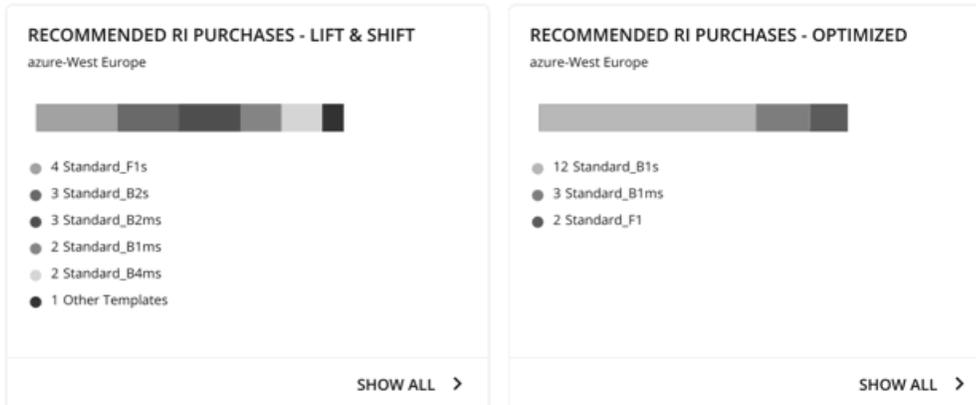
■ **Recommended RI Purchases Charts**

Intersight Workload Optimizer can recommend purchasing instance types at a discounted rate to help you increase the percentage of VMs covered by discounted pricing and reduce on-demand costs. This chart shows your pending purchases. Download the list of purchases and then send it your cloud provider or representative to initiate the purchase process.

NOTE:

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Currently, Intersight Workload Optimizer can recommend purchasing AWS EC2 RIs and Azure reservations.



To identify VMs that are good candidates for discounted pricing, Intersight Workload Optimizer analysis considers the history of a VM (by default, the last 21 days), and it looks for:

- Activity
If the VM's VCPU utilization percentile is 20% or higher, then Intersight Workload Optimizer considers it an active VM.
- Stability
If there have been no start, stop, or resize actions for the VM for 16 of the last 21 days, then Intersight Workload Optimizer considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Intersight Workload Optimizer can recommend purchasing additional discounts.

Click **Show All** to see a table with details for each discount.

The table shows the properties, up-front cost, and break-even period for each discount. The break-even period is the time at which savings will exceed the up-front cost, rounded to the month. The Cost Impact column indicates the monthly savings you would realize when you buy a specific discount.

When you choose one or more check boxes, the total count, up-front cost, and savings appear at the top.

Click **Details** under the **Actions** column to compare Lift & Shift and Optimized actions.

NOTE:

The plan assumes that a discount will always be less expensive than its on-demand counterpart. However, this is not always the case. There might be billing details from service providers that could lead to recommendations to move to a discounted instance type that is more expensive than running on demand.

Plan Actions

Intersight Workload Optimizer shows separate tabs for **Lift & Shift** and **Optimized** migration actions. You can download the list of actions as a CSV file.

| | | RESULTS OVERVIEW | | LIFT & SHIFT ACTIONS (2) | | OPTIMIZED ACTIONS (2) | |
|----------------------|----------------------|------------------|---------------|--------------------------|---------------------|-----------------------|------------|
| MOVE | Move Actions (1) | | | | | | |
| Volumes (1) | Type to search | | | | | | ADD FILTER |
| Virtual Machines (1) | Virtual Machine Name | Type | From | To | Risk | Action Category | Action |
| | pysavingAz1 | Computer Tier | Standard_B1ms | m5.large | Optimized migration | PERFORMANCE | DETAILS |

For *Optimized* migrations, when you expand an action on a VM, you will see charts that track VCPU and VMem utilization for that VM. With these charts, you can easily recognize the utilization trends that Intersight Workload Optimizer analyzed to determine the most efficient instance for the VM.



For more information about these charts, see [Utilization Charts \(on page 402\)](#).

Uploading Plan Results to Azure Migrate

Intersight Workload Optimizer can upload the plan results and additional plan information to the Azure Migrate portal as part of your migration process. This feature is only available for plans that simulate on-prem VM migration to an Azure region.

Uploaded information includes:

- Basic information for the on-prem VMs, including OS Name and Machine Name
- Target Azure region, VM size, and storage type

NOTE:

Azure Migrate does not support automatic selection of OS Disk or manual selection of Ultra Storage disk tiers as part of a migration plan.

- Discount recommendations
- OS license recommendations (based on the licensing option that you selected for the plan)

NOTE:

The Azure Migrate portal displays standardized information provided by third-party migration assessment solutions, including Intersight Workload Optimizer. Microsoft might not support displaying some information unique to Intersight Workload Optimizer.

Before uploading the results, be sure to complete the following tasks:

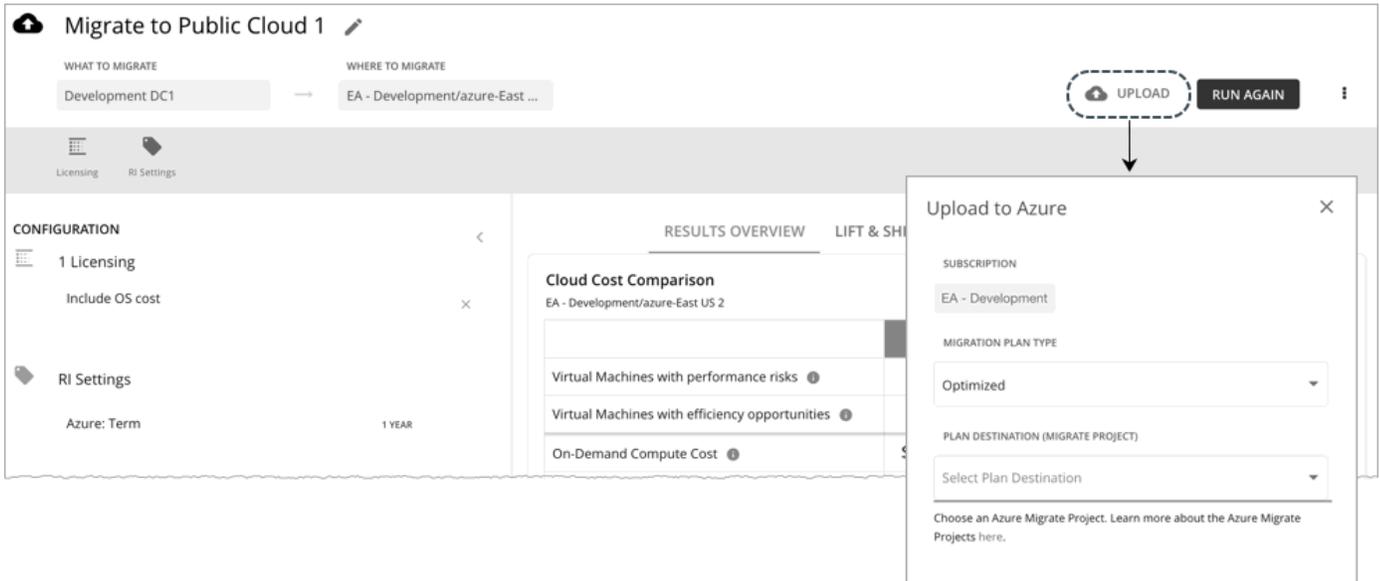
1. Create a project in the Azure Migrate portal.

2. Add Intersight Workload Optimizer as a migration assessment solution to the project.
3. Set the necessary permissions in the Azure Migrate portal. For details about permissions, see [Azure Service Principal and Subscription Permissions \(on page 86\)](#).

Consult the Azure documentation for information on completing these tasks.

When you are ready to upload:

1. Click **Upload** at the top-right corner of the Plan Page.



2. Specify the following:

- Migration Plan Type

Choose to migrate either the 'Lift & Shift' or 'Optimized' results.

- Plan Destination (Migrate Project)

Select from the list of Azure Migrate projects. These are the projects belonging to the Azure subscription that you selected for the plan. If you have not created a project for the subscription, go to the Azure Migrate portal and create one.

WARNING:

Uploading to a project with existing plan results overwrites those results.

The upload will fail if another upload targeted at the same destination is already in progress.

3. Click **Submit**.

The Plan Page updates to display the upload status. Refresh the page periodically to check:

- If the upload task completed without problems
- Any upload issues for individual entities

4. When the upload is complete, log in to the Azure Migrate portal and go to the project you selected as the plan destination.

The project should now display the uploaded information. Use the migration tools identified for the project to start the actual migration.

NOTE:

Repeat the upload procedure if you re-ran the plan and want to upload the new results.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

NOTE:

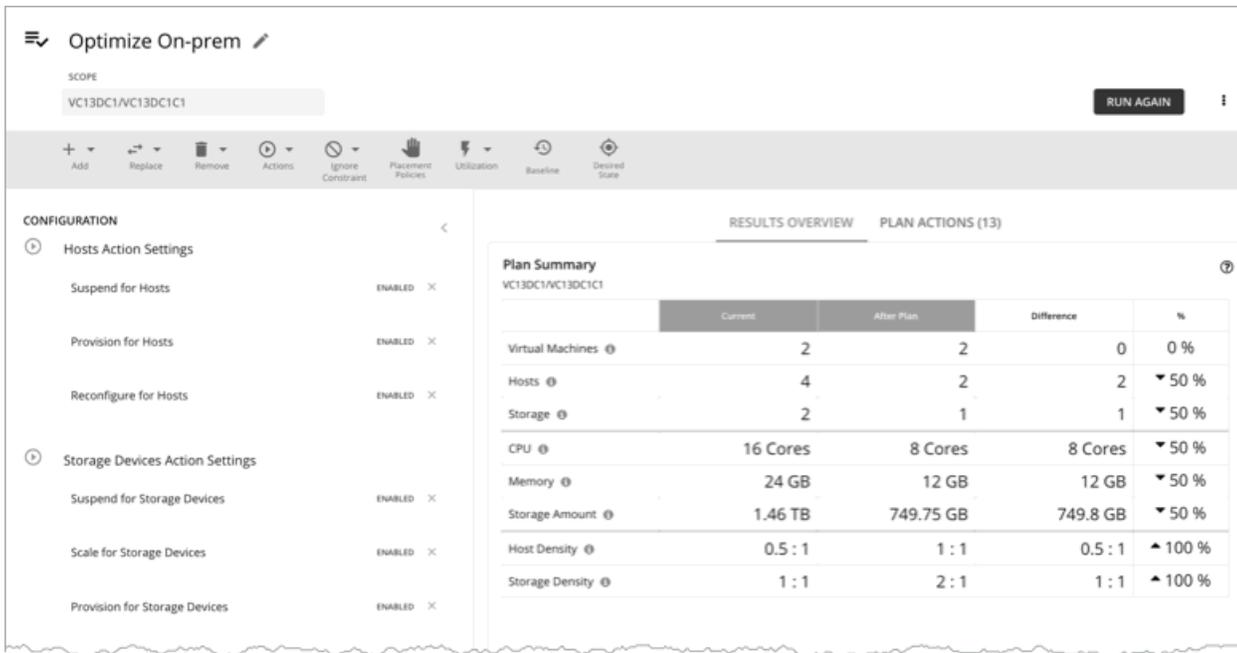
It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Optimize On-prem Plan

Run the Optimize On-prem plan to see the effects of executing certain actions, such as scaling virtual machines, suspending hosts, or provisioning storage, to your on-prem environment.

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).



The screenshot shows the 'Optimize On-prem' interface. At the top, the scope is set to 'VC13DC1/VC13DC1C1' and there is a 'RUN AGAIN' button. Below the scope, there are several action icons: Add, Replace, Remove, Actions, Ignore Constraints, Placement Policies, Utilization, Baseline, and Desired State. The left sidebar shows configuration settings for Hosts and Storage Devices, all of which are currently 'ENABLED'. The main area displays a 'Plan Summary' table comparing current and after-plan states.

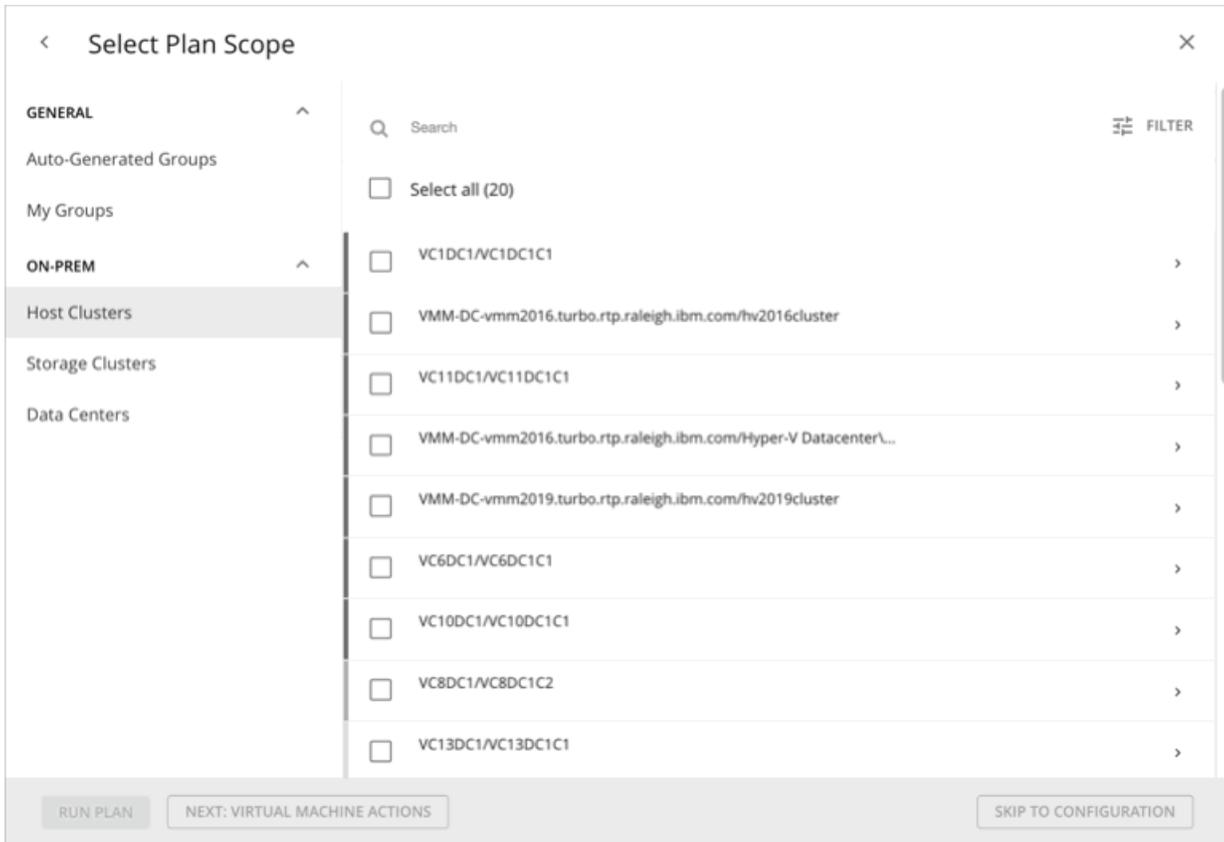
| | Current | After Plan | Difference | % |
|------------------|----------|------------|------------|---------|
| Virtual Machines | 2 | 2 | 0 | 0 % |
| Hosts | 4 | 2 | 2 | ▼ 50 % |
| Storage | 2 | 1 | 1 | ▼ 50 % |
| CPU | 16 Cores | 8 Cores | 8 Cores | ▼ 50 % |
| Memory | 24 GB | 12 GB | 12 GB | ▼ 50 % |
| Storage Amount | 1.46 TB | 749.75 GB | 749.8 GB | ▼ 50 % |
| Host Density | 0.5 : 1 | 1 : 1 | 0.5 : 1 | ▲ 100 % |
| Storage Density | 1 : 1 | 2 : 1 | 1 : 1 | ▲ 100 % |

Configuring an Optimize On-prem Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

1. Scope

If you set the scope to a specific Host Cluster, Storage Cluster, Data Center, or group, you can start any plan. You may need to go through additional steps, depending on your chosen plan type. For example, if you **Scope** to a cluster and choose the **Add Virtual Machines** plan type, the plan wizard prompts you to select the most suitable templates for the VMs you plan to add to the cluster.



2. Virtual Machine Actions

See the effect of enabling or disabling **Scale** actions on the entity included in the plan.

3. Host Actions

See the effect of enabling or disabling **Suspend**, **Provision**, or **Reconfigure** actions on the entity included in the plan.

< Host Actions
×

Suspend ⓘ

All Hosts in the plan scope
 All Hosts except for the following groups

Provision ⓘ

All Hosts in the plan scope
 All Hosts except for the following groups

Reconfigure

All Hosts in the plan scope
 All Hosts except for the following groups

RUN PLAN

NEXT: STORAGE ACTIONS

SKIP TO CONFIGURATION

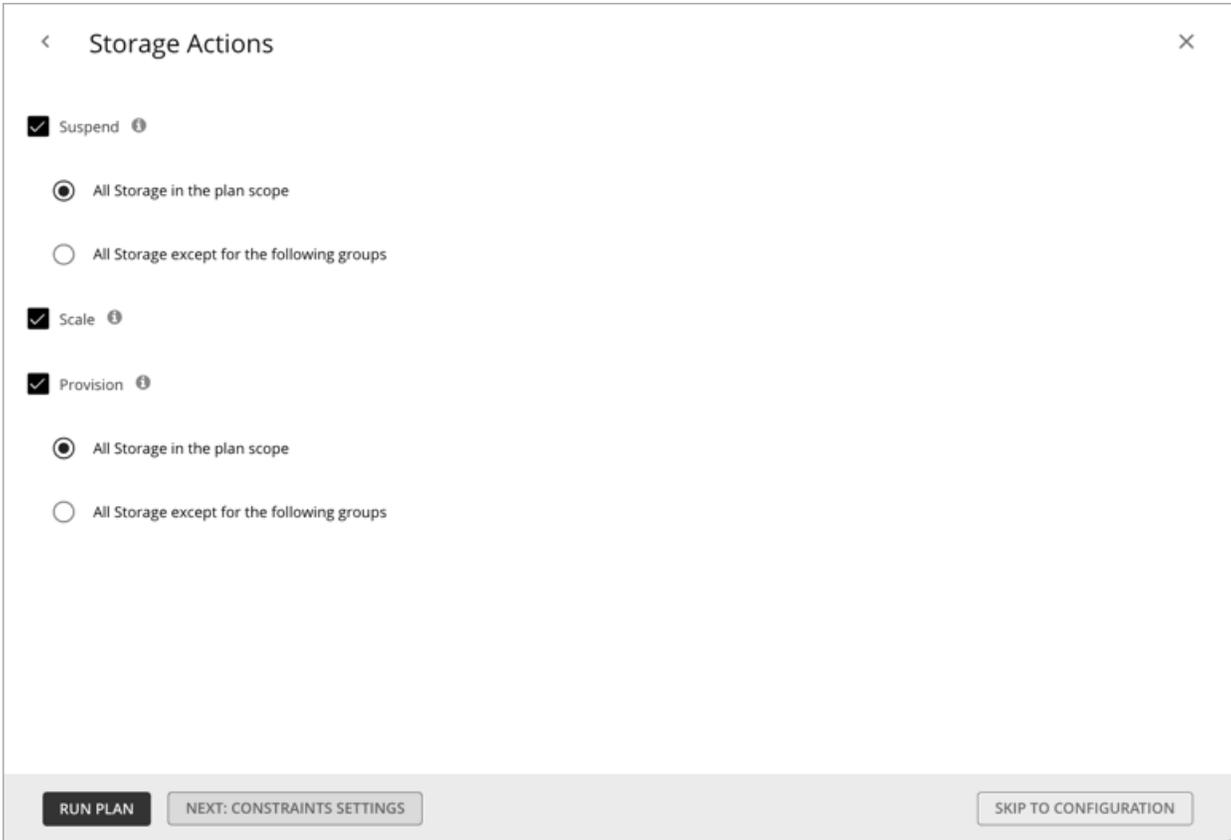
Suspend - Allows Intersight Workload Optimizer to remove hosts from your environment, if the capacity is not needed to support the workloads in scope.

Provision- Allows Intersight Workload Optimizer to add hosts to your environment, if the capacity is not enough to support the workloads in scope. You will have to purchase or provision hardware to meet the plan results.

Reconfigure- Allows Intersight Workload Optimizer to generate actions to guide you in changing your hosts (for example, licensing a host for a specific application or operating system).

4. Storage Actions

See the effect of enabling or disabling **Suspend**, **Scale**, or **Provision** actions on the entity included in the plan.



Suspend- Allows Intersight Workload Optimizer to remove data stores from your environment, if the capacity is not needed to support the workloads in scope.

Scale- Allows Intersight Workload Optimizer to modify your storage device to provide more IOPS or more capacity.

Provision- Allows Intersight Workload Optimizer to add storage devices to your environment if the capacity is not enough to support the workloads in scope. You will have to purchase or provision hardware to meet the plan results.

5. Constraints Settings

Choose whether to ignore constraints (such as placement policies) for VMs in your environment. By default, VMs are constrained to the cluster, network group, data center, or storage group that their hosts belong to. You can choose to ignore these boundaries.

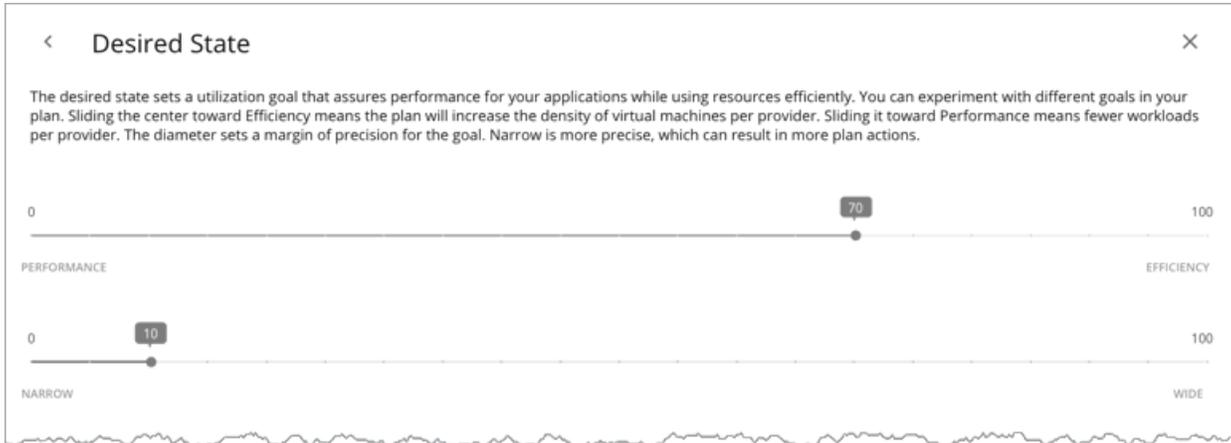
For example, by default a plan does not consider moving VMs to physical hosts outside of the current cluster. If you disable the Cluster constraint for a VM in your plan, then the plan can evaluate the results of hosting those VMs on any other physical machine within the scope of your plan. If the best results come from moving that VM to a different cluster, then the plan will show that result.

NOTE:

If you are adding hosts to a plan, and use host templates, then you must turn on **Ignore Constraints**.

6. Desired State

The desired state sets a utilization goal that assures performance for your applications while using resources efficiently. You can experiment with different goals in your plan. Sliding the center toward Efficiency means the plan will increase the density of virtual machines per provider. Sliding it toward Performance means fewer workloads per provider. The diameter sets a margin of precision for the goal. Narrow is more precise, which can result in more plan actions.



The desired state is a condition in your environment that assures performance for your workloads, while it utilizes your resources as efficiently as possible and you do not overprovision your infrastructure. Intersight Workload Optimizer uses default Desired State settings to drive its analysis. You should never change the settings for real-time analysis unless you are working directly with Technical support. However, you can change the settings in a plan to see what effect a more or less aggressive configuration would have in your environment.

You can think of the desired state as an n-dimensional sphere that encompasses the fittest conditions your environment can achieve. The multiple dimensions of this sphere are defined by the resource metrics in your environment. Metric dimensions include VMem, storage, CPU, etc. While the metrics on the entities in your environment can be any value, the desired state, this n-dimensional sphere, is the subset of metric values that assures the best performance while achieving the most efficient utilization of resources that is possible.

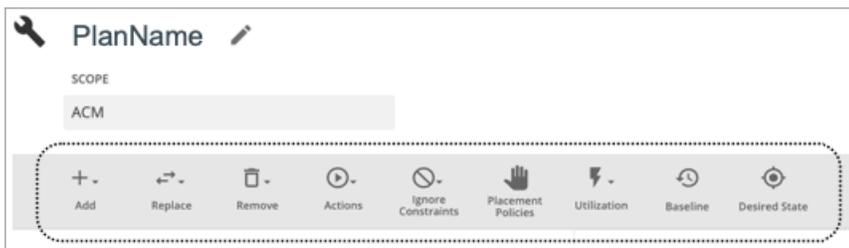
The Desired State settings center this sphere on Performance (more infrastructure to supply the workload demand), or on Efficiency (less investment in infrastructure to supply the workload demand). The settings also adjust the diameter of the sphere to determine the range of deviation from the center that can encompass the desired state. If you specify a large diameter, Intersight Workload Optimizer will have more variation in the way it distributes workload across hosting devices.

For more information, see [The Desired State \(on page 14\)](#)

Click **RUN PLAN** to view the results.

7. Plan Configuration

Use the Plan Configuration toolbar to fine-tune your plan settings. You can change workload demand and the supply of resources, and specify other changes to the plan market.



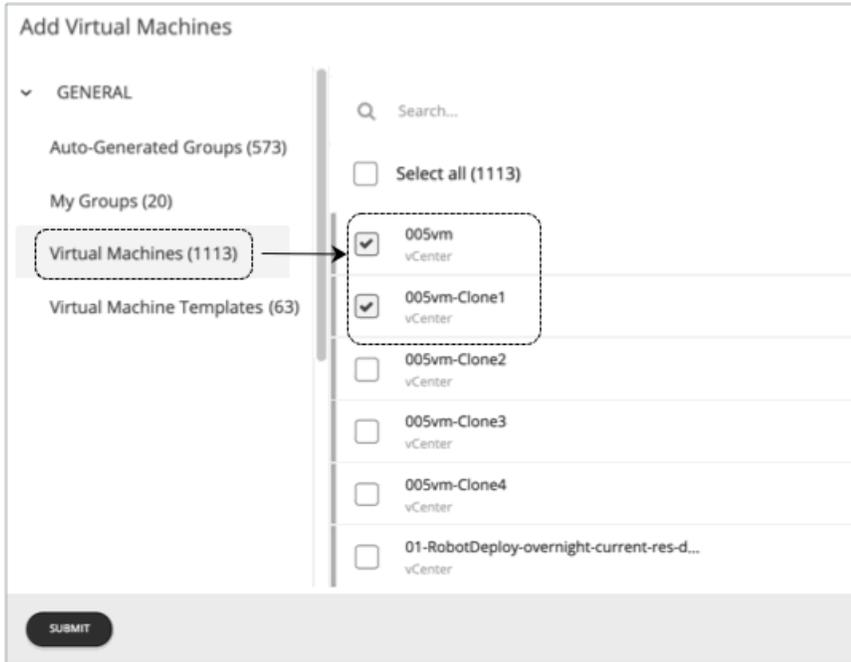
7.1. Add

Add virtual machines, hosts, or storage to your plan. For example, when you add hosts, you increase the compute resources for the plan.

Copy from an Entity or Template

Choose an entity or template to copy. This describes the new entities that Intersight Workload Optimizer will add to the plan. For example, you can run a plan that adds new VMs to a cluster. If you copy from a template, then the plan adds a new VM that matches the resource allocation you have specified for the given template.

- Option 1: Copy from an entity



■ Option 2: Copy from a template

If no existing template is satisfactory, create one by clicking **New Template**.



NOTE:

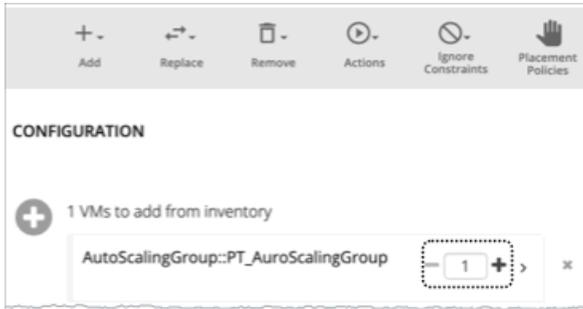
Intersight Workload Optimizer automatically adds any new template you create to the Template Catalog page (**Settings > Templates**).

It is not possible to use templates for containers or container pods.

Use the **Filter** option to show entities or templates with certain properties (name, number of CPUs, and so on). This makes it easier to sort through a long list.

Number of Copies to Add

After choosing an entity or template, it appears as an entry in the Configuration summary. Then you can set how many copies to add.



7.2. Replace

Replacing virtual machine is a way to change the properties of VMs in your plan market. When you replace workload, you select one or more VMs that you want to change, and then you select a template to use in their place. The list of changed VMs displays in the Configuration Summary. You can delete individual entries from the this summary if necessary.

Replacing hosts or storage is a way to plan for a hardware upgrade. For example, if you replace your hosts or data stores with a more powerful template, the plan might show that you can use fewer hosts or data stores, and it will show the best placement for workloads on those entities. You begin by selecting the entities you want to replace, and when you click **REPLACE** you can then choose a template that will replace them. Note that you can only choose a single template for each set of entities you want to have replaced. You can configure different replacements in the same plan, if you want to use more than one template.

7.3. Remove

Removing virtual machines frees up resources for other workloads to use. Removing hosts or storage means you have fewer compute or storage resources for your workloads.

If you think you have overprovisioned your environment, you can run a plan to see whether fewer hosts or less storage can still support the same workload.

7.4. Actions

See the effect of enabling or disabling actions on the entities included in the plan. See [Virtual Machine Actions \(on page 470\)](#), [Host Actions \(on page 470\)](#), and [Storage Actions \(on page 471\)](#).

7.5. Ignore Constraints

Choose to ignore constraints (such as placement policies) for VMs in your environment. See [Constraints Settings \(on page 472\)](#).

7.6. Placement Policies

By default, the plan includes all the placement policies that apply to the plan scope. Also, these policies are in their real-time state (enabled or disabled).

Placement Policies ✕

Set up placement polices for the plan ⓘ

🔍 Search... ⚙️ FILTER

10 Policies NEW PLACEMENT POLICY

| | | |
|-------------------------------------|--|---------|
| <input checked="" type="checkbox"/> | GROUP-DRS-AdityaNotOn48-rule/Cluster1/vsphere-dc7.eng.vmturbo.com Place | Enabled |
| <input checked="" type="checkbox"/> | GROUP-DRS-Ah-Rule-Snow-rule/Adv Eng/vsphere-dc12.eng.vmturbo.com Place | Enabled |
| <input checked="" type="checkbox"/> | GROUP-DRS-arsen-separate-vms-rule/Physical/vsphere-dc11.eng.vmturbo.com Place | Enabled |

You can use these settings to enable or disable existing policies, or you can create new policies to apply only to this plan scenario. For information about creating placement policies, see [Placement Policies \(on page 569\)](#).

7.7. Utilization

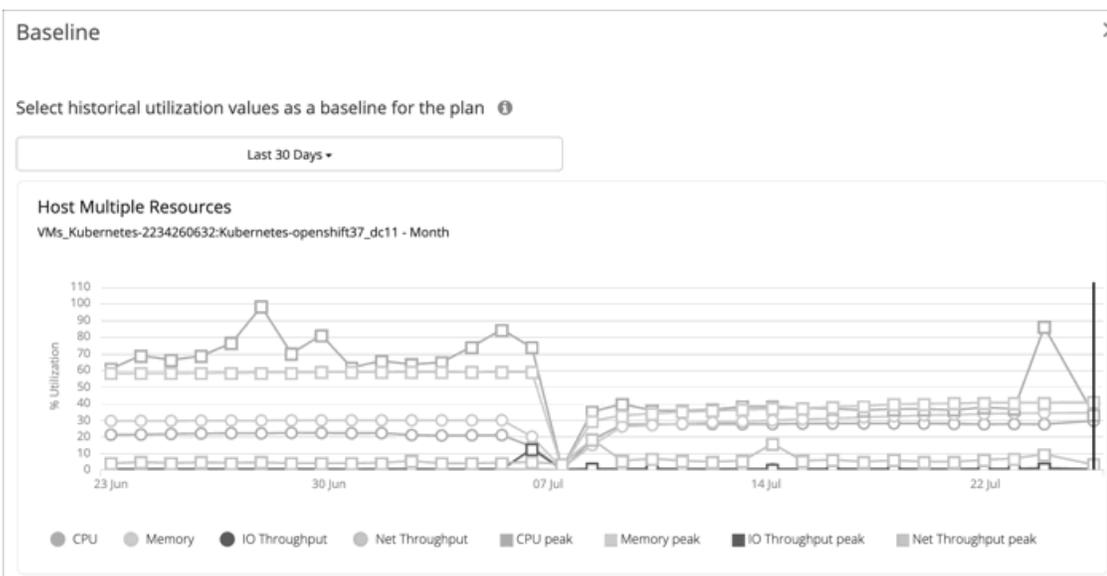
Setting utilization by a certain percentage is a way to increase or decrease the workload for the scope of your plan and any entity added to the plan, or for specific groups. Intersight Workload Optimizer uses the resulting utilization values as the baseline for the plan.

Max Host Utilization levels specify the percentage of the physical resource that you want to make available in the given plan. By default, hosts have utilization set to 100%. For a given plan, you can set the utilization to a lower value. For example, assume you want to simulate High Availability of 25% for some hosts in the plan. In that case, you can select these hosts and set their utilization levels to 75%.

Max Storage utilization levels specify the percentage of the physical resource that you want to make available in the given plan. By default, storage has utilization set to 100%. For a given plan, you can set the utilization to a lower value. For example, assume you have one data store that you want to share evenly for two clusters of VMs. Also assume that you are creating a plan for one of those clusters. In that case, you can set the data stores to 50% utilization. This saves storage resources for the other cluster that will use this storage.

7.8. Baseline

Use these settings to set up the baseline of utilization metrics for your plan.



By default, the plan runs against the current state of your environment. You can set up the plan to add or remove entities, or otherwise affect the plan calculations. But the utilization metrics will be based on the current state of the plan. If you run the same plan multiple times, each run begins with a fresh view of your inventory.

You can select from the list of snapshots to load the utilization statistics from a previous time period into the plan. Use this to run the plan against utilization that you experienced in the past. For example, assume a peak utilization period for the month before the winter holidays. During the holidays you want to plan to add new capacity that can better handle that peak. You would set the baseline to the utilization you saw during that peak before the holidays.

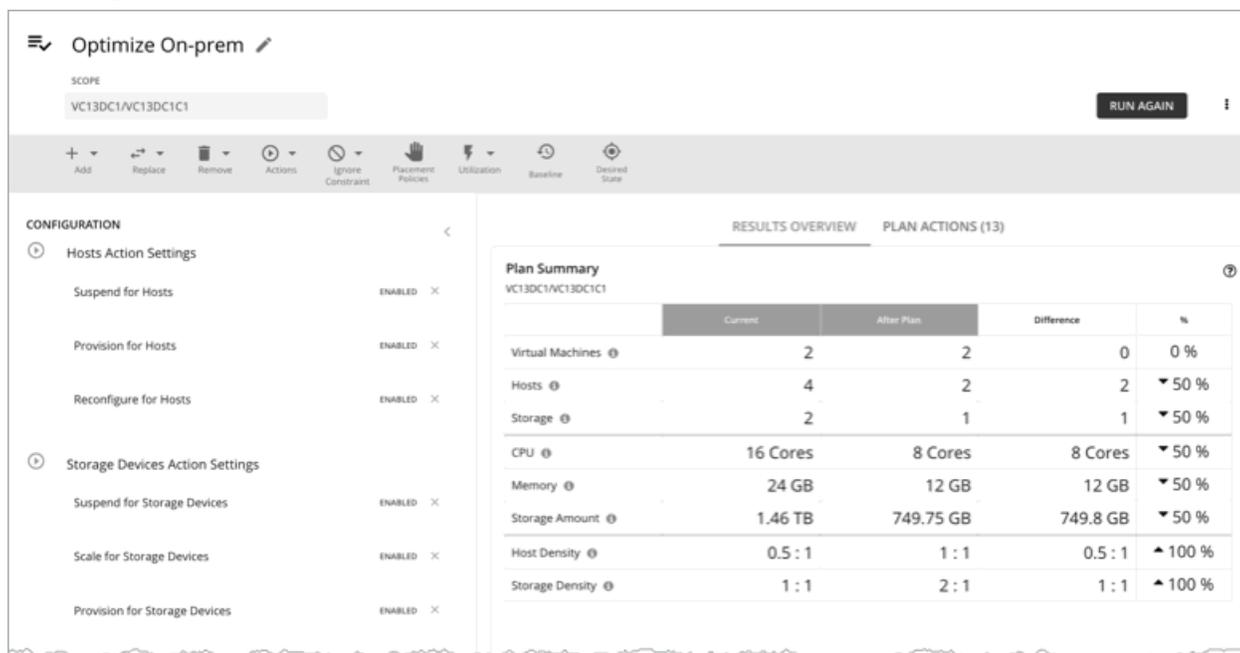
7.9. Desired State

The desired state is a condition in your environment that assures performance for your workloads, while it utilizes your resources as efficiently as possible and you do not over-provision your infrastructure. See [Desired State \(on page 472\)](#).

Working With Optimize On-prem Plan Results

After the plan runs, you can view the results to see how the plan settings you configured affect your environment.

Viewing the Results



The plan results include the following charts:

■ Plan Summary Chart

This chart compares your current resources to the resources you would get after executing the plan.

NOTE:

Under some circumstances, this chart might not count "non-participating" entities in the real-time market, such as suspended VMs or hosts in a failover state. The following charts, on the other hand, count all entities in the real-time market, regardless of state:

- Scope Preview chart (displays before you run the plan)
- Optimized Improvements and Comparison charts

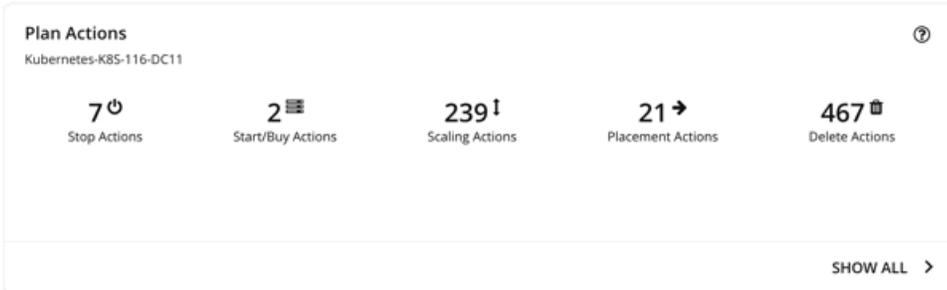
If the plan's scope includes VMs that cannot be placed, the results include a notification indicating the number of VMs. Click **Show Details** to see the list of VMs and the reasons for their non-placement.

Click **Show all** at the bottom of the chart to see savings or investment costs, or to download the chart as a CSV file.

■ Plan Actions Chart

This chart summarizes the actions that you need to execute to achieve the plan results. For example, if you run an Alleviate Pressure plan, you can see actions to move workloads from the hot cluster over to the cold cluster. If some VMs are overprovisioned, you might see actions to reduce the capacity for those workloads.

The text chart groups actions by [action type \(on page 413\)](#). The list chart shows a partial list of [actions \(on page 395\)](#).

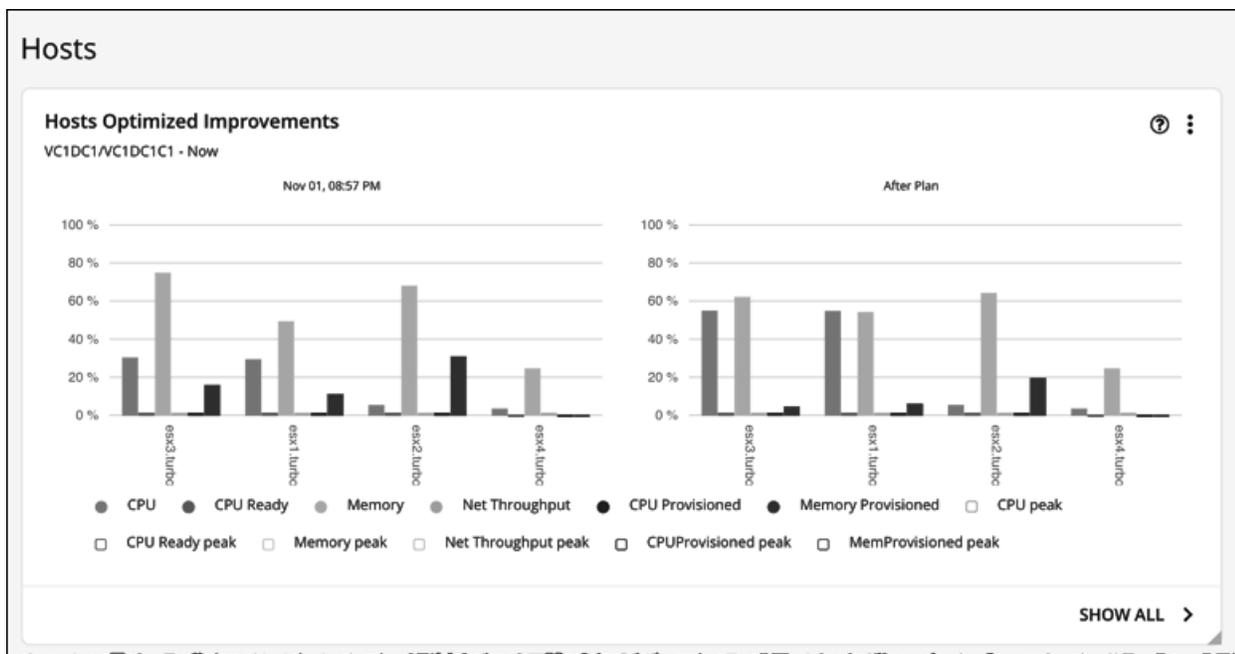


To view action details or download the list of actions as a CSV file:

- Click an action type in the text chart or an individual action in the list chart.
- Click **Show All** at the bottom of the chart.

■ **Optimized Improvements Charts for Hosts, Storage, and Virtual Machines**

The Optimized Improvements chart shows how the utilization of resources would change assuming you accept all of the actions listed in the Plan Actions chart.



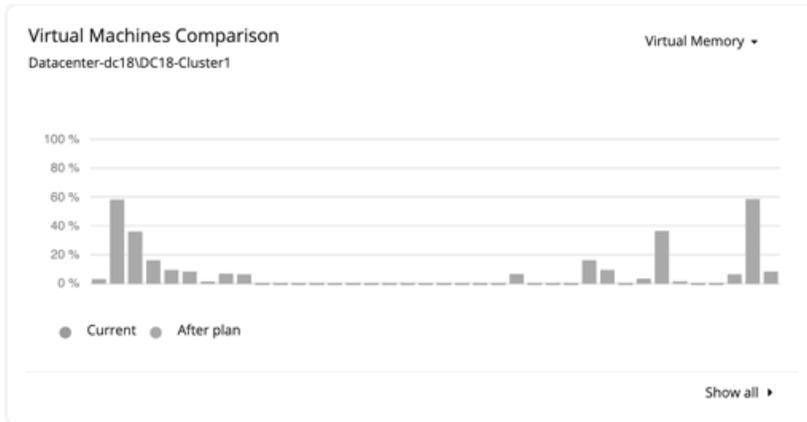
- In many of these charts, you can change the commodities on display. To do this, go to the top-right section of the chart, click the More options icon (⋮), and then select **Edit**. In the new screen that displays, go to the **Commodity** section and then add or remove commodities.

To restore the default commodities, use the **Reset view** option at the top-right section of the page.

- Click **Show all** at the bottom of the chart to see a breakdown of the current chart data by entity (for example, show CPU, Memory, and IO Throughput utilization for each host), or to download chart data as a CSV file.

■ **Comparison Charts for Hosts, Storage Devices, and Virtual Machines**

A Comparison chart shows how the utilization of a particular commodity (such as memory or CPU) for each entity in the plan would change if you execute the actions listed in the Plan Actions chart.



- To change the commodity displayed in the chart, go to the top-right section of a chart and then select from the list of commodities.

To restore the default commodity, go to the top-right section of the page, click the More options icon (), and then select **Reset view**.

- Click **Show all** at the bottom of the chart to show a breakdown of the current chart data by entity (for example, show Virtual Memory utilization for each virtual machine), or to download the chart as a CSV file.

NOTE:

For the Storage Devices Comparison chart, if you set the view to **VM Per Storage** and click **Show all**, the total number of VMs sometimes does not match the number in the Plan Summary chart. This happens if there are VMs in the plan that use multiple storage devices. The Storage Devices Comparison chart counts those VMs multiple times, depending on the number of storage devices they use, while the Plan Summary chart shows the actual number of VMs.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.

