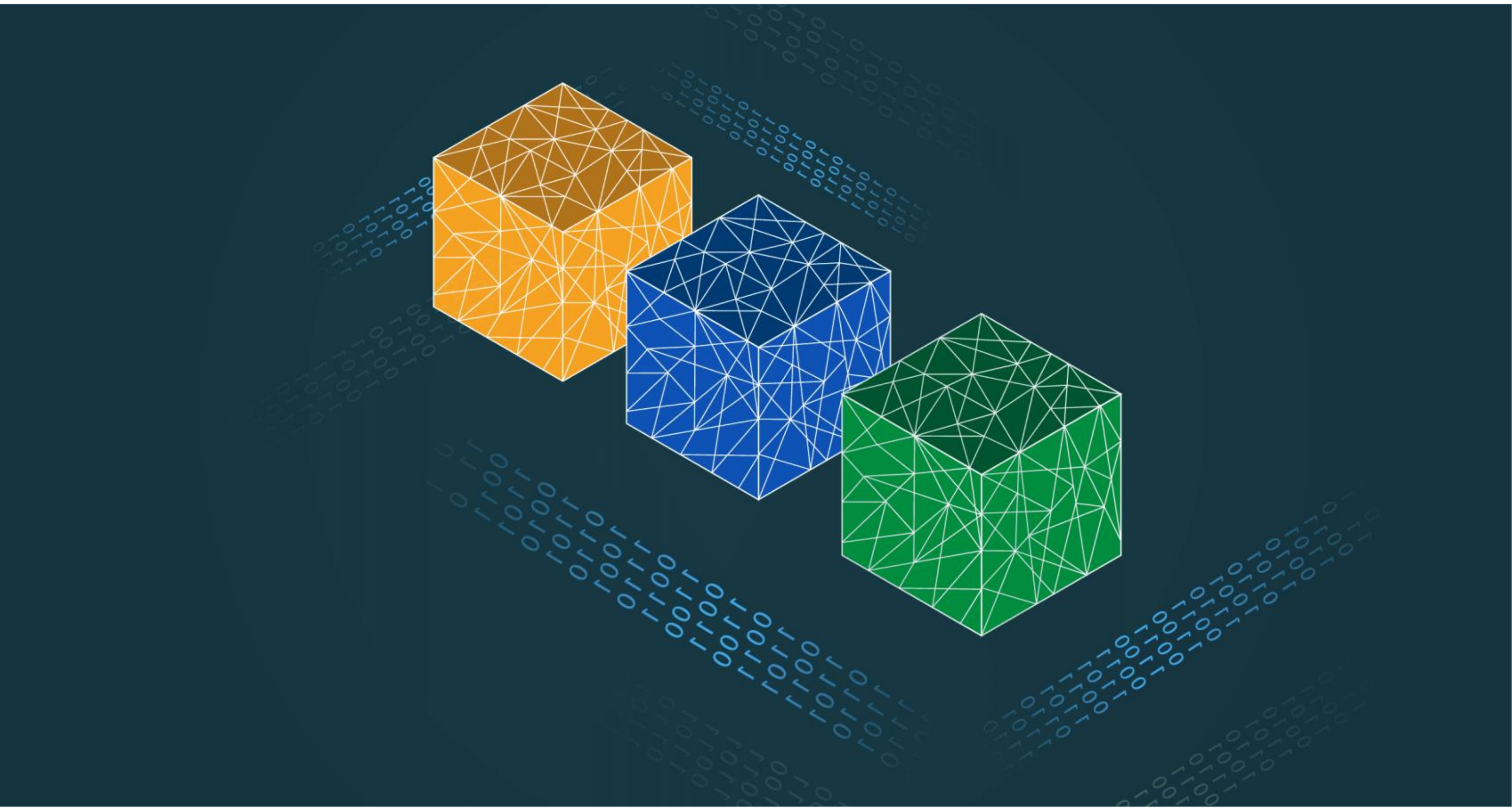
ProphetStor



Federator.ai GPU Booster User Guide

Federator.ai GPU Booster version 5.2.1

User Guide

ProphetStor Data Services, Inc.
830 Hillview Court, Suite 100
Milpitas, CA 95035 USA
Phone: 1.408.508.6255
Website: www.prophetstor.com

12.3.2024

Contents

Overview

Federator.ai GPU Booster leverages advanced multi-layer correlation and machine learning technologies on Kubernetes to address the challenges of managing GPU resources in competitive AI and ML environments. Designed for MultiTenant settings, it efficiently orchestrates AI/ML workloads, particularly in large language model training.

By recommending GPU allocations to accommodate the varying demands of AI training workloads, Federator.ai GPU Booster optimizes resource usage and enhances training efficiency, enabling organizations to fully harness their AI/ML capabilities, thereby accelerating progress in the field.

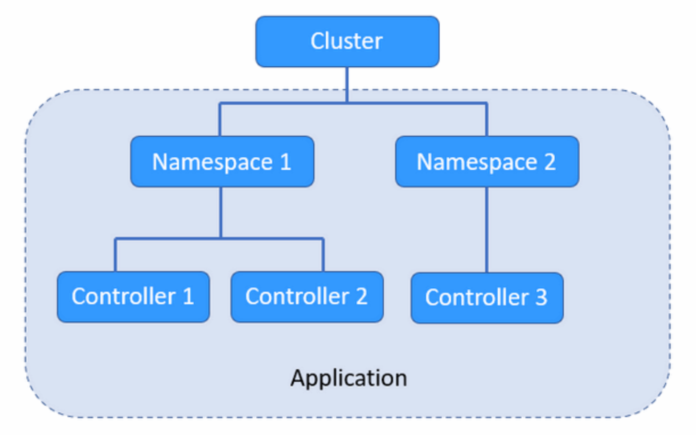
Using advanced machine learning algorithms to predict application workloads, Federator.ai GPU Booster offers:

- Efficient GPU resource allocation by leveraging Multi-Instance GPU (MIG) technology to run MultiTenant training sessions in parallel, efficiently allocating the necessary resources based on ML algorithms
- Maximize total throughput by strategically reallocating GPU resources among concurrent GPU workloads
- AI-based workload prediction for containerized applications in Kubernetes
- Resource recommendations based on workload prediction, application, Kubernetes, and other related metrics
- Automatic provisioning of CPU/memory for generic Kubernetes application controllers/namespaces
- Actual cost and potential savings based on recommendations for clusters, Kubernetes applications, and Kubernetes namespaces

If you have not installed Federator.ai GPU Booster yet, refer to your *Federator.ai GPU Booster Installation Guide* for information.

Terminology

Application – Defined by Federator.ai GPU Booster as a group of Kubernetes controllers that work together to serve tasks from the view of the end user. For example, an e-commerce web application consists of controllers for frontend and backend and can be considered as an application. An application is not a Kubernetes object.



Auto Provisioning – The ability to automatically deploy CPU and memory resource recommendations to controllers and namespaces of generic applications in Kubernetes clusters based on pre-defined profiles.

Container – An object that contains a software module with everything needed to run an application.

Controller – In Kubernetes, controllers are control loops that watch the state of your cluster, then make or request changes where needed. Each controller tries to move the current cluster state closer to the desired state. The types of controllers supported by Federator.ai GPU Booster are *Deployment* and *StatefulSet*. Additionally, Federator.ai GPU Booster supports *DeploymentConfig* controllers for OpenShift.

Cluster – A Kubernetes cluster with one or more nodes.

Deployment – A Deployment provides declarative updates for Pods and ReplicaSets. The user describes a desired state in a deployment and the deployment controller changes the actual state to the desired state at a controlled rate.

GPU Cluster – A Kubernetes cluster that includes GPU nodes

GPU Node – A Kubernetes cluster node that provides GPU resources

GPU Workload – Any deployment/job that utilizes GPUs in a Kubernetes cluster

Namespace – Kubernetes supports multiple virtual clusters backed by the same physical cluster. These virtual clusters are called namespaces.

Microservice – Also known as controllers, microservices are independent, modular Kubernetes components that work together as a single application.

Node – In Kubernetes, nodes are server-like machines, such as a virtual machine running complete systems and multiple applications. There can be master nodes and worker nodes.

Pod – A group of one or more containers with shared storage/network resources and a specification for how to run the containers. Typically, one container runs in each pod.

Replica – A copy of a pod running for an application.

StatefulSet – A Kubernetes object that manages stateful applications. Unlike a Deployment, a StatefulSet maintains a sticky identity for each of its pods that remains the same across any rescheduling.

Related topics:

Federator.ai GPU Booster Administration Portal

Configure Kubernetes Clusters

Configure Applications

Auto Provisioning

Autoscaling

Getting Started

After installation of Federator.ai GPU Booster, you must access the Federator.ai GPU Booster portal in order to configure your system.

Access the Federator.ai GPU Booster Portal

To access the Federator.ai GPU Booster administration portal, use the URL that is displayed at the end of the installation process.

You can also find the URL for the Federator.ai GPU Booster administration portal via the following methods:

Kubernetes

In a Kubernetes environment, use the `kubectl` command to find the administration portal service port number and node IP address.

```
# kubectl get svc -n federatorai |grep federatorai-dashboard-frontend-public
```

The output will look something like this:

```
federatorai-dashboard-frontend-public NodePort 10.103.181.133 <none> 9001:31012/TCP
```

Get the node’s IP to access (INTERNAL-IP).

```
$kubectl get nodes -o wide
```

For example:

```
# kubectl get nodes -o wide
NAME     STATUS   ROLES    AGE   VERSION   INTERNAL-IP    EXTERNAL-IP   OS-IMAGE
KERNEL-VERSION          CONTAINER-RUNTIME
h7-130   Ready    master   35d   v1.18.5   172.31.7.130   <none>        CentOS Linux 7 (Core)
3.10.0-957.el7.x86_64   docker://19.3.13
h7-131   Ready    <none>   35d   v1.18.5   172.31.7.131   <none>        CentOS Linux 7 (Core)
3.10.0-957.el7.x86_64   docker://19.3.13
h7-132   Ready    <none>   35d   v1.18.5   172.31.7.132   <none>        CentOS Linux 7 (Core)
3.10.0-957.el7.x86_64   docker://19.3.13
h7-133   Ready    <none>   35d   v1.18.5   172.31.7.133   <none>        CentOS Linux 7 (Core)
3.10.0-957.el7.x86_64   docker://19.3.13
```

The URL will be `https://172.31.7.130:31012`.

OpenShift

In an OpenShift environment, use the `oc get route` command to find the URL.

```
# oc get route -n federatorai|grep federatorai-dashboard-frontend
```

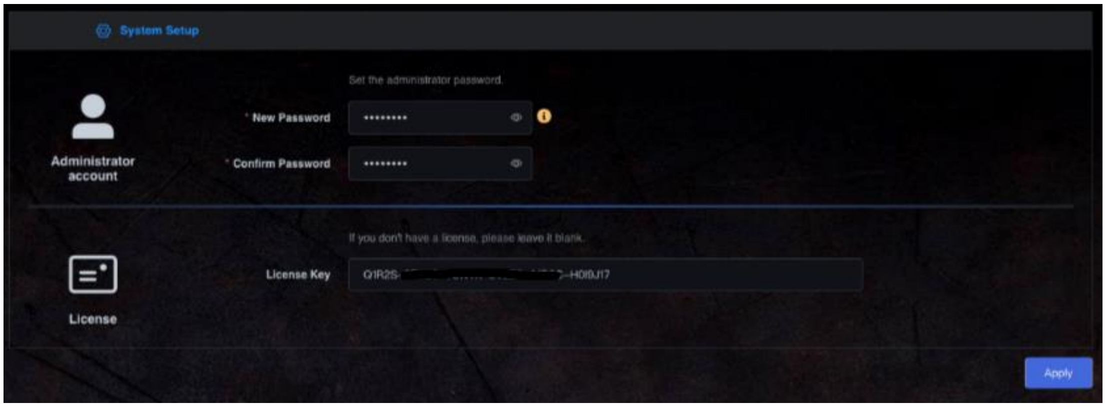
The output will look something like this:

```
federatorai-dashboard-frontend “federatorai-dashboard-frontend-federatorai.apps.ocp4.172-31-11-
30.nip.io”
```

The URL will be `https://federatorai-dashboard-frontend-federatorai.apps.ocp4.172-31-11-30.nip.io`

Setup Wizard

The first time you log in after installation of Federator.ai GPU Booster, a setup wizard launches that allows you to set up the new password and optionally apply a license key.



Set the administrator password.

* New Password: ••••••••

* Confirm Password: ••••••••

If you don't have a license, please leave it blank.

License Key: Q1R2S-[redacted]-H0I9J17

Related topics:

Terminology

Federator.ai GPU Booster Administration Portal

Configure Applications

Configure Kubernetes Clusters

Federator.ai GPU Booster Administration Portal

The Federator.ai GPU Booster administration portal displays the overall health of each cluster/GPU/GPU workload, as well as application workload and resource recommendations. Information is presented in tables and charts.

Portal Sections

The portal is separated into the following sections:

- Dashboard – Overall system information, including the number of monitored resources, as well as cluster and Kubernetes application workload predictions and recommendations.
- GPU Management – GPU health information, including GPU utilization, GPU memory utilization, temperature, and power usage, as well as GPU workloads and their operational metrics. Recommendations of GPU resources for GPU workloads and GPU resource configuration for GPU clusters are also included.
- Cluster Overview - Cluster and Kubernetes node health information, including CPU utilization, memory utilization, disk capacity.
- Predictions and Planning – Forecasting tools, including actual CPU and memory usage observations, predicted workload usage, utilization analysis, and recommendations for Kubernetes resources.
- Cost Management – Cost analysis, cost trends, and cost optimization for resources at different levels, including Kubernetes clusters and nodes, as well as Kubernetes namespaces and applications.
- Configuration – Configuration of clusters, Kubernetes applications and controllers, as well as system configuration, including resetting the admin password, metrics data source, system notifications, licensing, price books, and system configuration backup/restore.
- Events – System events that have occurred.
- Alerts – Triggered alerts and configuration of rules that monitor clusters, nodes, namespaces, applications, and controllers.

Portal Icons

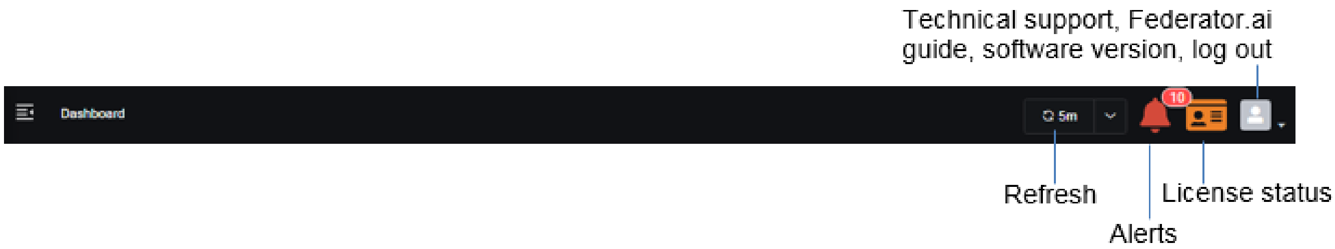
To make it easy to distinguish between cluster types and providers, the following icons are used throughout the portal:

Icon	Meaning
🗄	On-premises provider
☸	Kubernetes clusters

Common Administration Portal Functions

The administration portal presents information in tables and charts. At the top right of each portal page, you can do the following:

- Refresh statistics
- View system alerts
- Check license status
- Get technical support contact information
- View Federator.ai GPU Booster product documentation
- Display the product software version
- Log out


Technical support, Federator.ai guide, software version, log out
Dashboard
5m
10
Refresh
Alerts
License status

Refresh Statistics

By default, Federator.ai GPU Booster information is refreshed every five minutes. To change the interval, click the drop-down at the top right and select a 1, 5, 15, or 30-minute interval.

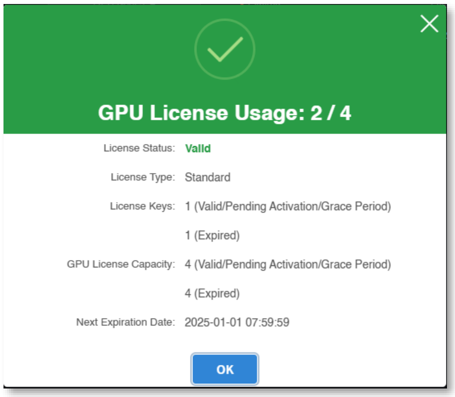
To force a refresh, click *Refresh Now* where the current interval is displayed.

System Alerts

Click the *Alerts* icon at the top right of the dashboard when it is red to jump to the *Alerts* page to view alerts that have been generated when a configured rule is triggered.

License Status

Click the *License* icon at the top right of the dashboard to see Federator.ai GPU Booster license information, including license status and type, number and type of license keys, licensed capacity and usage, as well as license expiration. When the icon is green, all licenses are valid. Orange indicates a trial license or a situation that requires attention, such as a license is near expiration, a license is in a grace period, or the number of monitored resources exceeds the license limit. Red requires immediate attention because it indicates a license has expired.



GPU License Usage: 2 / 4

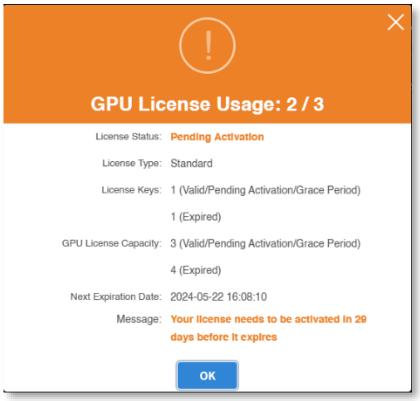
License Status: Valid

License Type: Standard

License Keys: 1 (Valid/Pending Activation/Grace Period)
1 (Expired)

GPU License Capacity: 4 (Valid/Pending Activation/Grace Period)
4 (Expired)

Next Expiration Date: 2025-01-01 07:59:59



GPU License Usage: 2 / 3

License Status: Pending Activation

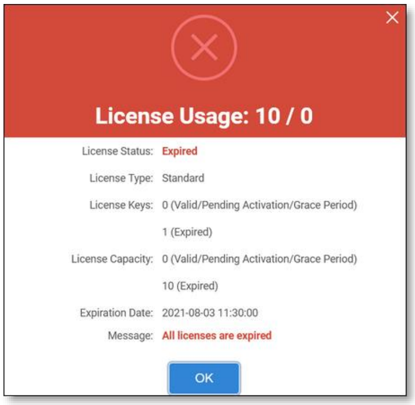
License Type: Standard

License Keys: 1 (Valid/Pending Activation/Grace Period)
1 (Expired)

GPU License Capacity: 3 (Valid/Pending Activation/Grace Period)
4 (Expired)

Next Expiration Date: 2024-05-22 16:08:10

Message: Your license needs to be activated in 29 days before it expires



License Usage: 10 / 0

License Status: Expired

License Type: Standard

License Keys: 0 (Valid/Pending Activation/Grace Period)
1 (Expired)

License Capacity: 0 (Valid/Pending Activation/Grace Period)
10 (Expired)

Expiration Date: 2021-08-03 11:30:00

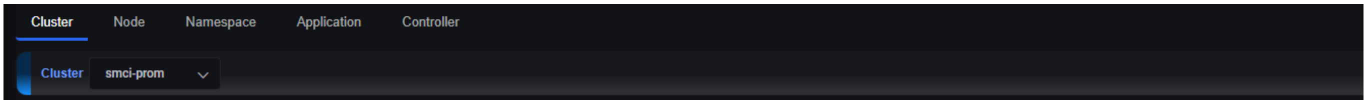
Message: All licenses are expired

User Functions

Click the *User* icon to contact technical support, view the Federator.ai GPU Booster product documentation, display the product software version, or log out from the system.

Filters

A panel appears on most pages that allows you to filter the display of information. Depending upon the page and cluster type, you may be able to select a cluster, application, namespace, controller, time range, etc.

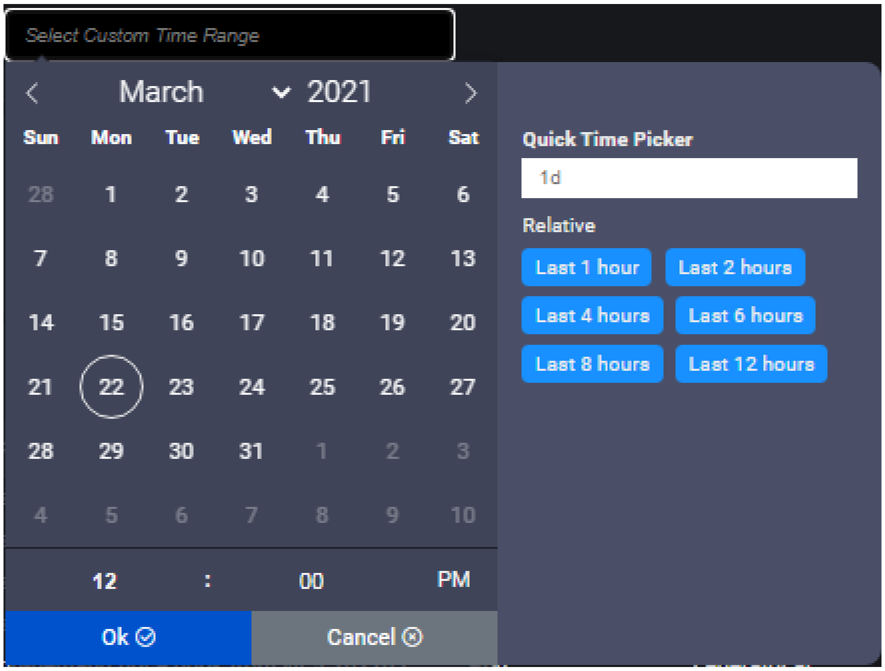


Cluster: smci-prom

Specify Time Range

When a chart allows you to specify a time range to display data, you can select a predefined time frame (e.g., last 1 hour, last 24 hours) from the drop-down box or you can specify a custom time range via one of the following methods:

- Use the calendar to select the start and end dates.
- Specify the number of hours (e.g., 5h), days (e.g., 5d), weeks (e.g., 3w), or months (e.g., 6m) in the *Quick Time Picker* box.
- Click a predefined relative time range.



Select Custom Time Range

March 2021

Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

12 : 00 PM

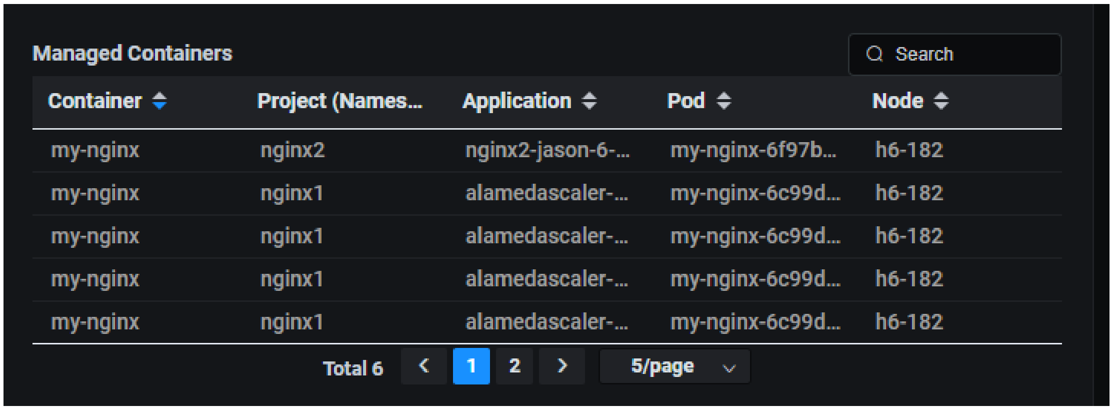
Quick Time Picker

1d

Relative

Search/Sort Information in Tables

Click a column heading to sort the entire list based upon values in that column. The blue highlighted triangle or inverted triangle on the column indicates the direction of the sort. To search, type a name or value in the *Search* box; clear the search field to return to the full view of the list. As soon as you start typing, only those items that have matching text are displayed. You can also determine how many rows to show per page (5, 10, or 20).



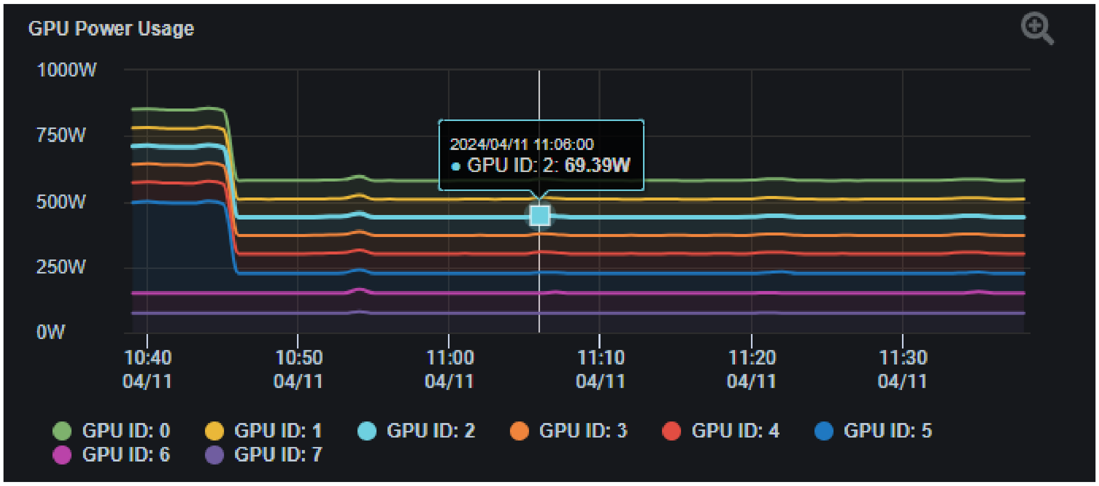
Managed Containers

Container	Project (Names...	Application	Pod	Node
my-nginx	nginx2	nginx2-jason-6-...	my-nginx-6f97b...	h6-182
my-nginx	nginx1	alamedascaler-...	my-nginx-6c99d...	h6-182
my-nginx	nginx1	alamedascaler-...	my-nginx-6c99d...	h6-182
my-nginx	nginx1	alamedascaler-...	my-nginx-6c99d...	h6-182
my-nginx	nginx1	alamedascaler-...	my-nginx-6c99d...	h6-182

Total 6 < 1 2 > 5/page

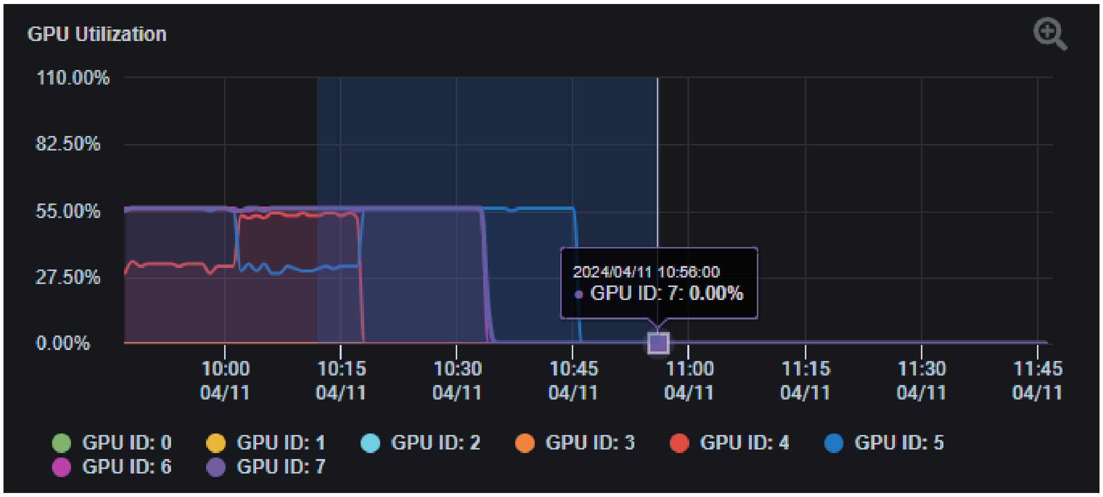
Show/Hide Metrics in Charts

Click anywhere on a chart to see values for a specific point in time. Highlight or click the key at the bottom of the chart to show/hide individual metrics.

