Workload Optimization Manager 3.6.3 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. 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)

© 2023 Cisco Systems, Inc. All rights reserved.

Contents

What's New 8

Product Overview 12

- How Workload Optimization Manager Works 12
  - The Desired State 13
  - The Market and Virtual Currency 13
  - Risk Index 15
  - The Workload Optimization Manager Supply Chain 15
- Workload Optimization Manager Targets 16
- Resource Descriptions 18

Getting Started 23

- Logging In to Workload Optimization Manager 23
- The Home Page 24
  - APPLICATION View 24
  - ON-PREM View 25
  - CLOUD View 26
- Configuring Targets 33
- Supply Chain of Entities 42
- Working With a Scoped View 43
  - Scoping the Workload Optimization Manager Session 44
  - Overview Charts 48
  - Details View 49
  - Scope Policies 51
  - List of Entities 55
  - Navigating With the Supply Chain 56
  - Viewing Cluster Headroom 57
- Workload Optimization Manager Actions 58
  - Actions by Entity Type 59
  - Action Types 67
  - Action Categories 71
  - Action Modes 72
  - Working With the Generated Actions 73
  - Pending Actions 75
  - Actions Tips and Best Practices 86
- Working With Policies 87
  - Placement Policies 87
  - Automation Policies 92

Entity Types - Applications....123
Business Application....123
Business Application Policies....125
Business Transaction....126
Business Transaction Policies....128
Service....129
Service Policies....139
Application Component....141
Application Component Policies....144
Application Topology....146
Entity Types - Container Platform....149
Container....151
Container Policies....157
Container Spec....158
Container Spec Policies....159
Workload Controller....163
Workload Controller Policies....164
Container Pod....165
Container Pod Policies....172
Namespace....173
Container Cluster....176
Virtual Machine (Kubernetes Node)....180
Kubernetes CPU Metrics....186
Entity Types - Cloud Infrastructure....190
Virtual Machine (Cloud)....191
Cloud VM Uptime....197
Estimated On-demand Costs for Cloud VMs....201
Cloud VM Policies....209
App Component Spec....215
Virtual Machine Spec....216
Estimated On-demand Monthly Costs for Azure App Service Plans....222
Estimated On-demand Monthly Savings for Empty Azure App Service Plans....224
Virtual Machine Spec Policies....225
Database Server (Cloud)....228
Estimated On-demand Costs for Cloud Database Servers....235
Cloud Database Server Policies....239
Volume (Cloud)....242
Cloud Volume Policies....249
Database (Cloud)....251

Estimated On-demand Costs for Cloud Databases..... 255
Cloud Database Policies..... 259
Zone..... 262
Region..... 264

Entity Types - On-prem Infrastructure..... 266
Virtual Machine (On-prem)..... 267
On-prem VM Policies..... 272
vCPU Scaling Controls..... 278
Database Server (On-prem)..... 292
On-prem Database Server Policies..... 294
Volume (On-prem)..... 296
On-prem Volume Policies..... 297
Virtual Datacenter (Private Cloud)..... 298
Provider Virtual Datacenters..... 298
Consumer Virtual Datacenters..... 300
Business User..... 301
Business User Policies..... 303
Desktop Pool..... 305
Desktop Pool Policies..... 306
View Pod..... 308
View Pod Policies..... 310
Host..... 310
Host Policies..... 313
Chassis..... 317
Datacenter..... 318
Storage..... 320
vSAN Storage..... 322
Storage Policies..... 324
Logical Pool..... 329
Logical Pool Policies..... 331
Disk Array..... 332
Disk Array Policies..... 334
Storage Controller..... 336
Storage Controller Policies..... 337
IO Module..... 338
Switch..... 339
Switch Policies..... 340

Plans: Looking to the Future..... 342
Plan Management..... 343

Setting Up Plan Scenarios....344
Plan Scenarios and Types....351
Optimize Container Cluster Plan....355
Migrate Container Workloads Plan....371
Optimize Cloud Plan....384
Migrate to Cloud Plan....391
Buy VM Reservations Plan....403
Alleviate Pressure Plan....410
Custom Plan....414
Configuring Nightly Plans....426

Park: Stop or Start Cloud Resources....428

Place: Reserve Workload Resources....434
Creating a Reservation....436
Managing Reservations....438

Dashboards: Focused Views....440
Built-in Dashboards....441
On-Prem Executive Dashboard....442
Cloud Executive Dashboard....443
Container Platform Dashboard....445
Creating and Editing Custom Dashboards....445
Creating and Editing Chart Widgets....448
Chart Types....452
Actions and Impact Chart Types....452
Status and Details Chart Types....459
Cloud Chart Types....488
On-Prem Chart Types....507

Creating Groups....509

Working With Schedules....513
Managing Calendar Schedules....513
Managing Parking Schedules....517

Templates: Resource Allocations for New Entities....520
Creating Templates....521
VM Template Settings....522
Host Template Settings....523
HCI Host Template Settings....524
Storage Template Settings....527

Billing and Costs....529
Reserved Instance Settings....530
Price Adjustments....531
Creating a Price Adjustment....531
AWS Price Override....533
AWS Billing Families....535
Azure Enterprise Agreements....537
Currency Settings....539
Administrative Tasks....540
Managing User Accounts....540
Configuring a Group for SSO Authentication....550
Maintenance: Logging and Troubleshooting....551
License Configuration....554
Email Settings....555

What's New

"What's New" is a running history of new features, as they have been added to different releases of Workload Optimization Manager.

Version 3.6.3

- **Parking Actions for Cloud VMs**

  This release introduces 'parking' actions for VMs in the public cloud to help you reduce your cloud expenses. With these actions, you can stop VMs for a period of time and then start them when you need them. You can enforce parking actions on demand or according to a schedule.

  For details, see Park: Stop or Start Cloud Resources *(on page 428)*.

- **Support for Microsoft Customer Agreement Accounts**

  The Azure Billing target now supports Microsoft Customer Agreement accounts. When you add an Azure Billing target and specify the Billing Account ID for an account, Workload Optimization Manager discovers pricing for the workloads billed under that account, and then uses pricing information when recommending actions for workloads.

  For details, see "Azure Billing" in the *Target Configuration Guide*.

- **Enforcement of Security Context Constraints (SCCs) for Pod Move Actions**

  Red Hat OpenShift uses SCCs to control permissions for pods. When executing pod move actions, Workload Optimization Manager now carries over the user-level SCCs of a pod to its new node. This prevents privilege escalation issues that occur when pods run with admin-level SCCs in their new nodes.

  For details, see Pod Move Actions *(on page 167)*.

- **Temporary Increases in Namespace Quotas**

  When a namespace has a defined quota and a Workload Controller in the namespace requires a resize, Workload Optimization Manager now increases the quota temporarily to ensure that the execution of the resize action complies with your rolling updates strategy.

  For details, see Namespace Actions *(on page 175)*.

- **New Operator Lifecycle Management (OLM) Feature**

  The new OLM feature automatically upgrades the Workload Optimization Manager client operator for customers utilizing the SaaS secure connector for hybrid support to provide a better SaaS experience with seamless access to new probe functionality, and reduced maintenance of SaaS client deployments.

- **VMware vCenter 8.0 Support**

  This release introduces support for VMware vCenter version 8.0. To submit Ideas for specific vCenter 8.0 features that Workload Optimization Manager should consider, use the Cisco Ideas Portal.

Version 3.6.2

- **New View of the Target Configuration Page**

  This release introduces a new view of the Target Configuration page to help you manage targets with ease. This new view includes the following improvements:

  - The page lists your targets in a sortable table and provides a new target editing experience.
  - For some targets, the page shows stages of validation and discovery, as well as any errors to help with troubleshooting.

  NOTE:
  Search and filter features will be introduced in the future.

  Try the new view by navigating to **Settings > Target Configuration** and turning on **New View** at the top-right section of the page.

  For details, see Configuring Targets *(on page 33)*.

- **Declarative Policy Configuration for Kubernetes Services**

  This release introduces a declarative approach to creating and managing policies for horizontally scalable Kubernetes services. These policies adjust the number of pod replicas that back those Services to help you meet Service Level Objectives (SLOs) for your applications.

  With this new approach, application owners who do not have access to the Workload Optimization Manager user interface can now use Custom Resource (CR) files in a Kubernetes cluster to create and manage policies. The settings in these policies are synced with Workload Optimization Manager every ten minutes to keep your environment up-to-date.

  For details, see Actions for Kubernetes Services *(on page 131)*.

Version 3.6.1

- **Azure VM Memory Metrics Collection via REST API**

  Workload Optimization Manager now supports collecting host OS based (agentless) memory metrics for Azure VMs, based on the REST API (Preview) recently made available by Azure. This mechanism is used as a fallback to collect VM memory when other agent-based VM memory collection mechanisms are not configured. Azure Memory Source Groups have been updated to better reflect the supported VM memory sources. For details, see "Memory Metrics Collection: Azure" in the *Target Configuration Guide*.

Version 3.6.0

This quarterly release includes the following new features and improvements.

CONTAINER RESOURCE MANAGEMENT

- **Migrate Container Workloads Plan**

  This release introduces a new plan type called Migrate Container Workloads. Run this 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 Workload Optimization Manager optimized plan. The results further highlight the actions you need to take to maintain and optimize workload performance in the new cluster.

  For details, see Migrate Container Workloads Plan *(on page 371)*.

- **Node Reconfigure Actions**

  For Kubernetes environments, Workload Optimization Manager can now generate reconfigure actions to notify you of nodes that are currently in the `NotReady` state. After a node's condition is addressed, and the state changes to `Ready`, Workload Optimization Manager can begin to monitor the health of the node and the associated container pods.

  For details, see Node Reconfigure Actions *(on page 184)*.

- **Unknown Container Pods Visibility**

  When a node is in the `NotReady` state, the associated container pods are in the `Unknown` state. These pods now display with a gray color in the user interface to help you differentiate them from other pods. In addition, you can now use the `Unknown` state as a container pod filter when you use Search or create groups.

- **Improvements**
  - Workload Optimization Manager now treats static pods as DaemonSets for the purpose of provisioning or suspending nodes. A static pod provides a node with a specific capability, and is therefore controlled by the node instead of the control plane.
    - If a node to be provisioned requires a static pod, Workload Optimization Manager generates actions to provision the node and the corresponding static pod.
    - If the only workload type left on a node is a static pod, Workload Optimization Manager generates actions to suspend the node and the corresponding static pod.
  - This release introduces the following memory usage optimizations for Kubeturbo to prevent out-of-memory (OOM) issues:
    - Proactively trigger garbage collection by taking advantage of the new Garbage Collector mechanism introduced in Go 1.19.
    - Enable pagination for Kubernetes API calls to allow iterating over large result sets in chunks. Page size is now dynamically calculated based on the cluster size and memory limit of Kubeturbo.
  - In earlier releases, default load balancers would timeout after 60 seconds if no activity was detected. This posed a challenge for Kubeturbo since it sometimes took over a minute before a heartbeat was sent after establishing a secure WebSocket communication with a server. Starting with this release, the Kubeturbo to server communication heartbeat is now configured at 30 seconds.

CLOUD RESOURCE MANAGEMENT

- **Azure App Service Optimization**

  Workload Optimization Manager can now recommend actions to scale your provisioned Azure App Service plans to optimize application performance, or delete empty plans to reduce your cloud expenditure. For scale actions, Workload Optimization Manager uses percentile calculations to measure application demand more accurately, and then picks the instance type that can meet demand at the lowest possible cost.

  For details, see "Support for Azure App Service" in the *Target Configuration Guide*.

ON-PREM RESOURCE MANAGEMENT

- **Improvements**
  - With the discovery of vCenter guest metrics now enabled by default, Workload Optimization Manager collects VM memory usage data at the OS level (via VMware tools), which is more accurate than the active memory data reported by the hypervisor. Customers may notice more frequent resize up vMem actions since the vMem usage values reported by the OS are almost always higher than those reported by the hypervisor.
  - When Workload Optimization Manager discovers that a vCenter datastore is in maintenance mode, it will stop recommending actions to move VM storage to that datastore.

USER INTERFACE MANAGEMENT

- **New Entity Filters**

  The following filters are now available when you use Search or create groups.

Entity	Filter
Virtual Machine	Storage Cluster Name
Database	Database Server Name
Container Pod	State

- **Improvements**
  - The Multiple Resources chart now includes the "Last 60 Days" timeframe.
  - When a Workload Optimization Manager instance manages a large number of targets, it could take several minutes for the Target Configuration page to load. With this improvement, the page now loads within seconds.
  - When a Workload Optimization Manager instance monitors a large number of cloud accounts and you set the scope to your global cloud environment, it could take several minutes for the Top Accounts chart to load. With this improvement, the chart now loads within seconds.

Product Overview

Thank you for choosing Workload Optimization Manager, the premier solution for Application Resource Management (ARM) of cloud and virtual environments.

Application Resource Management is a top-down, 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, Workload Optimization Manager represents your environment holistically as a *supply chain* of resource *buyers* and *sellers*, all working together to meet application demand. By empowering buyers (VMs, instances, containers, and services) with a budget to seek the resources that applications need to perform, and sellers to price their available resources (CPU, memory, storage, network) based on utilization in real-time, Workload Optimization Manager keeps your environment within the *desired state* – operating conditions that achieve the following conflicting goals at the same time:

- Assured application performance

  Prevent bottlenecks, upsize containers/VMs, prioritize workload, and reduce storage latency.

- Efficient use of resources

  Consolidate workloads to reduce infrastructure usage to the minimum, downsize containers, prevent sprawl, and use the most economical cloud offerings.

Workload Optimization Manager is a containerized, microservices architected application running in a Kubernetes environment (or within a VM) on your network or a public cloud VPC. You then assign services running on your network to be Workload Optimization Manager *targets*. Workload Optimization Manager discovers the entities (physical devices, virtual components and software components) that each target manages, and then performs analysis, anticipates risks to performance or efficiency, and recommends actions you can take to avoid problems before they occur.

How Workload Optimization Manager Works

To keep your infrastructure in the desired state, Workload Optimization Manager 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. On top of that, 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, Workload Optimization Manager 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, Workload Optimization Manager provides closed-loop management of resources, from the datacenter, through to the application.

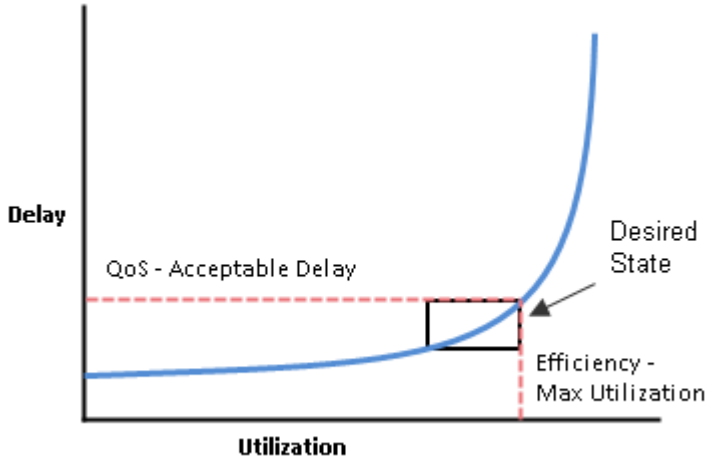
See Supply Chain of Entities *(on page 42)* for a visual layout of the buyer and seller relationships.

Workload Optimization Manager 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. Workload Optimization Manager 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 Workload Optimization Manager, 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 upper 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 upper 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, Workload Optimization Manager analyzes the operating conditions and constantly recommends actions to keep the entire environment within the desired state. If you execute these actions (or let Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager can use economic solutions to efficiently redistribute the demand or increase the supply.

Workload Optimization Manager uses two sets of abstraction to model the environment:

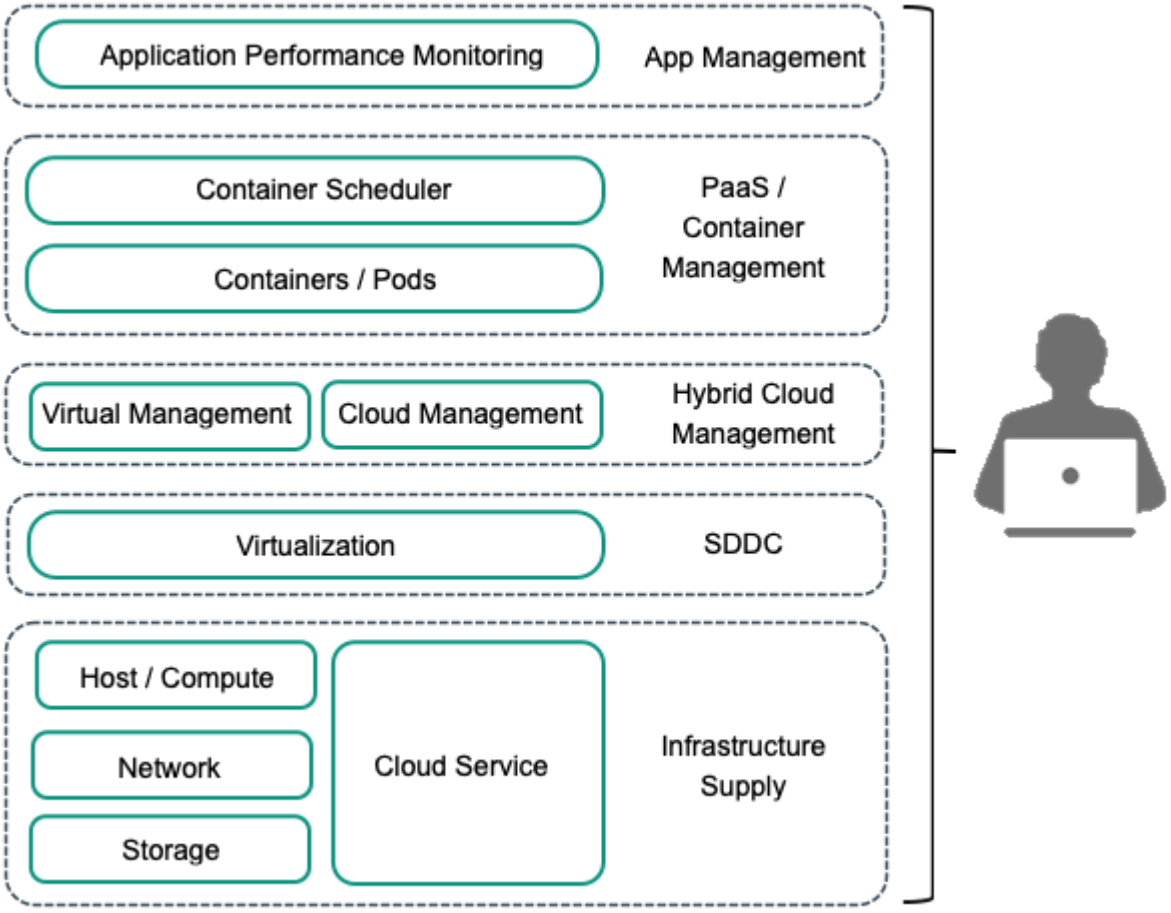
- Modeling the physical and virtual IT stack as a service supply chain

  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 42)* 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

Workload Optimization Manager 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. Workload Optimization Manager constantly works to keep the Risk Index within acceptable bounds.

You can think of Risk Index as the cost for a resource — Workload Optimization Manager works to keep the cost at a competitive level. This is not simply a matter of responding to threshold conditions. Workload Optimization Manager 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 Workload Optimization Manager. 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, Workload Optimization Manager 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 Workload Optimization Manager configuration is ready to use in many environments. However, you can fine-tune the configuration to address special services and resources in your environment. Workload Optimization Manager 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 Workload Optimization Manager operation. For more information about policies, see Working With Policies *(on page 87)*.

The Workload Optimization Manager Supply Chain

Workload Optimization Manager 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, Workload Optimization Manager discovers VMs, the hosts and datastores that provide resources to the VMs, and the applications that use VM resources. For a Kubernetes target, it discovers services, namespaces, containers, container pods, and nodes. The entities in your environment form a chain of supply and demand where some entities provide resources while others consume the supplied resources. Workload Optimization Manager *stitches* these entities together, for example, by connecting the discovered Kubernetes nodes with the discovered VMs in vCenter.

For information about specific members of the supply chain, see Supply Chain of Entities *(on page 42)*.

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	The basic building block of Workload Optimization Manager supply and demand. All the resources that Workload Optimization Manager monitors are commodities. For example, the CPU capacity or memory that a host can provide are commodities. Workload Optimization Manager can also represent clusters and segments as commodities. When the user interface shows *commodities*, it's showing the resources a service provides. When the interface shows *commodities bought*, it's showing what that service consumes.
Composed Of	The resources or commodities that make up the given service. For example, in the user interface you might see that a certain VM is *composed of* commodities such as one or more physical CPUs, an Ethernet interface, and physical memory. Contrast *Composed Of* with *Consumes*, where consumption refers to the commodities the VM has bought. Also contrast *Composed Of* 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.

Term:	Definition:
Consumes	The services and commodities a service has bought. A service *consumes* 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.
Entity	A buyer or seller in the market. For example, a VM or a datastore is an entity.
Environment	The totality of data center, network, host, storage, VM, and application resources that you are monitoring.
Inventory	The list of all entities in your environment.
Risk Index	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. 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. In most cases, for optimal operation the Risk Index on a provider should not go into double digits.

Workload Optimization Manager Targets

You can assign instances of the following technologies as Workload Optimization Manager targets.

- Applications and Databases
  - Apache Tomcat 7.x, 8.x, and 8.5.x
  - AppDynamics 4.1+
  - AppInsights
  - Dynatrace 1.1+
  - IBM WebSphere Application Server 8.5+
  - Instana, release-209 or later
  - JBoss Application Server 6.3+
  - JVM 6.0+
  - Microsoft SQL Server 2012, 2014, 2016, 2017, and 2019
  - MySQL 5.6.x and 5.7.x
  - NewRelic
  - Oracle 11g R2, 12c, 18c, and 19c
  - Oracle WebLogic 12c
- Cloud Native
  - Kubernetes, including any compliant k8s distribution (Rancher, Tanzu, open source, etc.)
  - Cloud-hosted k8s services (AKS, EKS, GKE, IBM, Cisco IKS, ROKS, ROSA, etc.)
  - Red Hat OpenShift 3.11 and higher (OCP 4.x)

- Fabric and Network
  - Cisco UCS Manager 3.1+
  - HPE OneView 3.00.04
- Guest OS Processes
  - SNMP
  - WMI: Windows versions 8 / 8.1, 10, 2008 R2, 2012 / 2012 R2, 2016, 2019 and 7
- Hyperconverged
  - Cisco HyperFlex 3.5
  - Nutanix Community Edition
  - VMware vSAN
- Hypervisors
  - Microsoft Hyper-V 2008 R2, Hyper-V 2012/2012 R2, Hyper-V 2016, Hyper-V 2019
  - VMware vCenter 7.0 and 8.0
- Orchestrator
  - ActionScript
  - Flexera One
  - ServiceNow
- Private Cloud
  - Microsoft System Center 2012/2012 R2 Virtual Machine Manager, System Center 2016 Virtual Machine Manager, and System Center Virtual Machine Manager 2019
- Public Cloud
  - Amazon AWS
  - Amazon AWS Billing
  - Google Cloud Platform (GCP)
  - GCP Billing
  - Microsoft Azure Service Principal
  - Azure Billing
  - Microsoft Enterprise Agreement
- Storage
  - EMC ScaleIO 2.x and 3.x
  - EMC VMAX using SMI-S 8.1+
  - EMC VPLEX Local Architecture with 1:1 mapping of virtual volumes and LUNs
  - EMC XtremIO XMS 4.0+
  - HPE 3PAR InForm OS 3.2.2+, 3PAR SMI-S, 3PAR WSAPI
  - IBM FlashSystem running on Spectrum Virtualize 8.3.1.2 or later (8.4.2.0 or later recommended)
  - NetApp Cluster Mode using ONTAP 8.0+ (excluding AFF and SolidFire)
  - Pure Storage F-series and M-series arrays
- Virtual Desktop Infrastructure
  - VMware Horizon

Resource Descriptions

To perform intelligent workload balancing, Workload Optimization Manager collects raw data from its target servers – hypervisors, cloud management stacks, public cloud accounts, etc. Workload Optimization Manager polls its targets at 10-minute intervals to collect the latest data samples. It then uses these 10-minute data points for analysis and to display data in the GUI.

The way Workload Optimization Manager collects host memory data from vCenter Server illustrates how this works. vCenter Server collects peak metrics from its managed VMs at 20-second intervals. Every ten minutes Workload Optimization Manager polls vCenter Server to collect its last round of data samples (30 samples in 10 minutes). To track a VM's utilization of host memory, Workload Optimization Manager requests *memory.active* data samples from vCenter. From that polling, Workload Optimization Manager can track:

- Peak Memory Utilization - Workload Optimization Manager uses the greatest value in each polling sample. This gives the highest percentage of active memory utilization for the selected VM (or group of VMs), calculated over the selected time period. For a maximum value, Workload Optimization Manager uses the highest observed active memory value in the data sample.
- Average Memory Utilization - Workload Optimization Manager averages all the values in each polling sample.

NOTE:
The above example describes utilization calculations for on-prem entities. For workloads on the public cloud, Workload Optimization Manager includes the **Aggressiveness** and **Max Observation Period** settings to calculate a percentile of utilization. By using a percentile, Workload Optimization Manager can recommend more relevant actions to take advantage of elasticity on the public cloud.

The following table lists the metrics Workload Optimization Manager collects, and includes details about how they are collected or measured. When the Workload Optimization Manager user interface plots charts of clusters or groups of devices, these charts show the average of the percentage of allocated resources that are used.

Resource:	Description:
1- 2- 4-CPU Rdy	Wait time in the ready queue on the host, measured in ms. Workload Optimization Manager monitors 1-CPU, 2-CPU, 4-CPU, up to 32-CPU ready queues on hosts. Charts show 1 - 4 CPU values. The charts show the percentage allocated ready queue capacity that is in use on the host. For host charts, this is a measure of the total ready queue wait time for all the VMs running on that host.
Balloon	Ballooning capacity on the PM, measured in KBytes. This capacity is the greater of: ▪ 65% of the VMem configured for all powered-on VMs that the PM hosts ▪ The physical memory capacity of the PM Charts show the percentage of the PM's ballooning capacity that is in use.
Buffer	For network environments that support buffered switch ports (Arista networks), this resource measures utilization of a port buffer. For example, if a host connects to the network through port 1 on a switch, and that port has enough traffic to cause packet buffering, this resource will show utilization.
Connection	Connection is the measurement of Database Server connections utilized by applications. Workload Optimization Manager collects connection data from Database Servers discovered via Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager collected displays in the Connection chart. For details, see Connection Chart *(on page 463)*.

Resource:	Description:
Cooling	Allocated cooling indicates the highest acceptable running temperature for a physical device, such as a chassis in a compute fabric.
CPU	Host CPU capacity, measured in MHz. This shows what percentage of CPU cycles are devoted to processing instructions. ▪ Host charts show the percentage of the host's CPU capacity that is in use. ▪ VM charts show the percentage of the host's CPU capacity that is consumed by the given 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. Workload Optimization Manager collects cache hit rate data from Database Servers discovered via Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager collected displays in the DB Cache Hit Rate chart. For details, see DB Cache Hit Rate Chart *(on page 464)*.
Database Memory (DBMem)	Database memory (or DBMem) is the measurement of memory utilized by a Database Server. Workload Optimization Manager collects memory data from Database Servers discovered via Databases and APM targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager collected displays in the DB Memory chart. For details, see DB Memory Chart *(on page 466)*.
Flow0 – InProvider Flow	For measuring network flow, the flow that is within a single provider – For example, the network flow between VMs that are hosted by the same physical machine. This measures network flow between consumers that are on the same set of closely connected providers. Charts show the percentage of capacity that is utilized. Note that Workload Optimization Manager assumes an unlimited supply of InProvider Flow because this flow does not go across the physical network.
Flow1 – InDPOD Flow	For measuring network flow, the flow that is local to the given DPOD. This measures network flow between consumers that are on the same set of closely connected providers. Charts show the percentage of capacity that is utilized.
Flow2 – CrossDPOD Flow	For measuring network flow, the flow that is between different DPODs. This measures network flow between consumers that are on different sets of closely connected providers. Charts show the percentage of capacity that is utilized.
Heap	Heap is the portion of a VM or container's memory allocated to individual applications. Workload Optimization Manager collects heap data from Application Components discovered via Applications and APM targets. When you set the scope to one or several Application Components, the data that Workload Optimization Manager collected displays in the Heap chart. For details, see Heap Chart *(on page 467)*.
HotStorage	For Nutanix platforms, the storage capacity on the server-attached flash.

Resource:	Description:
IO	Data rate through the host's IO adapter, measured in KBytes/sec. ▪ Datacenter charts show the average percentage of the host IO capacity that is in use, for all the hosts in the datacenter. ▪ Host charts show the percentage of the host's total IO capacity that is in use.
IOPS	See 'Storage access'.
Latency	Allocated capacity for latency on a datastore. This measures the latency experienced by all VMs and hosts that access the datastore. Charts show the percentage of allocated latency that is in use on the datastore.
Mem	Host memory, measured in Kbytes. ▪ Host charts show the percentage of the host's memory that is in use. ▪ VM charts show the percentage of the host's memory that is consumed by the given VM.
NET	Data rate through the host's Network adapter, measured in Kbytes/sec. ▪ Datacenter charts show the average percentage of the host NET capacity that is used for all the hosts in the datacenter. ▪ Host charts show the percentage of the host's total NET capacity that is in use.
Normalization factor (AWS only)	Normalization factor is a measure of RI capacity that you can use to compare or combine the capacity for different instance families. Workload Optimization Manager 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 Workload Optimization Manager scope. Each workload is assigned normalized units depending on its instance type. AWS normalization factor and Azure reservation ratio are equivalent concepts.
Power	A measure of the power that is consumed by a physical device.
Reservation ratio (Azure only)	Ratio refers to the number of Azure reservation units that cover workload capacity compared to the total number of reservation units for a given Workload Optimization Manager scope. Each workload is assigned reservation units based on its instance type. Reservation ratio information appears in the tooltips of cloud discount charts. Information about the Azure instance types and their reservation workloads is provided in the Discount Inventory chart. Azure reservation ratio and AWS normalization factor are equivalent concepts.
Remaining GC Capacity	Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC). Workload Optimization Manager collects GC data from Application Components discovered via 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 Workload Optimization Manager calculated displays in the Remaining GC Capacity chart. For details, see Remaining GC Capacity Chart *(on page 469)*.
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).

Resource:	Description:
	Workload Optimization Manager collects response time data from entities discovered via 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 Workload Optimization Manager collected displays in the Response Time chart. For details, see Response Time Chart *(on page 470)*.
Risk Index	A measure of the impact on 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. For all the resources that impact performance or risk, charts show the Risk Index for the most utilized resource of a given entity. For example, if a host has a Risk Index of 6 for MEM and 12 for CPU, the chart will show the higher value.
Storage access	Storage access, also known as IOPS, is the per-second measurement of read and write access operations on a storage entity. Workload Optimization Manager collects storage access data from VMs, Database Servers, and storage entities discovered via cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Workload Optimization Manager collected displays in the Storage Access chart. For details, see Storage Access Chart *(on page 473)*.
Storage amount	Storage amount is the measurement of storage capacity utilized by a given entity. Workload Optimization Manager collects storage amount data from VMs, Database Servers, and storage entities discovered via cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Workload Optimization Manager collected displays in the Storage Amount chart. For details, see Storage Amount Chart *(on page 475)*.
Swap	The rate of memory swapping to disk, in bytes per second. The default capacity is 5,000,000 Byte/sec.
Threads	Threads is the measurement of thread capacity utilized by applications. Workload Optimization Manager collects thread data from Application Components discovered via Applications and APM targets. When you set the scope to one or several Application Components, the data that Workload Optimization Manager collected displays in the Threads chart. For details, see Threads Chart *(on page 476)*.
Transaction log	Transaction log is the measurement of storage capacity utilized by Database Servers for transaction logging. Workload Optimization Manager collects transaction log data from Database Servers discovered via Databases and APM targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager collected displays in the Transaction Log chart. For details, see Transaction Log Chart *(on page 481)*.
Transactions	Transaction is a value that represents the per-second utilization of the transactions allocated to a given entity. Workload Optimization Manager collects transaction data from entities discovered via Applications, Databases, and APM targets. Entities include Business Applications, Business Transac-

Resource:	Description:
	tions, Services, Application Components, and self-hosted Database Servers. When you set the scope to any of these entities, the data that Workload Optimization Manager collected displays in the Transaction chart. For details, see Transaction Chart *(on page 478)*.
VCPU	The allocated CPU capacity, measured in MHz. Charts show the percentage of VCPU cycles that are devoted to processing instructions.
Virtual storage	Virtual storage is the measurement of virtual storage capacity utilized by VMs. Workload Optimization Manager collects virtual storage data from VMs discovered via on-prem and APM targets. When you set the scope to one or several VMs, the data that Workload Optimization Manager collected displays in the Virtual Storage chart. For details, see Virtual Storage Chart *(on page 482)*.
VMem	The allocated memory capacity, measured in Kbytes. Charts show the percentage of VMem that is in use. Note that percentages of allocated VMem are measured against whichever is the less of: The VMem limit (if set) or the allocated VMem capacity. This is also true in reports and recommended actions. For example, assume a VM with allocated VMem of 8 GB, but a limit of 4 GB. In this case, the percentage in a chart shows the percentage utilized of 4GB.

Getting Started

To get started with the platform, open a web browser to your Workload Optimization Manager installation. The Workload Optimization Manager 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 Workload Optimization Manager from any internet connection.

Logging In to Workload Optimization Manager

To get started with the platform, open a web browser to your Workload Optimization Manager installation. The Workload Optimization Manager 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 Workload Optimization Manager from any internet connection.

Before you can log in, your enterprise must have a valid Workload Optimization Manager account, or an instance of Workload Optimization Manager must be installed in your environment. To get the IP address of your Workload Optimization Manager installation, contact your system administrator.

To log in to Workload Optimization Manager:

1. Navigate your Web browser to the Workload Optimization Manager installation.

   For the URL, provide the IP address or machine name for the installation. This URL opens the Workload Optimization Manager Login page. You should bookmark this URL for future use.

2. Provide the user name and password for your account.

   Your system administrator creates user accounts. Contact your system administrator for login information.

After you log in, the browser opens to the Home Page *(on page 24)*. This page is your starting point for sessions with the Workload Optimization Manager platform. From the Home Page you can see the overviews of your environment.

To display this information, Workload Optimization Manager communicates with *target services* such as hypervisors, storage controllers, and public cloud accounts. Note that your Workload Optimization Manager administrator sets up the target configuration. For information about supported targets and how to configure them, see "Target Configuration" in the *Target Configuration Guide*.

The Home Page

When you launch Workload Optimization Manager, the **Home Page** is the first view you see. 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 123)*.
  - 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

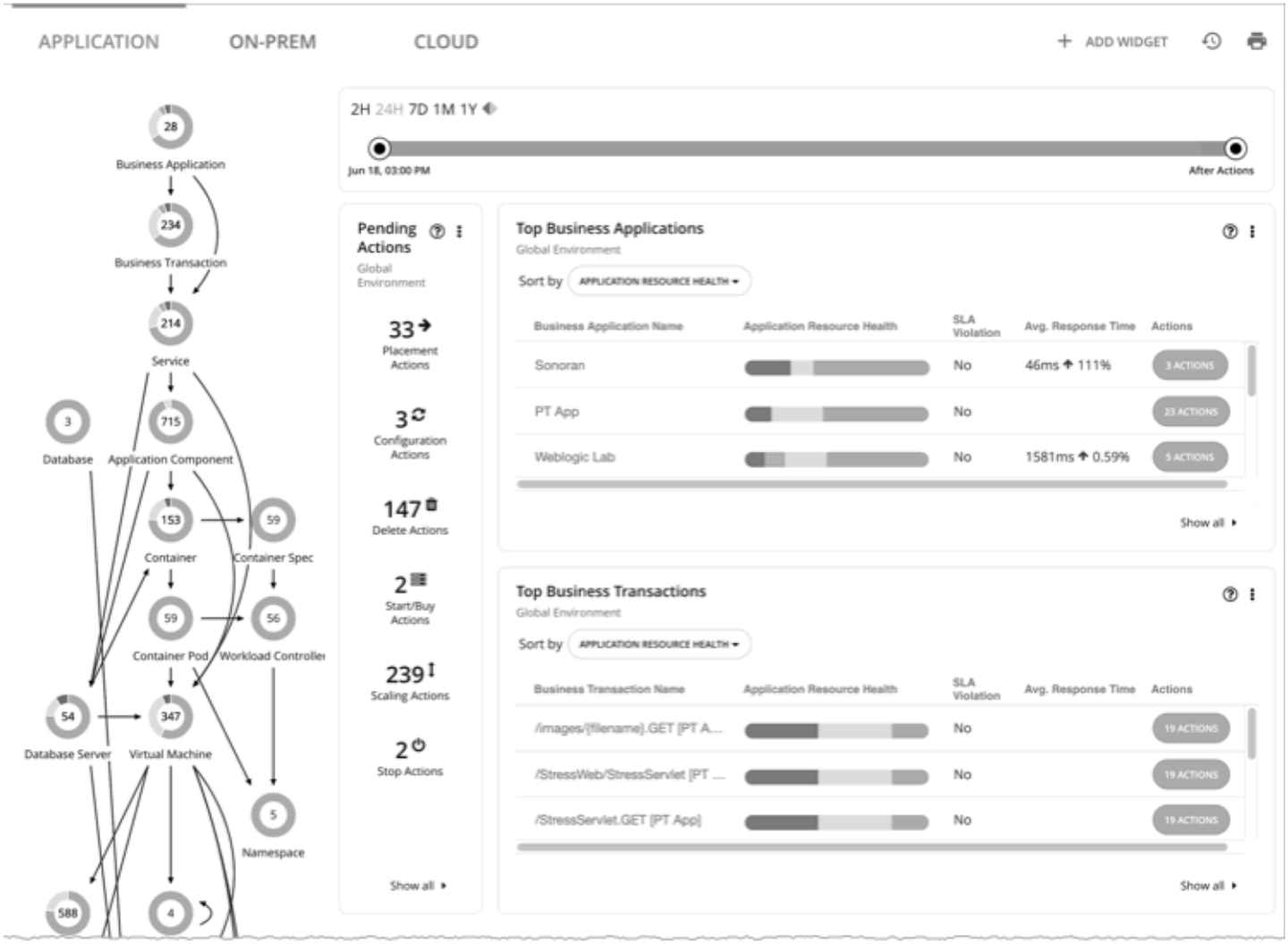
  Click an entity tier in the Supply Chain to see a list of those entities. For example, click Virtual Machine to see a list of all the VMs in your environment.
- Navigate to other Workload Optimization Manager pages, including:
  - Search – Set the session scope to drill down to details about your environment
  - Plan – Run what-if scenarios
  - Place – Use Workload Optimization Manager 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 Workload Optimization Manager to set up business rules and policies, configure targets, define groups, and perform other administrative tasks

Getting Home

Wherever you are in your Workload Optimization Manager session, you can always click the Home icon to return to the **Home Page**.

APPLICATION View

The **APPLICATION** view presents your environment in the context of your Business Applications *(on page 123)*. See the overall health of your applications, examine any performance and compliance risks, and execute the actions that Workload Optimization Manager recommends to address these risks.



This view also shows the Business Transactions *(on page 126)* and Services *(on page 129)* 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, Workload Optimization Manager displays them in both the ON-PREM and the CLOUD views. Once Workload Optimization Manager 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 Workload Optimization Manager 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 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 Workload

Optimization Manager 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 Workload Optimization Manager installation. For information about setting up public cloud targets, see "Cloud Targets" in the *Target Configuration Guide*.

In addition, to view full cost information in AWS, you must have created a Cost and Usage report in your AWS account and you must store it in an S3 bucket.

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 30)*.
- 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

Workload Optimization Manager tracks your cloud spend based on the cost information it discovers from targets (for example, accounts, billing reports, and on-demand or discount costs), price adjustments *(on page 531)*, and rate cards.

Cost for Services

Workload Optimization Manager uses the billing reports from your cloud service providers, as they are associated with your cloud targets. Workload Optimization Manager parses these reports to get cost breakdowns by service, service provider, Azure Resource Group, and cloud account. You can see cost data in charts such as:

- Cloud Estimated Cost
- Cost Breakdown by Cloud Accounts, Component, or Service Provider
- Expenses

Workload Expenses

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

- Compute

  For compute expenses Workload Optimization Manager uses hourly expense per template as specified in the associated public cloud account.
- Storage

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

- License

  For AWS environments, Workload Optimization Manager can calculate OS costs. To calculate the OS cost for a VM, Workload Optimization Manager 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.

  For Azure environments, Workload Optimization Manager can track OS costs for existing VMs. For actions to purchase reservations, Workload Optimization Manager does not include the OS cost. For more information about Azure reservations, see Azure Enterprise Agreements *(on page 537)*.

- IP

  For some workloads, you might use IP services that incur a cost. For example, your cloud provider might charge to grant a static IP to a VM. On AWS environments Workload Optimization Manager can include that cost in its calculation and analysis.

Workload Optimization Manager 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.

Workload Optimization Manager uses this cost information when making VM resize and placement decisions. You can see this information in Expenses charts.

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 Workload Optimization Manager, you should consider:

- For AWS, the Workload Optimization Manager 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, Workload Optimization Manager does not discover or show the costs for dedicated hosting of your workloads.
- Pricing for DT workloads is different than pricing for Shared Tenancy. Workload Optimization Manager 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.
- Workload Optimization Manager discovers the true costs of RIs for DT workloads. However, because the on-demand VM costs are based on Shared Tenancy, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager generates an action to change that workload to a different instance type. Because Workload Optimization Manager 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

  Workload Optimization Manager 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, Workload Optimization Manager 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.

Resizing Cloud Workloads

To resize a workload (for example, a VM or an RDS instance) on the cloud, Workload Optimization Manager chooses the cloud tier that best matches the workload requirements. This can be to reduce cost by choosing a smaller tier, or it can be to assure performance by choosing a larger tier. To accomplish the resize, Workload Optimization Manager actually moves the workload to the new tier. This can include moving to a new availability zone.

Note that resize decisions also take into account the savings you can realize through discounts *(on page 30)*. When considering workload resize actions, Workload Optimization Manager can recommend resizing to a an instance type that takes advantage of discounted pricing because the overall cost will be less.

As it considers a resize, Workload Optimization Manager also considers the storage and network requirements. Even if the compute resources are underutilized on a workload, if the available tiers cannot support the workload's storage or network requirements then Workload Optimization Manager will not recommend the change.

> **NOTE:**
> In AWS environments, under certain circumstances VM resizing can fail. If the restart of the VM initially fails, Workload Optimization Manager waits 30 seconds and tries to restart again. Workload Optimization Manager will try to restart up to four times. If the restart still fails, Workload Optimization Manager assumes the VM cannot start up on the new tier, and it restarts the VM on the old tier.

Scaling on the Public Cloud

On the cloud, scaling actions change the VM to a different instance type. These can include:

- Changing a VM to an instance type with different capacity
- Changing a VM to an instance type that is charged a discounted rate

For these actions, the action list shows the current cost for the source workload, and also the projected cost given the change. To show the current cost, Workload Optimization Manager uses the actual costs for that workload. However, to show the projected cost it uses an estimate based on average utilization for the VM, for the costs of the given tier.

Note that scaling to an instance type that is charged a discounted rate can result in running the VM on a larger instance when the cost is lower. This might occur even though the VM does not need that capacity and there are other smaller instance types available.

In Azure environments, there are circumstances where a VM resize can be especially disruptive. In a given region, the infrastructure can be made up of different clusters that have different sets of underlying hardware. Further, some tiers that are available in the given region are only available on different clusters. If Workload Optimization Manager recommends resizing from a tier on one cluster, to a tier on another cluster, then the resize action can take longer to complete than usual.

In both Azure and AWS environments, Workload Optimization Manager conforms to specific instance requirements as it generates resize actions. For more information, see:

- Azure Instance Requirements *(on page 196)*
- AWS Instance Requirements *(on page 194)*

Discounts

Workload Optimization Manager 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. Workload Optimization Manager discovers the following discounts:

- AWS Reserved Instances (RIs) and Savings Plans
- Azure reservations
- GCP committed use discounts

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

- Potential Savings or Necessary Investments Charts *(on page 458)*

  If Workload Optimization Manager has found actions you can take to improve performance or to reduce cost, then you can see an overview of them in the Potential Savings or Necessary Investments charts. To see a listing of the specific actions, click **Show All** at the bottom of the chart. For more about actions, see Workload Optimization Manager Actions *(on page 58)*.

- Discount Utilization *(on page 502)*

  This chart shows how well you have utilized your current discount inventory *(on page 499)*. 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 497)*

  This chart shows the percentage of VMs covered by discounts. If you have a high percentage of on-demand VMs, you should be able to reduce your monthly costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

- Discount Inventory *(on page 499)*

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

- Recommended RI Purchases *(on page 495)*

  Workload Optimization Manager 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 403)*.

  Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.

Support for Government Workloads

AWS GovCloud (US) and Azure Government provide dedicated regions for US government customers and their partners to architect secure cloud solutions and meet regulatory and compliance requirements.

Workload Optimization Manager discovers workloads in these regions when you add the required accounts as targets. For details on the required accounts, see "AWS GovCloud Targets" in the *Target Configuration Guide* and "Azure Government Targets" in the *Target Configuration Guide*.

Discovered workloads include:

- AWS VMs (including auto-scaling groups), volumes, database servers, and spot instances
- Azure VMs (including availability/scale sets), volumes, and SQL databases

Workload Optimization Manager recommends actions on VMs, volumes, and SQL databases to address performance issues and optimize costs.

> **NOTE:**
> Workload Optimization Manager currently does not support Azure Government integration with Application Insights. You can add accounts for Azure Government and Application Insights as targets, but Application Insights will only return performance data for non-government workloads.

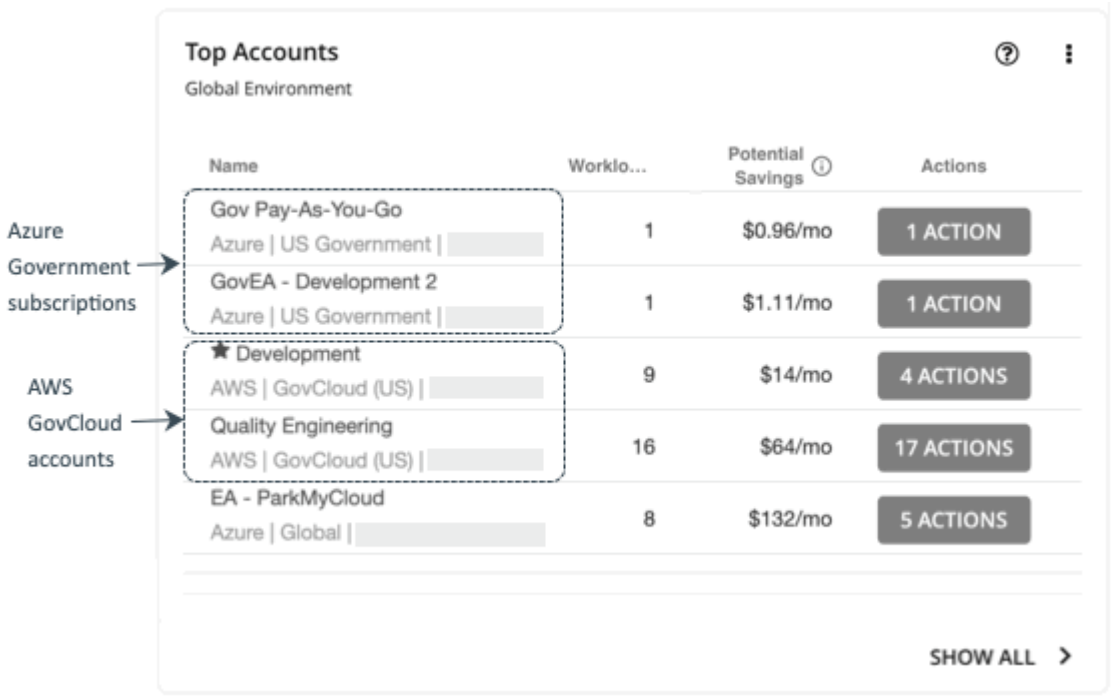
Information in Charts

Use the following charts to view information about your government accounts and workloads.

- **Top Accounts** chart

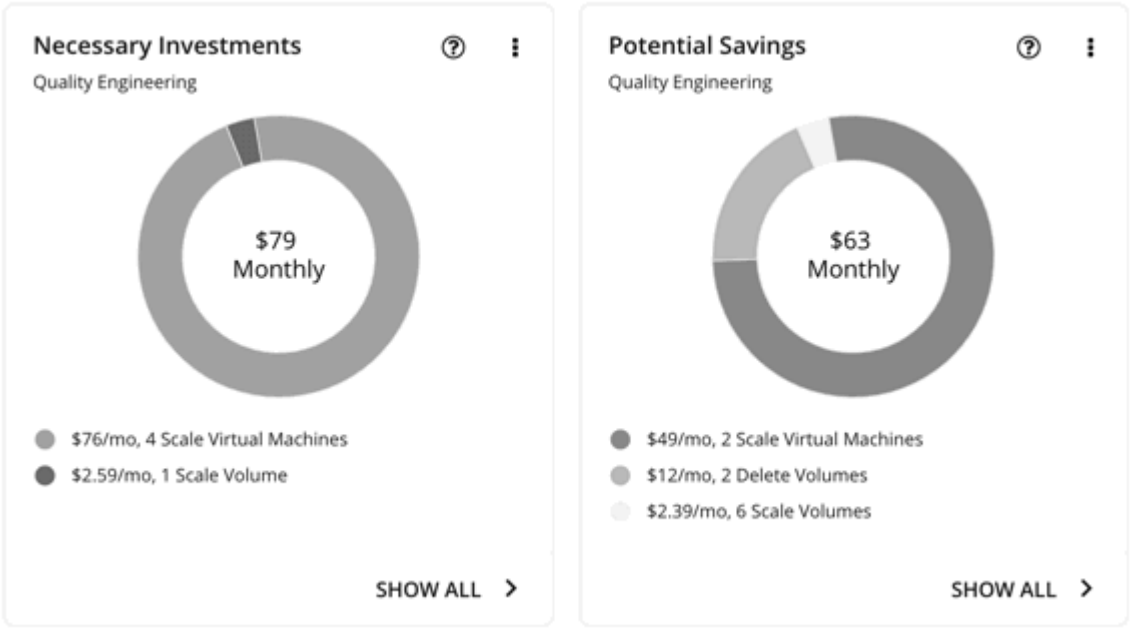
  Use the Top Accounts chart as a starting point. This chart shows the following:

  - Azure Government subscriptions discovered via the service principal and EA accounts that you have added as targets
  - AWS GovCloud master and member accounts that you have added as targets. Accounts with a star symbol are master accounts.



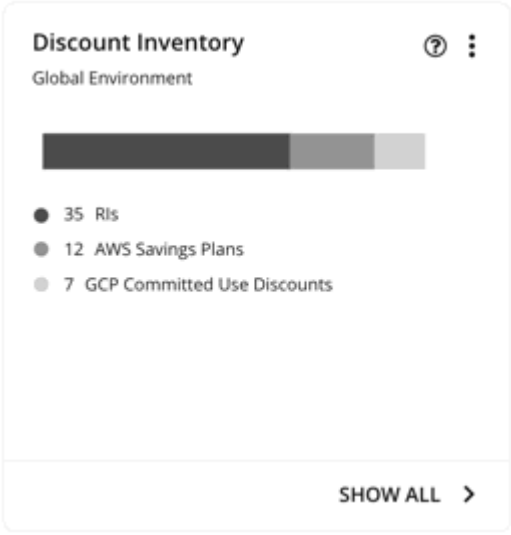
- **Necessary Investments** and **Potential Savings** charts

  Set the scope to a government account or subscription, and then see the Necessary Investments and Potential Savings charts to evaluate the costs you would incur or save if you execute all the pending actions for your government workloads.



- **Discount Inventory** chart

  The government accounts that you added as targets enable Workload Optimization Manager to gain full insight into the discounts *(on page 30)* that you have purchased for your government workloads. Even as you selectively add secondary targets, Workload Optimization Manager remains aware of all discounts, and how they are utilized across the board. This increases the accuracy of the allocation and purchase recommendations that Workload Optimization Manager generates for your government workloads.



Workload Planning

You can run an Optimize Cloud plan to identify performance and efficiency opportunities for existing government workloads, or a Migrate to Cloud plan to migrate government VM groups to another cloud provider.

For on-prem clusters, you can run a Migrate to Cloud plan to see how you can safely migrate the VMs in these clusters to a government account/subscription and region.

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:

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

  For details about scale and delete actions, see Virtual Machine Spec *(on page 216)*.

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 Service Principal and Subscription Permissions" in the *Target Configuration Guide*.

Configuring Targets

A target is a service that performs management in your virtual environment. Workload Optimization Manager 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, Workload Optimization Manager communicates with the service via the management protocol that it exposes – The REST API, SMI-S, XML, or some other management transport. Workload Optimization Manager 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. Workload Optimization Manager then discovers and validates the target, and then updates the supply chain with the entities that the target manages.

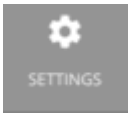
> **NOTE:**
> Workload Optimization Manager regularly checks the status of your targets. If target discovery or validation fails, the Target Configuration page updates the status. Under some circumstances, the target can become discoverable or valid again, but the status does not update. In this case,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

For a list of supported targets and configuration requirements, see "Target Configuration" in the *Target Configuration Guide*.

Starting with Workload Optimization Manager version 3.6.2, you have the option of using the enhanced target settings experience using the new target view. In this view, Target Manager lists your targets in a sortable table and includes a new edit experience. Some target integrations communicate the stages of validation and discovery, as well as any encountered error conditions to aid in troubleshooting your target configurations. Follow the instructions below to configure your target using either the new or old view.

Configuring a target using the new view

1. Navigate to the Settings page.

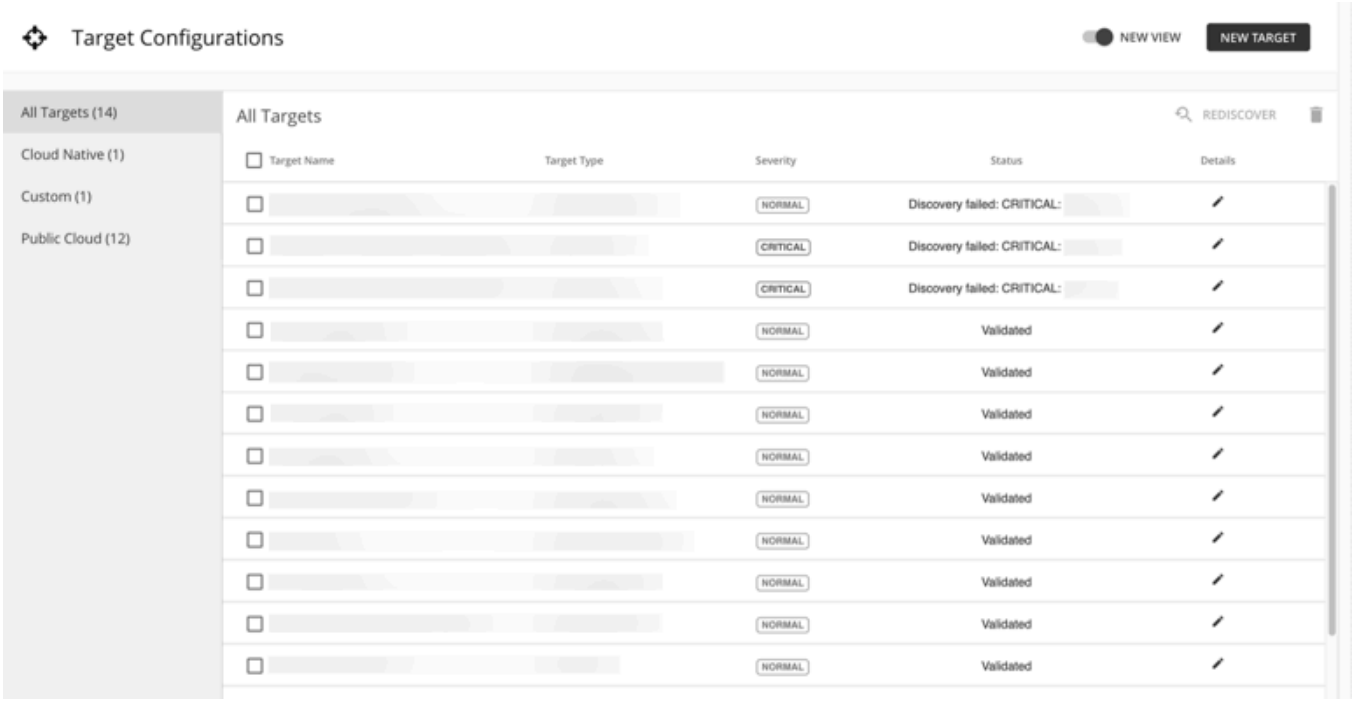


Click to navigate to the Settings page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.

2. Choose Target Configuration.

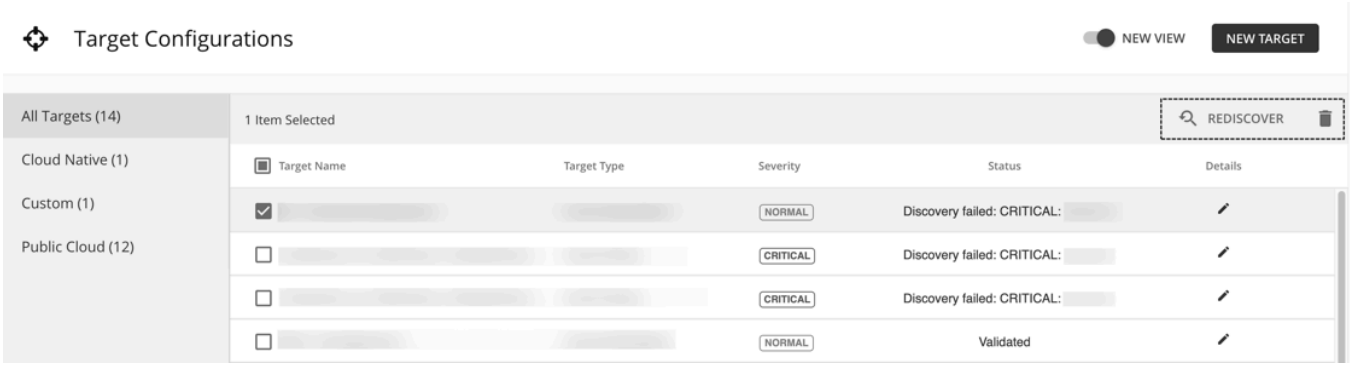
Click to navigate to the Target Configuration Page.

3. Click **New View** to enable to new target view.
4. Review the list of targets.



This page lists all the targets that you currently have configured for Workload Optimization Manager in a sortable table. You can inspect or edit these targets, or add a new target.

5. Select one or more targets to work with.



When you select a target you can:

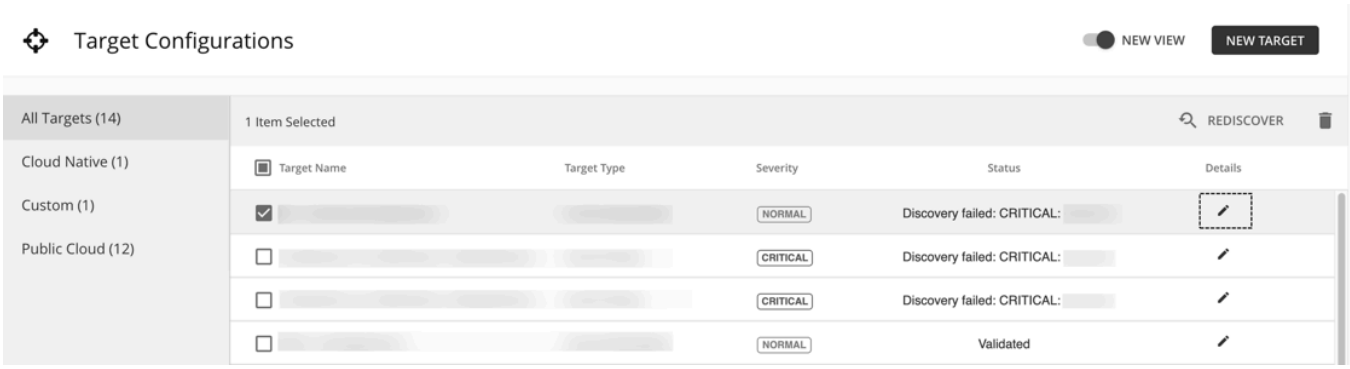
- Rediscover

  Direct Workload Optimization Manager to fully discover the entities that this target manages. This will rebuild the topology that is associated with this target.

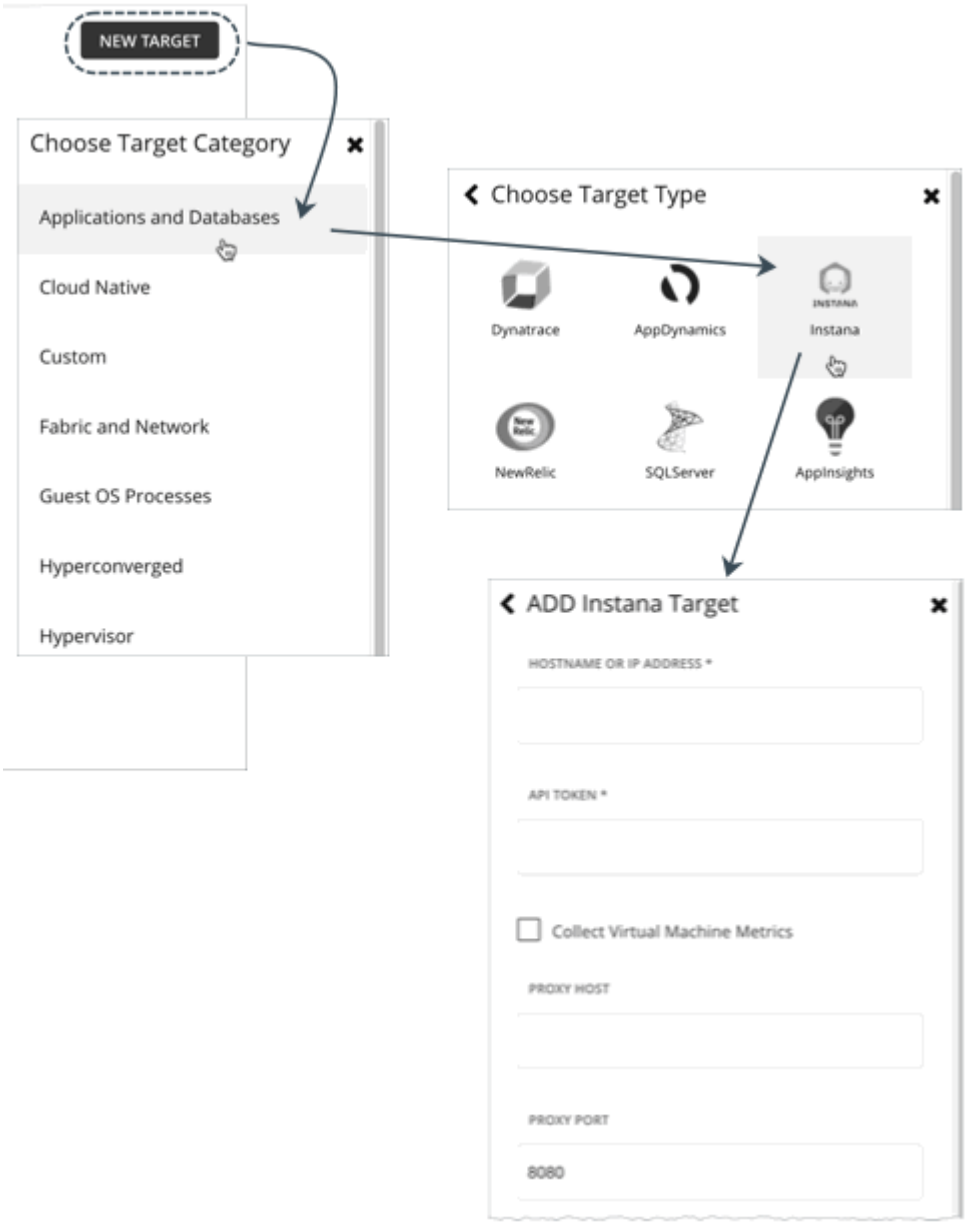
- Delete

  When you delete a target, Workload Optimization Manager removes all the associated entities from the supply chain.

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



7. Create a new target and add it to Workload Optimization Manager.



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

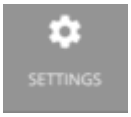
  Validation is in progress.
- Validated

  Validation was successful. Workload Optimization Manager 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.

Configuring a target using the old view

1. Navigate to the Settings page.

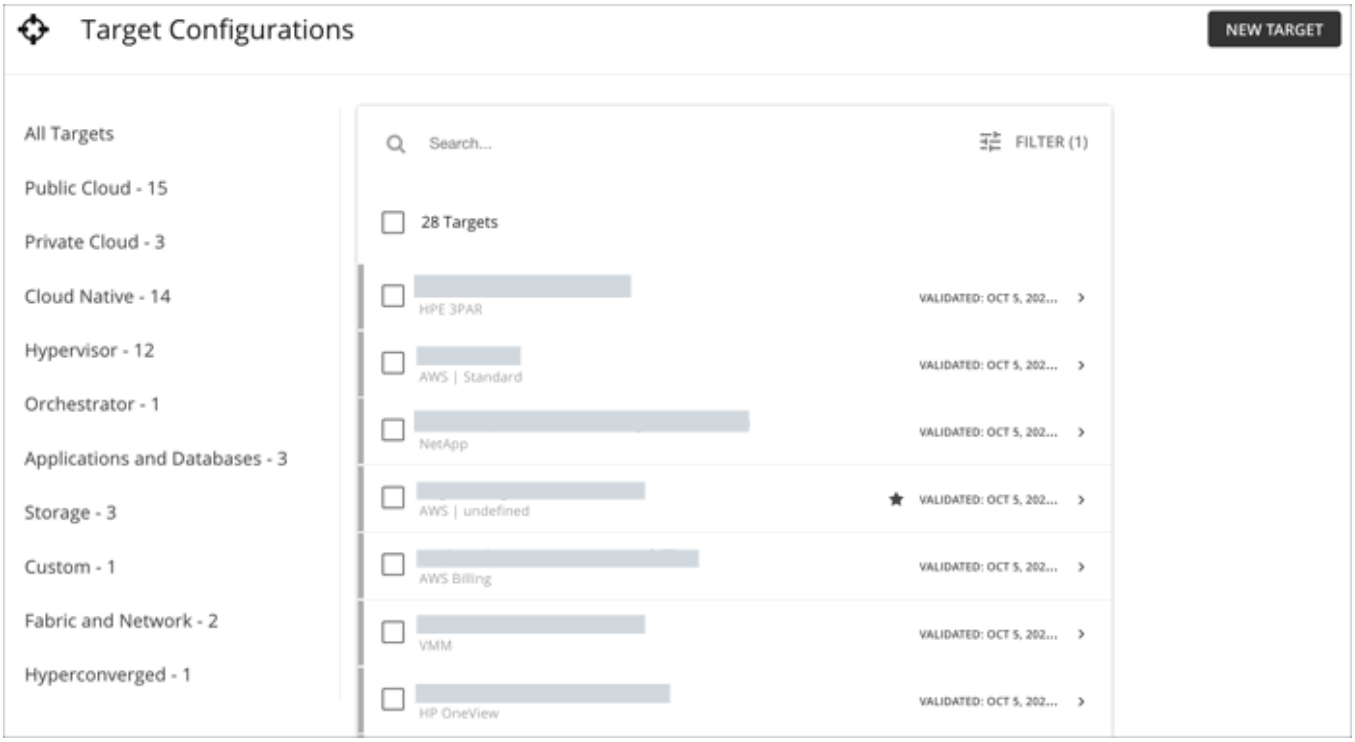


Click to navigate to the Settings page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.

2. Choose Target Configuration.

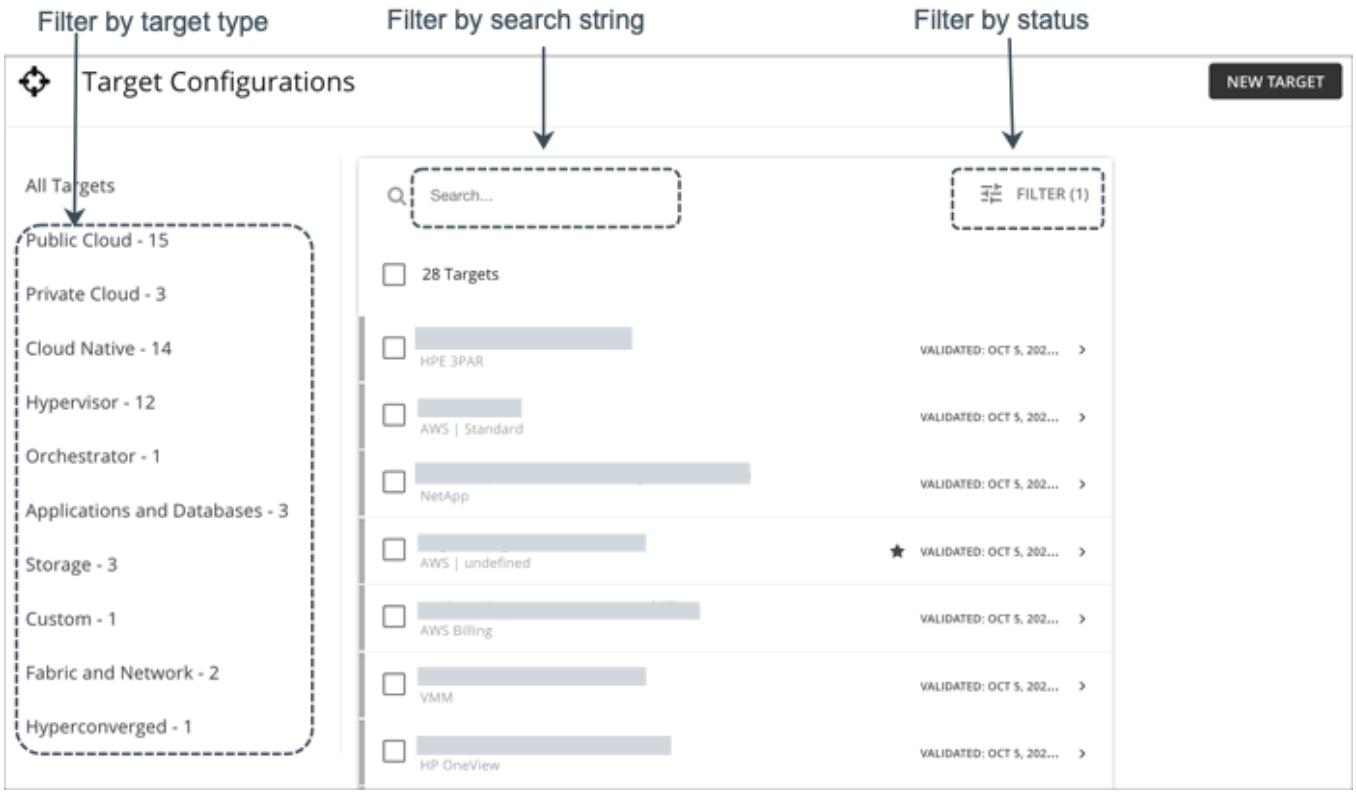
Click to navigate to the Target Configuration page.

3. Review the list of targets.



This page lists all the targets that you currently have configured for Workload Optimization Manager. You can inspect or edit these targets, or add a new target.

4. Filter the list of targets.



For a long list of targets, you can:

- Filter targets by target type.
- Use Search to filter targets by 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 sort targets by name or status.

5. Select one or more targets to work with.

When you select a target you can:

- Rediscover

  Direct Workload Optimization Manager to fully discover the entities that this target manages. This will rebuild the topology that is associated with this target.

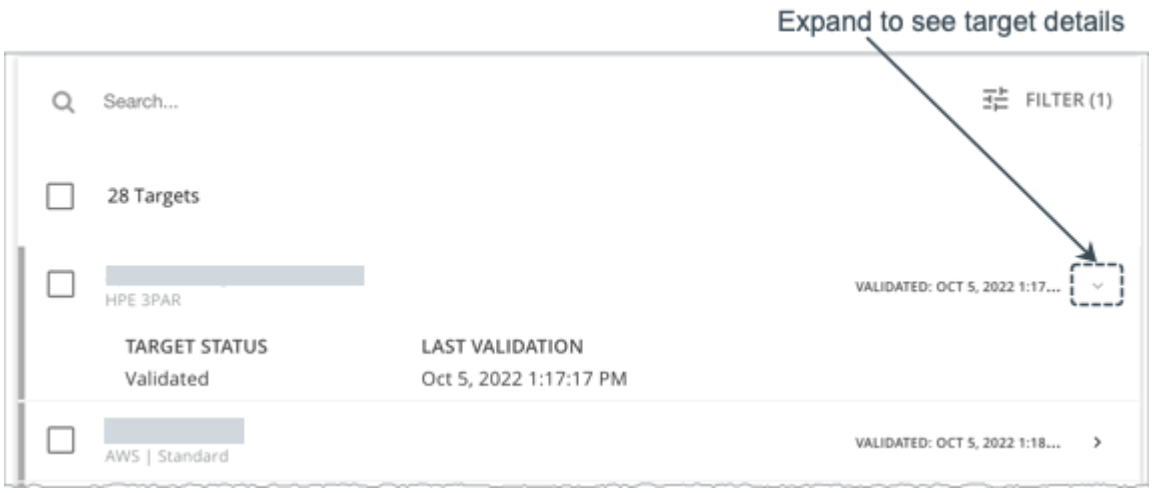
- Validate

  Direct Workload Optimization Manager to validate its connection with the target. For example, if you create a new user account on the target, you can edit the target connection to use that account, and then revalidate.

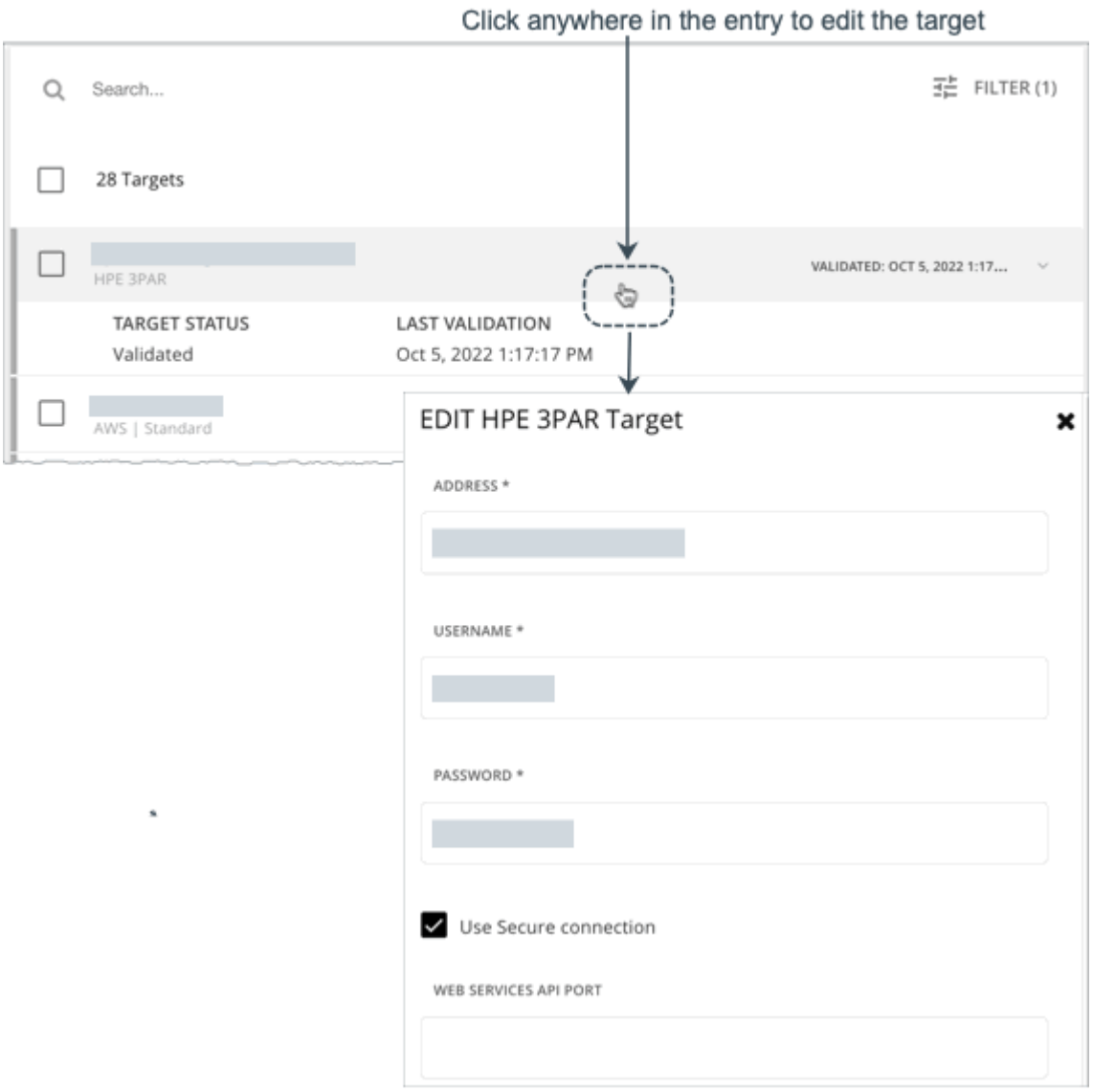
- Delete (delete icon)

  When you delete a target, Workload Optimization Manager removes all the associated entities from the supply chain.

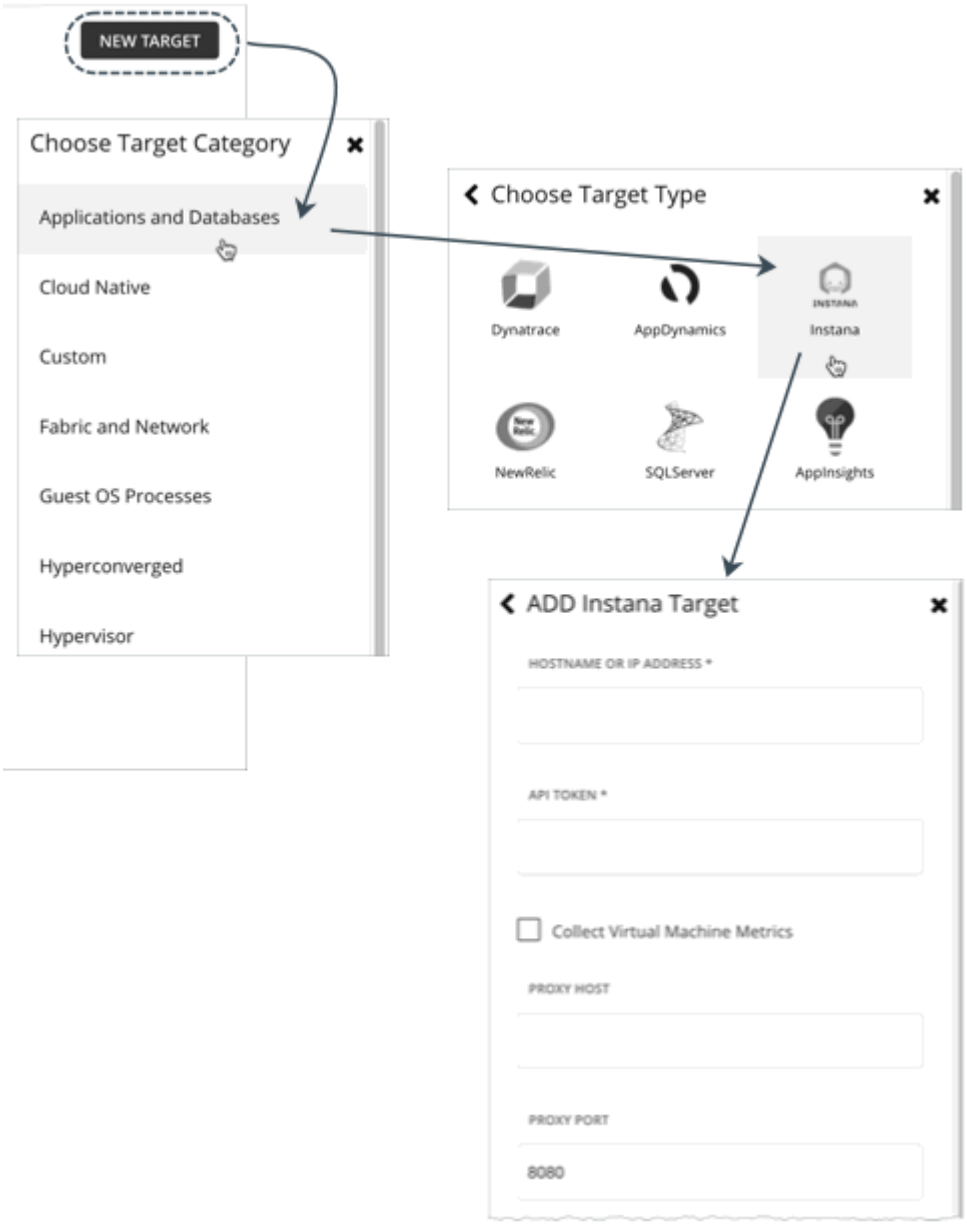
6. Expand an entry to see details.



You can also click anywhere in the entry to edit the target's configuration. For example, if you entered the wrong username or password, you can change those credentials and validate the target again.



7. Create a new target and add it to Workload Optimization Manager.



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.

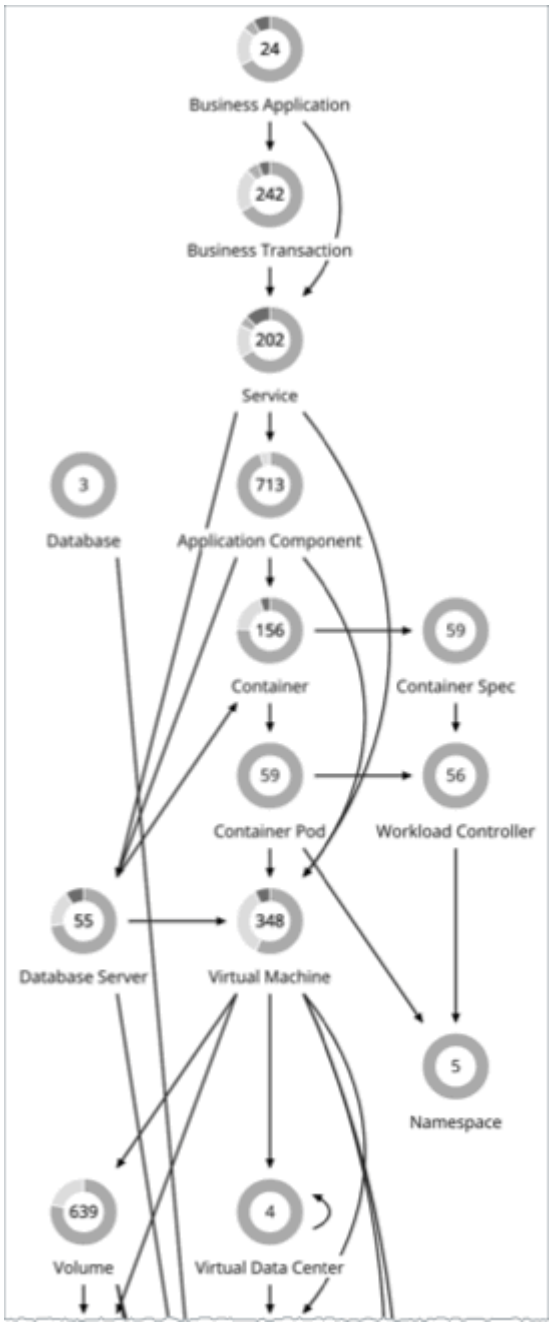
- Validating

  Validation is in progress.
- Validated

  Validation was successful. Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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

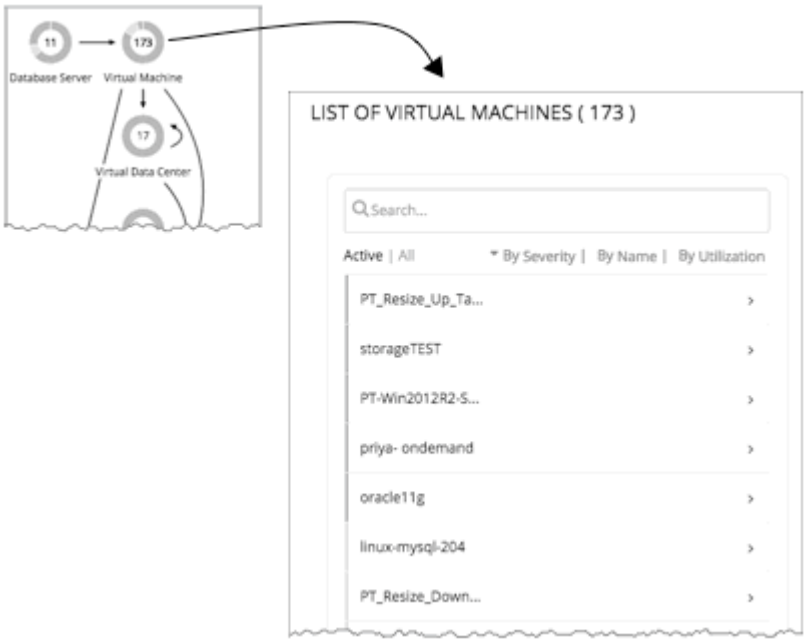
  The arrow from one entry to another indicates the flow of resources. For example, the Virtual Machine entry has arrows to Hosts and to Storage. If the VMs are running in a Virtual Data Center, it will have another arrow to that as well. This means that your VMs consume resources from hosts, storage,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

Listing Entities From the Home Page

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

- APPLICATIONS – All your Business Applications *(on page 123)*
- 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 **Home Page** shows a Global view of your environment. To drill down into specifics of your environment, you can set a scope to your Workload Optimization Manager 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.

Things You Can Do

- Scoping the Workload Optimization Manager Session *(on page 44)*
- Navigating With the Supply Chain *(on page 56)*
- Viewing Cluster Headroom *(on page 57)*

Scoping the Workload Optimization Manager Session

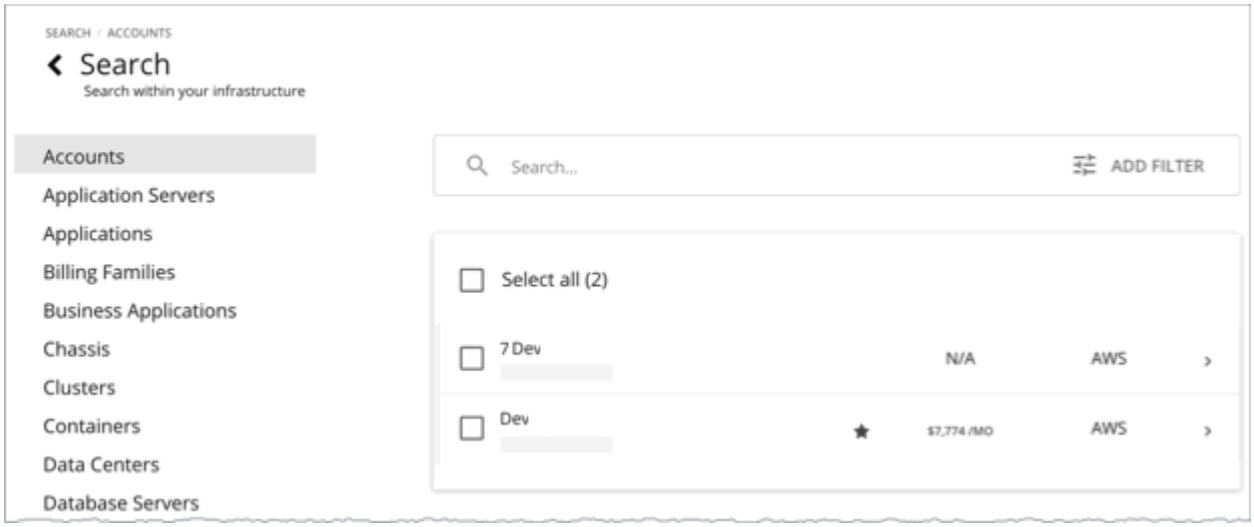
The default scope for the **Home Page** shows an overview of the global environment. What if you want to focus on less than the global environment? Assume you are responsible for a subset of workloads in your environment. This could be:

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

It's easy to set the session scope so that Workload Optimization Manager 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.

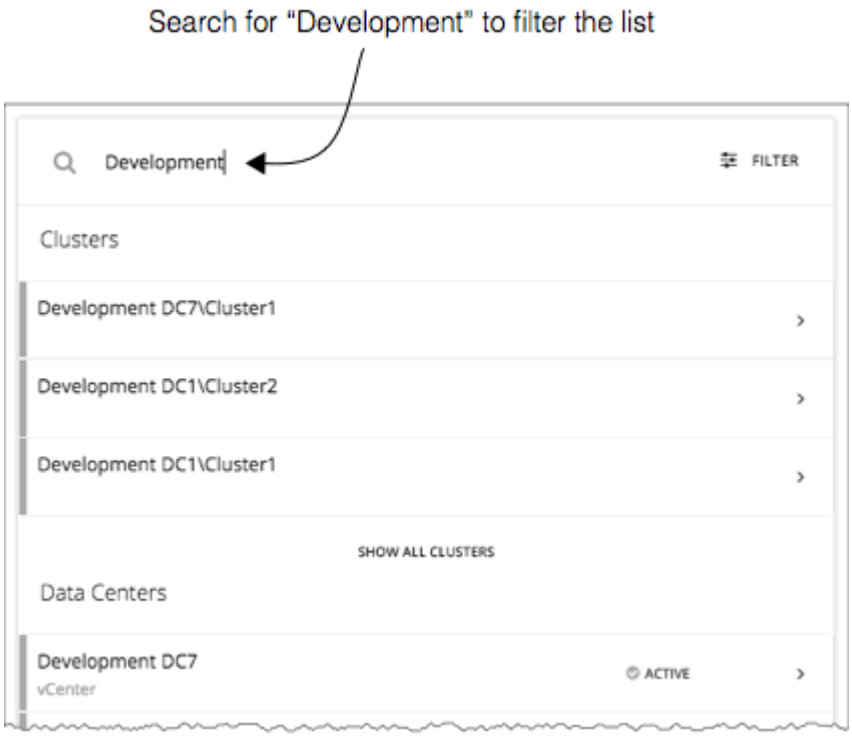
1. Navigate to the Search Page.

   Click to navigate to 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 type of entities that you want to search through. Find the list of entity types on the left. Select **All** to search the complete environment. Or 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.

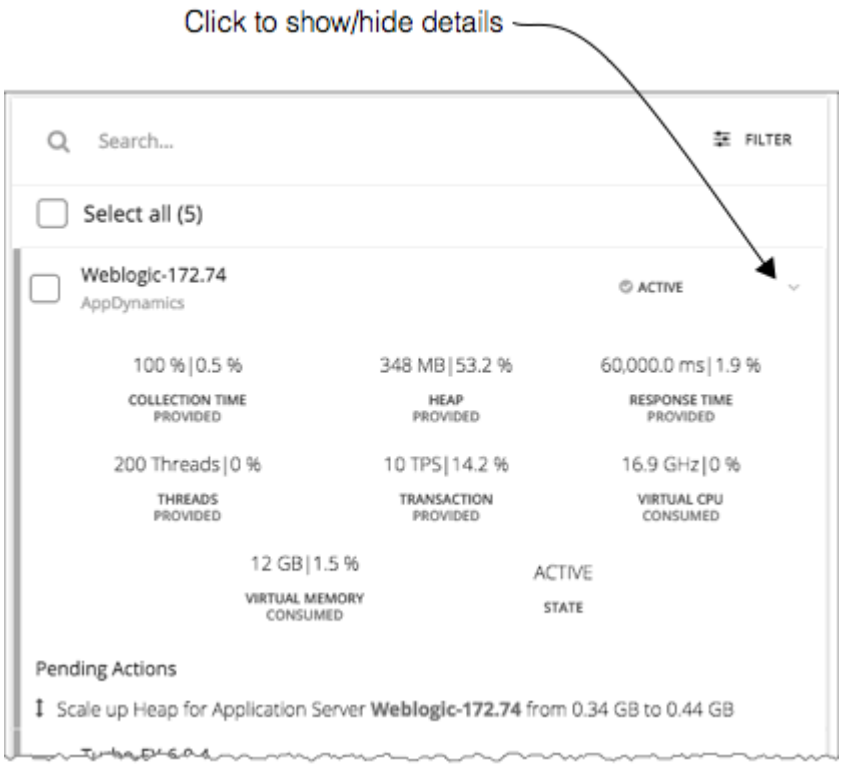
   For example, if you're showing **All** and you search for "Development", then you will see all clusters, groups, and entities with "Development" in their names.



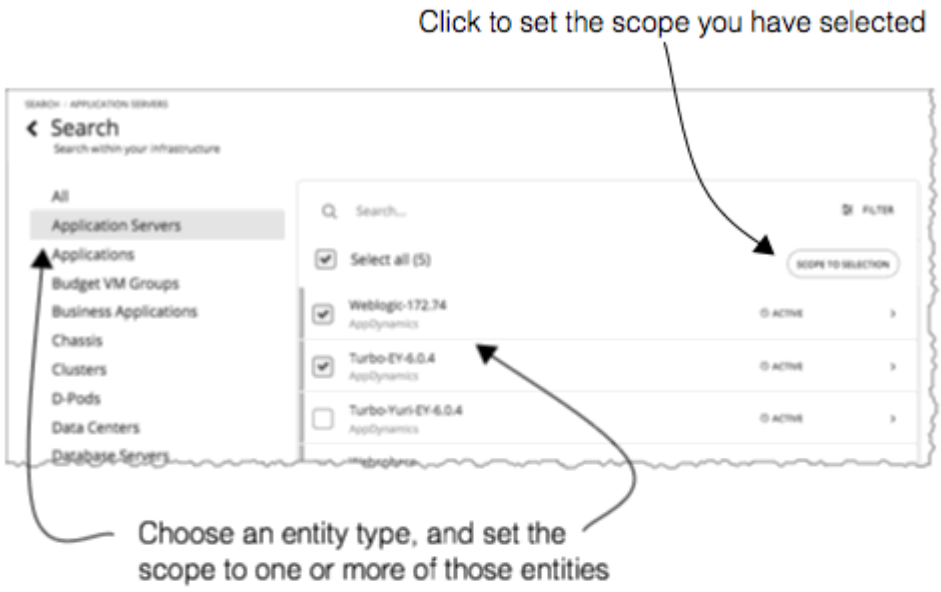
4. Expand an entry to see details.

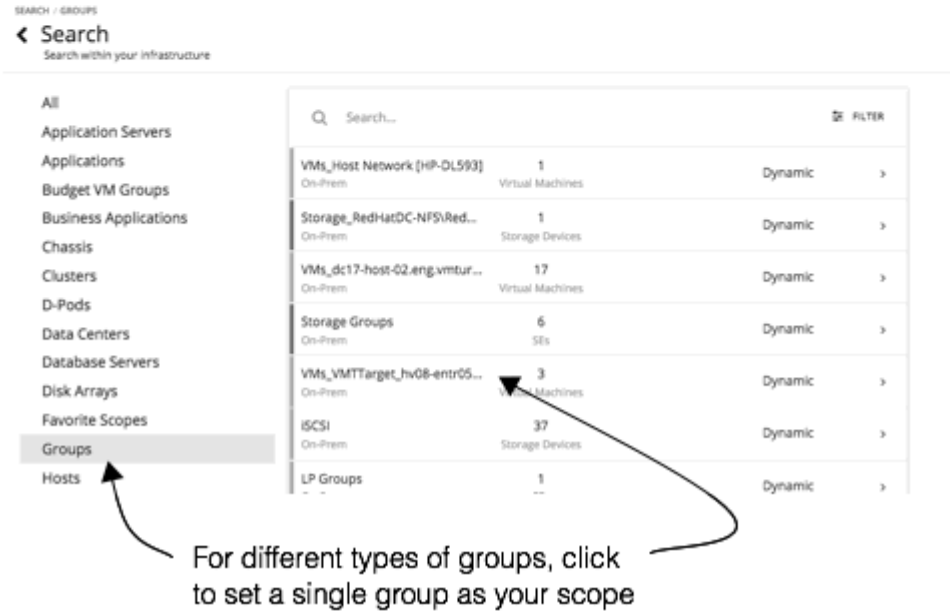
   For example, expand a group or an entity to see utilization details and pending actions.

   NOTE:
   For hosts in the public cloud, utilization and capacity for host and datacenter resources don't affect Workload Optimization Manager 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 **Home Page**.





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 **Home Page**. For example, if you select a cluster in the **Search** listing, you set the **Home Page** focus to that cluster. Use the **Home Page** 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 Workload Optimization Manager 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 Workload Optimization Manager Session *(on page 44)*
- Create new charts: See Creating and Editing Chart Widgets *(on page 448)*

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:

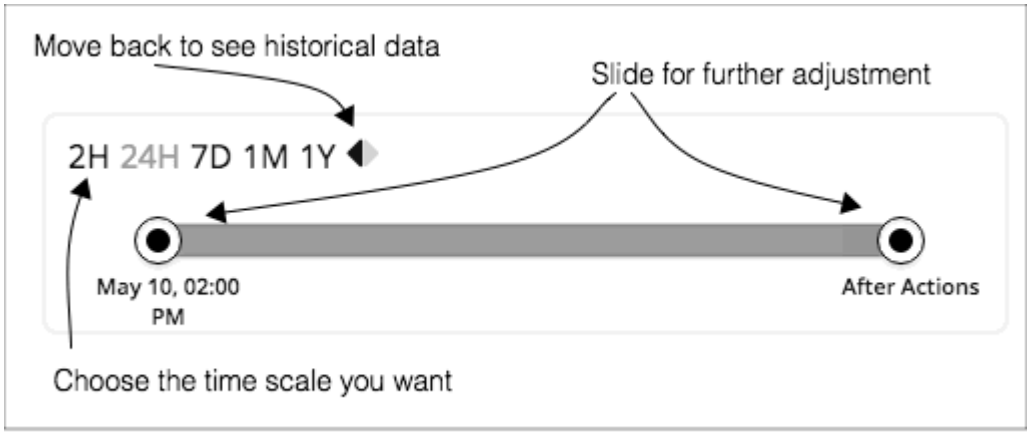
- Set Supply Chain Focus

  Choose a tier in the supply chain to set the view focus - see Navigating With the Supply Chain *(on page 56)*

- Set Scope

  Use **Search** to set the scope of the viewing session - see Scoping the Workload Optimization Manager Session *(on page 44)*

Chart Time Frame



You can set a time frame from recent hours to the past year, and set that to the charts in the view. Use the Time Slider to set specific start and end times within that range. 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 Workload Optimization Manager stores historical data in its database. As you run Workload Optimization Manager 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 Workload Optimization Manager Session *(on page 44)*
- Create new charts: See Creating and Editing Chart Widgets *(on page 448)*

Setting Chart Focus

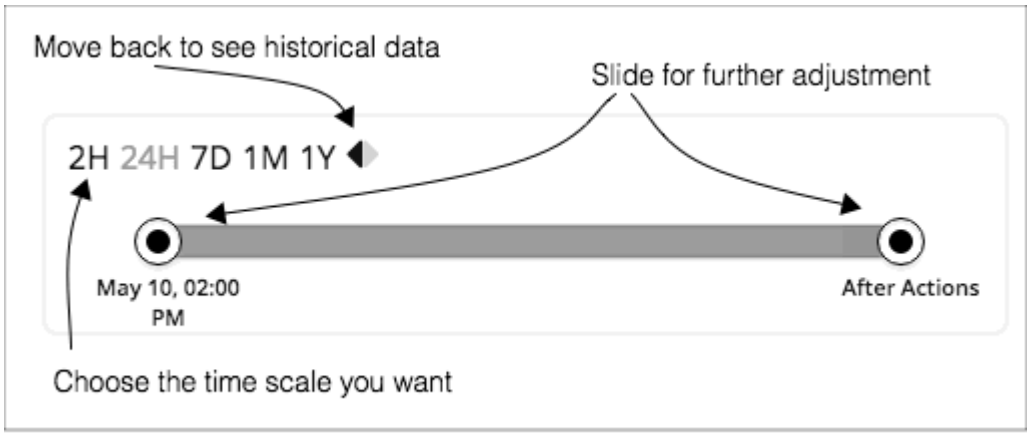
The charts update to reflect the focus that you have set for your viewing session. While viewing the Overview Charts, you can set the focus in different ways:

- Set Supply Chain Focus

  Choose a tier in the supply chain to set the view focus – see Navigating With the Supply Chain *(on page 56)*
- Set Scope

  Use **Search** to set the scope of the viewing session – see Scoping the Workload Optimization Manager Session *(on page 44)*

Chart Time Frame

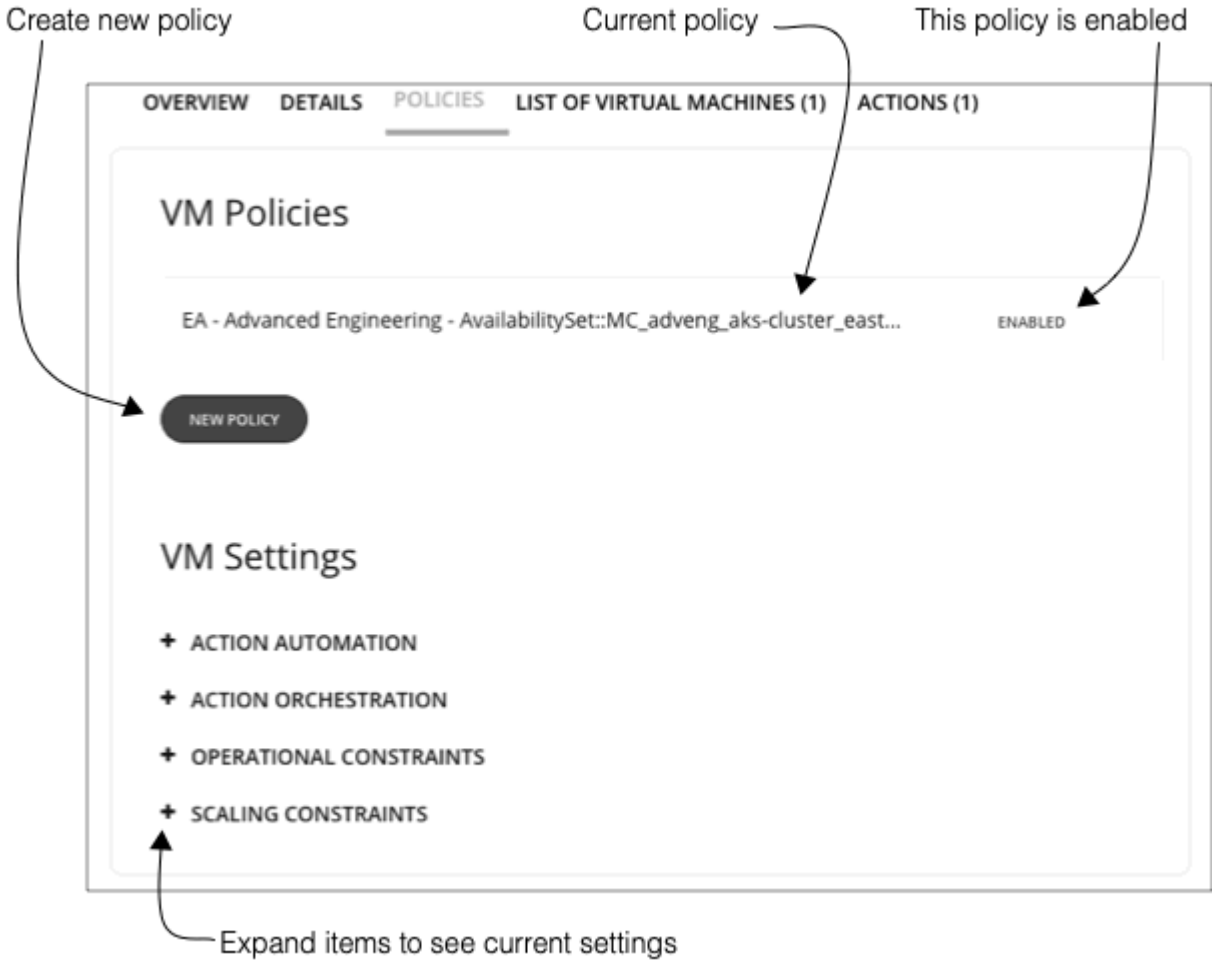


You can set a time frame from recent hours to the past year, and set that to the charts in the view. Use the Time Slider to set specific start and end times within that range. The green section in the slider shows that you can set the time range to include a projection into the future. For this part 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 Workload Optimization Manager stores historical data in its database. As you run Workload Optimization Manager 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 has been enabled or disabled. In addition, you can create new policies and apply them to that scope.

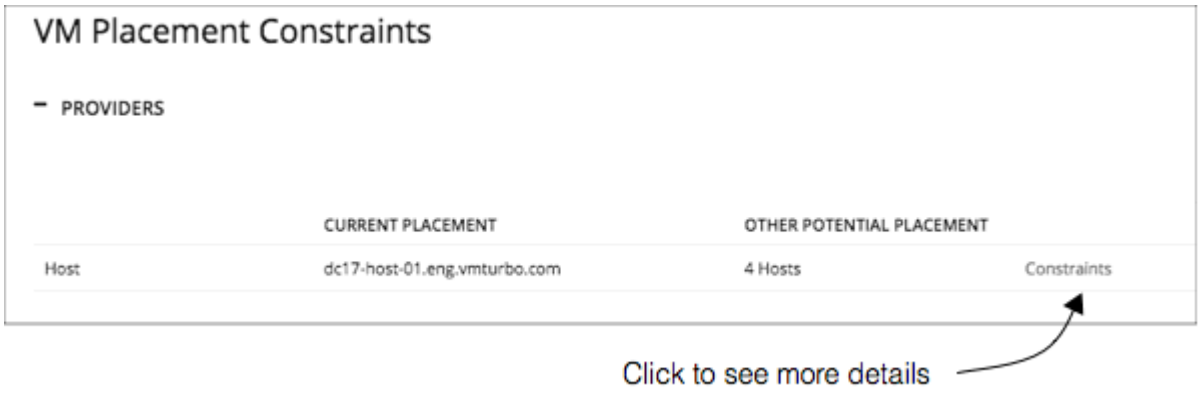
To edit a policy, click the policy name. You can then change the policy settings, or enable/disable the policy.

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 has been applied to this scope.

When you create a new policy, it automatically includes the current scope. You can add other groups to the policy scope if you like. Note that 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.

For more information, see Automation Policies *(on page 92)*.

Entity Placement Constraints



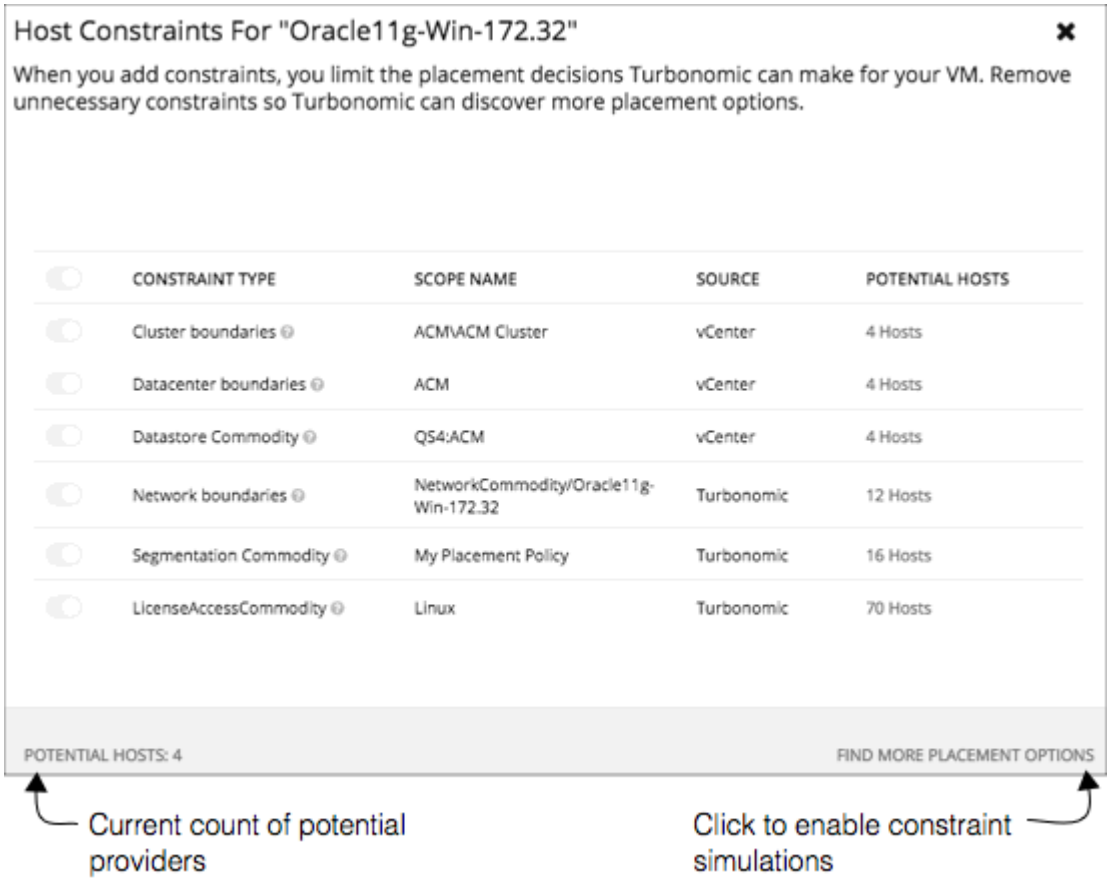
When you drill down to a single entity, you can see details about the entity's relationships in the supply chain. This shows you which entities provide resources to this entity. When considering providers for this entity, you can see the name of each current provider, and how many alternative providers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

Reviewing the constraints on an entity helps you understand the actions that Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 for this VM.



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 Workload Optimization Manager 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.

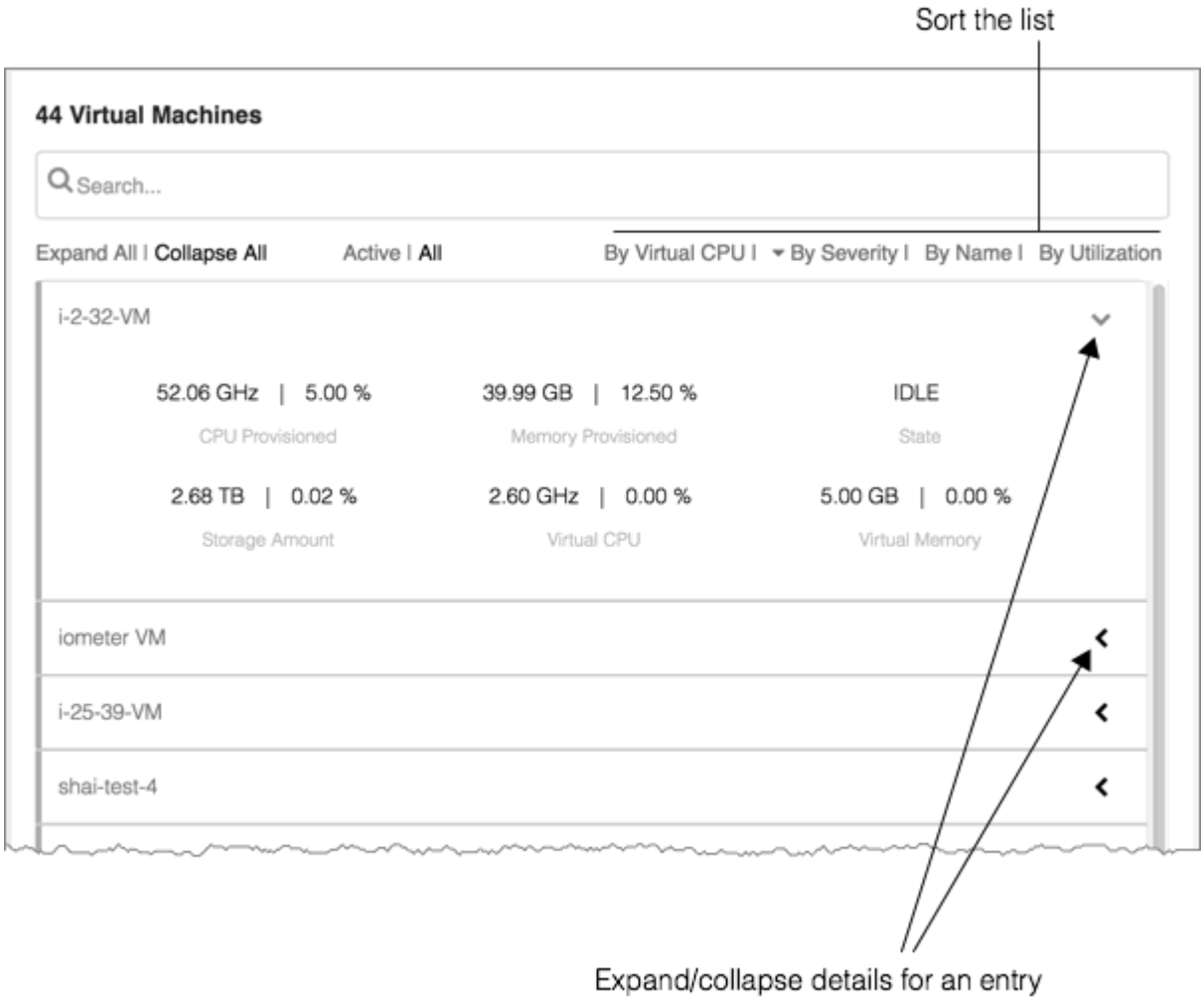
- POTENTIAL PROVIDERS

  For the given constraint, how many providers that constraint allows. To see a list of the potential providers, click the POTENTIAL PROVIDERS value.

To dig deeper into how these constraints affect your entity, click **FIND MORE PLACEMENT OPTIONS**. 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.

In this mode you can enable and disable different combinations of constraints. As you do, the **POTENTIAL PROVIDERS** label updates to show how many providers are available to your entity. To see the resulting list of providers, click the **POTENTIAL PROVIDERS** label.

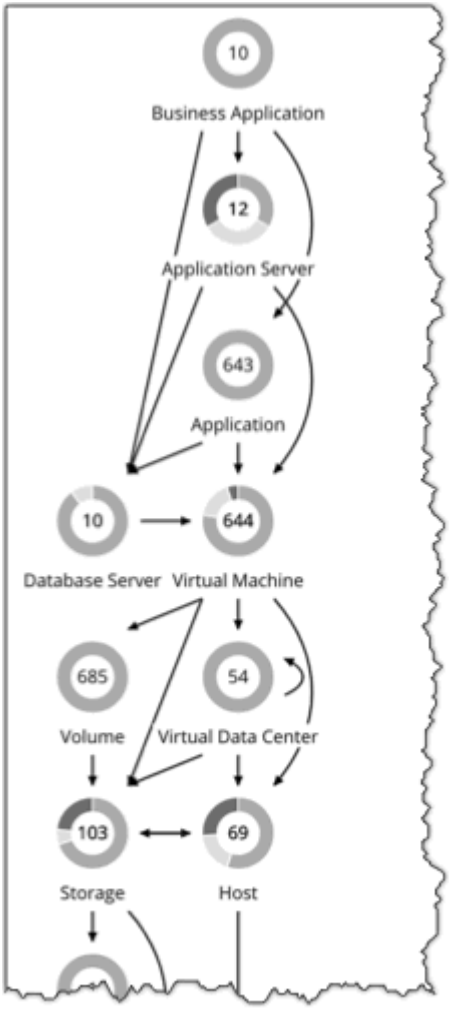
List of Entities



The list of entities is a quick way to drill down to details about your environment, so 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 56)*

Navigating With the Supply Chain



After you have set the scope of your Workload Optimization Manager 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 Workload Optimization Manager session, the **Home Page** 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. The 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 **Home Page** then includes charts that show headroom for that cluster, as well as time to exhaustion of the cluster resources.

To view cluster headroom:

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

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

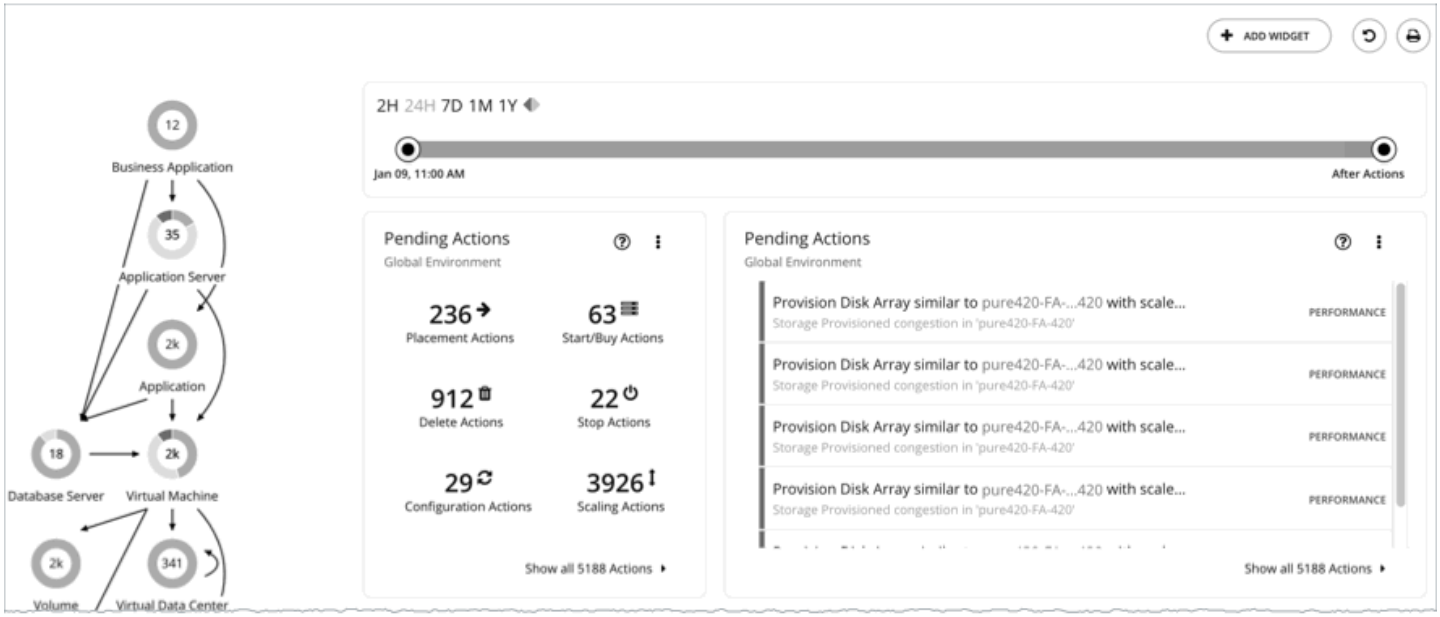
To calculate cluster capacity and headroom, Workload Optimization Manager runs nightly plans that take into account the conditions in your current environment. The plans use the Economic Scheduling Engine to identify the optimal workload distribution for your clusters. This can include moving your current 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 426)*

Workload Optimization Manager Actions

After you deploy your targets, Workload Optimization Manager starts to perform market analysis as part of its Application Resource Management process. This holistic analysis identifies problems in your environment and the actions you can take to *resolve* and *avoid* these problems. Workload Optimization Manager then generates a set of actions for that particular analysis and displays it in the Pending Actions charts for you to investigate.



Workload Optimization Manager can generate the following actions:

Action	Description
Provision	Introduce new resource providers to update the environment's capacity. For example: ▪ Provisioning a host adds more compute capacity that is available to VMs. ▪ Provisioning a VM adds capacity to run applications.
Start	Start a suspended entity to add capacity to the environment.
Resize	Re-allocate resource capacity on an entity. For example, reduce vCPUs or vMem on a VM, or add volumes to a disk array.
Increase discount coverage	Scale cloud VMs to instance types that are charged discounted rates and have existing capacity, to reduce your costs.
Buy discounts *(on page 30)*	Purchase additional discount capacity to move your environment toward the discount coverage that you desire.
Reconfigure	Reconfigure an entity that violates a policy. For example, reconfigure an on-prem VM that violates a vCPU scaling policy.

Action	Description
Move	Change a consumer to use a different provider, such as moving a VM to a different host. Moving a VM to a different storage means relocating any file-based component that belongs to a virtual machine.
Suspend	Stop and set resources aside without removing them from the environment. For example, you might consider suspending a virtual machine to save money.
Delete	Remove storage (for example, datastores on disk arrays or unattached volumes).

Actions by Entity Type

Workload Optimization Manager 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	**None** Workload Optimization Manager 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 Transaction	**None** Workload Optimization Manager 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.
Service	*For non-Kubernetes Services:* **None** Workload Optimization Manager does not recommend actions for non-Kubernetes Services, 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. *For Kubernetes Services:* **Provision or Suspend** For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications. For details, see Actions for Kubernetes Services *(on page 131)*.

Entity Type	Supported Actions
Application Component	**Resize** Resize the following resources to maintain performance: ▪ Thread Pool Workload Optimization Manager generates thread pool resize actions. These actions are recommend-only and can only be executed outside Workload Optimization Manager. ▪ Connections Workload Optimization Manager uses connection data to generate memory resize actions for on-prem Database Servers. ▪ Heap Workload Optimization Manager 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 Workload Optimization Manager. **NOTE:** Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC). The resources that Workload Optimization Manager can resize depend on the processes that it discovers from your Applications and Databases targets. Refer to the topic for a specific target to see a list of resources that can be resized.

Container Platform Entity Types

Entity Type	Supported Actions
Container	**Resize** Resize containers to assure optimal utilization of resources. By default, containers resize consistently, which allows all replicas of the same container for the same workload type to resize any resource consistently. For details, see Container Actions *(on page 153)*.
Container Spec	**None** A Container Spec retains the historical utilization data of ephemeral containers. Workload Optimization Manager uses this data to make accurate container resize decisions, but does not recommend actions for the Container Spec itself.
Namespace	**Resize Quota** Workload Optimization Manager treats quotas defined in a namespace as constraints when making container resize decisions. If existing container actions would exceed the namespace quotas, Workload Optimization Manager recommends actions to resize up the affected namespace quota. Note that Workload Optimization Manager does not recommend actions to resize *down* a namespace quota. Such an action reduces the capacity that is already allocat-

Entity Type	Supported Actions
	ed to an application – The decision to resize down a namespace quota should include the application owner.
Workload Controller	**None** A Workload Controller executes container actions. When you set the scope to a Workload Controller and view the actions list, the actions apply to containers. Workload Optimization Manager does not recommend actions for the Workload Controller itself. **NOTE:** Workload Optimization Manager uses namespace or organization/space quotas as constraints when making resize decisions. The Workload Controller aggregates container actions. If those container resizes exceed current namespace quotas, Workload Optimization Manager blocks execution of container resize actions until the namespace quotas are sufficient. For more information about namespace quotas, see Resource Quotas *(on page 173)*.
Container Pod	▪ **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** For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, provision or suspend pods associated with those Services to maintain SLOs for your applications. When recommending node provision or suspend actions, Workload Optimization Manager will also recommend provisioning pods (based on demand from DaemonSets) or suspending the related pods. For details, see Container Pod Actions *(on page 167)*.
Container Cluster	**None** Workload Optimization Manager does not recommend actions for a Container Cluster. Instead, it recommends actions for the containers, pods, nodes (VMs), and volumes in the cluster. Workload Optimization Manager shows all of these actions when you scope to a Container Cluster and view the Pending Actions chart.
Kubernetes node (VM)	A Kubernetes 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.

Entity Type	Supported Actions
	NOTE: For nodes in the public cloud, Workload Optimization Manager reports the cost savings or investments attached to these actions. For details, see Node Actions *(on page 182)*.

Cloud Infrastructure Entity Types

Entity Type	Supported Actions
Virtual Machine (Cloud)	▪ **Scale** Change the VM instance to use a different instance type or tier to optimize performance and costs. ▪ **Stop** and **Start** (also known as 'parking' actions) Stop a VM for a given period of time to reduce your cloud expenses, and then start it at a later time. ▪ **Discount-related actions** If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing discount coverage. To increase coverage, you scale VMs to instance types that have existing capacity. If you need more capacity, then Workload Optimization Manager will recommend actions to purchase additional discounts. For details, see Cloud VM Actions *(on page 193)* and Parking: Stop or Start Cloud Resources *(on page 428)*.
Virtual Machine Spec	▪ **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 218)*.
App Component Spec	**None** Workload Optimization Manager 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 218)*.
Database (Cloud)	**Scale** ▪ DTU Model Scale DTU and storage resources to optimize performance and costs. ▪ vCore Model Scale vCPU, vMem, IOPS, throughput and storage resources to optimize performance and costs. For details, see Cloud Database Actions *(on page 253)*.

Entity Type	Supported Actions
Database Server (Cloud)	**Scale** Scale compute and storage resources to optimize performance and costs. For details, see Cloud Database Server Actions *(on page 230)*.
Volume (Cloud)	▪ **Scale** Scale attached volumes to optimize performance and costs. ▪ **Delete** Delete unattached volumes as a cost-saving measure. For details, see Cloud Volume Actions *(on page 243)*.
Zone	**None** Workload Optimization Manager does not recommend actions for a cloud zone.
Region	**None** Workload Optimization Manager does not recommend actions for a cloud region.

On-prem Infrastructure Entity Types

Entity Type	Supported Actions
Virtual Machine (On-prem)	▪ **Resize** – 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, Workload Optimization Manager verifies that the VM's cluster can adequately support the new size. If the cluster is highly utilized, Workload Optimization Manager will recommend a move action, taking into consideration the capacity of the new cluster and compliance with existing placement policies. For hypervisor targets, Workload Optimization Manager can resize vCPU by changing the VM's socket or cores per socket count. For details, see VCPU Scaling Controls *(on page 278)*. – 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. ▪ **Move**

Entity Type	Supported Actions
	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, Workload Optimization Manager can reconfigure VMs that violate vCPU scaling policies. For details, see VCPU Scaling Controls *(on page 278)*. For details, see On-prem VM Actions *(on page 269)*.
Volume (On-prem)	▪ **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 Workload Optimization Manager to use relevant metrics when analyzing and recommending actions for volumes. For details, see Enable Analysis of On-prem Volumes *(on page 95)*. – Workload Optimization Manager 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.
Database Server (On-prem)	**Resize** Resize the following resources: ▪ Connections Workload Optimization Manager 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. Workload Optimization Manager 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

Entity Type	Supported Actions
	Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology. Currently, Workload Optimization Manager 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 Datacenter	**None** Workload Optimization Manager does not recommend actions for a Virtual Datacenter. Instead, it recommends actions for the entities that provide resources to the Virtual Datacenter.
Business User	**Move** Move a Business User between desktop pools to address: ▪ Resource congestion on the image When utilization is consistently near capacity for image resources, Workload Optimization Manager can recommend moving a Business User to a desktop pool that serves larger images. ▪ Resource congestion on the desktop pool When utilization is consistently near capacity for the desktop pool, Workload Optimization Manager can recommend moving a Business User to a desktop pool that has more available resources. **NOTE:** To support moves, you must configure placement policies that merge *similarly configured* desktop pools. For details, see Desktop Pool Placement Policies *(on page 308)*.
Desktop Pool	**None** Workload Optimization Manager does not recommend actions for a desktop pool. It recommends actions for the Business Users running active sessions in the pool.
View Pod	**None** Workload Optimization Manager does not recommend actions for a view pod. Instead, it recommends actions for the Business Users that are running active sessions.
Host	▪ **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. Workload Optimization Manager 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.

Entity Type	Supported Actions
	▪ **Reconfigure** Workload Optimization Manager generates this action in response to changing demand for software licenses. For details, see License Policy *(on page 91)*. For details, see Host Actions *(on page 312)*.
Chassis	**None** Workload Optimization Manager does not recommend actions for a chassis.
Datacenter	**None** Workload Optimization Manager does not recommend actions for a datacenter. Instead, it recommends actions for the entities running in the datacenter.
Storage	▪ **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** Delete a datastore or volume that has been suspended for a period of time. For details, see Storage Actions *(on page 321)*.
Logical Pool	▪ Resize ▪ Provision ▪ Move ▪ Start ▪ Suspend
Disk Array	▪ **Provision** For high utilization of the disk array's storage, provision a new disk array (recommendation, only). ▪ **Start** For high utilization of disk array, start a suspended disk array (recommendation, only). ▪ **Suspend** For low utilization of the disk array's storage, move VMs to other datastores and suspend volumes on the disk array (recommendation, only). ▪ **Move**

Entity Type	Supported Actions
	(Only for NetApp Cluster-Mode) For high utilization of Storage Controller resources, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager can move a datastore to another array.
Storage Controller	**Provision** For high utilization of the storage controller's CPU, provision a new storage controller, and then move disk arrays to it.
IO Module	**None** Workload Optimization Manager does not recommend actions for an IO Module.
Switch	**Resize** Resize PortChannel for a switch to increase bandwidth.

Action Types

Workload Optimization Manager performs the following general types of actions:

- Placement – Place a consumer on a specific provider
- Scaling – Resize allocation of resources, based on profitability
  - Resize up, shown as a required investment
  - Resize down, shown as savings
- Discount Optimization – Increase discount *(on page 30)* coverage and reduce costs by scaling VMs to instance types that are charged discounted rates
- Configuration – Correct a misconfiguration
- Start/Buy – Start a new instance to add capacity to the environment, shown as a required investment. For cloud environments, purchase discounts *(on page 30)* to reduce costs.
- Stop – Suspend an instance to increase efficient use of resources, shown as savings
- Delete – Remove storage (for example, datastores on disk arrays or unattached volumes).

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, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager from recommending a more obvious action. For details, see Entity Placement Constraints *(on page 52)*.

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, Workload Optimization Manager 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, Workload Optimization Manager uses benchmark data from spec.org. This benchmark data maps to effective capacity settings that Workload Optimization Manager 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 523)*.

Shared-Nothing Migration Actions

If you have enabled both storage and VM moves, Workload Optimization Manager can perform shared-nothing migrations, which move the VM and the stored VM files simultaneously. For details, see Shared-Nothing Migration *(on page 274)*.

Cross-vCenter vMotion

VMware vSphere 6.0 introduces functionality that enables migration of virtual machines between different vCenter Server instances. Workload Optimization Manager supports this capability through *Merge* placement policies (see Creating Placement Policies *(on page 88)*). 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 Workload Optimization Manager Supply Chain, the Host nodes represent availability zones. Workload Optimization Manager 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 30)* 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 Workload Optimization Manager can recommend a VM move. However, because Instance Storage is ephemeral and a move would lose the stored data, Workload Optimization Manager does not recommend moving a VM that has Instance Storage as its root storage.

If a VM is running within a billing family, then Workload Optimization Manager only recommends moving that VM to other regions within that billing family.

In AWS environments that use RIs, Workload Optimization Manager recognizes Availability Zones that you have specified for your RI purchases. For move and resize actions, Workload Optimization Manager 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, Workload Optimization Manager will use the Zone RI first.

Scaling

Scaling actions update capacity in your environment. For vertical scaling, Workload Optimization Manager 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.

Workload Optimization Manager can provision the following:

- Containers
- VMs
- Hosts
- Storage
- Storage Controllers (only for planning scenarios)
- Disk Arrays

Under certain circumstances, Workload Optimization Manager can also recommend that you provision a virtual datacenter.

Storage Resize Actions

Any storage resize action impacts both the storage entities and the entities managed by the given hypervisor. However, not all hypervisors recognize changes to the storage capacity. After executing a storage resize, Workload Optimization Manager indicates that the resize action has succeeded but a hypervisor might not show the corresponding change in storage capacity. If this occurs, then you must refresh the hypervisor target so Workload Optimization Manager can discover the storage changes.

To avoid this situation, you can set the action mode to *Manual* or *Recommend* for storage resize actions. In that way, you can perform the resizes yourself, and then manually refresh your hypervisors.

Scaling on the Public Cloud

On the cloud, scaling actions change the VM to a different instance type. These can include:

- Changing a VM to an instance type with different capacity
- Changing a VM to an instance type that is charged a discounted rate

For these actions, the action list shows the current cost for the source workload, and also the projected cost given the change. To show the current cost, Workload Optimization Manager uses the actual costs for that workload. However, to show the projected cost it uses an estimate based on average utilization for the VM, for the costs of the given tier.

Note that scaling to an instance type that is charged a discounted rate can result in running the VM on a larger instance when the cost is lower. This might occur even though the VM does not need that capacity and there are other smaller instance types available.

In Azure environments, there are circumstances where a VM resize can be especially disruptive. In a given region, the infrastructure can be made up of different clusters that have different sets of underlying hardware. Further, some tiers that are available in the given region are only available on different clusters. If Workload Optimization Manager recommends resizing from a tier on one cluster, to a tier on another cluster, then the resize action can take longer to complete than usual.

In both Azure and AWS environments, Workload Optimization Manager conforms to specific instance requirements as it generates resize actions. For more information, see:

- Azure Instance Requirements *(on page 196)*
- AWS Instance Requirements *(on page 194)*

Discount Optimization

To reduce your cloud costs, Workload Optimization Manager can recommend scaling VMs to instance types that are charged discounted rates.

- Discount Utilization *(on page 502)*

  This chart shows how well you have utilized your current discount inventory *(on page 499)*. 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 497)*

  This chart shows the percentage of VMs covered by discounts. If you have a high percentage of on-demand VMs, you should be able to reduce your monthly costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

- Discount Inventory *(on page 499)*

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

Discount optimization actions are not executed by Workload Optimization Manager 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.

Workload Optimization Manager can reconfigure the following:

- VMs
- Containers
- Storage
- Disk Arrays
- Virtual Datacenters

Start/Buy

Workload Optimization Manager can recommend that you start a suspended entity to add capacity to the environment. It can also recommend purchasing cloud provider discounts *(on page 30)* 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.

Workload Optimization Manager can suspend the following:

- Application Components
- Container Pods
- Disk Arrays
- Hosts

- Storage (on-prem)
- Virtual datacenter

Delete

Delete actions affect storage. For example, Workload Optimization Manager might recommend that you delete wasted files to free up storage space, or delete unused storage in your cloud environment to reduce storage costs.

Wasted Storage in Azure Environments

In Azure environments, Workload Optimization Manager can identify unmanaged storage as unattached volumes, recommend that you remove this unused storage, and then show *estimated* savings after you remove this storage and no longer pay for it. The savings that Workload Optimization Manager shows are estimates based on the overall cost for that storage, since Azure does not provide specific values for the cost per volume or cost for the amount of storage that is in use for a given volume. If the estimated savings appear unusually high, then you should identify which storage the actions will remove, and review your billing to calculate the costs with more precision.

Action Categories

Workload Optimization Manager 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.

Performance Assurance

Ultimately, the reason to manage workloads in your environment is to assure performance and meet QoS goals. When Workload Optimization Manager detects conditions that directly put QoS at risk, it recommends associated actions in the Performance category. You can consider these critical conditions, and you should execute the recommended actions as soon as possible.

Actions	Risks/Opportunities
▪ Provision a new VM, Host, Datastore ▪ Increase or decrease the number of VCPUs ▪ Provision a new container or container pod ▪ Resize heap for an Application Component ▪ Scale the resource capacity on an entity	▪ <Resource> Congestion High utilization of application resources. High utilization of resources on workload, host, or datastore.

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 Workload Optimization Manager discovers underutilized resources, it recommends actions to consolidate your operations. For example, it can recommend that you move certain VMs onto a different host. This can free a physical machine to be shut down.

Actions	Risks/Opportunities
▪ Move VM ▪ Start or suspend VM ▪ Buy discounts *(on page 30)* ▪ Scale down resource allocation	▪ Overprovisioning Excess resource capacity in a provider.

Prevention

Workload Optimization Manager 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 associated actions in the Prevention category. 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 the QoS for some services may be put at risk.

Actions	Risks/Opportunities
▪ Resize vCPU and vMem ▪ Move VM or storage ▪ Start VM or host	▪ <Resource> Congestion High resource utilization on the named VM or datastore. For example, CPU congestion or memory congestion can occur on a VM, or an IOPS bottleneck can occur on a datastore. ▪ Workload Balancing Excess workload on a given physical machine that can be addressed by moving a VM to another host.

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, Workload Optimization Manager identifies the violation and recommends actions that bring the entity back into compliance.

Actions	Risks/Opportunities
▪ Move VM ▪ Move container ▪ Provision VM, Host, Datastore	▪ Misconfiguration Container configuration is in violation of a policy. ▪ Placement Violation The placement of a VM is in violation of a Workload Optimization Manager policy or an imported Placement Policy. ▪ Misconfiguration The configuration violates discovered requirements. For example, a VM is configured to access a network that is not available from the current cluster.

Action Modes

Action 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 modes in a policy to set that business rule.

Workload Optimization Manager supports the following action modes:

- Recommend – Recommend the action so a user can execute it via the given hypervisor or by other means
- Manual – Recommend the action, and provide the option to execute that action through the Workload Optimization Manager user interface
- Automatic – Execute the action automatically

For automated resize or move actions on the same entity, Workload Optimization Manager 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, Workload Optimization Manager could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.

The Pending Actions charts only count actions in *Recommend* or *Manual* mode.

Automated actions appear in the following charts:

- **All Actions** chart on the **Home Page** and the On-prem Executive Dashboard
- **Accepted Actions** chart on the **Home Page**

Setting Action Modes

To set action modes for specific entities, you can edit the Workload Optimization Manager automation policies. This is how you specify the default action modes, or set special action modes for a given group or cluster. For more information, see Automation Policies *(on page 92)*.

Action Orchestration

Workload Optimization Manager policies can also include Action Orchestration settings. These settings determine whether Workload Optimization Manager executes the actions, or whether to map the actions to workflows managed by external orchestrators. If you want to execute via an orchestrator workflow, you must set the action mode to *Manual* or *Automatic*. For more information about action orchestration, see Setting Up Action Orchestration *(on page 107)*.

Action Mode Overrides

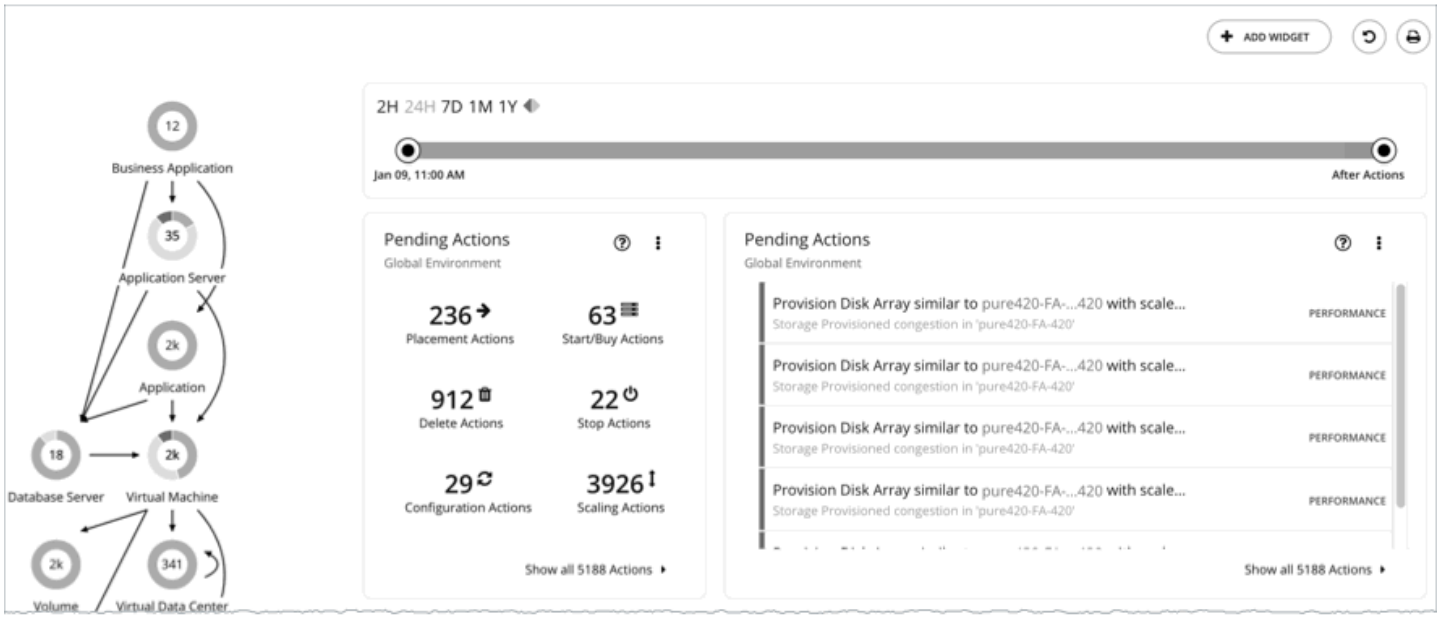
Under some conditions, Workload Optimization Manager changes the action mode of an action from *Manual* to *Recommend*.

Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager changes actions to move workloads *to* the host from *Manual* to *Recommend*.