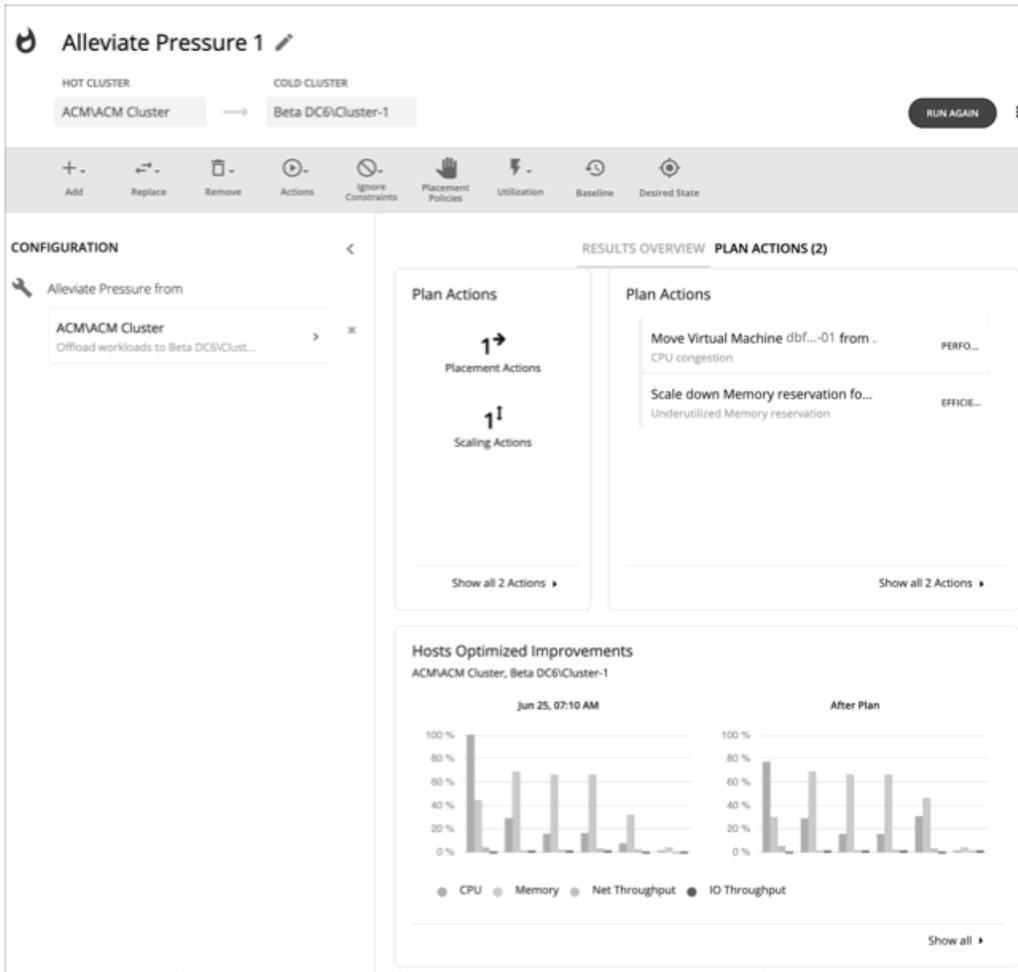


NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Alleviate Pressure Plan



Use the Alleviate Pressure plan to find out how to migrate workloads from a stressed or *hot* cluster over to a cluster with more headroom. This plan shows the minimal changes you need to make to reduce risks on the hot cluster.

The plan results:

- Show the actions to migrate workloads from the hot cluster to the cold one
- Compare the current state of your clusters to the optimized state
- Show resulting headroom for both the hot and the cold clusters
- Show trends of workload-to-inventory over time for both clusters

Alleviate Pressure plans make use of the headroom in your clusters. Headroom is the number of VMs the cluster can support, for CPU, Memory and Storage.

To calculate cluster capacity and headroom, Intersight Workload Optimizer runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current VMs to other hosts within the given cluster, if such moves would result in a more desirable workload distribution. The result of the plan is a calculation of how many more VMs the cluster can support.

To calculate VM headroom, the plan simulates adding VMs to your cluster. The plan assumes a certain capacity for these VMs, based on a specific VM template. For this reason, the count of VMs given for the headroom is an approximation based on that VM template.

To specify the templates these plans use, you can configure the nightly plans for each cluster. For more information, see [Configuring Nightly Plans \(on page 492\)](#)

NOTE:

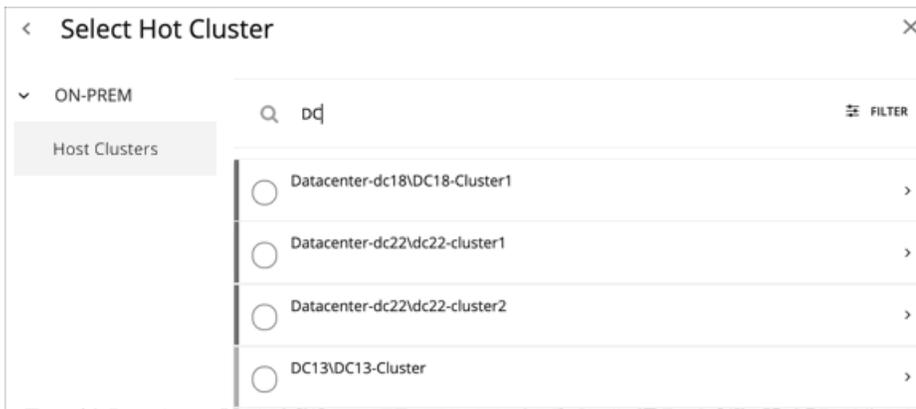
To execute, this plan must ignore certain constraints. The plan ignores cluster constraints to allow migrating workloads from the hot cluster to the cold one. It also ignores network constraints, imported DRS policies, and any Intersight Workload Optimizer that would ordinarily be in effect.

Configuring an Alleviate Pressure Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

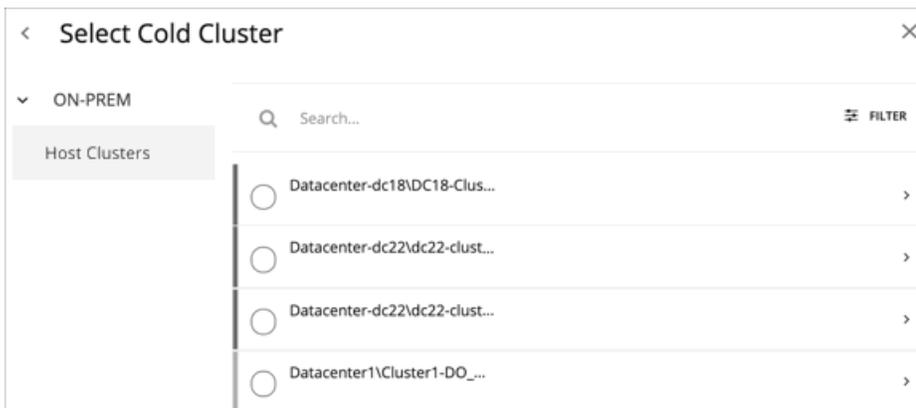
1. Scope

The wizard first gives you a list for you to choose the hot cluster. This is the cluster that shows risks to performance. The list sorts with the most critical clusters first, and it includes the calculated headroom for CPU, Memory, and Storage in each cluster.



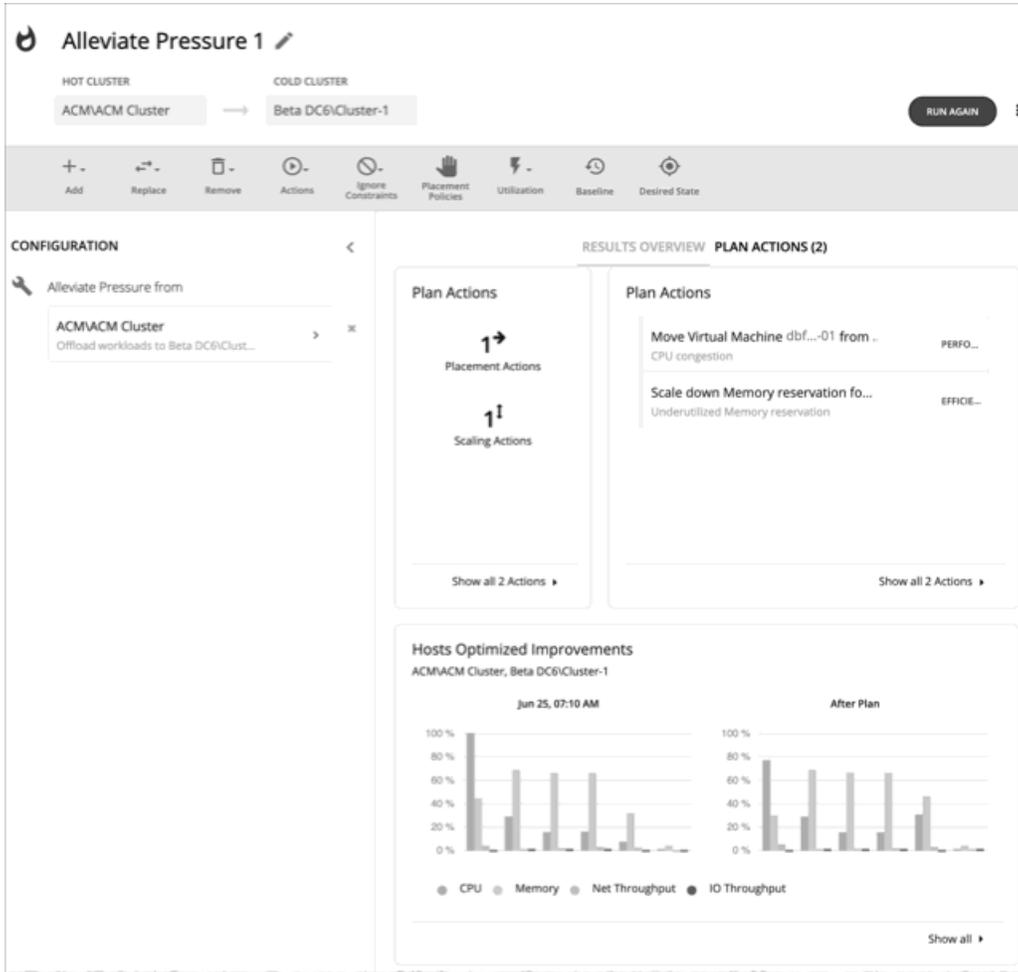
2. Cold Cluster

After you select the hot cluster, choose the cold cluster.



Working With Alleviate Pressure Plan Results

After the plan runs, you can view the results to see how the migration of workloads off of your hot cluster affects your environment.



Viewing the Results

The results include the following charts:

- **Plan Actions**

You can see a list of actions to reduce the pressure on the hot cluster. It's typical to see actions to move workloads from the hot cluster over to the cold cluster. If some VMs are overprovisioned, you might see actions to reduce the capacity for those workloads.
- **Hosts Optimized Improvements**

This chart compares the current state of the hot cluster to its state after executing the plan actions. It displays the resource utilization of the cluster's hosts both before and after the plan.
- **Headroom**

With these charts, you can compare the headroom between the hot and cold clusters.
- **Virtual Machines vs Hosts and Storage**

This chart shows the total number of virtual machines, hosts, and storage in your on-prem environment, and tracks the data over time. Chart information helps you understand and make decisions around capacity and utilization, based on historical and projected demand.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.



The toolbar items that display are similar to the toolbar items for a custom plan. For details, see [Configuring a Custom Plan \(on page 483\)](#).

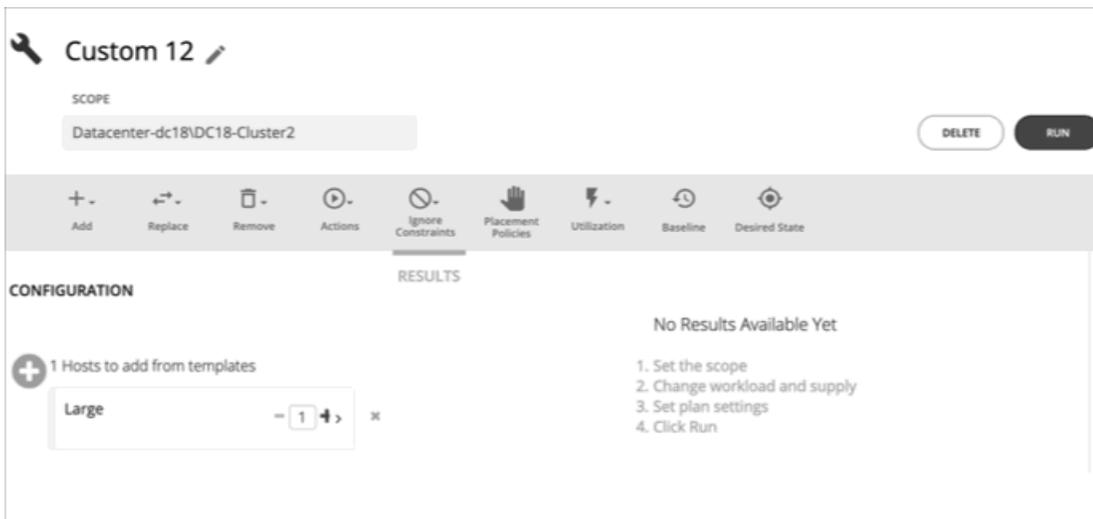
NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Custom Plan

For an overview of setting up plan scenarios, see [Settings Up User Plan Scenarios \(on page 418\)](#).



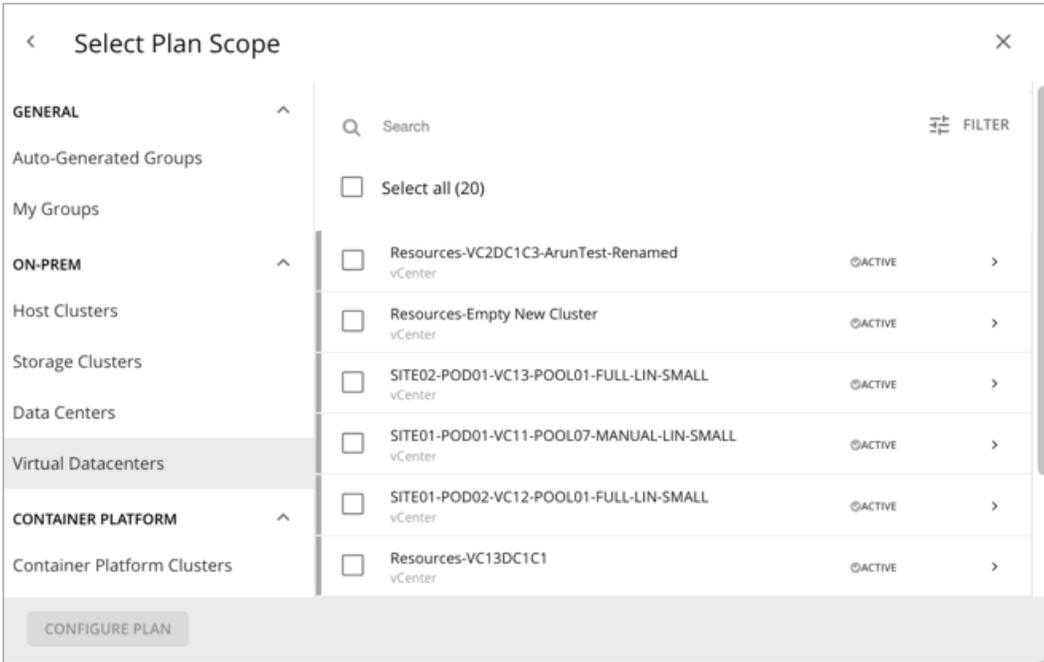
When you create a custom scenario, you specify the plan scope as an initial step, and then skip the plan wizards and jump straight into setting up the plan parameters. You can name the plan, change workload demand and the supply of resources, and specify other changes to the plan market.

Configuring a Custom Plan

For an overview of setting up plan scenarios, see [Setting Up Plan Scenarios \(on page 418\)](#).

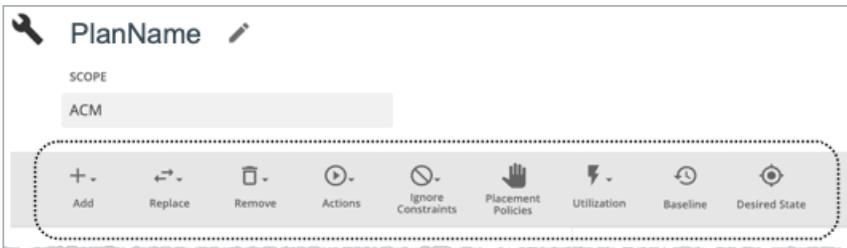
1. Scope

Specify the plan scope and then click **Configure Plan**.



2. Plan Configuration

Use the Plan Configuration toolbar to fine-tune your plan settings. You can change workload demand and the supply of resources, and specify other changes to the plan market.



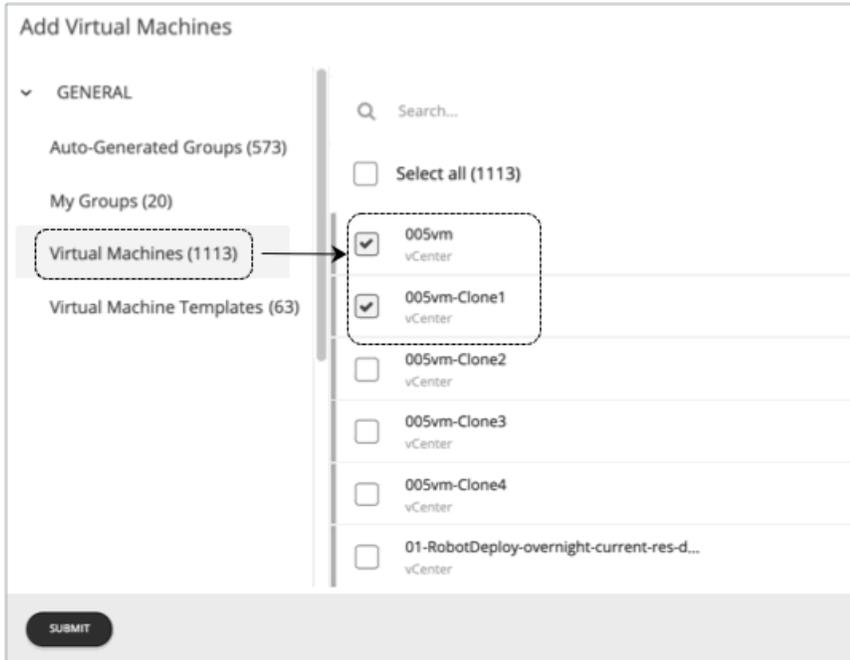
2.1. Add

Add virtual machines, hosts, or storage to your plan. For example, when you add hosts, you increase the compute resources for the plan.

Copy from an Entity or Template

Choose an entity or template to copy. This describes the new entities that Intersight Workload Optimizer will add to the plan. For example, you can run a plan that adds new VMs to a cluster. If you copy from a template, then the plan adds a new VM that matches the resource allocation you have specified for the given template.

- Option 1: Copy from an entity



- Option 2: Copy from a template

If no existing template is satisfactory, create one by clicking **New Template**.



NOTE:

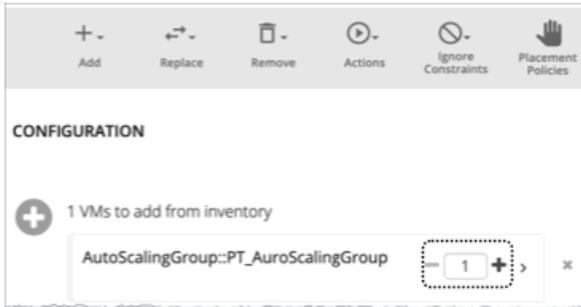
Intersight Workload Optimizer automatically adds any new template you create to the Template Catalog page (**Settings > Templates**).

It is not possible to use templates for containers or container pods.

Use the **Filter** option to show entities or templates with certain properties (name, number of CPUs, etc.). This makes it easier to sort through a long list.

Number of Copies to Add

After choosing an entity or template, it appears as an entry in the Configuration summary. Then you can set how many copies to add.



2.2. Replace

Replacing virtual machine is a way to change the properties of VMs in your plan market. When you replace workload, you select one or more VMs that you want to change, and then you select a template to use in their place. The list of changed VMs displays in the Configuration Summary. You can delete individual entries from the this summary if necessary.

Replacing hosts or storage is a way to plan for a hardware upgrade. For example, if you replace your hosts or datastores with a more powerful template, the plan might show that you can use fewer hosts or datastores, and it will show the best placement for workloads on those entities. You begin by selecting the entities you want to replace, and when you click **REPLACE** you can then choose a template that will replace them. Note that you can only choose a single template for each set of entities you want to have replaced. You can configure different replacements in the same plan, if you want to use more than one template.

2.3. Remove

Removing virtual machines frees up resources for other workloads to use.

Removing hosts or storage means you have fewer compute or storage resources for your workloads. If you think you have overprovisioned your environment, you can run a plan to see whether fewer hosts or less storage can still support the same workload.

2.4. Actions

See the effect of enabling or disabling actions on the entities included in the plan. For example, you might plan for more workload but know that you don't want to add more hardware, so you disable Provision of hosts for your plan. The results would then indicate if the environment can support the additional workload.

2.5. Ignore Constraints

Choose to ignore constraints (such as placement policies) for VMs in your environment. By default, VMs are constrained to the cluster, network group, datacenter, or storage group that their hosts belong to. You can choose to ignore these boundaries.

For example, by default a plan does not consider moving VMs to physical hosts outside of the current cluster. If you disable the Cluster constraint for a VM in your plan, then the plan can evaluate the results of hosting those VMs on any other physical machine within the scope of your plan. If the best results come from moving that VM to a different cluster, then the plan will show that result.

NOTE:

If you are adding hosts to a plan, and use host templates, then you must turn on **Ignore Constraints**.

2.6. Placement Policies

By default, the plan includes all the placement policies that apply to the plan scope. Also, these policies are in their real-time state (enabled or disabled).

Placement Policies ✕

Set up placement policies for the plan ⓘ

FILTER

10 Policies
NEW PLACEMENT POLICY

| | | |
|-------------------------------------|---|---------|
| <input checked="" type="checkbox"/> | GROUP-DRS-AdityaNotOn48-rule/Cluster1/vsphere-dc7.eng.vmturbo.com <small>Place</small> | Enabled |
| <input checked="" type="checkbox"/> | GROUP-DRS-Ah-Rule-Snow-rule/Adv Eng/vsphere-dc12.eng.vmturbo.com <small>Place</small> | Enabled |
| <input checked="" type="checkbox"/> | GROUP-DRS-arsen-separate-vms-rule/Physical/vsphere-dc11.eng.vmturbo.com <small>Place</small> | Enabled |

You can use these settings to enable or disable existing policies, or you can create new policies to apply only to this plan scenario. For information about creating placement policies, see [Placement Policies \(on page 569\)](#).

2.7. Utilization

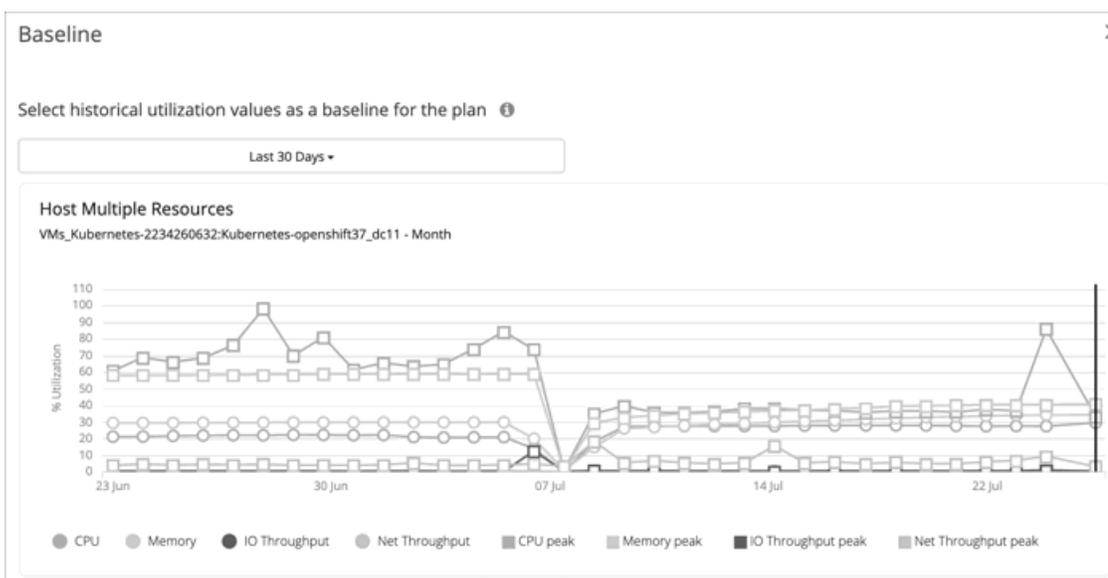
Setting utilization by a certain percentage is a way to increase or decrease the workload for the scope of your plan and any entity added to the plan, or for specific groups. Intersight Workload Optimizer uses the resulting utilization values as the baseline for the plan.

Max Host Utilization levels specify the percentage of the physical resource that you want to make available in the given plan. By default, hosts have utilization set to 100%. For a given plan, you can set the utilization to a lower value. For example, assume you want to simulate High Availability of 25% for some hosts in the plan. In that case, you can select these hosts and set their utilization levels to 75%.

Max Storage utilization levels specify the percentage of the physical resource that you want to make available in the given plan. By default, storage has utilization set to 100%. For a given plan, you can set the utilization to a lower value. For example, assume you have one data store that you want to share evenly for two clusters of VMs. Also assume that you are creating a plan for one of those clusters. In that case, you can set the datastores to 50% utilization. This saves storage resources for the other cluster that will use this storage.

2.8. Baseline

Use these settings to set up the baseline of utilization metrics for your plan.

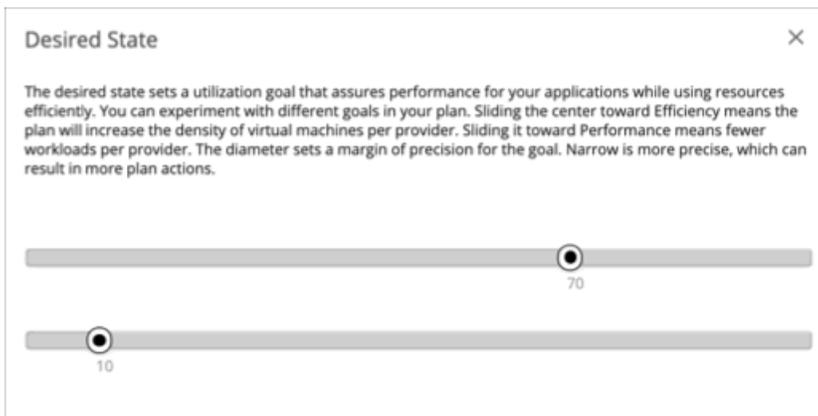


By default, the plan runs against the current state of your environment. You can set up the plan to add or remove entities, or otherwise affect the plan calculations. But the utilization metrics will be based on the current state of the plan. If you run the same plan multiple times, each run begins with a fresh view of your inventory.

You can select from the list of snapshots to load the utilization statistics from a previous time period into the plan. Use this to run the plan against utilization that you experienced in the past. For example, assume a peak utilization period for the month before the winter holidays. During the holidays you want to plan to add new capacity that can better handle that peak. You would set the baseline to the utilization you saw during that pre-holiday peak.

2.9. Desired State

The desired state is a condition in your environment that assures performance for your workloads, while it utilizes your resources as efficiently as possible and you do not overprovision your infrastructure. Intersight Workload Optimizer uses default Desired State settings to drive its analysis. You should never change the settings for real-time analysis unless you are working directly with Technical support. However, you can change the settings in a plan to see what effect a more or less aggressive configuration would have in your environment.



You can think of the desired state as an n-dimensional sphere that encompasses the fittest conditions your environment can achieve. The multiple dimensions of this sphere are defined by the resource metrics in your environment. Metric dimensions include VMem, storage, CPU, etc. While the metrics on the entities in your environment can be any value, the desired state, this n-dimensional sphere, is the subset of metric values that assures the best performance while achieving the most efficient utilization of resources that is possible.

The Desired State settings center this sphere on Performance (more infrastructure to supply the workload demand), or on Efficiency (less investment in infrastructure to supply the workload demand). The settings also adjust the diameter of the sphere to determine the range of deviation from the center that can encompass the desired state. If you specify a large diameter, Intersight Workload Optimizer will have more variation in the way it distributes workload across hosting devices.

For more information, see [The Desired State \(on page 14\)](#).

Working With Custom Plan Results

After the plan runs, you can view the results to see how the plan settings you configured affect your environment.

The screenshot shows the 'Custom 1' configuration page in the Cisco Intersight Workload Optimizer. The 'SCOPE' is set to 'ACM'. A toolbar contains icons for Add, Replace, Remove, Actions, Ignore Constraints, Placement Policies, Utilization, Baseline, and Desired State. The 'CONFIGURATION' pane on the left shows 'Add 1 Storage from templates' with a 'Small' instance selected. The main area displays 'RESULTS OVERVIEW' and 'PLAN ACTIONS (107)'. A 'Plan Summary' table compares current resources to resources after the plan is executed.

| | Current | After Plan | Difference | % |
|------------------|-----------|------------|------------|----------|
| Virtual Machines | 86 | 86 | 0 | 0 % |
| Hosts | 4 | 3 | 1 | ▼ 25 % |
| Storage | 3 | 4 | 1 | ▲ 33.3 % |
| CPU | 64 Cores | 64 Cores | 0 | 0 % |
| Memory | 512 GB | 512 GB | 0 GB | 0 % |
| Storage Amount | 8316.5 TB | 8317.5 TB | 1 TB | 0 % |
| Host Density | 22:1 | 29:1 | 7:1 | ▲ 31.8 % |
| Storage Density | 29:1 | 22:1 | 7:1 | ▼ 24.1 % |

Viewing the Results

The results include the following charts:

■ Plan Summary Chart

This chart compares your current resources to the resources you would get after executing the plan.

NOTE:

Under some circumstances, this chart might not count "non-participating" entities in the real-time market, such as suspended VMs or hosts in a failover state. The following charts, on the other hand, count all entities in the real-time market, regardless of state:

- Scope Preview chart (displays before you run the plan)
- Optimized Improvements and Comparison charts

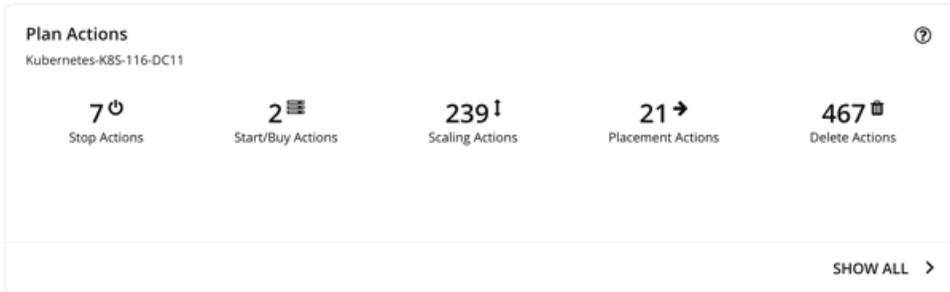
If the plan's scope includes VMs that cannot be placed, the results include a notification indicating the number of VMs. Click **Show Details** to see the list of VMs and the reasons for their non-placement.

Click **Show all** at the bottom of the chart to see savings or investment costs, or to download the chart as a CSV file.

■ Plan Actions Chart

This chart summarizes the actions that you need to execute to achieve the plan results. For example, if you run an Alleviate Pressure plan, you can see actions to move workloads from the hot cluster over to the cold cluster. If some VMs are overprovisioned, you might see actions to reduce the capacity for those workloads.

The text chart groups actions by [action type \(on page 413\)](#). The list chart shows a partial list of [actions \(on page 395\)](#).

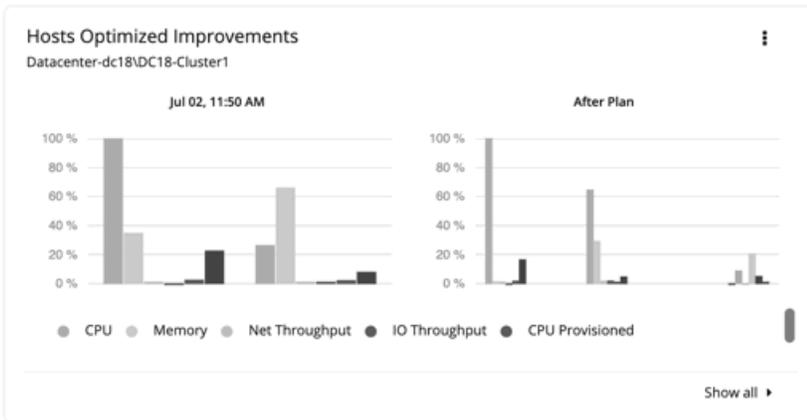


To view action details or download the list of actions as a CSV file:

- Click an action type in the text chart or an individual action in the list chart.
- Click **Show All** at the bottom of the chart.

■ **Optimized Improvements Charts for Hosts, Storage, and Virtual Machines**

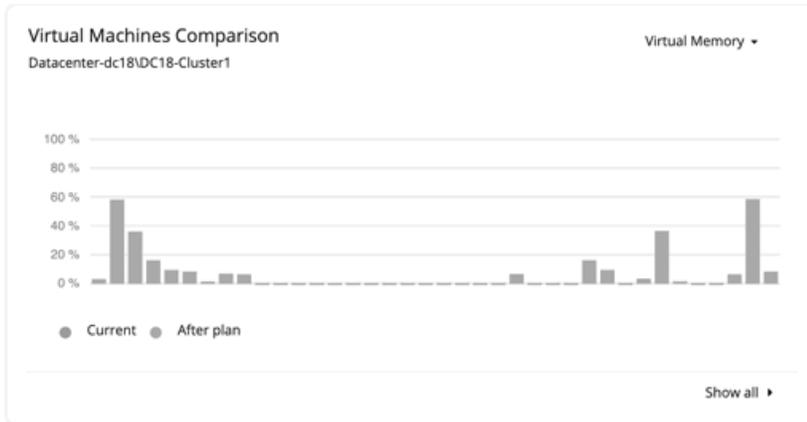
The Optimized Improvements chart shows how the utilization of resources would change assuming you accept all of the actions listed in the Plan Actions chart.



- In many of these charts, you can change the commodities on display. To do this, go to the top-right section of the chart, click the More options icon (⋮), and then select **Edit**. In the new screen that displays, go to the **Commodity** section and then add or remove commodities.
To restore the default commodities, use the **Reset view** option at the top-right section of the page.
- Click **Show all** at the bottom of the chart to see a breakdown of the current chart data by entity (for example, show CPU, Memory, and IO Throughput utilization for each host), or to download chart data as a CSV file.

■ **Comparison Charts for Hosts, Storage Devices, and Virtual Machines**

A Comparison chart shows how the utilization of a particular commodity (such as memory or CPU) for each entity in the plan would change if you execute the actions listed in the Plan Actions chart.



- To change the commodity displayed in the chart, go to the top-right section of a chart and then select from the list of commodities.
To restore the default commodity, go to the top-right section of the page, click the More options icon (), and then select **Reset view**.
- Click **Show all** at the bottom of the chart to show a breakdown of the current chart data by entity (for example, show Virtual Memory utilization for each virtual machine), or to download the chart as a CSV file.

NOTE:

For the Storage Devices Comparison chart, if you set the view to **VM Per Storage** and click **Show all**, the total number of VMs sometimes does not match the number in the Plan Summary chart. This happens if there are VMs in the plan that use multiple storage devices. The Storage Devices Comparison chart counts those VMs multiple times, depending on the number of storage devices they use, while the Plan Summary chart shows the actual number of VMs.

Re-Running the Plan

You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

Use the toolbar on top of the Configuration section to change the configuration settings.



For details about these settings, see [Configuring a Custom Plan \(on page 483\)](#).

NOTE:

It is not possible to change the scope of the plan in the Plan Page. You will need to start over if you want a different scope. To start over, go to the top-right section of the page, click the More options icon (), and then select **New Plan**.

When you are ready to re-run the plan, click **Run Again** on the top-right section of the page.

Configuring Nightly Plans



Intersight Workload Optimizer runs nightly plans to calculate headroom for the clusters in your on-prem environment. For each cluster plan, you can:

1. View the plan and its template.
2. Expand to see template details.
3. Set which VM template to use in these calculations.

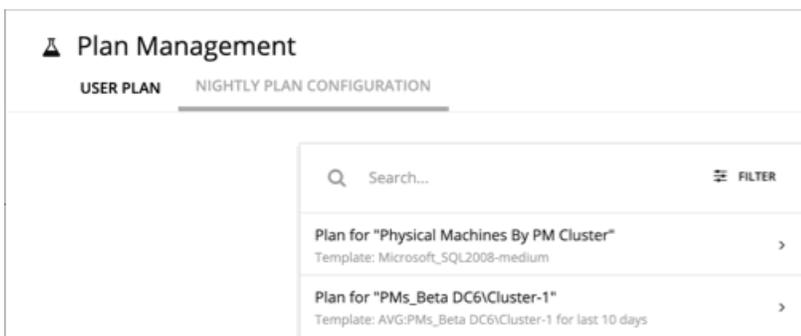
For information about viewing cluster headroom, see [Viewing Cluster Headroom \(on page 41\)](#).

To calculate cluster capacity and headroom, Intersight Workload Optimizer runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current VMs to other hosts within the given cluster, if such moves would result in a more desirable workload distribution. The result of the plan is a calculation of how many more VMs the cluster can support.

To calculate VM headroom, the plan simulates adding VMs to your cluster. The plan assumes a certain capacity for these VMs, based on a specific VM template. For this reason, the count of VMs given for the headroom is an approximation based on that VM template.

To set templates to use for the nightly plans:

1. Navigate to the Plan Page and click **NIGHTLY PLAN CONFIGURATION**.



This displays a list of all the nightly plans. Intersight Workload Optimizer creates a nightly plan for each cluster.

2. Click the plan that you want to configure.
A fly-out appears that lists all the available templates.
3. Select the template you want for this plan.
Choose the template and click **Select**.

Placement: Reserve Workload Resources

From the Workload Placement Page, you can set up reservations to save the resources you will need to deploy VMs at a future date. Intersight Workload Optimizer calculates optimal placement for these VMs and then reserves the host and storage resources that they need.

To reserve VMs, you will need to choose a VM template, specify any placement constraints, set how many instances to reserve, and then indicate whether to reserve now or in the future. Because reserved VMs do not yet exist, they do not participate in the real-time market.

About VM Templates for Reservations

VM templates specify the resource requirements for each reserved VM, including:

- Compute and storage resources that are allocated to each VM
- Consumed factor. This is the percentage of allocated CPU, memory, or storage that the reserved VM will utilize.

For more information about these templates, see [VM Template Settings \(on page 588\)](#).

About Placement of Reserved VMs

To determine the best placement for the VMs you want to reserve, Intersight Workload Optimizer runs a plan that uses the last-generated data in nightly-run headroom plans.

NOTE:

If you change your environment by adding targets or changing policies, wait until the next run of headroom plans for the affected scope before you create reservations.

When making placement decisions, Intersight Workload Optimizer considers the following:

- Placement constraints set in the reservation
- Demand capacity

Intersight Workload Optimizer calculates demand based on the *resource allocation* and *consumed factor* set in VM templates. For example, to create a reserved VM from a template that assigns 3 GB of virtual memory and a consumed factor of 50%, Intersight Workload Optimizer calculates 1.5 GB of demand capacity for the reservation.

- Overprovisioned capacity

For reserved VMs, this corresponds to the resource allocation set in VM templates. Continuing from the previous example, Intersight Workload Optimizer assumes 3 GB of overprovisioned capacity for a reserved VM created from a template that assigns 3 GB of virtual memory.

For providers (hosts and storage), Intersight Workload Optimizer calculates overprovisioned capacity. The default overprovisioned capacity is 1000% for host Mem and CPU, and 200% for storage. A host with 512 GB of memory has an overprovisioned capacity of 5 TB (5120 GB).

Providers must have sufficient *demand* and *overprovisioned* capacity to place a reservation. Intersight Workload Optimizer analyzes the current and historical utilization of cluster, host, and storage resources to identify viable providers for the VMs when they are deployed to your on-prem environment. In this way, Intersight Workload Optimizer can prevent congestion issues after you deploy the VMs.

NOTE:

Intersight Workload Optimizer persists historical utilization data in its database so it can continue to calculate placements accurately when market analysis restarts.

The initial placement attempt either succeeds or fails.

- **Successful Initial Placements**

If the initial placement attempt is successful, Intersight Workload Optimizer adds the reserved VM to your inventory.

In the previous example, a reserved VM that requires 1.5 GB of demand capacity and 3 GB of overprovisioned capacity can be placed on a host with 512 GB of memory (5 TB of overprovisioned capacity), assuming no constraints prevent the placement.

Note that *actual* and *reserved* VMs share the same resources on providers. This means that provider capacity changes as demand from the actual VMs changes. Intersight Workload Optimizer polls your environment once per day to identify changes in provider capacity. It then evaluates if it can continue to place the reserved VMs *within the same cluster*, and then shows the latest placement status.

For example, if the host for a reserved VM is congested at the time of polling, Intersight Workload Optimizer might decide to move the VM to another host in the cluster that has sufficient capacity. In this case, the placement status stays the same (**Reserved**). Should you decide to deploy the VM at that point, you need to deploy it to the new host. If, on the other hand, there is no longer a suitable host in the cluster, the placement fails and the status changes to **Placement Failed**. Deploying the VM at that point results in congestion. Intersight Workload Optimizer does *not* retry fulfilling the reservation.

Reserved VMs are listed on the Workload Placement page. You can also get a list of reserved VMs, with information about each reservation, by making the following API request to the `/reservations` endpoint:

```
GET https://10.10.10.10/api/v3/reservations?status=RESERVED
```

However, reserved VMs are not visible in your application topology in the Intersight Workload Optimizer user interface.

- **Failed Initial Placements**

If the initial placement attempt is unsuccessful (for example, if all providers have seen historical congestion), the Workload Placement page shows that the placement has failed and Intersight Workload Optimizer does *not* retry fulfilling the reservation. You can get a list of VMs for which the placement failed by making the following API request to the `/reservations` endpoint:

```
GET https://10.10.10.10/api/v3/reservations?status=PLACEMENT_FAILED
```

Current and Future Reservations

You can create a current or future reservation from the Workload Placement Page.

- **Current Reservation**

Intersight Workload Optimizer calculates placement immediately and then adds the reserved VMs to your inventory if placement is successful.

This reservation stays in effect for 24 hours, or until you delete it.

- **Future Reservation**

Set the reservation for some time in the future.

Intersight Workload Optimizer does not calculate placement at this time – the future reservation saves the definition, and Intersight Workload Optimizer will calculate placement at the time of the reservation start date.

This reservation stays in effect for the duration that you set, or until you delete it.

Displaying the Workload Placement Page

To see the reservations that are in effect and to create new reservations, click the **PLACE** button in the Navigation Menu.

ALL RESERVATIONS

Workload Placement
Workload Placement

CREATE RESERVATION

| <input type="text" value="search..."/> FILTER | |
|--|---|
| <input type="checkbox"/> | 5 Reservations |
| <input type="checkbox"/> | Cud_GiantVM Cud_GiantVM 2/7/2020 - 3/7/2020 PLACEMENT_FAILED "GiantVM" PLACEMENT F... > |
| <input type="checkbox"/> | CudMultipleVMs 10 "Hatice_VM" placed on HawthorneDev RESERVED > |
| <input type="checkbox"/> | CudRes4 2 "Hatice_VM" placed on HawthorneDev RESERVED > |
| <input type="checkbox"/> | CudRes5 1 "Hatice_VM" placed on HawthorneDev RESERVED > |
| <input type="checkbox"/> | MyReservation MyReservation 2/7/2020 - 3/7/2020 RESERVED "Hatice_VM" RESERVED > |

Creating a Reservation

Reservations set aside resources for anticipated workload. While a reservation is in the RESERVED state, Intersight Workload Optimizer continually calculates placement for the reserved VMs.

To create a reservation:

1. Navigate to the Workload Placement page.
2. Create a new reservation.

CREATE RESERVATION

In the Workload Placement page, click **CREATE RESERVATION**.

Intersight Workload Optimizer displays a list of templates. Choose the template you want, and click **NEXT: CONSTRAINTS**.

3. Optionally, specify placement constraints.

In the **Constraints** section and choose which constraints to apply to this reservation.

Constraints are optional, but note that these constraints are how you ensure that the template you have chosen is viable in the given locations that Intersight Workload Optimizer will choose.

The constraints you can choose include:

- **Scope**
Choose the datacenter or host cluster that you will limit the reservation to.
- **Placement Policy**
This list shows all the placement policies have been created as **Intersight Workload Optimizer Segments**. Choose which placement policies the reservation will respect.
- **Networks**
Intersight Workload Optimizer discovers the different networks in your environment. Use this constraint to limit workload placement to the networks you choose.

When you are done setting constraints, click **NEXT: RESERVATION SETTINGS**.

4. Make the reservation settings, and create the reservation.

To finalize the reservation, make these settings:

- **RESERVATION NAME**

The name for the reservation. You should use unique names for all your current reservations. This name also determines the names of the reservation VMs that Intersight Workload Optimizer creates to reserve resources in your environment. For example, assume the name *MyReservation*. If you reserve three VMs, then Intersight Workload Optimizer creates three reservation VMs named *MyReservation_0*, *MyReservation_1*, and *MyReservation_2*.

- VIRTUL MACHINES COUNT

How many VMs to reserve.

NOTE:

You can include up to 100 VMs in a single reservation.

- RESERVATION DATE

The time period that you want the reservation to be active. Can be one of:

- Reserve Now

Use this to calculate the ideal placement for a workload that you want to deploy today. Intersight Workload Optimizer begins planning the reservation immediately when you click **CREATE RESERVATION**. The reservation stays in effect for 24 hours – At that time Intersight Workload Optimizer deletes the reservation.

- Future Reservation

This executes the reservation for the date range you specify. Intersight Workload Optimizer begins planning the reservation on the day you set for START DATE. The END DATE determines when the reservation is no longer valid. At that time, Intersight Workload Optimizer deletes the reservation.

When you are finished with the reservation settings, click **CREATE RESERVATION**. Intersight Workload Optimizer displays the new reservation in the Workload Placement page. Depending on the reservation settings and your environment, the reservation can be in one of the one of the following states:

- UNFULFILLED

The reservation request is in the queue, waiting for an ongoing reservation request to complete.

- INPROGRESS

Intersight Workload Optimizer is planning the placement of the reservation workloads.

- FUTURE

Intersight Workload Optimizer is waiting for the START DATE before it will start to plan the reservation.

- RESERVED

Intersight Workload Optimizer has planned the reservation, and it found providers for all the VMs in the reservation. As your environment changes, Intersight Workload Optimizer continues to calculate the placement for the reservation VMs. If at any time it finds that it cannot place all the VMs, it changes the reservation to PLACEMENT FAILED.

- PLACEMENT FAILED

Intersight Workload Optimizer cannot place all the reservation VMs. As your environment changes, Intersight Workload Optimizer continues to calculate placement for the VMs. If at any time it finds that it can place all the VMs, it changes the reservation to RESERVED.