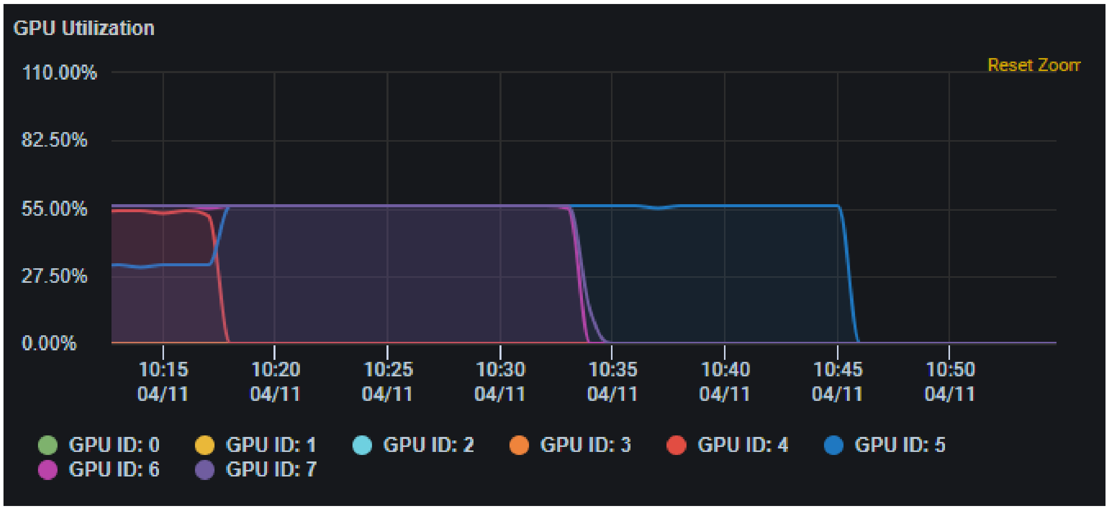


Zoom In/Out of Charts

Click and drag a pointer in a chart to zoom into a specific time frame of interest.



Click *Reset Zoom* to return to the original time frame.



Related topics:

Common Administration Portal Functions

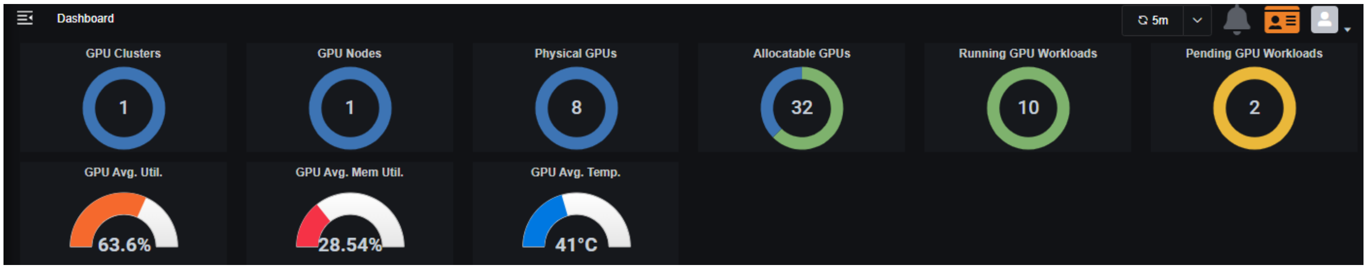
Dashboard

Licenses

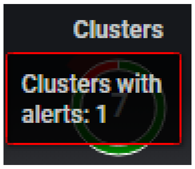
Alerts

Dashboard

The *Dashboard* displays the number of monitored resources in Kubernetes GPU clusters, as well as cluster and GPU resource utilization and GPU workloads.



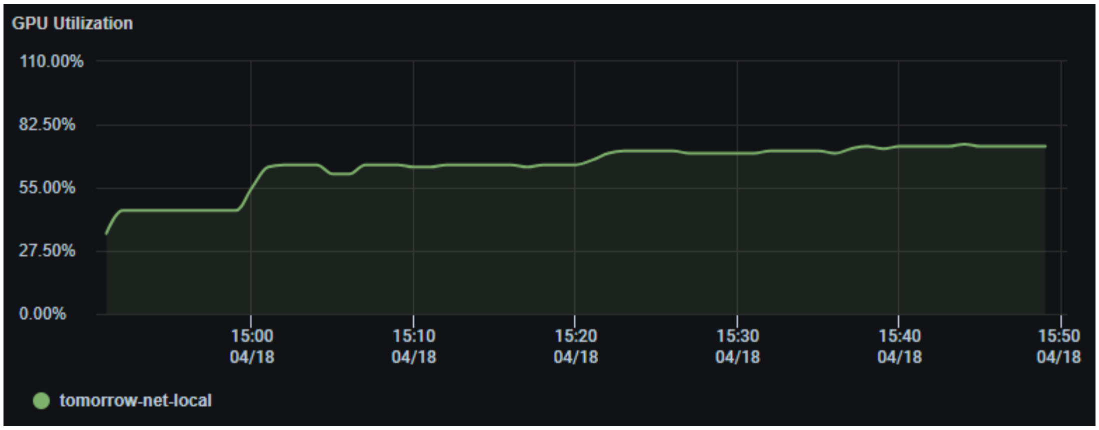
The *Monitored Resources* section displays red to indicate there are alerts for resources. Highlight the resource to see how many alerts are active.



Click to jump to the Alerts / *Activity* Page to view active and resolved alerts for all related resources (e.g., nodes).

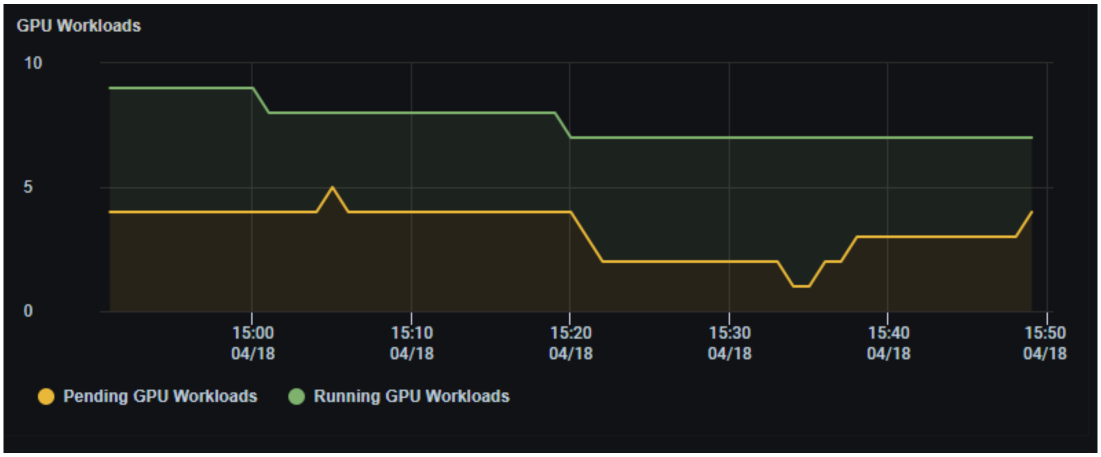
GPU Utilization

This chart displays the average GPU utilization of each cluster managed by Federator.ai GPU Booster. Select a cluster from a drop-down list to filter out other clusters.



GPU Workloads

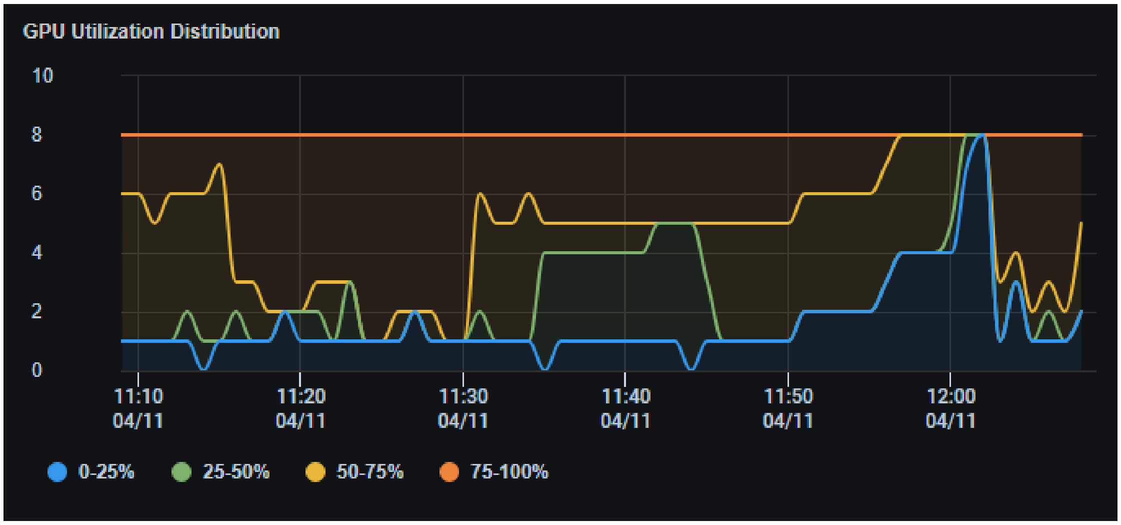
The *GPU Workloads* chart displays the number of running GPU workloads and pending GPU workloads from all clusters through time. Select a cluster from a drop-down list to see only the GPU workloads from that cluster.



GPU Utilization Distribution

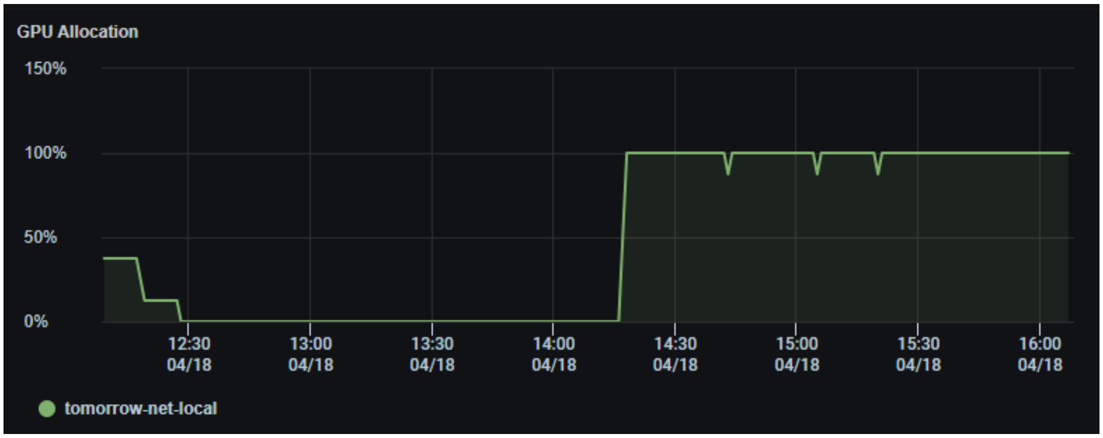
The *GPU Utilization Distribution* chart displays the number of GPUs in different ranges of utilization. There are four ranges of GPU utilizations: 0-25%, 25-50%, 50-75%, and 75-100%. This chart shows how GPUs are utilized in a GPU cluster through time. The more GPUs are in higher utilization ranges, the better overall GPU utilization of the entire cluster.

Select a cluster from the drop-down list to see the GPU utilization distribution of a specific cluster.



GPU Allocation

The *GPU Allocation* chart displays the percentage of GPU resources are allocated for active GPU workloads of each cluster. Select a cluster from the drop-down list to see GPU allocation of a specific cluster.



Related topics:

Common Administration Portal Functions

Refresh Statistics

License Status

User Functions

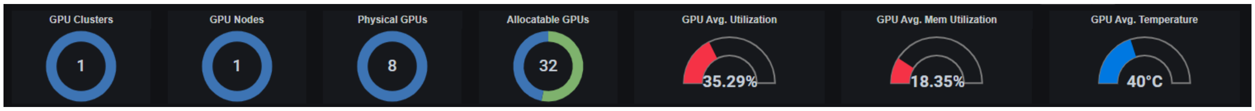
Search/Sort Information in Tables

Show/ Hide Information in Charts

Zoom In/Out of Charts

GPU Management - GPU Health

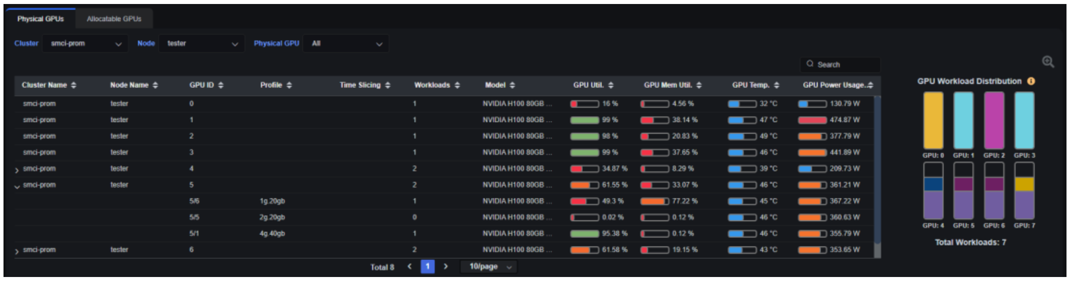
The *GPU Health* page displays the current and past history of utilization of GPUs in the GPU clusters that are managed by Federator.ai GPU Booster. At the top of the page, it displays the current number of GPU clusters, GPU nodes, physical GPUs and logical/allocatable GPUs in the system. It also shows the average GPU utilization, GPU memory utilization and the average temperature of all GPUs.



GPU Clusters	GPU Nodes	Physical GPUs	Allocatable GPUs	GPU Avg. Utilization	GPU Avg. Mem Utilization	GPU Avg. Temperature
1	1	8	32	35.29%	18.35%	40°C

Physical GPUs

This tab shows the list of the physical GPUs and the current status. Using the dropdown menu to view specific cluster or cluster node with GPUs. In physical GPU table, detail information of each GPU such as GPU ID, GPU model, GPU Profile, and current operational metrics are display. It also shows the current GPU workload distribution on the GPU's of a specific GPU node.



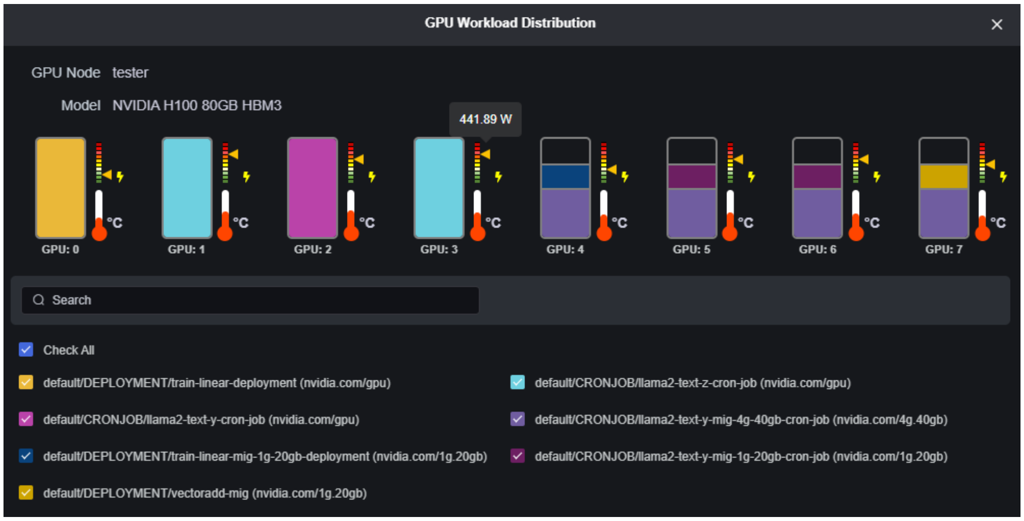
Cluster: smci-prom   Node: tester   Physical GPU: All

Cluster Name	Node Name	GPU ID	Profile	Time Slicing	Workloads	Model	GPU Util.	GPU Mem Util.	GPU Temp.	GPU Power Usage..
smci-prom	tester	0			1	NVIDIA H100 80GB ...	16 %	4.56 %	32 °C	130.79 W
smci-prom	tester	1			1	NVIDIA H100 80GB ...	99 %	38.14 %	47 °C	474.87 W
smci-prom	tester	2			1	NVIDIA H100 80GB ...	98 %	20.83 %	49 °C	377.79 W
smci-prom	tester	3			1	NVIDIA H100 80GB ...	99 %	37.65 %	46 °C	441.89 W
smci-prom	tester	4			2	NVIDIA H100 80GB ...	34.87 %	8.29 %	39 °C	209.73 W
smci-prom	tester	5			2	NVIDIA H100 80GB ...	61.55 %	33.07 %	46 °C	361.21 W
		5/6	1g.20gb		1	NVIDIA H100 80GB ...	49.3 %	77.22 %	45 °C	367.22 W
		5/5	2g.20gb		0	NVIDIA H100 80GB ...	0.02 %	0.12 %	46 °C	360.63 W
		5/1	4g.40gb		1	NVIDIA H100 80GB ...	95.38 %	0.12 %	46 °C	355.79 W
smci-prom	tester	6			2	NVIDIA H100 80GB ...	61.58 %	19.15 %	43 °C	353.65 W

Total 8 ‹ 1 › 10/page

Total Workloads: 7

Click on the "zoom" icon of the GPU Workload Distrbution widget to see more detailed information about the workloads that allocate each GPU.



GPU Node tester

Model NVIDIA H100 80GB HBM3

☑ Check All

☑ default/DEPLOYMENT/train-linear-deployment (nvidia.com/gpu)

☑ default/CRONJOB/llama2-text-z-cron-job (nvidia.com/gpu)

☑ default/CRONJOB/llama2-text-y-cron-job (nvidia.com/gpu)

☑ default/CRONJOB/llama2-text-y-mig-4g-40gb-cron-job (nvidia.com/4g.40gb)

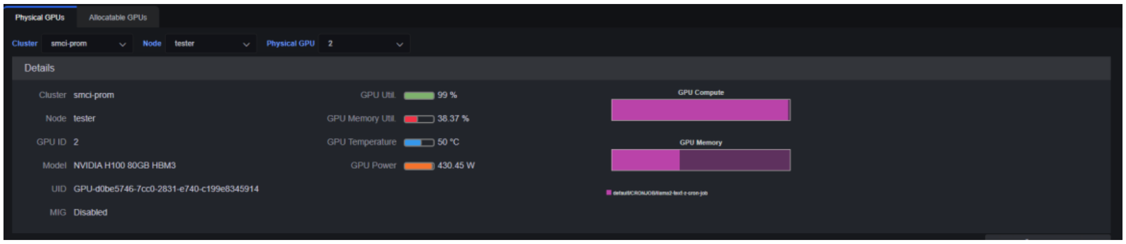
☑ default/DEPLOYMENT/train-linear-mig-1g-20gb-deployment (nvidia.com/1g.20gb)

☑ default/CRONJOB/llama2-text-y-mig-1g-20gb-cron-job (nvidia.com/1g.20gb)

☑ default/DEPLOYMENT/vectoradd-mig (nvidia.com/1g.20gb)

If a physical GPU is configured with *Multiple Instance GPU* (MIG), click the expansion icon (>) to see the MIG configuration details and their utilization status. If a specific GPU is selected, a detailed status of the GPU will be displayed.

Here is an example of information shown of a physical GPU allocated by an active GPU workload.



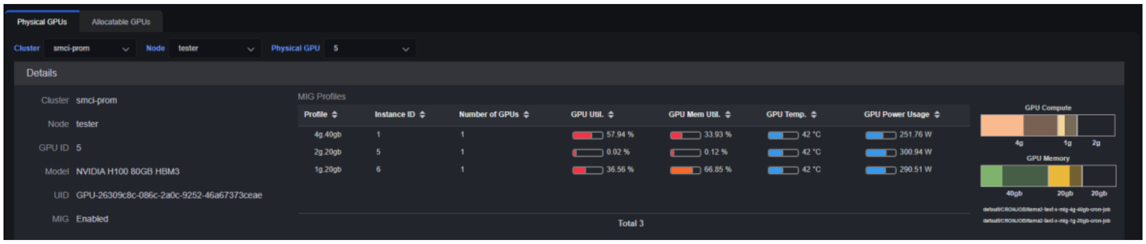
Cluster smci-prom
Node tester
Physical GPU 2

Details

Cluster smci-prom
Node tester
GPU ID 2
Model NVIDIA H100 80GB HBM3
UID GPU-d0be5746-7cc0-2831-e740-c199e8345914
MIG Disabled

GPU Util. 99 %
GPU Memory Util. 38.37 %
GPU Temperature 50 °C
GPU Power 430.45 W

Here is an example of two GPU workloads using two different MIG instances from the same physical GPU.



Cluster smci-prom
Node tester
Physical GPU 5

Details

Cluster smci-prom
Node tester
GPU ID 5
Model NVIDIA H100 80GB HBM3
UID GPU-26309c8c-086c-2a0c-9252-46a67373ceae
MIG Enabled

MIG Profiles

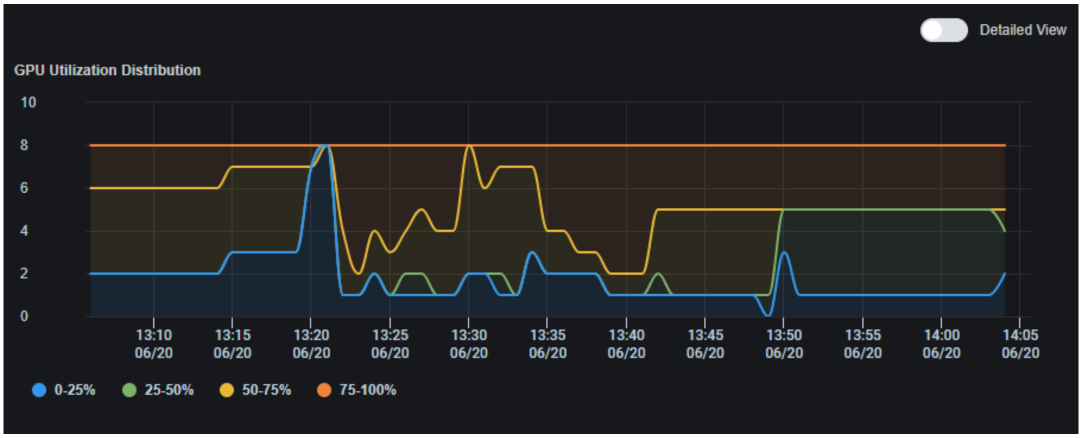
Profile	Instance ID	Number of GPUs	GPU Util.	GPU Mem Util.	GPU Temp.	GPU Power Usage
4g.40gb	1	1	57.94 %	33.93 %	42 °C	251.76 W
2g.20gb	5	1	0.02 %	0.12 %	42 °C	300.94 W
1g.20gb	6	1	36.56 %	66.85 %	42 °C	290.51 W

Total 3

GPU Compute
4g
1g
2g
GPU Memory
40gb
20gb
20gb
default/CRONJOB/llama2-text-x-mig-4g-40gb-cron-job
default/CRONJOB/llama2-text-x-mig-1g-20gb-cron-job

GPU Utilization Distribution

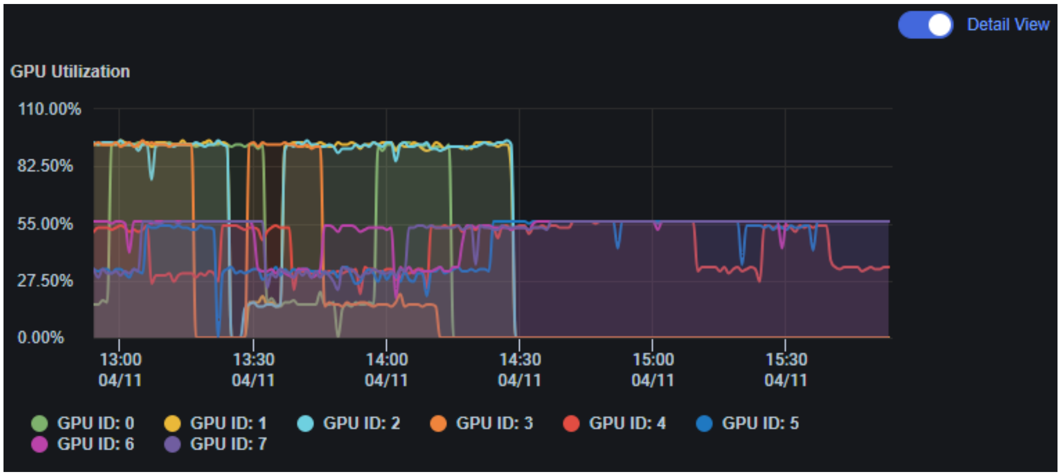
This chart displays the number of GPUs of selected GPU clusters or GPU cluster nodes in different ranges of GPU utilization. This provides a quick look on how GPUs are utilized in the past.



To see individual GPU's utilization history, click on the "Detailed View" button.