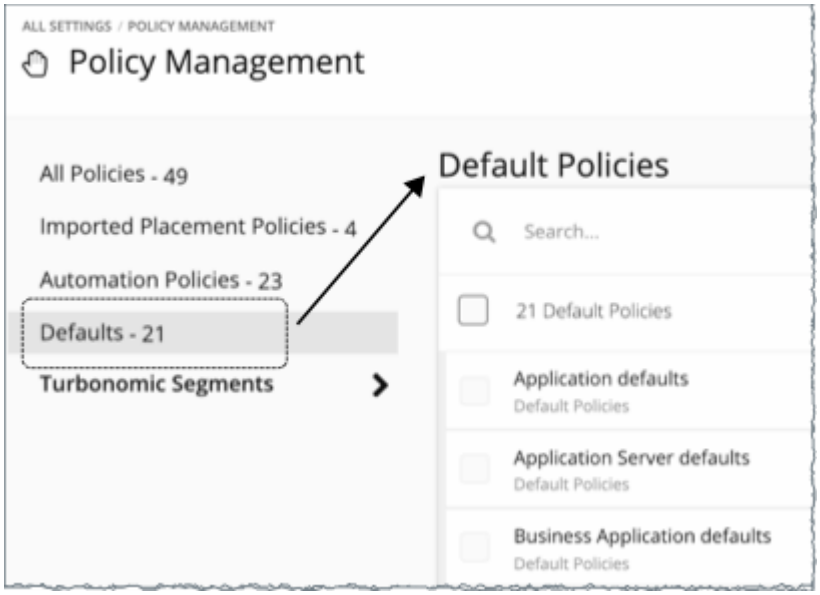
For cloud environments, some instances require workloads to be configured in specific ways before they can move to those instance types. If Workload Optimization Manager recommends moving a workload that is not suitably configured onto one of these instances, then it changes the action mode from *Manual* to *Recommend*, and then describes the reason.

Working With the Generated Actions

When you start using Workload Optimization Manager, all the actions that the product generates appear as pending. You can view them in the Pending Actions charts and then decide whether to execute and/or automate them. You can also disable them.



Workload Optimization Manager 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. Workload Optimization Manager gives you full control over all automation decisions.



When you first see the pending actions, you execute many of them to see immediate improvements in performance and 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

  Workload Optimization Manager will not consider recommending disabled actions when it performs its analysis.
- Allowing certain actions to execute automatically, such as those that assure QoS 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 Workload Optimization Manager 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 Workload Optimization Manager to post those actions for review and only execute the actions that receive an approval.

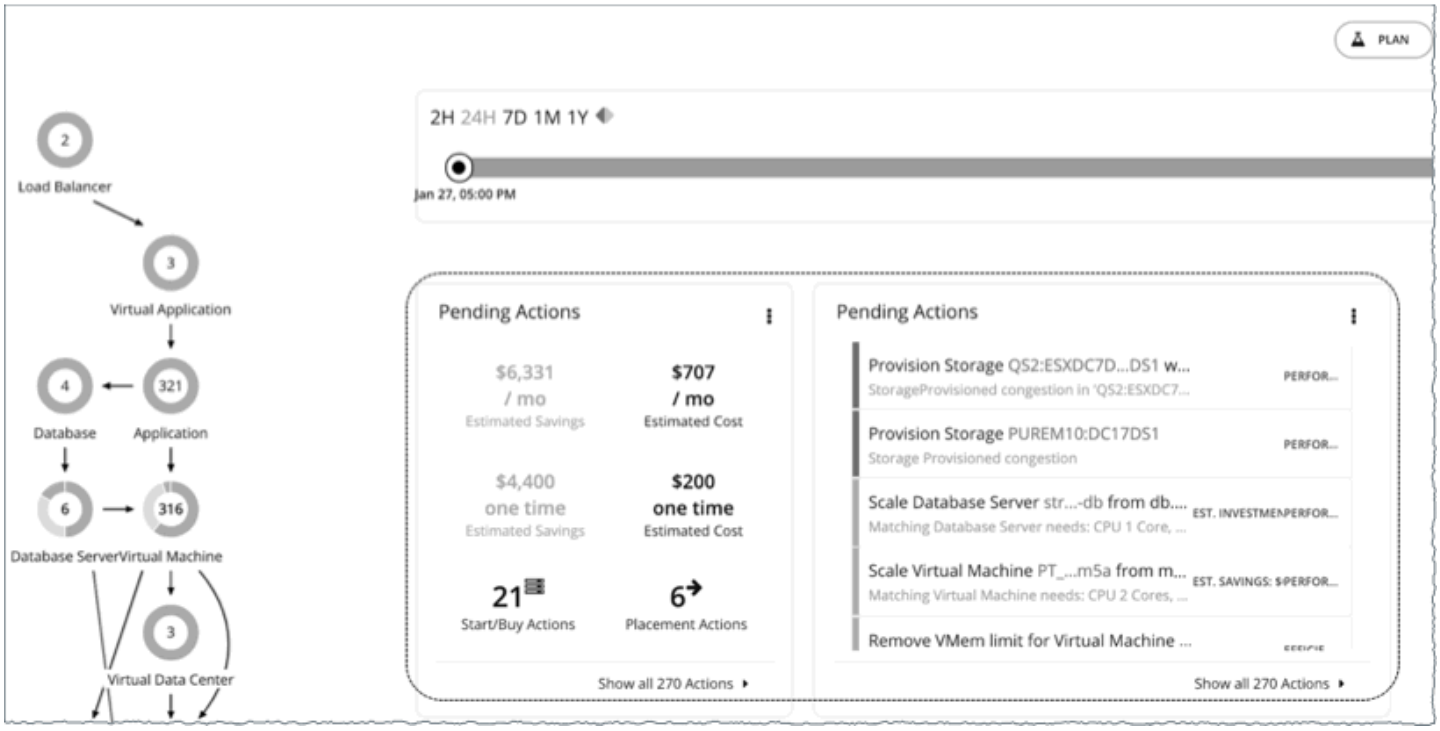
These are the actions that you would look for in the Pending Actions charts. They no longer show after you execute them, if you disable or automate them, or if the environment changes in the next market analysis such that the actions are no longer needed.

What You Can Do:

- View and execute pending actions: See Pending Actions *(on page 75)*.
- See the different display views for the pending actions charts: See Pending Actions Charts *(on page 452)*.
- Scope pending actions in the **Home Page**: See Pending Actions Scope *(on page 77)*.
- See a running history of generated and executed actions: See Actions Charts *(on page 454)*.
- Review the default policies that drive the actions the product generates.
- Create and run plans to simulate different conditions, and see what actions will keep things healthy under those conditions: See Plan Management *(on page 343)*.

Pending Actions

Workload Optimization Manager treats all the non-automated actions that it generates as pending and shows them in the Pending Actions charts.



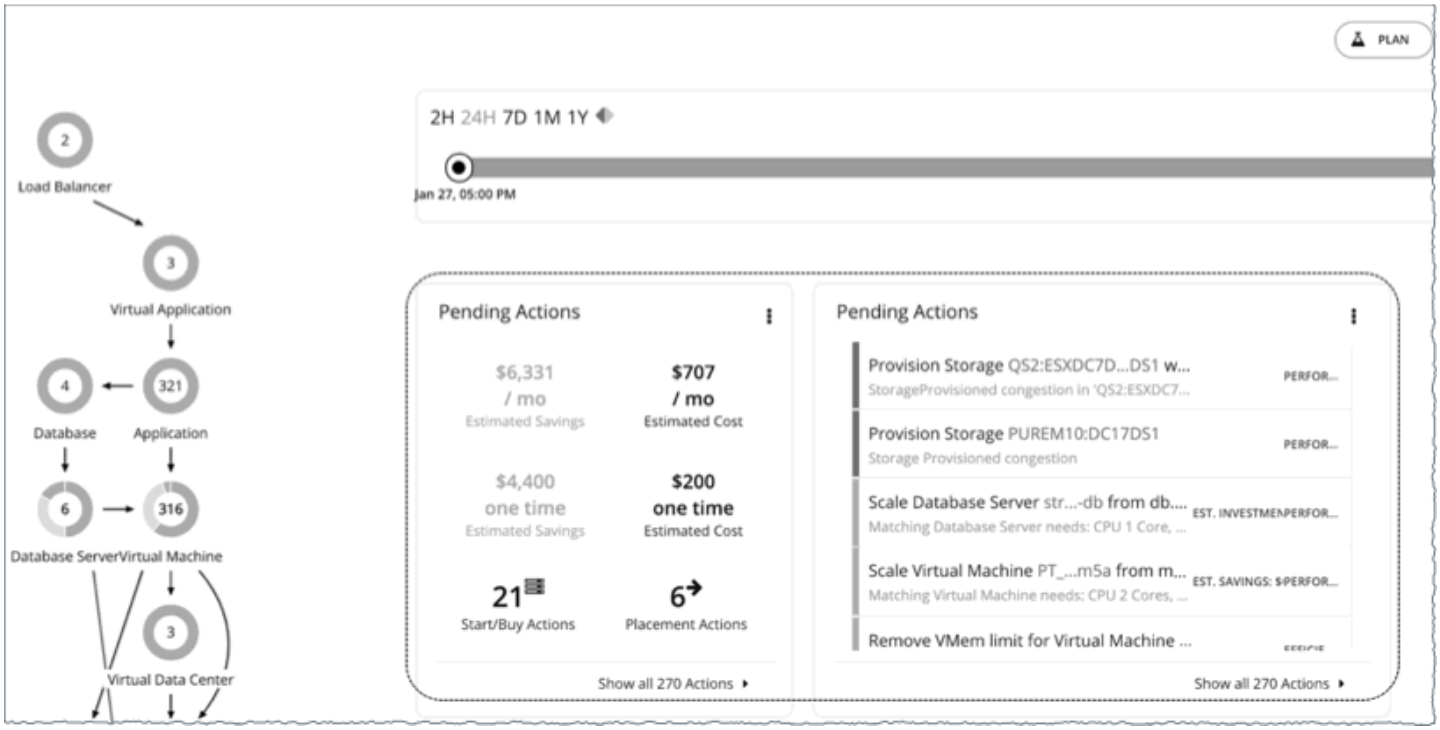
To get the best results from Workload Optimization Manager, execute these actions promptly and consider automating as many of them as possible. You can execute these actions from the user interface or outside Workload Optimization Manager. To automate these actions, create an automation policy *(on page 97)* or change the action mode to *Automatic* in the default policies *(on page 93)*.

Workload Optimization Manager can execute up to five actions at a time, and queues any new incoming actions for later execution.

Default Pending Actions Charts

Each time you log in to the user interface, Workload Optimization Manager immediately shows the Pending Actions charts on the **Home Page**'s **HYBRID** view. These charts provide a summary of the actions that require your attention, and entry points to the Pending Actions List *(on page 79)*.

NOTE:
You can also add these charts to any of your dashboards *(on page 440)*.

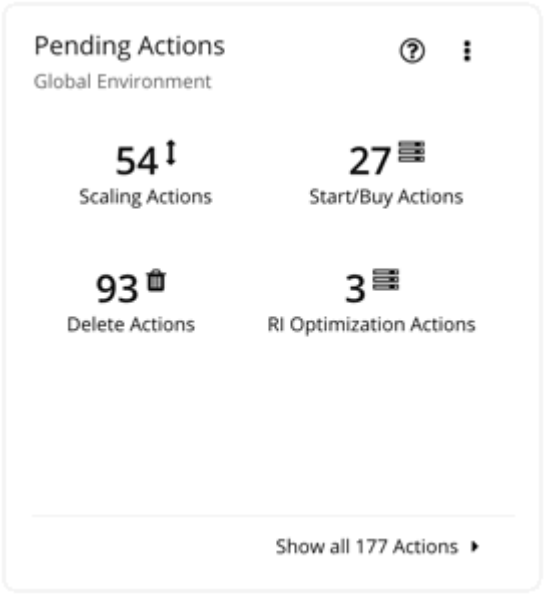


By default, a text chart and a list chart display in the **Home Page**, with the scope set to *Global Environment*.

You can change the chart type by clicking the icon on the upper-right corner of the chart. For details about the available chart types, see Pending Actions Charts *(on page 452)*.

Pending Actions - Text Chart

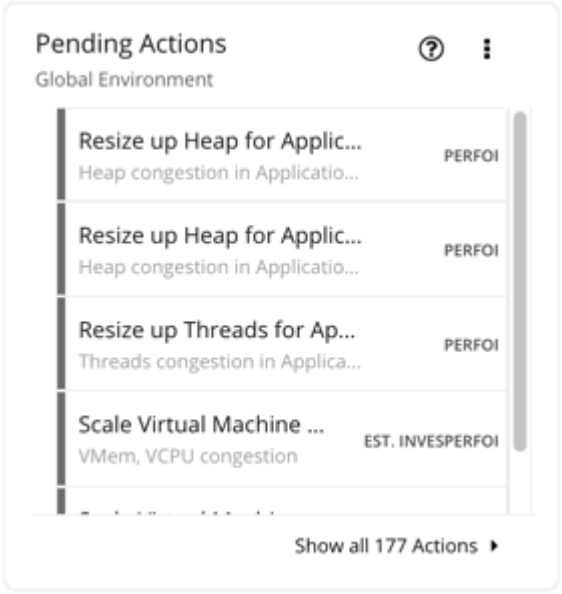
The **text chart** shows the estimated costs or savings associated with the pending actions, and the number of actions for each action type *(on page 67)*.



NOTE:
The text chart is also available in the **ON-PREM** or **CLOUD** view, with data scoped to the selected environment.

Pending Actions - List Chart

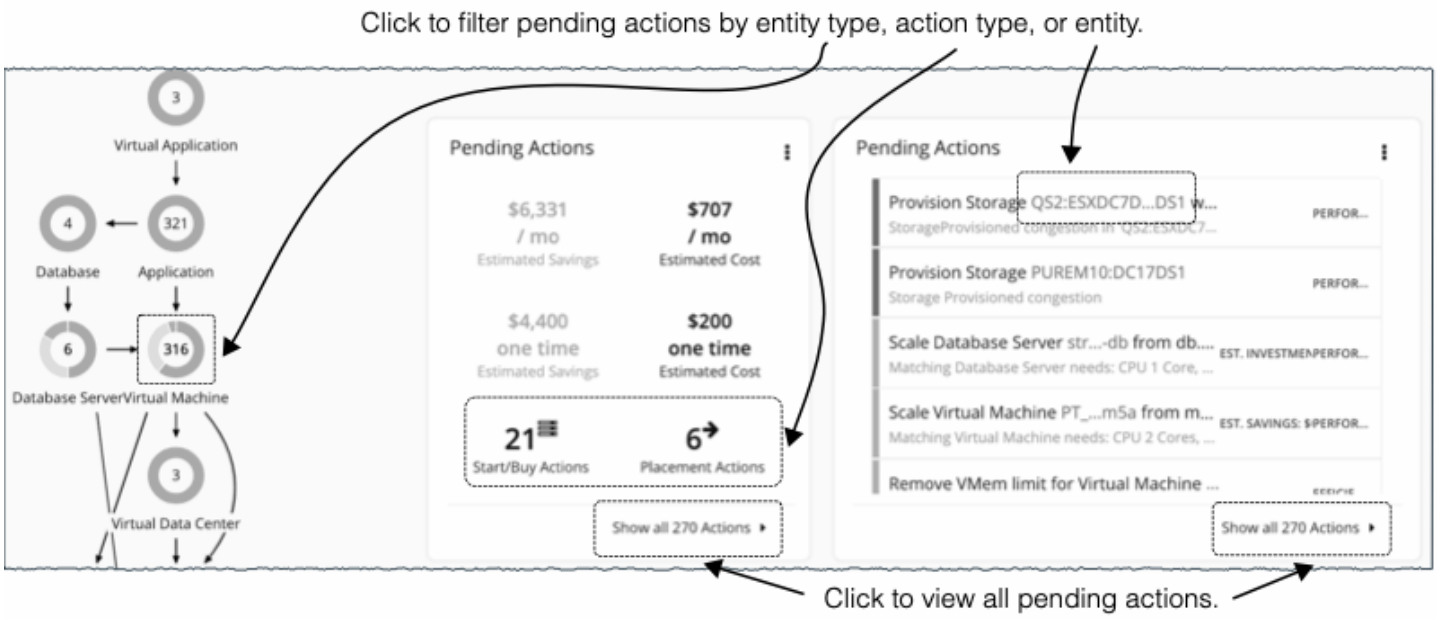
The **list chart** shows a partial list of pending actions, ordered by the severity of the associated problems.



Pending Actions Scope

To perform Application Resource Management, Workload Optimization Manager identifies actions you can take to *avoid* problems before they occur. You can perform these actions manually, direct Workload Optimization Manager to perform the actions on command, or direct Workload Optimization Manager to perform actions automatically as they arise.

There are several ways to scope pending actions in the **Home Page**.



To view all pending actions, click **Show all Actions** in the Pending Actions chart.

Click one of the following to narrow the scope of pending actions:

- An entity type in the supply chain.

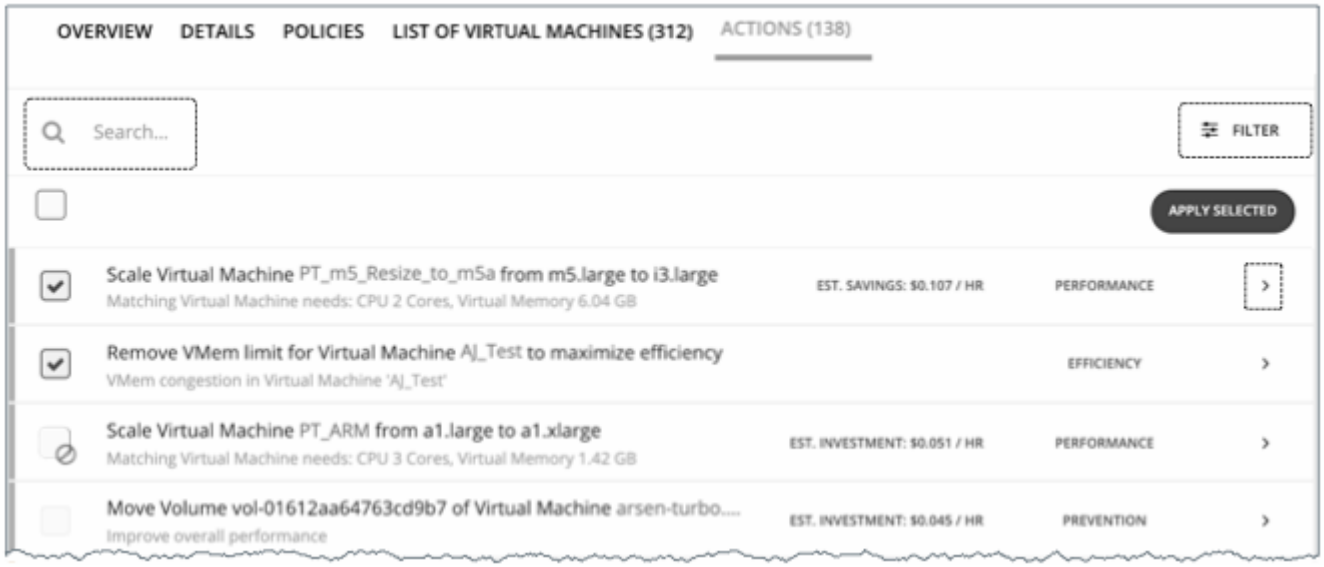
  Workload Optimization Manager generates actions based on how entity types use or provide resources, and what each entity type supports. For details on the actions that each entity type supports, see Actions by Entity Type *(on page 59)*.

  Only entity types with risks (critical, major, or minor) have pending actions. Hover on the entity type to see a breakdown of risks.
- An action type in the text chart
- An entity name in the list chart

> **NOTE:**
> If you are in the **ON-PREM** or **CLOUD** view, the text chart displays by default. Switch to the list chart to see the entity names.

If you clicked **Show all Actions** or an action type, the Pending Actions List *(on page 79)* displays immediately.

If you clicked an entity type or an entity name, an Overview page displays first. In that page, click the **Actions** tab to view the Pending Actions List.

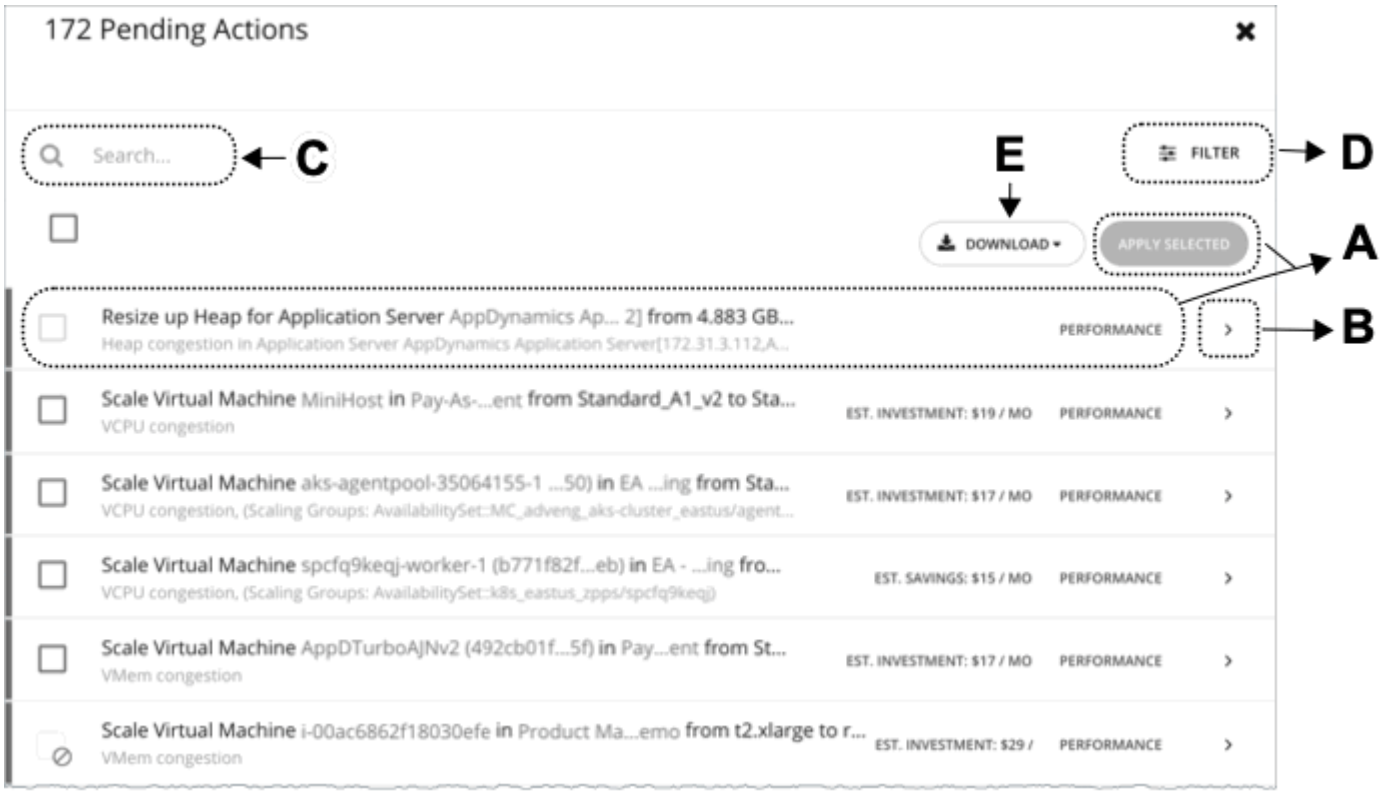


The Pending Actions List includes additional features to narrow the scope further. You can search for specific actions using meaningful keywords or use filters. For details, see Pending Actions List *(on page 79)*.

Pending Actions List

The Pending Actions List includes all the actions that Workload Optimization Manager currently recommends for the given scope (for details, see Pending Actions Scope *(on page 77)*).

You can select actions to execute, and you can expand action items to see more details.



A. Actions List

Each row in the actions list shows:

- The specific action that Workload Optimization Manager recommends.
- If applicable, the estimated investment needed to successfully execute the action or the resulting savings after performing the action
- The action category *(on page 71)*.

By default, actions display by the severity of the associated problems, indicated by the thin colored line before the checkbox. Use the Filter functionality to change the order by other categories.

Select one or several actions to execute and click **Apply Selected**.

If you see an action with:

- A grayed-out checkbox ( )

  The action is recommended-only.

  Possible reasons:

  – The action mode is *Recommend* or the underlying technology for the entity does not support automation. This means you have to perform the action outside Workload Optimization Manager.

  The Action Details page indicates that the action is blocked by a policy.

  – An action that is otherwise executable cannot be executed currently due to prerequisite actions.

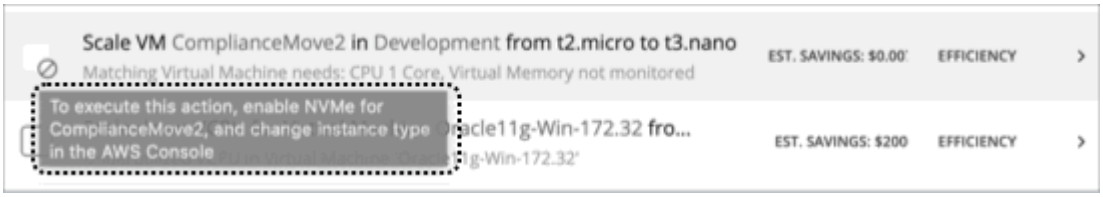
  For example, in order to suspend Host A, VM_01 in the host must first move to Host B. However, Host B only has capacity for one VM and is currently hosting VM_02. In this case, Host A suspension is blocked by two prerequisite actions – VM_02 moving to another host and VM_01 moving to Host B.

  The Action Details page for the main action (Host A suspension in the example) indicates that there are actions on the target or destination that need to be executed first.

  When all the prerequisite actions have been executed, the main action becomes executable.

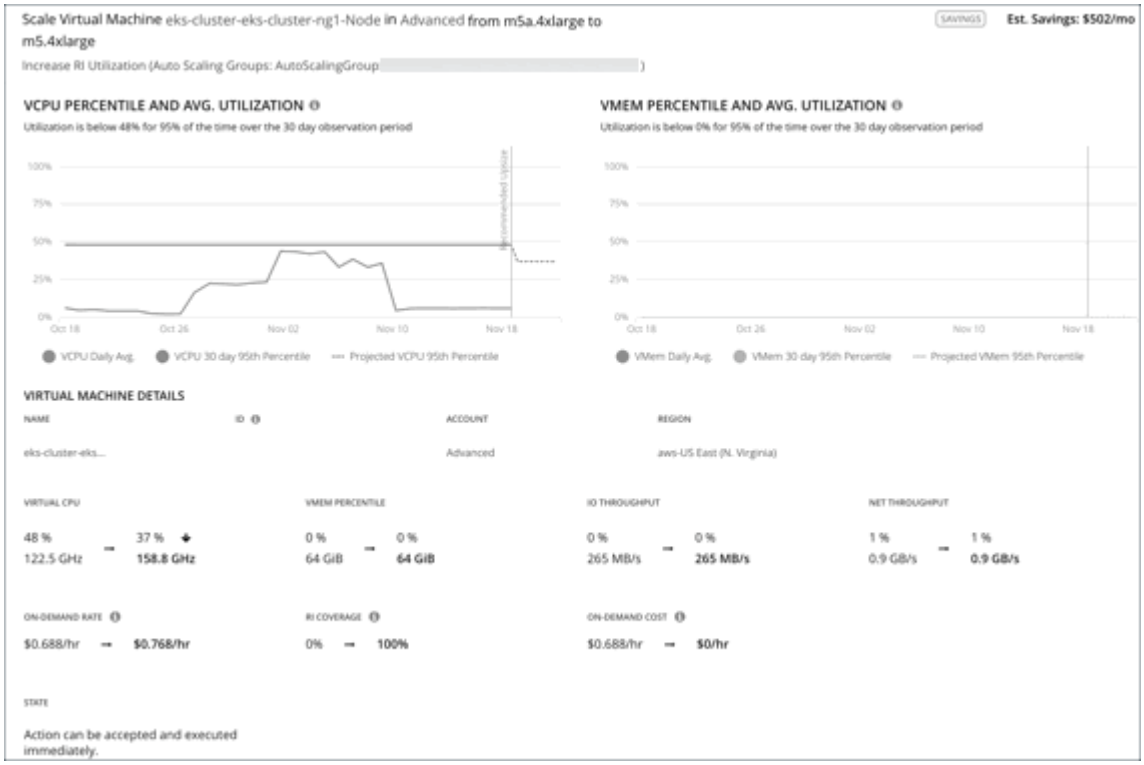
- A grayed-out checkbox and a prohibition symbol ( )

  You need to perform some prerequisite steps outside Workload Optimization Manager before you can execute the action. Hover on the checkbox to see the prerequisite steps.



B. Action Details

Click the arrow icon to expand the entry and view action details.



Action details include:

- A description of the recommended action, such as **Scale Virtual Machine...**.

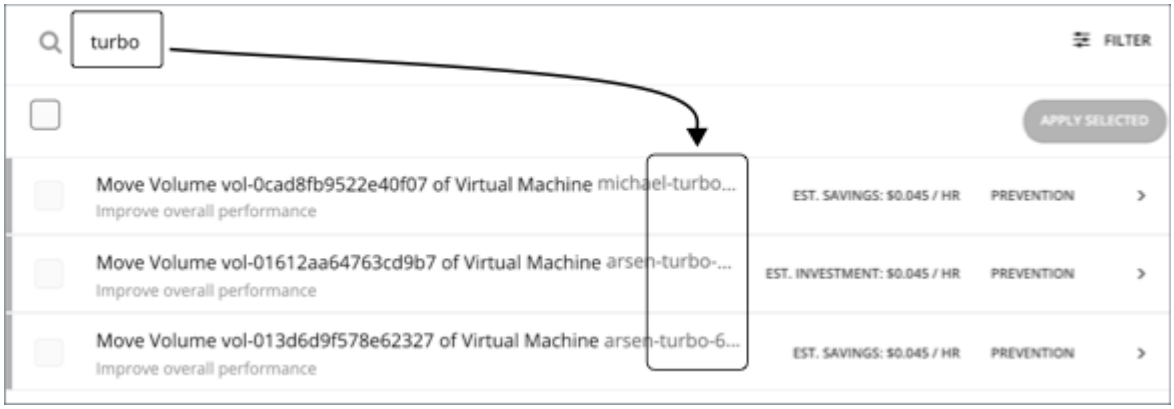
  NOTE:
  The action item gives the names of the affected entities. You can click on these entity names to drill down and set the **Home Page** scope to that specific entity. To return after drilling down to an entity in the action details, use the browser's **Back** button.

- Immediately below the description, a summary of requirements, risks, opportunities, or reasons for the recommended action
- The impact of executing the action.

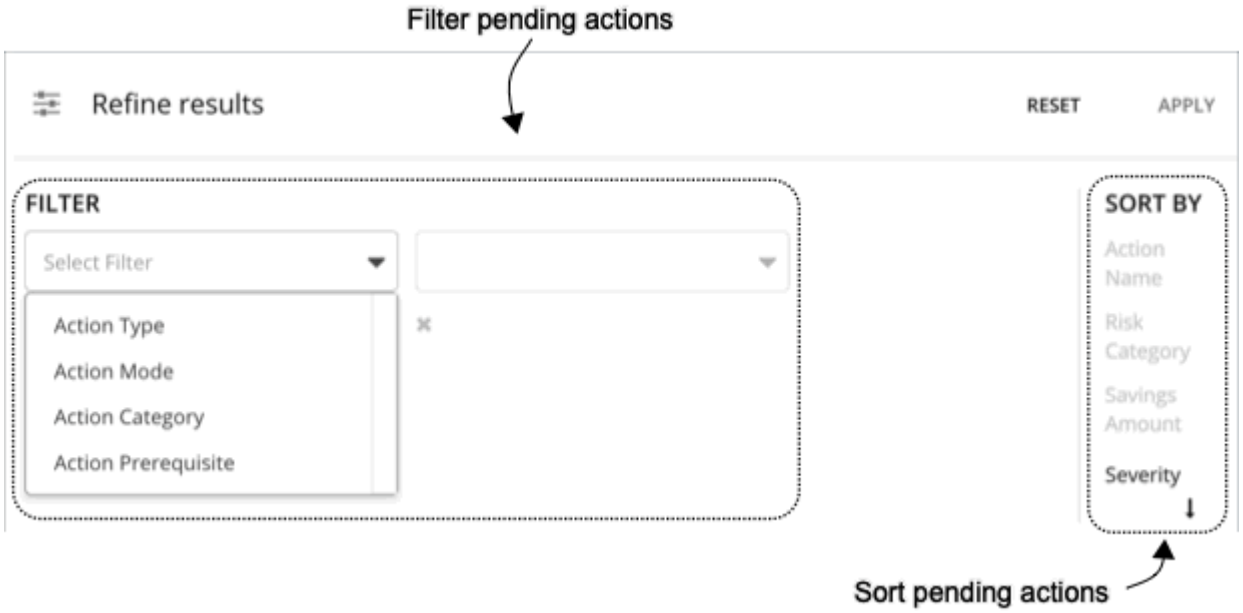
For more information, see Action Details *(on page 83)*.

C. Search

For a long list of pending actions, use search to narrow the results.



D. Filter and Sort



When you click **Filter**, you can:

- Filter the list by action type *(on page 67)*, action mode *(on page 72)*, action category *(on page 71)*, action prerequisite, or any combination of these items.
- Sort the actions in ascending or descending order by severity, name of the action target, risk category, or savings amount.

  Workload Optimization Manager determines action severity by the amount of improvement the affected entities will gain by executing the action. Action severities are:

  - Minor – Issues that affect cost or workload distribution, but not impact the QoS your users will experience
  - Major – Issues that can affect QoS and should be addressed
  - Critical – Issues that affect the QoS that your environment can deliver, and you are strongly advised to address them

For example:

- To see only the actions that you can execute through the Workload Optimization Manager user interface, filter the list by action mode and select **Manually executable**.

- To see only resize actions that are manually executable and that give efficiency improvements, set the filter as follows:

E. Download

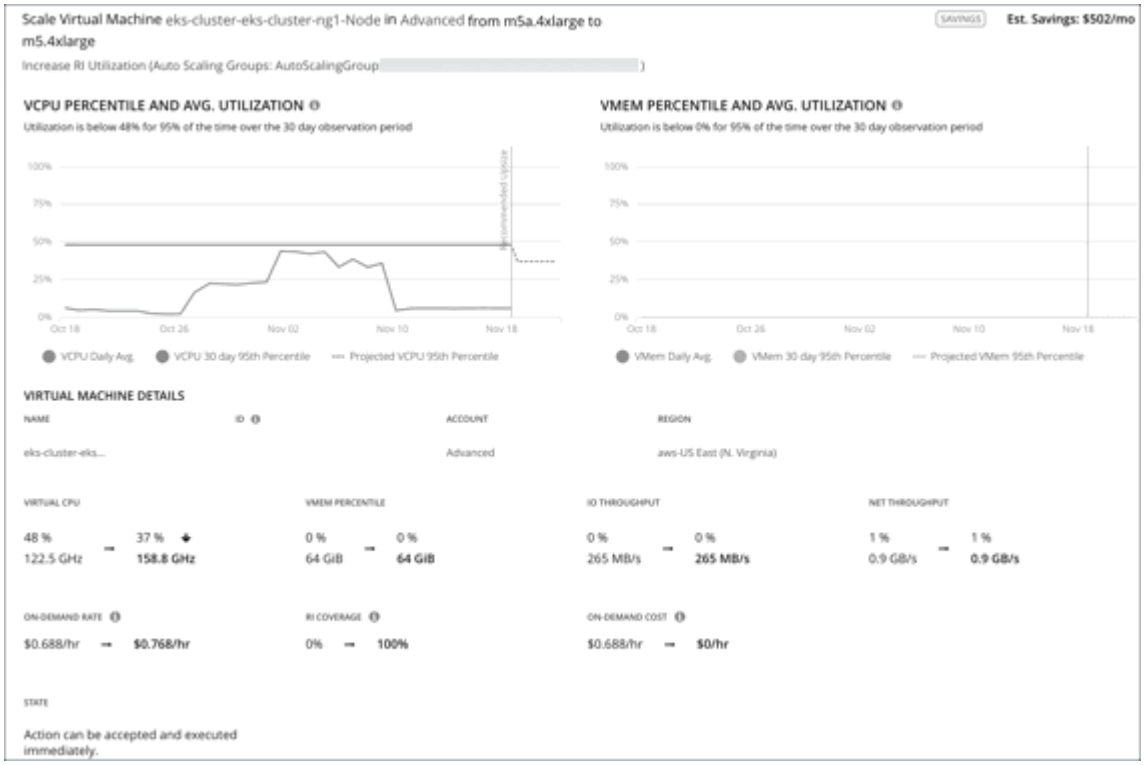
Download the pending actions list as a CSV file.

Action Details

Each action in the Pending Actions list comes with a description and additional details to help you understand why Workload Optimization Manager recommends it and what you would gain if you execute it.

At first glance, some 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.

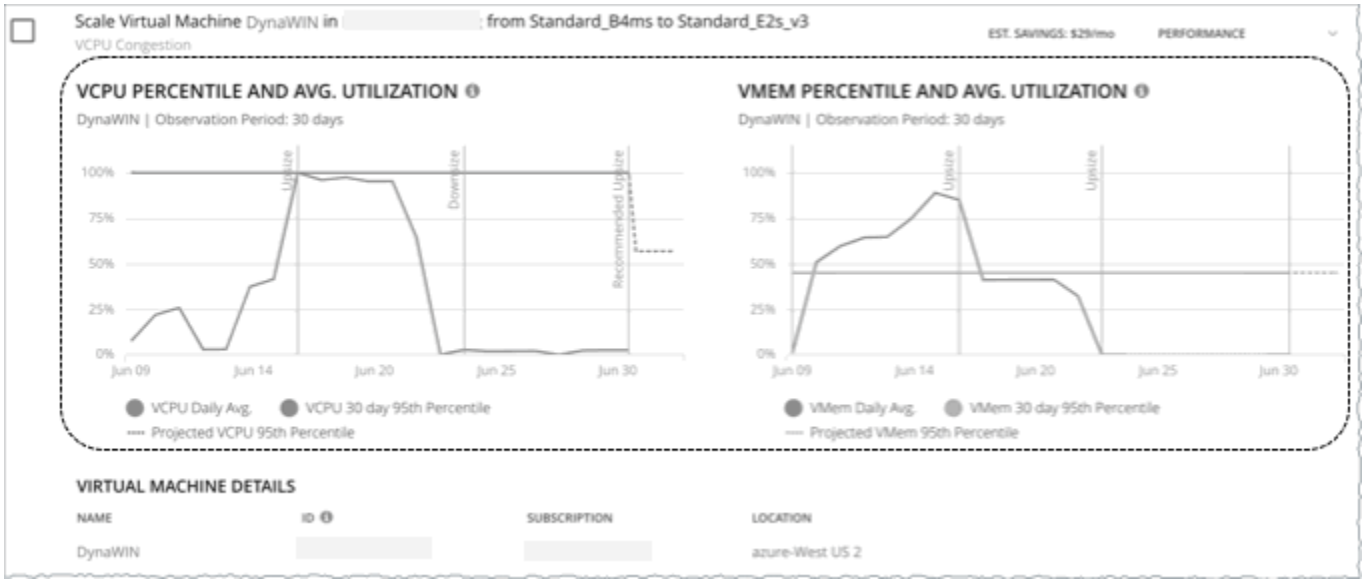
Example



In the image shown above, the action details indicate that scaling the virtual machine to a different instance type impacts discount coverage in a meaningful way. By increasing discount coverage from 0% to 100%, the projected hourly on-demand cost drops to $0, bringing estimated savings of $502 per month.

Utilization Charts

Workload Optimization Manager 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 Workload Optimization Manager'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	
Virtual Machine	▪ vCPU ▪ vMem	▪ vCPU ▪ vMem	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. 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.
Virtual Machine Spec	▪ vCPU ▪ vMem	▪ vCPU ▪ vMem ▪ Storage ▪ Number of replicas	See Virtual Machine Spec Actions *(on page 218)*.
Database (cloud)	▪ DTU Pricing Model – DTU ▪ vCore Pricing Model – vCPU – vMem – IOPS – Throughput	▪ DTU Pricing Model – DTU – Storage ▪ vCore Pricing Model – vCPU – vMem – IOPS – Throughput – Storage	See Cloud Database Actions *(on page 253)*.
Database Server (cloud)	▪ vCPU ▪ vMem ▪ IOPS	▪ vCPU ▪ vMem ▪ IOPS	See Cloud Database Server Actions *(on page 230)*.
Volume (cloud)	▪ IOPS ▪ Throughput	▪ IOPS ▪ Throughput	These charts also appear when you scope to a given

Entity Type	Monitored Resources		Notes
	Percentile Utilization	Average Utilization	
			volume and view the Details page. See Cloud Volume Actions *(on page 243)*.
Workload Controller	▪ vCPU limits and requests ▪ vMem limits and requests	▪ vCPU limits, throttling, and requests ▪ vMem limits and requests	See Container Actions *(on page 153)*.

Actions Tips and Best Practices

To get the best results from Workload Optimization Manager's Application Resource Management, you should set as many actions as possible to *Automated*. If you want to approve any changes, set the actions to *Manual*.

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, Workload Optimization Manager 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, scope the group in an automation policy, set the action mode to *Automatic*, and then set the schedule to off-peak hours or weekends. For details on setting schedules, see Setting Policy Schedules *(on page 105)*.

Resize Actions

Allow VMs that have hot-add enabled to automatically resize up.

Use Tuned Scaling to automatically resize VM and storage resources when the resize amount falls within an acceptable range, and for Workload Optimization Manager to notify you when the amount falls outside the range so you can take the most appropriate action. For details, see Tuned Scaling for On-prem VMs *(on page 270)*.

After executing a storage resize, Workload Optimization Manager indicates that the resize action has succeeded but the hypervisor might not show the corresponding change in storage capacity. If this occurs, perform a manual refresh of the hypervisor so it can discover the storage changes.

Move Actions

Workload Optimization Manager recommends automating host and storage migration.

Use placement constraints if you have placement requirements for specific workloads in your environment (for example, all production virtual machines moving only to specific clusters). Workload Optimization Manager can automatically import placement policies when you add a target, or you can create new placement policies. For more information, see Placement Policies *(on page 87)*.

Working With Policies

Policies set business rules to control how Workload Optimization Manager analyzes resource allocation, how it displays resource status, and how it recommends or executes actions. Workload Optimization Manager includes two fundamental types of policies:

- Placement Policies

  To modify workload placement decisions, Workload Optimization Manager divides its market into segments that constrain the valid placement of workloads. Workload Optimization Manager discovers placement rules that are defined by the targets in your environment, and you can create your own segments.

- Automation Policies

  Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

The Policy Management page shows all the currently defined policies. From this page you can:

- Create new policies.
- Delete a user-created policy.
- Edit a default or user-created policy.
- Enable or disable discovered placement policies. For a Workload Optimization Manager segment (a placement policy that was created in Workload Optimization Manager), you can edit the policy definition as well as enable/disable it.

To see the policies that are applied to a scope, go to the Search page and set the Workload Optimization Manager session to that scope. Then show the Policy view. For more information, see Scope Policies *(on page 51)*.

Things You Can Do

- Manage Imported Placement Policies – Importing Workload Placement Policies *(on page 88)*
- Create a Placement Policy – Creating Placement Policies *(on page 88)*
- Create a Scoped Automation Policy – Creating Scoped Automation Policies *(on page 97)*
- Create an Orchestration Policy – Action Orchestration *(on page 107)*

Placement Policies

To optimize your environment, Workload Optimization Manager recommends actions to place workloads such as applications, containers, or VMs on their providers. Workload Optimization Manager can recommend these actions, or execute them automatically.

When calculating workload placement, Workload Optimization Manager respects cluster boundaries, networks, and provisioned data stores. In addition, the configuration of your environment can specify logical boundaries, and within Workload Optimization Manager you can create even more boundaries. These boundaries impose segments on the market that Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager calculates moves. When managing segments you can:

- Review the placement policies that Workload Optimization Manager has discovered. These are policies that have been defined in your environment, outside of Workload Optimization Manager. See Importing Workload Placement Policies *(on page 88)*.
- Create placement segments that restrict workload placement according to specific rules. See Creating Placement Policies *(on page 88)*.

NOTE:
You can enable or disable any imported policy or created workload placement segment to affect placement calculations in the real-time environment or in plans.

Importing Workload Placement Policies

The hypervisors that you set as targets can include placement policies of their own. Workload Optimization Manager 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.

Workload Optimization Manager imports:

- vCenter Server DRS Rules

  See " Other Information Imported from vCenter " in the *Target Configuration Guide*
- Virtual Machine Manager Availability Sets

  See " " in the *Target Configuration Guide*
- Flexera One License Specifications

  See " Flexera" in the *Target Configuration Guide*

NOTE:
In vCenter environments, Workload Optimization Manager does not import DRS rules if DRS is disabled on the hypervisor. Further, if Workload Optimization Manager did import an enabled DRS rule, and somebody subsequently disables that DRS rule, then Workload Optimization Manager will discover that the rule was disabled and will remove the imported placement policy.

Creating Placement Policies

Placement Policies set up constraints to affect how Workload Optimization Manager calculates the placement of workloads in your environment. In this way, you can direct Workload Optimization Manager to recommend actions that satisfy business rules for your enterprise.

Workload Optimization Manager discovers Placement policies that have been defined in your environment, and you can also create Placement policies through the Workload Optimization Manager user interface. Note that you can enable or disable any Placement policy, both for real-time analysis and for planning scenarios.

Workload Optimization Manager supports 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

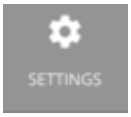
  For example, the VMs in a consumer group can never run on a host that is in the provider group. You can use such a segment to reserve specialized hardware for certain workloads.
- **Merge** – Merge clusters into a single provider group

  For example, you can merge three host clusters in a single provider group. This enables Workload Optimization Manager 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. Workload Optimization Manager can then recommend consolidating VMs or reconfiguring hosts in response to changing demand for licenses.

1. Navigate to the Settings Page.

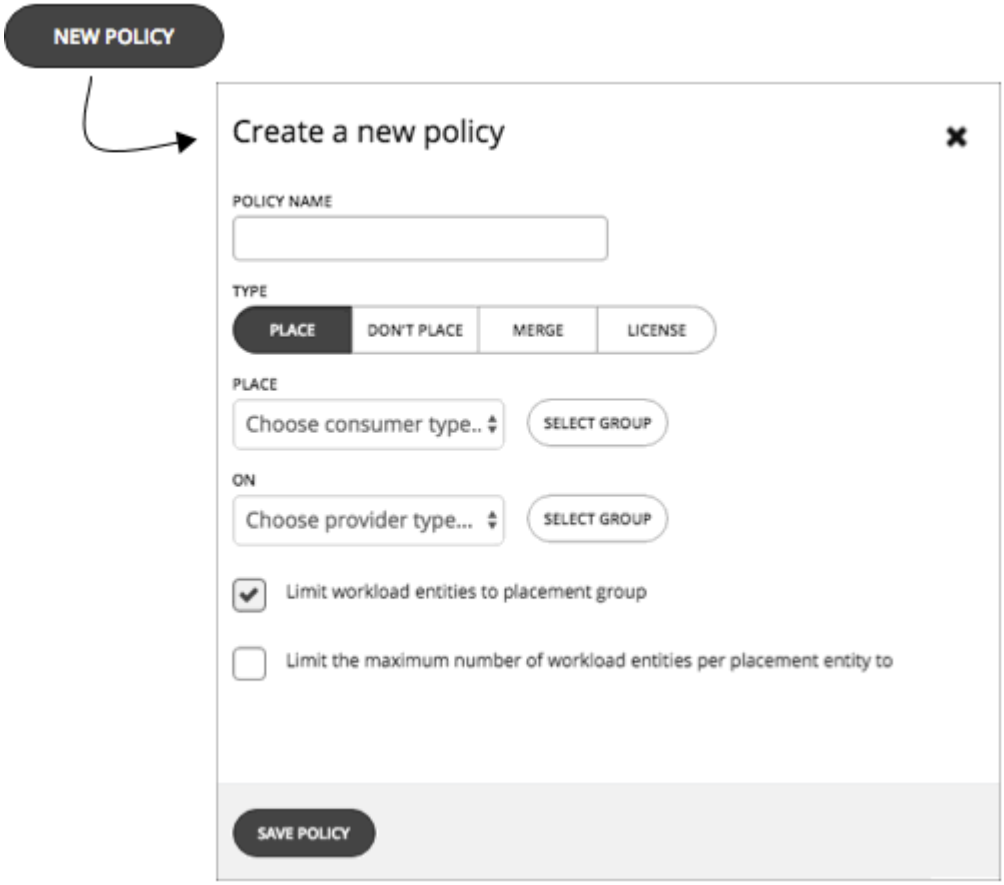
   Click to navigate to the Settings Page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.

2. Choose Policies.

   Click to navigate to the Policy Management Page.

   This page lists all the policies that you currently have configured for Workload Optimization Manager.

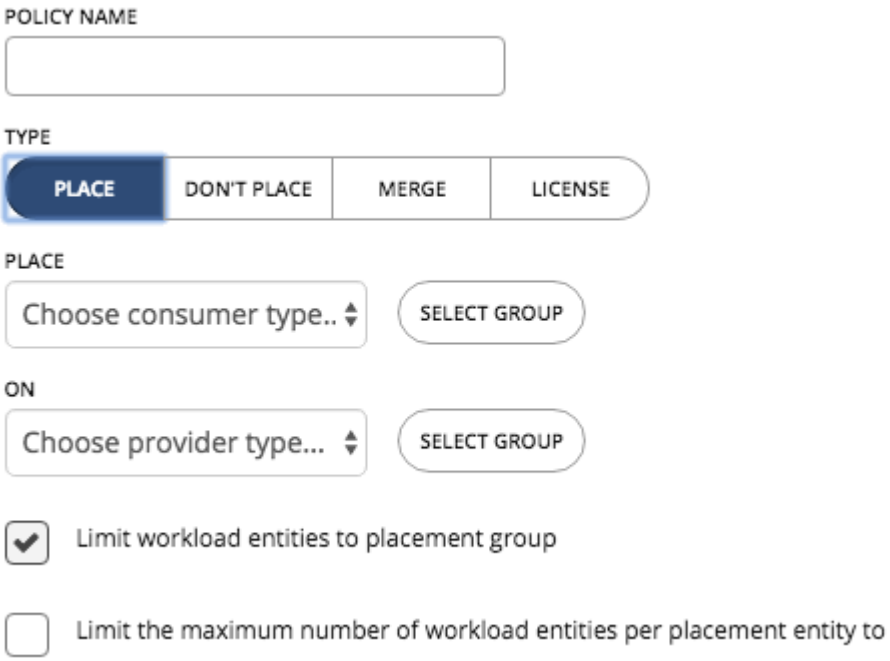
3. Create a new Placement policy.

   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

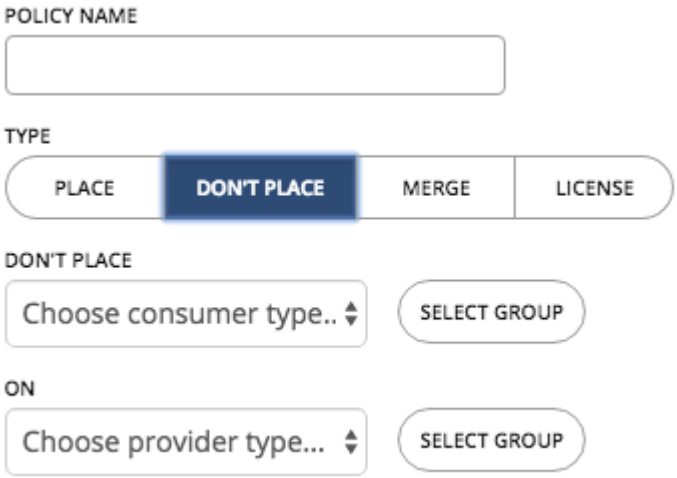
4. Create a **Place** policy.



These policies control where workload can be placed. For example, you can specify that a VM will only be placed on a host that is a member of a specific cluster. Or you could specify that any applications in a specific group can only be placed on a datastore 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 will provide resources to the consumers
- **Limit workload entities to placement group** – Set the policy to only place consumer entities 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

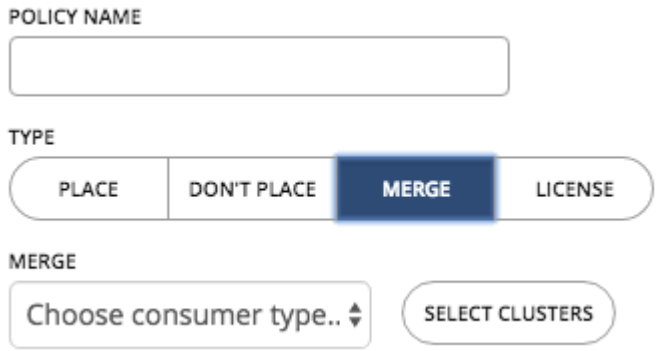
5. Create a **Don't Place** policy.



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. Or you can specify that a set of non-critical 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.



You can create placement policies that merge multiple clusters into a single logical group for the purpose of 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, Workload Optimization Manager has even more opportunities to increase efficiency in your environment.

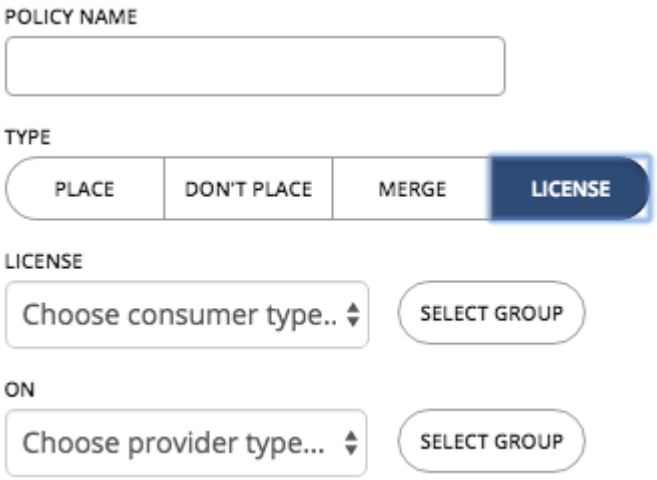
For merge policies, keep the following considerations in mind:

- For most policies that merge host and storage clusters, the clusters you place in the Merge segment must be members of the same datacenter.
- For vCenter environments, you can create placement policies that merge datacenters to support cross-vCenter moves. In this case, where a datacenter corresponds to a given vCenter target, the merged clusters can be in different datacenters. In this case you must create two merge policies; one to merge the affected datacenters, and another to merge the specific clusters.

  Also note that the clusters you merge must use the same network names on their respective datacenters.

To create a Merge policy, choose the type of entity to merge, and then select the groups you will merge.

7. Create a **License** policy.



Assume you have purchased a number of licenses for a database – you pay for the right to run that database on a certain number of 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, Workload Optimization Manager can recommend the following actions in response to changing demand for licenses:

- When demand is low, Workload Optimization Manager 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 Workload Optimization Manager 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 will see two recommendations – move VM_02 to Host_01 and then remove the license in Host_01. You will not see a recommendation to suspend Host_01.
- When demand exceeds capacity, and there are hosts in the policy that currently do not provide licenses, Workload Optimization Manager recommends reconfiguring those hosts to become providers and then moving VMs to those hosts. If all hosts are currently providing licenses, Workload Optimization Manager 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 Workload Optimization Manager to recommend reconfigure actions on hosts. In the automation policies, add the license-providing hosts and then enable the *Reconfigure* action.

8. When you have made all your settings, be sure to save the Policy.

Automation Policies

As Workload Optimization Manager gathers metrics, it compares the metric values against specified constraint and capacity settings to determine whether a metric exhibits a problem, and what actions to recommend or execute to avoid a problem. Workload Optimization Manager uses Automation Policies to guide its analysis and resulting actions. These policies can specify:

- Action Automation

  Whether to execute automatically or manually, or whether to just recommend the action. For more information, see Action Automation *(on page 106)*.
- Action Orchestration and Action Workflows

  Whether to have Workload Optimization Manager execute the action, have Workload Optimization Manager direct an orchestrator to execute the action, or execute the action with a workflow. For more information, see Action Orchestration *(on page 107)*.
- Constraints and Other Settings

  Settings that affect the Workload Optimization Manager analysis of the state of your environment. These include operational, utilization, and scaling constraints.

  For more information, see Constraints and Other Settings *(on page 122)*.

Default and Scoped Automation Policies

Workload Optimization Manager ships with default Automation Policy setting for the different types of entities it can discover in your environment. The settings for these default policies should be adequate to meet your initial business requirements. These policies apply to the global scope – Unless you override them, they affect all the entities in your environment. For more information, see Working With Default Automation Policies *(on page 93)*.

Workload Optimization Manager can include scoped Action Policies, which override the default settings for certain entities. With these policies you specify one or more groups of entities as the policy scope. You can also set a schedule to the policy to

specify maintenance windows, or to support orchestration workflows that require approval before executing the given action. For more information, see Working With Scoped Automation Policies *(on page 96)* and Setting Policy Schedules *(on page 105)*.

Working With Default Automation Policies

Workload Optimization Manager ships with default Automation Policy settings for the different types of entities it can discover in your environment. The settings for these default policies should be adequate to meet your initial business requirements. These policies apply to the global scope – Unless you override their settings, they affect all the entities in your environment.

Over time you might learn that you want to make global changes to certain policy settings. For example, **Enforce Non Disruptive Mode** is turned off by default. You might learn that in most cases you want to turn it on, and only turn it off for select scopes. In that case, you would turn it on in the default Automation Policy for VMs, and then set scoped policies for those groups of VMs for which you want to turn it off.

Relationships Between Default and Scoped Policies

Your default Automation Policies and scoped Automation Policies take effect in relation to each other. A default policy has a global effect, while a scoped policy overrides the default policy for the entities within the indicated scope. You should keep the following points in mind:

- Scoped policies override a subset of settings.

  A scoped policy can override a subset of settings for the entity type, and for the remainder Workload Optimization Manager will use the default policy settings on the indicated scope.
- When an entity applies conflicting scoped policies, Workload Optimization Manager 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.
- Scoped policies always take precedence over default policies.

  Even if the default policy has a more conservative setting, the setting in the scoped policy wins for entities in that scope.
- For a global effect, *always* use default policies.

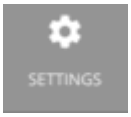
  Because of the *conservative setting wins* rule for scoped policies, you should never use a scoped policy to set a global effect. For example, you can create a scoped policy for the **All VMs** group. If you then specify a conservative setting for that policy, no other scoped policy can specify a more aggressive setting – the conservative setting will always win.

  For this reason, you should always use default Automation Policies whenever you want to achieve a global effect.

Viewing and Editing Default Automation Policies

To view or edit your default policies:

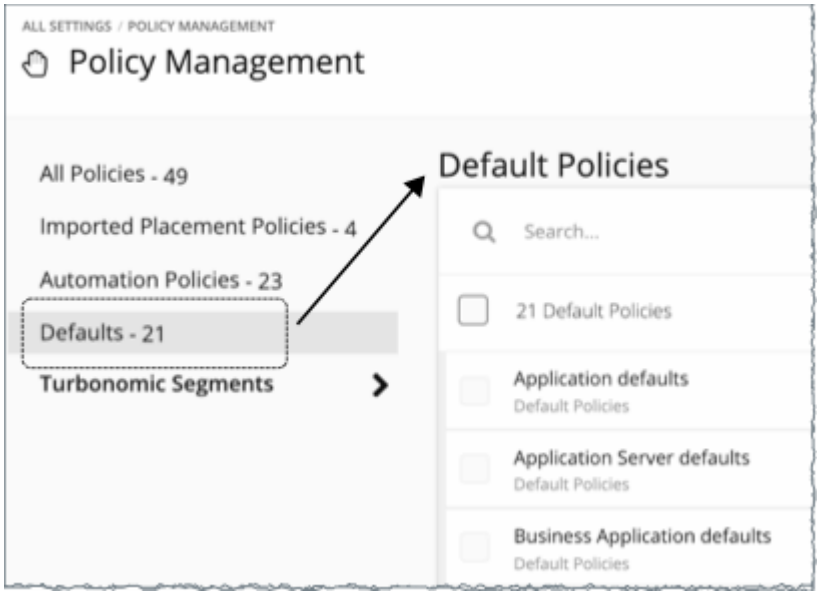
1. Navigate to the Settings Page.

   Click to navigate to the Settings Page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.
2. Choose Policies.

   Click to navigate to the Policy Management Page.

   This page lists all the policies that you currently have configured for Workload Optimization Manager.
3. On the Policy Management page, click **Defaults**.

   The page displays a list of all the default policies, by entity type.
4. Click the entity type whose default settings you wish to view or change.

   A fly-out appears with all the settings for that default policy. You can navigate to view different settings
5. Optionally, edit settings for this default policy.

   Navigate to the settings you want to change, and enter a different value for each.
6. When you're done, click **Save and Apply**.

Global Default Policy

Use these settings to modify Workload Optimization Manager analysis globally for any scope of your environment. These defaults affect both scoped automation policies and default automation policies.

ACTION AUTOMATION

Disable All Actions

Attribute	Default Setting
Disable All Actions	OFF

When this is ON, Workload Optimization Manager dos not generate any actions for your environment. For example, assume you have configured a number of polices 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 Workload Optimization Manager analysis will use to calculate time to exhaustion of your cluster resources.

Workload Optimization Manager 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, Workload Optimization Manager allows overprovisioning hosts up to 10 times their memory capacity, and up to 30 times their CPU capacity. When this setting is ON, Workload Optimization Manager removes these overprovisioning limits to allow VM placements on already overprovisioned hosts.

This setting does not stop Workload Optimization Manager 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 296)* 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)

  Workload Optimization Manager 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 391)* 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**

  Workload Optimization Manager 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 Workload Optimization Manager 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 272)* and verify that Storage Move actions are in *Recommend* or *Manual* mode. In addition, review your storage placement policies *(on page 87)* to ensure that individual VM volumes can be placed on the expected storage.

Working With Scoped Automation Policies

To override the current default Automation Policies, you can create scoped policies. These specify settings you want to change for certain entities in your environment. For these policies, you assign the policy to one or more groups of entities. In addition, you can assign a schedule to a scoped policy to set up maintenance windows or other scheduled actions in your environment.

Reasons to create scoped Automation Policies include:

- Change the Analysis Settings for Certain Entities

  Workload Optimization Manager 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. You can configure scoped policies to modify Operational Constraints or Scaling Constraints. For more information, see Constraints and Other Settings *(on page 122)*.

- Phase In Action Automation

  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 mode to Automatic for different actions on those VMs (see Action Modes *(on page 72)*).

- Trigger external workflows

  If actions require change approval, or integrations into DevOps pipelines to perform tasks before, instead of, or after action execution, you should scope those entities to a group and apply an automation policy.

For the steps to create a scoped policy, see Creating Scoped Automation Policies *(on page 97)*.

Discovered Scoped Automation Policies

As Workload Optimization Manager discovers your environment, it can find configurations that set up scopes that need specific policies. For example:

- HA Configurations

  For vCenter Server environments, Workload Optimization Manager 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

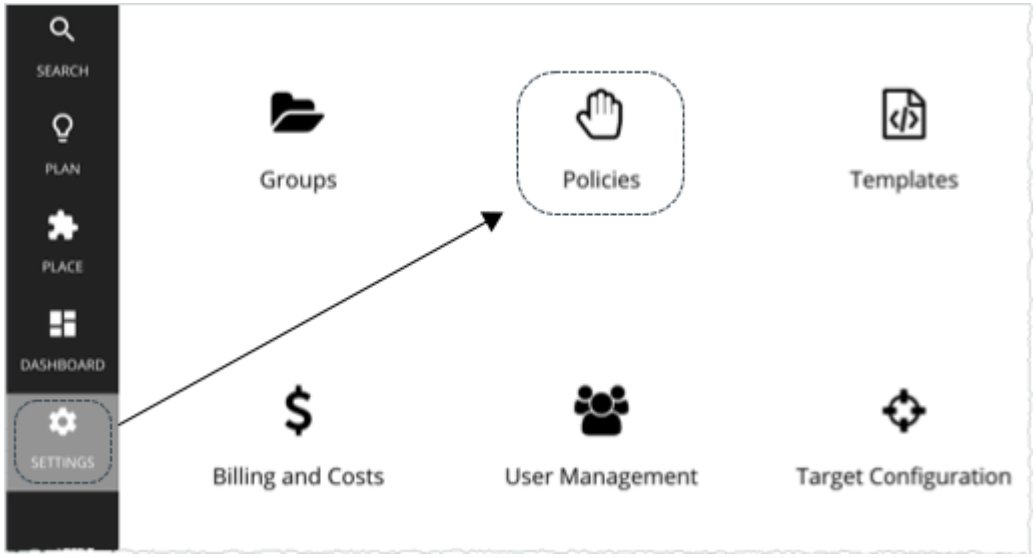
  In public cloud environments, Workload Optimization Manager discovers groups of VMs that should keep all their VMs on the same template. In the Automation Policies list, these appear with the prefix `AvailabilitySet::` on the policy names. You can enable Consistent Resizing for the VMs in each group so Workload Optimization Manager can resize them to the same size.

Creating Scoped Automation Policies

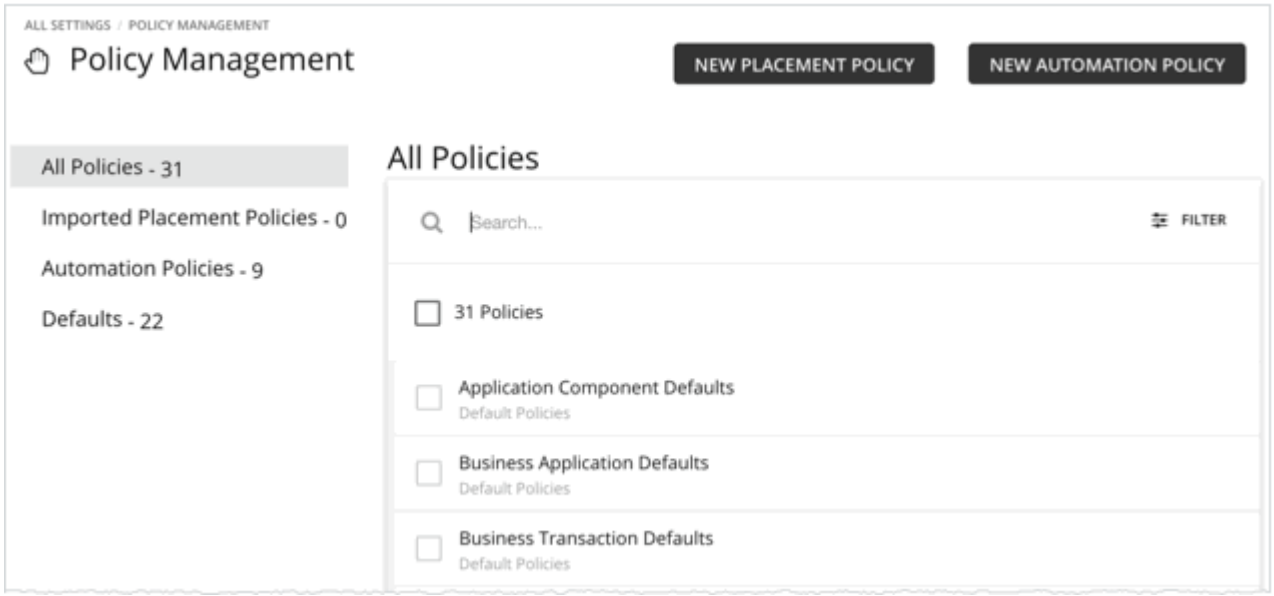
Create a new scoped Automation Policy from the Policy Management Page.

1. Entry Point

Navigate to the Settings Page and then choose **Policies**.



This opens the Policy Management Page, which lists all the currently available policies.

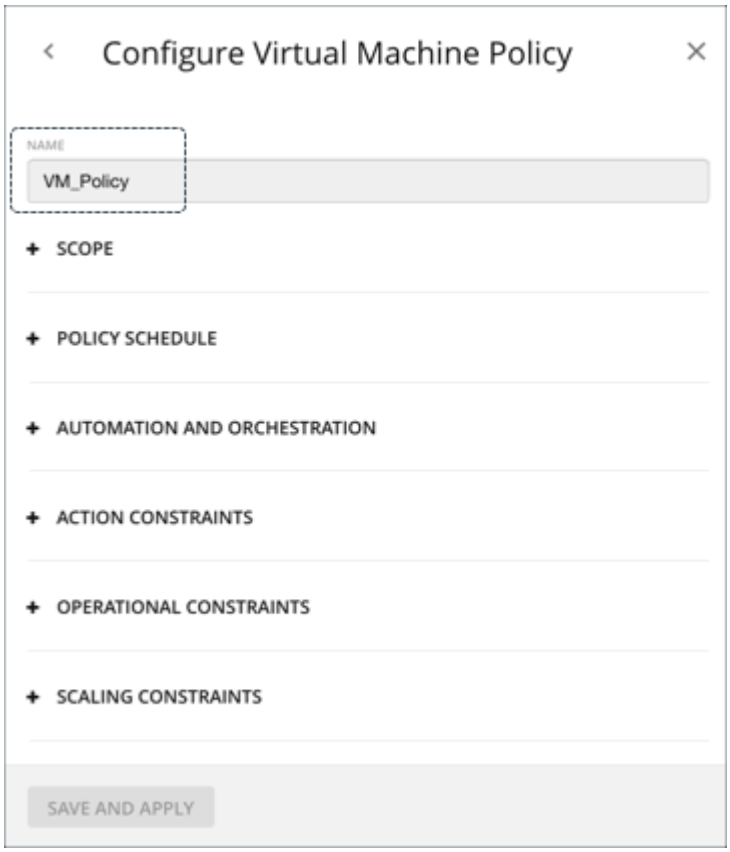


Click **NEW AUTOMATION POLICY** and then select the policy type (such as Virtual Machine).

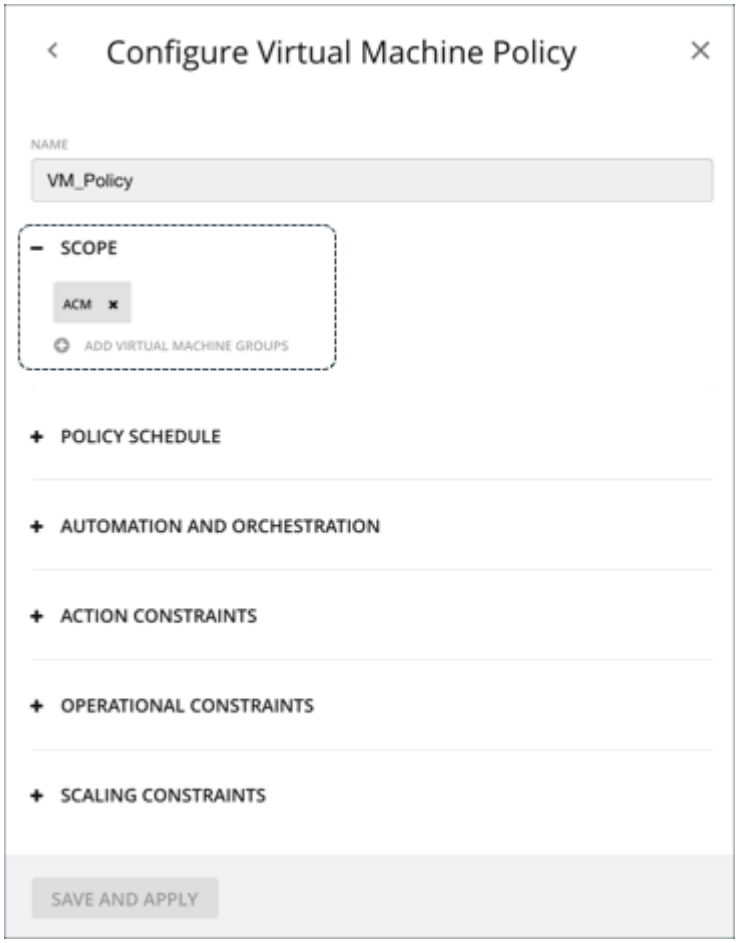
This sets the type of entity that your policy will affect. Note that Workload Optimization Manager supports different actions for different types of entities. For example, you cannot add VMem to a storage device. Setting policy type is the first step you take to focus on which actions you want to map to your workflows.

2. Policy Name

Name the policy.



3. Scope

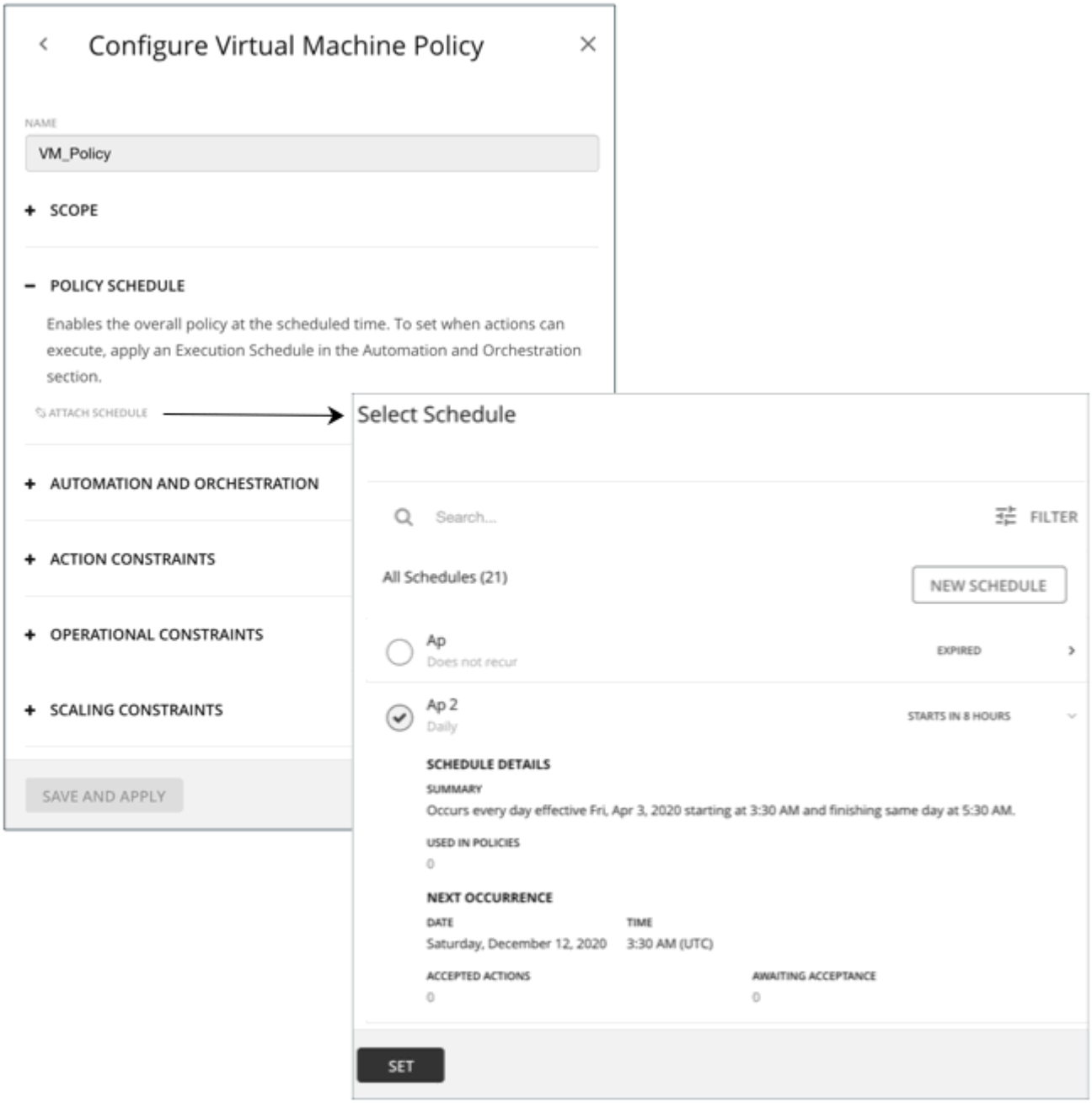


The scope determines which entities this policy will affect. Choose one or more groups, or create new groups and add them to the policy scope. These groups match the type of entity you have set for the policy.

In Workload Optimization Manager 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 Action Policy settings. For conflicts among scoped policy settings, the most conservative setting will take effect. For more details, see Policy Scope *(on page 104)*.

4. Policy Schedule



For use cases and information about how schedules affect policies, see Policy Schedules *(on page 105)*.

The **Select Schedule** fly-out lists all the schedules that are currently defined for your instance of Workload Optimization Manager.

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**

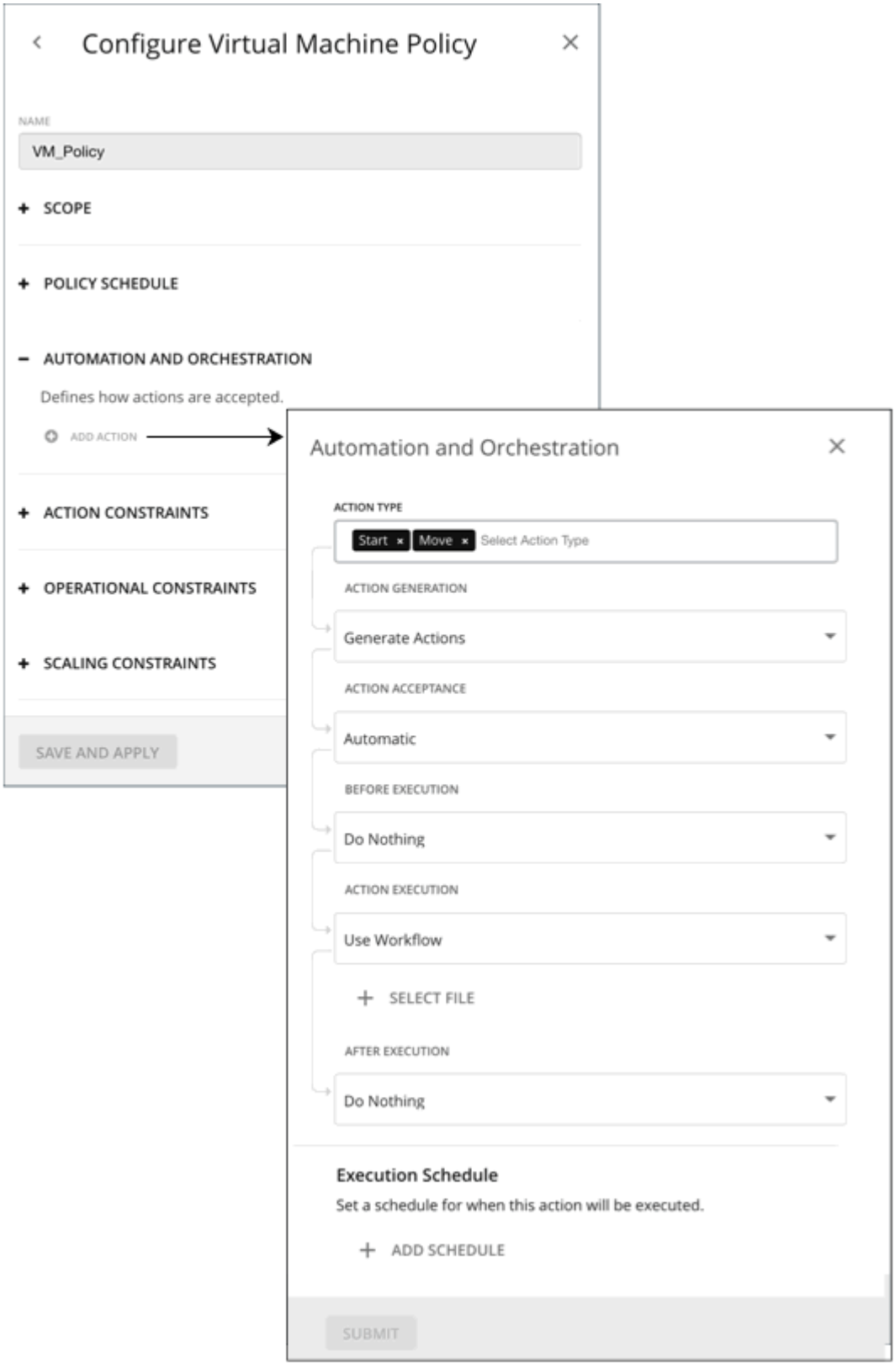
  The number of Manual actions affected by this schedule that are in the Pending Actions list, and have not been accepted. Click the number for a list of these actions.

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 513)*.

NOTE:
When you configure a schedule window for a VM resize action, to ensure Workload Optimization Manager 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 Workload Optimization Manager will not execute the resize action.

5. Automation and Orchestration

You can define automation and orchestration 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

  Workload Optimization Manager 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

  Workload Optimization Manager 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 – Recommend the action so a user can execute it via the given hypervisor or by other means
  - Manual – Recommend the action, and provide the option to execute that action through the Workload Optimization Manager user interface
  - Automatic – Execute the action automatically

    For automated resize or move actions on the same entity, Workload Optimization Manager 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, Workload Optimization Manager could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.

If you have an orchestrator target (such as ServiceNow), and that target includes an installation of the *Workload Optimization Manager Actions* application, you can send the action to the orchestrator. Choose from the following options:

- Generate Action then Send Record to Orchestrator
- Generate Action then Request Approval from Orchestrator

For more information, see Action Orchestration *(on page 107)*.

5.3. Before Execution, **Action Execution**, and **After Execution**

By default, generated actions execute without the need for orchestration. Workload Optimization Manager gives you the ability to set up orchestration to affect the execution of actions.

For more information, see Action Orchestration *(on page 107)*.

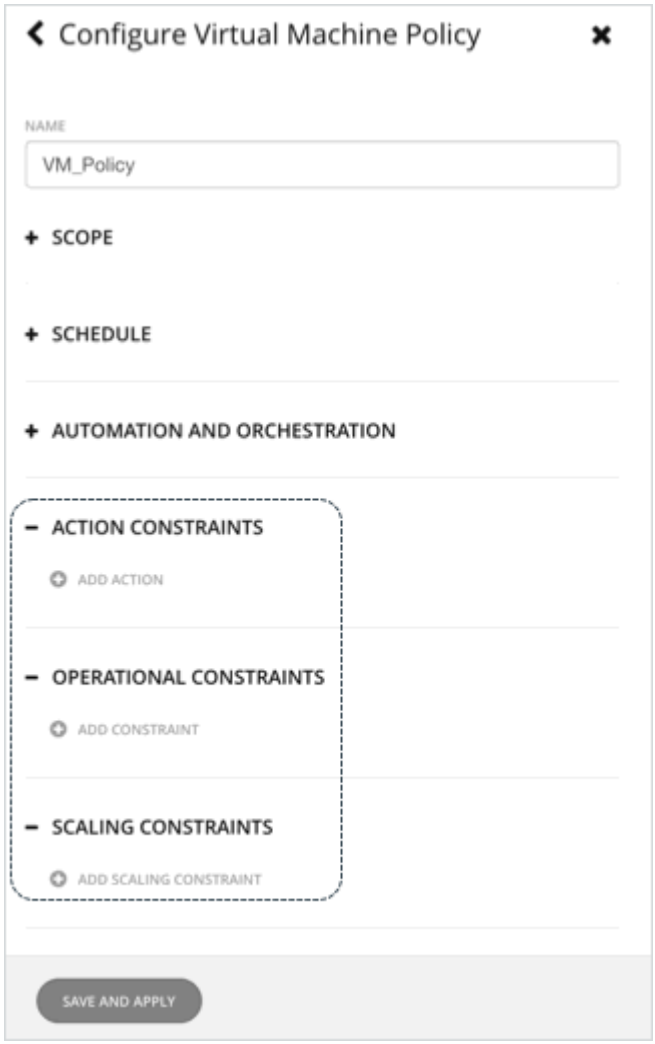
5.4. Execution Schedule

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, Workload Optimization Manager can recognize the need to resize, and will execute the action.

For more information, see Action Execution Schedules *(on page 105)*.

6. Constraints and Other Settings

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. For information about the settings you can make, see Constraints and Other Settings *(on page 122)*.



Policy Scope

You must declare a scope whenever you make a scoped Automation Policy. The scope determines which entities will be affected by the policy settings. To set scope, you assign one or more groups to the policy. You can use discovered groups, or you can create your own groups. For information about creating groups, see Creating Groups *(on page 509)*.

Relationships Between Default and Scoped Policies

Your default Automation Policies and scoped Automation Policies take effect in relation to each other. A default policy has a global effect, while a scoped policy overrides the default policy for the entities within the indicated scope. You should keep the following points in mind:

- Scoped policies override a subset of settings.

  A scoped policy can override a subset of settings for the entity type, and for the remainder Workload Optimization Manager will use the default policy settings on the indicated scope.
- When an entity applies conflicting scoped policies, Workload Optimization Manager 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.

- Scoped policies always take precedence over default policies.

  Even if the default policy has a more conservative setting, the setting in the scoped policy wins for entities in that scope.

- For a global effect, *always* use default policies.

  Because of the *conservative setting wins* rule for scoped policies, you should never use a scoped policy to set a global effect. For example, you can create a scoped policy for the **All VMs** group. If you then specify a conservative setting for that policy, no other scoped policy can specify a more aggressive setting – the conservative setting will always win.

  For this reason, you should always use default Automation Policies whenever you want to achieve a global effect.

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 Workload Optimization Manager analysis, and the actions it generates. You can set up scheduled times when you want to change those settings.

Remember that for scoped automation policies, it is possible that one entity can be in two different scopes – This means the entity can be under the effect of two different policies. For this reason, scoped policies keep the rule, *the most conservative setting wins*. However, a more aggressive scoped policy takes precedence over the corresponding default automation policy. For more details, see Policy Scope *(on page 104)*.

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 scoped policy, then the conservative settings *win*, and the scheduled changes do not take effect.

Policy Schedule and Action Execution Schedule

A scheduled policy can include *actions*. When the policy is in effect, Workload Optimization Manager 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 105)*.

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, Workload Optimization Manager can recognize the need to resize, and will execute resize actions. For this particular example, you will need to:

1. Create a scoped 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

Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager adds it to the queue of actions to be executed during the maintenance window. The action details show the action state as `AWAITING EXECUTION`. Workload Optimization Manager 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 Workload Optimization Manager removes it from the list of pending actions. This action will not be executed.

Workload Optimization Manager 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 default or scoped policies.

NOTE:
When you configure an execution schedule for a resize action, to ensure Workload Optimization Manager 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 Workload Optimization Manager will not execute the resize action. For information about non disruptive mode, see Non-disruptive Mode *(on page 274)*.

Action Automation

To avoid problems in your environment, Workload Optimization Manager analysis identifies actions that you can execute to keep things in optimal running order. You can specify the **degree of automation** you want for these given 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 modes** in a policy to set that business rule.

Action 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 modes in a policy to set that business rule.

Workload Optimization Manager supports the following action modes:

- Recommend – Recommend the action so a user can execute it via the given hypervisor or by other means
- Manual – Recommend the action, and provide the option to execute that action through the Workload Optimization Manager user interface
- Automatic – Execute the action automatically

For automated resize or move actions on the same entity, Workload Optimization Manager 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, Workload Optimization Manager could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.

Action Mode Configuration

There are two ways to configure action modes:

- Change the action mode in a default policy. For details, see Working With Default Automation Policies *(on page 93)*.
- Create an automation policy, scope the policy to specific entities or groups, and then select the action mode for each action.

  Workload Optimization Manager 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.

  For details, see Creating Scoped Automation Policies *(on page 97)*.

Action Orchestration

Action Orchestration specifies whether Workload Optimization Manager will execute an action, or whether Workload Optimization Manager will pass the action request to an orchestrator or an action workflow to effect the change in your environment. In this way, you can integrate supported orchestrators to execute of actions for specific scopes of entities in your environment.

About Orchestrators

Action Orchestration targets assign workflows that execute multiple actions to make changes in your environment. Workload Optimization Manager 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 mode is *Manual* or *Automatic*, then when Workload Optimization Manager recommends the action, it will direct the orchestrator to use the mapped workflow to execute it.

Workload Optimization Manager supports integration with ServiceNow. You can configure policies that log Workload Optimization Manager actions in your ServiceNow instance, and that submit actions for approval in ServiceNow workflows.

This section shows how to link orchestration workflows to automation policies. It assumes you have already configured an appropriate Orchestration target. It also assumes that you have configured workflows on that target in such a way that Workload Optimization Manager can discover the workflows and map them to automation policies. For information about Orchestration target requirements, see "Orchestrator Targets" in the *Target Configuration Guide*.

NOTE:
For some orchestration workflows, it is necessary to schedule an action to execute only during a specific maintenance window. Workload Optimization Manager policies can include schedules to enable this use case. However, you must be sure that you do not set the schedule to the policy that declares the orchestration you want. Instead, you should use two policies for the same scope – one to set up the orchestration, and another to schedule the time window during which the action mode will be `Automatic` (to set up the maintenance window). For more information, see Setting Policy Schedules *(on page 105)*.

About Action Workflows

Action workflows can add custom processing to Workload Optimization Manager actions at different entry points. For example, you can create an action workflow that sends an email whenever Workload Optimization Manager recommends moving a VM, or

you can create an action workflow that runs as a replacement for the action that Workload Optimization Manager would execute. For example:

- Action Script workflow

  You can deploy action scripts on a remote machine and configure an Action Script target that communicates with this server. Workload Optimization Manager discovers the exposed scripts and displays them as options you can choose when you specify a workflow in your automation policy. For more information about Action Scripts, see Deploying Action Scripts *(on page 112)*.

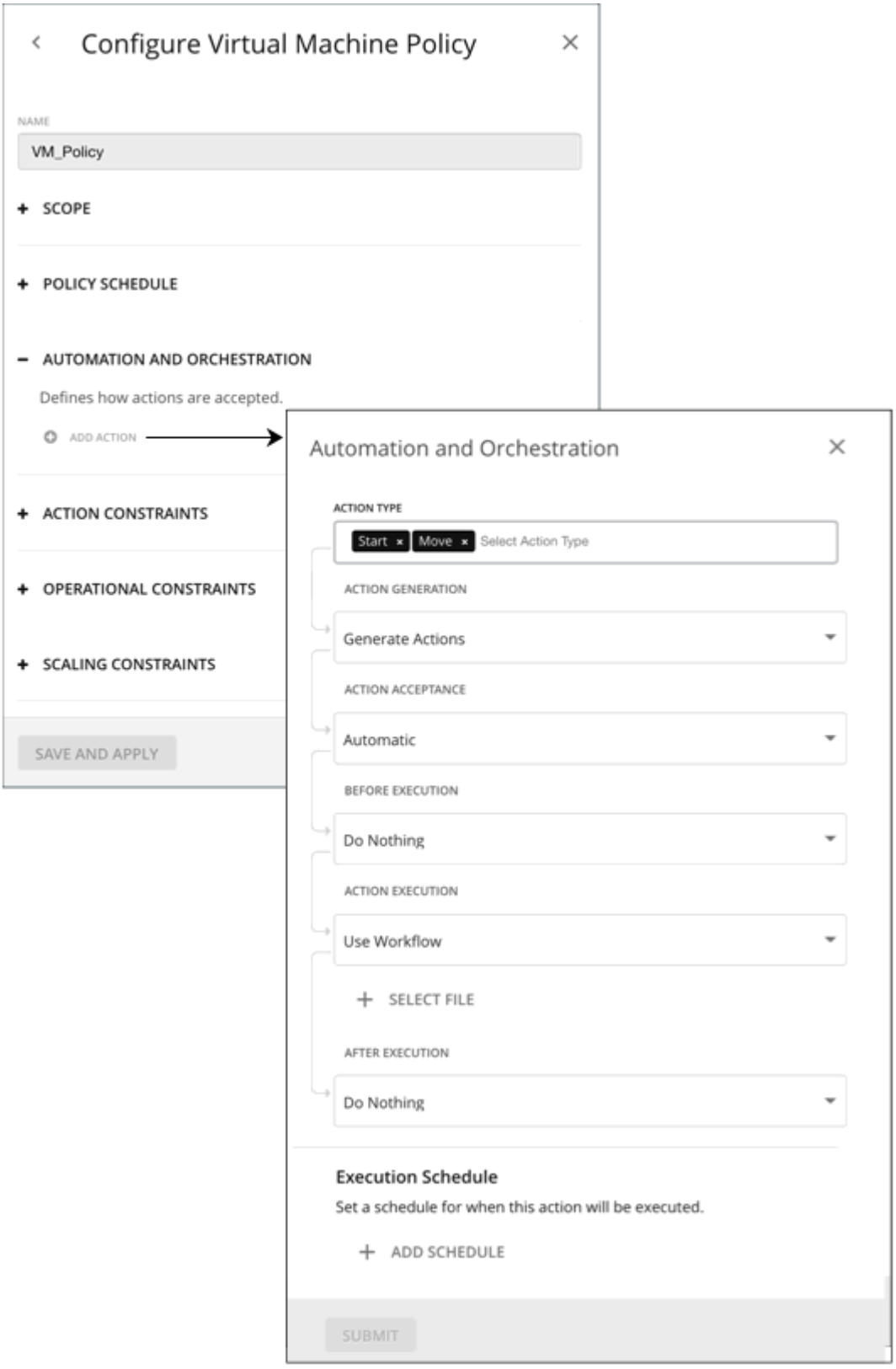
- Webhook workflow

  To implement a webhook, you create a workflow that specifies parameters such as the HTTP URL, HTTP method, and payload template. You can then use this workflow in your automation policy to orchestrate how actions execute. For more information about webhooks, see Creating Webhook Workflows *(on page 119)*.

Specifying Action Orchestration

As you create a policy, you specify the entity type and the scope of entities the policy affects. You can also set modes for specific actions. For example, you can set a mode of Manual for the Resize action for a given scope of VMs.

1. Expand **Automation and Orchestration** and click **Add Action**. Then select the action type you want to orchestrate.



There is no orchestration for this action by default. The following table describes the supported orchestrations.

	Generate Actions	Generate Action then Send Record to Orchestrator	Generate Action then Request Approval from Orchestrator
Description	Generate actions as usual, but use a workflow to control action execution.	Send a record of the generated actions to the orchestrator (such as ServiceNow).	Defer the generated actions to your orchestration workflow for approval. Workload Optimization Manager passes control for this action to your orchestration workflow as a Change Request (CR).
Prerequisites	An Orchestration target	An Orchestration target	An Orchestration target
Action Acceptance	Choose from the following: ▪ Recommend – Recommend the action so a user can execute it via the given hypervisor or by other means ▪ Manual – Recommend the action, and provide the option to execute that action through the Workload Optimization Manager user interface ▪ Automatic – Execute the action automatically For automated resize or move actions on the same entity, Workload Optimization Manager 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, Workload Optimization Manager could resize vCPU first. After this resize completes, it waits five minutes before resizing vMem.		Action acceptance automatically changes to "External Approval." If the action is approved, the action executes using the default Action Acceptance mode.
Before Execution	The default is Do Nothing. Select **Use Workflow** to trigger an external orchestration workflow such as a webhook to complete pre-execution tasks. Failure of the workflow indicates an action failure. For example, you can post an email notification to your team that an action has been generated.		
Action Execution	The default is Native. Workload Optimization Manager executes the action with its default action processing. Select **Use Workflow** to trigger an external orchestration workflow to execute the action. Failure of the workflow indicates an action failure.		
Execution Schedule	There is no execution schedule by default. Workload Optimization Manager executes the action immediately. If the policy includes a schedule, Workload Optimization Manager executes the action at the scheduled time.		There is no execution schedule by default. Workload Optimization Manager executes the action immediately. If the policy includes a schedule, Workload Optimization Manager executes the action at the scheduled time.

	Generate Actions	Generate Action then Send Record to Orchestrator	Generate Action then Request Approval from Orchestrator
			NOTE: Workload Optimization Manager discovers and enforces execution schedules defined in orchestration approval workflows. To avoid potential issues with schedules, set the execution schedule either in the orchestrator (such as ServiceNow) or Workload Optimization Manager.
After Execution	The default is Do Nothing. Select **Finalize action workflow** to trigger an external orchestration workflow such as a webhook to complete post-execution tasks. Failure of the workflow indicates an action failure.	The default is Do Nothing. Other options include: ▪ **Notify action workflow** Trigger an external orchestration workflow such as a webhook. For example, you can post a message to your collaboration platform to let your team know that an action has executed. Action status is not affected by the status of the notification workflow. ▪ **Finalize action workflow** Trigger an external orchestration workflow such as a webhook to complete post-execution tasks. Failure of the workflow indicates an action failure.	

2. When you have made all your settings, be sure to save the action policy.

Deploying Action Scripts

Action Scripts provide an interface that can add custom processing to Workload Optimization Manager actions.

Action scripts execute on a remote server (a VM or a container) that you have configured as a Workload Optimization Manager target. That server includes a manifest file that identifies the scripts you have deployed, as well the entities and actions they can respond to. Workload Optimization Manager discovers these scripts via the manifest and presents them as orchestration options for actions in automation policies.

For example, assume you have defined a script with:

- `name: MyVmMoveAction`
- `entityType: VIRTUAL_MACHINE`
- `actionType: MOVE`

Following this example, you can use the API to add orchestration to a policy for move actions on VMs. Because you have defined a script for that action, you can specify Action Script as the orchestration type, and you can choose the `MyVmMoveAction` script as the orchestration workflow to perform.

To deploy your action scripts, you will:

- Set up the remote Action Script Server (see Setting Up the Action Script Server *(on page 113)*)
- Create the action script executables on the remote server (see Creating Action Scripts *(on page 114)*)
- Deploy the Action Script Manifest on the remote server (see Deploying the Action Script Manifest *(on page 115)*)

Setting Up the Action Script Server

Workload Optimization Manager uses remote servers to execute action scripts. Managing the processes remotely means that you do not install custom code on the Workload Optimization Manager server, which eliminates associated security risks there. However, you are responsible for maintaining the security of your Action Script Server to ensure the integrity of your custom code. To accomplish this, the configuration of the remote server must meet certain requirements.

Resource Requirements for the Server

The remote server can be a VM or a container. The capacity you configure for the server depends entirely on the processes you intend to run on it. Workload Optimization Manager does not impose any special resource requirements on the server.

Configuring Command Execution

To support execution of your scripts, you must install any software that is necessary to run the scripts. This includes libraries, language processors, or other processes that your scripts will invoke.

Workload Optimization Manager invokes the scripts as commands on the server. The server must run an SSH service that you have configured to support command execution and SFTP operations. At this time, Cisco has tested action scripts with the OpenSSH sshd daemon.

The standard port for SSH is 22. You can configure a different port, and provide that for admins who configure the server as an Action Script target.

An action script can invoke any process you have deployed on the remote server. You do not have to run scripts; however, you must be able to invoke the processes from the command line. The script manifest gives Workload Optimization Manager the details it needs to build the command line invocation of each script.

Configuring the Action Script User Account

To execute the scripts on your server, Workload Optimization Manager logs on via a user account that is authorized to execute the scripts from the command line. You provide the user credentials when you configure the Action Script target. To support this interaction, the user account must meet the following requirements:

- Public key

  The user must have a public key in the `.ssh/authorized_keys` file. When you configure the Action Script target, you provide this as the Private Token for the target.
- Security for the `.ssh` directory

  The Action Script User should be the only user with authorized access. You should set file permissions to `600`.
- Supported shells

  The Action Script User shell can be either the Bourne shell (usually at `/bin/sh`) or the Bourne-Again shell (usually at `/bin/bash`). Workload Optimization Manager passes parameters as it invokes your scripts. At this time it only supports script execution through these shells.

Handling Action Script Timeouts

Workload Optimization Manager limits script execution to 30 minutes. If a script exceeds this limit, Workload Optimization Manager sends a `SIGTERM` to terminate the execution of the process.

Note that Workload Optimization Manager does not make any other attempt to terminate a process. For example you could implement the script so it traps the `SIGTERM` and continues to run. The process should terminate at the soonest safe opportunity. However, if the process does not terminate, then you must implement some way to terminate it outside of Workload

Optimization Manager. A runaway process continues to use its execution thread, which can block other processes (action scripts or primary processes) if there are no more threads in the pool.

Creating Action Scripts

An action script can be any executable that a user can invoke from a command line. You can save these executable files anywhere on the server – The Manifest indicates the path to the file (see Deploying the Action Script Manifest *(on page 115)*). The Action Script user that you have configured for the script server must have access to your script files, with read and execution privileges.

To execute a script, Workload Optimization Manager builds the appropriate SSH command from the manifest information it has discovered. It grants a timeout limit of 30 minutes by default, or the manifest entry can declare a different limit. If the execution exceeds the limit, Workload Optimization Manager sends a `SIGTERM` to terminate the process.

Passing Information to the Action Script

Workload Optimization Manager uses two techniques to pass information about an action to the associated action script:

- Pass general information via environment variables
- Pass full action data via `stdin`

To pass general information into the script, Workload Optimization Manager sets environment variables on the Action Script Server. You can reference these environment variables in your scripts. For example, assume you want to send an email that includes the name of the VM that is an action target. You can get that name via the `VMT_TARGET_NAME` environment variable.

The following list shows the environment variables that Workload Optimization Manager can set when it executes a script. Note that not all of these variables apply for every action. For example, an action to scale VMEM does not include providers, so the action does not include values for the `VMT_CURRENT_INTERNAL`, `VMT_CURRENT_NAME`, `VMT_NEW_INTERNAL`, or `VMT_NEW_NAME` variables. If a given variable does not apply, Workload Optimization Manager sets it to an empty string.

- `VMT_ACTION_INTERNAL`

  The UUID for the proposed action. You can use this to access the action via the REST API. For example, your script could accept or cancel the action according to its own criteria.
- `VMT_ACTION_NAME`

  The name of the action.
- `VMT_CURRENT_INTERNAL`

  The internal name for the current provider.
- `VMT_CURRENT_NAME`

  The display name for the current provider.
- `VMT_NEW_INTERNAL`

  The internal name for the new provider.
- `VMT_NEW_NAME`

  The display name for the new provider.
- `VMT_TARGET_INTERNAL`

  The internal name of the entity this action will affect. You can use this to access the target entity via the REST API. For example, you can get historical statistics or you can change settings for the entity.
- `VMT_TARGET_NAME`

  The display name of the entity this action will affect.
- `VMT_TARGET_UUID`

  The UUID of the entity this action will affect.

For some scripts, you might need a complete description of the associated action. For example, assume you want to analyze the utilization metrics for a given resource. The environment variables for passing general information do not include this information.

When it invokes an action script, Workload Optimization Manager passes the complete data for the associated action via `stdin`. Your script can load this into a variable to access the specific data it needs. For example, the following loads `stdin` into `myActionData`:

```
myActionData=$(cat -)
```

`stdin` contains a JSON string that represents of the full data associated with this action. For example, the `myActionData` variable could contain a string similar to:

```
{"actionType":"RIGHT_SIZE","actionItem":[{"actionType":"RIGHT_SIZE","uuid":"143688943343760","targetS
E":{"entityType":"VIRTUAL_MACHINE","id":"4200fcdb-eafe-2a4a-abf5-a7ad2b00555c"...
```

Deploying the Action Script Manifest

The Action Script Manifest identifies the scripts that you want to expose to Workload Optimization Manager. You provide the location of the manifest as part of the Action Script Target configuration – After Workload Optimization Manager validates the target, it then discovers these scripts and presents them in the Orchestration Policy user interface.

Creating the Scripts Manifest File

The Scripts Manifest is a file that declares an array of Script Objects for each script you want to expose. You can create the manifest as either a JSON or a YAML file.

For example, following are two examples of the same manifest – One in YAML and the other in JSON. Notice that in either case, the manifest is an array of two Script objects:

- YAML Manifest:

```
scripts:
    - name: MyVmMovePrep
      description: Execute this script in preperation to a VM Move
      scriptPath: vmScripts/movePrep.sh
      entityType: VIRTUAL_MACHINE
      actionType: MOVE
      actionPhase: PRE
    - name: MyVmSuspendReplace
      description: Execute this instead of a VM Suspend action
      scriptPath: vmScripts/suspendReplace.sh
      entityType: VIRTUAL_MACHINE
      actionType: SUSPEND
      actionPhase: REPLACE
```

- JSON Manifest:

```
{
    "scripts": [
     {
      "name": "MyVmMovePrep",
      "description": "Execute this script in preperation to a VM Move",
      "scriptPath": "vmScripts/movePrep.sh",
      "entityType": "VIRTUAL_MACHINE",
      "actionType": "MOVE",
      "actionPhase": "PRE"
     },
     {
      "name": "MyVmSuspendReplace",
      "description": "Execute this instead of a VM Suspend action",
      "scriptPath": "vmScripts/suspendReplace.sh",
      "entityType": "VIRTUAL_MACHINE",
      "actionType": "SUSPEND",
      "actionPhase": "REPLACE"
     }
    ]
}
```

You can save the Scripts Manifest file to any location on your server, so long as the Scripts User has access to that location, and has read and execute privileges. You will provide this location as the **Script Path**, which the Workload Optimization Manager administrator will give as part of the Action Script target configuration.

Note that the filename extension for the manifest must match the file format (either YAML or JSON). For example, you should name the file either `MyManifest.yaml` or `MyManifest.json`, respectively.

Declaring Script Objects

Each script object in the manifest can contain the following fields:

- **`name`**

  Required – The name for this action script. After Workload Optimization Manager discovers your scripts, it displays this name as a Orchestration Workflow choice in the user interface for creating orchestration policies.

- **`description`**

Optional – A description of the script. The Workload Optimization Manager user interface does not display this description.

- **`scriptPath`**

  Required – The path to the executable for this entry. You can give an absolute path, or a path that is relative to the location of the Scripts Manifest. The Action Script User that you set up for the Action Script server must have read and execute privileges for the executable file.

- **`entityType`**

  Required – The type of entity this script responds to. Can be one of:

  - `Switch`
  - `VIRTUAL_DATACENTER`
  - `STORAGE`
  - `DATABASE_SERVER`
  - `WEB_SERVER`
  - `VIRTUAL_MACHINE`
  - `DISK_ARRAY`
  - `DATA_CENTER`
  - `PHYSICAL_MACHINE`
  - `CHASSIS`
  - `BUSINESS_USER`
  - `STORAGE_CONTROLLER`
  - `IO_MODULE`
  - `APPLICATION_SERVER`
  - `APPLICATION`
  - `CONTAINeR`
  - `CONTAINER_POD`
  - `LOGICAL_POOL`
  - `STORAGE_VOLUME`
  - `DATABASE`
  - `VIEW_POD`
  - `DESKTOP_POOL`

  To configure the same script to respond to actions on different entity types, declare separate entries for that script, one for each entity type.

- **`actionType`**

  Required – The type of action this script responds to. Note that different entity types can support different actions. Can be one of:

  - `START`
  - `MOVE`
  - `SCALE`

    Resize on cloud – move workload from one cloud template or tier to another.
  - `SUSPEND`
  - `PROVISION`
  - `RECONFIGURE`
  - `RESIZE`
  - `DELETE`

  - RIGHT_SIZE
  - ACTIVATE
  - DEACTIVATE
  - BUY_RI
- **actionPhase**

  Required – Where in the life cycle of an action that you want your script to execute.

  Can be one of:

  - PRE

    For an action that has been accepted, or an AUTOMATED action before it executes, this state is a preparation phase where your script can execute just before the action itself executes.

    Run your script to set up conditions just before the action executes.
  - REPLACE

    For action execution, your script executes *in stead of* the execution that Workload Optimization Manager would perform.

    Run your script as a replacement for the Workload Optimization Manager action.
  - POST

    The action has completed execution, either in a SUCCEEDED, FAILING, or FAILED state.

    FAILING means that the status was checked after the action execution fails, but before the POST script has finished execution.

    Run your script after the action has completed execution.
- **timeLimitSeconds**

  Optional – How long to run the action before assuming a timeout. When execution exceeds this limit, Workload Optimization Manager sends a SIGTERM to terminate the execution of the process.

  If you do not provide a value, Workload Optimization Manager assumes a limit of 30 minutes (1800 seconds).

Webhooks

You can configure automation policies in Workload Optimization Manager to send data via webhooks to external web servers. A webhook is an automated message that Workload Optimization Manager can use to send data to external applications. Things you can do with webhooks include:

- Send notifications to collaboration platforms such as Slack
- Integrate Workload Optimization Manager with workflow management systems
- Override Workload Optimization Manager actions with your own logic

The webhook implementation supports HTTP messaging. In addition, to implement a webhook you create a workflow via the Workload Optimization Manager API.

To set up a webhook, you will:

- Identify the application to receive the webhook

  Possible applications can include collaboration platforms such as Slack, orchestration platforms such as ServiceNow, cloud provider APIs, or you can create a custom application that responds to HTTP methods.
- Create a webhook workflow in your Workload Optimization Manager instance

  For this version of Workload Optimization Manager you define webhook workflows via the API.

  A webhook definition can include:

  - The URL to the application to which you are sending the webhook
  - An HTTP method

  - A template for the webhook payload
  - Authentication credentials to access the application

  For information about creating a webhook workflow, see Creating Webhook Workflows *(on page 119)*.

- Create an Automation Policy that uses the webhook

  Automation Policies include orchestration settings where you can choose to execute a webhook for given actions. Workload Optimization Manager can execute a webhook when it creates an action, before it executes the action, instead of executing the action, and after it executes the action.

  For information about creating policies that use webhook workflows for orchestration, see Action Orchestration *(on page 107)*.

After you set up a webhook, when Workload Optimization Manager generates or executes the action you identified in the policy, it sends a message to the url that you specified in the webhook.

Creating Webhook Workflows

To implement a webhook, you create a workflow that specifies parameters such as the HTTP URL, HTTP method, and payload template. You can then use this workflow in your automation policy to orchestrate how actions execute.

To create a workflow, use the API to POST a Workflow object to Workload Optimization Manager instance. For example, the following `curl` commands get authorization to access a Workload Optimization Manager server, and then add a simple webhook workflow to that server:

- Authenticate on the server

  This command requests authentication credentials and stores them in a variable you can set to a cookie in subsequent curl headers, where:

  - <T8c_IP_ADDRESS> is the address of the Workload Optimization Manager server
  - <ADMIN_ACCOUNT_NAME> is the name of an account with admin privileges
  - <ADMIN_PWD> is the admin account password

```
JSESSIONID=$(curl \
    --silent \
    --cookie-jar - \
    --insecure \
    https://<T8c_IP_ADDRESS>/vmturbo/rest/login \
    --data "username=<ADMIN_ACCOUNT_NAME>&password=<ADMIN_PWD>" \
    | awk '/JSESSIONID/{print $7}')
```

- Create the workflow

  This command creates the workflow on the server, where:

  - <T8c_IP_ADDRESS> is the address of the Workload Optimization Manager server
  - <WEBHOOK_ADDRESS> is the address of the webhook server

```
curl \
    "https://<T8c_IP_ADDRESS>/api/v3/workflows" \
    --insecure \
    --compressed \
    --header 'Accept: application/json' \
    --header 'Content-Type: application/json' \
    --header "cookie: JSESSIONID=$JSESSIONID" \
    --request POST \
    --data '
    {
        "displayName": "My_WebHook",
        "className": "Workflow",
        "description": "First webhook attempt.",
        "discoveredBy":
        {
            "readonly": false
```

```
       },
      "type": "WEBHOOK",
      "typeSpecificDetails": {
      "url": "http://<WEBHOOK_ADDRESS>",
         "method": "POST",
         "template": "{ \"text\":\"My Webhook Template -- DATA: Action Details:
$action.details\" }",
         "type": "WebhookApiDTO"
      }
   }
   '
```

This is a simple webhook that sends its template to the indicated url. For a listing of the parameters you can set in the workflow, see WebhookApiDTO *(on page 120)* or the "API Reference" in the *API Guide*.

The template payload is the string `My Webhook Template -- DATA: Action Details:`, plus the action details that are included in the action's data object. The variable `$action.details` is a reference to a field in the `ActionApiDTO` object that represents the current action. Your template can reference any of the fields in this DTO, starting with `action` as the object name. For example, `$action.createTime` gives you the time the action was created. For a full listing of the `ActionApiDTO` object, see "API Reference" in the *API Guide* or the API Swagger UI.

Sample Webhook Application

A webhook workflow sends a message to an application via HTTP. You express the message as a template that can include values from the action data in its payload. This template can express text, JSON, or any other payload that you application can accept.

You can use webhooks to send messages to a number of existing applications, including Slack, Amazon Web Services, and others.

To deploy a simple example, and to test your webhook templates, you can implement a node.js server that receives the webhook message and prints out the template data. If you install this server on a machine in your network, then you can give its URL in the webhook workflow, and test your response to specific actions.

Following is a listing for a node.js web server that you can use.

```
let port = 9090;
const http = require("http");
console.log(`Starting server on port ${port}`);
```

```
http.createServer((request, response) => {
    request.setEncoding('utf8');
    console.log('REQUEST METHOD: ', request.method);
```

```
    let datStr = '';
    request.on('data', chunk => {datStr = datStr + chunk});
    request.on('end', () => {console.log('End of DATA: ', datStr)})
}).listen(port);
```

When you run this program, it prints a message to the console to say that it is running, and to identify the port it listens on.

When the server receives a message, it prints out the request method, and then prints out the message payload, as specified in the workflow's `template` field.

If you have configured an Automation Policy to use this workflow, then this server will log a message for each action that Workload Optimization Manager executes on an entity within the policy's scope.

WebhookApiDTO

The WebhookApiDTO inherits from `WorkflowAspect`

Required Parameters:

method

- **type:** string
- **description:** The http method used to make the request.
- **enum:** ['GET', 'POST', 'PUT', 'DELETE', 'PATCH']

url

- **type:** string
- **description:** The URL that HTTP request is made to.

Optional Parameters:

template

- **type:** string
- **description:** The template for the body of request.

authenticationMethod

- **type:** string
- **description:** The authentication method to use for the request.
- **enum:** ['NONE', 'BASIC', 'OAUTH']

username

- **type:** string
- **description:** The username for the authenticated request.

password

- **type:** string
- **description:** The password for the authenticated request.

trustSelfSignedCertificates

- **type:** boolean
- **description:** If true, self-signed certificates will be trusted when using HTTPS connections. Defaults to 'false'.

headers

- **type:** array
- **description:** The request headers.

oauthData

- **type:** object
- **description:** Model to define the oAuth data.

  Required Parameters:

  - `clientId`: *string* The client id used for oAuth authorization.
  - `clientSecret`: *string* The client secret used for oAuth authorization.
  - `authorizationServerUrl`: *string* The URL of the authorization server.
  - `grantType`: *enum* `["CLIENT_CREDENTIALS"]`

  Optional Parameters:

  - `scope`: *string* The oAuth scope.

Constraints and Other Settings

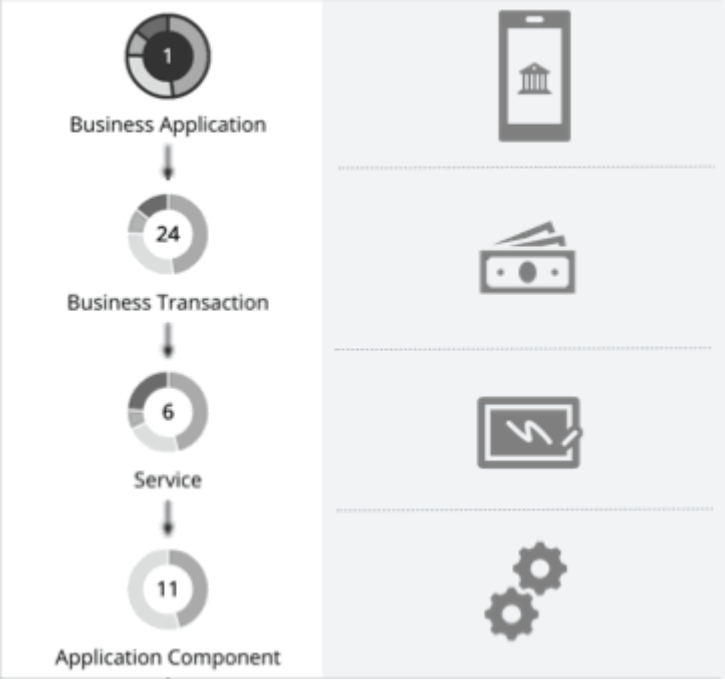
Workload Optimization Manager collects metrics to drive the analysis that it uses when it calculates actions for your environment. It compares current utilization and demand against allocated capacities for resources, so it can recommend actions that keep your environment in optimal running condition.

Action policies include constraints and other settings that you can make to adjust the analysis that Workload Optimization Manager performs. For example, you can set different levels of overprovisioning for host or VM resources, and Workload Optimization Manager will consider that as a factor when deciding on actions.

Workload Optimization Manager ships with default policies that show all the constraints and settings you can make for each policy. These take effect until you create and apply a policy with different values. For the steps in creating a new policy, see Creating Scoped Automation Policies *(on page 97)*. You can edit the defaults if you want to change analysis globally.

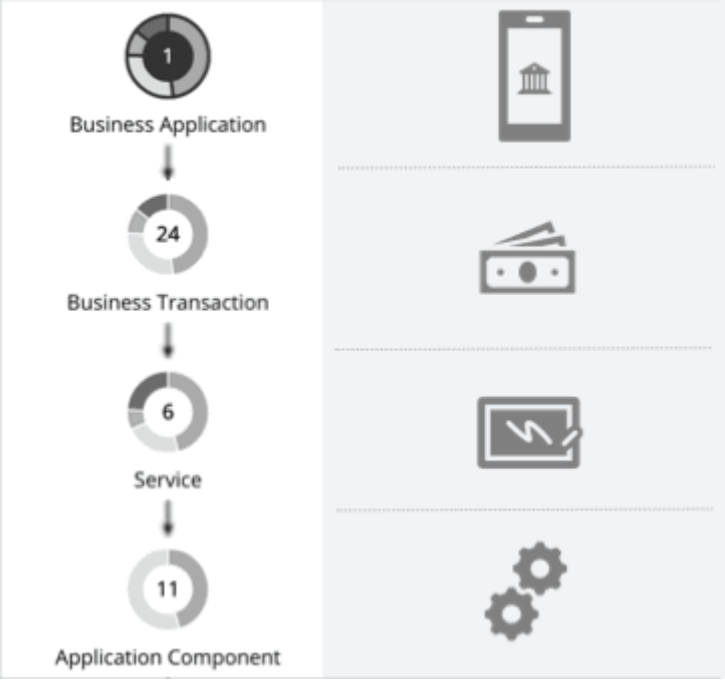
Entity Types - Applications

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.



Business Application

A Business Application is a logical grouping of Business Transactions *(on page 126)*, Services *(on page 129)*, Application Components *(on page 141)*, 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	
Budget:	Business Applications have unlimited budget.
Provides:	The complete application to end users
Consumes from:	Business Transactions, Services, Application Components, Database Servers, and the underlying nodes
Discovery:	Workload Optimization Manager discovers the following: ▪ AppDynamics Business Applications ▪ Dynatrace Applications 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 146)*.

Monitored Resources

Workload Optimization Manager monitors the following:

- 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).

- Transactions

  Transaction is a value that represents the per-second utilization of the transactions 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, Workload Optimization Manager estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

None

Workload Optimization Manager 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 Workload Optimization Manager estimates SLO based on monitored values.
Transaction SLO	None If you enable SLO, Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager estimates SLO based on monitored values.
Response Time SLO [ms]	None

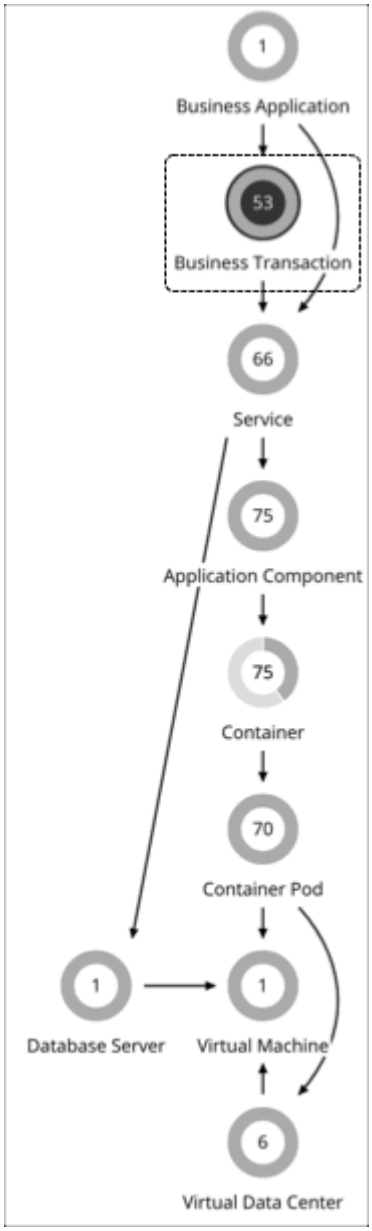
Attribute	Default Setting/Value
	If you enable SLO, Workload Optimization Manager uses the default value of 2000. You can change this to a different value.

Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Workload Optimization Manager 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. For more information, see Business Application *(on page 123)*.

Synopsis



Synopsis	
Budget:	Business Transactions have unlimited budget.
Provides:	Response time and transactions to Business Applications
Consumes from:	Services *(on page 129)*, Application Components *(on page 141)*, Database Servers, and the underlying nodes
Discovery:	Workload Optimization Manager discovers the following: ▪ AppDynamics Business Transactions ▪ NewRelic Key Transactions

Synopsis	
	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 146)*.

Monitored Resources

Workload Optimization Manager monitors the following:

- 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).

- Transactions

  Transaction is a value that represents the per-second utilization of the transactions 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, Workload Optimization Manager estimates SLOs based on monitored values. You can set your own SLO values in policies.

Actions

None

Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

None

Workload Optimization Manager 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 Workload Optimization Manager estimates SLO based on monitored values.
Transaction SLO	None If you enable SLO, Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager estimates SLO based on monitored values.
Response Time SLO [ms]	None If you enable SLO, Workload Optimization Manager uses the default value of 2000. You can change this to a different value.

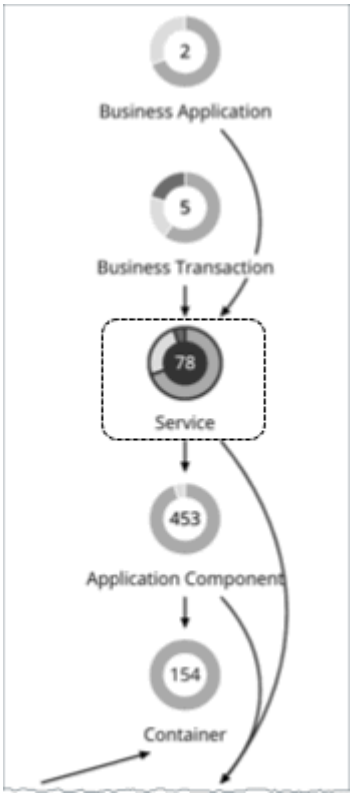
Response time SLO determines the upper limit for acceptable response time (in milliseconds). If response time reaches the given value, Workload Optimization Manager sets the risk index to 100%.

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.

For more information, see Application Component *(on page 141)* and Business Application *(on page 123)*.

Synopsis



Synopsis	
Budget:	Services have unlimited budget.
Provides:	Response time and transactions to Business Transactions *(on page 126)* and Business Applications
Consumes from:	Application Components, Database Servers, and the underlying nodes
Discovery:	For APM targets, Workload Optimization Manager discovers the following: ▪ AppDynamics Tiers ▪ Dynatrace Services ▪ Instana Services ▪ NewRelic APM Applications / NewRelic Services (New Relic ONE) **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 146)*. For Kubernetes, Workload Optimization Manager discovers Kubernetes Services through the Kubeturbo pod that you have deployed in your environment.

Monitored Resources

Workload Optimization Manager monitors the following:

- 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 Kubernetes, this is the desired *weighted average* response time of all Application Component replicas associated with a Service.

- Transactions

  Transaction is a value that represents the per-second utilization of the transactions allocated to a given entity.

  For Kubernetes, this is the maximum number of transactions per second that each Application Component replica can handle.

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, Workload Optimization Manager estimates SLOs based on monitored values. You can set your own SLO values in policies.

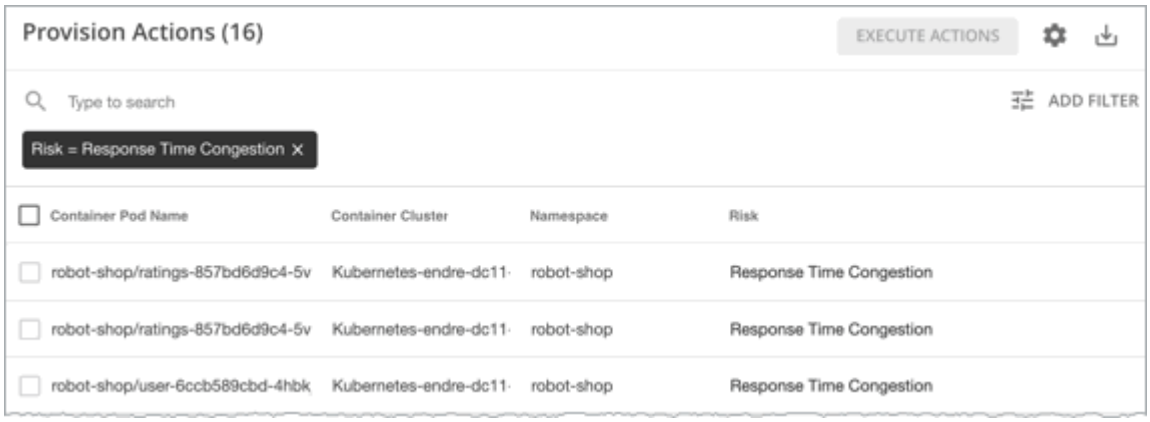
Actions for non-Kubernetes Services

Workload Optimization Manager does not recommend actions for non-Kubernetes Services, 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.

Actions for Kubernetes Services

For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications.

For example, when current response time for an application is in direct violation of SLO, Workload Optimization Manager will recommend provisioning pods to improve response time. When you examine a pending action to provision pods, you will see *Response Time Congestion* as the reason for the action.



If applications can meet response time SLOs using less resources, Workload Optimization Manager will recommend suspending pods to improve infrastructure efficiency.

Action Generation Requirements

Workload Optimization Manager generates actions for Kubernetes Services under the following conditions:

- Services are discovered by the Kubeturbo pod that you have deployed in your environment.
- Services collect performance metrics via the Instana target or DIF (Data Ingestion Framework).
- You have created policies for the Services.

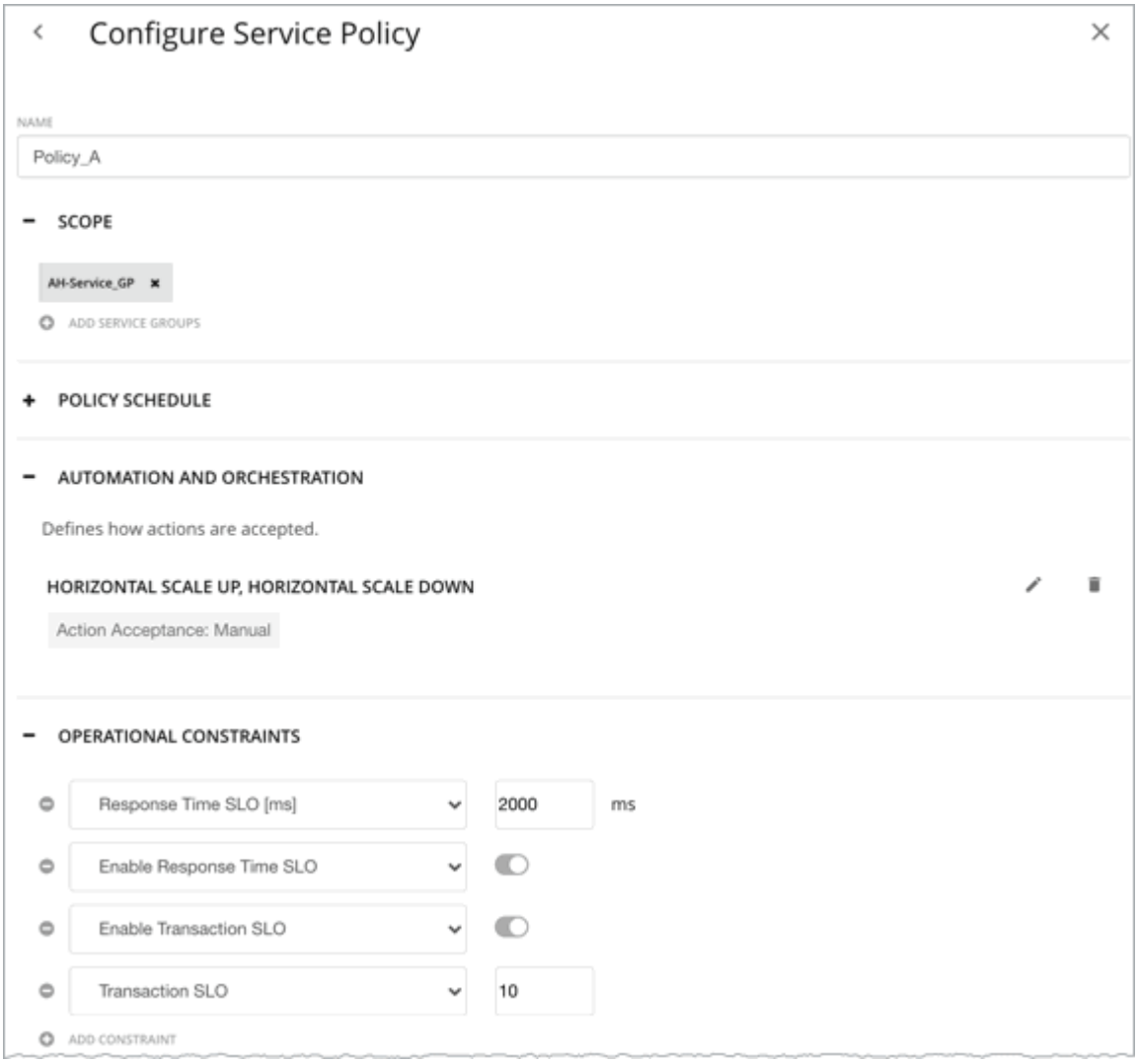
Creating Policies for Kubernetes Services

You can create policies from two places.

- Workload Optimization Manager user interface

  Navigate to **Settings > Policies**, and then click **New Automation Policy > Service**. The next section discusses how to configure the policy correctly.

  The policy should be similar to the following:



- Kubernetes cluster

  Application owners who do not have access to the Workload Optimization Manager user interface can create the following CR files in their Kubernetes clusters. Kubeturbo discovers the settings in these files and then syncs them with the user interface every ten minutes.

  - `ServicePolicy` CR file

    This is a YAML file that specifies the policy settings that apply to a group of Kubernetes Services (download a sample here). The next section discusses how to configure this file correctly.

    The file should be similar to the following:

```
apiVersion: policy.turbonomic.io/v1alpha1
kind: SLOHorizontalScale
metadata:
  name: Policy_A
spec:
  minReplicas: 1
  maxReplicas: 10
  objectives:
  - name: ResponseTime
    value: 300
  - name: Transaction
    value: 20
  behavior:
```

```
    scaleUp: Automatic
    scaleDown: Disabled
```

IMPORTANT:
Before creating the `ServicePolicy` CR file, be sure to add this CRD file to the Kubernetes cluster. This CRD file is mandatory and is intended to ensure the validity of the `ServicePolicy` CR file.

– `ServicePolicyBinding` CR file

This is a YAML file that binds the policy settings specified in `ServicePolicy` to the group of Kubernetes services specified in this CR file (download a sample here). The next section discusses how to configure this file correctly.

The file should be similar to the following:

```
apiVersion: policy.turbonomic.io/v1alpha1
kind: PolicyBinding
metadata:
  name: AH-Service_GP
policyRef:
  kind: SLOHorizontalScale
  name: Policy_A
targets:
- apiVersion: apps/v1
  kind: Service
  name: cass-api
- apiVersion: apps/v1
  kind: Service
  name: cass-(cass|user)
```

IMPORTANT:
Before creating the `ServicePolicyBinding` CR file, be sure to add this CRD file to the Kubernetes cluster. This CRD file is mandatory and is intended to ensure the validity of the `ServicePolicyBinding` CR file.

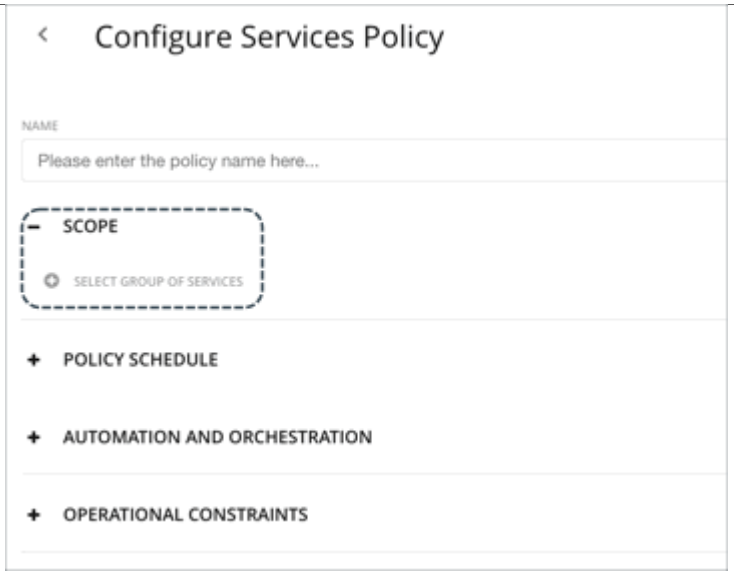
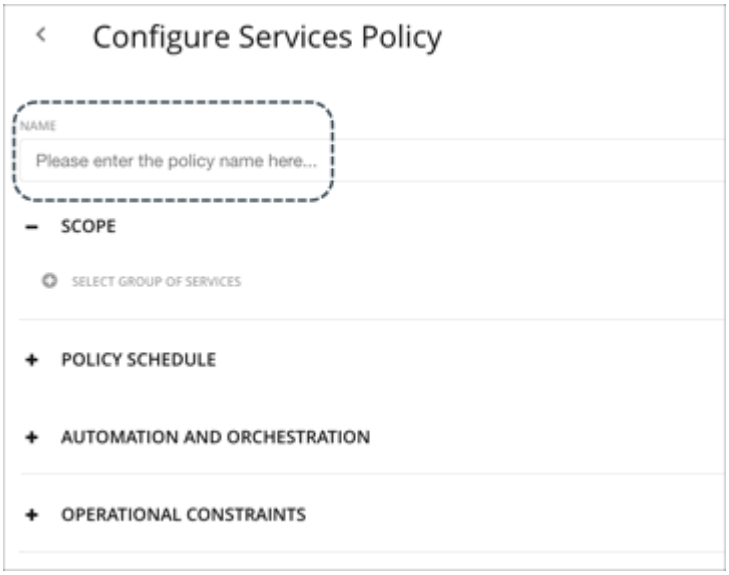
Settings for Kubernetes Service Policies

Configure the following settings for the policies.

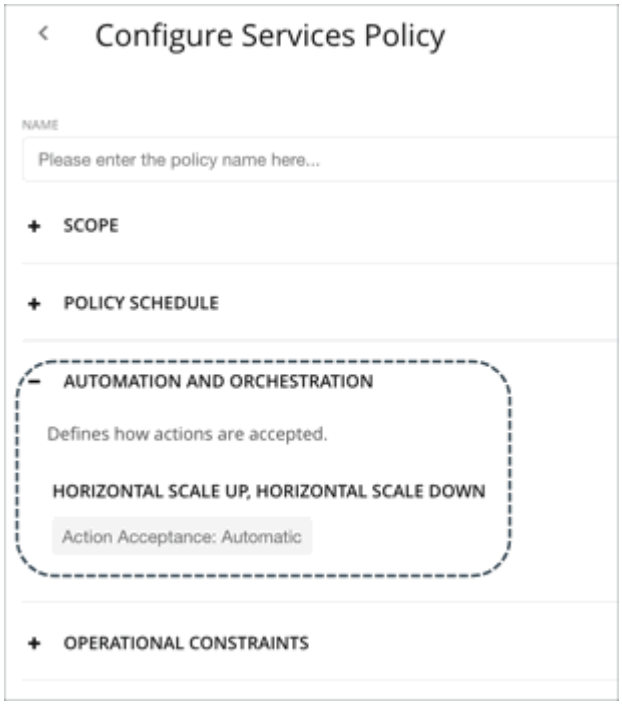
NOTE:
When a group of Services applies multiple conflicting policies, Workload Optimization Manager enforces the policy with the most conservative settings.

Setting	User Interface	Kubernetes Cluster CR Files
Scope	Specify a group of Kubernetes Services that will apply the policy.	

	Select a group of Services.	Configure the following in the `ServicePolicyBinding` CR file:
		`metadata:`
		`  name: <Group_Name>`
		`policyRef:`
		`  kind: SLOHorizontalScale`
		`  name: <Policy_Name>`
		`targets:`

<table>
<tr><th>Set-ting</th><th>User Interface</th><th>Kubernetes Cluster CR Files</th></tr>
<tr><td></td><td> 
</td><td><pre>- apiVersion: apps/v1
  kind: Service
  name: <Service_Name_1>
- apiVersion: apps/v1
  kind: Service
  name: <Service_Name_2></pre>
Where:
▪ <Group_Name> is the name of the group of Services that will apply the policy. In the user interface, this group will be added to Search > Groups and will be shown in the format <Name-space>/<Group_Name>[Target_-Name].
▪ <Policy_Name> is the policy name defined in the ServicePolicy CR file. The name appears as follows:
<pre>metadata:
  name: <Policy_Name></pre>
In the user interface, this policy will be added to Settings > Policies.
▪ Service_Name_1 and Service_-Name_2 are the Services belonging to the group. Add more entries as need-ed. You can specify actual names or use regular expressions.</td></tr>
<tr><td>Policy name</td><td colspan="2">Give the policy a name.</td></tr>
<tr><td></td><td>Type a name.
 
</td><td>Configure the following in the ServicePoli-cy CR file:
<pre>metadata:
  name: <Policy_Name></pre></td></tr>
</table>

<table>
<tr><th>Setting</th><th>User Interface</th><th>Kubernetes Cluster CR Files</th></tr>
<tr><td rowspan="2">Horizontal scaling actions</td><td colspan="2">Choose from the following:
Turn on scale up and scale down.
Turn on scale up, turn off scale down.
Turn on scale down, turn off scale up.
To turn on, set the action mode to Automatic, Manual, or Recommend. To turn off, set the action mode to Disabled. See Action Modes (on page 72) for details.

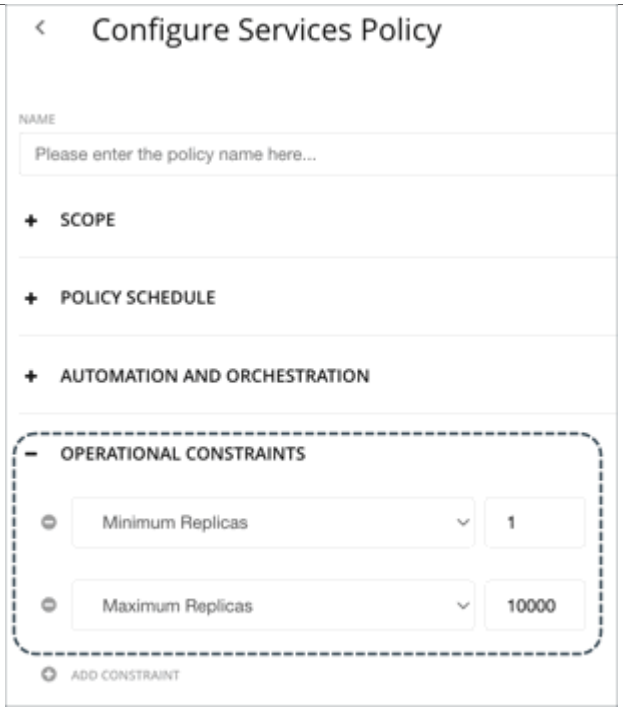
NOTE:
If you turned off horizontal scaling but specified SLOs values in the policy, actions do not generate but SLO values will continue to display in the Response Time and Transaction charts for Services, for your reference. This allows you to gauge performance against those SLOs.</td></tr>
<tr><td>Specify the scale actions and their action modes.
 
</td><td>Configure the following in the <code>ServicePolicy</code> CR file:
<pre>behavior:
  scaleUp: <Action_Mode>
  scaleDown: <Action_Mode></pre></td></tr>
<tr><td rowspan="2">SLOs</td><td colspan="2">Set SLOs and your desired SLO values.
Response Time SLO
Response Time SLO is the desired weighted average response time (in milliseconds) of all Application Component replicas associated with a Service.
Transaction SLO
Transaction SLO is the maximum number of transactions per second that each Application Component replica can handle.
If an SLO value is not specified, the policy will use the default value in the global policy for Services.</td></tr>
<tr><td>Enable Response Time and/or Transaction SLO, and then specify SLO values.</td><td>Configure the following in the <code>ServicePolicy</code> CR file:</td></tr>
</table>

Setting	User Interface	Kubernetes Cluster CR Files

		`objectives:`
		`- name: ResponseTime`
		`value: <ResponseTimeSLO>`
		`- name: Transaction`
		`value: <TransactionSLO>`
		
		To disable a particular SLO, remove the `name-value` pair from the file.

(Optional) 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.	