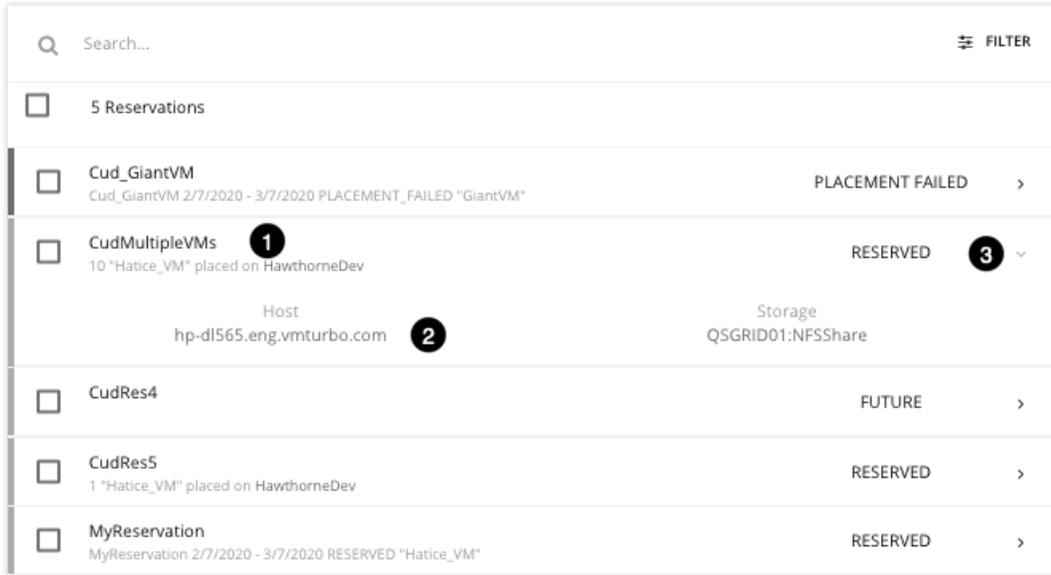
- INVALID

An error occurred while planning the placement of the reservation VMs.

NOTE:

The list of reservations refreshes whenever you open the Workload Placement page. To see changes in reservation state, navigate away from the page, and navigate back to it again.

Managing Reservations



| Search... | FILTER |
|--|--------------------|
| 5 Reservations | |
| <input type="checkbox"/> Cud_GiantVM Cud_GiantVM 2/7/2020 - 3/7/2020 PLACEMENT_FAILED "GiantVM" | PLACEMENT FAILED > |
| <input type="checkbox"/> CudMultipleVMs 10 "Hatice_VM" placed on HawthorneDev | RESERVED > |
| Host: hp-dl565.eng.vmturbo.com Storage: QSGRID01:NFSShare | |
| <input type="checkbox"/> CudRes4 | FUTURE > |
| <input type="checkbox"/> CudRes5 1 "Hatice_VM" placed on HawthorneDev | RESERVED > |
| <input type="checkbox"/> MyReservation MyReservation 2/7/2020 - 3/7/2020 RESERVED "Hatice_VM" | RESERVED > |

The PLACE page displays the current list of reservations.

1. For an entry in the RESERVED state, you can click the entry name to open the Reservation Settings flyout.
2. To see details about the provider entities or the data center that is hosting the reserved VMs, click the entity name.
3. You can expand items in the list to see some details, or you can click to view the full details.
4. To delete a reservation, select it in the list and click the DELETE icon. This cancels the reservation or deployment.

Deploying Workloads to the Reserved Resources

When you reserve resources, you know that they are available for you to deploy actual VMs in your environment.

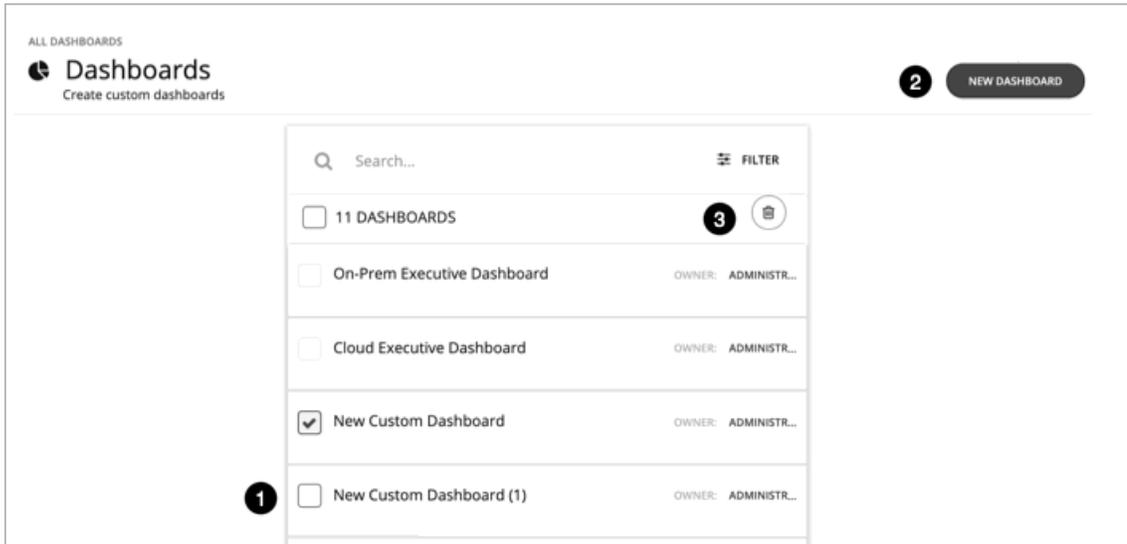
1. Note the placement that your reservation calculated.
Expand the reservation entry in the Workload Placement page and note the hosts and storage that provide resources for your VMs.
2. Delete the reservation.
Before you deploy the reserved VMs, delete the reservation to free up the Intersight Workload Optimizer market to manage the placement of the VMs you are about to deploy.

NOTE:

When you delete a reservation from the user interface or API, Intersight Workload Optimizer only marks the reservation for deletion and waits 48 hours before permanently deleting it. You can permanently delete a reservation by using the API's `reservation_force_delete` parameter along with a DELETE call to a specific reservation. When `reservation_force_delete = true`, the system removes the reservation permanently, no matter what state it is in.

3. Deploy the actual VMs.
In your Hypervisor user interface, deploy the VMs to the hosts and storage that you noted. When you are done, Intersight Workload Optimizer manages their placement the same as it manages the rest of your environment.

Dashboards: Focused Views



Custom views are dashboards to give you views of your environment that focus on different aspects of the environment's health. At a glance, you can gain insights into service performance health, workload improvements over time, actions performed and risks avoided, and savings in cost. For cloud environments, you can see utilization of discounts, potential savings, required investments, and the cost/performance of specific cloud accounts.

The Dashboards page lists all the dashboards that are available to you, including built-in and custom dashboards that your account can access.

1. To view a dashboard, click its name in the list.
2. To create a custom dashboard, click NEW DASHBOARD.
3. To delete a dashboard, select the dashboard and click the delete button

Built-in dashboards give you overviews of your on-prem, cloud, and container environments, showing how you have improved your environment over time.

NOTE:

In charts that show tables, if the table contains more than 500 cells, then the User Interface disables the option to export the chart as PDF. You can still export the chart as a CSV file to load in a spreadsheet.

Built-in Dashboards

Built-in Dashboards are scorecards of your environment. They demonstrate how well you are improving performance, cost, and compliance, as well as opportunities for further improvements that are available.

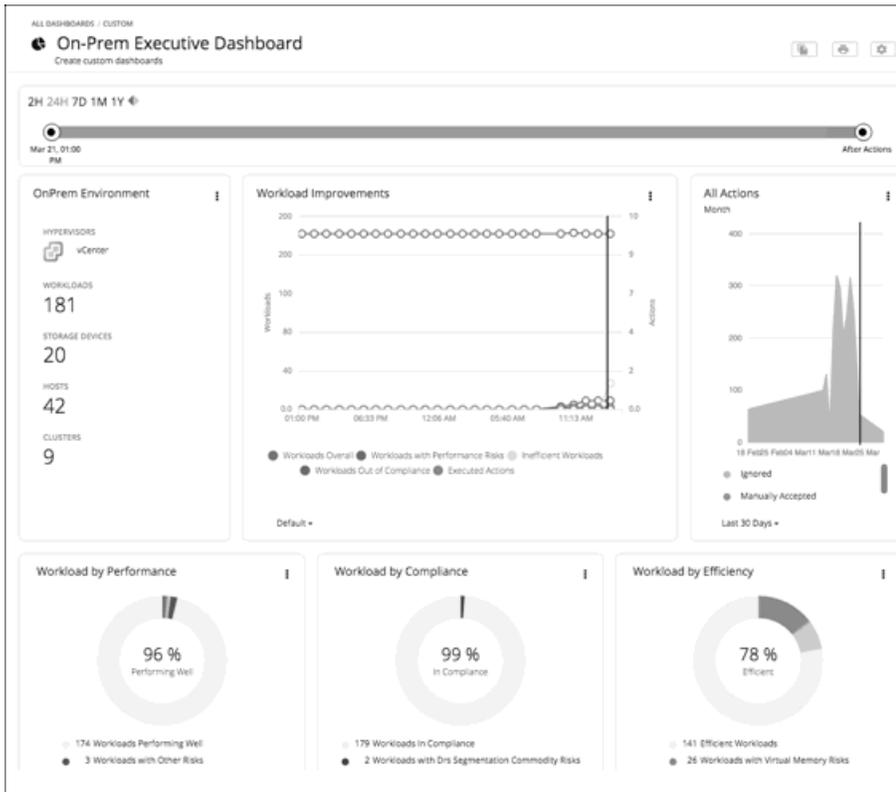
Intersight Workload Optimizer ships with these dashboards:

- On-Prem Executive Dashboard
- On-Prem CPU Ready Dashboard
- Cloud Executive Dashboard
- Container Platform Dashboard

NOTE:

Intersight Workload Optimizer ships these dashboards with default configurations. To edit a dashboard, you must log in with the administrator user account. Users logged in with that account can add or remove chart widgets, and change widget scopes. For information about editing dashboards, see [Creating and Editing Custom Dashboards \(on page 502\)](#).

On-Prem Executive Dashboard



The On-Prem Executive Dashboard shows the overall performance, capacity, and compliance in your on-prem infrastructure. This includes insights into:

- **Actions History**
 - The **On-Prem Environment** chart widget shows you an overview of your on-prem environment that Intersight Workload Optimizer is managing and controlling. The chart displays the workloads and the infrastructure that Intersight Workload Optimizer discovered.
 - The **Workload Improvements** chart widget shows how the efficiency, performance, and policy risks associated with your workloads have disappeared as you have increased your adoption of Intersight Workload Optimizer Workload Automation. The chart tracks how your workloads have grown as your execution of actions have increased or decreased as your environment achieves and maintains its desired states over time.
 - The **All Actions** chart widget shows the number of actions that Intersight Workload Optimizer has generated versus the ones executed. This gives you an understanding of where there were more opportunities for improvement that were not taken in the past versus those that are available today.
- **Opportunities**
 - The **Workload by Performance**, **Workload by Compliance**, and **Workload by Efficiency** chart widgets indicate workload health by showing the risks that are currently in your environment and each classification of those risks. You can click **Show Action** on the chart to reveal all of the outstanding actions that need to be taken to resolve those risks on your workloads.
 - The **Necessary Investments** and **Potential Savings** chart widgets together project how the current actions to improve performance, efficiency, and compliance will impact your costs.
- **Current State**
 - This chart shows the top host clusters in your on-prem environment by CPU, memory, and storage capacity or utilization. In the default view, the chart shows the top clusters by CPU headroom (available capacity). It also shows time to exhaustion of cluster resources, which is useful for future planning (for example, you might need to buy more hardware).

- The **Virtual Machines vs Hosts and Storage** and the **Virtual Machines vs Hosts and Storage -Density** chart widgets show how your overall density has improved in your on-prem environment. A high count of VMs per host or storage means that your workloads are densely packed.

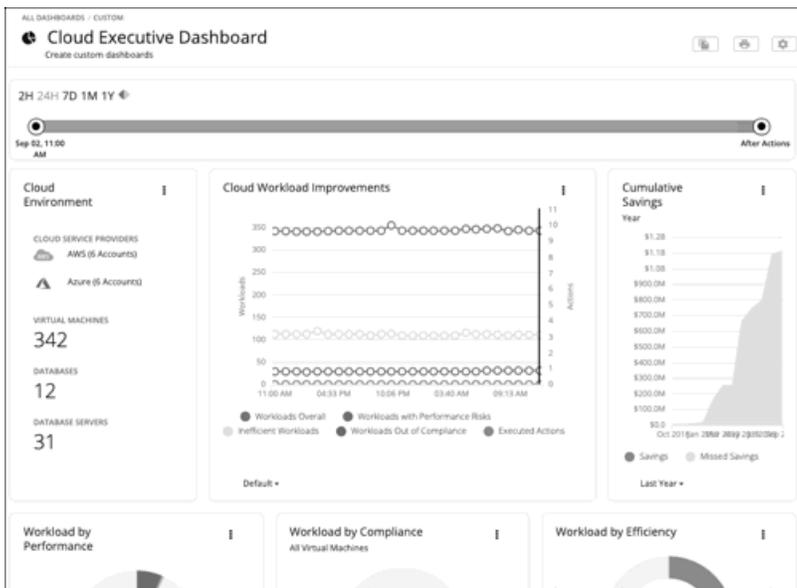
On-Prem CPU Ready Dashboard

The On-Prem CPU Ready Dashboard shows the Host Clusters, Hosts, and VMs with the highest CPU Ready values. Reviewing the charts in this dashboard can help you tune your CPU Ready settings for the specific workloads that are running in your environment. It includes insights into the following details:

- **Host CPU Ready Utilization**
 - The **CPU Ready** chart widget shows the overall average utilization of CPU Ready for all hosts that are discovered by vCenter targets in your environment.
 - The **Top Hosts** chart widget shows the hosts with the highest CPU Ready in your environment.
 - The **Top Host Clusters, BY UTILIZATION** chart widget shows the host clusters with the highest CPU Ready in your environment.
- **Virtual Machine CPU Ready Utilization**
 - The **Top Virtual Machines** chart widget shows the virtual machines with the highest CPU Ready in your environment.
- **Effects of Actions on CPU Ready**
 - The **Hosts Optimized Improvements** chart widget shows a comparison of the CPU Ready peaks for hosts in your environment before and after actions are executed.

For more information about CPU Ready, see [CPU Ready Chart \(on page 520\)](#).

Cloud Executive Dashboard



The Cloud Executive Dashboard shows your overall cloud expenditures and how you can improve performance and reduce cost. This includes insights into:

- **Actions History**
 - The **Cloud Environment** chart widget shows you an overview of your cloud environment that Intersight Workload Optimizer is managing and controlling. The chart displays the workloads, cloud service providers, and cloud accounts that you currently have set up as Intersight Workload Optimizer targets.
 - The **Workload Improvements** chart widget shows how the efficiency, performance, and policy risks associated with your workloads have disappeared as you have increased your adoption of Intersight Workload Optimizer Workload

Automation. The chart tracks how your workloads have grown as your execution of actions have increased or decreased as your environment achieves and maintains its desired states over time.

- The **Cumulative Savings** chart widget shows you the cost savings for executed cloud actions compared to the cloud actions that you have not executed (missed savings).
- Opportunities
 - The **Workload by Performance**, **Workload by Compliance**, and **Workload by Efficiency** chart widgets indicate workload health by showing the risks that are currently in your environment and each classification of those risks. You can click **Show Action** on the chart to reveal all of the outstanding actions that need to be taken to resolve those risks on your workloads.
 - The **Necessary Investments** and **Potential Savings** chart widgets together project how the current actions to improve performance, efficiency, and compliance will impact your costs.
 - **Cloud Estimated Cost** chart widget shows estimated monthly costs and investments for the cloud. Monthly cost amounts are summarized as amounts with and without actions.
- Current State
 - The **Top Accounts** chart widget shows all of the cloud accounts in your cloud environment and what the utilization is for each account. You can see the number of workloads, estimated monthly costs, saved by actions, and actions taken. In the default view, the chart shows the top cloud accounts and you can click **Show All** button to see all of the accounts. In the Show All list, you can also download the account cost data as a CSV file or PDF.
 - The **Cost Breakdown by Tag** chart widget shows the tags you have assigned to your cloud resources and the costs associated with each of these tagged categories. The **Cost Breakdown by Cloud Service Provider** chart widget is an Expenses chart widget that shows your expenses for each cloud service provider.
 - Usage of Discounts

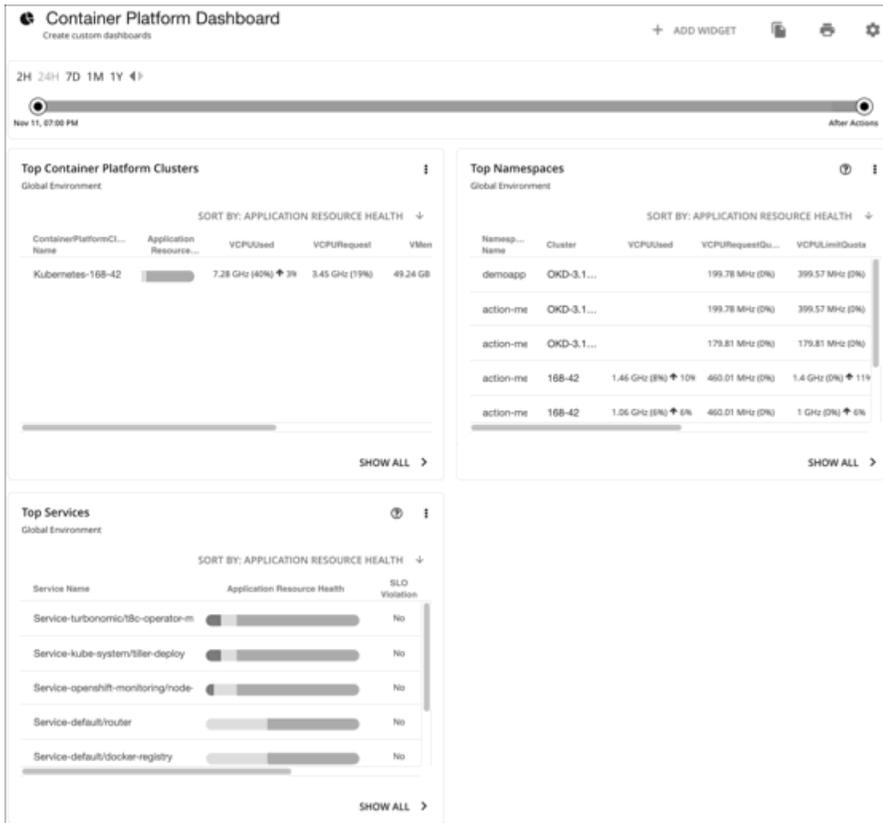
Discounts reduce cost by offering a subscription-based payment plan. Intersight Workload Optimizer discovers these discounts and tracks usage patterns to identify workloads that can take advantage of discounted pricing. The Cloud Executive Dashboard shows whether you are getting the most out of your current discounts.

 - [Discount Coverage \(on page 557\)](#)

This chart shows the percentage of cloud workloads (VMs and RDS database servers) covered by discounts. For VMs covered by discounts, you can reduce your costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.
 - [Discount Inventory \(on page 560\)](#)

This chart lists the cloud provider discounts discovered in your environment.

Container Platform Dashboard



The Container Platform Dashboard shows the overall performance, capacity, and health of your container infrastructure. This includes insights into:

- **Top Container Platform Clusters**
Assess the health of your clusters and sort them by risk level.
- **Top Namespaces**
Identify namespaces that are running out of quota, and how much resources each namespace is using in both quotas and actual utilization.
- **Top Services**
Assess the impact of Services on the performance of your applications.

Creating and Editing Custom Dashboards

A custom dashboard is a view that you create to focus on specific aspects of your environment. You can create dashboards that are private to your user account, or dashboards that are visible to any user who logs into your Intersight Workload Optimizer deployment.

Two common approaches exist for creating custom dashboards:

- **Scope First**
You can create a dashboard in which all of the chart widgets focus on the same scope of your environment. For example, you might want to create a dashboard that focuses on costs for a single public cloud account. In that case, as you add chart widgets to the dashboard, you give them all the same scope.
- **Data First**
You might be interested in a single type of data for all groups of entities in your environment. For example, each chart widget in the dashboard can focus on Cost Breakdown by Cloud Service, but you set the scope of each chart widget to a different cloud region or zone.

Of course, you can mix and match, according to your needs. You can set any scopes or data sources to the chart widgets in a dashboard to set up whatever organization and focus that you want.

NOTE:

If you set a scope to your Intersight Workload Optimizer session, the specified scope does not affect your custom dashboards. For information about scoped views, see [Working With a Scoped View \(on page 31\)](#).

Creating a Dashboard

To create a custom dashboard:

1. Navigate to the Dashboards Page.

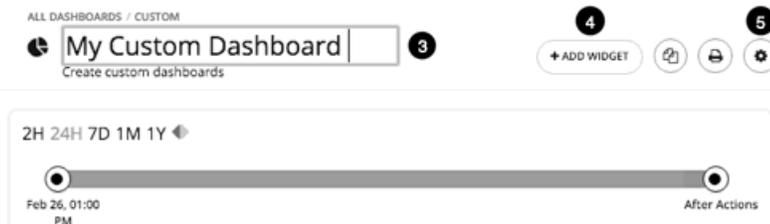
Click **More**, then display the Dashboards Page.

This page lists all dashboards that are available to you. To view a dashboard, click its name in the list.

2. Create a new dashboard.



Click **NEW DASHBOARD** to add a new dashboard to your Intersight Workload Optimizer session. The dashboard appears with a default name and without chart widgets. The time range in the Time Slider is set to 24 hours by default.



3. Name the dashboard.

Give a name that describes the dashboard. If you will share the dashboard with all Intersight Workload Optimizer users, the name will help them decide whether to view it.

4. Add chart widgets to the dashboard.



Add as many chart widgets to the dashboard as you want. See [Creating and Editing Chart Widgets \(on page 504\)](#).

5. Optionally, set the dashboard access.

Click **Gear** to change the setting.

Dashboard access can be:

- **Only Me** – (default) The dashboard is only available to your Intersight Workload Optimizer user account.
- **All Users** – Every Intersight Workload Optimizer user can see this dashboard.

As soon as you create a new dashboard, it appears in the list on the Dashboard Page. Users with access to it can click the dashboard name in the list to view it, create a copy of the dashboard, or print the dashboard.

At any time, if you are an administrator or the dashboard owner, you can view and make the following changes to the dashboard:

- Add, edit, or delete widgets
- Change the dashboard name
- Change the dashboard access setting

For executive dashboards, only an administrator (username=administrator) can edit an executive dashboard.

Editing a Dashboard

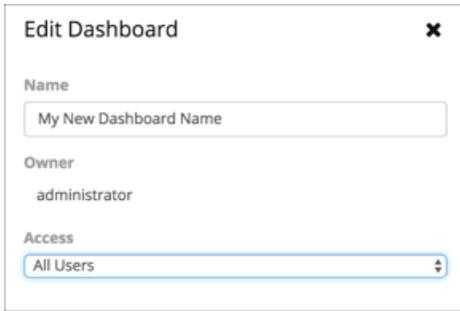
If you have created a dashboard, you can change the name of the dashboard, its access settings, and its chart widgets. To change the chart widgets, see [Creating and Editing Chart Widgets \(on page 504\)](#).

1. Navigate to the Dashboards Page.

Click **More**, then display the Dashboards Page.

2. Click the name of the dashboard that you want to edit.
3. Click **Gear** in the dashboard.

In the Edit Dashboard flyout, change the dashboard name or set the dashboard access.



For the dashboard's access, you can set:

- **Only Me** – The dashboard is only available to your Intersight Workload Optimizer user account.
 - **All Users** – Every Intersight Workload Optimizer user can see this dashboard.
4. When you are done, close the panel.
Your changes take effect when you close the panel.

Deleting a Dashboard

If you are an administrator or the dashboard owner, you can delete a custom dashboard. You cannot delete executive dashboards.

1. Navigate to the Dashboards Page.

Click **More**, then display the Dashboards Page.

This page lists all dashboards that are available to you.

2. Delete one or more dashboards.

In the list, choose the check box for each dashboard to delete and click **Trash can**.

Creating and Editing Chart Widgets

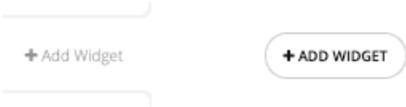
Intersight Workload Optimizer displays information about your environment in various chart widgets. To focus on the information you need, you can add new chart widgets to scoped views and dashboards, and you can edit existing chart widgets. You can also pull the corners of chart widgets to resize them and change the display order of chart widgets in dashboards.

When you create or edit a chart widget, you can choose various settings. For example, in the Top Utilized chart widget, if you choose Host Clusters as the Entity Type, you can then choose Utilization as the Data Type and Storage Provisioned as the Commodity.

Creating a Chart Widget

To create a new chart widget:

1. Click **Add Widget** to open the Widget Gallery.



On a dashboard, click **Add Widget** at the upper-right corner. In a scoped view, click **Add Widget** on the right above the charts.

2. Choose a chart widget in the Widget Gallery.

The Widget Gallery is a list of thumbnail previews of chart widgets.

You can scroll through the gallery or search it. For example, if you type "Health" in the **Search** field, the results are two chart widgets, Health and Workload Health. You can choose chart widgets from these categories:

- Actions and Impact
- Status and Details
- Cloud
- On-Prem

To see the possible displays of a specific chart widget, use the horizontal scroll bar at the bottom of the thumbnail to scroll through the display choices.

To choose a chart widget to add it to your dashboard, click the thumbnail preview.

The Widget Preview window with the Edit flyout opens.

3. Configure the settings for your chart widget.

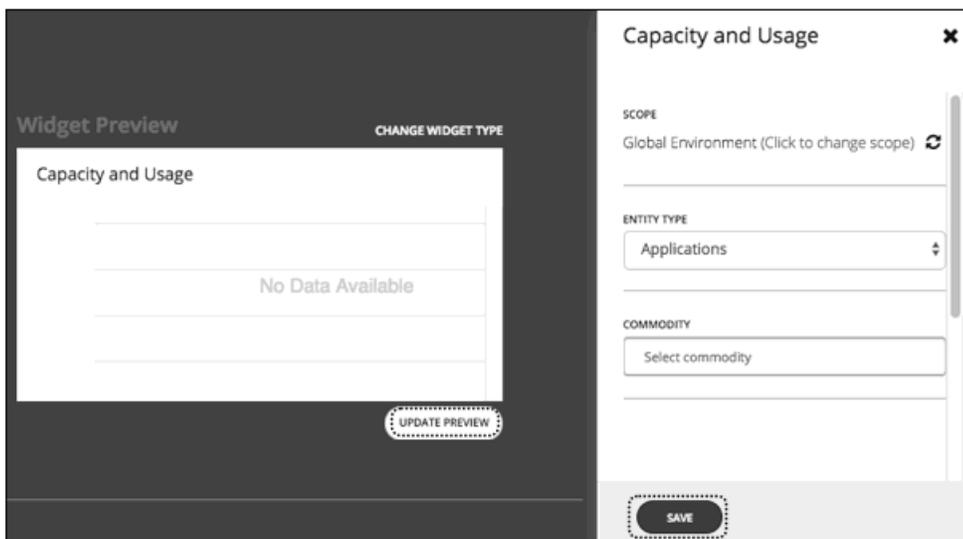
Chart widget settings determine the data that the chart widget shows.

In the Edit flyout, choose the settings and click **Update Preview** to display the result in the Widget Preview window.

When you are satisfied with your settings, click **Save**. The chart widget is added to your dashboard.

For information about settings, see [Chart Widget Settings \(on page 506\)](#).

For example:



To delete a chart widget from your dashboard, choose **Delete** in the More options menu at the upper-right corner of the chart widget.

Methods to Access Chart Widget Settings

Two methods exist for accessing the chart widget settings in the Edit flyout:

- You can access the settings in the Edit flyout when you add a chart widget to your dashboard after you click a thumbnail preview.

- For an existing chart widget in a dashboard, you can choose **Edit** in the More options menu.

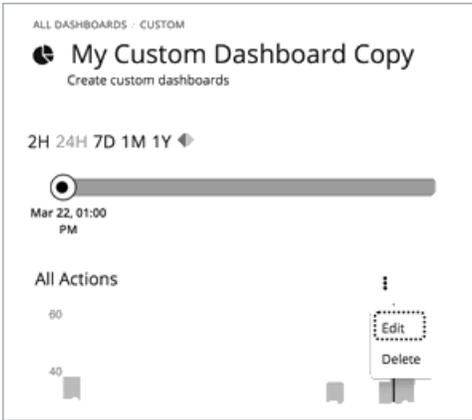


Chart Widget Settings

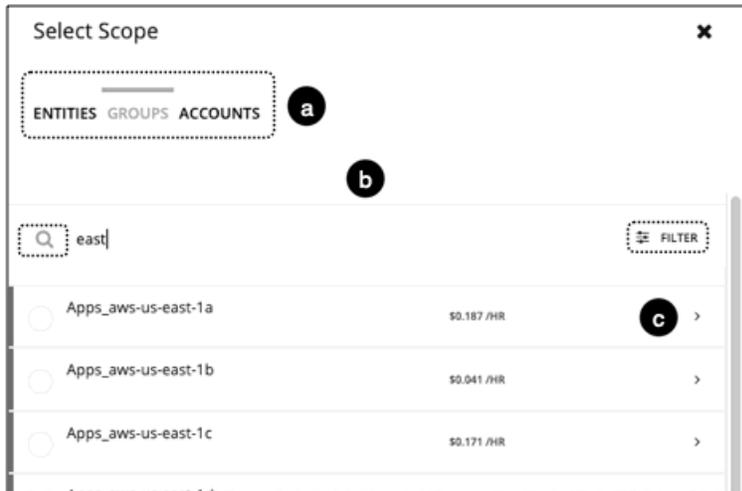
Chart widget settings vary according to the type of chart widget. Also, depending on the value that you choose for a setting, more settings appear. The following is a list of frequently used chart widget settings:

- Scope**

The set of entities in your environment that this chart widget represents. By default, the chart widget scope is set to **Global Environment**.

For every type of chart widget, you have the option to set the chart's scope. To do so:

- Click **Click to change scope** to open the Select Scope flyout.
- In the Select Scope flyout you can:
 - Select the scope for the chart: entities, groups, or accounts.
The ACCOUNTS tab is available depending on the type of chart widget.
Your choice appears in the **Scope** field.
 - Search or refine results with a filter.
 - Display details for the list of entities, groups or accounts.



- Timeframe**

The timeframe for historical data or projections in the chart. Choices for the chart's timeframe are: Default, Last 2 Hours, Last 24 Hours, Last 7 Days, Last 30 Days, and Last Year.

If you set the timeframe to **Default**, the dashboard Time Slider controls the timeframe setting. For example, if your dashboard Time Slider is set to one month (1M), then all chart widgets with the Default timeframe in that dashboard are

set to one month and show information for one month. Note that the dashboard Time Slider does not override the other specific timeframe settings.

- Chart Type

The chart widget display type. Most chart widgets can display horizontal bar or ring charts. Other display choices can include tabular data, band chart, stacked bar, line, or area charts.

NOTE:

For summary charts like horizontal bar and ring charts, when the legend has more than four categories, the remaining categories are represented as a fifth category named "Other."

- Entity Type

The type of entities or their data that you want to display in this chart widget. Choices vary (for example, Applications, Hosts, Virtual Data Centers, Storage Devices, and so on).

- Commodity

The resources that you want this chart widget to monitor. Some charts can monitor multiple commodities. Choices vary (for example, CPU, Memory, Virtual Storage, and so on).

Chart Types

Intersight Workload Optimizer provides many different types of charts in the Widget Gallery. To design dashboards, you should be familiar with the data each chart presents. These charts provide information on actions, impact, status of your environment, and details about specific entities, cloud, and on-prem environments.

Actions and Impact Chart Types

These chart widgets provide information on actions, pending actions, risks that you avoided, improvements, and potential savings or investments.

Pending Actions Charts

Pending Actions charts show the actions that Intersight Workload Optimizer recommends to improve the current state of your environment.

Chart Type

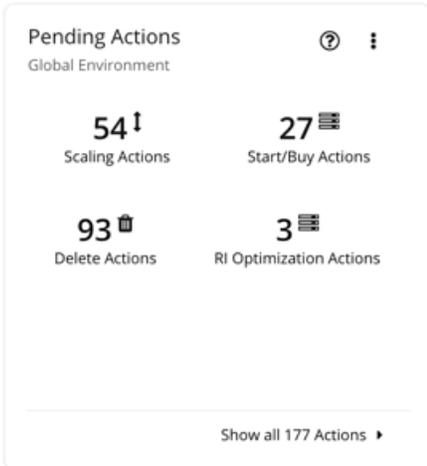
You can set the display to:

- Text
- Ring Chart
- Horizontal Bar
- List

Examples:

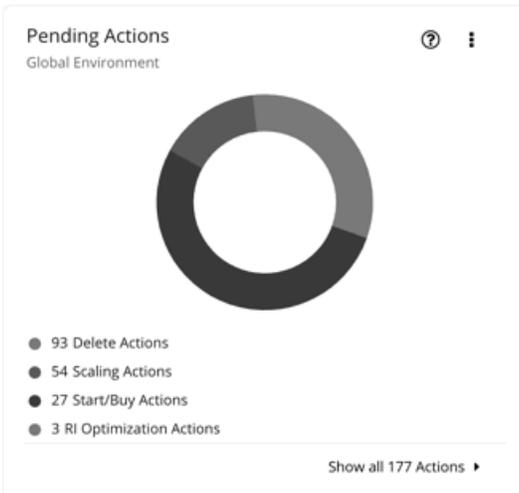
- Text

The text chart shows the number of actions for each action type. It gives a quick visual indication of the kinds of actions that are pending. For details, see [Action Types \(on page 413\)](#).



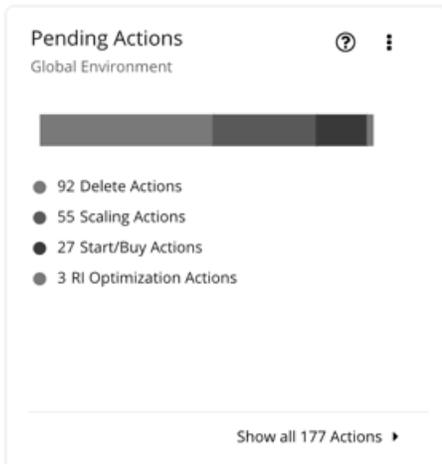
■ Ring Chart

The ring chart counts the number of actions for each action type. It gives a quick visual indication of the kinds of actions that are pending. For details, see [Action Types \(on page 413\)](#).



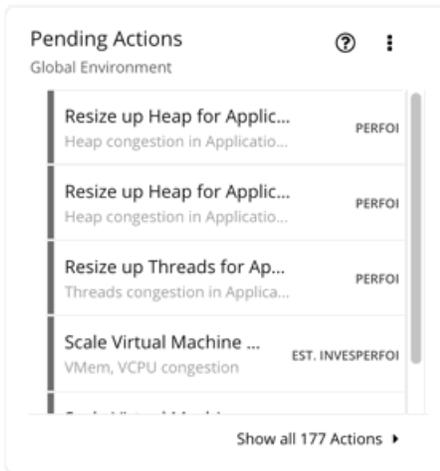
■ Horizontal Bar

The horizontal bar chart counts the number of actions for each action type. It gives a quick visual indication of the kinds of actions that are pending. For details, see [Action Types \(on page 413\)](#)



■ List

The list chart shows an abbreviated listing of the actions for the chart's scope. For details about the different actions generated by the product, see [Actions \(on page 395\)](#).



At the bottom of the chart, click **Show All** to see a full list of pending actions that are in the scope of the chart, along with action details and controls to execute actions. For details, see [Working with Action Center \(on page 396\)](#).

Actions Charts

Actions charts keep a history of pending (not executed) and executed actions. These charts use historical data from the Intersight Workload Optimizer database. You can set the chart to show hourly, daily, or monthly data points.

Filter

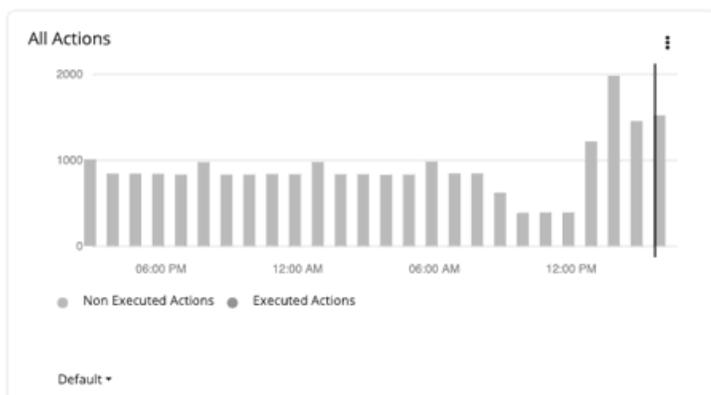
You can filter the chart to show **All Actions** (pending and executed actions) or only **Executed Actions**.

Chart Type

You can set the display to:

- Tabular
- Area Chart
- Text
- Stacked Bar Chart

Stacked Bar Chart



For the Stacked Bar Chart, each bar represents a time period. Hover over the bar to see the number of *unique* actions for that time period. In the default view, the bars represent actions per hour in the last 24 hours. The 2:00 PM bar, for example, shows actions between 2:00 PM and 2:59 PM.

A pending action that remains valid for an extended period of time is counted *once* for each hour, day, and month it remains pending. This also applies to pending actions that go away as conditions in the market change, but are generated again in the future. Once a pending action is executed, it is counted once (this time, as an executed action) on the hour, day, and month of execution.

Consider the following scenarios.

- An action was generated at 1:25 PM and then executed two hours later at 3:25 PM.
 - For per-hour views (Last 24 Hours or Default), the action will be counted three times – as a *pending* action in the 1:00 PM and 2:00 PM bars, and as an *executed* action in the 3:00 PM bar.
 - For per-day (Last 7 or 30 Days) or per-month (Last Year) views, the action will be counted once (as an executed action) on the day or month of execution.
- An action was generated at 6:20 PM but went away (without being executed) in the next hour. The same action was generated again the next day at 9:10 AM and was executed immediately.
 - For per-hour views, the action will be counted twice – as a pending action in the 6:00 PM bar and as an executed action in next day's 9:00 AM bar.
 - For per-day views, the action will also be counted twice – as a pending action on Day 1 and an executed action on Day 2.
 - For per-month views, the action will be counted once (as an executed action) on the month of execution.

Use the chart to evaluate the rate of action execution, which underscores the importance of executing actions in a timely manner. As pending actions persist, they become more challenging to track and your environment stays in a risky state longer. To reduce potential delays in executing actions, consider action automation.

Tabular Chart

To see the full list of actions, click **Show All** at the bottom of the chart.

| All Actions | | | | |
|-------------------------|--|-----------------------|-------------|---------------|
| DATE CREATED | ACTION DESCRIPTION | RISK TYPE | EXECUTION | DATE EXECUTED |
| 19 Oct 2018 17:25 PM | Provision PhysicalMachine dc17-host-01.eng.vmturbo.com | Performance Assurance | Recommended | N/A |
| 19 Oct 2018 17:25 PM | Provision PhysicalMachine dc17-host-01.eng.vmturbo.com | Performance Assurance | Recommended | N/A |
| 19 Oct 2018 17:25 PM | Provision PhysicalMachine dc17-host-01.eng.vmturbo.com | Performance Assurance | Recommended | N/A |
| 19 Oct 2018 17:25 PM | Provision PhysicalMachine dc17-host-01.eng.vmturbo.com | Performance Assurance | Recommended | N/A |

Default ▾ Show all ▶

Risks Avoided Charts

As you execute the actions Intersight Workload Optimizer has recommended, you improve your environment's health and avoid risks to performance or cost. These charts show how many risks you have avoided over time. For example, the charts can show how many over-provisioning and congestion risks you avoided.

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Optimized Improvements Charts

Intersight Workload Optimizer automatically executes or recommends actions, depending on the policies that you set up. For the recommended actions, you can use Optimized Improvements charts to show how utilization of resources would change assuming you accept all of the [pending actions \(on page 507\)](#).

Entity Type

Entity types you can choose include:

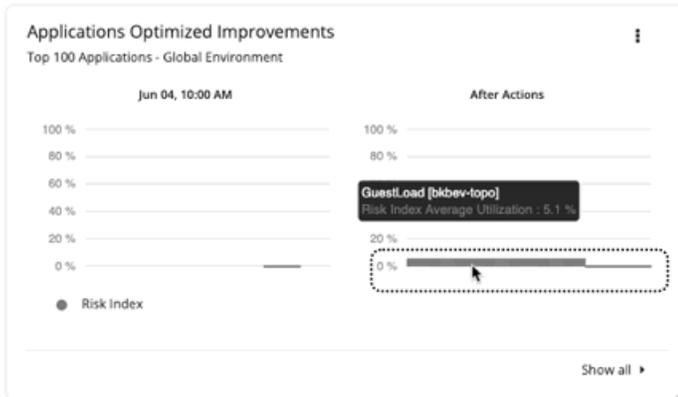
- Business Applications
- Business Transactions
- Services
- Application Components
- Chassis
- Containers
- Container Pods
- Container specs
- Namespaces
- Workload Controllers
- Data Centers
- Databases
- Database Servers
- Disk Arrays
- IO Modules
- Internet
- Logical Pool
- Networks
- Hosts
- Regions
- Storage Devices
- Storage Controllers
- Switches
- Virtual Data Centers
- Virtual Machines
- Volumes
- Zones

Commodity

Depending on the entity type, you can add different resource commodities that you want to measure. For a chart of Hosts, you can measure commodities such as CPU and Memory.

Display

Optimized Improvements charts show two bar charts for the entities that are in scope – one for current consumption, and the other for the consumption you would expect to see if you execute all the actions. You can hover on the graph for details in a tooltip.



Potential Savings or Investments Charts

These charts show potential savings or necessary investments in your cloud expenditure, assuming you execute all the pending actions that Intersight Workload Optimizer identifies as a result of its analysis.

For example, if some workloads risk losing performance, Intersight Workload Optimizer might recommend scaling actions for the virtual machine to increase resources. The Necessary Investments chart shows how these actions translate to an increase in expenditure.

On the other hand, if there are pending actions to scale a virtual machine, which result in reduced monthly costs, the Potential Savings chart shows the reduced cost that would result from those actions.

This chart also tracks discount optimization actions. VM scaling actions may result in freed up capacity on a discounted instance type, which can now be applied to a different VM. Discount optimization actions reflect the potential savings resulting from reassigning the capacity to a different VM. These actions are not executed by Intersight Workload Optimizer users. They reflect capacity reassignment performed by your cloud provider.

The projected amounts include on-demand costs for VMs. For information about on-demand cost calculations, see [Estimated On-demand Monthly Costs for Cloud VMs \(on page 273\)](#).

Type

You can choose **Potential Savings** or **Necessary Investments**.

Chart Type

You can set the display to:

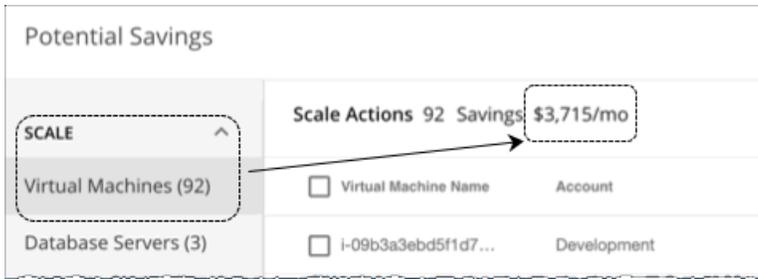
- Text
- Ring Chart
- Horizontal Bar

For the ring chart, you can click an action type (for example, **Scale Volumes**) in the chart or legend to display a filtered view of the actions list.

Show All

Click **Show all** at the bottom of the chart to see a breakdown of savings or investments by action/entity type and entity. By default, the actions are shown in order of largest amounts so you can easily identify which ones will incur the highest costs or introduce the most savings.

For example, you can see the savings you would realize if you execute all *Scale* actions on the *virtual machines* included in the chart's scope.



The table then breaks down the total savings by individual virtual machines, and includes links to the specific actions that you need to perform to realize those savings.

| Potential Savings | | Scale Actions 92 Savings \$3,715/mo | | | | | | | | DOWNLOAD |
|-----------------------|---|-------------------------------------|---------------|-------------|----------------|-------------------|-----------------|--------------------|----------|----------|
| SCALE ^ | Virtual Machine Name | Account | Instance Type | RI Coverage | On-Demand Cost | New Instance Type | New RI Coverage | New On-Demand Cost | Savings | Action |
| Virtual Machines (92) | <input type="checkbox"/> | | | | | | | | | |
| Database Servers (3) | <input type="checkbox"/> i-09b3a3ebd5f1d7... | Dev | t2.xlarge | 0% | \$0.371/hr | m5.2xlarge | 100% | \$0/hr | \$271/mo | DETAILS |
| BUY ^ | <input type="checkbox"/> eks-cluster-eks-w... | Advanced | t2.xlarge | 0% | \$0.371/hr | m5.2xlarge | 100% | \$0/hr | \$271/mo | DETAILS |

You can also compare instance types, costs, and discount coverage before and after executing the actions, allowing you to easily identify actions with the most savings.

Status and Details Chart Types

These chart widgets provide information on the status of your environment and details about specific entities.

Health Charts

Health charts show the current status of your environment, by entity type. For example, you can choose to show the health of all hosts in your environment, or the health of all the workloads running on a public cloud region.

Entity Type

Entity types you can choose include:

- Business Applications
- Business Transactions
- Services
- Application Components
- Chassis
- Containers
- Container Pods
- Container specs
- Namespaces
- Workload Controllers
- Data Centers
- Databases
- Database Servers
- Disk Arrays
- IO Modules
- Internet
- Logical Pool

- Networks
- Hosts
- Regions
- Storage Devices
- Storage Controllers
- Switches
- Virtual Data Centers
- Virtual Machines
- Volumes
- Zones

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Basic Info Charts

The Basic Info charts provide an overview of a single entity or individual Azure resource group that you have chosen as your scope.