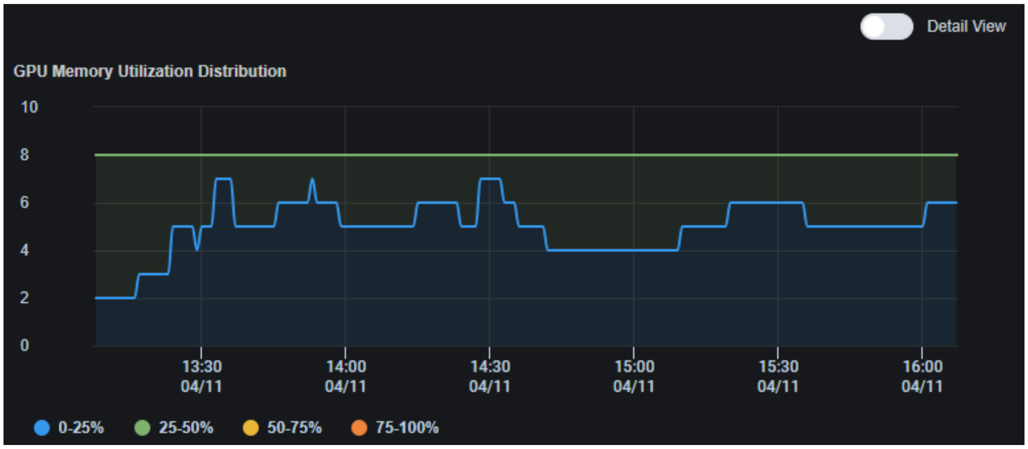
GPU Utilization

This chart displays the utilization history of each GPU.



GPU Memory Utilization Distribution

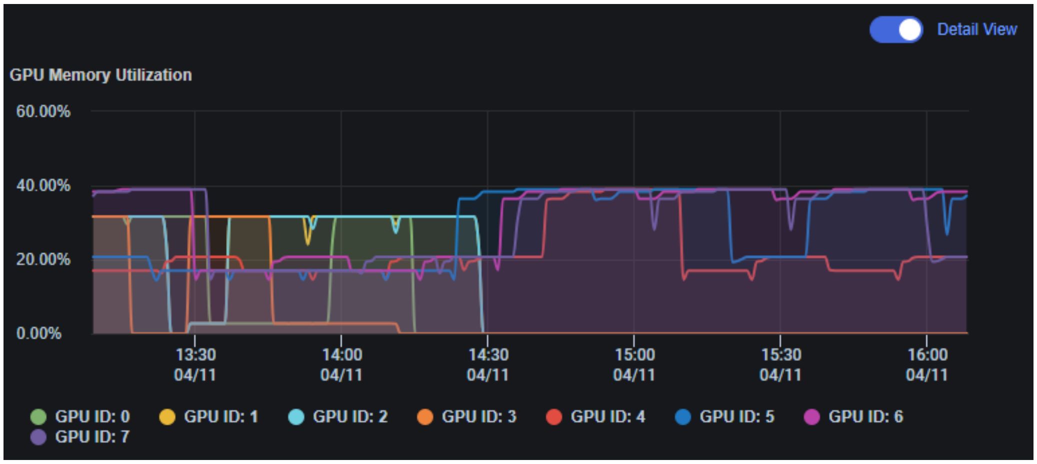
This chart displays the number of GPUs of selected GPU clusters or GPU cluster nodes in different ranges of GPU memory utilization. This provides a quick look on how GPU memory is utilized in the past.



To see individual GPU memory utilization history, click on the "Detailed View" button.

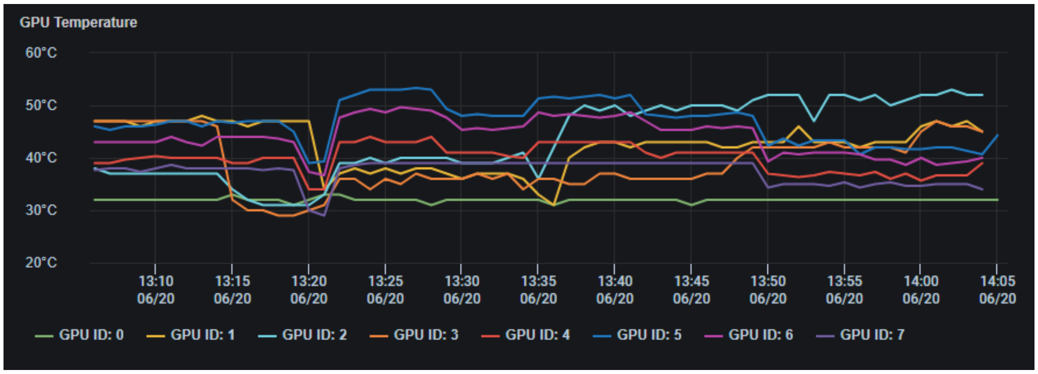
GPU Memory Utilization

This chart displays the memory utilization history of each GPU.



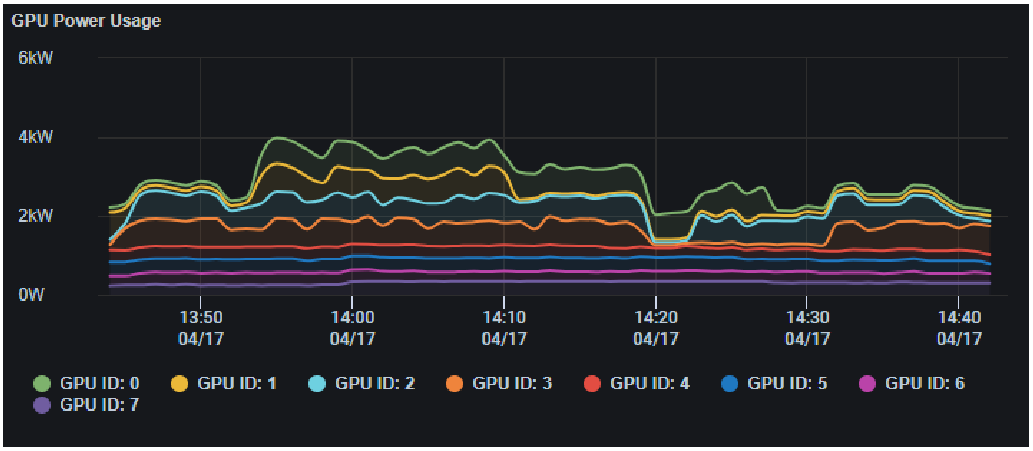
GPU Temperature

This chart displays the history of temperature of each individual GPU.



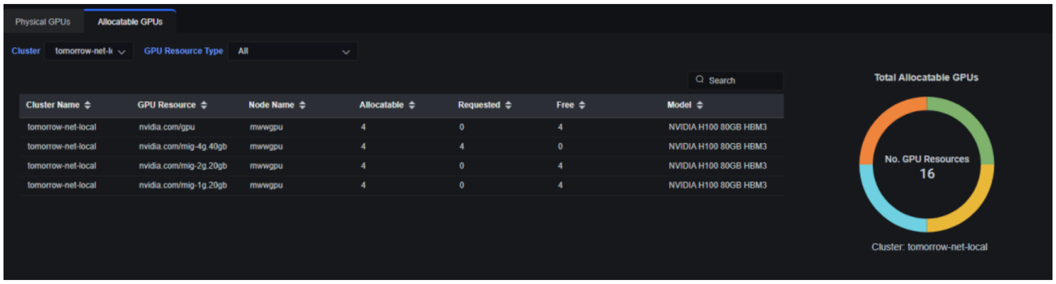
GPU Power Usage

This chart displays the power usage history of each individual GPU.



Allocatable GPUs

Some newer Nvidia GPU models support Multi-Instance GPU feature (MIG). This feature allows GPUs to be securely partitioned into up to seven separate GPU instances. Each MIG instance is an allocatable GPU resource in a Kubernetes cluster. The *Alloctable GPUs* tab shows the current list of GPU resources available in a cluster and the current usage of each GPU resource type.

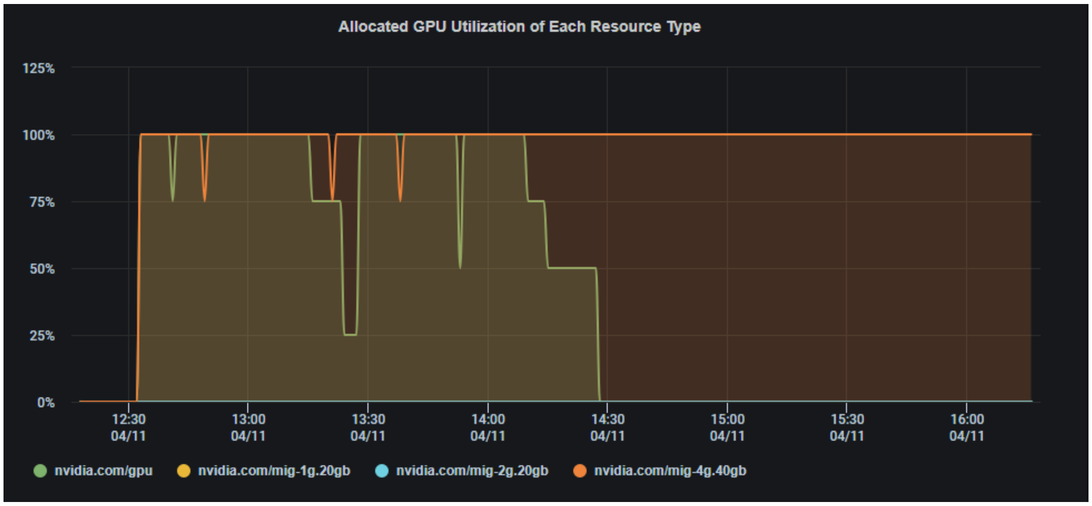


Cluster tomorrow-net-l

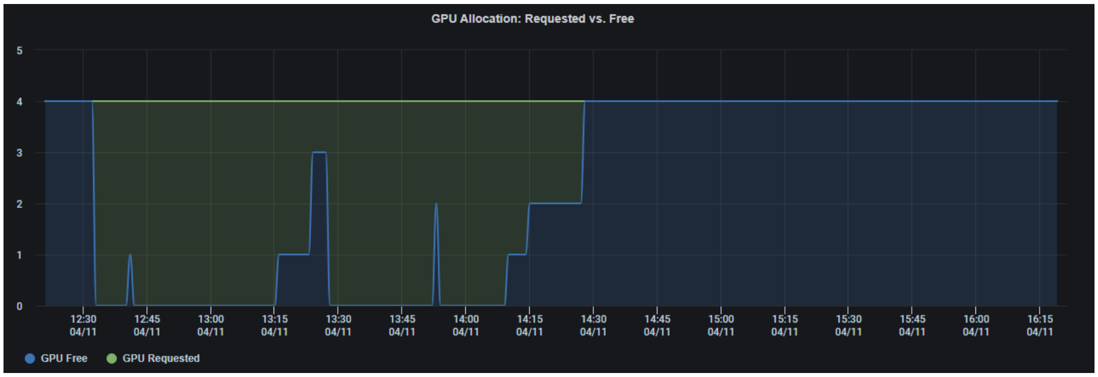
GPU Resource Type All

Cluster Name	GPU Resource	Node Name	Allocatable	Requested	Free	Model
tomorrow-net-local	nvidia.com/gpu	mwwgpu	4	0	4	NVIDIA H100 80GB HBM3
tomorrow-net-local	nvidia.com/mig-4g.40gb	mwwgpu	4	4	0	NVIDIA H100 80GB HBM3
tomorrow-net-local	nvidia.com/mig-2g.20gb	mwwgpu	4	0	4	NVIDIA H100 80GB HBM3
tomorrow-net-local	nvidia.com/mig-1g.20gb	mwwgpu	4	0	4	NVIDIA H100 80GB HBM3

The *GPU Allocation Per Resource Type* chart displays the history of allocation for each GPU resource type.

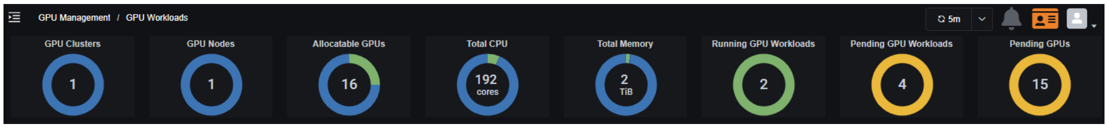


When a specific GPU resource type is selected, the chart *GPU Allocation: Requested vs. Free* displays the history of number of requested and free GPU resources.



GPU Management – GPU Workloads

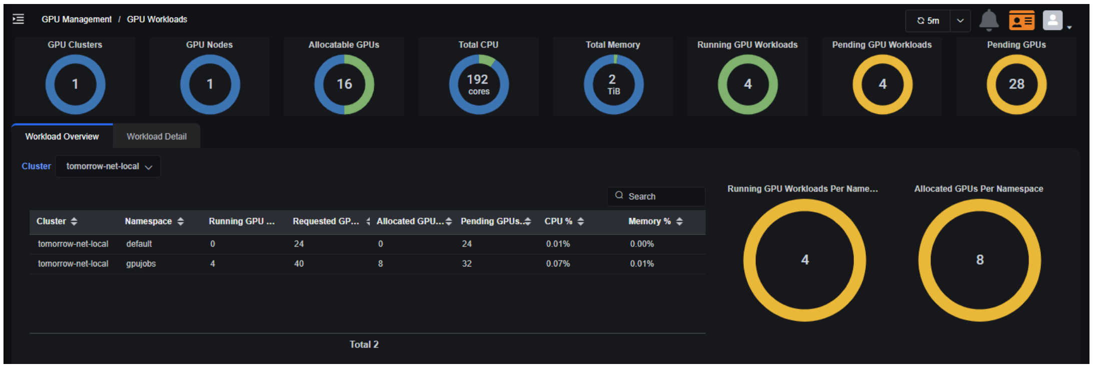
A *GPU Workload* is a deployment, job, or a cron job deployed in Kubernetes that utilizes GPUs. GPU workloads are automatically discovered after Federator.ai GPU Booster is deployed in a Kubernetes cluster. The top banner of the *GPU Workloads* page displays the information of number of clusters, number of cluster nodes, number of allocatable GPU resources, CPU/memory usage by all GPU workloads, total number of GPU workloads that are currently running or pending, and the total number of pending GPU resources.



GPU Clusters	GPU Nodes	Allocatable GPUs	Total CPU	Total Memory	Running GPU Workloads	Pending GPU Workloads	Pending GPUs
1	1	16	192 cores	2 TiB	2	4	15

Overview of GPU Workload

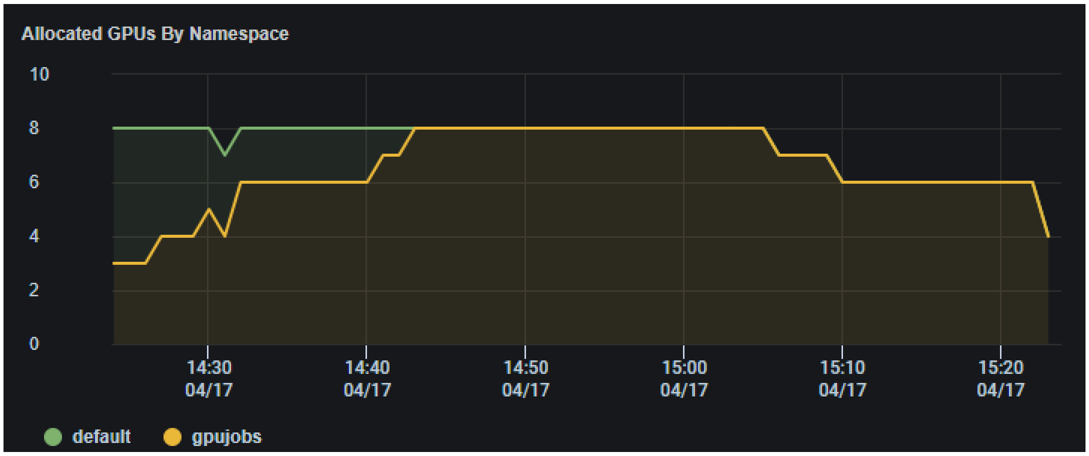
The *Workload Overview* tab displays a summary of GPU workloads in each GPU cluster: the namespace of the workloads, total number of Running GPU workloads, total number of requested GPUs, total number of allocated GPUs, total number of pending GPUs, percentage of CPU usage and memory usage for each namespace.



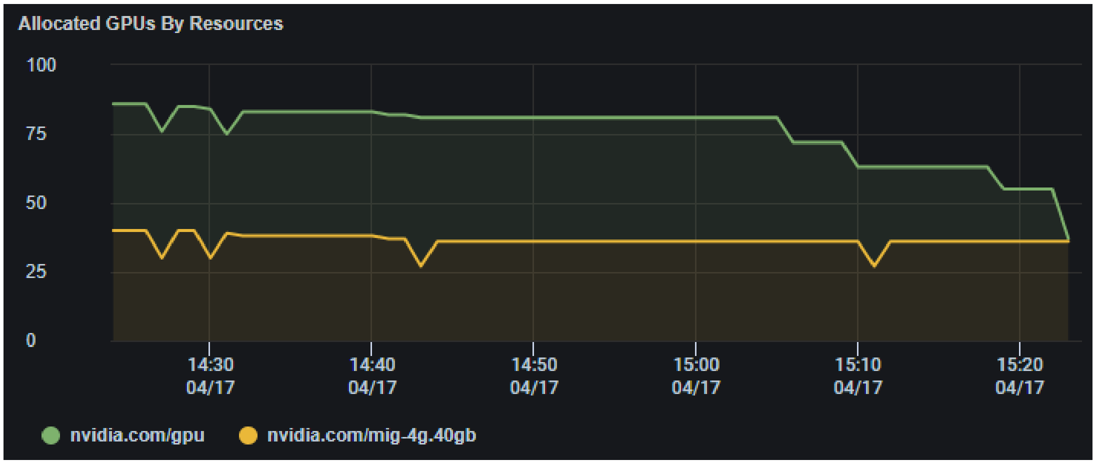
Cluster	Namespace	Running GPU ...	Requested GP...	Allocated GPU...	Pending GPUs..	CPU %	Memory %
tomorrow-net-local	default	0	24	0	24	0.01%	0.00%
tomorrow-net-local	gpujobs	4	40	8	32	0.07%	0.01%

Total 2

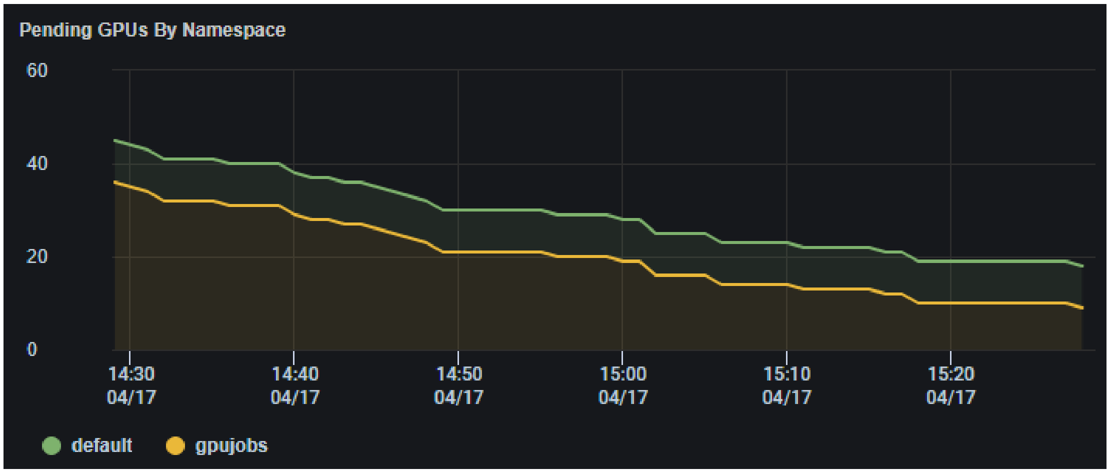
There are several charts that show the history of GPU workloads related metrics at the cluster level. The *Allocated GPUs By Namespace* chart shows the history of the total number of GPUs requested by all GPU workloads of each namespace.



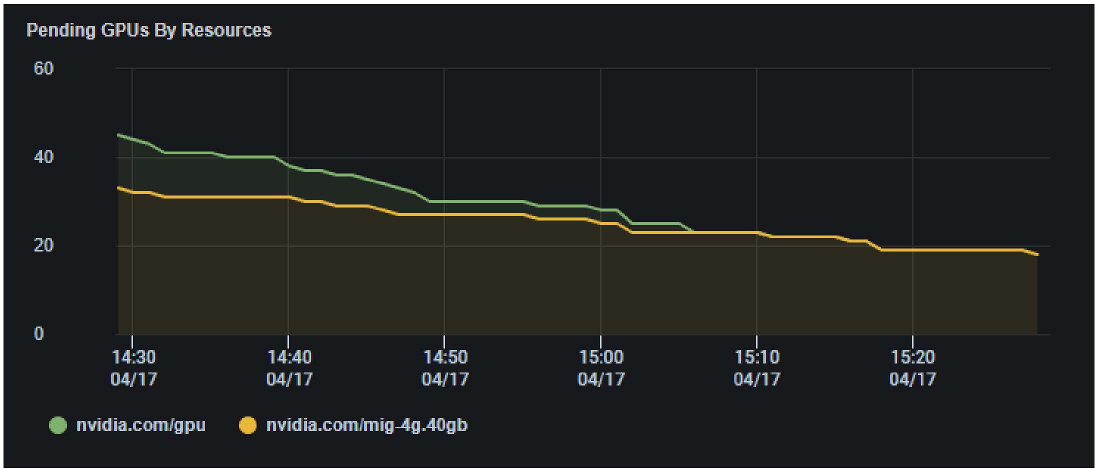
The *Allocated GPUs By Resources* chart displays the history of GPU allocations by all GPU workloads for each GPU resource.



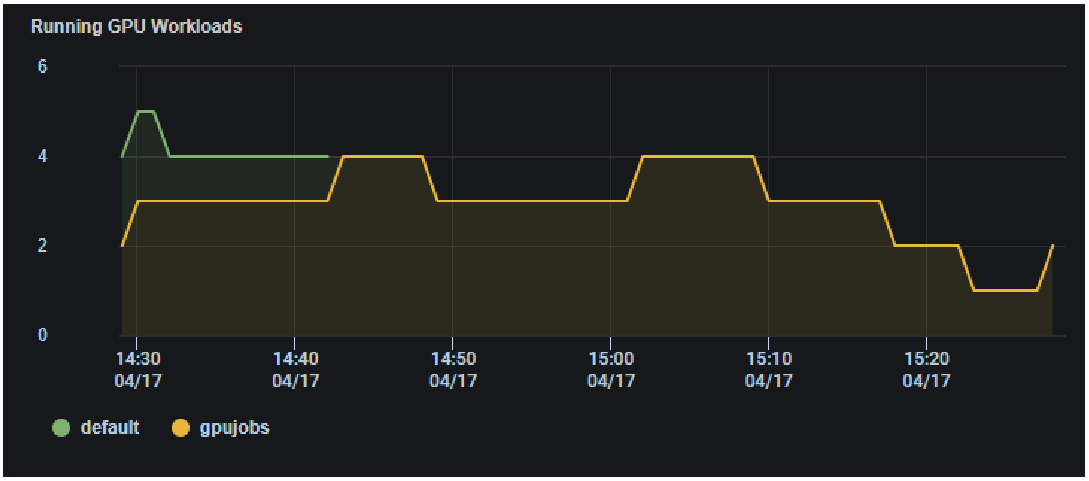
The *Pending GPUs By Namespace* chart displays the history of the number of pending GPUs by all GPU workloads of each namespace. This shows the administrators how many GPUs are not available to workloads from a specific namespace.



The *Pending GPUs by Resources* chart displays the history of number of pending GPUs by all GPU workloads of each GPU resource type. This gives the administrators a clear picture of how often a specific GPU resource was requested by but not assigned to GPU workloads.



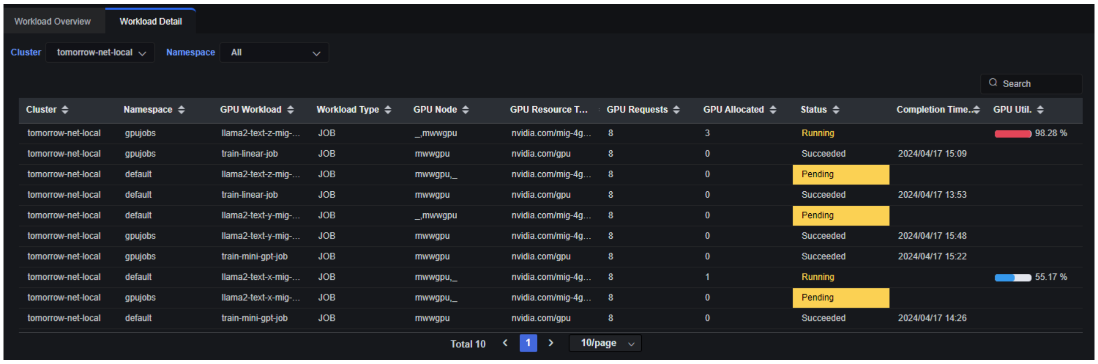
The *Running GPU Workloads* chart displays the history of running workloads from each namespace.



GPU Workload Detail

In the *Workload Detail* tab, users can view the list of all GPU workloads filtered by each GPU cluster. When selecting *All* in the *Namespace* dropdown menu, the table shows all GPU workloads in a cluster. A table containing detail information of each GPU workload is displayed. The information includes the name of a GPU workload, the cluster and namespace of a GPU workload, the GPU node where a GPU workload runs, and the GPU resource types a workload requested. Numbers of GPU requests and allocated of a GPU workload are also displayed.

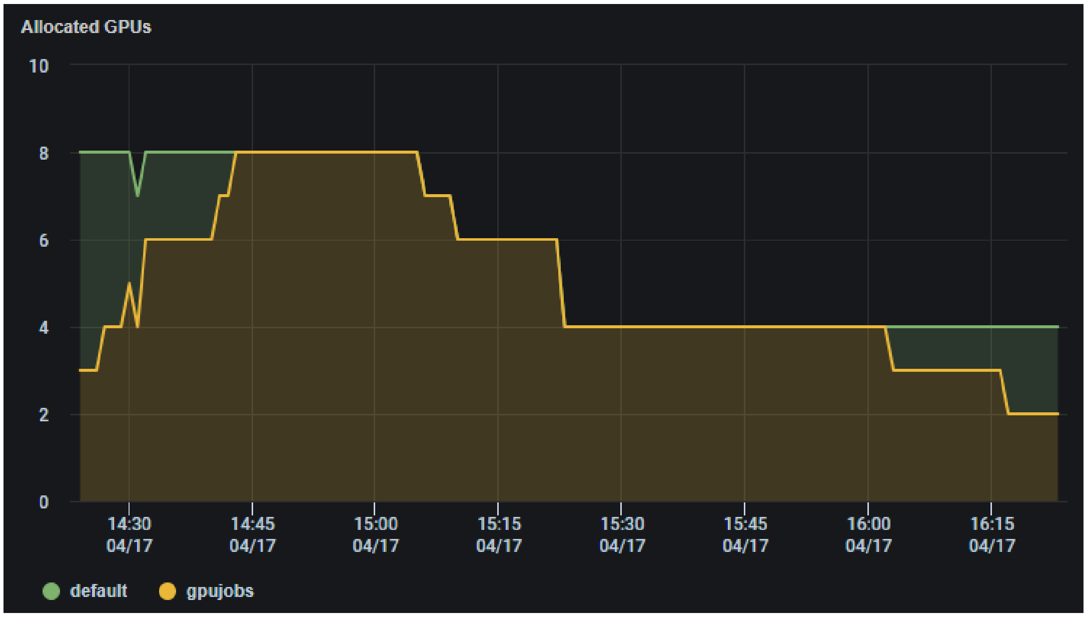
The status of a GPU workload is either *Running, Pending,* or *Succeeded.* When a GPU workload status is *Succeeded*, the completion timestamp is shown. A GPU workload is in the *Pending* state if GPU resources are not yet allocated for this workload. Otherwise, a GPU workload will be in the *Running* state.