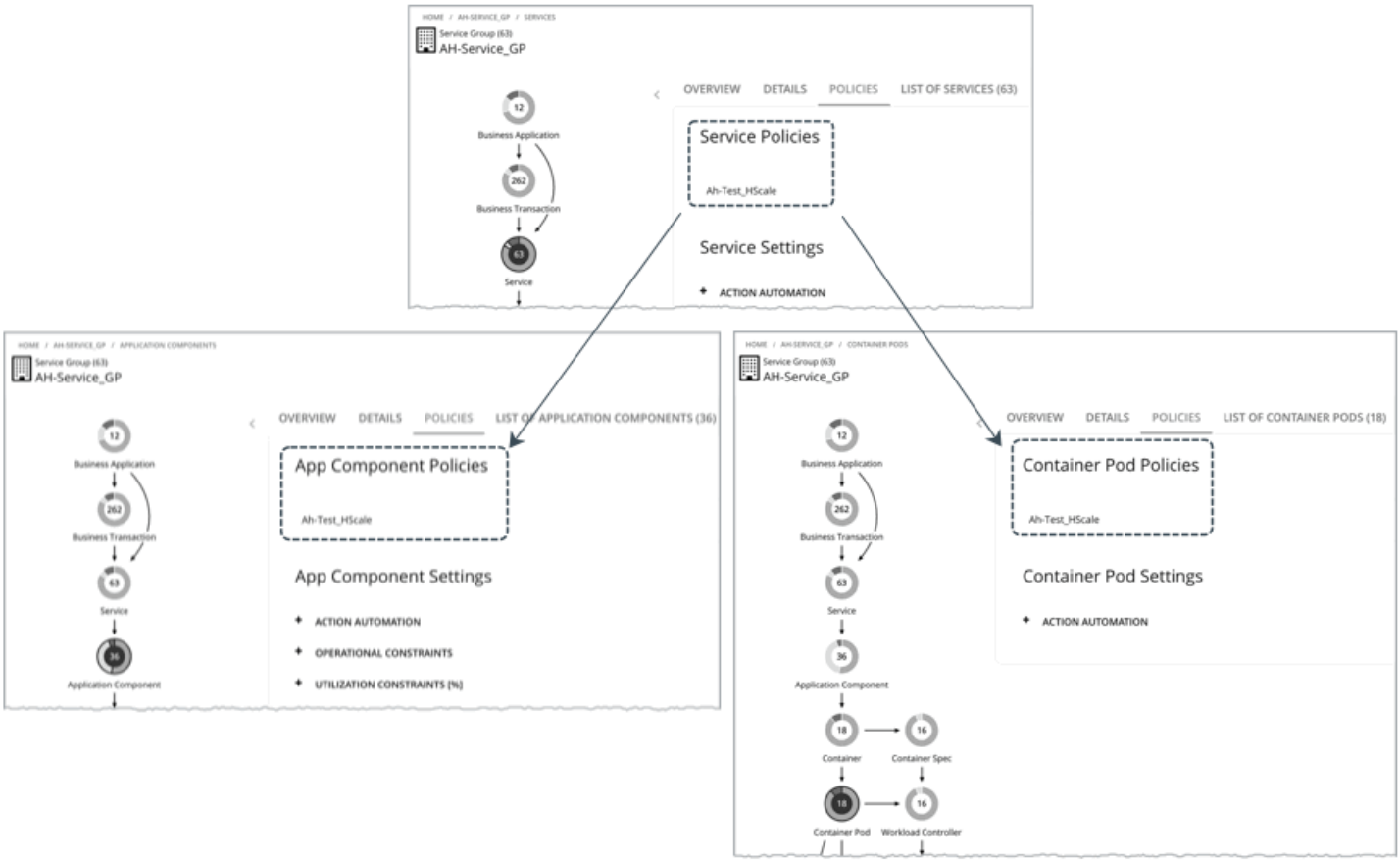
	Specify your desired values.	Configure the following in the `ServicePolicy` CR file:
		
		`spec:`
		`minReplicas: <minReplicas>`
		`maxReplicas: <maxRelicas>`

Set-ting	User Interface	Kubernetes Cluster CR Files
		

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.

Action Automation Considerations

We recommend action automation under the following circumstances:

- Your applications run as a set of Kubernetes 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 Kubernetes HPA (Horizontal Pod Autoscaling) enabled for the same workload.

We recommend 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.

Service Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Policies for Non-Kubernetes Services

- **Action Automation and Orchestration**

  Workload Optimization Manager does not recommend actions for non-Kubernetes Services, 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.

- **Transaction SLO**

  Enable this SLO if you are monitoring performance through Services.

Attribute	Default Setting/Value
Enable Transaction SLO	Off Workload Optimization Manager estimates SLO based on monitored values.
Transaction SLO	None If you enable SLO, Workload Optimization Manager 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, Workload Optimization Manager sets the risk index to 100%.

- **Response Time SLO**

  Enable this SLO if you are monitoring performance through Services.

Attribute	Default Setting/Value
Enable Response Time SLO	Off Workload Optimization Manager estimates SLO based on monitored values.
Response Time SLO [ms]	None If you enable SLO, Workload Optimization Manager uses the default value of 2000. 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, Workload Optimization Manager sets the risk index to 100%.

Policies for Kubernetes Services

These are the policy settings that you configure from the user interface. Workload Optimization Manager also supports policy settings configured in CR files in your Kubernetes clusters. For details, see Actions for Kubernetes Services *(on page 131)*.

- **Action Automation and Orchestration**

  For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications.

  To generate these actions, you must turn on horizontal scaling (up and/or down) and specify your desired SLOs in a Service policy.

  For details about these actions and the environments that are suitable for automating these actions, see Actions for Kubernetes Services *(on page 131)*.

Attribute	Default Setting/Value
Horizontal Scale Down	Off (Disabled)
Horizontal Scale Up	Off (Disabled)

- **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, Workload Optimization Manager 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, Workload Optimization Manager 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.

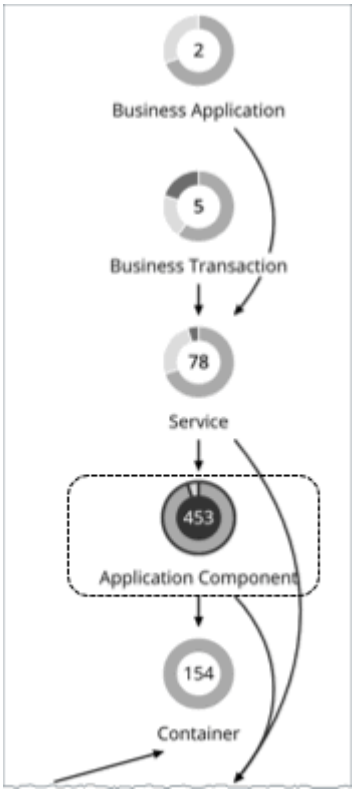
Attribute	Default Setting/Value
Minimum Replicas	1
Maximum Replicas	10000

Application Component

An Application Component is a software component, application code, or a unit of processing within a Service *(on page 129)* that consumes resources to enable it to perform its function for the Business Application *(on page 123)*. For example, Apache Tomcat is a Java Servlet container that hosts a range of Java applications on the web.

Workload Optimization Manager can recommend actions to adjust the amount of resources available to Application Components.

Synopsis



Synopsis	
Budget:	Application Components have unlimited budget.
Provides:	▪ Response Time and Transactions to Services, Business Transactions *(on page 126)*, and Business Applications ▪ Response Time, Transactions, Heap, Remaining GC Capacity, and Threads to end users
Consumes:	Compute resources from nodes
Discovery:	Workload Optimization Manager discovers the following: ▪ Apache Tomcat ▪ AppDynamics Nodes ▪ Dynatrace Processes ▪ NewRelic APM Application Instances ▪ Application Insights Components ▪ SNMP ▪ WMI ▪ Data Ingestion Framework metrics for Kubernetes environments

Monitored Resources

The exact resources monitored will differ based on application type. This list includes all resources you may see.

- Virtual CPU

  Virtual CPU is the measurement of CPU utilized by the entity.
- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.
- Transactions

  Transaction is a value that represents the per-second utilization of the transactions 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 Server 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, Workload Optimization Manager 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 Kubernetes 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 Actions for Kubernetes Services *(on page 131)*.

Actions

Resize

Resize the following resources to maintain performance:

- Thread Pool

  Workload Optimization Manager generates thread pool resize actions. These actions are recommend-only and can only be executed outside Workload Optimization Manager.
- Connections

  Workload Optimization Manager uses connection data to generate memory resize actions for on-prem Database Servers.
- Heap

Workload Optimization Manager 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 Workload Optimization Manager.

NOTE:
Remaining GC capacity is the measurement of Application Component uptime that is *not* spent on garbage collection (GC).

The resources that Workload Optimization Manager can resize depend on the processes that it discovers from your Applications and Databases targets. Refer to the topic for a specific target to see a list of resources that can be resized.

Application Component Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about Application Component actions, see Application Component Actions. *(on page 143)*

Action	Default Mode
Resize heap (up or down)	Recommend
Resize thread pool (up or down)	Recommend
Resize connections (up or down)	Recommend

You can use Action Scripts *(on page 112)* for action orchestration. Third-party orchestrators (such as ServiceNow) are not supported.

Transaction SLO

Enable this SLO to monitor the performance of your Application Components.

NOTE:
In Kubernetes 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 Actions for Kubernetes Services *(on page 131)*.

Attribute	Default Setting/Value
Enable Transaction SLO	Off Workload Optimization Manager estimates SLO based on monitored values.
Transaction SLO	None

Attribute	Default Setting/Value
	If you enable SLO, Workload Optimization Manager 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, Workload Optimization Manager sets the risk index to 100%.

Response Time SLO

Enable this SLO to monitor the performance of your Application Components.

> **NOTE:**
> In Kubernetes 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 Actions for Kubernetes Services *(on page 131)*.

Attribute	Default Setting/Value
Enable Response Time SLO	Off Workload Optimization Manager estimates SLO based on monitored values.
Response Time SLO [ms]	None If you enable SLO, Workload Optimization Manager uses the default value of 2000. 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, Workload Optimization Manager sets the risk index to 100%.

Heap Utilization

The Heap utilization that you set here specifies the percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity. For example, a value of 80 means that Workload Optimization Manager considers 80% utilization to be 100% of capacity.

Attribute	Default Value
Heap Utilization (%)	80

Workload Optimization Manager 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, Workload Optimization Manager will not recommend Heap scaling.

If Heap utilization is low and Remaining GC Capacity is high, Workload Optimization Manager will recommend resizing down Heap. If the opposite is true, then Workload Optimization Manager 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.

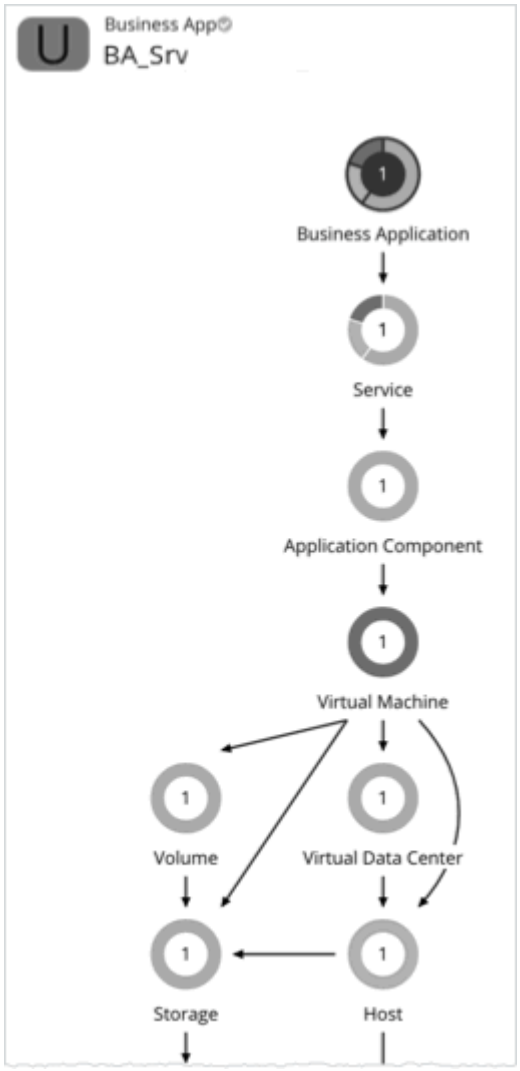
Attribute	Default Value
Heap Scaling Increment (MB)	128

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, Workload Optimization Manager will not leave an Application Component with less than the increment value. For example, if you use the default 128, then Workload Optimization Manager cannot reduce the Heap to less than 128 MB.

Application Topology

Workload Optimization Manager gives you the ability to create your own Business Applications *(on page 123)*, Business Transactions *(on page 126)*, and Services *(on page 129)* without the need to ingest additional application data into the platform. This is especially useful in environments where there are gaps in the application stack shown in the Workload Optimization Manager supply chain. For example, in the absence of an application monitoring target such as AppDynamics or Dynatrace, you will 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 and nodes (i.e., the infrastructure that backs the application entities) in your environment for which you want to measure performance. Workload Optimization Manager 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, and drill down to the individual entities and nodes for finer details.



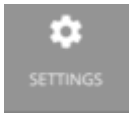
Workload Optimization Manager does not perform analysis on any user-created application entity, but it aggregates the underlying risks the same way it does for auto-discovered entities.

After you create an application entity, Workload Optimization Manager counts it in the global supply chain and adds it to the relevant charts (for example, if you created a new Service that has performance risks, you might see it listed in the Top Services chart). Drill down to the newly created entity to monitor its performance. You can also use Search to find the application entity and set it as your scope.

NOTE:
It could take up to 10 minutes to see newly created entities in the supply chain.

Creating Application Entities

1. Navigate to the **Settings** Page.



2. Choose **Application Topology**.

3. Click **New Application Topology** and then choose *Automatic* or *Manual*.
   - Automatic

     Create a new application entity composed of tagged entities. For example, create a new Business Application composed of VMs with the "Production" tag.

     a. Select the application entity type that you want to create.

     b. Type an entity name prefix to help you easily identify the application entities that Workload Optimization Manager will create for you.

     c. Specify the tags that will identify the underlying entities.
   - Manual

     Create a new application entity composed of a specific set of application entities and nodes.

     a. Select the application entity type that you want to create.

     b. Give the application entity a name.

     c. Select the underlying application entities and nodes.

     d. Enable or disable **Direct Link**.

        – Disabled (default)

          When **Direct Link** is disabled, Workload Optimization Manager creates a context-based definition of the application entity you are creating and automatically updates that definition as the entity evolves. This allows you to create flexible definitions with minimal effort.

          The underlying application entities and nodes that you specified act as "seed entities" for creating the definition. Workload Optimization Manager 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 could impact performance, making it more time-consuming to identify and resolve performance issues outside of the selected scope. With **Direct Link** disabled, Workload Optimization Manager might discover Application Components and VMs (leaf entities) that back the Services, and then show them in the supply chain. The result is a more 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, Workload Optimization Manager automatically updates the definition so your supply chain view remains current.

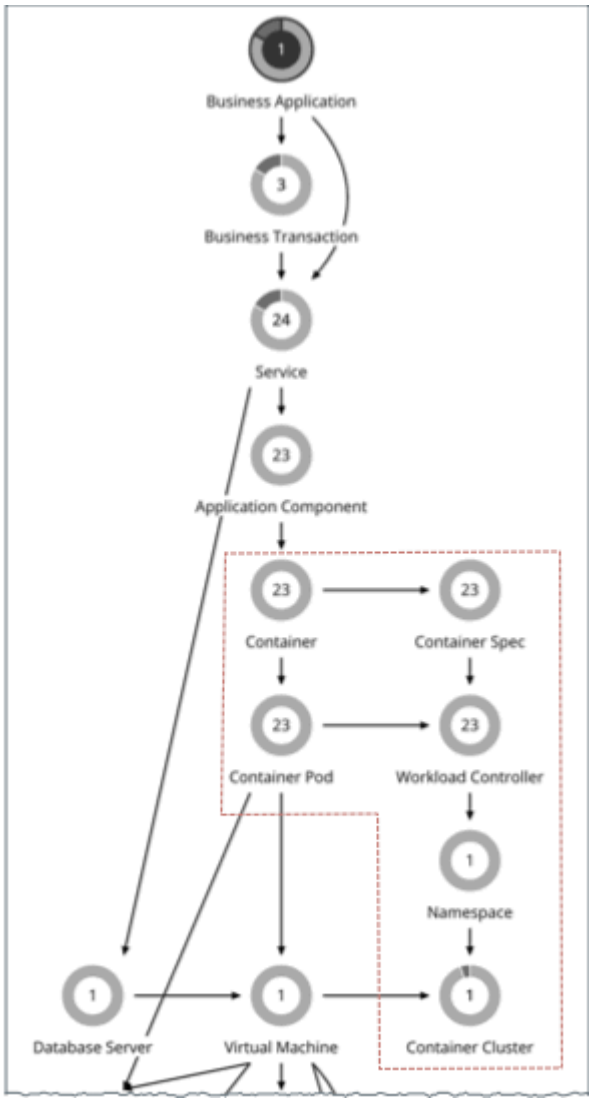
        – Enabled

          When **Direct Link** is enabled, Workload Optimization Manager 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

Workload Optimization Manager 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.



For a Cloud Native environment, Workload Optimization Manager discovers:

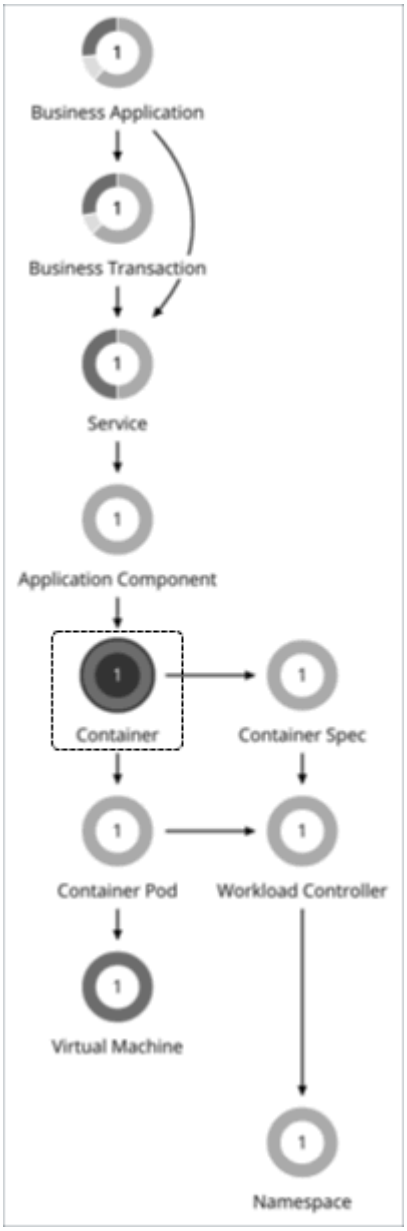
Entity Type	Kubernetes Object or Reference	Notes
Service *(on page 129)*	Service	A logical set of pods that represents a given application. In Kubernetes, the Service exposes a single entry point for the application process. While the Pods that comprise the service are ephemeral, the service is persistent. The Service entity also gives historical tracking of the number of replicas that run to support the Service.
Container *(on page 151)*	Container	The individual containers that deploy in your environment. Because the container instances that support a service can change at any time, these are considered *ephemeral*.
Container Pod *(on page 165)*	Pod	These are the smallest deployable units of computing that you can create and manage in Kubernetes. One Container Pod can contain multiple Container entities. These are also considered *ephemeral*.
Container Spec *(on page 158)*	A container's Spec	Persistent entities that collect containers with like properties. In Kubernetes the container's Spec includes the size specifications of limits and requests. In the Workload Optimization Manager supply chain, the count of replicas maps to the count of Container entities that a Container Spec encompasses. In Workload Optimization Manager, the persistent Container Spec maintains historical data for its ephemeral containers, and all the replicas that have run in the past.
Workload Controller *(on page 163)*	Controller	A persistent entity that maps to the different controllers in your Kubernetes environment, such as Deployments or Stateful Sets. A single Workload Controller can contain one or more Container Spec entities, and it can be related to one or more running replica pods. In the Supply Chain, the Workload Controller exposes the impact of Namespace quotas on Container Spec resize actions. The Workload Controller aggregates resize actions for the containers that are in its supply chain. In this way, a single action on a Workload Controller can encompass multiple Container actions.

Entity Type	Kubernetes Object or Reference	Notes
Namespace *(on page 173)*	Namespace	A logical group of workloads each namespace must be unique within a given Container Cluster. You can specify Resource Quotas for a Namespace, which limit compute resource capacity available to its workloads. Workload Optimization Manager will block execution of resize actions that would exceed Namespace quotas, and identify the quota increase you need to accommodate the workload resize. For Red Hat OpenShift, a Namespace is equivalent to a Project.
Container Cluster *(on page 176)*	Cluster	A collection of VMs (referred to as Nodes in Kubernetes). The Container Cluster scope aggregates actions so you can see cluster health in one view. This gives you an idea of cluster health from the perspective of your workloads.
Virtual Machine (Kubernetes Node) *(on page 180)*	Node	In Kubernetes environments, a node is a virtual or physical machine that contains the services necessary to run pods. Workload Optimization Manager represents nodes as Virtual Machine entities in the supply chain. Workload Optimization Manager 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. Workload Optimization Manager also discovers and displays Node Pools, and Red Hat OpenShift Machine Sets.
Volume *(on page 242)*	PV	If a Container Pod is attached to a volume, Workload Optimization Manager discovers it as a Persistent Volume (PV), and shows which Pods are connected to the PV.

Container

An application container is a standalone, executable image of software that includes components to host an application.

Synopsis



Synopsis	
Budget:	A container obtains its budget by selling resources to the hosted application.
Provides:	Resources for the applications to use: ▪ Virtual CPU ▪ Virtual Memory
Consumes:	Resources from container pods, virtual machines, and virtual datacenters.
Discovered through:	For Kubernetes, Workload Optimization Manager discovers containers through the Kubeturbo pod that you have deployed in your environment.

Synopsis	
	For Dynatrace and AppDynamics hosted on containers: ▪ Dynatrace: Workload Optimization Manager discovers containers through the metadata of processes. ▪ AppDynamics: Workload Optimization Manager discovers containers through container objects.

Monitored Resources

Workload Optimization Manager monitors the following resources for a container:

VMem

The virtual memory utilized by the container against the memory limit (if no limit is set, then node capacity is used).

VMem Request

If applicable, the virtual memory utilized by the container against the memory request.

VCPU

The virtual CPU (in mCores) utilized by the container against the CPU limit (if no limit is set, then node capacity is used).

VCPU Request

If applicable, the virtual CPU (in mCores) utilized by the container against the CPU request.

VCPU Throttling

The throttling of container vCPU 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

Resize

Resize containers to assure optimal utilization of resources. By default, containers resize consistently, which allows all replicas of the same container for the same workload type to resize any resource consistently.

vCPU Limit Resizing Due to Throttling

For vCPU limit resizes, Workload Optimization Manager 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, and this is exactly how Kubernetes was designed to work. Learn more about CPU throttling here.

Especially for sudden throttling spikes, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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.

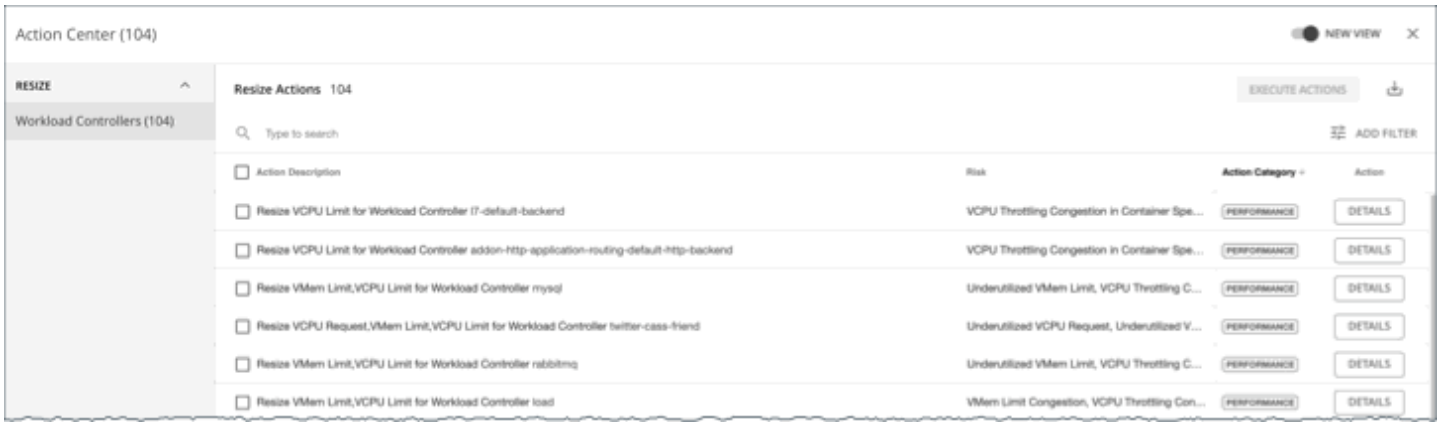
Action Visibility, Merging, and Execution

Workload Optimization Manager shows and executes container resize actions via Workload Controllers *(on page 163)*. You will *not* see actions when you set the scope to containers.

Actions also 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.

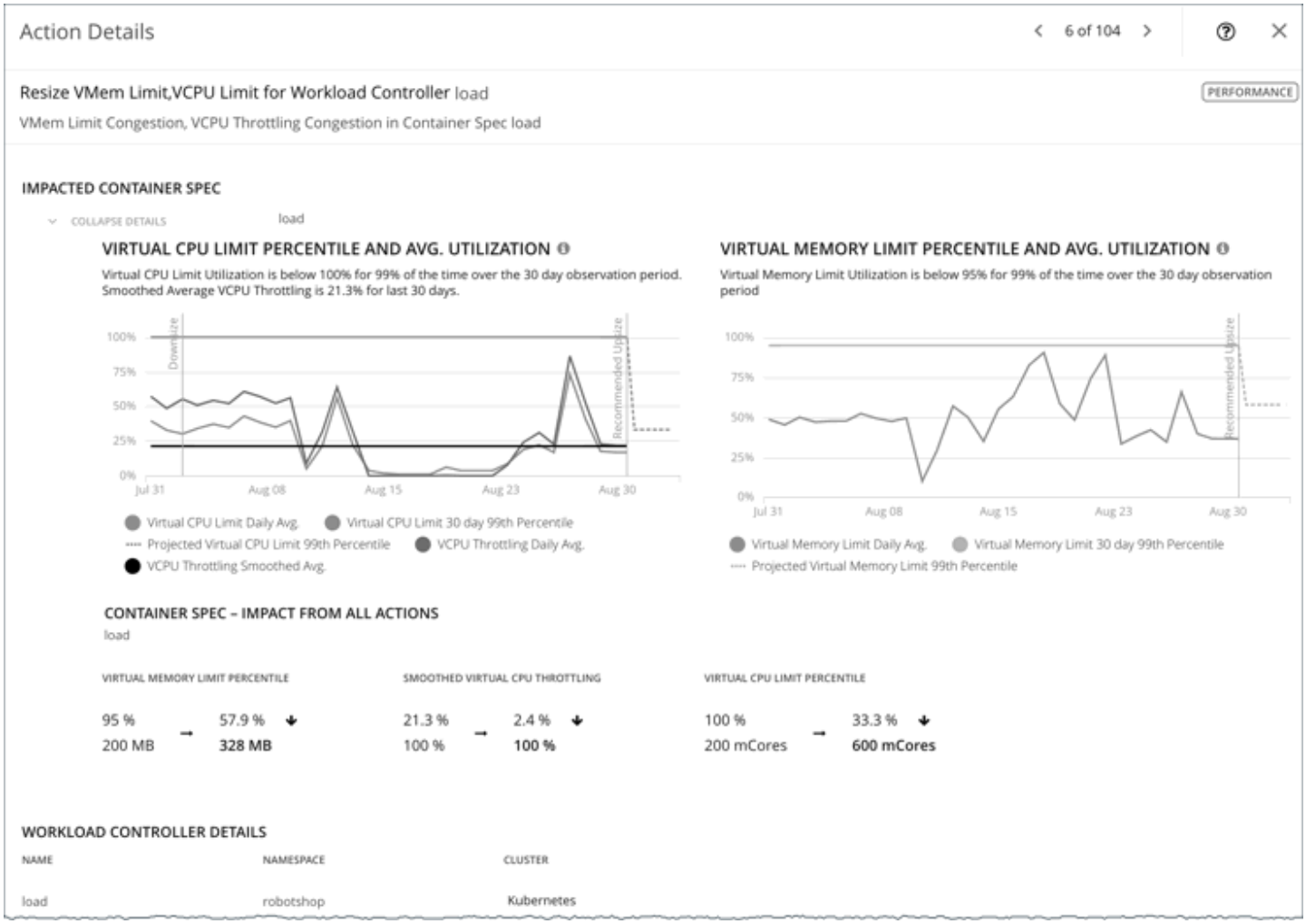
Executing several container resize actions can be very disruptive since pods need to restart with each resize. For replicas of the container scale group(s) related to a single Workload Controller, Workload Optimization Manager consolidates resize actions into one *merged action* to minimize disruptions. 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.

After you set the scope to Workload Controllers, go to the Pending Actions chart and then click **Show All** 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.



By default, container resize actions are set in *Manual* mode at the Workload Controller level. This means that Workload Optimization Manager 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 Workload Optimization Manager, create Workload Controller policies and set the resize action mode to *Recommend*. To automate actions, create Workload Controller policies and set the resize action 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. Workload Optimization Manager 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 Workload Optimization Manager's resize recommendations.

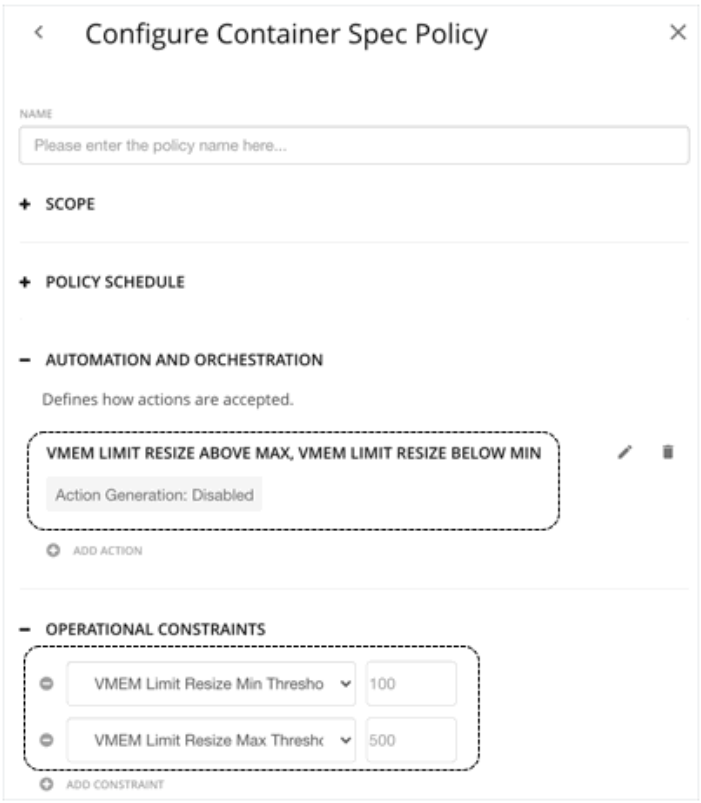
> **NOTE:**
> You can set scaling constraints in Container Spec policies to refine the percentile calculations. For details, see Aggressiveness and Observation Periods *(on page 161)*.

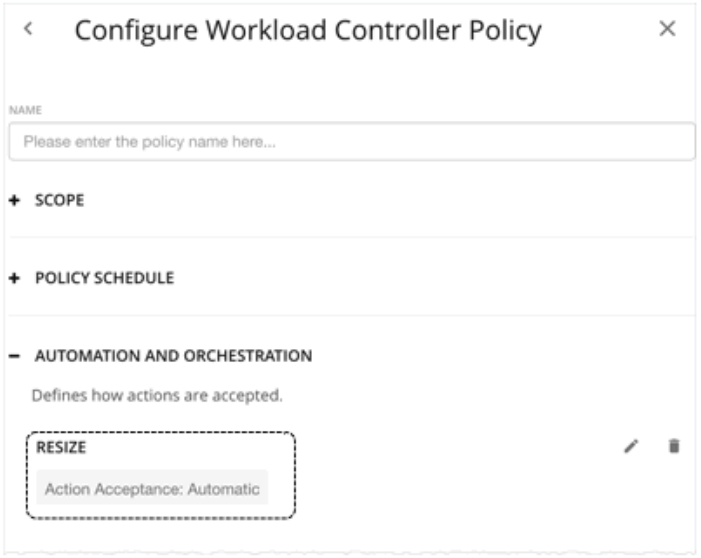
Tuned Scaling for Containers

Workload Optimization Manager can automate resizes if the resize values fall within a normal range, and then post more conservative actions when resize values fall outside the range. To do this, you would set specific *action modes* and *tuned scaling* values in policies.

For example, consider resizing vMem limits. As memory demand increases, Workload Optimization Manager can automatically execute vMem limit resizes that fall within the normal range. If the Container Spec requests memory beyond the normal range, Workload Optimization Manager will either ignore the action or post it for you to review, depending on the tuned scaling settings that you configured.

Assume the following tuned scaling settings in policies:





- The **Operational Constraints** settings in the Container Spec policy specify 100 MB to 500 MB as the normal range.
- With the **Resize** action mode in the Workload Controller policy set to *Automatic*, Workload Optimization Manager will automate *resize up* actions that are below the *max* threshold and *resize down* actions that are above the *min* threshold.

> **NOTE:**
> If the action mode is *Recommend*, Workload Optimization Manager will post the actions for you to review. You can only execute these actions outside Workload Optimization Manager.

- In the Container Spec policy, since the action mode for **vMem Limit Resize Above Max** and **vMem Limit Resize Below Min** is set to *Disabled*, Workload Optimization Manager will not generate resize actions that fall outside the normal range.
- Since vMEM increment constant is *not* defined in the Container Spec policy, Workload Optimization Manager uses the default value of 128 MB.

With these two policies in effect:

- If a Container Spec with 200 MB of vMEM limit needs to resize to 328 MB, Workload Optimization Manager automatically resizes to 328 MB.
- If a Container Spec with 200 MB of vMEM limit needs to resize to 72 MB, Workload Optimization Manager does not generate the action. vMEM limit remains at 200 MB.

NOTE:
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, Workload Optimization Manager uses the most conservative settings for the affected entities.

For tuned scaling, the effective action 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 Policy Scope *(on page 104)*.

Container Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

Resize

Resize containers to assure optimal utilization of resources. By default, containers resize consistently, which allows all replicas of the same container for the same workload type to resize any resource consistently.

Workload Optimization Manager shows and executes container resize actions via Workload Controllers *(on page 163)*. You will *not* see actions when you set the scope to containers.

Action	Default Mode	
	Container	Workload Controller
Resize	N/A	Manual (automatable)

For details, see Container Actions *(on page 153)*.

Consistent Resizing

- *For groups in scoped policies*:

Attribute	Default Setting
Consistent Resizing	Off

When you create a policy for a group of containers and turn on Consistent Resizing, Workload Optimization Manager 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, Workload Optimization Manager executes each resize individually in a way that avoids disruption to your workloads.

- *For auto-discovered groups*:

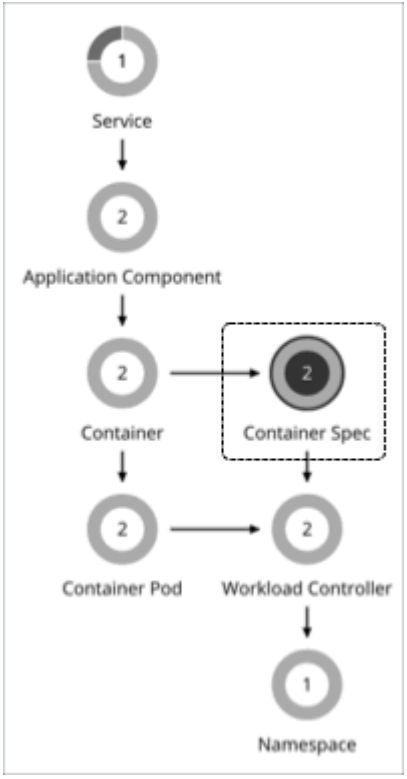
Workload Optimization Manager discovers Kubernetes groups such as Deployments, ReplicationControllers, ReplicaSets, DaemonSets, and StatefulSets, and automatically enables Consistent Resizing in a read-only policy for each group. If you do not need to resize all the members consistently, create another policy for the group and turn off Consistent Resizing.

Container Spec

A Container Spec is a shared definition for all ephemeral container replicas. It is a persistent entity that retains the historical utilization data of containers, which Workload Optimization Manager leverages to make container sizing decisions. Utilization data includes:

- vCPU used by all container replicas
- vCPU request capacity (if applicable)
- vMem used by all container replicas
- vMem request capacity (if applicable)

Synopsis



Synopsis	
Budget:	N/A
Provides:	N/A
Consumes:	N/A

Synopsis	
Discovered through:	Kubeturbo Mediation Pod

Monitored Resources

When you view the resources for a Container Spec, you will see the historical usage of any instance of a container running for the workload (assuming the workload name stays the same). The chart shows the trend of usage even with restarts or redeployments.

Actions

None

A Container Spec retains the historical utilization data of ephemeral containers. Workload Optimization Manager uses this data to make accurate container resize decisions, but does not recommend actions for the Container Spec itself.

> **NOTE:**
> To view container resize actions, set the scope to the Workload Controller for related containers. Go to the Pending Actions chart and click **Show All** to see the full list. For more information about container actions, see Container Actions *(on page 153)*.

Constraint for Sidecar Container Specs

A Kubernetes 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, Workload Optimization Manager adds them to a group called "Injected Sidecars/All ContainerSpecs". This group applies a read-only policy that sets the action mode for resizes to *Recommend*. This means that you can only execute resizes outside of Workload Optimization Manager. The parent Workload Controller will continue to resize non-sidecar Container Specs as usual.

Container Spec Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

The following settings affect tuned scaling:

Setting	Default Mode
vCPU Request Resize Below Min	Recommend
vCPU Limit Resize Above Max	Recommend
vCPU Limit Resize Below Min	Recommend

Setting	Default Mode
vMem Request Resize Below Min	Recommend
vMem Limit Resize Above Max	Recommend
vMem Limit Resize Below Min	Recommend

The default mode of *Recommend* means that when resize values in actions fall outside the normal range (as defined in Container Spec policies), Workload Optimization Manager will post the actions for you to review. You can only execute these actions outside Workload Optimization Manager. If you set the action mode to *Disabled*, Workload Optimization Manager will not generate the actions.

For an overview of tuned scaling, see Tuned Scaling for Containers *(on page 155)*.

Resize Thresholds

Workload Optimization Manager uses resize thresholds as operational constraints to set up tuned scaling for Container Specs. For an overview of tuned scaling, see Tuned Scaling for Containers *(on page 155)*.

Attribute	Default Value
VCPU Request Resize Min Threshold (mCores)	10
VCPU Limit Resize Min Threshold (mCores)	500
VCPU Limit Resize Max Threshold (mCores)	64000
VMEM Request Resize Min Threshold (MB)	10
VMEM Resize Min Threshold (MB)	10
VMEM Resize Max Threshold (MB)	1048576

Max CPU Throttling Tolerance (%)

This value defines your acceptable level of throttling and directly impacts the resize actions generated on CPU Limits.

Attribute	Default Value
Max CPU Throttling Tolerance (%)	20

A low percentage value indicates more sensitivity to throttling, while a high value indicates more tolerance for throttling and a higher risk of congestion. Workload Optimization Manager generates resize actions to accommodate the value that you defined. A default of up to 20% throttling will be tolerated unless defined otherwise.

Learn more about CPU throttling here.

Increment Constants

Workload Optimization Manager 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. Workload Optimization Manager could recommend to reduce the request by 100, down to 700 mCores.

For vMem, you should not set the increment value to be lower than what is necessary for the container to operate. If the vMem increment is too low, then it's possible that Workload Optimization Manager would allocate insufficient vMem. For a container that is underutilized, Workload Optimization Manager 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 Workload Optimization Manager cannot reduce the vMem to less than 128 MB.

Rate of Resize

(For the *default* policy only)

Attribute	Default Value
Rate of Resize	High

When resizing resources for a container, Workload Optimization Manager calculates the optimal values for vMem and vCPU. But it does not necessarily make a change to that value in one action. Workload Optimization Manager uses the Rate of Resize setting to determine how to make the change in a single action, as follows:

- **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, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager will raise vMem to 8 GB (or as close to that as the increment constant will allow).

Aggressiveness and Observation Periods

Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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 container. Without using a percentile, Workload Optimization Manager never resizes below the recognized peak utilization. For most containers there are moments when peak CPU reaches high levels. Assume utilization for a container peaked at 100% just once. Without the benefit of a percentile, Workload Optimization Manager will not reduce allocated CPU for that container.

With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager uses the percentile you set. For the above example, assume a single CPU burst to 100%, but for 99% of the samples CPU never exceeded 50%. If you set **Aggressiveness** to 99th Percentile, then Workload Optimization Manager can see this as an opportunity to reduce CPU 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 – The least aggressive, recommended for critical workloads that need maximum guaranteed performance at all times.
- 99th Percentile (Default) – The recommended setting to achieve maximum performance and savings.
- 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. Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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"

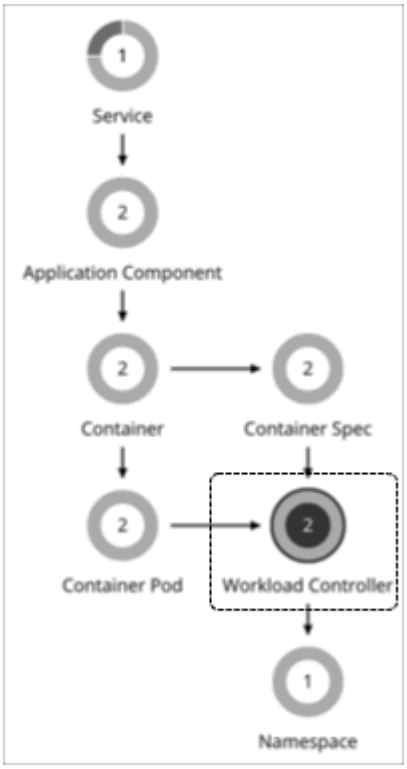
time, when utilization is low. If analysis uses enough historical data for an action, then the action is more likely to remain viable during the maintenance window.

- More Elastic – None
- Recommended – 1 Day
- Less Elastic – 3 or 7 Days

Workload Controller

A Workload Controller is a Kubernetes controller that watches the state of your pods and then requests changes where needed. You can execute container resize actions when you set the scope to a Workload Controller.

Synopsis



Synopsis	
Budget:	N/A
Provides:	N/A
Consumes:	N/A
Discovered through:	Kubeturbo Mediation Pod

Monitored Resources

Workload Optimization Manager does not monitor resources for Workload Controllers.

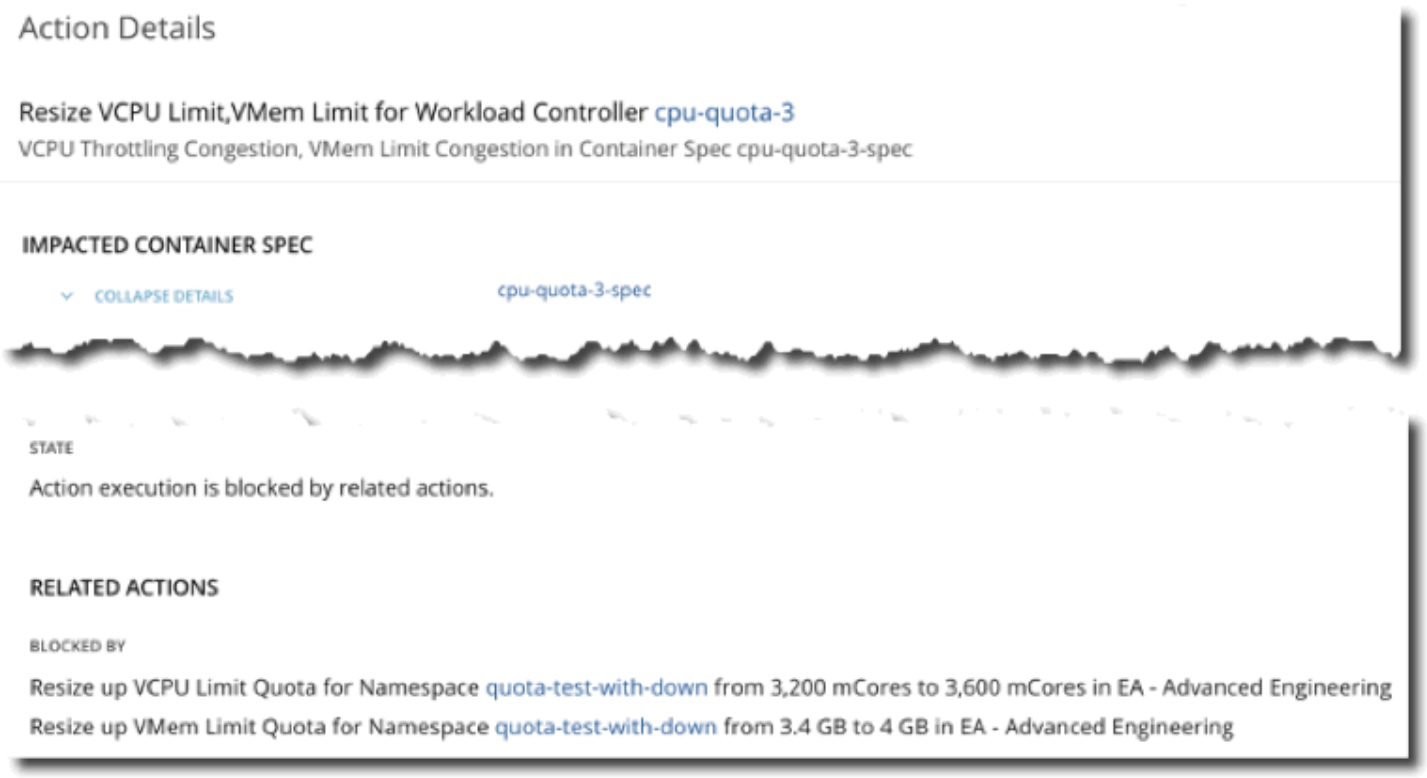
Actions

None

A Workload Controller executes container actions. When you set the scope to a Workload Controller and view the actions list, the actions apply to containers. Workload Optimization Manager does not recommend actions for the Workload Controller itself.

> **NOTE:**
> Workload Optimization Manager uses namespace or organization/space quotas as constraints when making resize decisions. The Workload Controller aggregates container actions. If those container resizes exceed current namespace quotas, Workload Optimization Manager blocks execution of container resize actions until the namespace quotas are sufficient. For more information about namespace quotas, see Resource Quotas *(on page 173)*.

For resize actions on a Workload Controller, the actions details include descriptions of the affected Container Spec entities, and how the resources will change for each. If the resize exceeds current namespace quotas, then Workload Optimization Manager blocks the Wokload Controller action. The action details list the Namespace actions that block execution of this resize in the **Related Actions** list.



For more information about container actions, see Container Actions *(on page 153)*.

Workload Controller Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

Workload Optimization Manager shows and executes container resize actions via Workload Controllers *(on page 163)*. You will *not* see actions when you set the scope to containers.

Action	Default Mode	
	Container	Workload Controller
Resize	N/A	Manual (automatable)

For details, see Container Actions *(on page 153)*.

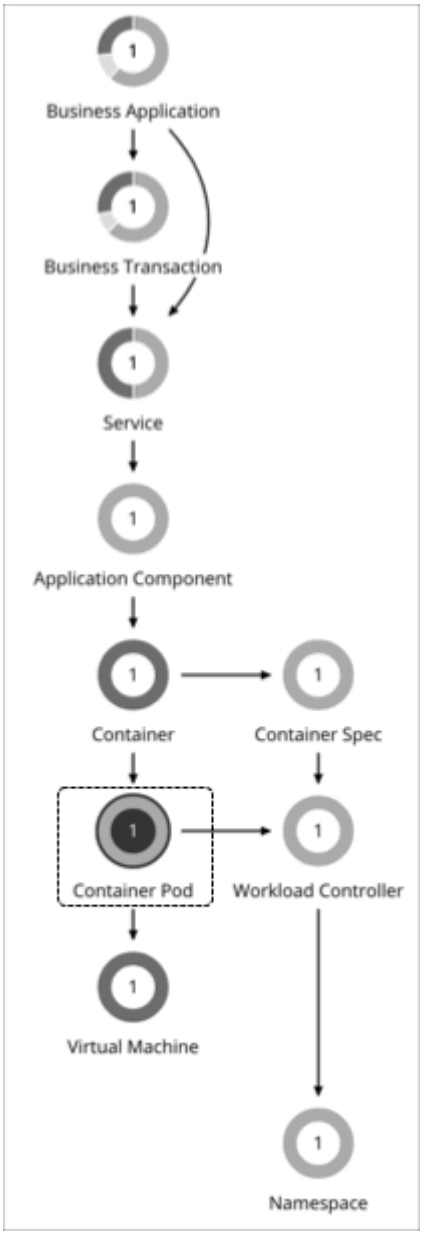
Executing several container resize actions can be very disruptive since pods need to restart with each resize. For replicas of the container scale group(s) related to a single Workload Controller, Workload Optimization Manager consolidates resize actions into one *merged action* to minimize disruptions. 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 orchestration is currently not supported.

Container Pod

A container pod is a Kubernetes pod, which is a group of one or more containers with shared storage or network resources and a specification for how to run the containers together.

Synopsis



Synopsis	
Budget:	A container pod obtains its budget by selling resources to containers.
Provides:	Resources for containers to use: ▪ Virtual CPU ▪ Virtual Memory
Consumes:	Resources from virtual machines and namespaces.
Discovered through:	Workload Optimization Manager discovers Kubernetes pods through the Kubeturbo pod that you have deployed in your environment.

Monitored Resources

Workload Optimization Manager monitors the following resources for a container pod:

VMem

The virtual memory utilized by the pod against the node physical capacity.

VCPU

The virtual CPU (in mCores) utilized by the pod against the node physical capacity.

VMem Request

The virtual memory request allocated by the pod against the node allocatable capacity.

VCPU Request

The virtual CPU (in mCores) request allocated by the pod against the node allocatable capacity.

VMem Request Quota

If applicable, The amount of virtual memory request the pod has allocated against the namespace quota.

VCPU Request Quota

If applicable, The amount of virtual CPU request (in mCores) the pod has allocated against the namespace quota.

VMem Limit Quota

If applicable, The amount of virtual memory limit the pod has allocated against the namespace quota.

VCPU Limit Quota

If applicable, The amount of virtual CPU limit (in mCores) the pod has allocated against the namespace quota.

Pod Move Actions

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.

The following items impact the generation and execution of pod move actions:

- Constraints

  Workload Optimization Manager respects the following constraints when making placement decisions for pods:

  - Kubernetes taints for nodes and tolerations for pods are treated as constraints. For example, if a pod has a toleration attribute that restricts it from moving to a certain node, Workload Optimization Manager will not move that pod to the restricted node.
  - Workload Optimization Manager imports Kubernetes node labels and treats them as constraints. For example, if a pod has a defined node label, Workload Optimization Manager will move that pod to a node with a matching label.
  - Workload Optimization Manager 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 88)*.

- Eviction Thresholds

  Workload Optimization Manager 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, Workload Optimization Manager temporarily increases the quota to allow a pod to move, while maintaining that one replica continues to run. You can disable temporary increases in quotas, but be aware that this will result in failure to move pods. To disable increases, set the following in the `yaml` resource for Kubeturbo deployment:

```
update-quota-to-allow-moves=false
```

- 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 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, we recommend adding 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 and Suspension Actions In Response to SLOs

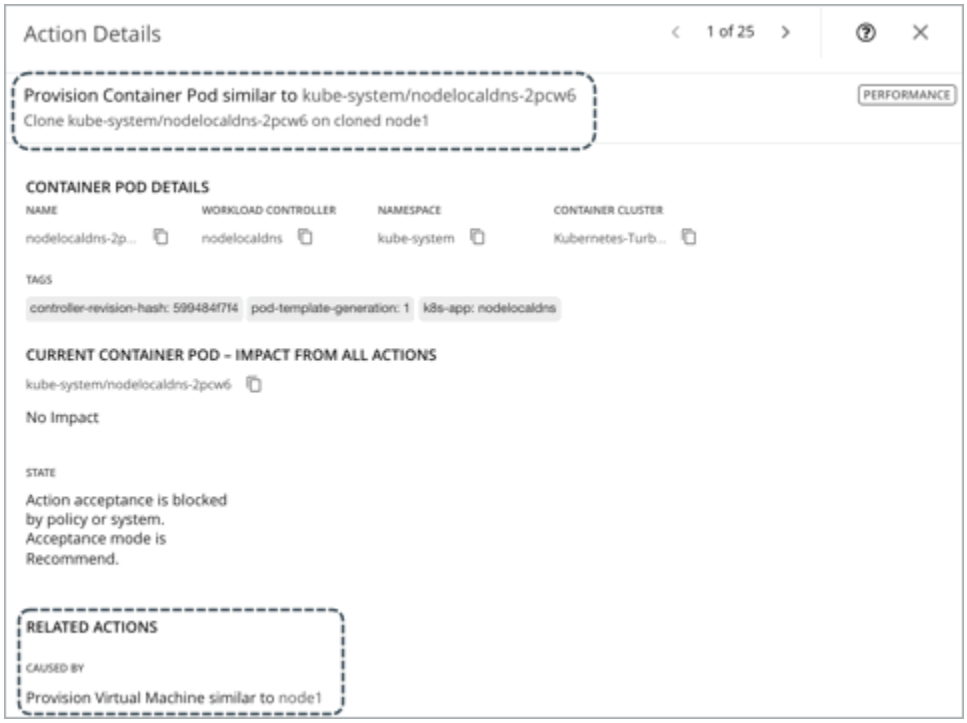
For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications.

For details, see Actions for Kubernetes Services *(on page 131)*.

Pod Provision Action in Response to Node Provision

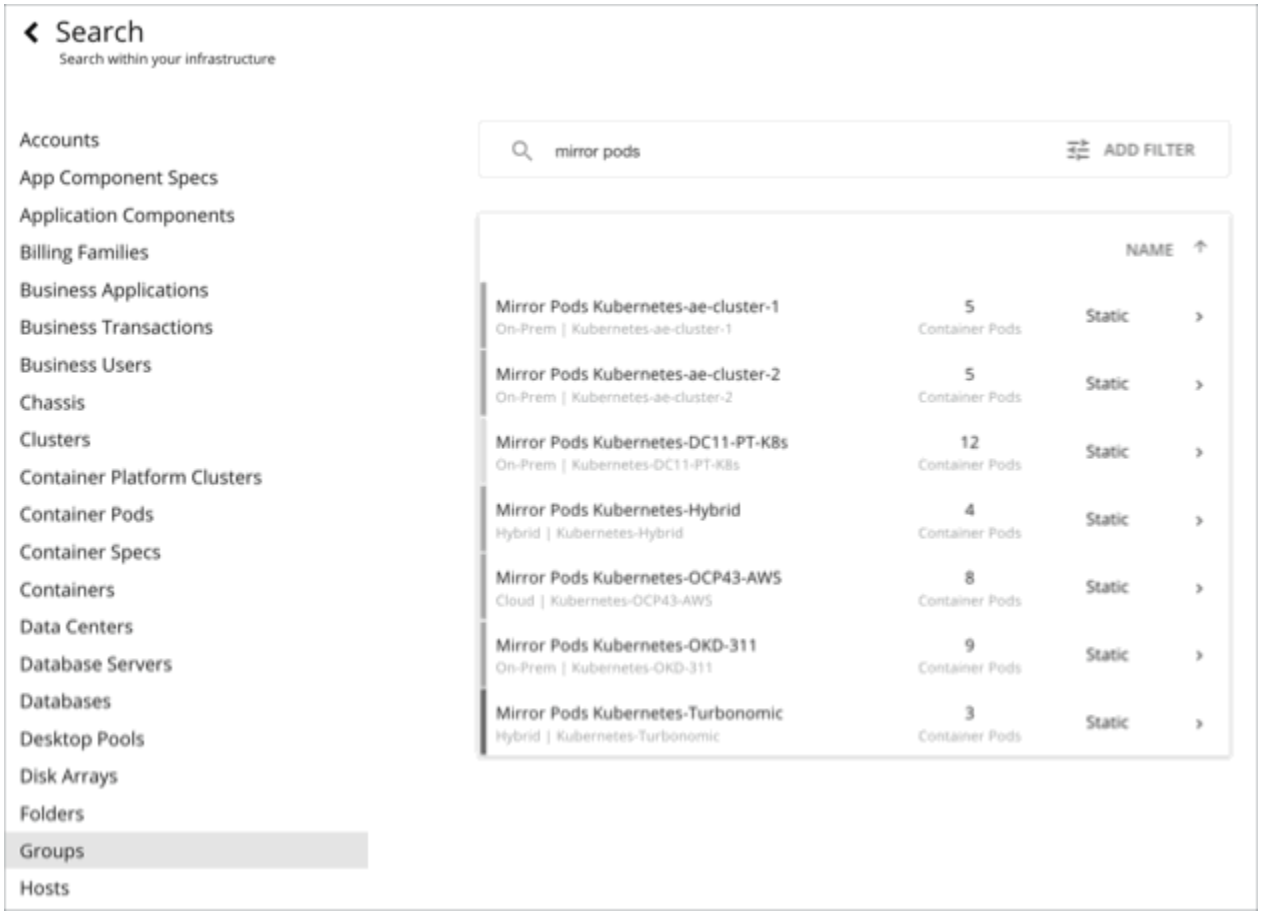
When recommending node provision actions, Workload Optimization Manager also recommends pod provision actions that reflect the projected demand from required DaemonSet pods, and respects the maximum number of pods allowed for a node. This ensures that any application workload can be placed on the new 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.



Workload Optimization Manager 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, Workload Optimization Manager generates actions to provision the node and the corresponding static pod.

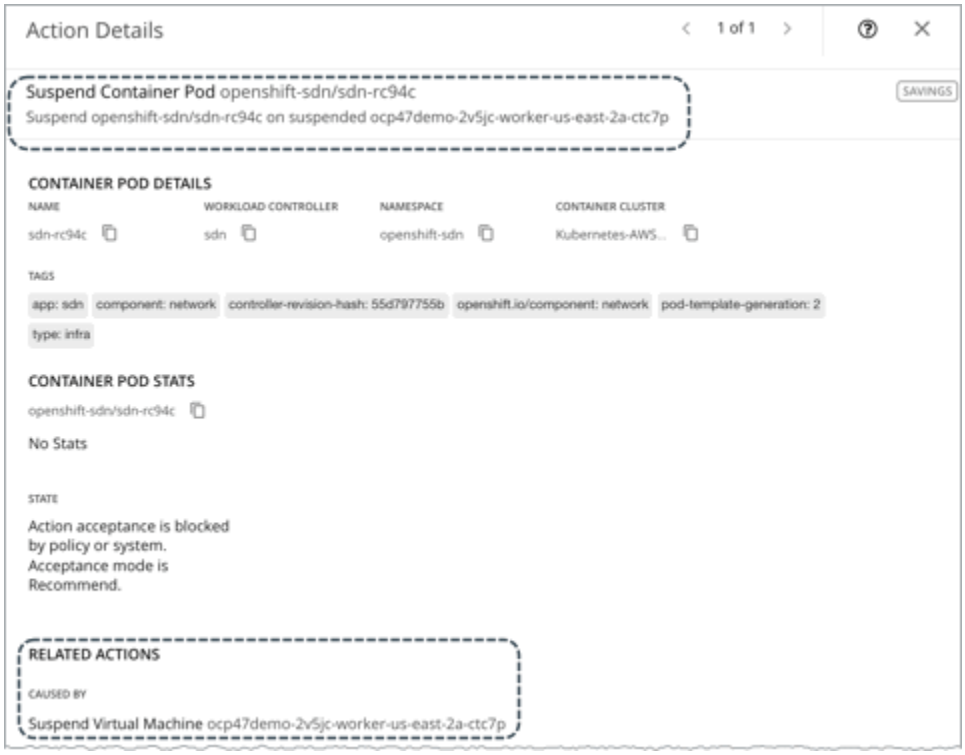
Workload Optimization Manager creates an auto-generated group of static pods when it discovers a static pod on each node in a cluster. To view all the auto-generated groups, go to Search, select Groups, and then type `mirror pods` as your search keyword.



Pod Suspension Action in Response to Node Suspension

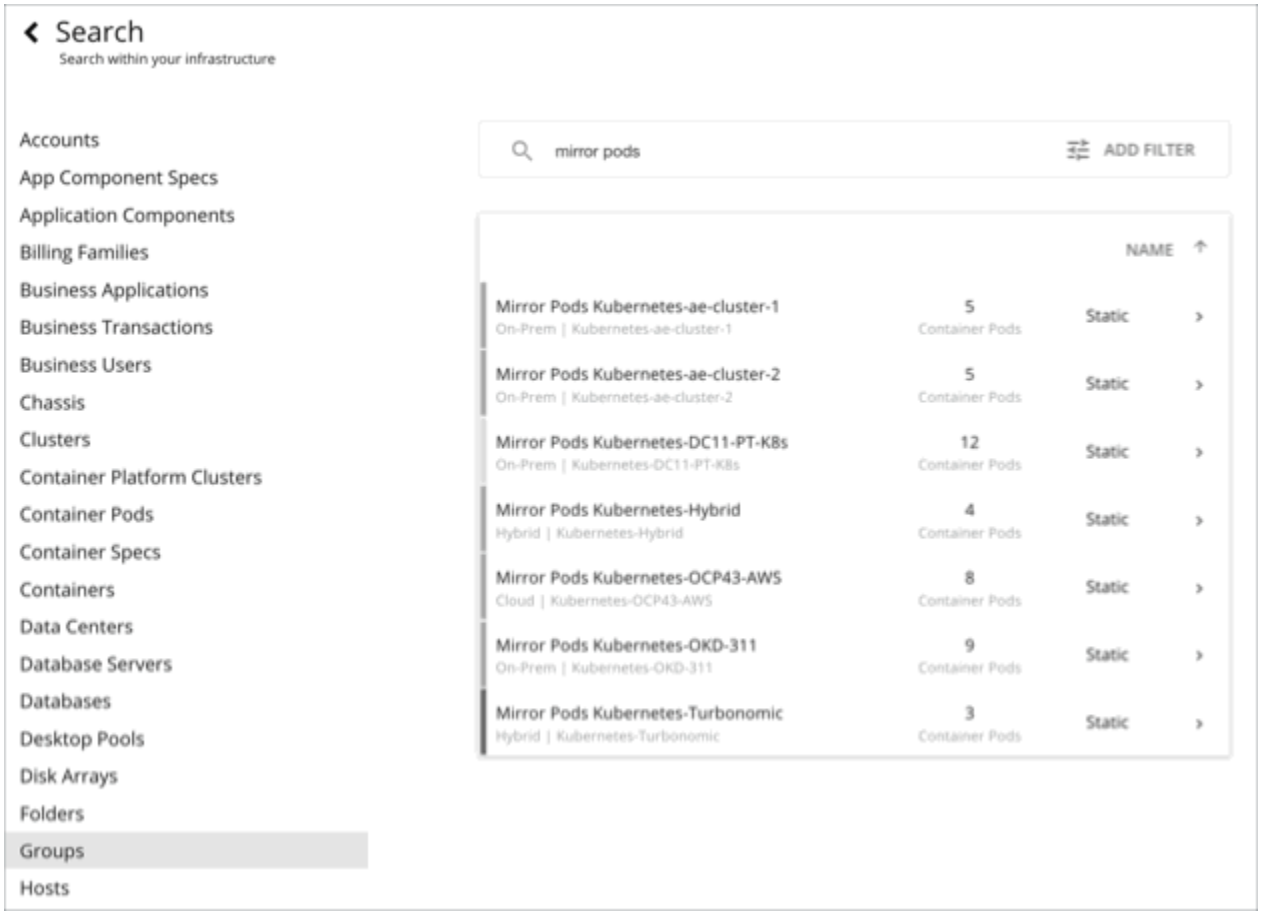
When recommending node suspension actions, Workload Optimization Manager also recommends suspending the DaemonSet pods that are no longer required to run the suspended nodes.

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.



Workload Optimization Manager 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, Workload Optimization Manager generates actions to suspend the node and the corresponding static pod.

Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about container pod actions, see Container Pod Actions *(on page 167)*.

Action	Default Mode
Move	Manual

Action orchestration is currently not supported.

Placement Policies

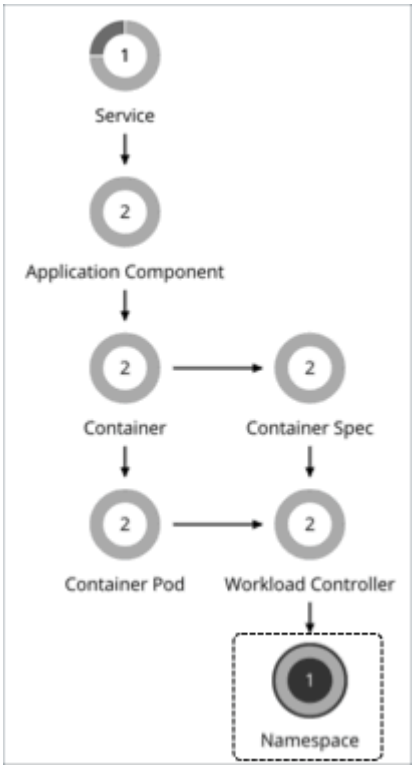
You can create placement policies to enforce constraints for pod move actions. For example, you can have a policy that allows pods to only move to certain nodes, or a policy that prevents pods from moving to certain nodes.

For more information, see Creating Placement Policies *(on page 88)*.

Namespace

A namespace is a logical pool of resources in a Kubernetes environment 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	
Budget:	N/A
Provides:	N/A
Consumes:	N/A
Discovered through:	Kubeturbo Mediation Pod

Resource Quotas

A namespace can include the following compute resource quotas:

VMem Request Quota

The total amount of virtual memory request for all pods allocated to the namespace against the namespace quota.

VCPU Request Quota

The total amount of virtual CPU request (in mCores) for all pods allocated to the namespace against the namespace quota.

VMem Limit Quota

The total amount of virtual memory limit for all pods allocated to the namespace against the namespace quota.

VCPU Limit Quota

The total amount of virtual CPU limit (in mCores) for all pods allocated to the namespace against the namespace quota.

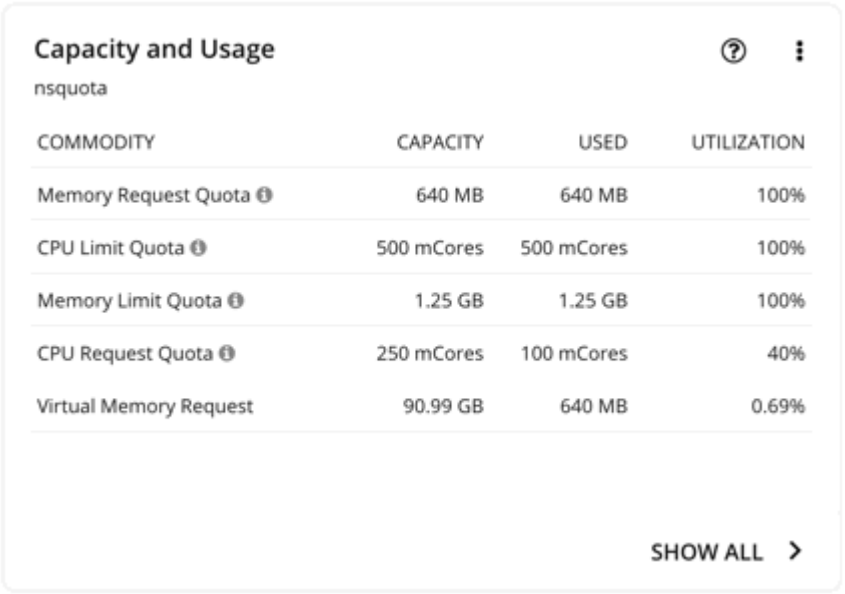
When they are configured, these quotas define the capacity for the given namespace. Workload Optimization Manager recognizes these quotas as it calculates actions in your environment.

If containers in the namespace require more compute resources, and those requirements exceed the namespace quotas, then Workload Optimization Manager recommends increasing the quotas. It will *block* execution of the underlying container actions until the namespace quotas are sufficient. In the details for a quota resize action, you can see the list of blocked container actions.

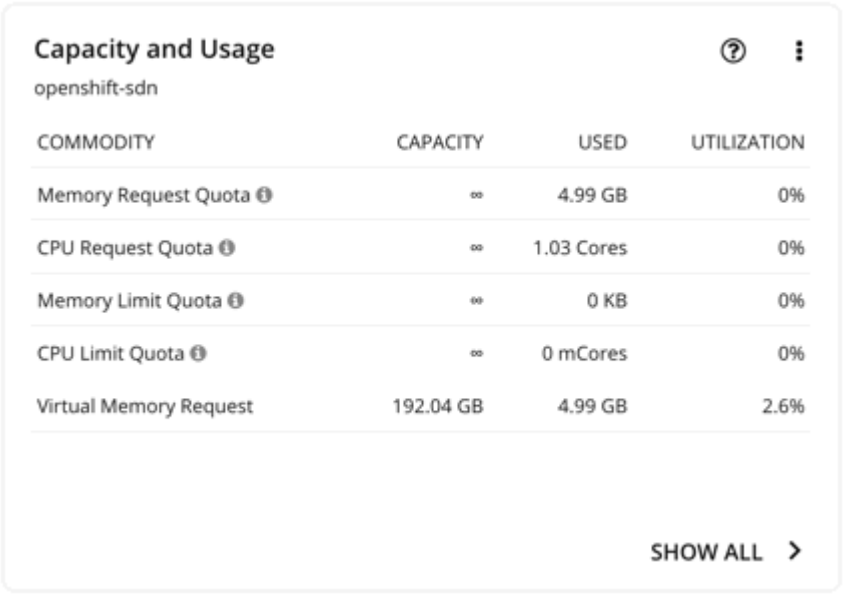
For more about actions to increase namespace quotas, see Actions *(on page 175)*.

When you run Optimize Container Cluster plans, Workload Optimization Manager can calculate increased namespace quotas in the plan results. For more information, see Optimize Container Cluster Plan *(on page 355)*.

Workload Optimization Manager treats quotas defined in namespaces as constraints when making sizing decisions for containers. When you scope to a namespace in the supply chain, 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.



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).



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).

Labels and Annotations

Workload Optimization Manager 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

Workload Optimization Manager monitors actual utilization of VMem, VCPU, VMem Requests and VCPU Requests against cluster capacity.

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.

To see which namespaces use the most cluster resources, set the scope to a container cluster and see the **Top Namespaces** chart. You can use the data in the chart for showback analysis.

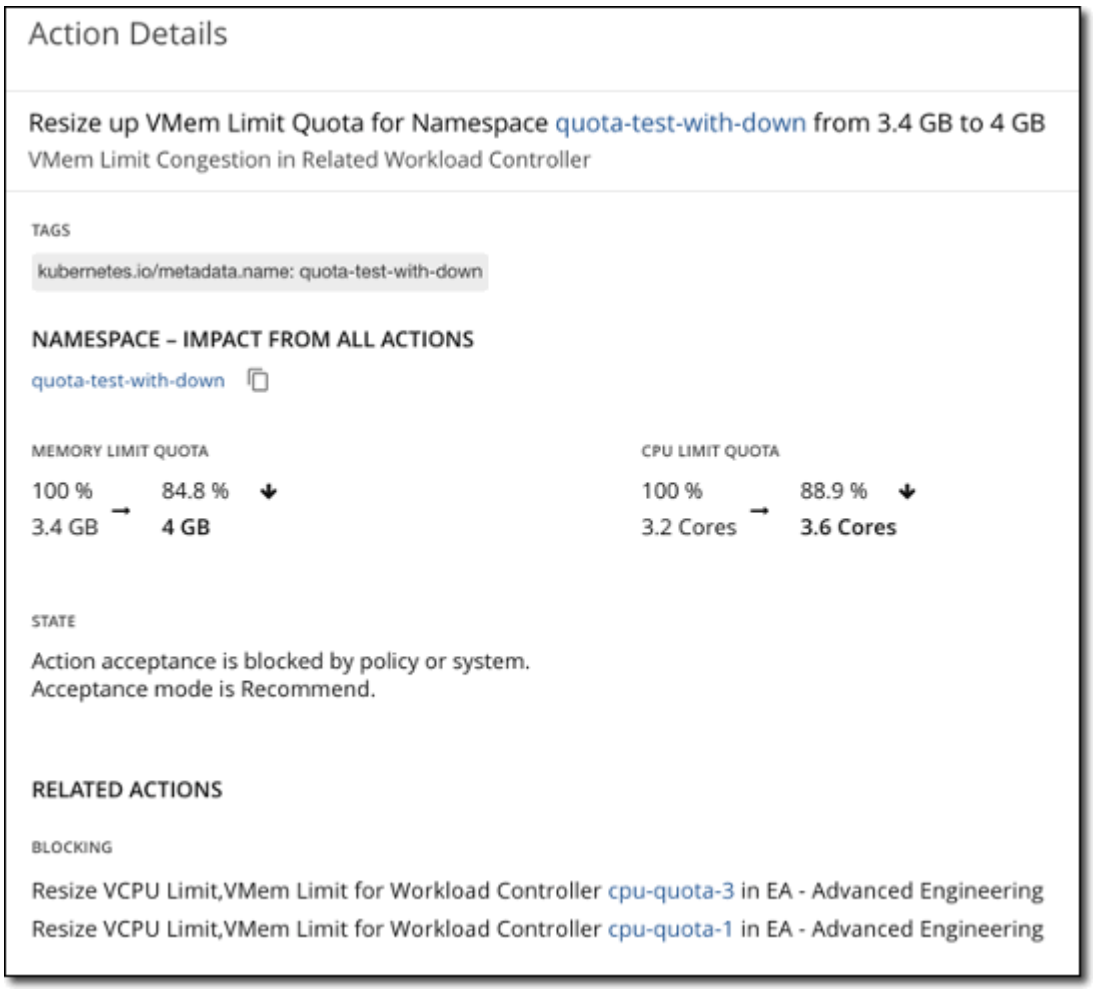
Namespace Actions

Resize Quota

Workload Optimization Manager treats quotas defined in a namespace as constraints when making container resize decisions. If existing container actions would exceed the namespace quotas, Workload Optimization Manager recommends actions to resize up the affected namespace quota.

Note that Workload Optimization Manager 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.

Workload Optimization Manager only recommends a resize for namespace quotas if underlying actions to resize containers require the increased quota. Note that Workload Optimization Manager aggregates container actions in Workload Controller entities. When you have a recommendation to resize namespace quotas, Workload Optimization Manager blocks execution of the resize actions for the affected Workload Containers. The action details show these blocked actions in the **Related Actions** list.



NOTE:
For more information about container resize actions, see Workload Controller Actions *(on page 164)* and Container Actions *(on page 153)*.

Temporary Increases in Namespace Quotas

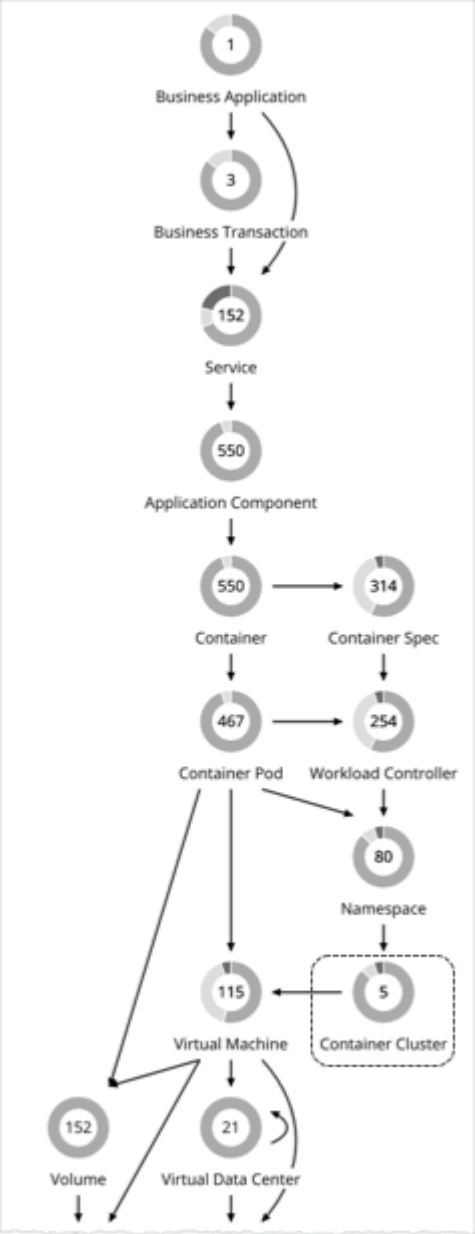
When executing a resize action on a Workload Controller in a namespace with a defined quota, Workload Optimization Manager can increase the namespace quota temporarily to accommodate new replicas.

When the execution of a resize action requires new replicas, a Kubernetes 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 Cluster

A Container Cluster is a Kubernetes cluster that Workload Optimization Manager discovers through Kubeturbo. With this entity type, Workload Optimization Manager can fully link the entire container infrastructure with the underlying nodes, and then present all actions on containers and nodes in a single view. This gives you full visibility into the actions that impact the health of your container environment.

Synopsis



Synopsis	
Budget:	N/A
Provides:	N/A
Consumes:	N/A
Discovered through:	Kubeturbo Mediation Pod

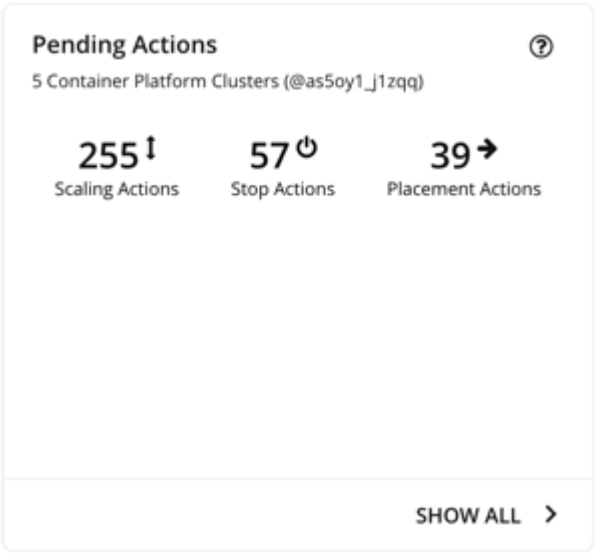
Monitored Resources

Workload Optimization Manager does not monitor resources for Kubernetes clusters. Instead, it monitors resources for the containers, pods, nodes (VMs), and volumes in the cluster.

Actions

None

Workload Optimization Manager does not recommend actions for a Container Cluster. Instead, it recommends actions for the containers, pods, nodes (VMs), and volumes in the cluster. Workload Optimization Manager shows all of these actions when you scope to a Container Cluster and view the Pending Actions chart.



For actions on nodes:

- For actions to suspend or provision nodes in the public cloud, Workload Optimization Manager includes cost information (investments or savings) attached to those actions. Note that Workload Optimization Manager generates these actions *not* to optimize costs, but to assure performance and efficiency for your container infrastructure. Workload Optimization Manager reports costs to help you track your cloud spend.

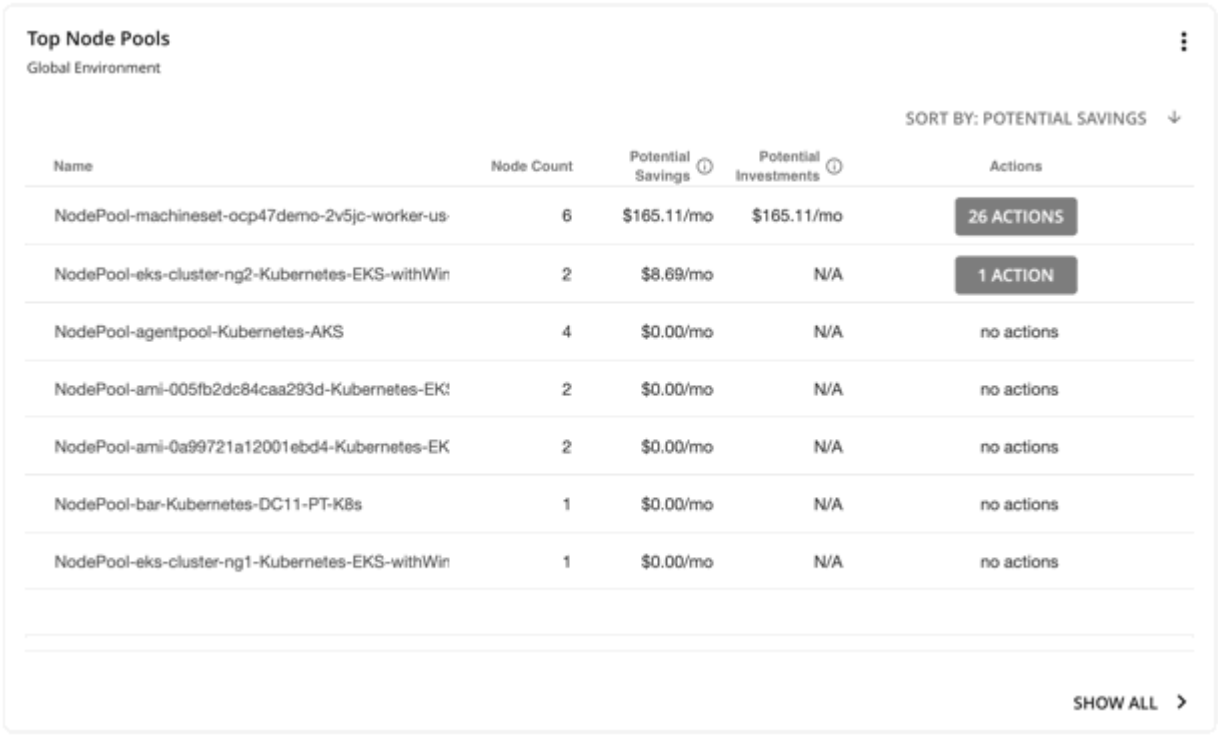
  To view cost information, set the scope to a cluster in the public cloud and view the Necessary Investments or Potential Savings charts. You can also set the scope to the global cloud environment to see total costs, or to individual container clusters or nodes.
- For VMs/nodes that make up an Azure Kubernetes Service (AKS) cluster, you can manually execute recommended VM Provision and VM Suspend actions. This adjusts the count of nodes in a given node pool, where Provision raises the node count, and Suspend lowers it. You can execute these actions if the cluster is also discovered through an Azure target (along with the KubeTurbo target).
- Node pools and machine sets are ways to deploy and scale compute resources for Kubernetes services hosted in the public cloud and the Red Hat OpenShift 4.x container platform on any infrastructure.

  For Kubernetes services in the public cloud, Workload Optimization Manager uses default labels with the following patterns to discover the node pool types within each cluster:

  - Azure Kubernetes Service (AKS): `agentpool`
  - Amazon Elastic Kubernetes Service (EKS):
    `//alpha.eksctl.io/nodegroup-name`
    `eks.amazonaws.com/nodegroup`
  - Google Kubernetes Engine (GKE): `cloud.google.com/gke-nodepool`

  For Red Hat OpenShift 4.x, Workload Optimization Manager creates node pools based on machine sets.

For both discovered and auto-created node pools, Workload Optimization Manager 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 Cluster entity in the supply chain.



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 Workload Optimization Manager with Red Hat OpenShift 4.x Machine Operator, or via an Action Script. You can also manually execute node actions for AKS, EKS, or GKE via the cloud provider.

NOTE:
The following capabilities will be introduced in a future release:

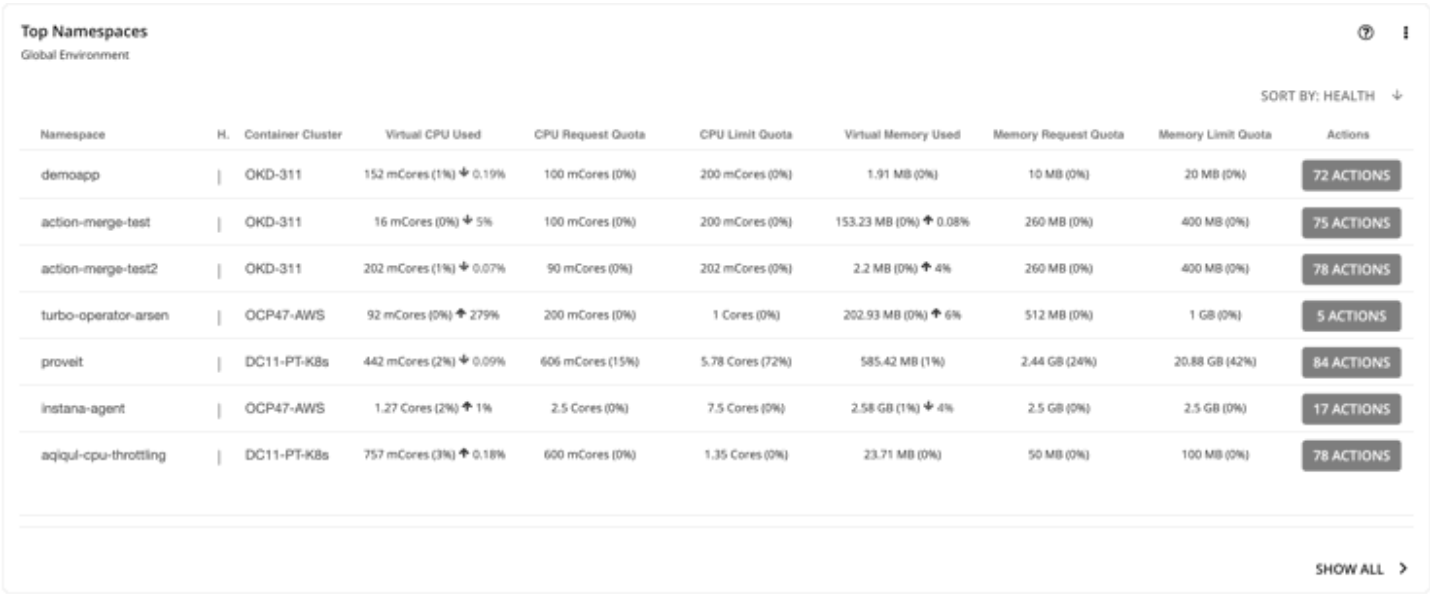
- Actions to provision or suspend nodes via a plan simulation
- Policies for node pools
- Execution of node actions for AKS, EKS and GKE through Workload Optimization Manager

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.

The **Top Namespaces** chart shows the namespaces that use the most cluster resources. You can use the data in the chart for showback analysis.

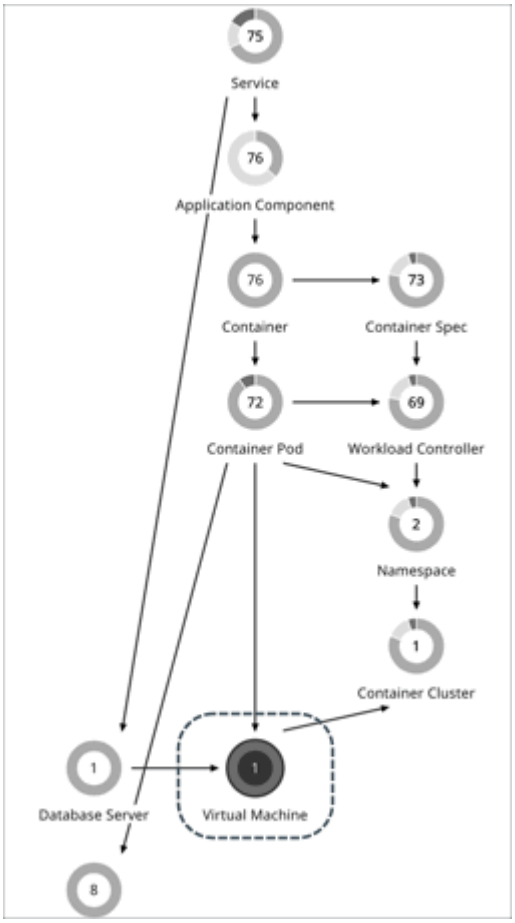


Virtual Machine (Kubernetes Node)

In Kubernetes environments, a node is a virtual or physical machine that contains the services necessary to run pods. Workload Optimization Manager represents nodes as Virtual Machine entities in the supply chain.

Workload Optimization Manager 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. Workload Optimization Manager also discovers and displays Node Pools, and Red Hat OpenShift Machine Sets.

Synopsis



Synopsis	
Provides:	Resources to pods
Consumes:	Resources from container clusters
Discovered through:	Kubeturbo Mediation Pod

Monitored Resources

Workload Optimization Manager monitors the following resources for nodes that host Kubernetes pods. These resources are monitored along with the resources from the infrastructure probes, such as vCenter or a public cloud mediation probe.

- Virtual Memory

  The memory currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- Virtual CPU

  The CPU currently used by all containers on the node. The capacity for this resource is the Node Physical capacity.
- Memory Request Allocation

  The memory available to the node to support the ResourceQuota request parameter for a given namespace (Kubernetes namespace or Red Hat OpenShift project).
- CPU Request Allocation

The CPU available to the node to support the ResourceQuota request parameter for a given namespace (Kubernetes namespace or Red Hat OpenShift project).

- Virtual Memory Request

  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

  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.

- MemAllocation

  The memory ResourceQuota limit parameter for a given namespace (Kubernetes namespace or Red Hat OpenShift project).

- CPUAllocation

  The CPU ResourceQuota limit parameter for a given namespace (Kubernetes namespace or Red Hat OpenShift project).

Actions

Workload Optimization Manager can recommend 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.

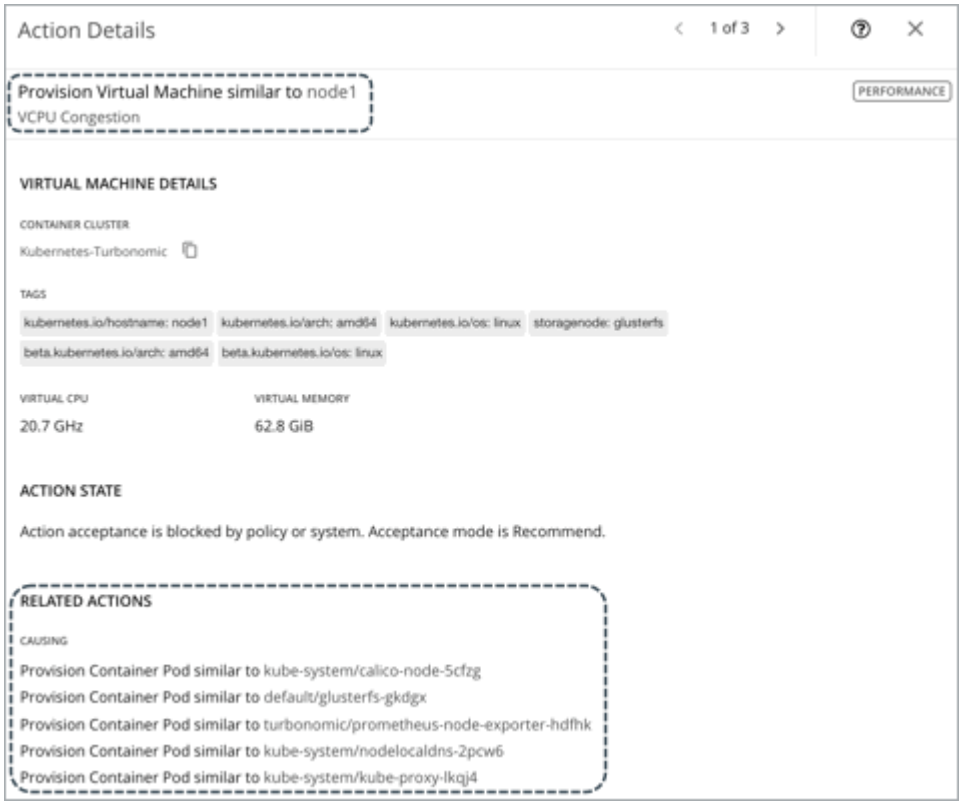
NOTE:
For nodes in the public cloud, Workload Optimization Manager reports the cost savings or investments attached to node and provision actions. For example, you can see the additional costs you would incur if you provision nodes and then scale their volumes, or the savings you would realize if you 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 cluster *(on page 176)* or the global cloud environment to view aggregated cost information.

Node Provision Actions

When recommending node provision actions, Workload Optimization Manager 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.

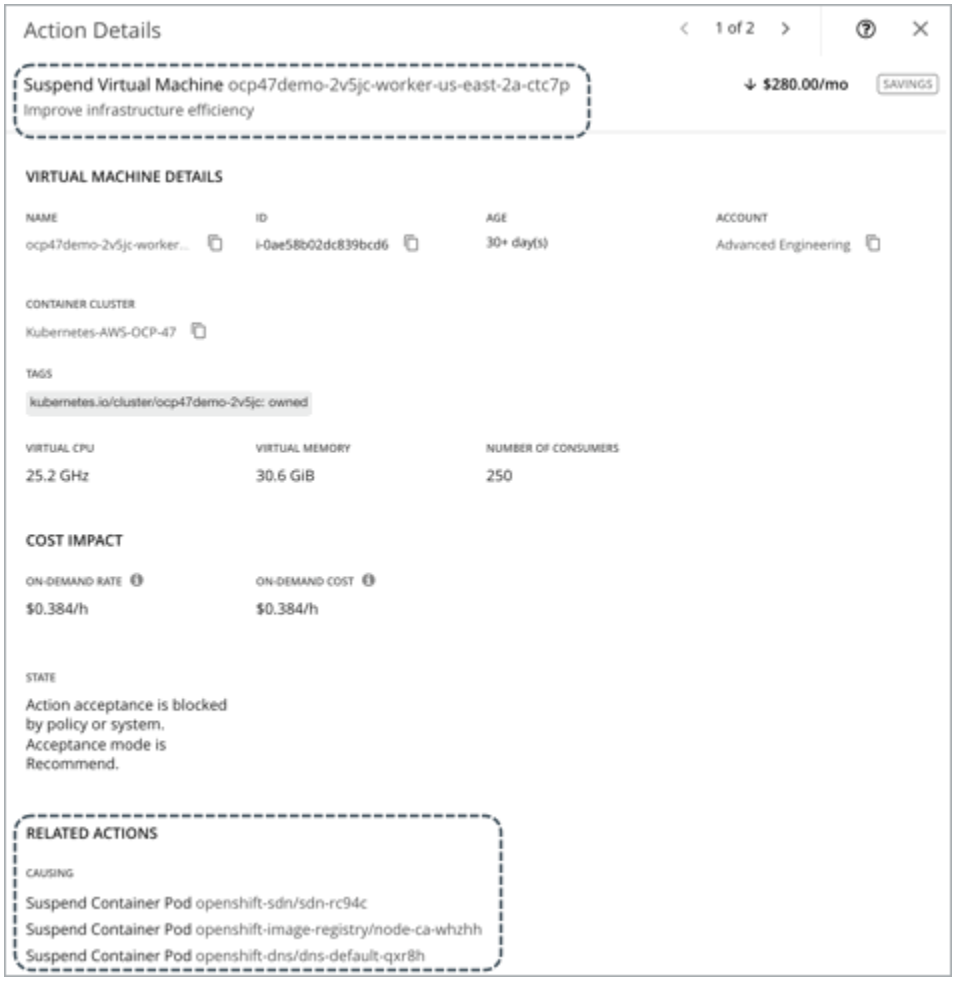


Workload Optimization Manager treats static pods as DaemonSets for the purpose of provisioning nodes. Because a static pod provides a node with a specific capability, it is controlled by the node and is not accessible through the API server. If a node to be provisioned requires a static pod, Workload Optimization Manager generates actions to provision the node and the corresponding static pod.

Node Suspension Actions

When recommending node suspension actions, Workload Optimization Manager 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.



Workload Optimization Manager 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, Workload Optimization Manager generates actions to suspend the node and the corresponding static pod.

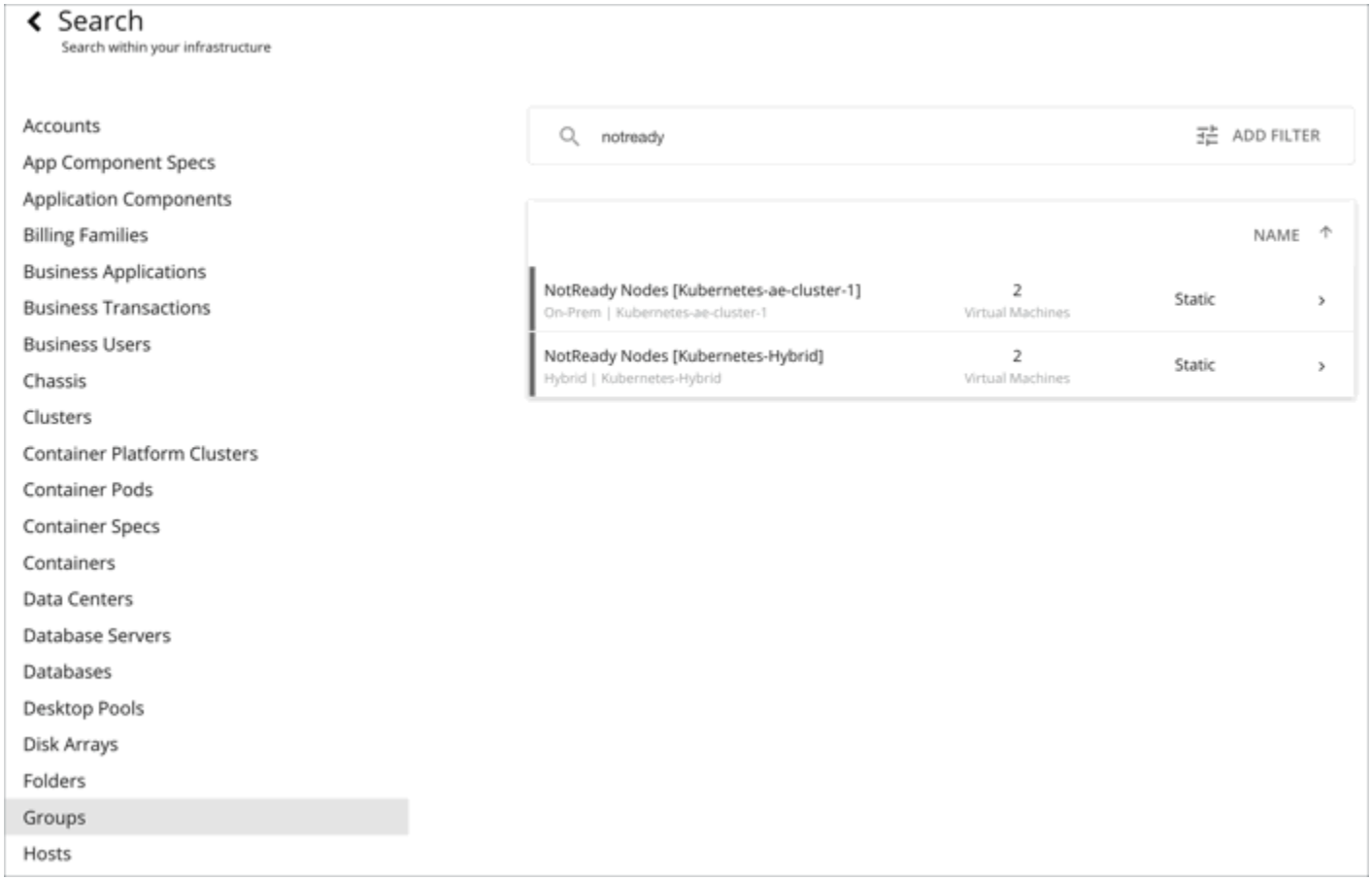
Node Reconfigure Actions

Workload Optimization Manager generates node reconfigure actions to notify you of nodes that are currently in the `NotReady` state.

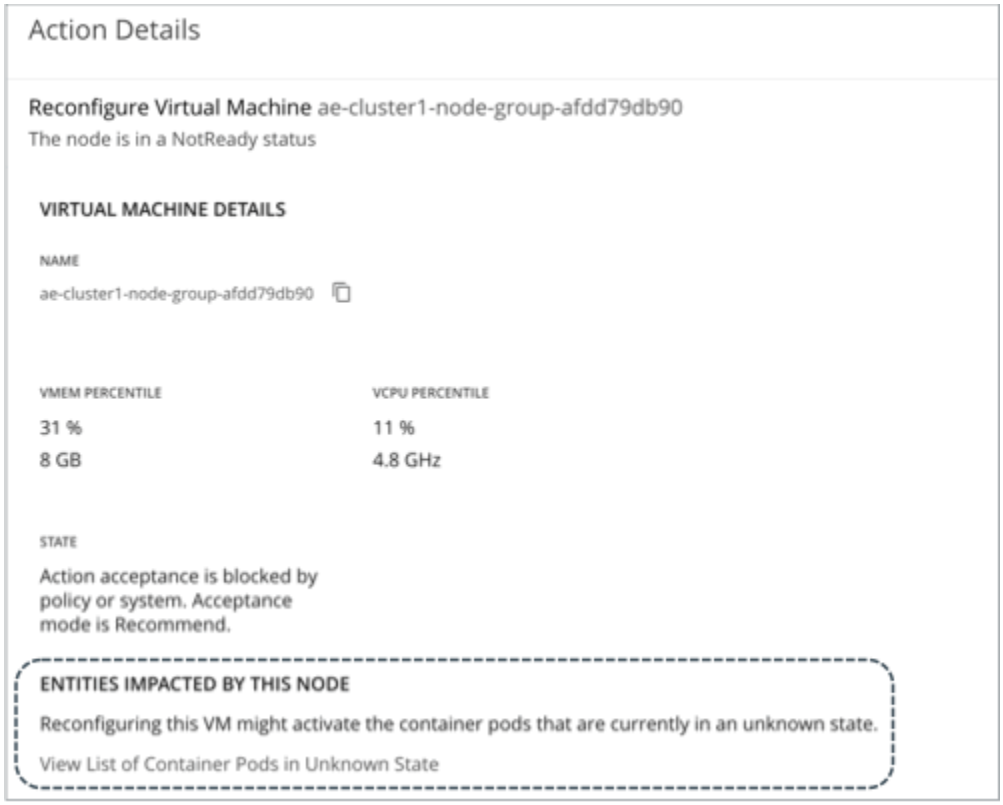
A reconfigure action should change a node's state to `Ready` so that Workload Optimization Manager can begin to monitor the health of the node and the associated container pods. This action is read-only and must be executed outside Workload Optimization Manager. As part of action execution, you might need to restart the node or the kubelet agent on the node.

> **NOTE:**
> Workload Optimization Manager 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 Kubernetes cluster, Workload Optimization Manager 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.



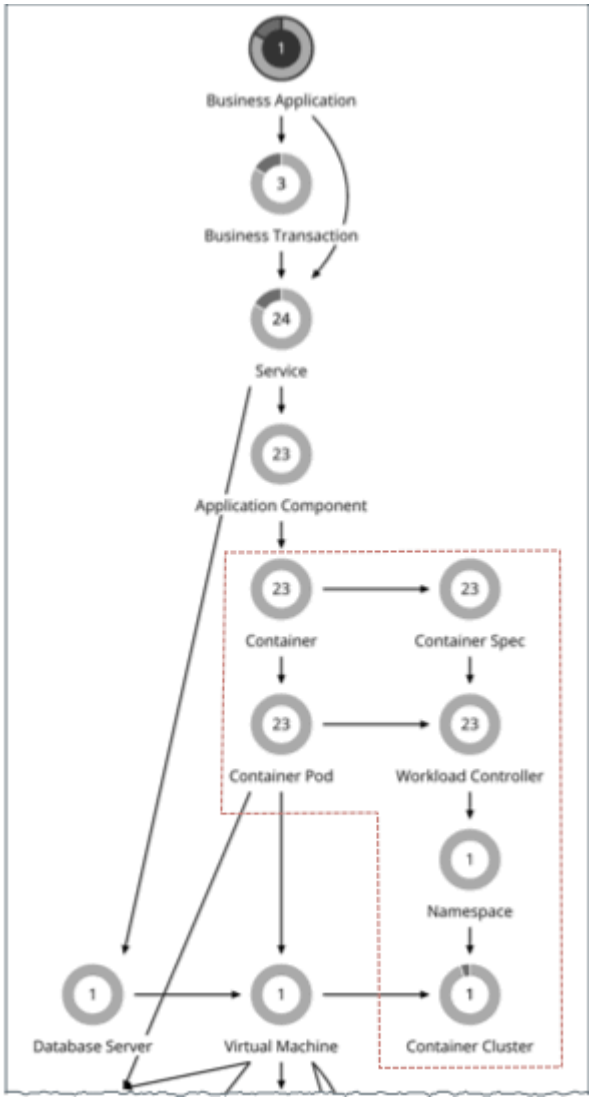
When you examine a pending reconfigure action, you can click the link in the 'Entities Impacted by this Node' section to view a list of impacted pods.



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.

Kubernetes CPU Metrics

To meet user requirements and align with Kubernetes specifications, Workload Optimization Manager uses millicore (mCore) as the base unit for CPU metrics for your Kubernetes platform.



These include metrics for the following CPU-related commodities:

- vCPU
- vCPU Request
- vCPU Limit Quota
- vCPU Request Quota

Workload Optimization Manager 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 153)* for a container, you will see utilization and resize values in mCores.
- When you create Container Spec policies *(on page 159)*, resize thresholds and increment constants for CPU-related commodities are set in mCores.
- For an Optimize Container Cluster plan, the plan results *(on page 358)* for CPU-related commodities are shown in mCores.

For nodes (VMs) and Application Components:

- For nodes stitched to your Kubernetes platform, the base unit for *vCPU Request* is also mCore, since this commodity is provided only to Kubernetes.
- For both nodes and Application Components (standalone or stitched to your Kubernetes 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 Workload Optimization Manager uses.

Entity	CPU Commodity			
	vCPU	vCPU Request	vCPU Limit Quota	vCPU Request Quota
Container	mCore	mCore	mCore	mCore
Container Spec	mCore	mCore	N/A	N/A
Workload Controller	N/A	N/A	mCore	mCore
Container Pod	mCore	mCore	mCore	mCore
Namespace	mCore	mCore	mCore	mCore
Container Cluster	mCore	mCore	N/A	N/A
Node (VM)	MHz	mCore	N/A	N/A
Application Component	MHz	N/A	N/A	N/A

This feature is available starting in version 3.0.5. For customers updating to version 3.0.5 or later:

- This feature does *not* require you to update your Kubeturbo image after the update.
- For time series charts, metrics generated *after* the update are actual mCore values, but pre-update metrics are the same (unconverted) values in MHz displayed in mCore units. This results in unexpected data in charts immediately after the update.

  For example:

  If vCPU Limit for a Container Spec was resized from 1300 MHz to 1200 MHz *before* you updated Workload Optimization Manager, data points in charts correctly show these values in MHz.

  Immediately after the update:

  - When you view the Virtual CPU chart for the Container Spec, Workload Optimization Manager will show a capacity value of 1200 mCores (which is 1200 MHz in reality) for the last data point *before* the update, and the equivalent value of 500 mCores for the first data point *after* the update. This gives the impression of a resize down action between the data points, even if no such action was executed.

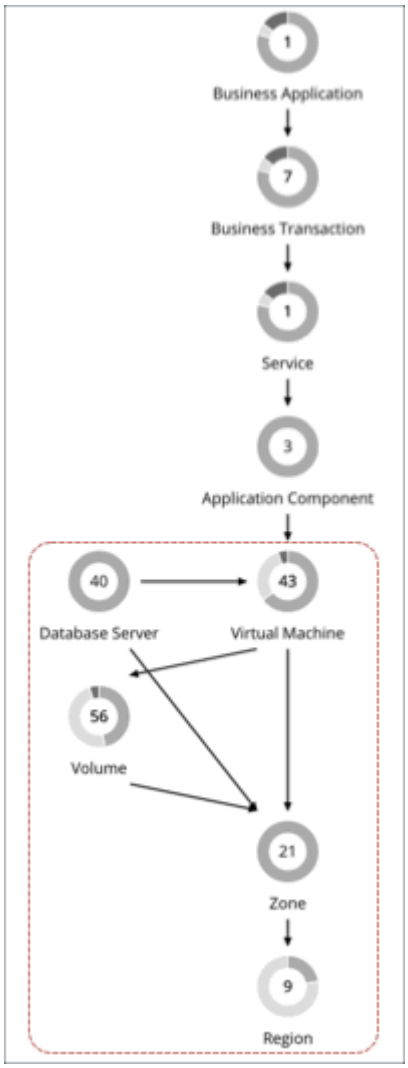
  - Assume Workload Optimization Manager recommends an action to resize VCPU Limit for the Container Spec from 500 to 700 mCores. When you view the details for this action via the associated Workload Controller, the time series chart will show unexpected data.
    - For the actual recommended action, the data point shows current capacity as 1200 mCores, instead of 500 mCores. The new value after executing the action correctly shows as 700 mCores.
    - For the last resize action before the update, the data point shows the same MHz values (1300 and 1200), but in mCore units.
    - One day after the update, a new data point displays in the chart, indicating that capacity was resized from 1200 mCores to 500 mCores, even if no actual resize action was executed.

  Over time, data points with unexpected values will begin to fall out of range and newer data points will reflect actual mCore values.

- For increment constants in Container Spec policies, the default value of 100 remains unchanged, but the unit changes from MHz to mCores. This means that each resize action will now increase or decrease capacity by 100 mCores, instead of 100 MHz.

Entity Types - Cloud Infrastructure

Workload Optimization Manager discovers and monitors the entities that make up your cloud infrastructure, and recommends actions to assure application performance at the lowest possible cost.

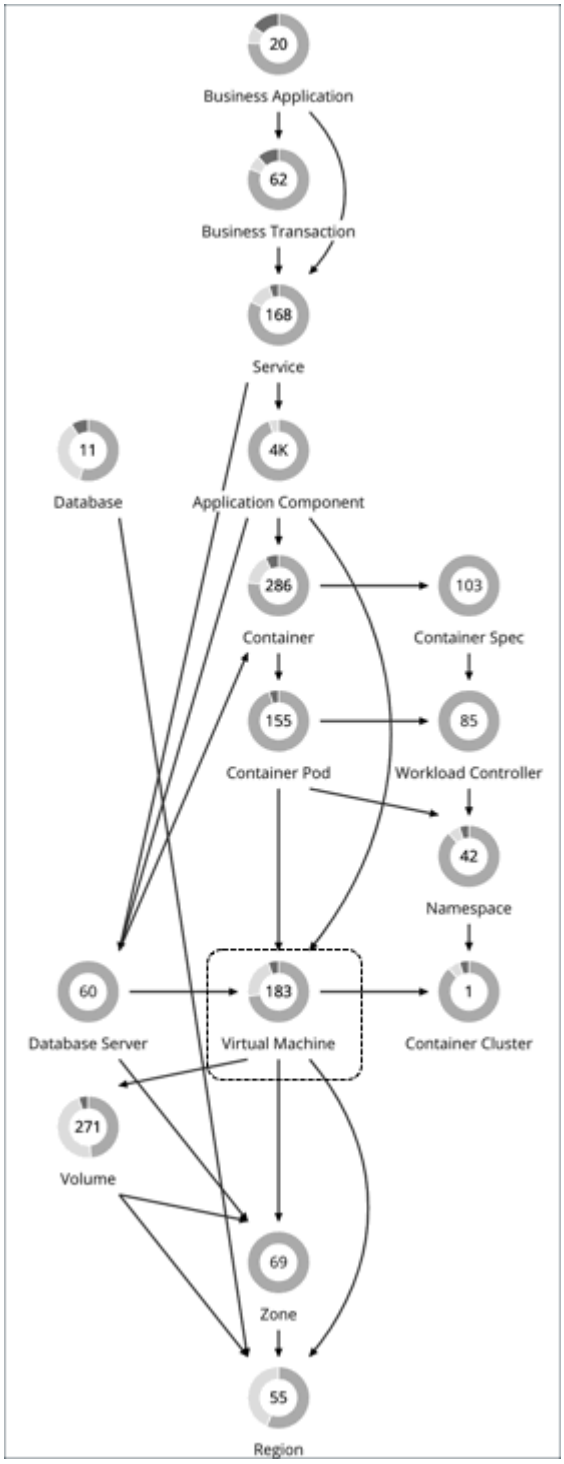


Virtual Machine (Cloud)

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:
Kubernetes nodes are represented as Virtual Machines in the Workload Optimization Manager supply chain. For details about nodes, see Virtual Machine (Kubernetes Node) *(on page 180)*.

Synopsis



Synopsis	
Budget:	A VM gains its budget by selling resources to the applications it hosts.

Synopsis	
Provides:	Resources for hosted applications to use: ▪ VMEM (Kbytes) ▪ VCPU (MHz) ▪ VStorage ▪ IOPS (storage access operations per second) ▪ Latency (capacity for disk latency in ms) ▪ Memory and CPU Requests (for Kubernetes environments)
Consumes:	Resources from cloud zones
Discovered through:	Cloud targets

Monitored Resources

Workload Optimization Manager monitors the following resources for a cloud VM:

- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.
- Virtual CPU

  Virtual CPU is the measurement of CPU utilized by the entity.
- Storage Amount

  The utilization of the datastore's capacity
- Storage Access Operations Per Second (IOPS)

  The utilization of IOPS allocated for the VStorage on the VM
- Net Throughput

  Rate of message delivery over a port
- Net Throughput Inbound

  Rate of message received over a port
- Net Throughput Outbound

  Rate of message sent over a port
- I/O Throughput

  The throughput to the underlying storage for the entity
- Latency

  The utilization of latency allocated for the VStorage on the VM

Cloud VM Actions

- **Scale**

  Change the VM instance to use a different instance type or tier to optimize performance and costs.
- **Stop** and **Start** (also known as 'parking' actions)

  Stop a VM for a given period of time to reduce your cloud expenses, and then start it at a later time.
- **Discount-related actions**

  If you have a high percentage of on-demand VMs, you can reduce your monthly costs by increasing discount coverage. To increase coverage, you scale VMs to instance types that have existing capacity. If you need more capacity, then Workload Optimization Manager will recommend actions to purchase additional discounts.

Additional information:

- For scale actions, you can choose **Cloud Scale All**, **Cloud Scale for Performance**, or **Cloud Scale for Savings**.

  You can direct Workload Optimization Manager to only execute cloud VM scaling actions that improve performance (*Cloud Scale for Performance*) or reduce costs (*Cloud Scale for Savings*). The default action 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.

  Cloud Scale All enables all scaling actions, including those that result in efficiency improvements and increased costs.

  When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see Relationship Between Scoped and Default Policies *(on page 93)*.
- Stop and start actions are also referred to as 'parking' actions. You can enforce these actions on demand or according to a schedule. Parking schedules are separate from the automation schedules that you can set for scale actions. For details, see Parking: Stop or Start Cloud Resources *(on page 428)*.
- 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 403)*.

  Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.

Cloud VMs with Failed Sizing

For workload on the public cloud, if Workload Optimization Manager tries to execute a scale action but the action fails, then Workload Optimization Manager places the affected VM in a special group named *Cloud VMs with Failed Sizing*. Under normal circumstances this group will be empty. But in case some actions have failed, you can review the contents of this group to inspect the individual VMs. As soon as Workload Optimization Manager successfully executes a scale action on a VM in this group, it then removes the VM from the group.

> **NOTE:**
> When Workload Optimization Manager places a VM in this group, it restarts the VM to ensure that it is running correctly with its original configuration.

By default Workload Optimization Manager does not include any action policies for this group. Whatever action mode is set to the given VMs remains in effect while the VMs are in this group. You can create a policy and scope the policy to this group. For example, assume you see typical failures for actions that Workload Optimization Manager tries to execute during working hours. In that case, you can create a scheduling window that enables scale actions during off hours. That can help to automatically execute the actions and remove the VMs from this group.

Note that the VMs in this group could already be in a scope that is affected by another actions policy. Remember that with competing policies, the most conservative policy wins. When working with the Cloud VMs with Failed Sizing group, this can have unintended consequences. Assume you have VMs with automated scale actions, and you create a policy the sets the action mode to Manual for this group. Assume a failed scale action places a VM into this group. In that case the more conservative action mode takes effect, and the VM will use Manual mode. Because of a failed scale action, the VM does not automate subsequent scale actions.

AWS VMs

AWS Instance Requirements

In AWS some instances require workloads to be configured in specific ways before they can move to those instance types. If Workload Optimization Manager recommends moving a workload that is not suitably configured onto one of these instances, then it sets the action to Recommend Only, and describes the reason. Workload Optimization Manager will not automate the move, even if you have set the action mode for that scope to `Automatic`. You can execute the move manually, after you have properly configured the instance.

Note that if you have workloads that you cannot configure to support these requirements, then you can set up a policy to keep Workload Optimization Manager from making these recommendations. Create a group that contains these workloads, and

then create a placement policy for that scope. In the policy, **Excluded Templates** to exclude the instance types that do require ENA support. For information about placement policies, see Automation Policies *(on page 92)*. For information about excluding instance types, see Cloud Instance Types *(on page 213)*.

The instance requirements that Workload Optimization Manager recognizes are:

- Enhanced Network Adapters

  Some workloads 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. Workload Optimization Manager can recommend moving a workload that does not support ENA onto an instance that does. To make that move, you must perform the required configuration of the workload before you can execute the move. If you move a non-ENA VM to an instance that requires ENA, then AWS cannot start up the VM after the move. Before executing the move, you must enable ENA on the VM.

  For information about ENA configuration, see "Enabling Enhanced Networking with the Elastic Network Adapter (ENA) on Windows Instances" in the AWS documentation.

- Linux AMI Virtualization Type

  An Amazon Linux AMI can use ParaVirtual (PV) or Hardware Virtual Machine (HVM) virtualization. Workload Optimization Manager can recommend moving a PV workload 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 workload. Workload Optimization Manager can recommend moving a 32-bit workload to an instance that only supports a 64-bit platform.

- NVMe Block

  Some instances expose EBS volumes as NVMe block devices, but not all workloads are configured with NVMe drivers. Workload Optimization Manager can recommend moving such a workload to an instance that supports NVMe. Before executing the move, you must install the NVMe drivers on the workload.

In addition, Workload Optimization Manager recognizes processor types that you currently use for your workloads. For move or resize actions, Workload Optimization Manager keeps your workloads on instance types with compatible processors:

- GPU-based instances:

  Workload Optimization Manager recognizes when your workload is on a GPU-based instance. To ensure the workload always stays on a compatible processor, Workload Optimization Manager does not recommend resize actions.

- ARM-based instances

  If your workload is on an ARM-based instance, then Workload Optimization Manager will only recommend resizes to other compatible ARM-based instance types.

Resizing Storage Capacity in AWS Environments

When a VM needs more storage capacity Workload Optimization Manager recommends actions to move it to an instance that provides more storage. Note that AWS supports both Elastic Block Store (EBS) and Instance storage. Workload Optimization Manager recognizes these storage types as it recommends storage actions.

If the root storage for your workload is Instance Storage, then Workload Optimization Manager will not recommend a storage action. This is because Instance Storage is ephemeral, and such an action would cause the workload to lose all the stored data.

If the root storage is EBS, then Workload Optimization Manager recommends storage actions. EBS is persistent, and the data will remain after the action. However, if the workload uses Instance Storage for extra storage, then Workload Optimization Manager does not include that storage in its calculations or actions.

Action Details for AWS Workloads

In AWS environments, Workload Optimization Manager considers a VM's used and reserved memory to calculate virtual memory utilization, and drives actions based on the calculated value. This may not always match the values seen in CloudWatch or at the OS level of the VM.

According to the AWS FAQ, *"In C5, portions of the total memory for an instance are reserved from use by the Operating System including areas used by the virtual BIOS for things like ACPI tables and for devices like the virtual video RAM."*. When Workload Optimization Manager recommends moving to one of these instances, the action details use the capacity that is reported by the instance template. However, subsequent reporting of the Mem capacity for the given instance uses the values that Workload Optimization Manager discovers in the environment.

Nodes in AWS EMR Clusters

Workload Optimization Manager treats nodes in AWS EMR clusters like regular VMs. As such, it could incorrectly generate scaling actions for such nodes. After a node scaling 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, we recommend that you disable scaling actions for nodes in EMR clusters.

AWS automatically assigns system tags to EMR clusters. To disable scaling 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.

Azure VMs

Azure Resource Group Discovery

To discover Azure Resource Groups, you can set up the following targets:

- Microsoft Azure service principle targets
- Microsoft Azure Enterprise Agreement (EA) targets

For Azure environments that include Resource Groups, Workload Optimization Manager discovers the Azure Resource Groups and the tags that are used to identify these groups.

In the Workload Optimization Manager user interface, to search for a specific Azure Resource Group, choose **Resource Groups** in the Search Page.

You can set the scope of your Workload Optimization Manager 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 Workload Optimization Manager 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 460)*.

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 Workload Optimization Manager 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. Workload Optimization Manager 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 Workload Optimization Manager will not recommend an action that would move the workload to a non-AN instance.

- Azure Premium Storage

  Workload Optimization Manager 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, Workload Optimization Manager recognizes processor types that you currently use for your workloads. If your workload is on a GPU-based instance, then Workload Optimization Manager will only recommend moves to other compatible GPU-based instance types. For these workloads, Workload Optimization Manager does not recommend resize actions.

IOPS-aware Scaling for Azure VMs

Workload Optimization Manager considers IOPS utilization when making scaling decisions for Azure VMs. To measure utilization, Workload Optimization Manager 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 211)*.

Analysis impacts VM scaling decisions in different ways. For example:

- If your instance experiences IOPS bottlenecks, Workload Optimization Manager 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, Workload Optimization Manager 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.

Cloud VM Uptime

For cloud VMs, Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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

Workload Optimization Manager 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 201)*.

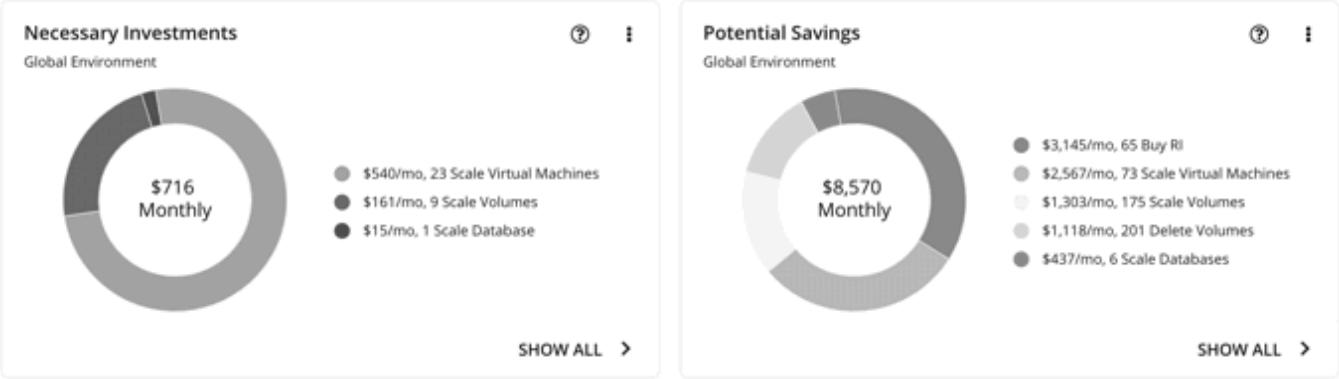
Uptime data impacts cost calculations, but not the actual scaling decisions that Workload Optimization Manager makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Uptime Data in Charts

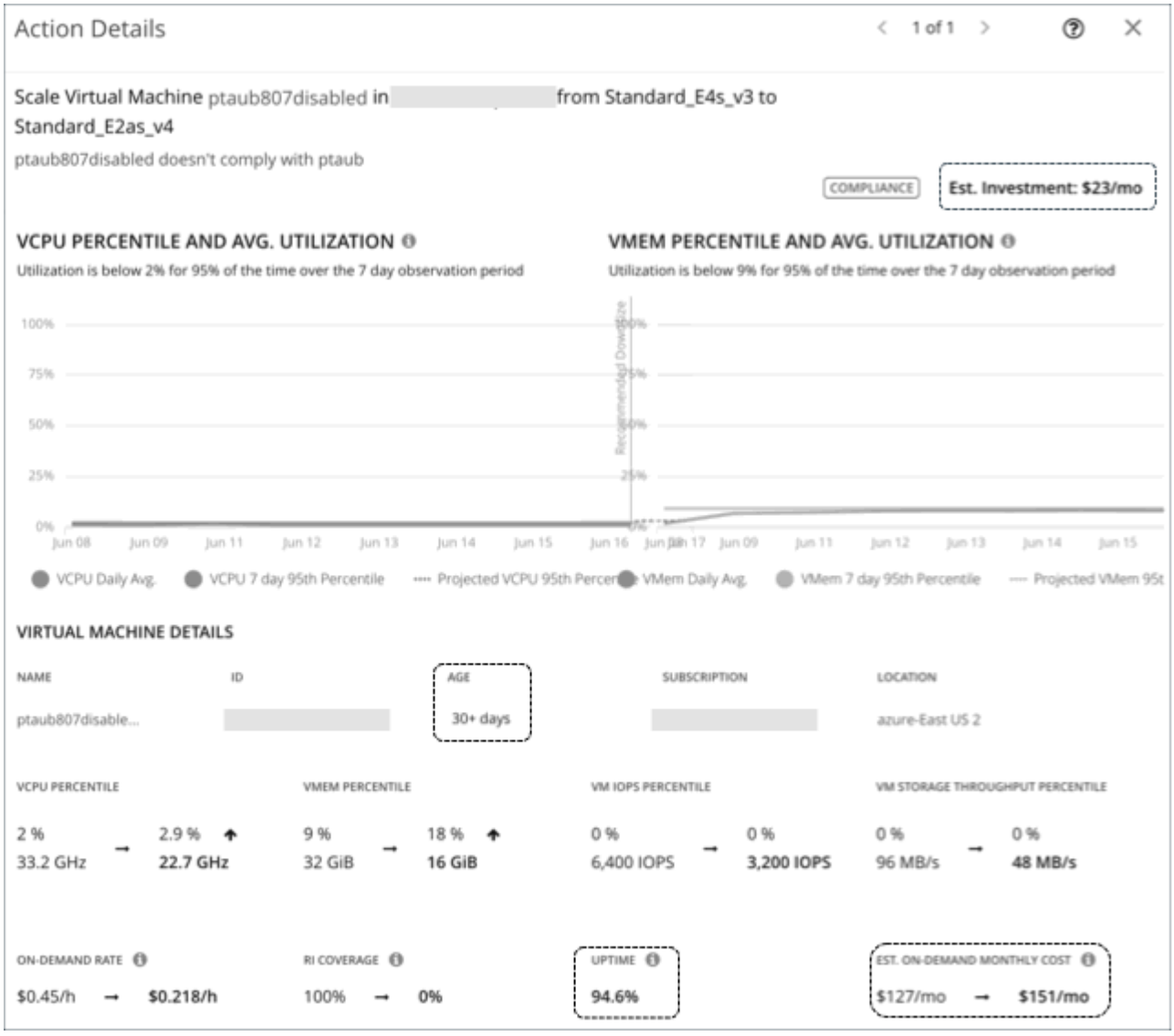
Workload Optimization Manager recalculates uptime data every hour and then updates the values shown in charts. The following charts reflect the cost impact of uptime-based calculations:

- Potential Savings and Necessary Investment charts

  The projected amounts in these charts include on-demand costs for cloud VMs.

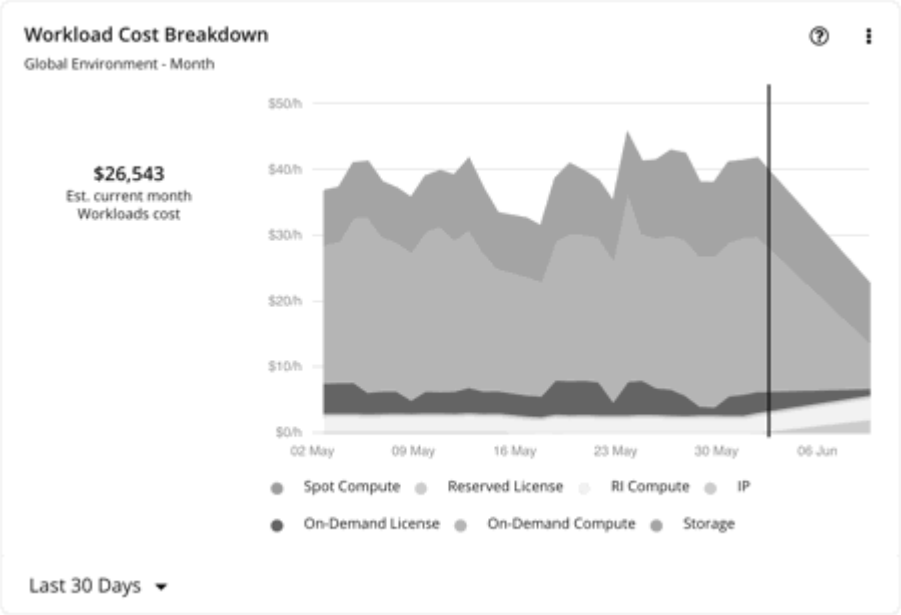


  When you click **Show All** in these charts and view details for a pending VM action, the **Action Details** page shows on-demand costs before and after you execute the action, factoring in the VM's uptime value. The page also shows the VM's age.

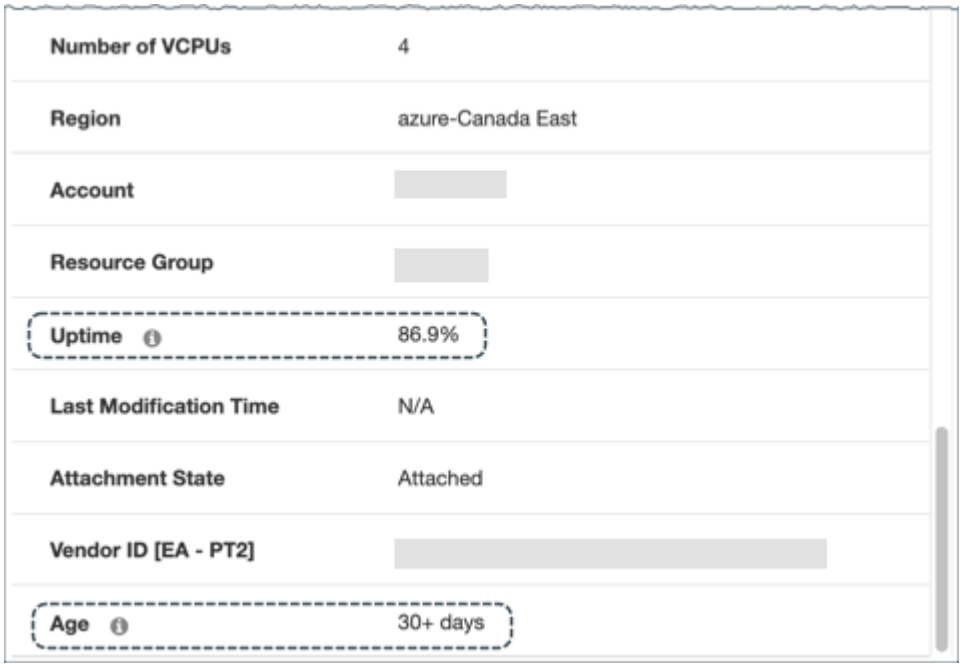


- Workload Cost Breakdown chart

  This chart shows estimated costs over time, including on-demand costs for VMs.

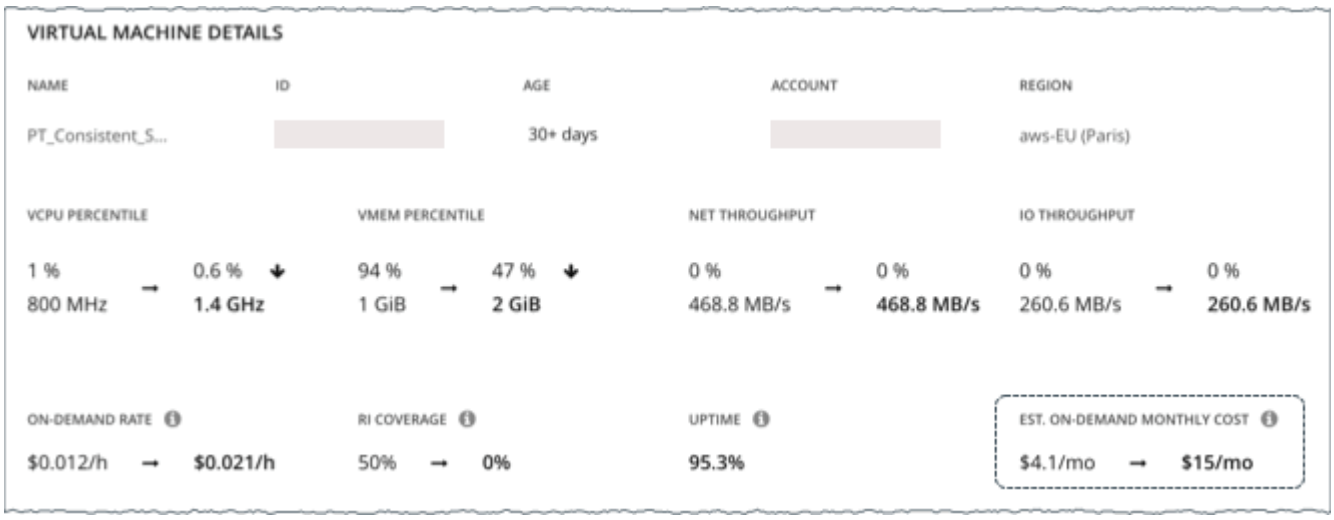


- The **Entity Information** chart shows the latest uptime and age data for a specific cloud VM.



Estimated On-demand Costs for Cloud VMs

Workload Optimization Manager considers a variety of factors when calculating *Estimated On-demand Monthly Cost* for a cloud VM.