Type

You can choose:

- **Entity Information**

This chart shows basic information (ID, Name, Type, State, Severity, Target Name, and so on) for the scoped entity or Azure resource group.

- **Related Tag Information**

This chart lists any available tag information for the scoped entity or Azure resource group. For example, in a cloud environment, if a virtual machine has tags applied to it, the chart shows those tags for the virtual machine.

Display

The chart shows the information as Tabular.

Capacity and Usage Charts

These charts list the resources for the selected entity type, showing the source of the target metric data, their allocated capacity, and how much of that capacity is in use.

Entity Type

Entity types that support this chart include:

- Business Applications
- Business Transactions
- Services
- Application Components
- Containers
- Container Pods
- Container specs
- Namespaces

- Workload controllers
- Data Centers
- Database Servers
- Disk Arrays
- Logical Pool
- Networks
- Hosts
- Regions
- Zones
- Storage Devices
- Storage Controllers
- Virtual Machines
- Volumes

Commodity

Depending on the entity type, you can add different resource commodities that you want to measure. For example, for a chart of Virtual Machines, you can measure commodities such as virtual CPU, memory, and storage.

NOTE:

For a cloud database server, the chart might show incorrect *used* values for vMem and Storage Amount after an action executes. It could take up to 40 minutes for the correct values to display.

Commodity Source

You can view the source of every commodity collected by a target in Intersight Workload Optimizer. This chart shows the source of both the available capacity and the utilization of these commodities across all entities in the Intersight Workload Optimizer environment in the **Capacity Source** and **Used Source** columns. The available sources for the commodity components include:

- Target or Probe name - The source is its parent target. You will see the parent target name (such as vCenter) or "Target" if the data source is not stitched and it originates from one target source.
- Calculation - The source is calculated. See [Calculation \(on page 515\)](#) for more details.
- Policy - The source is a policy setting. See [Policy \(on page 516\)](#) for more details.

Calculation

Review the descriptions in the table below to learn more about the algorithm used to derive the calculated value for the commodity.

| Calculation | Description |
|---|--|
| Used value for the number of used virtual cores | Sum of all used values from the consumers of the commodity. |
| Used value of the VCPU cores commodity | Average of all used values from the consumers of the commodity. The peak value is derived from the maximum values from the consumers of the commodity. |
| Used value of the Memory Provisioned and CPU Provisioned commodities | Sum of all used values from the consumers of the commodities. |
| Capacity of the Storage Provisioned, CPU Provisioned, Memory Provisioned, Memory Allocation, and CPU Allocation commodities | <p>Multiply the overprovisioned percentage of a commodity by the source commodity capacity.</p> <p>The calculation can be expressed as follows:</p> $\text{Commodity Overprovisioned Percentage} * \text{Source Commodity Capacity}$ |
| Used value of the Storage Access and Storage Latency commodities (without a storage target) | For Storage Access: Sum of all used values from the consumers of the commodity without a storage target. |

| Calculation | Description |
|---|---|
| | For Storage Latency: Average of all used IOPS-weighted values from the consumers of the commodity without a storage target. |
| Used value of the Storage Access and Storage Latency commodities | For Storage Access: Sum of all used values from the consumers of the commodity. For Storage Latency: Average of all used IOPS-weighted values from the consumers of the commodity. |
| Capacity of the Storage Access and Storage Latency commodities | Derived from the provider of the commodity. |
| Used value of the Response Time commodity | Average of all used values of the number of replicas of a service. |
| Capacity of the Response Time and Transaction commodities | Derived from the database; if the data is not present in the database, the commodity's capacity value is set to the commodity's used value. |
| Used value of the Connection commodity | Derived from the sold commodity. |
| Capacity of the Response Time commodity | Derived from the entity disk counts. |
| Used value of the Storage Provisioned commodity | Sets the used value of the commodity to the capacity value of the Storage Amount commodity if the entity buys a Storage Provisioned commodity and sells the Storage Amount commodity. |
| Used value of the Virtual Memory commodity | Verifies that the sum of all used values of consumers of the commodities is lower than the VMEM provided by the virtual machine. |
| Capacity of the CPU Allocation commodity | Sum of all capacity values of all of the providers of the commodity; if the entity has exactly one provider, it uses the capacity value of a single provider. |
| Used value of the Storage Access, Storage Provisioned, and Storage Latency commodities | Merges the identical entities. |
| Conversion for Virtual CPU, Virtual CPU Request, Virtual CPU Limit Quota, and Virtual CPU Request Quota commodities | Converts the CPU units from MHz to millicore. |
| Capacity of the GPU Memory commodity | For cloud VMs, the amount of GPU memory provided by an individual GPU card (not the sum of all cards). |
| Percentile of the GPU Memory commodity | For cloud VMs, at each point in time, all GPUs are surveyed and the utilization is taken from the card with the highest memory consumption. |
| Used value of the GPU Memory commodity | For cloud VMs, the average memory in use across all GPUs. |

Policy

Review the descriptions in the table below to learn more about the source used to set the policy value for the commodity.

| Policy Setting | Description |
|--|--|
| Capacity of the Response Time and Transaction commodities | Derived from the "Enable SLO" value set in the default policy. |
| Capacity of the Response Time commodity | Derived from the value set in the policy. |
| Capacity of the Response Time commodity for Azure cloud target | Derived from the value set in the policy |
| Capacity of the Storage Access and Storage Latency commodities | Derived from the value set in the policy. |

Multiple Resources Charts

Multiple Resources charts show the historical utilization of commodities for the scoped entity or a group of entities. The vertical bar shows the current moment – plots that extend to the right project utilization into the future.

Entity Type

Entity types you can choose include:

- Business Applications
- Business Transactions
- Services
- Application Components
- Containers
- Container Pods
- Container specs
- Namespaces
- Workload controllers
- Data Centers
- Database Servers
- Disk Arrays
- Logical Pool
- Networks
- Hosts
- Regions
- Zones
- Storage Devices
- Storage Controllers
- Virtual Machines
- Volumes

Commodity

Depending on the entity type, you can add different resource commodities that you want to measure. For example, for a chart of volumes, you can measure commodities such as IO throughput, storage access, and storage amount.

Show Peaks

Edit the chart and choose the **Show Peaks** checkbox to include peak information in the chart.

Display

The chart shows the historical utilization and, if chosen, the peak information as a Line chart.

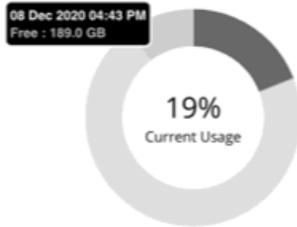
Resources Charts

Resources charts show the utilization of a resource over time, for the entities in the chart's scope. The chart title shows the resource that you are plotting, as well as the chart's current scope.

To see finer details about your environment, you can set up charts that show utilization of specific commodities. For example, you can set up a dashboard with a number of Resources charts with their scopes set to the same cluster. Such a dashboard gives you a detailed look at the health of that cluster. Or you could make a dashboard with each chart scoped to a different cluster, but have all the charts show the same resource utilization.

Ring Chart

For certain entity types (such as hosts, storage, and disk arrays), you will see a ring chart on the left that indicates the current overall utilization of a particular resource. Hover over the ring chart to see the following information:



- Free: Available capacity
- Used: Utilized capacity
- Reserved: Unavailable capacity due to utilization constraints

The sum of *Free* and *Used* capacity equals the total allocated capacity.

In addition to showing the currently discovered free and used capacity, Intersight Workload Optimizer also calculates *Reserved* capacity based on utilization constraints set in policies.

For example, for a cluster with 100 GB of allocated storage, Intersight Workload Optimizer might discover 80 GB of free capacity, and 20 GB of used capacity. If the cluster is currently applying a storage policy that has a utilization constraint of 90%, then Intersight Workload Optimizer will show 10 GB of reserved capacity.

Options

Choose **Show Utilization** to see averages and peaks/lowes, or **Show Capacity** to see averages and peaks/lowes versus capacity.

The **Show Summary** option adds a ring chart to the view, showing the current utilization of the selected commodity.

Chart Type

You can set the following types of display:

- Line Chart
 - A line plot showing resource utilization over time. The vertical green bar shows the current moment – Plots that extend to the right project utilization into the future.
- Band Chart
 - Lines plot average capacity and average used. The chart shows a band where its thickness indicates peaks and lows.

Carbon Footprint Chart

Carbon footprint is the measurement of carbon dioxide equivalent (CO₂e) emissions for a given entity. Intersight Workload Optimizer measures carbon footprint in grams.

Intersight Workload Optimizer collects energy-related data from hosts and VMs at 10-minute intervals, and then uses that data to calculate carbon footprint. When you set the scope to one or several hosts or VM discovered from supported targets, the data that Intersight Workload Optimizer calculated displays in the Carbon Footprint chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Intersight Workload Optimizer also relies on host power data for the supported targets. This data is mandatory for Energy and Carbon Footprint calculations.

Supported Entities

Data is available in the Carbon Footprint chart for hosts and VMs discovered via vCenter targets.

Carbon Footprint Calculation

The calculation for carbon footprint can be expressed as follows:

$$(\text{Energy Consumption}) * [(\text{Power Usage Effectiveness}) * (\text{Carbon Intensity})] = \text{Carbon Footprint}$$

Where:

- Energy Consumption is the consumption data collected from the entity.
- Power Usage Effectiveness (PUE) is a ratio that describes how efficiently a computer data center uses energy; specifically, how much energy is used by the computing equipment. PUE is the ratio of the total amount of energy used by a computer data center facility to the energy delivered to computing equipment. The closer PUE is to 1, the more efficient the computer data center.
- Carbon Intensity (CI) is a measurement of how 'clean' electricity is. It refers to how many grams of carbon dioxide (CO₂) are released to produce 1 watt-hour (Wh) of electricity. Electricity that is generated using fossil fuels is more carbon intensive, as the process by which it is generated creates CO₂ emissions. Renewable energy sources, such as wind, hydro, or solar power produce next to no CO₂ emissions, so their carbon intensity value is much lower and often zero.

By default, PUE is set to 1.5, while CI is set to 0.25 g/Wh. These values appear in the [default policy \(on page 377\)](#) for Data Center entities. You can modify these values directly for a global effect, or set specific values in custom automation policies for Data Centers.

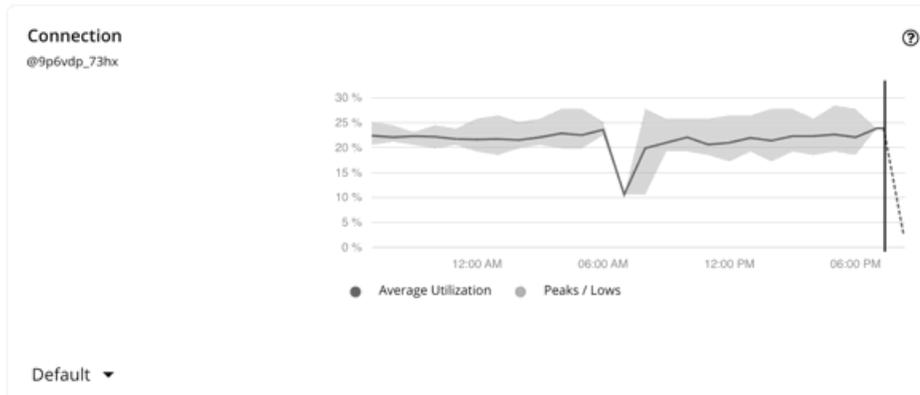
For example, if energy consumption for a host is 50 Wh, carbon footprint will be calculated as follows:

$$(50) * [(1.5) * (.25)] = 18.75 \text{ grams of CO}_2\text{e}$$

Connection Chart

Connection is the measurement of database connections utilized by applications.

Intersight Workload Optimizer collects connection data from Database Servers discovered by Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Intersight Workload Optimizer collected displays in the Connection chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Database Server (see the next section for a list of supported Database Servers).

Supported Database Servers

Data is available in the Connection chart for Database Servers discovered via the following targets:

| Target | Supported Database Servers |
|-------------|---|
| AWS | RDS |
| Azure | SQL |
| AppDynamics | MongoDB |
| Instana | Oracle |
| MySQL | MySQL |
| Oracle | Oracle |
| JBoss | All Database Servers discovered from the target |
| SQL | SQL |
| Tomcat | All Database Servers discovered from the target |
| WebLogic | All Database Servers discovered from the target |
| WebSphere | All Database Servers discovered from the target |

Effect on Memory Resize/Scale Actions

Intersight Workload Optimizer uses connection data to generate memory resize actions for on-prem Database Servers.

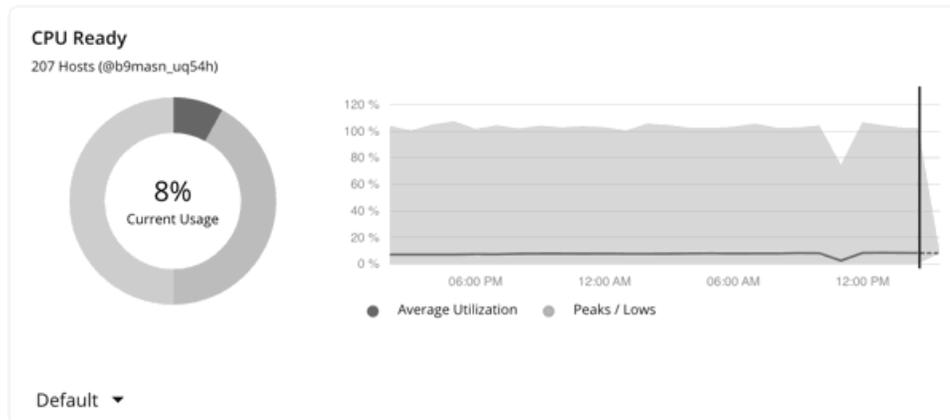
For cloud Database Servers, Intersight Workload Optimizer uses connection data as a constraint when generating scale actions. For details about scale actions, see [Database Server Actions \(on page 297\)](#).

CPU Ready Chart

CPU Ready is the measurement of time that a VM is ready to use CPU, but is unable to schedule physical CPU time because host CPU resources are busy.

Intersight Workload Optimizer collects CPU Ready data from hosts and VMs that are discovered by vCenter targets. It then calculates CPU Ready capacity and utilization to make accurate VM move recommendations.

When you set the scope to one or several VMs or hosts, the CPU Ready chart displays CPU Ready data.



Host CPU Ready Capacity

Intersight Workload Optimizer calculates host CPU Ready capacity by using following formula:

$$\text{Host Logical Processors} * 20 = \text{Host CPU Ready Capacity}$$

The following values are represented in the formula:

- `Host Logical Processors` is the number of logical CPU cores on a host.
- `20` is the standard Ready Queue interval (in seconds) at which the hypervisor measures metrics.

For example, if a host has 10 cores and applies the default host policy, Intersight Workload Optimizer calculates capacity according to the following formula:

$$10 * 20 = 200$$

Host CPU Ready Utilization

CPU Ready utilization is the measurement of capacity that is in use. Intersight Workload Optimizer calculates host CPU Ready utilization by using following formula:

$$\text{Host CPU Ready Average} / \text{Host CPU Ready Capacity} = \text{Host CPU Ready Utilization}$$

The following values are represented in the formula:

- `Host CPU Ready Average` is the average of the 20-second CPU Ready summation values that are collected from vCenter every 10 minutes. Each value is expressed in number of milliseconds.
- `Host CPU Ready Capacity` is the capacity that is calculated by Intersight Workload Optimizer. For more information, see the previous section.

For example, if raw utilization on a host is 40 and CPU Ready capacity is 100, Intersight Workload Optimizer uses the following calculation for host CPU Ready utilization.

First, to get the `Host CPU Ready Average` value, Intersight Workload Optimizer collects 20-second CPU Ready summation values from vCenter during the poll period, for example:

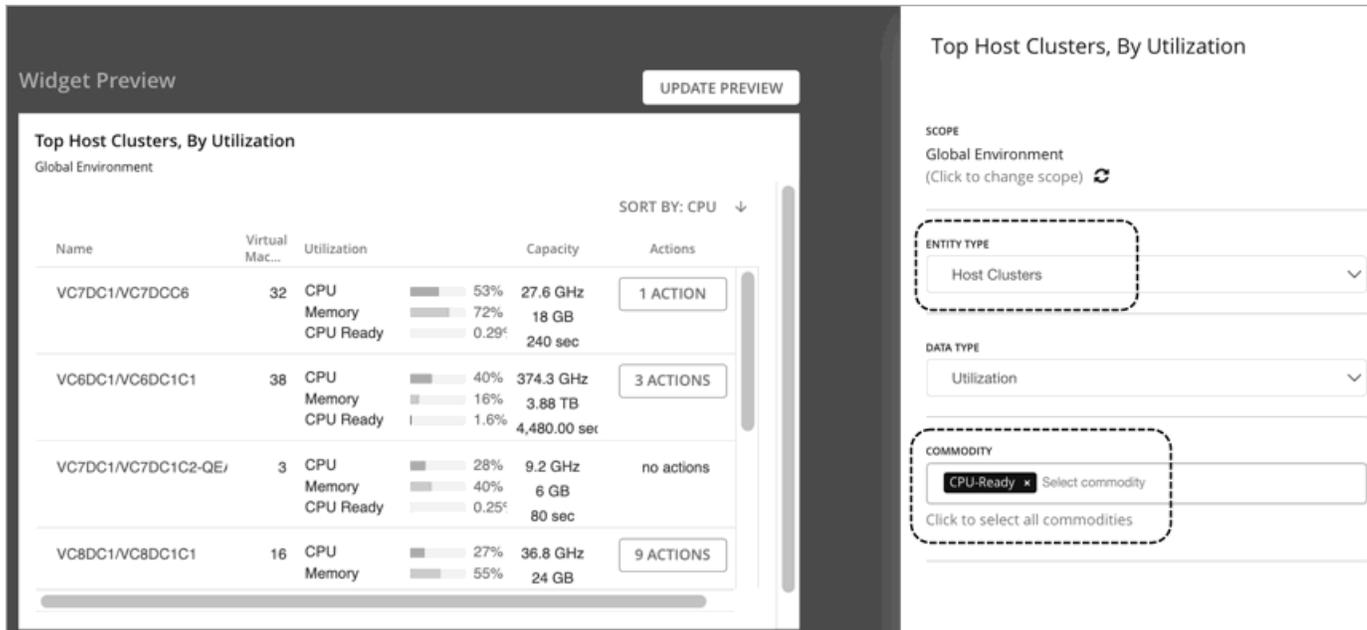
$$\text{AVERAGE}(500, 100, 1000, 1500, 1500, 1500, 500, 100, 1000, 1500, 1500, 1500, 500, 100, 1000, 1500, 1500, 1500, 500, 100, 1000, 1500, 1500, 1500) = 920 \text{ ms} = 0.92 \text{ s}$$

Intersight Workload Optimizer then divides the `Host CPU Ready Average` value by `Host CPU Ready Capacity`, which in this example is 100, to get a host CPU Ready utilization value of 0.92, or about 9%

$$0.92/100 = 0.0092 \text{ (0.9\%)}$$

To view the hosts with the highest CPU Ready, add the **Top Utilized** chart to your dashboard. When you configure the chart, select **Hosts** as the entity type, and **CPUReady** as the commodity.

To view the host clusters with the highest CPU Ready, add the **Top Utilized** chart to your dashboard. When you configure the chart, select **Host Clusters** as the entity type and **CPUReady** as the commodity.



This chart shows any pending action to move VMs out of a host cluster due to CPU Ready congestion.

VM CPU Ready Capacity

Intersight Workload Optimizer calculates VM CPU Ready capacity by using the following formula:

$$\text{VM Logical Processors} * 20 = \text{VM CPU Ready Capacity}$$

The following values are represented in the formula:

- VM Logical Processors is the number of vCPUs on a VM.
- 20 is the standard Ready Queue interval in seconds at which the hypervisor measures metrics.

For example, for a VM with 2 vCPUs, Intersight Workload Optimizer uses the following calculation for VM CPU Ready capacity:

$$2 * 20 = 40$$

VM CPU Ready Utilization

CPU Ready utilization is the measurement of capacity that is in use. Intersight Workload Optimizer calculates VM CPU Ready utilization by using following formula:

$$\text{Raw Utilization} / \text{VM CPU Ready Capacity} = \text{VM CPU Ready Utilization}$$

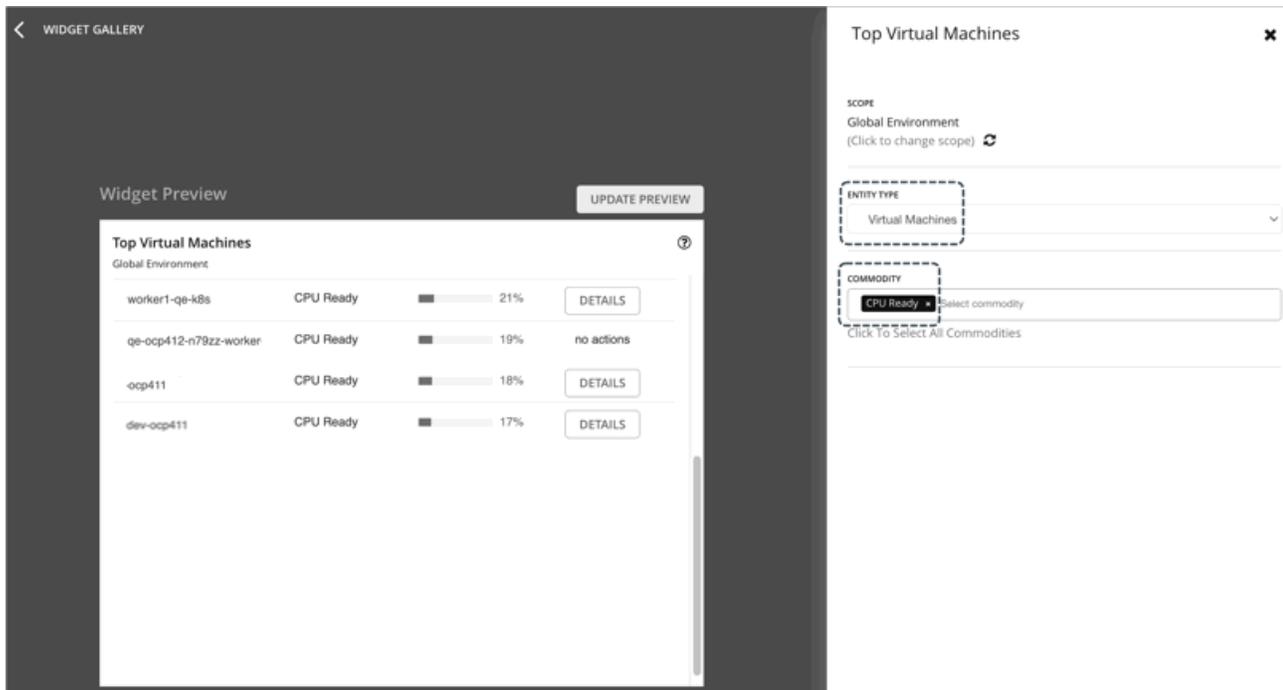
The following values are represented in the formula:

- Raw Utilization is the utilization value (in seconds) that is collected from vCenter.
- VM CPU Ready Capacity is the capacity that is calculated by Intersight Workload Optimizer. For more information, see the previous section.

For example, if raw utilization on a VM is 40 and CPU Ready capacity is 100, Intersight Workload Optimizer uses the following calculation for VM CPU Ready utilization:

40/40 = 1 (100%)

To view the most utilized VMs, add the **Top Utilized** chart to your dashboard. When you configure the chart, select **virtual machines** as the entity type, and **CPUReady** as the commodity.



This chart shows any pending action to move a VM to a different host due to CPU Ready congestion.

Effect on VM Move Actions

Intersight Workload Optimizer considers host CPU Ready utilization when it makes placement decisions for VMs. Ready Queue Utilization is a host policy setting for the percentage of utilization that Intersight Workload Optimizer considers as full utilization. For example, if utilization reaches 50%, Intersight Workload Optimizer considers Ready Queue to be fully utilized and the market might generate a move action to remedy the high Ready Queue utilization condition. The default value for this policy setting is 50%. This percentage is suitable for most environments. For environments where applications are sensitive to latency, you can reduce the percentage so that Intersight Workload Optimizer is more sensitive to CPU Ready when it places VMs. However, CPU Ready utilization is only one among many factors that go into move decisions. In some environments, other factors might offset CPU Ready utilization concerns.

Logical Processors are considered when Intersight Workload Optimizer places VMs on hosts. For example: If a VM has 64 vCPUs, it must run on a host with at least 64 logical cores. Intersight Workload Optimizer does not attempt to move a VM to a host with fewer Logical Processors. For VMs, Logical Processors Capacity and Used values equal the number of vCPUs that are configured on the VM. For hosts, Logical Processor Capacity equals the number of host logical cores and the Used value equals the sum of all VM Logical Processors (vCPUs). In this way, the Host Logical Processor Utilization represents the ratio of VM vCPUs to the host logical cores. Overprovisioned hosts are known to contribute to increased risk of CPU Ready latency. Be sure to take any resize down actions on VMs where CPU Ready is observed to reduce CPU overprovision.

NOTE: In VMware environments, the best practice is to keep CPU Ready values as low as possible. A CPU Ready value of 3% indicates a potential performance risk for most applications. If the CPU Ready value is 5% or greater, expect a significant performance impact.

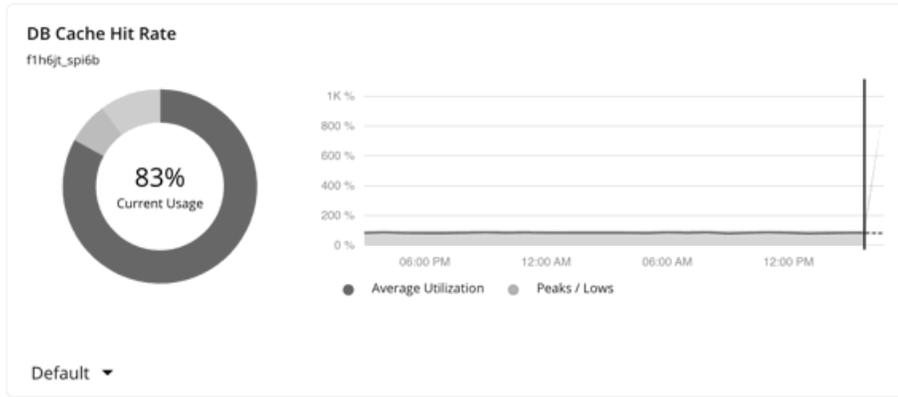
CPU Ready Dashboard: Identifying CPU Ready Risks

A [default dashboard \(on page 500\)](#) is available for reviewing CPU Ready in your environment. This dashboard shows the Host Clusters, Hosts, and VMs with the highest CPU Ready values. Reviewing the charts in this dashboard can help you tune your CPU Ready settings for the specific workloads that are running in your environment. To view this dashboard, select **DASHBOARD** from the main navigation menu and click **On-Prem CPU Ready Dashboard**.

DB Cache Hit Rate Chart

DB cache hit rate is the measurement of Database Server accesses that result in cache hits, measured as a percentage of hits versus total attempts. A high cache hit rate indicates efficiency.

Intersight Workload Optimizer collects cache hit rate data from Database Servers discovered by Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Intersight Workload Optimizer collected displays in the DB Cache Hit Rate chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Database Server (see the next section for a list of supported Database Servers).

Supported Database Servers

Data is available in the DB Cache Hit Rate chart for Database Servers discovered via the following targets:

| Target | Supported Database Servers |
|-------------|----------------------------|
| AWS | RDS |
| Azure | SQL |
| AppDynamics | SQL, Oracle |
| Dynatrace | SQL |
| Instana | MySQL, SQL, Oracle |
| MySQL | MySQL |
| New Relic | SQL, MySQL |
| Oracle | Oracle |
| SQL | SQL |

Effect on Memory Resize Actions

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

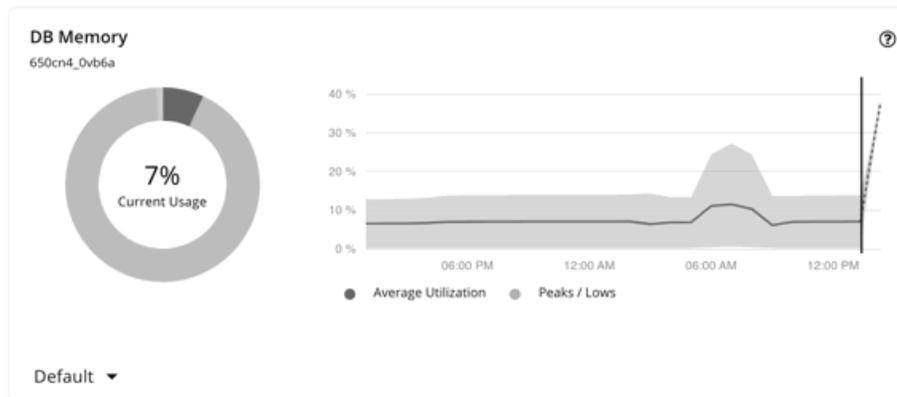
A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

DB Memory Chart

Database memory (or DBMem) is the measurement of memory that is utilized by a Database Server.

Intersight Workload Optimizer collects memory data from Database Servers discovered by Databases and APM targets. When you set the scope to one or several Database Servers, the data that Intersight Workload Optimizer collected displays in the DB Memory chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Database Server (see the next section for a list of supported Database Servers).

Supported Database Servers

Data is available in the DB Memory chart for Database Servers discovered via the following targets:

| Target | Supported Database Servers |
|-------------|----------------------------|
| AppDynamics | Oracle, MongoDB |
| Dynatrace | SQL, MySQL |
| Instana | MySQL, SQL |
| MySQL | MySQL |
| Oracle | Oracle |
| SQL | SQL |

Memory Resize Actions

Actions to resize database memory are driven by data on the Database Server, which is more accurate than data on the hosting VM. Intersight Workload Optimizer uses database memory and cache hit rate data to decide whether resize actions are necessary.

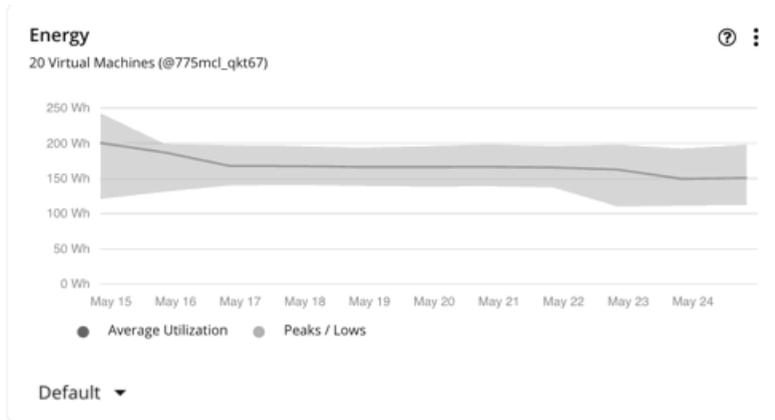
A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize down action generates.

When the cache hit rate is below the optimal value but database memory utilization remains low, no action generates. If utilization is high, a resize up action generates.

Energy Chart

Energy is the measurement of electricity consumed by a given entity over a period of time, expressed in watt-hours (Wh).

Intersight Workload Optimizer collects energy-related data from vCenter hosts and VMs at 10-minute intervals. For VMs, the data collected from hosts is used to calculate VM energy consumption. When you set the scope to one or several hosts or VM, the data that Intersight Workload Optimizer collected or calculated displays in the Energy chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Intersight Workload Optimizer also relies on host power data for the supported targets. This data is mandatory for Energy and Carbon Footprint calculations.

Supported Entities

Data is available in the Energy chart for hosts and VMs discovered via vCenter targets.

Calculation of Energy Consumed by VMs

Intersight Workload Optimizer calculates energy consumption for each VM on a host, based on two types of host energy:

- Host idle energy – this is the energy consumed by a host while in an idle state, when no processes are actively executing. This includes host overhead energy, which represents energy that is consumed by non-VM resources, such as hypervisors. Host idle energy is attributed to all member VMs based on each VM's allocated capacity.
- Host active energy – this is the energy consumed by the VMs' operating systems and active processes. Host active energy is attributed to all member VMs based on each VM's usage of host CPU.

The calculation can be expressed as follows:

$$\begin{aligned}
 & (\text{VM Size} / \text{Total Size for all VMs}) * \text{Host Idle Energy for VMs} + \\
 & (\text{VM CPU Utilization} / \text{Total CPU Utilization for all VMs}) * \text{Host Active Energy for VMs} \\
 & = \text{VM Energy Consumption (in Wh)}
 \end{aligned}$$

NOTE:

Intersight Workload Optimizer rounds the calculated values.

For example, consider a host with three VMs. Currently, host *idle* energy for VMs is 50 Wh, while host *active* energy for VMs is 20 Wh.

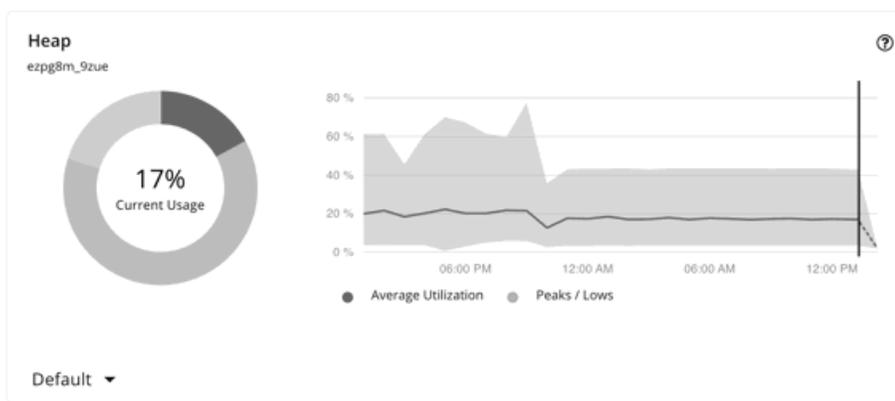
Intersight Workload Optimizer calculates VM energy consumption as follows:

| VM | Size (Cores) | CPU Utilization (%) | VM Energy Consumption (Wh) |
|--------------|--------------|----------------------|--|
| VM_01 | 2 | 50 (x 2 cores) = 100 | $(2 / 4) * 50 + (100 / 180) * 20 = 36.1$ |
| VM_02 | 1 | 80 (x 1 core) = 80 | $(1 / 4) * 50 + (80 / 180) * 20 = 21.4$ |
| VM_03 | 1 | 0 (idle) | $(1 / 4) * 50 + (0) * 20 = 12.5$ |
| Total | 4 | 180 | 70 |

Heap Chart

Heap is the portion of a VM or container's memory allocated to individual applications.

Intersight Workload Optimizer collects heap data from Application Components discovered by Applications and APM targets. When you set the scope to one or several Application Components, the data that Intersight Workload Optimizer collected displays in the Heap chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Application Component (see the next section for a list of supported Application Components).

Supported Application Components

Data is available in the Heap chart for Application Components discovered via the following targets:

| Target | Supported Application Components |
|-------------|----------------------------------|
| AppDynamics | Java applications, .NET, Node.js |
| Dynatrace | Java applications |
| JVM | Java applications |
| New Relic | Java applications, Node.js |
| Tomcat | Java applications |
| WebSphere | Java applications |

Heap Resize Actions

Intersight Workload Optimizer generates Heap resize actions if an Application Component provides Heap and Remaining GC Capacity, and the underlying VM or container provides VMem. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.

NOTE:

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

Number of Replicas Chart

This chart shows the replicas of Application Components running over a given time period.

Use this chart if:

- SLO-driven scaling is enabled for a Service, and *provision* or *suspend* actions are executed by Intersight Workload Optimizer. These actions adjust the number of replicas to help you meet your SLO goals.
- Or
- [Horizontal Pod Autoscaler](#) (HPA) is enabled for a *Deployment*, *ReplicaSet*, or *StatefulSet* that is exposed as a Service. Intersight Workload Optimizer discovers adjustments to the number of replicas made by HPA.

The chart shows following information:

■ Capacity values

The number of desired pod replicas configured in the workload controller that backs the Service. This can be configured in *Deployment*, *ReplicaSet*, *StatefulSet*, *ReplicationController* or *DeploymentConfig*.

The chart plots the *maximum* observed capacity within the given time period.

■ Used values

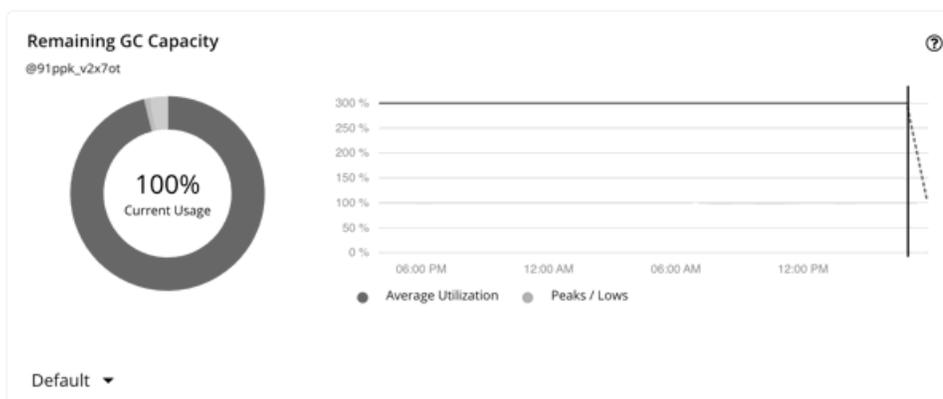
The number of *ready* pods owned by the workload controller. Pods in other states (for example, pending pods) are not counted.

The chart plots the *average* used values within the given time period. Hover over the chart to see the minimum and maximum used values.

Remaining GC Capacity Chart

Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

Intersight Workload Optimizer collects GC data from Application Components discovered by Applications and APM targets, and then uses that data to calculate remaining GC capacity. When you set the scope to one or several Application Components, the capacity that Intersight Workload Optimizer calculated displays in the Remaining GC Capacity chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Application Component (see the next section for a list of supported Application Components).

Supported Application Components

Data is available in the Remaining GC Capacity chart for Application Components discovered via the following targets:

| Target | Supported Application Components |
|-------------|----------------------------------|
| AppDynamics | Java applications, .NET |
| Dynatrace | Java applications |
| JVM | Java applications |
| New Relic | Java applications, Node.js |
| Tomcat | Java applications |
| WebSphere | Java applications |

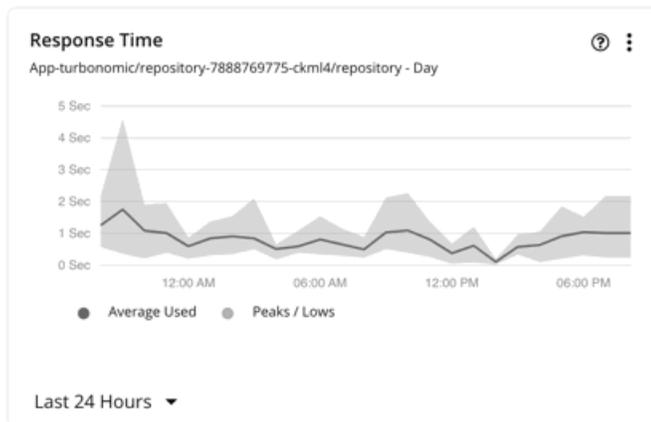
Effect on Heap Resize Actions

Intersight Workload Optimizer generates Heap resize actions if an Application Component provides Heap and Remaining GC Capacity, and the underlying VM or container provides VMem. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.

Response Time Chart

Response Time is the elapsed time between a request and the response to that request. Response Time is typically measured in seconds (s) or milliseconds (ms).

Intersight Workload Optimizer collects response time data from entities discovered by Applications, Databases, and APM targets. Entities include Business Applications, Business Transactions, Services, Application Components, and self-hosted Database Servers. When you set the scope to any of these entities, the data that Intersight Workload Optimizer collected displays in the Response Time chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Supported Entities

Data is available in the Response Time chart for the following entities:

| Target | Supported Entities |
|-------------|--|
| AppDynamics | Business Application, Business Transaction, Service, Application Component |
| Datadog | Business Application, Business Transaction, Service, Application Component |
| Dynatrace | Business Application, Service, Application Component |
| JVM | Application Component |
| MySQL | Database Server |
| New Relic | Business Transaction, Service, Application Component, Database Server |
| Oracle | Database Server |
| SQL | Database Server |
| Tomcat | Application Component |
| WebSphere | Application Component |

Response Time SLOs

To evaluate the performance of your applications and Database Servers, set Response Time SLOs (Service Level Objectives) as an operational constraint in policies. For applications, you can set the SLO at the Business Application, Business Transaction, Service, or Application Component level.

OPERATIONAL CONSTRAINTS

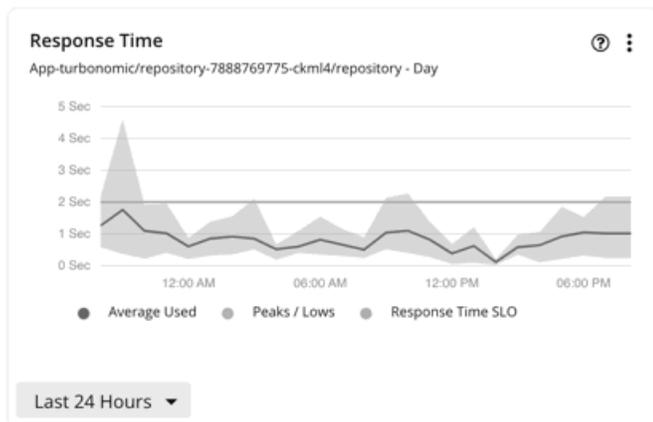
Transaction SLO 10

Enable Transaction SLO

Enable Response Time SLO

Response Time SLO [ms] 2000 ms

After you create a policy, the SLO value appears as a solid straight line in the Response Time chart. You can then gauge performance against the given SLO.



If you do not set an SLO, Intersight Workload Optimizer estimates SLO based on historical Response Time data collected from the target, and then displays the estimated value in the Capacity and Usage chart, as Response Time capacity. This estimated value is *not* reflected in the Response Time chart.

Capacity and Usage ? ⋮

SQLServer [win-dynatrace-mssql2017]

| Commodity | Capacity | Used | Utilization |
|-------------------|------------|------------|-------------|
| DB Cache Hit Rate | 100 % | 100 % | 100% |
| Response Time | 44.76 msec | 42.65 msec | 95.3% |
| TransactionLog | 347.21 MB | 71.36 MB | 20.55% |
| Transaction | 3.58 TPS | 0.38 TPS | 10.66% |
| DB Memory | 6.18 GB | 644.22 MB | 10.19% |

[SHOW ALL >](#)

NOTE:

When you set an SLO value, Response Time capacity in the Capacity and Usage chart shows as N/A.

Response Time SLOs for Container Platform Services

When you add a container platform target, Intersight Workload Optimizer discovers services monitored by AppDynamics, , Dynatrace, and New Relic.

To generate actions that adjust pod replicas, services must be discovered by the Kubeturbo agent that you deployed to your environment, as well as collect performance metrics through Instana or DIF (Data Ingestion Framework). In addition, Intersight Workload Optimizer requires that you turn on horizontal scaling and specify Response Time SLOs in policies for the affected services.

< Configure Service Policy
✕

NAME

— SCOPE

AH-Service_GP ✕

➕ ADD SERVICE GROUPS

+ POLICY SCHEDULE

— AUTOMATION AND ORCHESTRATION

Defines how actions are accepted.

HORIZONTAL SCALE UP, HORIZONTAL SCALE DOWN

Action Acceptance: Manual

— OPERATIONAL CONSTRAINTS

⊖ Response Time SLO [ms] 2000 ms

⊖ Enable Response Time SLO

⊖ Enable Transaction SLO

⊖ Transaction SLO 10

➕ ADD CONSTRAINT

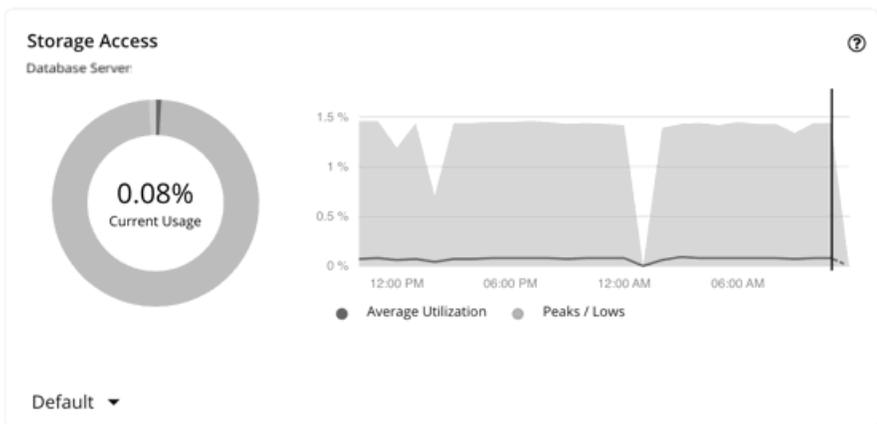
Response Time SLO is the desired *weighted average* response time (in milliseconds) of all Application Component replicas associated with a Service.

NOTE:

If you specified SLOs but turned off horizontal scaling in policies, no actions generate but SLO values will continue to display in the Response Time chart for Services, for your reference. This allows you to gauge performance against those SLOs.