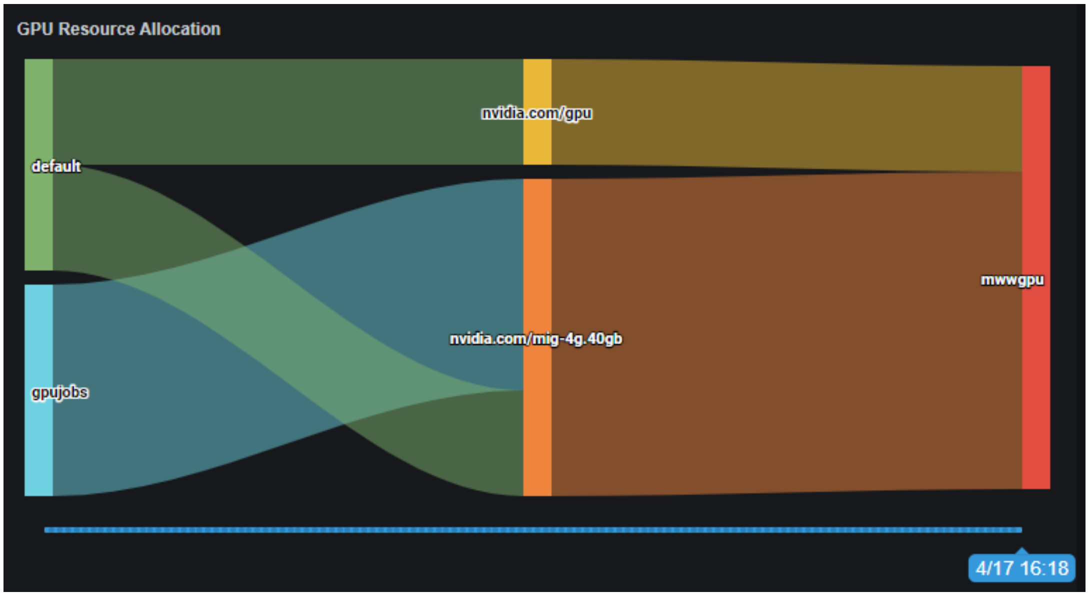
Cluster	Namespace	GPU Workload	Workload Type	GPU Node	GPU Resource T...	GPU Requests	GPU Allocated	Status	Completion Time...	GPU Util.
tomorrow-net-local	gpujobs	llama2-text-z-mig-...	JOB	_,mwwgpu	nvidia.com/mig-4g...	8	3	Running		98.28 %
tomorrow-net-local	gpujobs	train-linear-job	JOB	mwwgpu	nvidia.com/gpu	8	0	Succeeded	2024/04/17 15:09	
tomorrow-net-local	default	llama2-text-z-mig-...	JOB	mwwgpu,_	nvidia.com/mig-4g...	8	0	Pending		
tomorrow-net-local	default	train-linear-job	JOB	mwwgpu	nvidia.com/gpu	8	0	Succeeded	2024/04/17 13:53	
tomorrow-net-local	default	llama2-text-y-mig-...	JOB	_,mwwgpu	nvidia.com/mig-4g...	8	0	Pending		
tomorrow-net-local	gpujobs	llama2-text-y-mig-...	JOB	mwwgpu	nvidia.com/mig-4g...	8	0	Succeeded	2024/04/17 15:48	
tomorrow-net-local	gpujobs	train-mini-gpt-job	JOB	mwwgpu	nvidia.com/gpu	8	0	Succeeded	2024/04/17 15:22	
tomorrow-net-local	default	llama2-text-x-mig-...	JOB	mwwgpu,_	nvidia.com/mig-4g...	8	1	Running		55.17 %
tomorrow-net-local	gpujobs	llama2-text-x-mig-...	JOB	mwwgpu,_	nvidia.com/mig-4g...	8	0	Pending		
tomorrow-net-local	default	train-mini-gpt-job	JOB	mwwgpu	nvidia.com/gpu	8	0	Succeeded	2024/04/17 14:26	

When users select *All* from the *Namespace* drop-down menu, two charts are displayed:

- *Allocated GPUs:* This chart displays the history of total allocated GPUs by GPU workloads from each namespace.

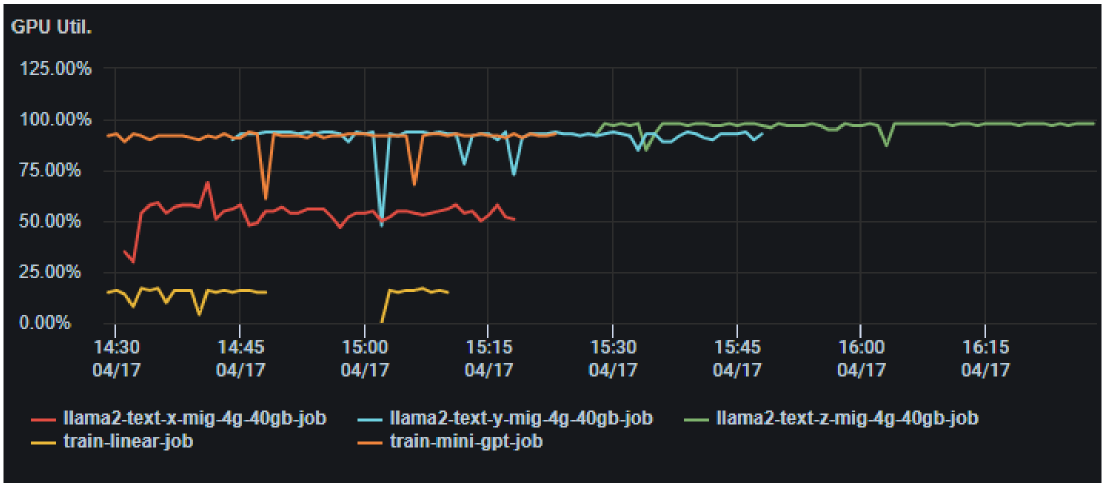


- *GPU Resources Allocation:* This chart displays the breakdown of GPU resources requested by GPU workloads from each namespace, and from which GPU nodes the GPU resources come. Sliding the time slide tab below the chart to see the GPU resource allocation at different time in the past.

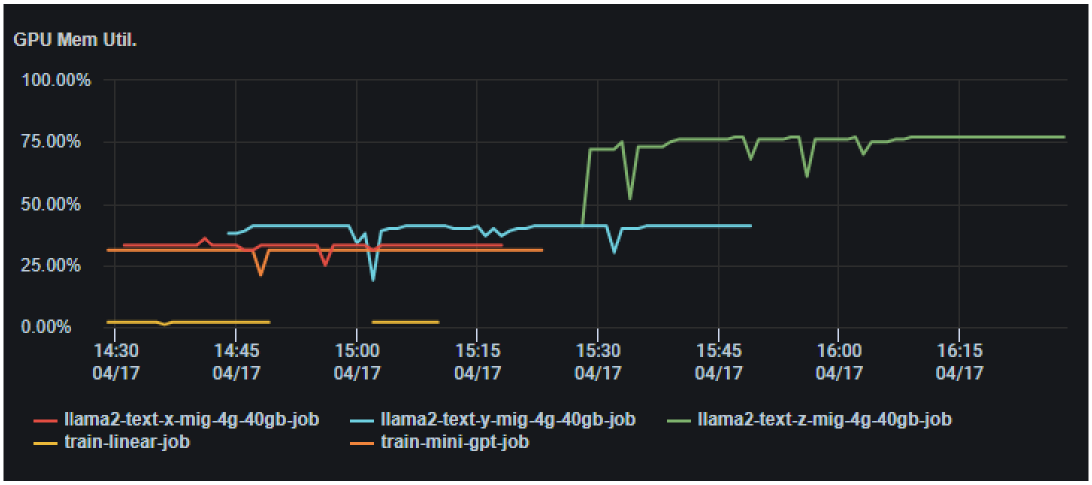


When users select a specific namespace from the *Namespace* drop-down menu, four charts are displayed:

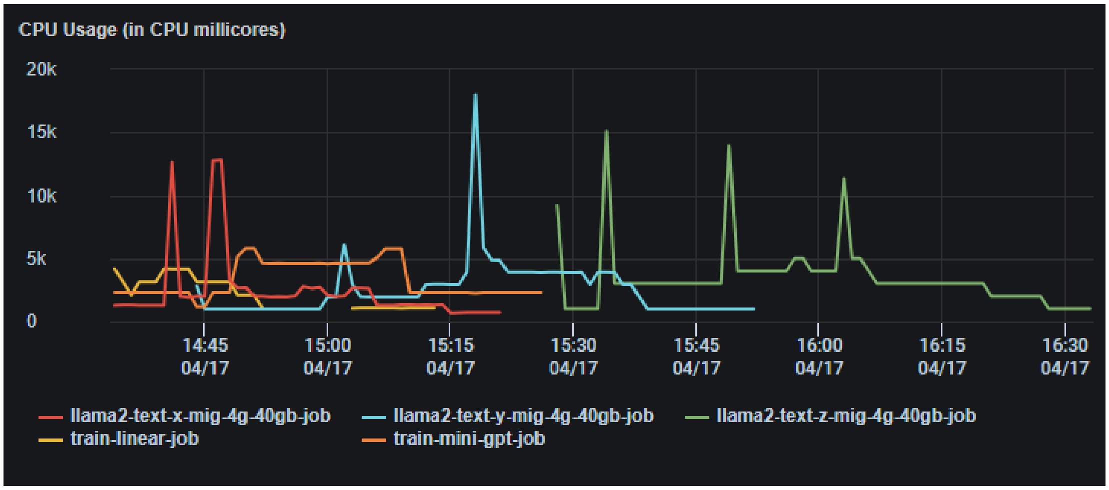
- *GPU Util:* This chart displays the history of GPU utilization of each GPU workload from the selected namespace.



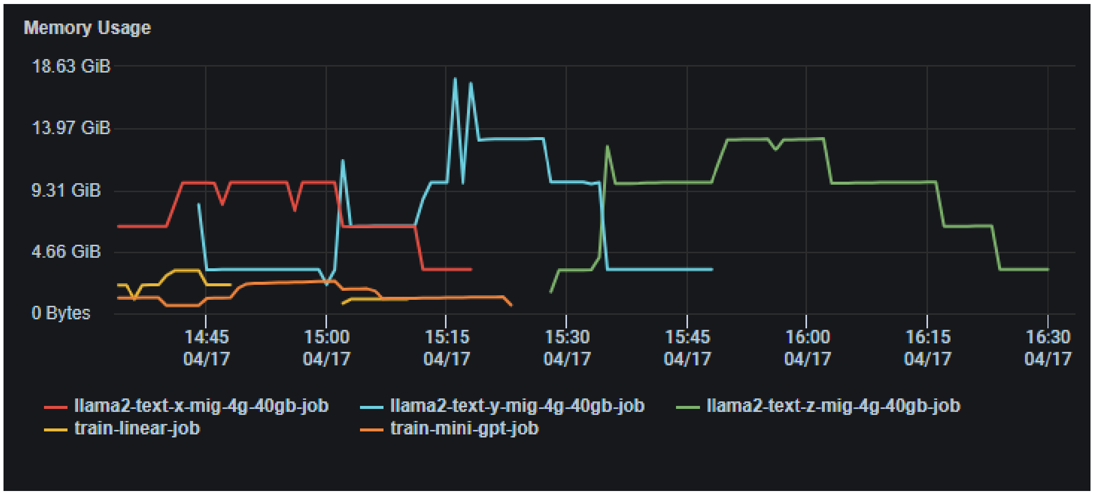
- *GPU Mem Util: GPU Util:* This chart displays the history of GPU memory utilization of each GPU workload from the selected namespace.



- *CPU Usage:* This chart displays the history of CPU usage in millicores of each GPU workload from the selected namespace.

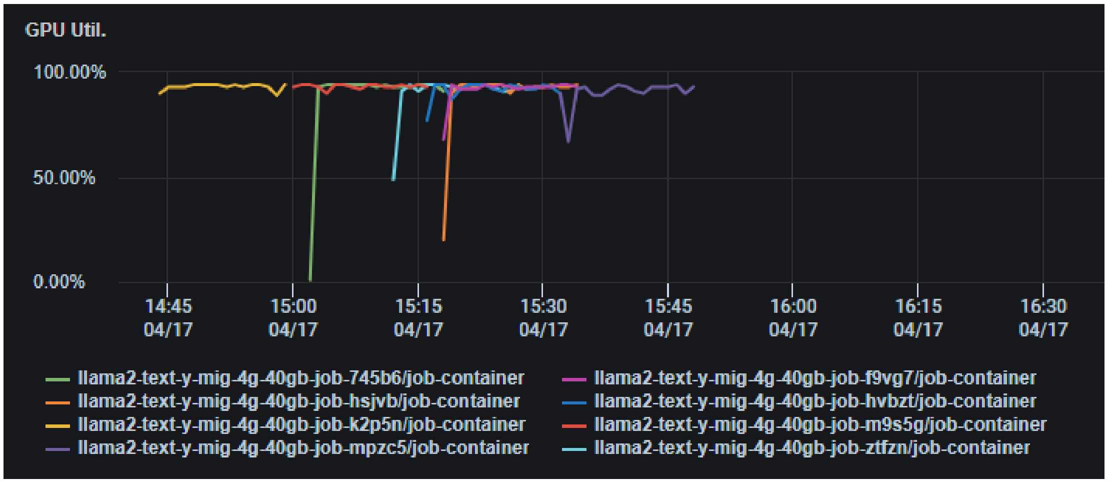


- *Memory Usage:* This chart displays the history of memory usage of each GPU workload from the selected namespace.

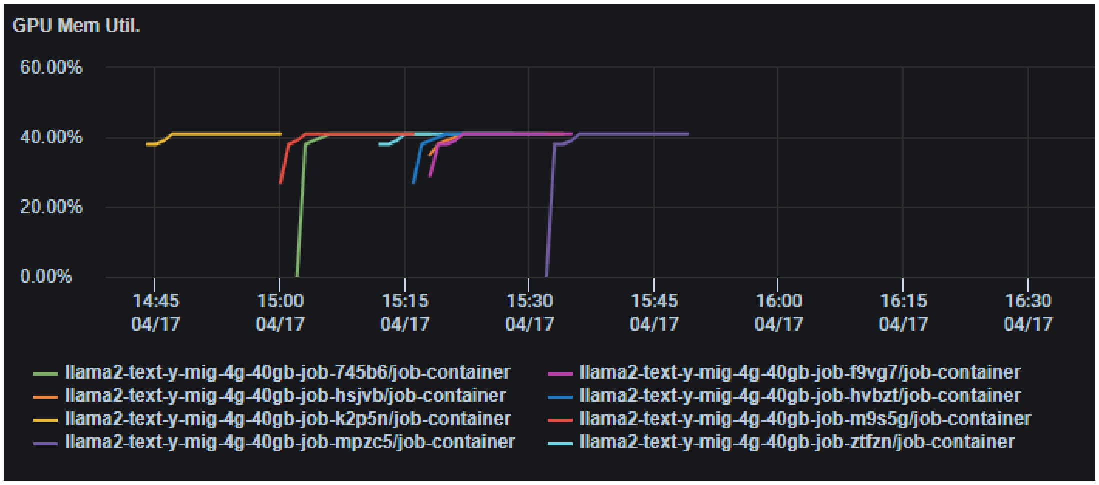


When users select a specific GPU workload from the *GPU Workload* drop-down menu, four charts are displayed:

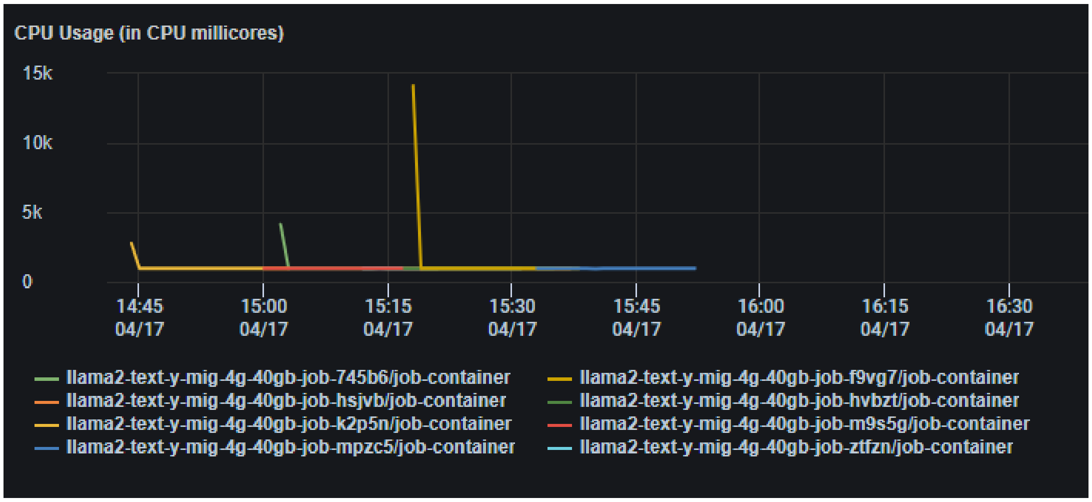
- *GPU Util:* This chart displays the history of GPU utilization of each pod/container from the selected GPU workload.



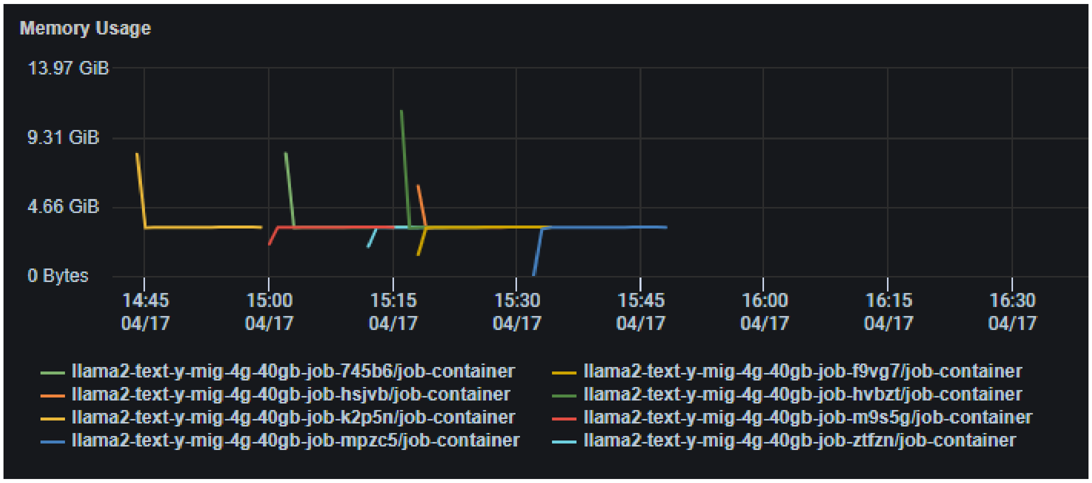
- *GPU Mem Util: GPU Util:* This chart displays the history of GPU memory utilization of each pod/container from the selected workload.



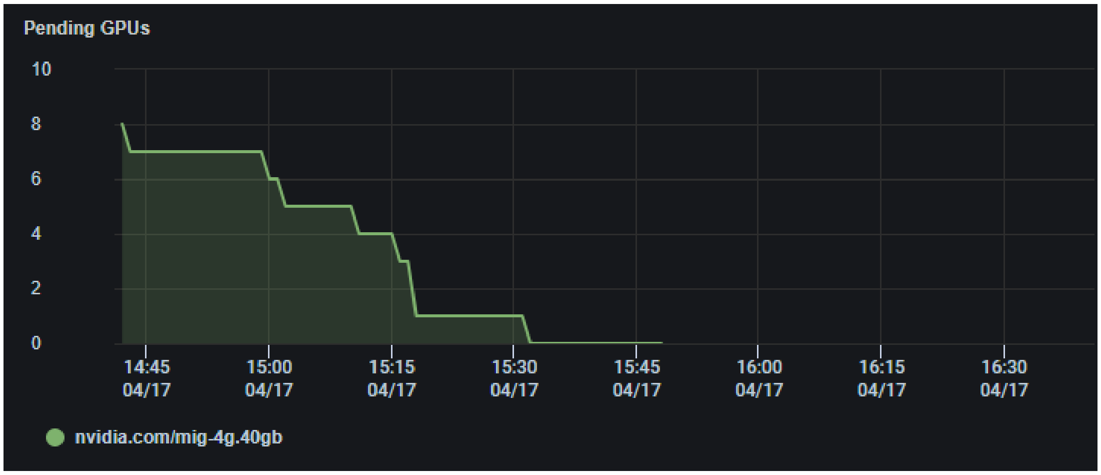
- *CPU Usage:* This chart displays the history of CPU usage in millicores of each pod/container from the selected GPU workload.



- *Memory Usage:* This chart displays the history of memory usage of each pod/container of the selected GPU workload.



- *Pending GPUs:* This chart displays the history of pending GPU resources by the selected GPU workload.

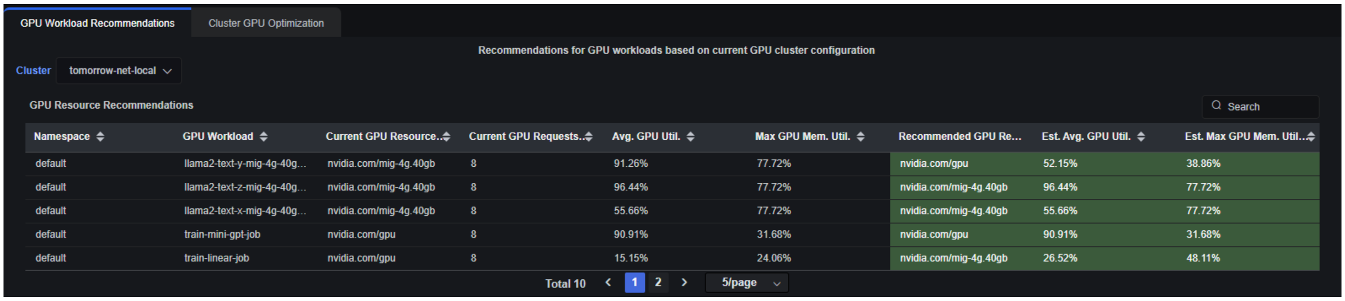


GPU Management – GPU Optimization

It is quite often that GPU workloads request GPU resources that are not fully utilized. The result is underutilized GPU resources and contention for same GPU resources from GPU workloads with different needs. For example, a GPU workload might request a whole GPU for the job when a MIG instance of 2g-20gb can easily do the job. Federator.ai GPU Booster analyzes the GPU utilization and GPU memory usage of all GPU workloads and recommends the appropriate MIG profile for each GPU workload to reduce the potential GPU resource contention and to increase the overall GPU utilization and throughput.

GPU Workload Recommendations

A combination of different GPU resources is usually configured in a GPU cluster. A typical configuration includes some complete GPUs and different sizes of MIG instances for various GPU workloads. In the *GPU Workload Recommendations* tab, Federator.ai GPU Booster displays the recommended GPU resources for each active GPU workload based on the existing GPU resource configuration in a cluster. The *GPU Resource Recommendations* table displays the current list of active GPU workloads, the current GPU resources requested by a GPU workload, the average GPU utilization of each run and the maximum GPU memory utilization of all runs. It also displays the recommended GPU resources for each active GPU workload and the estimated GPU utilization and memory utilization using the recommended GPU resource.



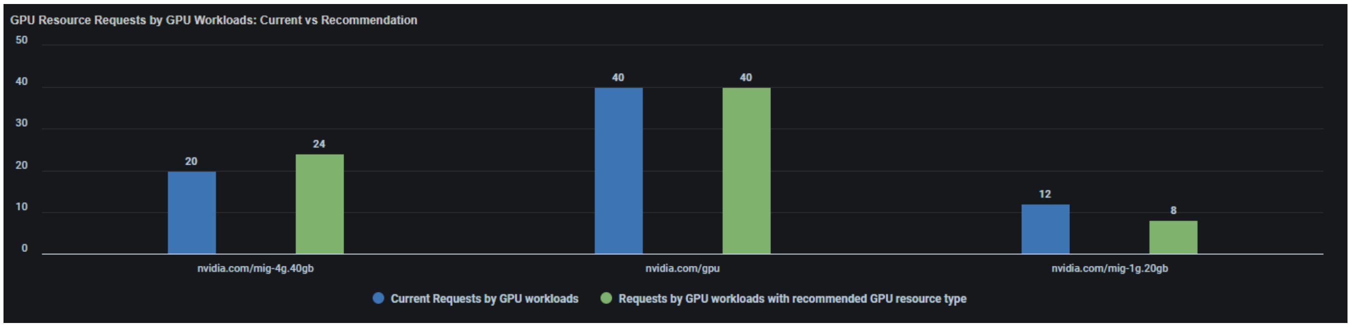
Recommendations for GPU workloads based on current GPU cluster configuration

Cluster: tomorrow-net-local

GPU Resource Recommendations

Namespace	GPU Workload	Current GPU Resource..	Current GPU Requests..	Avg. GPU Util.	Max GPU Mem. Util.	Recommended GPU Re...	Est. Avg. GPU Util.	Est. Max GPU Mem. Util...
default	llama2-text-y-mig-4g-40g...	nvidia.com/mig-4g.40gb	8	91.26%	77.72%	nvidia.com/gpu	52.15%	38.86%
default	llama2-text-z-mig-4g-40g...	nvidia.com/mig-4g.40gb	8	96.44%	77.72%	nvidia.com/mig-4g.40gb	96.44%	77.72%
default	llama2-text-x-mig-4g-40g...	nvidia.com/mig-4g.40gb	8	55.66%	77.72%	nvidia.com/mig-4g.40gb	55.66%	77.72%
default	train-mini-gpt-job	nvidia.com/gpu	8	90.91%	31.68%	nvidia.com/gpu	90.91%	31.68%
default	train-linear-job	nvidia.com/gpu	8	15.15%	24.06%	nvidia.com/mig-4g.40gb	26.52%	48.11%

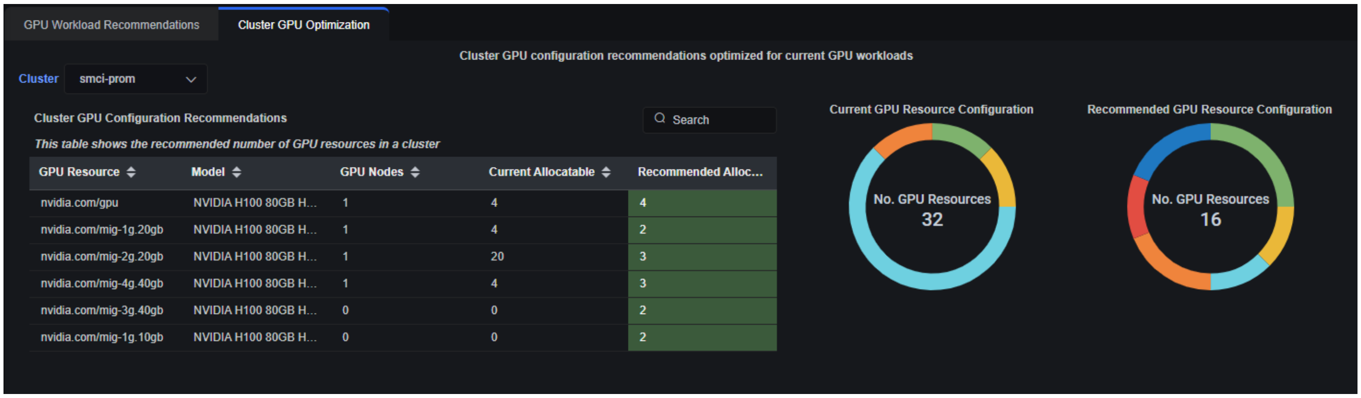
A comparison of current number of requests for different GPU resource types versus the number of requests after the recommendation is also displayed.