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:

```
On-demand Rate * Usage Not Covered by Discounts * Uptime * 730 =
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.

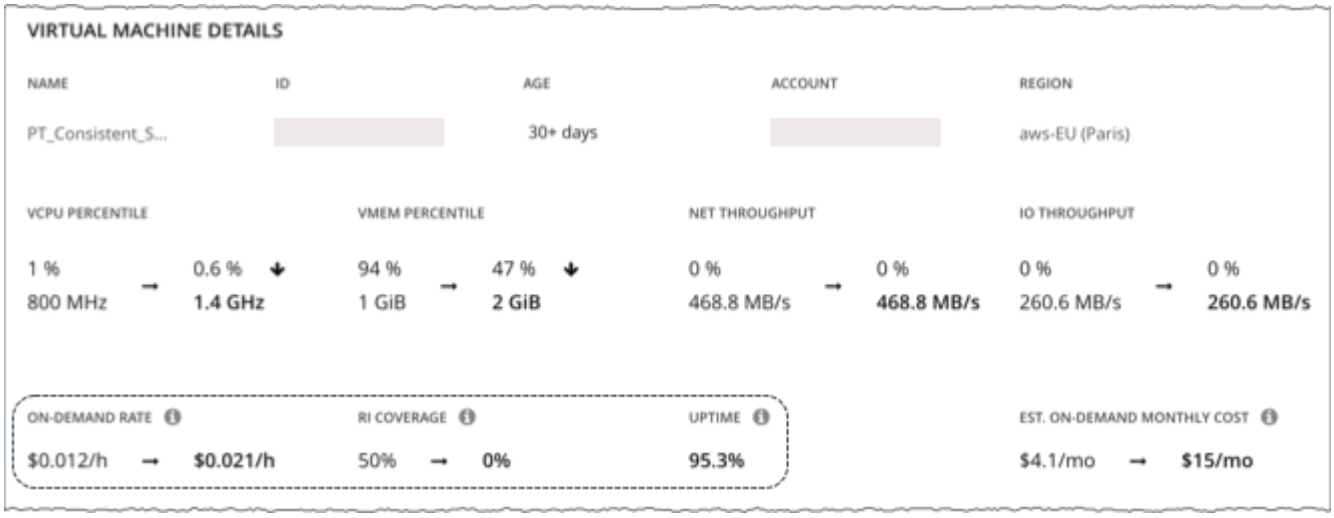
- **Usage Not Covered by Discounts** is the percentage of hourly VM usage not covered by any discount. For example:
  - Discount Coverage = 20% (0.2)
  - Usage Not Covered by Discounts = 80% (0.8)
- **Uptime** is a percentage value that indicates how long a VM has been running over a period of time (age). Age refers to the number of days that a VM has existed since first discovery. For VMs older than 30 days, Workload Optimization Manager only calculates uptime over the last 30 days.

  To estimate monthly on-demand costs, Workload Optimization Manager 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 Workload Optimization Manager uses to estimate monthly costs.

The listed items above impact cost calculations, but not the actual scaling decisions that Workload Optimization Manager makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

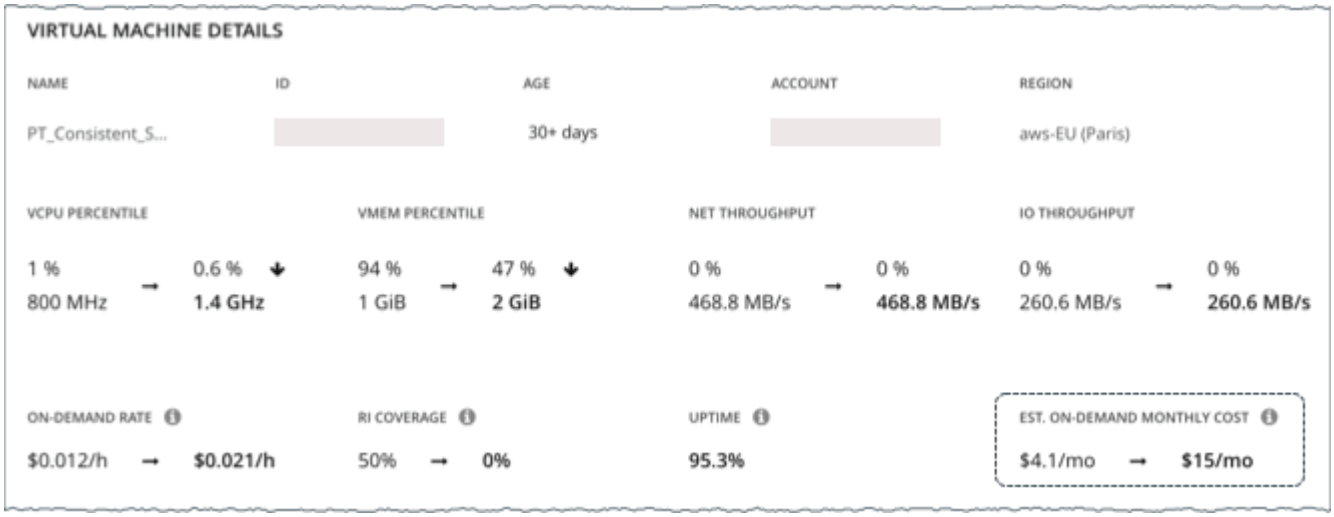
Example

Assume the following data for a pending scale action for an AWS VM:



	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 (*calculated based on discount coverage)*	50% (0.5)	100% (1.0)
Uptime	95.3% (.953)	

Workload Optimization Manager calculates the following:



- *Current* Estimated On-demand Monthly Cost:

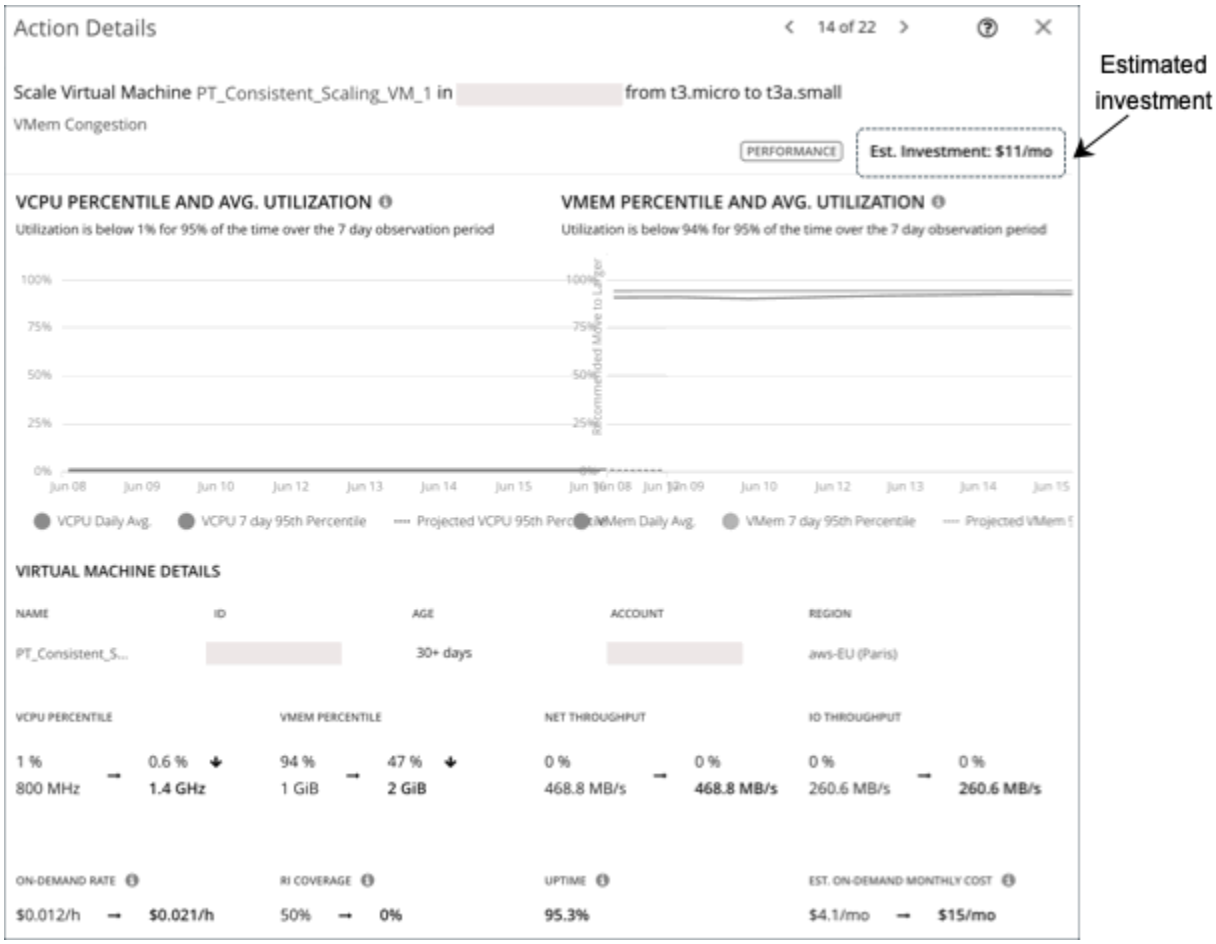
```
0.012 * 0.5 * 0.953 * 730 = 4.1
```

- Estimated On-demand Monthly Cost *after* executing the action:

```
0.021 * 1.0 * 0.953 * 730 = 15
```

NOTE:
Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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:

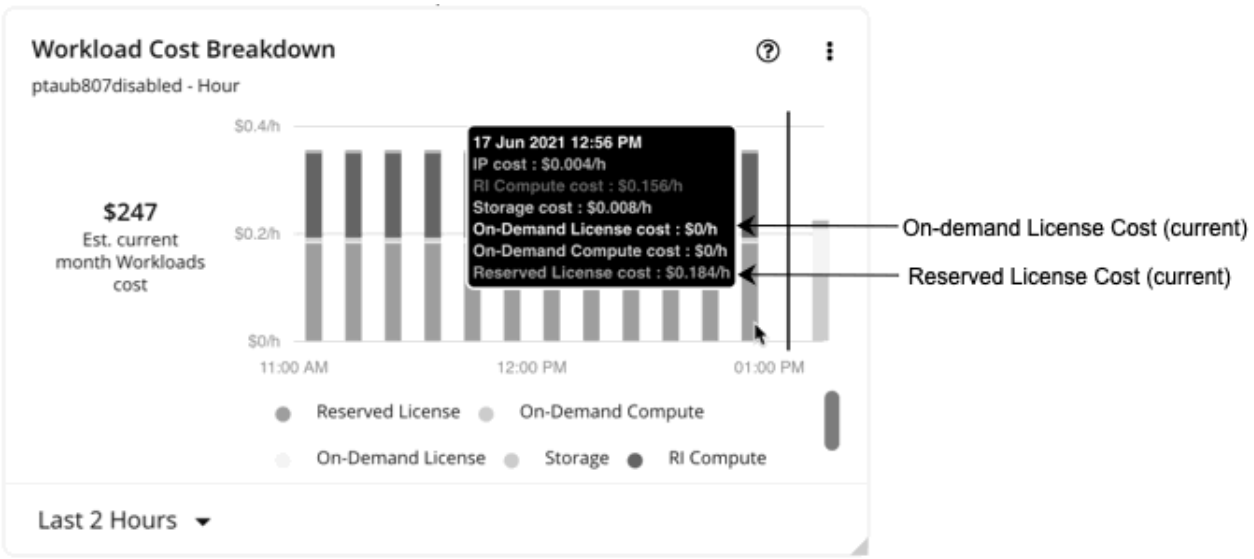
```
On-demand Rate - (Reserved License Cost + On-demand License Cost) =
On-demand Compute Rate
```

Where:

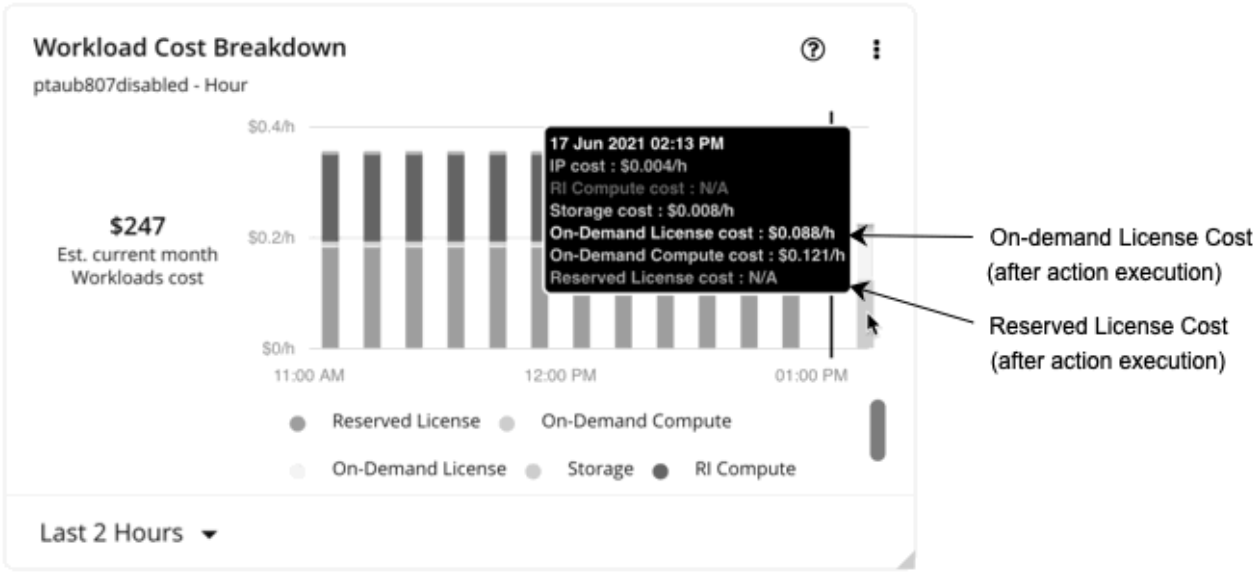
- **On-demand Rate** is the hourly cost for a VM's instance type *without* reservation coverage. This does *not* include license costs, storage, or IP. You can obtain on-demand rates via Azure Pricing Calculator.
- **Reserved License Cost** and **On-demand License Cost** are the hourly costs for the VM's licenses. You can obtain license costs via Azure Pricing Calculator or the Workload Optimization Manager 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:

```
(On-demand Compute Rate * Usage Not Covered by Reservations) + License Cost * Uptime * 730 =
Estimated On-demand Monthly Cost
```

Where:

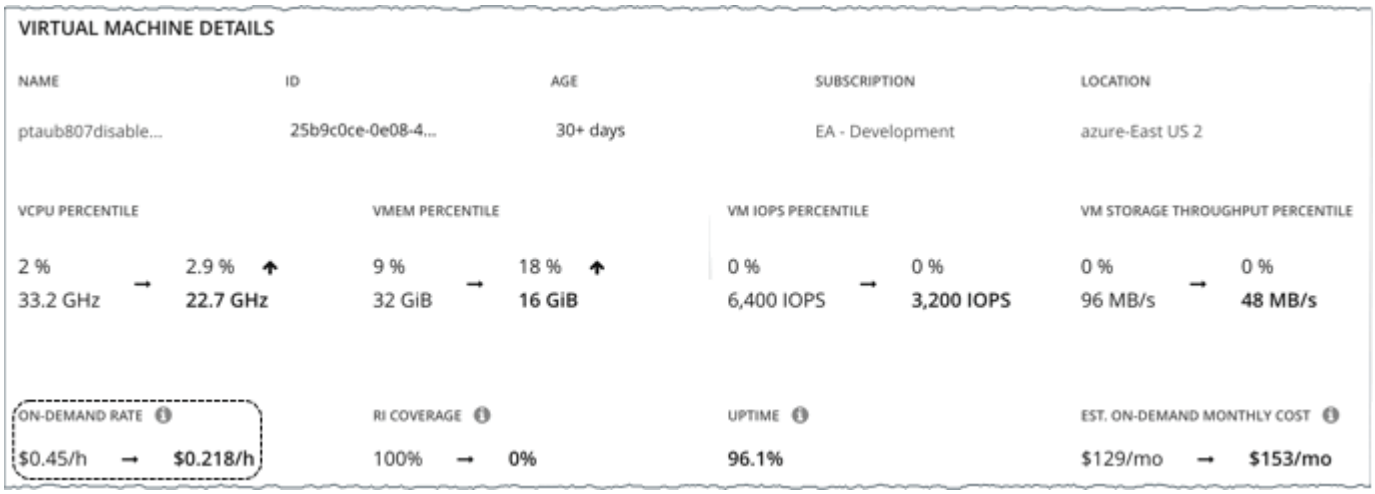
- **Usage Not Covered by Reservations** is the percentage of hourly VM usage not covered by any reservation. For example:
  - Reservation Coverage = 20% (0.2)
  - Usage Not Covered by Reservations = 80% (0.8)
- **License Cost** is the sum of On-demand License Cost and Reserved License Cost.
- **Uptime** is a percentage value that indicates how long a VM has been running over a period of time (age). Age refers to the number of days that a VM has existed since first discovery. For VMs older than 30 days, Workload Optimization Manager only calculates uptime over the last 30 days.

  To estimate monthly on-demand costs, Workload Optimization Manager 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 Workload Optimization Manager uses to estimate monthly costs.

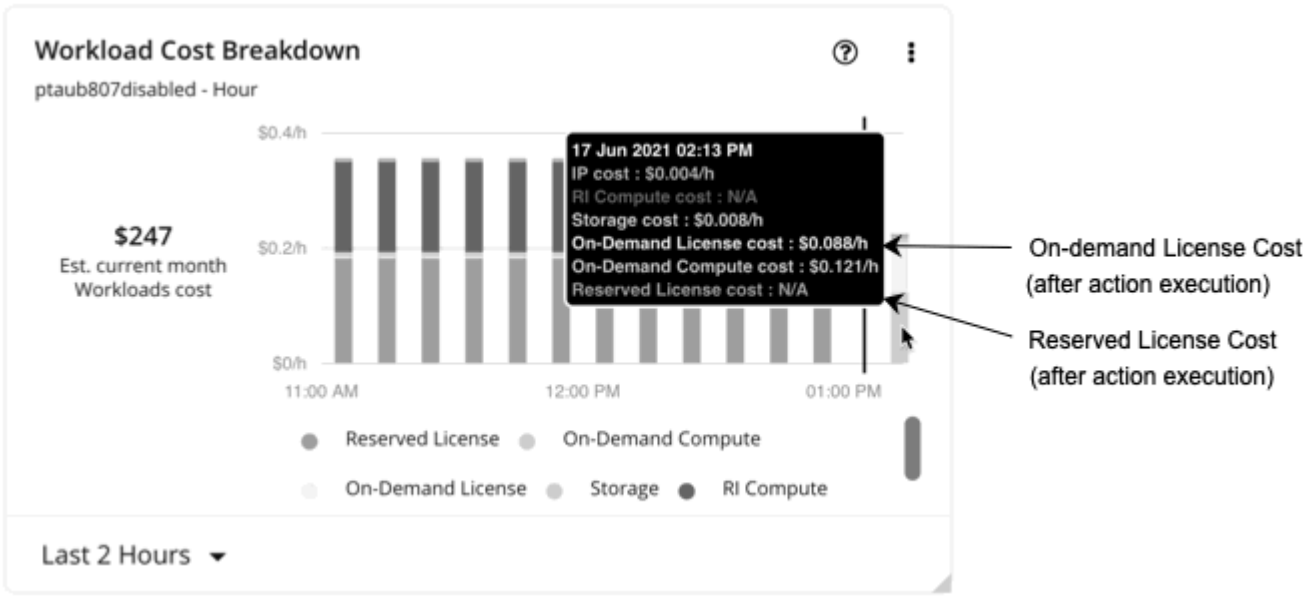
The listed items above impact cost calculations, but not the actual scaling decisions that Workload Optimization Manager makes. These decisions rely on other factors, such as resource utilization percentiles and scaling constraints set in policies.

Example

Assume the following data for a pending scale action for an Azure VM with license costs:



	Current Values	Values After Action Execution
On-demand Rate	$0.45/hr	$0.218/hr



	Current Values	Values After Action Execution
On-demand License Cost	$0/hr	$0.088/hr
Reserved License Cost	$0.184/hr	N/A

1. Workload Optimization Manager first calculates the following:
   - *Current* On-demand Compute Rate:

     ```
     0.45 - (0.184 + 0) = 0.266
     ```

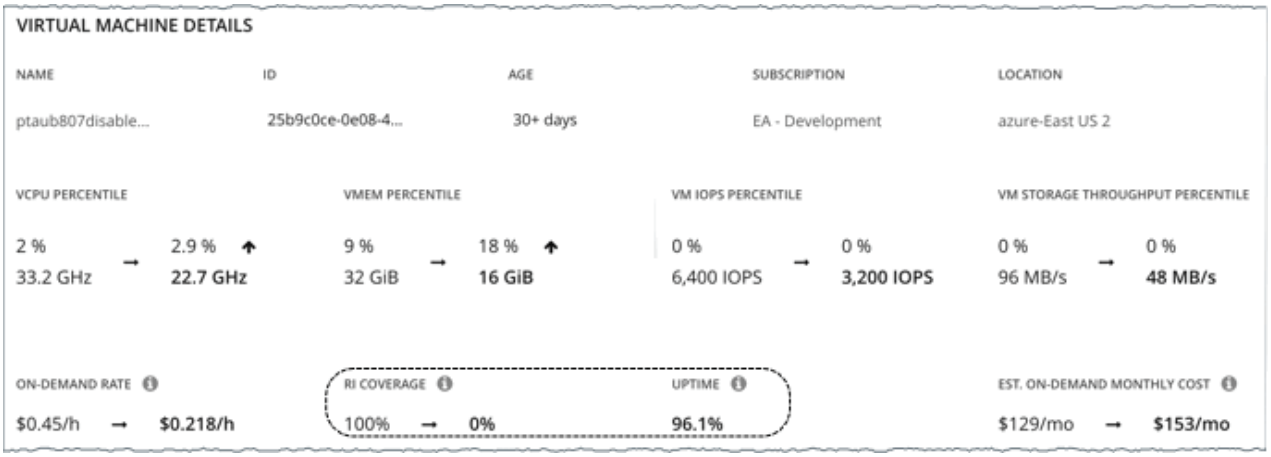
   - On-demand Compute Rate *after* executing the action:

     ```
     0.218 - (0 + 0.088) = 0.13
     ```

2. Workload Optimization Manager can now calculate Estimated On-demand Monthly Cost based on:
   - On-demand Compute Rate

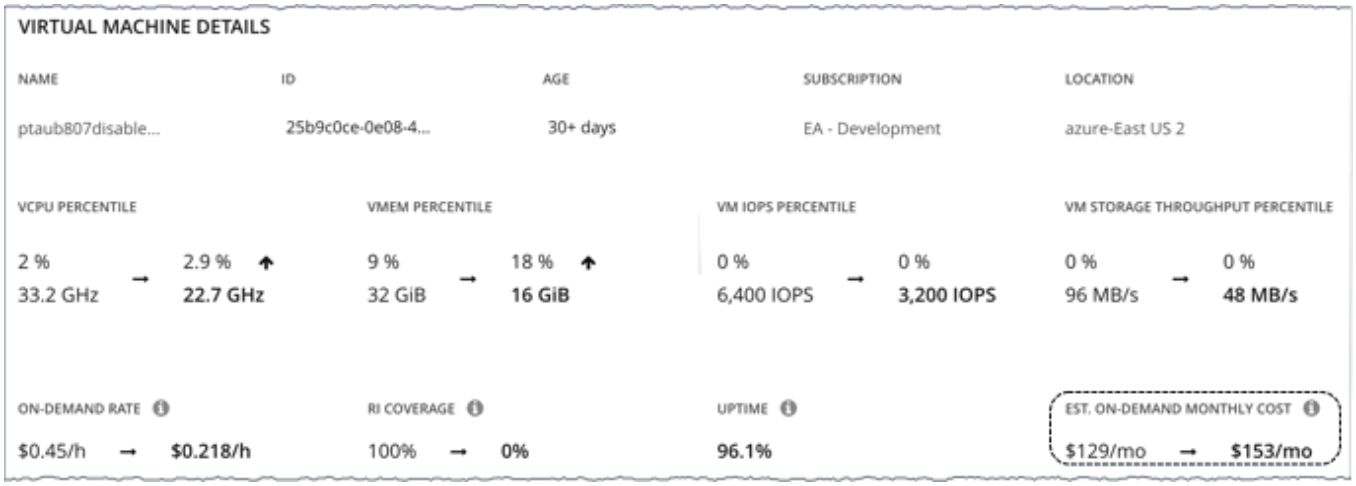
	Current Values	Values After Action Execution
On-demand Compute Rate	$0.266/hr	$0.13/hr

   - Usage Not Covered by Reservations and Uptime



	Current Values	Values After Action Execution
Reservation Coverage	100% (1.0)	0% (0.0)
Usage Not Covered by Reservations (*calculated based on reservation coverage)*	0% (0.0)	100% (1.0)
Uptime	96.1% (.961)	

Workload Optimization Manager calculates the following:



   - *Current* Estimated On-demand Monthly Cost:

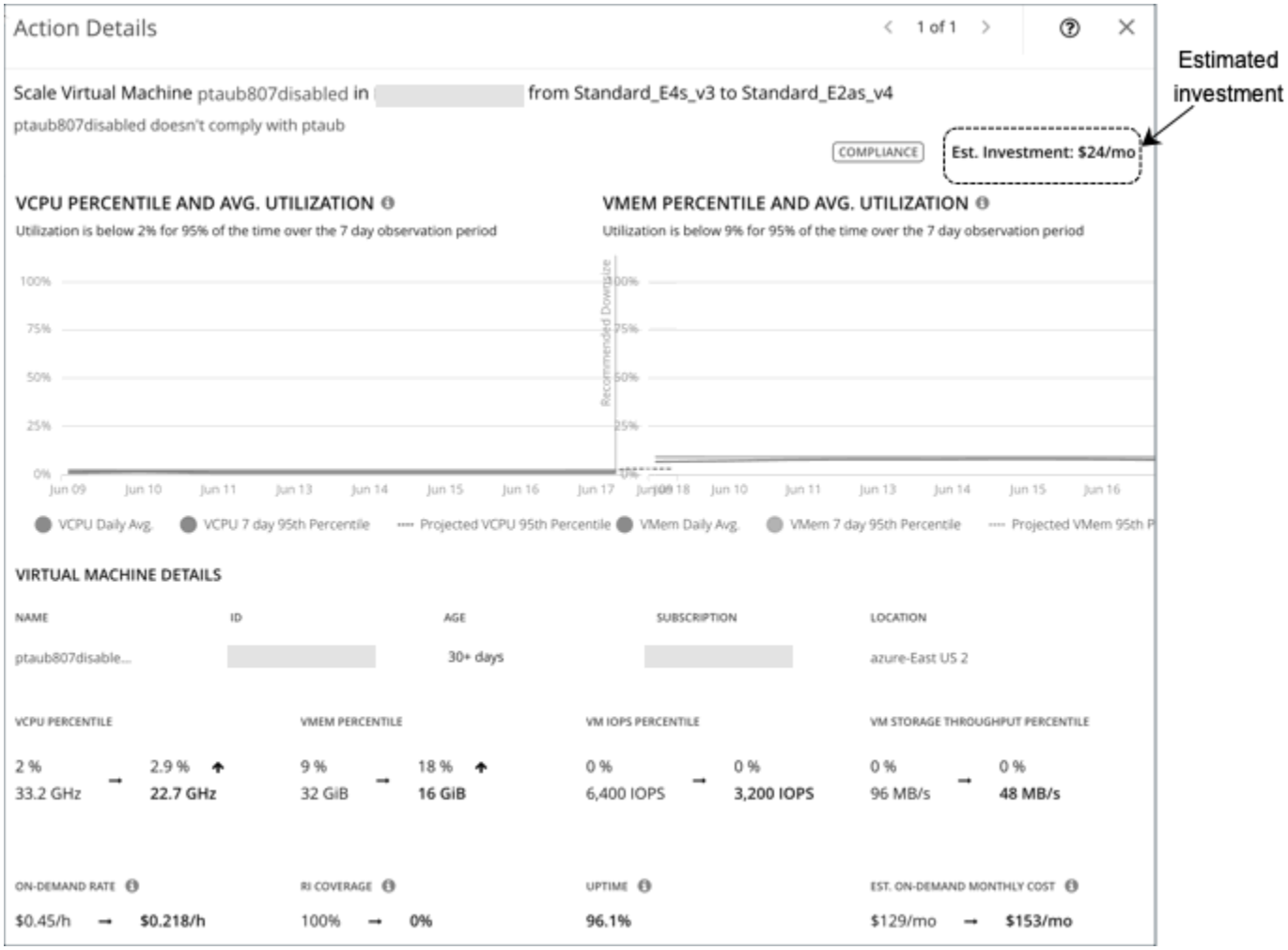
```
(0.266 * 0.0) + 0.184 * 0.961 * 730 = 129
```

- Estimated On-demand Monthly Cost *after* executing the action:

```
(0.13 * 1.0) + 0.088 * 0.961 * 730 = 153
```

NOTE:
Workload Optimization Manager 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, Workload Optimization Manager treats the action as an investment and shows an estimated investment of $24/month.



Cloud VM Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

NOTE:
The policy settings discussed in this topic only apply to cloud scale actions. For details about settings for parking actions (stop and start), see Parking: Stop or Start Cloud Resources *(on page 428)*.

Action Automation and Orchestration

For details about cloud VM actions, see Cloud VM Actions *(on page 193)*.

Cloud Scale

Action	Default Mode	AWS, Azure, and GCP
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	GCP
Buy discounts	Recommend	Rcmd	Not yet supported
Provision Kubernetes node (VM)	Manual	Rcmd	Rcmd
Suspend Kubernetes 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 Workload Optimization Manager 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 Workload Optimization Manager recommends actions, in most cases you should never need to use them. If you want to control how Workload Optimization Manager 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

Attribute	Default Value
	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:

Workload Optimization Manager 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 Workload Optimization Manager will attempt to match.

For details on how IOPS utilization affects scaling decisions, see IOPS-aware Scaling for Azure VMs *(on page 197)*.

Aggressiveness and Observation Periods

Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager will not reduce allocated CPU for that VM.

  With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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

Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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.

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
Enable Consistent Resizing	Off

For groups in scoped policies:

When you create a policy for a group of VMs and turn on Consistent Resizing, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager can enforce a rule to size all the VMs in a group equally.

For auto-discovered groups:

In public cloud environments, Workload Optimization Manager 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

  Workload Optimization Manager discovers Azure availability sets and scale sets.

  - For availability sets, Workload Optimization Manager 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. Workload Optimization Manager 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 mode for scale actions to *Recommend* and stays in effect for 730 hours (one month).

This means that for the duration of the policy, Workload Optimization Manager 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, Workload Optimization Manager automatically enables Consistent Resizing across all the VMs in the group. You can choose to execute all the actions for such a group, either manually or automatically. In that case, Workload Optimization Manager executes the resizes one VM at a time. 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

  Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager will change the Launch Configuration for the Auto Scaling Group but it will not terminate the EC2 instances.

Reasons to employ Consistent Resizing for a group include:

- Load Balancing

  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.

- High Availability (HA)

  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, Workload Optimization Manager 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 Workload Optimization Manager 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.

Instance Store Aware Scaling

Attribute	Default Setting
Instance Store Aware Scaling	Off

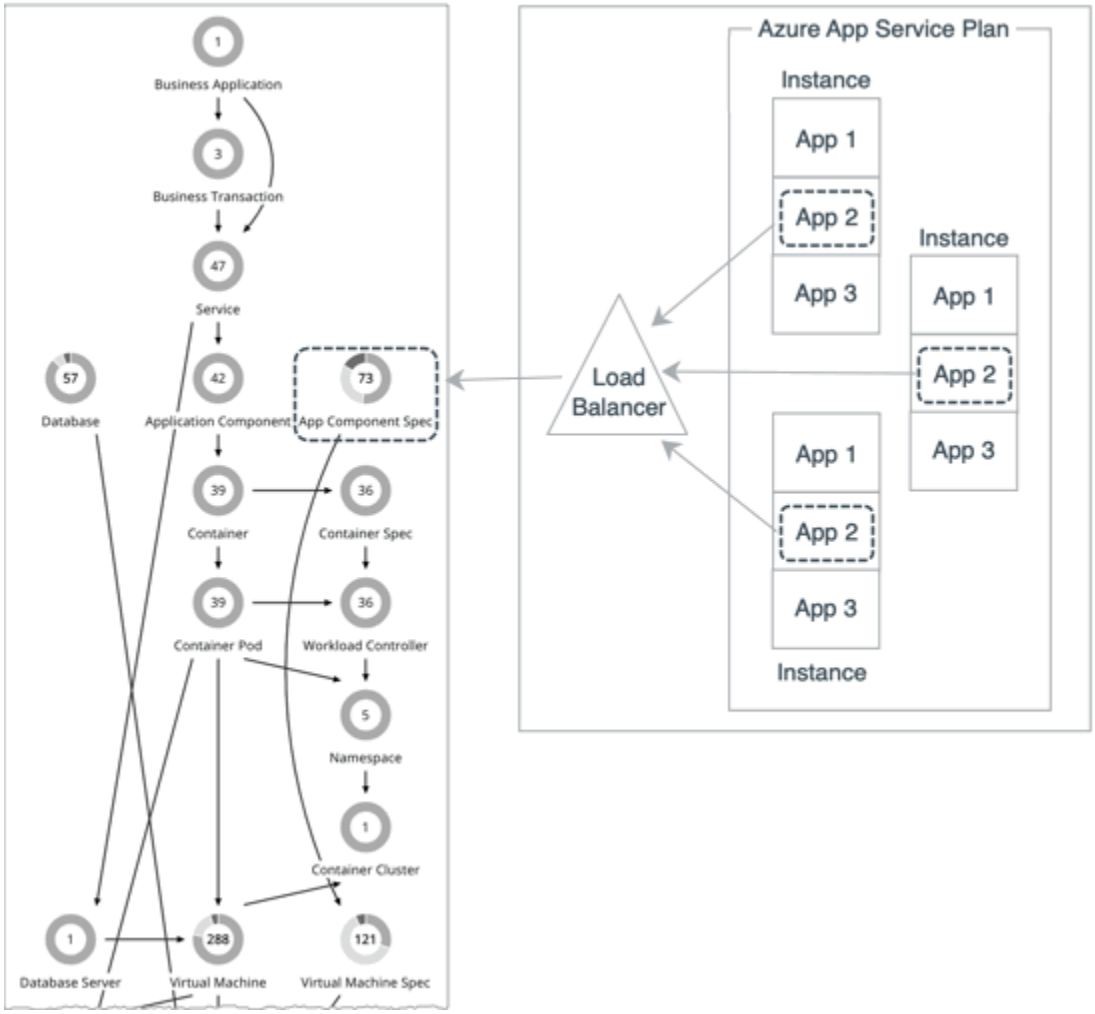
The template for your workload determines whether the workload can use an *instance store*, and it determines the instance store capacity. As Workload Optimization Manager calculates a resize or move action, it can recommend a new template that does not support instance stores, or that does not provide 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, Workload Optimization Manager 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 deployments, an App Component Spec represents a set of app instances comprising a single web application. Workload Optimization Manager discovers App Component Specs when you add an Azure target with the necessary permissions.



NOTE:
For a list of permissions, see "Azure Service Principal and Subscription Permissions" in the *Target Configuration Guide*.

Workload Optimization Manager also discovers the *plans* that provide resources to app instances. The supply chain shows these plans as Virtual Machine Specs *(on page 216)* 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

- 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

  Virtual CPU is the measurement of total CPU time utilized by a given app.

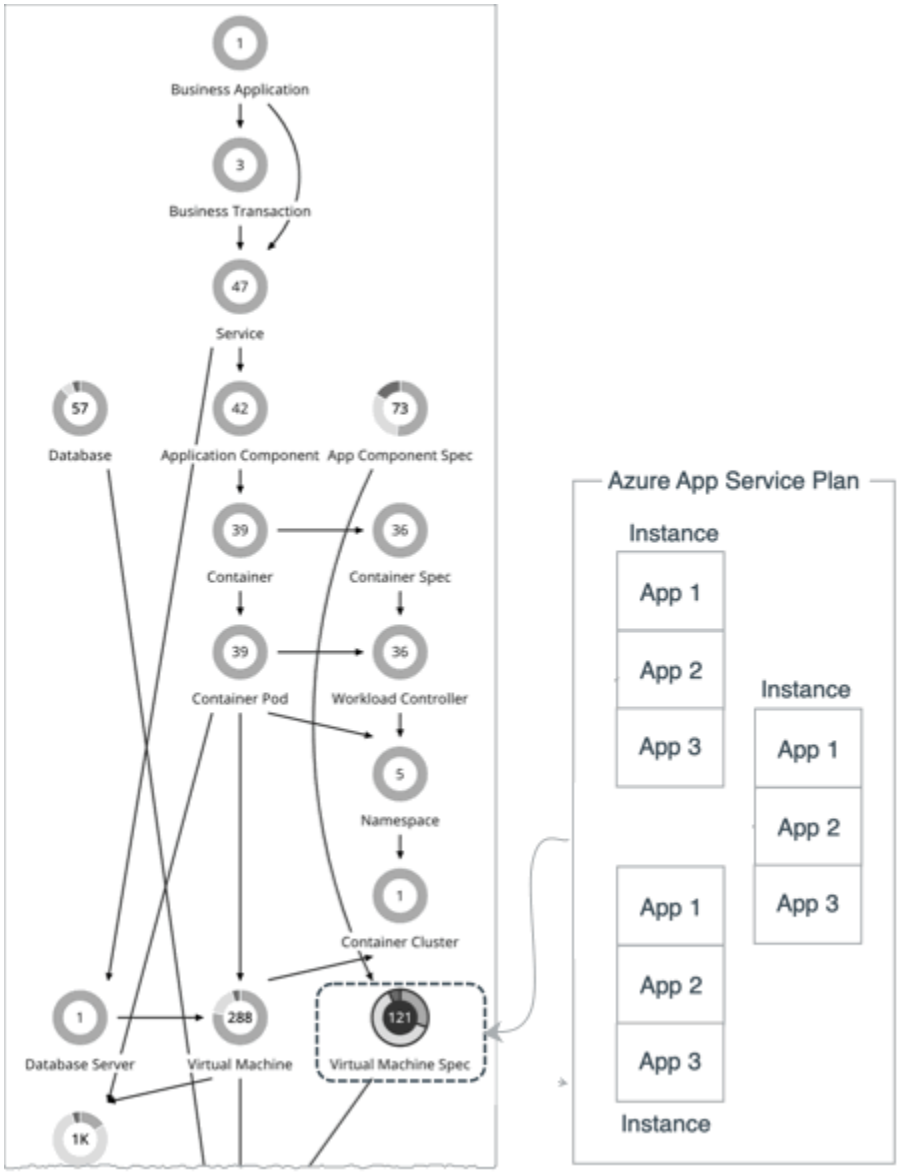
Actions

None

Workload Optimization Manager 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 218)*.

Virtual Machine Spec

In Azure App Service, *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, Workload Optimization Manager discovers the plans associated with apps, and shows them as Virtual Machine Specs in the supply chain. Currently, Workload Optimization Manager discovers all plans, except App Service Environment v3 I4, I5, and I6.



NOTE:
For a list of permissions, see "Azure Service Principal and Subscription Permissions" in the *Target Configuration Guide*.

Points to consider:

- Azure App Service offers several types of apps, including web apps, mobile apps, API apps, and logic apps. Workload Optimization Manager 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, Workload Optimization Manager will recommend that you delete it.
- For web apps, Workload Optimization Manager also discovers the app instances that consume resources from a plan, and shows them as App Component Specs *(on page 215)* 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, Workload Optimization Manager 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

- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.
- Virtual CPU

  Virtual CPU is the measurement of CPU utilized by the entity.
- Storage Amount

  Storage Amount is the measurement of Azure storage utilized by the entity.
- Number of Replicas

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

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

Workload Optimization Manager 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 220)*
- On-demand monthly costs *(on page 222)*
- VM instance count

  Workload Optimization Manager 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 Workload Optimization Manager policies *(on page 225)* for Virtual Machine Specs

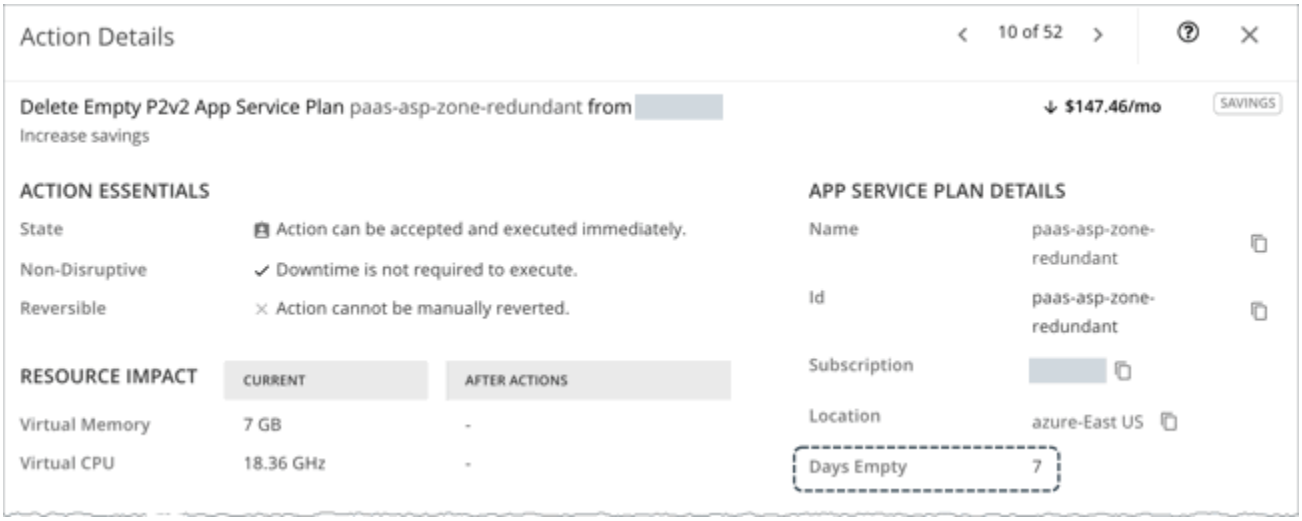
  For example, you can set a constraint if you want App Service plans to *only scale to* or *avoid* certain instance types.

Delete Actions

When Workload Optimization Manager 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. Workload Optimization Manager 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, Workload Optimization Manager removes the delete action attached to it.

Delete actions include the 'Days Empty' information that indicates how long a plan has been empty.

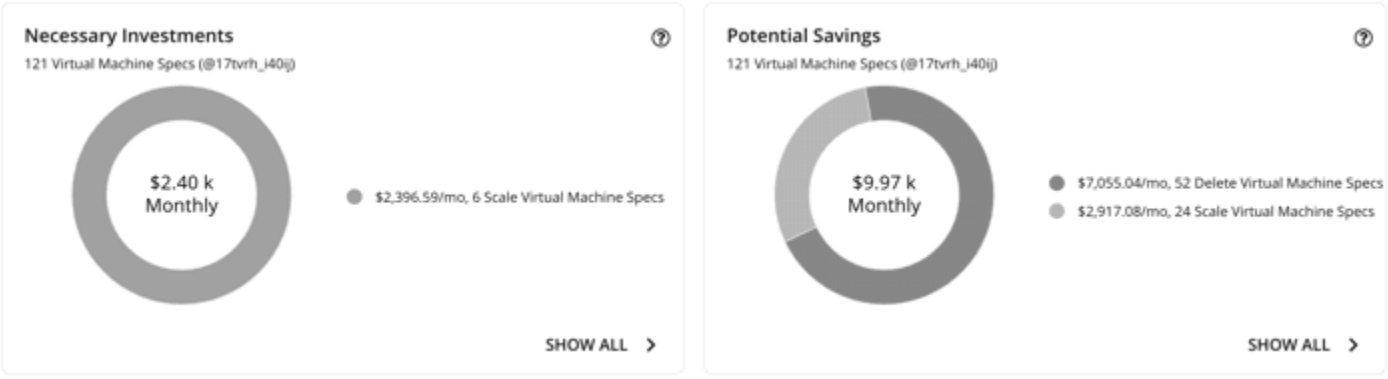


You can control the delete actions that Workload Optimization Manager recommends, based on the 'Days Empty' value that you set. For example, if you want Workload Optimization Manager to only generate delete actions for plans that have been empty for at least 5 days, perform these steps:

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.



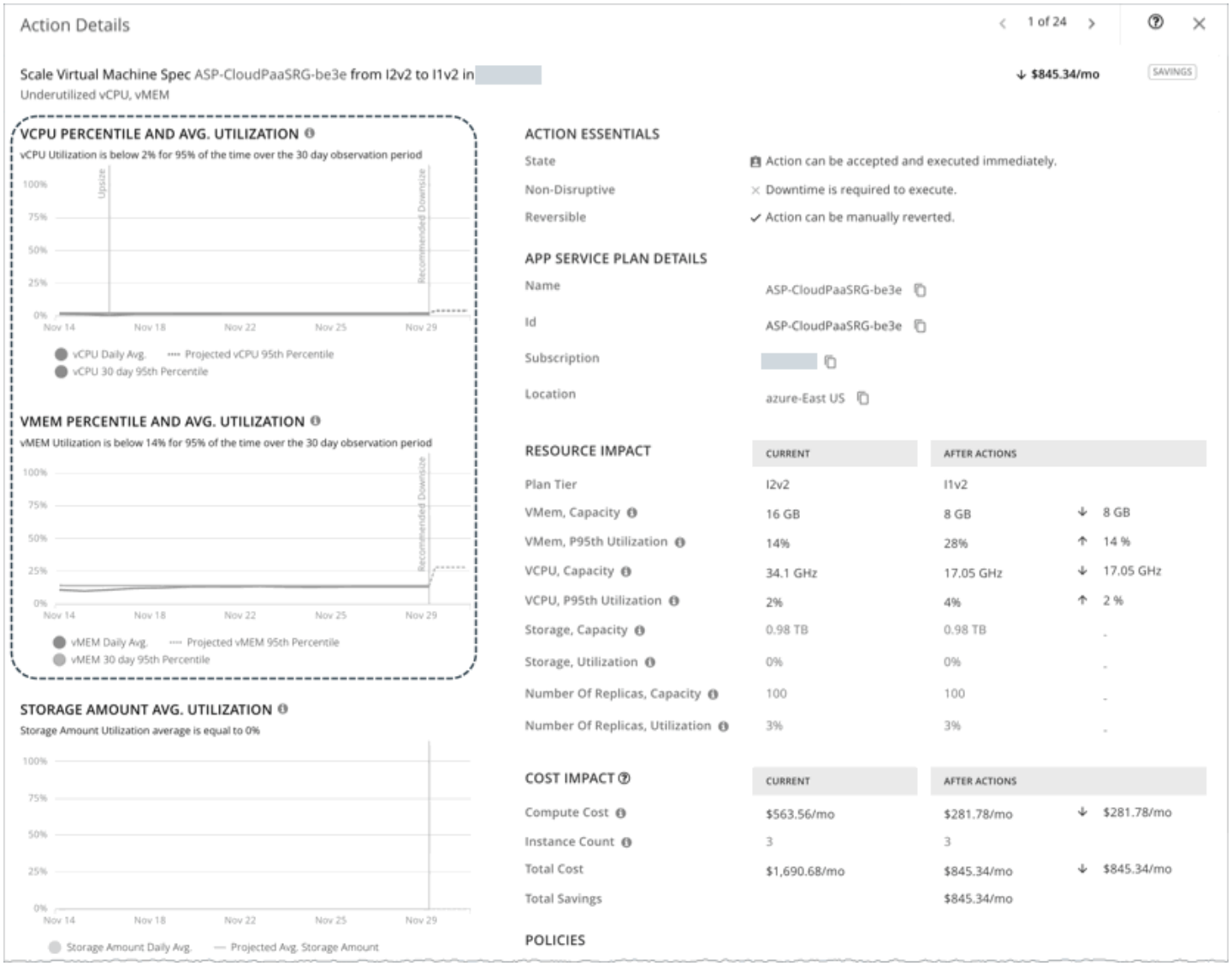
Click **Show All** for each chart to review and execute the actions.

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

Workload Optimization Manager 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.



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 Workload Optimization Manager'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 226)*.

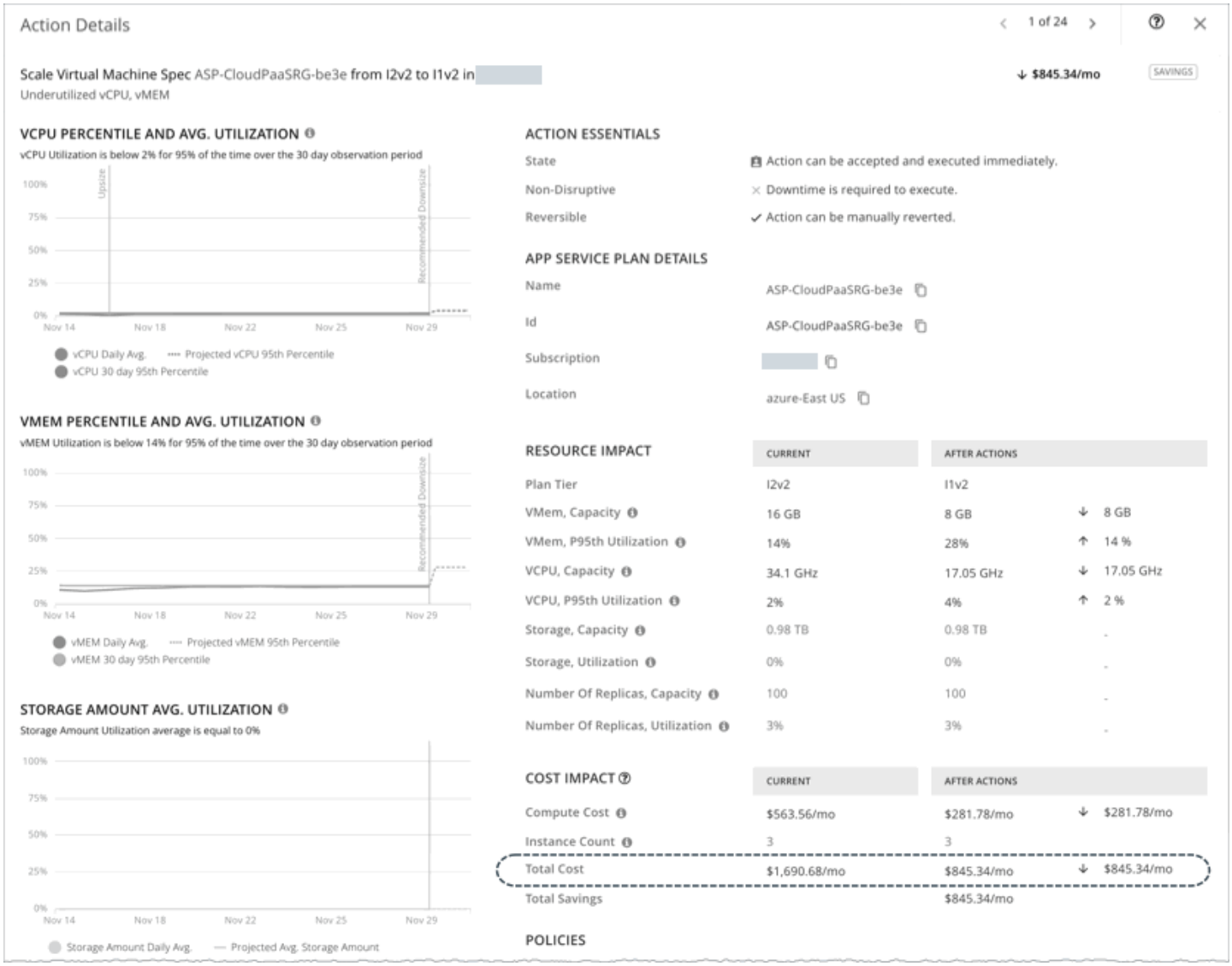
Disruptiveness and Reversibility of Scale Actions

Workload Optimization Manager 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

Workload Optimization Manager 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.



Cost Calculation

The calculation for estimated on-demand monthly cost can be expressed as follows:

```
(On-demand Compute Rate * 730) * Number of Instances = 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 Workload Optimization Manager 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 Workload Optimization Manager 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

Workload Optimization Manager calculates the following:

- *Current* estimated on-demand monthly cost:

```
($0.772 * 730) * 3 = $1690.68/Mo.
```

- Estimated on-demand monthly cost *after* executing the action:

```
($0.386 * 730) * 3 = $845.34/Mo.
```

NOTE:
Workload Optimization Manager rounds the calculated values that it displays in the user interface.

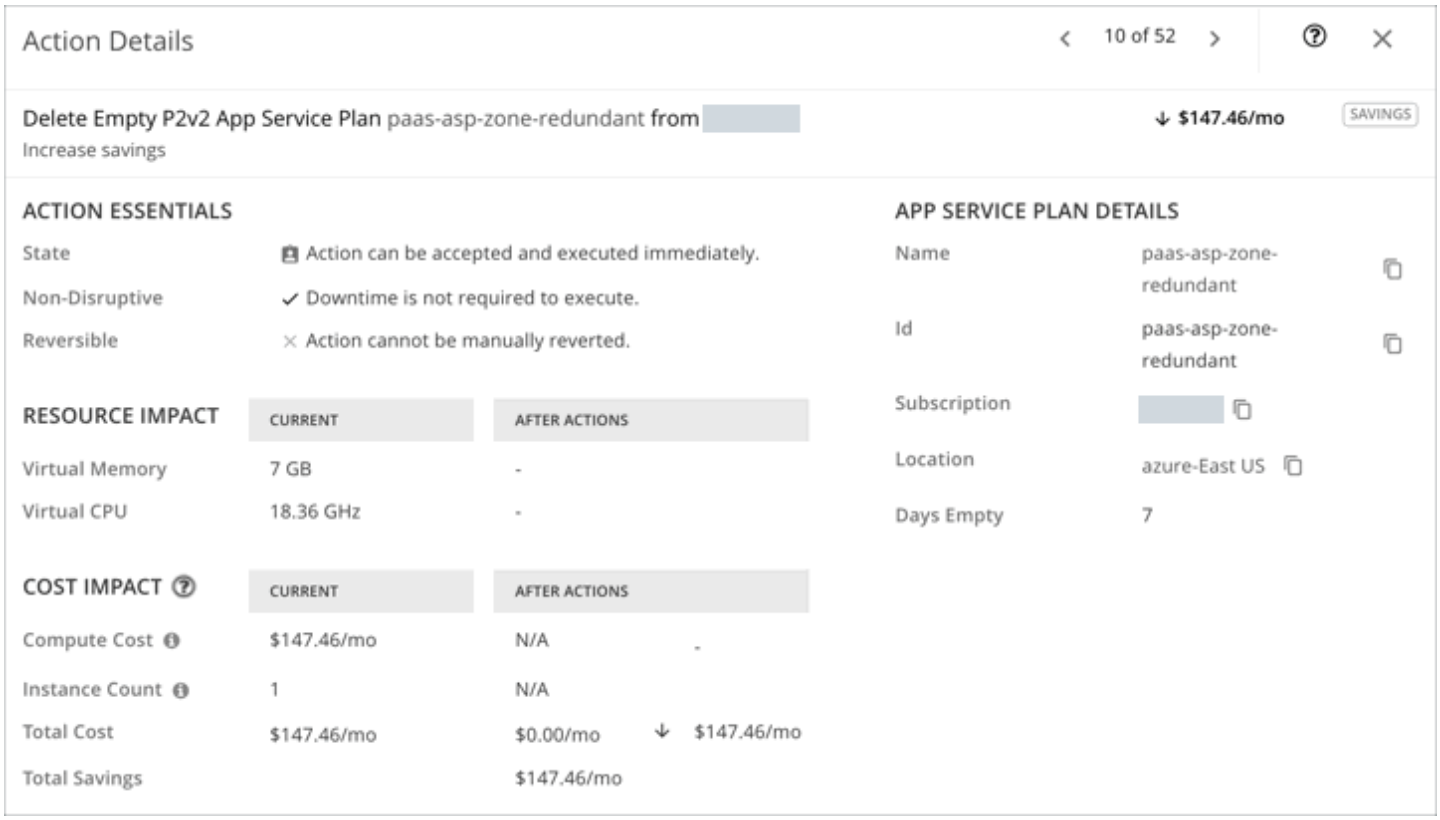
The estimated on-demand monthly cost is projected to decrease from $1690.68/month to $845.34/month, as shown in the Details section of the pending action.

Workload Optimization Manager treats the action as a cost-saving measure, and shows total savings of $845.34/month.

Estimated On-demand Monthly Savings for Empty Azure App Service Plans

Workload Optimization Manager 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.



Savings Calculation

The calculation for estimated on-demand monthly savings can be expressed as follows:

```
(On-demand Compute Rate * 730) * Number of Instances = 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 Workload Optimization Manager 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

Workload Optimization Manager calculates savings as follows:

```
($0.202 * 730) * 1 = $147.46/Mo.
```

> **NOTE:**
> Workload Optimization Manager rounds the calculated values that it displays in the user interface.

Workload Optimization Manager shows total savings of $147.46/month.

Virtual Machine Spec Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about Virtual Machine Spec actions, see Virtual Machine Spec Actions *(on page 218)*.

Action	Default Mode
Cloud Scale All	Manual

Action	Default Mode
Cloud Scale for Performance	Manual
Cloud Scale for Savings	Manual
Delete Virtual Machine Spec	Manual

When you create a policy, you can choose **Cloud Scale All**, **Cloud Scale for Performance**, or **Cloud Scale for Savings**.

- You can direct Workload Optimization Manager to only execute scaling actions that improve performance (*Cloud Scale for Performance*) or reduce costs (*Cloud Scale for Savings*). The default action mode for these actions is *Manual*. When you examine the pending actions for Virtual Machine Specs, only actions that satisfy the policies are allowed to execute. All other actions are read-only.
- *Cloud Scale All* enables all scaling actions, including those that result in efficiency improvements and increased costs.
- When policy conflicts arise, **Cloud Scale All** overrides the other two scaling options in most cases. For more information, see Relationship Between Scoped and Default Policies *(on page 93)*.

Scaling Sensitivity

Workload Optimization Manager uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

- **Aggressiveness**

Attribute	Default Value
Aggressiveness	95th Percentile

  Workload Optimization Manager uses these settings to calculate utilization percentiles for vCPU and vMEM.

  When evaluating performance, Workload Optimization Manager 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, Workload Optimization Manager 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 vCPU capacity. Without using a percentile, Workload Optimization Manager never scales below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Workload Optimization Manager will not reduce resources for the Virtual Machine Spec.

  With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager 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 Workload Optimization Manager 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 or 100th Percentile – More performance. Recommended for critical Virtual Machine Specs 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 Virtual Machine Specs that can stand higher resource utilization.

By default, Workload Optimization Manager 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. Workload Optimization Manager uses historical data from up to the number of days that you specify as a sample period. If the volume has fewer days' data then it uses all of the stored historical data.

Choose from the following settings:

- Less Elastic – Last 30 Days
- Recommended – Last 14 Days
- More Elastic – Last 3 or 7 Days

- **Min Observation Period**

Attribute	Default Value
Min Observation Period	None

This setting ensures historical data for a minimum number of days before Workload Optimization Manager 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.

Choose from the following settings:

- More Elastic – None
- Less Elastic – 1, 3, or 7 Days

Cloud Instance Types

Attribute	Default Value
Cloud Instance Types	None

By default, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

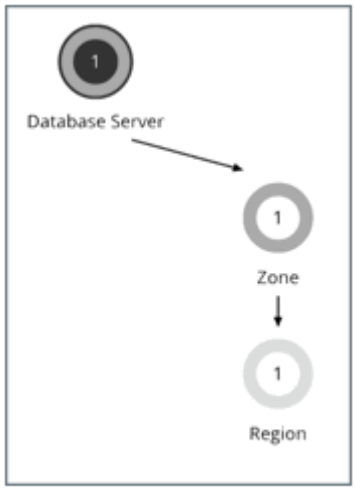
While these settings offer a way to modify how Workload Optimization Manager recommends actions, in most cases you should never need to use them. If you want to control how Workload Optimization Manager recommends actions to resize workloads, you can set the aggressiveness per the percentile of utilization, and set the length of the sample period for more or less elasticity on the cloud.

Database Server (Cloud)

In AWS public cloud environments, a Database Server is a relational database that you have configured using AWS Relational Database Service (RDS). Workload Optimization Manager discovers RDS instances through your AWS targets, and then generates scaling actions as needed.

> **NOTE:**
> Azure SQL Databases discovered by Workload Optimization Manager appear as *Database* entities in the supply chain. For details, see Database (Cloud) *(on page 251)*.

Synopsis



AWS RDS

Synopsis	
Budget:	A cloud Database Server has unlimited budget.
Provides:	Database services to cloud applications and end users
Consumes:	Compute and storage resources in the availability zone
Discovered through:	AWS targets

Permissions

Workload Optimization Manager requires the following permissions for AWS RDS Database Servers:

- Monitoring permissions:
  - cloudwatch:GetMetricData
  - pi:GetResourceMetrics
  - rds:DescribeDBInstances
  - rds:DescribeDBParameters
  - rds:ListTagsForResource
  - rds:DescribeOrderableDBInstanceOptions
- Action execution permissions:
  - rds:ModifyDBInstance

For a full list of permissions, see "Amazon Web Services" in the *Target Configuration Guide*.

Monitored Resources

Workload Optimization Manager monitors the following resources for a cloud Database Server:

- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.

- Virtual CPU

  Virtual CPU is the measurement of CPU utilized by the entity.
- Storage Amount

  Storage Amount is the amount of Amazon EBS storage utilized by the entity.
- Storage Access

  Storage Access is IOPS utilized by the entity.
- DB Cache Hit Rate (if available)

  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.
- Connections

  Connection is the measurement of Database Server connections utilized by applications.

Actions

Scale

Scale compute and storage resources to optimize performance and costs.

To recommend accurate scaling actions, Workload Optimization Manager 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 239)*.

Consider the following scenarios and actions:

- To address vCPU congestion, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager will not recommend new actions for the next six hours, in compliance with AWS's "cooldown" period for EBS storage.
- Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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%, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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.

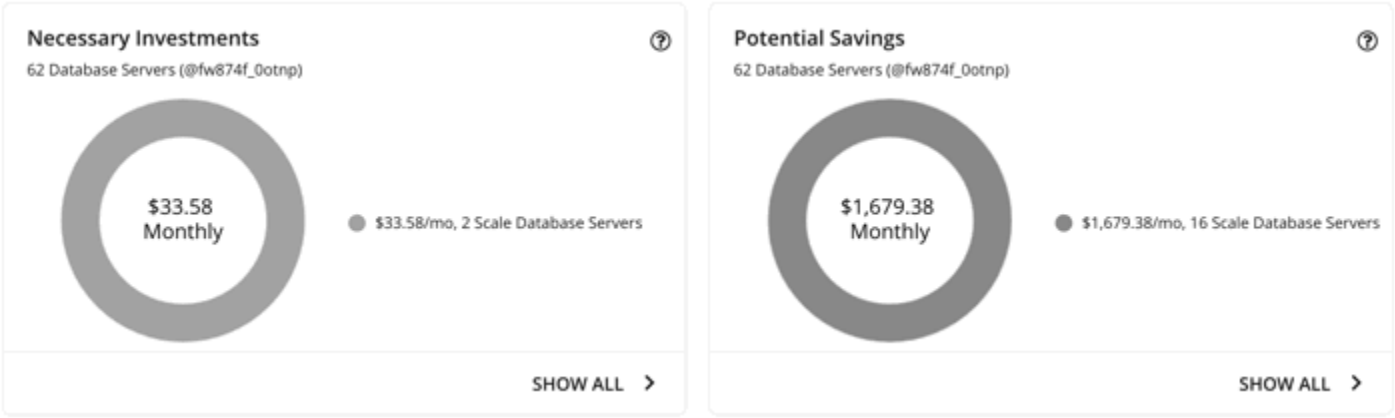
  – An action to scale to a different instance type in response to vCPU utilization might also include vMem changes, but vMem utilization alone (without cache hit rate metrics) will not drive actions.

Workload Optimization Manager 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. Workload Optimization Manager 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, Workload Optimization Manager calculates capacity as 682 connections (or `{8589934592/12582880}`). When Workload Optimization Manager 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.

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.



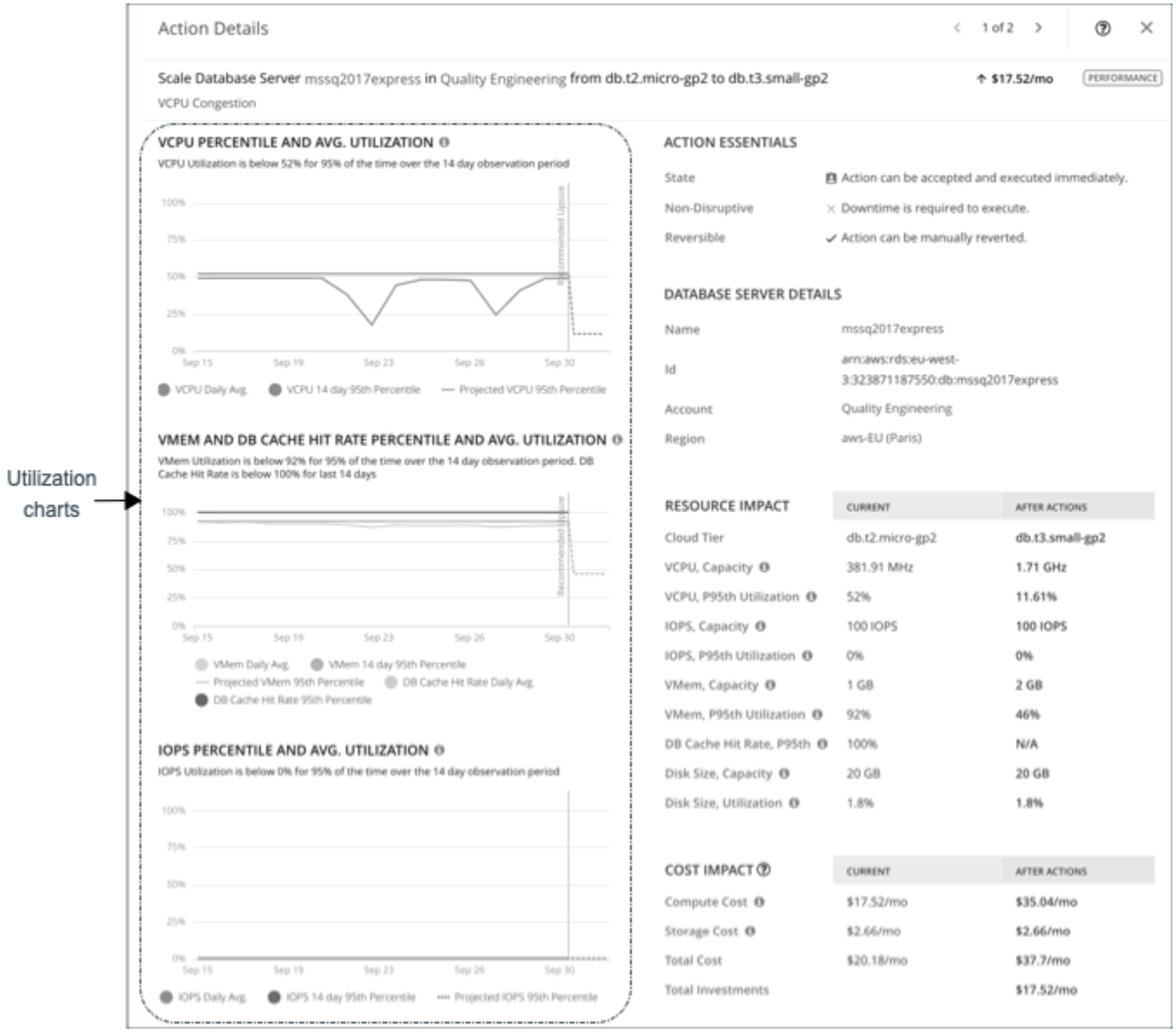
Click **Show All** for each chart to review and execute the actions.

The table lists all the actions that are pending for Database Servers, and the savings or investments for each action.

For details on how Workload Optimization Manager calculates savings or investments, see Estimated On-demand Costs for Cloud Database Servers *(on page 235)*.

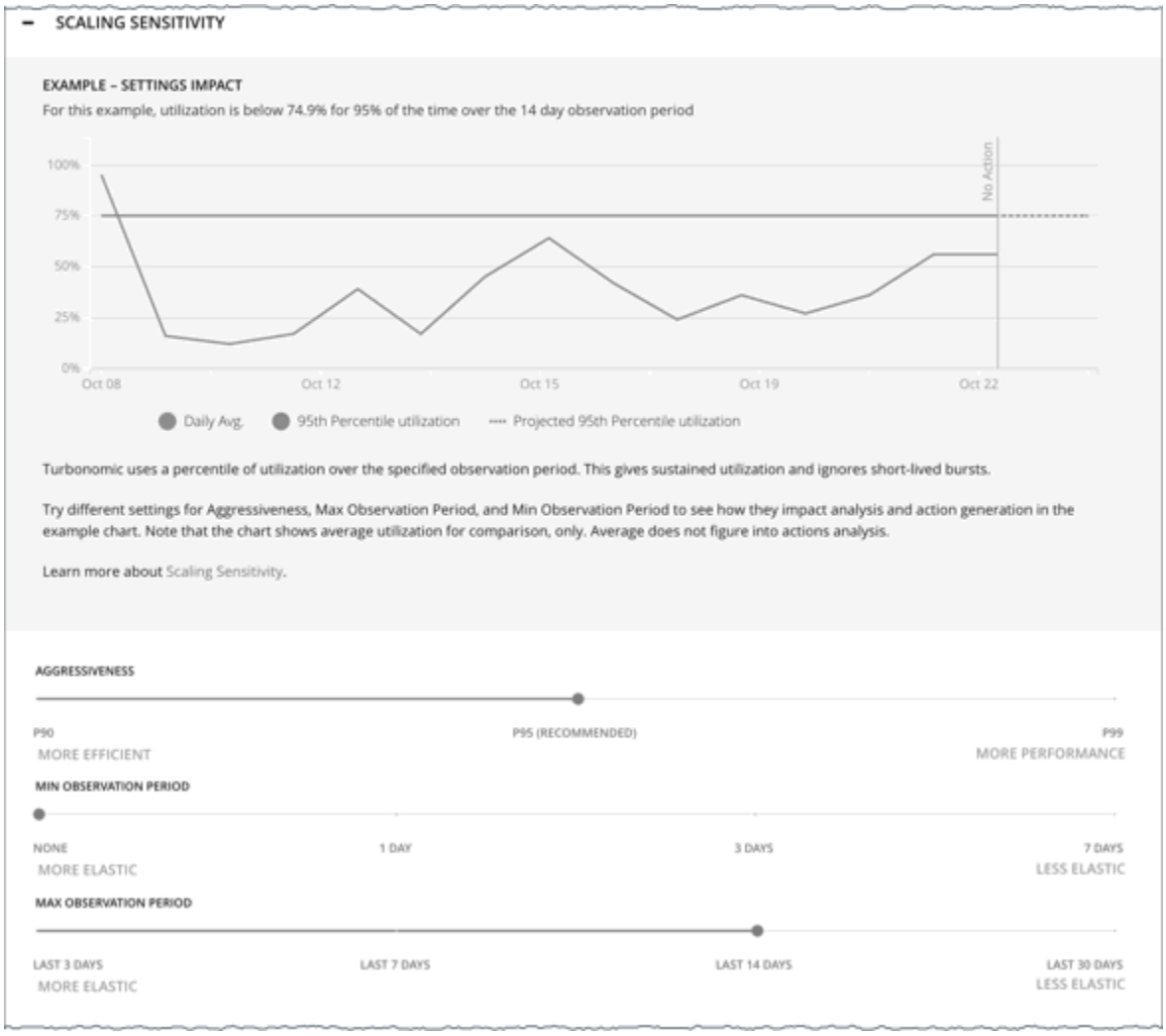
Utilization Charts

Workload Optimization Manager 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.



The charts also plot *daily average utilization* for your reference. If you have previously executed scaling actions on the Database Server, you can see the resulting improvements in daily average utilization. Put together, these charts allow you to easily recognize utilization trends that drive Workload Optimization Manager's scaling recommendations.

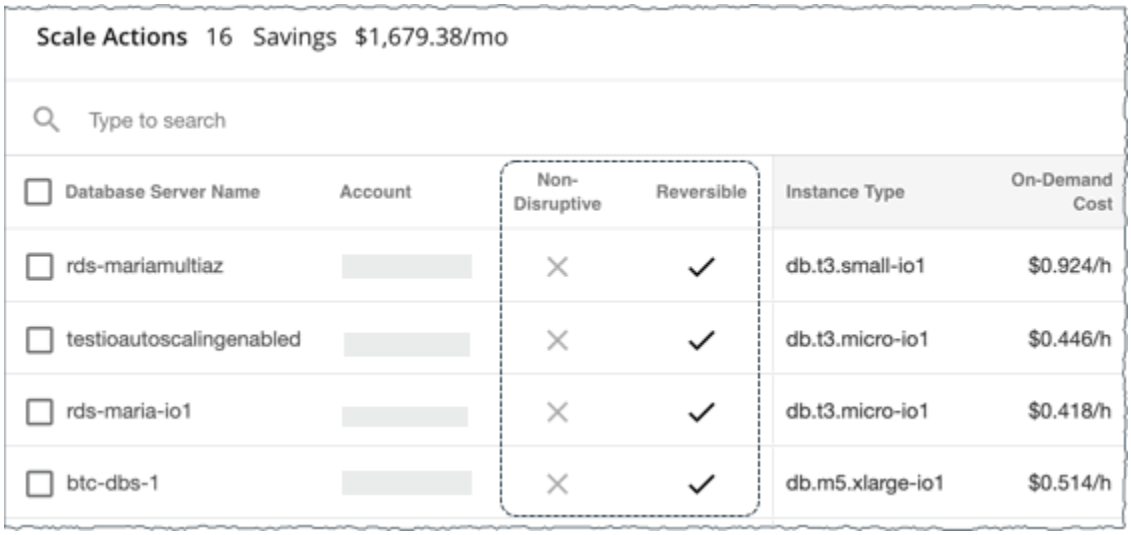
You can set scaling constraints in Database Server policies to refine the percentile calculations.



For details, see Scaling Sensitivity *(on page 239)*.

Non-disruptive and Reversible Scaling Actions

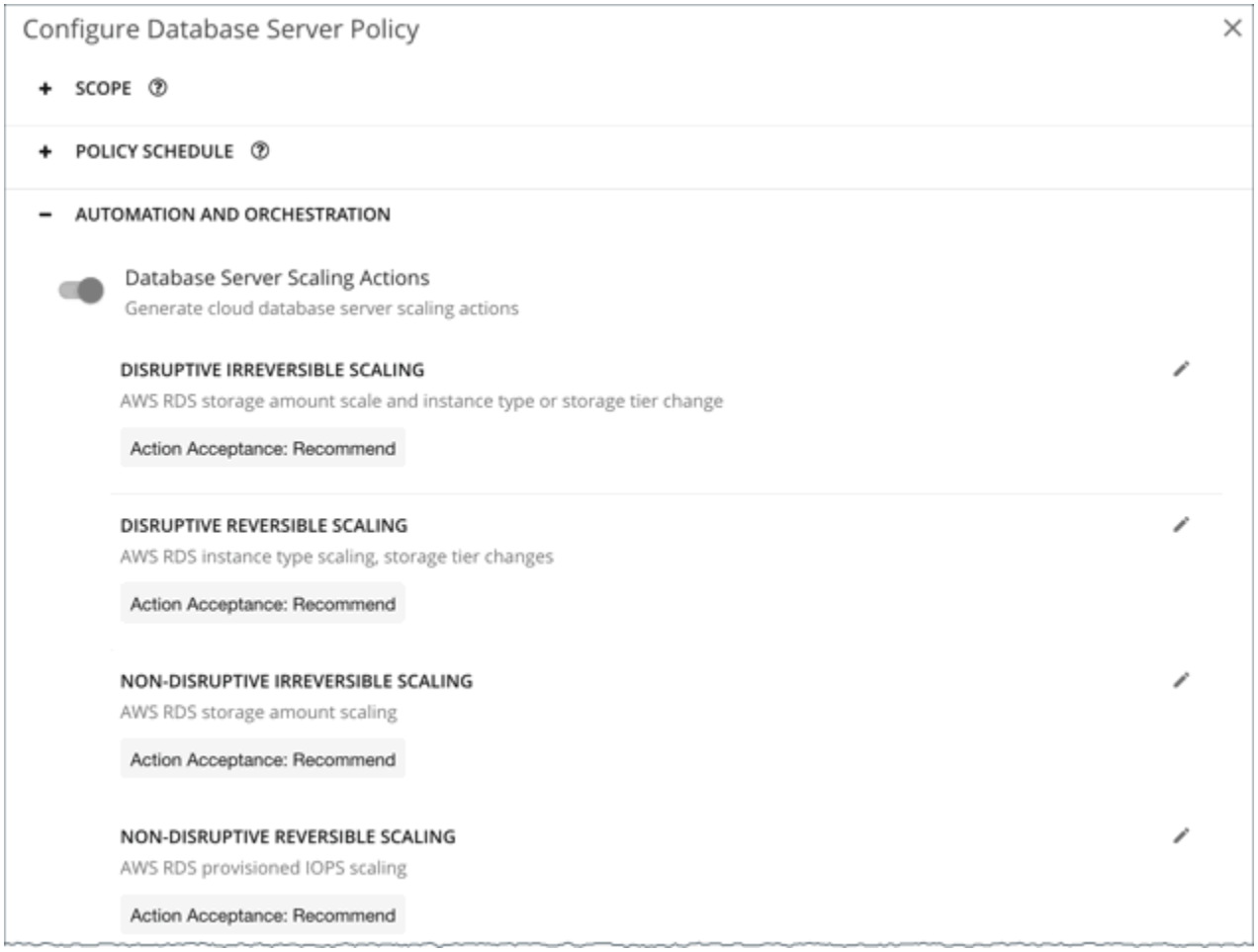
Workload Optimization Manager indicates whether a pending action is non-disruptive or reversible to help you decide how to handle the action.



The following table describes the disruptiveness and reversibility of the actions that Workload Optimization Manager 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
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 modes in policies to specify the degree of automation for these actions.



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:

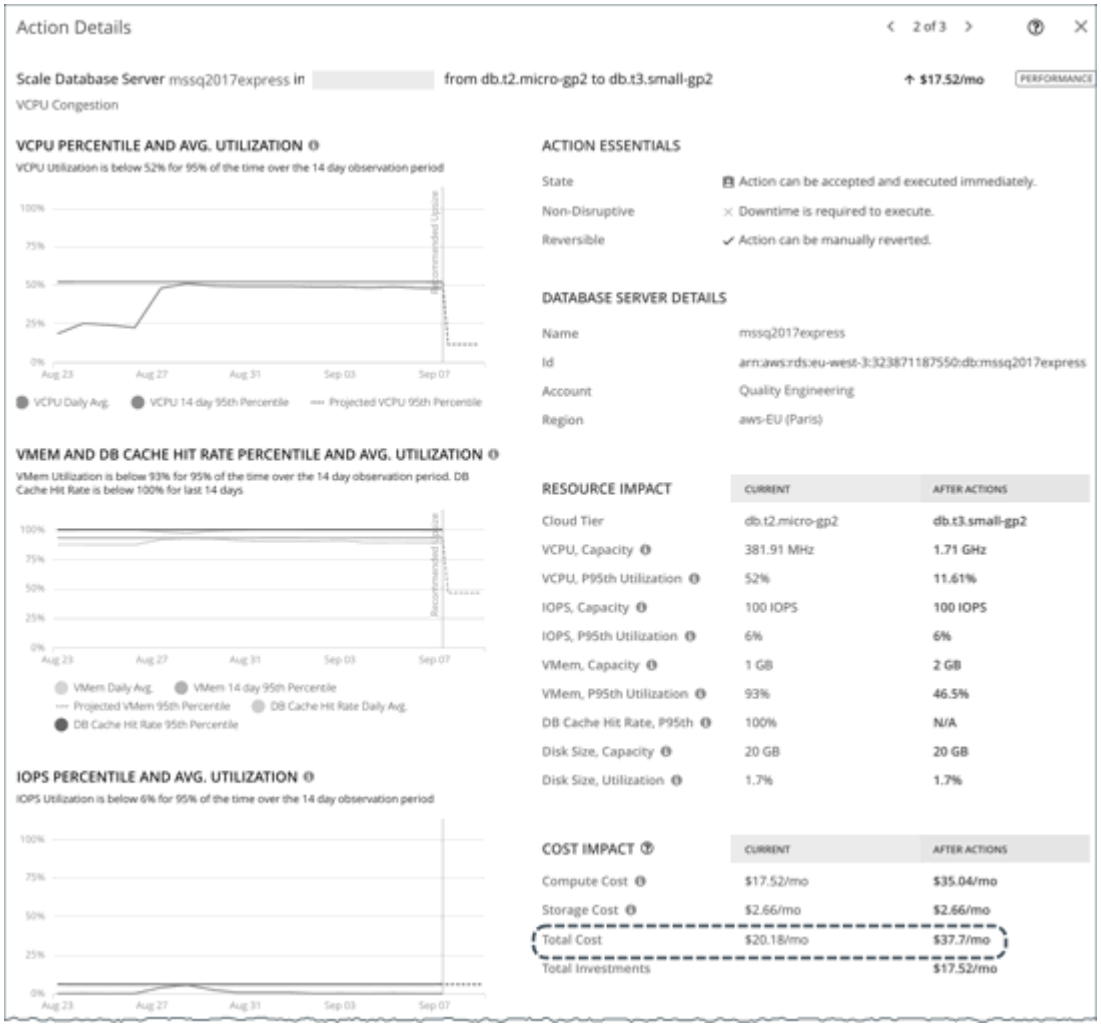
```
Cannot modify the instance class because there are no instances of the requested class available in the current instance's availability zone. Please try your request again at a later time.
```

> **NOTE:**
> For details about this error, see this AWS page.

When this error occurs, Workload Optimization Manager modifies the default Database Server policy to exclude the instance type from its scaling list. When the Database Server is available again, Workload Optimization Manager adds it back to the scaling list. For details about this list, see Cloud Instance Types *(on page 240)*.

Estimated On-demand Costs for Cloud Database Servers

Workload Optimization Manager 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:

```
(On-demand Compute Rate * 730) + (Provisioned Database Storage Rate * Provisioned Database Storage Amount) + (Provisioned IOPS Rate * Provisioned IOPS Amount) = 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 Workload Optimization Manager 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 Workload Optimization Manager 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

Workload Optimization Manager calculates the following:

- *Current* Estimated On-demand Monthly Cost:

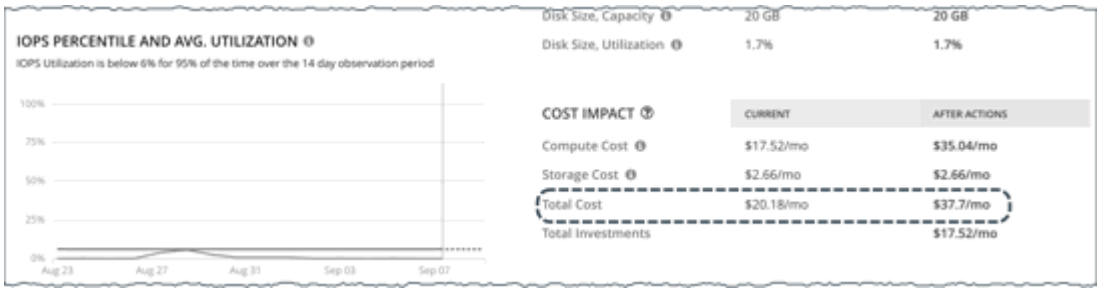
```
(0.024 * 730) + (0.133 * 20) + (0.00 * 0) = 20.18
```

- Estimated On-demand Monthly Cost *after* executing the action:

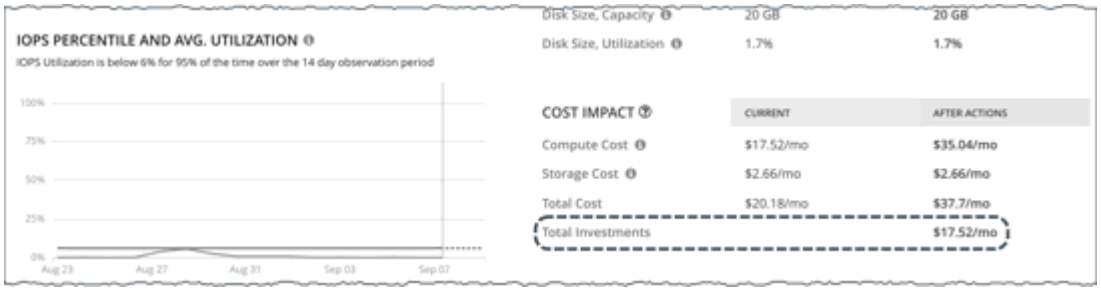
```
(0.048 * 730) + (0.133 * 20) + (0.00 * 0) = 37.7
```

NOTE:
Workload Optimization Manager 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.



Workload Optimization Manager 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:

```
(On-demand Compute Rate * 730) + (Provisioned Database Storage Rate * Provisioned
Database Storage Amount) + (I/O Request Rate * (Hourly Billed I/O Operation Count * 730))
= 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 Workload Optimization Manager 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 Workload Optimization Manager 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

Workload Optimization Manager 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:
Workload Optimization Manager 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, Workload Optimization Manager treats the action as a cost-saving measure and shows estimated savings of $89.79/month.

Cloud Database Server Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about cloud Database Server actions, see Cloud Database Server Actions *(on page 230)* and Non-disruptive and Reversible Scaling Actions *(on page 233)*.

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

Workload Optimization Manager uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Workload Optimization Manager uses these settings to calculate utilization percentiles for VCPU, VMEM and 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Workload Optimization Manager will not reduce resources for that Database Server.

With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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

Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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:**
> Workload Optimization Manager 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

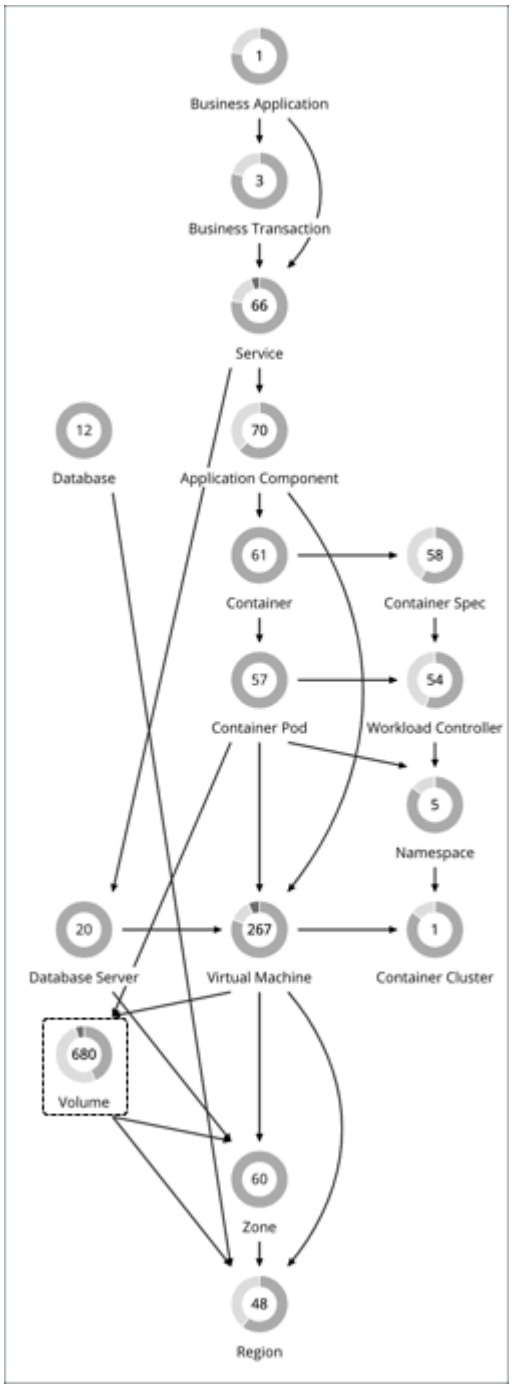
These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Workload Optimization Manager calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Workload Optimization Manager recommends actions, in most cases you should never need to use them. If you want to control how Workload Optimization Manager 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.

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	
Budget:	A cloud volume gains its budget by selling resources to the VMs that it serves.
Provides:	Storage resources for VMs to use: ▪ Storage Access ▪ Storage Amount ▪ IO Throughput
Consumes:	Storage services provided by Zones or Regions
Discovered through:	Cloud targets

Monitored Resources

Workload Optimization Manager monitors the following resources for AWS and Azure volumes:

- Storage Access

  The percentage of the volume's capacity for storage access operations (IOPS) that is in use.
- IO Throughput

  The percentage of the volume's capacity for IO throughput that is in use.
- IO Throughput Read

  The percentage of the volume's capacity for IO throughput Read that is in use.
- IO Throughput Write

  The percentage of the volume's capacity for IO throughput Write that is in use.

Notes:

- Workload Optimization Manager discovers Storage Amount (disk size) for AWS/Azure volumes, but does not monitor utilization.
- For a Kubeturbo (container) deployment that includes AWS/Azure volumes, Kubeturbo monitors Storage Amount utilization for the volumes. You can view utilization information in the Capacity and Usage chart.
- Currently, Workload Optimization Manager does not monitor resources for GCP volumes. It only monitors their attachment state and then generates delete actions for unattached volumes.

Actions

- **Scale**

  Scale attached volumes to optimize performance and costs.
- **Delete**

  Delete unattached volumes as a cost-saving measure.

Scale Actions

Scale attached AWS/Azure volumes to optimize performance and costs.

> **NOTE:**
> Currently, Workload Optimization Manager does not generate scale actions for GCP volumes.

Workload Optimization Manager 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 247)*.

- Scaling from one tier to another and then within the new tier, in a single action

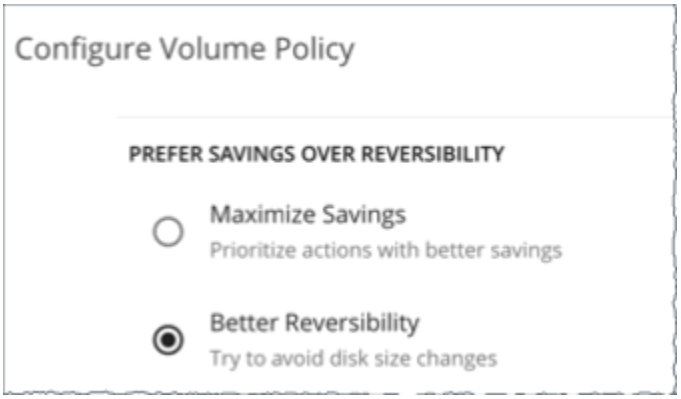
  For example, to achieve higher IOPS performance for VMs that are premium storage capable, you might see an action to scale the corresponding volume from *Standard* to *Premium*, and then scale up volume capacity from *32GB* to *256 GB*. This action increases costs and is *irreversible*, but is more cost effective than scaling up within the original tier.

  NOTE:
  Not all VM types are premium storage capable. For details, see the Azure Documentation.

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, Workload Optimization Manager 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 Workload Optimization Manager generates.

- Workload Optimization Manager 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, Workload Optimization Manager 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 251)*.

Delete Actions

Delete unattached volumes as a cost-saving measure. Workload Optimization Manager generates delete actions for AWS, Azure, and GCP volumes.

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

Workload Optimization Manager discovers Azure Site Recovery and Azure Backup volumes when you add Azure targets. Even though these volumes are always unattached, Workload Optimization Manager will never recommend deleting them because they are critical to business continuity and disaster recovery.

To identify Azure Site Recovery volumes, Workload Optimization Manager 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, Workload Optimization Manager checks for the RSV`aultBackup` 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

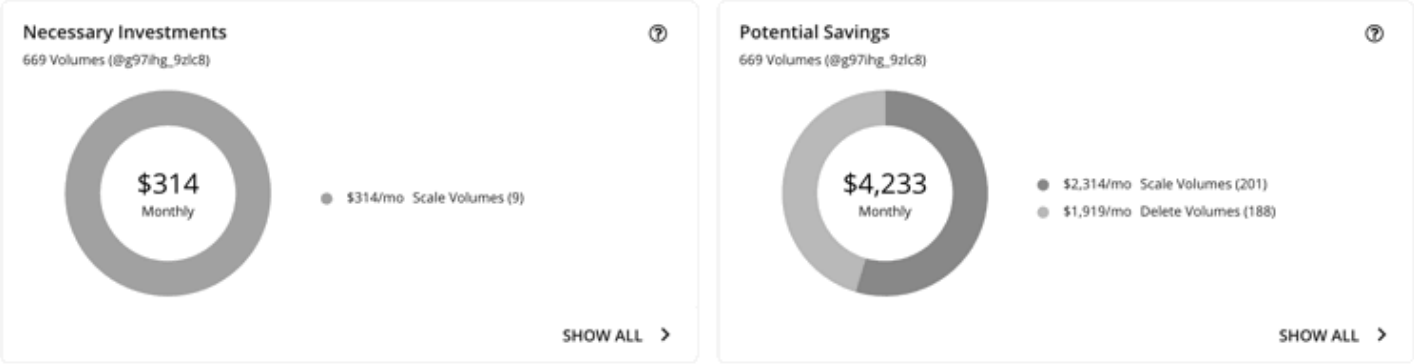
For Azure environments, Workload Optimization Manager can recommend scale and delete actions for locked volumes, but the lock level configured for the volumes may prevent some actions from executing.

- 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

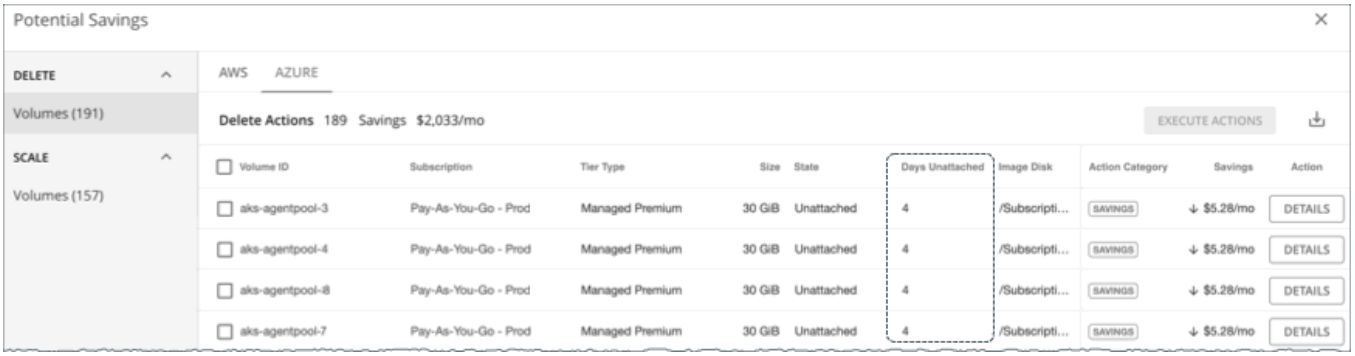
Use the Necessary Investments and Potential Savings charts to view pending volume actions. These charts show total monthly investments and savings, assuming you execute all the actions.



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:



– Number of days a volume has been unattached

  This information helps you decide whether to take the action.

  Workload Optimization Manager 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.

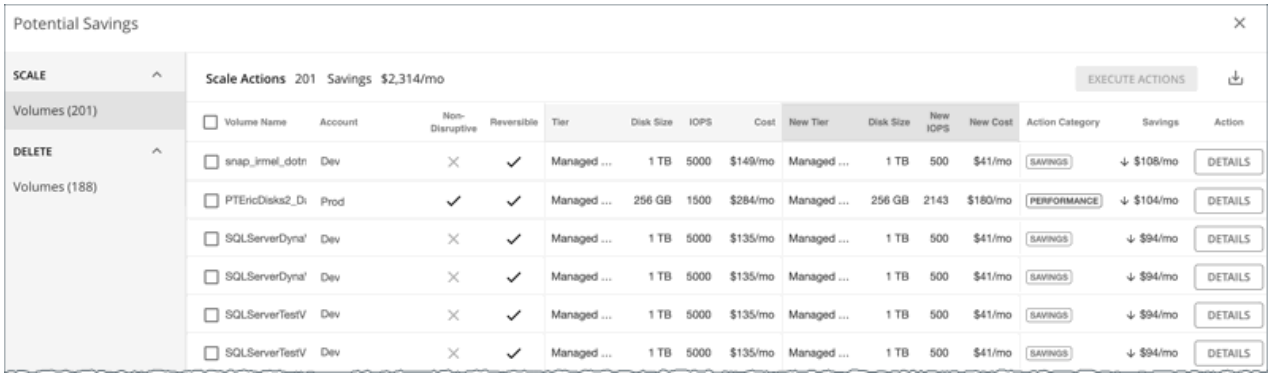
  For newly unattached volumes, Workload Optimization Manager shows a dash symbol (-) if a volume has been unattached within the last 6 hours. A value of 0 (zero) means that a volume has been unattached within the last 24 hours.

  Once Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager removes them from its records after 15 days.

  To see the last known VM attached to the volume, click **DETAILS**.

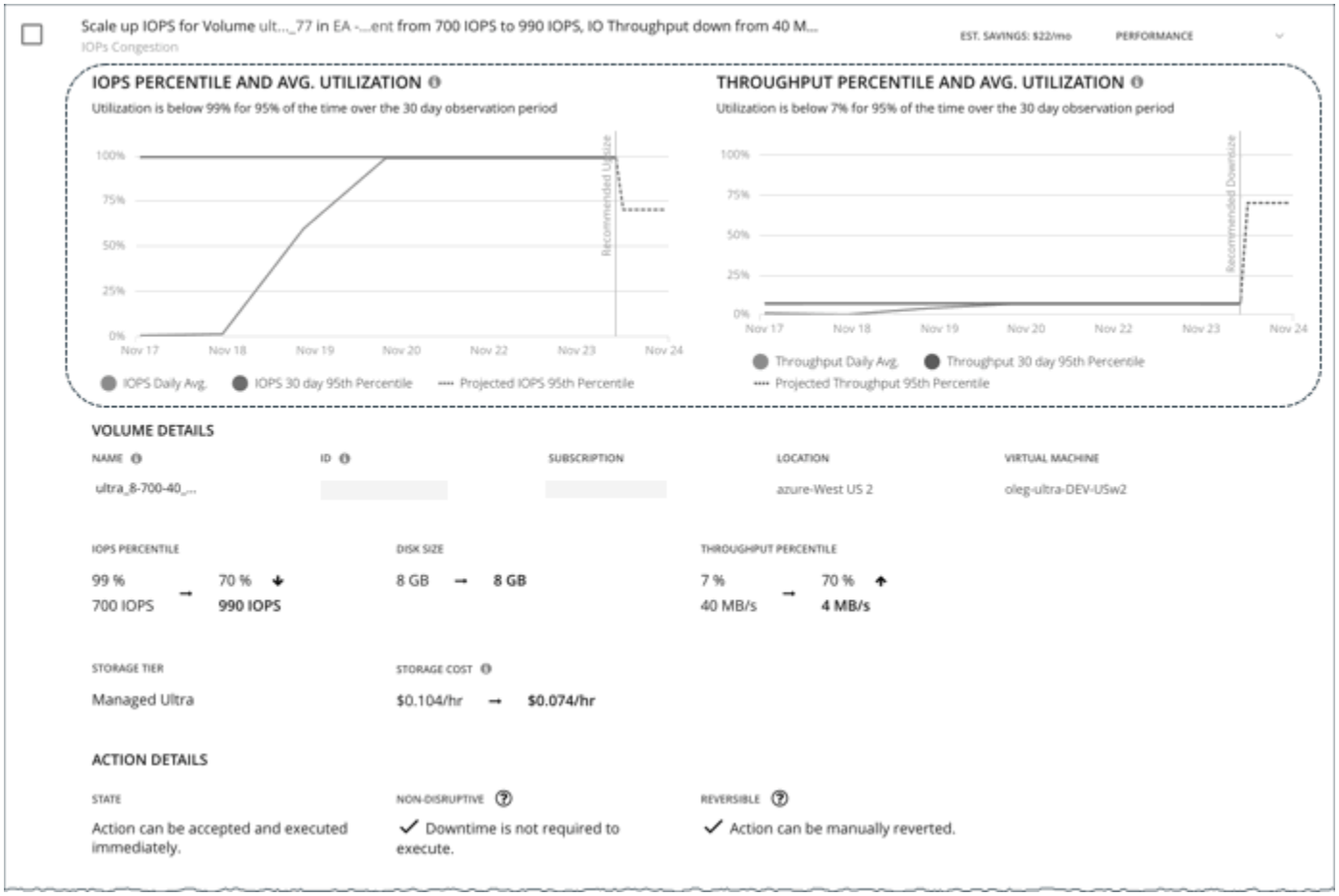
– For *Scale* actions in the Potential Savings or Necessary Investments chart:



  - 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



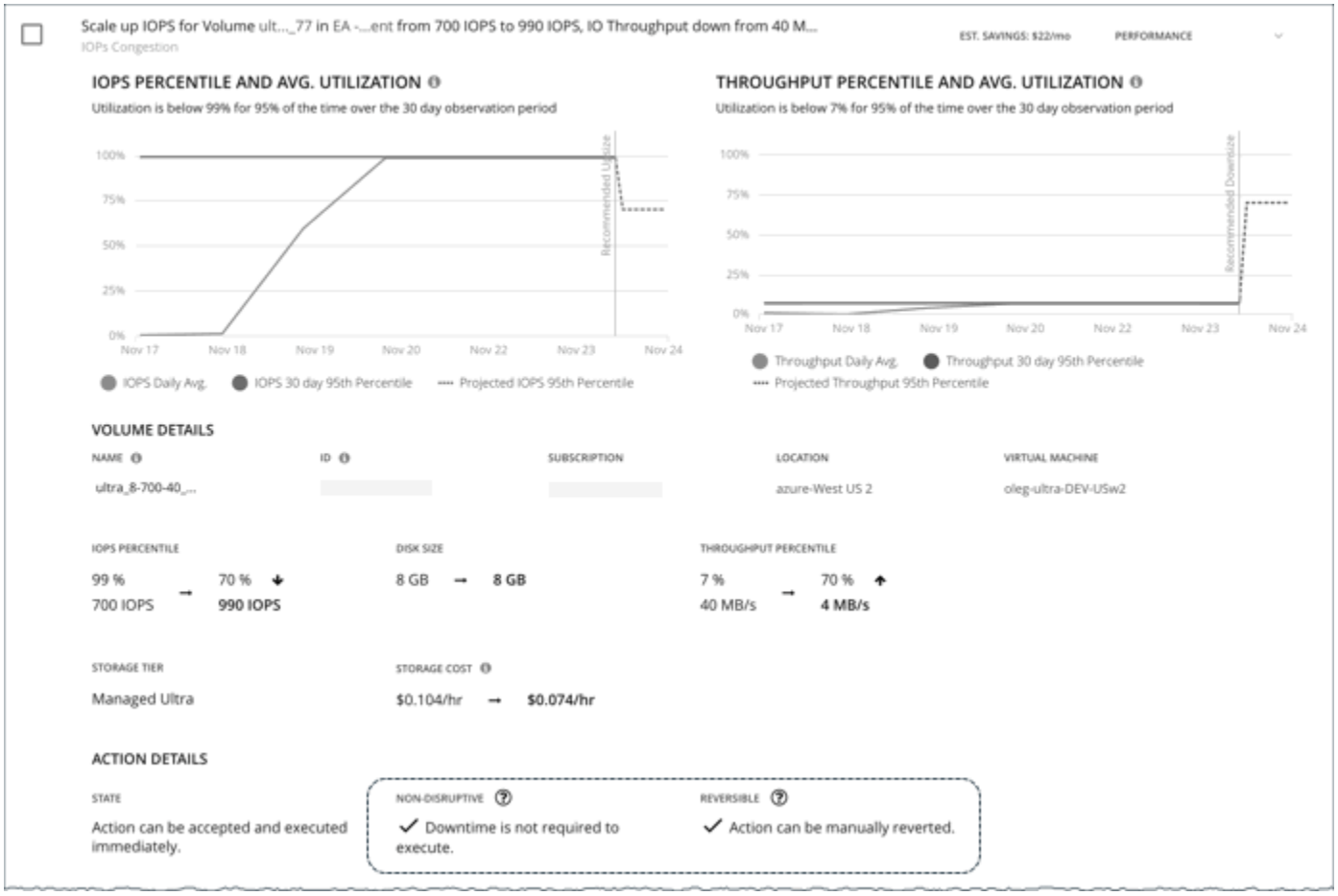
Workload Optimization Manager uses percentile calculations to measure IOPS and throughput more accurately, and drive scaling actions that improve overall utilization and reduce costs. When you examine the details for a pending scaling action on a volume, you will see charts that highlight resource *utilization 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 Workload Optimization Manager's scaling recommendations.

> **NOTE:**
> You can set scaling sensitivity in volume policies to refine the percentile calculations. For details, see Scaling Sensitivity *(on page 249)*.

Non-disruptive and Reversible Scaling Actions

Workload Optimization Manager indicates whether a pending action is non-disruptive or reversible.



- **Non-disruptive**

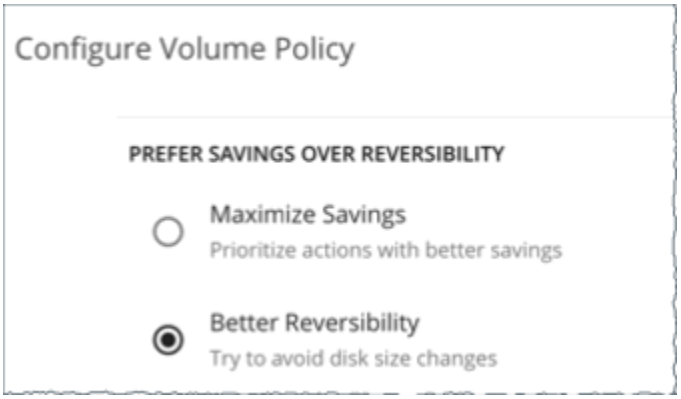
  Executing storage scaling actions can sometimes be disruptive if the VM must be rebooted to execute a storage change. For example, Azure Standard and Premium scaling actions are *disruptive*. When a storage action is disruptive, expect a single reboot (usually 2–3 minutes of downtime).

  The following scaling actions are *non-disruptive*:

  - Scaling IOPS and throughput on Azure Ultra storage
  - All scaling actions on AWS storage

- **Reversible**

  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**.



Cloud Volume Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about cloud volume actions, see Cloud Volume Actions *(on page 243)*.

Action	Default Mode
Scale	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

Workload Optimization Manager uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

- **Aggressiveness**

Attribute	Default Value
Aggressiveness	95th Percentile

  Workload Optimization Manager uses Aggressiveness when evaluating IOPS and throughput.

  When evaluating performance, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager never scales below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Workload Optimization Manager will not reduce resources for that volume.

  With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager 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 Workload Optimization Manager 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 or 100th Percentile – More performance. Recommended for critical volumes 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 volumes that can stand higher resource utilization.

By default, Workload Optimization Manager uses samples from the last 30 days. Use the **Max Observation Period** setting to adjust the number of days.

- **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. Workload Optimization Manager uses historical data from up to the number of days that you specify as a sample period. If the volume has fewer days' data then it uses all of the stored historical data.

Choose from 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	None

This setting ensures historical data for a minimum number of days before Workload Optimization Manager 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.

Choose from the following settings:

- More Elastic – None
- Less Elastic – 1, 3, or 7 Days

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, Workload Optimization Manager 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.

Database (Cloud)

Workload Optimization Manager discovers SQL Databases through your Azure targets. In particular, it discovers the resources on *individual databases* that are managed under both the DTU (Database Transaction Unit) and vCore pricing models.

- DTU Pricing Model

  In the DTU model, Azure bundles CPU, memory, and IOPS as a single DTU metric. Workload Optimization Manager actions on these databases consider both DTU and storage utilization.

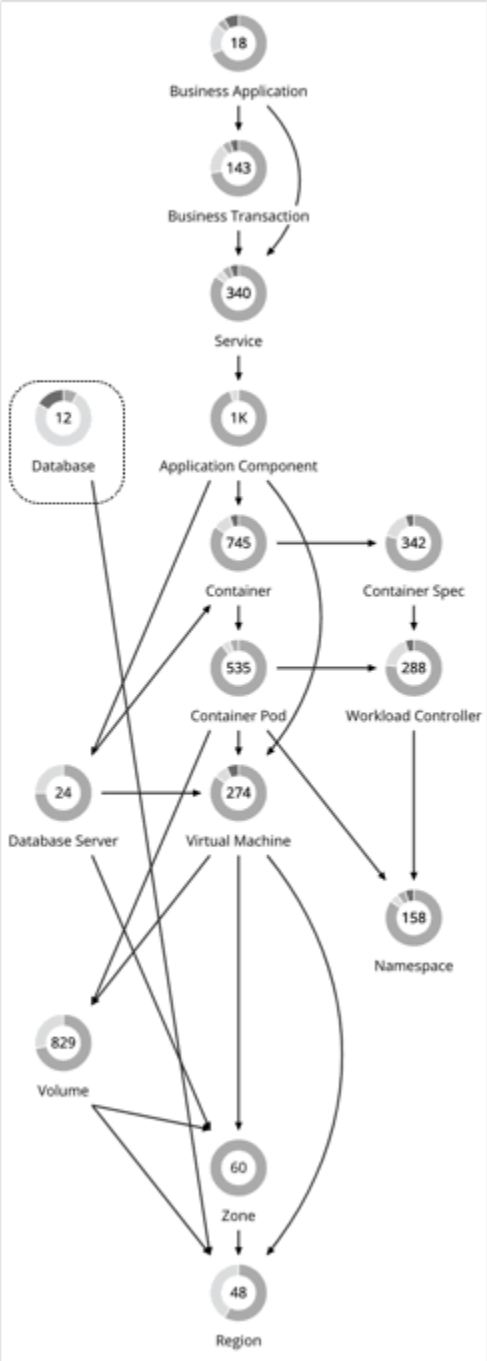
- vCore Pricing Model

  In the vCore model, analysis can track CPU, memory, IOPS, and throughput metrics in isolation. Workload Optimization Manager actions on these databases are driven by CPU, memory, IOPS, throughput and storage utilization.

NOTE:
For more information about the DTU and vCore models, see the Azure documentation.

AWS RDS databases appear as *Database Server* entities in the supply chain. For details, see Database Server (Cloud) *(on page 228)*.

Synopsis



Synopsis	
Budget:	A database has unlimited budget.
Provides:	Transactions to end users
Consumes:	▪ DTU Pricing Model: DTU and storage resources in an Azure region

Synopsis	
	▪ vCore Pricing Model: vCPU, vMem, IOPS, throughput, and storage resources in an Azure region
Discovered through:	Azure targets

Actions analysis also considers levels of concurrent workers and concurrent sessions, to constrain instance type selection. In all cases, Workload Optimization Manager database scaling actions aim to increase resource utilization and reduce costs while complying with business policies.

Monitored Resources

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

- DTU Pricing Model
  - DTU

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

    Storage is the storage capacity for the database.
- vCore Pricing Model
  - Virtual Memory

    Virtual Memory is the measurement of memory utilized by the entity.
  - Virtual CPU

    Virtual CPU is the measurement of CPU utilized by the entity.
  - Storage Access

    Storage Access is IOPS utilized by the entity.
  - Throughput

    Throughput is the utilization of transaction log write IO available to the entity.
  - Storage

    Storage is the storage capacity for the entity.

Workload Optimization Manager drives scaling actions based on the utilization of these 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.

Actions

Scale

- DTU Model

  Scale DTU and storage resources to optimize performance and costs.
- vCore Model

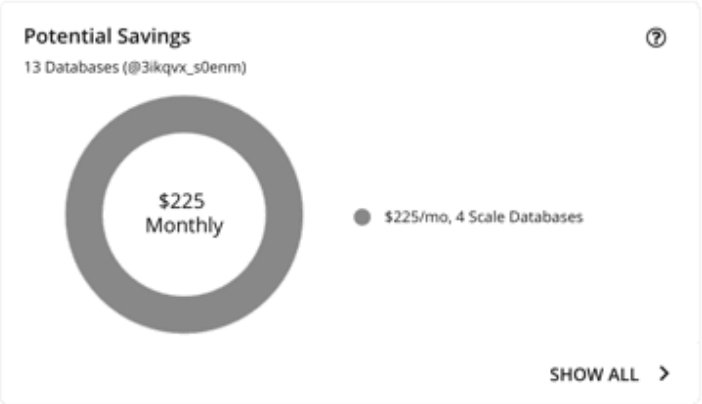
  Scale vCPU, vMem, IOPS, throughput and storage resources to optimize performance and costs.

Points to consider:

- Workload Optimization Manager will *not* recommend:
  - Scaling from one pricing model to another
  - Scaling vCore databases to instance types running Gen4 hardware. This hardware generation is nearing end-of-life 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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.

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.



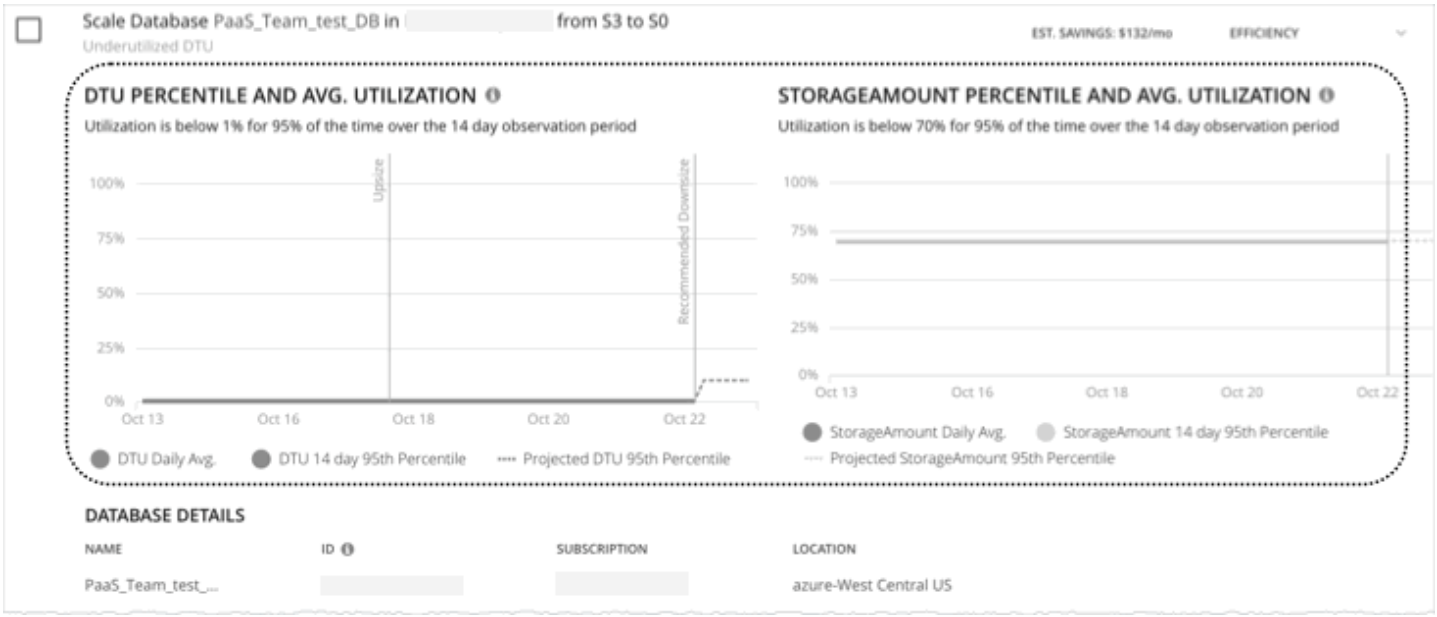
Click **Show All** for each chart to review and execute the actions.

The table shows the following:

- Actions that are pending for each database
- Savings or investments for each database

Utilization Charts for Scale Actions

Workload Optimization Manager 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 Workload Optimization Manager'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 259)*.

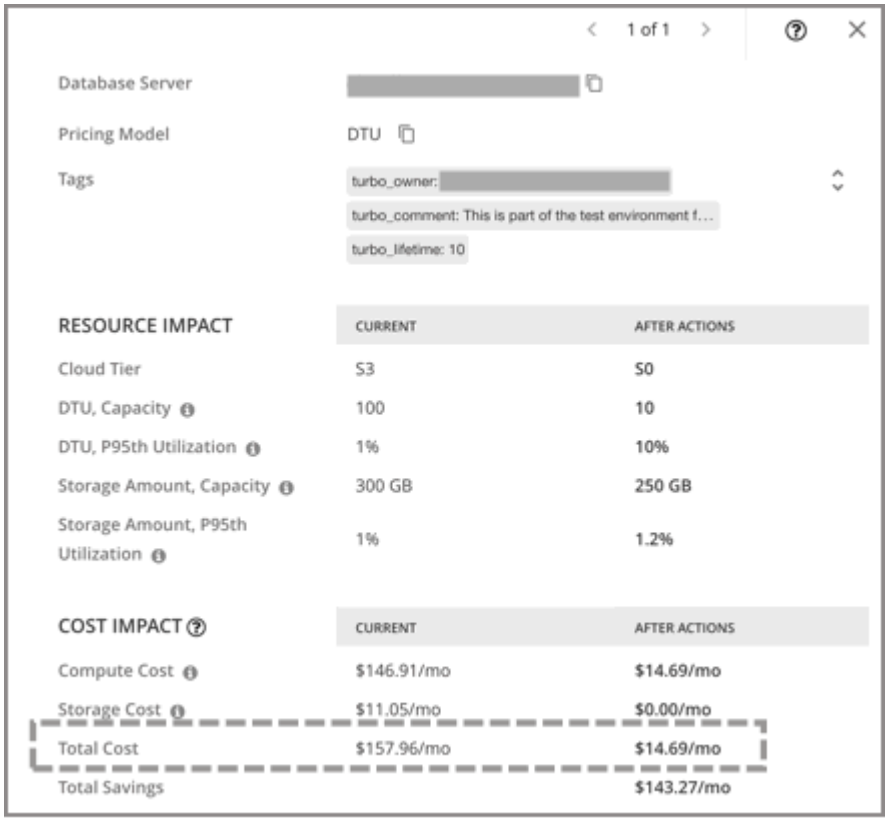
Non-disruptive and Reversible Scaling Actions

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.

Estimated On-demand Costs for Cloud Databases

Workload Optimization Manager considers a variety of factors when calculating *Estimated On-demand Monthly Cost* for an Azure SQL Database.



Azure SQL DTU Databases

Cost Calculation

For Azure SQL DTU Databases, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

```
(On-demand Compute Rate * 730)) + (Provisioned Database Storage Rate * (Provisioned Database Storage Amount - Performance Level Included Storage)) = 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 Workload Optimization Manager 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 Workload Optimization Manager 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

Workload Optimization Manager calculates the following:

- *Current* Estimated On-demand Monthly Cost:

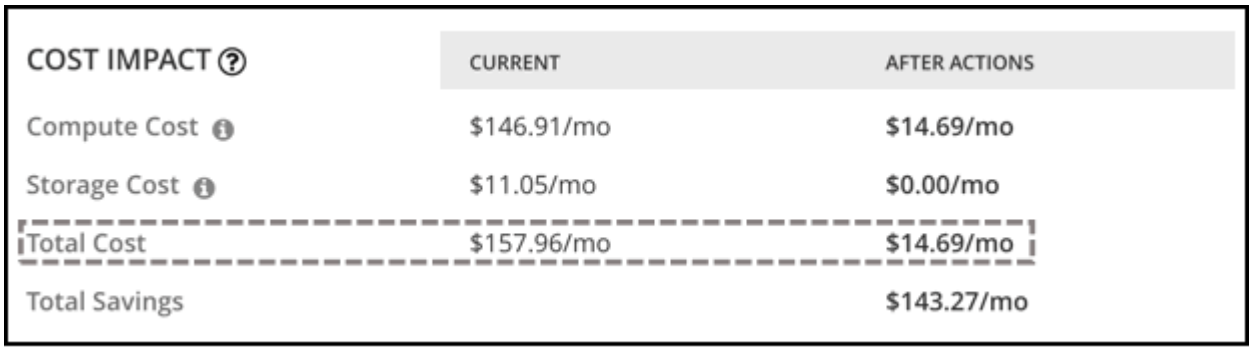
```
($0.20125 * 730) + ($0.221 * (300 - 250)) = $157.96/Mo.
```

- Estimated On-demand Monthly Cost *after* executing the action:

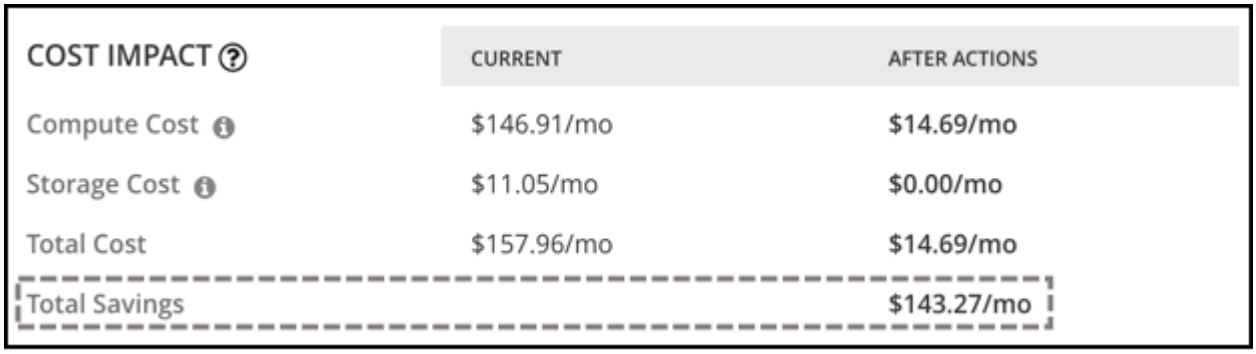
```
($0.020125 * 730) + ($0.221 * (250 - 250)) = $14.69/Mo.
```

NOTE:
Workload Optimization Manager 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.



Workload Optimization Manager treats the action as a saving, and shows an estimated savings of $143.27/month.



Azure SQL vCore Databases

Cost Calculation

For Azure SQL vCore Databases, the calculation for Estimated On-demand Monthly Cost can be expressed as follows:

```
(On-demand Compute Rate * 730) + (SQL License Rate * 730) + (Provisioned Database Storage Rate * (Provisioned Database Storage Amount + Log Space Allocated)) = 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 Workload Optimization Manager 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 Workload Optimization Manager 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

Workload Optimization Manager calculates the following:

- *Current* Estimated On-demand Monthly Cost:

  ```
  ($1.068 * 730) + ($0.799728 * 730) + ($0.115 * (32 + 9.6)) = $1368.23/Mo.
  ```

- Estimated On-demand Monthly Cost *after* executing the action:

  ```
  ($0.304 * 730) + ($0.199932 * 730) + ($0.115 * (5 + 1.5)) = $368.62/Mo.
  ```

NOTE:
Workload Optimization Manager 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, Workload Optimization Manager treats the action as a cost-saving measure and shows estimated savings of $999.61/month.

Cloud Database Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about cloud database actions, see Cloud Database Actions *(on page 253)*.

Action	Default Mode
Cloud DB Scale	Manual

Scaling Sensitivity

Workload Optimization Manager uses a percentile of utilization over the specified observation period. This gives sustained utilization and ignores short-lived bursts.

Workload Optimization Manager uses these settings to calculate utilization percentiles for DTU 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager never resizes below the recognized peak utilization. Assume utilization peaked at 100% just once. Without the benefit of a percentile, Workload Optimization Manager will not reduce resources for that database.

  With **Aggressiveness**, instead of using the single highest utilization value, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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

Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager considers all instance types currently available for scaling when making scaling decisions for databases. However, you may have set up your cloud 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 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 Workload Optimization Manager 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. To meet a target DTU utilization, the workloads must be members of a DTU pricing model. To meet individual VCPU, VMEM, or IOPs/Throughput targets, the workloads must be members of a vCore pricing model.

Attribute	DTU Pricing: Default Value	vCore Pricing: Default Value
Scaling Target DTU Utilization	70	N/A
VCPU	N/A	70
VMEM	N/A	90
IOPs/Throughput	N/A	70
Storage Amount Utilization	90	90

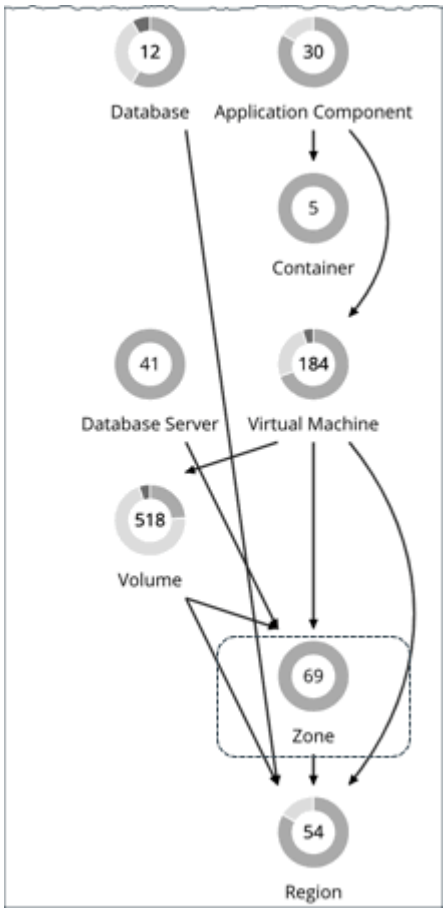
These advanced settings determine how much you would like a scope of workloads to utilize their resources. These are fixed settings that override the way Workload Optimization Manager calculates the optimal utilization of resources. You should only change these settings after consulting with Technical Support.

While these settings offer a way to modify how Workload Optimization Manager recommends actions, in most cases you should never need to use them. If you want to control how Workload Optimization Manager 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.

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	
Budget:	Workload Optimization Manager assumes a Zone has infinite resources.
Provides:	Compute and storage resources to VMs.
Consumes:	Region resources.
Discovered through:	Workload Optimization Manager discovers Zones through public cloud targets.

Monitored Resources

For public cloud environments, Workload Optimization Manager discovers the resources that an availability zone provides, including:

- 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.

- Relational Database Services (RDS)

  The RDS capabilities each cloud account provides.

- Storage Tiers

  Workload Optimization Manager discovers the storage tier that supports your workloads, and uses the tier pricing to calculate storage cost.

- Billing

  Workload Optimization Manager 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.

Workload Optimization Manager monitors the following resources for a Zone:

- 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.

- IO Throughput Read

  For environments that measure IO throughput read, the percentage utilized of throughput capacity for the zone.

- IO Throughput Write

  For environments that measure IO throughput write, 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.

- Net Throughput Inbound

  For environments that measure Net throughput Inbound, the percentage utilized of throughput inbound capacity for the zone.

- Net Throughput Outbound

  For environments that measure Net throughput Outbound, the percentage utilized of throughput outbound capacity for the zone.

Actions

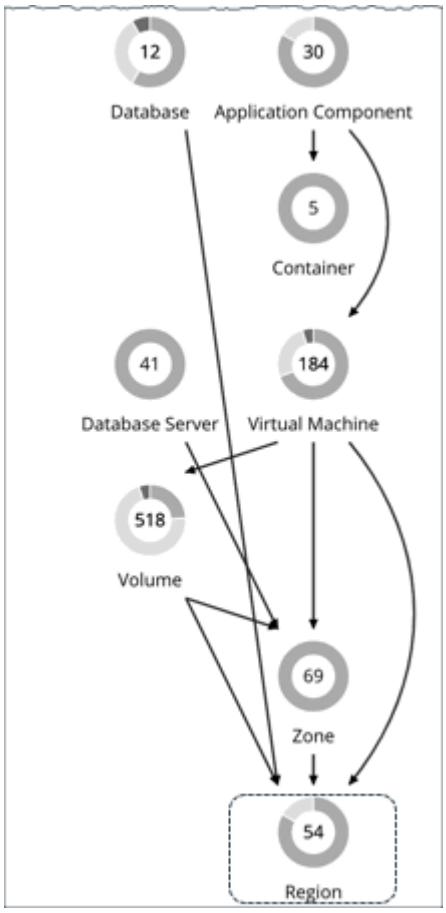
None

Workload Optimization Manager 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	
Budget:	Workload Optimization Manager assumes a Region has infinite resources.
Provides:	Hosting and storage resources to Zones.
Consumes:	NA
Discovered through:	Cloud service accounts, such as accounts on Amazon AWS, or subscriptions on Microsoft Azure.

Monitored Resources

Workload Optimization Manager 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.

- IO Throughput Read

  For environments that measure IO throughput read, the percentage utilized of throughput capacity for the zones.

- IO Throughput Write

  For environments that measure IO throughput write, 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.

- Net Throughput Inbound

  For environments that measure Net throughput Inbound, the percentage utilized of throughput inbound capacity for the zones.

- Net Throughput

  For environments that measure Net throughput Outbound, the percentage utilized of throughput outbound capacity for the zones.

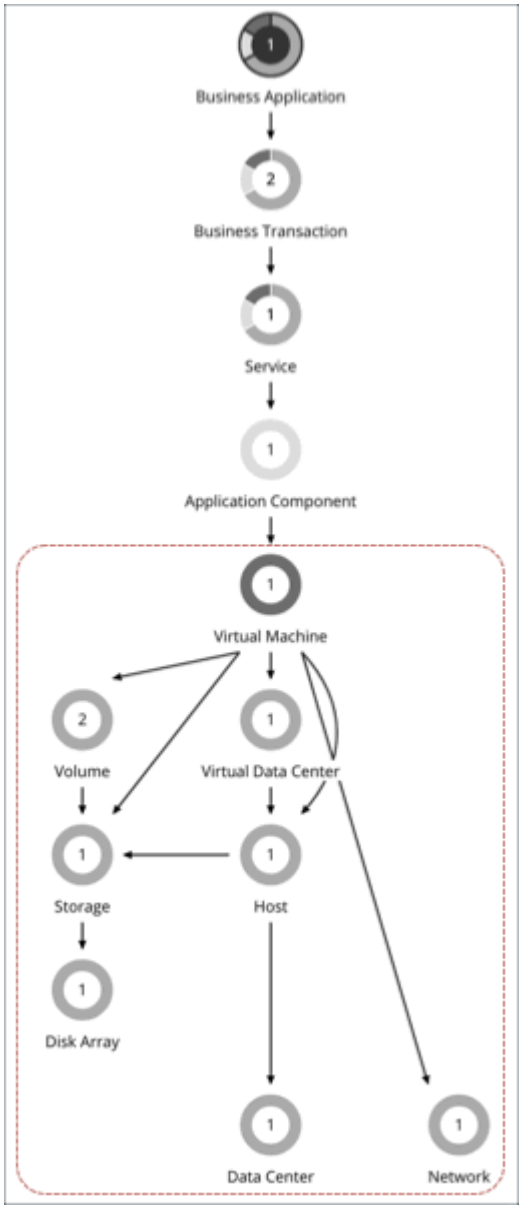
Actions

None

Workload Optimization Manager does not recommend actions for a cloud region.

Entity Types - On-prem Infrastructure

Workload Optimization Manager 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.

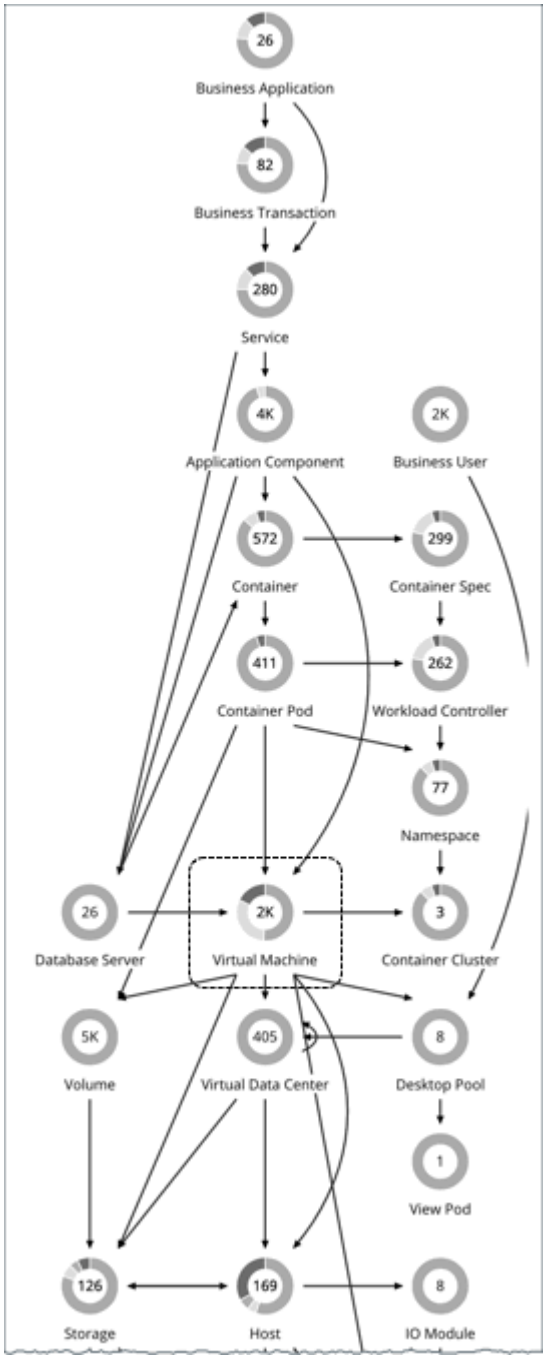


Virtual Machine (On-prem)

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:
Kubernetes nodes are represented as Virtual Machines in the Workload Optimization Manager supply chain. For details about nodes, see Virtual Machine (Kubernetes Node) *(on page 180)*.

Synopsis



Synopsis	
Budget:	A VM gains its budget by selling resources to the applications it hosts. If utilization is high enough, Workload Optimization Manager can allocate more resources to the VM, provision another instance, or move the VM to a host that has more resources. If utilization falls off, the VM loses budget. On the public cloud, if the budget isn't enough to pay for the host services, Workload Optimization Manager can post an action to suspend the VM.

Synopsis	
Provides:	Resources for hosted applications to use: ▪ VMEM ▪ VCPU ▪ VStorage ▪ IOPS (storage access operations per second) ▪ Latency (capacity for disk latency in ms) ▪ Memory and CPU Requests (for Kubernetes environments)
Consumes:	▪ Physical host resources, including CPU and Mem ▪ Storage
Discovered through:	Hypervisor targets

Monitored Resources

Workload Optimization Manager monitors the following resources for a VM:

- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.
- Virtual CPU

  Virtual CPU is the measurement of CPU utilized by the entity.
- Virtual Storage (VStorage)

  Virtual Storage is the measurement storage utilized by the entity.
- Storage Access

  Storage Access is the measurement of IOPS utilized by the entity.
- Latency

  Latency is the measurement of storage latency utilized by the entity.

Actions

- **Resize**
  - 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, Workload Optimization Manager verifies that the VM's cluster can adequately support the new size. If the cluster is highly utilized, Workload Optimization Manager will recommend a move action, taking into consideration the capacity of the new cluster and compliance with existing placement policies.

    For hypervisor targets, Workload Optimization Manager can resize vCPU by changing the VM's socket or cores per socket count. For details, see VCPU Scaling Controls *(on page 278)*.
  - 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.

- **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, Workload Optimization Manager can reconfigure VMs that violate vCPU scaling policies. For details, see VCPU Scaling Controls *(on page 278)*.

Tuned Scaling for On-prem VMs

For resizing VMs, Workload Optimization Manager includes tuned scaling action settings. These settings give 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 Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager will not automate memory resize for the VM.

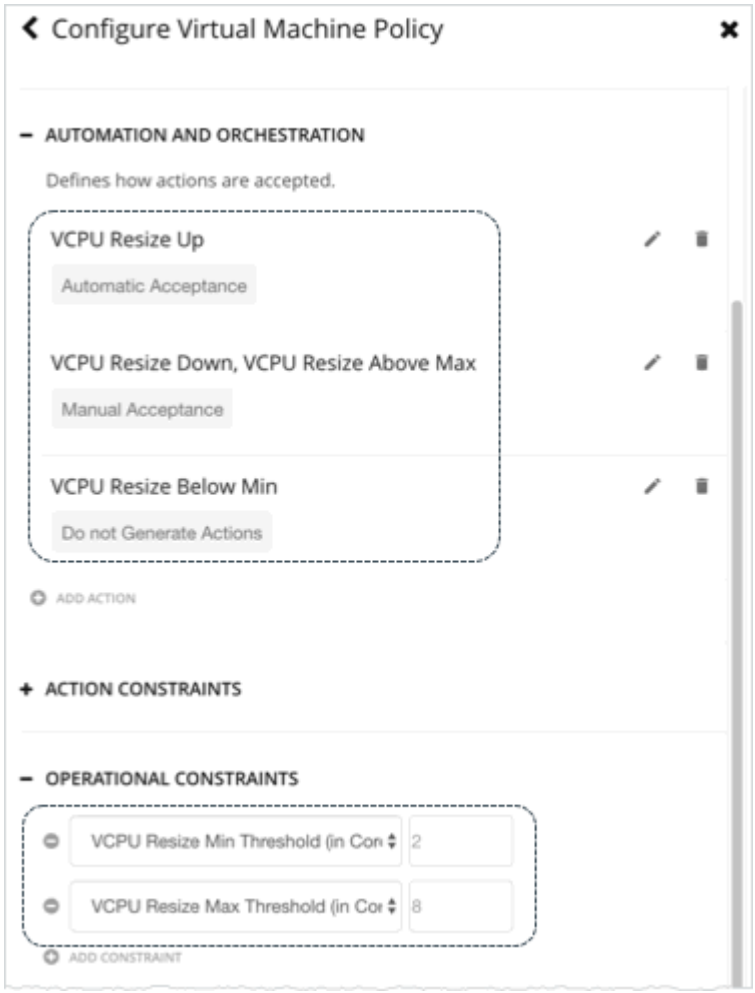
To configure tuned scaling:

1. Create a VM policy.
2. Under **Action Automation**, configure the action 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, Workload Optimization Manager handles resize actions as follows.

Current	Resize Request	Action 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, Workload Optimization Manager 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 resize, which is above the VCPU Resize Max threshold of 8, Workload Optimization Manager posts the VCPU Resize Up action (as a pending action) and pro-	10 VCPUs (if you executed the pending action)

Current	Resize Request	Action Mode	Result
		vides 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, Workload Optimization Manager 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, Workload Optimization Manager 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 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 Policy Scope *(on page 104)*.

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 105)*.

On-prem VM Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about on-prem VM actions, see On-prem VM Actions *(on page 269)*.

Action	Default Mode	vCenter	Hyper-V
Move	Manual	Auto	Auto
Reconfigure	Recommend	Rcmd	Rcmd
Start	Manual	Auto	Auto
Storage Move	Recommend	Auto	With VMM: Auto Otherwise: Rcmd
Provision (Kubernetes nodes only)	Manual	Rcmd	Rcmd
Suspend (Kubernetes nodes only)	Manual	Auto	Auto
vCPU Resize Up*	Manual	Auto	Auto
vCPU Resize Above Max*	Recommend	Auto	Auto
vCPU Resize Down*	Manual	Auto	Auto
vCPU Resize Below Min*	Recommend	Auto	Auto
vMem Resize Up*	Manual	Auto	Auto
vMem Resize Above Max*	Recommend	Auto	Auto
vMem Resize Down*	Manual	Auto	Auto
vMem Resize Below Min*	Recommend	Auto	Auto

** Workload Optimization Manager uses these settings, in conjunction with resize operational constraints (on page 275), to set up tuned scaling (on page 270) for on-prem VMs.*

You can use Action Scripts *(on page 112)* and third-party orchestrators (such as ServiceNow) for action orchestration.

Non-disruptive Mode

VM actions include the modifier, **Enforce Non Disruptive Mode**. When you enable this modifier, Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager will not execute the resize action.

Hypervisor VMEM for Resize

For on-prem environments, Workload Optimization Manager discovers VMEM utilization and can recommend actions to resize the VMEM capacity on a VM. For environments that do not include Applications and Databases as targets, the data that analysis uses to make these recommendations comes from the underlying hypervisors. However, that data is not always sufficient to result in accurate resize recommendations. Use the **Use Hypervisor VMEM for Resize** setting to determine how to generate VMEM recommendations.

Attribute	Default Setting
Hypervisor VMEM for Resize	On

- **On**

  When your environment includes Applications and Databases as targets, Workload Optimization Manager uses the VMEM metrics those targets discover. If a scope of VMs does not fall under those targets, then analysis *will* generate VMEM resize actions for that scope. In this case, analysis uses the VMEM metrics it discovers from the underlying hypervisors.

- **Off**

  When your environment includes Applications and Databases as targets, Workload Optimization Manager uses the VMEM metrics those targets discover. If a scope of VMs does not fall under those targets, then analysis *will not* generate VMEM resize actions for that scope.

Shared Nothing Migration

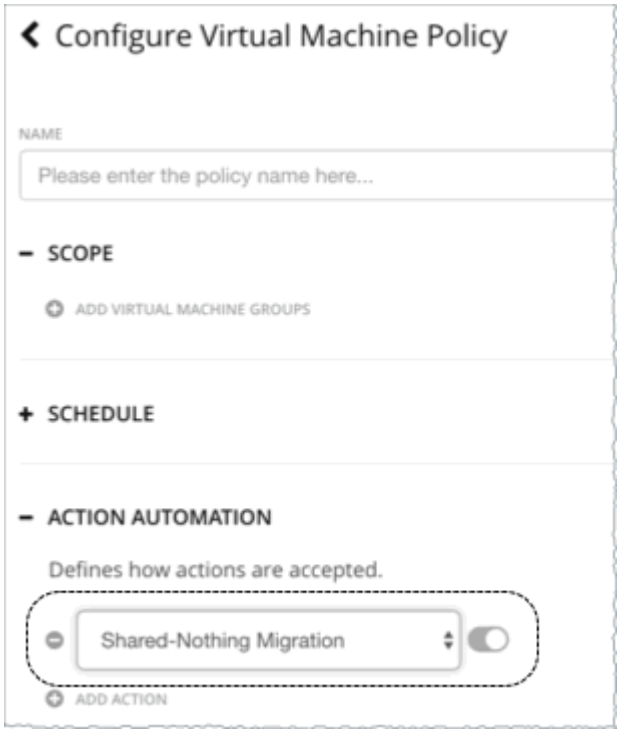
If you have enabled both storage and VM moves, Workload Optimization Manager can perform shared-nothing migrations, which move the VM and the stored VM files simultaneously. For example, assume a VM on a host also uses local storage on that host. In that case, Workload Optimization Manager can move that 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. Workload Optimization Manager recommends enabling it only on VMs that need it. To do this, you must first set the action 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:**
> Workload Optimization Manager does not simulate shared-nothing migrations in plans.

Resize Thresholds (Operational Constraints)

Workload Optimization Manager 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 Workload Optimization Manager to take more conservative actions when resizes are outside the range.

For details about tuned scaling, see Tuned Scaling for On-prem VMs *(on page 270)*.

Attribute	Default Value
VCPU Resize Max Threshold (in Cores)	64 Tuned Scaling Range Upper Limit
VCPU Resize Min Threshold (in Cores)	1 Tuned Scaling Range Lower Limit
VMEM Resize Max Threshold (MB)	131072 Tuned Scaling Range Upper Limit
VMEM Resize Min Threshold (MB)	512 Tuned Scaling Range Lower Limit

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 [GB]	None If you enable resize, Workload Optimization Manager uses the default value of 1024. You can change this to a different value.

vCPU Scaling Controls

For details, see VCPU Scaling Controls *(on page 278)*.

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 [MB]	1024
Increment constant for VStorage [GB]	1024

NOTE:
vCPU resize increments are configured in conjunction with vCPU scaling controls. For details, see VCPU Scaling Controls *(on page 278)*.

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 Workload Optimization Manager would allocate insufficient VMem for the machine to

operate. For a VM that is under utilized, Workload Optimization Manager 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 Workload Optimization Manager cannot reduce the VMem to less than 1024 MB.

Rate of Resize

When resizing resources for a VM, Workload Optimization Manager calculates the optimal values for VMem, VCPU and VStorage. But it does not necessarily make a change to that value in one action. Workload Optimization Manager uses the Rate of Resize setting to determine how to make the change in a single action.

Attribute	Default Value
Rate of Resize	Medium (2)

- **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, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager will raise VMem to 8 GB (or as close to that as the increment constant will allow).

Consistent Resizing

Attribute	Default Setting
Enable Consistent Resizing	Off

For groups in scoped policies:

When you create a policy for a group of VMs and turn on Consistent Resizing, Workload Optimization Manager 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.

> **NOTE:**
> For consistent resizing in on-prem environments, 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, Workload Optimization Manager 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.

Reasons to employ Consistent Resizing for a group include:

- Load Balancing

  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.

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.

Ignore NVMe Constraints

Workload Optimization Manager recognizes when a VM instance includes an NVMe driver. To respect NVMe constraints, it will not recommend a move or resize to an instance type that does not also include an NVMe driver. If you ignore NVMe constraints, then Workload Optimization Manager is free to resize or move the instance to a type that does not include an NVMe driver.

Attribute	Default Setting
Ignore NVMe Constraints	Off

Placement Policies

Workload Optimization Manager 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. Workload Optimization Manager can then recommend consolidating VMs or reconfiguring hosts in response to changing demand for licenses.

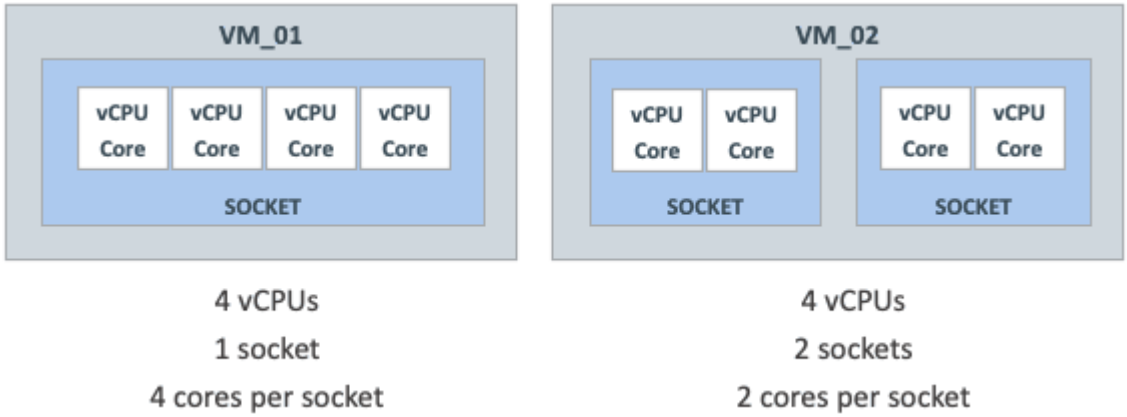
For more information, see Creating Placement Policies *(on page 88)*.

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 88)*.

vCPU Scaling Controls

Workload Optimization Manager 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.



Workload Optimization Manager 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 272)* 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

Workload Optimization Manager 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	Workload Optimization Manager decides	Reconfigured to 1 core per socket	▪ Non-disruptive if hot-add is enabled and VM sockets are increasing ▪ Disruptive if VM	Disruptive

vCPU Scaling Option	Unit	Sockets	Cores Per Socket	Resize Action	Reconfigure Action
				cores per socket does not equal 1, even if 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	Workload Optimization Manager decides	Preserve VM cores per socket	Non-disruptive if hot-add is enabled and VM sockets are increasing	Not generated
Change sockets	1 socket	Workload Optimization Manager decides	User-specified VM cores per socket	▪ 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	Workload Optimization Manager decides	Disruptive	Not generated
Change cores per socket	1 core per socket	Match host sockets	Workload Optimization Manager decides	Disruptive	▪ Non-disruptive if hot-add is enabled and VM sockets are increasing ▪ Disruptive if

vCPU Scaling Option	Unit	Sockets	Cores Per Socket	Resize Action	Reconfigure Action
					cores per socket is changed
Change cores per socket	1 core per socket	User-specified VM sockets	Workload Optimization Manager decides	Disruptive	▪ 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	Workload Optimization Manager decides	▪ 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 274)* 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 Workload Optimization Manager knowledge of the application domain and Guest OS to scope them out of policies.

Scaling Option: Change Virtual CPUs

In this scaling option, Workload Optimization Manager 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).

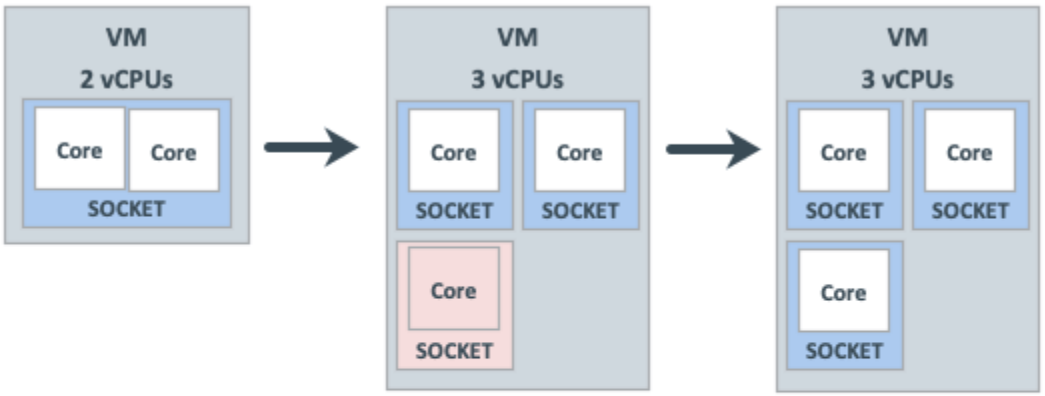
- If a VM requires a change to compute resources, Workload Optimization Manager generates a resize vCPU action that assumes 1 core per socket. If the VM currently does not have 1 core per socket, Workload Optimization Manager reconfigures it to 1 core per socket as part of action execution.
- If a VM is already optimally sized, but its current cores per socket is not 1, Workload Optimization Manager 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 Workload Optimization Manager 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.

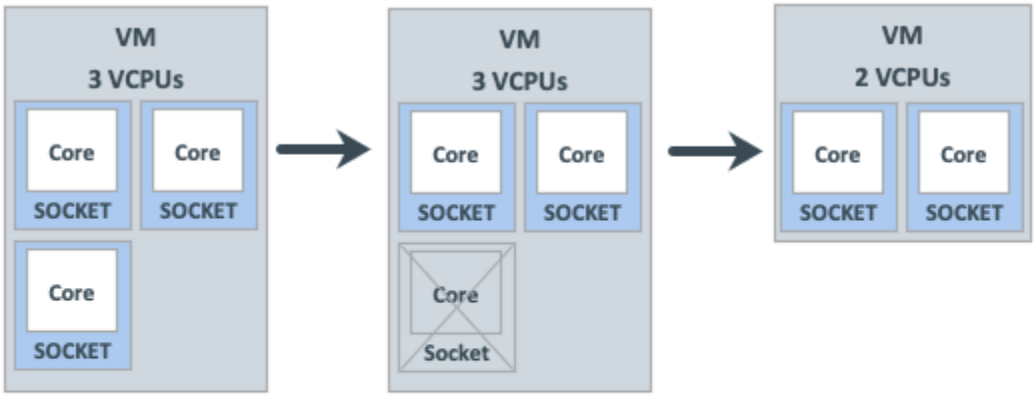


Current:
2 vCPUs
1 socket
2 cores per socket

Recommended Action:
Resize to 3 vCPUs
by changing to:
- 3 sockets
- 1 core per socket

After Action Execution:
3 vCPUs
3 sockets
1 core per socket

- 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.



Current:
3 vCPUs
3 sockets
1 core per socket

Recommended Action:
Resize to 2 vCPUs
by changing to
2 sockets

After Action Execution:
2 VCPUs
2 sockets
1 core per socket

Scaling Option: Change Sockets

In this scaling option, Workload Optimization Manager adds or removes compute resources by changing VM sockets.

- If a VM requires a change to compute resources, Workload Optimization Manager 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), Workload Optimization Manager 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), Workload Optimization Manager 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

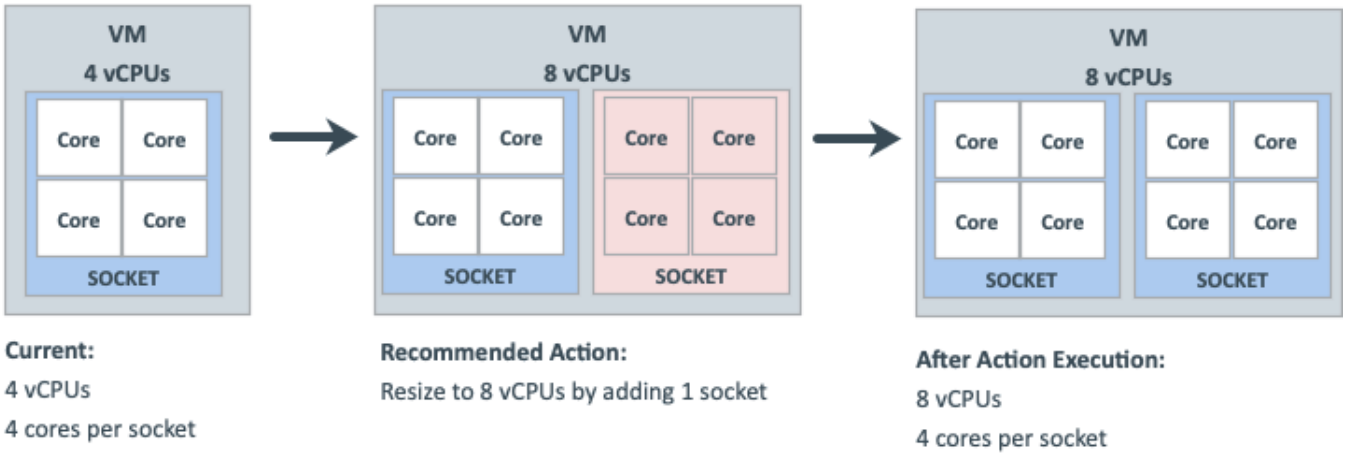
In this scaling option, Workload Optimization Manager adds or removes compute resources by changing VM sockets in increments of 1, and preserves VM cores per socket.

This scaling option is ideal under the following scenarios:

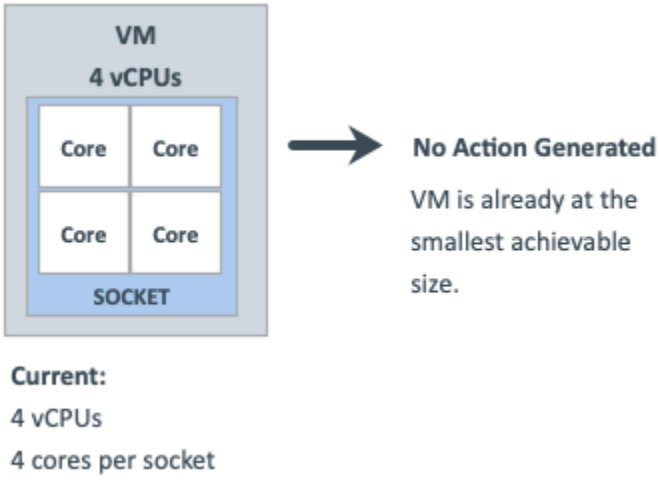
- You require Workload Optimization Manager to leave the VM cores per socket configuration unchanged for operational policy reasons (such as compliance with an application support contract policy).
- You have VMs that need to upsize non-disruptively to meet rising application demand.
- You have VMs with 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. Workload Optimization Manager 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.



Change Sockets and Specify Cores Per Socket

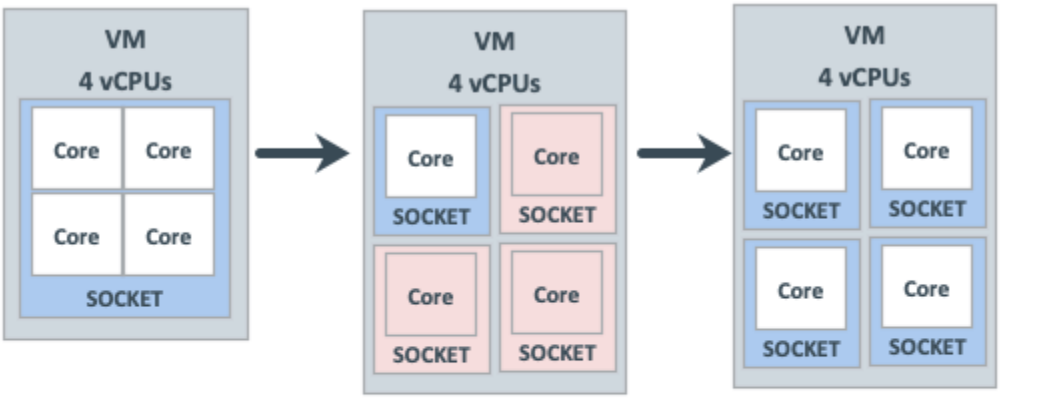
In this scaling option, Workload Optimization Manager adds or removes compute resources by changing VM sockets in increments of 1, and reconfigures VM cores per socket according to the value that you specify.

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 Workload Optimization Manager 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. Workload Optimization Manager has determined that the VM is already optimally sized, so a resize action is not necessary.

- Since the VM is in violation of policy, Workload Optimization Manager changes sockets from 1 to 4, and cores per socket from 4 to 1.

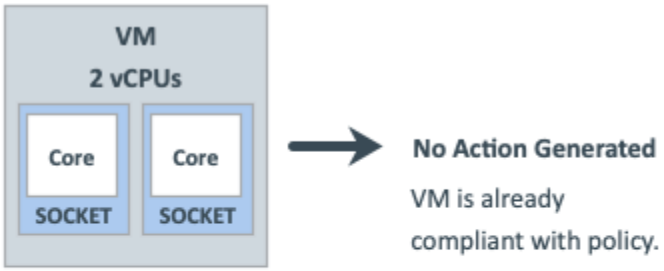


Current:
4 vCPUs
1 socket
4 cores per socket (non-compliant)

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.



No Action Generated
VM is already compliant with policy.

Current:
2 vCPUs
2 sockets
1 core per socket (compliant)

Scaling Option: Change Cores Per Socket

In this scaling option, Workload Optimization Manager adds or removes compute resources by changing the VM cores per socket.

- If a VM requires a change to compute resources, Workload Optimization Manager 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), Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager adds or removes compute resources by changing the VM cores per socket in increments of 1, and preserves VM sockets.

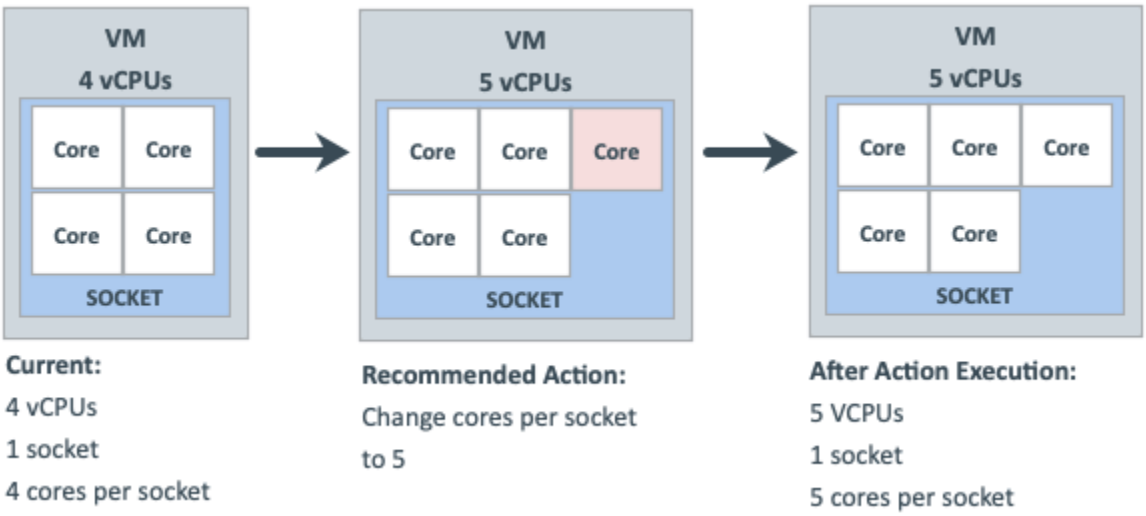
This scaling option is ideal under the following scenarios:

- You require Workload Optimization Manager 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.

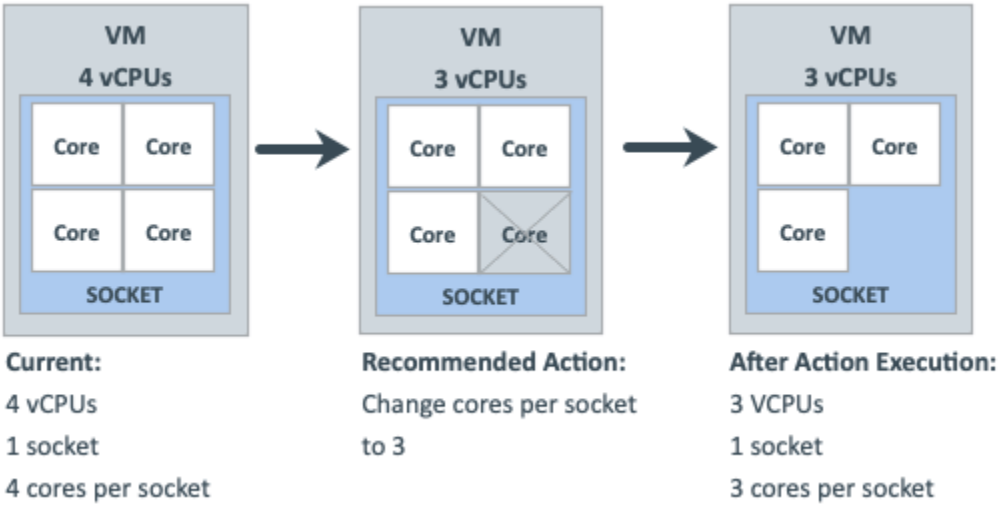
> **NOTE:**
> You can also use the 'Match Host Sockets' scaling option (discussed below) for NUMA sensitive VMs.

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. Workload Optimization Manager 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

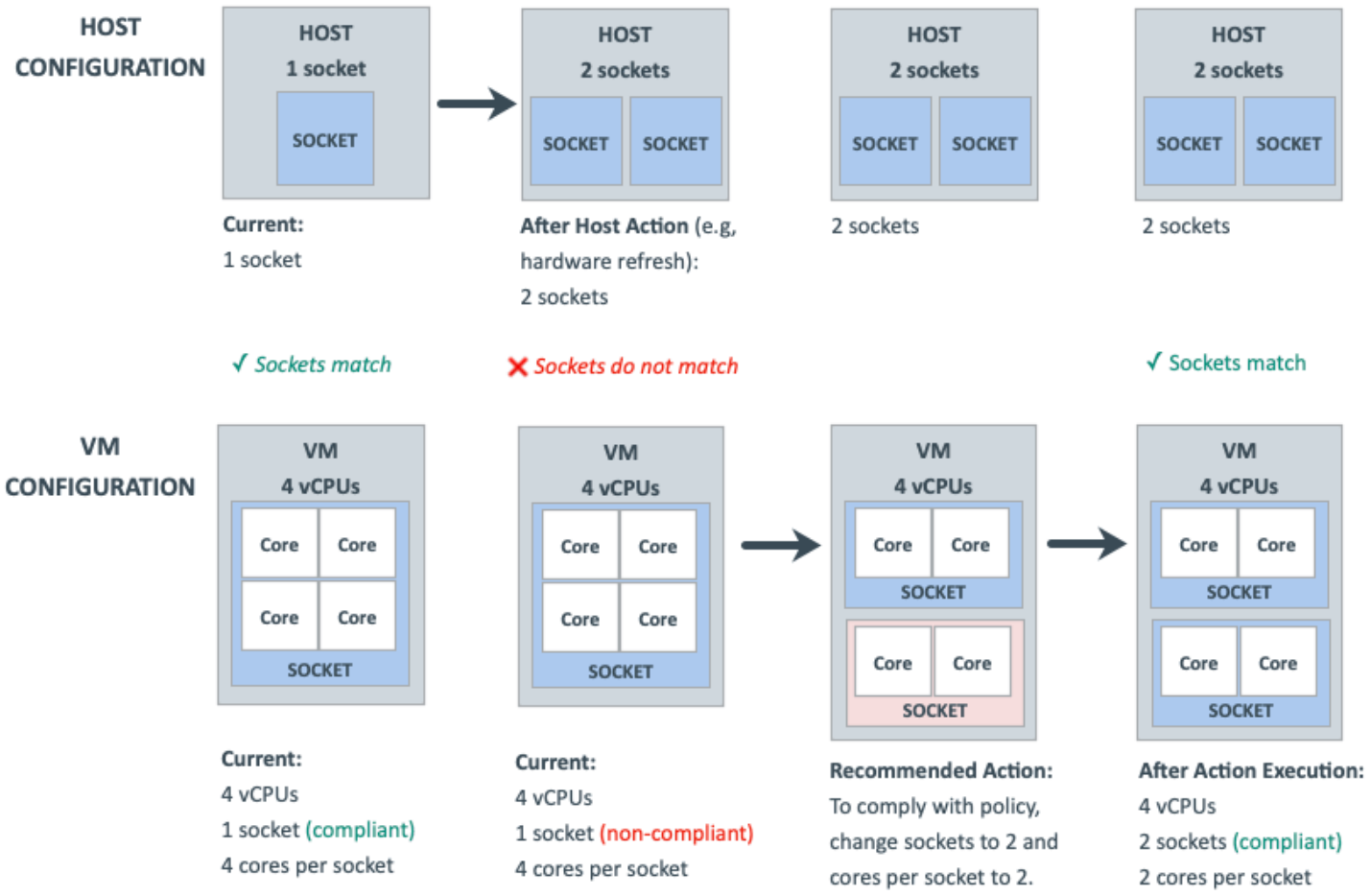
In this scaling option, Workload Optimization Manager reconfigures VM sockets to match the number of host sockets, thereby balancing vCPUs evenly across physical sockets. It also changes VM cores per socket to maintain the same compute capacity (vCPU).

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. Workload Optimization Manager 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. Workload Optimization Manager 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), Workload Optimization Manager must distribute 4 cores between 2 sockets. The end result is 2 sockets and 2 cores per socket.



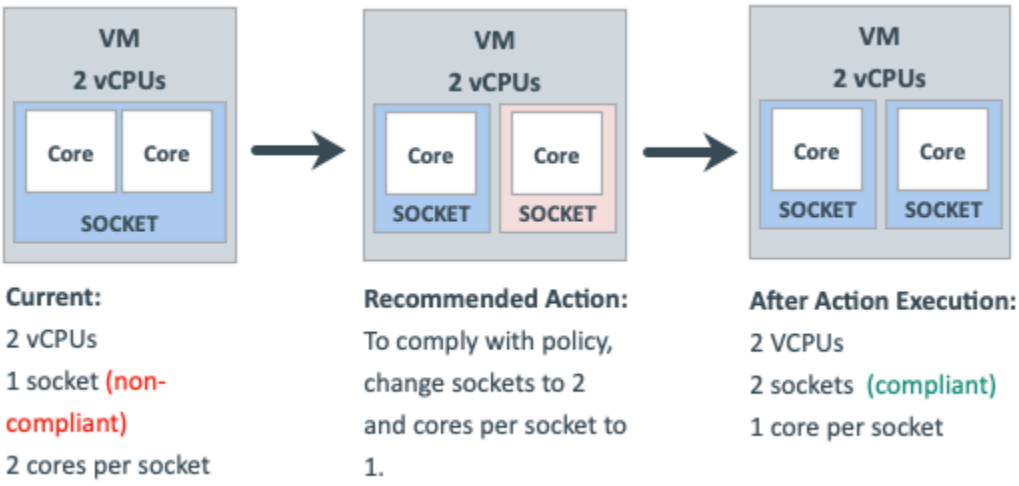
Change Cores Per Socket and Specify Sockets

In this scaling option, Workload Optimization Manager reconfigures VM sockets according to the value that you specify, and changes VM cores per socket to maintain the same compute capacity (vCPU).

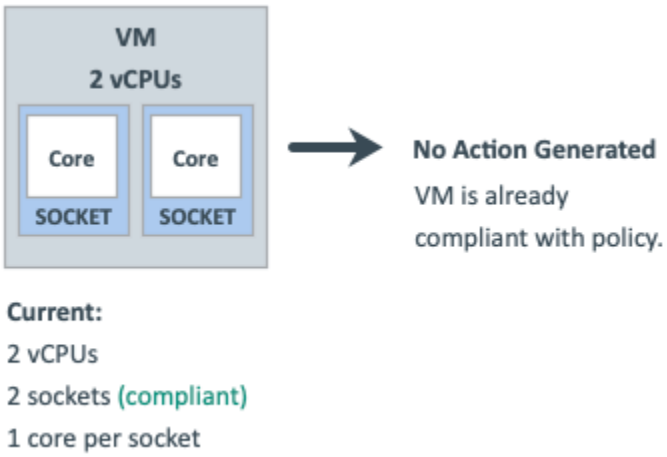
This scaling option is ideal if you have VMs that require a specific socket value for operational policy reasons (such as socket-based licensing or compliance with an application support contract policy).

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. Workload Optimization Manager has determined that the VM is already optimally sized, so a resize action is not necessary.

- Since the VM is in violation of policy, Workload Optimization Manager 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, Workload Optimization Manager adds or removes compute resources in increments of MHz (1800 MHz by default).

If a VM requires a change to compute resources, Workload Optimization Manager generates a resize vCPU action that assumes 1 core per socket, regardless of the VM's actual cores per socket.

If Workload Optimization Manager 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. Workload Optimization Manager 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**, Workload Optimization Manager 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:** Workload Optimization Manager assumes one core per socket and only changes sockets.	Not supported by Workload Optimization Manager	Not supported by Workload Optimization Manager
Change cores per socket – Match host sockets				
Change cores per socket – User specified sockets				

Tie Breakers

When a single VM applies multiple conflicting policies, Workload Optimization Manager 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 3.3.7) 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 identity 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.

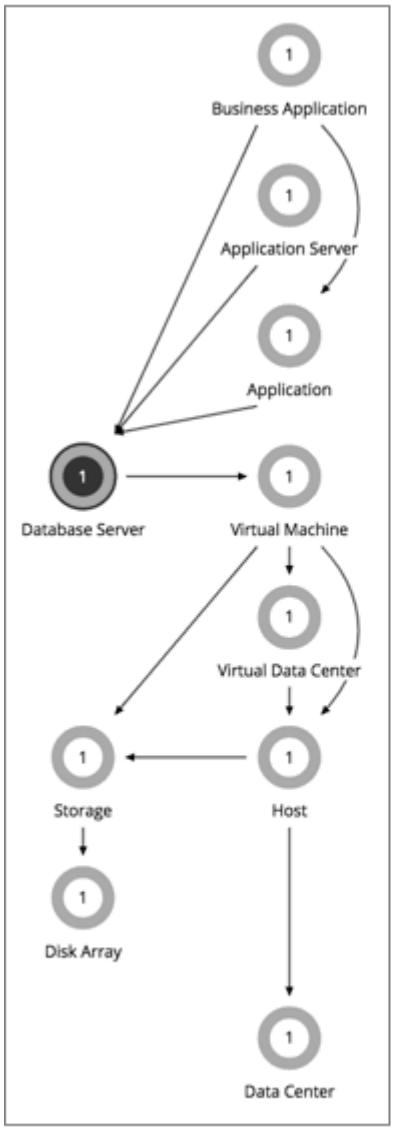
  VMs in the group will be reconfigured to 2 sockets if required, and resized by changing the cores per socket count while keeping the sockets fixed at 2, thus ensuring compliance with socket-based licensing. Resize actions are disruptive because cores per socket will change. Reconfigure actions are non-disruptive 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

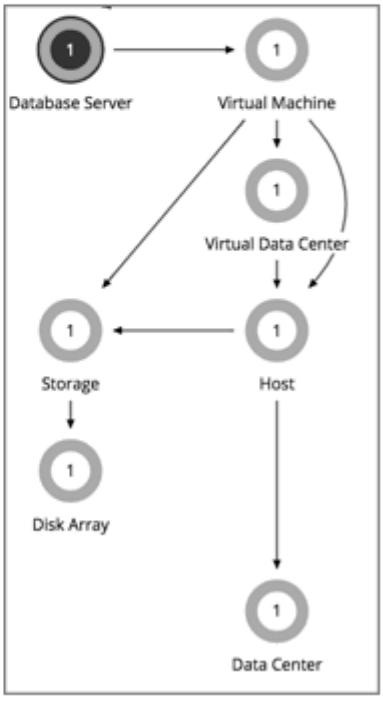
Database Server (On-prem)

For on-prem, a Database Server is a database discovered through one of the associated database application targets or through APM solutions.

Synopsis



AppDynamics



Database Server Target

Synopsis	
Budget:	On-prem Database Servers have unlimited budget.
Provides:	▪ 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:	▪ AppDynamics targets ▪ Database Server targets

Synopsis	
	▪ Dynatrace MySQL and MSSQL processes ▪ NewRelic Infrastructure Integration (NRI): MySql, MsSql, MongoDB, OracleDB

Monitored Resources

Workload Optimization Manager monitors the following resources for an on-prem Database Server:

- Virtual Memory

  Virtual Memory is the measurement of memory utilized by the entity.
- Transactions

  Transaction is a value that represents the per-second utilization of the transactions allocated to a given entity.
- Database memory

  Database memory (or DBMem) is the measurement of memory utilized by a Database Server.
- Connections

  Connection is the measurement of Database Server 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

Resize

Resize the following resources:

- Connections

  Workload Optimization Manager 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. Workload Optimization Manager 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, Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

Action	Default Mode
Resize	Manual
Resize DBMem (Up/ Down)	Manual

You can use Action Scripts *(on page 112)* and third-party orchestrators (such as ServiceNow) for action orchestration.

Transaction SLO

Transaction SLO determines the upper limit for acceptable transactions per second. When the number of transactions reaches the given value, Workload Optimization Manager sets the risk index to 100%.

Attribute	Default Setting/Value
Enable Transaction SLO	Off
Transaction SLO	None If you enable SLO, Workload Optimization Manager 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, Workload Optimization Manager sets the risk index to 100%.

Attribute	Default Setting/Value
Enable Response Time SLO	Off Workload Optimization Manager estimates SLO based on monitored values.
Response Time SLO [ms]	None If you enable SLO, Workload Optimization Manager uses the default value of 2000. You can change this to a different value.

DBMem Utilization

The utilization that you set here specifies the percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity.

Attribute	Default Value
DBMem Utilization (%)	100

For example, a value of 80 means that Workload Optimization Manager considers 80% utilization to be 100% of capacity. Workload Optimization Manager recommends actions that avoid utilization beyond the given value.

DBMem Scaling Increment

This increment specifies how many units to add or subtract when scaling DBMem.

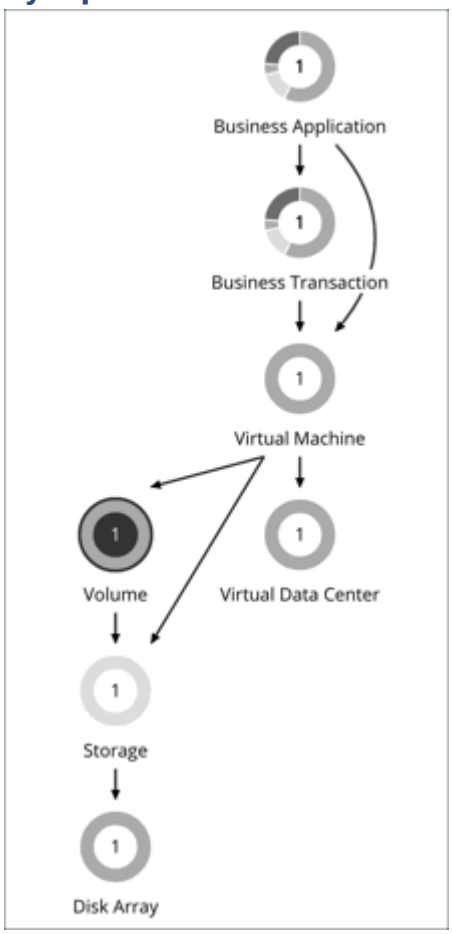
Attribute	Default Value
DBMem Scaling Increment (MB)	128

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, Workload Optimization Manager will not leave a database server with less than the increment value. For example, if you use the default 128, then Workload Optimization Manager cannot reduce DBMem to less than 128 MB.

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	
Budget:	An on-prem volume gains its budget by selling resources to the VMs that it serves.
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

- **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 Workload Optimization Manager to use relevant metrics when analyzing and recommending actions for volumes. For details, see Enable Analysis of On-prem Volumes *(on page 95)*.
  - Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

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 88)* and Enable Analysis of On-prem Volumes *(on page 95)*

Action Automation and Orchestration

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. Workload Optimization Manager 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 Datacenter (Private Cloud)

A virtual datacenter (vDC) is a collection or pool of resources that groups the resources around specific requirements or business needs. In private cloud environments, Workload Optimization Manager 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 Datacenters.

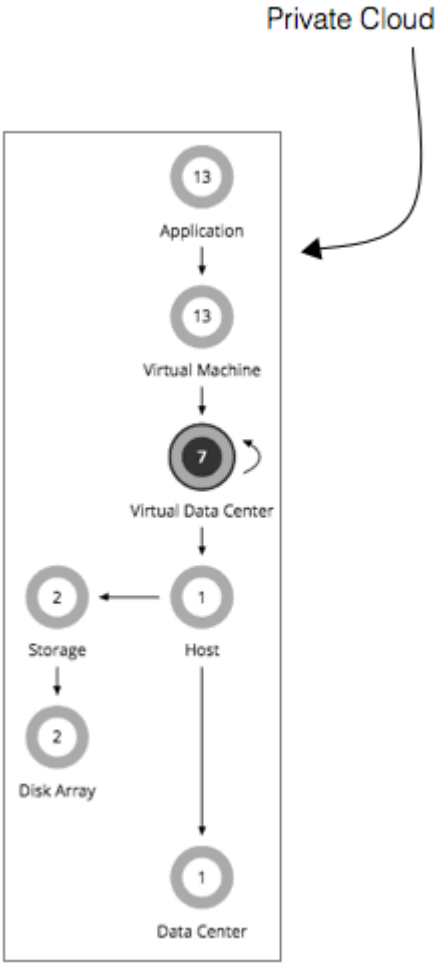
NOTE:
Different targets use different names to refer to Virtual Datacenters. In the Workload Optimization Manager supply chain, these entities are all represented by Consumer and Provider VDCs, as follows:

Workload Optimization Manager	vCenter Server	VMM
Consumer VDC	Resource Pool (Child)	Tenant or TenantQuota
Provider VDC	Resource Pool (Root)	Cloud

Provider Virtual Datacenters

A provider virtual datacenter (vDC) is a collection of physical resources (hosts and datastores) within a cloud stack. The cloud administrator has access to these resources, and defines the datacenter members. A Provider vDC is created to manage resources that will be allocated to external customers through one or more Consumer vDCs.

Synopsis



Synopsis	
Budget:	A Provider vDC gains its budget by selling resources to the Consumer vDCs that it hosts. If utilization falls off, the datacenter loses budget. Ultimately, if the budget isn't enough to pay for the services it consumes, Workload Optimization Manager will recommend decommissioning the Provider vDC.
Provides:	Physical resources such as hosts and datastores to Consumer vDCs.
Consumes:	Hosts and datastores from the physical infrastructure
Discovered through:	Workload Optimization Manager discovers vDCs through private cloud stack managers.

Monitored Resources

Workload Optimization Manager monitors the following resources for a Provider vDC:

- Memory (Mem)

  The utilization of the Datacenter's memory reserved or in use
- CPU

The utilization of the Datacenter's CPU reserved or in use

- Storage

  The utilization of the storage attached to the Provider vDC.

Actions

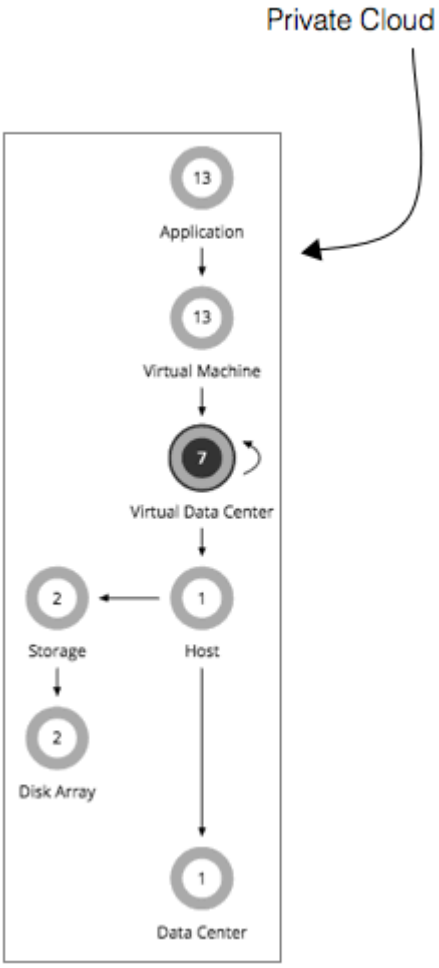
None

Workload Optimization Manager does not recommend actions for a Virtual Datacenter. Instead, it recommends actions for the entities that provide resources to the Virtual Datacenter.

Consumer Virtual Datacenters

A Consumer Virtual Datacenter (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 Datacenters use the resources supplied by a Provider Datacenter.

Synopsis



Synopsis	
Budget:	A Consumer vDC gains its budget as a function of its activity. The higher the utilization of the vDC, the more Workload Optimization Manager assumes the vDC is selling its services to a user. If utilization is high enough on a Consumer vDC, Workload Optimization Manager can increase resources for the vDC. If utilization falls off, Workload Optimization Manager can reduce resource capacity, or ultimately recommend terminating the vDC. Workload Optimization Manager can also resize VMs through the Consumer vDC in response to changes in VM utilization.
Provides:	Resources to host virtual systems.
Consumes:	Provider vDC
Discovered through:	Workload Optimization Manager discovers vDCs 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, Workload Optimization Manager 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

Workload Optimization Manager monitors the following resources for a Consumer vDC:

- Memory (Mem)

  The utilization of the Datacenter's memory reserved or in use
- CPU

  The utilization of the Datacenter's CPU reserved or in use
- Storage

  The utilization of the storage attached to the Consumer vDC.

Actions

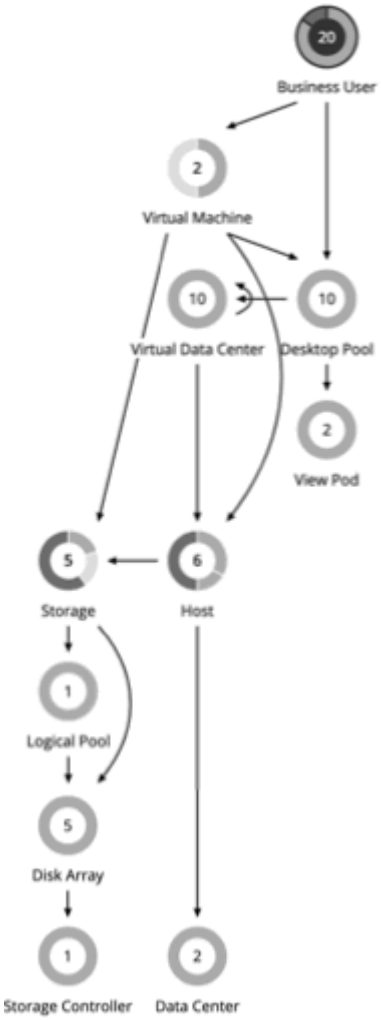
Workload Optimization Manager does not recommend actions to perform on a Consumer vDC. Instead, it recommends actions to perform on the entities running in the Provider vDC.

Business User

For Virtual Desktop Infrastructure (VDI) environments, a Business User is a user account that is entitled to launch one or more active VDI sessions. As it discovers desktop pools, Workload Optimization Manager creates Business User entities for each user that is entitled to a pool. One business user can be entitled to more than one desktop pool.

To properly work with Business User entities, Workload Optimization Manager discovers user information through the LDAP server that manages users for the VDI environment. Note that the account Workload Optimization Manager uses to connect to the LDAP server must be trusted for the same domains as are the users in your environment.

Synopsis



The Supply Chain shows relationships of Business Users to Desktop Pools and also to VMs. One Business User can have access to multiple Desktop Pools. When a Business User has an active session, the Supply Chain shows a direct link between the user and the VM that hosts the session. However, Workload Optimization Manager does not consider this direct connection when analyzing compute resources. Instead, Business Users utilize Desktop Pool resources, and the Desktop Pools use compute resources from the underlying Virtual Datacenters.

Synopsis	
Budget:	A Business User has unlimited budget.
Provides:	N/A

Synopsis	
Consumes:	Resources from the underlying desktop pools: ▪ Sessions ▪ Pool Memory ▪ Pool Storage ▪ Pool CPU When a Business User has an active session, the Supply Chain shows it in relation to the VM that hosts the session. The Business User consumes the VM's compute resources to support the session requirements for ImageCPU, ImageMem, and ImageStore resources.
Discovered through:	The LDAP server that manages these users. You can specify the LDAP server as part of the target configuration, or Workload Optimization Manager can discover it in association with the VDI target.

Monitored Resources

Workload Optimization Manager monitors the following resources for a Business User:

- ImageCPU

  CPU utilization, as a percentage of CPU capacity for the user's desktop image or images.
- ImageMem

  Memory utilization, as a percentage of Memory capacity for the user's desktop image or images.
- ImageStorage

  Storage utilization, as a percentage of storage capacity for the user's desktop image or images.

Business User Actions

Move

Move a Business User between desktop pools to address:

- Resource congestion on the image

  When utilization is consistently near capacity for image resources, Workload Optimization Manager can recommend moving a Business User to a desktop pool that serves larger images.
- Resource congestion on the desktop pool

  When utilization is consistently near capacity for the desktop pool, Workload Optimization Manager can recommend moving a Business User to a desktop pool that has more available resources.

NOTE:
To support moves, you must configure placement policies that merge *similarly configured* desktop pools. For details, see Desktop Pool Placement Policies *(on page 308)*.

Business User Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about Business User actions, see Business User Actions *(on page 303)*.

Action	Default Mode
Move	Recommend

Image Target Utilization

Workload Optimization Manager tracks utilization of desktop image resources for the Business Users in your Virtual Desktop Infrastructure (VDI) environment.

Attribute	Default Value
Image CPU Target Utilization	70 The target utilization as a percentage of CPU capacity.
Image MEM Target Utilization	90 The target utilization as a percentage of memory capacity.
Image Storage Target Utilization	90 The target utilization as a percentage of storage capacity.

Aggressiveness and Observation Period

Workload Optimization Manager uses these settings to calculate utilization percentiles. 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 utilization of compute and storage resources, Workload Optimization Manager considers a given utilization percentile. For example, assume a 95th percentile. The maximum utilization would be the highest value that 95% of the observed samples fall below.

  Using a percentile, Workload Optimization Manager can recommend more relevant actions, so that analysis can better exploit elasticity in your environment. 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 – The least aggressive, recommended for critical workloads that need maximum guaranteed performance at all times.
  - 95th Percentile (Default) – The recommended setting to achieve maximum performance and savings.
  - 90th Percentile – The most aggressive, recommended for non-production workloads that can stand higher resource utilization.

- **Max Observation Period**

Attribute	Default Value
Max Observation Period	Last 7 Days

To refine the calculation of resource utilization, you can set the sample time to consider. Workload Optimization Manager 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.)

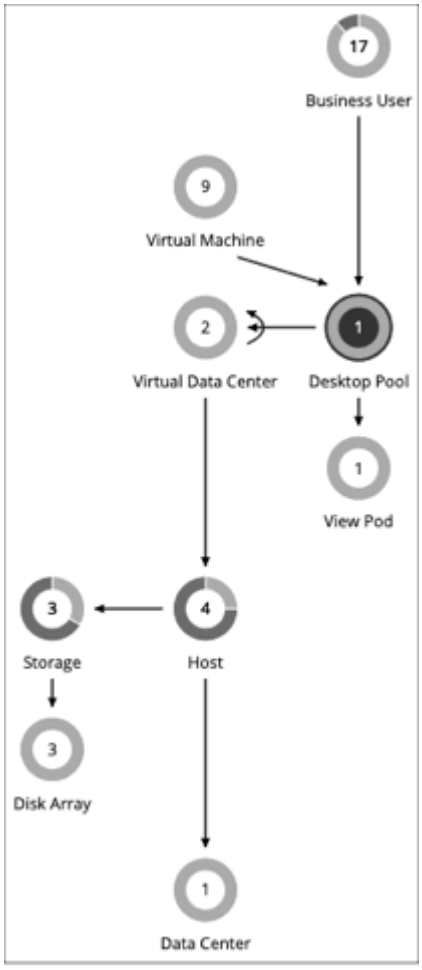
A shorter period means there are fewer data points to account for when Workload Optimization Manager calculates utilization percentiles. This results in more dynamic, elastic moves to different Desktop Pools, while a longer period results in more stable or less elastic moves. You can make the following settings:

- Less Elastic – Last 90 Days
- More Elastic – Last 30 Days
- (Default) Most Elastic – Last 7 Days

Desktop Pool

For Virtual Desktop Infrastructure (VDI) environments, a desktop pool is a collection of desktops that users can select from. The desktop pool can provide logical grouping of desktops according to user roles, assignment type (dedicated or floating), and the source of resources (physical host or VM).

Synopsis



The desktop pool gets compute and storage resources from the underlying Virtual Datacenter. For VMware Horizon View, the VDI architecture includes one or more vCenter Server instances. When it discovers the Horizon View target, Workload

Optimization Manager also discovers the supporting vCenter Server instances, and their corresponding Virtual Datacenters. These are the source of compute and storage resources for the associated desktop pools.

Synopsis	
Budget:	A Desktop Pool gets its budget by selling resources to Business Users.
Provides:	Resources for Business Users to use: ▪ PoolMEM ▪ PoolCPU ▪ Sessions
Consumes:	▪ Compute and storage resources from the associated Virtual Datacenters ▪ Sessions from the underlying View Pod
Discovered through:	The VDI management target. For VMware Horizon View, the target is the View Connection Server.

Monitored Resources

Workload Optimization Manager monitors the following resources for a Desktop Pool:

- Pool CPU

  The CPU available to the pool that is in use by active sessions.
- Pool Memory

  The memory available to the pool that is in use by active sessions.
- Pool Storage

  The storage capacity available to the pool that is in use by active sessions.
- Active Sessions

  How many active sessions are on the pool as a percentage of the pool's capacity as defined in the Workload Optimization Manager policy.
- Total Sessions

  How many active and disconnected (non-terminated) sessions are on the pool, as a percentage of the pool's capacity.

Actions

None

Workload Optimization Manager does not recommend actions for a desktop pool. It recommends actions for the Business Users running active sessions in the pool.

Desktop Pool Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

None

Workload Optimization Manager does not recommend actions for a desktop pool. It recommends actions for the Business Users running active sessions in the pool.

Observation Settings

Workload Optimization Manager uses these settings to decide whether to move Business Users from one desktop pool to another.

- **Daily Observation Windows**

Attribute	Default Value
Daily observation windows	3 windows per day

When evaluating utilization of pool resources, Workload Optimization Manager divides each day into different observation windows, calculates an average for each, and uses the highest value. In this way, Workload Optimization Manager can account for high-use periods in the day to base calculations off of the most representative usage of the desktop images.

Assume three observation windows:

Window	Time range	Average utilization
W1	00:00 – 08:00	10%
W2	08:00 – 16:00	80%
W3	16:00 – 24:00	40%

Average utilization for this day *without* the benefit of observation windows would be 44%. By using observation windows we can see that the representative utilization of pool resources is closer to 80%. That is because Workload Optimization Manager discovers an average utilization of 80% during the high-usage time of day.

When calculating whether to move business users from one desktop pool to another, Workload Optimization Manager averages the observation windows over the time you set for the Max Observation Period. For this reason, you should try to set up observation windows that capture the best representation of work habits amongst your business users.

- **Max Observation Period**

Attribute	Default Value
Max Observation Period	Last 7 Days

To refine the calculation of resource utilization, you can set the sample time to consider. Workload Optimization Manager uses historical data from up to the number of days that you specify as a sample period. (If the Workload Optimization Manager database has fewer days' data, then it uses all of its stored historical data.)

A shorter period means there are fewer data points to account for when Workload Optimization Manager calculates utilization. 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 30 Days
  - Recommended – Last 7 Days
  - More Elastic – Last 3 Days

Pool Utilization

These settings affect the actions Workload Optimization Manager recommends as it manages business users and active accounts on the desktop pool. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings.

Attribute	Default Value
Pool CPU Utilization	95
Pool Mem Utilization	95
Pool Storage Utilization	95

The values you set here specify what percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity. For example, setting 70 for Desktop Pool Pool CPU Utilization means that Workload Optimization Manager considers 70% utilization of that CPU to be 100% of capacity and 35% utilization to be 50% of capacity.

Placement Policies

Under some circumstances, you can have Business Users who need larger desktop images. This appears as users with high utilization of the image resources. In this case, Workload Optimization Manager can recommend moving the Business Users to a different desktop pool that serves up larger images.

To support moving Business Users, you must create a placement policy that merges desktop pools. Be sure to merge only desktop pools that are *similarly configured* – they should run the same operating system and applications, and differ only in allocated memory and/or CPU.

To merge desktop pools:

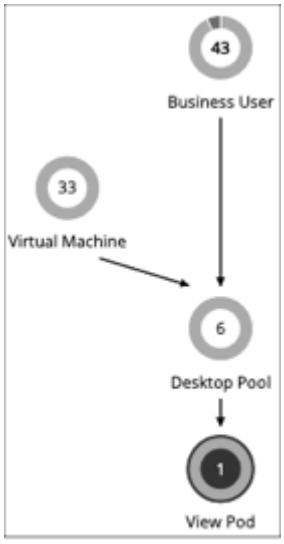
1. Create a new placement policy.
2. Choose **Merge** as the policy type.
3. For the consumer type to merge, choose **Desktop Pool**.
4. Choose the pools that you want to merge.
5. Save the policy.

For more information, see Creating Placement Policies *(on page 88)*.

View Pod

For Virtual Desktop Infrastructure (VDI) environments, a View Pod groups together a given set of Desktop Pools.

Synopsis



Synopsis	
Budget:	A View Pod has unlimited budget.
Provides:	Active Sessions.
Consumes:	N/A
Discovered through:	The VDI management target. For VMware Horizon View, the target is the View Connection Server.

Monitored Resources

Workload Optimization Manager monitors the following resources for a Desktop Pool:

- Active Sessions

  How many active sessions are on the pool as a percentage of the pool's capacity as defined in the Workload Optimization Manager policy.
- Total Sessions

  How many active and disconnected (non-terminated) sessions are on the pool, as a percentage of the pool's capacity.

Actions

None

Workload Optimization Manager does not recommend actions for a view pod. Instead, it recommends actions for the Business Users that are running active sessions.

Active Session Capacity for View Pods

Each View Pod entity has a set capacity of active sessions. By default, Workload Optimization Manager assumes a capacity of 8,000. So that Workload Optimization Manager can generate reliable actions for Business User entities, you must set this capacity to match the active session capacity that your Horizon administrator has deployed for the given view pod.

Once you know the correct active session capacity for your view pod, create an automation policy that sets the capacity. For complete information about creating automation policies, see Creating Scoped Automation Policies *(on page 97)*. For information about view pod policies, see View Pod Policies *(on page 310)*.

1. Create a new scoped automation policy.

   Navigate to the Settings Page and choose **Policies**. Then click **NEW AUTOMATION POLICY**, and select View Pod as the policy type. Be sure to name the new policy.

2. Set the policy scope to your view pod.

   To define its scope, you assign a group to the policy. You will have to create the group for this view pod:

   - Expand the **SCOPE** section and then click **ADD VIEW POD GROUPS**.
   - Choose the group that contains only the view pod you want to configure.

     If it has already been created, choose the group from the list. If the group does not appear, click **NEW GROUP** to create a static group that includes only the view pod you want to configure. For more information about creating groups, see Creating Groups *(on page 509)*.

     Choose the group you want and click **SELECT**. This returns you to the Configure View Pod Policy fly-out.

3. Set the view pod capacity.

   Expand the **UTILIZATION CONSTRAINTS** section and click **ADD UTILIZATION CONSTRAINT**. From the drop-down list, choose Active Sessions Capacity. In the capacity field, enter the capacity that you have calculated for your desktop pools.

4. Save your work

   When you're done, be sure to click **SAVE AND APPLY**.

View Pod Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

None

Workload Optimization Manager does not recommend actions for a view pod. Instead, it recommends actions for the Business Users that are running active sessions.

Active Sessions Capacity

This setting controls the number of active sessions a given view pod can support.

Attribute	Default Value
Active Sessions Capacity	8000

For each view pod, you should set this value to match the active session capacity that has been deployed in your VDI environment for the given view pod. For more information, see Active Session Capacity for View Pods *(on page 309)*.

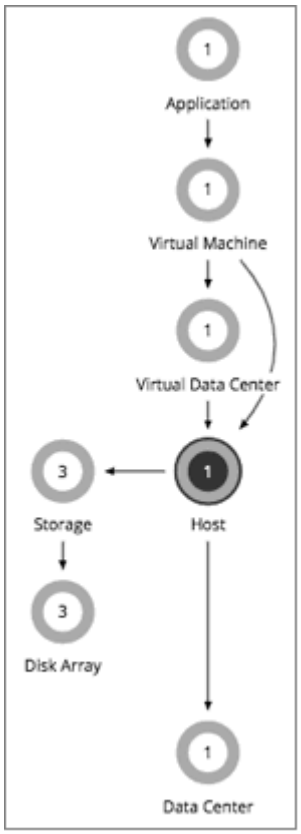
Host

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's most usual to use physical hardware as your hosts.

NOTE:
To support vSAN storage in your environment, you can deploy HCI Hosts. Workload Optimization Manager discovers the vSAN as a storage entity that consumes resources from the underlying hosts. For more information, see vSAN Storage *(on page 322)*.

On the public cloud a host is an availability zone. This is where your cloud workloads run. For details, see Zone *(on page 262)*.

Synopsis



Synopsis	
Budget:	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, Workload Optimization Manager can recommend that you provision a new one. If utilization falls off, the host loses budget. Ultimately, if the budget isn't enough to pay for the services it consumes, Workload Optimization Manager will recommend to suspend or power off the host.
Provides:	Host resources for VMs to use: ▪ Mem (Kbytes) ▪ CPU (MHz) ▪ IO (throughput on the I/O bus) ▪ Net (network throughput) ▪ Swap (swap rate capacity measured in bytes/sec)

Synopsis	
	▪ Ballooning (sharing of memory among hosted VMs) ▪ CPU Ready Queue (wait time on the queue in ms)
Consumes:	Datacenter resources (physical space, cooling, etc.) and storage.
Discovered through:	Workload Optimization Manager discovers hosts 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

Workload Optimization Manager monitors the following resources on a host:

- Memory (Mem)

  The utilization of the PM's memory reserved or in use
- CPU

  The utilization of the PM's CPU reserved or in use
- IO

  The utilization of the PM's IO adapters
- Net

  The utilization of data through the PM's network adapters
- Swap

  The utilization of the PM's swap space
- Balloon

  The utilization of shared memory among VMs running on the host. ESX-only
- CPU Ready

  The utilization of the PM's allocated ready queue capacity that is in use, for 1, 2, and 4 CPU ready queues. ESX-only

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. Workload Optimization Manager 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.
- **Reconfigure**

  Workload Optimization Manager generates this action in response to changing demand for software licenses. For details, see License Policy *(on page 91)*.

NOTE:
Workload Optimization Manager 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, Workload Optimization Manager automatically sets utilization constraints for that cluster. If you configure a failover host, Workload Optimization Manager reserves that host for HA and will not move VMs to it.

DRS Automation Settings

Workload Optimization Manager 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

  Workload Optimization Manager assigns a vendor migration level based on the migration level discovered from vCenter. The chart only shows the assigned migration level (i.e., the Workload Optimization Manager Vendor Migration Level).

Workload Optimization Manager 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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For details about host actions, see Host Actions *(on page 312)*.

Action	Default Mode	vCen-ter	Hy-per-V	UCS (blades only)
Start	Recommend	Auto	Rcmd	Auto
Suspend	Recommend	Auto	Rcmd	Auto
Provision	Recommend	Rcmd	Rcmd	Auto
Reconfig-ure	Recommend	Rcmd	Rcmd	Rcmd

You can use Action Scripts for action orchestration.

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. Workload Optimization Manager uses this setting when:

- Workload Optimization Manager actions to move VMs from one host to another are automated.
- The DRS automation level is *Fully Automated*, regardless of migration threshold.

  NOTE:
  Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager might recommend moving the VM from Host_02 to Host_03 as an alternative action.

In addition, Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager actions should be safe to execute.

The following table summarizes Workload Optimization Manager's response at various stages of maintenance.

Maintenance Status	DRS Activities	Host Status in Workload Optimization Manager	Workload Optimization Manager Pending Actions	Workload Optimization Manager New Actions
Host is entering maintenance mode.	Increased number of DRS activities moving VMs away from the	X Uncontrollable (Maintenance)	✓ Removed from the system	X Not generated

Maintenance Status	DRS Activities	Host Status in Workload Optimization Manager	Workload Optimization Manager Pending Actions	Workload Optimization Manager New Actions
	host entering maintenance			
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
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 Workload Optimization Manager 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, Workload Optimization Manager automates the first action to move a VM to another host, and then stops recommending actions. After the host leaves maintenance mode, Workload Optimization Manager automates actions to manage the performance of the cluster as normal.

Utilization Constraints

Utilization constraints affect the actions Workload Optimization Manager recommends as it manages your environment. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity.

Attribute	Default Value
Memory Overprovisioned Percentage	1000
Net Throughput	50
Ready Queue Utilization	50

Attribute	Default Value
Memory Utilization	100
IO Throughput	50
CPU Overprovisioned Percentage	1000
CPU Utilization	100
Swapping Utilization	20

For example:

- Setting 50 for Net Throughput means that Workload Optimization Manager considers 50% utilization of that throughput to be 100% of capacity and 25% utilization to be 50% of capacity
- Setting 1000 for Memory Overprovisioned Percentage means that overprovisioning memory by 5 times the physical capacity shows up as 50% utilization of the Mem Overprovisioned capacity in Workload Optimization Manager
- Setting 100 for Memory Utilization means that Workload Optimization Manager 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
Diameter	10
Center	70

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 Workload Optimization Manager considers to be the desired state.

Setting the center of the sphere chooses the priority for Workload Optimization Manager analysis. If you set the balance in favor of efficiency, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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.

Placement Policies

You can create placement policies that merge multiple clusters into a single logical group for the purpose of workload placement.

For example, you can merge three host clusters in a single provider group. This enables Workload Optimization Manager 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 88)*.

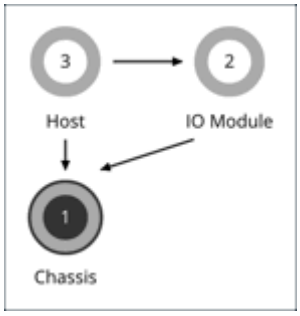
NOTE:
For vCenter, Cisco automatically imports any vSphere Host DRS rules when DRS is enabled, and displays them on the **Settings > Policies** page under **Imported Placement Policies**.

For more information, see Importing Workload Placement Policies *(on page 88)*.

Chassis

A chassis houses the servers that are part of a computing fabric. It provides compute, memory, storage, and bandwidth resources.

Synopsis



Synopsis	
Budget:	A Chassis has unlimited budget.
Provides:	Chassis resources (physical space, cooling, etc.).
Consumes:	N/A
Discovered through:	Workload Optimization Manager discovers Chassis through fabric manager targets.

NOTE:
When Workload Optimization Manager 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 datacenter to establish their relationship. When you set the scope to that datacenter and view the Health chart, you will see the blade servers in the list of hosts. In addition, when the datacenter is included in a merge policy (a policy that merges datacenters for the purpose of VM placement), the VMs in the blade servers apply the policy, allowing them to move between datacenters as necessary.

Monitored Resources

Workload Optimization Manager monitors the following resources for the servers in a chassis:

- Power

  Electricity being consumed by the Chassis

- Cooling

  The percentage of the acceptable temperature range that is utilized by this chassis. As the chassis temperature nears the high or low running temperature limits, this percentage increases.

Actions

None

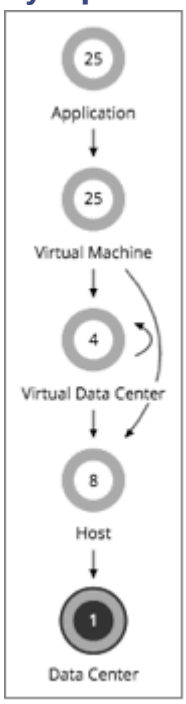
Workload Optimization Manager does not recommend actions for a chassis.

Datacenter

A datacenter is the sum of VMs, PMs, datastores, and network devices that are managed by a given hypervisor target. A datacenter provides compute, memory, storage, and bandwidth resources.

NOTE:
For public cloud environments, a datacenter is the cloud region. The hosts that get resources from the datacenter are availability zones within that region. For details, see Region *(on page 264)* and Zone *(on page 262)*.

Synopsis



Synopsis	
Budget:	A Datacenter has unlimited budget.
Provides:	Compute, memory, storage, and bandwidth resources
Consumes:	N/A
Discovered through:	Workload Optimization Manager discovers Datacenters through hypervisor targets.

NOTE:
When Workload Optimization Manager 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 datacenter to establish their relationship. When you set the scope to that datacenter and view the Health chart, you will see the blade servers in the list of hosts. In addition, when the datacenter is included in a merge policy (a policy that merges datacenters for the purpose of VM placement), the VMs in the blade servers apply the policy, allowing them to move between datacenters as necessary.

Monitored Resources

Workload Optimization Manager does not monitor resources directly from the datacenter, but it does monitor the following resources, aggregated for the hosts in a datacenter:

- Memory (Mem)

  The utilization of the PM's memory reserved or in use
- CPU

  The utilization of the PM's CPU reserved or in use

- IO

  The utilization of the PM's IO adapters
- Net

  The utilization of data through the PM's network adapters
- Swap

  The utilization of the PM's swap space
- Balloon

  The utilization of shared of memory among VMs running on the host. ESX-only
- CPU Ready

  The utilization of the PM's allocated ready queue capacity that is in use, for 1, 2, and 4 CPU ready queues. ESX-only

Actions

None

Workload Optimization Manager does not recommend actions for a datacenter. Instead, it recommends actions for the entities running in the datacenter.

Placement Policies

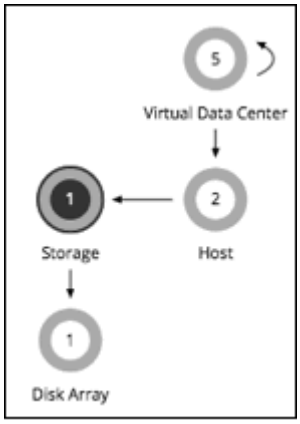
For vCenter environments, you can create placement policies that merge datacenters to support cross-vCenter moves. In this case, where a datacenter corresponds to a given vCenter target, the merged clusters can be in different datacenters. In this case you must create two merge policies; one to merge the affected datacenters, and another to merge the specific clusters.

For more information, see Creating Placement Policies *(on page 88)*.

Storage

Workload Optimization Manager represents storage as Datastores. A Datastore is a logical grouping of one or more physical storage devices that serve workload storage requirements.

Synopsis



Synopsis	
Budget:	A Datastore gains its budget by selling resources to the VMs it serves. If utilization of a Datastore is high enough, Workload Optimization Manager can recommend that you provision a new one.

Synopsis	
Provides:	Host resources for VMs to use: ▪ Storage amount ▪ IOPS (storage access operations per second) ▪ Latency (capacity for disk latency in ms)
Consumes:	Disk arrays (or aggregates)
Discovered through:	Workload Optimization Manager discovers on-prem Datastores through hypervisor targets and storage controllers.

Monitored Resources

Workload Optimization Manager monitors the following resources for a datastore:

- Storage Amount

  The utilization of the datastore's capacity
- Storage Provisioned

  The utilization of the datastore's capacity, including overprovisioning.
- Storage Access Operations Per Second (IOPS)

  The summation of the read and write access operations per second on the datastore

  NOTE:
  When it generates actions, Workload Optimization Manager 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

  The utilization of latency on the datastore

Storage 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**

  Delete a datastore or volume that has been suspended for a period of time.

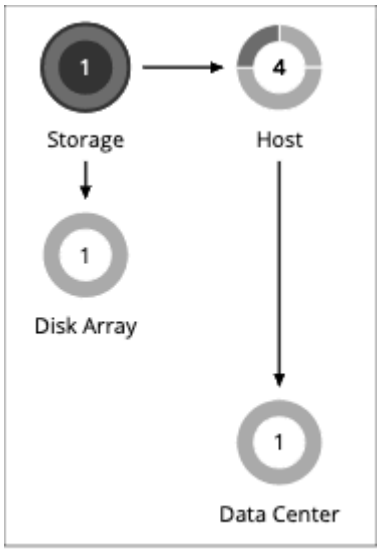
Storage resize actions use Workload Optimization Manager tuned scaling settings. This gives you increased control over the action mode Workload Optimization Manager will use for the affected actions. For an overview of tuned scaling, see Tuned Scaling for On-prem VMs *(on page 270)*.

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 88)*.

vSAN Storage

Overview



For environments that use hyperconverged infrastructure to provide storage on a vSAN, Workload Optimization Manager 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.

Workload Optimization Manager supports VMware vSAN, but does not support stretched VSAN clusters. Adding stretched clusters can cause the generation of incorrect storage recommendations and actions.

Workload Optimization Manager supports VMware vSAN.

vSAN Storage Capacity

When you consider vSAN capacity, you need to compare *Raw Capacity* with *Usable Capacity*.

- Raw Capacity
- Workload Optimization Manager discovers Raw Capacity configured in vCenter and uses it to calculate Usable Capacity. Raw Capacity displays in the Entity Information chart.
- Usable Capacity

  Workload Optimization Manager calculates Usable Capacity and then uses the calculated value to drive scaling actions. Workload Optimization Manager 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, Workload Optimization Manager considers a variety of attributes, including:

- Raw Capacity and Largest Host Capacity

  Workload Optimization Manager compares the Raw Capacity for all the hosts in the cluster and then uses the largest value as Largest Host Capacity.

- RAID Factor

  Workload Optimization Manager 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, Workload Optimization Manager 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 328)*.

The calculation for Usable Capacity can be expressed as:

```
Usable Capacity = (Raw Capacity - Largest Host Capacity * Host Capacity Reservation) * Slack Space
 Percentage * RAID Factor * Compression Ratio
```

If the result of the calculation is zero or a negative value, Workload Optimization Manager 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, Workload Optimization Manager supports VMware vSAN technology.

- Capacity

  Workload Optimization Manager 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 524)*.

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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

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 321)*.

Action	Default Mode	vCenter	Hyper-V
Delete (Volume)	Recommend	Rcmd	Rcmd
Suspend	Manual	Rcmd	Rcmd
Delete (Datastore)	Disabled	Rcmd	Rcmd
Move	Recommend	Rcmd	Rcmd
Provision	Recommend	Rcmd	Rcmd
Start	Recommend	Rcmd	Rcmd
Resize (Up, Down, Above Max, or Below Min - using tuned scaling)	Recommend	Rcmd	Rcmd

For datastores on disk arrays:

Action	Default Mode	Dell Compellent	HP 3Par	NetApp ONTAP	VNX	VMAX	Nutanix	Pure Storage
Delete (Volume)	Recommend	Rcmd	Not supported	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd
Suspend	Manual	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd
Delete (Datastore)	Disabled	Rcmd	Not supported	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd
Move	Recommend	Rcmd	Not supported	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd
Provision	Recommend	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd

Action	Default Mode	Dell Com-pellent	HP 3Par	NetApp ONTAP	VNX	VMAX	Nutanix	Pure Storage
Start	Rec-om-mend	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd
Resize (Up, Down, Above Max, or Below Min - using tuned scaling)	Rec-om-mend	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd

You can use Action Scripts for action orchestration.

For ServiceNow:

- Storage suspend and vSAN storage resize actions will not generate a CR.
- Currently Workload Optimization Manager can only execute a CR for storage provision actions on Pure and Dell Compellent storage.

Utilization Constraints

Utilization constraints affect the actions Workload Optimization Manager recommends as it manages your environment. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity.

Attribute	Default Value
Storage Amount Utilization	90
Storage Provisioned Utilization	100
IOPS Utilization	100
Latency Utilization	100

For example, setting 90 for Storage Amount Utilization means that Workload Optimization Manager considers 90% utilization of the physical storage to be 100% of capacity.

Storage Settings

Attribute	Default Setting/Value
Storage Overprovisioned Percentage	200
IOPS Capacity	50000
Storage Latency Capacity [ms]	100
Minimum Wasted Files [KB]	1000

Attribute	Default Setting/Value
Directories to Ignore	\.dvsData.*|\.snapshot.*|\.vSphere-HA.*|\.naa.*|\.etc.*|lost\+found.*
Files to Ignore	Empty String

- **Storage Overprovisioned Percentage**

  Storage Overprovisioned Percentage sets how much overprovisioning Workload Optimization Manager assumes when recommending actions for VM datastores. For example, if a datastore has a 30 GB capacity, and Storage Overprovisioned Percentage is set to 200, Workload Optimization Manager 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.

  Workload Optimization Manager 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**

  This sets the maximum storage latency to tolerate on a datastore, in ms. The default setting is 100 ms.

  Workload Optimization Manager 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 Workload Optimization Manager will show latency utilization of 50%.

  For VMAX environments, Workload Optimization Manager 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 Workload Optimization Manager policy, analysis will use that value instead.

- **Minimum Wasted Files**

  You can make settings to control how Workload Optimization Manager tracks and reports on wasted storage in your environment. Wasted storage is any disk space devoted to files that are not required for operations of the devices or applications in your environment. Wasted 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 don't want to track for wasted storage, set the given scope and disable datastore browsing there. If you prefer not to use Workload Optimization Manager resources to track wasted storage, leave the global setting checked.

  The settings for **Directories to Ignore** and **Files to Ignore** specify directories and files that Workload Optimization Manager will not consider when looking for wasted data storage space. Separate items in these lists with the OR bar ("|").

Scaling Constraints

Rate of Resize

Workload Optimization Manager uses the Rate of Resize setting to determine how to make storage resize changes in a single action.

Attribute	Default Value
Rate of Resize	High (3)

- **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.

Increment Constant for Storage Amount

This setting controls how many GB to add or subtract when resizing the allocation for a datastore.

Attribute	Default Value
Increment Constant for Storage Amount [GB]	100 GB

Hyperconverged Infrastructure Settings

Workload Optimization Manager considers these settings when calculating capacity and utilization for hyperconverged environments.

Attribute	Default Setting/Value
Host Capacity Reservation	1
Host IOPS Capacity	50000
Slack Space Percentage	25
Compression Ratio	1
Usable Space Includes Compression	Off

NOTE:
Workload Optimization Manager 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 322)*.

- **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.

Workload Optimization Manager 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, Workload Optimization Manager needs an estimate of datastore IOPS capacity (storage access). Workload Optimization Manager 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.

  Workload Optimization Manager 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. Workload Optimization Manager 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 Workload Optimization Manager to consider the compression ratio when calculating storage utilization and capacity. Whether this is on or off, Workload Optimization Manager always considers compression when calculating utilization of StorageProvisioned.

Placement Policies

Workload Optimization Manager supports placement policies for storage and storage clusters.

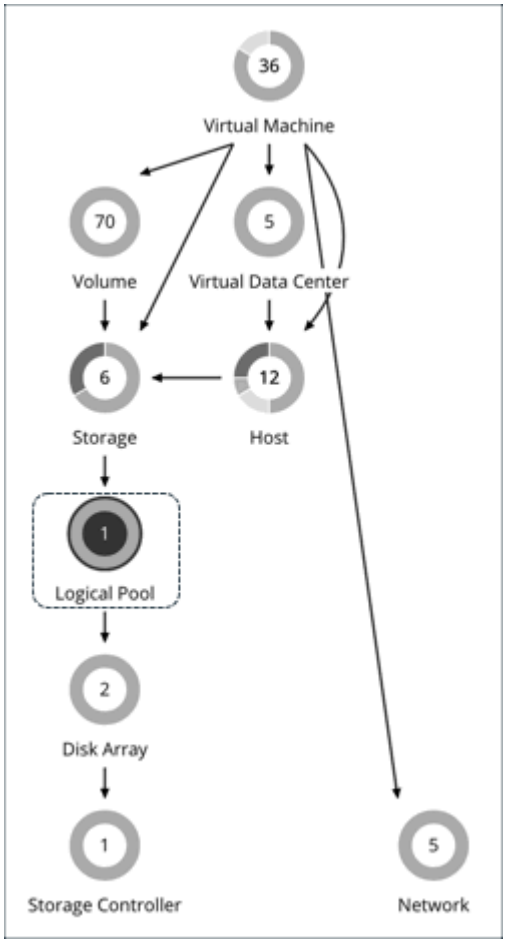
- You can create placement policies to enforce constraints for storage move actions. For example, you can have a policy that allows storage to only move to certain disk arrays, or a policy that prevents storage from moving to certain disk arrays.
- You can create placement policies that merge multiple clusters into a single logical group for the purpose of workload placement.

For more information, see Creating Placement Policies *(on page 88)*.

Logical Pool

A logical pool represents storage resources that are managed together and presented as a single storage system. Workload Optimization Manager 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	
Budget:	N/A
Provides:	Storage resources
Consumes:	Disk array resources
Discovered through:	Storage targets

Monitored Resources

Workload Optimization Manager monitors the following resources for a logical pool:

- Storage Amount

  The utilization of the logical pool's capacity.
- Storage Provisioned

  The utilization of the logical pool's capacity, including overprovisioning.
- Storage Access Operations Per Second (IOPS)

  The summation of the read and write access operations per second on the logical pool.

- Latency

  The utilization of latency on the logical pool.

Logical Pool Actions

- Resize
- Provision
- Move
- Start
- Suspend

Logical Pool Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

Action	Default Mode
Suspend	Disabled
Start	Disabled
Resize	Recommend
Move	Disabled
Provi-sion	Disabled

Storage Settings

Attribute	Default Value
Storage Latency Capacity [ms]	100
Storage Overprovisioned Percentage	200
IOPS Capacity	50000

- **Storage Latency Capacity**

  This 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 Workload Optimization Manager assumes when recommending actions for logical pools.

- **IOPS Capacity**

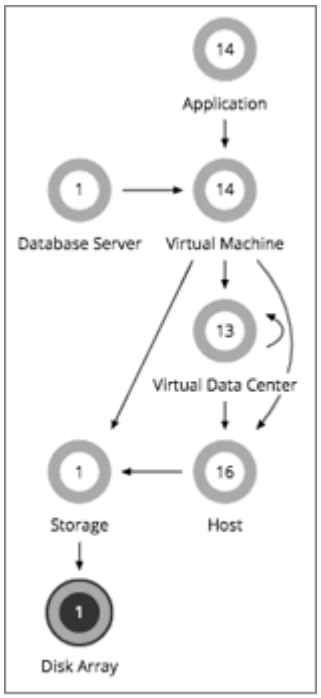
  IOPS Capacity is the IOPS setting for individual logical pools.

NOTE:
Workload Optimization Manager 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 Workload Optimization Manager 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	
Budget:	A disk array gains its budget by selling resources to the datastores it serves. If utilization of a disk array is high enough, Workload Optimization Manager can recommend that you provision a new one.
Provides:	Storage resources for datastores to use: ▪ Storage amount ▪ Storage Provisioned ▪ IOPS (storage access operations per second) ▪ Latency (capacity for disk latency in ms)

Synopsis	
Consumes:	Storage controllers
Discovered through:	Workload Optimization Manager discovers disk arrays through storage controller targets.

Monitored Resources

Workload Optimization Manager monitors the following resources for a disk array:

> **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 widget in the UI will display no data.

- Storage Amount

  The utilization of the Disk Array's capacity.
- Storage Provisioned

  The utilization of the Disk Array's capacity, including overprovisioning.
- Storage Access Operations Per Second (IOPS)

  The summation of the read and write access operations per second on the disk array
- Latency

  The utilization of latency, computed from the latency of each device in the disk array.

Disk Array Actions

- **Provision**

  For high utilization of the disk array’s storage, provision a new disk array (recommendation, only).
- **Start**

  For high utilization of disk array, start a suspended disk array (recommendation, only).
- **Suspend**

  For low utilization of the disk array’s storage, move VMs to other datastores and suspend volumes on the disk array (recommendation, only).
- **Move**

  (Only for NetApp Cluster-Mode) For high utilization of Storage Controller resources, Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager can move a datastore to another array.

Disk Array Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

The following table describes the default action mode for disk array actions and automation support for environments that have Disk Array Storage Controllers as targets.

Action	Default Mode	Dell Compellent	HP 3Par	NetApp ONTAP	VMAX	VNX	Nutanix	Pure Storage	XTremIO
Move	Disabled			Rcmd					
Provision	Recommend	Rcmd	Rcmd	Rcmd			Rcmd		
Resize (up)	Recommend	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd		
Start	Recommend								
Suspend	Disabled								

Action Automation for NetApp Storage Systems

For NetApp storage systems, the actions Workload Optimization Manager 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
Move Datastore between disk arrays on the same storage controller	Yes
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, Workload Optimization Manager can recommend moving an aggregate to another storage controller.

Utilization Constraints

Utilization constraints affect the actions Workload Optimization Manager recommends as it manages your environment. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Workload Optimization Manager 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
IOPS Capacity A generic setting for disk array IOPS capacity (see **Disk Array IOPS Capacity** below).	5000
VSeries LUN IOPS Capacity	5000
7.2k Disk IOPS Capacity	800
10k Disk IOPS Capacity	1200
15k Disk IOPS Capacity	1600
SSD Disk IOPS Capacity	50000
Disk Array IOPS Capacity	10000
Storage Overprovisioned Percentage	200
Storage Latency Capacity [ms]	100

NOTE:
Workload Optimization Manager 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 Workload Optimization Manager policies.

- **IOPS Capacity**

  The capacity of IOPS (IO operations per second) that your storage devices can support. Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager 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 Workload Optimization Manager 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.

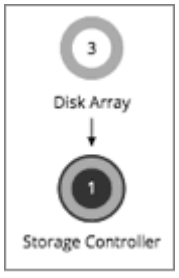
- **Storage Overprovisioned**

  This setting indicates how much overprovisioning Workload Optimization Manager assumes when recommending actions for disk arrays. For example, if a disk array has a 30 TB capacity, and DiskArray Overprovisioned Percentage is set to 200, Workload Optimization Manager will treat the datastore as though it has a capacity of 60 TB, or 200% of the actual disk array capacity.

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



Synopsis	
Budget:	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, Workload Optimization Manager 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:	Workload Optimization Manager directly accesses storage controller targets.

Monitored Resources

Workload Optimization Manager monitors the following resources for a storage controller:

- CPU

  The utilization of the Storage Controller's allocated CPU

- Storage Amount

  The utilization of the storage controller's capacity. The storage allocated to a storage controller is the total of all the physical space available to aggregates managed by that storage controller.

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

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

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

Actions for individual Disk Array Storage Controllers:

Action	Default Mode	Dell Compellent	HP 3Par	NetApp ONTAP	VNX	VMAX	Nutanix	Pure Storage	XTremIO
Provision	Disabled	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd	Rcmd

Utilization Constraints

Utilization constraints affect the actions Workload Optimization Manager recommends as it manages your environment. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity.

Attribute	Default Value
Storage Amount Utilization	90 Maximum allowed utilization of storage that is managed by the Storage Controller.
CPU Utilization	100 Maximum allowed utilization of Storage Controller CPU (from 20 to 100).

Storage Settings

Set capacity for specific storage resources.

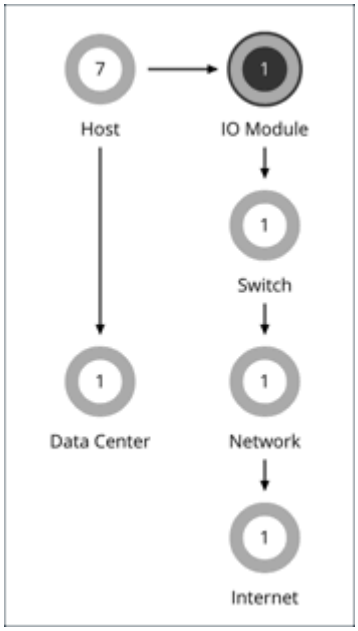
Attribute	Default Value
IOPS Capacity	5000
Storage Latency Capacity [ms]	100

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.

Workload Optimization Manager supports IO Modules when you have installed the Fabric Control Module license.

Synopsis



Synopsis	
Budget:	An IO Module gains its budget by selling Net resources to a physical machine.
Provides:	Net resources
Consumes:	Chassis and Fabric Interconnect
Discovered through:	Workload Optimization Manager discovers IO Modules through the fabric managers that use them.

Monitored Resources

Workload Optimization Manager monitors the following resources for an IO Module:

- NetThroughput

  Rate of message delivery over a port

Actions

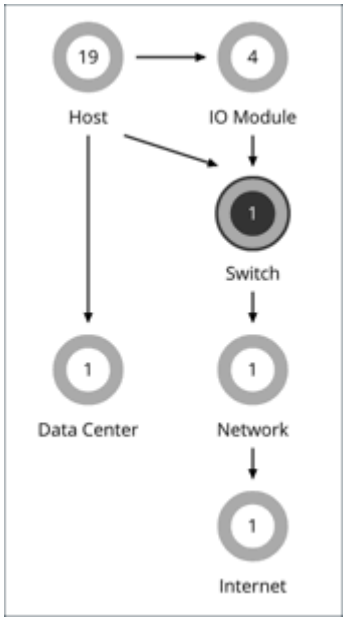
None

Workload Optimization Manager does not recommend actions for an 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	
Budget:	A switch gains its budget by selling Net resources to the IO Modules.
Provides:	Net resources
Consumes:	N/A
Discovered through:	Workload Optimization Manager discovers switches through managers of fabric platforms (such as UCS) that use them.

Monitored Resources

Workload Optimization Manager monitors the following resources for a switch:

- NetThroughput

  Rate of message delivery over a port
- PortChannel

  Amalgamation of ports with a shared net throughput and utilization

Actions

Resize

Resize PortChannel for a switch to increase bandwidth.

Switch Policies

Workload Optimization Manager ships with default settings that we believe will give you the best results from our analysis. These settings are specified in a set of default automation policies for each type of entity in your environment. For some scopes of your environment, you might want to change these settings. For example, you might want to change *action automation* or *constraints* for that scope. You can create policies that override the defaults for the scopes you specify.

Action Automation and Orchestration

For environments that have Fabric Managers as targets:

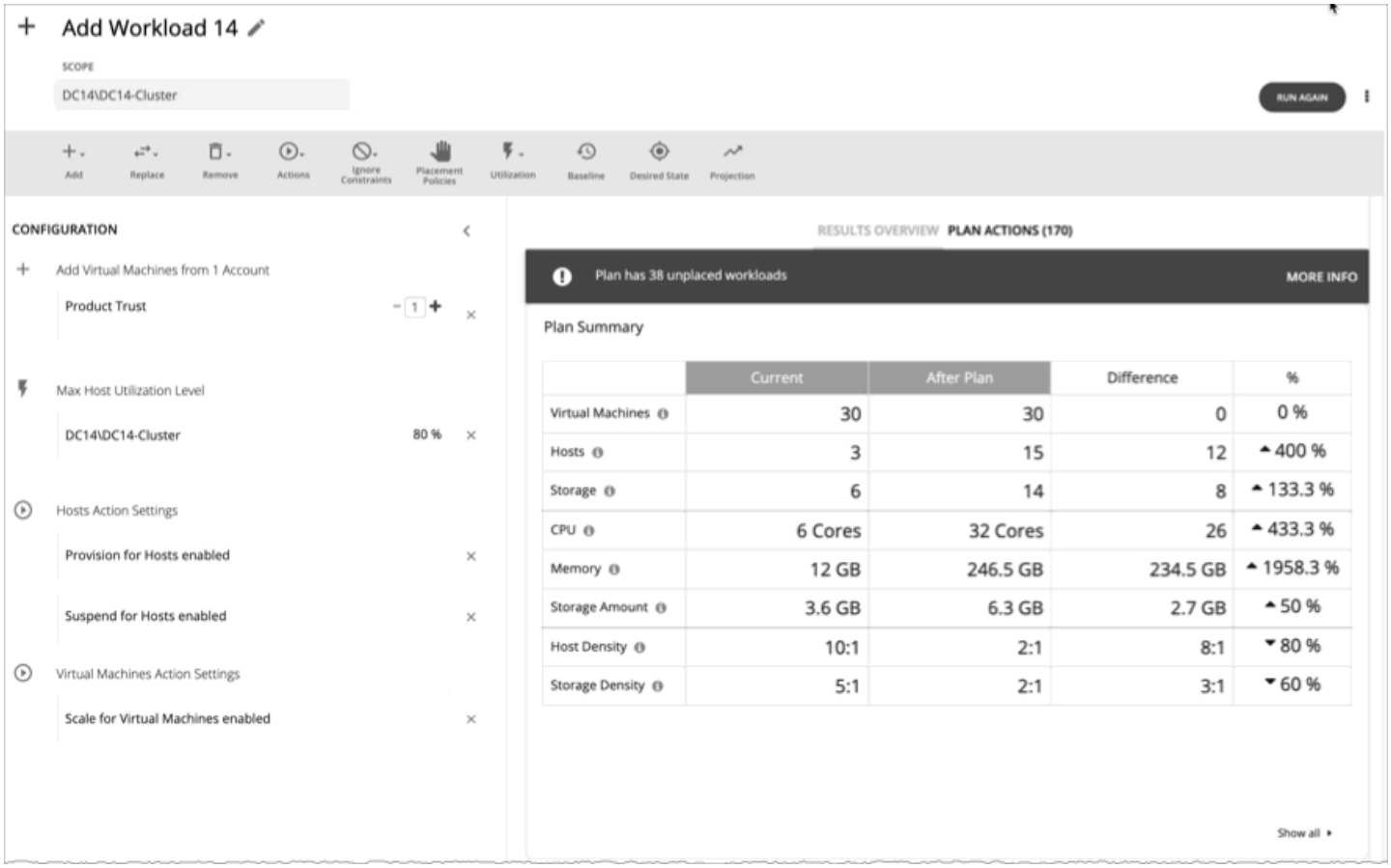
Action	Default Mode	Cisco UCS
Resize	Recommend	Rcmd
Start	Recommend	Rcmd
Provi-sion	Recommend	Rcmd
Suspend	Disabled	Rcmd
Move	Disabled	Rcmd

Utilization Constraints

Utilization constraints affect the actions Workload Optimization Manager recommends as it manages your environment. Workload Optimization Manager recommends actions that avoid using these resources beyond the given settings. The values you set here specify what percentage of the existing capacity that Workload Optimization Manager will consider to be 100% of capacity.

Attribute	Default Value
Switch Net Throughput	70

Plans: Looking to the Future



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, Workload Optimization Manager 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, Workload Optimization Manager 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.

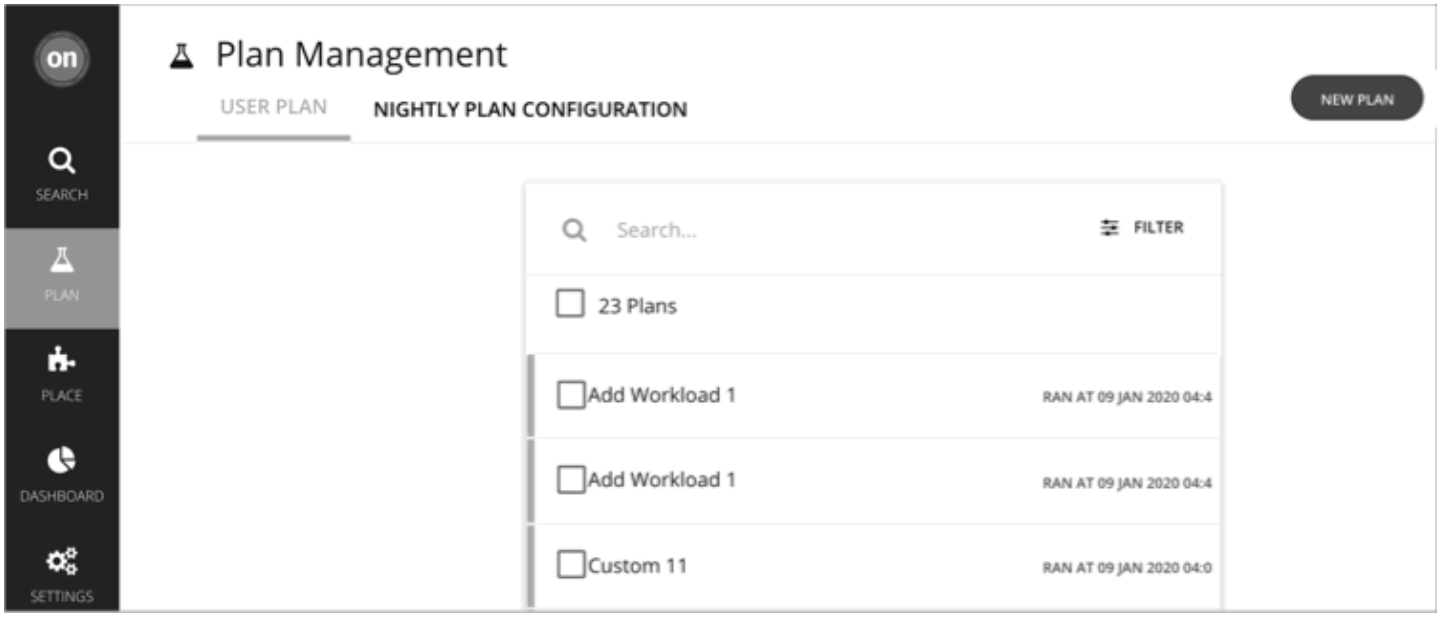
For example, assume a scenario that adds virtual machines to a cluster. To run the plan, Workload Optimization Manager takes a snapshot of the current market, and adds the VMs to the specified cluster. Workload Optimization Manager then runs analysis on the plan market, where each entity in the supply 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 Workload Optimization Manager navigation bar.

- Create new plans

  To create a new plan, click the **NEW PLAN** button. See Setting Up Plan Scenarios *(on page 344)*.

- View saved plans

  After you create and run a plan, Workload Optimization Manager 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

  Workload Optimization Manager 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 426)*.

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

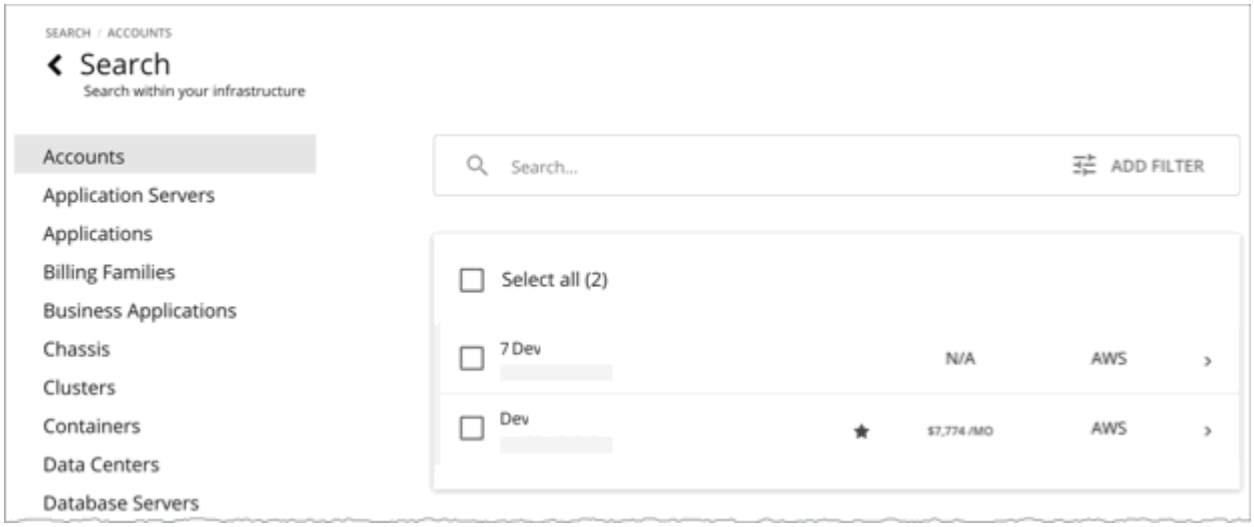
You can begin creating a plan scenario from different places in the user interface:

- 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 **Home Page**

  To start a plan scenario from the **Home Page**, 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

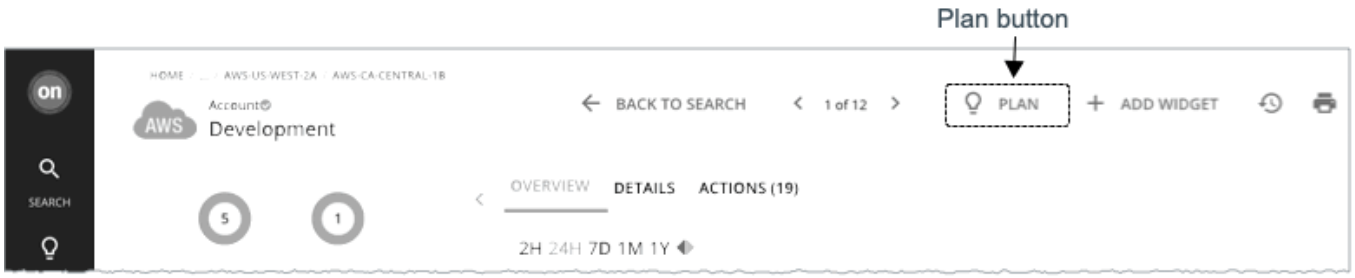
  If you set the scope to *a specific* Cluster, Datacenter, Group, Storage Cluster, or Virtual Datacenter, you can start any plan. You may need to go through additional steps, depending on your chosen plan type. For example, if you scope to a cluster and choose the Add Virtual Machines plan type, the plan wizard prompts you to select the most suitable templates for the VMs you plan to add to the cluster.

- Container 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.

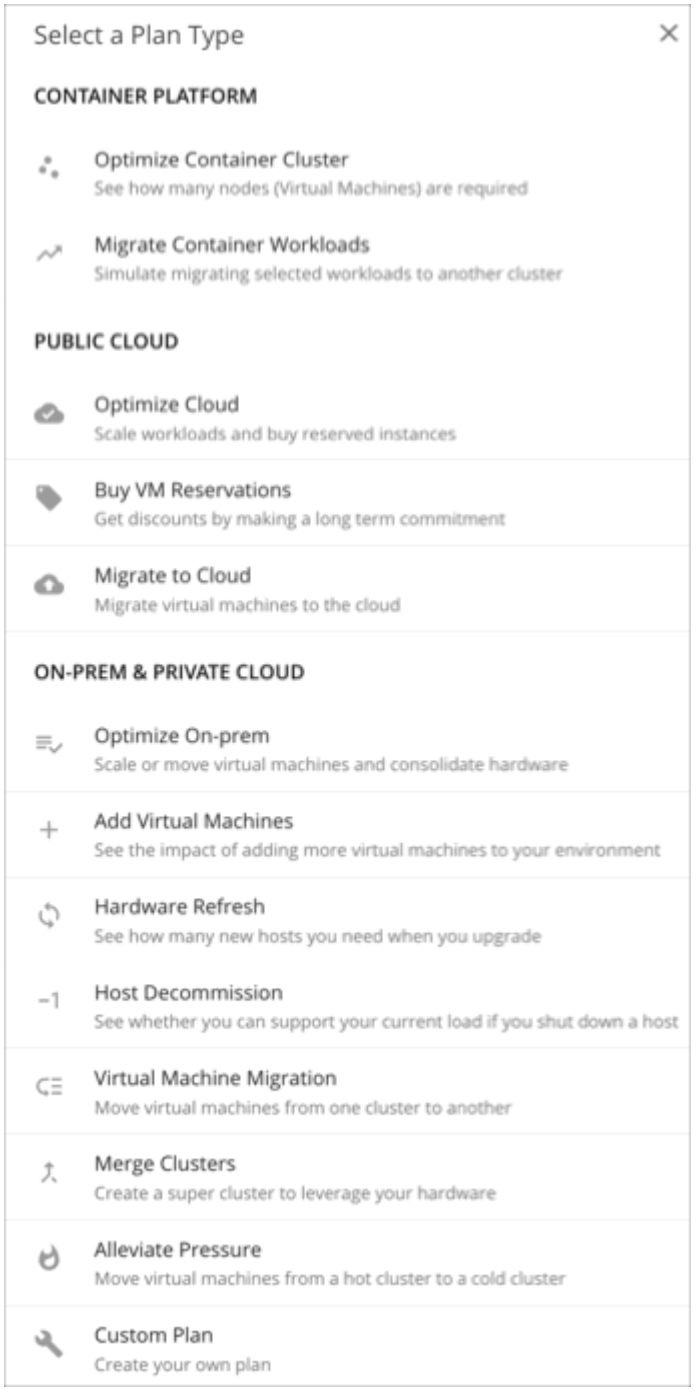
For details, see Scoping the Workload Optimization Manager Session *(on page 44)*.

After setting the scope, the **Plan** button appears in the **Home Page**.



2. Plan Types

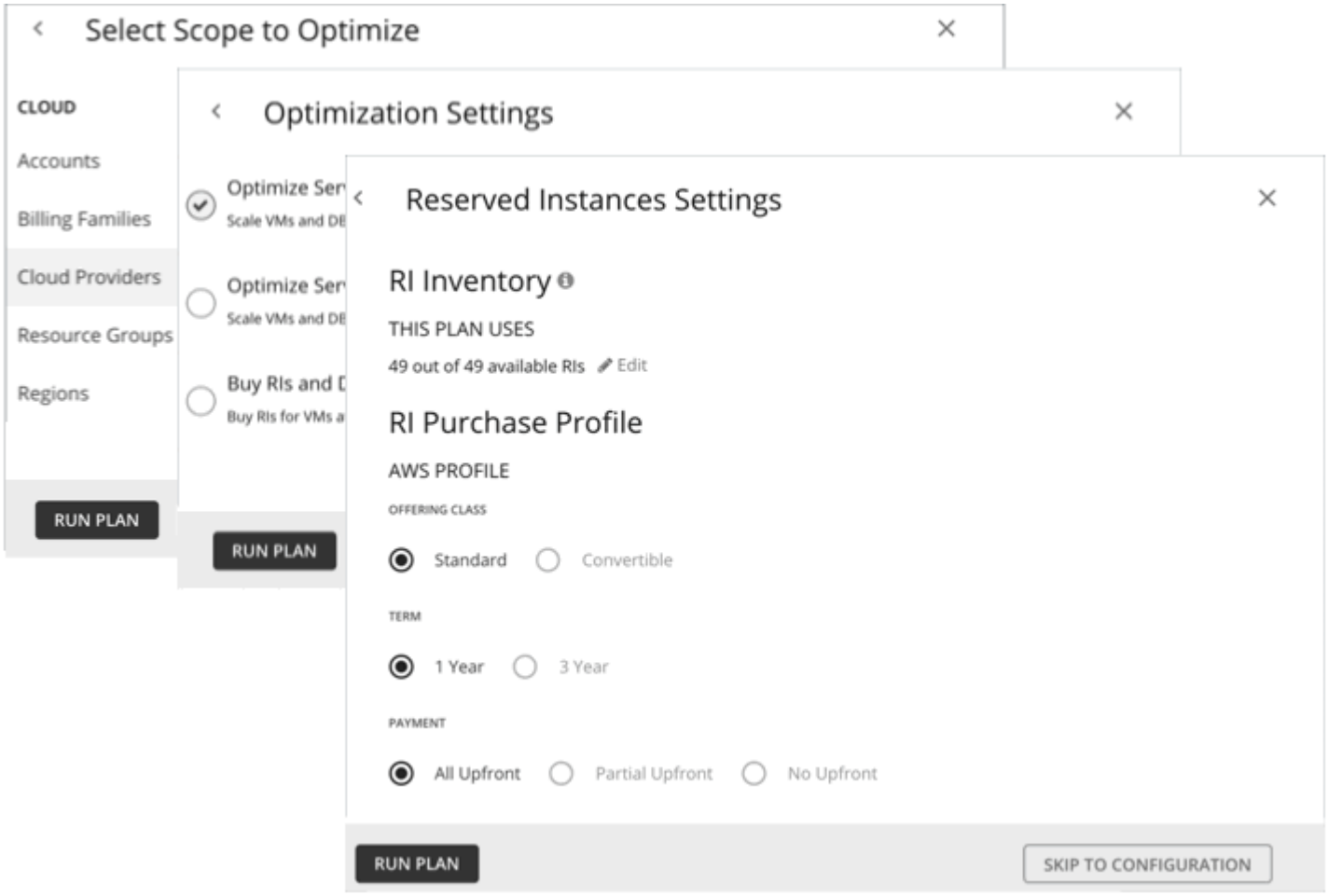
Select from the list of plan types. For more information, see Plan Scenarios and Types *(on page 351)*.



Workload Optimization Manager opens the appropriate plan wizard.

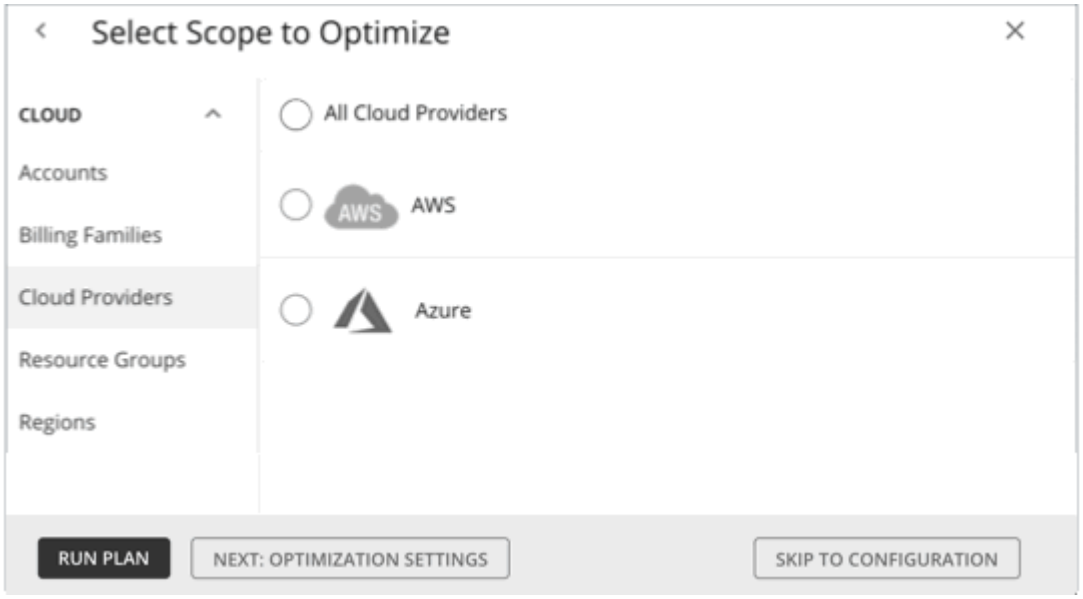
3. Plan Wizards

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.



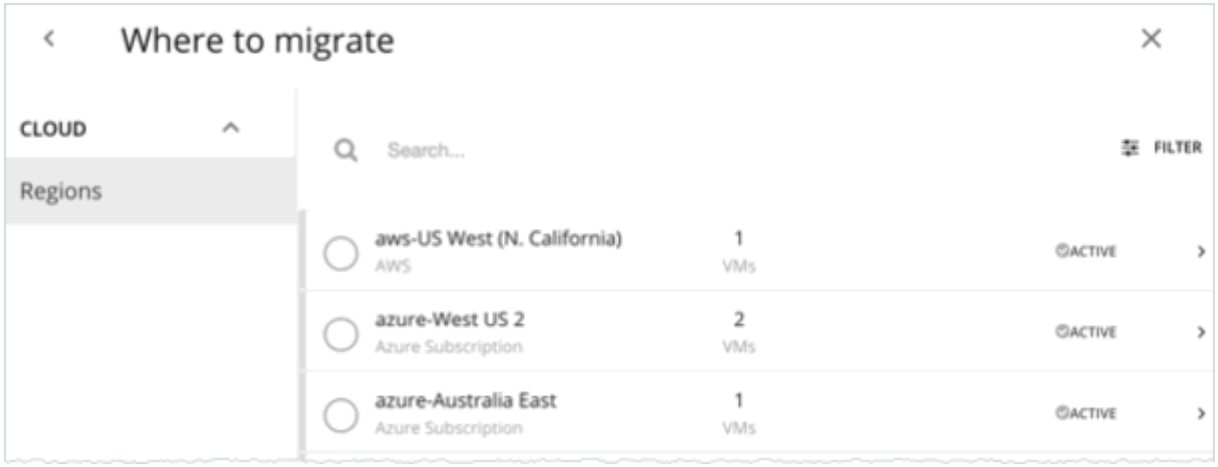
It usually helps to focus on a subset of your environment. For a very large environment, scoped plans run faster.

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.

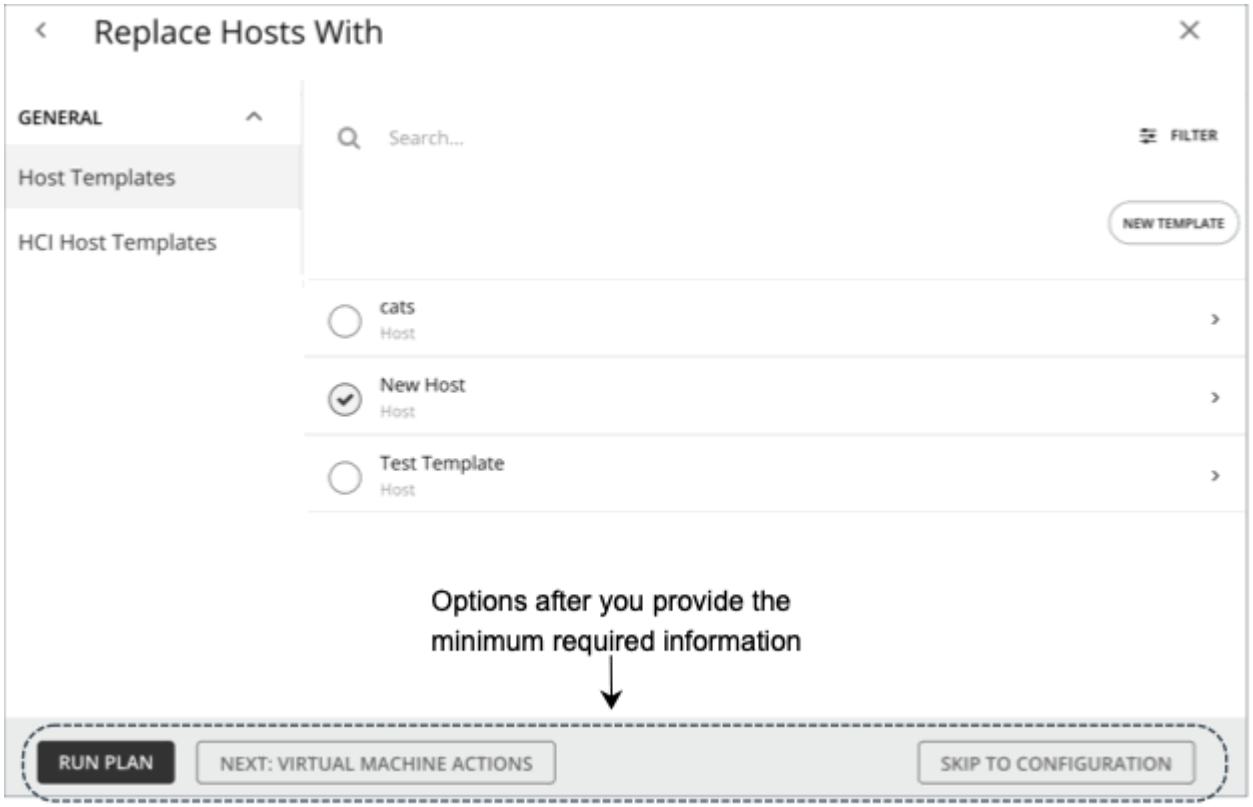
5. Additional Plan Information

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.

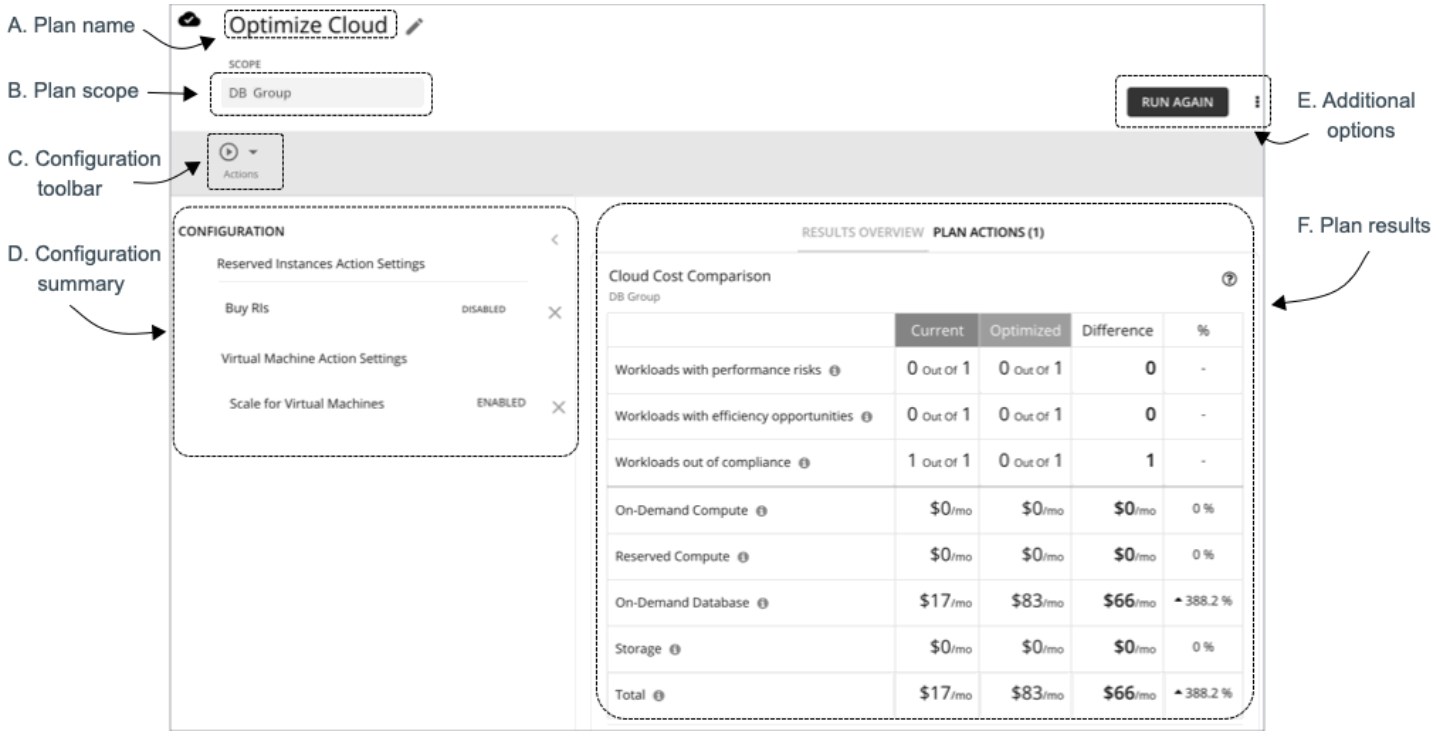
NOTE: For a custom plan, the only option available is **Configure Plan**. Click this button to open the Plan Page, configure the plan settings, and then run the plan.

7. The Plan Page

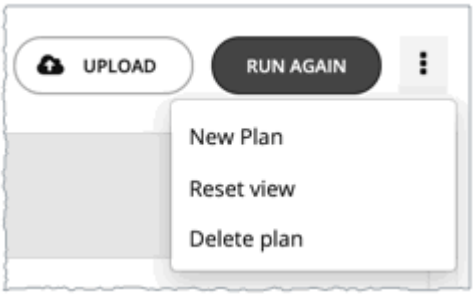
The Plan Page first displays if you skip the wizard or as soon as you run a plan.

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:



Plan Page Sections	Description
A. Plan name	Workload Optimization Manager 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.

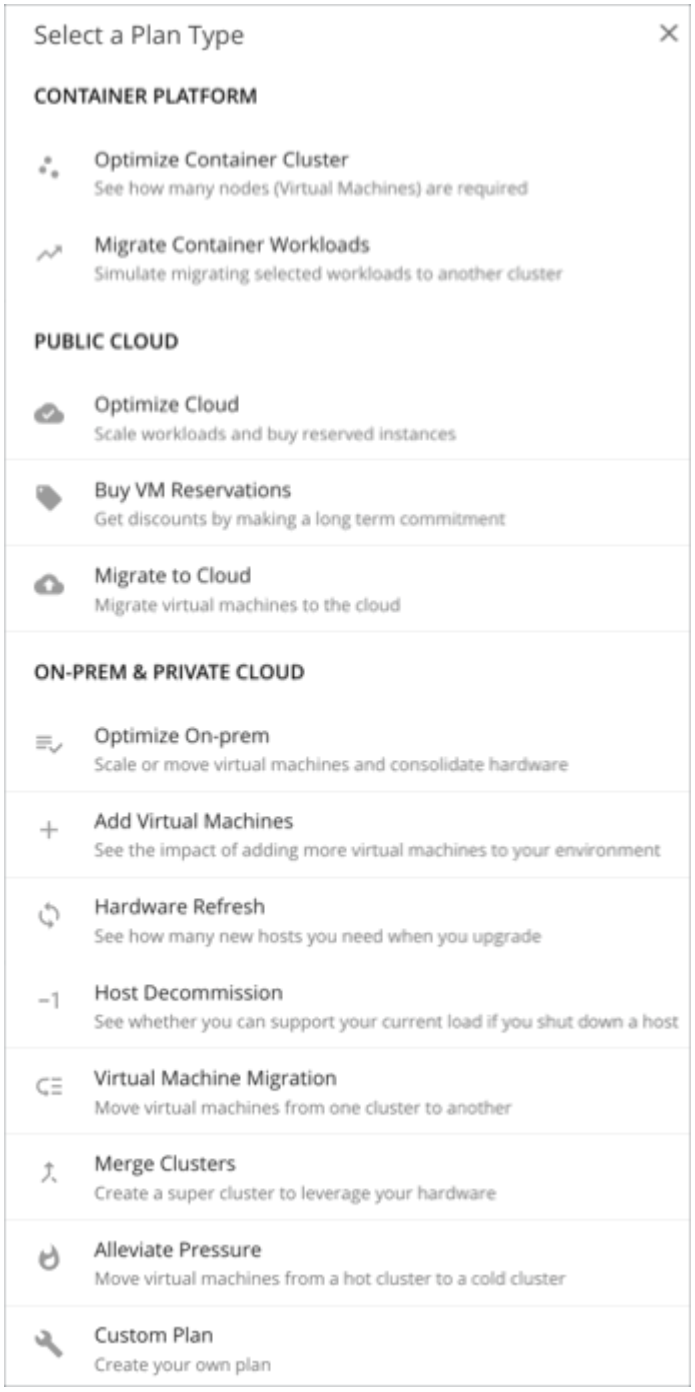
Plan Page Sections	Description
E. Additional options	See what else you can do with the plan. ▪ **Upload**: (For Azure only) Upload the results of a Migrate to Cloud plan to the Azure Migrate portal. For details, see Uploading the Results to Azure *(on page 401)*. ▪ **Run** / **Run Again**: – If a plan 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.  – **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	Review the results in the charts provided. 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.

8. Plan Management

All the plans you have created display in the Plan Management Page *(on page 343)*.

Plan Scenarios and Types

To simulate different plan scenarios, Workload Optimization Manager 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 Kubernetes 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 Workload Optimization Manager 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 30)*, 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 30)* 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, Workload Optimization Manager 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 datacenters (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 Clusters

See the effects of merging two or more clusters. For example, you can see if merging the clusters would require provisioning additional storage to support current demand, or if ignoring cluster boundaries would improve performance and efficiency.

Alleviate Pressure

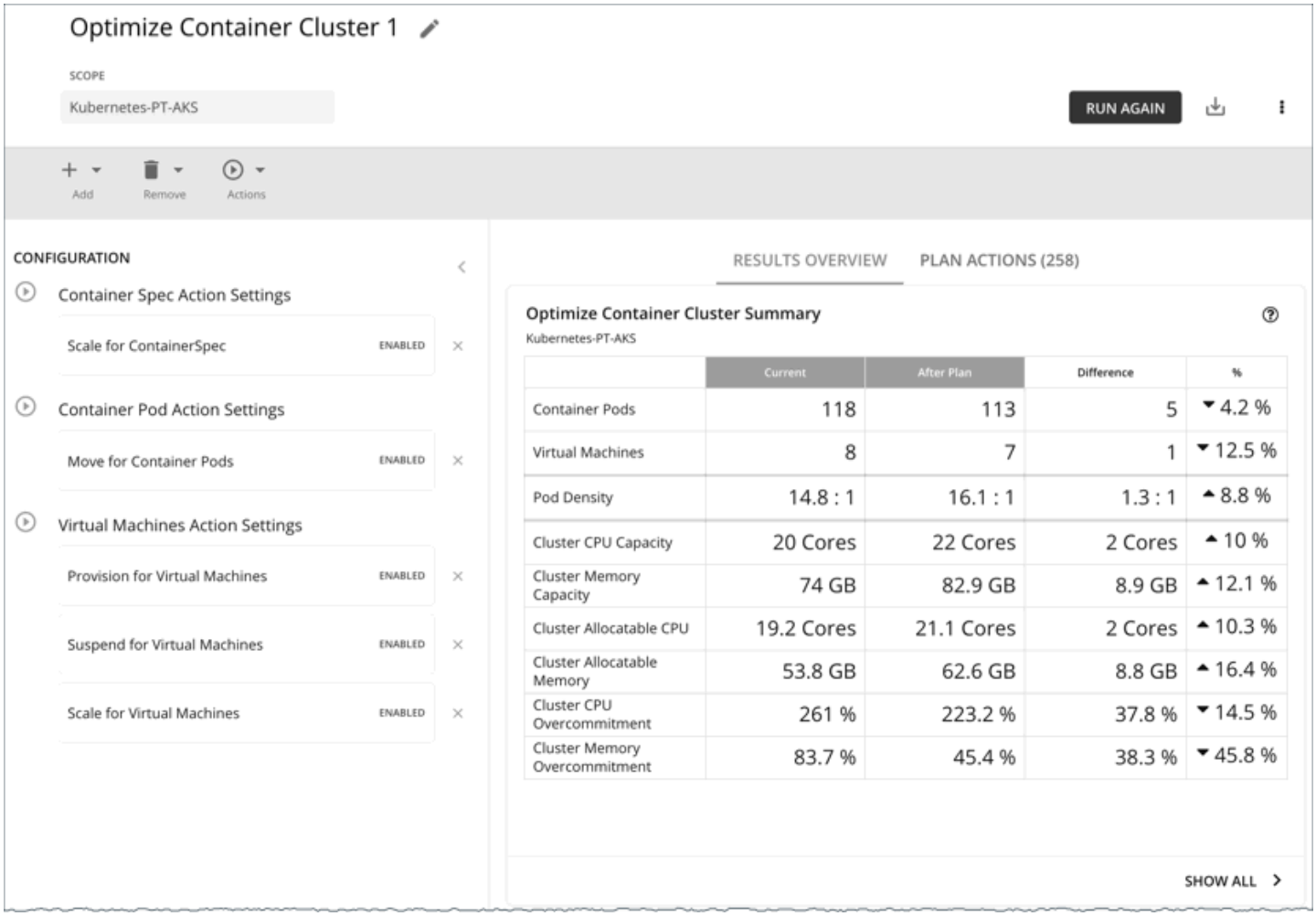
Choose a cluster that shows bottlenecks or other risks to performance, and check to see the minimal changes you can make by migrating some workloads to another cluster. The cluster that is showing risks is a *hot* cluster, and the cluster you will migrate to is a *cold* cluster.

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



Run an Optimize Container Cluster plan to identify performance and efficiency opportunities for a single Kubernetes 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:

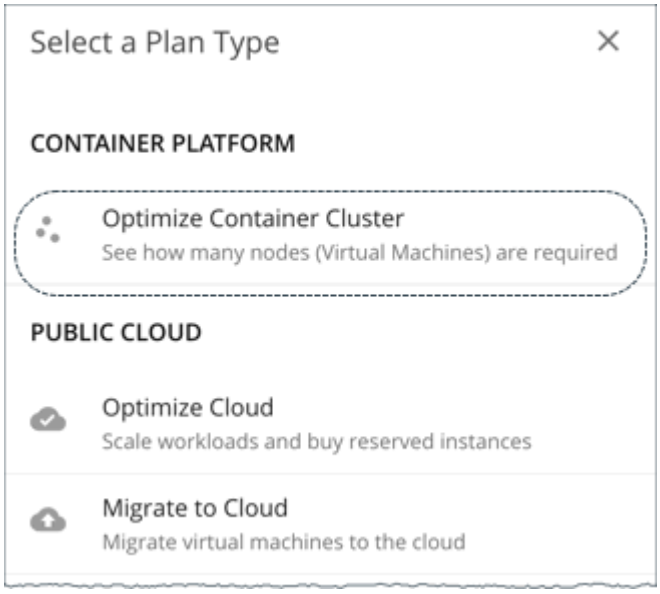
- Standalone container cluster
- Container cluster in an on-prem or public cloud environment
- Container cluster stitched to applications via Data Ingestion Framework (DIF)

Scoping to a group within a Kubernetes 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 Kubernetes cluster.

For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.



1. Scope

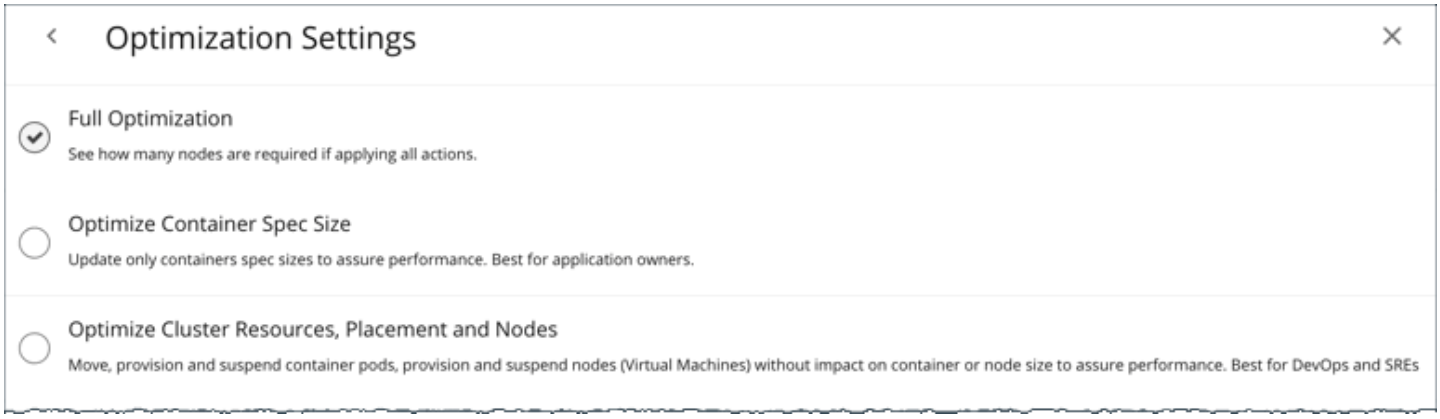
Select a Kubernetes cluster to optimize.

Scoping to a group within a Kubernetes 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. Workload Optimization Manager runs the **Full Optimization** scenario in this case.

2. Optimization Settings

Choose from the given optimization scenarios.



- **Full Optimization**

  Workload Optimization Manager 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.

  Workload Optimization Manager 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, Workload Optimization Manager shows the cost impact of actions on nodes and volumes, to help you track your cloud spend. Workload Optimization Manager 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, Workload Optimization Manager can also recommend the following actions:

  - Move VMs
  - Provision or suspend hosts
  - Provision or suspend storage

- **Optimize Container Spec Size**

  Workload Optimization Manager 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**

  Workload Optimization Manager 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.

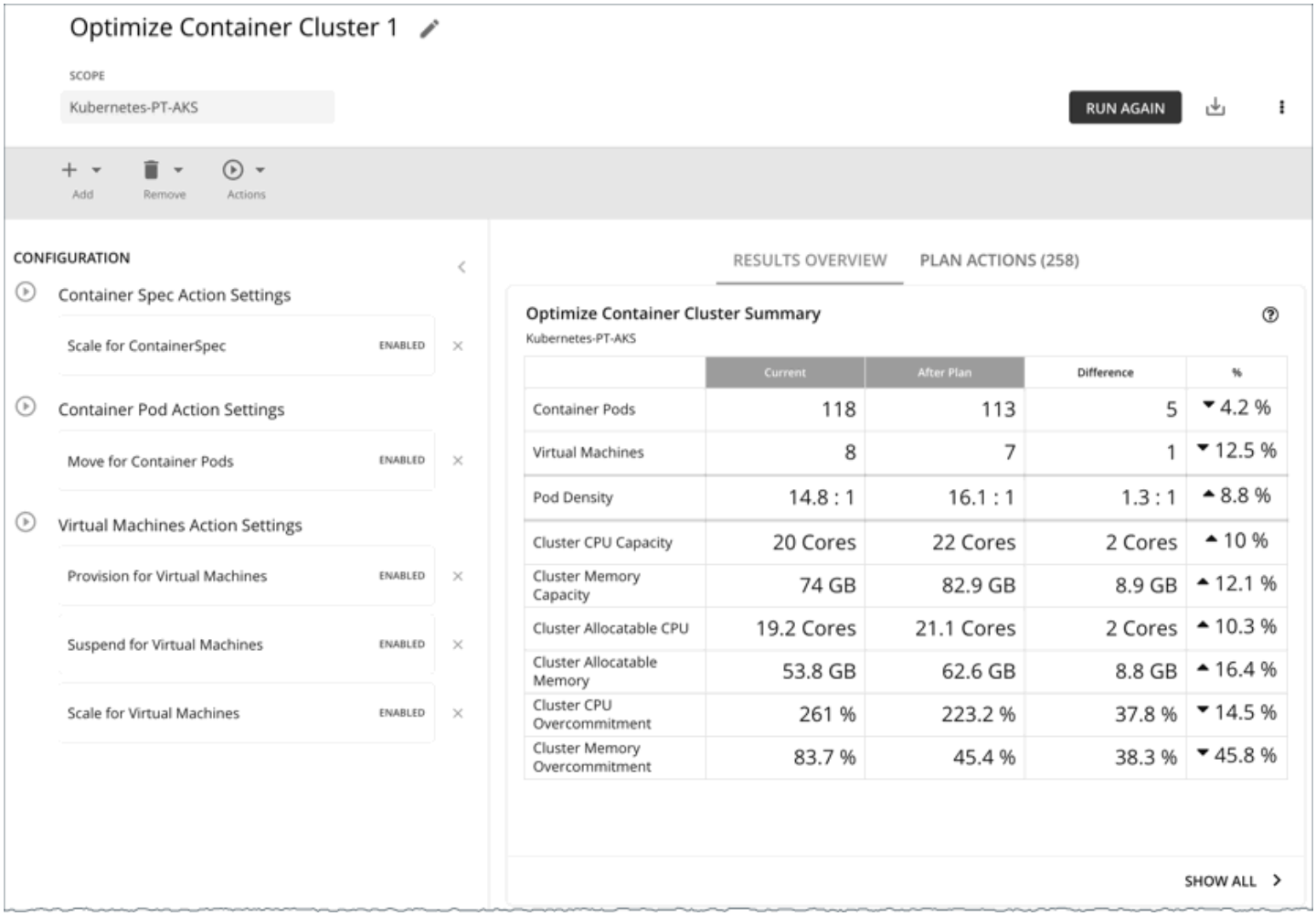
  Select an existing pod within or outside the selected Kubernetes cluster, and then specify how many copies to add. The plan simulates adding pods with the same resources as the selected pod.

- 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.

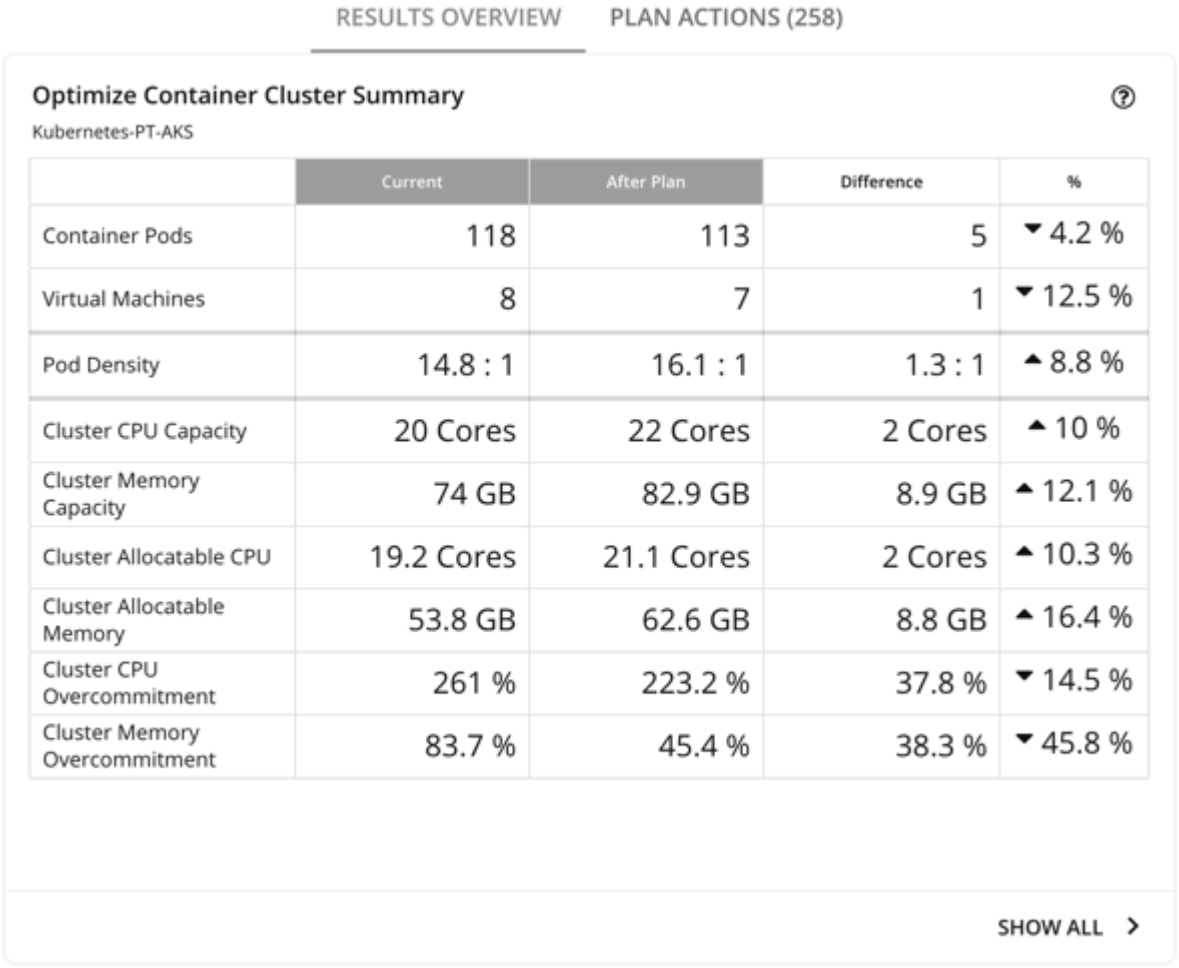


General Guidelines

Familiarize yourself with these common terms that appear in the plan results:

- A container pod represents the compute demand from a running pod.
- A Kubernetes 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



	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 %

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. Workload Optimization Manager manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.

  Workload Optimization Manager only calculates overcommitment in plans. The calculation can be expressed as:

  ```
  Overcommitment = Sum of CPU limits for all containers / Sum of CPU capacity for all nodes
  ```

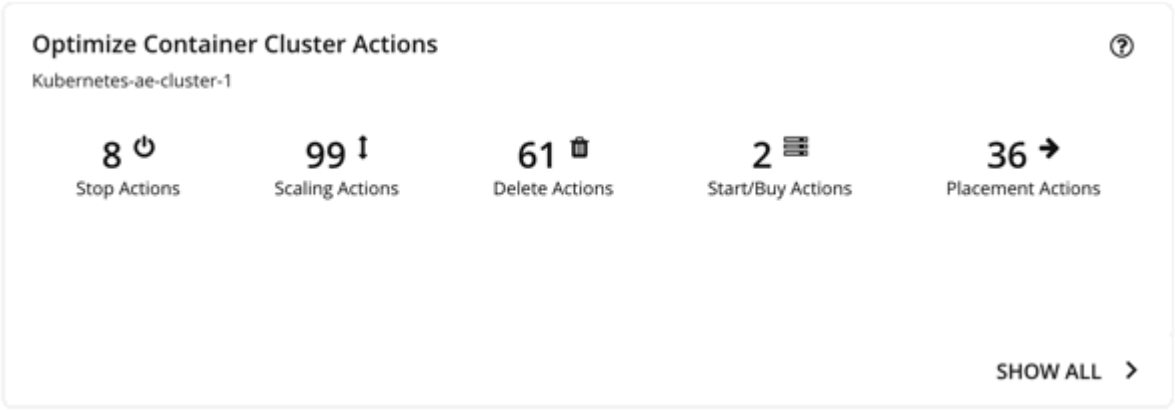
- **Cluster Memory Overcommitment**

  (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. Workload Optimization Manager manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk.

  Workload Optimization Manager only calculates overcommitment in plans. The calculation can be expressed as:

  ```
  Overcommitment = Sum of memory limits for all containers / 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.

Workload Optimization Manager 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, Workload Optimization Manager 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, Workload Optimization Manager shows the cost impact of actions on nodes and volumes, to help you track your cloud spend. Workload Optimization Manager 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, Workload Optimization Manager can also recommend the following actions:

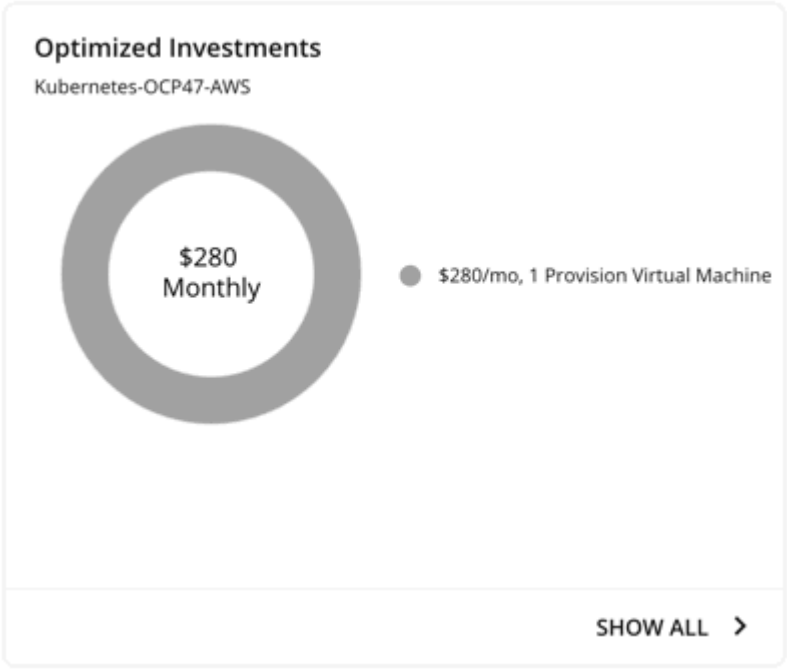
- Move VMs
- Provision or suspend hosts
- Provision or suspend storage

Optimized Savings

For a cluster in the public cloud, Workload Optimization Manager 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.

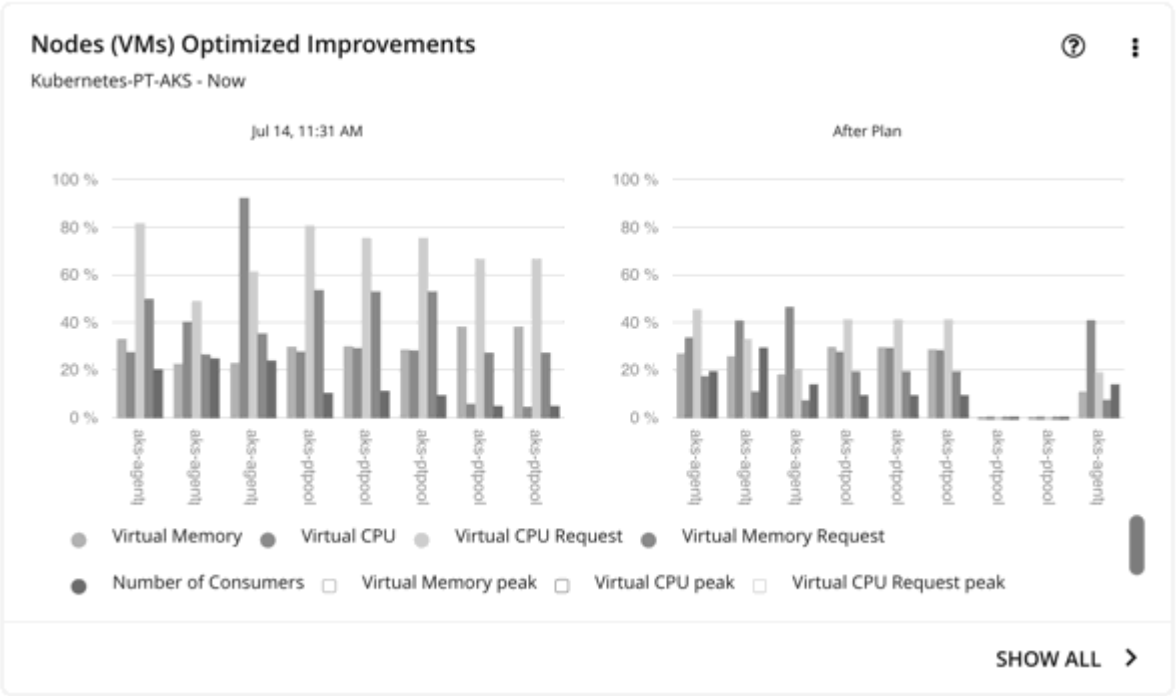
Optimized Investments



For a cluster in the public cloud, Workload Optimization Manager 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, Workload Optimization Manager 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.

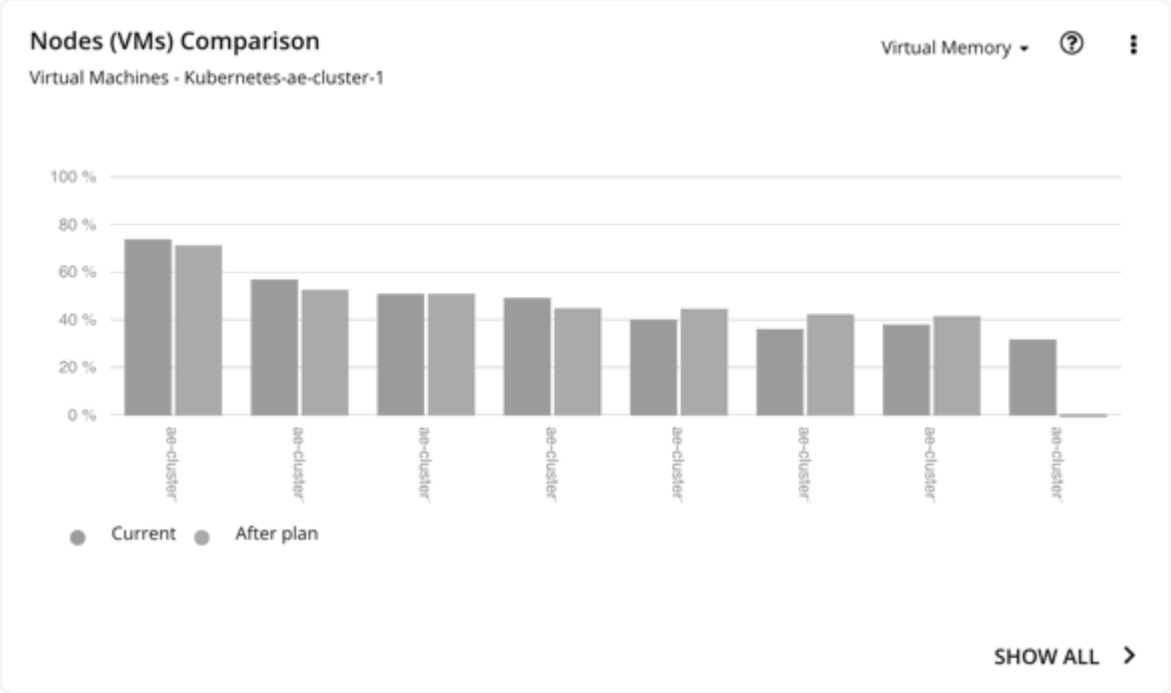
Nodes (VMs) Optimized Improvements



This chart compares the following before and after the plan:

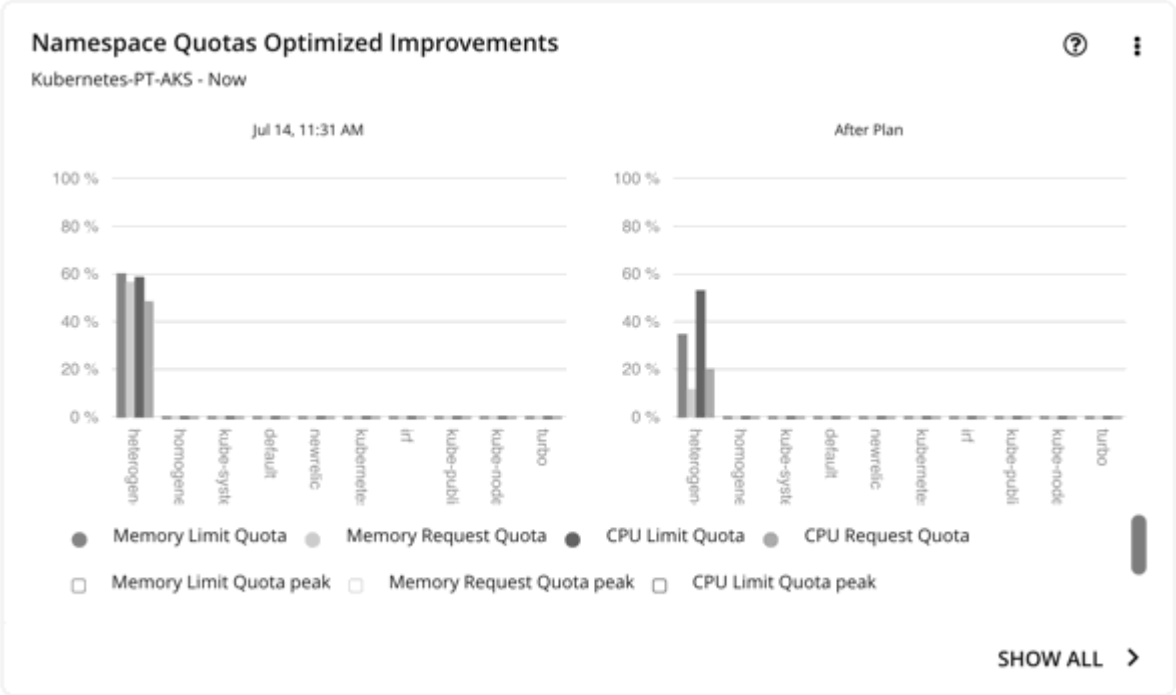
- Utilization of the following for all nodes:
  - vMem
  - vCPU
  - vMem Request
  - vCPU Request
- Number of pods consuming resources against the maximum pod capacity for all the nodes

Nodes (VMs) Comparison



This chart compares node resource utilization (one metric at a time) before and after the plan.

Namespace Quotas Optimized Improvements



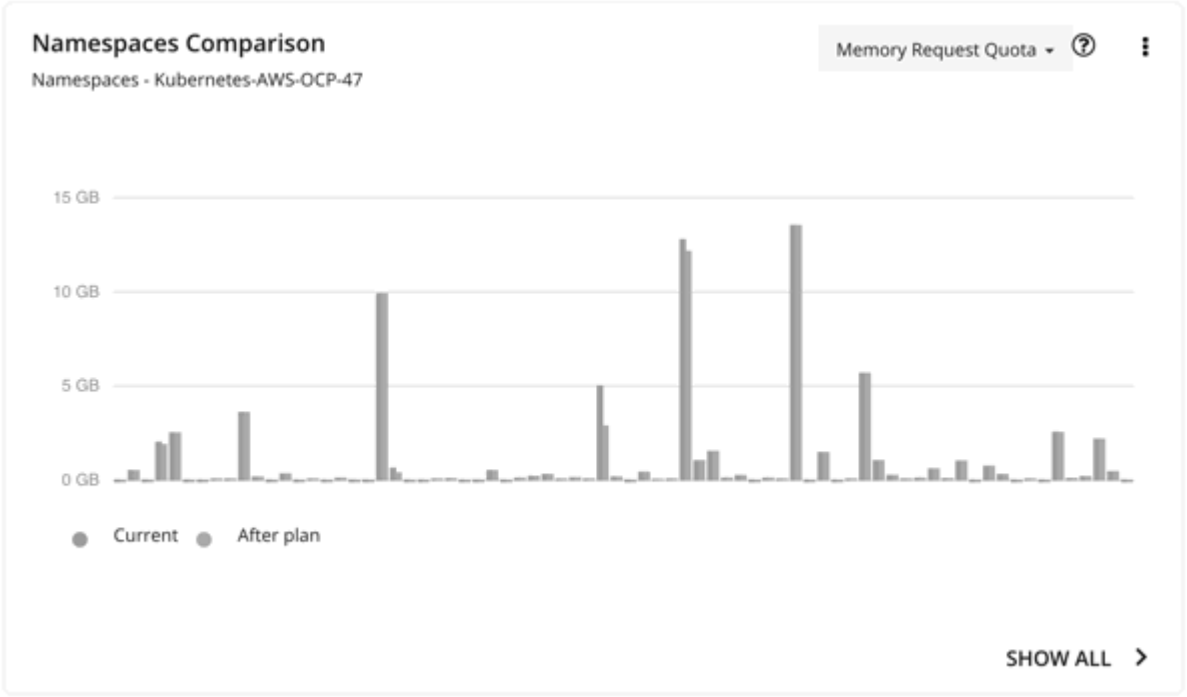
This chart shows pod utilization of resource quotas defined in namespaces. Resource quotas include:

- 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.