Storage Access Chart

Storage Access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity. Intersight Workload Optimizer collects storage access data from VMs, Database Servers, and storage entities discovered by cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Intersight Workload Optimizer collected displays in the Storage Access chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Supported Entities

Data is available in the Storage Access chart for entities discovered via the following targets:

| Target Type | Target | Supported Entities |
|----------------|-------------------------|--|
| Cloud | AWS | Virtual Machine, Database Server, Volume |
| | Azure | Virtual Machine, Database Server, Volume |
| | Google Cloud | Virtual Machine |
| Fabric | HPE OneView | Virtual Machine, Storage |
| Hyperconverged | HyperFlex | Storage, Disk Array |
| | Nutanix | Storage, Disk Array |
| Hypervisor | Hyper-V | Virtual Machine, Storage |
| | vCenter | Virtual Machine, Storage |
| Storage | EMC ScaleIO | Storage, Disk Array |
| | EMC VMAX | Storage, Disk Array, Logical Pool |
| | EMC XtremIO | Storage, Disk Array |
| | HPE 3PAR | Storage, Disk Array, Logical Pool |
| | IBM FlashSystem | Storage, Disk Array, Logical Pool |
| | NetApp | Storage, Disk Array |
| | Pure Storage FlashArray | Storage, Disk Array |

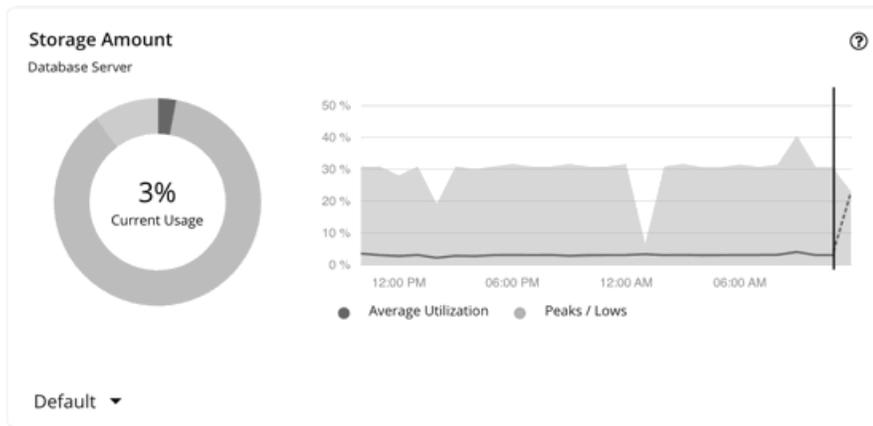
Scale Actions

Intersight Workload Optimizer considers storage access (IOPS) utilization when recommending scaling actions for [cloud VMs \(on page 265\)](#), [cloud Database Servers \(on page 297\)](#), and [volumes \(on page 327\)](#).

Storage Amount Chart

Storage Amount is the measurement of storage capacity that is in use.

Intersight Workload Optimizer collects storage amount data from VMs, Database Servers, and storage entities discovered by cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Intersight Workload Optimizer collected displays in the Storage Amount chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Supported Entities

Data is available in the Storage Amount chart for entities discovered via the following targets:

| Target Type | Target | Supported Entities |
|----------------|-------------------------|---|
| Cloud | AWS | Virtual Machine, Database Server |
| | Azure | Virtual Machine |
| | Google Cloud | Virtual Machine |
| Fabric | HPE OneView | Storage |
| Hyperconverged | HyperFlex | Storage, Disk Array |
| | Nutanix | Storage, Disk Array, Storage Controller |
| Hypervisor | Hyper-V | Storage |
| | vCenter | Storage |
| Storage | EMC ScaleIO | Storage, Disk Array, Storage Controller |
| | EMC VMAX | Storage, Disk Array, Logical Pool, Storage Controller |
| | EMC XtremIO | Storage, Disk Array, Storage Controller |
| | HPE 3PAR | Storage, Disk Array, Logical Pool, Storage Controller |
| | IBM FlashSystem | Storage, Disk Array, Logical Pool, Storage Controller |
| | NetApp | Storage, Disk Array, Storage Controller |
| | Pure Storage FlashArray | Storage, Disk Array, Storage Controller |

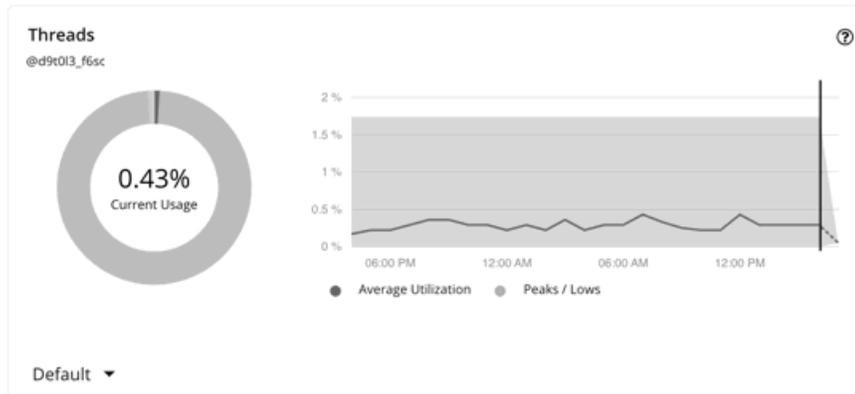
Scale Actions

Intersight Workload Optimizer can recommend scaling [cloud VMs \(on page 265\)](#) and [cloud Database Servers \(on page 297\)](#) to instance types that can adequately meet storage demand at the lowest possible cost. For cloud Database Servers, Intersight Workload Optimizer can also recommend scaling up storage amount within the same instance type. Note that scaling up storage amount is non-disruptive but irreversible.

Threads Chart

Threads is the measurement of thread capacity utilized by applications.

Intersight Workload Optimizer collects thread data from Application Components discovered by Applications and APM targets. When you set the scope to one or several Application Components, the data that Intersight Workload Optimizer collected displays in the Threads chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Application Component (see the next section for a list of supported Application Components).

Supported Application Components

Data is available in the Threads chart for Application Components discovered via the following targets:

| Target | Supported Application Components |
|-------------|----------------------------------|
| AppDynamics | Java applications, .NET |
| JVM | Java applications |
| New Relic | Java applications |
| Tomcat | Java applications |
| WebSphere | Java applications |

Thread Pool Resize Actions

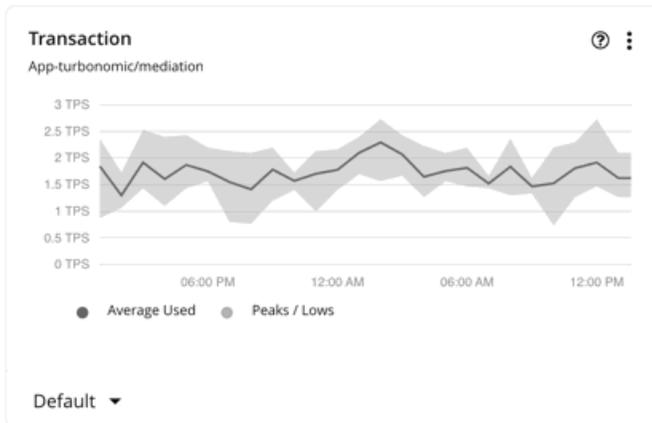
Intersight Workload Optimizer generates thread pool resize actions. These actions are recommend-only and can only be executed outside Intersight Workload Optimizer.

Transaction Chart

Transaction is a value that represents the per-second utilization of the transactions that are allocated to a given entity.

Intersight Workload Optimizer collects transaction data from entities discovered by Applications, Databases, and APM targets. Entities include Business Applications, Business Transactions, Services, Application Components, and self-hosted Database

Servers. When you set the scope to any of these entities, the data that Intersight Workload Optimizer collected displays in the Transaction chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Supported Entities

Data is available in the Transaction chart for the following entities:

| Target | Supported Entities |
|-------------|---|
| AppDynamics | Business Application, Business Transaction, Service, Application Component, Database Server |
| Dynatrace | Business Application, Service, Database Server, Application Component |
| MySQL | Database Server |
| New Relic | Business Transaction, Service, Application Component, Database Server |
| Oracle | Database Server |
| SQL | Database Server |
| Tomcat | Application Component |
| WebSphere | Application Component |

Transaction SLOs

To evaluate the performance of your applications and Database Servers, set Transaction SLOs as an operational constraint in policies. For applications, you can set the SLO at the Business Application, Business Transaction, Service, or Application Component level.

OPERATIONAL CONSTRAINTS

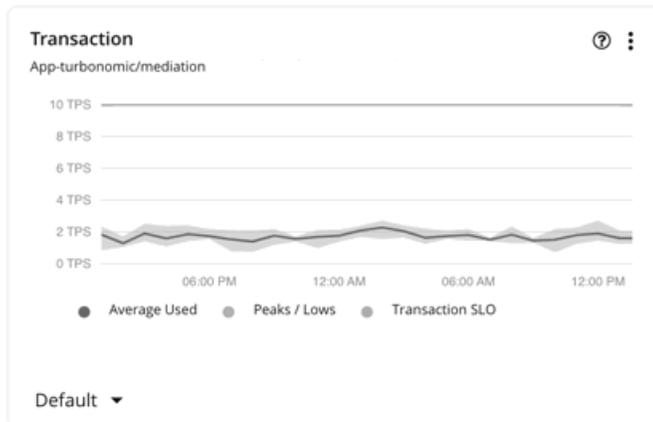
Enable Response Time SLO

Response Time SLO [ms] ms

Enable Transaction SLO

Transaction SLO

After you create a policy, the SLO value appears as a solid straight line in the Transaction chart. You can then gauge performance against the given SLO.



If you do not set an SLO, Intersight Workload Optimizer estimates SLO based on historical Transaction data collected from the target, and then displays the estimated value in the Capacity and Usage chart, as Transaction capacity. This estimated value is *not* reflected in the Transaction chart.

Capacity and Usage
App-turbonomic/mediation

| Commodity | Capacity | Used | Utilization |
|-----------------------|----------|------------|-------------|
| Remaining GC Capacity | 100 % | 99.59 % | 99.59% |
| Transaction | 3.2 TPS | 2.2 TPS | 68.79% |
| Heap | 24 GB | 0.93 GB | 3.87% |
| Virtual Memory | 32 GB | 1.54 GB | 4.82% |
| Virtual CPU | 17.6 GHz | 835.07 MHz | 4.74% |

SHOW ALL >

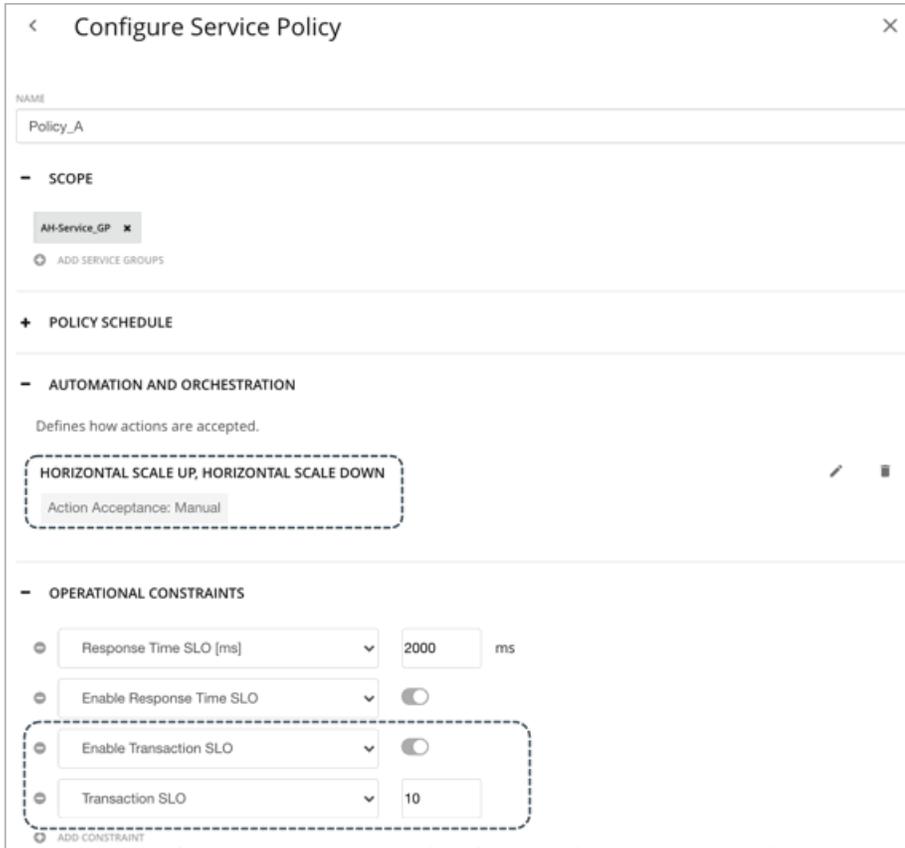
NOTE:

When you set an SLO value, Transaction capacity in the Capacity and Usage chart shows as N/A.

Transaction SLOs for Container Platform Services

When you add a container platform target, Intersight Workload Optimizer discovers container platform services managed by AppDynamics, Dynatrace, and New Relic.

To generate actions that adjust pod replicas, container platform services must be discovered by the KubeTurbo pod that you deployed to your environment, as well as collect performance metrics through Instana or DIF (Data Ingestion Framework). In addition, Intersight Workload Optimizer requires that you turn on horizontal scaling and specify Transaction SLOs in policies for the affected services.



Transaction SLO is the maximum number of transactions per second that each Application Component replica can handle.

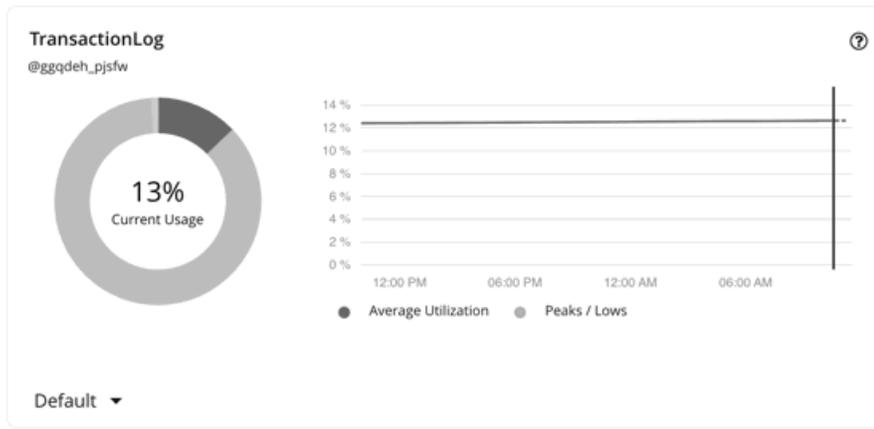
NOTE:

If you specified SLOs but turned off horizontal scaling in policies, no actions generate but SLO values will continue to display in the Transaction chart for services, for your reference. This allows you to gauge performance against those SLOs.

Transaction Log Chart

Transaction log is the measurement of storage capacity utilized by a Database Server for transaction logging.

Intersight Workload Optimizer collects transaction log data from Database Servers discovered by Databases and APM targets. When you set the scope to one or several Database Servers, the data that Intersight Workload Optimizer collected displays in the Transaction Log chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported Database Server (see the next section for a list of supported Database Servers).

Supported Database Servers

Data is available in the Transaction Log chart for Database Servers discovered via the following targets:

| Target | Supported Database Server |
|-------------|---------------------------|
| AppDynamics | SQL |
| Oracle | Oracle |
| SQL | SQL |

Transaction Log Resize Actions

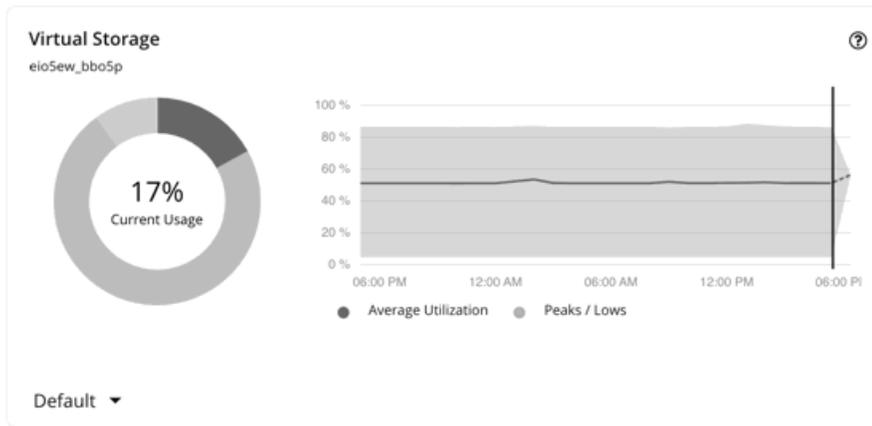
Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology.

Currently, Intersight Workload Optimizer does not support resize actions for Oracle and Database Servers on the Hyper-V platform (due to the lack of API support for virtual storage).

Virtual Storage Chart

Virtual storage is the measurement of virtual storage capacity that is in use.

Intersight Workload Optimizer collects virtual storage data from VMs discovered by on-prem and APM targets. When you set the scope to one or several VMs, the data that Intersight Workload Optimizer collected displays in the Virtual Storage chart.



The chart shows average and peak/low values over time. Use the selector at the bottom left section of the chart to change the time frame.

NOTE:

An empty chart could be the result of delayed discovery, target validation failure, unavailable data for the given time frame, or an unsupported entity (see the next section for a list of supported entities).

Supported Entities

Data is available in the Virtual Storage chart for entities discovered via the following targets:

| Target Type | Target | Supported Entities |
|----------------|-------------|--------------------|
| Database | MySQL | Virtual Machine |
| | Oracle | Virtual Machine |
| | SQL | Virtual Machine |
| Fabric | HPE OneView | Virtual Machine |
| Hyperconverged | HyperFlex | Virtual Machine |
| | Nutanix | Virtual Machine |
| Hypervisor | Hyper-V | Virtual Machine |
| | vCenter | Virtual Machine |

Virtual Storage Actions

Intersight Workload Optimizer can recommend the following virtual storage actions:

- Move a VM's volume (virtual storage) due to excess utilization of the current datastore, or for more efficient utilization of datastores in the environment.
- Reconfigure a VM's volume (virtual storage) to comply with placement policies.

Top Utilized Charts

Top Utilized charts show the entities or groups with the most utilization.

Entity Type

Entity types you can choose include:

- [Accounts \(on page 542\)](#) (public cloud)
- Application Components

- Business Applications
- Business Transactions
- Business Users
- Chassis
- [Host Clusters \(on page 542\)](#)
- Containers
- Container Pods
- Container specs
- Data Centers
- Database Servers
- Desktop Pools
- Disk Arrays
- IO Modules
- Internet
- Namespaces
- Networks
- Node Pools
- Hosts
- [Resource Groups \(on page 543\)](#)
- Services
- Storage Devices
- Storage Controllers
- Switches
- View Pods
- Virtual Data Centers
- Virtual Machines
- Volumes
- Wasted Files
- Workload controllers
- Regions
- Zones

Data Type

Depending on the entity type (for example, Host Clusters), you can choose **Headroom** or **Utilization** information in the chart.

Commodity

Depending on the entity type, you can add one or more different resource commodities that you want to measure.

Display

The chart lists the top entities by utilization of the commodities that you or the system has set. Depending on the entity type and scope, you can sort the information. To view the utilization details, hover over the entity to display the tooltip.

To drill down to an entity, click the entity name in the chart. This sets the scope to that entity.

Click the **ACTIONS** button for an entity to examine the actions that are pending for it, and then decide which ones are safe to execute.

Example: A top host clusters chart which can be sorted by CPU headroom or CPU exhaustion.

Top Host Clusters, By Headroom
Global Environment

SORT BY: CPU HEADROOM ↓

| Name | Virtual Machin... | Headroom | Time To Exhaustion | Actions |
|---------------------|-------------------|----------|--------------------|------------|
| VC7DC1/VC7DC1C3-QE | 1 CPU | 0% | More than 1 y... | no actions |
| VC3DC1/VC3DC1C1 | 2 CPU | 0% | More than 1 y... | 2 ACTIONS |
| RTP/HyperFlex_RTP_1 | 26 CPU | 2% | More than 1 y... | no actions |
| VC4DC1/VC4DC1C1 | 0 CPU | 2% | More than 1 y... | no actions |
| VC14DC1/VC14DC1C3- | 8 CPU | 2% | More than 1 y... | no actions |
| VC13DC1/VC13DC1C1 | 2 CPU | 3% | More than 1 y... | 1 ACTION |

SHOW ALL >

Top Host Clusters Chart

This chart shows the top host clusters in your on-prem environment by CPU, memory, and storage capacity or utilization. In the default view, the chart shows the top clusters by CPU headroom (available capacity). It also shows time to exhaustion of cluster resources, which is useful for future planning (for example, you might need to buy more hardware).

To calculate cluster capacity and headroom, Intersight Workload Optimizer runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current VMs to other hosts within the given cluster, if such moves would result in a more desirable workload distribution. The result of the plan is a calculation of how many more VMs the cluster can support.

Click the **ACTIONS** button for a given host cluster to see the actions that Intersight Workload Optimizer recommends to keep cluster resources in the desired state, and then decide which ones are safe to execute.

Click **Show All** to see all of the clusters. In the Show All list, you can also download capacity data as a CSV file. Click a cluster name to set the scope to that cluster and view more details about its current capacity and health.

Top Accounts Chart

This chart lists the cloud accounts with the largest potential savings. These are the savings you would realize if you execute pending actions for your cloud workloads. Click the **ACTIONS** button to examine these actions and decide which ones are safe to execute. The chart also shows billed costs in the last 30 days.

To set the scope to a particular account, click the account name.

Click **Show all** to view additional information, including the number of actions that have been executed for individual accounts or workloads, along with the resulting savings. If you have multiple cloud providers, each provider will have its own tab. You can download the accounts list as a CSV file.

AWS Accounts

The chart shows the AWS master and member accounts that you have added as targets. Accounts with a star symbol are master accounts.

NOTE:

Specific RIs can provide savings for multiple accounts. However, individual accounts show the full RI savings, which can result in exaggerated savings for that account.

Azure Accounts

The chart shows the subscriptions discovered via the service principal accounts that you have added as targets.

Google Cloud Accounts

The chart shows the folders and projects discovered via the Google Cloud service accounts that you have added as targets.

If a service account has access to a folder with projects and subfolders, the folder displays as the top-level account. Click **Show All** to see the full resource hierarchy and top-down data. If a service account has access to a project or subfolder but not its parent folder, the project or subfolder displays as the top-level account.

Top Resource Groups Chart

This chart highlights the estimated monthly cost for the top resource groups in your cloud environment and the savings you would realize if you execute the pending actions. Click the **ACTIONS** button to examine these actions and decide which ones are safe to execute. Click a resource group to set the scope to that group.

The chart also counts actions that have been executed for individual groups, and then shows the resulting savings.

Workload Health Charts

Workload Health charts show the health of workloads from the compliance, efficiency improvement, and performance assurance perspectives. These charts use current (real-time) data for the workloads chosen for the chart widget scope.

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Breakdown

You can choose:

- **Workload by Compliance**

A virtual environment can include policies that limit availability of resources. It's possible that the environment configuration violates these defined policies. In such cases, Intersight Workload Optimizer identifies the violation and recommends actions that bring the entity back into compliance.

- **Workload by Efficiency Improvement**

Efficient utilization of resources is an important part of running in the desired state. Running efficiently maximizes your investment and reduces cost. When Intersight Workload Optimizer discovers underutilized workloads, it recommends actions to optimize operations and save money.

- **Workload by Performance Assurance**

Ultimately, the reason to manage workloads in your environment is to assure performance and meet QoS goals. When Intersight Workload Optimizer detects conditions that directly put QoS at risk, it recommends associated actions to assure performance. You can consider these critical conditions, and you should execute the recommended actions as soon as possible.

Workload Health charts indicate actions that you should consider to improve the health of workloads. To see a list of actions, click **Show Actions** at the bottom of the chart.

Environment Charts

Environment charts provide an overview of your environment. They show the targets that you are managing and count the entities that Intersight Workload Optimizer has discovered through those targets. For example, you can display the cloud service providers, hypervisors, and the number of workloads.

Environment Type

You can choose one of the following views:

- Hybrid (both on-prem and cloud)
- Cloud
- On-Prem

Display

The chart shows the information as a Text chart type.

Workload Improvements Charts

Workload Improvements charts track the health of workloads in your environment over time, and map the health to the number of actions Intersight Workload Optimizer has executed in that time period.

In the chart, you can see the significance and value of executed actions:

- Workloads Overall
This is the total number of workloads over time.
- Workloads with Performance Risks
These are the workloads that are not performing well.
- Inefficient Workloads
These are the workloads that are running on under-utilized hosts or are not being utilized.
- Workloads Out of Compliance
These are the workloads that are violating a placement policy. Workloads that are not in compliance might be running on a host or placed on storage, for example, that violate a placement policy.
- Executed actions
Actions that Intersight Workload Optimizer executed.

The vertical line shows when the last data point was polled in your environment.

Environment Type

You can choose one of the following views:

- Hybrid (both on-prem and cloud)
- Cloud
- On-Prem

Display

The chart shows the information as a Line chart.

Cloud Chart Types

These chart widgets provide information on the status of your cloud environment.

For many cloud chart widgets that display costs and savings, Intersight Workload Optimizer uses the billing reports from your cloud service providers to build a picture of your overall costs. The data includes all costs that the service provider includes in the billing report. Intersight Workload Optimizer parses these reports into the formats that it uses for the cloud chart widgets.

Cost Breakdown Charts

Intersight Workload Optimizer displays the unbilled billing information from your cloud provider. Use this chart to [track \(on page 23\)](#) your expenses and see historical trends.

Categories

The chart can break down costs by the following categories.

- **Cost Breakdown by Service Provider**

See costs over time for each cloud service provider in your environment.

You can open more than one account from a single service provider. If you are running workloads on different service providers, this chart shows the distribution of costs across these providers.

- **Cost Breakdown by Account**

The chart shows the [cloud accounts \(on page 542\)](#) with the largest costs. The chart's legend displays up to 20 actual accounts and, if needed, an additional item labeled 'Others' that represents all accounts that are not in the top 20. Hover on a data point to see costs for individual accounts.

- **Cost Breakdown by Service**

The chart shows the services with the largest costs. The chart's legend displays up to 20 actual services and, if needed, an additional item labeled 'Others' that represents all services that are not in the top 20. Hover on a data point to see costs for individual services.

- **Cost Breakdown By Region**

The chart shows the regions with the largest costs. The chart's legend displays up to 20 actual regions and, if needed, an additional item labeled 'Others' that represents all regions that are not in the top 20. Hover on a data point to see costs for individual regions.

- **Cost Breakdown by Tag**

Intersight Workload Optimizer discovers the tags that you assigned to resources in your environment and then displays the tag keys when you click **Tag** at the top-right section of the chart. Select a tag key to see the tag values with the largest costs.

The chart's legend displays up to 20 actual tag values and, if needed, an additional item labeled 'Others' that represents all tag values that are not in the top 20. Hover on a data point to see costs for individual tag values.

- **Workload Cost Breakdown**

This chart breaks down costs by the following categories:

- On-Demand Compute
- Spot Compute
- On-Demand Compute License Bundle – only applies to Azure VMs and AWS VMs running Windows OS
- On-demand License
- Reserved License
- Storage
- Other cost – includes workload costs that are not categorized by Intersight Workload Optimizer, such as network charges

Points to consider:

- For VMs fully covered by AWS Reserved Instances, Azure reservations, or Google Cloud committed use discounts, on-demand cost is 0 (zero).
- For Azure VMs, the chart shows cost data from the day the VM was discovered by Intersight Workload Optimizer. Cost data before discovery is not available and therefore not reflected in the chart.

Chart Type

You can set the display to:

- Line Chart
- Stacked Bar Chart
- Area Chart

Chart Time Frame

Currently, the chart can display data from the last 7 or 30 days. As you change the time frame, Intersight Workload Optimizer divides the reported information into the appropriate time units to match that time frame. However, the source remains the same. Changing the time frame does not affect the source data or increase data polling.

Cloud Tier Breakdown Charts

Cloud Tier charts show the cloud tiers that Intersight Workload Optimizer discovers for the chart widget scope. For example, if the Chart Widget Scope is set to All Cloud VMs and the Entity Type is set to Virtual Machine, the chart shows all the cloud tiers that the workloads use.

Entity Type

You can choose any entity type in the list.

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Location Charts

Location charts show cloud provider regions in a world map for which there are discovered workloads. Click on any region to examine more detailed information in a scoped view.

Display

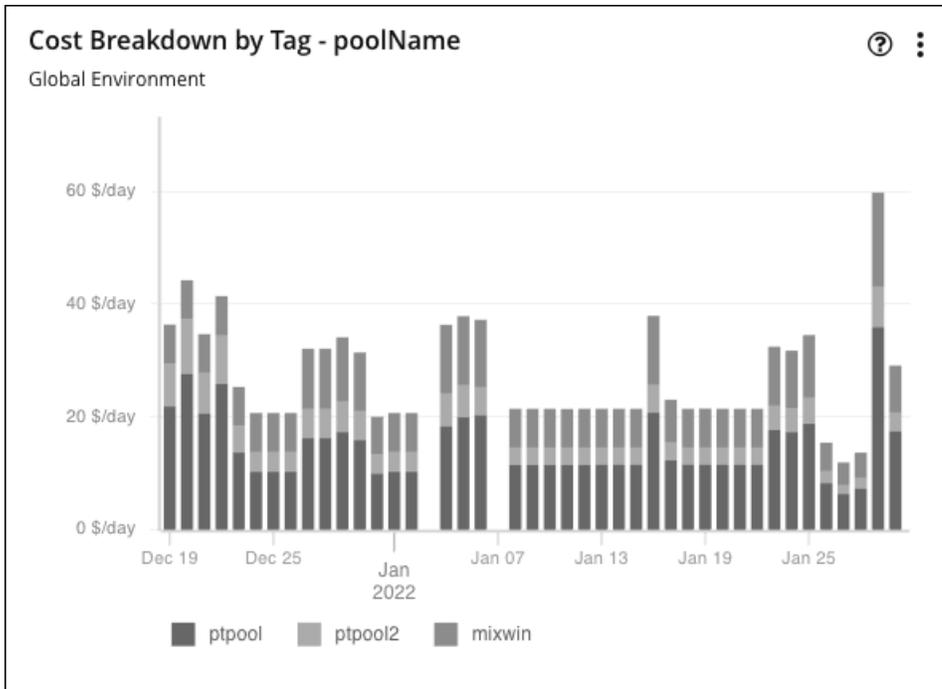
The chart shows the regions in countries in a Map chart.

Cost Breakdown by Tag Charts

Cost Breakdown By Tag charts show the costs for tagged cloud entities that Intersight Workload Optimizer discovered in your AWS, Azure, or Google Cloud environment. For the tagged entities in scope, the chart shows how daily costs change over time.

You choose a tag key to track, and then choose which tag values to include in the chart. Each data point aggregates the costs for all the entities with a given tag/value pair. You can display the cost breakdown in a stacked bar chart or an area chart.

For example, in this stacked bar chart, the tag **poolName** is *workload-type* and the tag **Values** are *ptpool*, *ptpool2*, and *mixwin*.



NOTE:

If you set up an AWS standard data export (CUR 2.0) for use with the AWS Billing target, be aware that AWS replaces tags with special characters and camel cases with underscores (_) in the data export. As a result, tags shown in AWS Cost Explorer and in the Cost Breakdown by Tag chart do not match.

Scope

To display these charts, add them to the default views in the Home Page or to your custom dashboards. By default, these charts are scoped to your global environment. You can change the scope to view granular data.

Timeframe

Currently, the chart can display data from the last 7 or 30 days.

Chart Type

You can set the display to:

- Area Chart
- Stacked Bar Chart

For more detail, hover over a data point. A tooltip appears to show specific values for that date. Click the legend items to show/hide data for specific values.

Tag Settings

Choose the Tag/Value pairs you want to display in the chart.

Note that tag Key and Value are case insensitive. Each data point in the chart aggregates the costs for all entities with the given tag Key/Value pair, regardless of case.

- **Key**

The tag name that you want to chart. Intersight Workload Optimizer discovers the tags you have configured in your environment.

You can choose one Key for the chart.
- **Values**

The values that you have configured in your environment for the given Key.

You can choose multiple values. To shorten the list of values, type a filter string in the Values field.

Cumulative Savings and Investments Charts

IMPORTANT:

The method for tracking savings and investments has been enhanced to accurately report the impact of actions on your cloud expenses. This enhanced tracking began in the September 2023 release. For cumulative savings and investments before this release, see the **Legacy Cumulative Savings** and **Legacy Cumulative Investments** charts.

Actions for your cloud workloads usually have cost savings or investments attached to them. For example, deleting unattached volumes can lower your costs significantly (savings), while scaling a VM to a different tier to improve performance could incur additional costs (investments). These charts highlight total savings and investments as a result of executing actions. Intersight Workload Optimizer uses billing data from your cloud provider to calculate savings and investments.

These charts display in the built-in Cloud Executive Dashboard. You can add these charts to the default views in the Home Page or to your custom dashboards.

NOTE:

Billing data from your cloud provider generally takes 1 to 2 days to update fully. Savings and investment figures may change until the data is fully updated.

In the Cloud Executive Dashboard, scoped users can see data for the entire cloud environment.

Scale Actions

Scale actions involve moving an entity to a different compute tier or adjusting the entity's allocated capacity. Currently, Intersight Workload Optimizer calculates savings and investments when scaling the following entities:

- VMs (AWS, Azure, and Google Cloud VMs)
- Volumes (AWS and Azure volumes)
- Databases (Azure SQL vCore/DTU and Cosmos DB)
- Document collections (for Cosmos DB)
- Virtual Machine Specs (Azure App Service plans)
- Database Servers (AWS RDS)

Intersight Workload Optimizer calculates savings and investment *per entity*. Calculation is based on the *before-action* and *after-action* costs.

- Before-action cost is the cost of an entity before an action was executed. This is based on the cost snapshot on the day of action execution.
- After-action cost is the cost of an entity as reflected in the daily billing report from the cloud provider.

Savings and investments are the total of all past scaling actions, with the exception that a scale in one direction (such as a scale up) reduces the amounts of previous actions in the opposite direction (such as a scale down), until one or more previous actions have no more effect.

To illustrate, consider three consecutive scale actions for a newly discovered VM and the effect of these actions on the calculated costs. Note that these actions are assumed to have been executed at midnight. Costs are calculated proportionately when actions are executed at other times of the day.

| Day | Executed Action | Before-action Cost | After-action Cost | Investments | Savings |
|-----|-----------------|--------------------|-------------------|------------------------|------------------------|
| 01 | Scale up | \$2 | \$5 | \$5 - \$2 = \$3 | \$0 |
| 02 | Scale down | \$5 | \$4 | \$4 - \$2 = \$2 | \$5 - \$4 = \$1 |
| 03 | Scale down | \$4 | \$2 | \$2 - \$2 = \$0 | \$5 - \$2 = \$3 |

Investment is the difference between the after-action cost on a particular day and the lowest ever before-action cost up to that day (\$2 for all three days in the example). Savings is the difference between the largest ever before-action cost up to a particular day (\$5 on Days 02 and 03) and the after-action cost on that day.

NOTE:

When the result of the calculation is a negative amount, Intersight Workload Optimizer considers savings or investments to be \$0.

Intersight Workload Optimizer reads the daily billing report from the cloud provider to determine the current daily cost of an entity. Calculation also takes into account workload uptime (for VMs) and the effect of consecutive scale actions on the same workload.

Points to consider:

- Calculation can adjust to varying VM uptime and discount coverage over time.
- Calculation stops for terminated entities or entities that Intersight Workload Optimizer no longer discovers.
- Calculation stops if there is a cost change in the entity that is *not* the result of a Intersight Workload Optimizer action.

Delete Actions

Delete actions always result in savings. Currently, Intersight Workload Optimizer calculates savings when deleting the following entities:

- Unattached volumes (AWS, Azure, and Google Cloud volumes)
- Empty Virtual Machine Specs (Azure App Service plans)
- Cosmos databases with provisioned throughput but without any underlying document collection

Intersight Workload Optimizer checks the daily cost of an entity on the day the entity was deleted, and uses that cost to calculate savings from that day onwards. The daily bill from the cloud provider is not used to calculate savings.

Stop and Suspend Actions

Stop and suspend actions always result in savings. Currently, Intersight Workload Optimizer calculates savings when:

- Stopping parkable VMs (AWS, Azure, and Google Cloud VMs)
- Suspending idle dedicated SQL pools (used in Azure Synapse Analytics)

NOTE:

Dedicated SQL pools are represented as Database entities in the supply chain.

Intersight Workload Optimizer checks the hourly cost of an entity's instance type at the time the action was executed, and then uses that cost to calculate savings for the duration of the action. Calculation stops as soon as the entity is powered on. The daily bill from the cloud provider is not used to calculate savings.

Chart Settings

Click the More options icon (), and then select **Edit** to modify the following settings:

■ Scope

By default, the charts are scoped to your global environment. You can change the scope to one or several accounts, billing families, groups, or entities.

■ Timeframe

- Last 7 Days or Last 30 Days – Each data point in the chart shows total savings or investments from previous days until the given day.
- Last Year – Each data point in the chart shows total savings or investments from previous months until the given month.

■ Chart Type

You can set the display to:

- Text and Bar Chart
- Text and Area Chart
- Stacked Bar Chart
- Area Chart

- Text
- Type

Switch between the **Cumulative Savings** and **Cumulative Investments** views. You can also change the displayed data to just **Savings** or **Investments** if you do not wish to see how the savings or investment costs accumulate over time.
- Group By

This setting breaks down data points by the selected group. By default, no group is selected. Each data point only shows the total cumulative values for a given day or month, depending on the selected timeframe.

The following groups are supported:

 - Account – Cloud accounts discovered in your environment
 - Action Type – Scale and delete actions
 - Cloud Service Provider – AWS, Azure and Google Cloud
 - Entity Type – Virtual Machine, Volume, Database (Azure only), Virtual Machine Spec (Azure only)
 - Region – Regions for the entities
 - Resource Group (Azure only) – Available resource groups in your environment
 - Tag – Tag values for a given tag key

Select a tag key from the list. For AWS, the tag keys are limited to cost allocation tags.

For example, if you selected **Account**:

 - The chart represents each account as a distinct color.
 - When you hover on a data point, the chart shows a list of accounts, ordered from largest to smallest cumulative values.
 - The chart legend shows the individual accounts. Click an account to show/hide values for that account.
 - The chart and legend show up to 20 accounts and, if needed, an additional item labeled 'Others' that represents all accounts that are not in the top 20.

Show All

Click **Show All** at the bottom of the chart to view and download data in tabular format.

Data in the table that opens is downloadable. In that table, you can further break down data by resource names. To do this, go to the Details column and click the icon for a particular row. A second table opens. Data in this table is also downloadable.

Legacy Cumulative Savings and Investments Charts

IMPORTANT:

These charts show cumulative savings and investments before the September 2023 release. The method for tracking savings and investments was enhanced in the September 2023 release, and data from this enhanced tracking is available in the **Cumulative Savings** and **Cumulative Investments** charts.

Actions for your cloud workloads usually have cost savings or investments attached to them. For example, deleting unattached volumes can lower your costs significantly (savings), while scaling a VM to a different tier to improve performance could incur additional costs (investments).

These charts highlight:

- Total *realized* savings and investments as a result of executing actions
- Total *missed* savings and investments when actions are not executed

Information in these charts can help shape your action handling policies. For example, you can start automating actions so you don't miss opportunities to assure performance at the lowest possible cost.

Scope

These charts display in the built-in Cloud Executive Dashboard and are scoped to your global environment. You can change the scope to view granular data. You can also add these charts to the default views in the Home Page or to your custom dashboards.

Another way to view granular data is to set the scope (in the supply chain or by using Search) to one or several accounts, billing families, groups, or workloads.

Scale Actions

For actions to scale workloads (VMs, Database Servers, databases, or volumes), Intersight Workload Optimizer calculates savings and investment *per workload* based on the hourly cost of the workload price difference, taking into account workload [uptime \(on page 271\)](#) and the effect of consecutive scale actions on the same workload.

- Calculated investments and savings are the total of all past scaling actions, with the exception that a scale in one direction reduces the amounts of previous actions in the opposite direction, until one or more previous actions have no more effect.

To illustrate:

Consider three consecutive scale actions for a VM and their effect on the calculation.

1. A cost increase of \$1.00 counts as an investment of \$1.00.
2. A subsequent cost decrease of \$0.25 is factored in as:
 - Savings of \$0.25 to the total amount in the Cumulative Savings chart
 - An investment of \$0.75 to the total amount in the Cumulative Investments chart
3. Another cost decrease of \$1.00 is factored in as:
 - Savings of \$1.25 to the total amount in the Cumulative Savings chart
 - An investment of \$0.00 to the total amount in the Cumulative Investments chart

By the time the third action was executed, the initial \$1.00 investment has been "undone" (investment amount is \$0.00) and is no longer considered when calculating savings and investments for the VM.

- Calculation can adjust to varying VM uptime and discount coverage over time.
- Calculation stops for terminated entities or entities that Intersight Workload Optimizer no longer discovers.
- Calculation stops if there is a cost change in the entity that is *not* the result of a Intersight Workload Optimizer action.

Volume Delete Actions

For actions to delete volumes, Intersight Workload Optimizer calculates savings accumulated over one year since volume deletion, based on the hourly cost of the deleted volume. It also estimates missed savings based on the hourly cost of the workload price difference and the number of hours that pending actions remain in the system.

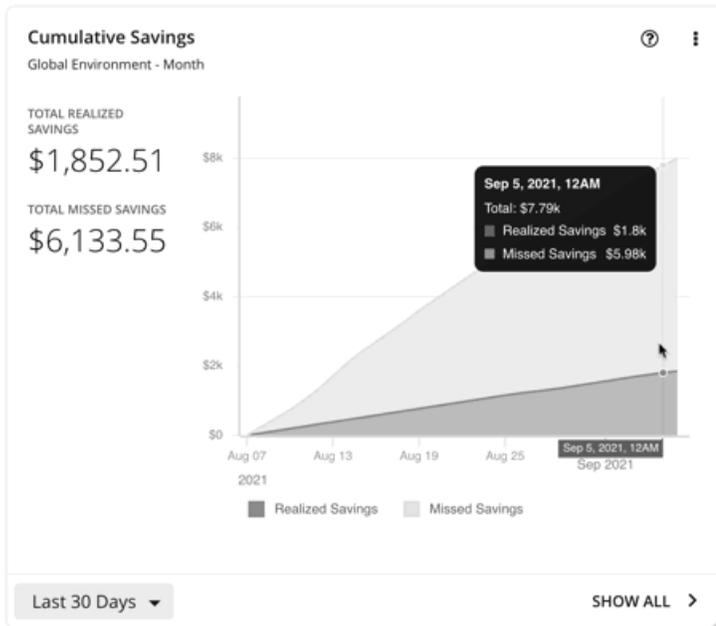
Chart Type

You can set the display to:

- Text and Bar Chart
- Text and Area Chart
- Stacked Bar Chart
- Area Chart
- Text

You can edit the chart to switch between the **Cumulative Savings** and **Cumulative Investments** views. You can also change the displayed data to just **Savings** or **Investments** if you do not wish to see how the savings or investment costs accumulate over time.

In this example, Intersight Workload Optimizer shows monthly realized and missed savings.



In the chart legend, you can click **Realized Savings** or **Missed Savings** to display a filtered view. Click the item again to reset the chart.

Click **Show All** at the bottom of the chart to view and download data in tabular format.

Savings and Investments Charts

IMPORTANT:

The method for tracking savings and investments has been enhanced to accurately report the impact of actions on your cloud expenses. This enhanced tracking began in the September 2023 release. For savings and investments before this release, see the **Legacy Cumulative Savings** and **Legacy Cumulative Investments** charts.

Actions for your cloud workloads usually have cost savings or investments attached to them. For example, deleting unattached volumes can lower your costs significantly (savings), while scaling a VM to a different tier to improve performance could incur additional costs (investments). These charts highlight total savings and investments as a result of executing actions. Intersight Workload Optimizer uses billing data from your cloud provider to calculate savings and investments.

You can add these charts to the default views in the Home Page or to your custom dashboards.

NOTE:

Billing data from your cloud provider generally takes 1 to 2 days to update fully. Savings and investment figures may change until the data is fully updated.

Scale Actions

Scale actions involve moving an entity to a different compute tier or adjusting the entity's allocated capacity. Currently, Intersight Workload Optimizer calculates savings and investments when scaling the following entities:

- VMs (AWS, Azure, and Google Cloud VMs)
- Volumes (AWS and Azure volumes)
- Databases (Azure SQL vCore/DTU and Cosmos DB)
- Document collections (for Cosmos DB)
- Virtual Machine Specs (Azure App Service plans)
- Database Servers (AWS RDS)

Intersight Workload Optimizer calculates savings and investment *per entity*. Calculation is based on the *before-action* and *after-action* costs.

- Before-action cost is the cost of an entity before an action was executed. This is based on the cost snapshot on the day of action execution.
- After-action cost is the cost of an entity as reflected in the daily billing report from the cloud provider.

NOTE:

When the result of the calculation is a negative amount, Intersight Workload Optimizer considers savings or investments to be \$0.

Intersight Workload Optimizer reads the daily billing report from the cloud provider to determine the current daily cost of an entity. Calculation also takes into account workload uptime (for VMs).

Points to consider:

- Calculation can adjust to varying VM uptime and discount coverage over time.
- Calculation stops for terminated entities or entities that Intersight Workload Optimizer no longer discovers.
- Calculation stops if there is a cost change in the entity that is *not* the result of a Intersight Workload Optimizer action.

Delete Actions

Delete actions always result in savings. Currently, Intersight Workload Optimizer calculates savings when deleting the following entities:

- Unattached volumes (AWS, Azure, and Google Cloud volumes)
- Empty Virtual Machine Specs (Azure App Service plans)
- Cosmos databases with provisioned throughput but without any underlying document collection

Intersight Workload Optimizer checks the daily cost of an entity on the day the entity was deleted. The daily bill from the cloud provider is not used to calculate savings.

Stop and Suspend Actions

Stop and suspend actions always result in savings. Currently, Intersight Workload Optimizer calculates savings when:

- Stopping parkable VMs (AWS, Azure, and Google Cloud VMs)
- Suspending idle dedicated SQL pools (used in Azure Synapse Analytics)

NOTE:

Dedicated SQL pools are represented as Database entities in the supply chain.

Intersight Workload Optimizer checks the hourly cost of an entity's instance type at the time the action was executed, and then uses that cost to calculate savings for the duration of the action. Calculation stops as soon as the entity is powered on. The daily bill from the cloud provider is not used to calculate savings.