Cluster GPU Optimization

Recommendations of appropriate GPU resources to different GPU workloads based on existing GPU resource configuration might not give you the optimized GPU utilization and throughput. Utilizing unique and patented algorithm, Federator.ai GPU Booster recommends a GPU resource configuration that is better suited for the active GPU workloads.

The *Cluster GPU Configuration Recommendations* table displays the GPU resources, the current quantities, and the recommended quantities of a GPU cluster.



Cluster GPU configuration recommendations optimized for current GPU workloads

Cluster: smci-prom

Cluster GPU Configuration Recommendations

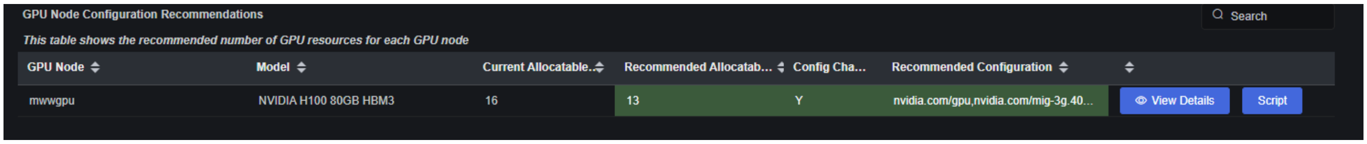
This table shows the recommended number of GPU resources in a cluster

GPU Resource	Model	GPU Nodes	Current Allocatable	Recommended Alloc...
nvidia.com/gpu	NVIDIA H100 80GB H...	1	4	4
nvidia.com/mig-1g.20gb	NVIDIA H100 80GB H...	1	4	2
nvidia.com/mig-2g.20gb	NVIDIA H100 80GB H...	1	20	3
nvidia.com/mig-4g.40gb	NVIDIA H100 80GB H...	1	4	3
nvidia.com/mig-3g.40gb	NVIDIA H100 80GB H...	0	0	2
nvidia.com/mig-1g.10gb	NVIDIA H100 80GB H...	0	0	2

Current GPU Resource Configuration: No. GPU Resources 32

Recommended GPU Resource Configuration: No. GPU Resources 16

Federator.ai GPU Booster also provides detailed information on how each GPU node should be configured based on the GPU configuration recommendations.

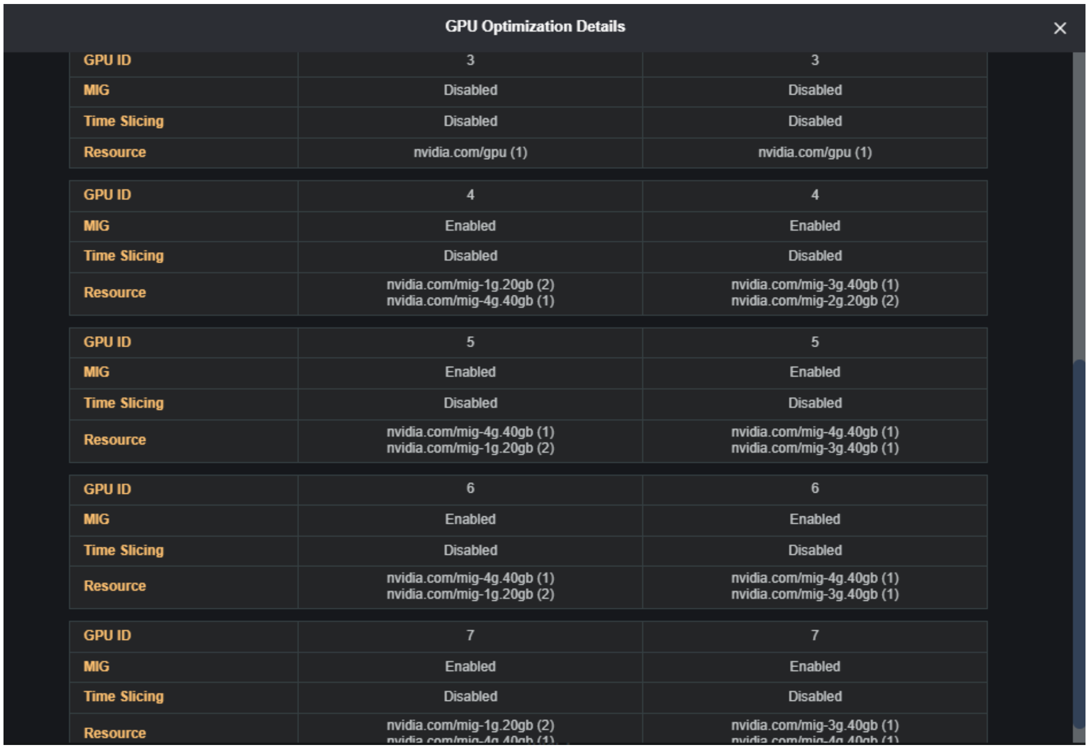


GPU Node Configuration Recommendations

This table shows the recommended number of GPU resources for each GPU node

GPU Node	Model	Current Allocatable..	Recommended Allocatab...	Config Cha...	Recommended Configuration	
mwwgpu	NVIDIA H100 80GB HBM3	16	13	Y	nvidia.com/gpu,nvidia.com/mig-3g.40...	View Details Script

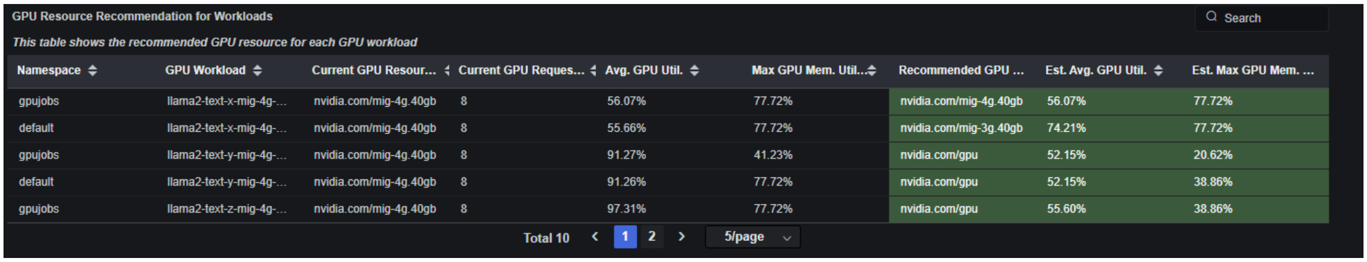
If a GPU node is marked with "Y" in the *Config Changes* column, you can click on the "*View Details*" button to see the configuration change for each GPU in the GPU node. Click on the "*Script"* button to see the sample shell script to configure the GPU node based on the recommendations.



GPU Optimization Details

GPU ID	3	3
MIG	Disabled	Disabled
Time Slicing	Disabled	Disabled
Resource	nvidia.com/gpu (1)	nvidia.com/gpu (1)
GPU ID	4	4
MIG	Enabled	Enabled
Time Slicing	Disabled	Disabled
Resource	nvidia.com/mig-1g.20gb (2) nvidia.com/mig-4g.40gb (1)	nvidia.com/mig-3g.40gb (1) nvidia.com/mig-2g.20gb (2)
GPU ID	5	5
MIG	Enabled	Enabled
Time Slicing	Disabled	Disabled
Resource	nvidia.com/mig-4g.40gb (1) nvidia.com/mig-1g.20gb (2)	nvidia.com/mig-4g.40gb (1) nvidia.com/mig-3g.40gb (1)
GPU ID	6	6
MIG	Enabled	Enabled
Time Slicing	Disabled	Disabled
Resource	nvidia.com/mig-4g.40gb (1) nvidia.com/mig-1g.20gb (2)	nvidia.com/mig-4g.40gb (1) nvidia.com/mig-3g.40gb (1)
GPU ID	7	7
MIG	Enabled	Enabled
Time Slicing	Disabled	Disabled
Resource	nvidia.com/mig-1g.20gb (2) nvidia.com/mig-4g.40gb (1)	nvidia.com/mig-3g.40gb (1) nvidia.com/mig-4g.40gb (1)

Finally, the Federator.ai GPU Booster displays the GPU resources recommended for each GPU workload based on the new GPU resource configuration.



GPU Resource Recommendation for Workloads

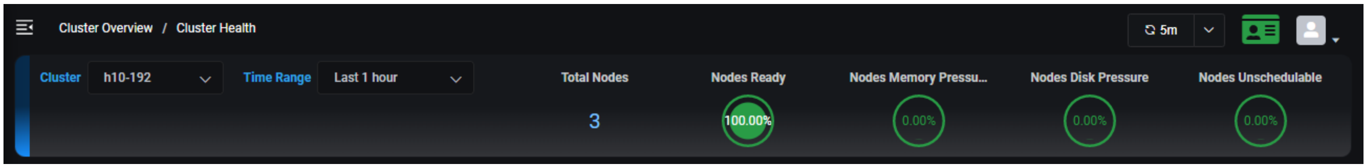
This table shows the recommended GPU resource for each GPU workload

Namespace	GPU Workload	Current GPU Resour...	Current GPU Reques...	Avg. GPU Util.	Max GPU Mem. Util...	Recommended GPU ...	Est. Avg. GPU Util.	Est. Max GPU Mem. ...
gpujobs	llama2-text-x-mig-4g-...	nvidia.com/mig-4g.40gb	8	56.07%	77.72%	nvidia.com/mig-4g.40gb	56.07%	77.72%
default	llama2-text-x-mig-4g-...	nvidia.com/mig-4g.40gb	8	55.66%	77.72%	nvidia.com/mig-3g.40gb	74.21%	77.72%
gpujobs	llama2-text-y-mig-4g-...	nvidia.com/mig-4g.40gb	8	91.27%	41.23%	nvidia.com/gpu	52.15%	20.62%
default	llama2-text-y-mig-4g-...	nvidia.com/mig-4g.40gb	8	91.26%	77.72%	nvidia.com/gpu	52.15%	38.86%
gpujobs	llama2-text-z-mig-4g-...	nvidia.com/mig-4g.40gb	8	97.31%	77.72%	nvidia.com/gpu	55.60%	38.86%

Cluster Overview - Cluster Health

The *Cluster Health* page displays actual usage observations about the nodes in a cluster. Usage for the last 1, 2, 4, 6, or 12 hours can be displayed and can further be filtered by selecting a range of dates.

For the selected cluster, you will see the total number of nodes and the percentage that are ready. For Kubernetes, you will also see the percentage of nodes that are under memory or disk pressure, as well as the percentage of nodes that are not schedulable. Memory pressure and disk pressure are defined by Kubernetes. Refer to https://kubernetes.io/docs/tasks/administer-cluster/out-of-resource/#node-conditions for more information.



Cluster Overview / Cluster Health

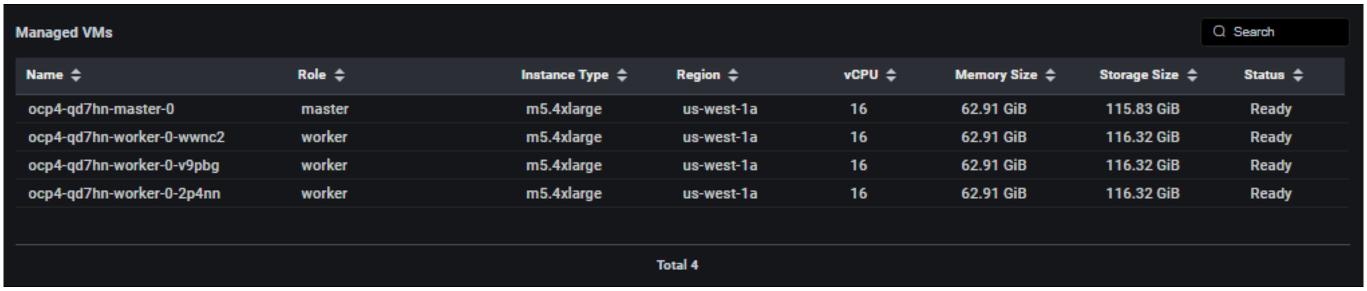
5m

Cluster: h10-192

Time Range: Last 1 hour

Total Nodes	Nodes Ready	Nodes Memory Pressu...	Nodes Disk Pressure	Nodes Unschedulable
3	100.00%	0.00%	0.00%	0.00%

The table below displays the current configuration for each cluster node, including Kubernetes role (master, worker), instance types being used at your cloud provider, cloud provider region, number of CPUs, memory size, storage size, and node status.



Managed VMs

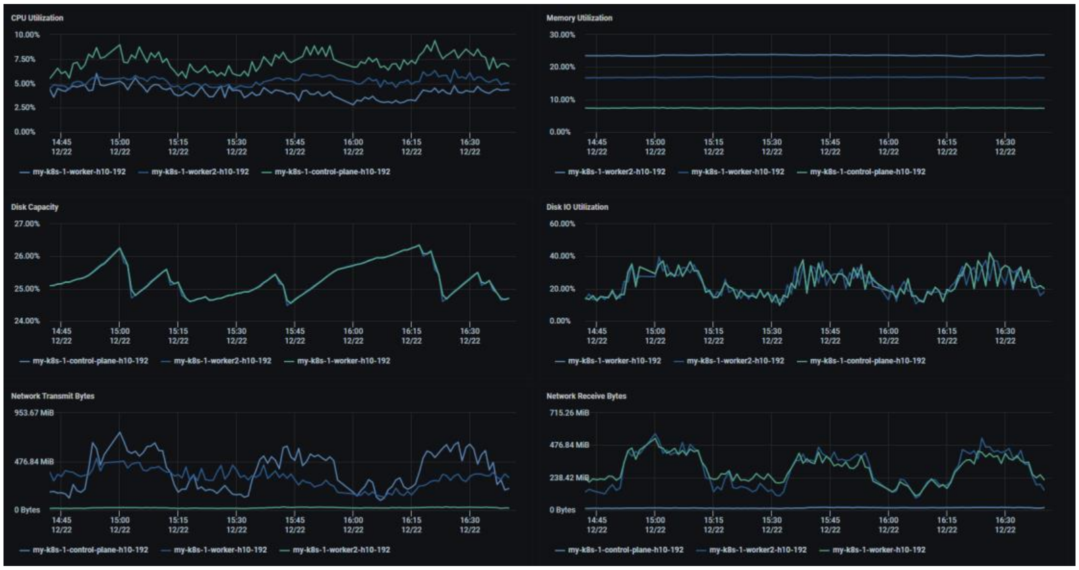
Search

Name	Role	Instance Type	Region	vCPU	Memory Size	Storage Size	Status
ocp4-qd7hn-master-0	master	m5.4xlarge	us-west-1a	16	62.91 GiB	115.83 GiB	Ready
ocp4-qd7hn-worker-0-wwnc2	worker	m5.4xlarge	us-west-1a	16	62.91 GiB	116.32 GiB	Ready
ocp4-qd7hn-worker-0-v9pbg	worker	m5.4xlarge	us-west-1a	16	62.91 GiB	116.32 GiB	Ready
ocp4-qd7hn-worker-0-2p4nn	worker	m5.4xlarge	us-west-1a	16	62.91 GiB	116.32 GiB	Ready

Total 4

Kubernetes cluster

CPU utilization, memory utilization, disk capacity, and disk IO utilization, network transmit and receive bytes charts are displayed for each Kubernetes node.



Related topics:

Search/Sort Information in Tables

Show/Hide Information in Charts

Zoom In/Out of Charts

Terminology

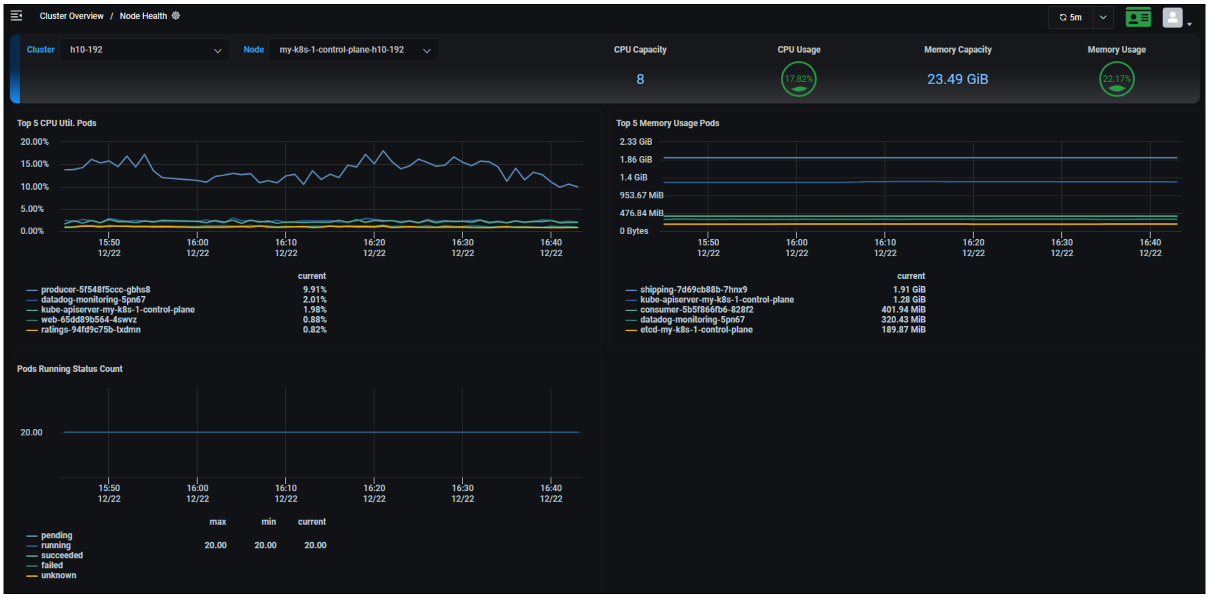
Cluster Overview - Node Health

The *Node Health* page displays actual usage observations about each node in a Kubernetes cluster. Select which cluster and node to display.

For the selected node, you will see the total CPU capacity and usage as well as memory capacity and usage.

The *Top 5* charts below display the CPU utilization of the top five pods on each node and the memory usage of the top five pods.

The *Pods Running Status Count* chart displays the number of pods running, the minimum and maximum number of pods that can run, along with the status of each pod.



Cluster: h10-192

Node: my-k8s-1-control-plane-h10-192

CPU Capacity	CPU Usage	Memory Capacity	Memory Usage
8	17.82%	23.49 GiB	22.17%

Top 5 CPU Util. Pods

	current
producer-5f548f5ccc-gbhs8	9.91%
datadog-monitoring-5pn67	2.01%
kube-apiserver-my-k8s-1-control-plane	1.98%
web-65dd89b564-4swvz	0.88%
ratings-94fd9c75b-txdmn	0.82%

Top 5 Memory Usage Pods

	current
shipping-7d69cb88b-7hnx9	1.91 GiB
kube-apiserver-my-k8s-1-control-plane	1.28 GiB
consumer-5b5f866fb6-828f2	401.94 MiB
datadog-monitoring-5pn67	320.43 MiB
etcd-my-k8s-1-control-plane	189.87 MiB

Pods Running Status Count

	max	min	current
pending			
running	20.00	20.00	20.00
succeeded			
failed			
unknown			

Related topics:

Search/Sort Information in Tables

Show/ Hide Information in Charts

Zoom In/Out of Charts

Terminology

Predictions and Planning – Kubernetes Resources

The *Kubernetes Resources* page displays actual CPU and memory usage observations, predicted usage, and recommendations for resources in a Kubernetes cluster.

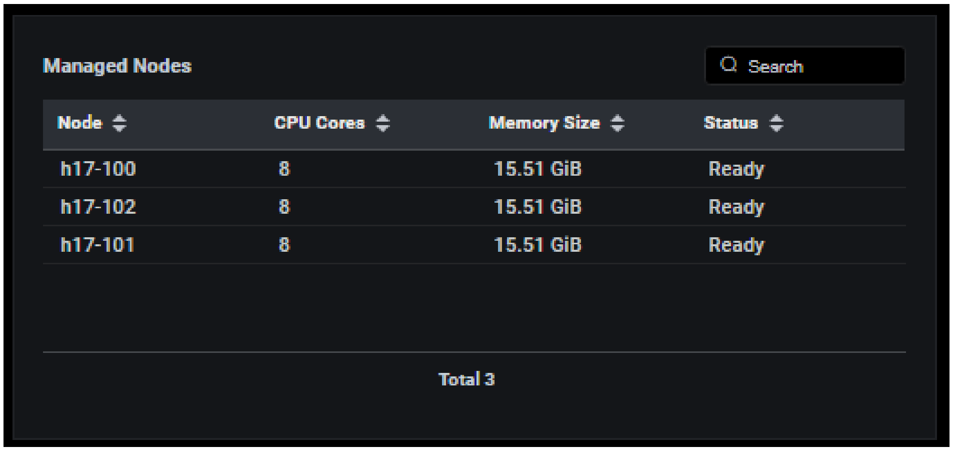
For Kubernetes, Federator.ai GPU Booster monitors resource usage for monitored clusters, nodes, namespaces, user-defined applications, and controllers and provides workload prediction, recommendations, and utilization analysis at each level.

With the analysis and recommendations, you can decide if a resource is over-provisioned (wasting resources), or if it is under-provisioned and will not sustain an increased workload.

Use the tabs at the top of the page to select the level of information you want to display. When you select a Kubernetes namespace, the namespace status will be displayed.

Managed Nodes Table

This table shows the list of nodes in the selected cluster with the number of CPU cores, the size of memory, and the status for each node.



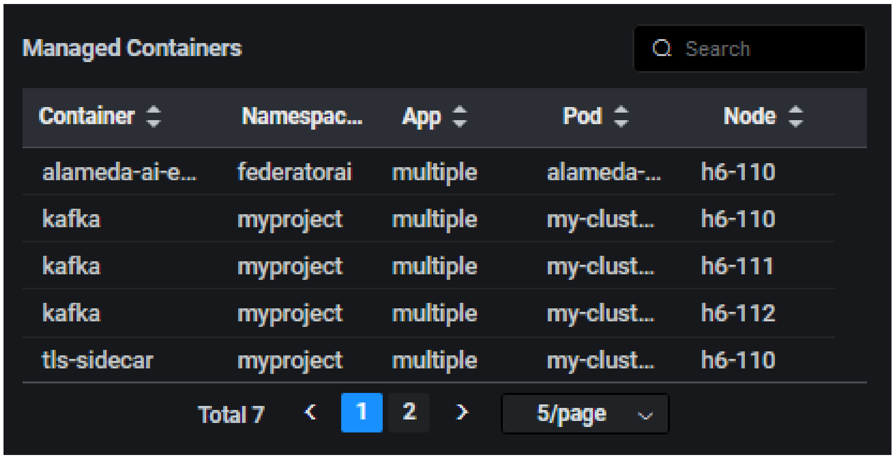
Managed Nodes

Node	CPU Cores	Memory Size	Status
h17-100	8	15.51 GiB	Ready
h17-102	8	15.51 GiB	Ready
h17-101	8	15.51 GiB	Ready

Total 3

Managed Containers Table

Based on the selected scope (cluster, node, namespace, application, or controller), this table lists the containers for the scope. Each container is listed along with its namespace, application, Kubernetes pod name, and the node where this container runs.



Managed Containers

Search

Container	Namespac...	App	Pod	Node
alameda-ai-e...	federatorai	multiple	alameda-...	h6-110
kafka	myproject	multiple	my-clust...	h6-110
kafka	myproject	multiple	my-clust...	h6-111
kafka	myproject	multiple	my-clust...	h6-112
tls-sidecar	myproject	multiple	my-clust...	h6-110

Total 7 < 1 2 > 5/page

Workload Prediction Table and Workload Observation and Prediction Charts

The *Workload Prediction* table displays daily, weekly, and monthly predictive CPU and memory data.

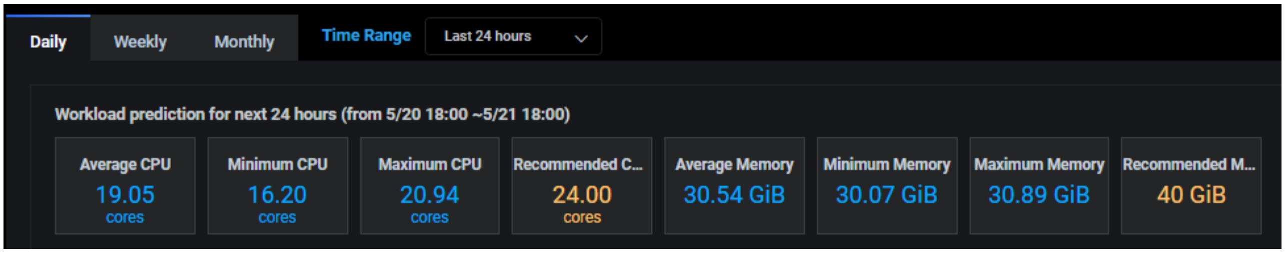
The *Workload Observation and Prediction* charts display observed actual usage for the selected time as well as predictive CPU and memory data:

- Daily – Predicts CPU and memory usage every hour for the next 24 hours.
- Weekly – Predicts CPU and memory usage every 6 hours for the next 7 days.
- Monthly – Predicts CPU and memory usage every day for the next 30 days.

Use the *Time Range* field to set a custom time period for observed CPU and memory usage.