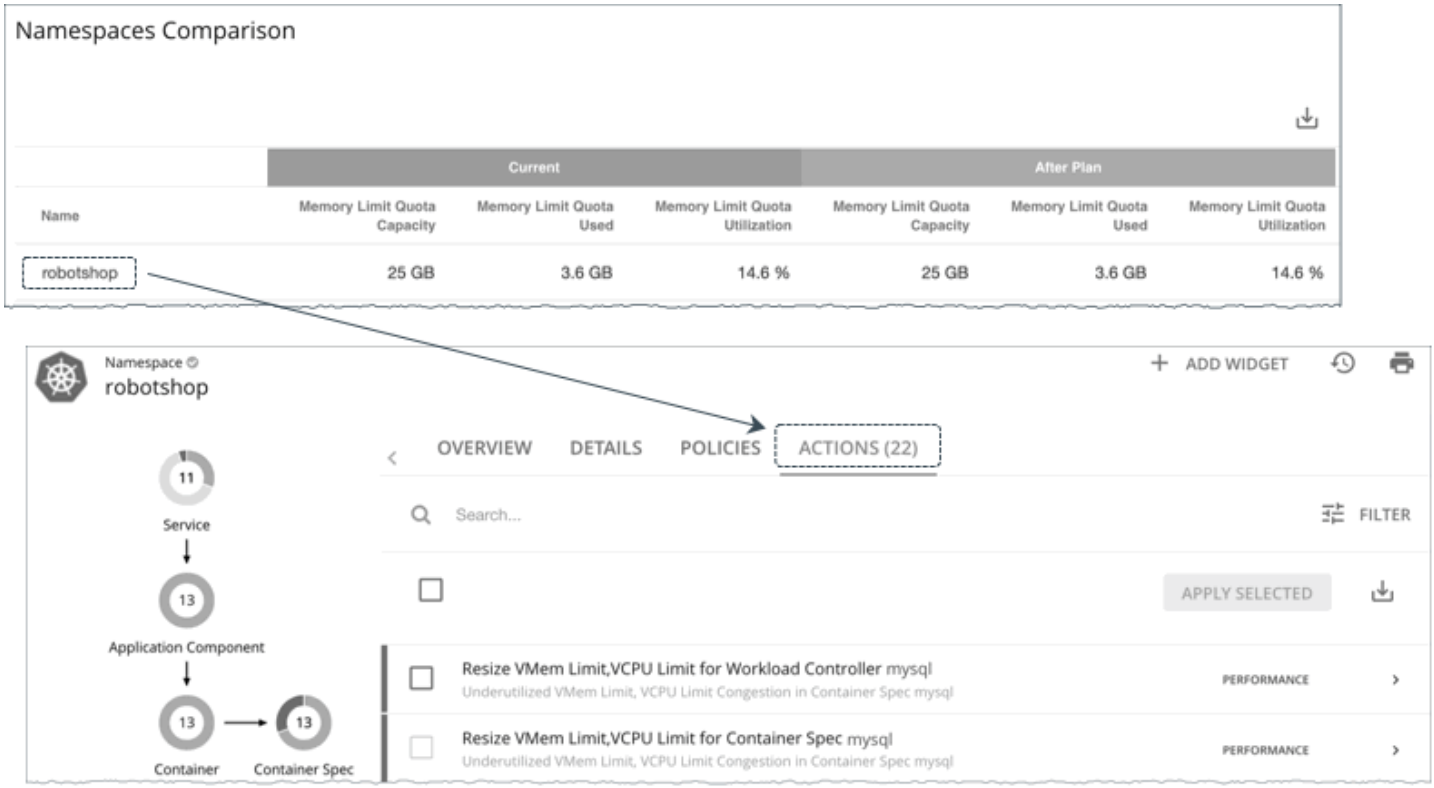
Namespaces Comparison



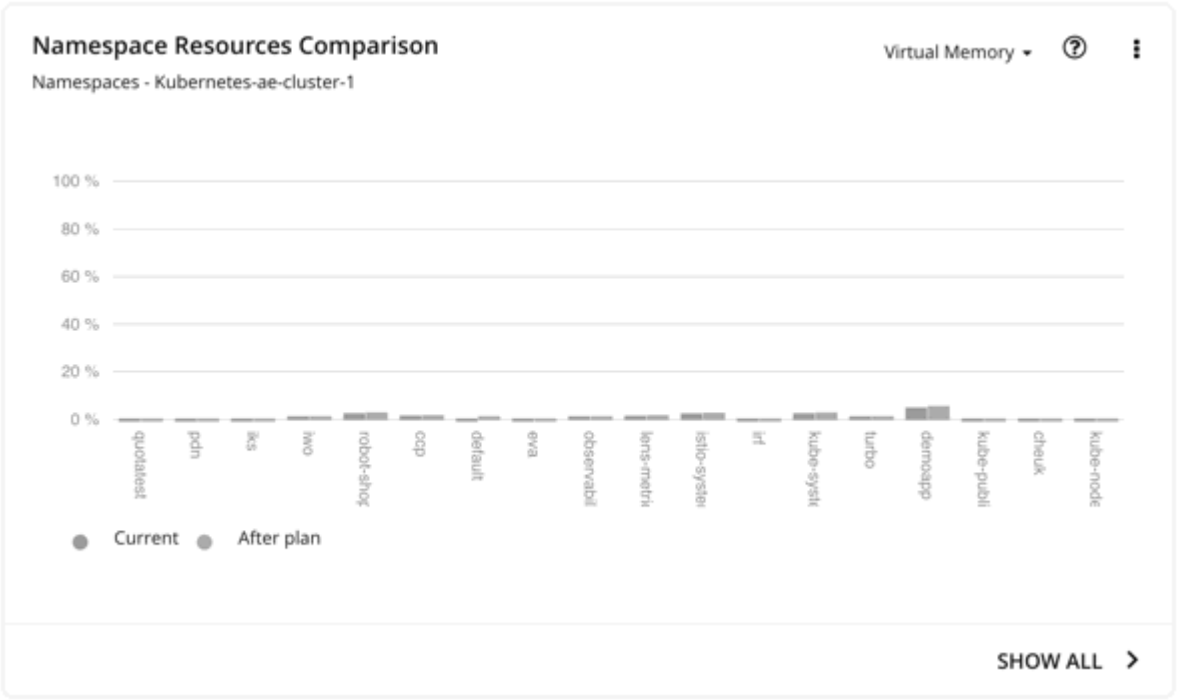
This chart compares namespace quota usage (one metric at a time) before and after the plan.

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.