Chart Settings

Click the More options icon (), and then select **Edit** to modify the following settings:

- Scope

By default, the charts are scoped to your global environment. You can change the scope to one or several accounts, billing families, groups, or entities.
- Timeframe
 - Last 7 Days or Last 30 Days – Each data point in the chart shows total savings or investments for the given day.
 - Last Year – Each data point in the chart shows total savings or investments for the given month.
- Chart Type

You can set the display to:

 - Text and Bar Chart
 - Text and Area Chart
 - Stacked Bar Chart
 - Area Chart

- Text
- **Type**
Switch between the **Savings** and **Investments** views. You can also change the displayed data to **Cumulative Savings** or **Cumulative Investments** to see how the savings or investment costs accumulate over time.
- **Group By**
This setting breaks down data points by the selected group. By default, no group is selected. Each data point only shows the total values for a given day or month, depending on the selected timeframe.
The following groups are supported:
 - Account – Cloud accounts discovered in your environment
 - Action Type – Scale and delete actions
 - Cloud Service Provider – AWS, Azure and Google Cloud
 - Entity Type – Virtual Machine, Volume, Database (Azure only), Virtual Machine Spec (Azure only)
 - Region – Regions for the entities
 - Resource Group (Azure only) – Available resource groups in your environment
 - Tag – Tag values for a given tag key
 Select a tag key from the list. For AWS, the tag keys are limited to cost allocation tags.
For example, if you selected **Account**:
 - The chart represents each account as a distinct color.
 - When you hover on a data point, the chart shows a list of accounts, ordered from largest to smallest values.
 - The chart legend shows the individual accounts. You can click an account to show/hide values for that account.
 - The chart and legend show up to 20 accounts and, if needed, an additional item labeled 'Others' that represents all accounts that are not in the top 20.

Show All

Click **Show All** at the bottom of the chart to view and download data in tabular format.

Data in the table that opens is downloadable. In that table, you can further break down data by resource names. To do this, go to the Details column and click the icon for a particular row. A second table opens. Data in this table is also downloadable.

Legacy Savings and Investments Charts

IMPORTANT:

These charts show savings and investments before the September 2023 release. The method for tracking savings and investments was enhanced in the September 2023 release, and data from this enhanced tracking is available in the **Savings** and **Investments** charts.

Actions for your cloud workloads usually have cost savings or investments attached to them. For example, deleting unattached volumes can lower your costs significantly (savings), while scaling a VM to a different tier to improve performance could incur additional costs (investments).

These charts highlight:

- Total *realized* savings and investments as a result of executing actions
- Total *missed* savings and investments when actions are not executed

Information in these charts can help shape your action handling policies. For example, you can start automating actions so you don't miss opportunities to assure performance at the lowest possible cost.

Scope

To display these charts, add them to the default views in the Home Page or to your custom dashboards. By default, these charts are scoped to your global environment. You can change the scope to view granular data.

Another way to view granular data is to set the scope (in the supply chain or by using Search) to one or several accounts, billing families, groups, or workloads.

Scale Actions

For actions to scale workloads (VMs, Database Servers, databases, or volumes), Intersight Workload Optimizer calculates savings and investment *per workload* based on the hourly cost of the workload price difference, taking into account workload [uptime \(on page 271\)](#).

- Calculation can adjust to varying VM uptime and discount coverage over time.
- Calculation stops for terminated entities or entities that Intersight Workload Optimizer no longer discovers.
- Calculation stops if there is a cost change in the entity that is *not* the result of a Intersight Workload Optimizer action.

Volume Delete Actions

For actions to delete volumes, Intersight Workload Optimizer calculates savings since volume deletion, based on the hourly cost of the deleted volume. It also estimates missed savings based on the hourly cost of the workload price difference and the number of hours that pending actions remain in the system.

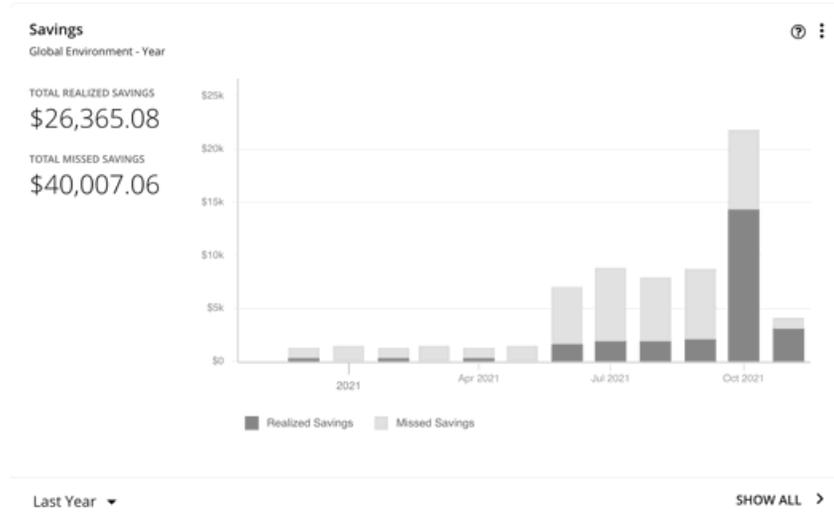
Chart Type

You can set the display to:

- Text and Bar Chart
- Text and Area Chart
- Stacked Bar Chart
- Area Chart
- Text

You can edit the chart to switch between the **Savings** and **Investments** views. You can also change the displayed data to **Cumulative Savings** or **Cumulative Investments** to see how the savings or investment costs accumulate over time.

In this example, the chart shows realized and missed savings per month over the last year. It indicates higher rates of realized savings in the last two months as more actions are executed rather than kept pending.



In the chart legend, you can click **Realized Savings** or **Missed Savings** to display a filtered view. Click the item again to reset the chart.

Click **Show All** at the bottom of the chart to view and download data in tabular format.

Recommended RI Purchases Charts

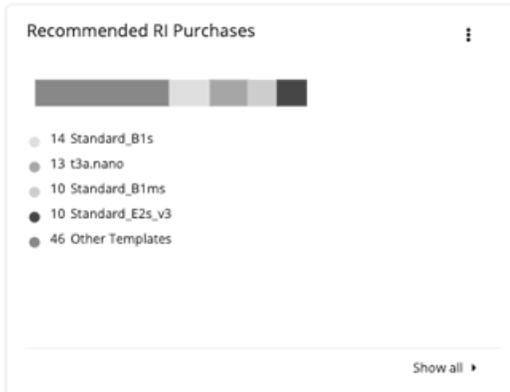
The Recommended RI Purchases chart will be removed from the cloud dashboard in a future release. Use [Action Center \(on page 396\)](#) to view RI purchase recommendations for your cloud workloads.

Intersight Workload Optimizer can recommend purchasing instance types at a discounted rate to help you increase the percentage of VMs covered by discounted pricing and reduce on-demand costs. This chart shows your pending purchases. Download the list of purchases and then send it your cloud provider or representative to initiate the purchase process.

NOTE:

Purchase actions should be taken along with the related VM scaling actions. To purchase discounts for VMs at their current sizes, run a [Buy VM Reservation Plan \(on page 455\)](#).

Currently, Intersight Workload Optimizer can recommend purchasing AWS EC2 RIs and Azure reservations.



Factors Affecting Recommendations

To identify VMs that are good candidates for discounted pricing, Intersight Workload Optimizer analysis considers the history of a VM (by default, the last 21 days), and it looks for:

- **Activity**
If the VM's VCPU utilization percentile is 20% or higher, then Intersight Workload Optimizer considers it an active VM.
- **Stability**
If there have been no start, stop, or resize actions for the VM for 16 of the last 21 days, then Intersight Workload Optimizer considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Intersight Workload Optimizer can recommend purchasing additional discounts.

Intersight Workload Optimizer generates purchase actions on a two-week cycle. It also generates a new set of actions if the discount inventory changes or after the platform restarts.

Be aware of the following:

- Different types of discounts have different costs, so the choice between using on-demand or discounted pricing can vary depending on your [purchase profile \(on page 594\)](#).
- Intersight Workload Optimizer can only estimate costs because the full data is only available after you complete the purchase. Estimates reflect costs you would see after scaling workloads to the newly purchased instance types. For scaling to already-purchased instance types, the chart reflects the actual costs.
- As Intersight Workload Optimizer generates purchase actions, it assumes that any other pending actions for the workload will also be executed. For example, assume a workload running on an r4.xlarge template. If Intersight Workload Optimizer recommends changing that instance type to an m5.medium, it can recommend that you purchase a discounted m5 to cover the workload and reduce costs. This purchase could be on a region that currently doesn't have any m5 workloads. The purchase recommendation assumes you will move the workload to that other region.
- For AWS EC2 RIs:
 - For environments that use the *Instance Size Flexible* rules, Intersight Workload Optimizer can recommend that you buy multiple RIs of smaller instance types to cover the resource requirements of larger instance types. For example, rather than buying one t2.small RI, Intersight Workload Optimizer can recommend that you buy four t2.nano RIs to offer an equivalent discount.
 - For environments that consolidate billing into Billing Families, Intersight Workload Optimizer recommends purchases that are within the given billing family. For more information, see [AWS Billing Families \(on page 599\)](#).

Show All

Click **Show All** to see a table with details for each discount.

The table shows the properties, up-front cost, and break-even period for each discount. The break-even period is the time at which savings will exceed the up-front cost, rounded to the month. The Cost Impact column indicates the monthly savings you would realize when you buy a specific discount.

When you choose one or more check boxes, the total count, up-front cost, and savings appear at the top.

| AWS | | AZURE | | | | | | | | | | | |
|--------------------------|----------|---------------|---------|----------|----------|-------------|--------------------|---------------|-------------------|-----------------|-------------|-----------------|---|
| Buy Actions | | 20 | Savings | | \$799/mo | | | | | | | EXECUTE ACTIONS | ↓ |
| Type to search | | | | | | | | | | | | ADD FILTER | |
| <input type="checkbox"/> | Account | Instance Type | Count | Platform | Term | Payment | Region | Up-Front Cost | Break Even Period | Action Category | Cost Impact | Action | |
| <input type="checkbox"/> | Advanced | r5a.large | 8 | Linux | 1 Year | All Upfront | aws-US West (N. C. | \$5,192 | 8 months | SAVINGS | ↓ \$210/mo | DETAILS | |
| <input type="checkbox"/> | Advanced | c5a.large | 6 | Linux | 1 Year | All Upfront | aws-US West (N. C. | \$2,934 | 7 months | SAVINGS | ↓ \$171/mo | DETAILS | |
| <input type="checkbox"/> | Advanced | r5a.large | 5 | Linux | 1 Year | All Upfront | aws-US East (Ohio) | \$2,910 | 8 months | SAVINGS | ↓ \$133/mo | DETAILS | |

| AWS | | AZURE | | | | | | | | | | | |
|--------------------------|--------------|-----------------|---------|--------|---------------------|---------------|-------------------|-----------------|-------------|---------|--|-----------------|---|
| Buy Actions | | 27 | Savings | | \$928/mo | | | | | | | EXECUTE ACTIONS | ↓ |
| Type to search | | | | | | | | | | | | ADD FILTER | |
| <input type="checkbox"/> | Subscription | Product Name | Quan... | Term | Region | Up-Front Cost | Break Even Period | Action Category | Cost Impact | Action | | | |
| <input type="checkbox"/> | Dev | Standard_D2s_v3 | 6 | 1 Year | azure-North Central | \$3,042 | 7 months | SAVINGS | ↓ \$173/mo | DETAILS | | | |
| <input type="checkbox"/> | Dev | Standard_B1ls | 49 | 1 Year | azure-East US | \$1,323 | 7 months | SAVINGS | ↓ \$73/mo | DETAILS | | | |
| <input type="checkbox"/> | Dev | Standard_DS1_v2 | 2 | 1 Year | azure-Brazil South | \$798 | 6 months | SAVINGS | ↓ \$57/mo | DETAILS | | | |

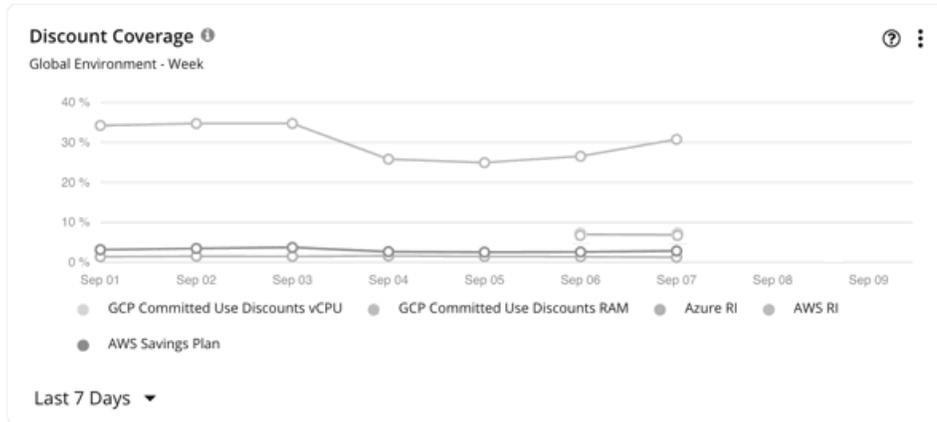
Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Discount Coverage Chart

This chart shows the percentage of cloud workloads (VMs and RDS database servers) covered by discounts. For VMs covered by discounts, you can reduce your costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.



To identify VMs that are good candidates for discounted pricing, Intersight Workload Optimizer analysis considers the history of a VM (by default, the last 21 days), and it looks for:

- **Activity**
If the VM's VCPU utilization percentile is 20% or higher, then Intersight Workload Optimizer considers it an active VM.
- **Stability**
If there have been no start, stop, or resize actions for the VM for 16 of the last 21 days, then Intersight Workload Optimizer considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Intersight Workload Optimizer can recommend purchasing additional discounts.

Timeframe

The chart shows daily data points and supports the following timeframes:

- Last 7 Days
- Last 30 Days

AWS EC2/RDS Reserved Instances and Compute Savings Plans

Intersight Workload Optimizer uses data exports from AWS to calculate Reserved Instance (RI) and Compute Savings Plan coverage. Intersight Workload Optimizer supports a data export created at the management account, but not member accounts. For Intersight Workload Optimizer to use these data exports, you must add a billing target in the Target Configuration page.

The chart shows historical data. Data for the current day may not be available until the cloud provider has exported it fully.

Hover on a data point in the chart to see the following information:

- The date and time for the data point
- The percentage of coverage. For RIs, coverage is based on normalization factors.

[Normalization factor](#) is a measure of RI capacity that you can use to compare or combine the capacity for different instance families.

Intersight Workload Optimizer measures RI coverage in terms of normalization factors. It compares the number of RIs calculated as normalization factors that cover workload capacity with the total number of normalization factors for a given Intersight Workload Optimizer scope. Each workload is assigned normalized units depending on its instance type.

Points to consider:

- **Scope**
 - If you set the scope to a specific AWS account, the chart shows the RI coverage for the workloads for the account and any RIs for the billing family.
 -
- **RIs**

In AWS, you can turn off RI discount sharing for specific accounts. These accounts will not share any discounts with other accounts. Intersight Workload Optimizer does not recognize RI coverage or utilization for these accounts. For example, the RI Coverage and RI Utilization charts will show zero values.

Azure Reservations

Intersight Workload Optimizer uses billed cost data from Azure to calculate reservation coverage for workloads. For Intersight Workload Optimizer to use billed cost data, you must add an Azure Billing target in the Target Configuration page.

The chart shows historical data. Data for the current day may not be available until the cloud provider has exported it fully.

Hover on a data point in the chart to see the following information:

- The date and time for the data point
- The percentage of coverage, based on ratios.

[Ratio](#) refers to the number of Azure reservation units that cover workload capacity compared to the total number of reservation units for a given Intersight Workload Optimizer scope. Each workload is assigned reservation units based on its instance type.

If you set the scope to a specific Azure subscription, this chart shows the reservation coverage for the workloads for the subscription, plus any shared reservations and single-scope reservations owned by this subscription.

Google Cloud Committed Use Discounts

Intersight Workload Optimizer uses billing data from Google Cloud to calculate CUD coverage for VM vCPU (cores) and VM memory. For Intersight Workload Optimizer to use billing data, you must add a Google Cloud Billing target in the Target Configuration page.

The chart shows historical data. Data for the current day may not be available until the cloud provider has exported it fully.

Hover on a data point in the chart to see the following information:

- The date and time for the data point
- The percentage of coverage

Points to consider:

- CUD coverage in this chart and in the Google Cloud console may not match for the following reasons:
 - Intersight Workload Optimizer uses UTC, while the Google Cloud console uses local time.
 - CUD coverage for VMs running [custom machine types](#) is reported by Google Cloud but not by Intersight Workload Optimizer. If the chart's scope includes these VMs, aggregated data may not match the data shown in Google Cloud. If the chart is scoped solely to these VMs, the chart will not display data.
 - Google Cloud reports CUD coverage for [resource reservations](#), which Intersight Workload Optimizer does not monitor. When CUDs cover both VMs and resource reservations, Google Cloud reports coverage for both, while Intersight Workload Optimizer only reports coverage for VMs.
- When you configure a billing export in Google Cloud, you can choose to export standard or detailed usage cost data. Detailed usage cost data is recommended because it includes granular data that Intersight Workload Optimizer can display in charts, such as discount coverage for individual VMs.

If you exported standard usage cost data and configured your Google Cloud Billing target to use this data:

- Coverage data in the chart is only available for scopes that are larger than individual VMs, such as projects or folders. This is because standard usage cost data does not include CUD coverage for individual VMs.
- Intersight Workload Optimizer uses retail pricing to calculate coverage. This may result in an underestimation of the coverage if the negotiated price for a machine type is lower than the retail price.

NOTE:

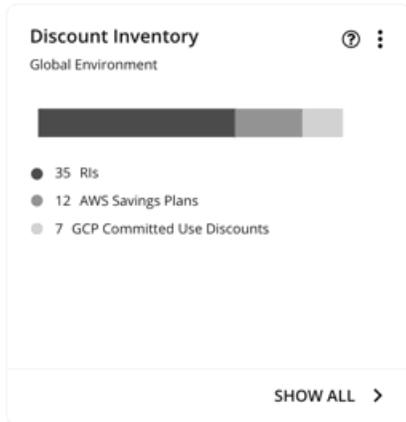
The Google Cloud Billing target uses standard usage cost data if you disabled the **Enable Resource Level Detail From Cost Export Table** option in the Target Configuration page for the target.

- When coverage reaches 100%, a rounding issue might result in a value that is slightly higher than 100% (such as 100.01%).

Discount Inventory Chart

This chart lists the cloud provider discounts discovered in your environment.

- AWS EC2 Reserved Instances (RIs)
- AWS Compute Savings Plans
- AWS RDS Reserved DB Instances
- Azure reservations
- Google Cloud committed use discounts



Timeframe

The chart shows daily data points and supports the following timeframes:

- Last 7 Days
- Last 30 Days

Scope

You can set the scope to your global cloud environment or to individual AWS accounts, Azure subscriptions, Google Cloud folders, billing families, or regions.

Show All

Click **Show All** at the bottom of the chart to see detailed information for the discounts in scope. If your scope includes multiple cloud providers, each provider will have its own tab.

| Discount Inventory | | | | | | | | | | | | | | | | | | | | | | | |
|--------------------------|---------------|---------------|-------|-------------------|----------|---------|-------------|--------------------------|---------------------|--------------|--------|-----------------------------|----------------|------|---------------|---------|---------------|--|--|--|--|--|--|
| AWS RESERVED INSTANCES | | | | AWS SAVINGS PLANS | | | | AZURE RESERVED INSTANCES | | | | GCP COMMITTED USE DISCOUNTS | | | | | | | | | | | |
| Total 8 | | | | | | | | | | | | Count | 12 | Cost | \$1,790.83/mo | Savings | \$1,434.16/mo | | | | | | |
| Reservation ID | Account | Instance Type | Count | Location | Platform | Tenancy | Class | Payment | Current Utilization | Est. Savings | Term | Exp. Date | Effective Cost | | | | | | | | | | |
| <input type="checkbox"/> | Quality Engin | t3.nano | 1 | aws-ap-north | Linux | Default | Convertible | All Upfront | 0% | \$0.00/mo | 1 Year | Nov 02, 2022 | \$3.67/mo | | | | | | | | | | |
| <input type="checkbox"/> | Quality Engin | t3.nano | 1 | aws-ap-north | Linux | Default | Convertible | All Upfront | 0% | \$0.00/mo | 1 Year | Nov 02, 2022 | \$3.67/mo | | | | | | | | | | |
| <input type="checkbox"/> | Development | t3.micro | 1 | aws-US East (| Linux | Default | Convertible | No Upfront | 100% | \$2.12/mo | 1 Year | Feb 28, 2023 | \$5.48/mo | | | | | | | | | | |

Each row in the table corresponds to a discount. Note that there can be several discounts for an Azure subscription or AWS/Google Cloud account, and each discount displays as its own row. Table columns show basic information obtained from the cloud provider, such as the name/ID of the discount, the subscription/account that uses that discount, instance type and location, term, and expiration dates. Click a subscription/account to narrow the scope.

The table supports the following general functionality:

- **Totals:** At the top of the page, you will see the total number of discovered discounts. For AWS RIs and Azure reservations, you will also see total costs and savings. As you select one or more checkboxes, the information changes to reflect the totals for your selections.
- **Column Sorting:** Click any column heading to sort the list.
- **Download:** Click the Download icon at the upper right section of the page to download the table as a CSV file.

Azure Reservations and AWS EC2 RIs

When you add an Azure EA account or an AWS management account as your primary cloud target, Intersight Workload Optimizer gains full insight into the discounts for your billing families. Even as you selectively add Azure subscriptions or AWS member accounts as secondary targets, Intersight Workload Optimizer remains aware of all discounts and how they are utilized across the board, and can thus recommend more accurate discount optimization and purchase actions.

Points to consider:

- For AWS, if you added some member accounts as targets, but not a management account, Intersight Workload Optimizer will not reflect discounts for member accounts that you have not added as targets.
- For Azure:
 - It could take Intersight Workload Optimizer up to a day to discover newly purchased Azure reservations.
 - There can be delays in billing information updates that Azure makes available to Intersight Workload Optimizer. If this happens, analysis might use partial billing data in its calculations and show incomplete costs for discounts from non-added Azure subscriptions.

Set the scope to your global environment to view the full inventory. When you click **Show All** at the bottom of the chart, pay attention to the following information shown in the table:

- For discounts in *added* accounts (Azure subscriptions or AWS member accounts):

| <input type="checkbox"/> | Reservation ID | Account | Instance Type | Count | Location | Platform | Tenancy | Class | Payment | Est. Current Utilization | Est. Savings | Term | Exp. Date | Effective Cost |
|--------------------------|----------------|-------------|---------------|-------|-----------------|----------|---------|-------------|-------------|--------------------------|---------------|---------|--------------|----------------|
| <input type="checkbox"/> | | Development | t3a.micro | 2 | aws-Asia Pacif | Linux | Default | Convertible | All Upfront | 54% | \$1.65/mo | 1 Year | Dec 05, 2023 | \$6.00/mo |
| <input type="checkbox"/> | | Development | m5.metal | 1 | aws-US East (I) | Linux | Default | Convertible | No Upfront | 100% | \$1,657.83/mo | 3 Years | Feb 23, 2025 | \$1,706.01/mo |

- **Account** column (for AWS) or **Subscription** column (for Azure)

This column shows the account name for the discount. Click the name to set the scope to that account. Note that there can be several discounts for an account, and each discount displays as its own row.

NOTE:

If there is a failure to re-validate the account for some reason, Intersight Workload Optimizer shows it as a *non-added* account in the Discount Inventory page.

- **Est. Current Utilization** column

This column shows the percentage of discount capacity currently used by VMs in all accounts. Intersight Workload Optimizer estimates the percentage if there are VMs in non-added accounts that use the discount (since the exact number of VMs is unknown).

- For discounts in *non-added* accounts:

| <input type="checkbox"/> | Reservation ID | Account | Instance Type | Count | Location | Platform | Tenancy | Class | Payment | Est. Current Utilization | Est. Savings | Term | Exp. Date | Effective Cost |
|--------------------------|----------------|---------|---------------|-------|----------------|----------|-----------|-------------|-------------|--------------------------|--------------|--------|--------------|----------------|
| <input type="checkbox"/> | | Quality | t3a.micro | 2 | aws-Asia Pacif | Linux | Default | Convertible | All Upfront | 54% | \$1.65/mo | 1 Year | Dec 05, 2023 | \$6.00/mo |
| <input type="checkbox"/> | | Quality | c5a.large | 1 | aws-Canada (C) | Linux | Dedicated | Convertible | No Upfront | 0% | \$0.00/mo | 1 Year | Jul 19, 2024 | \$51.82/mo |
| <input type="checkbox"/> | | Quality | t3.nano | 3 | aws-EU (Paris) | Linux | Default | Convertible | All Upfront | 54% | \$1.72/mo | 1 Year | Jan 09, 2024 | \$9.75/mo |

- **Account** column (for AWS) or **Subscription** column (for Azure)

This column shows a grayed-out, non-clickable name to indicate that you have not added the account as a target. Intersight Workload Optimizer is aware of this account and if the given discount is shared with other accounts because you have added a management or EA account.

- **Est. Current Utilization** column

This column shows the percentage of discount capacity currently used by VMs in all accounts. Intersight Workload Optimizer estimates the percentage if there are VMs in non-added accounts that use the discount (since the exact number of VMs is unknown).

AWS Savings Plans

If you added targets that are AWS accounts with read-only access to the AWS Savings Plans API, Intersight Workload Optimizer uses this chart to present the Savings Plans that it discovered in your cloud environment and the instance types they use.

| <input type="checkbox"/> Savings Plan ID | Account ⓘ | Type | Payment | Instance Family | Location | Commitment | Term | Start Date | Exp. Date |
|--|-----------|---------|-------------|-----------------|----------|------------|---------|--------------|--------------|
| <input type="checkbox"/> 55555 ... | Prod | Compute | All Upfront | All | All | \$0.001/hr | 1 Year | Dec 28, 2020 | Dec 28, 2021 |
| <input type="checkbox"/> 45555 ... | Prod | EC2 | No Upfront | t3 | aws-... | \$0.001/hr | 3 Years | Dec 29, 2020 | Dec 29, 2023 |

Google Cloud Committed Use Discounts

Intersight Workload Optimizer discovers committed use discounts (CUDs) for your workloads when you add a Google Cloud Billing target in the Target Configuration page.

| <input type="checkbox"/> Name | Account ⓘ | Status | Region | Type | Payment | Instance Family | Cores | Memory | Term | Start Date | End Date |
|---------------------------------------|-----------|--------|---------------------|---------------|-------------|-----------------|-------|--------|---------|--------------|--------------|
| <input type="checkbox"/> commitment-1 | | Active | us-west4 | Family Scoped | All upfront | N2 | 1 | 4 GB | 3 Years | Oct 15, 2021 | Oct 15, 2024 |
| <input type="checkbox"/> commitment-1 | | Active | eu-north1 | Family Scoped | All upfront | N2 | 1 | 2 GB | 1 Year | Oct 21, 2021 | Oct 21, 2022 |
| <input type="checkbox"/> commitment-1 | | Active | northamerica-northe | Family Scoped | All upfront | N2D | N/A | 2 GB | 1 Year | Jan 27, 2022 | Jan 27, 2023 |
| <input type="checkbox"/> commitment-2 | | Active | us-central1 | Family Scoped | All upfront | E2 | N/A | 4 GB | 1 Year | Oct 21, 2021 | Oct 21, 2022 |

This chart also shows the current overall utilization of your CUD inventory for VM vCPU (cores) and VM memory, as estimated by Intersight Workload Optimizer analysis.

Discount Utilization Chart

This chart shows how well you have utilized your current discount [inventory \(on page 560\)](#). The desired goal is to maximize the utilization of your inventory and thus take full advantage of the discounted pricing offered by your cloud provider.



Timeframe

The chart shows daily data points and supports the following timeframes:

- Last 7 Days
- Last 30 Days

AWS EC2/RDS Reserved Instances and Compute Savings Plans

Intersight Workload Optimizer uses data exports from AWS to calculate Reserved Instance (RI) and Compute Savings Plan utilization. Intersight Workload Optimizer supports a data export created at the management account, but not member accounts. In order for Intersight Workload Optimizer to use these data exports, you must add a billing target in the Target Configuration page.

The chart shows historical data. Data for the current day may not be available until the cloud provider has exported it fully.

Hover on a data point in the chart to see the following information:

- The date and time for the data point
- The percentage of utilization

For Compute Savings Plans, utilization is based on the total utilized and committed costs per day.

For RIs, utilization is based on normalization factors.

[Normalization factor](#) is a measure of RI capacity that you can use to compare or combine the capacity for different instance families.

Intersight Workload Optimizer measures RI coverage in terms of normalization factors. It compares the number of RIs calculated as normalization factors that cover workload capacity with the total number of normalization factors for a given Intersight Workload Optimizer scope. Each workload is assigned normalized units depending on its instance type.

Points to consider:

- Scope
 - You can set the scope to your global cloud environment or to individual accounts, billing families, or regions. Scoping to an account shows the RI utilization for the workloads for the entire billing family.
 -
 -
- RIs
 - In AWS, you can turn off RI discount sharing for specific accounts. These accounts will not share any discounts with other accounts. Intersight Workload Optimizer does not recognize RI coverage or utilization for these accounts. For example, the RI Coverage and RI Utilization charts will show zero values.
 - Under very rare circumstances, you can have RIs on payment plans that do not resolve to 1-year or 3-year terms. In this case, AWS does not return pricing data for those RIs. Intersight Workload Optimizer does not include such RIs in its calculations of RI utilization or RI cost.

Azure Reservations

Intersight Workload Optimizer uses billed cost data from Azure to calculate reservation utilization for workloads. In order for Intersight Workload Optimizer to use billed cost data, you must add an Azure Billing target in the Target Configuration page.

The chart shows historical data. Data for the current day may not be available until the cloud provider has exported it fully.

Hover on a data point in the chart to see the following information:

- The date and time for the data point
- The percentage of utilization, based on ratios.

[Ratio](#) refers to the number of Azure reservation units that cover workload capacity compared to the total number of reservation units for a given Intersight Workload Optimizer scope. Each workload is assigned reservation units based on its instance type.

You can set the scope to your global cloud environment or to individual subscriptions, billing families, or regions. Scoping to a subscription shows utilization for workloads for the entire billing family or for single and shared subscriptions.

Google Cloud Committed Use Discounts

This chart does not show *historical* utilization data for committed use discounts (CUDs) because Google Cloud does not track this data. If Intersight Workload Optimizer only manages Google Cloud targets, this chart will be empty.

NOTE:

Intersight Workload Optimizer uses billing data from Google Cloud to calculate the *current* utilization of your CUD inventory. To view current utilization data, see the Discount Inventory chart.

Cloud Estimated Cost Charts

Cloud Estimated Cost charts show estimated monthly costs and investments for the cloud. Monthly cost amounts are summarized as amounts with and without actions.

Display

The chart shows the information as a Text chart.

Volume Summary Charts

To help you manage your costs on the public cloud, these charts show the distribution and costs of volumes for the given scope. In this way, you can see how volume utilization affects your costs. For these charts, Intersight Workload Optimizer calculates the costs based on the cost information from the cloud targets.

These charts show the following data:

- Tiers

The chart breaks down volumes by tier (disk type) and shows the volume count and monthly cost for each tier.

- Volume State

The chart breaks down volumes by attachment state (attached or unattached) and shows the volume count and monthly cost for each state. For unattached volumes, you can reduce your cloud cost by the given amount if you delete these volumes. Click **Show All** and then click the **Details** button for an unattached volume to execute a delete action.

NOTE:

For an Optimize Cloud plan, the Volume Tier Summary chart shows 'Current' and 'Optimized' results. The 'Current' result includes currently unattached volumes that you can delete to reduce costs, while the 'Optimized' result assumes that unattached volumes have been deleted. To see a list of unattached volumes, click **Show Changes** at the bottom of the chart. For details about Optimize Cloud plans, see [Optimize Cloud Plan Results \(on page 452\)](#).

For a detailed breakdown, click **Show All** at the bottom of the chart. If you have multiple cloud providers, each provider will have its own tab. Click any column heading to sort the list. When you choose one or more check boxes, the total appears at the top.

NOTE:

For Azure environments with VMs in Scale Sets, for any VMs that are powered off, the associated volume shows a utilization of zero GB. This is an accurate presentation of the data that the Azure environment returns for such a powered-off VM. However, it is likely that some of the volume capacity is currently utilized.

Chart Unit

If you are scoped to a particular cloud provider, you can sort tiers and volumes by clicking the Edit option at the top-right corner of the chart, and then choosing one of the following units:

- **Count** – Sort by volume count, from largest to smallest.
- **Cost** – Sort by monthly cost, from highest to lowest.

On-Prem Chart Types

These chart widgets provide information on the status of your on-prem environment.

Density Charts

Density charts show the number of resource consumers (virtual machines or containers) per provider (host or storage). If available, choose the **Show Density** checkbox to see the ratio of consumers to providers.

These charts also show the desired count of virtual machines, assuming you want to fill the headroom completely. Note that the Desired Workloads values are the results of running plans. These plans can calculate workload moves within a cluster to gain more efficiency, but they always respect the cluster boundaries – the plans never move VMs to hosts on different clusters.

To display relevant data, you must set the scope to your global environment or a cluster group. Other scopes are not supported.
 To display relevant data, you must set the scope to your global environment or a cluster group. Other scopes are not supported.

Chart Type

You can set the display to:

- Stacked Bar Chart
- Line Chart

Ports Charts

Ports charts show the most utilized northbound or southbound ports in your on-prem environment over a given time period. These charts are useful in Fabric environments where you license port channels.

Display

The chart shows the information as Tabular.

Headroom Charts

Headroom charts show how much extra capacity your clusters have to host workloads.

To calculate cluster capacity and headroom, Intersight Workload Optimizer runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current VMs to other hosts within the given cluster, if such moves would result in a more desirable workload distribution. The result of the plan is a calculation of how many more VMs the cluster can support.

To calculate VM headroom, the plan simulates adding VMs to your cluster. The plan assumes a certain capacity for these VMs, based on a specific VM template. For this reason, the count of VMs given for the headroom is an approximation based on that VM template.

You can specify the following types of Headroom charts:

- CPU Headroom
- Memory Headroom
- Storage Headroom

Commodity

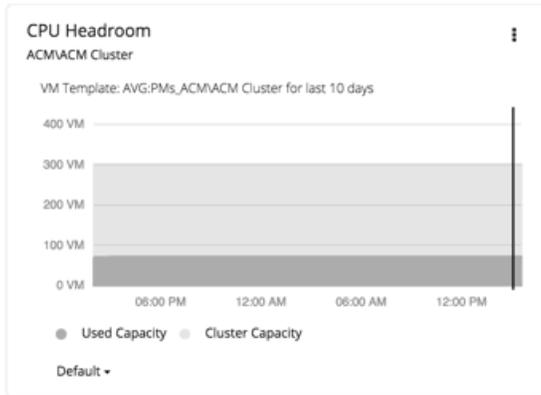
You can choose:

- CPU Headroom
- Memory Headroom
- Storage Headroom

Display

The chart shows the information as an Area chart.

Example:



Exhaustion Time Chart

This chart shows the current growth of workloads and projects when workloads will exceed the capacity of your current infrastructure. This is useful for future planning (for example, if you might need to buy more hardware).

You can track CPU, memory, and storage as well as the average monthly Virtual Machine growth and the average VM template. The amount of time is presented as days. For example, storage will be used up in 41 days.

Display

The chart shows the information as a Text chart.

Creating Groups

NOTE:

This page includes enhancements and a more modern look-and-feel that are only available when you enable the new design framework. To switch to the new framework, click the React icon  in the navigation bar of the user interface and then Turn ON the toggle. For more information, see "Design Framework for the User Interface" in the *User Guide*.

Groups assemble collections of resources for Intersight Workload Optimizer to monitor and manage. When setting scope for your Intersight Workload Optimizer session, you can select groups to focus on those specific resources. For example, if you have a number of VMs devoted to a single customer, you can create a group of just those VMs. When running a planning scenario you can set the scope to work with just that group.

Intersight Workload Optimizer discovers groups that exist in your environment and allows you to create custom groups.

Intersight Workload Optimizer supports two custom-grouping methods:

- **Dynamic** – You define these groups by specific criteria. You can group services according to naming conventions (all VM names that start with **ny**), resource characteristics (all hosts with four CPUs), or other criteria such as time zone or number of CPUs.
 - These groups are dynamic because Intersight Workload Optimizer updates the group as conditions change.
- **Static** – You create these groups by selecting the specific group members.

NOTE:

Do not use the Intersight Workload Optimizer user interface to delete discovered groups. If you do, later analysis cycles will discover them again, and add them to your environment. In the meantime, any analysis that relies on those groups can give unexpected results.

You *can* delete any custom group you have created. Before you do, verify that you do not have any charts, plans, or policies that use the group you want to delete. After you delete the group, such charts, plans, or policies will lose their scope. For example, a policy with no scope has no effect.

To create a group:

1. Navigate to the Settings Page.

Click **More**, then display the Settings Page. From there, you can perform a variety of Intersight Workload Optimizer configuration tasks.

2. Choose **Groups**.

Click to navigate to the Group Management Page.

This page lists all the custom groups that you currently have configured for Intersight Workload Optimizer.

3. Perform tasks in the Group Management page.

You can perform any of the following tasks:

- Filter groups by type
- Search for groups
- Click a group name to edit it

For a dynamic group, you can edit the set of criteria that select the group members. For a static group, you can add or subtract specific members.

- Expand an entry to see group details
- Select an entry to delete the group
- Create new groups

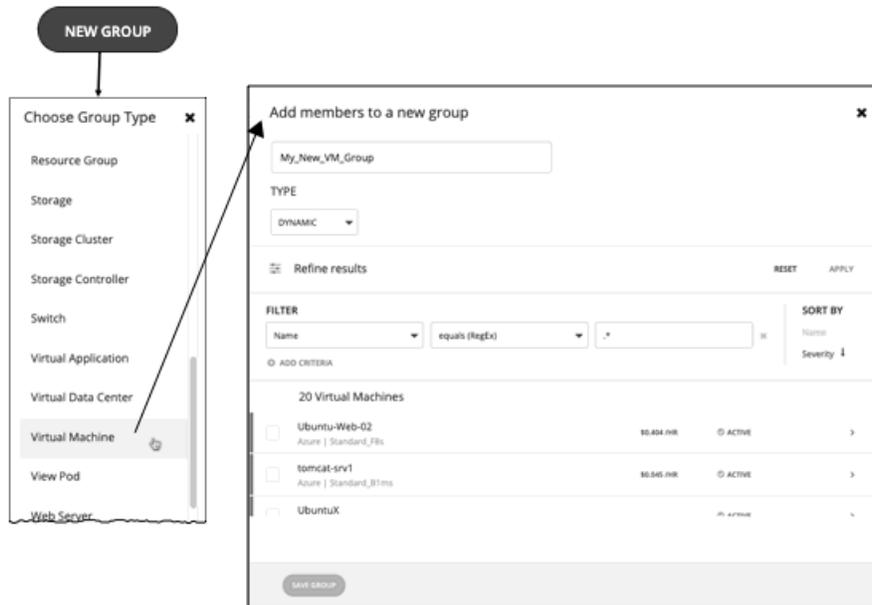
To work with a long list of groups, you can filter by group type. For example, only show groups of VMs, or groups of host machines. You can also type a string in the **Search** field to filter the list.

4. Expand an entry to see group details.

The details show you information about related entities such as how many hosts provide resources for a group of VMs. If there are any pending actions for the group, the details list those actions as well.

5. Create a new group.

Click **NEW GROUP**, choose a group type, and then specify the group settings.



- Give the group a unique name. To prevent issues, you should never use duplicate names for groups of the same entity type.
- Set whether the group will be static or dynamic.
 - To create a static group, select the member entities from the list. To filter the list, set group criteria.
 - To create a dynamic group, set group criteria. The list updates to show the resulting group members.
- Specify group criteria.
 - These criteria are entity attributes that determine group membership. You might create a group of all VMs that have 4 VCPUs. You can choose properties of the member entities, and you can choose properties of entities that are related to the members. For example, you can make a group of VMs that are hosted by PMs with the substring "Development" in their names.
 - As you set criteria, the list of entities updates to show the member entities. You also can sort the list by severity (per the most critical entity in group) or group name.
 - Note that you can use regular expression to express your match strings.
- When you are finished, save the group.
 - Save** adds this group to the **My Groups** collection.

Working With Policies

Policies set business rules to control how Intersight Workload Optimizer analyzes resource allocation, displays resource status, and recommends or executes actions. Intersight Workload Optimizer includes the following fundamental types of policies:

- Placement Policy
 - A placement policy is a set of rules that Intersight Workload Optimizer must satisfy when generating placement recommendations for on-prem and containerized workloads.
 - For details, see [Placement Policies \(on page 569\)](#).
- Automation Policy
 - An automation policy is a set of rules that Intersight Workload Optimizer must satisfy when executing [non-parking actions \(on page 405\)](#) on public cloud or on-prem workloads, or changing settings that affect analysis and action generation.
 - For details, see [Automation Policies \(on page 574\)](#).

Policy Management

Use the Policy Management page to view and manage policies. The page shows the following categories of policies:

- All Policies – All the currently defined policies
- Placement Policies
 - Imported Placement Policies – Placement policies [discovered \(on page 569\)](#) from your targets
 - Intersight Workload Optimizer Segments – Placement policies [created \(on page 569\)](#) from the Policy Management page
- Automation Policies
 - Imported Automation Policies – Automation policies [discovered \(on page 577\)](#) from your targets
 - User Defined Automation Policies – Automation policies [created \(on page 577\)](#) from the Policy Management page
 - Defaults – [Default \(on page 575\)](#) automation policies for the different entity types

NOTE:

To see the placement or automation policies that are applied by a particular entity, set the scope to that entity (from Search or the supply chain), and then click the Policies tab. For more information, see [Scope Policies \(on page 37\)](#).

Placement Policies

A placement policy is a set of rules that Intersight Workload Optimizer must satisfy when generating placement recommendations for on-prem and containerized workloads.

With these policies, Intersight Workload Optimizer can recommend placement actions that comply with your business rules. For example, a placement policy can constrain VMs to specific hosts. If a VM needs to move to a different host due to resource congestion, Intersight Workload Optimizer will generate an action to move the VM to one of the hosts defined in the policy.

Intersight Workload Optimizer discovers placement policies from your targets, and allows you to create your own policies.

NOTE:

You can enable or disable any placement policy to affect placement calculations in the real-time environment or in plans.

When calculating workload placement, Intersight Workload Optimizer respects cluster boundaries, networks, and provisioned data stores. In addition, the configuration of your environment can specify logical boundaries, and you can create even more boundaries within Intersight Workload Optimizer. These boundaries impose segments on the market that Intersight Workload Optimizer uses to model your application infrastructure.

In finance, a market segment divides the market according to the criteria different groups of people use when they buy or sell goods and services. Likewise in the Intersight Workload Optimizer market, a workload placement segment uses criteria to focus the buying and selling of resources within specific groups of entities. This gives you finer control over how Intersight Workload Optimizer calculates placements.

Imported Placement Policies

Your on-prem targets can include placement policies of their own. Intersight Workload Optimizer imports these placement policies, and considers them to be constraints on placement. You cannot disable these imported policies for real-time analysis, but you can disable them for plans.

Intersight Workload Optimizer imports the following:

- vCenter Server DRS Rules

See [Other Information Imported from vCenter \(on page 182\)](#)

To view imported placement policies, navigate to **Settings > Policies**, and then click **Imported Placement Policies**.

Creating Placement Policies

A placement policy is a set of rules that Intersight Workload Optimizer must satisfy when generating placement recommendations for on-prem and containerized workloads.

With these policies, Intersight Workload Optimizer can recommend placement actions that comply with your business rules. For example, a placement policy can constrain VMs to specific hosts. If a VM needs to move to a different host due to resource congestion, Intersight Workload Optimizer will generate an action to move the VM to one of the hosts defined in the policy.

Intersight Workload Optimizer discovers placement policies from your targets, and allows you to create your own policies.

You can create the following placement policies:

- **Place** - Determine which entities use specific providers.

For example, the VMs in a consumer group can only run on a host that is in the provider group. You can limit the number of consumers that can run on a single provider - for hosts in the provider group, only 2 instances of VMs in the consumer group can run on the same host. Or no more than the specified number of VMs can use the same storage device.

- **Don't Place** - Consumers must never run on specific providers.

For example, the VMs in a consumer group can never run on a host that is in the provider group. You can use such a segment to reserve specialized hardware for certain workloads.

- **Merge** - Merge clusters into a single provider group.

For example, you can merge three host clusters in a single provider group. This enables Intersight Workload Optimizer to move workload from a host in one of the clusters to a host in any of the merged clusters to increase efficiency in your environment.

- **License** - Set up hosts to provide licenses for VMs.

For VMs that require paid licenses, you can create placement policies that set up certain hosts to be the VMs' preferred license providers. Intersight Workload Optimizer can then recommend consolidating VMs or reconfiguring hosts in response to changing demand for licenses.

1. Open the Settings Page.

Click to open the Settings Page. From there, you can perform various Intersight Workload Optimizer configuration tasks.

2. Choose Policies.

Click to open the Policy Management Page.

This page lists all the policies that you currently configured for Intersight Workload Optimizer.

3. Create a Placement policy.

NEW POLICY

↓

Create a new policy

POLICY NAME *

TYPE

Place
▼

PLACE

Choose consumer type...
▼

ON

Choose provider type...
▼

Limit workload entities to placement group

Limit the maximum number of workload entities per placement entity to:

SAVE POLICY

First, select the type of Placement policy to create, then specify the settings:

- Give the policy a name.
- Choose the policy type and make the settings.
- Save the policy when you're done.

4. Create a **Place** policy.

TYPE

Place
▼

PLACE

Choose consumer type...
▼

ON

Choose provider type...
▼

Limit workload entities to placement group

Limit the maximum number of workload entities per placement entity to:

These policies control where the workload can be placed. For example, you can specify that a VM will be placed only on a host that is a member of a specific cluster. Alternatively, you might specify that any applications in a specific group can be placed only on a data store that is a member of a specific group.

- **Specify the consumer group** - The group or cluster of entities that will be placed on the identified providers.
- **Specify the provider group** - The group or cluster of entities that provide resources to the consumers.
- **Limit workload entities to placement group** - Set the policy to place consumer entities only on members of the provider group.
- **Limit the maximum number of workload entities per placement entity to** - Limit how many instances of the consumer entities can be placed on a single provider.