Workload Prediction Table

This table displays average/minimum/maximum CPU and memory usage and recommendations for the upcoming time selected - 24 hours (daily), 7 days (weekly), 30 days (monthly).



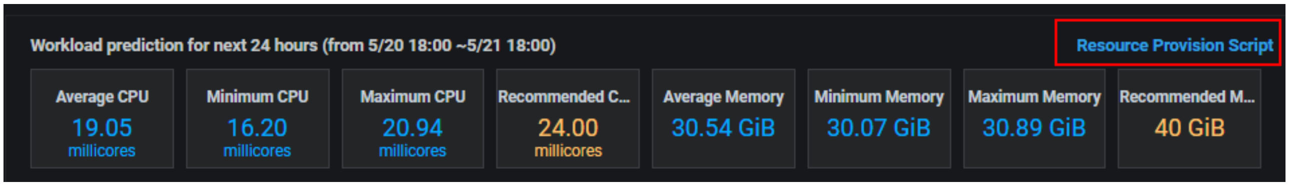
Time Range: Last 24 hours

Workload prediction for next 24 hours (from 5/20 18:00 ~5/21 18:00)

Average CPU	Minimum CPU	Maximum CPU	Recommended C...	Average Memory	Minimum Memory	Maximum Memory	Recommended M...
19.05 cores	16.20 cores	20.94 cores	24.00 cores	30.54 GiB	30.07 GiB	30.89 GiB	40 GiB

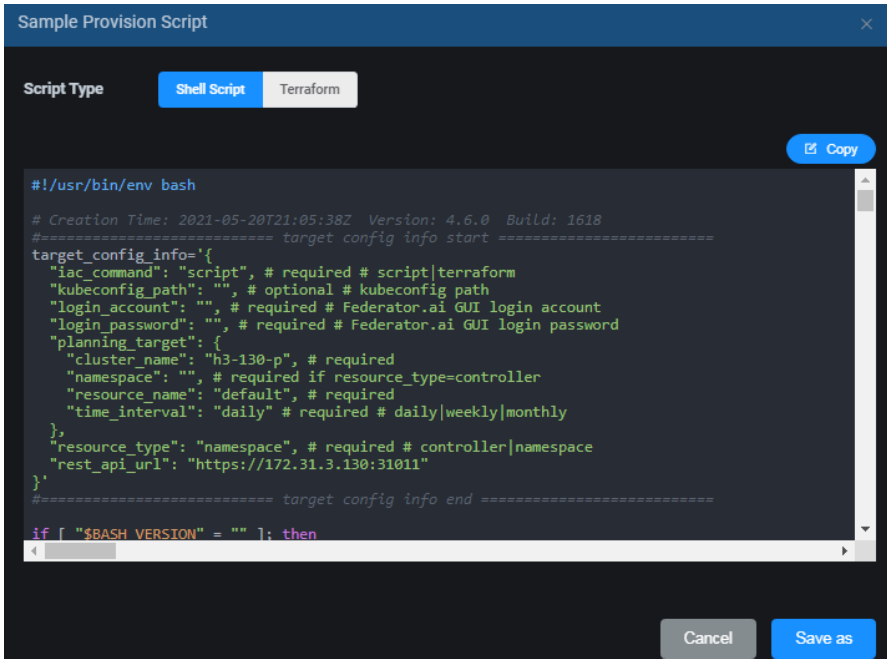
If you are displaying information for a Kubernetes namespace and the status is anything but *Monitoring*, this section will provide more information. For example, you will see the message, "Workload prediction is not configured for this namespace" or "Not enough information for predictions" for newly added, monitored namespaces.

For Kubernetes namespaces and controllers, Federator.ai GPU Booster provides a resource integration script that can be used to automatically apply the recommended CPU/memory for the namespace or controller. If an auto provisioning profile is assigned to a namespace or a controller, the resource integration script uses recommendations set by the auto provisioning profile. Otherwise, the resource integration script uses system recommendations based on the time frame you are viewing (daily/weekly/monthly). For remote Kubernetes clusters, you can copy a resource integration script to the remote cluster in order to run auto provisioning.



Workload prediction for next 24 hours (from 5/20 18:00 ~5/21 18:00)

Average CPU	Minimum CPU	Maximum CPU	Recommended C...	Average Memory	Minimum Memory	Maximum Memory	Recommended M...
19.05 millicores	16.20 millicores	20.94 millicores	24.00 millicores	30.54 GiB	30.07 GiB	30.89 GiB	40 GiB



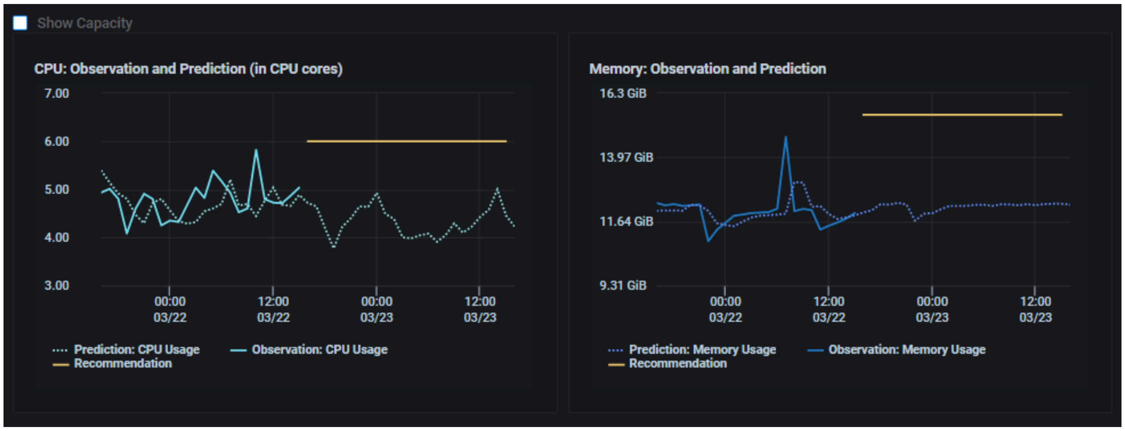
Sample Provision Script

Script Type: Shell Script

```
#!/usr/bin/env bash

# Creation Time: 2021-05-20T21:05:38Z  Version: 4.6.0  Build: 1618
#=========================== target config info start ==========================
target_config_info='{
  "iac_command": "script", # required # script|terraform
  "kubeconfig_path": "", # optional # kubeconfig path
  "login_account": "", # required # Federator.ai GUI login account
  "login_password": "", # required # Federator.ai GUI login password
  "planning_target": {
    "cluster_name": "h3-130-p", # required
    "namespace": "", # required if resource_type=controller
    "resource_name": "default", # required
    "time_interval": "daily" # required # daily|weekly|monthly
  },
  "resource_type": "namespace", # required # controller|namespace
  "rest_api_url": "https://172.31.3.130:31011"
}'
#=========================== target config info end ===========================

if [ "$BASH_VERSION" = "" ]; then
```

You can copy the script and run it in the Kubernetes cluster where the controller or the namespace is located. The script queries Federator.ai GPU Booster for the most recent recommendations and applies them to the controller or the namespace. Refer to Auto Provisioning Scripts for more information.

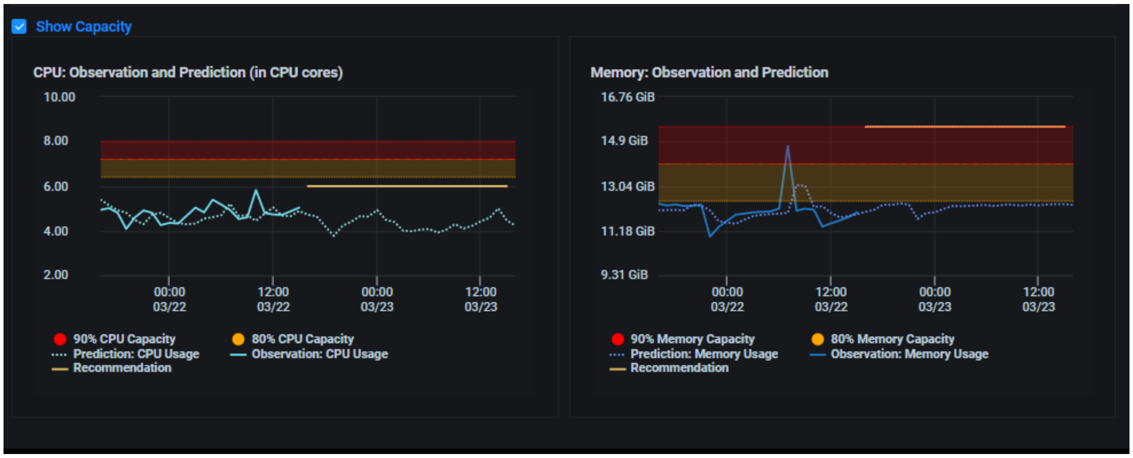
Workload Observation and Prediction Charts



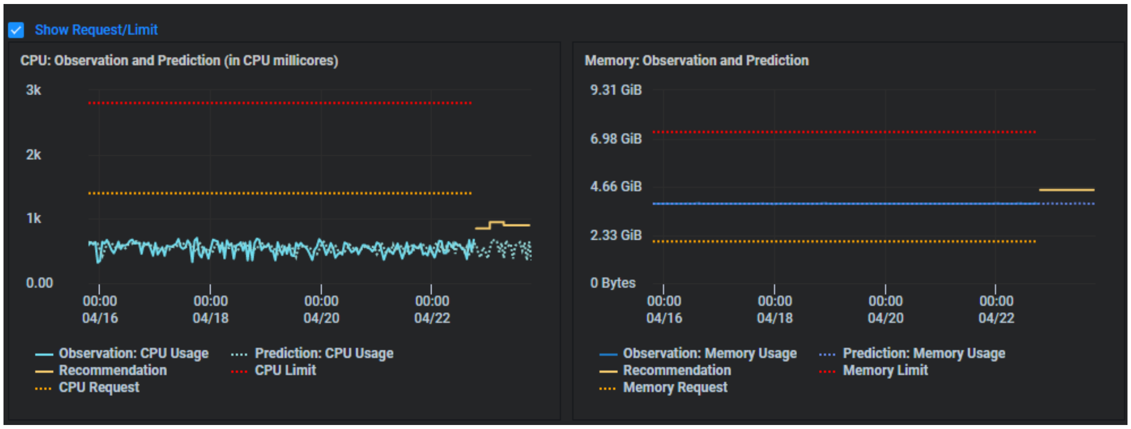
These charts display CPU and memory observations and predictions for all resources specified in the *Filter* panel.

- The solid line represents the observed actual usage.
- The dotted green line represents the past and future predicted usage.
- The solid yellow line represents the recommended usage, which can help you from over-provisioning resources.
- If the selected scope is a Kubernetes node, the solid line represents the node's total CPU and memory. A big difference between total resources and actual and predicted usage can indicate that you are over-provisioned. A small difference between total resources and actual and predicted usage can indicate that you might be under-provisioned.

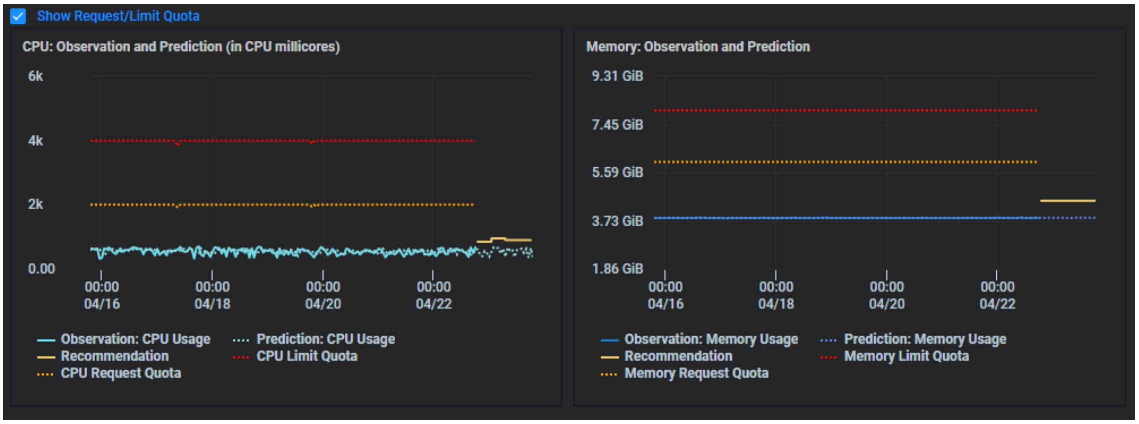
For clusters, check *Show Capacity* to see the maximum CPU and memory usage limits for the cluster, node. Orange represents 80-90% and red represents 90-100%. This is a useful way to see if the utilization of resources is approaching the overall capacity.



For applications and controllers, check *Show Request/Limit* to see the current usage/predictions and recommendations compared to the current resource allocations and limits.



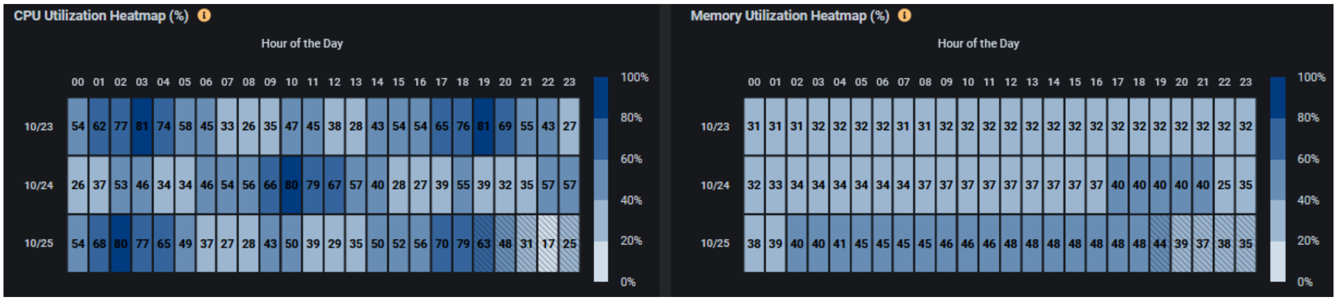
For namespaces, check *Show Request/Limit Quota* to see the current usage, predictions, and recommendations compared to the resource allocations and quota for a namespace.



Utilization Analysis Charts

The *Utilization Analysis* charts display daily, weekly, and monthly CPU and memory utilization data for Kubernetes clusters, nodes, applications, and controllers. Use filters to select the resources and the *Day/Week/Month* field to select a time period. For example, if *Weekly* is selected, use the *Week* field to select a different calendar week.

CPU and Memory Utilization Heatmap Charts


CPU Utilization Heatmap (%)

	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
10/23	54	62	77	81	74	58	45	33	26	35	47	45	38	28	43	54	54	65	76	81	69	55	43	27
10/24	26	37	53	46	34	34	46	54	56	66	80	79	67	57	40	28	27	39	55	39	32	35	57	57
10/25	54	68	80	77	65	49	37	27	28	43	50	39	29	35	50	52	56	70	79	63	48	31	17	25

Memory Utilization Heatmap (%)

	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
10/23	31	31	31	32	32	32	32	31	31	32	32	32	32	32	32	32	32	32	32	32	32	32	32	32
10/24	32	33	34	34	34	34	34	34	37	37	37	37	37	37	37	37	37	40	40	40	40	40	25	35
10/25	38	39	40	40	41	45	45	45	45	46	46	46	48	48	48	48	48	48	48	44	39	37	38	35

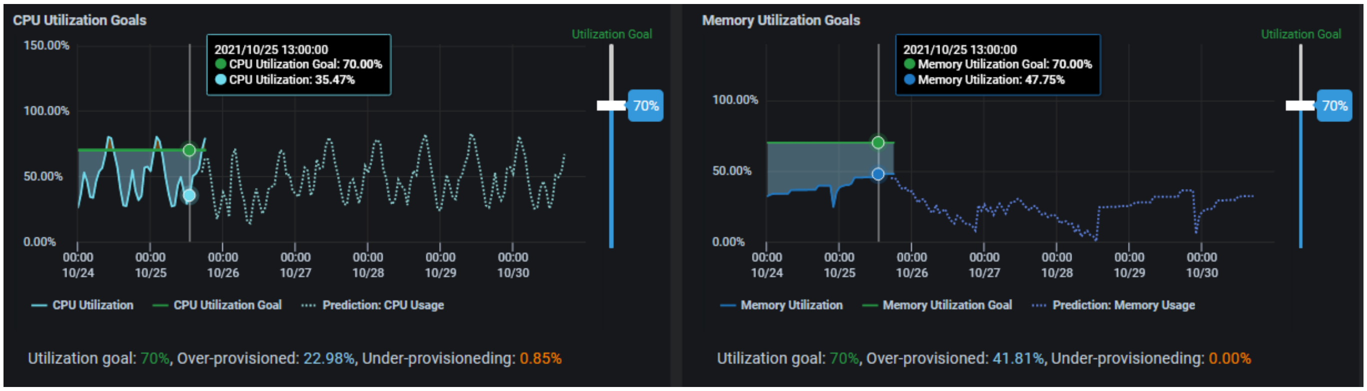
These charts display the actual and predicted CPU and memory usage percentage for all resources specified using filters for the selected time frame:

- Daily – Displays usage every hour for the last three days.
- Weekly – Displays usage each hour for a calendar week.
- Monthly – Displays usage every day for a calendar month.

For clusters and nodes, the percentage is calculated by actual usage divided by capacity. For applications and controllers, the percentage is calculated by actual usage divided by the requested (minimum) CPU/memory or the limit (maximum) CPU/memory.

The color gradient illustrates the percentage range, making it easy to see periods of high and low usage. Boxes with diagonal lines represent future predicted utilization.

CPU and Memory Utilization Goals Charts



Utilization goal: 70%, Over-provisioned: 22.98%, Under-provisioneding: 0.85%

Utilization goal: 70%, Over-provisioned: 41.81%, Under-provisioneding: 0.00%

These interactive charts display target goals along with actual and predicted CPU and memory usage for all resources specified by filters for the selected time frame.

- The blue line represents the actual CPU or memory utilization.
- The green line represents your utilization goal.
- The dotted blue line represents future predicted utilization.
- Gray areas represent periods of time when your resources were over-provisioned (wasted utilization).
- Orange areas represent periods of time when your resources were under-provisioned.

By comparing actual usage to your utilization goals, you can easily see where you are over- or under-provisioned, enabling you to adjust your cloud resources for more efficient usage. For example, if you see times when actual usage is consistently much lower than your utilization goals for a cluster, you may want to deploy additional applications in that cluster.

You can adjust your target utilization goals by using the slider on the right side of the chart.

Related topics:

Terminology

Search/Sort Information in Tables

Show/Hide Information in Charts

Zoom In/Out of Charts

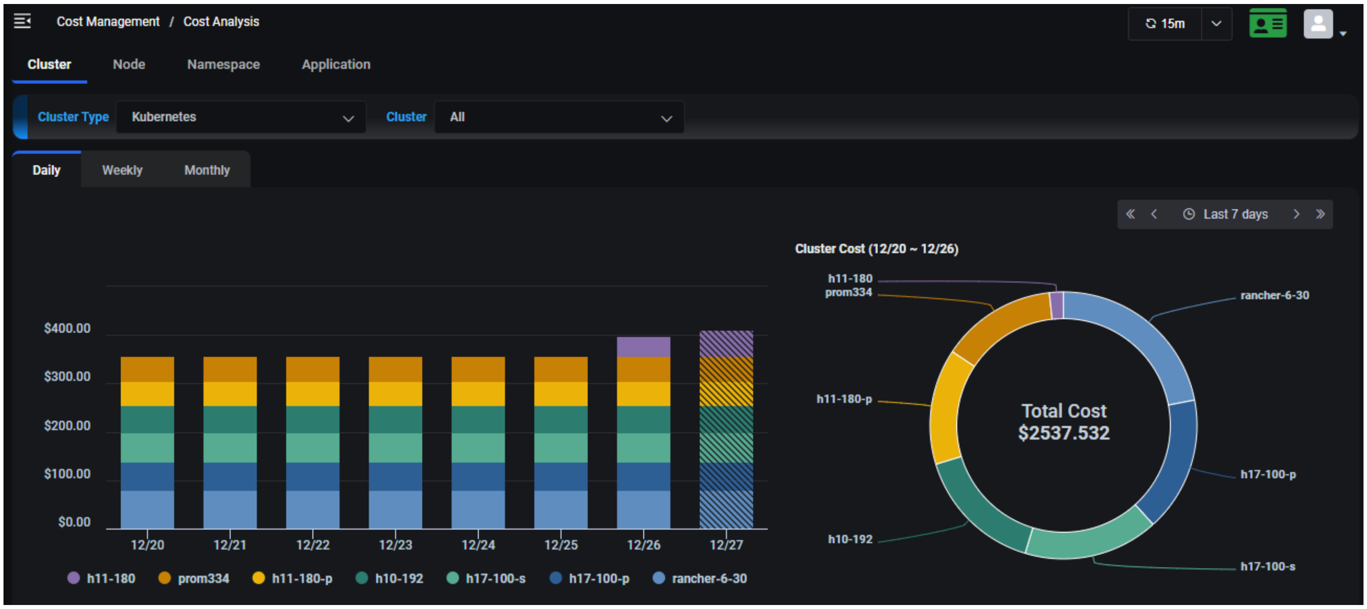
Cluster Configuration

Auto Provisioning

Cost Management – Cost Analysis

The *Cost Analysis* page displays daily, weekly, and monthly costs and predictions. By default, information is displayed for all clusters but can be changed to display one or more Kubernetes clusters or nodes, or one or more Kubernetes namespaces or applications.

Cost Analysis Charts



Cluster Type: Kubernetes

Cluster: All

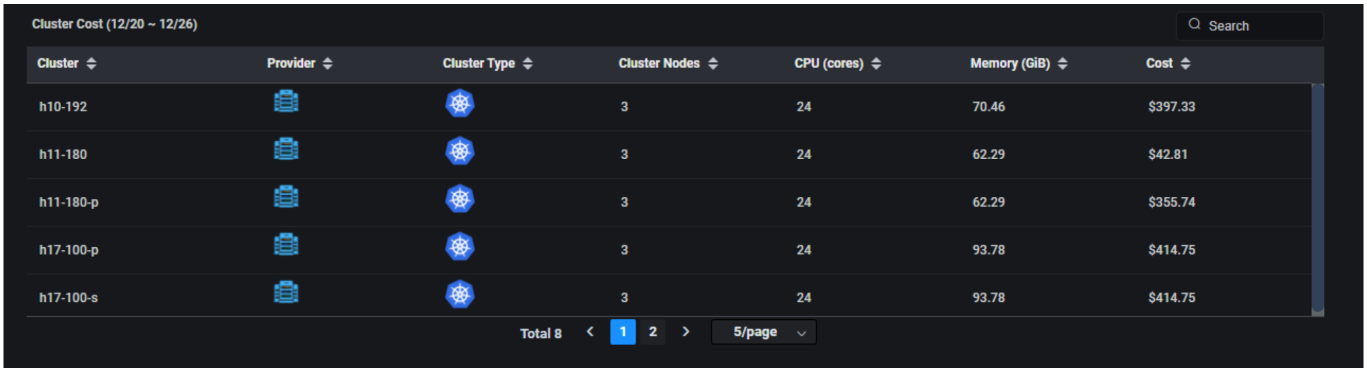
The left chart displays actual costs and predictions for the specified time frame:

- Daily – Displays costs for each day in the last seven days but can be changed to the last 14 or 28 days, or to a custom time range.
- Weekly – Displays costs for each week in the last four weeks but can be changed to the last eight or 12 weeks, or to a custom time range
- Monthly – Displays costs for the last three months but can be changed to the last six or nine months, or to a custom time range.

Click anywhere on the chart to see values for a specific point in time. Highlight or click the key at the bottom of the chart to show/hide individual metrics (i.e., clusters).

The right chart displays the total cost for all clusters. Highlight a section of the chart to see values for a specific cluster.

Cost Analysis Summary Table



Cluster Cost (12/20 ~ 12/26)

Cluster	Provider	Cluster Type	Cluster Nodes	CPU (cores)	Memory (GiB)	Cost
h10-192	🗄	☸	3	24	70.46	$397.33
h11-180	🗄	☸	3	24	62.29	$42.81
h11-180-p	🗄	☸	3	24	62.29	$355.74
h17-100-p	🗄	☸	3	24	93.78	$414.75
h17-100-s	🗄	☸	3	24	93.78	$414.75

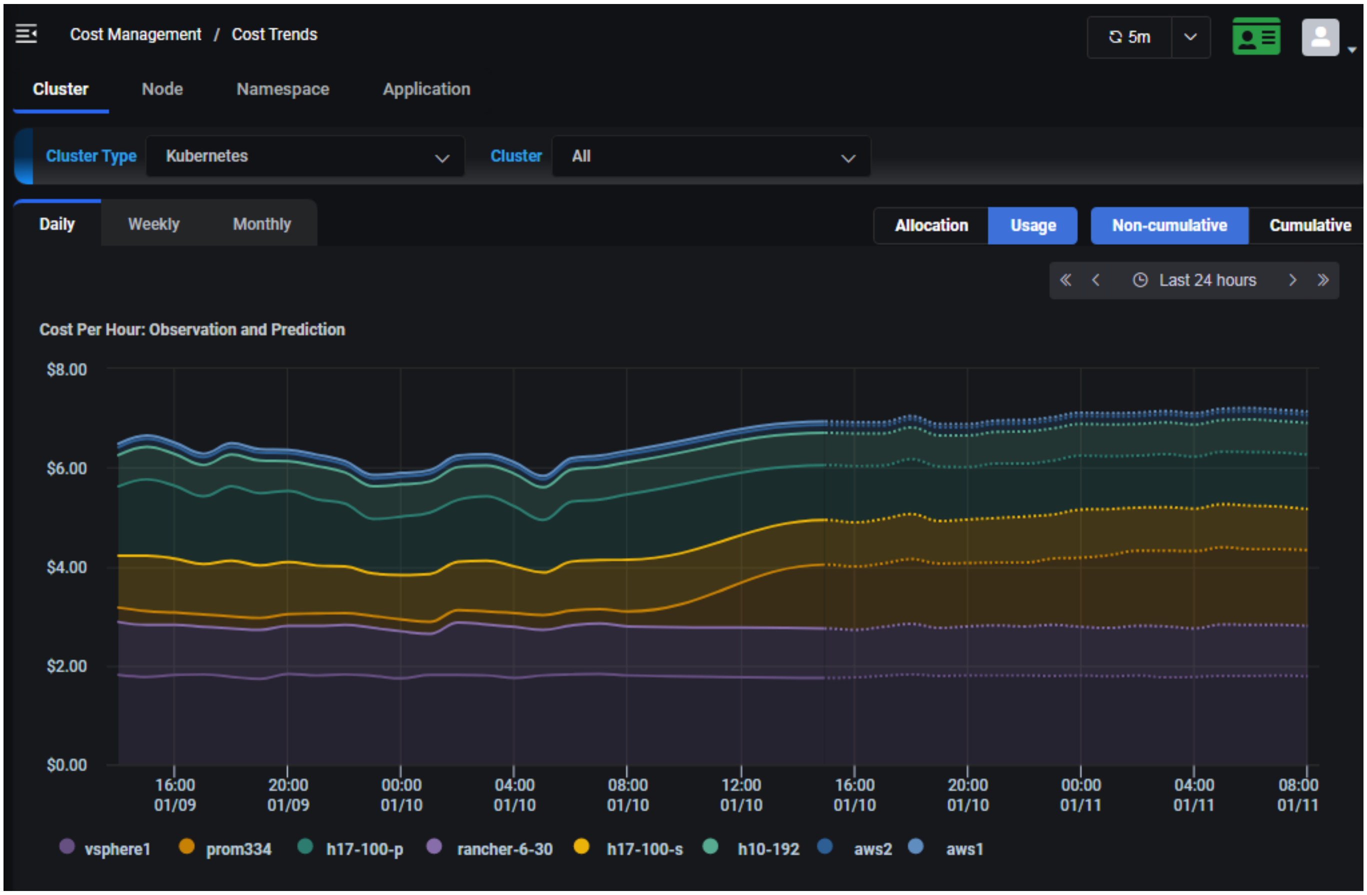
Summary information includes:

- Cluster – Cluster name, provider (AWS, Azure, Google, or on-premises), number of cluster nodes, CPU cores and memory capacity allocated for the cluster during the specified time frame, cost for each cluster for the specified time frame.
- Node – Node name, cluster name, provider, cluster type, instance type used at your cloud provider, CPU cores and memory capacity allocated for the node during the specified time frame, cost for each node for the specified time frame.
- Namespace – Kubernetes namespace name, cluster name, provider, cluster type, number of CPU millicores and amount of memory used during the specified time frame, cost for each namespace for the specified time frame.
- Application – Kubernetes application name, cluster name, provider, cluster type, namespace, CPU millicore and memory requests/limits, cost for each application for the specified time frame.