Namespace Resources Comparison



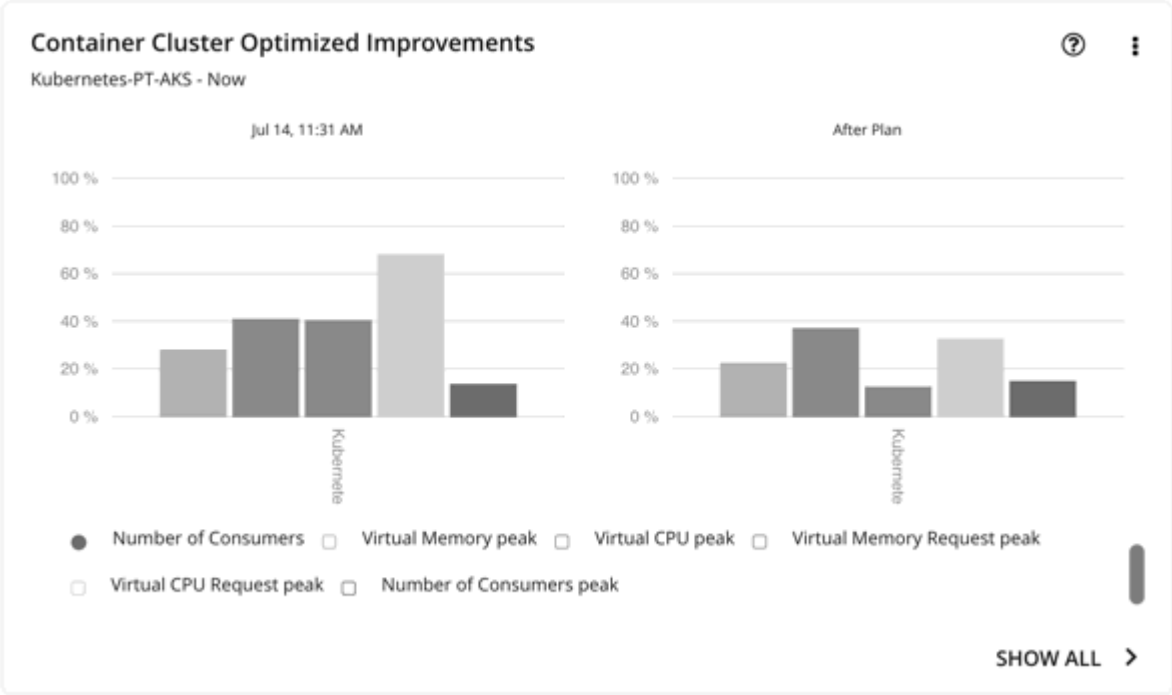
This chart shows how much cluster resources per namespace are utilized by pods. Utilization can be expressed as follows:

```
Utilization = Sum of actual vMem/vCPU used by pods / 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.

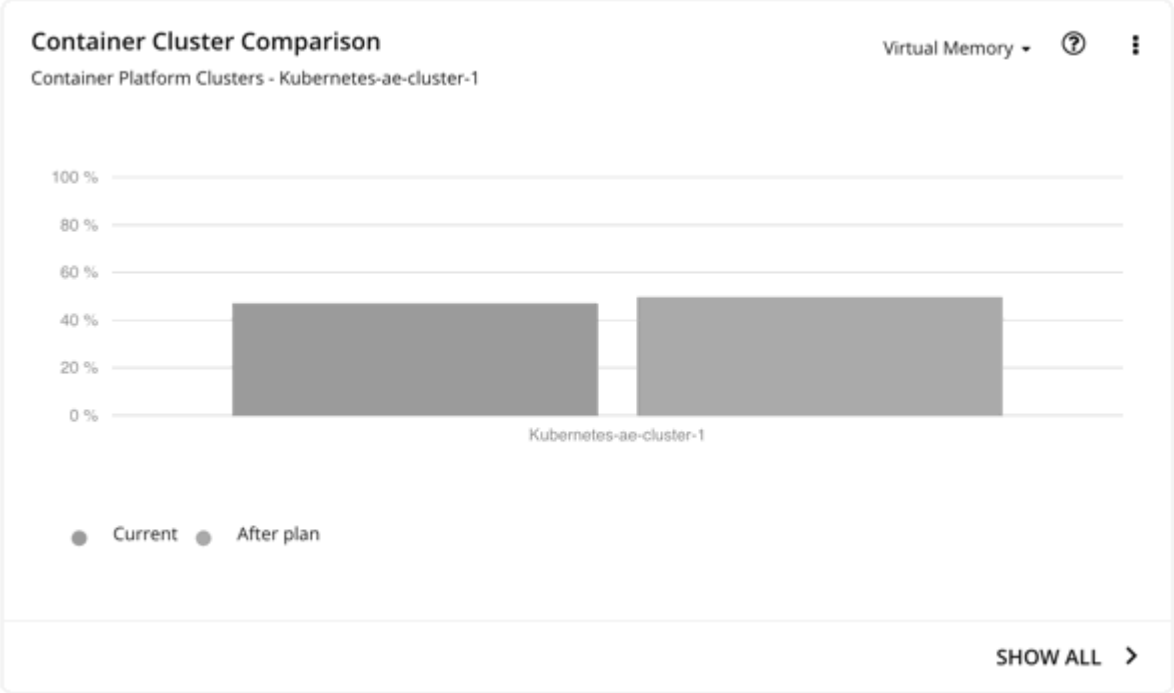
Container Cluster Optimized Improvements



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 on an on-prem Kubernetes cluster. These charts show how the utilization of resources would change assuming you accept all of the actions listed in the Plan Actions chart

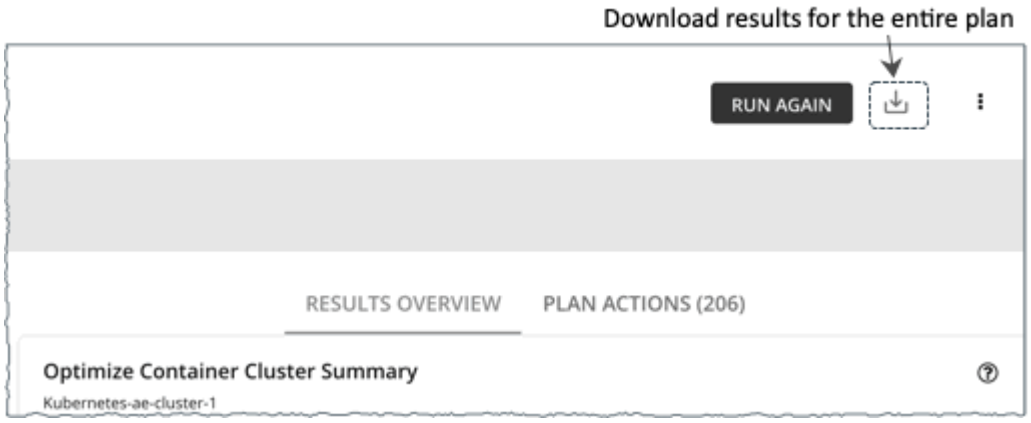
Hosts, Storage Devices, and Virtual Machines Comparison

Use these charts if you ran the plan on an on-prem Kubernetes 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.

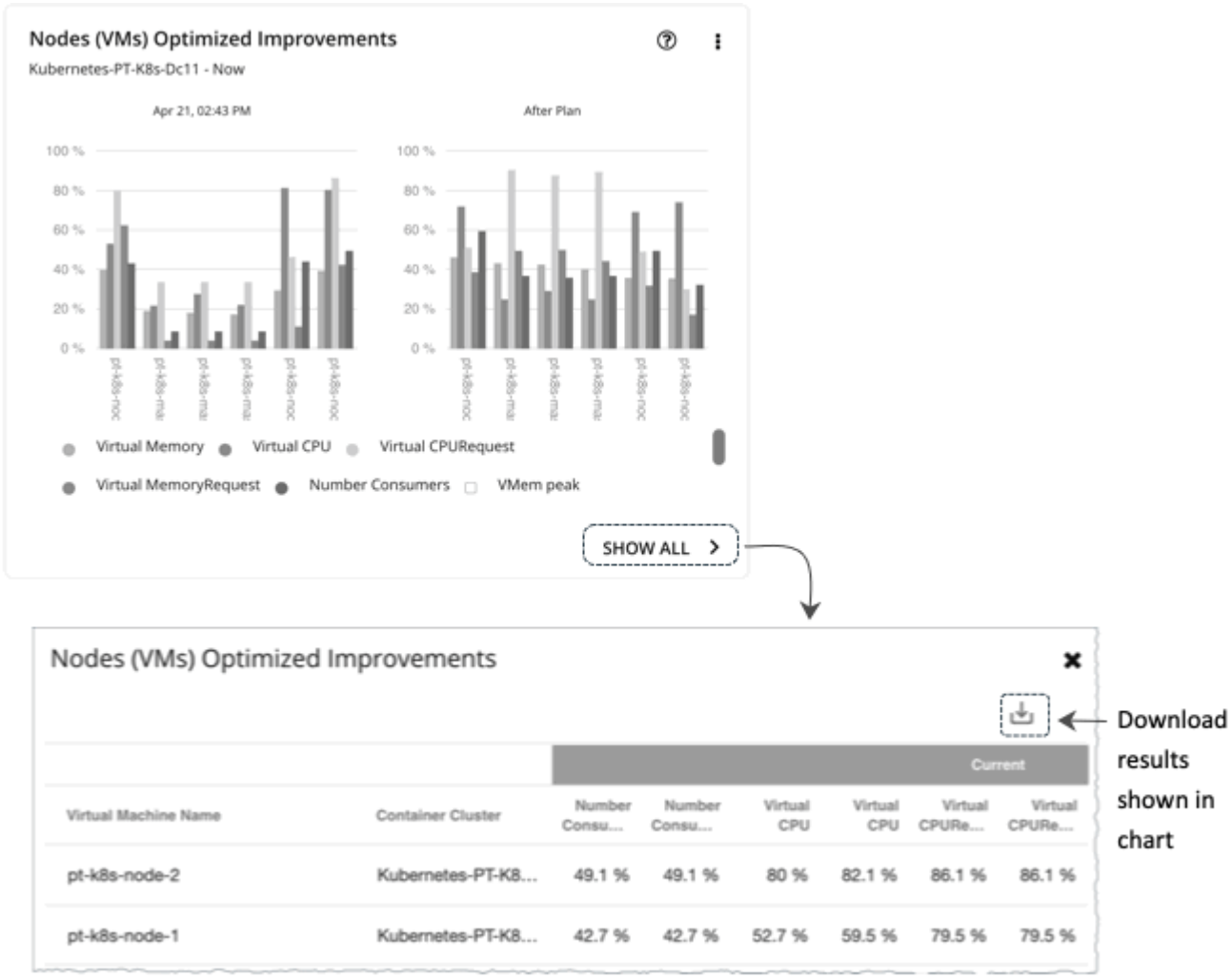
> **NOTE:**
> For the Storage Devices Comparison chart, if you set the view to **VM Per Storage** and click **Show all**, the total number of VMs sometimes does not match the number in the Summary chart. This happens if there are VMs in the plan that use multiple storage devices. The Storage Devices 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 container 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.



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.

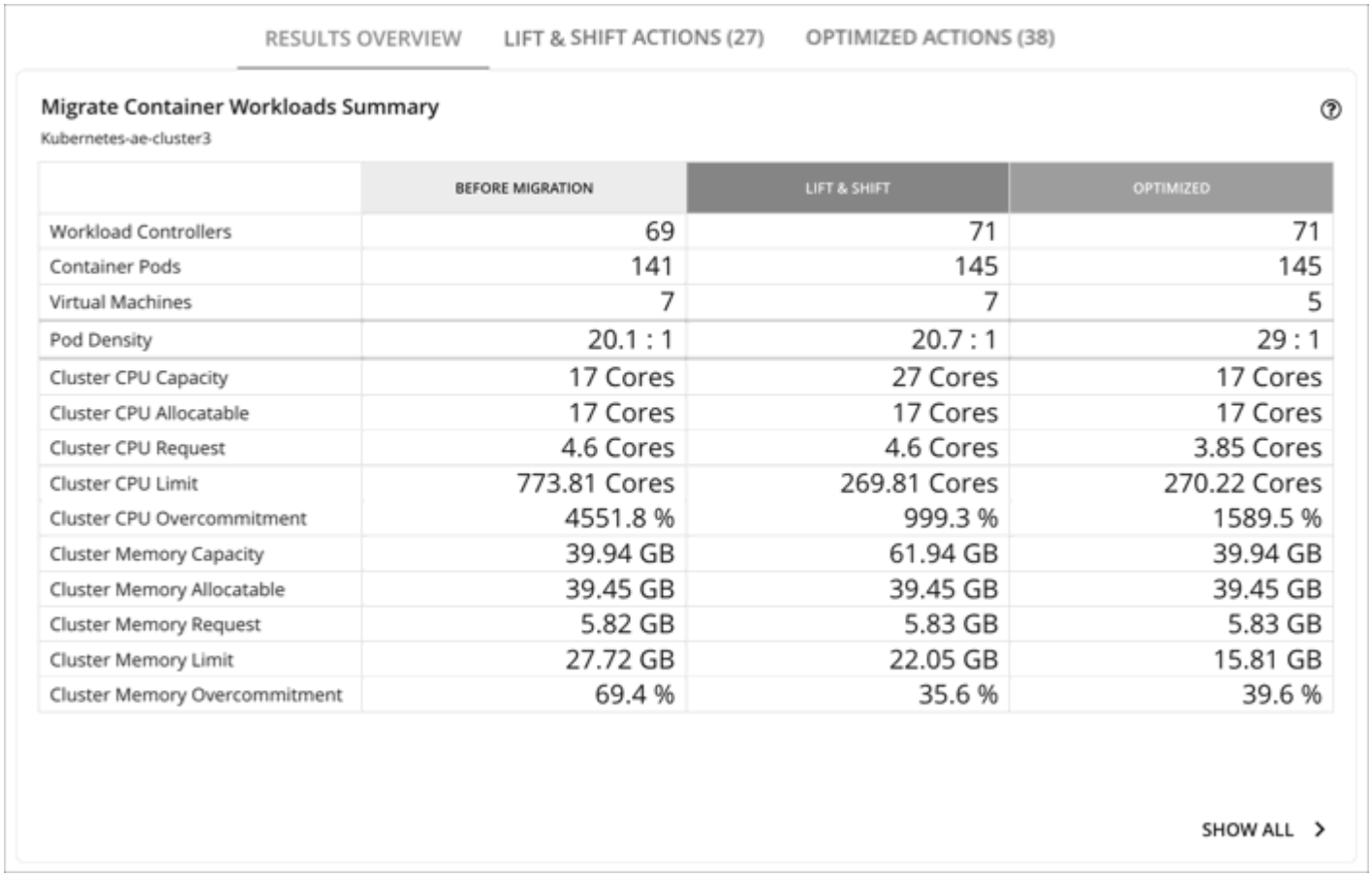
Migrate Container Workloads Plan

Run a Migrate Container Workloads plan to simulate the migration of container workloads from one 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 container cluster
- Container cluster in an on-prem or public cloud environment. For cloud environments, cost information is included in the plan results.
- Container cluster stitched to applications via Data Ingestion Framework (DIF)

The plan results show the impact of workload migration on the destination cluster, based on two migration scenarios:



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 >

- 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, Workload Optimization Manager 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**.

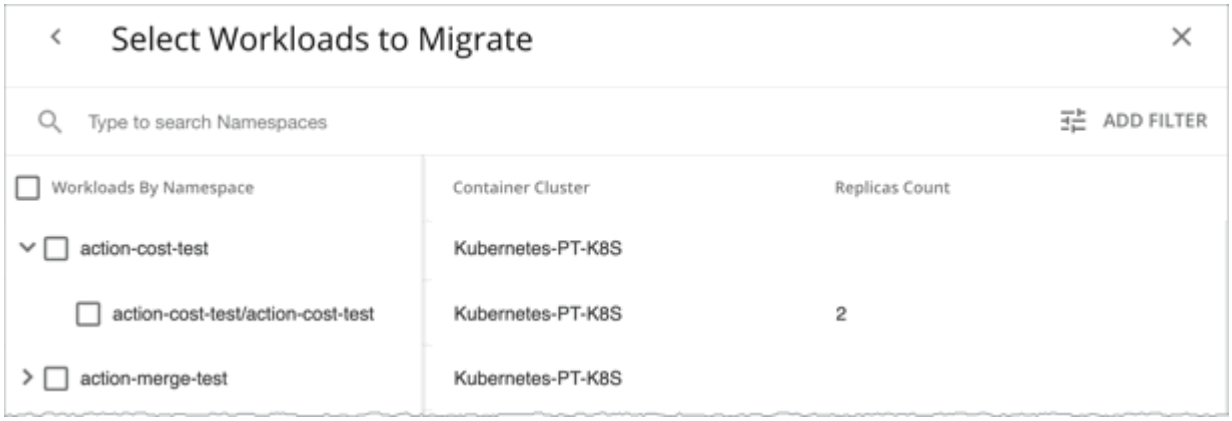
- From the supply chain

  To start a plan scenario from the supply chain, set the scope to a container 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 344)*.

1. Scope

Select the container workloads that you want to migrate.



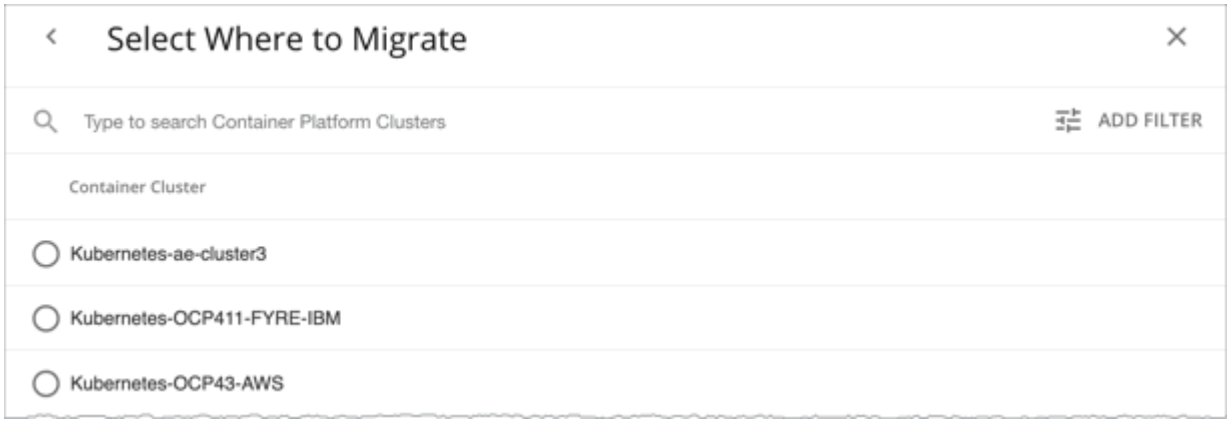
Container workloads are grouped by namespace. Expand a namespace to see individual Workload Controllers that manage workloads, and the number of container replicas for each Workload Controller. You can adjust the number of replicas in a later step. For example, you can add replicas to meet workload demand.

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 cluster, namespace, tag, or Workload Controller.

2. Where to Migrate

Choose the destination cluster for the container workloads you selected.



You can migrate to a:

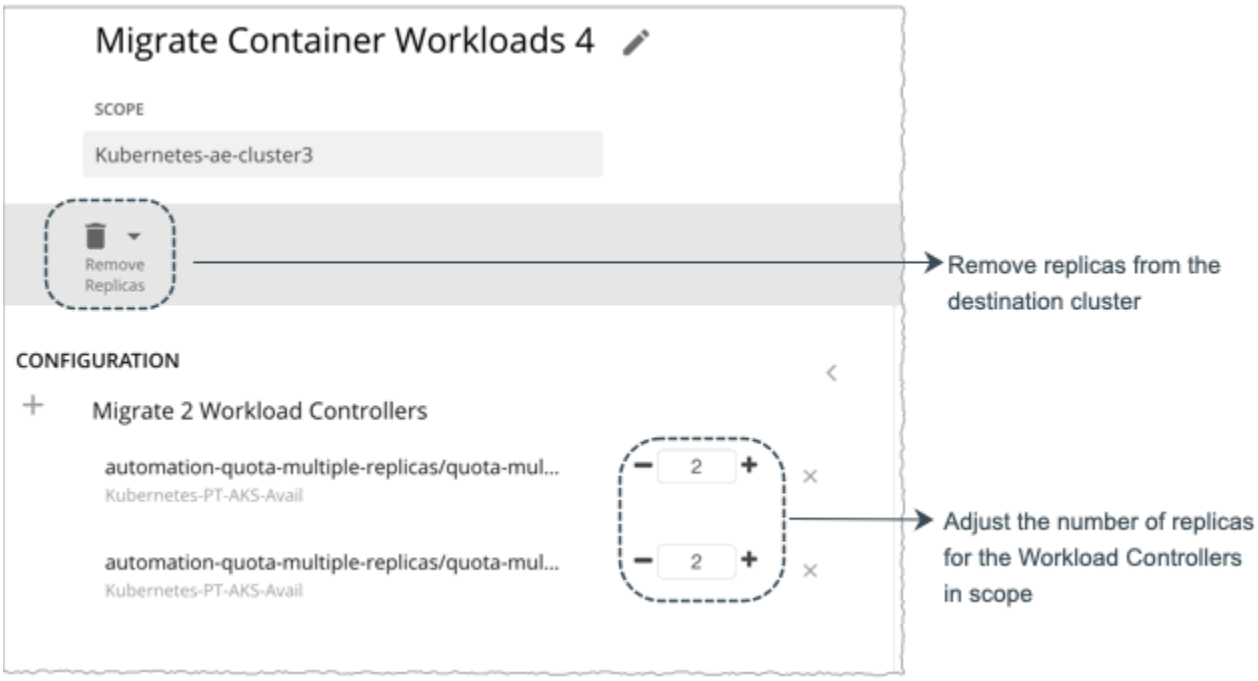
- Standalone container cluster
- Container cluster in an on-prem or public cloud environment
- Container cluster stitched to applications via Data Ingestion Framework (DIF)

After selecting a cluster, you can:

- Run the plan.

  Or

- Choose **Skip to Configuration** to configure additional settings, including:



  - Removing existing replicas from the destination cluster. You may need to do this to make room for the workloads you are migrating.
  - 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

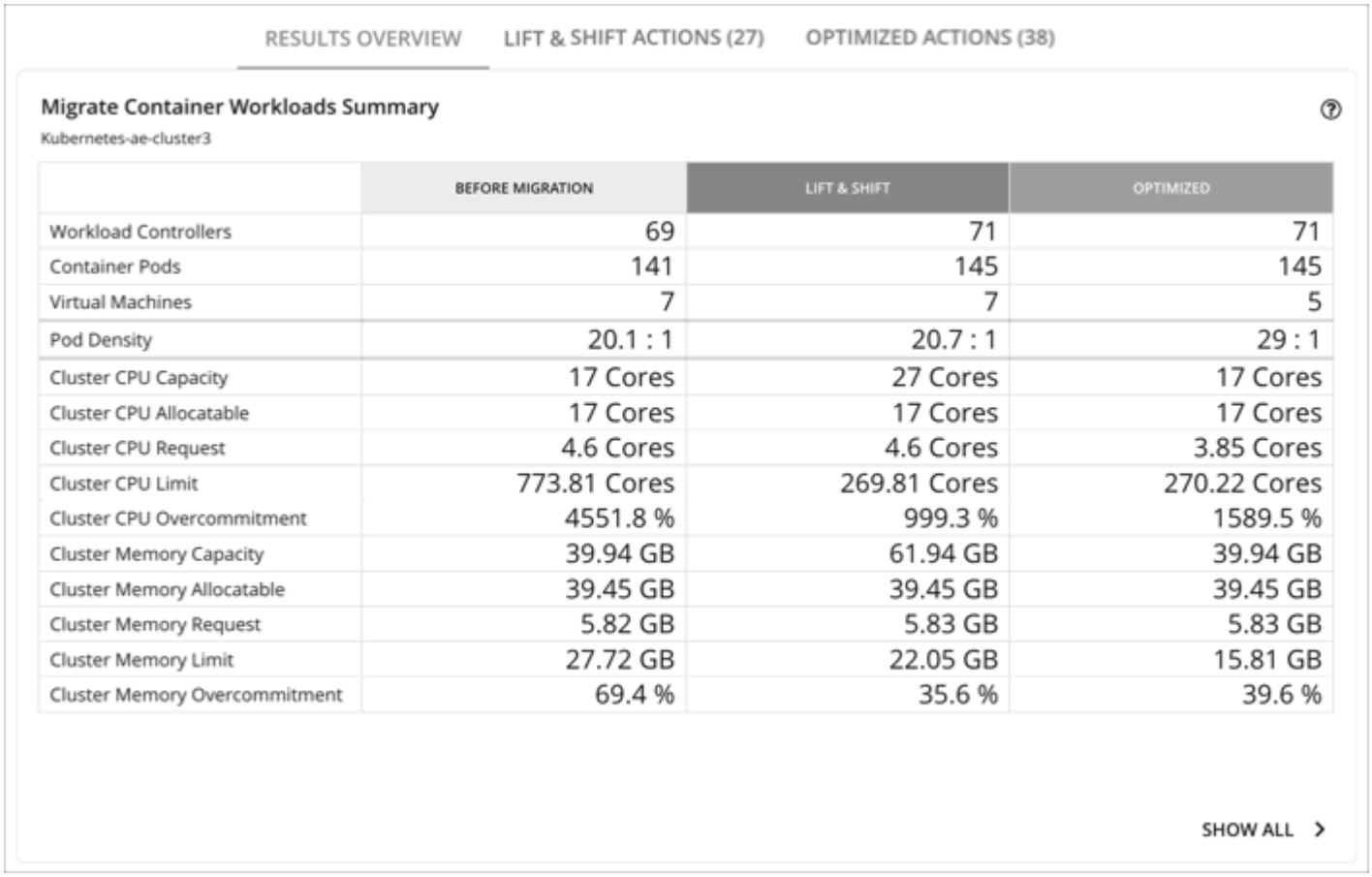
After the plan runs, you can view the results to see the impact of workload migration on the *destination* cluster. The results do *not* show data for the *source* cluster (i.e., the cluster that currently hosts the workloads).

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 Kubernetes 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



	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 %

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, Workload Optimization Manager 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

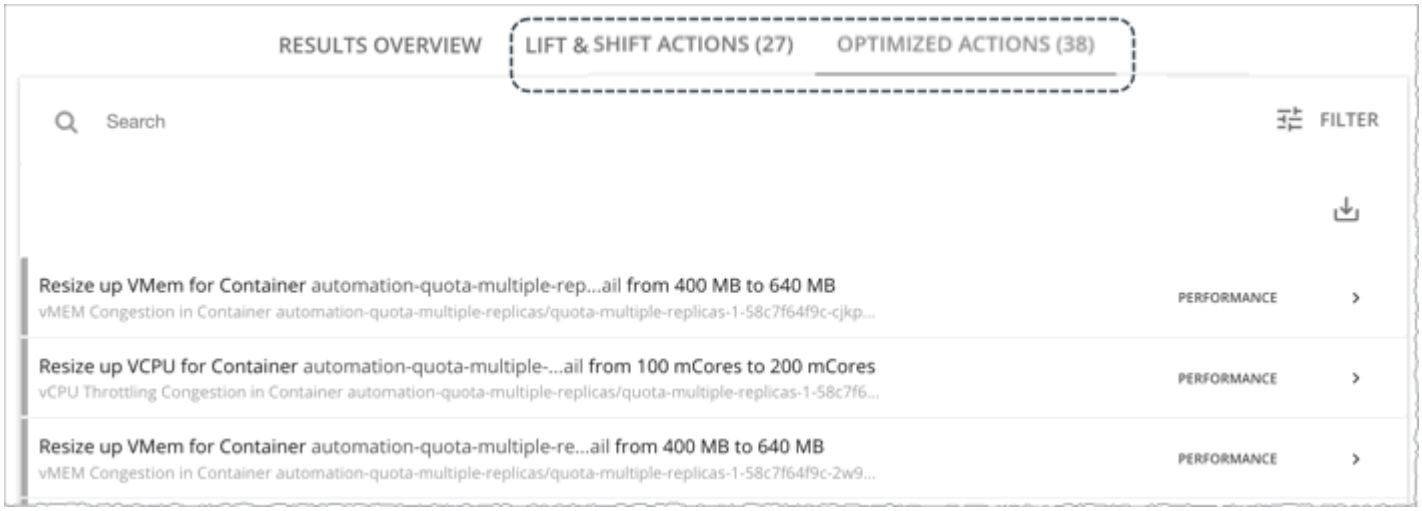
Table rows show the following information for the destination cluster:

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 Kubernetes 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. 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 ▪ 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	(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. Workload Optimization Manager manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk. Workload Optimization Manager only calculates overcommitment in plans. The calculation can be expressed as: `Overcommitment = Sum of CPU limits for all containers / 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.

Row	Description
Cluster Memory Request	Total Memory Request capacity for the cluster.
Cluster Memory Limit	Total Memory Limit capacity for the cluster.
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. Workload Optimization Manager manages cluster resources by actual utilization and limit rightsizing so that you can run more workloads with less risk. Workload Optimization Manager only calculates overcommitment in plans. The calculation can be expressed as: `Overcommitment = Sum of memory limits for all containers / Sum of memory capacity for all nodes`

Plan Actions

Workload Optimization Manager shows separate tabs for **Lift & Shift** and **Optimized** migration actions. You can download the list of actions as a CSV file.



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.

Workload Optimization Manager can recommend the following actions:

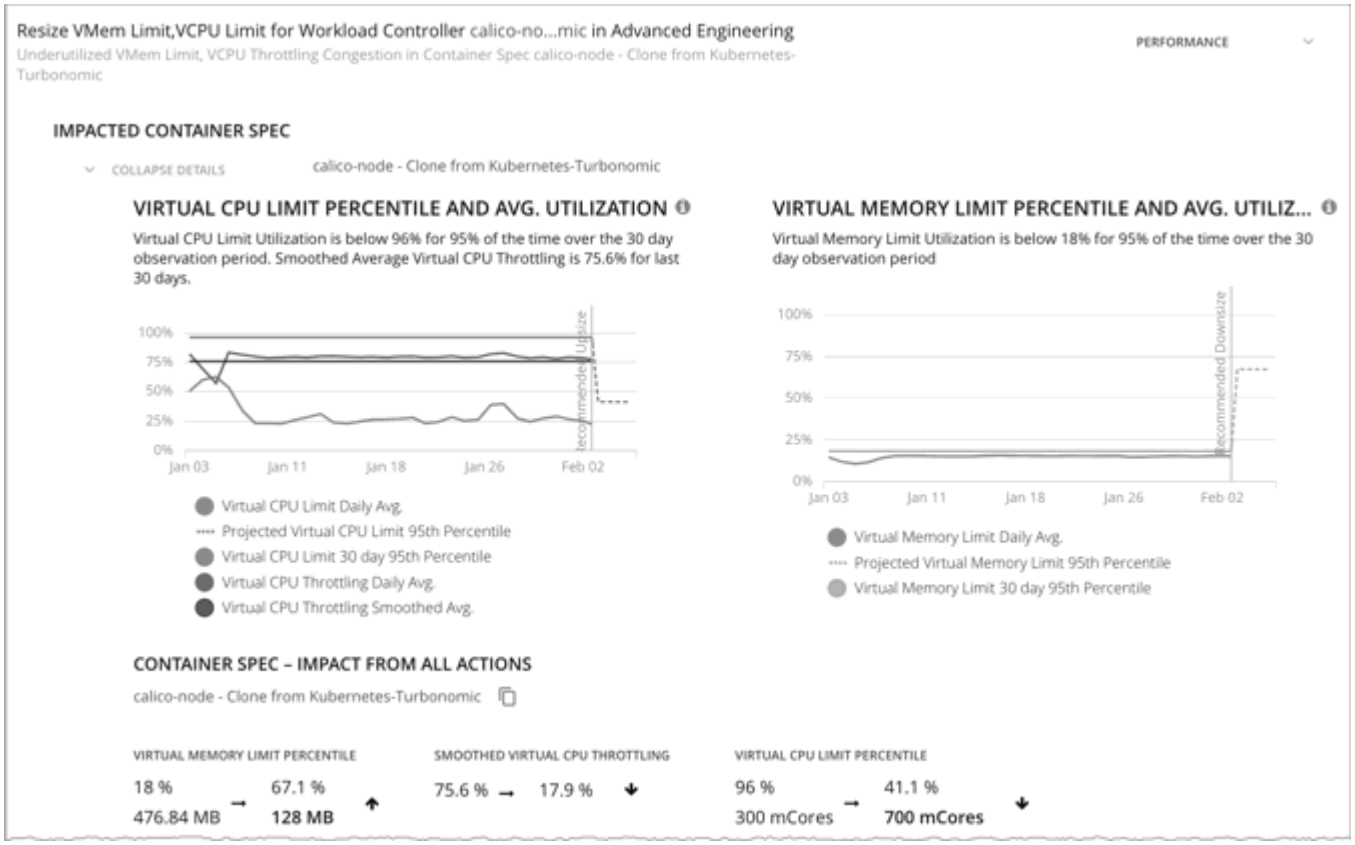
Entity	Lift & Shift Actions	Optimized Actions
Workload Controller	None	Resize container limits and requests

Entity	Lift & Shift Actions	Optimized Actions
		See "Workload Controller Resize Actions" below for additional details.
Namespace	None	Resize quotas (if container resize actions would exceed the namespace quotas)
Pod	▪ Move ▪ Provision (if node provision is required) ▪ Reconfigure	▪ Move ▪ Provision (if node provision is required) ▪ Suspend (if node suspension is required) ▪ Reconfigure
Node (VM)	▪ Provision ▪ Move on-prem nodes	▪ Provision ▪ Suspend ▪ Move on-prem nodes
Cloud volume	Scale	Scale
On-prem storage	▪ Provision ▪ Suspend	▪ Provision ▪ Suspend
On-prem host	▪ Provision ▪ Suspend	▪ 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 Workload Optimization Manager analyzed to make accurate resize decisions.



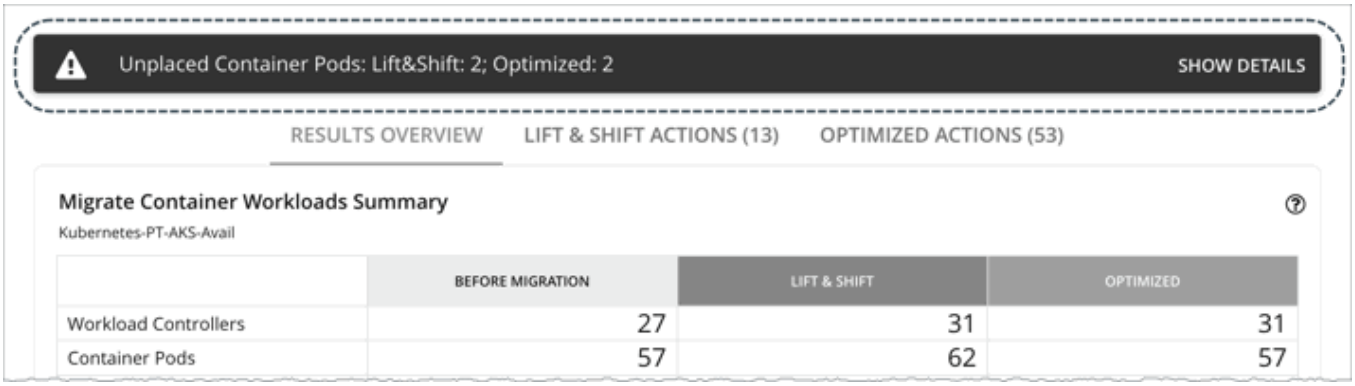
For more information about these charts, see Utilization Charts *(on page 84)*.

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, Workload Optimization Manager 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.

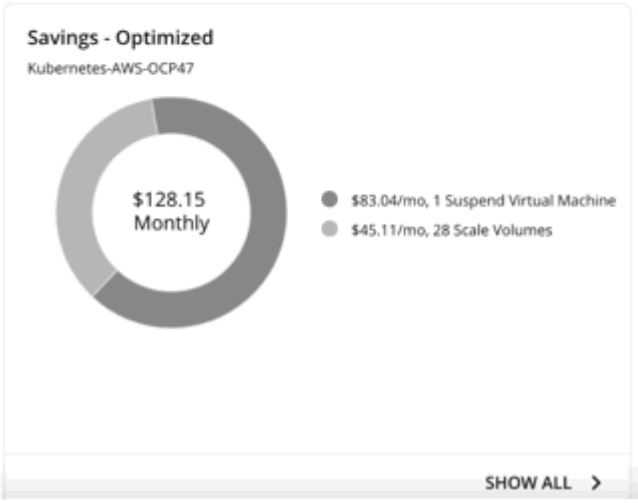
  The plan always enforces affinity and anti-affinity constraints, which could result in unplaced pods in the destination cluster. If this happens, the plan generates reconfigure actions for the unplaced pods and shows a notification in the plan results.



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, Workload Optimization Manager 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.

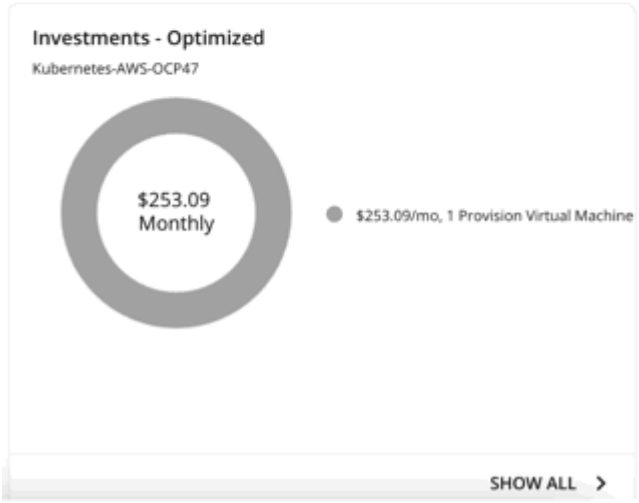
Note that application performance and infrastructure efficiency are the drivers of these actions, *not* cost optimization. Cost information is included to help you track your cloud expenses.

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, Workload Optimization Manager 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.



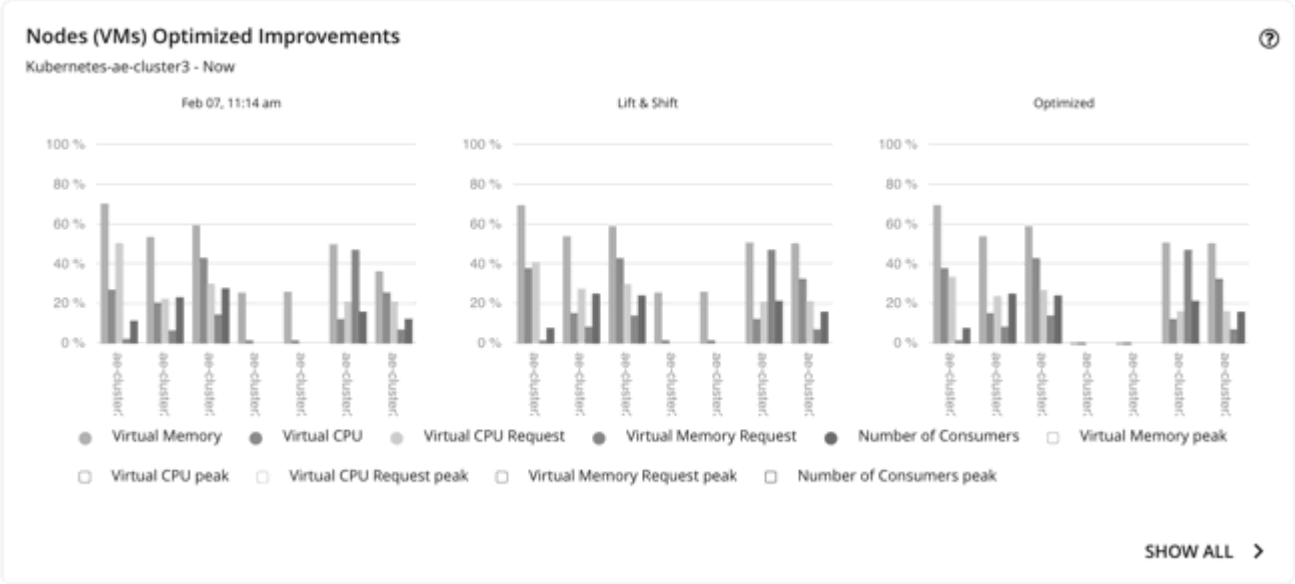
For example, if some applications risk losing performance, the plan might recommend provisioning nodes to increase capacity. The charts show how node provision actions translate to an increase in expenditure. You can use cost information in the charts as you seek approval to execute these actions.

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:

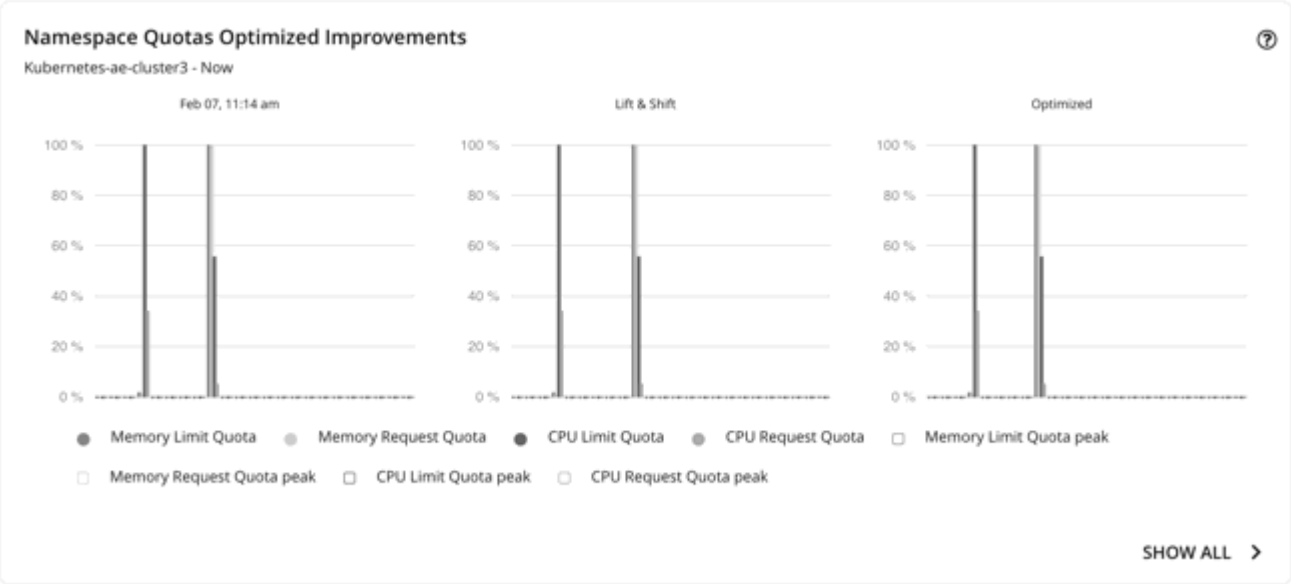


- Utilization of the following resources for all nodes:
  - vMem
  - vCPU
  - vMem Request
  - vCPU Request
- Number of pods consuming resources against the maximum pod capacity for all the nodes

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:



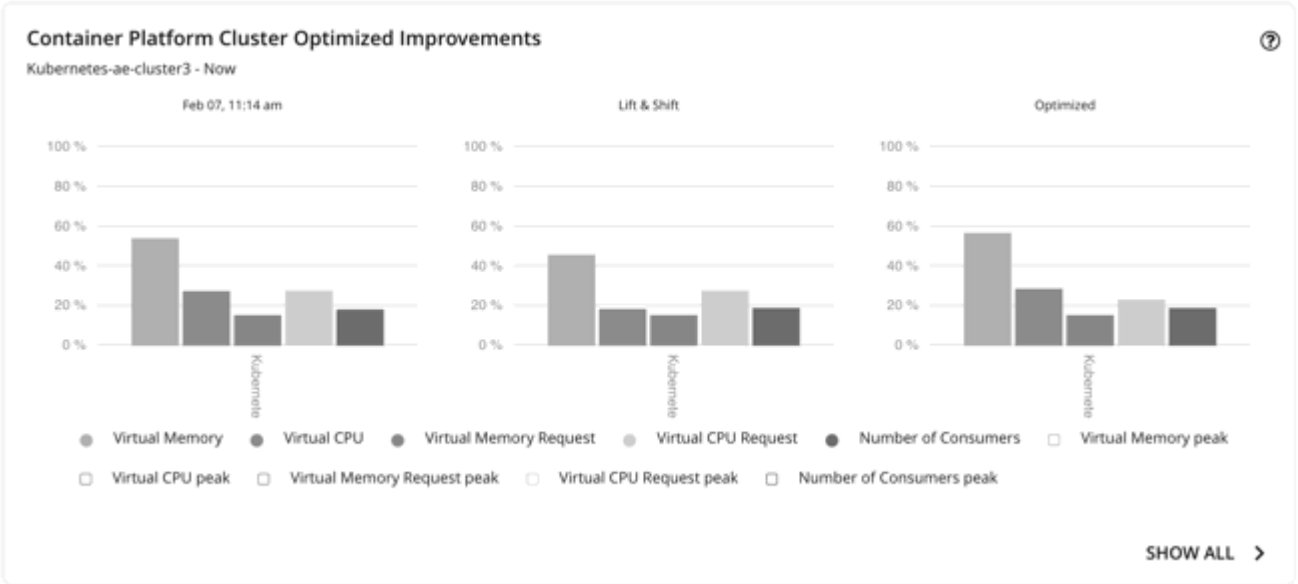
- CPU Limit Quota
- Memory Limit Quota
- CPU Request Quota
- Memory Request Quota

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.

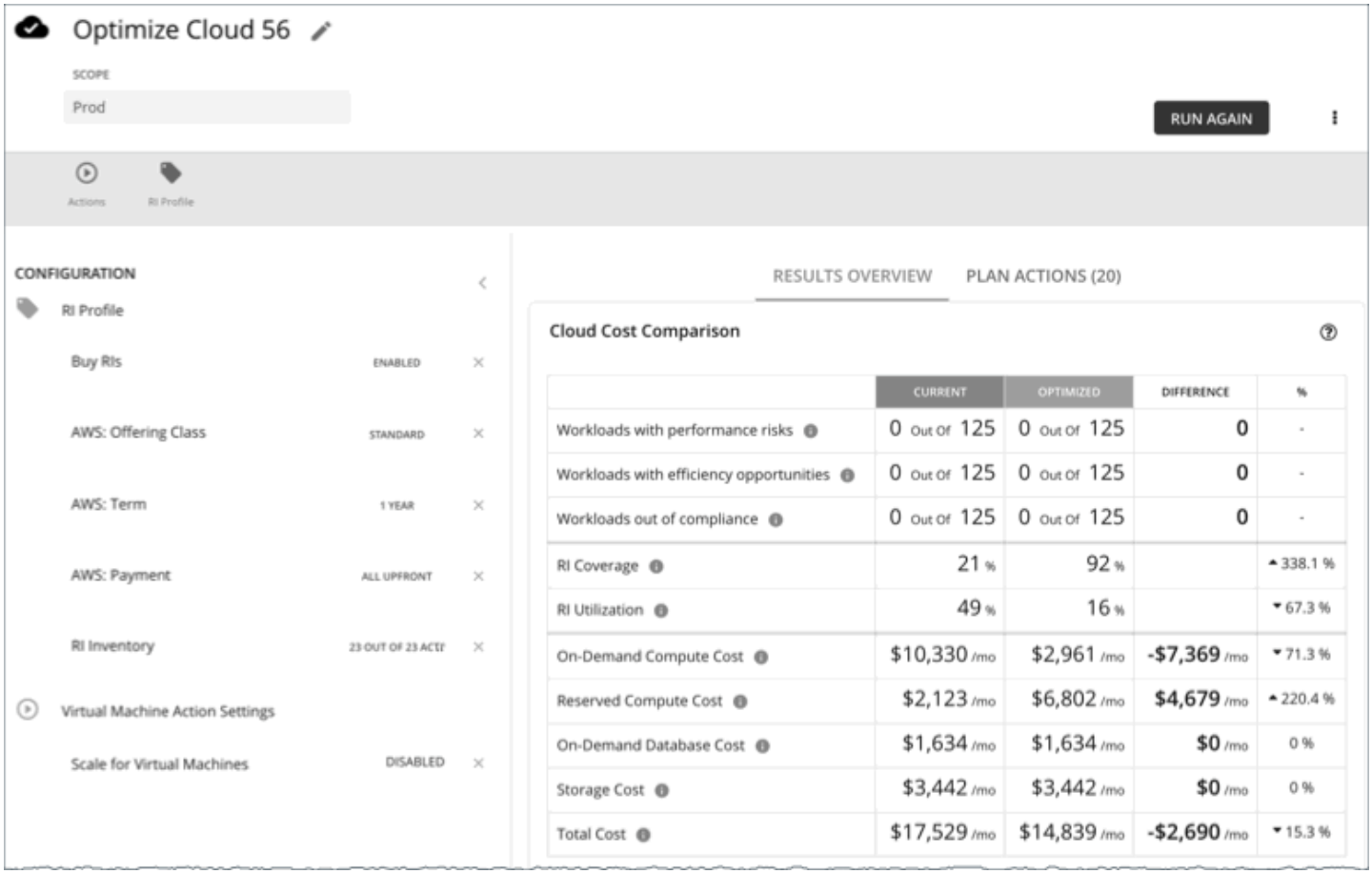
Use the toolbar on top of the Configuration section to change the configuration settings.

> **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.

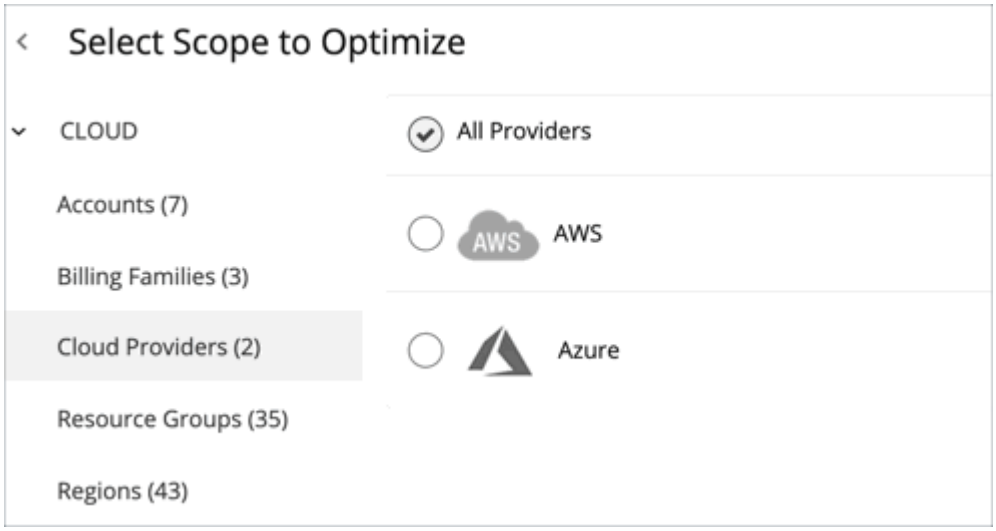


Configuring an Optimize Cloud Plan

For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.

1. Scope

You can scope by:



- Accounts

  Choose an AWS account or Azure subscription for the plan's scope. If you choose an Account for the scope, then the plan will not recommend discount purchases. To optimize discount purchases for a limited scope, choose a Billing Family.
- Billing Families

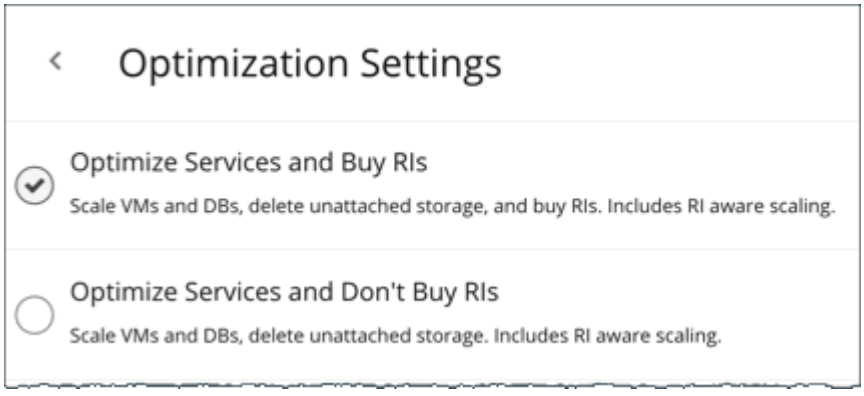
  Include discount purchases in the planning for a scope that is limited to a single billing family. The plan calculates discount purchases through the billing family's master account.
- Cloud Providers

  See how you can optimize all your AWS or Azure workloads.
- Resource Groups

  Workload Optimization Manager 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



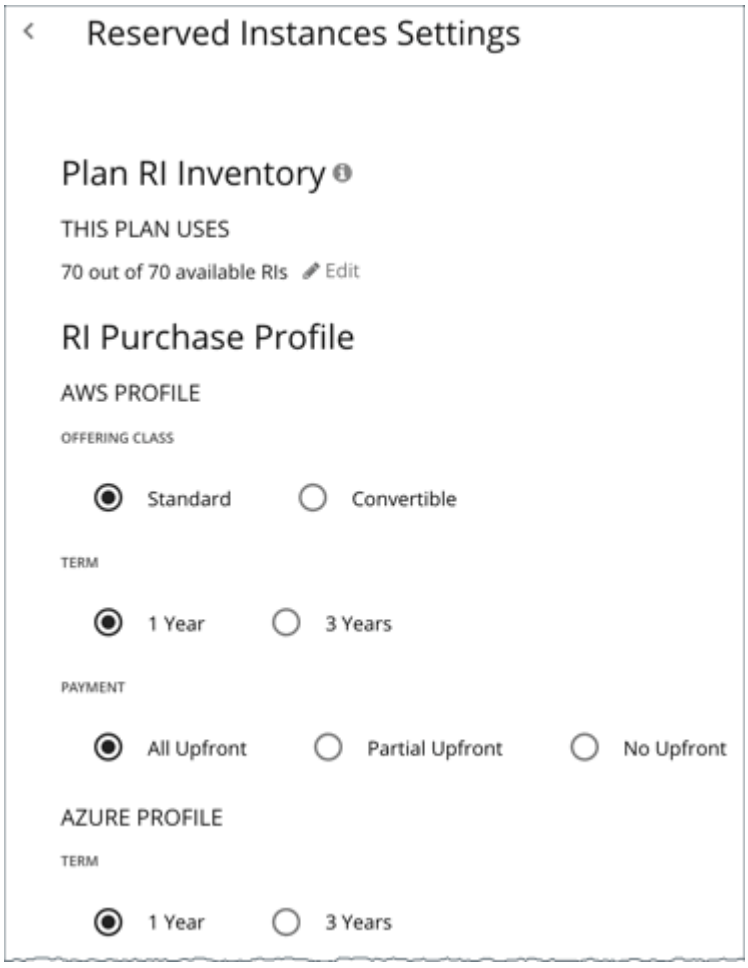
Choose from the given optimization options. Note that if you set a plan's scope to a resource group, Workload Optimization Manager will optimize services without recommending discount purchases.

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 403)*.

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 Workload Optimization Manager a week to reflect accurate results in Optimize Cloud plans.

3. Reserved Instances Settings



For **Plan RI Inventory**, the discounts for the current scope are selected by default. Click **Edit** to make changes.

For **RI Purchase 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.

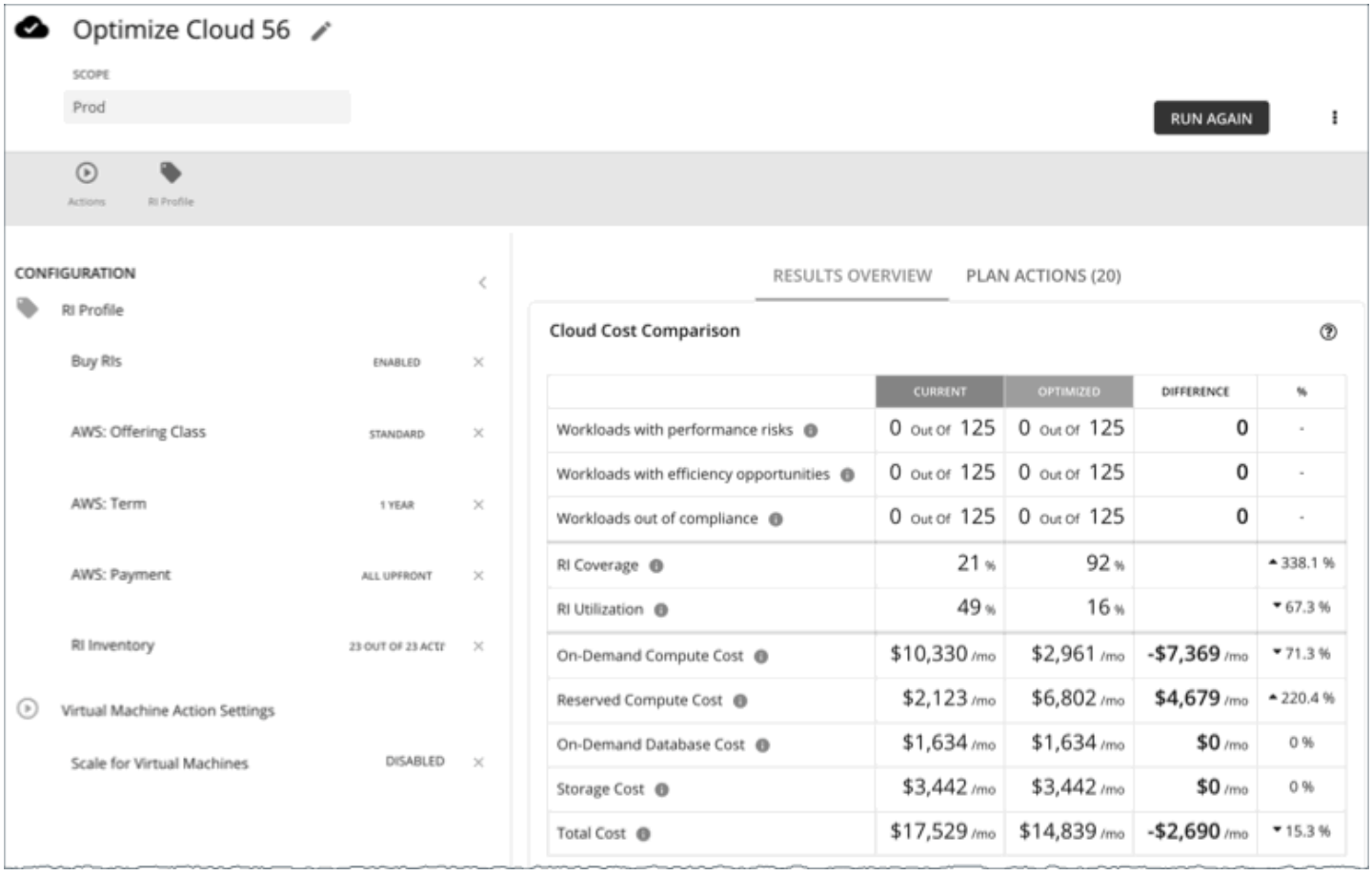
Working With Optimize Cloud Plan Results

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 templates used
- Compare breakdowns of storage tiers in use
- 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 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.

  Workload Optimization Manager 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, Workload Optimization Manager can get the information it needs to display license costs for database instances. For Azure clouds, Workload Optimization Manager 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.

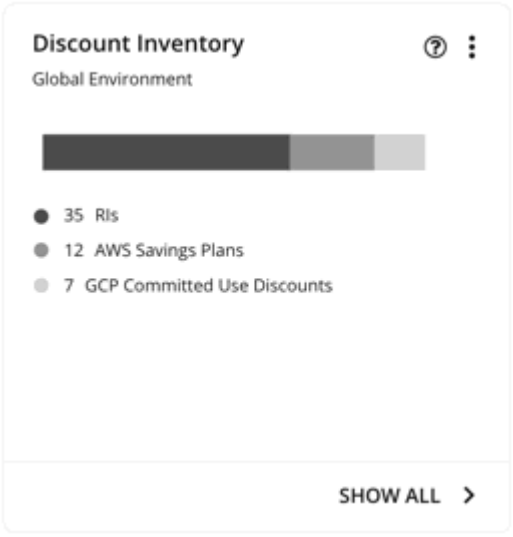
- **Volume Tier Summary**

  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.

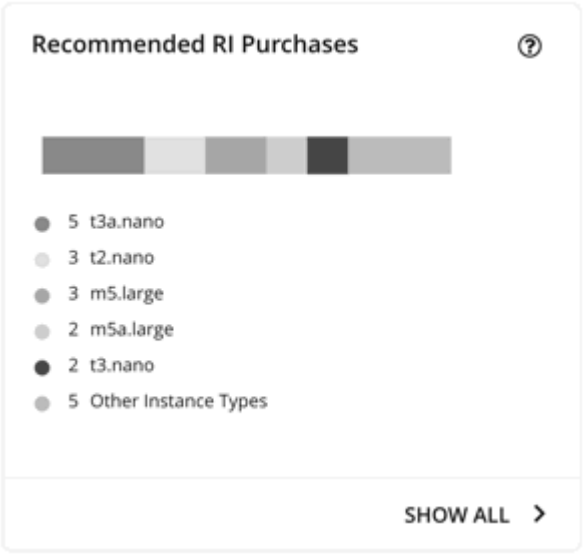


- **Recommended RI Purchases**

  Workload Optimization Manager 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 403)*.

Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.



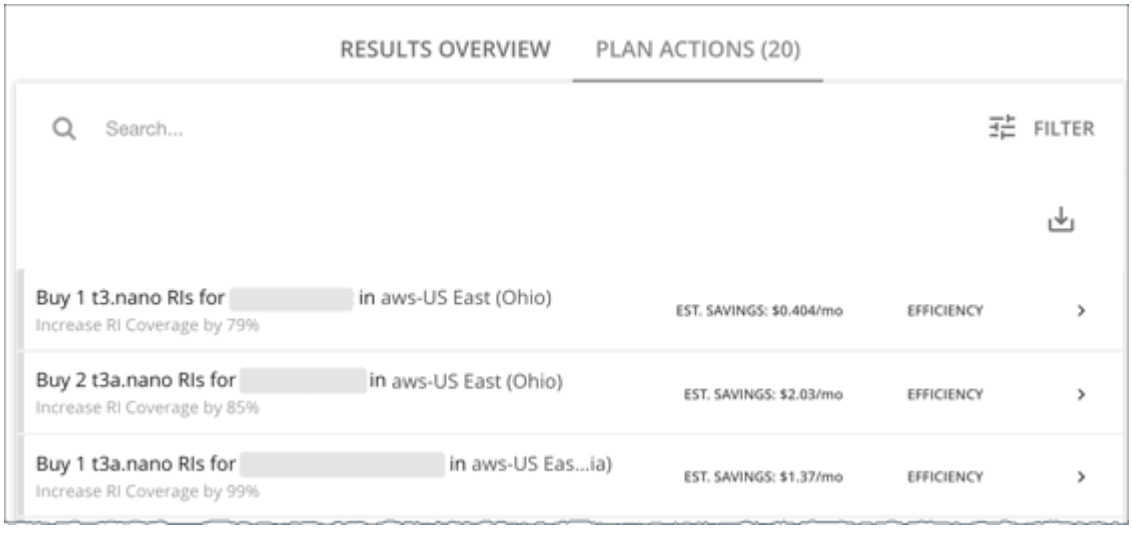
Click **Show All** to see a table with details for each discount.

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

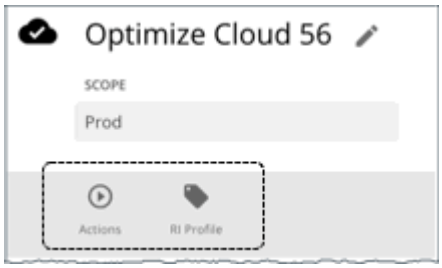
Click the **Plan Actions** tab on top of the page to view a list of actions that you need to execute to achieve the plan results.



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.



- Actions

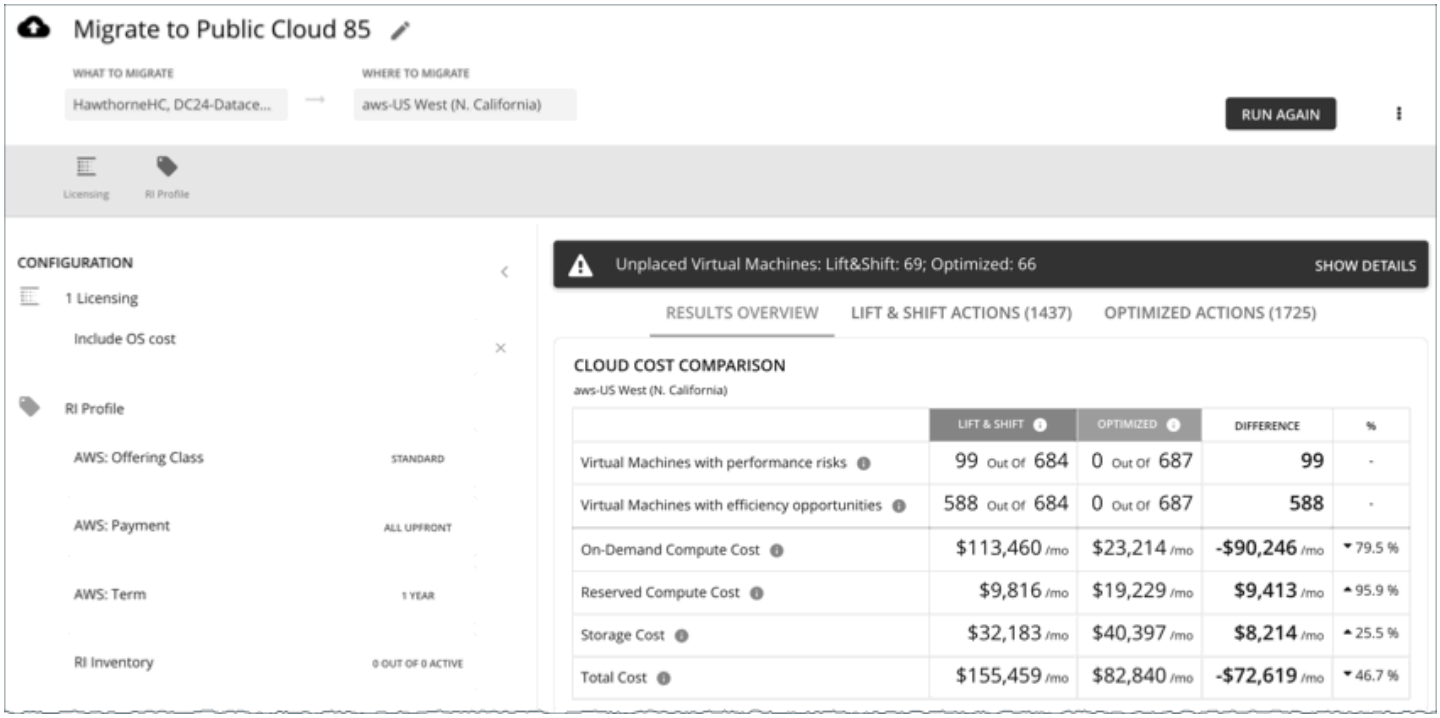
  Use this to enable or disable automatic Scale actions for the virtual machines in the plan.
- RI Settings

  See Reserved Instance Settings *(on page 387)*.

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



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 Workload Optimization Manager 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, Workload Optimization Manager considers the VM's history. For example, if the VM has been stable for 16 of the last 21 days, then Workload Optimization Manager will plan for that VM to use a discount. In this way, the plan calculates the best cost for the migration. However, if Workload Optimization Manager 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 Workload Optimization Manager 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 95)*.

The plan results show:

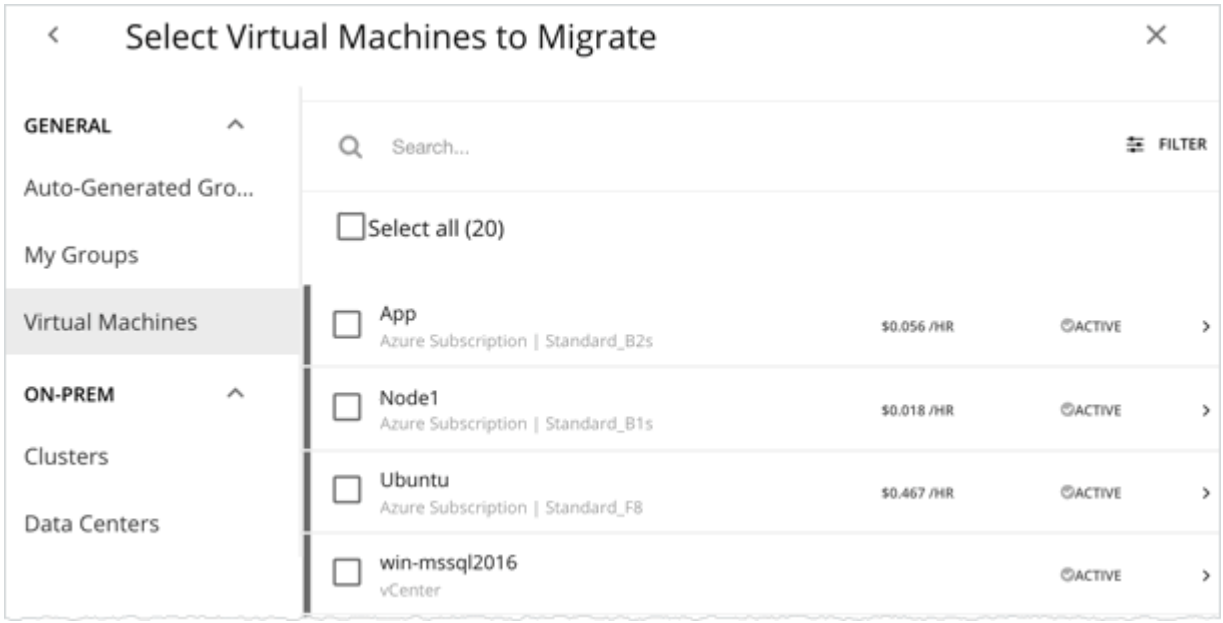
- Projected costs
- Actions to execute your migration and optimize costs and performance
- Optimal cloud instances to use, combining efficient purchase of resources with assured application performance
- The cost benefit of moving workloads from on-demand to discounted pricing
- Discounts you should purchase

Configuring a Migrate to Cloud Plan

For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.

1. Scope

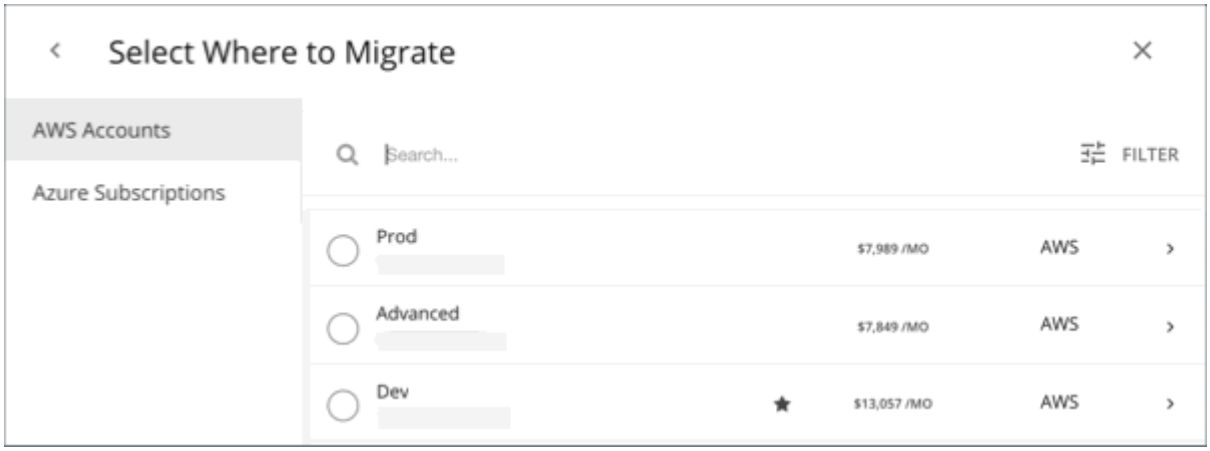
Select the VMs that you want to migrate. You can select VM groups and/or individual VMs.



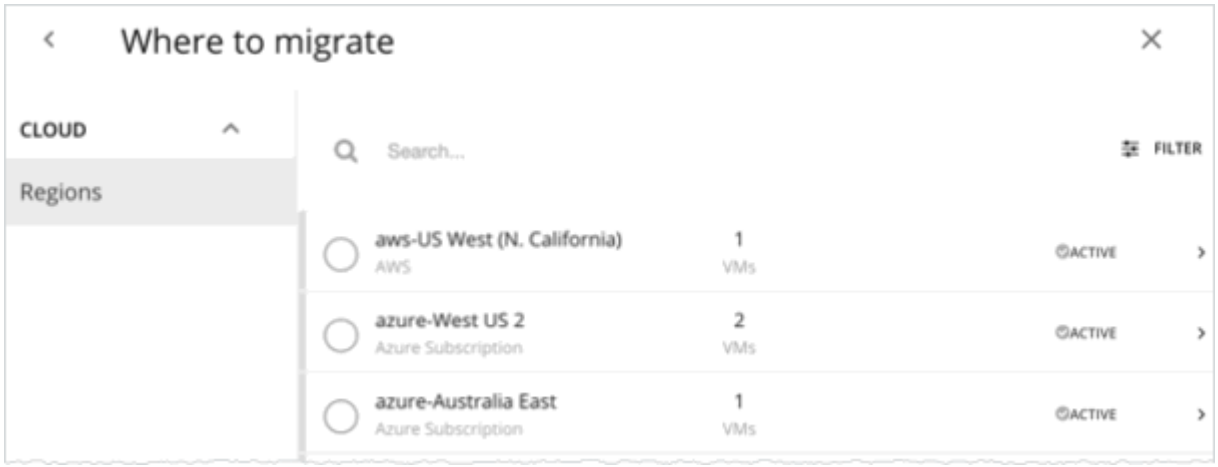
If you select an Auto Scaling Group, Workload Optimization Manager simulates migrating the VMs individually, and not as a group.

2. Where to Migrate

Choose a billing account (AWS account or Azure subscription).



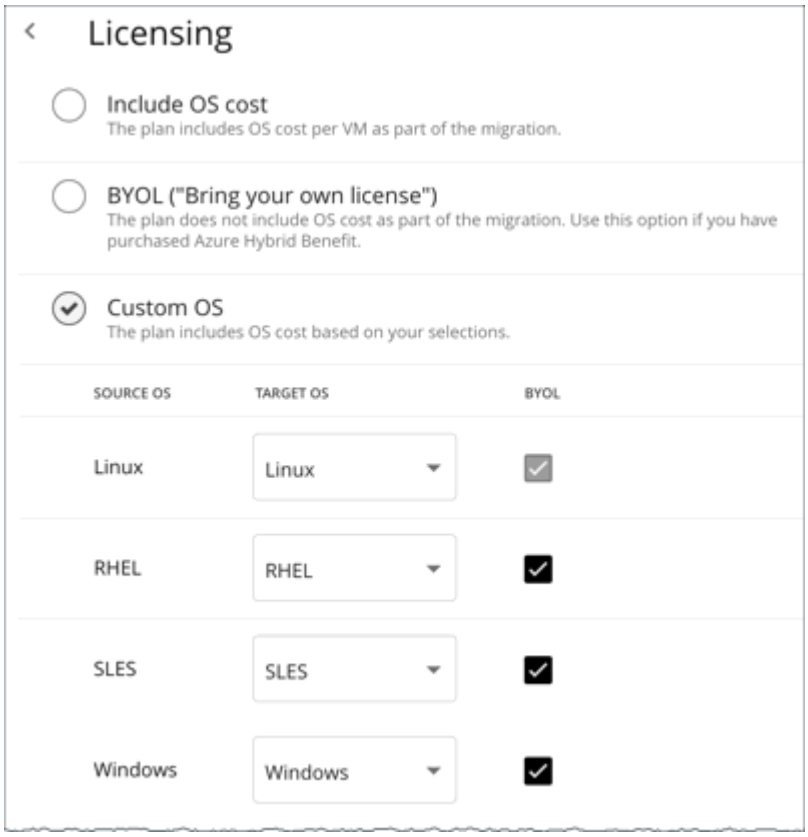
Choose a region. Workload Optimization Manager shows all the regions that you can access from your target cloud accounts.



By default, Workload Optimization Manager 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, Workload Optimization Manager will only consider the instance types defined in the policies.

3. Licensing (OS Migration Profile)

Select an OS Profile for this migration.



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 Workload Optimization Manager 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.

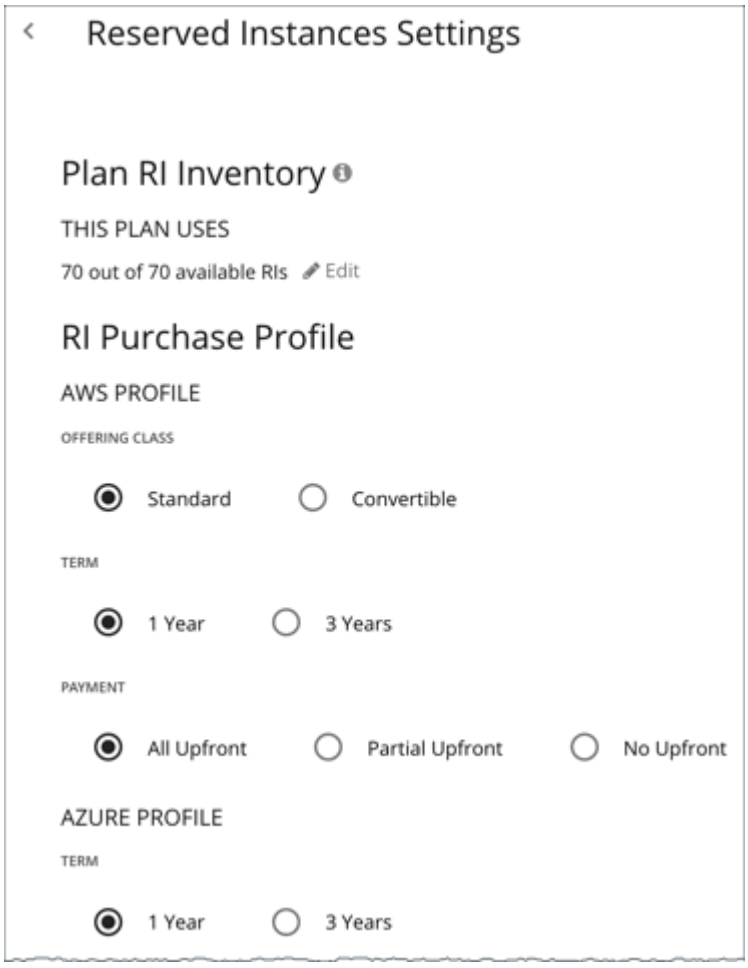
- Custom OS

  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, Workload Optimization Manager 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, Workload Optimization Manager 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**, Workload Optimization Manager gets the license cost from the service provider and includes that cost in the plan results.

4. Reserved Instances Settings



For **Plan RI Inventory**, the discounts for the current scope are selected by default. Click **Edit** to make changes.

For **RI Purchase 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:

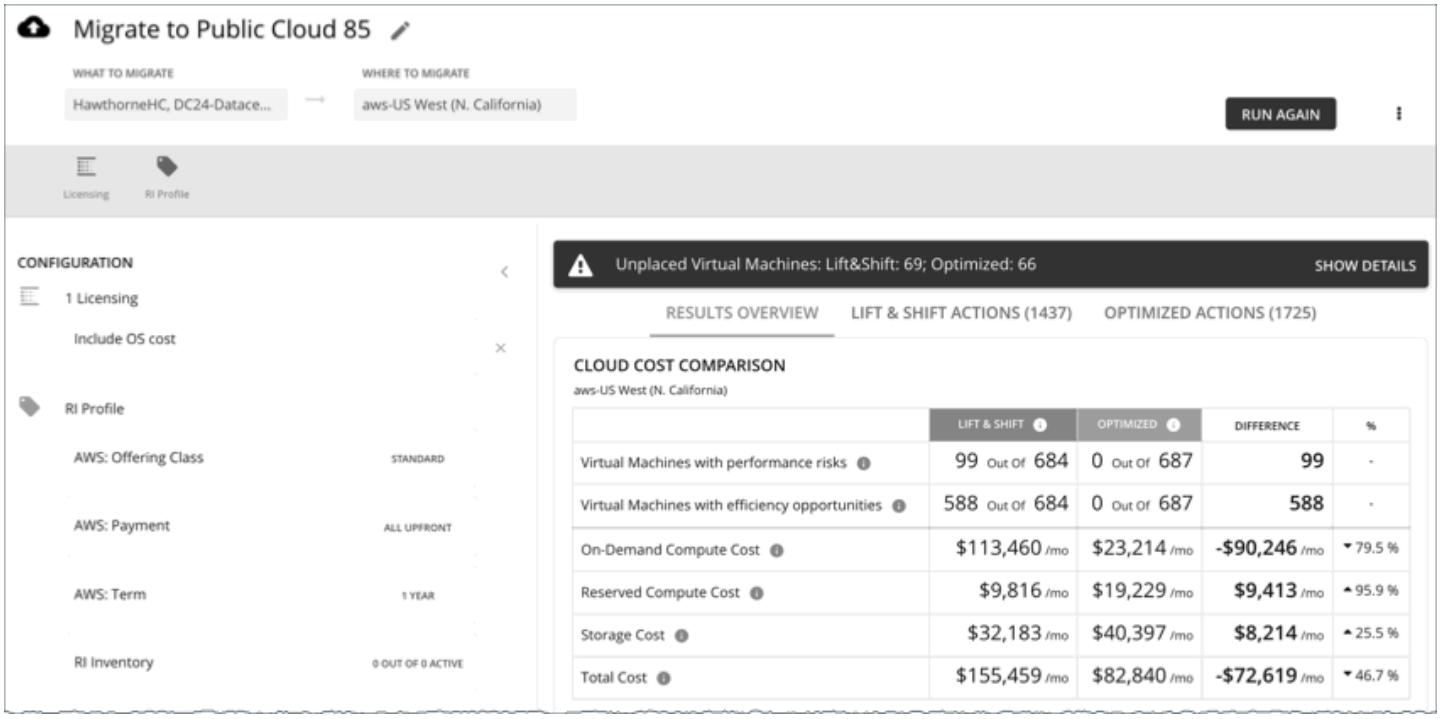
  – All Upfront – You make full payment at the start of the RI term.

  – 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.

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.



Workload Optimization Manager shows results for two migration scenarios:

- **Lift & Shift**

  Lift & Shift migrates your VMs to cloud instances that match their current resource allocations.

- **Optimized**

  As Workload Optimization Manager 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

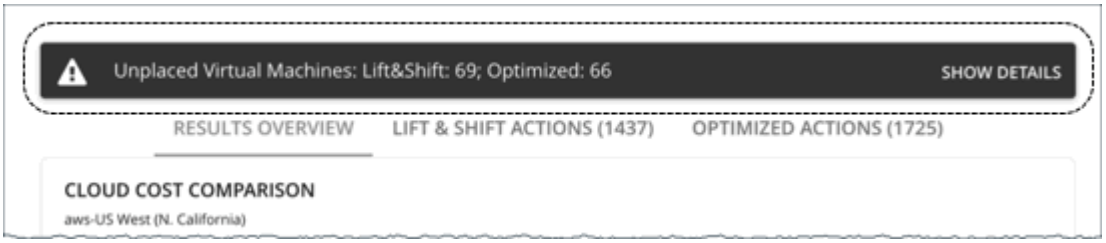
optimized migration, Workload Optimization Manager can recommend migrating to less expensive instances while still assuring performance, and then show the resulting savings. In addition, when you examine the actions for an optimized migration, you will see charts that plot the historical utilization 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.



  The charts in the plan results do not count these VMs.

  Workload Optimization Manager 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

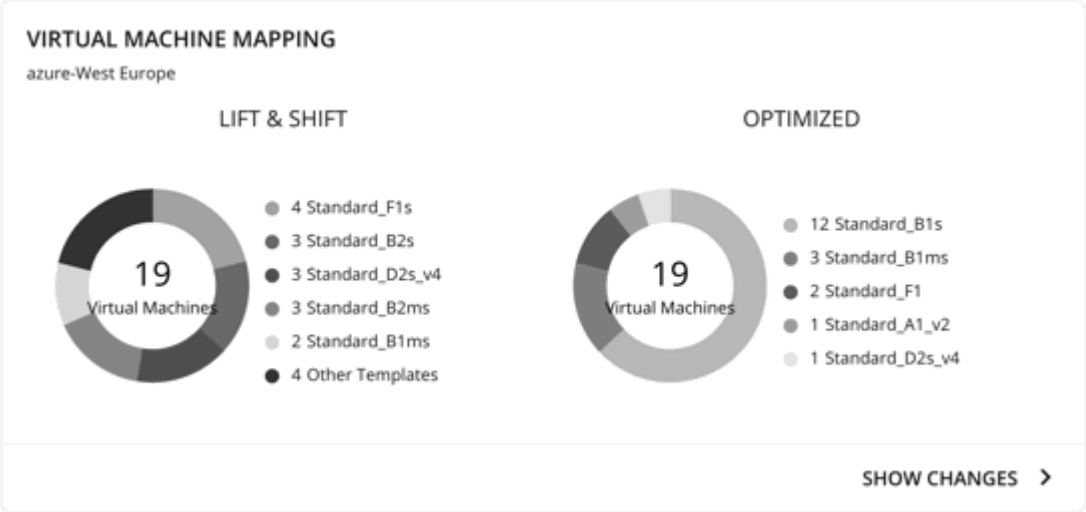
aws-US West (N. California)

	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 Workload Optimization Manager 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.

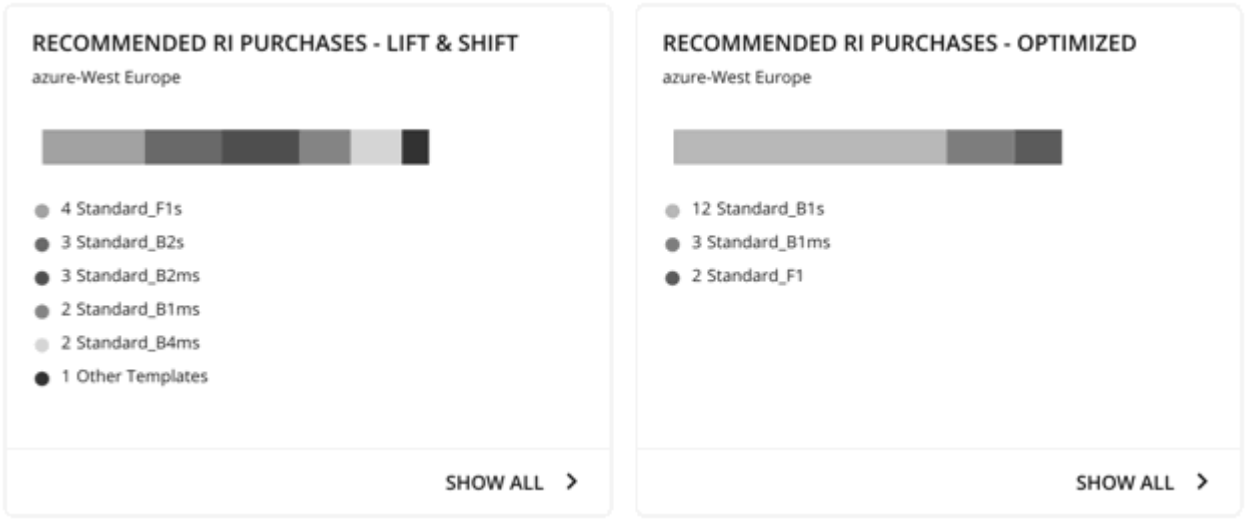
Click **Show Changes** to see a table with details for each volume in the plan. The table maps the volumes you plan to migrate to the volume types that Workload Optimization Manager recommends. It also shows the properties and monthly cost for each volume type. Under the **Actions** column, click **Details** to compare Lift & Shift and Optimized actions.

- **Recommended RI Purchases Charts**

  Workload Optimization Manager 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 403)*.

Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.



To identify VMs that are good candidates for discounted pricing, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Workload Optimization Manager 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.

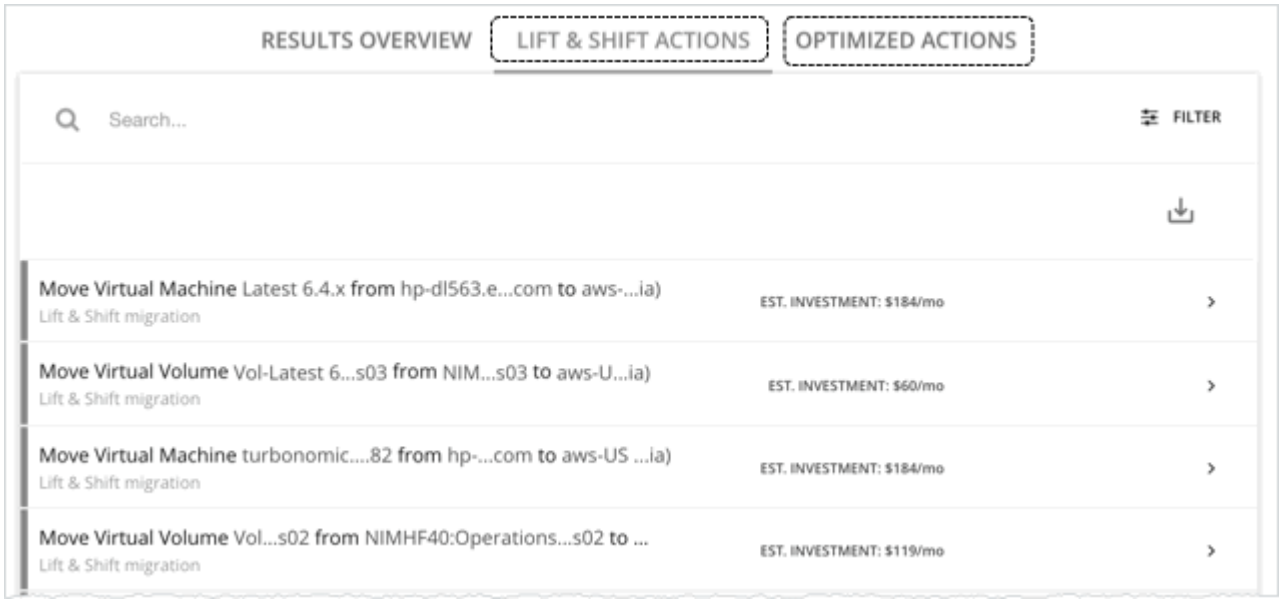
When you choose one or more check boxes, the total count, up-front cost, and savings appear at the top.

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

Workload Optimization Manager shows separate tabs for **Lift & Shift** and **Optimized** migration actions. You can download the list of actions as a CSV file.



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 Workload Optimization Manager analyzed to determine the most efficient instance for the VM.

For more information about these charts, see Utilization Charts *(on page 84)*.

Uploading Plan Results to Azure Migrate

Workload Optimization Manager 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 Workload Optimization Manager. Microsoft might not support displaying some information unique to Workload Optimization Manager.

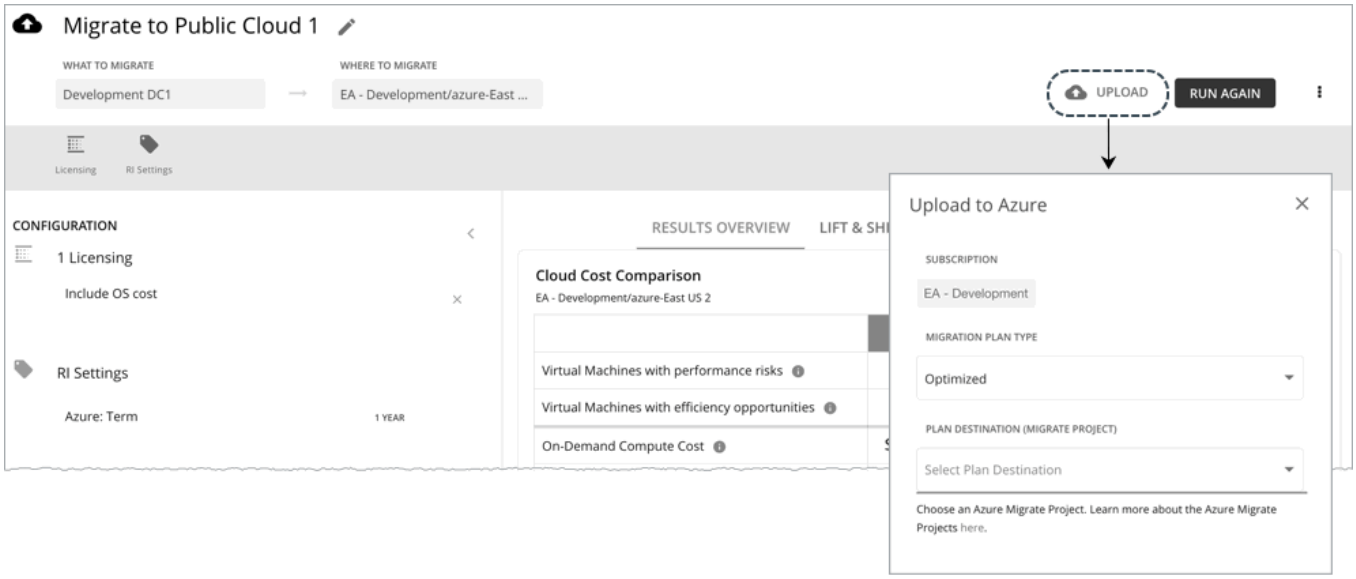
Before uploading the results, be sure to complete the following tasks:

1. Create a project in the Azure Migrate portal.
2. Add Workload Optimization Manager as a migration assessment solution to the project.
3. Set the necessary permissions in the Azure Migrate portal. Open Resource Explorer and configure the following operations:
   - Microsoft.Migrate/migrateprojects/read
   - Microsoft.Migrate/migrateprojects/solutions/read
   - Microsoft.Migrate/migrateprojects/solutions/getconfig/action

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

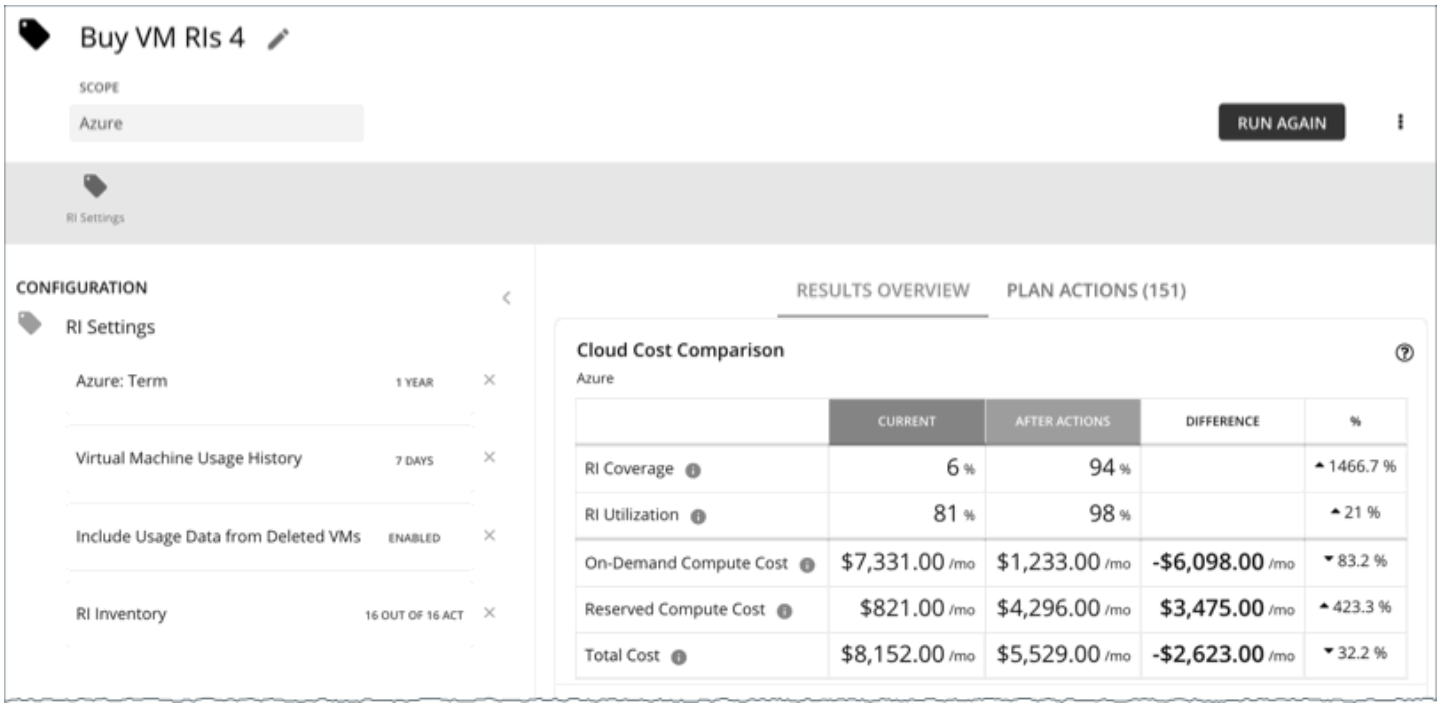
You can run the plan again with the same or a different set of configuration settings. This runs the plan scenario against the market in its current state, so the results you see might be different, even if you did not change the configuration settings.

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.

Buy VM Reservations Plan



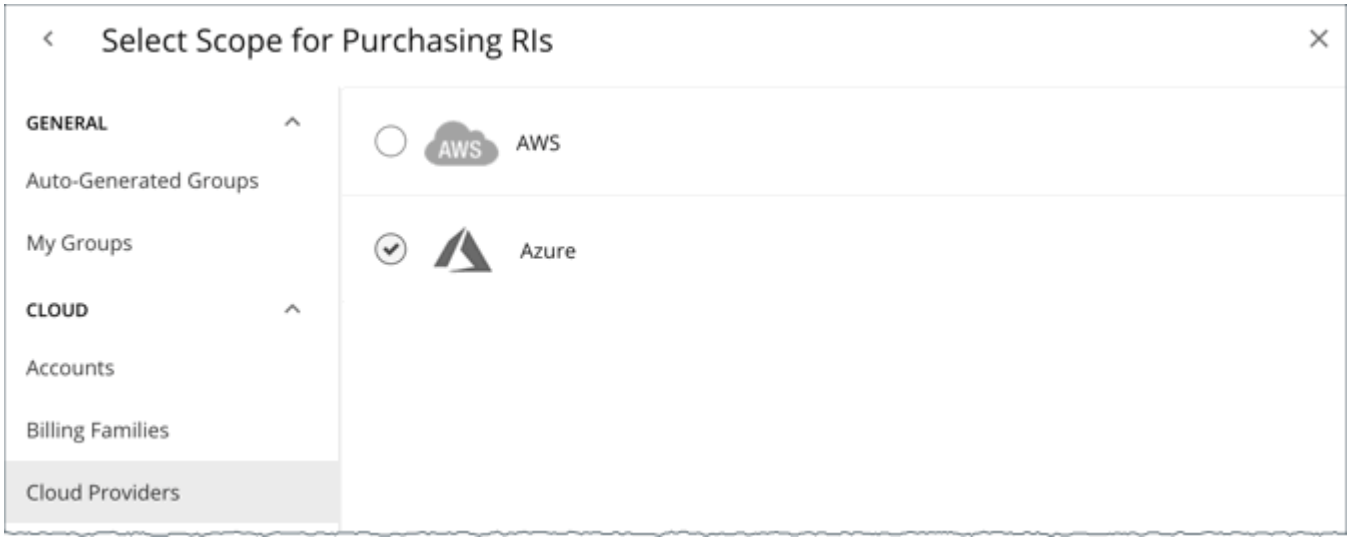
Run the Buy VM Reservations plan to see discount purchase opportunities that can significantly reduce on-demand costs for your cloud VMs. When calculating purchases, Workload Optimization Manager evaluates all purchasing options for your selected scope and usage data for the VMs in that scope. It then compares your current costs to the costs you would get after executing the plan recommendations.

Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.

Configuring a Buy VM Reservations Plan

For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.

1. Scope



You can scope by:

- Accounts

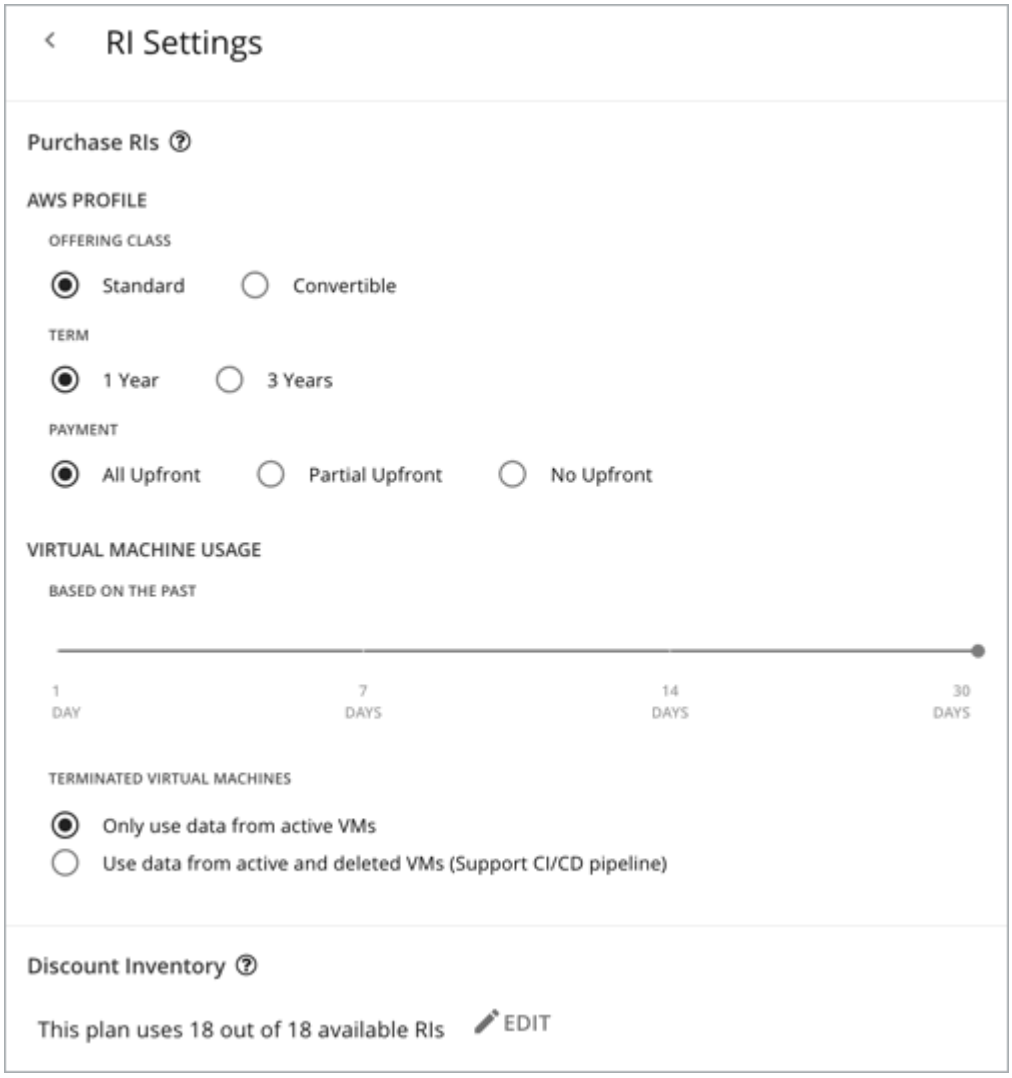
  Choose AWS accounts or Azure subscriptions for the plan's scope.
- Billing Families

  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.

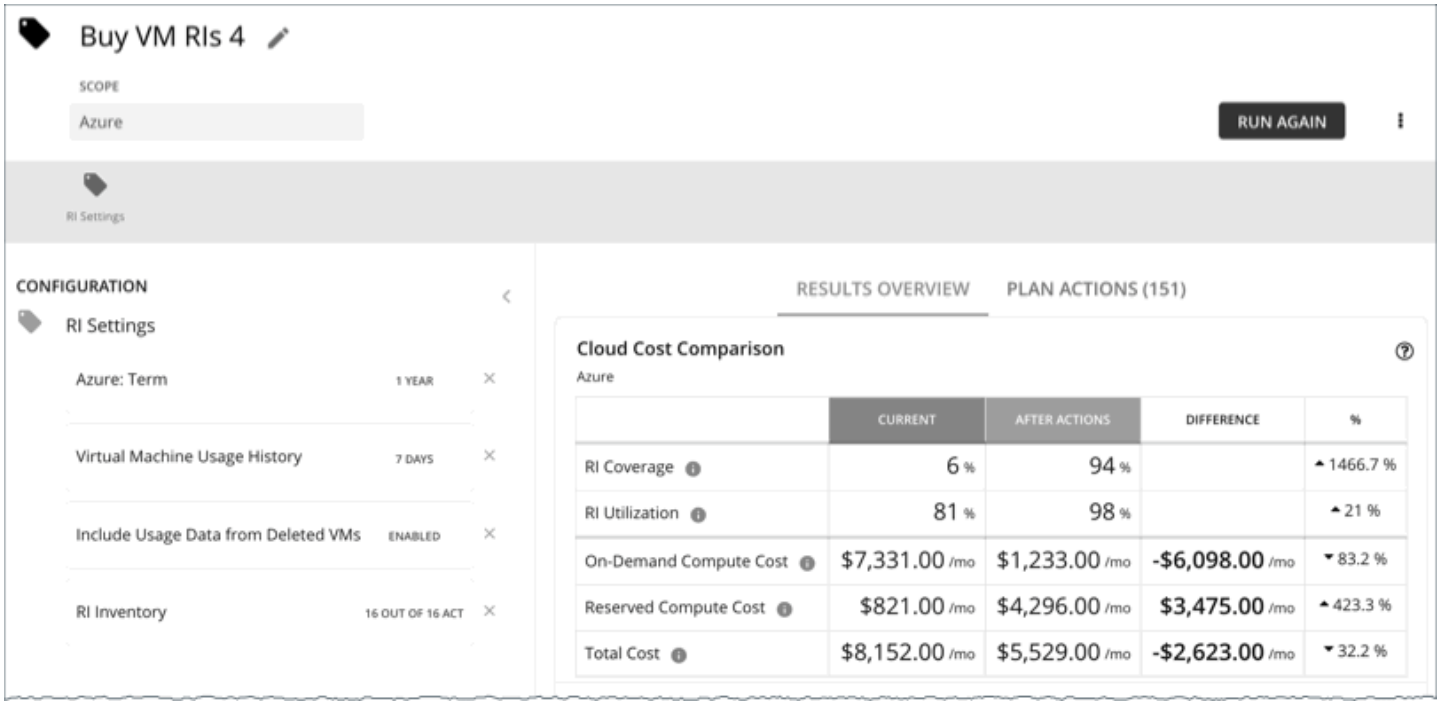
- **Terminated Virtual Machines**
  - **Only use data from active VMs** – Select this option if you terminate your VMs permanently.
  - **Use data from active and deleted VMs (Support CI/CD pipeline)** – Select this option if you want to use data from a CI/CD pipeline that regularly deploys and terminates 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**

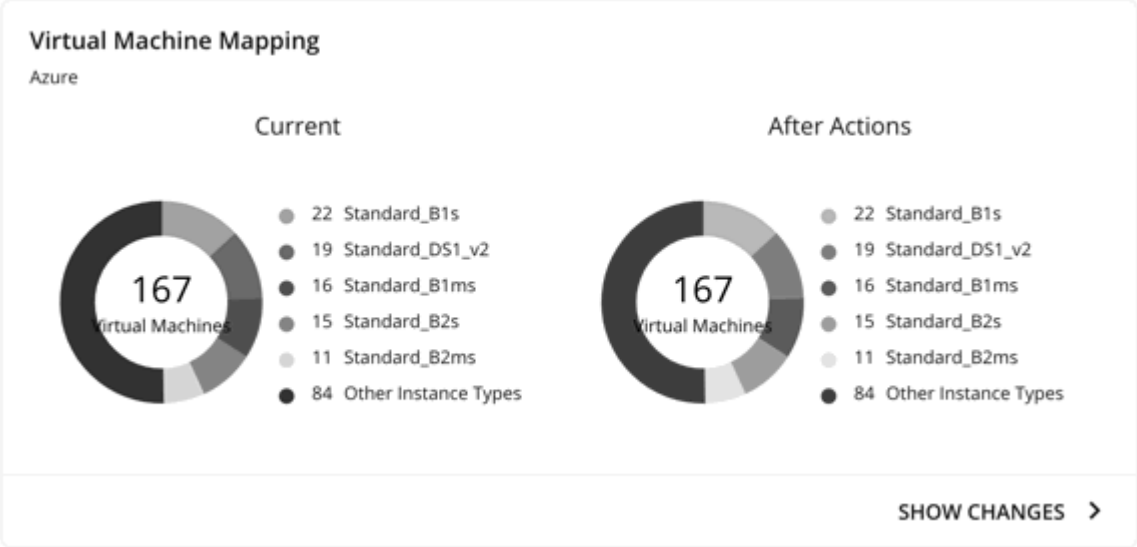
  This chart highlights changes to your existing discount coverage and utilization if you execute all the actions that the plan recommends. Actions include increasing coverage or purchasing additional instance types at a discounted rate. Your cloud provider will adjust discount allocations when the actions 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**

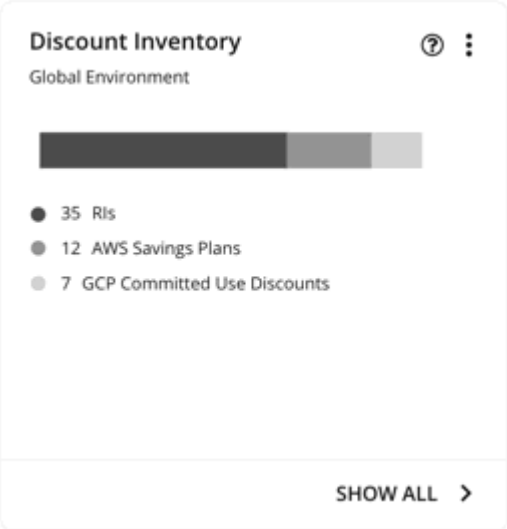
  This chart shows the instance types for the VMs included in the plan.



  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.

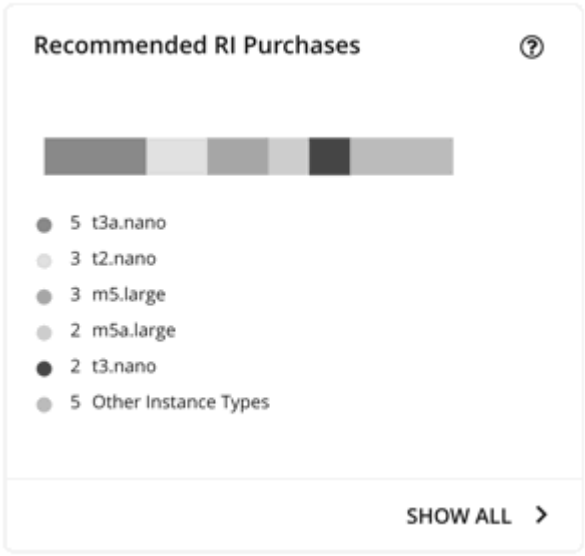
- **Discount Inventory**

  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**

  Workload Optimization Manager 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.



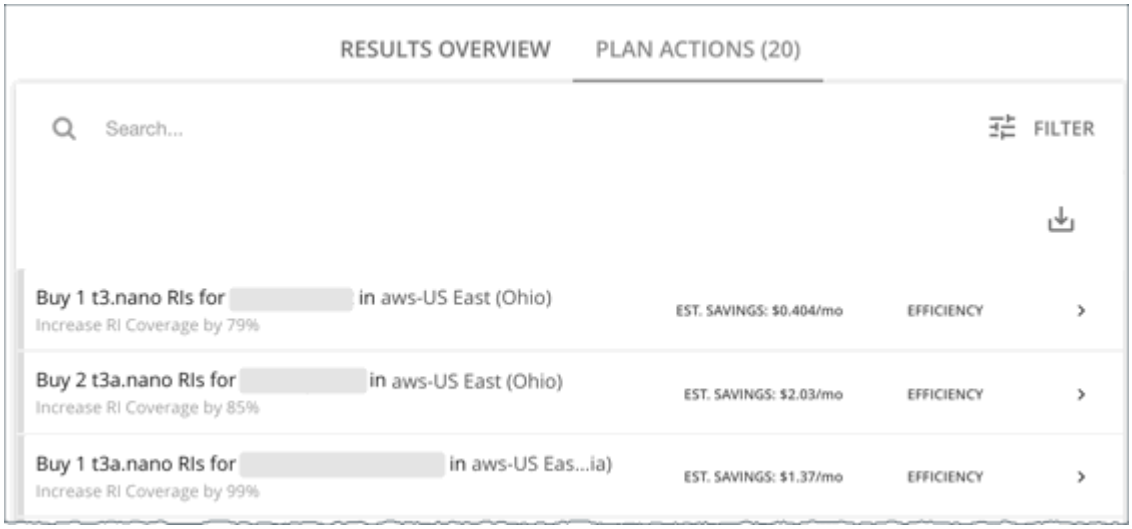
  Click **Show All** to see a table with details for each discount.

  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.



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 405)*.

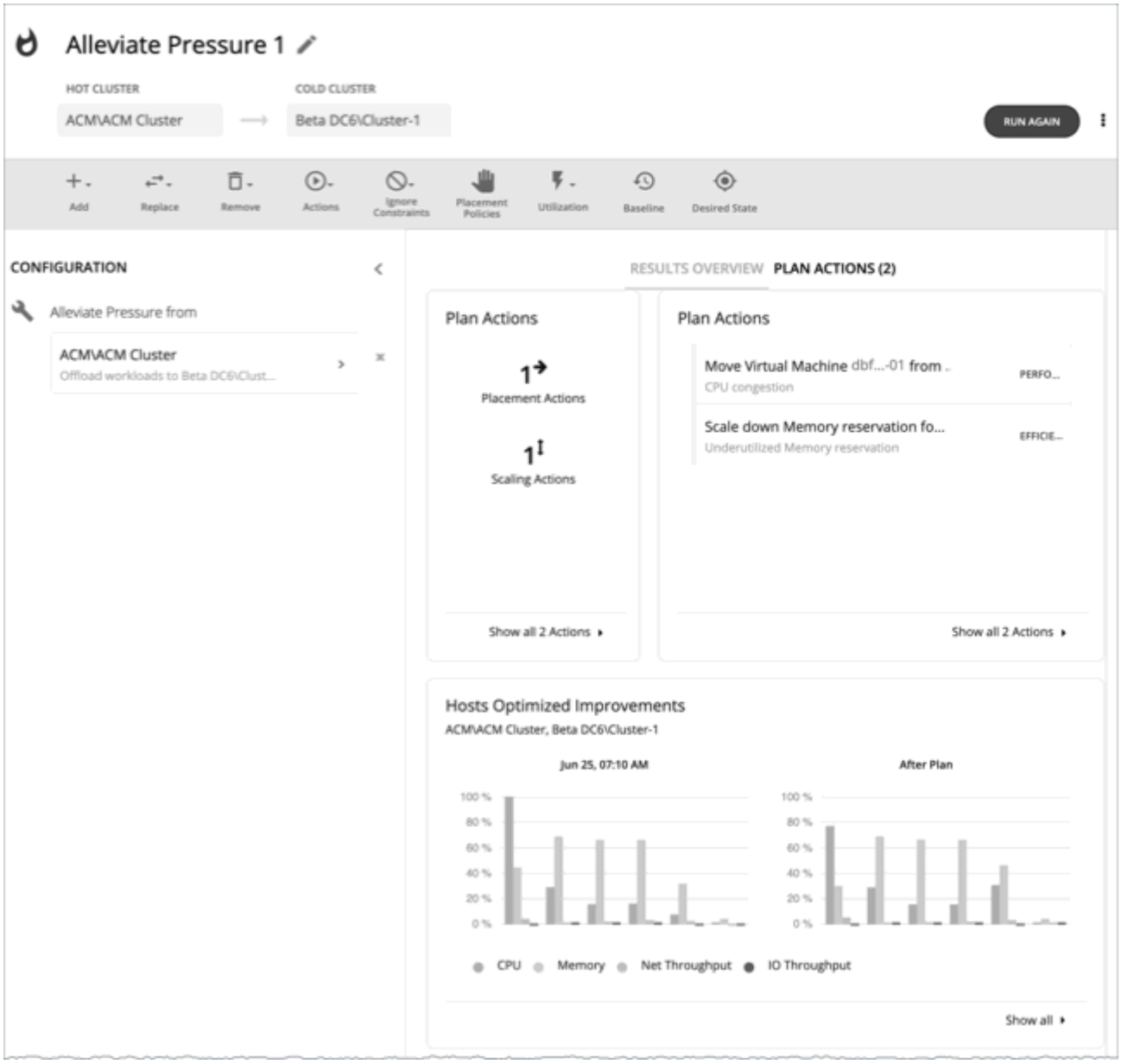
- Discount Inventory

  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.

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, Workload Optimization Manager 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 426)*

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 Workload Optimization Manager that would ordinarily be in effect.

Configuring an Alleviate Pressure Plan

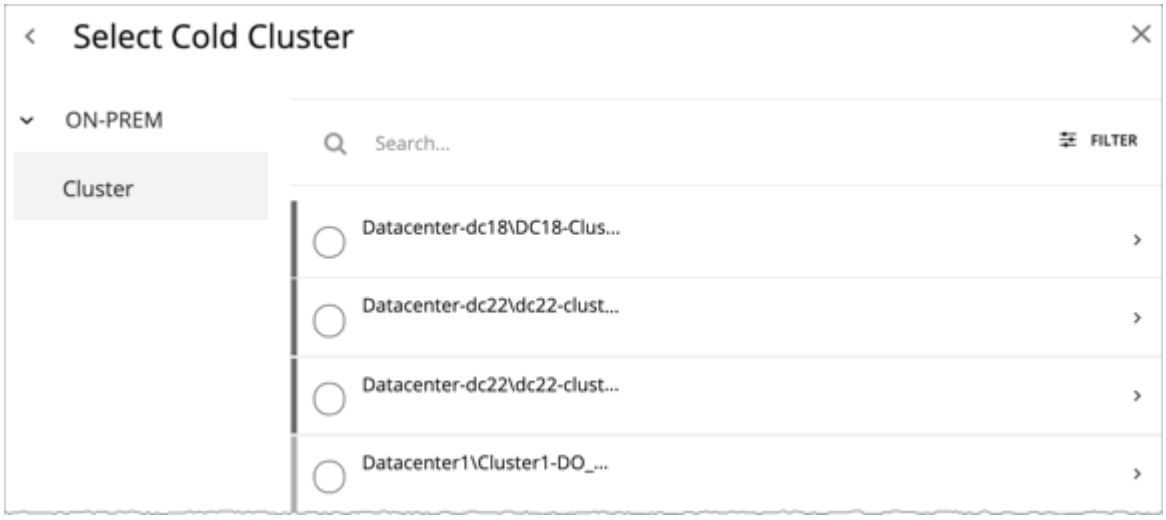
For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.

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.

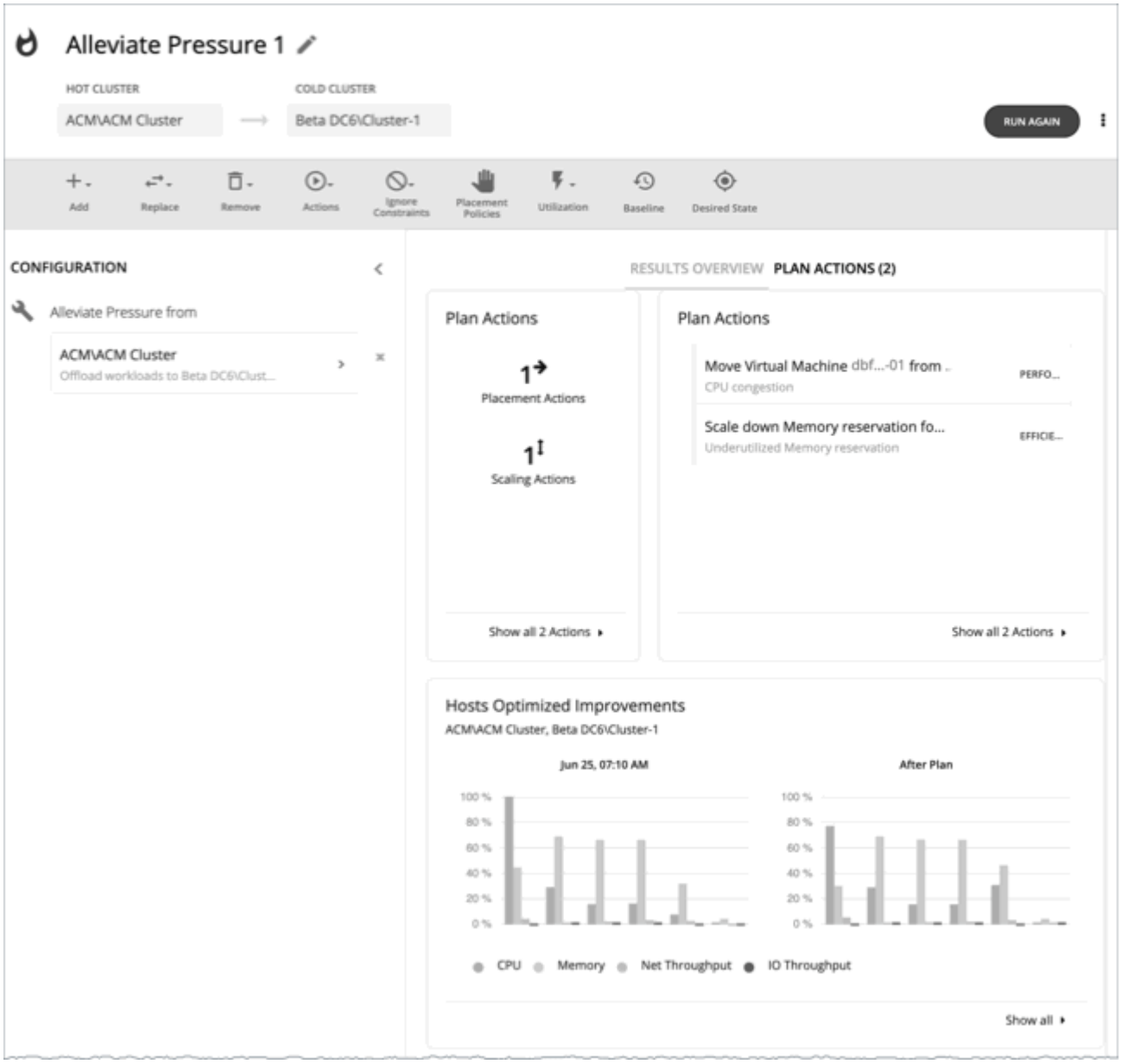
2. Cold Cluster

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.

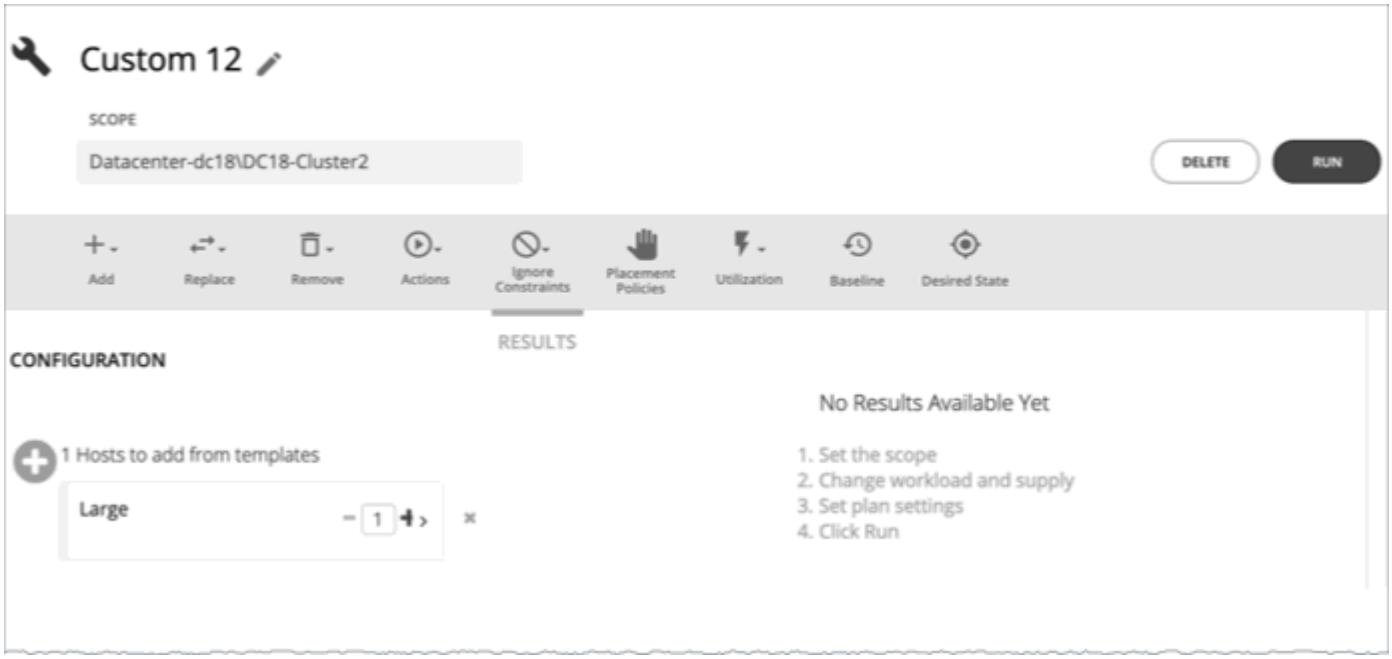
The toolbar items that display are similar to the toolbar items for a custom plan. For details, see Configuring a Custom Plan *(on page 415)*.

> **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 344)*.



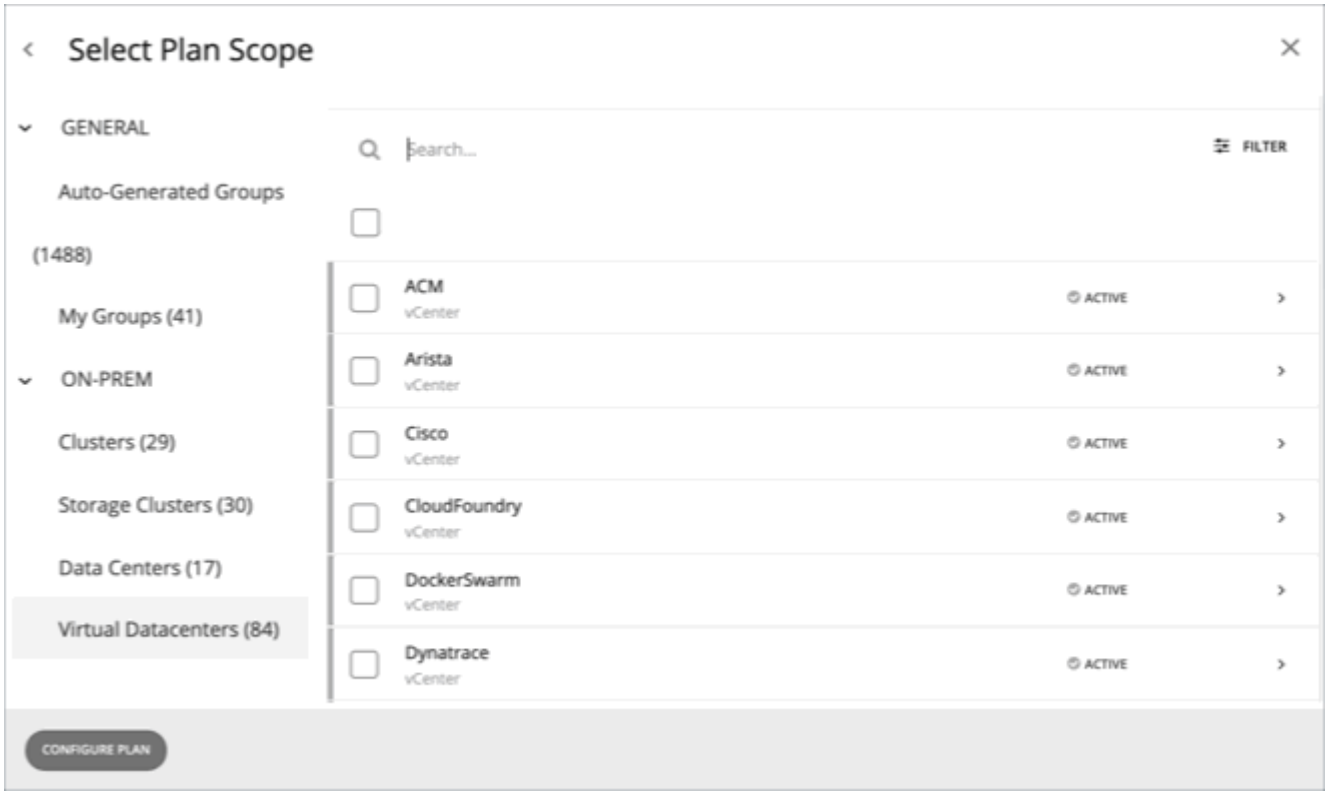
When you create a custom scenario, you specify the plan scope as an initial step, and then skip the plan wizards and jump straight into setting up the plan parameters. You can name the plan, change workload demand and the supply of resources, and specify other changes to the plan market.

Configuring a Custom Plan

For an overview of setting up plan scenarios, see Setting Up Plan Scenarios *(on page 344)*.

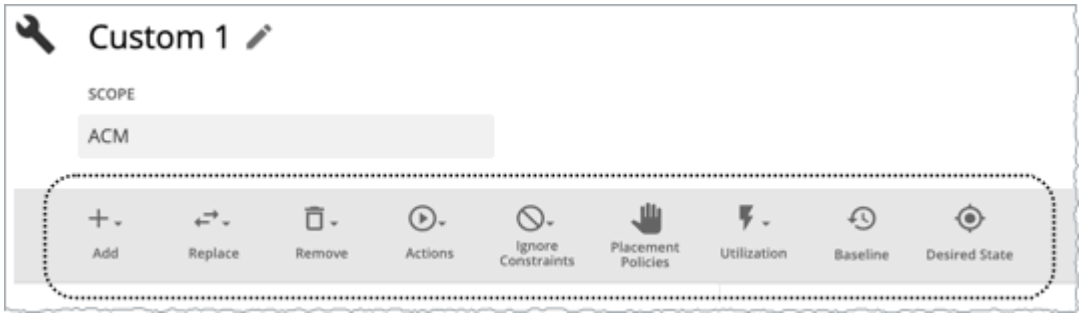
1. Scope

Specify the plan scope and then click **Configure Plan** at the bottom of the page.



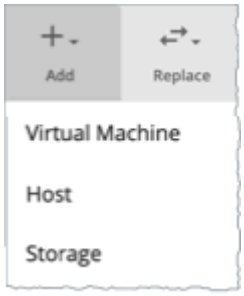
2. Plan Configuration

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

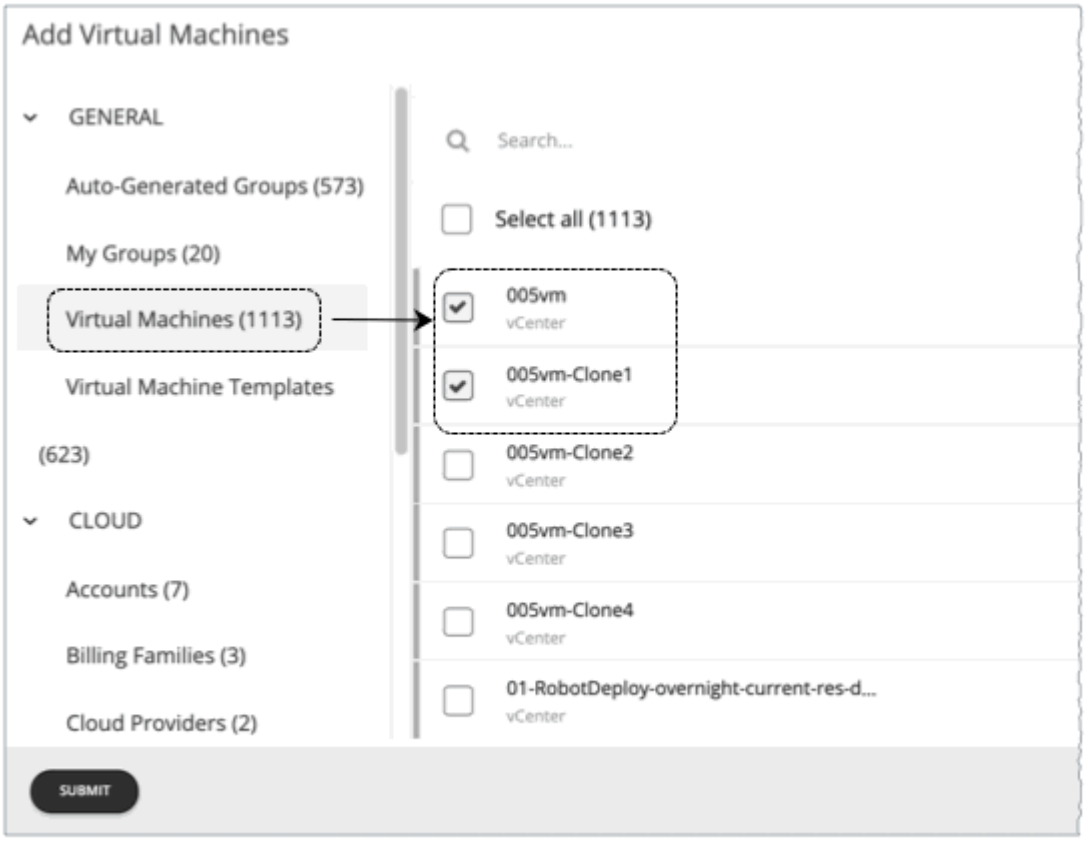
Add virtual machines, hosts, or storage to your plan. For example, when you add hosts, you increase the compute resources for the plan.



Copy from an Entity or Template

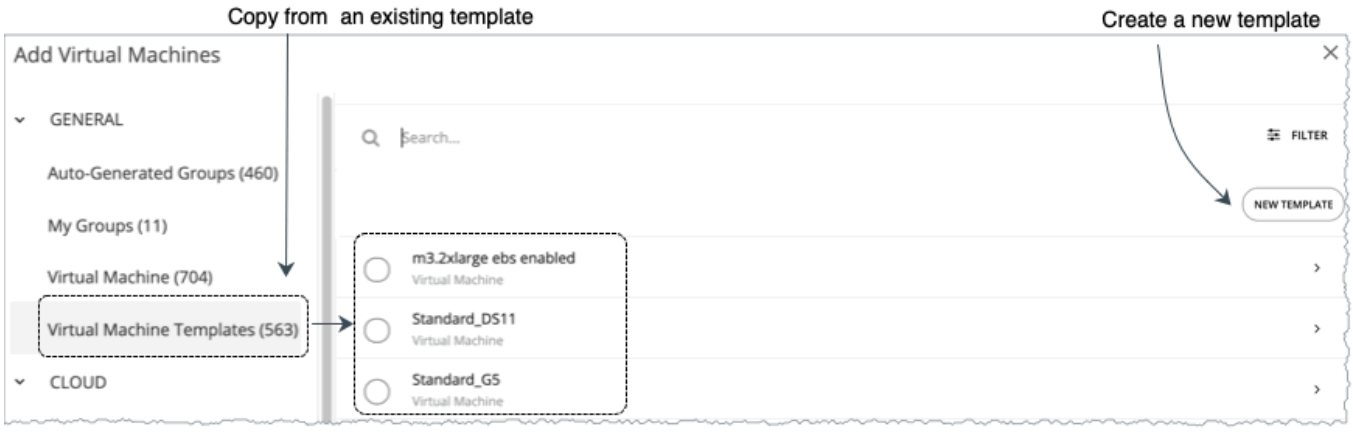
Choose an entity or template to copy. This describes the new entities that Workload Optimization Manager will add to the plan. For example, you can run a plan that adds new VMs to a cluster. If you copy from a template, then the plan adds a new VM that matches the resource allocation you have specified 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**.



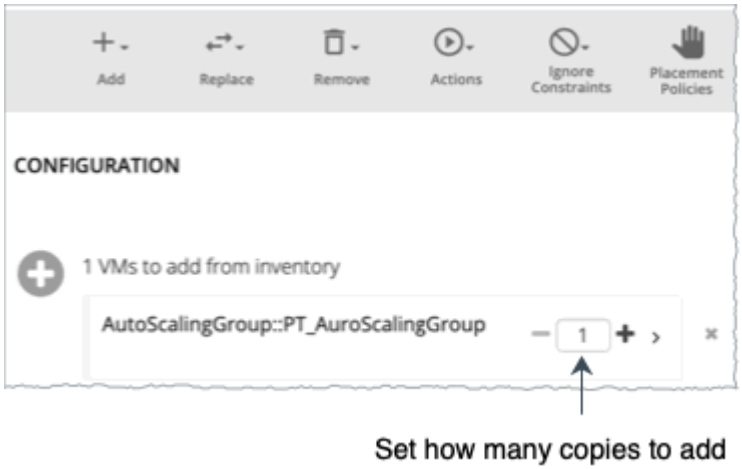
NOTE:
Workload Optimization Manager automatically adds any new template you create to the Template Catalog page (**Settings > Templates**).

It is not possible to use templates for containers or container pods.

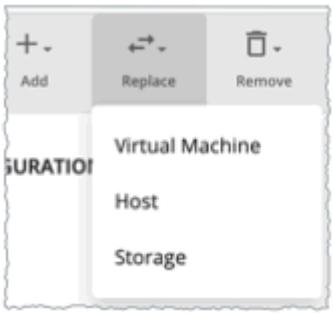
Use the **Filter** option to show entities or templates with certain properties (name, number of CPUs, etc.). This makes it easier to sort through a long list.

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.



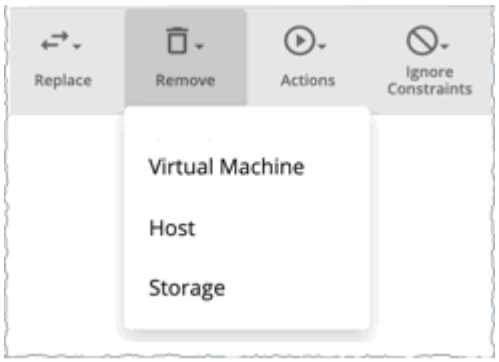
2.2. Replace



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.

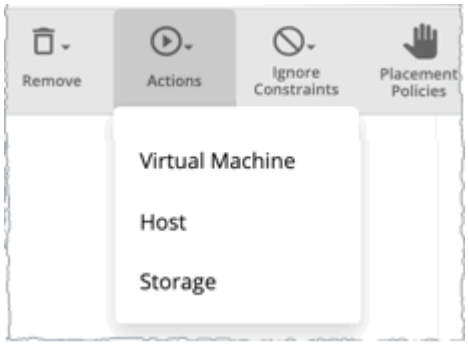
2.3. Remove



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.

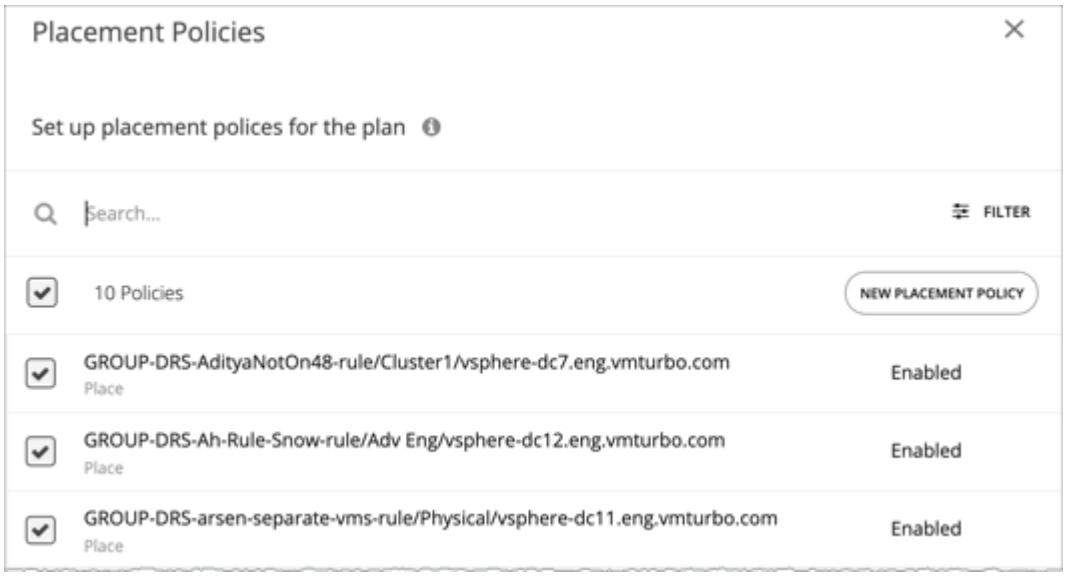
By default, VMs are constrained to the cluster, network group, datacenter, or storage group that their hosts belong to. You can choose to ignore these boundaries.

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

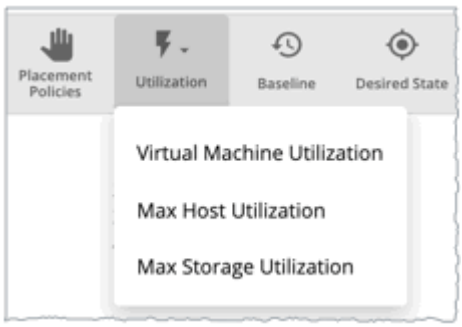
By default, the plan includes all the placement policies that apply to the plan scope. Also, these policies are in their real-time state (enabled or disabled).



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 88)*.

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. Workload Optimization Manager uses the resulting utilization values as the baseline for the plan.

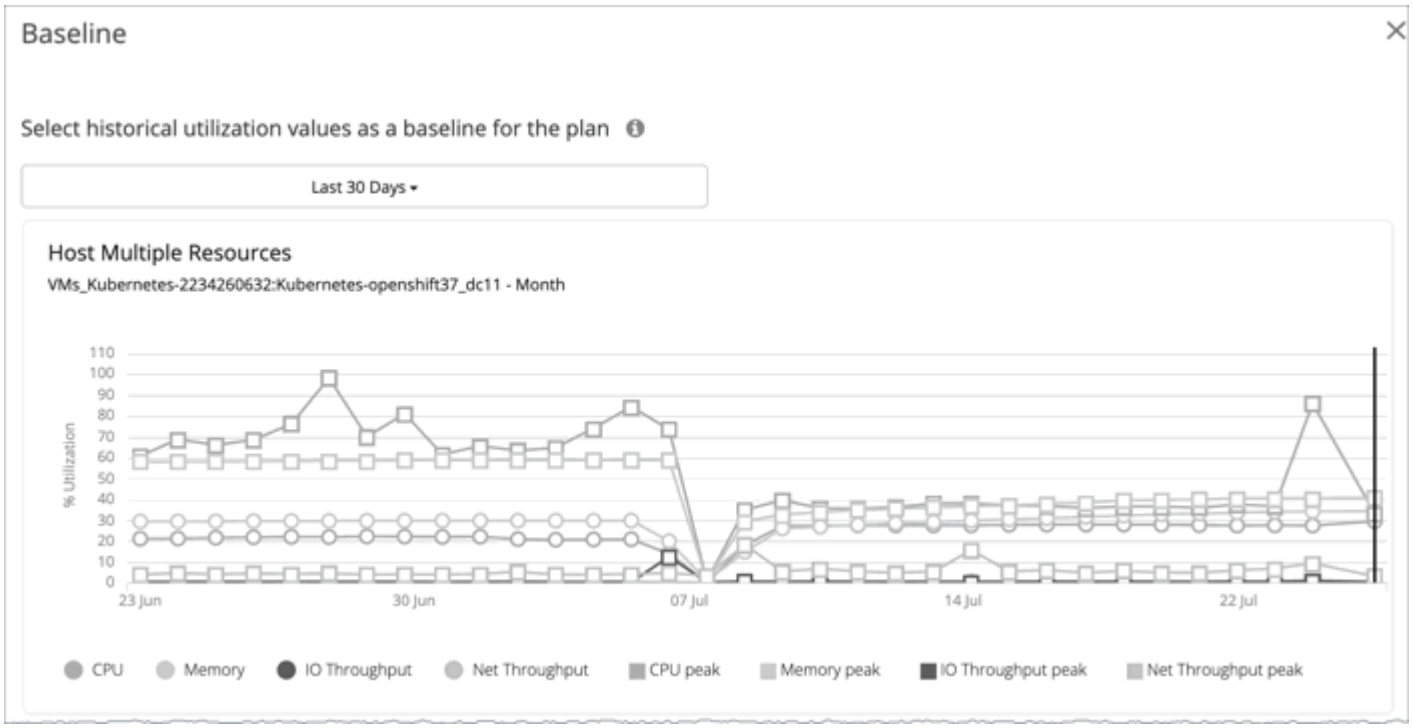


Max Host Utilization levels specify the percentage of the physical resource that you want to make available in the given plan. By default, hosts have utilization set to 100%. For a given plan, you can set the utilization to a lower value. For example, assume you want to simulate High Availability of 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.

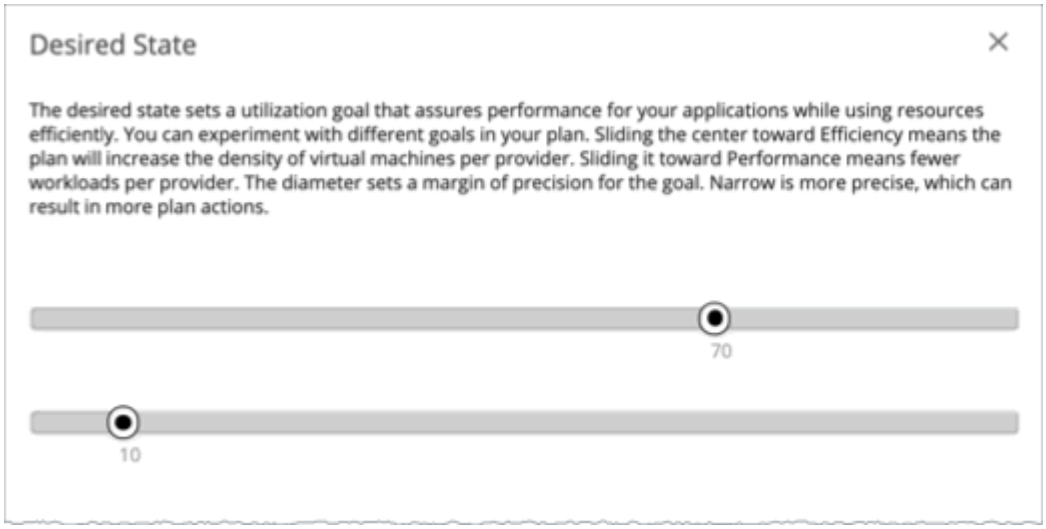


By default, the plan runs against the current state of your environment. You can set up the plan to add or remove entities, or otherwise affect the plan calculations. But the utilization metrics will be based on the current state of the plan. If you run the same plan multiple times, each run begins with 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. Workload Optimization Manager 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.



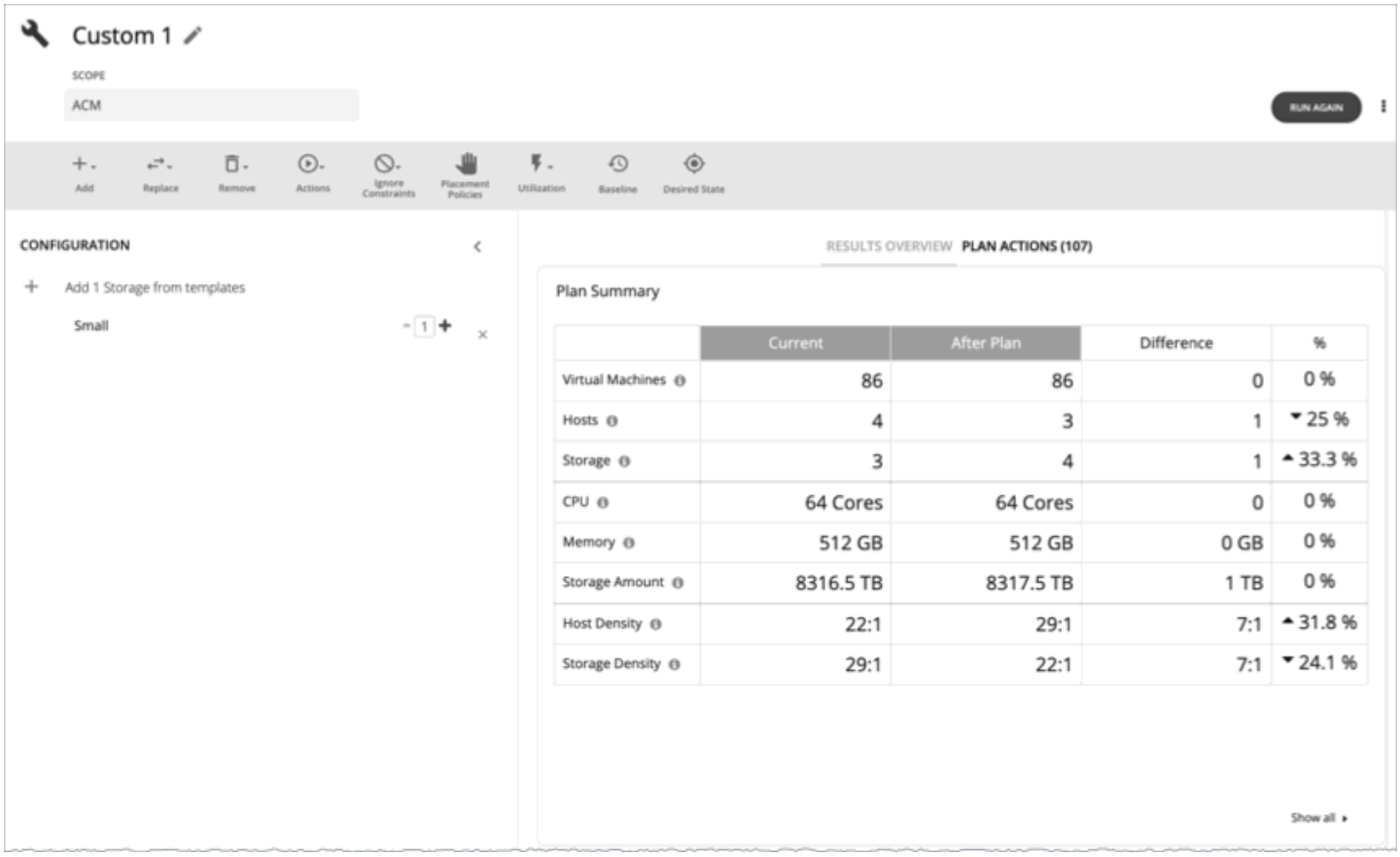
You can think of the desired state as an n-dimensional sphere that encompasses the fittest conditions your environment can achieve. The multiple dimensions of this sphere are defined by the resource metrics in your environment. Metric dimensions include VMem, storage, CPU, etc. While the metrics on the 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, Workload Optimization Manager will have more variation in the way it distributes workload across hosting devices.

For more information, see The Desired State *(on page 13)*.

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.



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

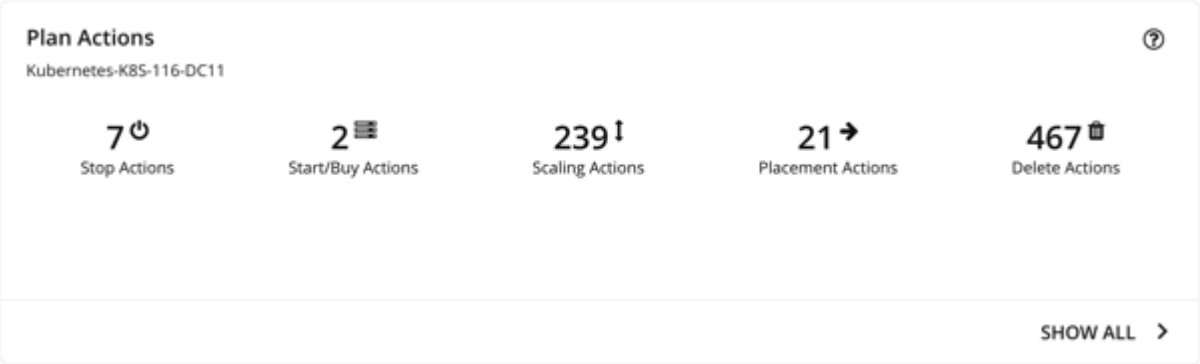
  If the plan's scope includes VMs that cannot be placed, the results include a notification indicating the number of VMs. Click **Show Details** to see the list of VMs and the reasons for their non-placement.

  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 67)*. The list chart shows a partial list of actions *(on page 58)*.

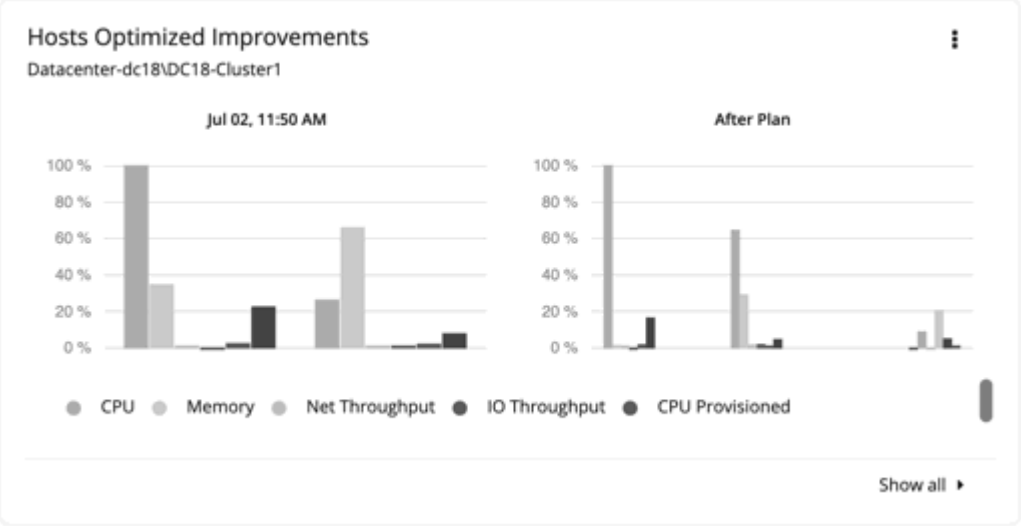


To view action details or download the list of actions as a CSV file:

- 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.

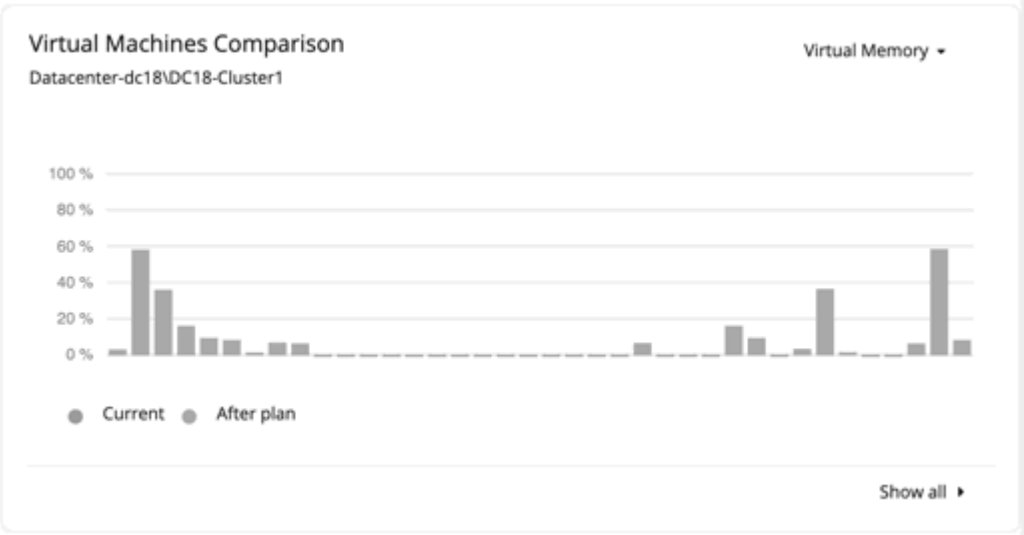


  - In many of these charts, you can change the commodities on display. To do this, go to the top-right section of the chart, click the More options icon ( ), and then select **Edit**. In the new screen that displays, go to the **Commodity** section and then add or remove commodities.

    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.

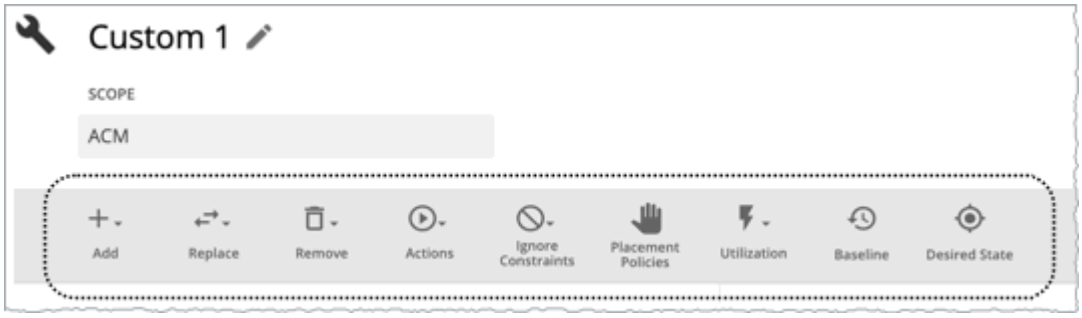
  To restore the default commodity, go to the top-right section of the page, click the More options icon ( ), and then select **Reset view**.
- Click **Show all** at the bottom of the chart to show a breakdown of the current chart data by entity (for example, show Virtual Memory utilization for each virtual 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.

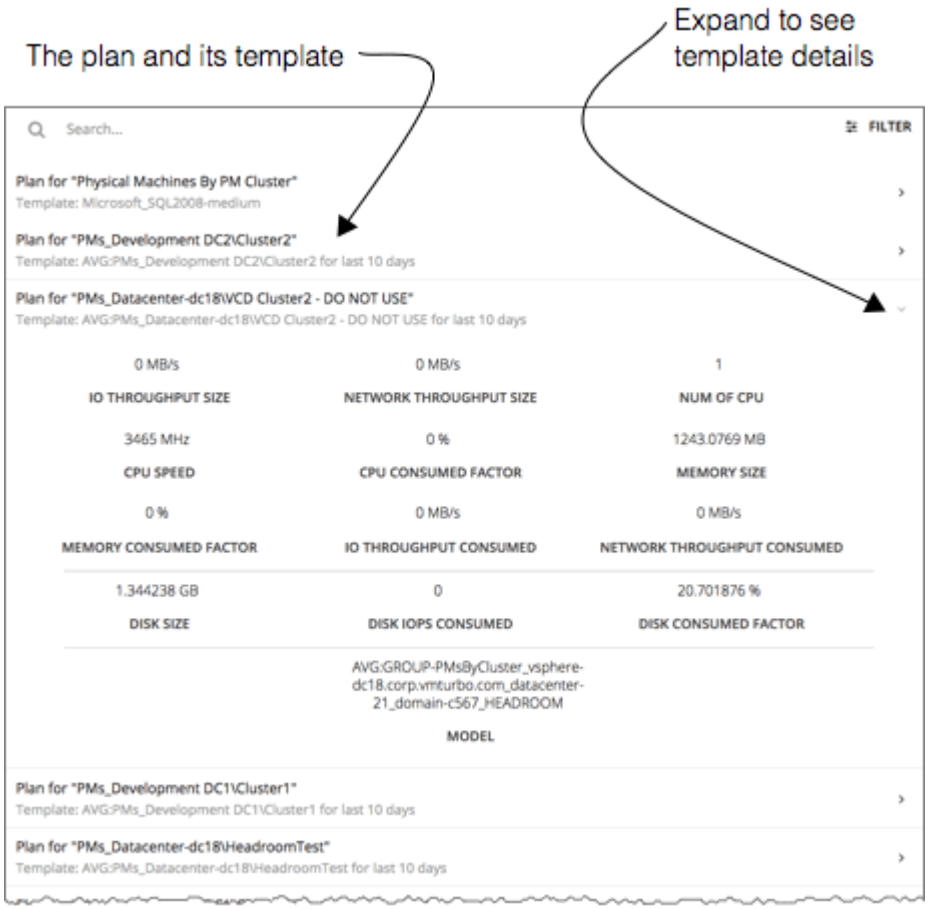


For details about these settings, see Configuring a Custom Plan *(on page 415)*.

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



Workload Optimization Manager 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.

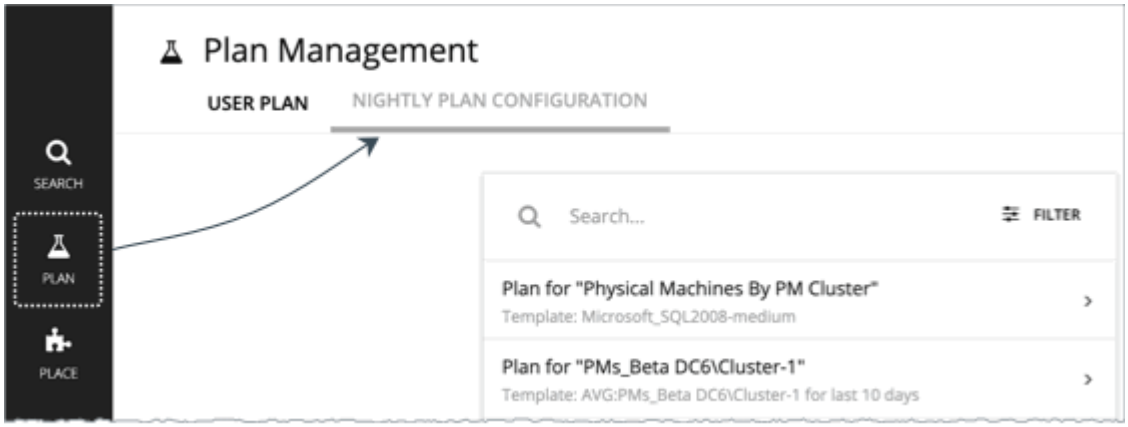
For information about viewing cluster headroom, see Viewing Cluster Headroom *(on page 57)*.

To calculate cluster capacity and headroom, Workload Optimization Manager 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**.

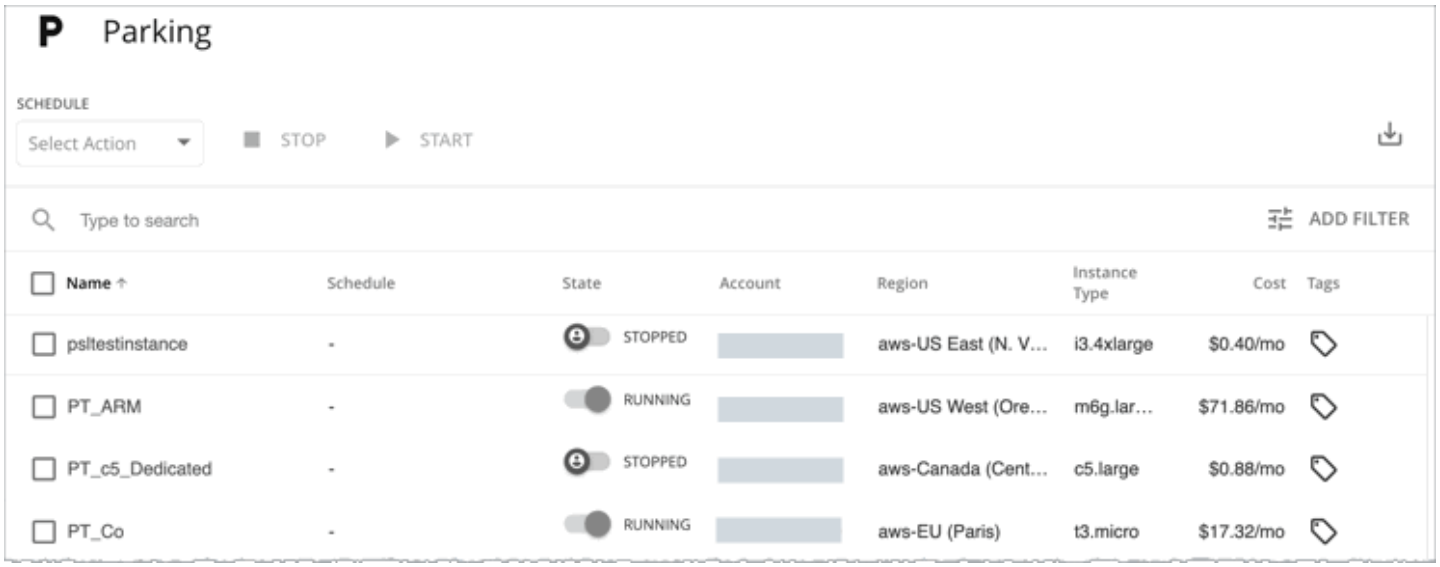


   This displays a list of all the nightly plans. Workload Optimization Manager creates a nightly plan for each cluster.
2. Click the plan that you want to configure.
   A fly-out appears that lists all the available templates.
3. Select the template you want for this plan.
   Choose the template and click **Select**.

Park: Stop or Start Cloud Resources

Workload Optimization Manager supports 'parking' actions for cloud resources. These actions *stop* your cloud resources for a given period of time to help you reduce your cloud expenses, and then *start* these resources later when you need them. You can enforce parking actions on demand or according to a schedule.

Currently, Workload Optimization Manager supports parking actions for **cloud VMs** discovered via AWS, Azure, and GCP targets. When you click **Park** on the main menu, the user interface displays a list of parkable VMs discovered from these targets.



Permissions

The permissions required to monitor cloud VMs and execute VM scaling actions should be sufficient to enforce parking actions. No additional permissions are required.

For a full list of permissions, see the following topics:

- BROKEN LINK!!! AWS Permissions
- BROKEN LINK!!! Azure Service Principal and Subscription Permissions
- BROKEN LINK!!! GCP Permissions

How Parking Works

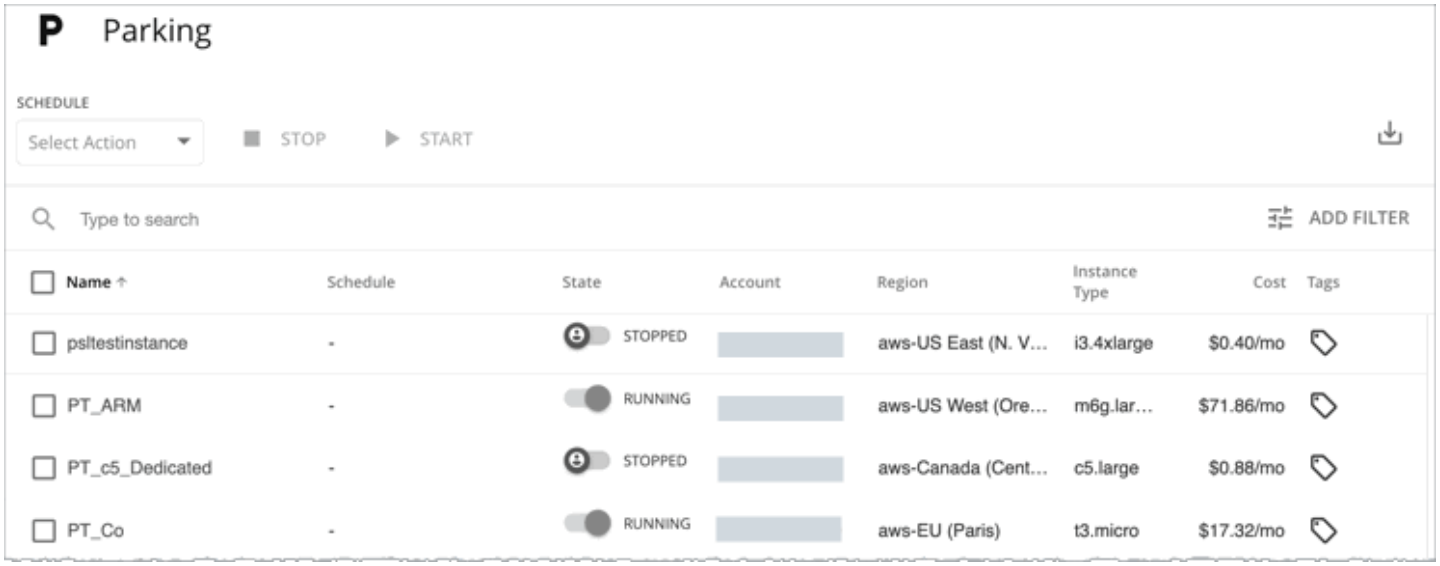
VMs that become idle for a period of time (for example, after business hours) and those that undergo regular maintenance are ideal for parking. When parking actions are enforced, Workload Optimization Manager reverses any conflicting actions on VMs. For example, when a VM has been started via Workload Optimization Manager and a user stops the VM via the cloud provider, Workload Optimization Manager will start the VM after a subsequent discovery cycle.

When a scaling action is in progress, Workload Optimization Manager postpones a stop or start action until the scaling action completes.

Be aware that VMs on consistent scaling groups, auto scaling groups, or scale sets are not parkable.

Working with the Parking Page

After discovering parkable VMs from a cloud provider, Workload Optimization Manager adds the VMs to the Parking page and shows their state (stopped or running) at the time of discovery. Use the Parking page to execute parking actions or attach parking schedules to VMs.



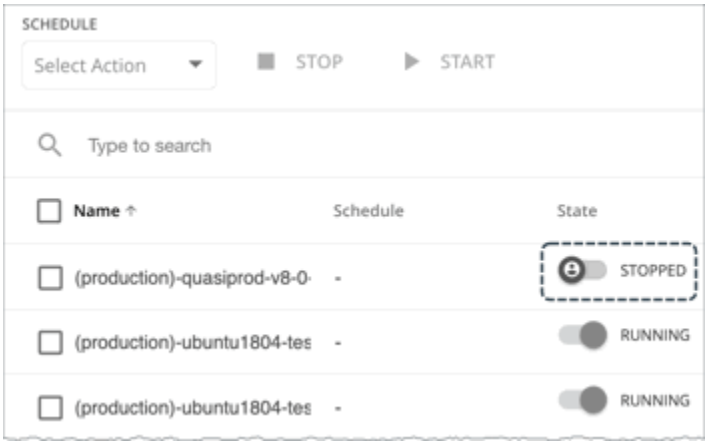
Points to consider

- Users with the Site Administrator, Administrator, or Automator role can manage the parking actions and settings in the Parking page. Users with other roles have read-only access to the page. For details about roles, see Managing User Accounts *(on page 540)*.
- Parking actions can only be executed from the Parking page. The Pending Actions, Potential Savings, and Top Accounts charts in the Home Page only show non-parking actions that you can execute.
- Add the All Actions chart to your dashboard and use the tabular view to see and download a list of parking (and non-parking) actions that have been executed successfully.
- Currently, actual savings associated with parking actions are not reflected in the Parking page or the Cumulative Savings chart. The **Cost** column in the Parking page shows the estimated on-demand monthly cost for a VM, for your reference.

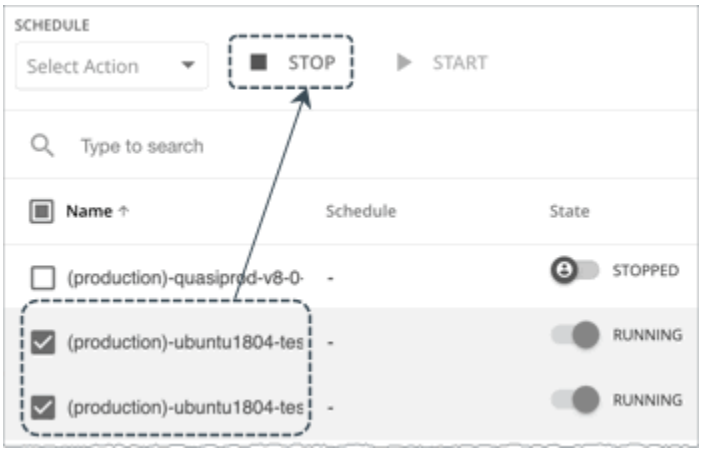
Executing Parking Actions

You can execute parking actions in the following ways:

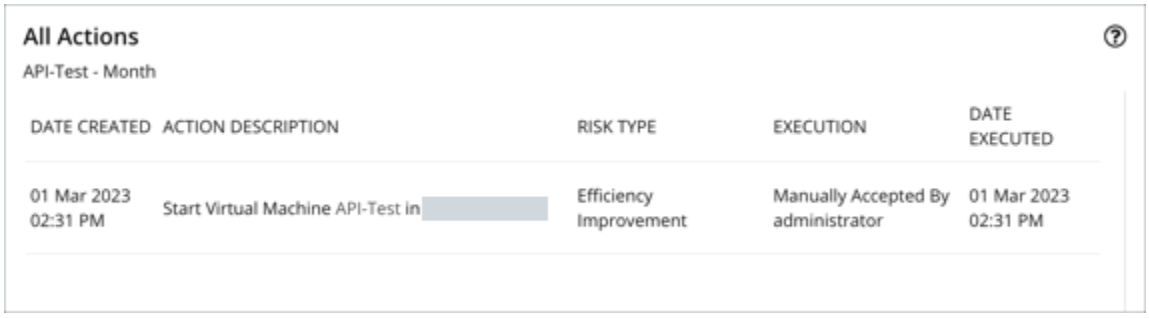
- For individual VMs, use the toggle under the **State** column to change the currently enforced parking action.



- Select one or several VMs with the same state, and then click **Stop** or **Start**.



In the All Actions chart, the tabular view shows the following information for the parking actions that you executed:



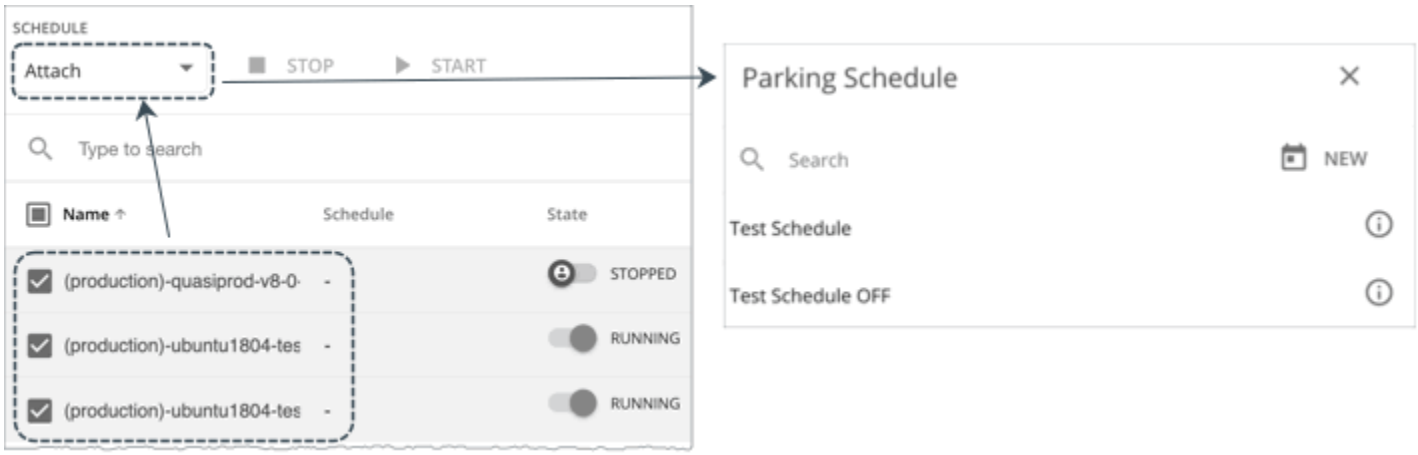
- Execution: Manually Accepted
- Risk Type: Savings (stop actions) or Efficiency Improvement (start actions)

Scheduling Parking Actions

Attach a parking schedule to VMs to ensure that parking actions occur automatically at the appropriate times.

NOTE:
Currently, parking actions cannot be scheduled using automation policies *(on page 209)* for cloud VMs. Automation policies (and the schedules configured in these policies) only apply to scale actions.

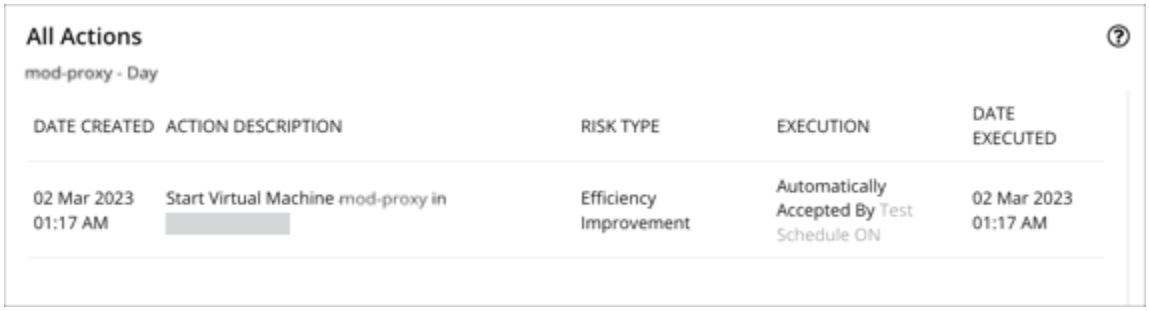
To attach a schedule, select one or several VMs without a schedule and then select **Attach** from the dropdown at the top-left section of the page. In the Parking Schedule page, choose from a list of previously configured schedules.



If no schedule is suitable or available, create one by clicking **New** at the top-right section of the Parking Schedule page. Be aware that a schedule created this way will immediately apply to the VMs in scope. If this is not your preference, create a new schedule by navigating to **Settings > Schedule** and then clicking **New Schedule > Parking Schedule**. For instructions on creating parking schedules, see Managing Parking Schedules *(on page 517)*.

After attaching a schedule, the name of the schedule appears under the **Schedule** column. Click the name to view the settings for that schedule. To edit the settings, navigate to **Settings > Schedule** and then open the schedule for editing.

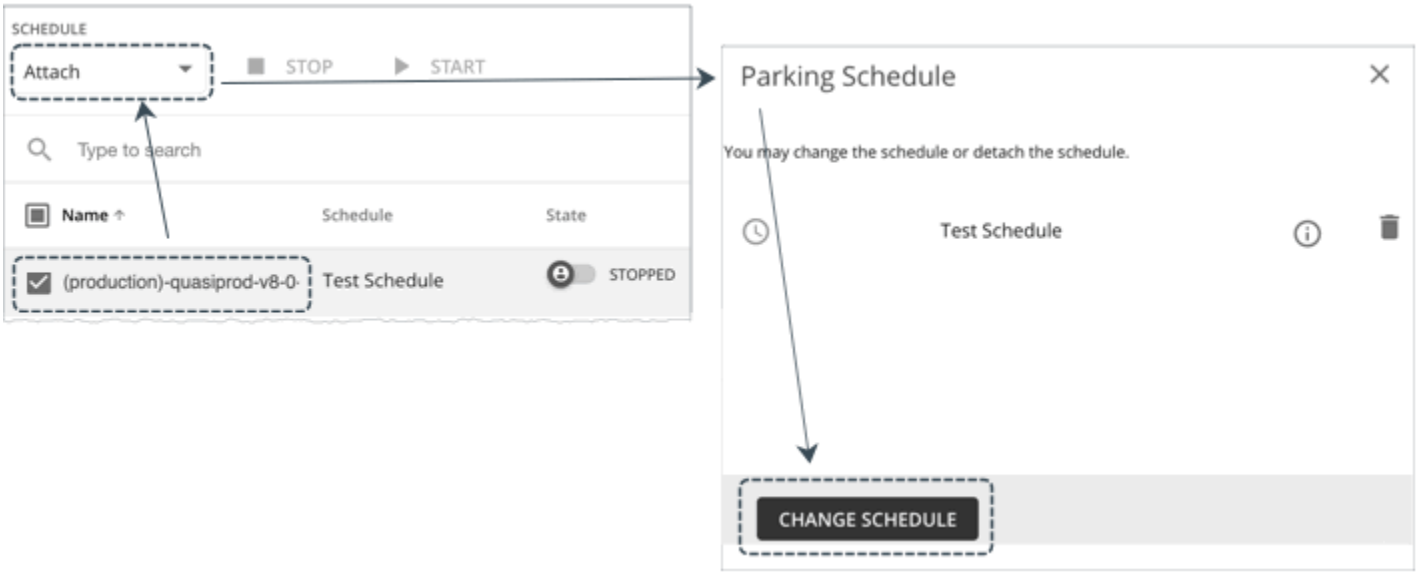
In the All Actions chart, the tabular view shows the following information for scheduled parking actions that have been executed:



- Execution: Automatically Accepted
- Risk Type: Savings (stop actions) or Efficiency Improvement (start actions)

Changing a Parking Schedule

To change the schedule currently attached to a VM, select the VM and then select **Attach** from the dropdown at the top-left section of the page. In the Parking Schedule page, click **Change Schedule** and then choose from a list of previously configured schedules (or create a new one). When you are done, the new schedule displays under the **Schedule** column.

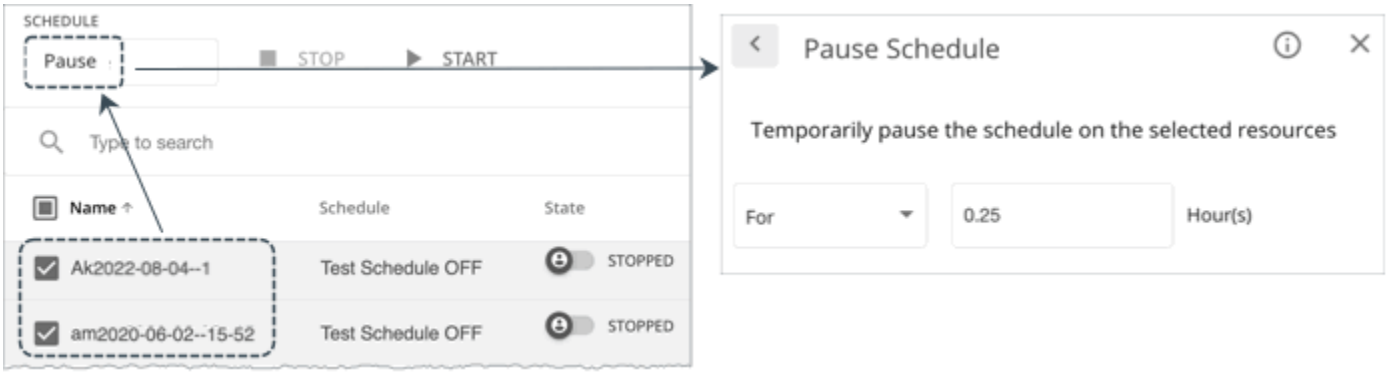


You can select multiple VMs, as long as those VMs share the same schedule.

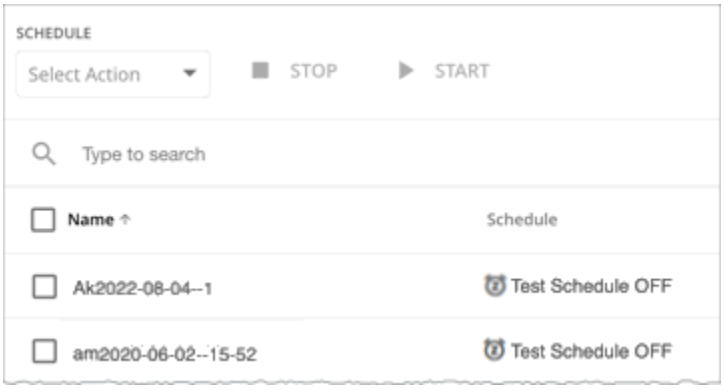
Pausing a Parking Schedule

You can pause schedule enforcement for a given period of time. If the VM was running, it keeps running. If it was stopped, it stays stopped.

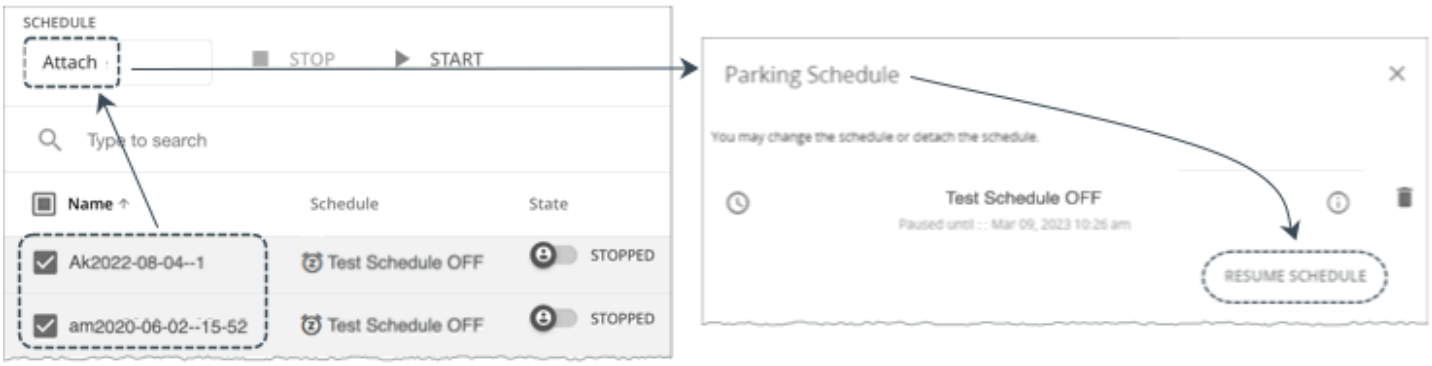
To pause a parking schedule, select one or several VMs with a schedule, and then select **Pause** from the dropdown at the top-left section of the page. In the Pause Schedule page, specify the duration of the pause.



An icon appears before the name of the schedule to indicate that it has been paused.



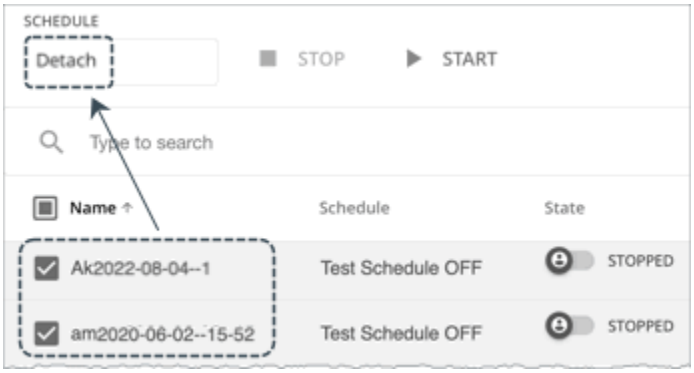
The parking schedule will resume automatically when the pause period is over. To manually resume the schedule, select the affected VM(s), and then select **Attach** from the dropdown at the top-left section of the page. In the Parking Schedule page, click **Resume Schedule**.



NOTE:
When you execute a parking action on demand and the affected VM has an existing schedule, a system prompt notifies you to pause the schedule before executing the action. This prevents potential conflicts between on-demand and scheduled parking actions.

Detaching a Parking Schedule

You can detach parking schedules from VMs anytime. Select one or several VMs with schedules, and then select **Detach** from the dropdown at the top-left section of the page.



General Tasks

- For a long list of VMs, you can type a search keyword or add filters to narrow the results.
- To set a VM as your scope, click the VM name.
- To download the data in the Parking page, click the download button at the top-right section of the page.

Place: 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. Workload Optimization Manager 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 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 522)*.

About Placement of Reserved VMs

To determine the best placement for the VMs you want to reserve, Workload Optimization Manager runs a plan using the last-generated data in nightly-run headroom plans.

> **NOTE:**
> If you have changed 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, Workload Optimization Manager considers the following:

- Placement constraints set in the reservation
- Demand capacity

  Workload Optimization Manager calculates demand based on the *resource allocation* and *consumed factor* set in VM templates. For example, if you want to create a reserved VM from a template that assigns 3 GB of virtual memory and a consumed factor of 50%, Workload Optimization Manager 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, Workload Optimization Manager 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), Workload Optimization Manager 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 will have an overprovisioned capacity of 5 TB (5120 GB).

Providers must have sufficient *demand* and *overprovisioned* capacity to place a reservation. Workload Optimization Manager 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, Workload Optimization Manager can prevent congestion issues after you deploy the VMs.

NOTE:
Workload Optimization Manager 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, Workload Optimization Manager 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 will prevent the placement.

  Note that *actual* and *reserved* VMs share the same resources on providers. This means that provider capacity will change as demand from the actual VMs changes. Workload Optimization Manager 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, Workload Optimization Manager 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 will result in congestion. Workload Optimization Manager will *not* retry fulfilling the reservation.

- Failed Initial Placements

  If the initial placement attempt is unsuccessful (for example, if all providers have seen historical congestion), Workload Optimization Manager shows that the placement has failed and will *not* retry fulfilling the reservation.

Current and Future Reservations

You can create a current or future reservation from the Workload Placement Page.

- Current Reservation

  Workload Optimization Manager 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

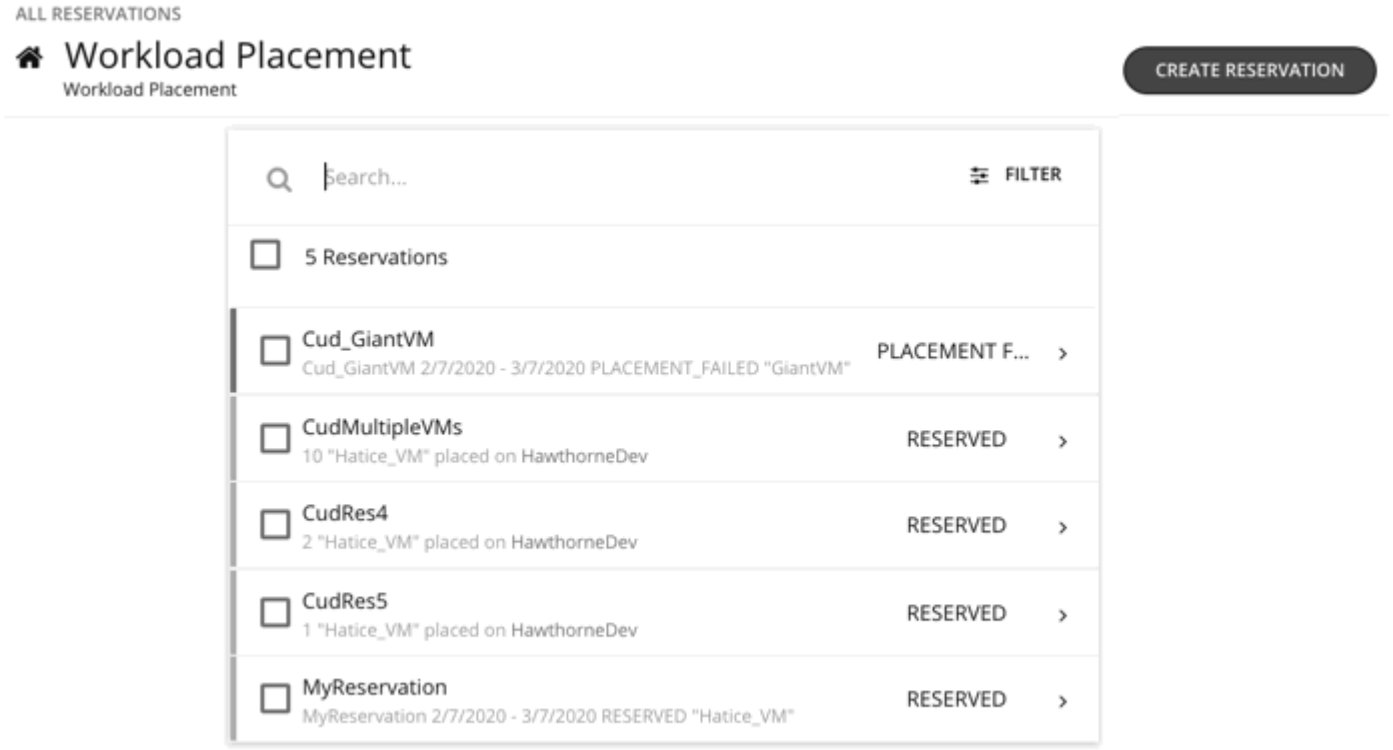
  Set the reservation for some time in the future.

  Workload Optimization Manager does not calculate placement at this time – the future reservation saves the definition, and Workload Optimization Manager 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 currently in effect and to create new reservations, click the **PLACE** button in the Navigation Menu.

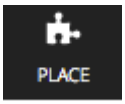


Creating a Reservation

Reservations set aside resources for anticipated workload. While a reservation is in the RESERVED state, Workload Optimization Manager continually calculates placement for the reserved VMs.

To create a reservation:

1. Navigate to the Workload Placement page.

2. Create a new reservation.

   In the Workload Placement page, click **CREATE RESERVATION**.

   Workload Optimization Manager 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 Workload Optimization Manager 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 **Workload Optimization Manager Segments**. Choose which placement policies the reservation will respect.

- Networks

  Workload Optimization Manager 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 Workload Optimization Manager creates to reserve resources in your environment. For example, assume the name *MyReservation*. If you reserve three VMs, then Workload Optimization Manager 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. Workload Optimization Manager begins planning the reservation immediately when you click **CREATE RESERVATION**. The reservation stays in effect for 24 hours – At that time Workload Optimization Manager deletes the reservation.

     – Future Reservation

       This executes the reservation for the date range you specify. Workload Optimization Manager 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, Workload Optimization Manager deletes the reservation.

   When you are finished with the reservation settings, click **CREATE RESERVATION**. Workload Optimization Manager 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

     Workload Optimization Manager is planning the placement of the reservation workloads.

   - FUTURE

     Workload Optimization Manager is waiting for the START DATE before it will start to plan the reservation.

- RESERVED

  Workload Optimization Manager has planned the reservation, and it found providers for all the VMs in the reservation. As your environment changes, Workload Optimization Manager 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

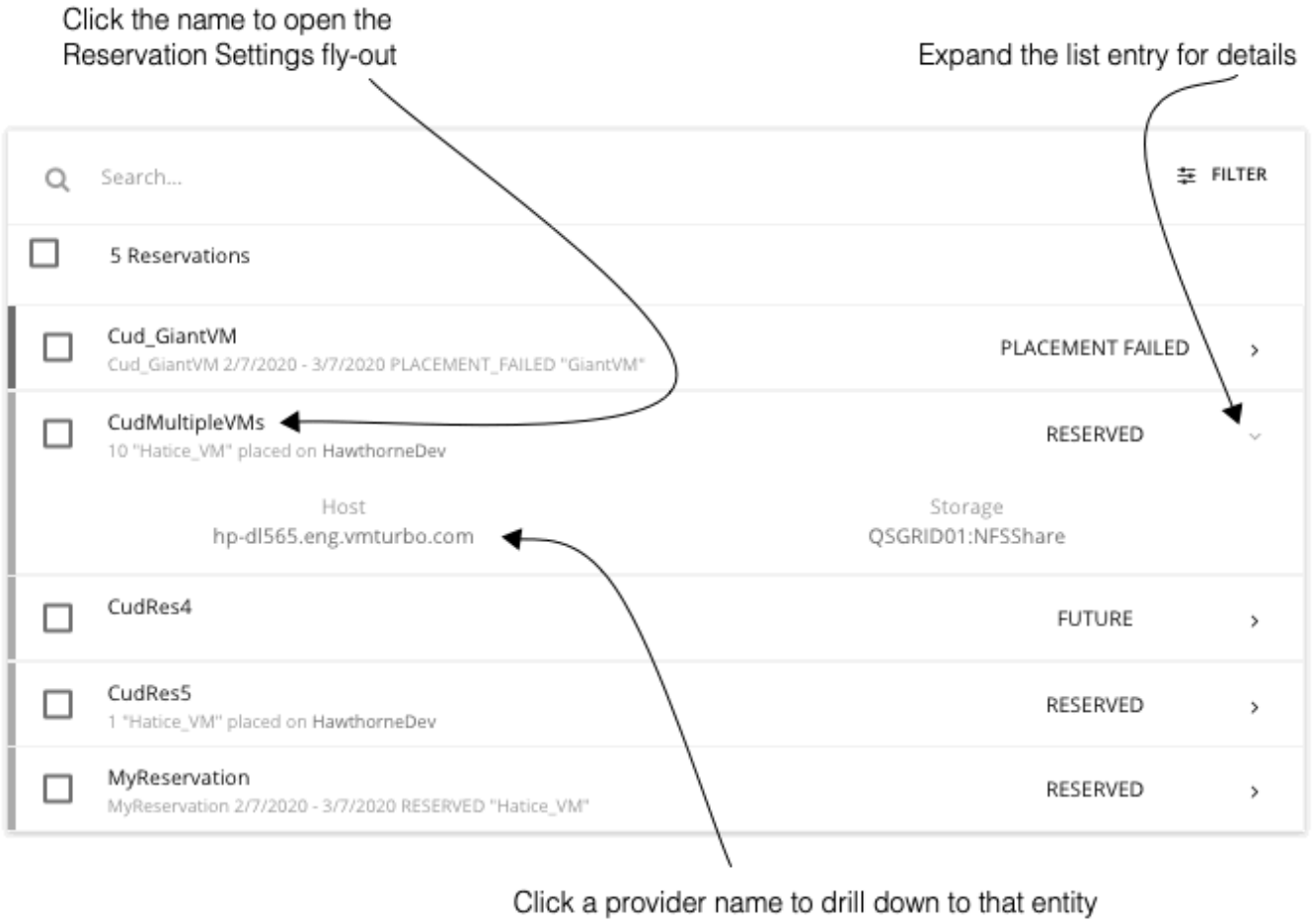
  Workload Optimization Manager cannot place all the reservation VMs. As your environment changes, Workload Optimization Manager continues to calculate placement for the VMs. If at any time it finds that it can place all the VMs, it changes the reservation to RESERVED.

- 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



The PLACE page displays the current list of reservations. You can expand items in the list to see some details, or you can click to view the full details. You can also select items to delete them, which cancels the reservation or deployment.

For an entry in the RESERVED state, you can click the entry name to open the Reservation Settings fly-out.

To delete a reservation, select it in the list and click the DELETE icon.

To see details about the provider entities, or the datacenter that is hosting the reserved VMs, click that entity name.

Deploying Workloads to the Reserved Resources

When you reserve resources, you know that they will be available for you to deploy actual VMs in your environment. To deploy these VMs, you should:

1. Note the placement that your reservation has calculated.

   Expand the reservation entry in the Workload Placement page and note the hosts and storage that will provide resources for your VMs.

2. Delete the reservation.

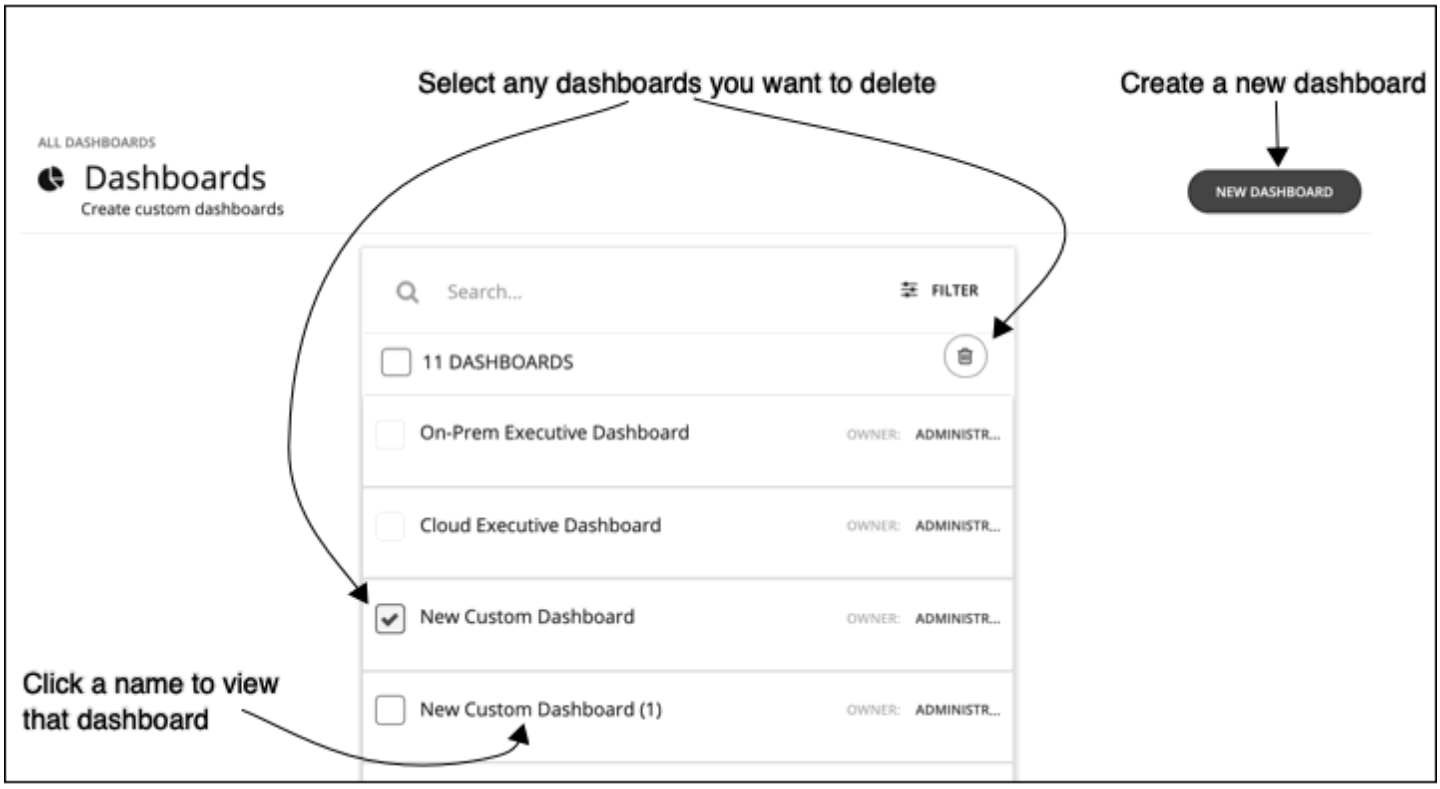
   Before you deploy the reserved VMs, you should delete the reservation. This frees up the Workload Optimization Manager market to manage the placement of the VMs you are about to deploy.

   NOTE:
   When you delete a reservation from the user interface or API, Workload Optimization Manager 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, Workload Optimization Manager will manage their placement the same as it manages the rest of your environment.

Dashboards: Focused Views



Dashboards 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. To view a dashboard, click its name in the list.

Built-in dashboards give you overviews of your on-prem, cloud, and container environments, showing how you have improved your environment over time.

From the Dashboard page, you can also create your own custom dashboards.

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

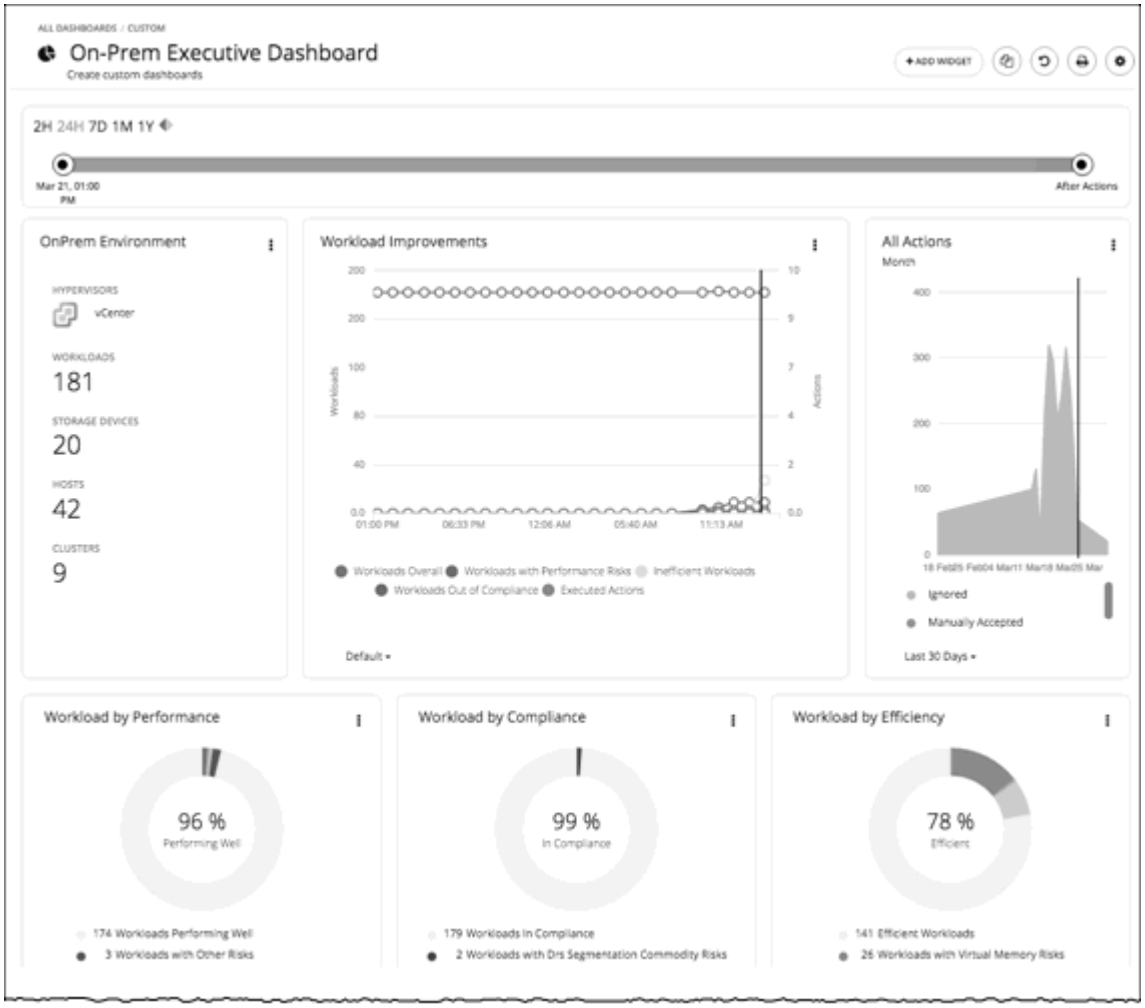
Built-in Dashboards are scorecards of your environment. They demonstrate how well you are improving performance, cost, and compliance, as well as opportunities for further improvements that are available.

Workload Optimization Manager ships with these dashboards:

- On-Prem Executive Dashboard
- Cloud Executive Dashboard
- Container Platform Dashboard

NOTE:
Workload Optimization Manager 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 445)*.

On-Prem Executive Dashboard

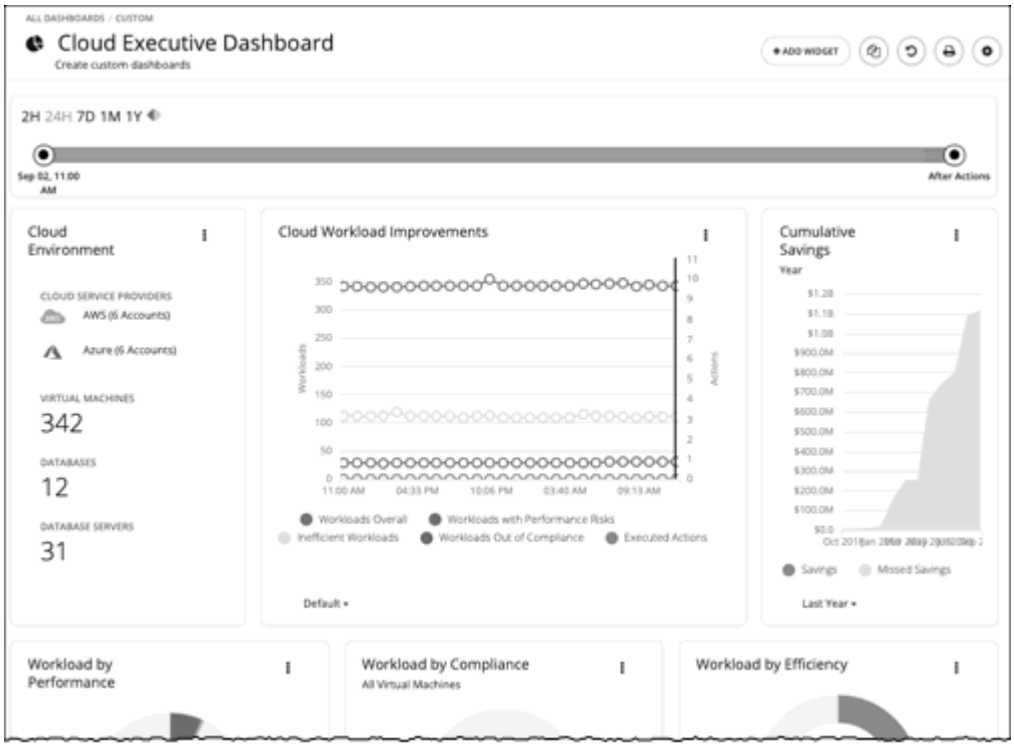


The On-Prem Executive Dashboard shows the overall performance, capacity, and compliance in your on-prem infrastructure. This includes insights into:

- Actions History
  - The **On-Prem Environment** chart widget shows you an overview of your on-prem environment that Workload Optimization Manager is managing and controlling. The chart displays the workloads and the infrastructure that Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 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.

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 Workload Optimization Manager is managing and controlling. The chart displays the workloads, cloud service providers, and cloud accounts that you currently have set up as Workload Optimization Manager 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 Workload Optimization Manager

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. Workload Optimization Manager 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.

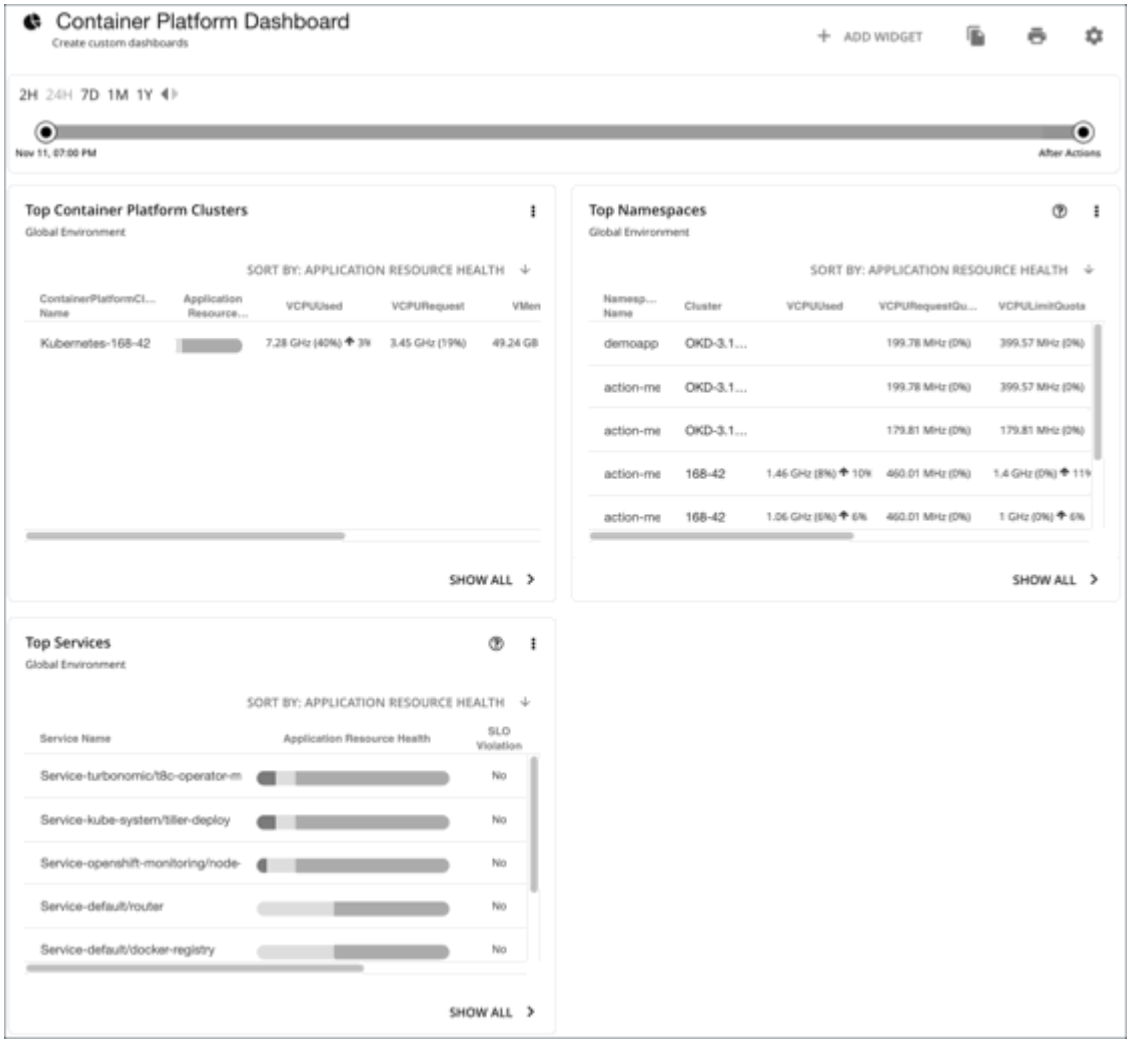
    - Discount Coverage *(on page 497)*

      This chart shows the percentage of VMs covered by discounts. If you have a high percentage of on-demand VMs, you should be able to reduce your monthly costs by increasing coverage. To increase coverage, you scale VMs to instance types that have existing capacity.

    - Discount Inventory *(on page 499)*

      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 Workload Optimization Manager 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.

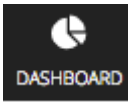
Of course, you can mix and match, according to your needs. You can set any scopes or data sources to the chart widgets in a dashboard to set up whatever organization and focus that you want.

NOTE:
If you set a scope to your Workload Optimization Manager session, the specified scope does not affect your custom dashboards. For information about scoped views, see Working With a Scoped View *(on page 43)*.

Creating a Dashboard

To create a custom dashboard:

1. Navigate to the Dashboards Page.

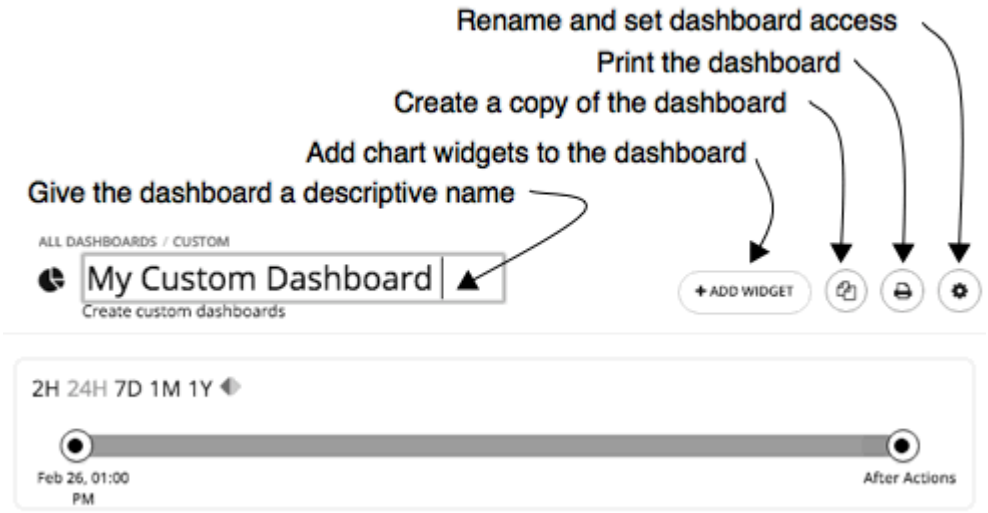
   Click to navigate to the Dashboard Page.

   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 Workload Optimization Manager 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 Workload Optimization Manager 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 448)*.

5. Optionally, set the dashboard access.

   Click **Gear** to change the setting.

   Dashboard access can be:

   - **Only Me** – The dashboard is only available to your Workload Optimization Manager user account.
   - **All Users** – Every Workload Optimization Manager user can see this dashboard.

   By default, access is set to **Only Me**.

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.

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

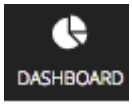
For executive dashboards, only an administrator (username=administrator) can edit an executive dashboard.

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 448)*.

To edit a dashboard's name or change its access settings:

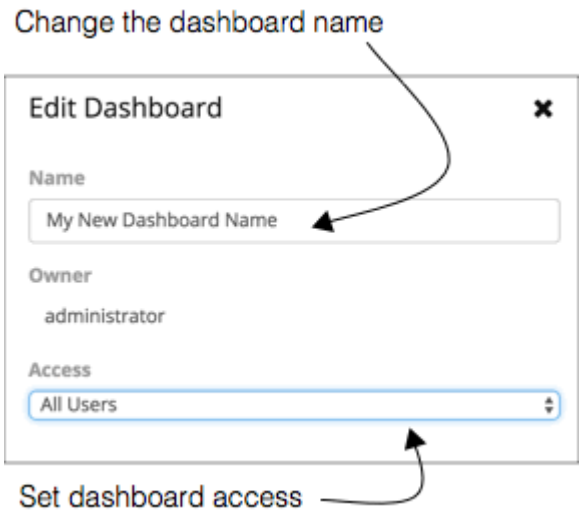
1. Navigate to the Dashboards Page.

   Click to navigate to the Dashboard Page.

2. Click the name of the dashboard that you want to edit.
3. Click **Gear** in the dashboard.

   In the dashboard's Edit fly-out, make your changes.



For the dashboard's access, you can set:

- **Only Me** – The dashboard is only available to your Workload Optimization Manager user account.
- **All Users** – Every Workload Optimization Manager user can see this dashboard.

4. When you are done, close the fly-out panel.

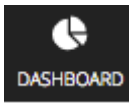
   Your changes take effect when you close the fly-out.

Deleting a Dashboard

If you are an administrator or the dashboard owner, you can delete a custom dashboard. You cannot delete executive dashboards.

To delete a custom dashboard:

1. Navigate to the Dashboards Page.

   Click to navigate to the Dashboard Page.

   This page lists all dashboards that are available to you.

2. Delete one or more dashboards.

   In the list, choose the checkbox for each dashboard you want to delete and click **Trash can**.

Creating and Editing Chart Widgets

Workload Optimization Manager 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 a variety of settings. For example, in the Top Utilized chart widget, if you choose 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 top-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 fly-out opens.

3. Configure the settings for your chart widget.

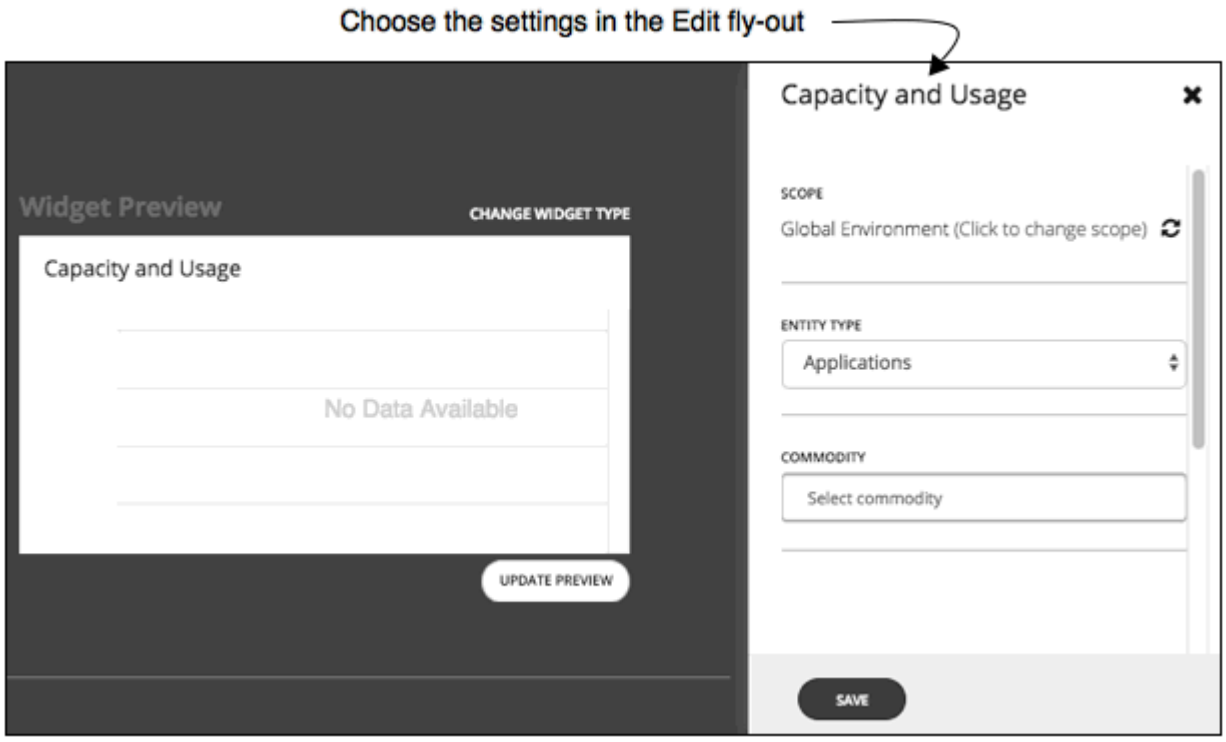
   Chart widget settings determine the data that the chart widget will show.

   In the Edit fly-out, choose the settings and click **Update Preview** to display the result in the Widget Preview pane.

   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 451)*.

   For example:



To delete a chart widget from your dashboard, choose **Delete** in the More options menu at the top-right corner of the chart widget.

Methods to Access Chart Widget Settings

Two methods exist for accessing the chart widget settings in the Edit fly-out:

- You can access the settings in the Edit fly-out 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 at the top-right corner.

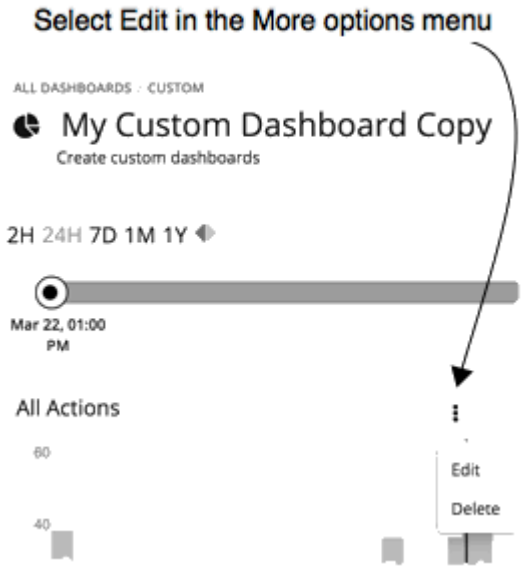


Chart Widget Settings

Chart widget settings vary according to the type of chart widget. Also, depending on the value that you choose for a setting, additional settings may appear. The following is a list of frequently-used chart widget settings:

- Scope

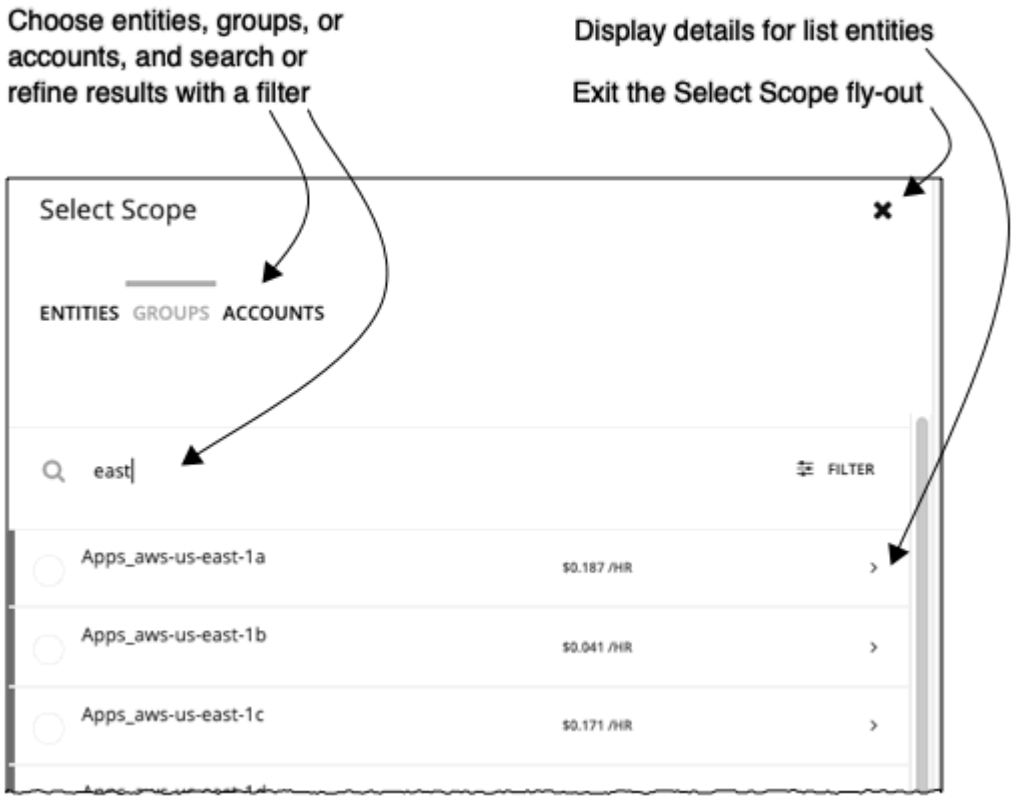
  The set of entities in your environment that this chart widget represents. By default, the chart widget scope is set to **Global Environment**.

  For every type of chart widget, you have the option to set the chart's scope. To do so:

  1. Click **Click to change scope** to open the Select Scope fly-out.
  2. In the Select Scope fly-out, choose the entity, group, or account that you want.

     The ACCOUNTS tab is available depending on the type of chart widget.

     Your choice appears in the **Scope** field.



- 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's 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

Workload Optimization Manager 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 Workload Optimization Manager recommends to improve the current state of your environment.

Chart Type

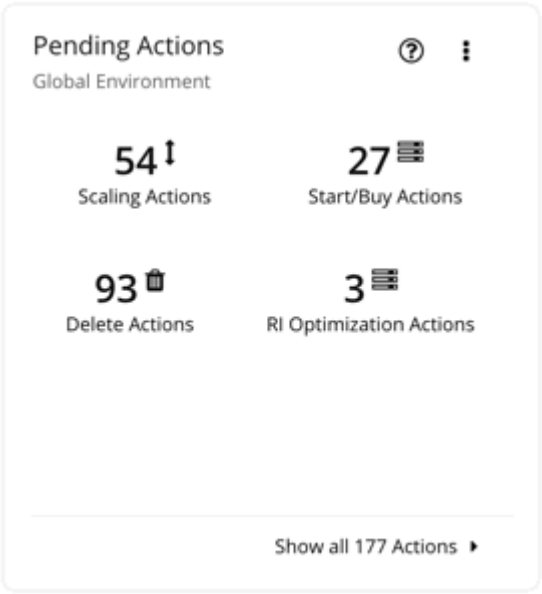
You can set the display to:

- Text
- Ring Chart
- Horizontal Bar
- List

Examples:

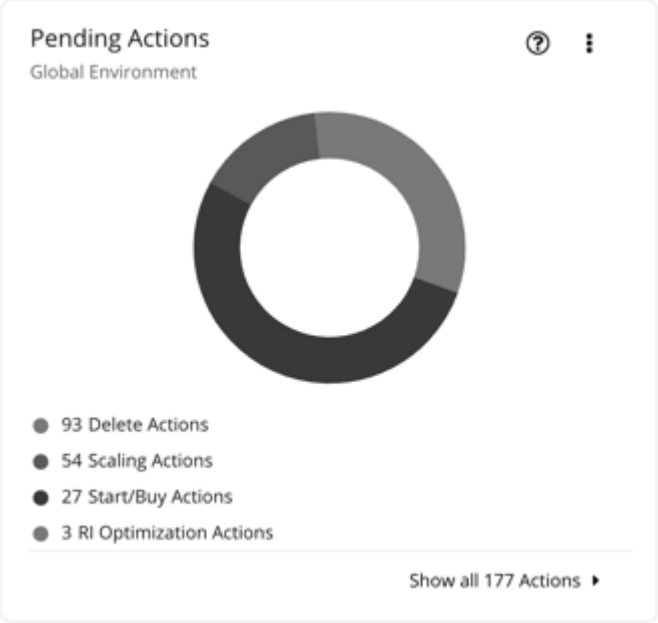
- Text

  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 67)*.



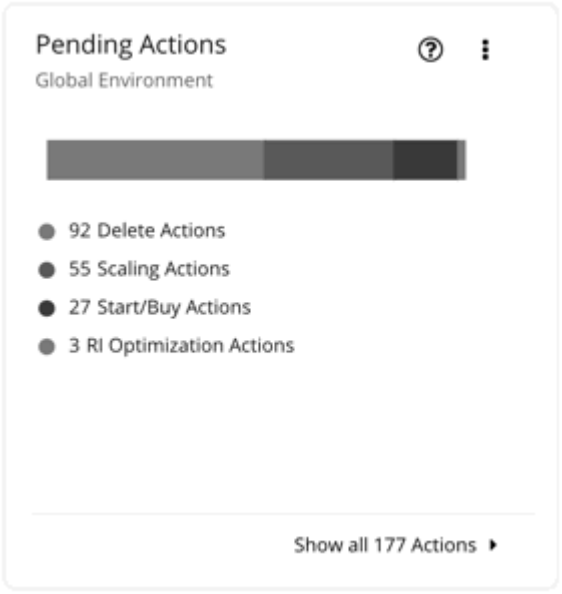
- Ring Chart

  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 67)*.



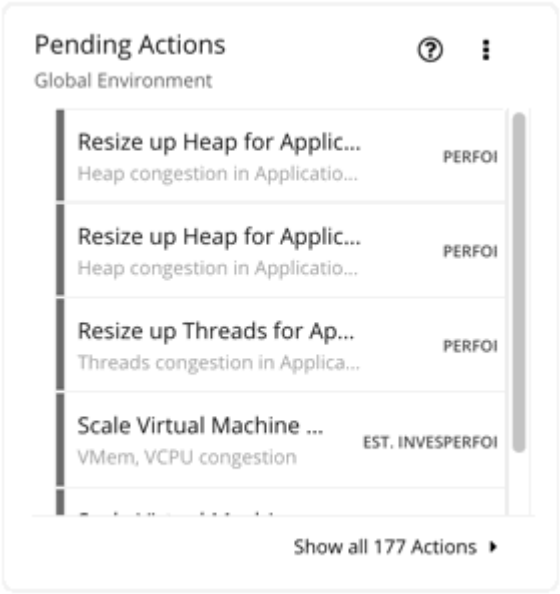
- Horizontal Bar

  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 67)*



- 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 58)*.



At the bottom of the chart, click **Show All Actions** 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 Pending Actions List *(on page 79)*.

Actions Charts

Actions charts keep a history of pending (not executed) and executed actions. These charts use historical data from the Workload Optimization Manager database. You can set the chart to show hourly, daily, or monthly data points.

Filter

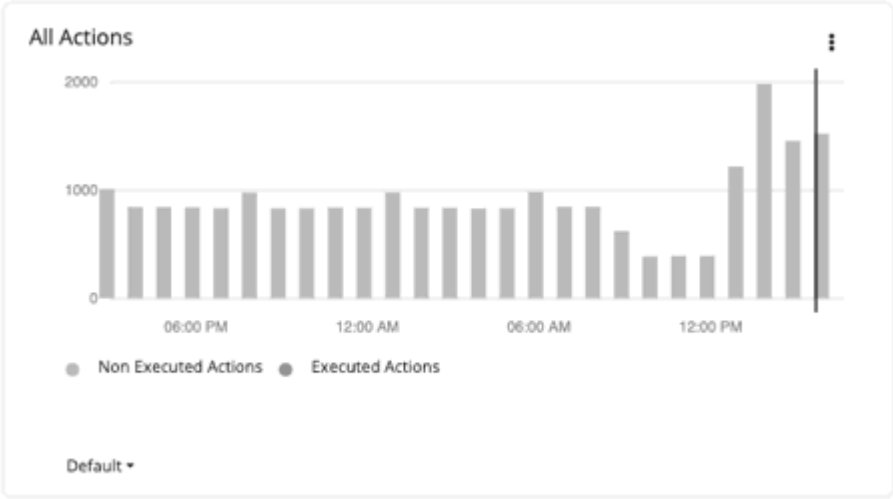
You can filter the chart to show **All Actions** (pending and executed actions) or only **Executed Actions**.

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.

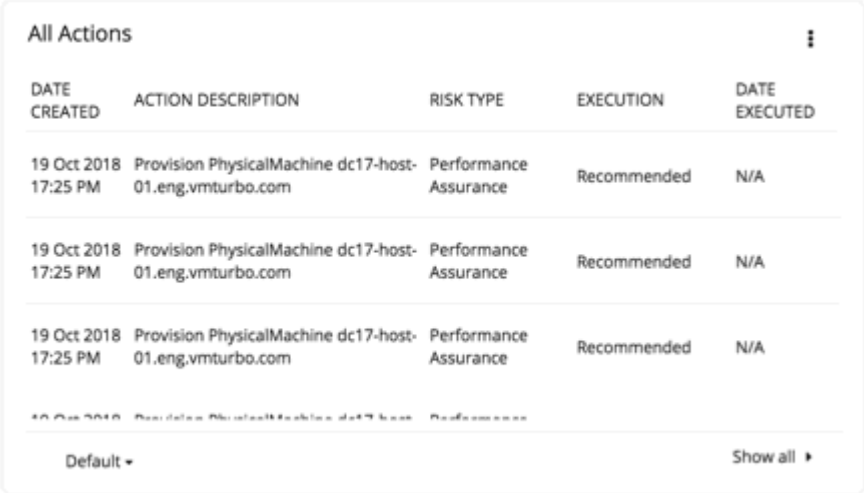
Consider the following scenarios.

- 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 *(on page 79)* of actions, click **Show All** at the bottom of the chart.



Risks Avoided Charts

As you execute the actions Workload Optimization Manager 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

Workload Optimization Manager 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 452)*.

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, Memory, and even network flow between VMs that are on the same host (In-Provider Flow) or on other hosts (In-DPOD or Cross-DPOD Flow).

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.

Example: An Optimized Improvements chart for applications

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 Workload Optimization Manager identifies as a result of its analysis.

For example, if some workloads risk losing performance, Workload Optimization Manager 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 track 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 Workload Optimization Manager 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 201)*.

Type

You can choose **Potential Savings** or **Necessary Investments**.

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

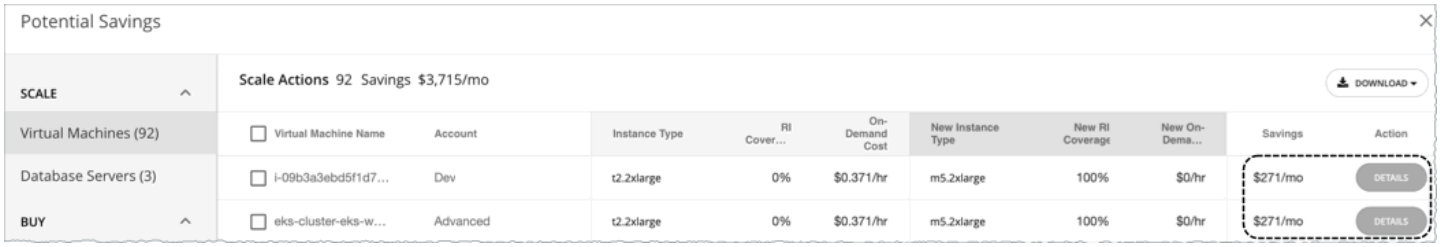
For the ring chart, you can click an action type (for example, **Scale Volumes**) in the chart or legend to display a filtered view of the actions list.

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.



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 their allocated capacity and how much of that capacity is in use.

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 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.

Display

The chart shows the information as Tabular.

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.

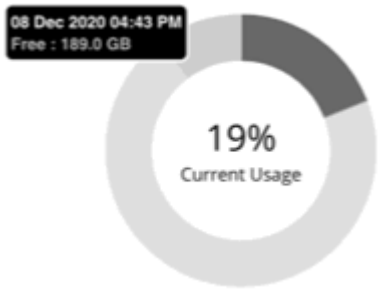
Resources Charts

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, Workload Optimization Manager also calculates *Reserved* capacity based on utilization constraints set in policies.

For example, for a cluster with 100 GB of allocated storage, Workload Optimization Manager 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 Workload Optimization Manager will show 10 GB of reserved capacity.

Options

Choose **Show Utilization** to see averages and peaks/lows, or **Show Capacity** to see averages and peaks/lows 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

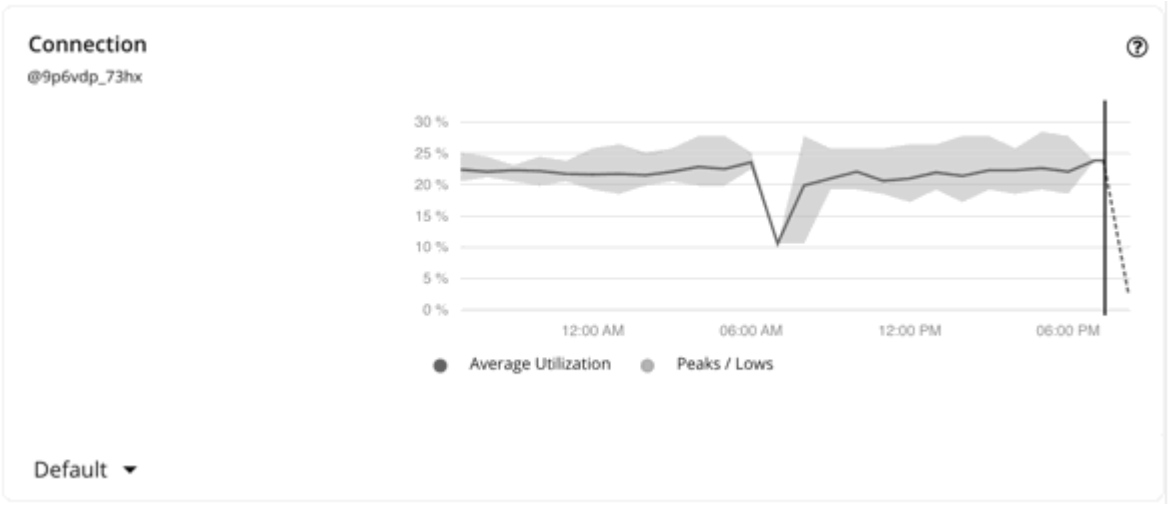
  A line plot showing resource utilization over time. The vertical green bar shows the current moment – Plots that extend to the right project utilization into the future.
- Band Chart

  Lines plot average capacity and average used. The chart shows a band where its thickness indicates peaks and lows.

Connection Chart

Connection is the measurement of Database Server connections utilized by applications.

Workload Optimization Manager collects connection data from Database Servers discovered via Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager 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

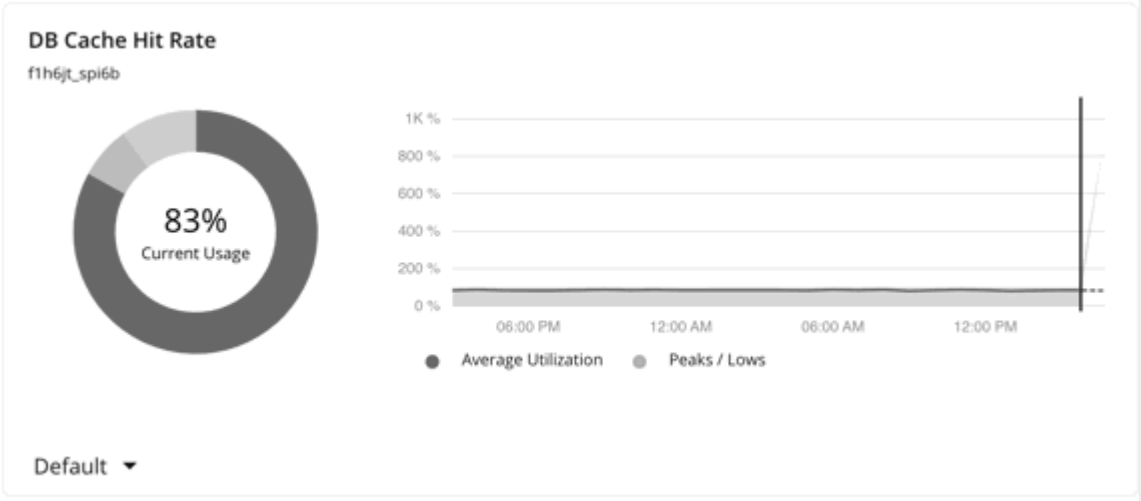
Workload Optimization Manager uses connection data to generate memory resize actions for on-prem Database Servers.

For cloud Database Servers, Workload Optimization Manager uses connection data as a constraint when generating scale actions. For details about scale actions, see Database Server Actions *(on page 230)*.

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.

Workload Optimization Manager collects cache hit rate data from Database Servers discovered via Databases, APM, and Cloud targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager 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, 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. Workload Optimization Manager 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 utilized by a Database Server.

Workload Optimization Manager collects memory data from Database Servers discovered via Databases and APM targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager 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	Oracle, MySQL
MySQL	MySQL
Oracle	Oracle

Target	Supported Database Servers
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. Workload Optimization Manager uses database memory and cache hit rate data to decide whether resize actions are necessary.

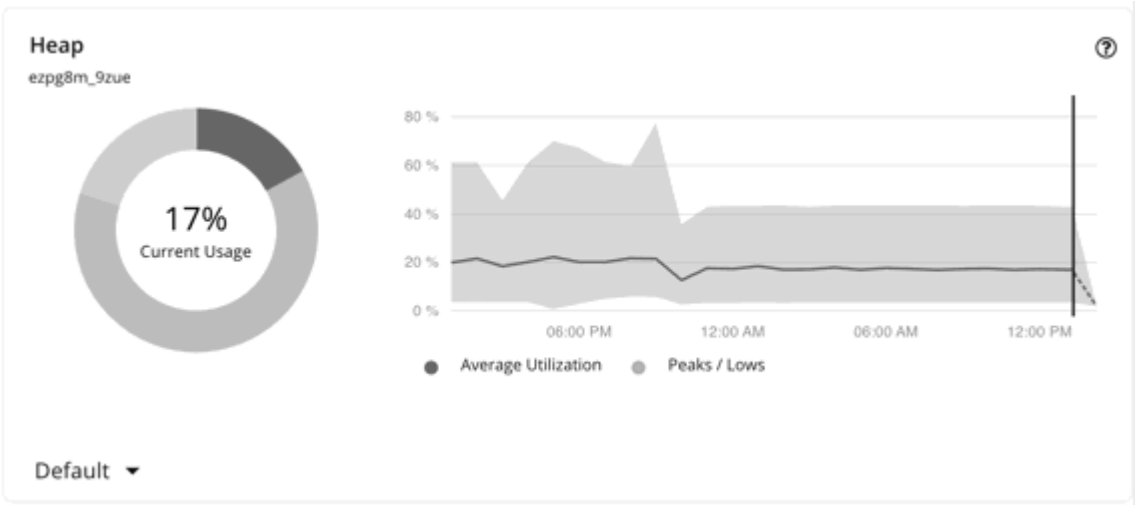
A high cache hit rate value indicates efficiency. The optimal value is 100% for on-prem (self-hosted) Database Servers, and 90% for cloud Database Servers. When the cache hit rate reaches the optimal value, no action generates even if database memory utilization is high. If utilization is low, a resize 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.

Heap Chart

Heap is the portion of a VM or container's memory allocated to individual applications.

Workload Optimization Manager collects heap data from Application Components discovered via Applications and APM targets. When you set the scope to one or several Application Components, the data that Workload Optimization Manager 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

Target	Supported Application Components
Dynatrace	Java applications
Instana	Java applications
JBoss	Java applications
JVM	Java applications
New Relic	Java applications, Node.js
Tomcat	Java applications
WebLogic	Java applications
WebSphere	Java applications

Heap Resize Actions

Workload Optimization Manager 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 Workload Optimization Manager.

> **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 Workload Optimization Manager. These actions adjust the number of replicas to help you meet your SLO goals.

  Or

- Kubernetes Horizontal Pod Autoscaler (HPA) is enabled for a *Deployment*, *ReplicaSet*, or *StatefulSet* that is exposed as a Service. Workload Optimization Manager 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

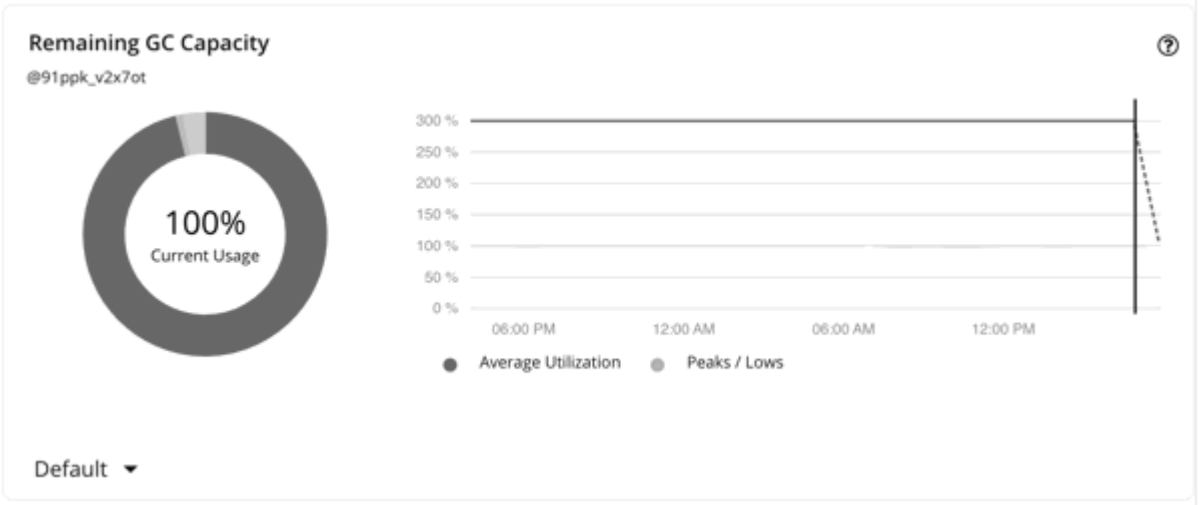
  The number of *ready* pods owned by the Workload Controller. Pods in other states (for example, pending pods) are not counted.

  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).

Workload Optimization Manager collects GC data from Application Components discovered via 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 Workload Optimization Manager 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
Instana	Java applications
JBoss	Java applications
JVM	Java applications
New Relic	Java applications, Node.js
Tomcat	Java applications
WebLogic	Java applications

Target	Supported Application Components
WebSphere	Java applications

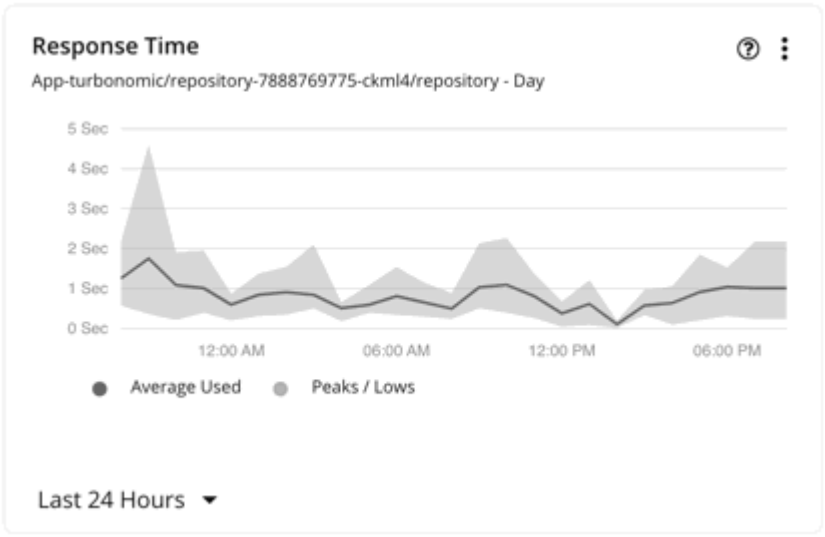
Effect on Heap Resize Actions

Workload Optimization Manager 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 Workload Optimization Manager.

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).

Workload Optimization Manager collects response time data from entities discovered via 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 Workload Optimization Manager 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

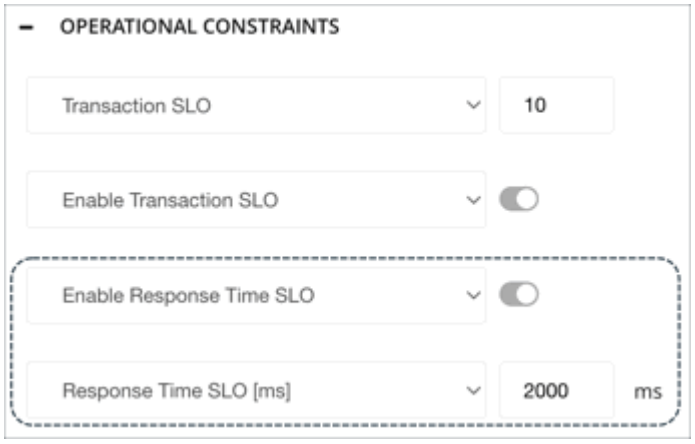
Data is available in the Response Time chart for the following entities:

Target	Supported Entities
AppDynamics	Business Application, Business Transaction, Service, Application Component

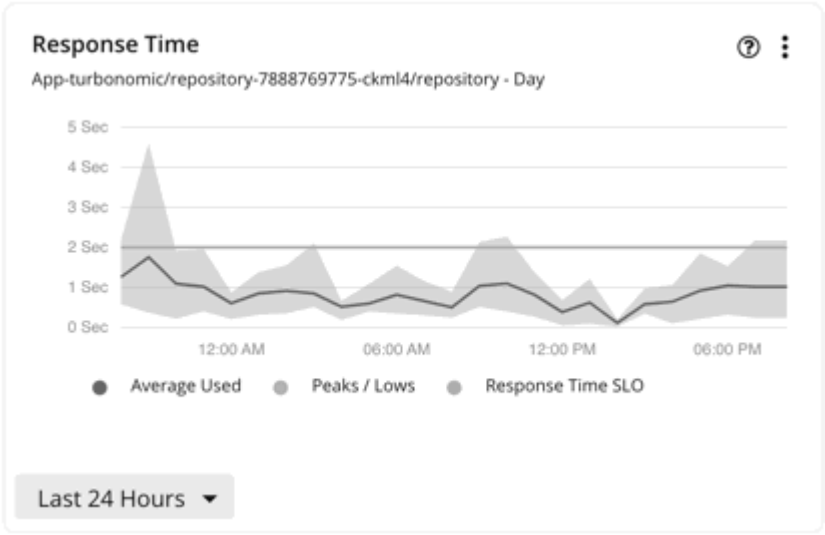
Target	Supported Entities
Dynatrace	Business Application, Service
Instana	Business Application, Business Transaction, Service, Application Component, Database Server
JBoss	Application Component
JVM	Application Component
MySQL	Database Server
New Relic	Business Transaction, Service, Application Component, Database Server
Oracle	Database Server
SQL	Database Server
Tomcat	Application Component
WebLogic	Application Component
WebSphere	Application Component

Response Time SLOs

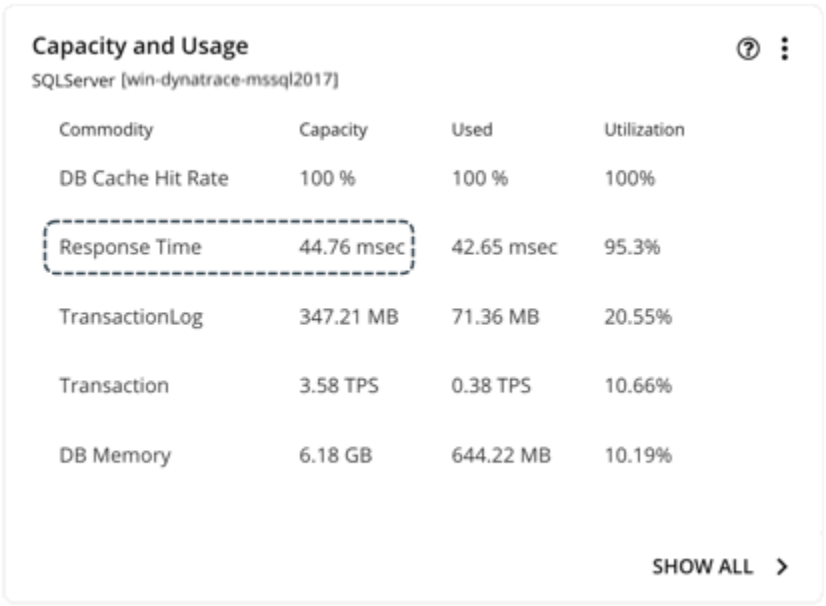
To evaluate the performance of your applications and Database Servers, set Response Time SLOs (Service Level Objectives) as an operational constraint in policies. For applications, you can set the SLO at the Business Application, Business Transaction, Service, or Application Component level.



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, Workload Optimization Manager 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.



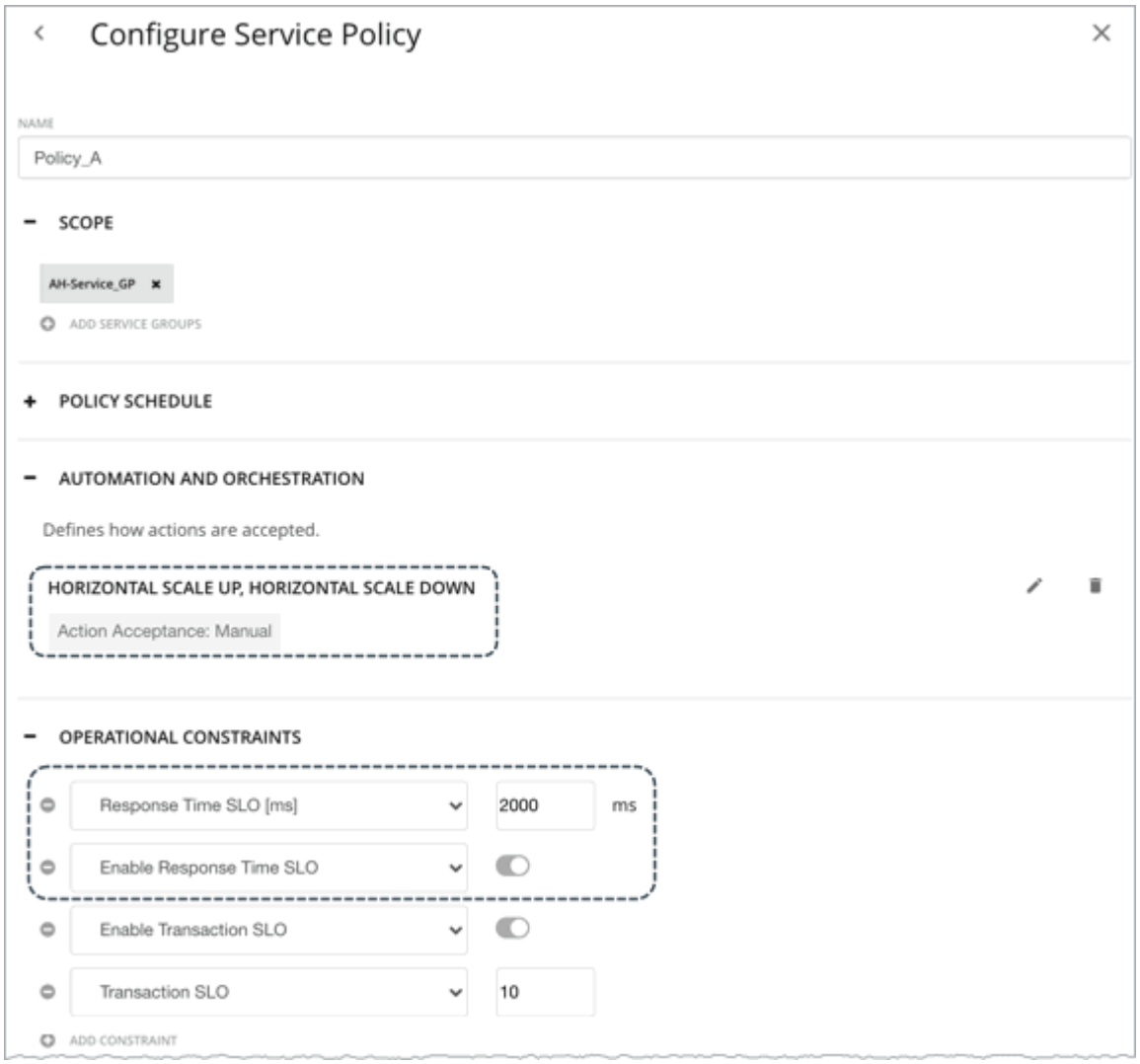
NOTE:
When you set an SLO value, Response Time capacity in the Capacity and Usage chart shows as N/A.

Response Time SLOs for Kubernetes Services

When you add a Kubernetes target, Workload Optimization Manager discovers container platform entities, including Kubernetes Services monitored by AppDynamics, Instana, Dynatrace, and New Relic.

For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications.

To generate actions that adjust pod replicas, Kubernetes Services must be discovered by the Kubeturbo pod that you have deployed in your environment, as well as collect performance metrics via Instana or DIF (Data Ingestion Framework). In addition, Workload Optimization Manager requires that you turn on horizontal scaling and specify Response Time SLOs in policies for the affected Services.



Response Time SLO is the desired *weighted average* response time (in milliseconds) of all Application Component replicas associated with a Service.

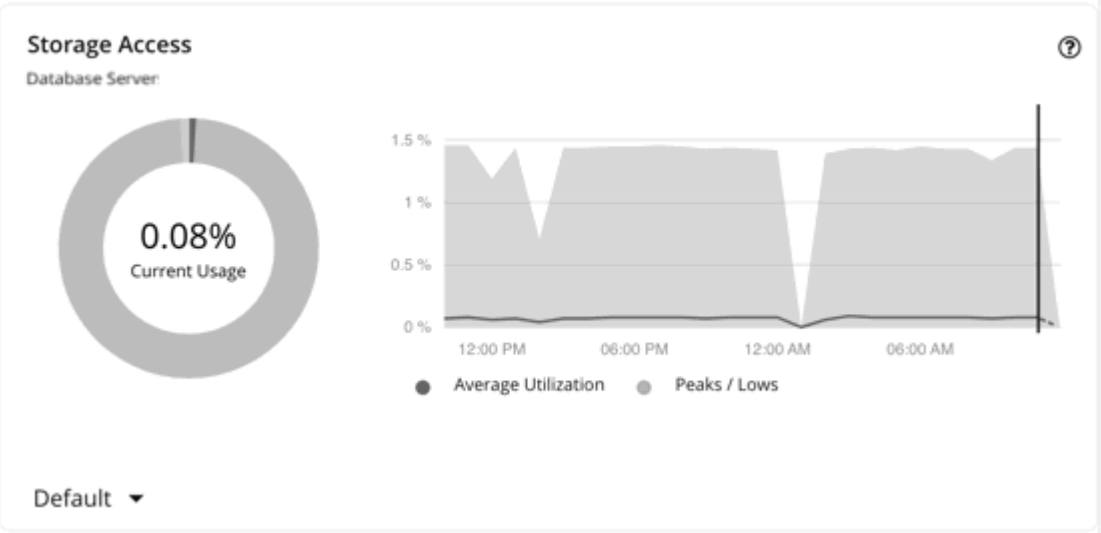
> **NOTE:**
> If you specified SLOs but turned off horizontal scaling in policies, no actions generate but SLO values will continue to display in the Response Time chart for Services, for your reference. This allows you to gauge performance against those SLOs.

For additional information, see Actions for Kubernetes Services *(on page 131)*.

Storage Access Chart

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

Workload Optimization Manager collects storage access data from VMs, Database Servers, and storage entities discovered via cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Workload Optimization Manager 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
	GCP	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

Target Type	Target	Supported Entities
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	Storage, Disk Array

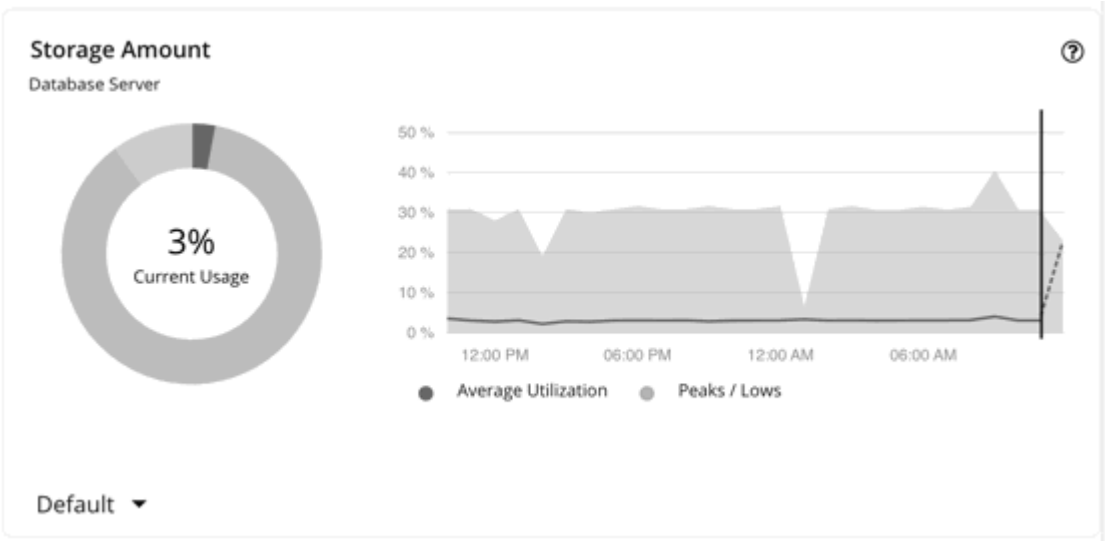
Scale Actions

Workload Optimization Manager considers storage access (IOPS) utilization when recommending scaling actions for cloud VMs *(on page 193)*, cloud Database Servers *(on page 230)*, and volumes *(on page 243)*.

Storage Amount Chart

Storage amount is the measurement of storage capacity utilized by a given entity.

Workload Optimization Manager collects storage amount data from VMs, Database Servers, and storage entities discovered via cloud, on-prem, and storage targets. When you set the scope to these entities, the data that Workload Optimization Manager 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
	GCP	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	Storage, Disk Array, Storage Controller

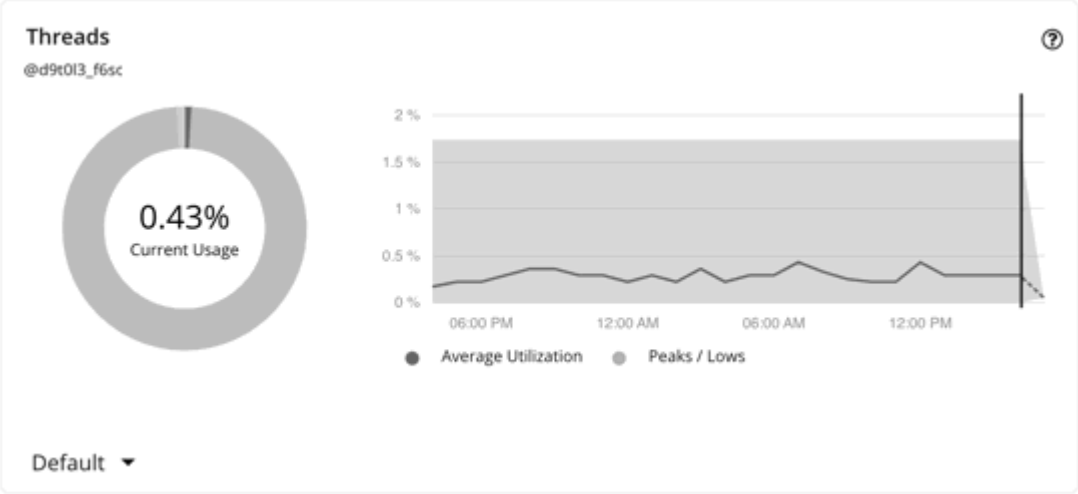
Scale Actions

Workload Optimization Manager can recommend scaling cloud VMs *(on page 193)* and cloud Database Servers *(on page 230)* to instance types that can adequately meet storage demand at the lowest possible cost. For cloud Database Servers, Workload Optimization Manager 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.

Workload Optimization Manager collects thread data from Application Components discovered via Applications and APM targets. When you set the scope to one or several Application Components, the data that Workload Optimization Manager 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
JBoss	Java applications
JVM	Java applications
New Relic	Java applications
Tomcat	Java applications
WebLogic	Java applications
WebSphere	Java applications

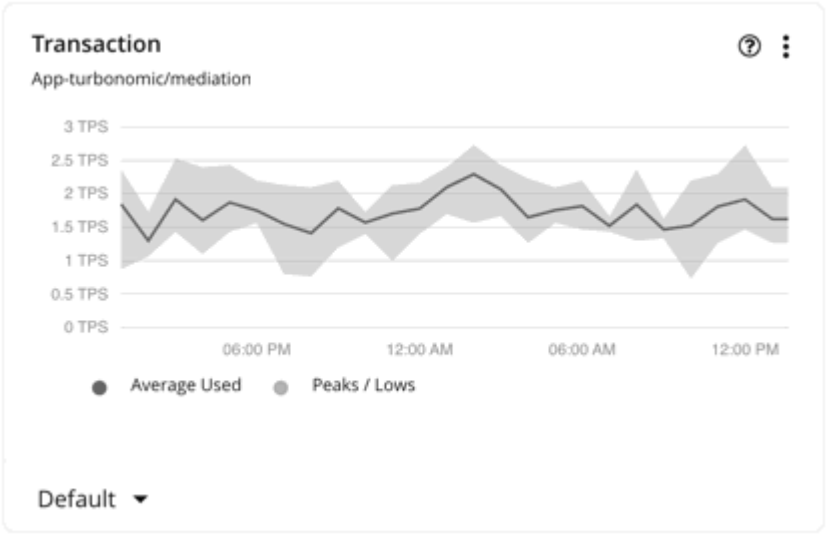
Thread Pool Resize Actions

Workload Optimization Manager generates thread pool resize actions. These actions are recommend-only and can only be executed outside Workload Optimization Manager.

Transaction Chart

Transaction is a value that represents the per-second utilization of the transactions allocated to a given entity.

Workload Optimization Manager collects transaction data from entities discovered via 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 Workload Optimization Manager 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

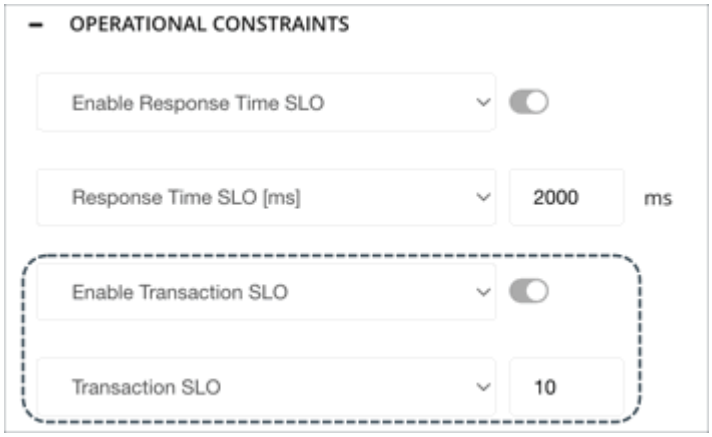
Data is available in the Transaction chart for the following entities:

Target	Supported Entities
AppDynamics	Business Application, Business Transaction, Service, Application Component, Database Server
Dynatrace	Business Application, Service, Database Server
Instana	Business Application, Business Transaction, Service, Database Server
JBoss	Application Component
MySQL	Database Server
New Relic	Business Transaction, Service, Application Component, Database Server

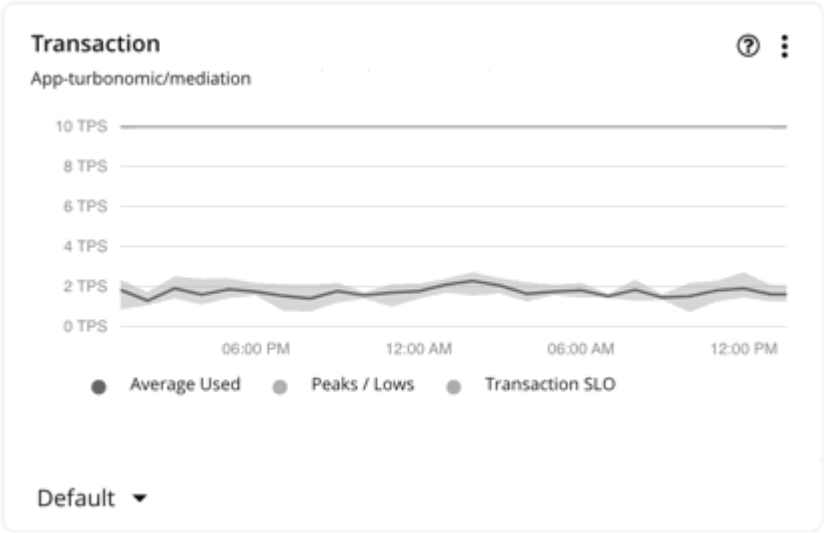
Target	Supported Entities
Oracle	Database Server
SQL	Database Server
Tomcat	Application Component
WebLogic	Application Component
WebSphere	Application Component

Transaction SLOs

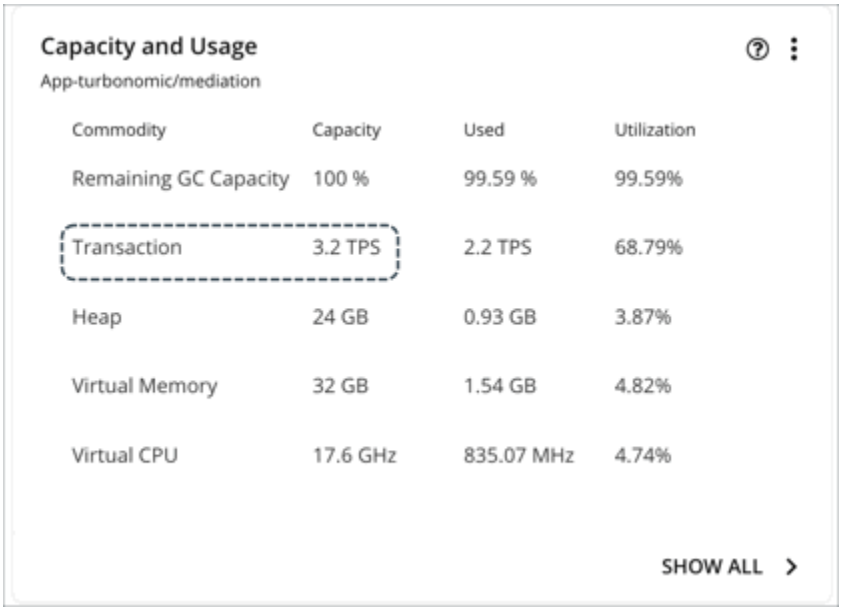
To evaluate the performance of your applications and Database Servers, set Transaction SLOs as an operational constraint in policies. For applications, you can set the SLO at the Business Application, Business Transaction, Service, or Application Component level.



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, Workload Optimization Manager 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.



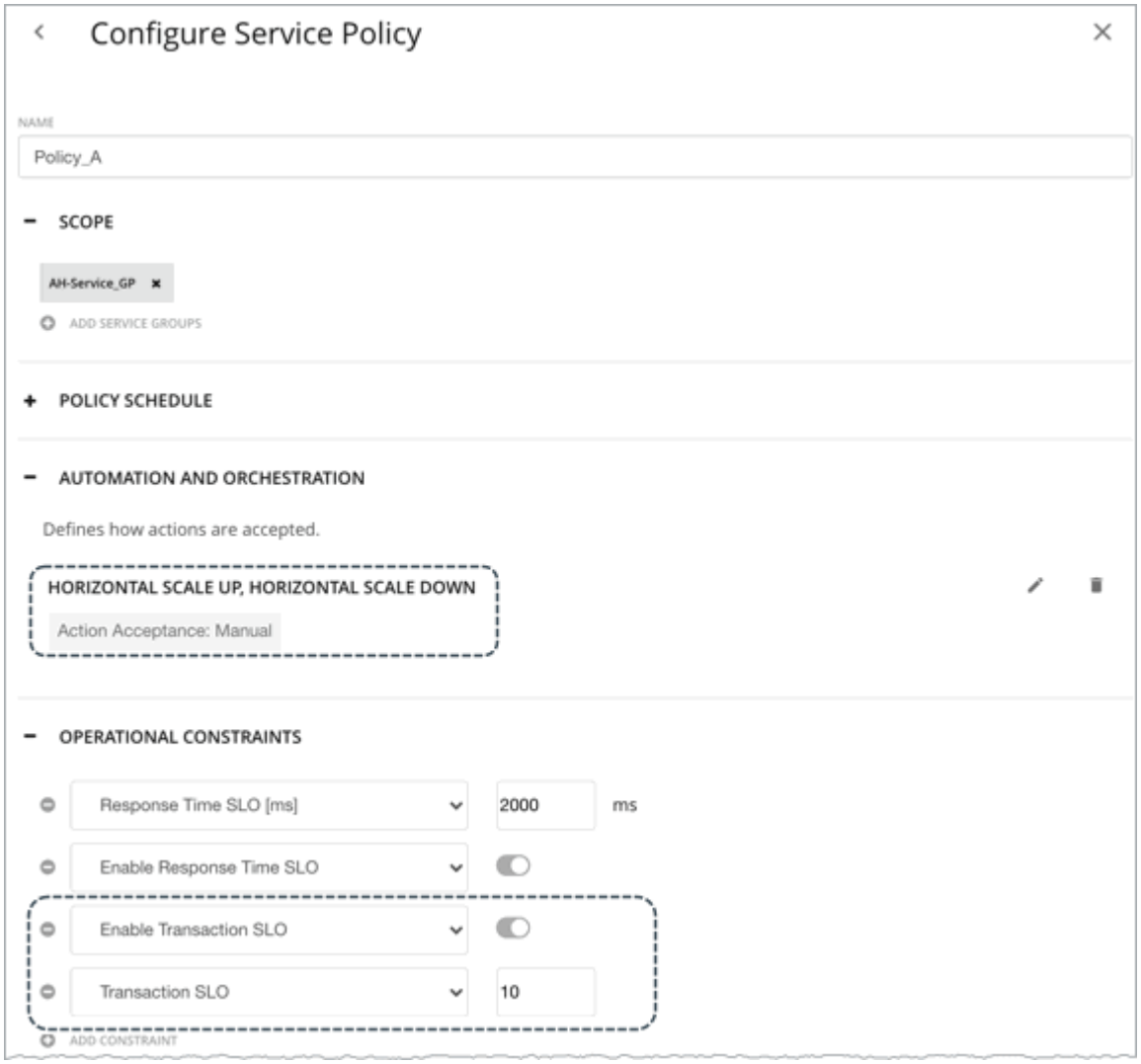
NOTE:
When you set an SLO value, Transaction capacity in the Capacity and Usage chart shows as N/A.

Transaction SLOs for Kubernetes Services

When you add a Kubernetes target, Workload Optimization Manager discovers container platform entities, including Kubernetes Services managed by AppDynamics, Instana, Dynatrace, and New Relic.

For horizontally scalable Kubernetes Services that collect performance metrics (or KPIs) for applications, Workload Optimization Manager can dynamically adjust the number of pod replicas that back those Services to help you meet SLOs (Service Level Objectives) for your applications.

To generate actions that adjust pod replicas, Kubernetes Services must be discovered by the Kubeturbo pod that you have deployed in your environment, as well as collect performance metrics via Instana or DIF (Data Ingestion Framework). In addition, Workload Optimization Manager 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.

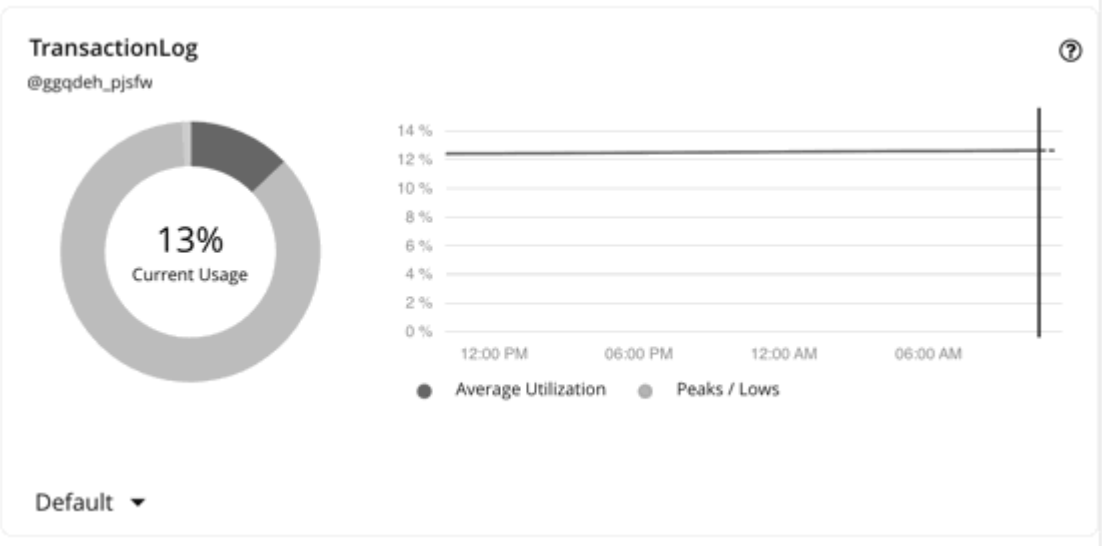
> **NOTE:**
> If you specified SLOs but turned off horizontal scaling in policies, no actions generate but SLO values will continue to display in the Transaction chart for Services, for your reference. This allows you to gauge performance against those SLOs.

For additional information, see Actions for Kubernetes Services *(on page 131)*.

Transaction Log Chart

Transaction log is the measurement of storage capacity utilized by Database Servers for transaction logging.

Workload Optimization Manager collects transaction log data from Database Servers discovered via Databases and APM targets. When you set the scope to one or several Database Servers, the data that Workload Optimization Manager 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
Instana	Oracle
Oracle	Oracle
SQL	SQL

Transaction Log Resize Actions

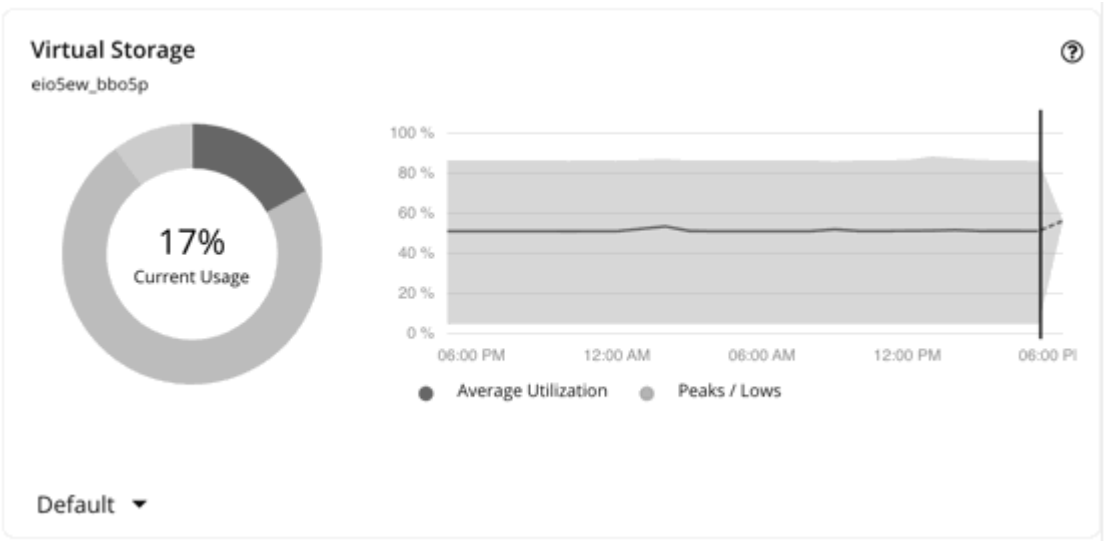
Resize actions based on the transaction log resource depend on support for virtual storage in the underlying hypervisor technology.

Currently, Workload Optimization Manager 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 utilized by VMs.

Workload Optimization Manager collects virtual storage data from VMs discovered via on-prem and APM targets. When you set the scope to one or several VMs, the data that Workload Optimization Manager 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
Guest OS Process	SNMP	Virtual Machine
	WMI	Virtual Machine
Hyperconverged	HyperFlex	Virtual Machine
	Nutanix	Virtual Machine
Hypervisor	Hyper-V	Virtual Machine
	vCenter	Virtual Machine

Virtual Storage Actions

Workload Optimization Manager 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 486)* (public cloud)
- Business Applications
- Business Transactions
- Services
- Application Components
- Zones
- Chassis
- Clusters *(on page 485)* (of hosts)
- Containers
- Container Pods
- Container Specs
- Namespaces
- Workload Controllers
- Data Centers
- Databases
- Database Servers
- Disk Arrays
- IO Modules
- Internet
- Logical Pool
- Networks
- Hosts
- Resource Groups *(on page 486)*
- Regions
- Storage Devices
- Storage Controllers
- Switches
- Virtual Data Centers
- Virtual Machines

- Volumes
- Wasted Files

Data Type

Depending on the entity type (for example, 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.

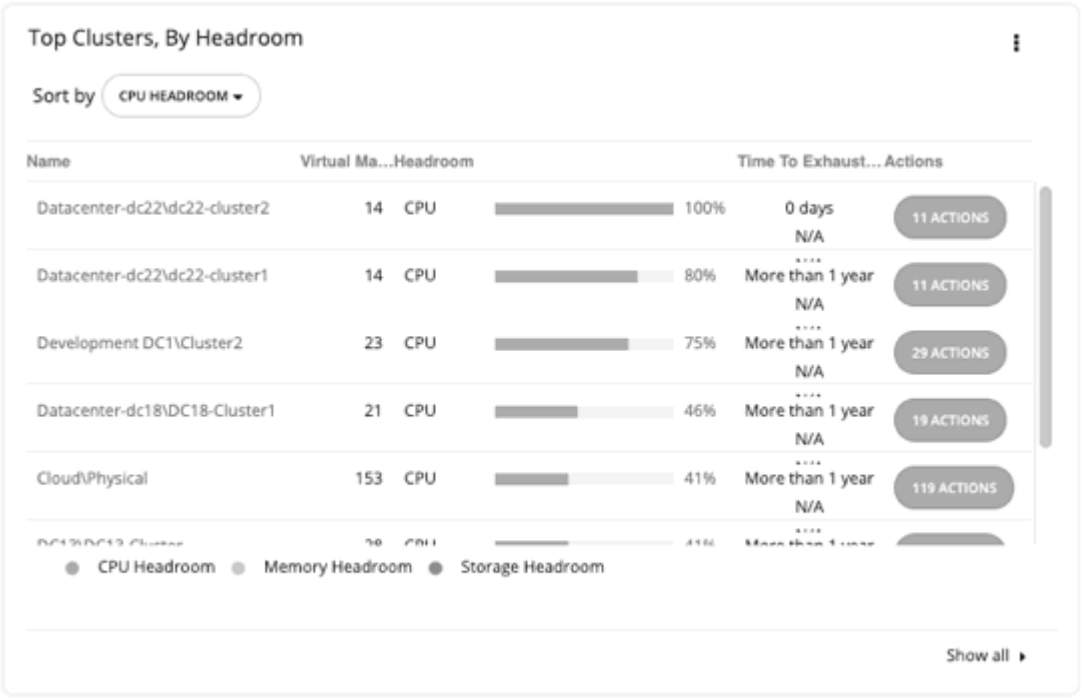
Display

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. For cloud entities, the estimated cost for those entities also display.

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 clusters chart which can be sorted by CPU headroom or CPU exhaustion.



Top Clusters Chart

This chart shows the top 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, Workload Optimization Manager 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 cluster to see the actions that Workload Optimization Manager 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 most pending actions. For each account, you can see 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. You can also click an account name to set the scope to that account.

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, , including AWS GovCloud *(on page 30)* accounts. 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 and EA accounts that you have added as targets, including Azure Government *(on page 30)* subscriptions.

GCP Accounts

The chart shows the folders and projects discovered via the GCP 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, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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. Workload Optimization Manager parses these reports into the formats that it uses for the cloud chart widgets.

> **NOTE:**
> In order for Workload Optimization Manager to access AWS monthly reports, you must have created a Cost and Usage report in your AWS account and you must store it in an S3 bucket.

Billing Breakdown Charts

Billing Breakdown charts show your expenditure on cloud services, so you can track overall cost, cost by region, or cost by cloud accounts. Workload Optimization Manager discovers pricing for cloud services through the cloud accounts and Azure subscriptions that you configured as targets.

Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Expenses Charts

Cisco uses billing information from your cloud provider to track *(on page 27)* compute, storage, database, IP, and data transfer costs for your workloads. Use this chart to manage your costs and see how they evolve over time.

Commodity

You can choose:

- **Billed Cost by Service Provider**

  See costs over time for each cloud service provider that you use in your cloud environment.

  You can open more than one account from a single service provider. If you are running workloads on different service providers, then this chart shows the distribution of costs across them.

- **Top Billed Cost by Account**

  The chart shows the cloud accounts *(on page 486)* 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.

  Currently, when adding the Azure Billing target, the Top Billed Cost by Account, Top Billed Cost by Service, and Top Billed Cost by Service Provider widgets do not display Azure billing data. This will be supported in a future release.

- **Top Billed Cost 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.

  Currently, when adding the Azure Billing target, the Top Billed Cost by Account, Top Billed Cost by Service, and Top Billed Cost by Service Provider widgets do not display Azure billing data. This will be supported in a future release.

- **Workload Cost Breakdown**

  This chart shows costs over time for each component of your cloud utilization. The vertical line indicates when the last data point was polled from your environment. Data points to the right of the vertical line are projections into the future.

  You can see costs for:

  - On-Demand Compute
  - Discounted Compute
  - Spot Compute
  - On-Demand License
  - Reserved License
  - Storage
  - IP (static IPs for workloads)

This chart reflects on-demand costs for VMs based on uptime and other factors. For details about on-demand cost calculations, see Estimated On-demand Monthly Costs for Cloud VMs *(on page 201)*.

Chart Type

You can set the display to:

- Line Chart
- Stacked Bar Chart
- Area Chart

Chart Time Frame

As you change the time frame, Workload Optimization Manager 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 Workload Optimization Manager 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

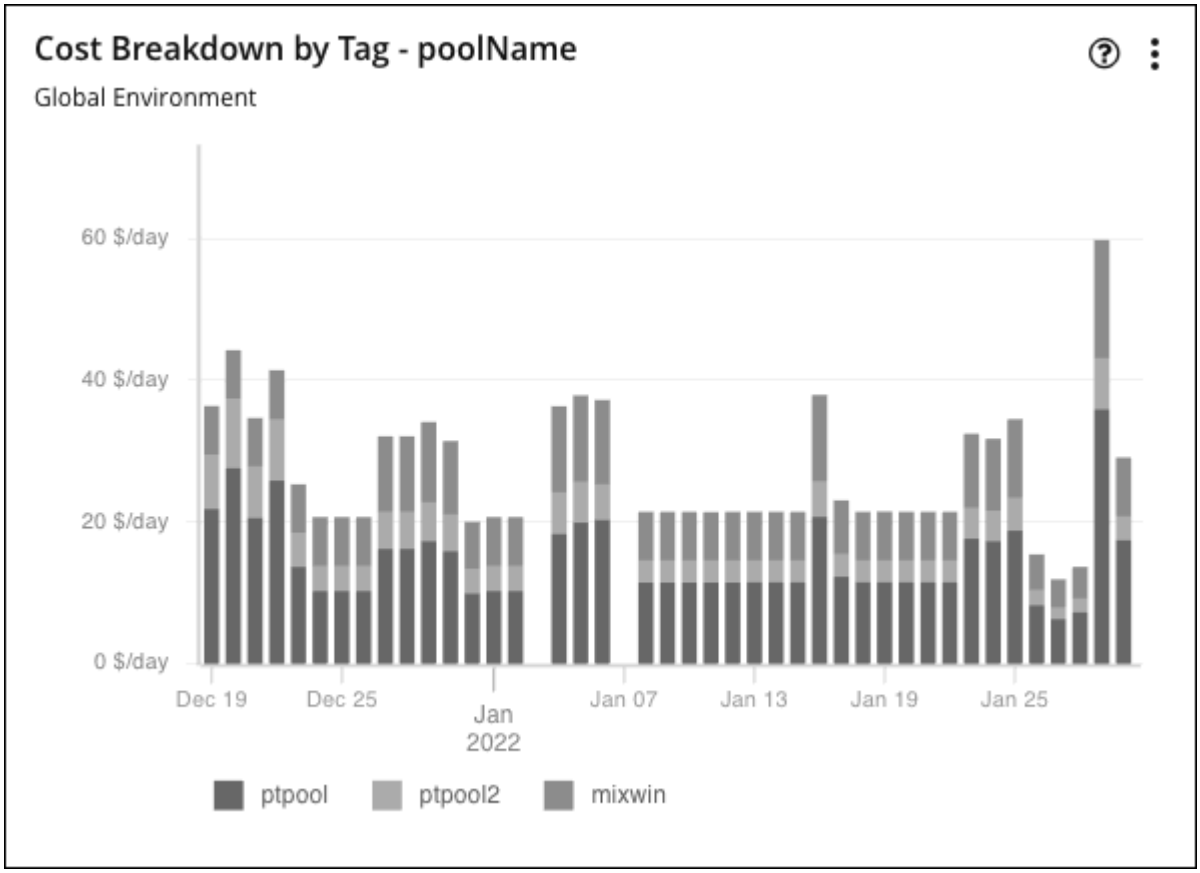
The chart shows the regions in countries in a Map chart.

Cost Breakdown by Tag Charts

Cost Breakdown By Tag charts show the costs for tagged cloud entities that Workload Optimization Manager has discovered in your Azure or GCP 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.

Example: For this stacked bar chart, the tag **poolName** is *workload-type* and the tag **Values** are *ptpool*, *ptpool2*, and *mixwin*.



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

Set the amount of historical data the chart will show.

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. Workload Optimization Manager 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

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), Workload Optimization Manager calculates savings and investment *per workload* based on the hourly cost of the workload price difference, taking into account workload uptime *(on page 197)* 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 temporarily stops for the hours that a workload is offline, and then resumes when the workload is online again and is running with the same configuration.
- Calculation stops for terminated workloads or workloads that Workload Optimization Manager no longer discovers.

Volume Delete Actions

For actions to delete volumes, Workload Optimization Manager 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.

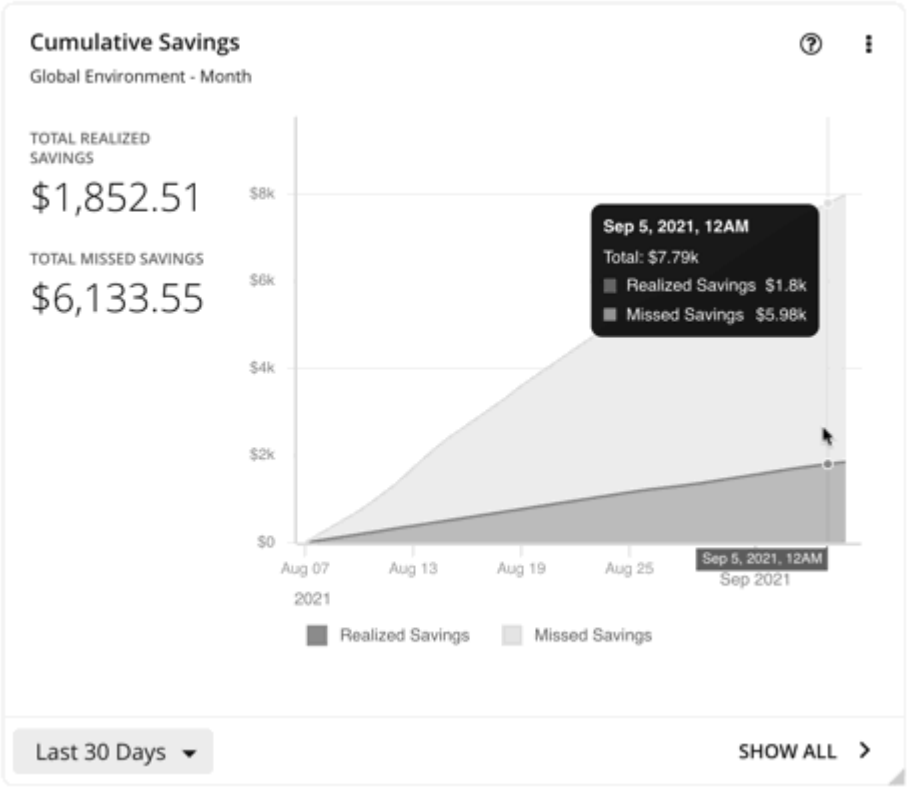
Chart Type

You can set the display to:

- Text and Area Chart
- Area Chart
- Text and Bar Chart
- Stacked Bar 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, Workload Optimization Manager 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

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), Workload Optimization Manager calculates savings and investment *per workload* based on the hourly cost of the workload price difference, taking into account workload uptime *(on page 197)*.

- Calculation temporarily stops for the hours that a workload is offline, and then resumes when the workload is online again and is running with the same configuration.
- Calculation stops for terminated workloads or workloads that Workload Optimization Manager no longer discovers.

Volume Delete Actions

For actions to delete volumes, Workload Optimization Manager 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.

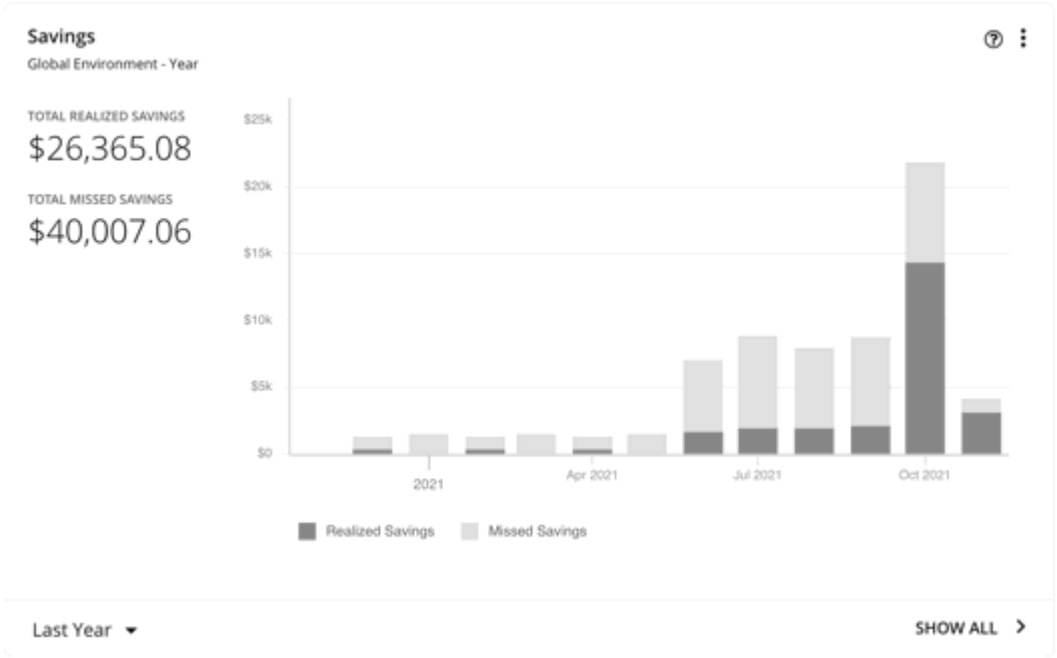
Chart Type

You can set the display to:

- Text and Area Chart
- Area Chart
- Text and Bar Chart
- Stacked Bar 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.



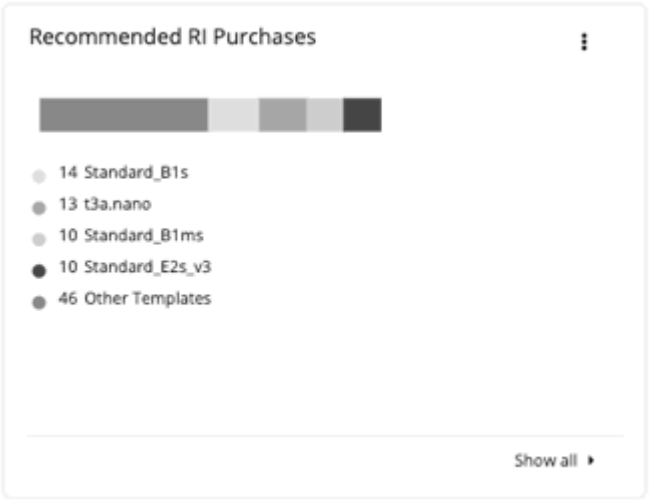
In the chart legend, you can click **Realized Savings** or **Missed Savings** to display a filtered view. Click the item again to reset the chart.

Click **Show All** at the bottom of the chart to view and download data in tabular format.

Recommended RI Purchases Charts

Workload Optimization Manager 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 403)*.
>
> Currently, Workload Optimization Manager can recommend purchase actions for AWS and Azure. Purchase actions for GCP will be introduced in a future release.



Factors Affecting Recommendations

To identify VMs that are good candidates for discounted pricing, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Workload Optimization Manager can recommend purchasing additional discounts.

Workload Optimization Manager 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 530)*.
- Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager 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 RIs:
  - For environments that use the *Instance Size Flexible* rules, Workload Optimization Manager 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, Workload Optimization Manager can recommend that you buy four t2.nano RIs to offer an equivalent discount.
  - For environments that consolidate billing into Billing Families, Workload Optimization Manager recommends purchases that are within the given billing family. For more information, see AWS Billing Families *(on page 535)*.

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.

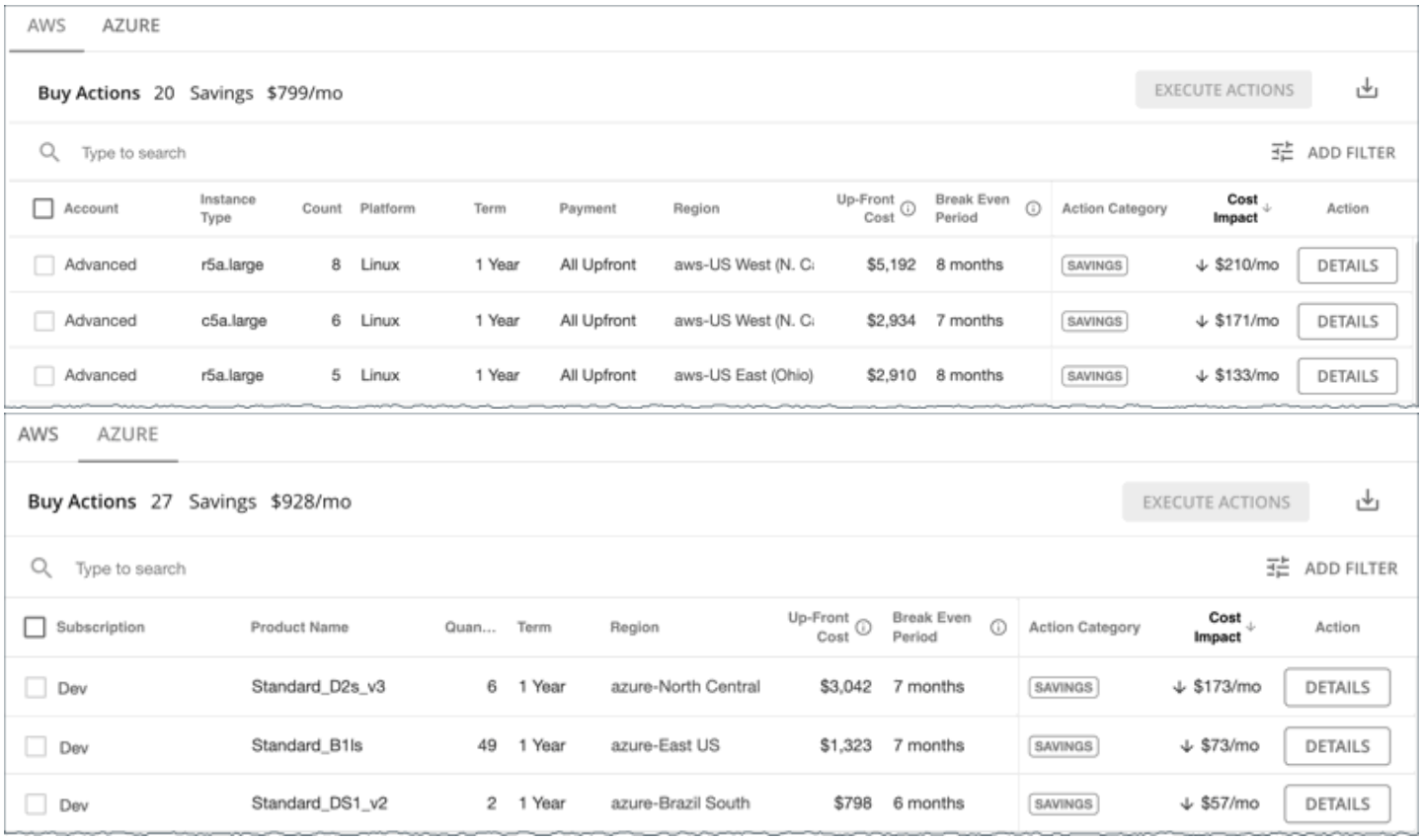


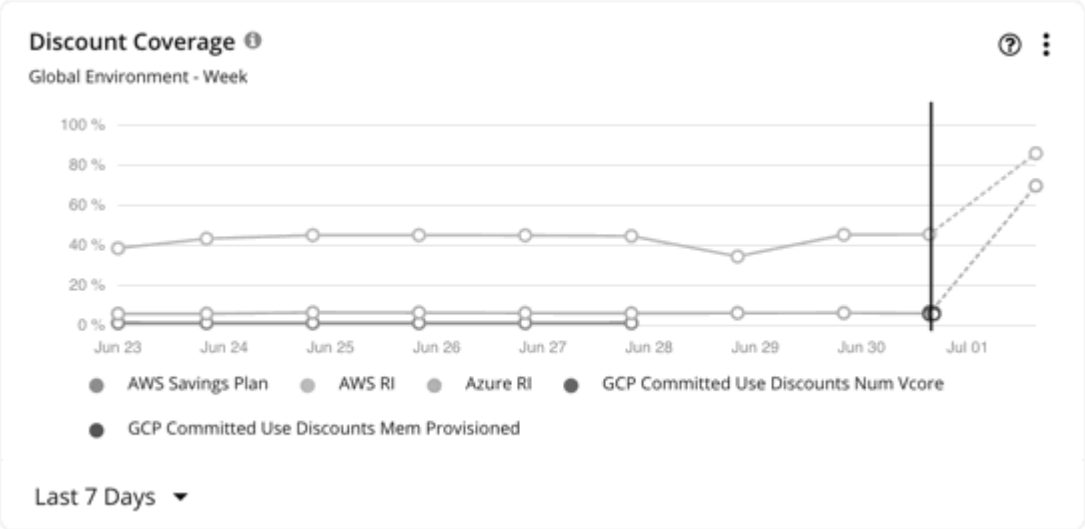
Chart Type

You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

Discount Coverage Chart

This chart shows the percentage of VMs covered by discounts. If you have a high percentage of on-demand VMs, you should be able to reduce your monthly 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, Workload Optimization Manager 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 Workload Optimization Manager 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 Workload Optimization Manager considers it stable.

If the current discount inventory cannot support the VM, or if supporting it would exceed your desired coverage, then Workload Optimization Manager can recommend purchasing additional discounts.

AWS RIs

If you set the scope to a specific AWS account, the chart shows the RI coverage for the workloads for the account, plus any RIs for the billing family.

The data point on the solid vertical line shows the latest data that was polled from your environment. Data points to the left of the vertical line represent historical data, while data points to the right are projections into the future.

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 normalization factors

  Normalization factor is a measure of RI capacity that you can use to compare or combine the capacity for different instance families.

  Workload Optimization Manager 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 Workload Optimization Manager scope. Each workload is assigned normalized units depending on its instance type.

AWS Savings Plans

To view data for AWS Savings Plans, you must:

- Set the scope to your global environment.
- Choose a timeframe that shows daily or monthly data points (Last 7 Days, Last 30 Days, or Last Year)

  AWS measures your Savings Plans commitment in $/hour, but shows daily and monthly costs in Cost Explorer. Workload Optimization Manager polls Cost Explorer periodically to obtain the latest cost data, and then uses that data to calculate

Savings Plans utilization or coverage *per day*. For this reason, you will not see Savings Plans data if you choose a timeframe that shows hourly data points (Last 2 Hours or Last 24 Hours).

NOTE:
AWS timestamps data in UTC, but the chart presents data in your local time. This difference could result in the appearance of a missing day in the chart, but has no effect on data completeness (the chart always reflects the complete data set).

The chart shows historical data, represented by data points to the left of the solid vertical line. Data for the current day is not available since the latest data that AWS provides is always a few days old. In addition, Workload Optimization Manager does not project coverage into the future.

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

Azure Reservations

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.

The data point on the solid vertical line shows the latest data that was polled from your environment. Data points to the left of the vertical line represent historical data, while data points to the right are projections into the future.

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 Workload Optimization Manager scope. Each workload is assigned reservation units based on its instance type.

GCP Committed Use Discounts

The chart displays the latest data that was polled from your environment, but does *not* show historical data or project coverage into the future.

Hover on a data point in the chart to see the following information:

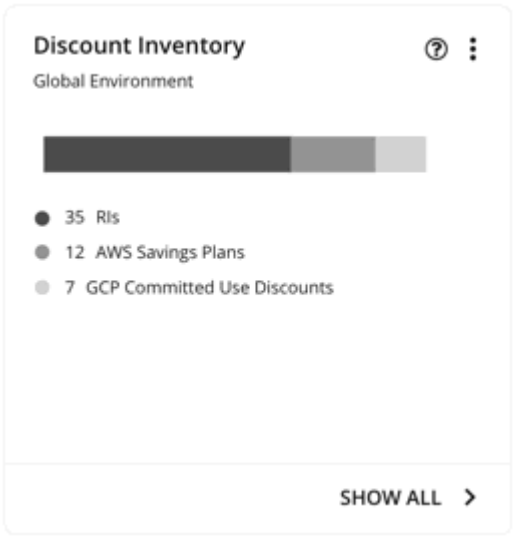
- The date and time for the data point
- The percentage of coverage

Discount Inventory Chart

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

- AWS Reserved Instances (RIs) and Savings Plans for regular and GovCloud *(on page 30)* workloads
- Azure reservations for regular and Azure Government *(on page 30)* workloads
- GCP committed use discounts

Chart Type

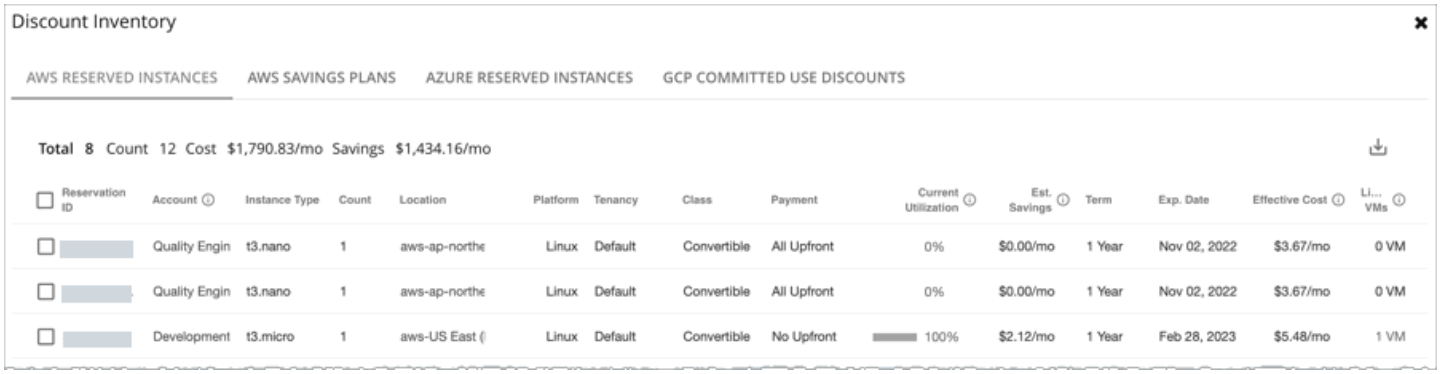


You can set the display to:

- Text
- Ring Chart
- Horizontal Bar

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.



Each row in the table corresponds to a discount. Note that there can be several discounts for an Azure subscription or AWS/GCP 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 RIs

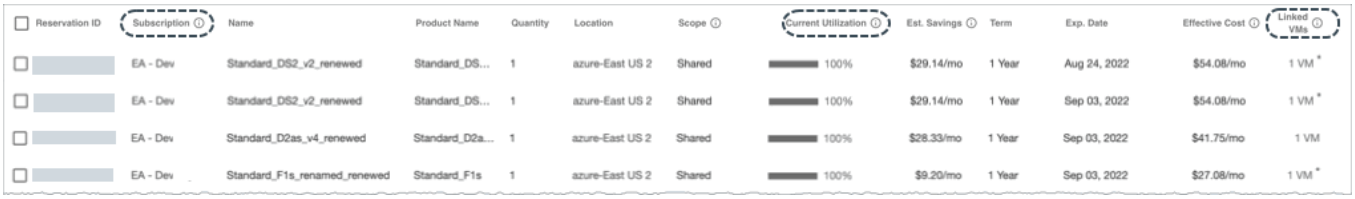
When you add an Azure EA account or an AWS master account as your primary cloud target, Workload Optimization Manager gains full insight into the discounts for your billing families. Even as you selectively add Azure subscriptions or AWS member accounts as secondary targets, Workload Optimization Manager 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 master account, Workload Optimization Manager will not reflect discounts for member accounts that you have not added as targets.
- For Azure:
  - It could take Workload Optimization Manager up to a day to discover newly purchased Azure reservations.
  - There can be delays in billing information updates that Azure makes available to Workload Optimization Manager. 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):



Reservation ID	Subscription	Name	Product Name	Quantity	Location	Scope	Current Utilization	Est. Savings	Term	Exp. Date	Effective Cost	Linked VMs
	EA - Dev	Standard_DS2_v2_renewed	Standard_DS...	1	azure-East US 2	Shared	100%	$29.14/mo	1 Year	Aug 24, 2022	$54.08/mo	1 VM *
	EA - Dev	Standard_DS2_v2_renewed	Standard_DS...	1	azure-East US 2	Shared	100%	$29.14/mo	1 Year	Sep 03, 2022	$54.08/mo	1 VM *
	EA - Dev	Standard_D2as_v4_renewed	Standard_D2a...	1	azure-East US 2	Shared	100%	$28.33/mo	1 Year	Sep 03, 2022	$41.75/mo	1 VM
	EA - Dev	Standard_F1s_renamed_renewed	Standard_F1s	1	azure-East US 2	Shared	100%	$9.20/mo	1 Year	Sep 03, 2022	$27.08/mo	1 VM *

  - **Subscription** column (for Azure) or **Account** column (for AWS)

    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, Workload Optimization Manager shows it as a *non-added* account in the Discount Inventory page.

  - **Current Utilization** column

    This column shows the percentage of discount capacity currently used by VMs in all accounts. Workload Optimization Manager estimates the percentage if there are VMs in non-added accounts that use the discount (since the exact number of VMs is unknown).

  - **Linked VMs** column

    This column shows how many VMs in the account use the discount. Click the number to view the VMs.

    A star symbol (*) after the number indicates that there are VMs from *non-added* accounts that also use the discount. Since you have not added those accounts, the exact number of VMs is unknown.

- For discounts in *non-added* accounts:

Reservation ID	Subscription	Name	Product Name	Quantity	Location	Scope	Current Utilization	Est. Savings	Term	Exp. Date	Effective Cost	Linked VMs
	9224a	Standard_F1_r...	Standard_F1	1	azure-East US 2	Shared	0%	$0.00/mo	1 Year	Aug 24, 2022	$27.08/mo	0 VM
	9224a	Standard_F1s_...	Standard_F1s	1	azure-East US 2	Shared	0%	$0.00/mo	1 Year	Aug 24, 2022	$27.08/mo	0 VM
	acbc8	Standard_D2s_...	Standard_D2a_v4	2	azure-East US 2	acbc86dd-2d	38%	$21.25/mo	1 Year	Aug 24, 2022	$83.50/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. Workload Optimization Manager is aware of this account and if the given discount is shared with other accounts because you have added a master or EA account.

- **Current Utilization** column

  This column shows the percentage of discount capacity currently used by VMs in all accounts. Workload Optimization Manager estimates the percentage if there are VMs in non-added accounts that use the discount (since the exact number of VMs is unknown).

- **Linked VMs**

  If the number under this column is 1 or more, the number indicates VMs from *added accounts* that use the discount. Click the number to view the VMs.

  If the number is 0 (zero), then the discount is currently not used anywhere.

  A hyphen symbol (-) indicates that there are VMs from other *non-added* accounts that also use the discount. Since you have not added those accounts, 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, Workload Optimization Manager uses this chart to present the Savings Plans that it discovered in your cloud environment (including GovCloud *(on page 30)*) and the instance types they use.

Savings Plan ID	Account	Type	Payment	Instance Family	Location	Commitment	Term	Start Date	Exp. Date
55555 ...	Prod	Compute	All Upfront	All	All	$0.001/hr	1 Year	Dec 28, 2020	Dec 28, 2021
45555 ...	Prod	EC2	No Upfront	t3	aws-...	$0.001/hr	3 Years	Dec 29, 2020	Dec 29, 2023

GCP Committed Use Discounts

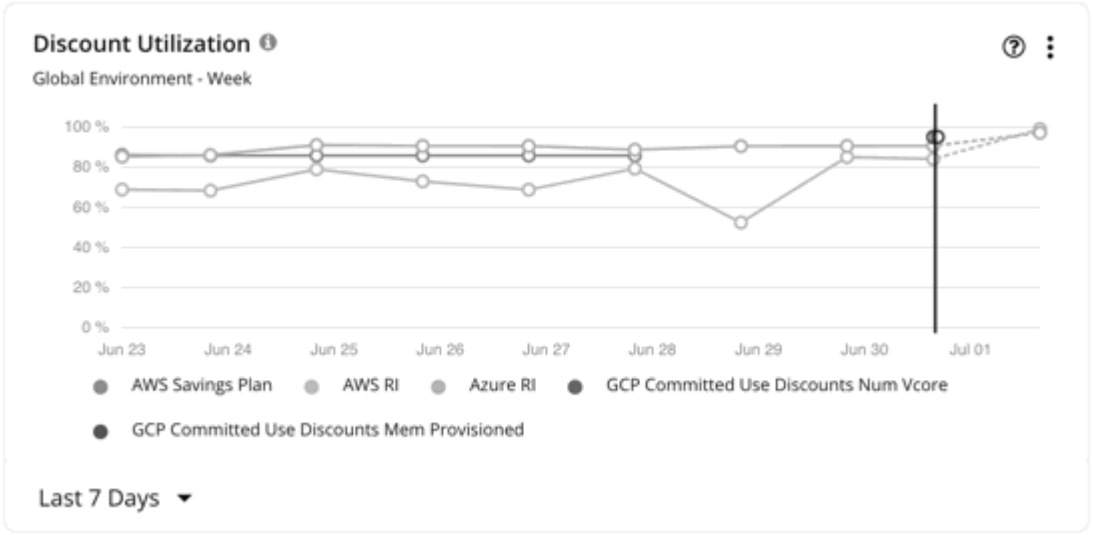
Workload Optimization Manager discovers committed use discounts for your workloads when you add the following as targets:

- A service account with the "Billing Account Viewer" role to the related billing accounts
- Billing accounts

Name	Account	Status	Region	Type	Payment	Instance Family	Cores	Memory	Term	Start Date	End Date
commitment-1		Active	us-west4	Family Scoped	All upfront	N2	1	4 GB	3 Years	Oct 15, 2021	Oct 15, 2024
commitment-1		Active	europe-north1	Family Scoped	All upfront	N2	1	2 GB	1 Year	Oct 21, 2021	Oct 21, 2022
commitment-1		Active	northamerica-northea	Family Scoped	All upfront	N2D	N/A	2 GB	1 Year	Jan 27, 2022	Jan 27, 2023
commitment-2		Active	us-central1	Family Scoped	All upfront	E2	N/A	4 GB	1 Year	Oct 21, 2021	Oct 21, 2022

Discount Utilization Chart

This chart shows how well you have utilized your current discount inventory *(on page 499)*. 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.



AWS RIs

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.

> **NOTE:**
> 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. Workload Optimization Manager does not include such RIs in its calculations of RI utilization or RI cost.

The data point on the solid vertical line shows the latest data that was polled from your environment. Data points to the left of the vertical line represent historical data, while data points to the right are projections into the future.

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 normalization factors

  Normalization factor is a measure of RI capacity that you can use to compare or combine the capacity for different instance families.

  Workload Optimization Manager 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 Workload Optimization Manager scope. Each workload is assigned normalized units depending on its instance type.

AWS Savings Plans

To view data for AWS Savings Plans, you must:

- Set the scope to your global environment.
- Choose a timeframe that shows daily or monthly data points (Last 7 Days, Last 30 Days, or Last Year)

  AWS measures your Savings Plans commitment in $/hour, but shows daily and monthly costs in Cost Explorer. Workload Optimization Manager polls Cost Explorer periodically to obtain the latest cost data, and then uses that data to calculate

Savings Plans utilization or coverage *per day*. For this reason, you will not see Savings Plans data if you choose a timeframe that shows hourly data points (Last 2 Hours or Last 24 Hours).

NOTE:
AWS timestamps data in UTC, but the chart presents data in your local time. This difference could result in the appearance of a missing day in the chart, but has no effect on data completeness (the chart always reflects the complete data set).

The chart shows historical data, represented by data points to the left of the solid vertical line. Data for the current day is not available since the latest data that AWS provides is always a few days old. In addition, Workload Optimization Manager does not project coverage into the future.

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 the total utilized and committed costs per day

Azure Reservations

You can set the scope to your global cloud environment or to individual subscriptions, billing families, or regions. Scoping to a subscription shows the reservations utilization for workloads for the entire billing family or for single and shared subscriptions.

The data point on the solid vertical line shows the latest data that was polled from your environment. Data points to the left of the vertical line represent historical data, while data points to the right are projections into the future.

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 Workload Optimization Manager scope. Each workload is assigned reservation units based on its instance type.

GCP Committed Use Discounts

The chart displays the latest data that was polled from your environment, but does *not* show historical data or project utilization into the future.

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

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, Workload Optimization Manager 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 388)*.

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.

Monthly Savings or Investments Totals Charts

Monthly Savings or Investments Totals charts help you examine the monthly savings or investments for executed cloud actions. For example, if an executed action causes an increase in the price, this is an investment. These charts also show the missed monthly savings or missed performance investments that you could have achieved for recommended cloud actions, if you executed them.

For this chart's scope, you can choose an account or subscription, a group of accounts or subscriptions, or use the default, Global Environment. If you use the default Global Environment, the chart will automatically use all cloud accounts for its scope. Other examples of scope settings are: An AWS Billing family, an Azure subscription, the All AWS Accounts predefined group, or the All Azure Accounts predefined group.

For all actions except Suspend, savings and investments are estimated based on the hourly cost of workload price differences and 730 hours per month of workload usage. Savings from Suspend actions are estimated based on the hourly cost of workload price differences and actual suspend times as defined in the suspension policy.

Missed savings and investments are estimated based on the hourly cost of workload price differences and the number of hours that recommended actions exist in the system.

Monthly Savings or Investments Totals charts calculate data on a monthly basis since your update of Workload Optimization Manager to version 2.3.0. Historical data stored in the database prior to version 2.3.0 is not included.

Chart Type

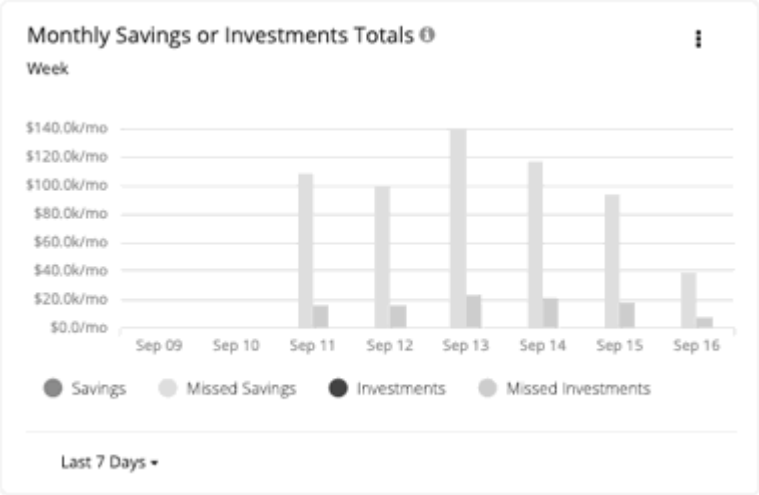
You can set the display to:

- Stacked Bar Chart
- Tabular

Examples:

- Stacked Bar

  This chart shows the monthly totals of savings or investments for each of the last seven days. It also shows the missed monthly savings or performance investments that you could achieve by executing recommended cloud actions.



  In the chart legend, you can also choose an item to change the display of the chart. Click the item again to reset the chart. For example, if you want to examine investment information, click **Investments** in the legend.

- Tabular

  This chart shows the monthly totals of savings or investments for each of the last seven days. It also shows the missed monthly savings or performance investments that you could achieve by executing recommended cloud actions.

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, Workload Optimization Manager 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

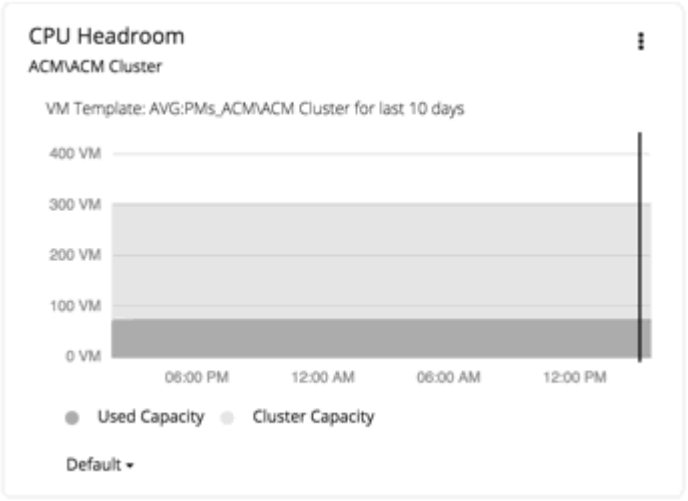
You can choose:

- 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

Groups assemble collections of resources for Workload Optimization Manager to monitor and manage. When setting scope for your Workload Optimization Manager 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.

Workload Optimization Manager discovers groups that exist in your environment. These groups include PM clusters, and entities grouped by different logical boundaries. For example, Workload Optimization Manager discovers Storage by Disk Array, Physical Machines by Datacenter, and VMs by Network. In addition, Workload Optimization Manager discovers pools such as virtual datacenters, or folders that implement specific HA policies.

You can also create custom groups. Workload Optimization Manager 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 physical machines with four CPUs), or other criteria such as time zone or number of CPUs.

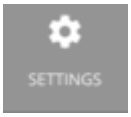
  These groups are dynamic because Workload Optimization Manager updates the group as conditions change.
- Static – You create these groups by selecting the specific group members.

NOTE:
You should never use the Workload Optimization Manager user interface to delete discovered groups. If you do, later analysis cycles will discover them again, and add them to your environment. But until it recreates those groups, any analysis that relies on those groups can give unexpected results.

You *can* delete any custom group you have created. Before you do, you should verify that you do not have any charts, plans, or policies that use the group you want to delete to set their scopes. After you delete the group, such charts, plans, or policies will loose their scope. For example, a policy with no scope has no effect.

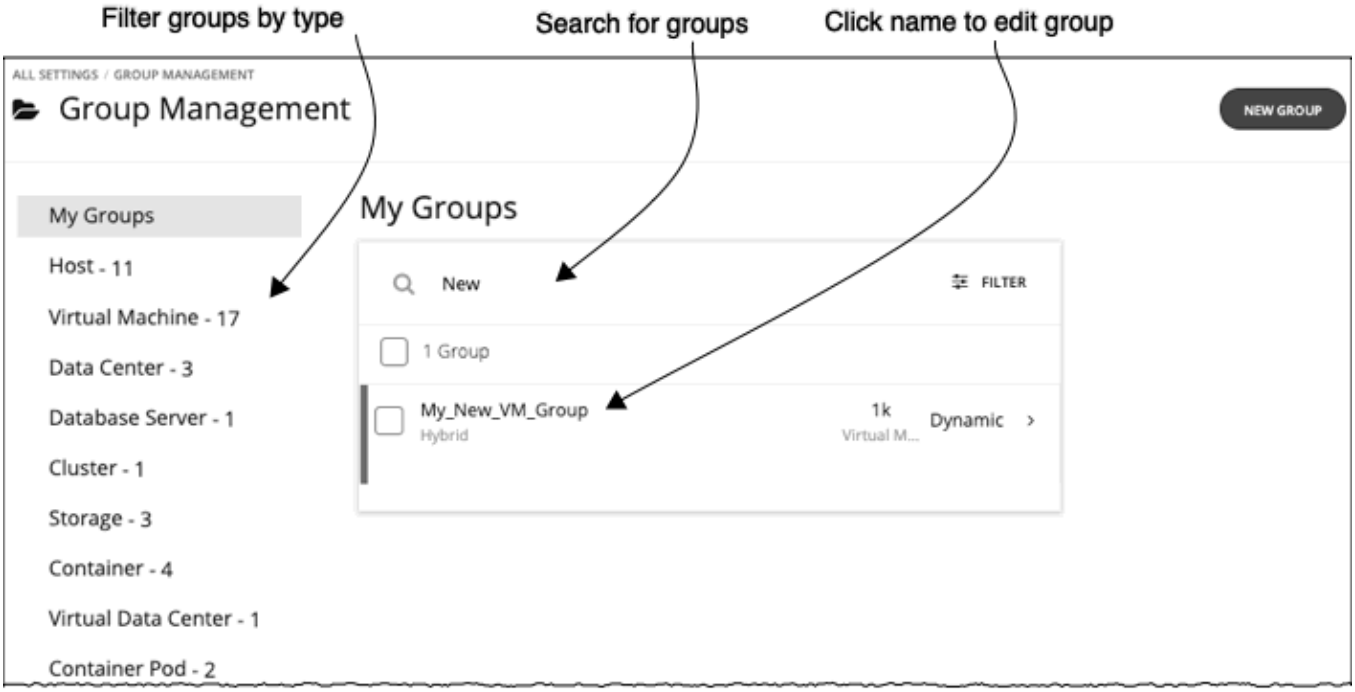
To create a group:

1. Navigate to the Settings Page.

   Click to navigate to the Settings Page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.
2. Choose **Groups**.

Click to navigate to the Group Management Page.



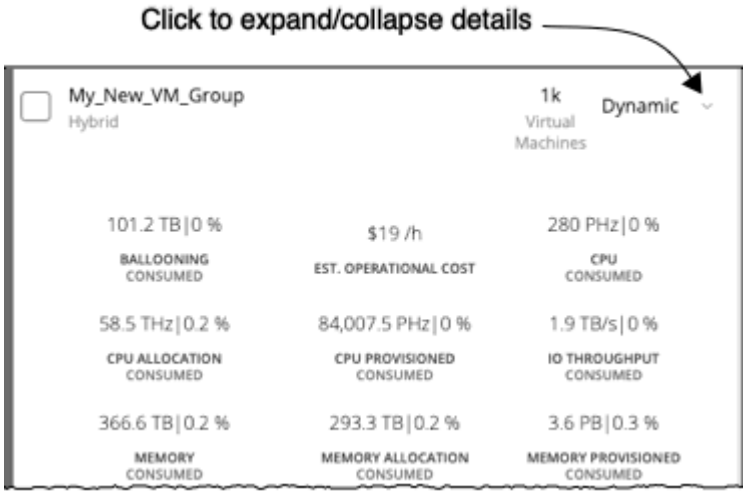
This page lists all the custom groups that you currently have configured for Workload Optimization Manager. You can:

- Expand an entry to see group details
- Select an entry to delete the group
- 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.
- 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.

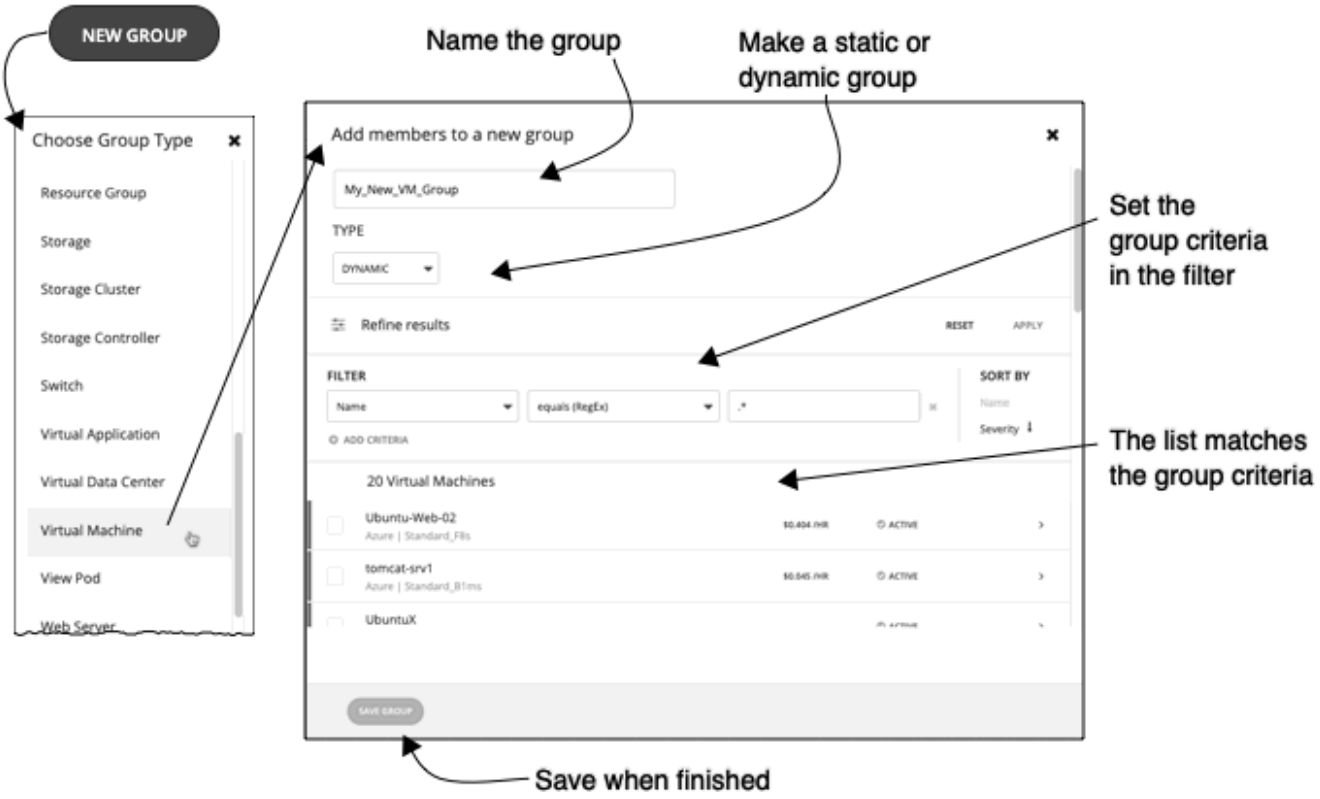
3. Expand an entry to see group details.



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.

4. Create a new group.

   Click **NEW GROUP**.



Next, choose a group type.

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 Schedules

Schedules specify a period of time during which certain events can occur. Workload Optimization Manager uses the following types of schedules:

- Calendar Schedule

  A calendar schedule is a setting that sets up a window of time when an automation policy *(on page 97)* takes effect. This policy can execute non-parking actions *(on page 59)* on public cloud or on-prem entities, or change settings that affect analysis and action generation.

- Parking Schedule

  A parking schedule is a setting that sets up a window of time when parking actions (stop or start) are enforced on virtual machines in your public cloud environment. Click **Park** on the main menu to see a list of parkable virtual machines.

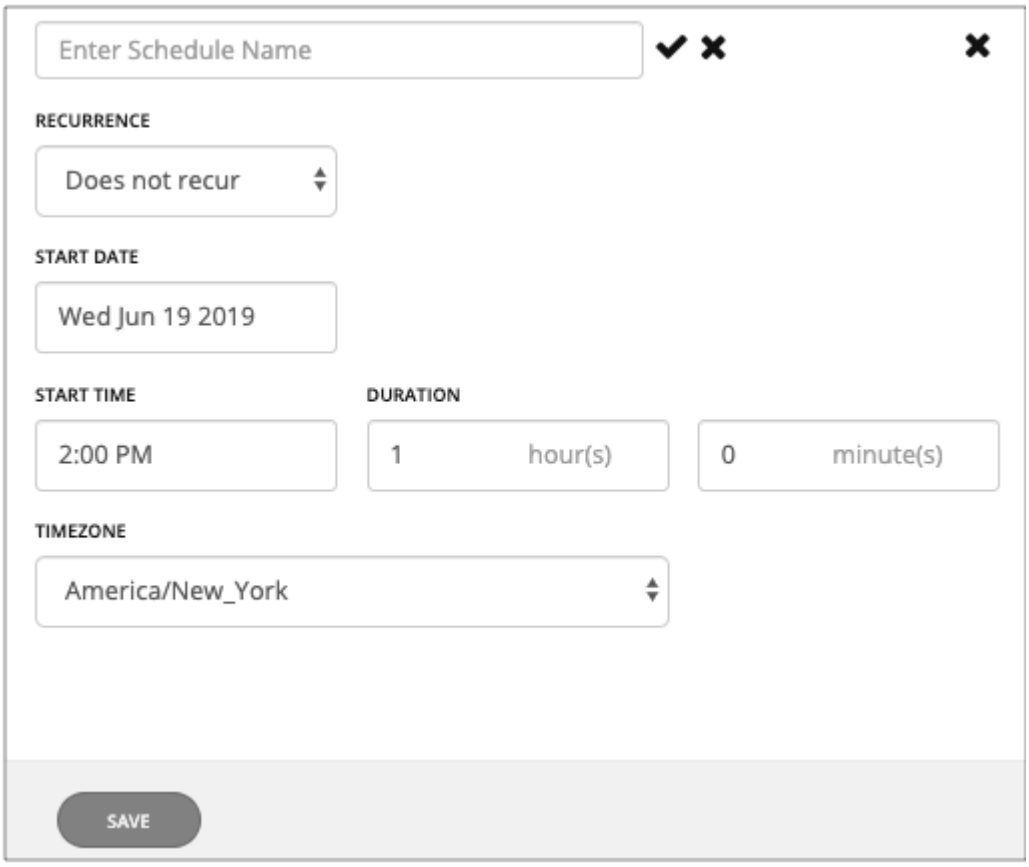
Managing Calendar Schedules

A calendar schedule is a setting that sets up a window of time when an automation policy *(on page 97)* takes effect. This policy can execute non-parking actions *(on page 59)* on public cloud or on-prem entities, or change settings that affect analysis and action generation.

Creating Calendar Schedules

1. Navigate to **Settings > Schedule**.
2. Click **New Schedule > Calendar Schedule**.

3. Type a name for the schedule.



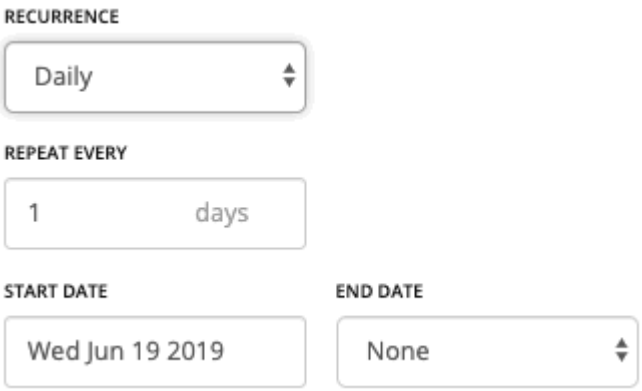
4. Set the recurrence for the schedule.

   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.

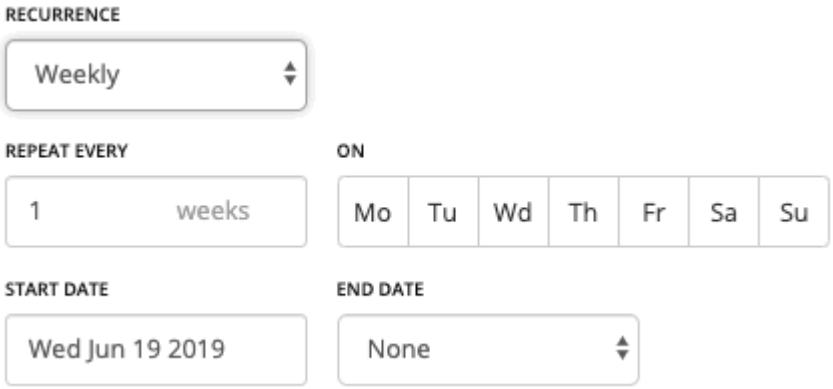
   - Daily

     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



  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

  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.

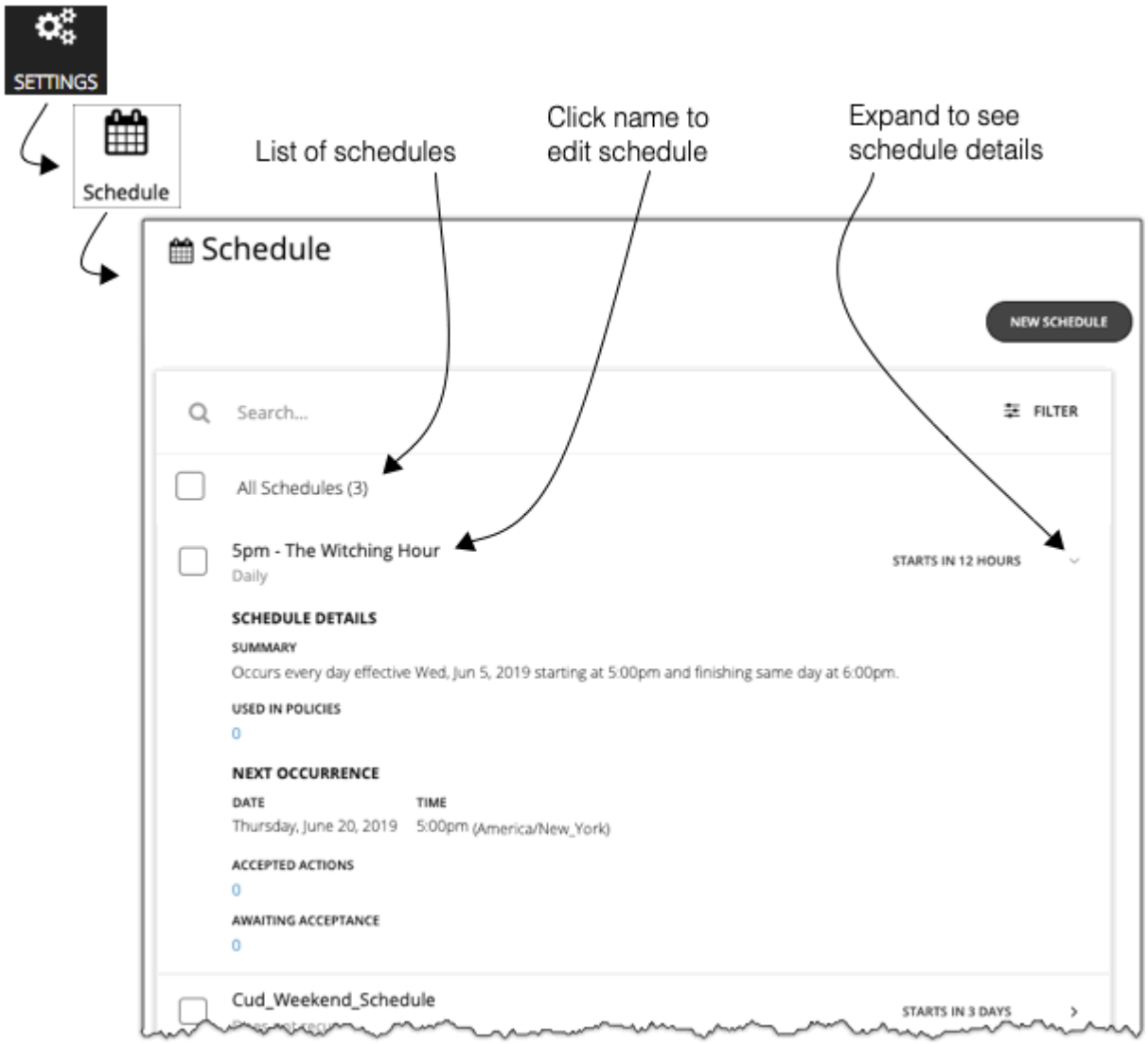
6. Set the time zone.

   This gives a reference for the schedule's start time. Workload Optimization Manager 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



The Schedules page lists all the currently defined schedules. From this page you can:

- Select an entry to delete the schedule.
- Select an entry to defer the next occurrence.

  Workload Optimization Manager 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.
- 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.

- Create new schedules.

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, Workload Optimization Manager disables the affected policies until you edit them to either:

- Apply a different schedule to the policy and save the change, or...
- 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 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.

Workload Optimization Manager posts a confirmation dialog before deleting a schedule that is currently in use.

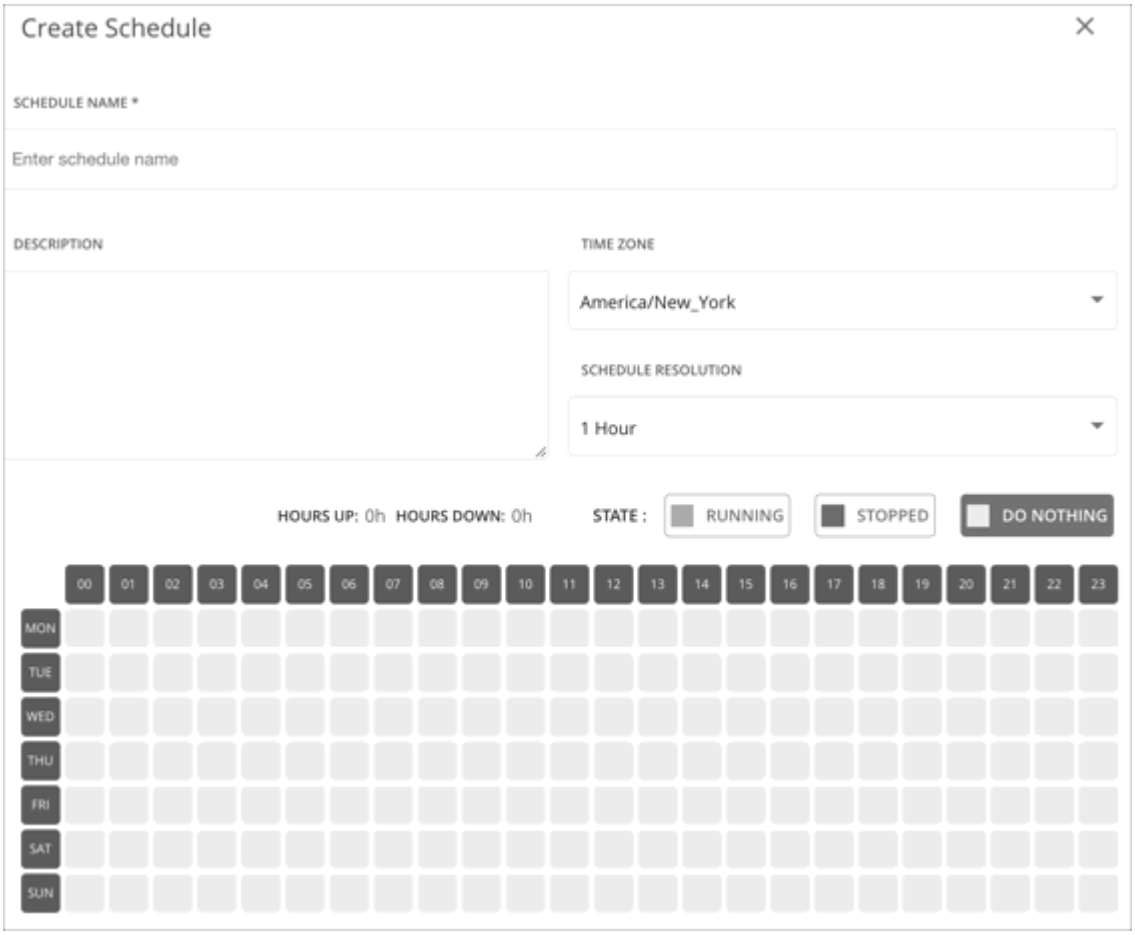
Managing Parking Schedules

A parking schedule is a setting that sets up a window of time when parking actions (stop or start) are enforced on virtual machines in your public cloud environment. Click **Park** on the main menu to see a list of parkable virtual machines.

Creating Parking Schedules

1. Navigate to **Settings > Schedule**.
2. Click **New Schedule > Parking Schedule**.

   The Create Schedule page displays.



3. Provide a name and description.

   When attaching a parking schedule to one or more cloud VMs, you will see the *name* of the schedule. It is therefore important to use a name that will help you recognize each schedule and choose the appropriate one.

4. Set the time zone.

   This gives a reference for the schedule's start time. Workload Optimization Manager 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.

5. Set the schedule resolution.

   By default, each block in the schedule configuration section corresponds to an hour of the day. You can change this to 15-minute increments to give you more flexibility with scheduling. For example, with 15-minute increments, you can enforce parking actions between 8:15 and 20:45. This is not possible with the default schedule resolution.

6. Set a parking schedule.

   Parking schedules recur every week, on the days and times that you specify. These schedules include the following states that define the parking policies that you can enforce.

State	Policy
Running (green)	A VM must be running on the selected days and times.

State	Policy
	Workload Optimization Manager will start the VM immediately if: ▪ The VM is not running at the beginning of the schedule. ▪ A user stops the VM via the cloud provider.
Stopped (red)	A VM must not be running on the selected days and times. Workload Optimization Manager will stop the VM immediately if: ▪ The VM is running at the beginning of the schedule. ▪ A user starts the VM via the cloud provider.
Do Nothing (gray)	Do not enforce parking actions on the selected days and times. This is the default state for all the blocks.

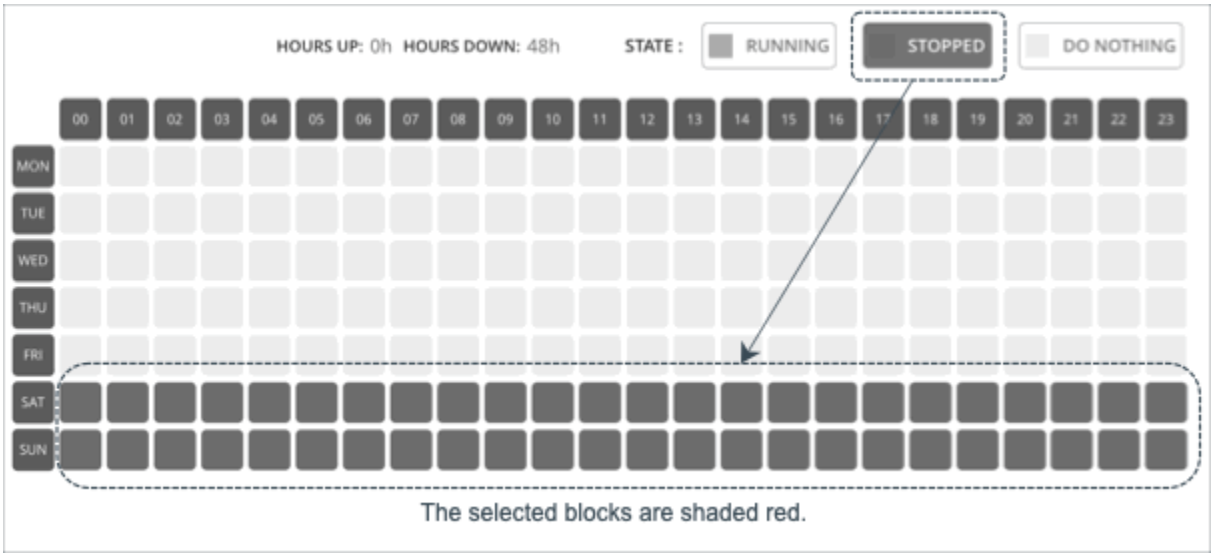
Basic steps:

a. Click either the **Running** (green) or **Stopped** (red) button.

b. Click the blocks for the days and times the action will be enforced. To select multiple contiguous blocks, click and hold, and then drag in any direction.

c. To undo your selections, click the **Do Nothing** (gray) button and then click the affected blocks.

In the following example, stop actions are enforced every Saturday and Sunday. No parking actions are enforced on all the other days of the week.



7. Click **Create**.

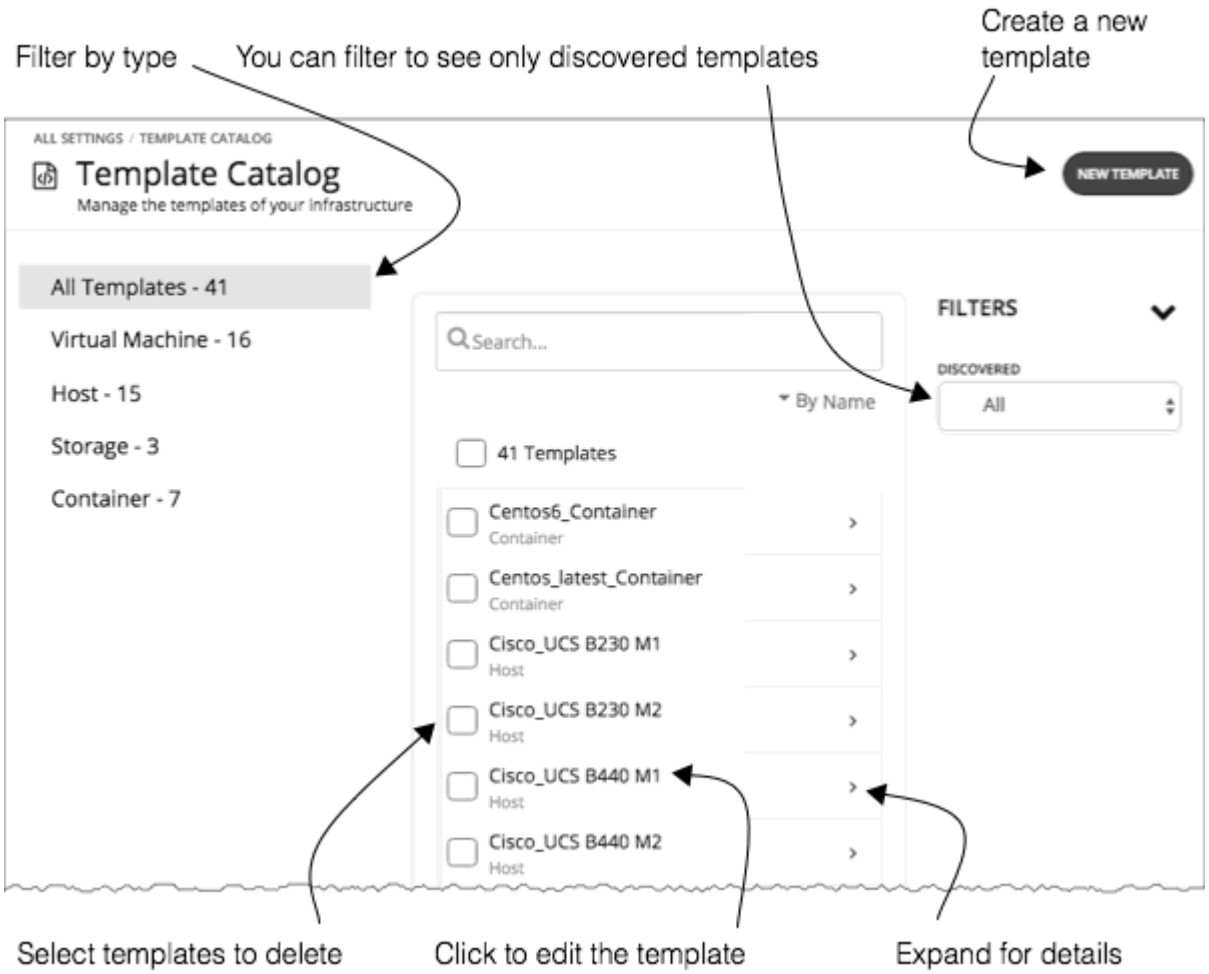
The schedule can now be attached to the VMs listed in the Parking page. For details about the Parking page, see Park: Stop or Start Cloud Resources *(on page 428)*.

Deleting Parking Schedules

Only unattached parking schedules can be deleted.

1. Select one or several unattached schedules to delete.
2. Click the delete button at the top-right section of the page.

Templates: Resource Allocations for New Entities



Workload Optimization Manager uses templates to describe new entities that it will deploy 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 given 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 Workload Optimization Manager 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, Workload Optimization Manager chooses the best image for that instance.

The Template Catalog shows all of the templates that have been specified or discovered for your installation of Workload Optimization Manager. From this page, you an also create new templates and edit existing ones.

Creating Templates

Templates specify the resources for entities that Workload Optimization Manager can deploy in your environment, or in plans.

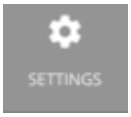
A VM template definition can include one or more images that Workload Optimization Manager uses to deploy the VM in your environment. The image identifies the actual deployment package, including a path to the physical files (for example an OVA).

The Template Catalog shows all of the templates that have been specified or discovered for your installation of Workload Optimization Manager. From this page, you an 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.

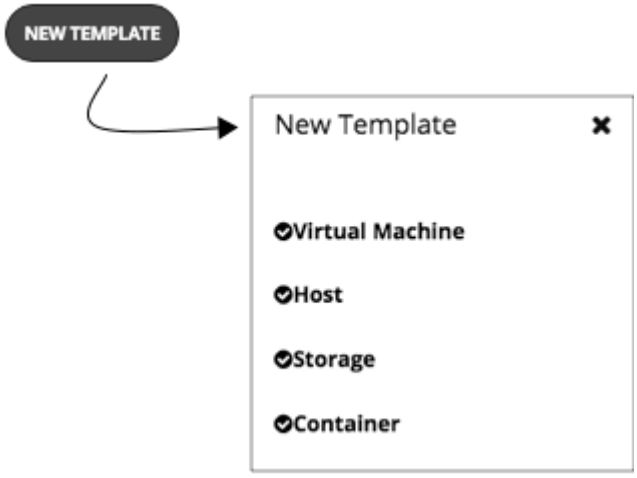
   

2. Choose Templates.

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 Workload Optimization Manager deploys the associated VM to your environment or in a plan, it uses these values to determine the size of the VM. Workload Optimization Manager 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 Workload Optimization Manager 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:
Workload Optimization Manager 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 Workload Optimization Manager 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 – Workload Optimization Manager 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 datacenter. 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 running hardware replacement plans.

The host template is a collection of these settings:

- CPU

  The processor for this host model. Note that CPU size and speed are not the only factors to determine processing power. To address this, you can 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 Workload Optimization Manager 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 – Workload Optimization Manager multiplies these values to calculate the host CPU resources.
- Memory

  The amount of memory to allocate for the VM, in MB.
- 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 running a plan, Workload Optimization Manager can use the price to calculate costs or savings when adding or removing host machines in an on-prem datacenter.

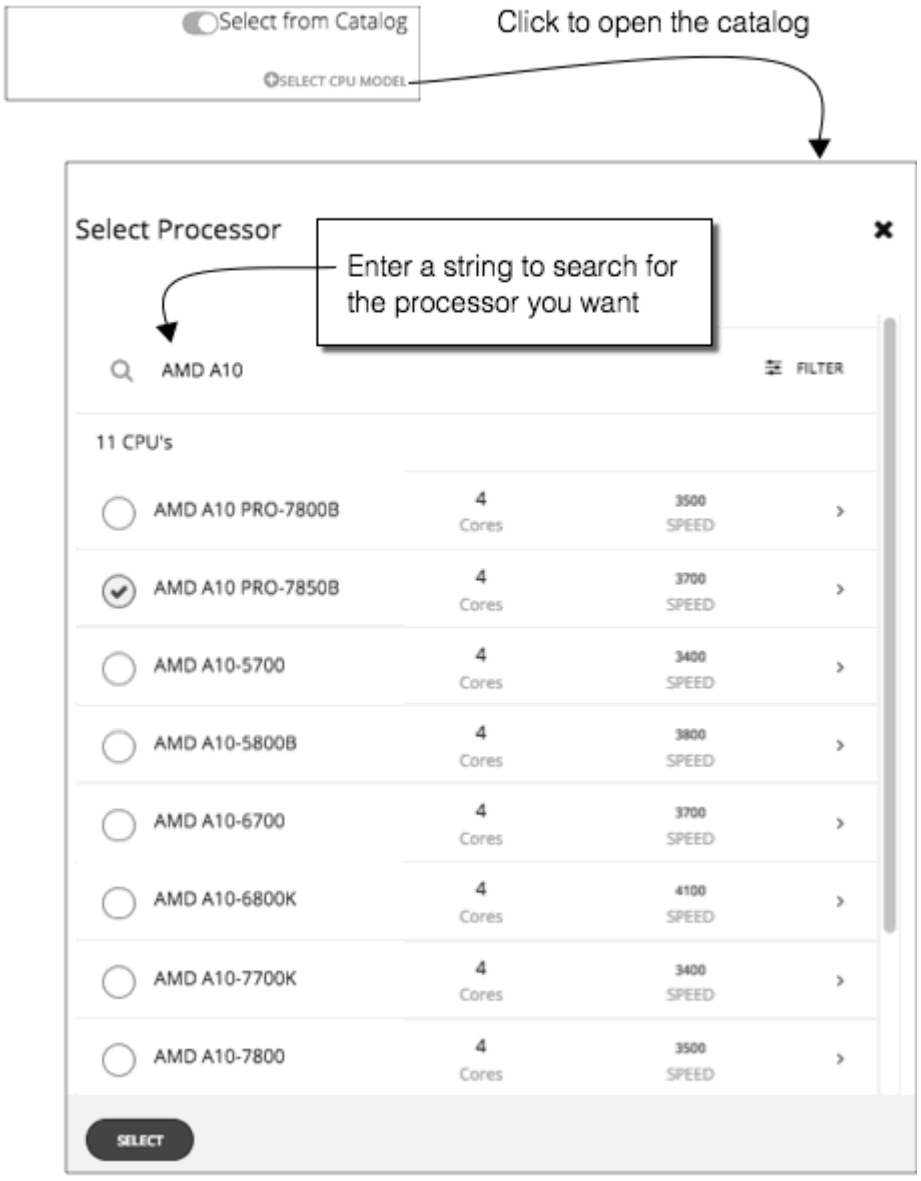
Selecting CPUs from the Catalog

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. This can affect planning in two ways:

- When planning hardware replacement, the plan knows the template's effective capacity. This means the plan knows how to best place workloads on the new hardware.
- For already deployed hosts, Workload Optimization Manager discovers the effective capacity and uses that information when calculating workload placement.

To build the catalog of CPU capacity, Workload Optimization Manager 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:
Workload Optimization Manager also uses the effective processor capacity when calculating workload placement in real-time. For more information, see Effective CPU Capacity *(on page 68)*.



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 Workload Optimization Manager detects a configuration issue with VMM or Hyper-V. As a result,Workload Optimization Manager treats the VMs as not controllable and will not attempt to place them.

The HCI Host template is a collection of these settings:

- CPU

  The processor for this host model. Note that CPU size and speed are not the only factors to determine processing power. To address this, you can 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 Workload Optimization Manager 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 – Workload Optimization Manager multiplies these values to calculate the host CPU resources.

- Memory

  The amount of memory to allocate for the VM, in MB.

- 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 will compute the effective storage capacity.

- Redundancy

  The redundancy method for this storage 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 running a plan, Workload Optimization Manager can use the price to calculate costs or savings when adding or removing host machines in an on-prem datacenter.

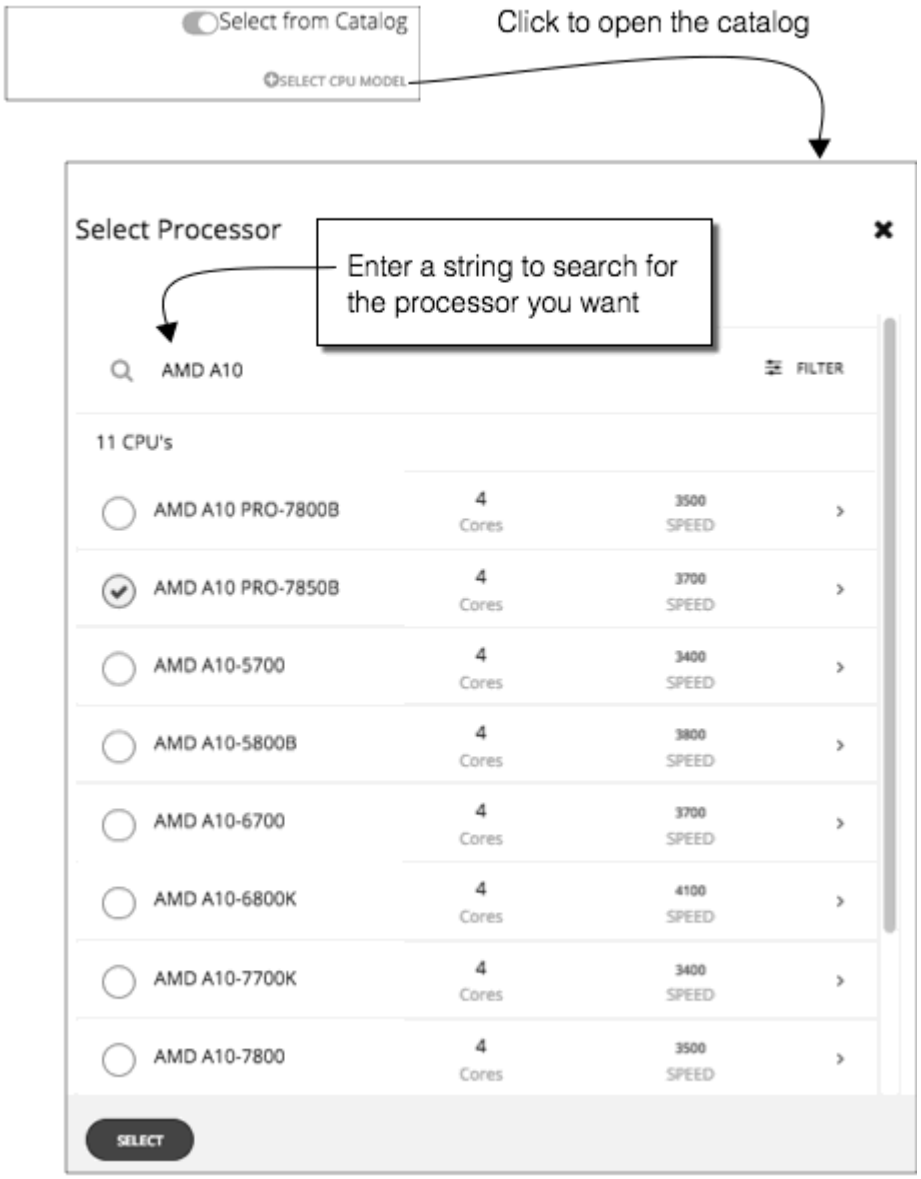
Selecting CPUs from the Catalog

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. This can affect planning in two ways:

- When planning hardware replacement, the plan knows the template's effective capacity. This means the plan knows how to best place workloads on the new hardware.
- For already deployed hosts, Workload Optimization Manager discovers the effective capacity and uses that information when calculating workload placement.

To build the catalog of CPU capacity, Workload Optimization Manager 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:
Workload Optimization Manager also uses the effective processor capacity when calculating workload placement in real-time. For more information, see Effective CPU Capacity *(on page 68)*.



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, Workload Optimization Manager 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 Workload Optimization Manager, you can set up costs that Workload Optimization Manager uses in its calculations. This setup includes:

- Reserved Instance Settings

  To recommend placing workloads on instance types that take advantage of discounted pricing, Workload Optimization Manager 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 Workload Optimization Manager uses in its calculations.

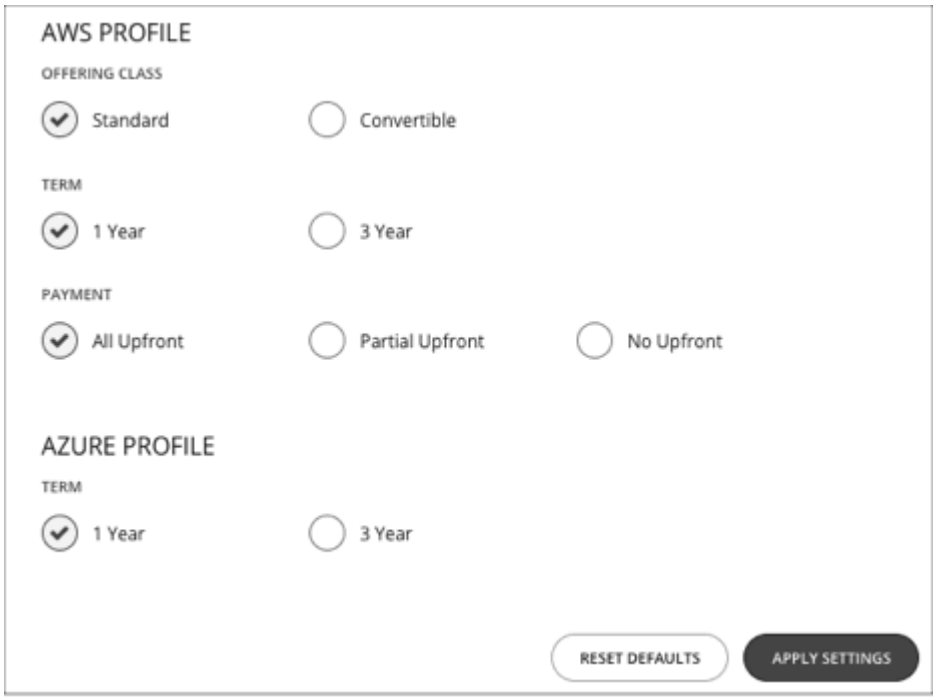
- Price Adjustment

  Cloud service providers can offer their own price lists, including special costs for services or discounts for workloads. However, Workload Optimization Manager does not discover these adjustments. For example, to reflect any discounted prices in the Workload Optimization Manager display and in Workload Optimization Manager analysis, you must manually configure those discounts. In Workload Optimization Manager, you configure such discounts via **Price Adjustments** for specific billing groups in your cloud environment.

- Currency

  By default, Workload Optimization Manager 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, Workload Optimization Manager 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 Workload Optimization Manager uses in its calculations.

A purchase profile determines the costs that Workload Optimization Manager will use for all discount purchase decisions in your environment. As it sees opportunities to move workloads to another term, Workload Optimization Manager determines the costs based on the profile, and includes cost information in action descriptions. Workload Optimization Manager 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 service providers can offer their own price lists, including special costs for services or discounts for workloads. However, Workload Optimization Manager does not discover these adjustments. For example, to reflect any discounted prices in the Workload Optimization Manager display and in Workload Optimization Manager analysis, you must manually configure those discounts. In Workload Optimization Manager, you configure such discounts via **Price Adjustments** for specific billing groups in your cloud environment.

Workload Optimization Manager 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, Workload Optimization Manager 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:
Workload Optimization Manager 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

A price adjustment configures adjusted workload pricing that you have negotiated with your Cloud Provider. After you configure an adjustment, Workload Optimization Manager applies it to pricing in the affected cloud scope.

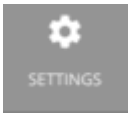
To create a price adjustment in Workload Optimization Manager, you identify the adjustment's scope – the subscriptions or billing families the adjustment applies to – and then set the type and percentage for the price adjustment. This specifies an overall adjustment for the workloads that fall within the billing group. For AWS, you can later drill into the adjustment to specify overrides for specific template families or services.

Notes:

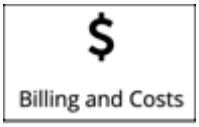
- To use a price adjustment with a given billing group, you must increase the memory allocated to the VM that hosts your Workload Optimization Manager instance. Workload Optimization Manager requires that you provide a minimum amount of memory when you install the product. To use price adjustments, Cisco 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 Workload Optimization Manager 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 to navigate to 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 cloud service provider accounts that are consolidated into a single billing schedule. Billing group details depend on your service provider:
     - AWS: To consolidate billing, AWS supports billing families of AWS accounts, where there is a *master* account and other *member* accounts. Workload Optimization Manager lists each billing family as a billing group. You can choose a billing family to set the scope of this adjustment.

     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, Workload Optimization Manager will apply this percentage as it calculates the optimal workload capacity and placement.

  NOTE:
  If you set an overall adjustment of 0%, then Workload Optimization Manager 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.

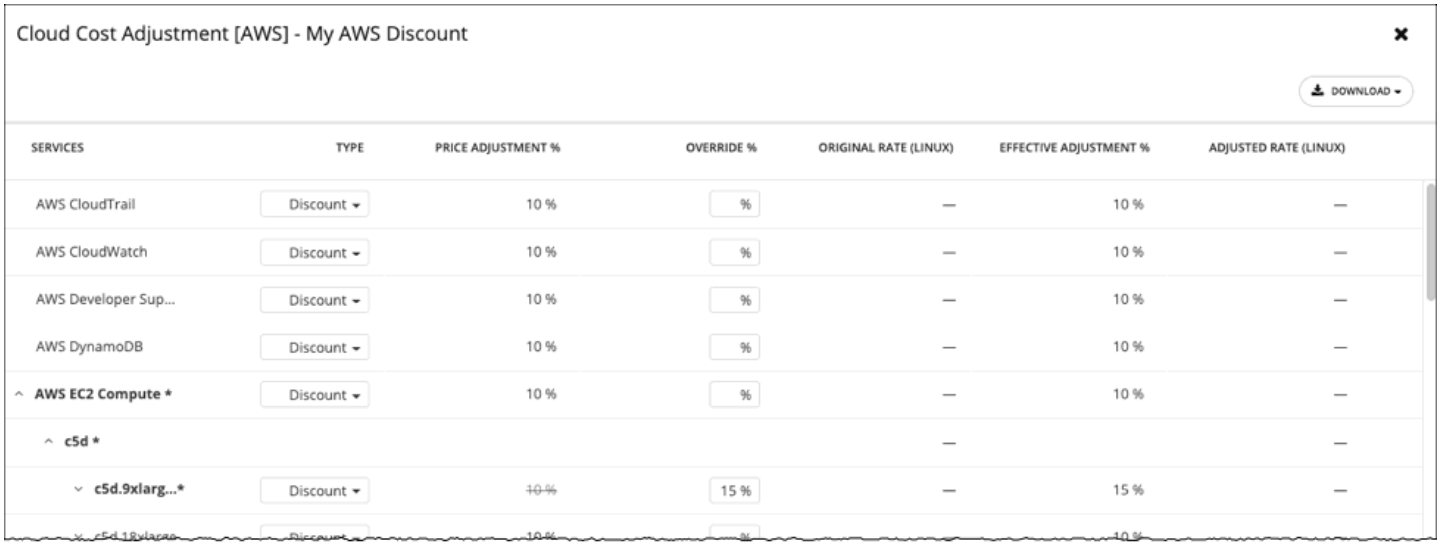
   The PRICE ADJUSTMENT percentage you just specified applies as a default in the adjustment scope. However, you might have negotiated different prices for specific services or template families in your cloud environment. To configure these special prices, click **PRICE OVERRIDES** to open the Cloud Cost Adjustment fly-out.

   For details, see AWS Price Override *(on page 533)*.

6. Save your work.

   After you have configured the price adjustment, click **SAVE**.

AWS Price Override



To override the PRICE ADJUSTMENT setting for AWS Billing groups, Workload Optimization Manager 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. Workload Optimization Manager treats the AWS Billing family as a Billing Group. For more information about billing families and accounts, see AWS Billing Families *(on page 535)*.

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.

Workload Optimization Manager 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 Workload Optimization Manager 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, Workload Optimization Manager 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 Cost and Usage report 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 Workload Optimization Manager 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 Workload Optimization Manager applies by default to the given service.
- OVERRIDE %

  If you have entered a value, this is the price adjustment Workload Optimization Manager 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.

AWS Billing Families

As you configure AWS targets, Workload Optimization Manager discovers AWS accounts that are consolidated into *billing families*. A billing family has one *master* account, and zero or more *member* accounts. By recognizing billing families, Workload Optimization Manager more accurately calculates cloud investments and savings, and makes more accurate recommendations for RI coverage.

In the Targets user interface, master accounts appear in bold, with a star next to them. You can expand the account entry to see the related member accounts. If you expand the entry for a member account, then the related accounts includes the family master, indicated by a star.

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 Workload Optimization Manager, if you enable Billing Family Recognition, then you can see the billing family master and member accounts in the Targets user interface, and Workload Optimization Manager 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 Workload Optimization Manager credentials for an AWS role that includes the permission, `organizations:DescribeOrganization`. With that permission, Workload Optimization Manager can:

  - Discover master 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 Workload Optimization Manager 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 Workload Optimization Manager cannot discover complete billing information for that family, and its analysis will be based on incomplete information.

  Workload Optimization Manager displays member accounts that have been configured as targets in regular text. For members that Workload Optimization Manager discovers but have not been configured as targets, Workload Optimization Manager 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

  Workload Optimization Manager 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 Workload Optimization Manager includes the new member in its analysis of billing families. If the new member is not already a target, then Workload Optimization Manager lists the new member in grayed text.

- You can configure discounts per billing family

  Workload Optimization Manager 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 531)* and Discount Override: AWS *(on page 533)*.

- You might see master 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 Workload Optimization Manager user interface as a master account, but it has no member accounts.

Azure Enterprise Agreements

You can configure Workload Optimization Manager to manage Azure subscriptions within the context of an Enterprise Agreement (EA). An EA defines specific pricing, including the pricing for reservations. When you configure an EA target, and set the EA key to your Azure targets, Workload Optimization Manager uses that richer pricing information to calculate workload placement and reservations coverage for your Azure environment.

To enable Workload Optimization Manager management of Azure EA environments, you must configure:

- One Microsoft Enterprise Agreement target
- At least one Service Principal target that can discover the underlying Azure subscriptions

For information about Azure targets, see "Microsoft Azure" in the *Target Configuraton Guide*.

In the Targets View, you can identify the targets related to Azure EA as follows:

- EA Targets

  The target that discovers the EA to track pricing and reservations. You can have one EA target per Workload Optimization Manager deployment.

- Azure Subscription Targets

The targets that manage the workloads in your Azure environment. These are discovered by Service Principal targets. Note that not all subscription targets *necessarily* participate in the EA. Expand these entries to see the related Service Principal target. For members of the EA, you can see the related EA target as well.

Subscriptions that do not participate in the EA appear as Standalone targets.

NOTE:
In rare circumstances, you can have a subscription that is not in use – The subscription has no workloads associated with it. In this case, Workload Optimization Manager identifies the subscription as Standalone. This is because the target cannot discover any cost or usage information that would relate the subscription to its EA.

Empty Azure EA subscriptions that are not incurring any charges will not stitch with the Azure Billing target or the Azure EA target, and a discrepancy will occur in the offer ID of the subscription. Once the subscription incurs a charge, the stitching occurs and the subscription should correctly associate with the Azure Billing target with the correct offer ID.

- Service Principal Targets

  The Azure target that you configure to discover Azure subscription targets. Expand the entry to see the discovered targets. If you have configured an EA target, the entry lists that as well, along with the EA enrollment number.

Reservations and Azure EA

For Azure environments, Workload Optimization Manager can only discover and use reservations if you have configured a Microsoft Enterprise Account target, and if one or more subscriptions participate in that EA.

To discover and manage reservations in Azure environments, Workload Optimization Manager uses both the EA target and the associated subscription targets. On its own, a subscription target exposes costs for pay-as-you-go pricing. The EA target discovers pricing for the available reservations. Workload Optimization Manager combines this information to track:

- Utilization of reservations
- VMs covered by reservations
- VM costs (accounting for reservations)
- Purchase recommendations

NOTE:
This release of Workload Optimization Manager does not support discovery and management of reservations for Classic VMs, Classic Cloud Services, and Suppressed Core VMs.

Cost Calculations for Azure Environments

To understand the reported costs in your Azure environment, consider these points:

- For targets that participate in the EA, Workload Optimization Manager uses the terms of the given EA, and bases costs on the Offer ID that is effective for the given subscription.
- For VMs in Azure, reservations pricing does not include the cost of the OS license. However pricing for on-demand VMs does include the license cost.

NOTE:
For Microsoft Azure EA environments, the projected cost for actions to purchase reservations might not match associated costs you find in the Microsoft Pricing Calculator.

Workload Optimization Manager actions can recommend purchases. For these recommendations, the action assumes a free Linux OS, so the cost estimate does not include the OS cost. However, The Microsoft Pricing Calculator does include costs for OS licenses. As a result, when you compare the Workload Optimization Manager cost estimates to the values in the Pricing Calculator, it's likely that the two estimates will not match. This difference also affects the Break Even Point that appears in the Recommended RI Purchases chart. Because the recommended purchases do not include Azure costs for OS licenses, the listed Break Even Point can be optimistic.

- For on-prem workloads you migrated to Azure, Workload Optimization Manager recognizes Azure Hybrid Benefit (AHUB) savings for reservations and on-demand workloads. The costs you see in Workload Optimization Manager charts include this benefit. However, remember that recommended actions do not include any license cost, so the actions will not reflect any proposed AHUB savings (see above).

Currency Settings

By default, Workload Optimization Manager 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.

Workload Optimization Manager 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. Workload Optimization Manager does not convert monetary amounts when you switch symbols.

Administrative Tasks

To perform Workload Optimization Manager administrative tasks, you will navigate to different pages from **Settings**.The different tasks you can perform for Workload Optimization Manager include:

- Managing User Accounts *(on page 540)*

  Create and manage user accounts for Workload Optimization Manager.
- Maintenance: Logging, and Troubleshooting *(on page 551)*

  Perform general tasks such as setting log levels or sending troubleshooting data to Technical Support.
- License Configuration *(on page 554)*

  Review the status of your current license, and apply any license upgrades.

Managing User Accounts

As an administrator, you specify accounts that grant users specific access to Workload Optimization Manager. User accounts determine the following for a given user login:

- User Authentication

  To configure an account, you set the type of authentication the account will use:
  - Local User – Configure the username and password and save those credentials on the Workload Optimization Manager server.
  - External User – Single user accounts that authenticate through Single Sign-on (SSO) or through Microsoft Active Directory (AD).
  - External Group – A group of user accounts that authenticate through SSO or AD.
- User Authorization

  Properties that determine the range of access and features for a given user:
  - Role – Access to specific Workload Optimization Manager features
  - Scope – How much of the environment this user can manage

As you configure user accounts, you can set up access to specific clusters in your environment. You can even set up accounts for tenant customers, and only show them the virtual workloads they own in their specific virtual datacenters.

IMPORTANT:
For self-hosted Workload Optimization Manager instances, you can configure Workload Optimization Manager to use SSO authentication. When SSO is enabled, Workload Optimization Manager only permits logins via the SSO IdP. Whenever you navigate to your Workload Optimization Manager installation, it redirects you to the SSO Identity Provider (IdP) for authentication before displaying the Workload Optimization Manager user interface.

Before you enable SSO for your Workload Optimization Manager installation, *you must configure at least one SSO user with Workload Optimization Manager administrator privileges*. If you do not, then once you enable SSO you will not be able to configure any SSO users in Workload Optimization Manager. To authorize an SSO user as an administrator, use **EXTERNAL AUTHENTICATION** to do one of the following:

- Configure a single SSO user with administrator authorization.

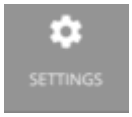
  Add an external user. The username must match an account that is managed by the IdP.
- Configure an SSO user group with administrator authorization.

  Add an external group. The group name must match a user group on the IdP, and that group must have at least one member.

For information about configuring SSO user groups in SAML, see Configuring a Group for SSO Authentication *(on page 550)*. For information about configuring SSO authentication for Workload Optimization Manager, see "Single Sign-On Authentication" in the *Installation Guide*.

To work with Workload Optimization Manager accounts:

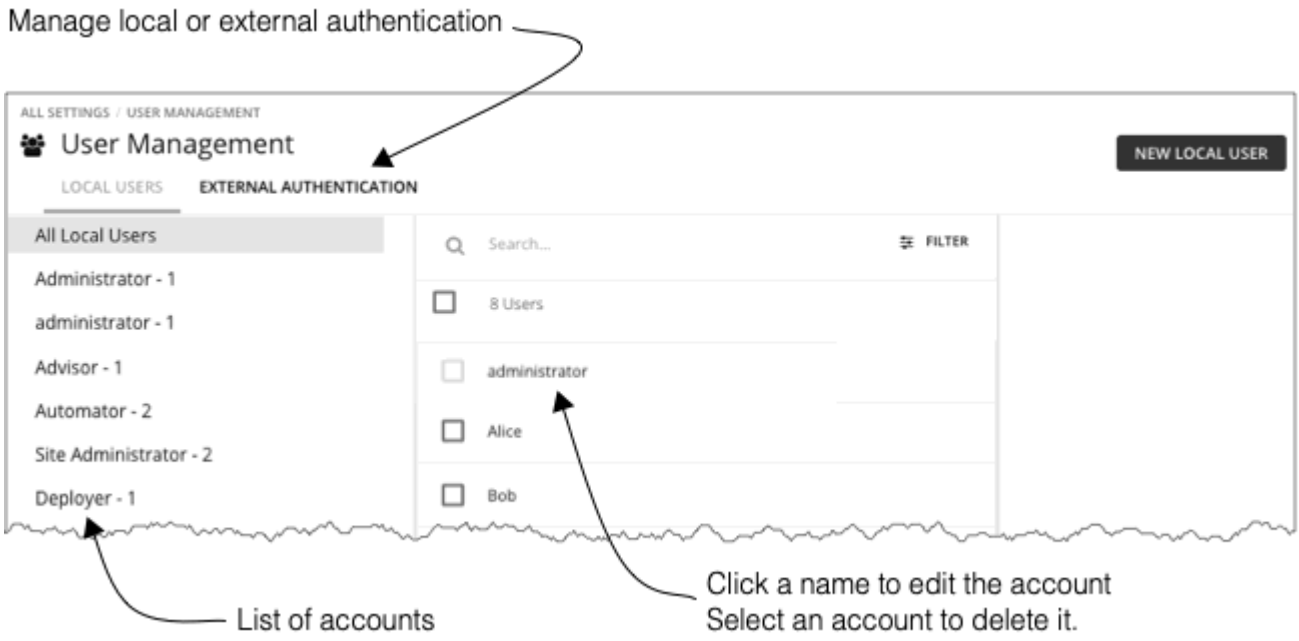
1. Navigate to the Settings Page.

   Click to navigate to the Settings Page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.
2. Choose User Management.

   User Management

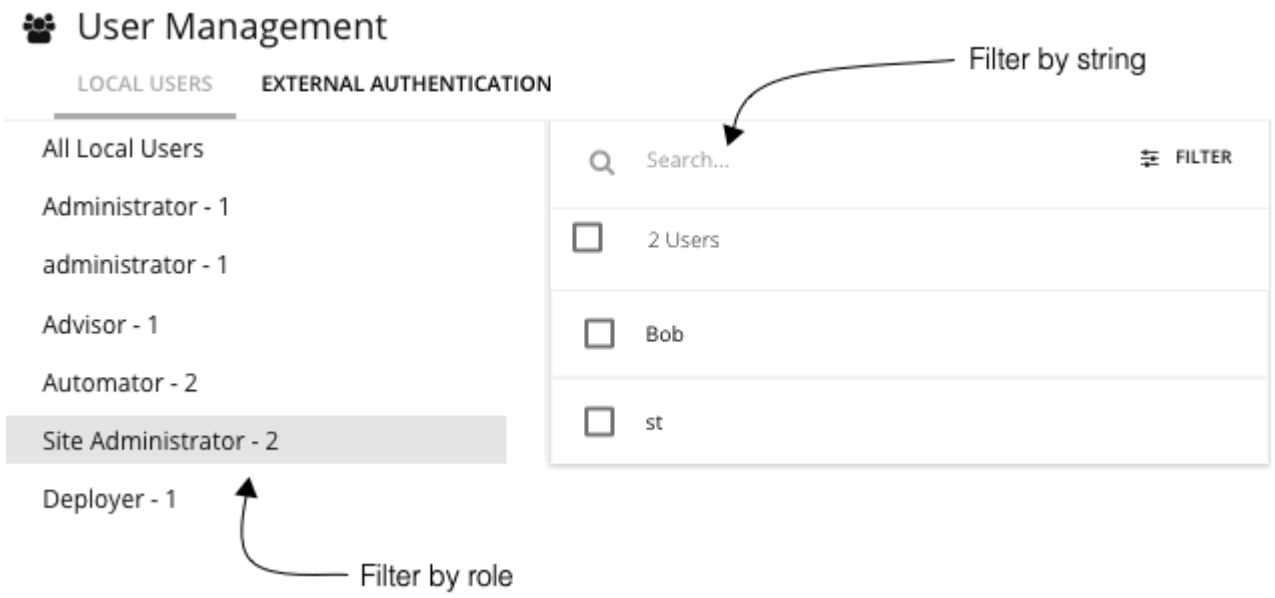
   Click to navigate to the User Management Page.



This page lists all the user accounts that you currently have configured for Workload Optimization Manager. You can:

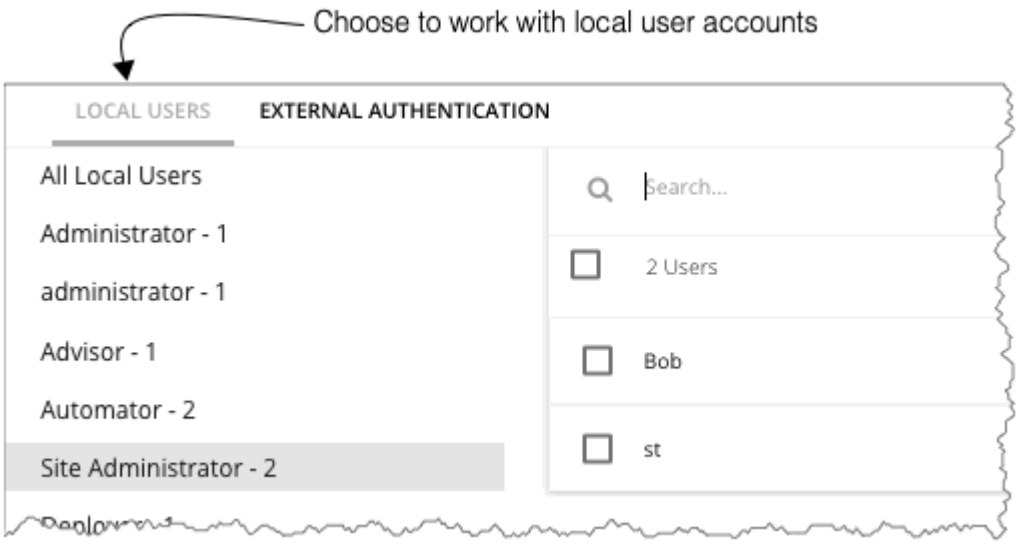
- Click to manage LOCAL USERS or EXTERNAL AUTHENTICATION
- Select an entry to delete the account
- Click a name to edit the account
- Create new user or group account
- Configure Active Directory settings

3. Filter the list of users.



To work with a long list of users, you can filter by role (for example, only show administrator or only show observer users). You can also type a string in the **Search** field to filter the list, and you can sort the list by name.

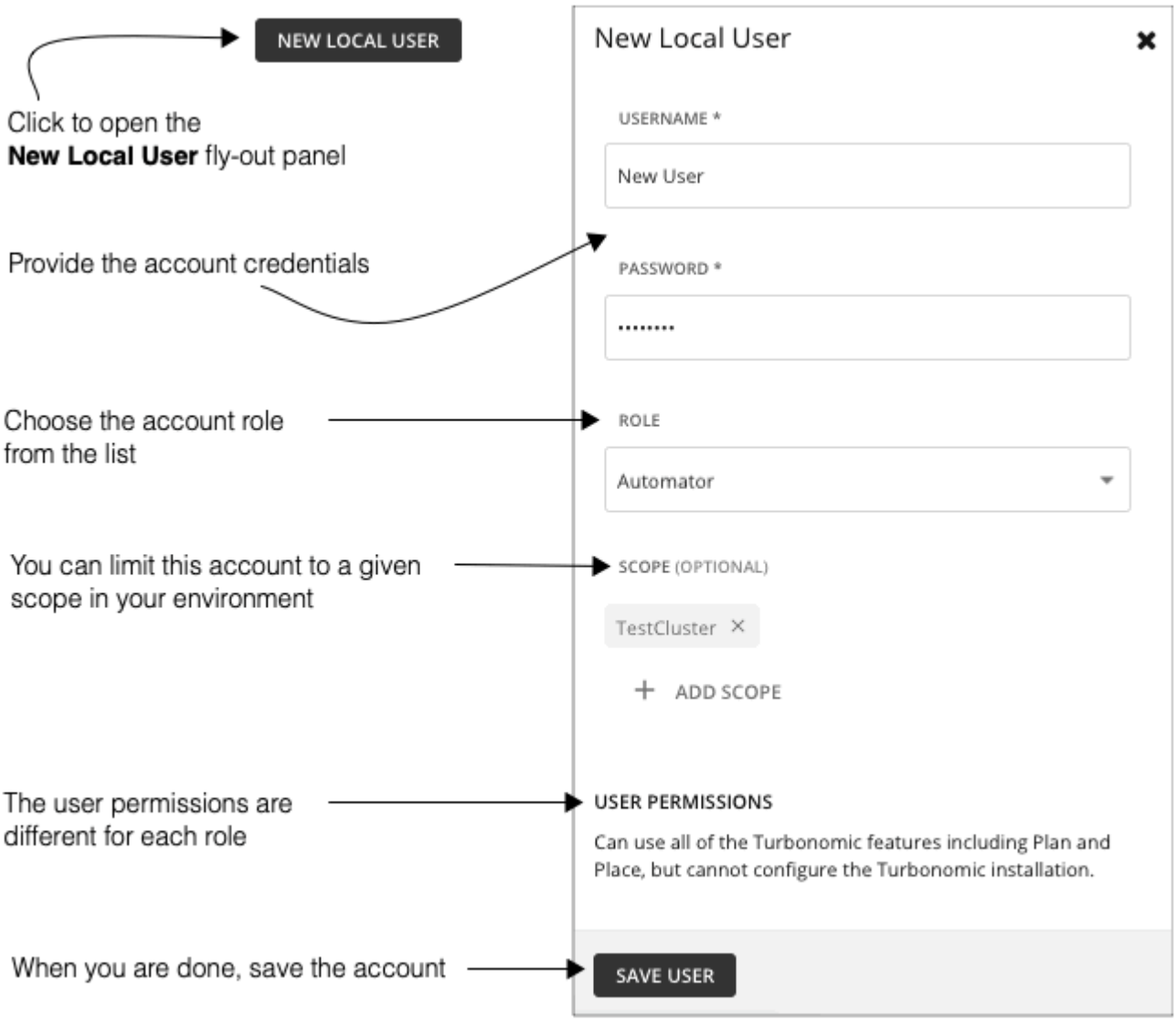
4. Work with Local user accounts.



Workload Optimization Manager stores local accounts and their credentials on the Workload Optimization Manager platform. Local authentication is for individual users, only.

When you choose **LOCAL USERS**, Workload Optimization Manager displays a list of all the local user accounts you have configured for this installation.

5. Create or edit a local user account.



To add a new local user, click **NEW LOCAL USER**. To edit an existing account, click the account name in the list. To configure a local account, specify:

- Authentication:

  Provide the username and password. Workload Optimization Manager stores these credentials on the local server.

- Authorization – User Role:
  - Administrator

    Users with this role can use all Workload Optimization Manager features and modify settings to configure the Workload Optimization Manager installation. For Workload Optimization Manager instances hosted in the public cloud, this role is limited to the Workload Optimization Manager representative that manages the instances.
  - Site Administrator

    Users with this role can use all Workload Optimization Manager features and modify site-specific settings to configure the Workload Optimization Manager installation. Users can also administer Groups, Policies, Templates, Billing/Costs, and Target Configuration, but not Email, Licenses, Updates, and Maintenance. Users can create other user accounts, except accounts with the Administrator role.

- Automator

  Users with this role can use all Workload Optimization Manager features (including Plan, Park, and Place), but cannot configure the Workload Optimization Manager installation or create policies.

- Deployer

  Users with this role can view all Workload Optimization Manager charts and data, use Place to reserve workloads, and create placement policies and templates. However, users cannot run plans or execute any recommended actions.

- Advisor

  Users with this role can view all Workload Optimization Manager charts and data, and run plans. However, users cannot use Place to reserve workloads, create policies, or execute any recommended actions.

- Observer

  Users with this role can view the environment, including the Home Page and Dashboards. Users can also use Search to set a scope to the session. For scope, only VM groups and Resource Groups are supported.

- Operational Observer

  Users with this role can view the environment, including the Home Page, Dashboards, Groups, and Policies. Users can also use Search to set a scope to the session.

- Shared Advisor

  Users with this role are scoped users. They can view the Home Page and Dashboards, but only see VMs and Applications. Users cannot execute Workload Optimization Manager actions.

- Shared Observer

  Users with this role are scoped users. They can view the Home Page and custom Dashboards, but only see VMs and Applications. Users cannot see Executive Dashboards or execute Workload Optimization Manager actions. This is the most restricted user.

- Report Editor

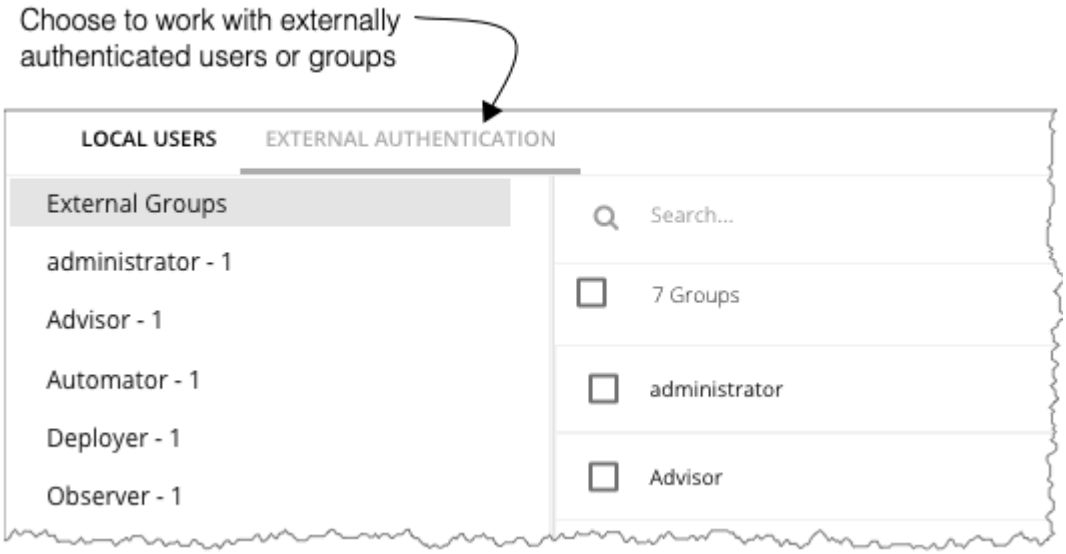
  Users with this role can create, edit, and delete reports. Due to limits to the reporting license, only one user per instance is allowed to have this role (by default, the local **administrator** user). To assign this role to another user, you must first remove it from the current user. Be sure that the new user is *not* a scoped user.

- Authorization – Scope (optional)

  The scope limits what the user can monitor. For example, you can scope to a group that contains only the physical machines that support this user's VMs or applications. Click **ADD SCOPE** and choose which groups or clusters this user can see.

  NOTE:
  Under most circumstances, a scoped user cannot see actions for entities that are outside of the configured scope. However, when zooming in to Host entities, the user can see actions for storage that is outside of the user's scope if the hosts use that storage.

6. Work with EXTERNAL AUTHENTICATION to set up SSO or AD accounts.



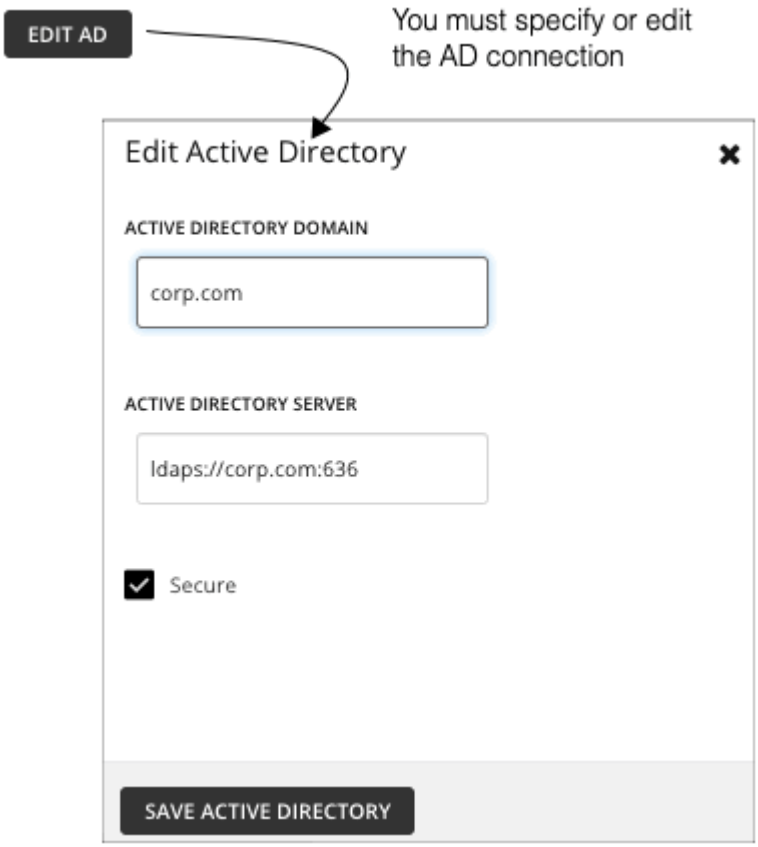
For External Authentication, you configure Workload Optimization Manager to use SSO or AD services to manage the credentials and authentication of users. You can create external accounts to authorize user groups or individual users.

> **NOTE:**
> If a user is a member of multiple groups, then Workload Optimization Manager logs the user on via the first SSO or AD that successfully authenticates the user. Also note that Workload Optimization Manager does not support nested AD groups – AD logins must be for users in a top-level group.

To enable SSO, you must configure access to the given IdP. For information about configuring SSO, see "Single Sign-On Authentication" in the *Installation Guide*.

To enable AD you must specify either an AD domain, an AD server, or both. Workload Optimization Manager uses this connection for all AD users.

7. Enable AD authentication.



To enable AD, click **CONNECT TO AD** and configure:

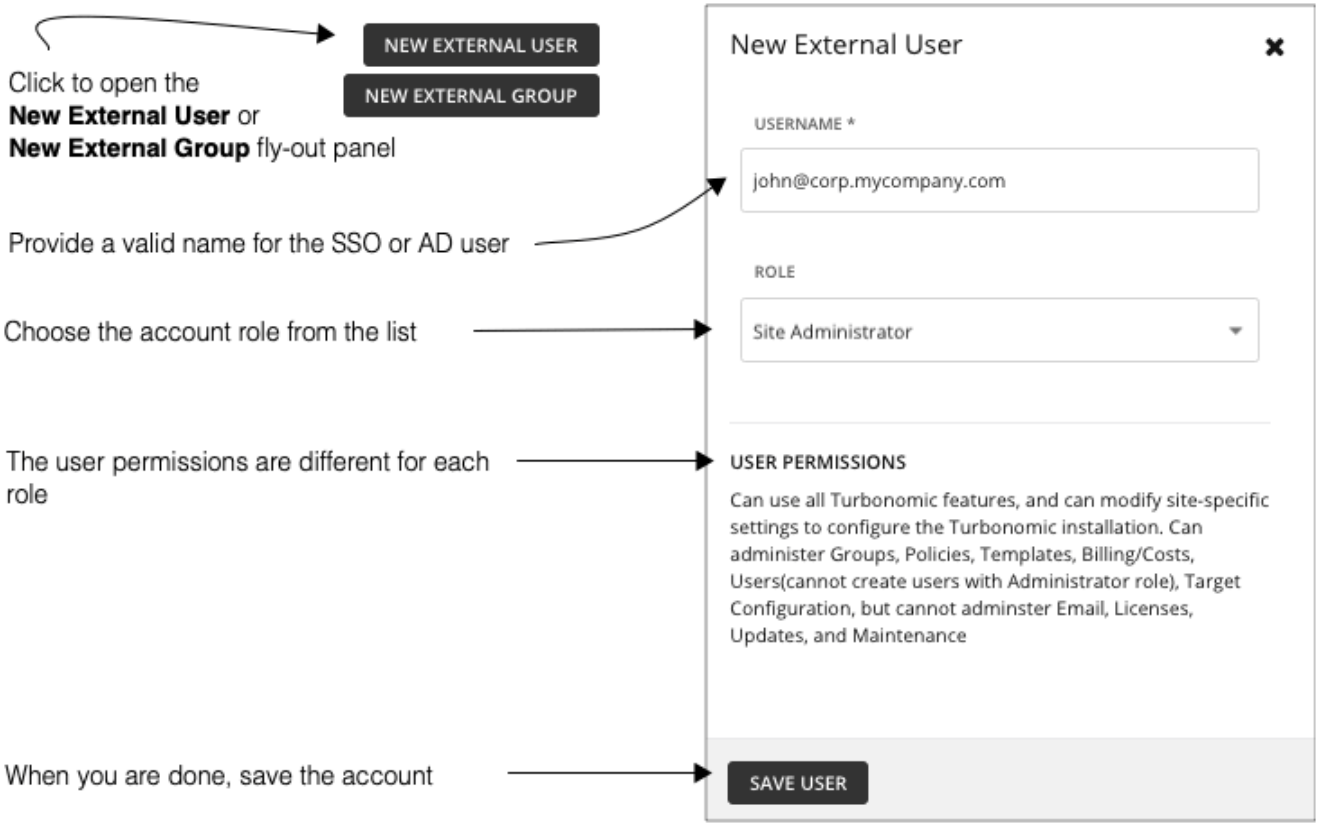
- Active Directory Domain – To authenticate AD groups, specify a domain so that AD can find a given user via the User Principal Name (UPN). If you specify a domain, but not a server, authentication uses any AD server from that domain.
- Active Directory Server – To disable AD groups, specify a server but do not specify a domain. If you specify a domain and a server, authentication will use that server, and will also support groups.

  When you configure an AD server, by default Workload Optimization Manager assumes the AD server port to be 389 or 636. To specify a custom port for the AD server, add the port number to the AD server IP address. For example, `10.10.10.123:444` sets port `444`.
- Secure – Use a secure connection when communicating with AD servers. Note that the AD domain must be configured to use LDAPS, and you must have imported a certificate into the Workload Optimization Manager server.

  Workload Optimization Manager can support LDAP channel binding and LDAP signing. To support these Active Directory features, you must configure secure access.

  For more information, see "Enforcing Secure Access" in the *Installation Guide*.

8. Create or edit an SSO or AD account.


NEW EXTERNAL USER
NEW EXTERNAL GROUP
Click to open the
New External User or
New External Group fly-out panel
Provide a valid name for the SSO or AD user
Choose the account role from the list
The user permissions are different for each role
When you are done, save the account
New External User
USERNAME *
john@corp.mycompany.com
ROLE
Site Administrator
USER PERMISSIONS
Can use all Turbonomic features, and can modify site-specific settings to configure the Turbonomic installation. Can administer Groups, Policies, Templates, Billing/Costs, Users(cannot create users with Administrator role), Target Configuration, but cannot adminster Email, Licenses, Updates, and Maintenance
SAVE USER

This account can be for a user group or for a single user. To add a new account, click **NEW EXTERNAL GROUP** or **NEW EXTERNAL USER**. To edit an existing account, click the account name. To configure an external account, specify:

- Authentication:

  Provide the group or user name for this account. The name you provide must meet certain requirements, depending on the type of account you are creating:

  - **External Group - SSO**

    Provide a name that matches a group the IdP manages.

  - **External Group - AD**

    The group name must match a group that is accessible from the domain and servers that you configured in **EDIT AD**.

  - **External User - SSO**

    Provide a user name that matches a user managed by the IdP.

  - **External User - AD**

    The username must be a valid User Principal Name (UPN). For example, john@corp.mycompany.com.

- Authorization – User Role:

  - Administrator

    Users with this role can use all Workload Optimization Manager features and modify settings to configure the Workload Optimization Manager installation. For Workload Optimization Manager instances hosted in the public cloud, this role is limited to the Workload Optimization Manager representative that manages the instances.

  - Site Administrator

    Users with this role can use all Workload Optimization Manager features and modify site-specific settings to configure the Workload Optimization Manager installation. Users can also administer Groups, Policies, Templates, Billing/Costs, and Target Configuration, but not Email, Licenses, Updates, and Maintenance. Users can create other user accounts, except accounts with the Administrator role.

  - Automator

    Users with this role can use all Workload Optimization Manager features (including Plan, Park, and Place), but cannot configure the Workload Optimization Manager installation or create policies.

  - Deployer

    Users with this role can view all Workload Optimization Manager charts and data, use Place to reserve workloads, and create placement policies and templates. However, users cannot run plans or execute any recommended actions.

  - Advisor

    Users with this role can view all Workload Optimization Manager charts and data, and run plans. However, users cannot use Place to reserve workloads, create policies, or execute any recommended actions.

  - Observer

    Users with this role can view the environment, including the Home Page and Dashboards. Users can also use Search to set a scope to the session. For scope, only VM groups and Resource Groups are supported.

  - Operational Observer

    Users with this role can view the environment, including the Home Page, Dashboards, Groups, and Policies. Users can also use Search to set a scope to the session.

  - Shared Advisor

    Users with this role are scoped users. They can view the Home Page and Dashboards, but only see VMs and Applications. Users cannot execute Workload Optimization Manager actions.

  - Shared Observer

    Users with this role are scoped users. They can view the Home Page and custom Dashboards, but only see VMs and Applications. Users cannot see Executive Dashboards or execute Workload Optimization Manager actions. This is the most restricted user.

  - Report Editor

    Users with this role can create, edit, and delete reports. Due to limits to the reporting license, only one user per instance is allowed to have this role (by default, the local **administrator** user). To assign this role to another user, you must first remove it from the current user. Be sure that the new user is *not* a scoped user.

- Authorization – Scope (optional)

  The scope limits what members of this group can monitor. For example, you can scope for access to only the hosts that support this group's VMs or applications. Click **DEFINE SCOPE** and choose which entities this members of this group can see.

Configuring a Group for SSO Authentication

To use SSO authentication in Workload Optimization Manager, you should configure user groups on the IdP. The IdP can authenticate the group members, and then Workload Optimization Manager can assign the user role and scope according to that group's authentication. To manage personnel changes, you only need to manage the membership in the IdP group. For example, if a user leaves your organization, you only need to remove the member from the group on the IdP. Because authorization on Workload Optimization Manager is by group, that user will not have any authorization settings stored on the Workload Optimization Manager server.

IMPORTANT:
Before you enable SSO for your Workload Optimization Manager installation, *you must configure at least one SSO user with Workload Optimization Manager administrator privileges*. If you do not, then once you enable SSO you will not be able to configure any SSO users in Workload Optimization Manager. To authorize an SSO user as an administrator, use **EXTERNAL AUTHENTICATION** to do one of the following:

- Configure a single SSO user with administrator authorization.

  Add an external user. The username must match an account that is managed by the IdP.

- Configure an SSO user group with administrator authorization.

  Add an external group. The group name must match a user group on the IdP, and that group must have at least one member.

For more information about configuring SSO authentication, see "Single Sign-On Authentication" in the *Installation Guide*.

Specifying a Group in the SAML Response

To support SSO, Workload Optimization Manager recognizes IdP responses that comply with SAML 2.0. To create user groups, for each user response you include an attribute named group, and give the group name as the attribute value. For example, assuming the following users, setting the group attribute for each user assigns that user to the appropriate group.

Users:	Group Attribute:
• George • Paul	Attribute Name=group, AttributeValue=Beatles

Users:	Group Attribute:
▪ John ▪ Ringo	
▪ Smokey ▪ Pete ▪ Ronnie ▪ Claudette ▪ Bobby ▪ Marv	Attribute Name=group, AttributeValue=Miracles

As you specify the user response, to add the user to a group you include a group attribute. For example, to add a user to a group named turbo_admin_group, you would include the following attribute in that user's SAML response:

```
<saml2:Attribute
  Name="group"
  NameFormat="urn:oasis:names:tc:SAML:2.0:attrname-format:unspecified">
  <saml2:AttributeValue
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:type="xs:string">
        turbo_admin_group
  </saml2:AttributeValue>
</saml2:Attribute>
```

Setting Group Authorization in Workload Optimization Manager

To set an account role and scope to a user group, you must use the group name that you specify as the value in the given SAML group attribute. In the above example, the group value is turbo_admin_group. To set authorization for that group:

1. Open the User Management page to EXTERNAL AUTHENTICATION.

   Navigate to **Settings > User Management**, and display the **EXTERNAL AUTHENTICATION** view.

2. Create a new External Group

   Click **NEW EXTERNAL GROUP**.

3. Provide the group name.

   Be sure to use the name that you specify in the group attribute of the SAML response. For the above example, use the name turbo_admin_group.

4. Specify the group's authorization

   For the above example, since this is turbo_admin_group, you should set the **ADMINISTRATOR** role, and you should not set any scope (grant full access to the environment).

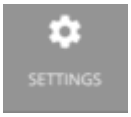
   After you configure this group in Workload Optimization Manager, then any member of turbo_admin_group that the IdP returns will have full administrator privileges on your Workload Optimization Manager installation.

Maintenance: Logging and Troubleshooting

The Maintenance Options Page provides tools to set logging levels and to export data for technical support, and import diagnostic files from Technical Support. Many of these tools are for advanced users. You should contact Cisco technical support before you use them.

To execute these actions, navigate to the Maintenance Options page:

1. Navigate to the Settings Page.

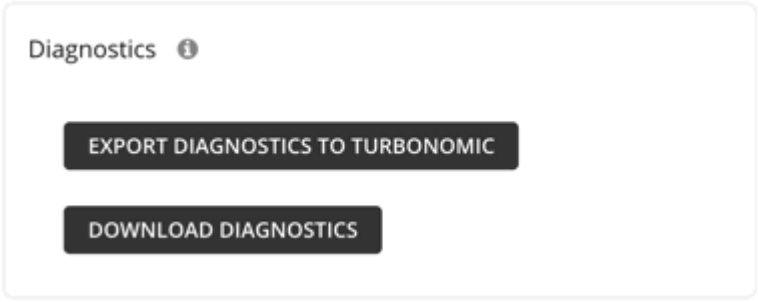


Click to navigate to the Settings Page.

2. Choose Maintenance Options.

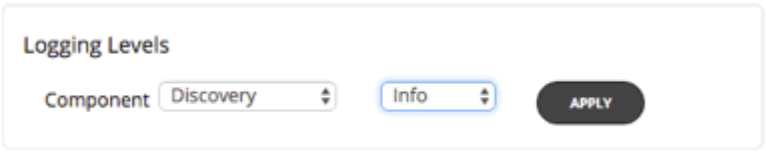
Maintenance Options

Diagnostics



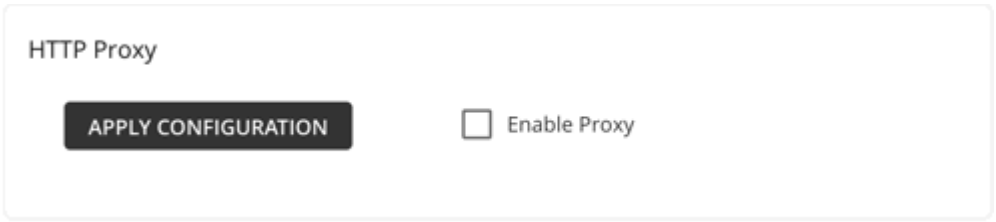
If you are experiencing problems with Workload Optimization Manager, your support engineer might request that you export diagnostic data. You can export the data and then send it to the support engineer as requested.

Logging Levels



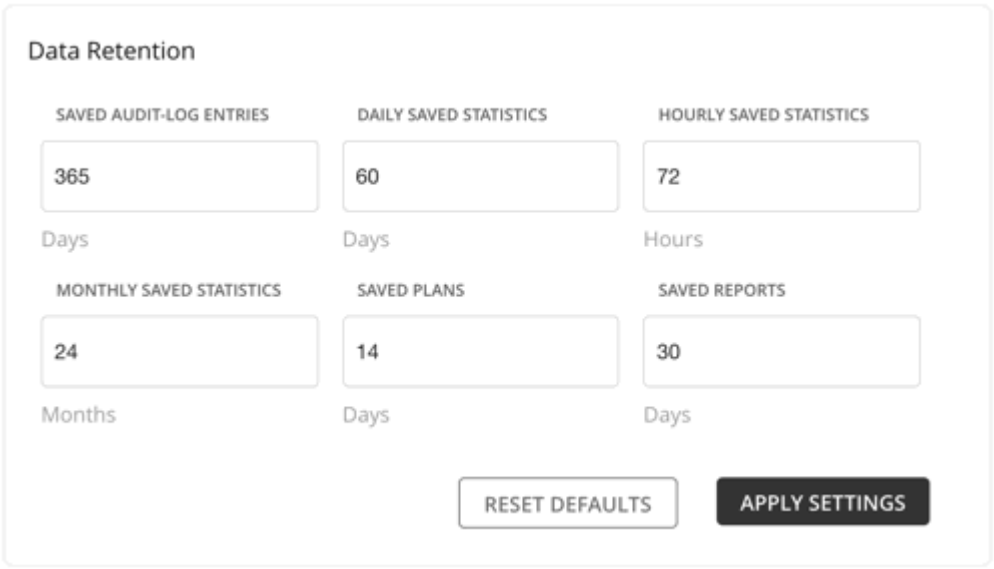
You can set the level of logging for different components of the Workload Optimization Manager platform. You should be aware that setting more verbose logging levels increases the disk space required to store the log files. You normally change these settings only while you're working with a Workload Optimization Manager support engineer.

HTTP Proxy



If your environment requires an HTTP proxy for Workload Optimization Manager to access the web, provide the credentials here.

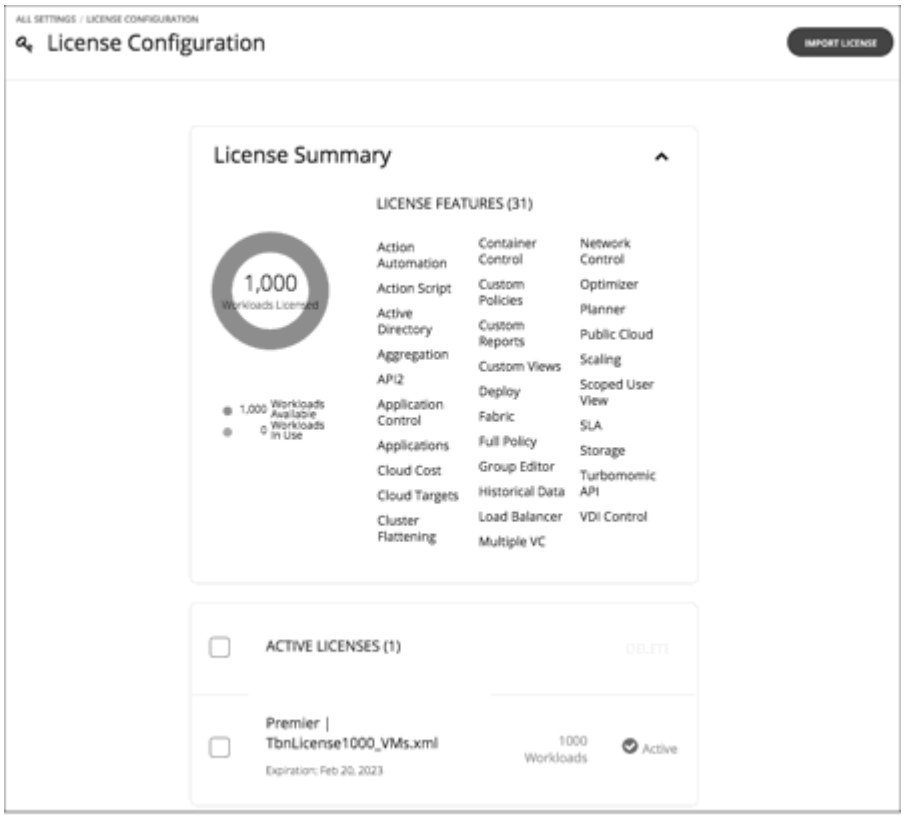
Data Retention



Workload Optimization Manager 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. Workload Optimization Manager 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.

License Configuration



To activate the full range of Workload Optimization Manager features, you must purchase the appropriate license. When you purchase the license, Cisco sends an e-mail message with instructions on how to obtain the license key.

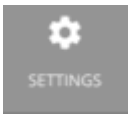
A product license enables specific features as well as a specific number of workloads that you can manage. You can add additional licenses to Workload Optimization Manager as a way to increase the number of workloads your installation can manage. Note that as you add more licenses, they must all support the same feature set.

The License Configuration page shows you:

- The number of workloads you can manage under this license
- How many workloads are currently in use
- The set of features this license enables
- A list of each license and it status

To navigate to the License Configuration page:

1. Navigate to the Settings Page.

2. Choose License.

To activate a license or to update your current license:

1. Obtain your license.

   Cisco sends an e-mail message with instructions on how to obtain the license key. Save the license file on your local machine so you can upload it to your Workload Optimization Manager installation.

2. Apply the license to your Workload Optimization Manager installation.

   First click **IMPORT LICENSE**. Then browse to the license file that you saved and open it. Or you can drag the file into the **Enter License** fly-out.

   After you have uploaded the file, click **SAVE**.

After you have activated your license, you can then add more licenses to increase your workload coverage, or you can license a higher feature set.

> **NOTE:**
> This only applies only to legacy Workload Optimization Manager customers. As you apply new licenses to Workload Optimization Manager, you must be sure that they are for the same edition or feature set. If you try to apply an incompatible license file, Workload Optimization Manager displays an *Invalid Feature Set* error. To apply the new license you must either delete your current license so you can install the new feature set, or you must obtain a different license file that matches your current feature set.
>
> After you install a new license, you should clear your browser cache and reload the Workload Optimization Manager user interface.

To increase your licensed workload coverage:

1. Obtain your additional license.

   Note that your additional licenses must match the feature set of your current license.

2. Apply the license to your Workload Optimization Manager installation.

To upgrade your license to a higher feature set:

1. Obtain your new license for the new features.

   You should obtain a license that supports at least the same number of workloads as your current license.

2. Delete your current license from Workload Optimization Manager.

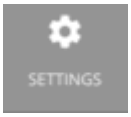
   On the license page, select all the licenses that you currently have installed, then click **DELETE**.

3. Apply the license to your Workload Optimization Manager installation.

Email Settings

Configure email settings to enable email communication from Workload Optimization Manager.

1. Navigate to the Settings Page.

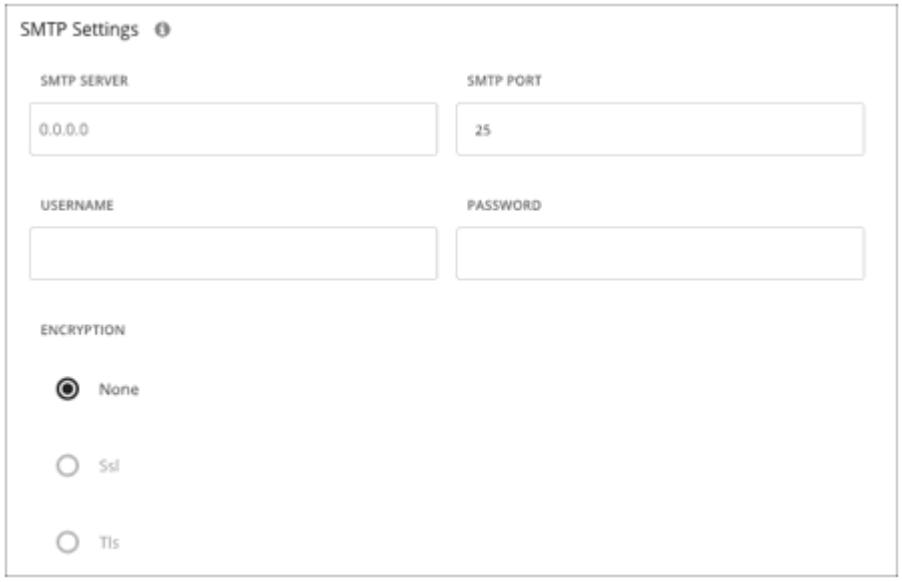
   Click to navigate to the Settings Page. From there, you can perform a variety of Workload Optimization Manager configuration tasks.

2. Navigate to the Email Settings Page.

   From here, you can configure:

   - SMTP Settings
   - General Email Settings

SMTP Settings



The SMTP Settings fields identify the mail relay server you use on your network to enable email communication from Workload Optimization Manager.

If the server requires authentication, provide the username and password here. You can also choose the following encryption options for notifications:

- None
- Ssl
- Tls

General Email Settings

Use this setting to specify the return address (the FROM address) for emails that Workload Optimization Manager generates and sends.