5. Create a **Don't Place** policy.

TYPE

Don't Place ▼

DON'T PLACE

Containers ▼

+ SELECT GROUP OF CONTAINERS

ON

Virtual Machines ▼

+ SELECT GROUP OF VIRTUAL MACHINES

These policies identify groups or clusters that will never host the consumer entities. For example, you can specify that a VM will never be placed on a host that is a member of a specific cluster. Alternatively, you can specify that a set of noncritical applications will never be placed on specialized hardware, as a way to ensure availability for critical applications.

- **Specify the consumer group** - The group or cluster of entities that will be excluded from the identified providers.
- **Specify the provider group** - The group or cluster of entities that will not provide resources to the consumers.

6. Create a **Merge** policy.

TYPE

Merge ▼

MERGE

Host Clusters ▼

+ SELECT CLUSTERS

You can create placement policies that merge multiple clusters into a single logical group for workload placement.

For example, your environment might divide hosts into clusters according to hardware vendor, or by some other criteria. Workload placement typically does not cross such cluster boundaries. However, there might be no technical reason to apply these boundaries to workload placement. By creating a larger pool of provider resources, Intersight Workload Optimizer has even more opportunities to increase efficiency in your environment.

Combining merge and placement policies can provide significant operational advantages during hardware refreshes or data center migrations.

For merge policies, keep the following considerations in mind:

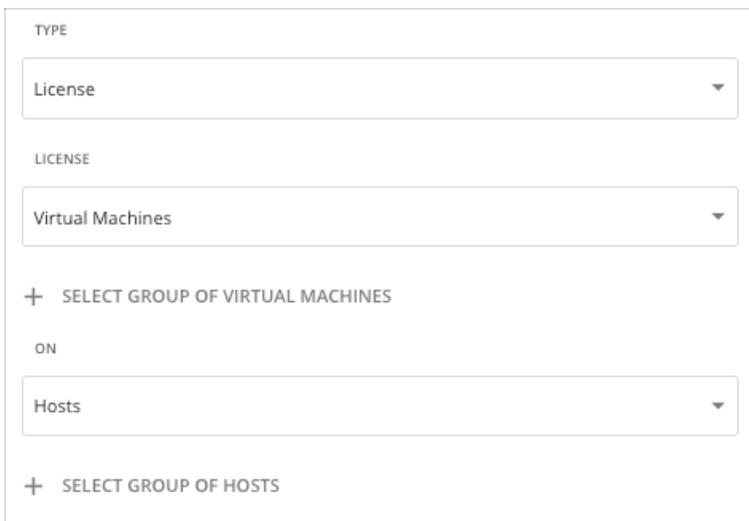
- For most policies that merge host and storage clusters, the clusters you place in the Merge segment must be members of the same data center.
- For vCenter environments, you can create placement policies that merge data centers to support cross-vCenter moves. In this case, where a data center corresponds to a vCenter target, the merged clusters can be in different data centers. In this case, you must create two merge policies; one to merge the affected data centers, and another to merge the specific clusters.
If the clusters you merge do not use the same network names on their respective data center, configure a network merge policy to define the compatible networks.
- For network merges, if you select to merge "Network A" and "Network B," all networks across data centers with name "Network A" and "Network B" are merged.

To create a merge policy, choose the type of entity to merge, and then select the groups to merge.

WARNING:

If you have reservations defined, network merge policies do not currently respect these reservations.

7. Create a **License** policy.



Assume that you have purchased several licenses for a database – you pay for the right to run that database on some hosts. You can create a license policy to identify the hosts that provide the license, and the VMs that can consume that license.

After you create the policy, Intersight Workload Optimizer can recommend the following actions in response to changing demand for licenses:

- When demand is low, Intersight Workload Optimizer recommends consolidating VMs on as few license-providing hosts as possible to reduce your license costs. To consolidate, you move VMs to another host and then reconfigure the original hosts to remove their licenses. Note that Intersight Workload Optimizer will *not* recommend suspending these hosts. Since they remain active, they can be reconfigured to become providers when demand starts to exceed capacity.

For example, if you have Host_01 providing a license to VM_01 and Host_02 providing a license to VM_02, you see two recommendations – move VM_02 to Host_01 and then remove the license in Host_02. You will not see a recommendation to suspend Host_02.

- When demand exceeds capacity, and there are hosts in the policy that currently do not provide licenses, Intersight Workload Optimizer recommends reconfiguring those hosts to become providers and then moving VMs to those hosts. If all hosts are providing licenses, Intersight Workload Optimizer recommends adding licenses to the hosts to meet demand.

These actions are more efficient than provisioning new hosts.

To create a license policy:

- Specify the license consumers (VMs).
- Specify the license providers (hosts).

In addition to creating a license policy, you must also create host automation policies to allow Intersight Workload Optimizer to recommend reconfigure actions on hosts. In the automation policies, add the license-providing hosts and then enable the *Reconfigure* action.

8. Save the policy.

Automation Policies

An automation policy is a set of rules that Intersight Workload Optimizer must satisfy when executing [non-parking actions \(on page 405\)](#) on public cloud or on-prem workloads, or changing settings that affect analysis and action generation.

Automation policies include the following settings:

- **Action Generation**
This setting specifies whether a specific action will be generated.
- **Action Acceptance**
If a specific action will be generated, this setting specifies the degree of automation for the action. For more information, see [Action Acceptance Modes \(on page 415\)](#).
- **Constraints and Other Settings**
These settings affect the Intersight Workload Optimizer analysis of the state of your environment. These include operational, utilization, and scaling constraints.
The settings you can make are different according to the type of entity this policy will affect. Each setting you add to the policy takes precedence over the default value for that setting.

Default and User-defined Automation Policies

Intersight Workload Optimizer ships with default automation policies that are believed to give you the best results based on our analysis. For certain entities in your environment, you can create automation policies as a way to override the defaults.

For example, **Enforce Non Disruptive Mode** is turned off in the default automation policy for on-prem VMs. In most cases, you might want to turn on the setting, and only turn it off for select VMs. In that case, you would turn it on in the default automation policy for VMs, and then create policies for those groups of VMs for which you want to turn it off.

The default and user-defined automation policies take effect in relation to each other. A default policy has a global effect, while a user-defined policy overrides the default policy for the entities within the indicated scope. You should keep the following points in mind:

- User-defined policies override a subset of settings.
A user-defined policy can override a subset of settings for the entity type. For the remainder, Intersight Workload Optimizer will use the default policy settings on the indicated scope.
- When an entity applies conflicting user-defined policies, Intersight Workload Optimizer applies the following tie breakers:
 - A scheduled policy always takes precedence over a non-scheduled policy, even if the non-scheduled policy is more conservative.
 - Among scheduled policies with *identical* schedules, the most conservative setting wins.
 - Among non-scheduled policies, the most conservative setting wins.

For example, a VM currently belongs to four groups with different policy settings.

- Group A policy: Resize VM in *Manual* mode every Saturday.
- Group B policy: Resize VM in *Automatic* mode every Saturday.
- Group C policy: Resize VM in *Manual* mode (no schedule).
- Group D policy: Resize VM in *Recommend* mode (no schedule).

Results:

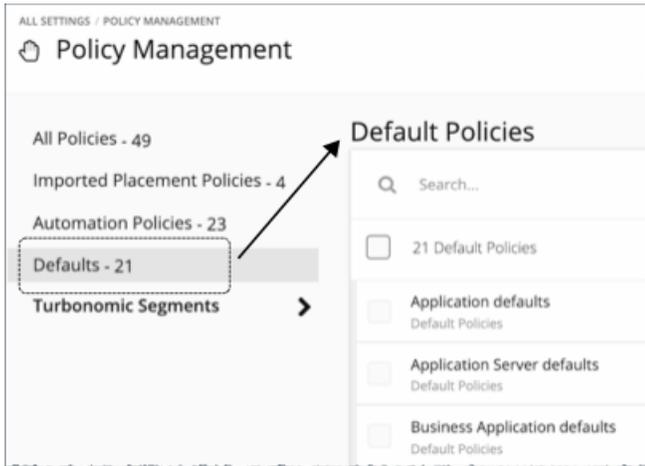
- On a Saturday, Groups A and B policies take precedence over Groups C and D policies. The VM ultimately applies the Group A setting because it is more conservative.
- On all the other days, only Groups C and D policies are active. The VM applies the Group D setting because it is more conservative.
- User-defined policies always take precedence over default policies.

Even if the default policy has a more conservative setting, the setting in the user-defined policy wins for entities in that scope.

- For a global effect, *always* use default policies.

Default Automation Policies

To view the default automation policies, navigate to **Settings > Policies**, and then click **Defaults**.



The page displays a list of all the default policies, by entity type. You can click the entity type to view or change the default settings.

Global Default Policy

Use these settings to modify Intersight Workload Optimizer analysis globally for any scope of your environment. These defaults affect both default and user-defined automation policies.

AUTOMATION WORKFLOW

Disable All Actions

| Attribute | Default Setting |
|---------------------|-----------------|
| Disable All Actions | OFF |

When this is ON, Intersight Workload Optimizer does not generate any actions for your environment. For example, assume you have configured a number of policies that automate actions, but you want to stop making changes to the entire environment for a period of time. Turn this ON to stop all execution with a single setting.

OPERATIONAL CONSTRAINTS

VM Growth Observation Period

| Attribute | Default Value |
|------------------------------|---------------|
| VM Growth Observation Period | 1 month |

Use this setting to specify how much historical data the Intersight Workload Optimizer analysis will use to calculate time to exhaustion of your cluster resources.

Intersight Workload Optimizer runs nightly plans to calculate headroom for the clusters in your on-prem environment. To review your cluster headroom in dashboards, set the view scope to a cluster. With that scope, the view includes charts to show headroom for that cluster, as well as time to exhaustion of the cluster resources.

To calculate cluster growth trends, analysis uses historical data for the given clusters. With **VM Growth Observation Period**, you can specify how much historical data the headroom analysis will use to calculate time to exhaustion of your cluster resources. For example, if cluster usage is growing slowly, then you can set the observation to a period that is long enough to capture that rate of growth.

If the historical database does not include at least two entries in the monthly data for the cluster, then analysis uses daily historical data.

Allow Unlimited Host Provisioning

| Attribute | Default Setting |
|-----------------------------------|-----------------|
| Allow Unlimited Host Provisioning | OFF |

By default, Intersight Workload Optimizer allows overprovisioning hosts up to 10 times their memory capacity, and up to 30 times their CPU capacity. When this setting is ON, Intersight Workload Optimizer removes these overprovisioning limits to allow VM placements on already overprovisioned hosts.

This setting does not stop Intersight Workload Optimizer from recommending actions to provision new hosts in clusters.

Enable Analysis of On-prem Volumes

| Attribute | Default Setting |
|------------------------------------|-----------------|
| Enable Analysis of On-prem Volumes | OFF |

[On-prem volumes \(on page 363\)](#) represent VM Disks discovered by hypervisor targets. A VM will have one volume for each configured disk and another volume (representing the configuration) that always moves with Disk 1.

■ OFF (default)

Intersight Workload Optimizer analyzes volume resources as part of VM analysis. In the real-time market and on-prem plans, any action to move VM storage ensures that volumes stay together on the underlying datastore. A [Migrate to Cloud plan \(on page 460\)](#) will recommend storage per datastore to hold all the VM Disks currently on the datastore.

For example, assume a VM with three disks. Disks 1 and 3 are on Datastore A, while Disk 2 is on Datastore B.

- During a storage migration, VM Disk volumes 1 and 3 will stay on the same datastore.
- A Migrate to Cloud plan will recommend a storage disk for VM Disk volumes 1 and 3, and another storage disk for VM Disk volume 2.

■ ON

Intersight Workload Optimizer analyzes resources on each volume independently. In the real-time market and on-prem plans, any action to move VM storage migrates volumes to the most optimal datastore. A Migrate to Cloud plan will recommend storage for each volume.

For example, assume a VM with three disks. Disks 1 and 3 are on Datastore A, while Disk 2 is on Datastore B.

- During a storage migration, VM Disk volumes 1, 2, and 3 can migrate to different datastores.
- A Migrate to Cloud plan will recommend three separate storage disks for VM Disk volumes 1, 2, and 3.

IMPORTANT:

When you turn on this setting, your Intersight Workload Optimizer instance will start to use more memory and storage to perform its analysis. For example an environment with 10,000 VMs and an average of three disks per VM represents a three-fold increase in entities that require analysis. Currently, instances that monitor more than 50,000 VMs will experience a significant drop in performance. For this reason, this setting is turned off by default.

Before turning on this setting, review your [VM automation policies \(on page 341\)](#) and verify that Storage Move actions are in *Recommend* or *Manual* mode. In addition, review your [storage placement policies \(on page 569\)](#) to ensure that individual VM volumes can be placed on the expected storage.

Imported Automation Policies

As Intersight Workload Optimizer discovers your environment, it can find configurations that set up scopes that need specific policies. For example:

- HA Configurations

For vCenter Server environments, Intersight Workload Optimizer discovers HA cluster settings and translates them into CPU and memory utilization constraints. The discovery creates a group of type *folder* for each HA cluster, and creates a policy that sets the appropriate CPU and memory constraints to that policy.

- Availability Sets

In public cloud environments, Intersight Workload Optimizer discovers groups of VMs that should keep all their VMs on the same template. In the Automation Policies list, these appear with the prefix `AvailabilitySet::` on the policy names. You can enable Consistent Resizing for the VMs in each group so Intersight Workload Optimizer can resize them to the same size.

To view imported automation policies, navigate to **Settings > Policies**, and then click **Imported Automation Policies**.

Creating Automation Policies

To override the default automation policies, you can create your own policies. These policies specify the settings you want to change for certain entities in your environment. You can assign a schedule to your policy to set up maintenance windows or other scheduled actions in your environment.

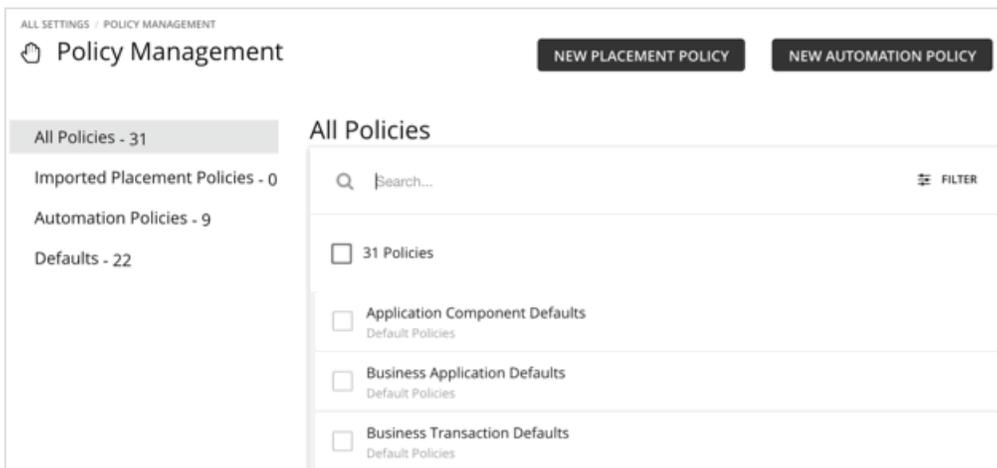
Below are some use cases for creating automation policies.

- Intersight Workload Optimizer uses a number of settings to guide its analysis of the entities in your environment. The default settings might be fine in most cases, but you might want different analysis for some groups of entities.
- Assume you want to automate scaling and placement actions for the VMs in your environment. It is common to take a cautious approach, and start by automating clusters that are not critical or in production. You can scope the policy to those clusters, and set the action acceptance mode to Automatic for different actions on those VMs (see [Action Acceptance Modes \(on page 415\)](#)).

1. Entry Point

Click **More** and display the Settings Page. Then choose **Policies**.

This opens the Policy Management Page, which lists all the currently available policies.

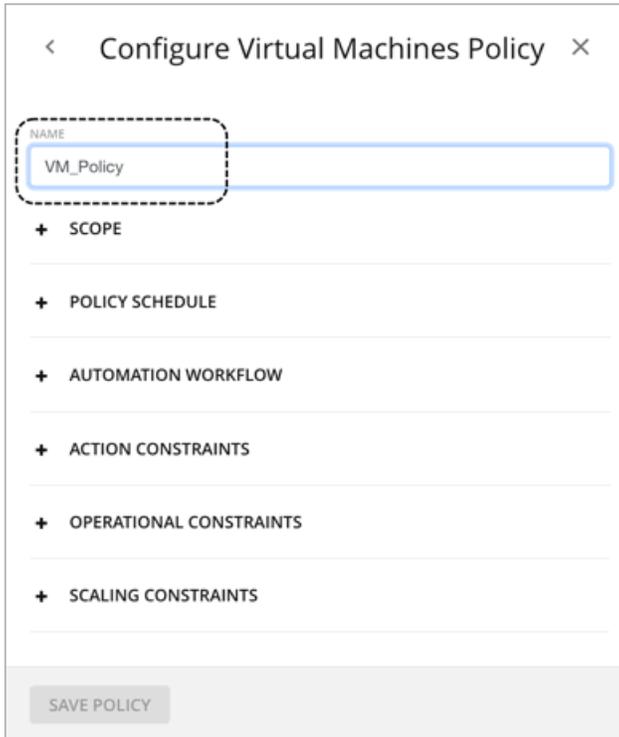


Click **NEW POLICY > Automation Policy** and then select the entity type (such as Virtual Machine).

This sets the type of entity that your policy will affect. Note that Intersight Workload Optimizer supports different actions for different types of entities. For example, you cannot add VMem to a storage device. Setting policy type is the first step you take to focus on which actions you want to map to your workflows.

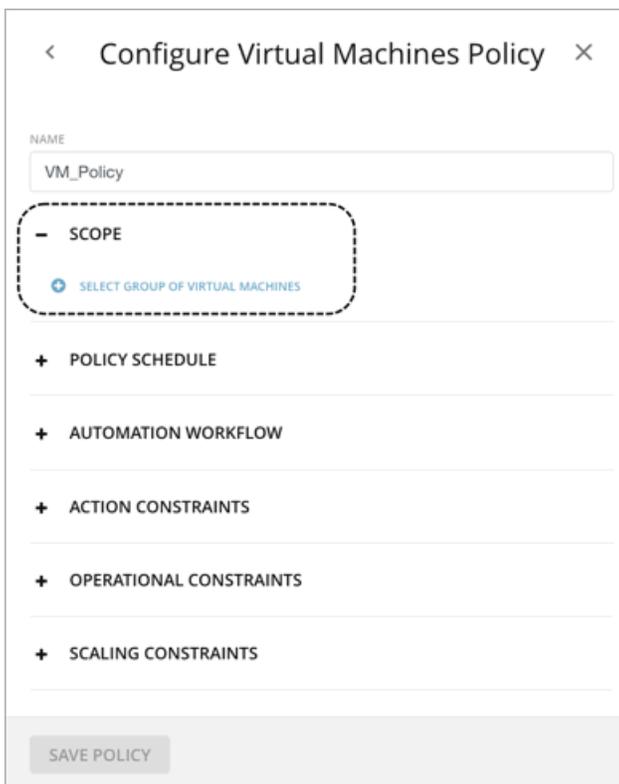
2. Policy Name

Name the policy.



The screenshot shows the 'Configure Virtual Machines Policy' interface. At the top, there is a title bar with a back arrow, the text 'Configure Virtual Machines Policy', and a close 'X' icon. Below the title bar is a 'NAME' field containing the text 'VM_Policy'. This field is highlighted with a dashed blue border. Below the name field are several expandable sections, each with a plus sign and a label: '+ SCOPE', '+ POLICY SCHEDULE', '+ AUTOMATION WORKFLOW', '+ ACTION CONSTRAINTS', '+ OPERATIONAL CONSTRAINTS', and '+ SCALING CONSTRAINTS'. At the bottom of the form is a 'SAVE POLICY' button.

3. Scope



The screenshot shows the 'Configure Virtual Machines Policy' interface. At the top, there is a title bar with a back arrow, the text 'Configure Virtual Machines Policy', and a close 'X' icon. Below the title bar is a 'NAME' field containing the text 'VM_Policy'. Below the name field is the 'SCOPE' section, which is expanded and highlighted with a dashed blue border. Inside the 'SCOPE' section, there is a plus sign and the text 'SELECT GROUP OF VIRTUAL MACHINES'. Below the 'SCOPE' section are several other expandable sections, each with a plus sign and a label: '+ POLICY SCHEDULE', '+ AUTOMATION WORKFLOW', '+ ACTION CONSTRAINTS', '+ OPERATIONAL CONSTRAINTS', and '+ SCALING CONSTRAINTS'. At the bottom of the form is a 'SAVE POLICY' button.

The scope determines which entities this policy will affect. Choose one or more groups, or create new groups and add them to the policy scope. These groups match the type of entity you have set for the policy.

In Intersight Workload Optimizer you can find nested groups (groups of groups). For example, the "By PM Cluster" group contains host clusters, and each host cluster is a group. Do not set the policy scope to a parent of nested groups. When setting up policies, be sure you set them to individual groups. If necessary, create a custom group for the settings you want to apply.

NOTE:

A single entity can be a member of multiple groups. This can result in a conflict of settings, where the same entity can have different policy settings. For conflicts among user-defined policy settings, the most conservative setting will take effect. For details, see [Default and User-defined Policies \(on page 574\)](#).

4. Policy Schedule

The screenshot shows the 'Configure Virtual Machines Policy' configuration page. The 'POLICY SCHEDULE' section is expanded, showing a description: 'Enables the overall policy at the scheduled time. To set when actions can execute, apply an Execution Schedule in the Automation and Orchestration section.' Below this is an 'ATTACH SCHEDULE' link with an arrow pointing to a 'Select Schedule' fly-out menu. The fly-out menu lists 'All Schedules (21)' and includes a search bar, a filter icon, and a 'NEW SCHEDULE' button. Two schedules are listed: 'Ap' (Expired) and 'Ap 2' (Daily, Starts in 8 hours). The 'Ap 2' schedule is selected, and its details are shown, including a summary, 'USED IN POLICIES', 'NEXT OCCURRENCE' (Saturday, December 12, 2020 at 3:30 AM UTC), and 'ACCEPTED ACTIONS' (Awaiting Acceptance). A 'SET' button is at the bottom of the fly-out.

For use cases and information about how schedules affect policies, see [Policy Schedules \(on page 583\)](#).

The **Select Schedule** fly-out lists all the schedules that are currently defined for your instance of Intersight Workload Optimizer. Expand a schedule entry to see its details. The details include a summary of the schedule definition, as well as:

- **Used in Policies**

The number of policies that use this schedule. Click the number to review the policies.

- **Next Occurrence**

When the schedule will next come into effect.

- **Accepted Actions**

How many scheduled actions have been accepted to be executed in the next schedule occurrence. Click the number for a list of these actions.

- **Awaiting Acceptance**

The number of Manual actions affected by this schedule that are in the Pending Actions list, and have not been accepted. Click the number for a list of these actions.

If none of the listed schedules is suitable for your policy (or if none exists), click **New Schedule**. For details, see [Managing Calendar Schedules \(on page 584\)](#).

NOTE:

When you configure a schedule window for a VM resize action, to ensure Intersight Workload Optimizer will execute the action during the scheduled time, you must turn off the **Enforce Non Disruptive Mode** setting for that scheduled policy. Even if you turn the setting off for the global policy, you still must turn the setting off for your scheduled policy. Otherwise Intersight Workload Optimizer will not execute the resize action.

5. Automation Workflow

NOTE:

If your installation supports both automation *and* orchestration, you can use policy settings to integrate actions with orchestration workflows. If your installation does *not* support orchestration, the following settings will have no effect:

- Before Execution
- Action Execution
- After Execution

You can define automation workflow settings for different action types within the same policy. For example, for a group of VMs in a policy, you can automate all *Resize* actions, but require *Suspend* actions to go through an approval process via an Orchestrator (such as ServiceNow).

5.1. Action Type

See a list of actions that are viable for the policy, and then make your selections.

5.2. Action Generation and Acceptance

- Do not Generate Actions

Intersight Workload Optimizer never considers your selected actions in its calculations. For example, if you do not want to generate *Resize* actions for VMs in the policy, analysis will still drive toward the desired state, but will do so without considering resizes.

- Generate Actions

Intersight Workload Optimizer generates your selected actions to address or prevent problems. Choose from the following *Action Acceptance* modes to indicate how you would like the actions to execute:

- Recommend Only – Recommend the action so that a user can execute it outside Intersight Workload Optimizer
- Manual – Recommend the action, and provide the option to execute that action through the Intersight Workload Optimizer user interface
- Automated – Execute the action automatically.

For automated resize or move actions on the same entity, Intersight Workload Optimizer waits five minutes between each action to avoid failures associated with trying to execute all actions at once. Any action awaiting execution stays in queue. For example, if a VM has both vCPU and vMem resize actions, Intersight Workload Optimizer could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.

- Automated when approved - Execute the action automatically only after they are externally approved. This option is available only after adding a Service Now target.

5.3. Execution Schedule

You can defer the execution of generated actions to a non-critical time window. For example, if a workload experiences memory bottlenecks during the week, you can defer the necessary resize to the weekend. Even if the workload has minimal utilization over the weekend, Intersight Workload Optimizer can recognize the need to resize, and will execute the action.

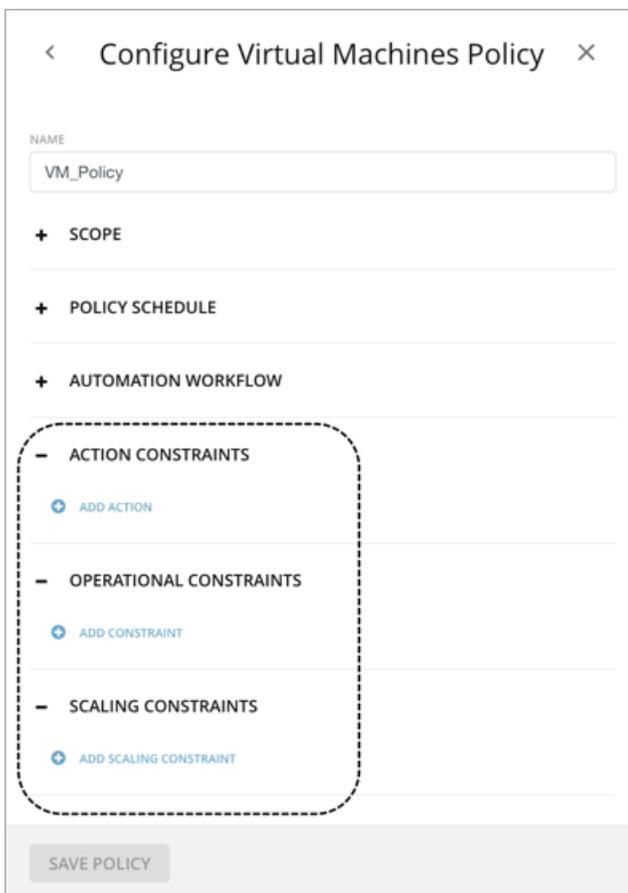
For more information, see [Action Execution Schedules \(on page 583\)](#).

6. Constraints and Other Settings

Intersight Workload Optimizer collects metrics to drive the analysis that it uses when it calculates actions for your environment. It compares current utilization and demand against allocated capacities for resources, so it can recommend actions that keep your environment in optimal running condition.

Automation policies include constraints and other settings that you can make to adjust the analysis that Intersight Workload Optimizer performs. For example, you can set different levels of overprovisioning for host or VM resources, and Intersight Workload Optimizer will consider that as a factor when deciding on actions.

The settings you can make are different according to the type of entity this policy will affect. Each setting you add to the policy takes precedence over the default value for that setting.



Configure Virtual Machines Policy

NAME
VM_Policy

+ SCOPE

+ POLICY SCHEDULE

+ AUTOMATION WORKFLOW

- ACTION CONSTRAINTS

+ ADD ACTION

- OPERATIONAL CONSTRAINTS

+ ADD CONSTRAINT

- SCALING CONSTRAINTS

+ ADD SCALING CONSTRAINT

SAVE POLICY

Policy Schedules

You can set a schedule for an automation policy, which sets a window of time when the policy takes effect. For example, you can modify the operational or scaling Constraints for a given period of time. These settings affect Intersight Workload Optimizer analysis, and the actions that the platform generates. You can set up scheduled times when you want to change those settings.

Remember that for user-defined automation policies, it is possible that one entity can be in two different scopes. This means that the entity can be under the effect of two different policies. For this reason, user-defined policies keep the rule, *the most conservative setting wins*. However, a more aggressive user-defined policy takes precedence over the corresponding default automation policy.

You must consider these rules when you add schedules to policies. If the more conservative settings are in a default automation policy, then the scheduled change takes effect. However, if the more conservative settings are in another user-defined policy, then the conservative settings *win*, and the scheduled changes do not take effect.

For details, see [Default and User-defined Policies \(on page 574\)](#).

Policy Schedule and Action Execution Schedule

A scheduled policy can include *actions*. When the policy is in effect, Intersight Workload Optimizer recommends or automatically executes those actions as they are generated. Some of those actions could be disruptive so you may want to defer their execution to a non-critical time window. In this case, you will need to set an *action execution schedule* within the scheduled policy. For example, you can set a policy that automatically resizes or starts VMs for your customer-facing apps for the entire month of December, in anticipation of an increase in demand. Within this same policy, you can set the resize execution schedule to Monday, from midnight to 7:00 AM, when demand is expected to be minimal.

For more information, see [Action Execution Schedules \(on page 583\)](#).

Action Execution Schedules

You can defer the execution of generated actions to a non-critical time window. For example, if mission-critical VMs experience memory bottlenecks during the week, you can defer the necessary memory resizes to the weekend. Even if the VMs have minimal utilization over the weekend, Intersight Workload Optimizer can recognize the need to resize, and will execute resize actions. For this particular example, you will need to:

1. Create a policy for the VMs.
2. Select *VMem Resize Up* from the list of actions and then set the action mode to either *Automatic* or *Manual*.

NOTE:

Execution schedules have no effect on recommended actions. It is therefore not necessary to set up an execution schedule if all the actions in your policy will be in *Recommend* mode.

3. Set an Execution Schedule that starts on Saturday at 8:00 AM and lasts 48 hours.

Execution of Scheduled Actions

Intersight Workload Optimizer posts an action at the time that the conditions warrant it, which means that you might see the action in the Pending Actions list even before the execution schedule takes effect. The action details show what schedule affects the given action, and shows the next occurrence of that schedule.

■ Automatic

When the schedule takes effect, Intersight Workload Optimizer executes any pending automated actions.

■ Manual

Before the execution schedule, the action details for manually executable actions show the action state as `PENDING ACCEPT`. If you accept the action (select it and click **Apply Selected**), then Intersight Workload Optimizer adds it to the queue of actions to be executed during the maintenance window. The action details show the action state as `AWAITING EXECUTION`. Intersight Workload Optimizer executes the actions when the schedule takes effect.

Keeping Actions Valid Until the Scheduled Time

If you have scheduled action execution for a later time, then conditions could change enough that the action is no longer valid. If this happens, and the action remains invalid for 24 hours, then Intersight Workload Optimizer removes it from the list of pending actions. This action will not be executed.

Intersight Workload Optimizer includes scaling constraints that work to stabilize action decisions for VMs. The resulting actions are more likely to remain valid up until their scheduled window for execution. You can make these settings in the default or user-defined automation policies.

NOTE:

When you configure an execution schedule for a resize action, to ensure Intersight Workload Optimizer will execute the action during the scheduled time, you must turn off the **Enforce Non Disruptive Mode** setting for the policy. Even if you turn the setting off for the global policy, you still must turn the setting off for your policy. Otherwise Intersight Workload Optimizer will not execute the resize action. For information about non disruptive mode, see [Non-disruptive Mode \(on page 342\)](#).

Working With Schedules

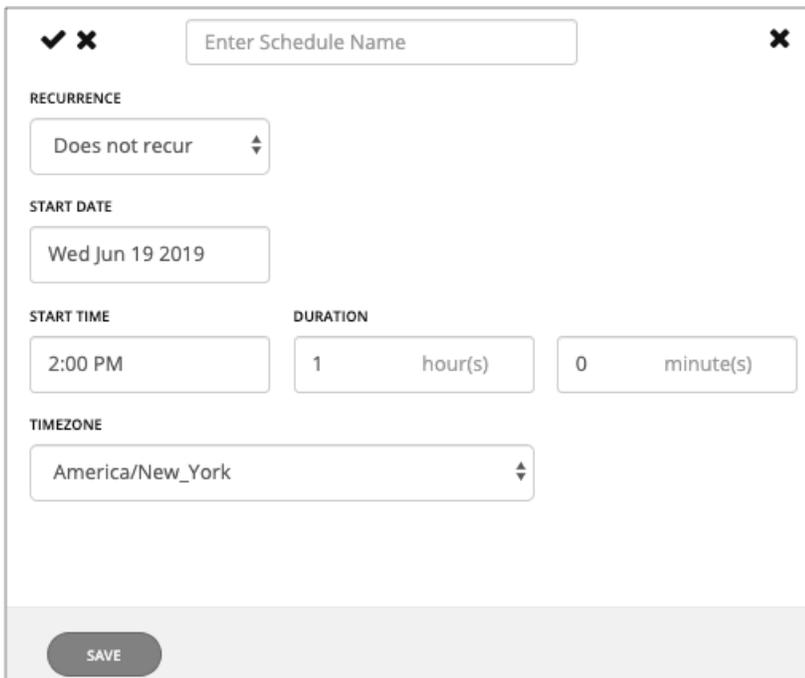
Schedules specify a period of time during which certain events can occur.

Managing Calendar Schedules

A calendar schedule is a setting that sets up a window of time when an [automation policy \(on page 577\)](#) takes effect. This policy can execute [non-parking actions \(on page 405\)](#) on public cloud or on-prem entities, or change settings that affect analysis and action generation.

Creating Calendar Schedules

1. Click **More**, then display the Settings Page.
2. Click **New Schedule > Calendar Schedule**.
3. Type a name for the schedule.



4. Set the recurrence for the schedule.

Choose whether the scheduled period occurs just once, or whether it repeats over time. The settings vary according to the recurrence you choose:

- Does Not Recur

This is a one-time schedule window. A non-recurring window has a start date, and no end date. The window starts on the day and time you specify, and remains open for the given duration.

- Daily

RECURRENCE

Daily

REPEAT EVERY

1 days

START DATE **END DATE**

Wed Jun 19 2019 None

Repeat this schedule every given number of days. For example, repeating 30 days is similar to repeating monthly, except it repeats by the count of days, not by the calendar month.

The schedule begins on the **Start Date**, and continues repeating until the **End Date**. If **End Date** is "None", the schedule repeats perpetually.

- Weekly

RECURRENCE

Weekly

REPEAT EVERY **ON**

1 weeks Mo Tu Wd Th Fr Sa Su

START DATE **END DATE**

Wed Jun 19 2019 None

Repeat this schedule every given number of weeks, on the week days you specify. For example, to repeat every weekend, set it to repeat every one week on Saturday and Sunday.

The schedule begins on the **Start Date**, and continues repeating until the **End Date**. If **End Date** is "None", the schedule repeats perpetually.

- Monthly

RECURRENCE

Monthly

REPEAT EVERY **ON**

1 months First Saturday

START DATE **END DATE**

Wed Jun 19 2019 None

Repeat this schedule every given number of months, to begin on a given day in the month. For example, you can schedule a maintenance window to begin on the first Saturday of each month.

The schedule begins on the **Start Date**, and continues repeating until the **End Date**. If **End Date** is "None", the schedule repeats perpetually.

5. Set the start time and duration.

These settings specify how long the scheduled window remains open. You set the duration in terms of hours and minutes. Using a duration instead of an end time removes ambiguities such as starting before midnight and ending after. However, you should make sure the duration is not longer than the recurrence.

6. Set the time zone.

This gives a reference for the schedule's start time. Intersight Workload Optimizer uses that reference when it opens and closes the schedule window.

You see the same time zone setting no matter where you are located. Convert the schedule time to your local time to track when the schedule opens in your working day.

7. Save the schedule.

Viewing Calendar Schedules

To view calendar schedules, navigate to **Schedule**.

The Schedules page lists all the currently defined schedules. From this page you can:

1. View list of schedules.
2. Click name to edit schedule.
3. Select an entry to delete the schedule.
4. Expand an entry to see schedule details.

The details include a summary of the schedule definition, as well as:

- **Used in Policies**

The number of policies that use this schedule. Click the number to review the policies.

- **Next Occurrence**

When the schedule will next come into effect.

- **Accepted Actions**

How many scheduled actions have been accepted to be executed in the next schedule occurrence. Click the number for a list of these actions.

- **Awaiting Acceptance**

The number of Manual actions affected by this schedule that are in the Pending Actions list, and have not been accepted. Click the number for a list of these actions.

5. Create a schedule.
6. Select an entry to defer the next occurrence.

Intersight Workload Optimizer calculates when the next scheduled window will open. If you want cancel the scheduled occurrence one time, you can select the schedule and defer the upcoming occurrence. This defers the schedule wherever it is applied. If the schedule is applied to more than one policy, this will defer all the policies that use this schedule. Before you defer a schedule, you should expand the details and review all the policies that use this schedule.

Deleting Calendar Schedules

Before you delete a schedule, you should view its details to make sure no policies use it. If you delete a schedule that is in use by any policies, Intersight Workload Optimizer disables the affected policies until you edit them to either:

- Apply a different schedule to the policy and save the change
- Save the policy with no schedule

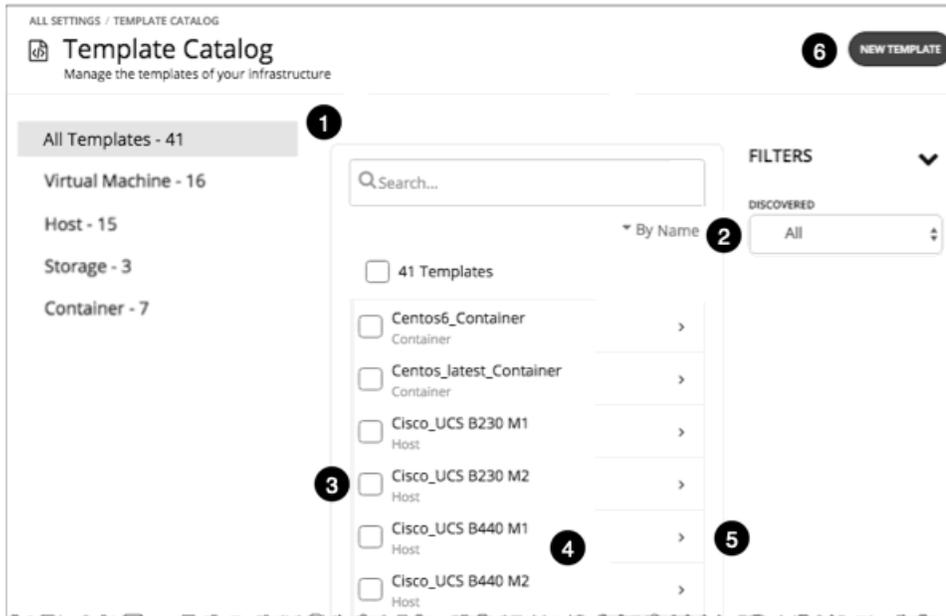
Saving with no schedule confirms that you intend for this policy to apply at all times. Because scheduled policies are for special cases, this is usually not what you intend. For example, a scheduled maintenance window can have aggressive action acceptance modes that you do not want to enable during peak hours. If you save the policy with no schedule, then the aggressive settings will take effect at all times.

Intersight Workload Optimizer posts a confirmation dialog before deleting a schedule that is currently in use.

Templates: Resource Allocations for New Entities

NOTE:

This page includes enhancements and a more modern look-and-feel that are only available when you enable the new design framework. To switch to the new framework, click the React icon  in the navigation bar of the user interface and then Turn ON the toggle. For more information, see "Design Framework for the User Interface" in the *User Guide*.



Intersight Workload Optimizer uses templates to describe new entities that it deploys in your environment or in plans. The templates specify resource allocations for these entities. For example, you can run a plan that adds new VMs to a cluster. If you add ten copies of a template, then the plan places ten new VMs that match the resource allocation you have specified for the template. For your cloud environment, you can see templates to match the instance types in your cloud accounts and subscriptions.

A VM template definition can include one or more images that Intersight Workload Optimizer uses to deploy the VM in your environment. The image identifies the actual deployment package, including a path to the physical files (for example an OVA).

As you deploy an instance of a VM template, Intersight Workload Optimizer chooses the best image for that instance.

The Template Catalog shows the templates that are specified or discovered for your installation of Intersight Workload Optimizer. From this page, you can:

1. Filter by type.
2. Filter to see only discovered templates.
3. Select templates to delete.
4. Click to edit the template.
5. Expand for details.
6. Create a template.

Creating Templates

NOTE:

This page includes enhancements and a more modern look-and-feel that are only available when you enable the new design framework. To switch to the new framework, click the React icon  in the navigation bar of the user interface and then Turn ON the toggle. For more information, see "Design Framework for the User Interface" in the *User Guide*.

Templates specify the resources for entities that Intersight Workload Optimizer can deploy in your environment, or in plans.

A VM template definition can include one or more images that Intersight Workload Optimizer uses to deploy the VM in your environment. The image identifies the actual deployment package, including a path to the physical files (for example an OVA).

The Template Catalog shows all of the templates that have been specified or discovered for your installation of Intersight Workload Optimizer. From this page, you can also create new templates and edit existing ones.

Creating and Editing Templates

To create a new template, navigate to the Template Catalog and click **NEW TEMPLATE**. To edit a template, click the template's name. When you create a new template, the first step is to choose the entity type.

1. Navigate to the Settings Page.

Click **More** and then display the Settings Page.

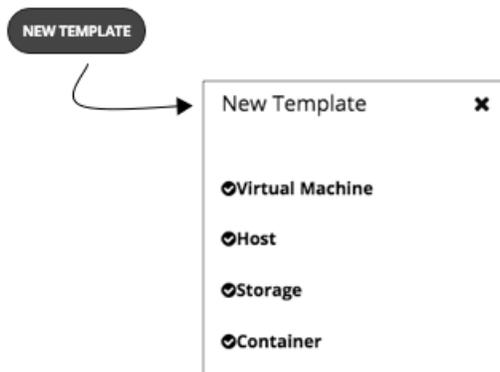
2. Choose Templates.



3. Create or edit a template

To create a new template, navigate to the Template Catalog and click **NEW TEMPLATE**. To edit a template, click the template's name.

4. If you're creating a new template, choose the entity type.



5. Make the settings for your template.

For each type of template, you set allocations for different resources. You can make templates of the following types:

- Virtual Machine
- Host
- Storage
- Container

6. Make the settings for your template, and then save your changes.

When the template window opens, it displays the most common resource settings. You can expand the settings to see the full collection for that template type.

7. Save your changes.

After you have made your settings and named the template, click **CREATE** or **SAVE**.

VM Template Settings

A VM template describes the resource allocation that you want to provide for a type of VMs. When Intersight Workload Optimizer deploys the associated VM to your environment or in a plan, it uses these values to determine the size of the VM. Intersight Workload Optimizer uses the Size settings to calculate the best placement for a VM of this type.

A VM template can optionally include an image description. When Intersight Workload Optimizer uses the template to deploy a VM to your environment, it uses the image to access the actual bits that install as the VM instance.

NOTE:

Intersight Workload Optimizer generates a special template called *headroomVM*, which it uses to calculate cluster headroom. The Template Catalog shows the template as editable, but you should not edit it because Intersight Workload Optimizer will overwrite your changes the next time it generates the template.

VM Size

- CPU

The virtual CPUs assigned to the VM. Specify the number of **Cores** and the **VCPU** clock speed – Intersight Workload Optimizer multiplies these values to calculate the host CPU resources it will allocate when placing the VM.

The **Utilization** value sets the percentage of allocated CPU that the placed VM will consume. To ensure the host has left over resources for infrastructure tasks, you should assign less than 100%.

- Memory

The amount of memory to allocate for the VM, in MB.

The **Utilization** value sets the percentage of allocated memory that the placed VM will consume. To ensure the host has left over resources for infrastructure tasks, you should assign less than 100%.

Note that you should never allocate less memory than is required for the VM's guest OS.

- Storage

The storage resources to allocate for this VM.

- **disk/rdm** – If you choose **rdm**, then the VM can use VMware Raw Device Mapping for its storage.
- **IOPS** – The capacity for IO operations you give the VM for this datastore.
- **Size** – The amount of storage capacity, in GB.

The **Utilization** value sets the percentage of allocated memory that the placed VM will consume. To ensure the storage has left over resources for infrastructure tasks, you should assign less than 100%.

Note that you can allocate multiple datastores to the VM.

- Network

The amount of the host's network throughput to assign to the VM, in Mb/s.

- IO

The amount of throughput on the host's IO bus to assign to the VM, in Mb/s

Host Template Settings

Host templates describe models of physical hosts that you can deploy in the on-prem data center. As part of capacity planning, you might want to see how to replace your current hosts with different models. To do that, you create templates to represent the hosts you want, and then use those templates when you run hardware replacement plans.

The host template is a collection of these settings:

- CPU

The processor for this host model. CPU size and speed are not the only factors to determine processing power. Specify the host CPU in the following ways:

- Select from Catalog



When you enable **Select from Catalog**, you can open up a catalog of CPU models that Intersight Workload Optimizer uses to map the model to an effective capacity for the CPU.

- Cores and CPU Speed



When you disable **Select from Catalog**, you can specify the number of **Cores** and the **CPU** clock speed – Intersight Workload Optimizer multiplies these values to calculate the host CPU resources.

- **Memory**

The amount of memory to allocate for the VM, in GB.

- **Network**

The host's network throughput, in MB/s.

- **IO**

The host's IO bus throughput, in MB/s

- **Price**

If you know the price of the host model that you're specifying for the template, you can enter it here. When a plan runs, Intersight Workload Optimizer uses the price to calculate costs or savings when it adds or removes host machines in an on-prem data center.