Cost Management – Cost Trends

The *Cost Trends* page displays daily, weekly, and monthly costs and predictions based on your expected workload. By default, information is displayed for all clusters but can be changed to display one or more Kubernetes clusters or nodes, or one or more Kubernetes namespaces or applications.

Cost Trends Chart



The chart displays actual costs and predictions for the specified time frame:

- Daily – Displays hourly costs for the last 24 hours and predictions for the next 24 hours but can be changed to display costs for the last two or four days, or to a custom time range.
- Weekly – Displays hourly costs for the last seven days and predictions for the next seven days but can be changed to display costs for the last two or four weeks, or to a custom time range.
- Monthly – Displays daily costs for the last 30 days and predictions for the next month but can be changed to display costs for the last two, three, or four months, or to a custom time range.

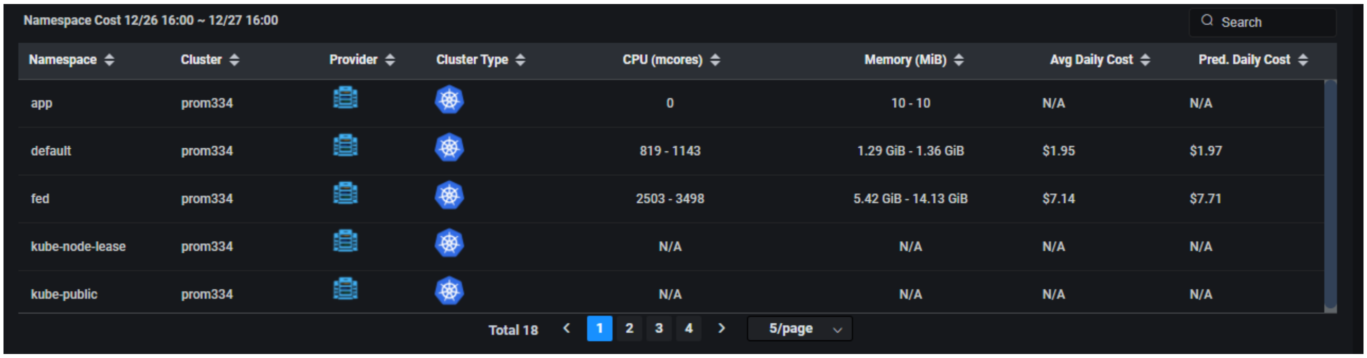
For clusters and nodes, when all clusters/nodes are displayed, you can select to view cost based on allocation or usage. Click *Allocation* to see fixed costs for cluster/node capacity, which will be consistent, as they do not fluctuate much unless a cluster node is added or removed. Click *Usage* to see the cost based on actual usage. When an individual cluster is selected, allocation and usage will both be displayed.

For namespaces and applications, the cost trends are based on the actual usage of resources.

Click the *Cumulative* button to view trends cumulatively through the end of the selected time period.

Click anywhere on the chart to see values for a specific point in time. Click the key at the bottom of the chart to show/hide individual metrics (i.e., namespaces).

Cost Trends Summary Table



Namespace Cost 12/26 16:00 ~ 12/27 16:00

Search

Namespace	Cluster	Provider	Cluster Type	CPU (mcores)	Memory (MiB)	Avg Daily Cost	Pred. Daily Cost
app	prom334	▤	☸	0	10 - 10	N/A	N/A
default	prom334	▤	☸	819 - 1143	1.29 GiB - 1.36 GiB	$1.95	$1.97
fed	prom334	▤	☸	2503 - 3498	5.42 GiB - 14.13 GiB	$7.14	$7.71
kube-node-lease	prom334	▤	☸	N/A	N/A	N/A	N/A
kube-public	prom334	▤	☸	N/A	N/A	N/A	N/A

Total 18 ‹ 1 2 3 4 › 5/page

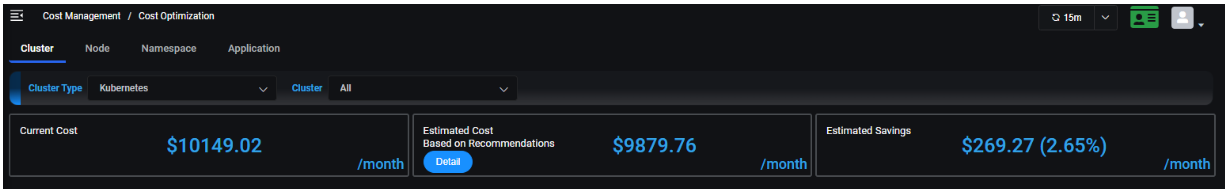
Summary information includes:

- Cluster – Cluster name, provider (AWS, Azure, Google, or on-premises), cluster type (Kubernetes), number of cluster nodes, CPU cores and memory capacity allocated for the cluster during the specified time frame, average and predicted cost for each cluster for the specified time frame (costs are typically the same, unless a cluster node is added or removed).
- Node – Node name, cluster name, provider, cluster type, instance type used at your cloud provider, CPU cores and memory capacity allocated for the node during the specified time frame, average and predicted cost for each node for the specified time frame (costs are typically the same, unless a node is added or removed).
- Namespace – Kubernetes namespace name, cluster name, provider, cluster type, amount CPU/memory used during the specified timeframe, average and predicted cost for each namespace for the specified time frame.
- Application – Kubernetes application name, cluster name, provider, cluster type, CPU millicore and memory requests/limits, average and predicted cost for each application for the specified time frame.

Cost Management – Cost Optimization

The *Cost Optimization* page displays daily, weekly, and monthly recommendations and potential savings based on your expected workload. By default, information is displayed for all clusters but can be changed to display one or more Kubernetes clusters or nodes, or one or more Kubernetes namespaces or applications.

Cost Optimization Chart



Cluster Type: Kubernetes
Cluster: All

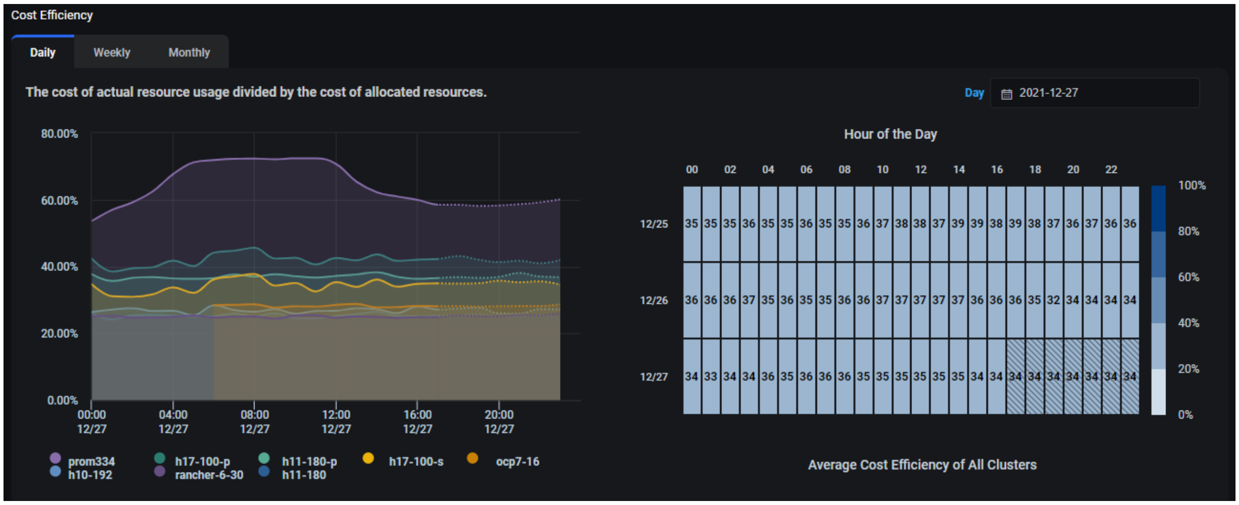
Current Cost $10149.02 /month

Estimated Cost Based on Recommendations $9879.76 /month

Estimated Savings $269.27 (2.65%) /month

This chart displays your current monthly cost with your existing resource configuration, the estimated monthly cost based on recommendations, and the potential monthly savings. Click the *Detail* link to view the recommendations (further down the page) to see how these savings are calculated. The savings typically come from reducing idle resources. The information continually refreshes itself as new data becomes available.

Cost Efficiency Charts



The cost of actual resource usage divided by the cost of allocated resources.

Day: 2021-12-27

	00		02		04		06		08		10		12		14		16		18		20		22	
12/25	35	35	35	36	35	35	36	35	35	36	37	38	38	37	39	39	38	39	38	37	36	37	36	36
12/26	36	36	36	37	35	36	35	35	36	36	37	37	37	37	37	36	36	36	35	32	34	34	34	34
12/27	34	33	34	34	36	35	36	36	36	35	35	35	35	35	35	34	34	34	34	34	34	34	34	34

The *Cost Efficiency* charts display the actual cost of resource usage divided by the cost of allocated resources for the specified time frame as well as cost efficiency predictions going forward. Information is displayed in a time-series graph and heatmap for each cluster/node/namespace/application or an average for all clusters/nodes/namespaces/applications, depending upon what was selected. The lower the percentage, the more wasteful your configuration. The goal is to get to as close to 100% cost efficiency without performance risk as possible.

- Daily – Displays hourly cost efficiency from the start of today and predictions going forward for the day but can be changed to display cost efficiency for a different day.
- Weekly – Displays hourly cost efficiency for the current week and predictions going forward for the week but can be changed to display cost efficiency for a different week.

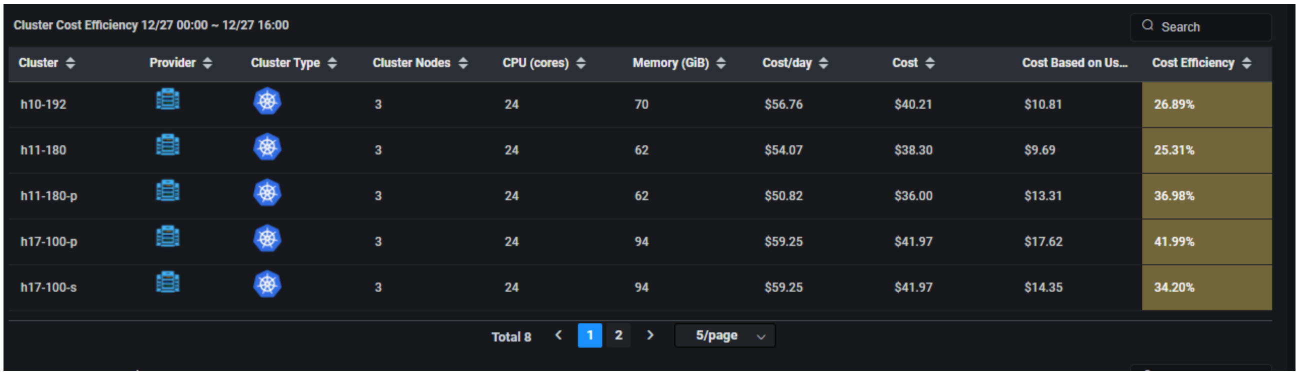
- Monthly – Displays daily cost efficiency for the current month and predictions going forward for the month but can be changed to display cost efficiency for a different month.

Click anywhere on either chart to see values for a specific point in time. Click the key at the bottom of the graph to show/hide individual metrics (i.e., clusters).

The color gradient in the heatmap illustrates the percentage range, making it easy to see periods of high and low usage. Boxes with diagonal lines represent future predicted cost efficiency values.

Cluster Cost Optimization

Cluster Cost Efficiency Table

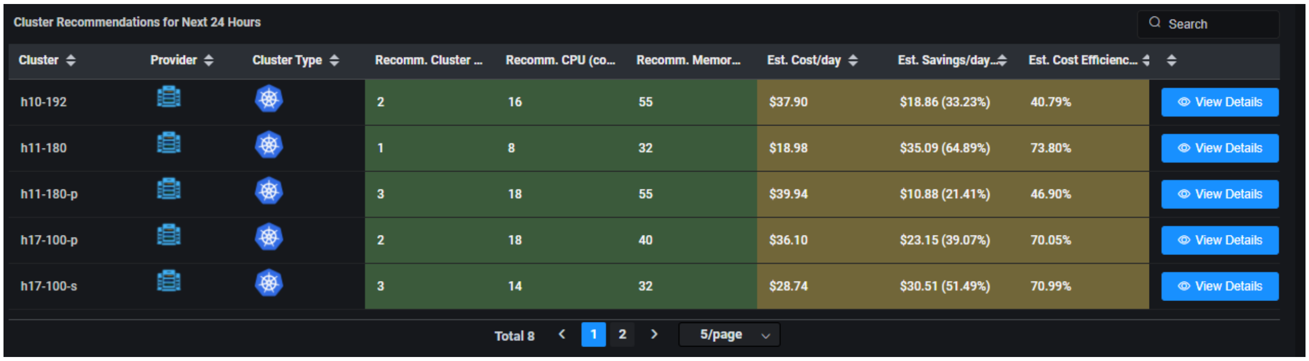


Cluster Cost Efficiency 12/27 00:00 ~ 12/27 16:00

Cluster	Provider	Cluster Type	Cluster Nodes	CPU (cores)	Memory (GiB)	Cost/day	Cost	Cost Based on Us...	Cost Efficiency
h10-192	🗄	☸	3	24	70	$56.76	$40.21	$10.81	26.89%
h11-180	🗄	☸	3	24	62	$54.07	$38.30	$9.69	25.31%
h11-180-p	🗄	☸	3	24	62	$50.82	$36.00	$13.31	36.98%
h17-100-p	🗄	☸	3	24	94	$59.25	$41.97	$17.62	41.99%
h17-100-s	🗄	☸	3	24	94	$59.25	$41.97	$14.35	34.20%

Cluster Cost Efficiency information includes cluster name, provider (AWS, Azure, Google, or on-premises), cluster type (Kubernetes), number of cluster nodes, CPU cores and memory capacity allocated for the cluster during the specified time frame, cost for each cluster for the specified time frame, cost from the actual cluster usage, and the cost efficiency percentage (the cost from actual cluster usage divided by the allocated cost) for the specified time frame.

Cluster Recommendations Table



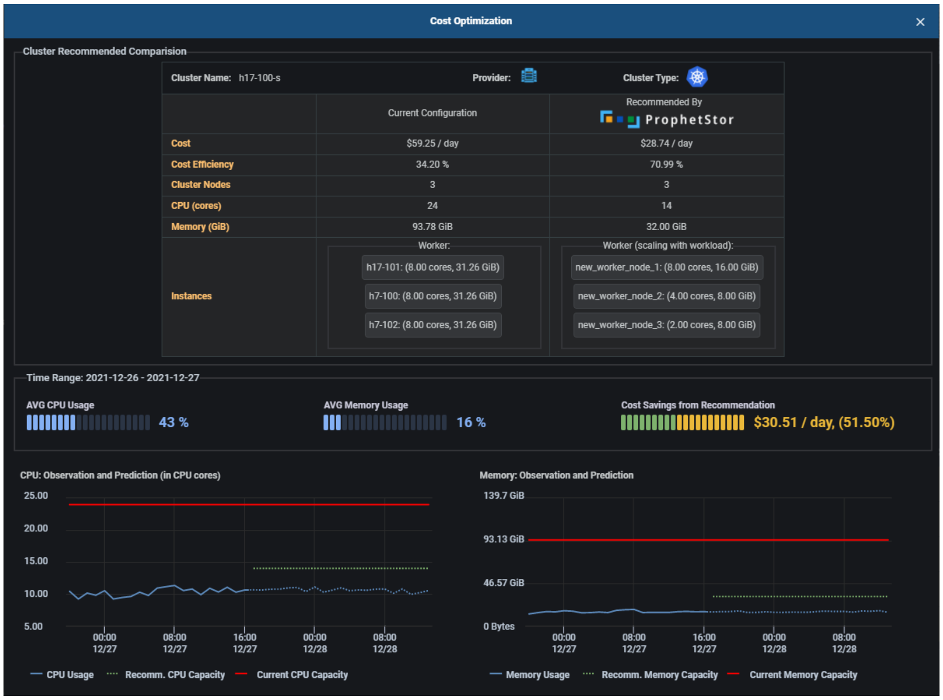
Cluster Recommendations for Next 24 Hours

Cluster	Provider	Cluster Type	Recomm. Cluster ...	Recomm. CPU (co...	Recomm. Memor...	Est. Cost/day	Est. Savings/day...	Est. Cost Efficienc...	
h10-192	🗄	☸	2	16	55	$37.90	$18.86 (33.23%)	40.79%	View Details
h11-180	🗄	☸	1	8	32	$18.98	$35.09 (64.89%)	73.80%	View Details
h11-180-p	🗄	☸	3	18	55	$39.94	$10.88 (21.41%)	46.90%	View Details
h17-100-p	🗄	☸	2	18	40	$36.10	$23.15 (39.07%)	70.05%	View Details
h17-100-s	🗄	☸	3	14	32	$28.74	$30.51 (51.49%)	70.99%	View Details

The *Cluster Recommendations* table shows the changes you can make to save money. For example, it may show that your cost efficiency is 60%, meaning you are only using 60% of your currently allocated resources, and based on your predicted workload, four nodes are sufficient instead of the five current ones.

The table displays cluster name, provider (AWS, Azure, Google, or on-premises), cluster type (Kubernetes), recommended number of cluster nodes, CPU cores, and memory, as well as the estimated cost, savings, and cost efficiency percentage (the cost from predicted cluster usage divided by the allocated cost) per day/week/month for the cluster if the recommendations are followed. Click the *View Details* link to view the recommendations to see how these savings are calculated.

Cluster Cost Optimization Details



Cost Optimization

Cluster Recommended Comparision

Cluster Name: h17-100-s	Provider:	Cluster Type:
	Current Configuration	Recommended By ProphetStor
Cost	$59.25 / day	$28.74 / day
Cost Efficiency	34.20 %	70.99 %
Cluster Nodes	3	3
CPU (cores)	24	14
Memory (GiB)	93.78 GiB	32.00 GiB
Instances	Worker: h17-101: (8.00 cores, 31.26 GiB) h7-100: (8.00 cores, 31.26 GiB) h7-102: (8.00 cores, 31.26 GiB)	Worker (scaling with workload): new_worker_node_1: (8.00 cores, 16.00 GiB) new_worker_node_2: (4.00 cores, 8.00 GiB) new_worker_node_3: (2.00 cores, 8.00 GiB)

Time Range: 2021-12-26 - 2021-12-27

AVG CPU Usage 43 %

AVG Memory Usage 16 %

Cost Savings from Recommendation $30.51 / day, (51.50%)

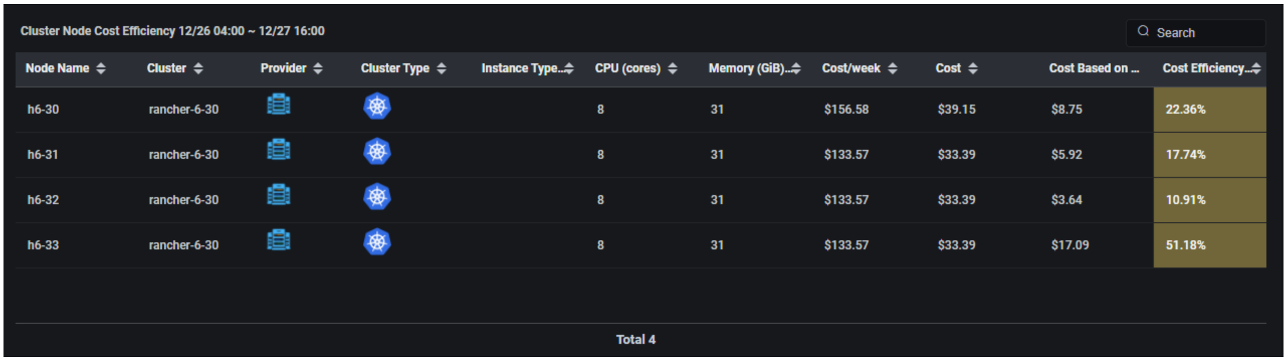
The *Cost Optimization* details page for a cluster shows the current configuration vs. the recommended number of cluster nodes, CPU cores, memory, and specific CPU/memory of each instance for the specified time frame (day/week/month).

Average actual CPU and memory usage is displayed along with the cost savings if the recommendations are followed.

Current CPU and memory capacity, recommended CPU and memory capacity, along with observed and predicted CPU and memory usage are also shown for Kubernetes.

Cluster Node Cost Optimization

Cluster Node Cost Efficiency Table



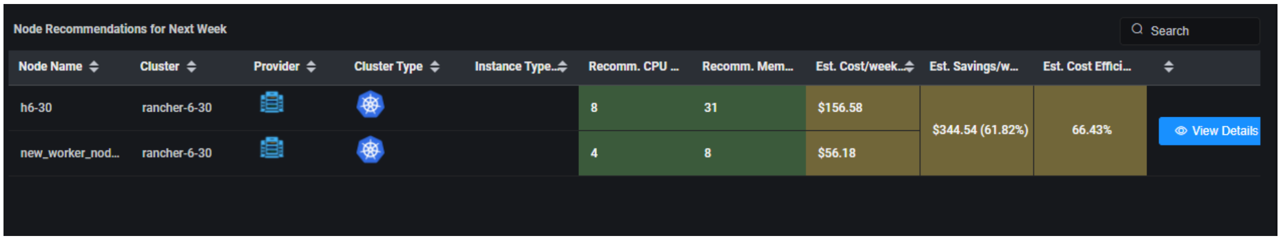
Cluster Node Cost Efficiency 12/26 04:00 ~ 12/27 16:00

Node Name	Cluster	Provider	Cluster Type	Instance Type...	CPU (cores)	Memory (GiB)...	Cost/week	Cost	Cost Based on ...	Cost Efficiency...
h6-30	rancher-6-30	▤	☸		8	31	$156.58	$39.15	$8.75	22.36%
h6-31	rancher-6-30	▤	☸		8	31	$133.57	$33.39	$5.92	17.74%
h6-32	rancher-6-30	▤	☸		8	31	$133.57	$33.39	$3.64	10.91%
h6-33	rancher-6-30	▤	☸		8	31	$133.57	$33.39	$17.09	51.18%

Total 4

Cluster Node Cost Efficiency information includes node name, cluster name, provider (AWS, Azure, Google, or on-premises), cluster type (Kubernetes), instance type used at your cloud provider, CPU cores and memory capacity allocated for the node during the specified time frame, cost for each node and cost based on actual usage for the specified time frame, and the cost efficiency percentage (the cost from actual node usage divided by the allocated cost) for the specified time frame.

Cluster Node Recommendations Table



Node Recommendations for Next Week

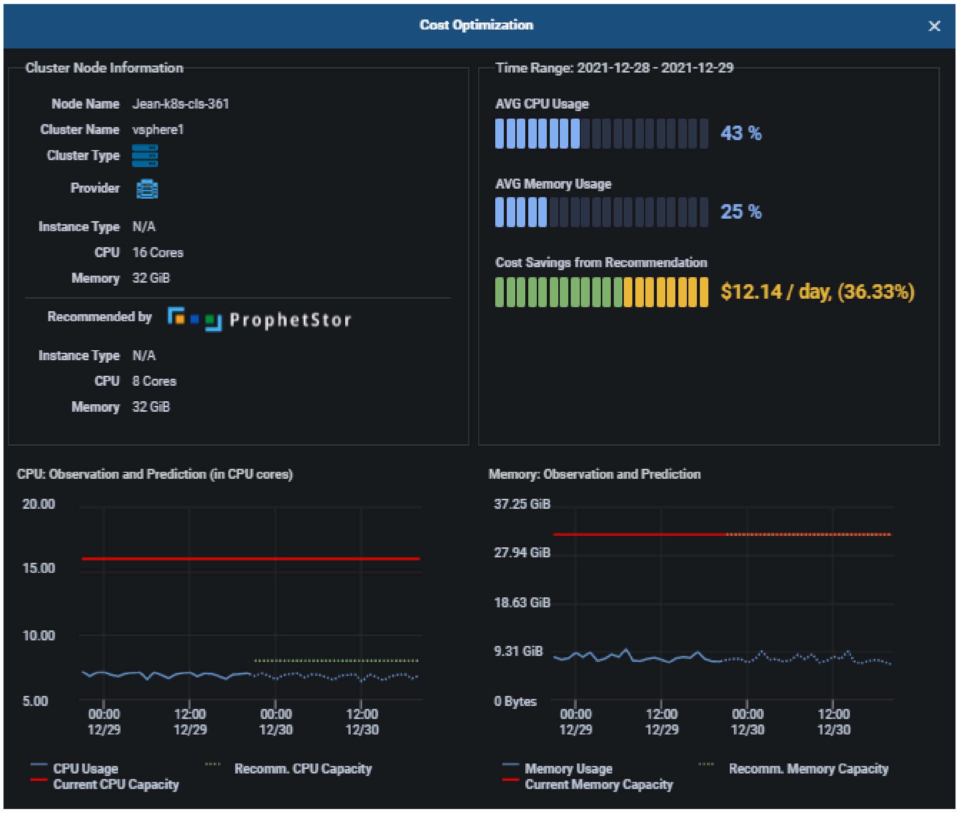
Node Name	Cluster	Provider	Cluster Type	Instance Type...	Recomm. CPU ...	Recomm. Mem...	Est. Cost/week...	Est. Savings/w...	Est. Cost Effici...	
h6-30	rancher-6-30	▤	☸		8	31	$156.58	$344.54 (61.82%)	66.43%	View Details
new_worker_nod...	rancher-6-30	▤	☸		4	8	$56.18			

The *Cluster Node Recommendations* table displays cluster name, provider (AWS, Azure, Google, or on-premises), cluster type (Kubernetes), instance type, recommended CPU cores and memory, as well as the estimated cost, savings, and cost efficiency percentage (the cost from predicted usage divided by the allocated cost) per day/week/month if the recommendations are followed. Note that since containers are deployed to different cluster nodes by Kubernetes, there is no individual recommendation for a cluster node. Instead, the cost optimization and recommendations are applied to the entire cluster. Click the *View Details* link to view the recommendations to see how these savings are calculated.