Selecting CPUs from the Catalog

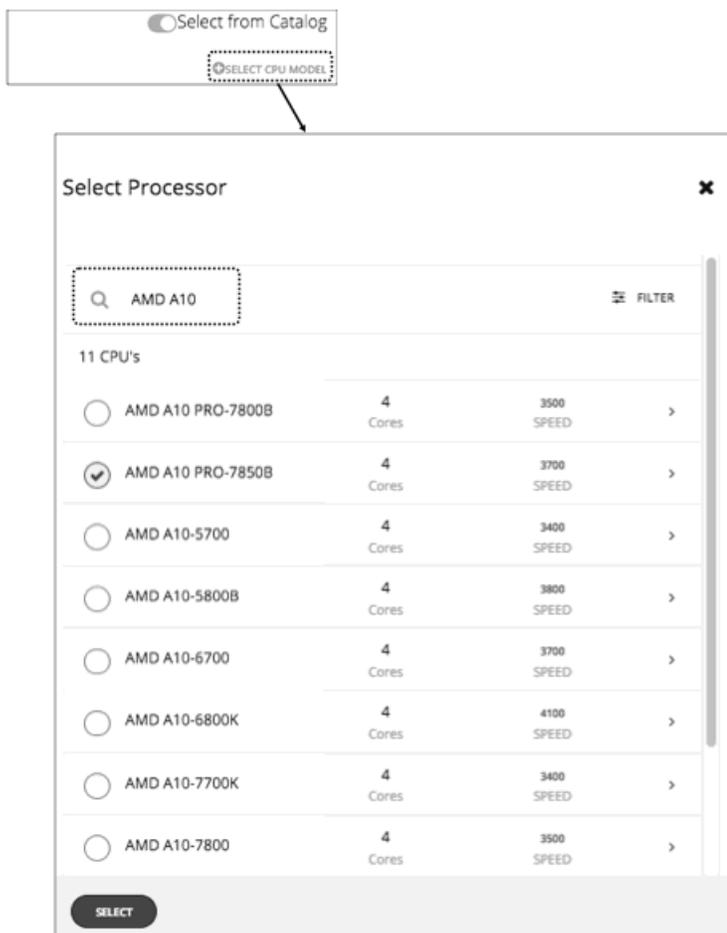
CPU processor speed is not necessarily an effective indicator of CPU capacity. For example, processor architecture can make a slower CPU have a greater effective capacity. Newer models of machines can often have fewer cores or less clock speed, but still have a higher effective capacity. Planning is affected in two ways:

- When planning hardware replacement, the plan knows the template's effective capacity. This means that the plan knows how to best place workloads on the new hardware.
- For deployed hosts, Intersight Workload Optimizer discovers the effective capacity and uses that information when it calculates workload placement.

To build the catalog of CPU capacity, Intersight Workload Optimizer uses benchmark data from spec.org. When you set up the CPU for a host template, you can search this catalog for the processor you want, and set it to the template.

NOTE:

Intersight Workload Optimizer also uses the effective processor capacity when it calculates workload placement in real-time. For more information, see [Effective CPU Capacity \(on page 413\)](#).



HCI Host Template Settings

HCI host templates describe models of physical hosts that support participation in a vSAN. Along with the host compute specifications, you also include specifications for storage capacity and redundancy (RAID level and failover). You can use these templates to plan for changes to your vSAN capacity.

NOTE:

For Hyper-V environments, if you run a Hardware Replace plan that replaces hosts with HCI Host templates, the results can be inconsistent or the plan can fail to place all the VMs in the plan scope. This typically occurs when Intersight Workload Optimizer detects a configuration issue with VMM or Hyper-V. As a result, Intersight Workload Optimizer treats the VMs as not controllable and does not attempt to place them.

The HCI Host template is a collection of these settings:

- CPU

The processor for this host model. CPU size and speed are not the only factors to determine processing power. Specify the host CPU in the following ways:

- Select from Catalog



When you enable **Select from Catalog**, you can open up a catalog of CPU models that Intersight Workload Optimizer uses to map the model to an effective capacity for the CPU.

– Cores and CPU Speed



When you disable **Select from Catalog**, you can specify the number of **Cores** and the **CPU** clock speed – Intersight Workload Optimizer multiplies these values to calculate the host CPU resources.

- **Memory**
The amount of memory to allocate for the VM, in GB.
- **Network**
The host's network throughput, in MB/s.
- **IO**
The host's IO bus throughput, in MB/s
- **Storage**
The capacity for this storage.
 - **IOPS** – The effective IOPS capacity.
 - **Size** – Raw storage capacity, in GB. A plan that uses this template computes the effective storage capacity.
- **Redundancy**
The redundancy method for this storage is on the virtualized SAN. This combines the RAID level and the number of host failures to tolerate.
- **Price**
If you know the price of the host model that you're specifying for the template, you can enter it here. When a plan runs, Intersight Workload Optimizer uses the price to calculate costs or savings when adding or removing host machines in an on-prem data center.

Selecting CPUs from the Catalog

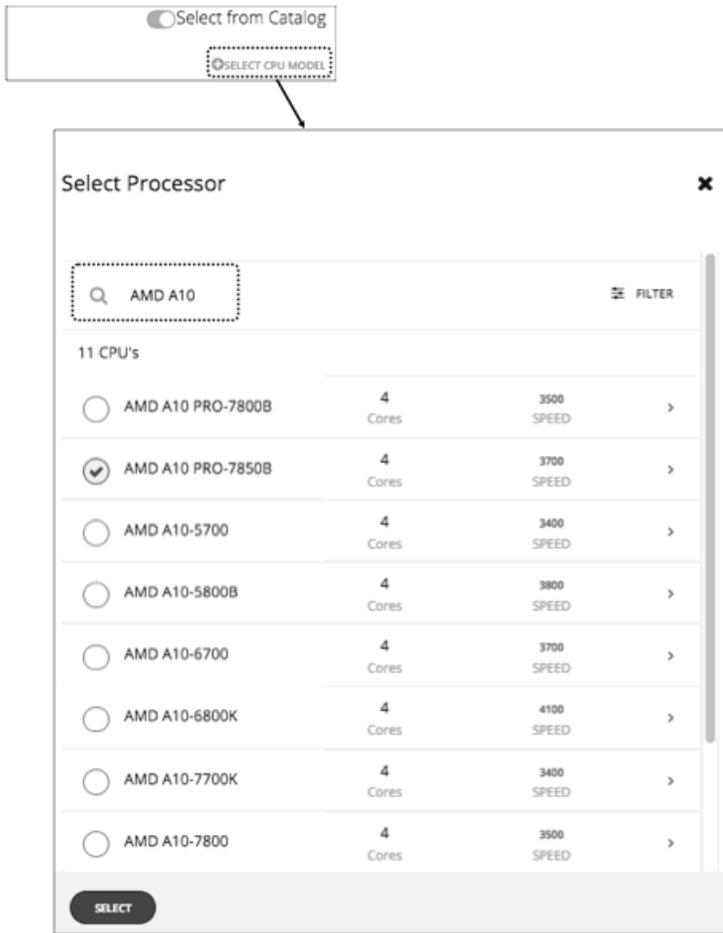
CPU processor speed is not necessarily an effective indicator of CPU capacity. For example, processor architecture can make a slower CPU have a greater effective capacity. Newer models of machines can often have fewer cores or less clock speed, but still have a higher effective capacity. Planning is affected in two ways:

- When planning hardware replacement, the plan knows the template's effective capacity. The plan knows how to best place workloads on the new hardware.
- For already deployed hosts, Intersight Workload Optimizer discovers the effective capacity and uses that information when it calculates workload placement.

To build the catalog of CPU capacity, Intersight Workload Optimizer uses benchmark data from spec.org. When you set up the CPU for a host template, you can search this catalog for the processor you want, and set it to the template.

NOTE:

Intersight Workload Optimizer also uses the effective processor capacity when it calculates workload placement in real-time. For more information, see [Effective CPU Capacity \(on page 413\)](#).



Storage Template Settings

Storage templates describe models of storage that you can deploy in the on-prem datacenter. As part of capacity planning, you might want to see how to replace your current storage with different models. To do that, you create templates to represent the storage you want, and then use those templates when running hardware replacement plans.

The storage template is a collection of these settings:

- Storage

The capacity for this storage.

- **IOPS** – The capacity for IO operations on this storage.
- **Size** – The amount of storage capacity, in GB.

- Price

If you know the price of the storage model that you're specifying for the template, you can enter it here. When running a plan, Intersight Workload Optimizer can use the price to calculate costs or savings when adding or removing storage in an on-prem datacenter.

Billing and Costs

As you work with Intersight Workload Optimizer, you can set up costs that Intersight Workload Optimizer uses in its calculations. This setup includes:

- **Reserved Instance Settings**

To recommend placing workloads on instance types that take advantage of discounted pricing, Intersight Workload Optimizer uses the real pricing plans that are available to the targets public cloud accounts. Setting up a purchase profile adds even more detail to the pricing structure that Intersight Workload Optimizer uses in its calculations.

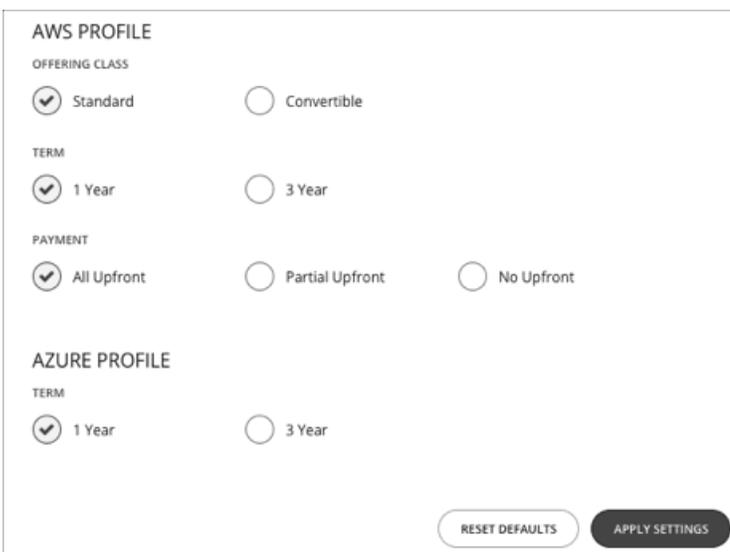
- **Price Adjustment**

Cloud providers can offer their own price lists, including special costs for services or discounts for workloads. However, Intersight Workload Optimizer does not discover these adjustments. For example, to reflect any AWS discounted prices in Intersight Workload Optimizer, you must manually configure price adjustments in the user interface, in **Settings > Billing and Costs > Price Adjustments**. To configure price adjustments for a specific billing family, you must add the corresponding master account as an AWS target in the Target Configuration page.

- **Currency**

By default, Intersight Workload Optimizer uses the dollar symbol (\$) when displaying the costs and savings that it discovers or calculates for your cloud workloads. You can set a different symbol to match your preferred currency. For example, if your cloud provider bills you in euros, change the currency symbol to €.

Reserved Instance Settings



To recommend placing workloads on instance types that take advantage of discounted pricing, Intersight Workload Optimizer uses the real pricing plans that are available to the targets public cloud accounts. Setting up a purchase profile adds even more detail to the pricing structure that Intersight Workload Optimizer uses in its calculations.

A purchase profile determines the costs that Intersight Workload Optimizer will use for all discount purchase decisions in your environment. As it sees opportunities to move workloads to another term, Intersight Workload Optimizer determines the costs based on the profile, and includes cost information in action descriptions. Intersight Workload Optimizer also uses this information to calculate projected changes in costs.

Note that the settings you configure apply to your global public cloud environment.

To set up a profile, navigate to **Settings > Billing and Costs**, and display the **RESERVED INSTANCE SETTINGS** tab. Then make the settings for your purchase profile:

- **Offering Class**

For AWS environments, choose the offering class that corresponds to the RI types that you typically use in your environment.

- Term

For AWS and Azure environments, choose the payment terms you contract for your discounts. TERM can be one of **1 Year** or **3 Year**. Typically, longer term payment plans cost less per year.

- Payment

The payment option that you prefer for your AWS RIs:

- All Upfront – You make full payment at the start of the RI term.
- Partial Upfront – You make a portion of the payment at the start of the term, with the remain cost paid at an hourly rate.
- No Upfront – You pay for the RIs at an hourly rate, for the duration of the term.

When you are satisfied with your RI Purchase Profile settings, click **APPLY SETTINGS**. Or to reset the form, click **RESET DEFAULTS**.

Price Adjustments

Cloud providers can offer their own price lists, including special costs for services or discounts for workloads. However, Intersight Workload Optimizer does not discover these adjustments. For example, to reflect any AWS discounted prices in Intersight Workload Optimizer, you must manually configure price adjustments in the user interface, in **Settings > Billing and Costs > Price Adjustments**. To configure price adjustments for a specific billing family, you must add the corresponding master account as an AWS target in the Target Configuration page.

Intersight Workload Optimizer applies these price adjustments to:

- Costs for workload template families, including:
 - Compute
 - Discount Compute
- Costs for services, including:
 - Bandwidth
 - VM Licenses
 - AWS CloudWatch
 - AWS DynamoDB
 - And others

Note that in AWS environments, Intersight Workload Optimizer does not apply any discounts or other price adjustments to Spot Compute costs.

The general steps to configure a price adjustment are:

- Create the price adjustment:
 - Specify the adjustment scope

To do this, you choose which cloud service provider is giving you the adjustment, and then choose a billing group to set the scope of the adjustment.
 - Choose the Type

The price adjustment can be a Discount or an Increase. In most cases you will specify discounts for the price adjustment. While this sets the type for the overall adjustment, you can override the type for specific line items.
 - Specify a Price Adjustment setting

The Price Adjustment is the overall adjustment that your cloud service provider offers for the billing groups in your current scope. For example, AWS might offer you a 10% discount for a given account. For that billing group, you would specify a 10% Discount for the Price Adjustment setting.
- Specify Price Overrides (AWS only)

While your service provider might offer a general price adjustment for the billing group you chose, it might also offer further discounts for select services or template families. Or it might offer discounts for some template families, but price increases for some other services. You can configure these differences as Price Overrides.

NOTE:

Intersight Workload Optimizer uses the adjustments that you configure to display costs in the user interface. However, the values for hourly cost per entity, total hourly cost, total monthly cost, or total yearly cost can show inaccuracies on the order of a fraction of a percent. This is due to rounding when calculating the adjusted cost per entity.

Creating a Price Adjustment

Cloud providers can offer their own price lists, including special costs for services or discounts for workloads. However, Intersight Workload Optimizer does not discover these adjustments. For example, to reflect any AWS discounted prices in Intersight Workload Optimizer, you must manually configure price adjustments in the user interface, in **Settings > Billing and Costs > Price Adjustments**. To configure price adjustments for a specific billing family, you must add the corresponding master account as an AWS target in the Target Configuration page.

If price adjustments are not set, Intersight Workload Optimizer will show on-demand pricing, which could result in incorrect cost information in Intersight Workload Optimizer.

After you configure an adjustment, Intersight Workload Optimizer applies it to pricing in the affected cloud scope.

To create a price adjustment in Intersight Workload Optimizer, you identify the adjustment's scope – the subscriptions or billing families the adjustment applies to – and then set the type and percentage for the price adjustment. This specifies an overall adjustment for the workloads that fall within the billing group. For AWS, you can later drill into the adjustment to specify overrides for specific template families or services.

Notes:

- To use a price adjustment with a given billing group, you must increase the memory allocated to the VM that hosts your Intersight Workload Optimizer instance. Intersight Workload Optimizer requires that you provide a minimum amount of memory when you install the product. To use price adjustments, Cisco recommends that you increase the allocated memory as follows:
 - For the first price adjustment assigned to one or more billing groups, increase by 4 GB.
 - For each subsequent price adjustment assigned to one or more billing groups, increase by an additional 1 GB.
- Whenever you add, edit, or remove a Price Adjustment that is in use, you must allow sufficient time for Intersight Workload Optimizer to fully discover all of the affected environment, and to propagate the changes throughout that environment. In an average environment, this can take up to 30 minutes. As an alternative, you can manually execute rediscovery for the affected cloud subscription or account.

To create a price adjustment:

1. Navigate to the Settings Page.

Click **More**, then display the Settings Page. From there, you can perform a variety of configuration tasks.

2. Choose Billing and Costs.

Click to navigate to the Billing and Costs page.

3. Display the PRICE ADJUSTMENT tab.

Click the **PRICE ADJUSTMENT** tab to see all of the adjustments that have been configured for your environment. In this list you can:

- Click an entry to see details and edit the adjustment
- Select an entry to delete the adjustment
- Create new price adjustments

4. Create the price adjustment.

First click **NEW PRICE ADJUSTMENT**, then specify the following settings to configure a price adjustment:

- Give the adjustment a name.
- To set the scope for this adjustment, choose its Billing Groups.

Click in the **BILLING GROUPS** field to display the Billing Groups fly-out.

In the Billing Groups fly-out, choose the cloud service provider you want to work with and then choose the billing group for the scope of this adjustment.

A Billing Group is a set of AWS accounts that are consolidated into a single billing schedule.

To consolidate billing, AWS supports billing families of AWS accounts, where there is a *master* account and other *member* accounts. Intersight Workload Optimizer lists each billing family as a billing group. You can choose a billing family to set the scope of this adjustment.

NOTE:

If a particular billing family is missing in the list of billing groups, check if you have added the corresponding master account as an AWS target in the Target Configuration page. This master account must be added in order for the billing family to display in the list.

After you have chosen your billing group, click **SAVE** to return to the Add New Price Adjustment fly-out.

- Set the Type for this price adjustment – Choose either **Discount** or **Increase**.
- Specify a percentage of adjustment as the Price Adjustment.

Enter the percentage in the **PRICE ADJUSTMENT** field. The acceptable value depends on the type of adjustment:

- For a discount: 0 - 99.99%
- For an increase: 0 - 999.99%

This is the general percentage of adjustment (increase or discount) for the current scope. For any costs within the adjustment scope, Intersight Workload Optimizer will apply this percentage as it calculates the optimal workload capacity and placement.

NOTE:

If you set an overall adjustment of 0%, then Intersight Workload Optimizer enforces a Type setting of Discount. The end result is the same, because an increase or a discount of 0% is the same.

5. (AWS only) Specify any price overrides for this price adjustment.

The PRICE ADJUSTMENT percentage you just specified applies as a default in the adjustment scope. However, you might have negotiated different prices for specific services or template families in your cloud environment. To configure these special prices, click **PRICE OVERRIDES** to open the Cloud Cost Adjustment fly-out.

For details, see [AWS Price Override \(on page 597\)](#).

6. Save your work.

After you have configured the price adjustment, click **SAVE**.

AWS Price Override

| Cloud Cost Adjustment [AWS] - My AWS Discount | | | | | | |
|---|----------|--------------------|------------|-----------------------|------------------------|-----------------------|
| SERVICES | TYPE | PRICE ADJUSTMENT % | OVERRIDE % | ORIGINAL RATE (LINUX) | EFFECTIVE ADJUSTMENT % | ADJUSTED RATE (LINUX) |
| AWS CloudTrail | Discount | 10 % | % | — | 10 % | — |
| AWS CloudWatch | Discount | 10 % | % | — | 10 % | — |
| AWS Developer Sup... | Discount | 10 % | % | — | 10 % | — |
| AWS DynamoDB | Discount | 10 % | % | — | 10 % | — |
| ^ AWS EC2 Compute * | Discount | 10 % | % | — | 10 % | — |
| ^ c5d * | | | | — | | — |
| v c5d.9xlarge...* | Discount | +0 % | 15 % | — | 15 % | — |
| v c5d.18xlarge... | Discount | 10 % | % | — | 10 % | — |

To override the PRICE ADJUSTMENT setting for AWS Billing groups, Intersight Workload Optimizer analysis can use settings for different services that AWS provides to your accounts.

In AWS, you can set up a billing family that includes a *master* account and a given set of *member* accounts. Intersight Workload Optimizer treats the AWS Billing family as a Billing Group. For more information about billing families and accounts, see [AWS Billing Families \(on page 599\)](#).

Assume you have configured a price adjustment with a discount of 10% for a billing family, to match the overall discount that AWS offers you for that scope. But then assume the account includes extra discounts for some of the services your billing families provide. Then you can create overrides to add the extra discounts to those services.

Intersight Workload Optimizer uses the adjusted costs in its analysis as it calculates actions. For example, assume a price adjustment of 10% for a billing group, and a discount of 20% for the M4.Large family of templates. As Intersight Workload Optimizer places a workload, it will consider both the template capacity and the template cost. Even if an M4 template is larger than the workload actually needs, the M4 template could be less expensive because of the added discount. In that case, Intersight Workload Optimizer will place the workload on the less expensive template.

NOTE:

The Cloud Cost Adjustment table lists the services that are available to you for the AWS Billing family that you have set up as the discount scope. The services this table displays depend on whether the billing family uses the given service, and whether there is any recorded cost at the time that you display the table. For this reason, under some circumstances you might see different services listed in the table.

Under all circumstances, the table lists the services, AWS EC2 Compute, AWS EC2 Reserved Instance, and AWS RDS.

Also, for the Cloud Cost Adjustment table to display CSP Cost and Effective Cost, you must have created a data export in AWS, and you must store it in an S3 bucket.

In the Cloud Cost Adjustment table, you can perform the following:

- Override the price adjustment for a service or template family.

To add an override, choose the line item for a service, or expand the row for a template family and:

- Set the Type. Double-click and then choose **Discount** or **Increase**. Press **Enter** to confirm your setting.
- Specify the percentage for this override, and then press **Enter** to confirm your override. The value you enter here is an absolute value for the discount or increase Intersight Workload Optimizer will apply for this line item.

When you're done setting these overrides, click **Save**.

- To remove all overrides and revert back to the PRICE ADJUSTMENT Discount, click **CLEAR ALL OVERRIDES**.
- To download a report of the discounts for each service, click **DOWNLOAD** and choose CSV or PDF.

The table lists the following information about your discounts:

- SERVICES

The different cloud services to which you can set an override discount. To see individual workload templates:

- For Azure, expand **Virtual Machines**
- For AWS, expand **AWS EC2 Compute** or **EC2 Reserved Instance**

- TYPE

Whether this price adjustment will be an increase or a discount. By default, this field shows the setting that you have made for the Price Adjustment. However, you can change it as an override for an individual entry.

- PRICE ADJUSTMENT %

The percentage that you have specified for the Price Adjustment setting. This is the general adjustment that Intersight Workload Optimizer applies by default to the given service.

- OVERRIDE %

If you have entered a value, this is the price adjustment Intersight Workload Optimizer applies to the given service.

- ORIGINAL RATE (LINUX)

The Cloud Service Provider's cost for VM templates, per hour. To see these costs, expand the workload services to show specific templates. The cost assumes no charge for the OS license, as though the VM runs Linux.

- EFFECTIVE ADJUSTMENT %

The actual adjustment for the given service.

- ADJUSTED RATE (LINUX)

The discounted cost for VM templates, per hour. To see these costs, expand **Virtual Machines** to show specific templates. The cost assumes no charge for the OS license, as though the VM runs Linux.

Enabling AWS Billing Family Recognition

As you configure AWS targets, Intersight Workload Optimizer discovers AWS accounts that are consolidated into *billing families*. A billing family has one *management* account, and zero or more *member* accounts. By recognizing billing families, Intersight Workload Optimizer more accurately calculates cloud investments and savings, and makes more accurate recommendations for RI coverage.

For RI purchases, different accounts in a billing family can share the same RI resources. At the same time, accounts in other billing families cannot use those RIs. This adds flexibility to your RI coverage, while maintaining order over the billing.

In Intersight Workload Optimizer, if you enable Billing Family Recognition, then you can see the billing family management and member accounts in the user interface, and Intersight Workload Optimizer can recommend proper RI purchases within the correct billing families.

To enable Billing Family Recognition, ensure the following as you configure your AWS targets:

- Use the proper role for each AWS target

To properly discover billing family information for a target, you must give Intersight Workload Optimizer credentials for an AWS role that includes the permission, `organizations:DescribeOrganization`. With that permission, Intersight Workload Optimizer can:

- Discover management accounts and member accounts in different billing families
- Display the account names in the user interface
- Discover billing information for each family and account
- Recommend RI actions that respect billing family boundaries

- Configure targets for the complete billing family

One billing family can consolidate a number of AWS accounts. For Intersight Workload Optimizer to include these accounts in its analysis, you must configure each one as a separate target. If you do not configure all the accounts in a billing family, then Intersight Workload Optimizer cannot discover complete billing information for that family, and its analysis will be based on incomplete information.

Intersight Workload Optimizer displays member accounts that have been configured as targets in regular text. For members that Intersight Workload Optimizer discovers but have not been configured as targets, Intersight Workload Optimizer displays their names in grayed text.

If you have enabled Billing Family Recognition, you should keep the following points in mind:

- Billing families can grow.

Intersight Workload Optimizer regularly checks the membership of your billing families. If it discovers a new member account, it adds that account to the list of members. If you have already configured the account as a target, then Intersight Workload Optimizer includes the new member in its analysis of billing families. If the new member is not already a target, then Intersight Workload Optimizer lists the new member in grayed text.

- You can configure discounts per billing family.

Intersight Workload Optimizer includes a feature to set a discount for a billing group, and to override that discount for specific template families within that scope. For more information, see [Cloud Discounts \(on page 595\)](#) and [Discount Override: AWS \(on page 597\)](#).

- You might see management accounts that have no member accounts

AWS treats every account you create as a part of a billing family. Assume you created an account, but you had no reason to consolidate its billing with any other accounts. In that case, the account appears in the Intersight Workload Optimizer user interface as a management account, but it has no member accounts.

Currency Settings

By default, Intersight Workload Optimizer uses the dollar symbol (\$) when displaying the costs and savings that it discovers or calculates for your cloud workloads. You can set a different symbol to match your preferred currency. For example, if your cloud provider bills you in euros, change the currency symbol to €.

To change the currency symbol, go to **Settings > Billing and Costs** and then click the **Currency** tab.

Intersight Workload Optimizer saves your preference in the local storage of the browser that you used to access the user interface. It reverts to the default symbol if you use another browser or view the user interface in incognito/private mode.

Currency symbols are for display purposes only. Intersight Workload Optimizer does not convert monetary amounts when you switch symbols.

Maintenance Options

NOTE:

This page includes enhancements and a more modern look-and-feel that are only available when you enable the new design framework. To switch to the new framework, click the React icon  in the navigation bar of the user interface and then Turn ON the toggle. For more information, see "Design Framework for the User Interface" in the *User Guide*.

The Maintenance Options Page provides data retention settings.

Data Retention

Data Retention

| | | |
|----------------------------------|---------------------------------|---------------------------------|
| SAVED AUDIT-LOG ENTRIES | DAILY SAVED STATISTICS | HOURLY SAVED STATISTICS |
| <input type="text" value="365"/> | <input type="text" value="60"/> | <input type="text" value="72"/> |
| Days | Days | Hours |
| MONTHLY SAVED STATISTICS | SAVED PLANS | SAVED REPORTS |
| <input type="text" value="24"/> | <input type="text" value="14"/> | <input type="text" value="30"/> |
| Months | Days | Days |

Intersight Workload Optimizer gathers metrics from your environment to provide historical reports. To optimize data storage, it consolidates the data into three groups - Hourly, Daily, and Monthly. Daily statistics consolidate Hourly data, and Monthly statistics consolidate Daily data. Intersight Workload Optimizer also saves plans, reports, and audit log entries.

You can always modify the default values to meet your requirements. Remember that the longer the retention period, the more storage is required.



Intersight Workload Optimizer Data Exports

You can export the data in Intersight Workload Optimizer to an external database and then use your favorite reporting tool to generate custom reports. This topic describes how to set up Amazon Redshift with Apache Kafka and then use the Intersight API to enable the data export. While this approach applies to Redshift, you can use it for other databases and data warehouse services.

This feature is available to select customers. Contact your Cisco representative or Cisco TAC for more information

Task Overview

To set up data exports, perform the following tasks:

1. [Set up Redshift with Kafka \(on page 601\)](#).
2. [Manage data exports using Intersight API Docs \(on page 610\)](#).
3. [View exported data \(on page 611\)](#).

Setting Up Redshift with Kafka

The setup process may take significant time to complete. In the future, the process may be captured in a template for ease of execution.

IMPORTANT:

Use the names specified in this documentation while creating the resources. The CloudFormation templates that will be used later refer to these resources by names. If you want to override the names, remember to make the corresponding changes in the template parameters during execution.

Setup can be broken down into the following sub-tasks:

1. [Set up a VPC, subnet, and NAT gateway \(on page 602\)](#).

This one-time configuration sets up Redshift.

If Redshift is already in place, skip this sub-task. Ensure that the VPC/Redshift setup is consistent with the names of the resources specified in this documentation.

2. [Provision common resources \(on page 607\)](#).

This sub-task requires a CloudFormation template to provision additional common resources. Resources created from this template are common for all tenants hosted on the Redshift cluster.

3. [Provision a new tenant \(on page 608\)](#).

This sub-task requires a CloudFormation template and is executed for each Intersight Workload Optimizer tenant. This will generate a tenant schema in Redshift, where data will be persisted.

Setting Up a VPC, Subnet, and NAT Gateway

The VPC that you create will be used exclusively for Intersight Workload Optimizer reporting of tenants running on Intersight SaaS.

NOTE:

If Redshift is already in place, skip this sub-task. Ensure that the VPC/Redshift setup is consistent with the names of the resources specified in this documentation.

Overview

To set up a VPC, subnet, and NAT gateway, perform the following steps:

1. Create a VPC.
2. Create security groups for the VPC.
3. Create the Redshift cluster subnet group.
4. Create the Redshift workload management.
5. Create the Redshift cluster.
6. Set up a Redshift secret.
7. Create parameters in the AWS Systems Manager Parameter Store.
8. Set up an AWS S3 bucket and upload the reporting files.

Creating a VPC

1. Sign in to the Amazon VPC console.
<https://console.aws.amazon.com/vpc/>
2. On the VPC dashboard, choose **Create VPC**.
3. Configure the following settings:

| Setting | Instructions |
|--------------------------|---|
| Resources to create | Choose VPC and more to automatically create the VPC and associated resources, such as subnets and route tables. |
| Name tag auto-generation | Specify <code>saas-reporting-vpc</code> . |
| NAT gateways (\$) | Choose in 1 AZ . NAT gateway is required by AWS CodeBuild to access the internet and pull libraries, such as the Flyway command line and the Redshift driver used for building the schema for the tenant. |

For all other settings, use the default configurations.

4. Choose **Create VPC**.

Amazon creates a VPC consisting of two public subnets, two private subnets, and one NAT gateway.

Creating Security Groups for the VPC

A default security group named `default` is created as part of the VPC creation. This security group is sufficient to allow CodeBuild access to Redshift.

To allow access to other sources, such as Amazon Firehose and ThoughtSpot, additional security groups with inbound rules are needed.

1. In the navigation pane of the VPC dashboard, expand **Security** and then choose **Security groups**.
2. Choose **Create security group**.
3. Configure the following settings for Firehose.
 - Basic details

| Setting | Instructions |
|---------------------|---|
| Security group name | Specify <code>RedshiftFirehoseSG</code> . |
| Description | Specify <code>RedshiftFirehoseSG</code> . |
| VPC | Specify <code>saas-reporting-vpc</code> . |

- Inbound rules

Click **Add rule** and then configure the following settings:

| Setting | Instructions |
|---------|---|
| Type | Choose Redshift . After choosing Redshift, the Protocol and Port range fields automatically update with the TCP and 5439 values, respectively. These values are not configurable. |
| Source | Choose Custom and then specify <code>52.70.63.192/27</code> . |

4. (Optional) If ThoughtSpot is enabled, choose **Create security group** and then configure the following settings:

- Basic details

| Setting | Instructions |
|---------------------|--|
| Security group name | Specify <code>RedshiftThoughtSpotSG</code> . |
| Description | Specify <code>RedshiftThoughtSpotSG</code> . |
| VPC | Specify <code>saas-reporting-vpc</code> . |

- Inbound rules

Click **Add rule** and then configure the following settings:

| Setting | Instructions |
|---------|---|
| Type | Choose Redshift . After choosing Redshift, the Protocol and Port range fields automatically update with the TCP and 5439 values, respectively. These values are not configurable. |
| Source | Choose Custom and then specify <code>54.188.23.248/32</code> . |

5. Choose **Create security group**.

Creating the Redshift Cluster Subnet Group

1. Sign in to the Amazon Redshift console.

<https://console.aws.amazon.com/redshiftv2/>

2. In the navigation pane, expand **Configurations** and then choose **Subnet groups**.

3. Choose **Create cluster subnet group**.

4. Configure the following settings:

- Cluster subnet group details

| Setting | Instructions |
|-------------|------------------------------|
| Name | Specify your preferred name. |
| Description | Specify a description. |

- Add subnets

Click **Add rule** and then configure the following settings:

| Setting | Instructions |
|---------|--|
| VPC | Choose <code>saas-reporting-vpc</code> and then choose Add all the subnets for this VPC . Be sure to remove private subnets from the list. |

5. Choose **Create cluster subnet group**.

Creating the Redshift Workload Management

1. In the navigation pane of the Redshift console, expand **Configurations** and then choose **Workload management**.
2. Choose **Create parameter group**.
3. Configure the following settings:

| Setting | Instructions |
|----------------------|---|
| Parameter group name | Specify <code>saas-reporting-param-group</code> . |
| Description | Specify <code>saas-reporting-param-group</code> . |

4. Choose **Create**.
5. Choose the parameter group that you created.
6. In the **Parameters** tab, choose **Edit parameters** and then update the following parameters:

| Parameter | Instructions |
|---|----------------------|
| <code>enable_case_sensitive_identifier</code> | Choose true . |
| <code>enable_user_activity_logging</code> | Choose true . |
| <code>max_concurrency_scaling_clusters</code> | Specify 2. |
| <code>require_ssl</code> | Choose true . |
| <code>use_fips_ssl</code> | Choose true . |

7. Choose **Save**.

Creating the Redshift Cluster

This cluster will host the tenant schema. For Intersight Workload Optimizer, Redshift must be provisioned as a cluster. Redshift serverless is not supported.

1. In the Redshift console, choose **Create cluster**.
2. Configure the following settings:
 - Cluster subnet group details

| Setting | Instructions |
|--------------------|---|
| Cluster identifier | Specify <code>saas-reporting-cluster</code> . |
| Use case | Specify <code>Production</code> . |
| Node type | Choose ra3.xlplus . |
| Number of nodes | Specify 2. |

- Database configurations

| Setting | Instructions |
|-----------------|--|
| Admin user name | Specify the user name for the cluster admin. |

| Setting | Instructions |
|----------------|--|
| Admin password | Choose Manually add the admin password and then specify the password for the cluster admin. |

- Cluster permissions

| Setting | Instructions |
|----------------------|--|
| Associated IAM roles | Choose Manage IAM roles and Create IAM Role . In the Create the default IAM role window, choose Create IAM role as default . |

- Additional configurations

| Setting | Instructions |
|--------------|----------------------|
| Use defaults | Turn off the toggle. |

- Network and security

| Setting | Instructions |
|-----------------------------|---|
| Virtual private cloud (VPC) | Choose saas-reporting-vpc . |
| VPC security groups | Choose All . |
| Cluster subnet group | Choose cluster-subnet-group-1 . |
| Enhanced VPC routing | Choose Turn on . |
| Publicly accessible | Choose Turn on Publicly accessible . |

- Database configurations

| Setting | Instructions |
|-----------------|--|
| Database name | Specify <code>demo</code> or <code>production</code> . |
| Port | Specify <code>5439</code> . |
| Parameter group | Choose saas-reporting-param-group . |
| Encryption | Choose Use AWS Key Management Service . |
| AWS KMS | Choose Default Redshift Key . |

- Backup

| Setting | Instructions |
|--------------------|-------------------------|
| Cluster relocation | Choose Enabled . |

3. Choose **Create cluster**.

Setting up a Redshift Secret

Create a secret named `redshift_secret` with the following key/value pairs. The values that you set will be used to connect to Redshift during tenant provisioning.

1. Sign in to the Secrets Manager console.

<https://console.aws.amazon.com/secretsmanager/>

2. Choose **Store a new secret**.
3. Configure the following settings:
 - Secret type

| Setting | Instructions |
|-------------|--------------------------------------|
| Secret type | Choose Other Type of Secret . |

- Credentials

| Setting | Instructions |
|-------------------|---|
| Key/value | Specify the key/value for the Redshift cluster admin. |
| ClusterJDBCURL | Specify the JDBC URL for accessing the Redshift cluster. You can copy the URL from the cluster details page. |
| dbname | Specify the name of the database to which the tenant will be provisioned. This should match the name indicated in ClusterJDBCURL. |
| ClusterIdentifier | Specify the name of the cluster. |

- Secret name

| Setting | Instructions |
|-------------|--------------------------|
| Secret name | Specify redshift_secret. |

4. Choose **Store**.

Creating Parameters in the AWS Systems Manager Parameter Store

Create three parameters in the AWS Systems Manager Parameter Store. These will be referenced in the CloudFormation template. The first two IDs must be for private subnets since they will be selected for use by Amazon CodeBuild or Kinesis Data Analytics in the template. If any one of the first two are public subnets, there may be timeout issues with CodeBuild or Kinesis Data Analytics.

1. Sign in to the AWS Systems Manager console.
<https://console.aws.amazon.com/systems-manager/>
2. In the navigation pane, expand **Application Management** and then choose **Parameter Store**.
3. For each parameter, choose **Create parameter** and then configure settings for the parameter. Choose **Create Parameter** when you are done.

The following are the required parameters:

- VpcId

| Setting | Instructions |
|---------|--|
| Name | Specify VpcId. |
| Type | Choose String . |
| Value | Specify the ID of the VPC that you created in a previous task. |

- SubnetIds

| Setting | Instructions |
|---------|--|
| Name | Specify SubnetIds. |
| Type | Choose StringList . |
| Value | Specify the IDs of the two private subnets for the VPC that you created in a previous task. Separate IDs by a comma. |

| Setting | Instructions |
|---------|--|
| | Be sure to specify only the private subnet IDs or <code>flink</code> may not work. |

- SecurityGroupIds

| Setting | Instructions |
|---------|--|
| Name | Specify SecurityGroupIds. |
| Type | Choose StringList . |
| Value | Specify the IDs of all the security groups that you created in a previous task. Separate IDs by a comma. |

Setting up an AWS S3 Bucket and Uploading the Reporting Files

The S3 bucket will be used to host all the reporting files for each release and the CloudFormation templates used to provision a tenant. Release files include the schema zip file, `etl` apps, and other dependencies.

AWS requires the S3 bucket name to be unique. If the bucket name changes, remember to use the new name (and not the default bucket name) in the CloudFormation template parameters while provisioning the tenant.

For details on setting a S3 bucket and the naming conventions, see the [AWS documentation](#).

- Download the following files:
 - [provision_common_resources.yaml](#)
 - [provision_tenant_galaxy_schema.yaml](#)
 - [schema-and-tmls-8.10.1.zip](#)
 - [etl-0.11.1.jar](#)
- Create a folder in your local machine. The folder should reflect the schema version (for example: 8.10.1). Copy `etl-0.11.1.jar` and `schema-and-tmls-8.10.1.zip` to the folder.
- Sign in to the AWS S3 Management console.

<https://console.aws.amazon.com/s3/home>
- Choose **Create bucket**. Be sure that the bucket name is unique. For example, name the S3 bucket `iwo-reporting-templates`.
- After creating the bucket, open the bucket for editing.
- Choose **Upload**.
- Choose **Add files** and then upload the YAML files you downloaded.
- Choose **Add folder** and then upload the folder that you created in a previous step.
- Choose **Upload**.

Provisioning Common Resources

Set up common resources (such as MSK cluster) from the CloudFormation template `provision_common_resources.yaml` stored in your S3 bucket. This creates a galaxy schema in the Redshift cluster.

- Sign in to the AWS CloudFormation console.

<https://console.aws.amazon.com/cloudformation>
- In the navigation pane, choose **Stacks**.
- At the top right section of the page, choose **Create stack > With new resources (standard)**.
- In the Create stack page, configure the following settings:
 - Prerequisite - Prepare template

| Setting | Instructions |
|------------------|-----------------------------------|
| Prepare template | Choose Template is ready . |

- Specify template

| Setting | Instructions |
|-----------------|---|
| Template source | <p>Specify the Amazon S3 URL for the <code>provision_common_resources.yaml</code> file that you uploaded to your S3 bucket in the previous task.</p> <p>To get the URL:</p> <ol style="list-style-type: none"> In the AWS S3 Management console, open your S3 bucket. In the Objects tab, find and right-click <code>provision_common_resources.yaml</code>. Choose Copy Link Address. |

5. In the Specify stack details page, configure the following settings:

- Stack name

| Setting | Instructions |
|------------|------------------------------|
| Stack name | Specify your preferred name. |

- Parameters

| Setting | Instructions |
|------------------|--|
| ClusterName | Change the default MSK cluster name if needed. If you change the name, remember to use that name during tenant provisioning. |
| (Optional) Email | Specify a valid email address. An email will be sent to this address when there is a failure during stack execution. |

6. In the Configure stack options page, leave the default configuration.

7. In the Review and create page, choose **Submit**.

Common resources are now created. If you encounter errors, investigate the specific resource in the **Resources** tab for the failed stack. If you specified an email address, review the error details in the email notification.

To send data to MSK cluster provisioned from the template, the Intersight account needs access to the MSK end point. For more details on setting up the connectivity, see the [AWS documentation](#).

Provisioning a New Tenant

Set up resources to enable data streaming from the CloudFormation template `provision_tenant_galaxy_schema.yaml` stored in your S3 bucket. This creates a galaxy schema in the Redshift cluster.

1. Sign in to the AWS CloudFormation console.

<https://console.aws.amazon.com/cloudformation>

2. In the navigation pane, choose **Stacks**.

3. At the top right section of the page, choose **Create stack > With new resources (standard)**.

4. In the Create stack page, configure the following settings:

- Prerequisite - Prepare template

| Setting | Instructions |
|------------------|-----------------------------------|
| Prepare template | Choose Template is ready . |

- Specify template

| Setting | Instructions |
|-----------------|---|
| Template source | <p>Specify the Amazon S3 URL for the <code>provision_tenant_galaxy_schema.yaml</code> file that you uploaded to your S3 bucket in the previous task.</p> <p>To get the URL:</p> <ol style="list-style-type: none"> In the AWS S3 Management console, open your S3 bucket. In the Objects tab, find and right-click <code>provision_tenant_galaxy_schema.yaml</code>. Choose Copy Link Address. |

5. In the Specify stack details page, configure the following settings:

- Stack name

| Setting | Instructions |
|------------|------------------------------|
| Stack name | Specify your preferred name. |

- Parameters

| Setting | Instructions |
|---------------------------------|---|
| TenantName | <p>Specify the tenant name. The name should not contain hyphens (-), underscores (_) or uppercase characters. Hyphens are not allowed in the name of a Redshift database user. Underscores and uppercase characters are not allowed in the name of an S3 bucket.</p> <p>In addition, the name should not start with a numeric value. AWS will accept this value, but Intersight Workload Optimizer does not support it as it requires double quoting the schema name in Redshift. The recommended naming is <code>wo{accountId}</code>, such as <code>wo649637c5756461301eb41640</code>.</p> <p>Ensure that the <code>FlinkApplicationJar</code>, <code>PipelineZip</code> and <code>ReleaseVersion</code> match the version already uploaded to the S3 bucket.</p> |
| (Optional) PipelineFailureEmail | Specify a valid email address. An email will be sent to this address when there is a failure during stack execution. |

NOTE:

The template also has the capability to provision ThoughtSpot related resources, but they are disabled by default. ThoughtSpot is a third-party solution that facilitates visualization and analysis of tenant data stored in Redshift. To take advantage of this feature, a ThoughtSpot account is required. For the rest of the parameters, use the default values.

6. In the Configure stack options page, leave the default configuration.

7. In the Review and create page, choose **Submit**.

Resources needed for data streaming are now created, including new Redshift schema with a tenant name and tables. If you encounter errors, investigate the specific resource in the **Resources** tab for the failed stack. If you specified an email address, review the error details in the email notification.

You have now completed all the tasks that you need to perform in AWS.

To establish connectivity between the Intersight Workload Optimizer tenant running in the cloud and the corresponding Redshift schema, the `WriteUserAccessKeyId` and `WriteUserSecretAccessKey` credentials from the stack output are needed, along with the MSK cluster public endpoints. Refer to Intersight API Docs to enable the data export feature.

Managing Data Exports Using Intersight API Docs

Perform this task using the Intersight using API Docs. You need the account administrator role to initiate a data export, which is not enabled by default.

If data export is not enabled, the API returns the following error message.

```
Feature IwoDataExporter is not enabled for this account.
```

Enabling Data Exports

In [API Docs](#), create a new management object (MO) called TenantCustomization.

Here is an example payload:

```
{
  "Account": {
    "ClassId": "mo.MoRef",
    "Moid": "12345abe7564613201a67890",
    "ObjectType": "iam.Account"
  },
  "ClassId": "iwotenant.TenantCustomization",
  "IwoId": "12345abe7564613201a67890",
  "EnableDataExtractor": true,
  "MskServerForDataExtractor": "b-2-public.myskcluster.rrrrrr.c9.kafka.us-east-1.amazonaws.com:9198,
b-1-public.myskcluster.rrrrrr.c9.kafka.us-east-1.amazonaws.com:9198",
  "WriteUserAccessKeyId": "*****",
  "WriteUserSecretAccessKey": "*****",
  "ObjectType": "iwotenant.TenantCustomization"
}
```

The MskServerForDataExtractor, WriteUserAccessKeyId and WriteUserSecretAccessKey fields must reflect the output from the AWS reporting configuration. Record the Moid of the newly created MO, which is needed for tenant updates or deletion. After you perform this step, successive data should start streaming from Intersight Workload Optimizer to the Redshift schema created for this account.

Updating Data Export Settings

In [API Docs](#), apply a patch to update the MO. Use the Moid of the TenantCustomization MO.

Disabling Data Exports

In [API Docs](#), execute the Delete command to disable data export. Use the Moid of the TenantCustomization MO.

Here is an example:

```
{
  "EnableDataExtractor": true,
  "MskServerForDataExtractor": "b-2-public.myskcluster.rrrrrr.c9.kafka.us-east-1.amazonaws.com:9198,
b-1-public.myskcluster.rrrrrr.c9.kafka.us-east-1.amazonaws.com:9198",
  "WriteUserAccessKeyId": "*****",
  "WriteUserSecretAccessKey": "*****"
}
```

Viewing Exported Data

Data such as probe state, business expense, action state, and action savings are written to the Redshift schema. Reports can be generated by joining the fact tables with the appropriate dimension tables.