This table shows the changes you can make to save money. However, the information presented is based on the cluster type:

- Kubernetes – Since containers are deployed to different cluster nodes by Kubernetes, there is no individual recommendation of individual cluster node. Recommendations for the optimal number of nodes in Kubernetes clusters are based on the cluster, even if an individual node is selected for display. Recommendations for CPU and memory are per node. For example, the system may recommend fewer nodes for the cluster but more CPU and memory for each node.

Cluster Node Cost Optimization Details



Cost Optimization

Cluster Node Information

Node Name Jean-k8s-cls-361
Cluster Name vsphere1
Cluster Type
Provider
Instance Type N/A
CPU 16 Cores
Memory 32 GiB

Recommended by ProphetStor

Instance Type N/A
CPU 8 Cores
Memory 32 GiB

Time Range: 2021-12-28 - 2021-12-29

AVG CPU Usage 43 %

AVG Memory Usage 25 %

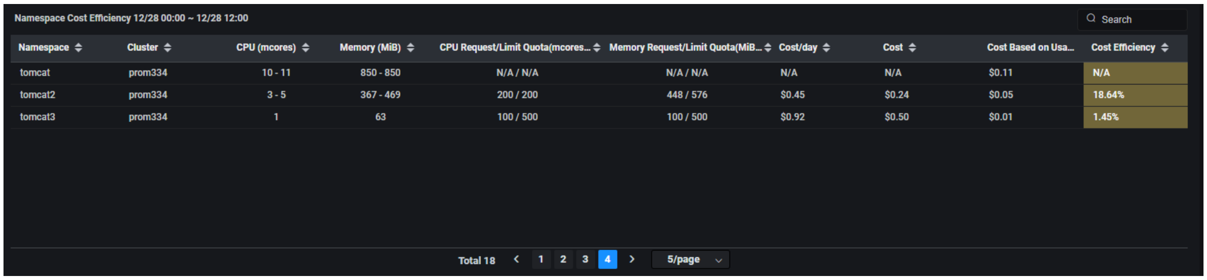
Cost Savings from Recommendation $12.14 / day, (36.33%)

The *Cost Optimization* details page shows the current configuration vs. the recommended instance type, number of CPU cores and memory for the specified time frame (day/week/month) of a specific node. For Kubernetes clusters, recommendations are displayed for the entire cluster.

Cost from usage, cost from recommendations, and potential savings are also displayed. Click *Cumulative* to view information cumulatively through the end of the selected time period.

Namespace Cost Optimization

Namespace Cost Efficiency Table



Namespace Cost Efficiency 12/28 00:00 ~ 12/28 12:00

Search

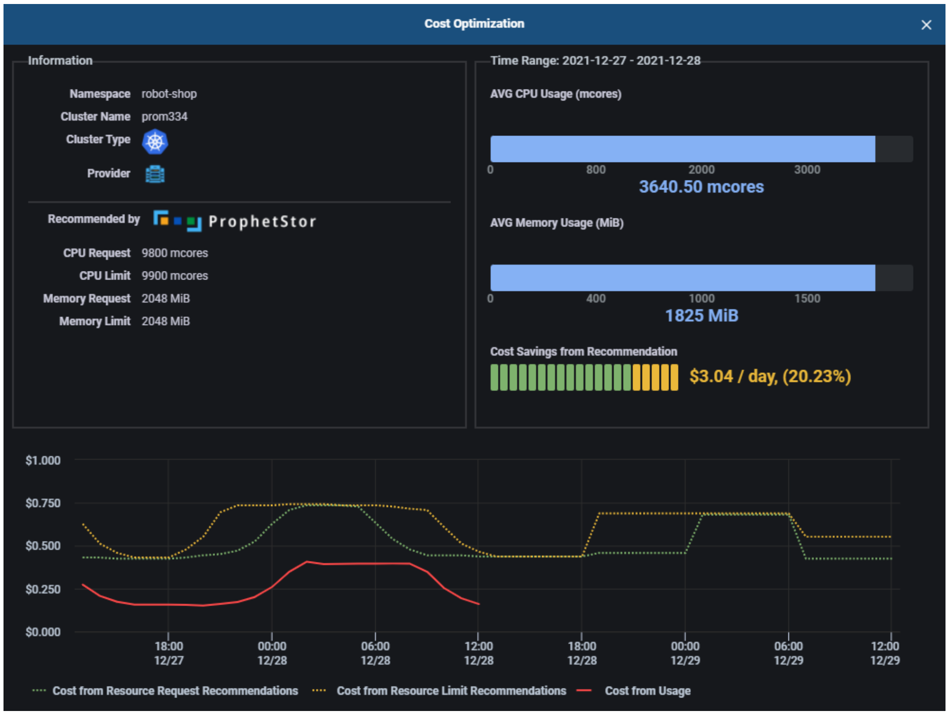
Namespace	Cluster	CPU (mcores)	Memory (MiB)	CPU Request/Limit Quota(mcores...	Memory Request/Limit Quota(MiB...	Cost/day	Cost	Cost Based on Usa...	Cost Efficiency
tomcat	prom334	10 - 11	850 - 850	N/A / N/A	N/A / N/A	N/A	N/A	$0.11	N/A
tomcat2	prom334	3 - 5	367 - 469	200 / 200	448 / 576	$0.45	$0.24	$0.05	18.64%
tomcat3	prom334	1	63	100 / 500	100 / 500	$0.92	$0.50	$0.01	1.45%

Total 18 ‹ 1 2 3 4 › 5/page

Namespace Cost Efficiency information includes Kubernetes namespace, cluster, number of CPU millicores and amount of memory used during the specified time frame, CPU and memory request/limit quota, as well as the daily/weekly/monthly cost based on resource quota, current up-to-date cost based on quota, cost based on actual usage, and cost efficiency percentage (the cost from actual resource usage by this namespace divided by the namespace resource quota) per day/week/month for the namespace.

Namespace Recommendation information includes Kubernetes namespace, cluster, recommended CPU and memory request/limit quota, as well as the estimated cost, estimated savings, and estimated cost efficiency percentage based on recommended resource quota per day/week/month for the namespace. Click the *View Details* link to view the recommendations to see how these savings are calculated.

Namespace Optimization Details



Cost Optimization

Information

Namespace: robot-shop
Cluster Name: prom334
Cluster Type
Provider

Recommended by ProphetStor

CPU Request: 9800 mcores
CPU Limit: 9900 mcores
Memory Request: 2048 MiB
Memory Limit: 2048 MiB

Time Range: 2021-12-27 - 2021-12-28

AVG CPU Usage (mcores): 3640.50 mcores

AVG Memory Usage (MiB): 1825 MiB

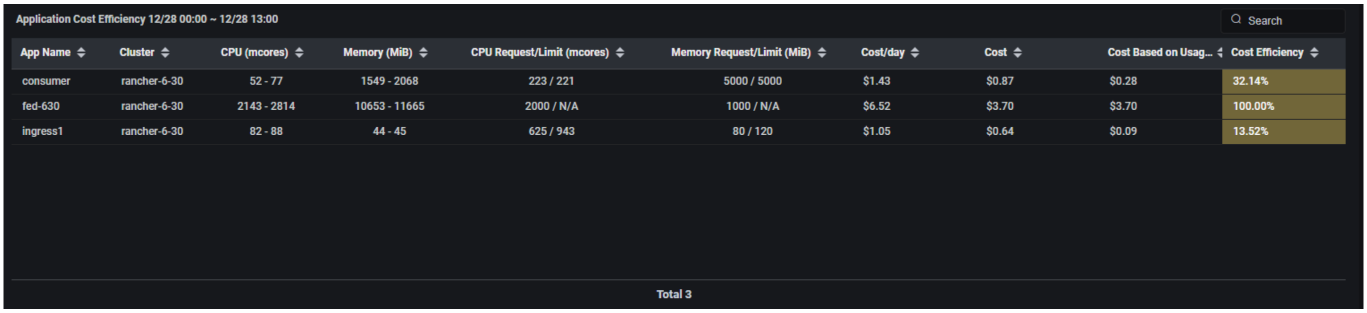
Cost Savings from Recommendation: $3.04 / day, (20.23%)

The *Namespace Optimization* details page shows the current configuration vs. the recommended amount of CPU and memory requests/limits quota for the specified time frame (day/week/month).

Average actual CPU and memory usage is displayed along with the cost savings if the recommendations are followed. Cost from usage and cost from request/limit recommendations are also displayed. Note that the calculation of cost efficiency and estimated savings requires namespace resource quota information. If the information is not available, the current cost efficiency and estimated cost savings will not be available.

Application Cost Optimization

Application Cost Efficiency Table



Application Cost Efficiency 12/28 00:00 ~ 12/28 13:00

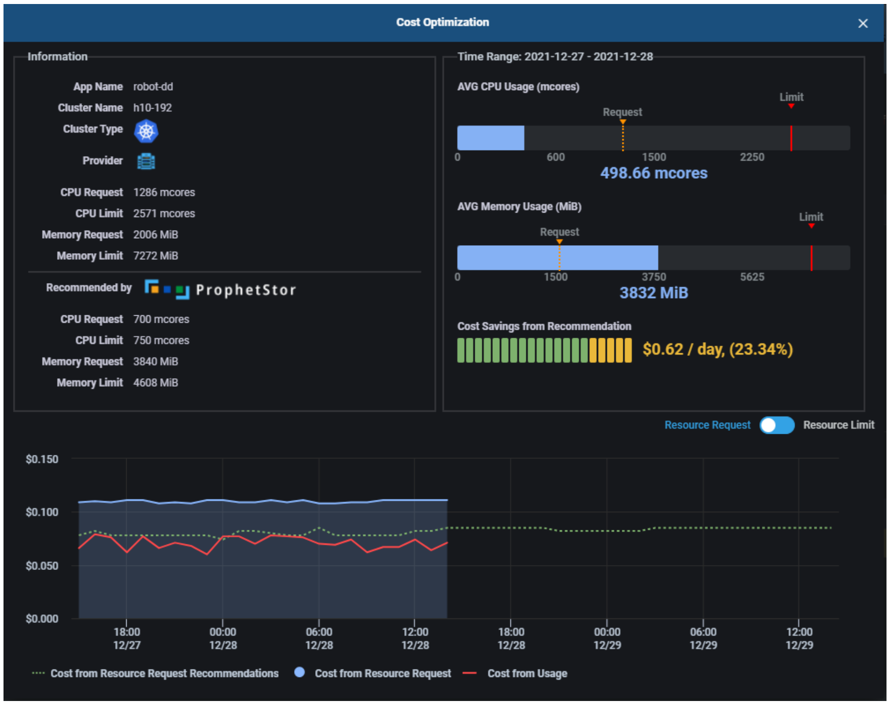
App Name	Cluster	CPU (mcores)	Memory (MiB)	CPU Request/Limit (mcores)	Memory Request/Limit (MiB)	Cost/day	Cost	Cost Based on Usag...	Cost Efficiency
consumer	rancher-6-30	52 - 77	1549 - 2068	223 / 221	5000 / 5000	$1.43	$0.87	$0.28	32.14%
fed-630	rancher-6-30	2143 - 2814	10653 - 11665	2000 / N/A	1000 / N/A	$6.52	$3.70	$3.70	100.00%
ingress1	rancher-6-30	82 - 88	44 - 45	625 / 943	80 / 120	$1.05	$0.64	$0.09	13.52%

Total 3

Cost Efficiency information includes Kubernetes application name, cluster, CPU and memory usage during the time period, CPU and memory requests/limits, as well as the daily/weekly/monthly cost, up-to-date cost, actual cost based on usage, and cost efficiency percentage (the cost from actual application usage divided by the allocated cost) per day/week/month.

Application Recommendations information includes Kubernetes application name, cluster, recommended CPU and memory requests/limits, as well as the estimated cost, savings, and cost efficiency percentage (the cost from actual application usage divided by the allocated cost) per day/week/month for the application if the recommendations are followed. Each application has its own recommendations. Click the *View Details* link to view the recommendations to see how these savings are calculated.

Application Optimization Details



Cost Optimization

Information

App Name robot-dd
Cluster Name h10-192
Cluster Type
Provider

CPU Request 1286 mcores
CPU Limit 2571 mcores
Memory Request 2006 MiB
Memory Limit 7272 MiB

Recommended by ProphetStor

CPU Request 700 mcores
CPU Limit 750 mcores
Memory Request 3840 MiB
Memory Limit 4608 MiB

Time Range: 2021-12-27 - 2021-12-28

AVG CPU Usage (mcores)

498.66 mcores

AVG Memory Usage (MiB)

3832 MiB

Cost Savings from Recommendation

$0.62 / day, (23.34%)

The *Application Optimization* details page shows the current configuration vs. the recommended amount of CPU and memory requests/limits for the specified time frame (day/week/month).

Average actual CPU and memory usage is displayed along with the cost savings if the recommendations are followed.

Cost from usage, cost from current resource requests/limits, and cost from recommendations are also displayed. You can switch between *Resource Request and Resource Limit* to see the savings comparison.

Related topics:

Application Cost Analysis

Cost Allocation Kubernetes Namespaces

Common Administration Portal Functions

Terminology

Search/Sort Information in Tables

Show/Hide Information in Charts

Zoom In/Out of Charts

Price Books

Configuration - Clusters

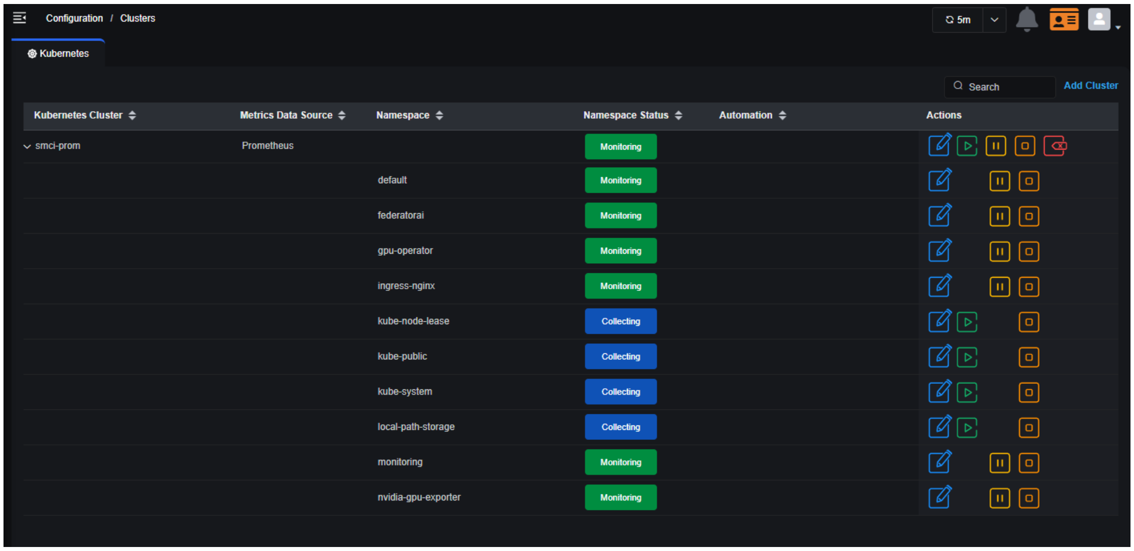
The *Clusters* page displays all Kubernetes clusters being monitored by Federator.ai GPU Booster. You can add, manage, and remove clusters from this page.

Kubernetes Clusters

Select the *Kubernetes* tab to see the list of monitored Kubernetes clusters. Expand a cluster to see the namespaces for a cluster.

You can also see namespace status for the cluster and individual namespaces. The status will be *Monitoring* when the system is collecting metrics and providing workload predictions. The status will be *Collecting* when all the namespaces in the cluster are in collecting state. When a namespace is in collecting state, the system will still collect metrics, but prediction tasks are paused. When a namespace is added to an existing cluster, it will collect data but will not be automatically monitored until it is manually set for monitoring.

A status of *Stopped* means there is no collecting of metrics and no predictions. A cluster is *Stopped* if all of its namespaces are stopped.



Kubernetes Cluster	Metrics Data Source	Namespace	Namespace Status	Automation	Actions
smci-prom	Prometheus		Monitoring		✎ ▷ ⏸ □ ⌫
		default	Monitoring		✎ ⏸ □
		federatorai	Monitoring		✎ ⏸ □
		gpu-operator	Monitoring		✎ ⏸ □
		ingress-nginx	Monitoring		✎ ⏸ □
		kube-node-lease	Collecting		✎ ▷ □
		kube-public	Collecting		✎ ▷ □
		kube-system	Collecting		✎ ▷ □
		local-path-storage	Collecting		✎ ▷ □
		monitoring	Monitoring		✎ ⏸ □
		nvidia-gpu-exporter	Monitoring		✎ ⏸ □

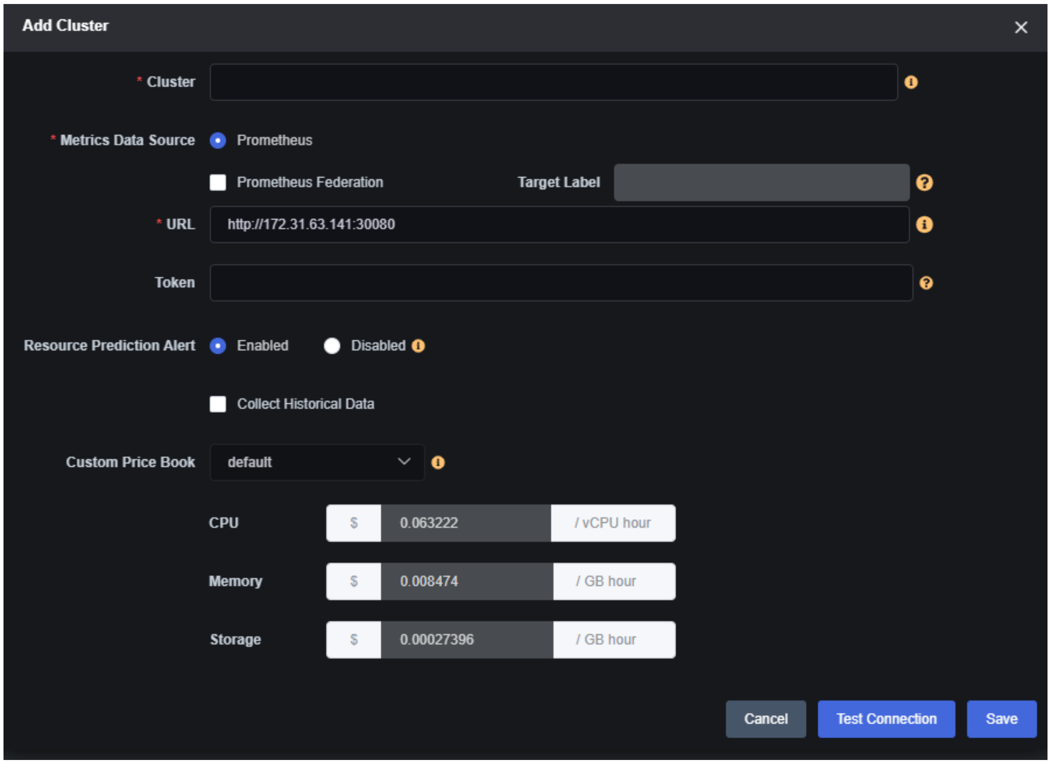
In addition to adding a cluster, you can perform the following functions for existing clusters and namespaces:

Icon	Function
✎	Edit settings for clusters and namespaces.
▷	Start monitoring and predictions for all namespaces or a specific namespace.
⏸	Pause monitoring and predictions for all namespaces or a specific namespace.
□	Stop collecting metrics and making predictions for all namespaces or a specific namespace.

	Remove a cluster.

Add a Kubernetes Cluster

1. On the *Configuration* / *Clusters* page, click *Add Cluster*.



Add Cluster

* Cluster

* Metrics Data Source ● Prometheus

□ Prometheus Federation Target Label

* URL http://172.31.63.141:30080

Token

Resource Prediction Alert ● Enabled ○ Disabled

□ Collect Historical Data

Custom Price Book default

CPU	$	0.063222	/ vCPU hour
Memory	$	0.008474	/ GB hour
Storage	$	0.00027396	/ GB hour

Cluster – Specify the name of a cluster to be managed. There is a maximum of 253 lowercase characters,"-", or "." allowed. The name must start and end with an alphanumeric character.

Metrics Data Source – Select the source of metrics for this cluster.

Prometheus Federation – If *Prometheus* is your data source, specify if you are using Federation, which is a group of Prometheus servers that send metrics to a centralized Prometheus server. You will need to specify the target label of the centralized Prometheus server. The format is: `<label-name>:<label-value>` (e.g., clusterID:host-1).

URL/Token – For the Prometheus open-source monitoring system, the URL is required but the token is optional. Each cluster can use different values.

- Authenticate Prometheus by using basic authentication with a username and password.

  Use the following command to generate the token:

  ```
  # echo -n “<username>:<password>” | base64
  ```

  Refer to the following for information about securing the Prometheus API using basic authentication (Basic Auth): https://prometheus.io/docs/guides/basic-auth/

- Authenticate Prometheus by using a service account token in OpenShift:

  Use the following commands to get the service account name for Prometheus:

  ```
  # oc get prometheus -n openshift-monitoring
  NAME AGE
  k8s 169d
  # oc get prometheus -n openshift-monitoring k8s -oyaml | grep serviceAccount
  serviceAccountName: prometheus-k8s
  ```

  Use the following command to get the token for the Prometheus service account:

  ```
   # oc serviceaccounts get-token prometheus-k8s -n openshift-monitoring
  ```

Collect Historical Data – Specify if you want the system to collect up to three months' worth of historical data for existing nodes and namespaces in this cluster. This will enable weekly and monthly predictions, recommendations, and cost analysis for newly added clusters without waiting to collect weeks’ or months’ worth of data. If less than three months’ worth of data exists, the system will collect the maximum data that is available. Typically, collection of historical data will complete in about 2-3 hours. If you need to pause collection, you can edit cluster settings.

Note that Federator.ai GPU Booster uses the APIs provided by your metrics data source to access historical data. The data source imposes limits on how many calls can be made to their service per hour. If the rate limit is too low, the queries for historical data may exceed the limit and the API will return an error. You will need to contact your metrics data service provider to raise your API rate limit.

Custom Price Book – Specify the price book to use for this cluster if it is located on-premises.

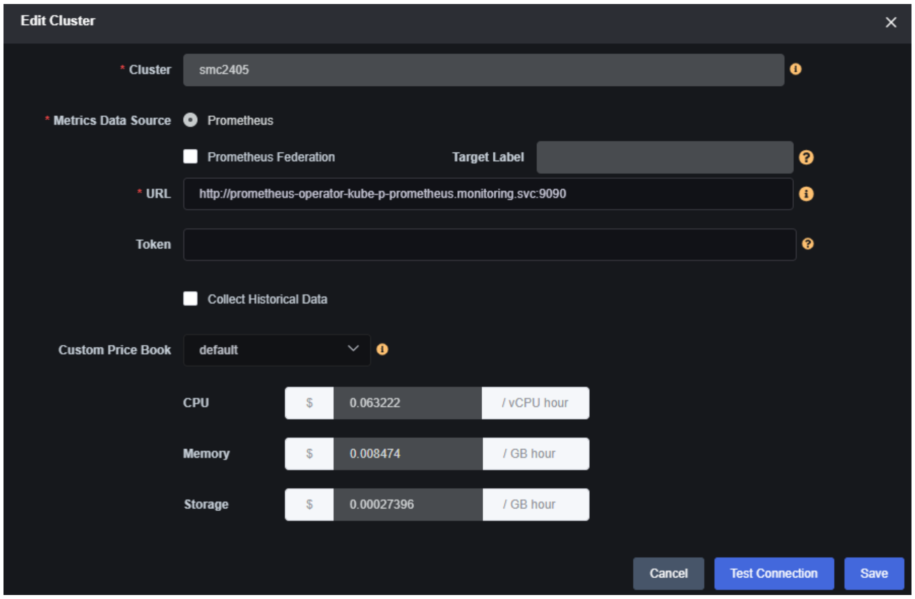
2. Click *Test Connection* to confirm that all information is correct.

3. Click *Save* when the system can connect to the cluster.

Manage Kubernetes Clusters

You can do the following from the *Configuration / Clusters* page:

- Edit cluster settings – Click the *Edit Cluster* icon to configure Prometheus Federation (Prometheus), test the connection to the cluster, or change the custom price book (local clusters). For historical data collection, you can:
  - Start the collection of up to three months' worth of historical data (from the current time) for existing nodes and namespaces in the cluster, if it was not enabled when the cluster was added.
  - See the status of data collection, including the time period collected.
  - Pause/resume historical data collection that is in progress.



Edit Cluster

* Cluster: smc2405

* Metrics Data Source: ● Prometheus

□ Prometheus Federation Target Label:

* URL: http://prometheus-operator-kube-p-prometheus.monitoring.svc:9090

Token:

□ Collect Historical Data

Custom Price Book: default

CPU: $ 0.063222 / vCPU hour

Memory: $ 0.008474 / GB hour

Storage: $ 0.00027396 / GB hour

- Edit namespace settings – Change monitoring status and configure auto provisioning for the namespace. To do this, click the *Edit Namespace* icon.
- Start monitoring and prediction for all namespaces in the cluster or a specific namespace. To do this, click the *Start Monitoring and Predictions* icon.
- Pause monitoring and prediction for all namespaces or a specific namespace. To do this, click the *Pause Monitoring and Predictions* icon for a cluster or for a namespace.
- Stop collecting metrics and making predictions for all namespaces or a specific namespace. To do this, click the *Stop Collecting Metrics and Predictions* icon for a cluster or for a namespace.
- Remove a cluster that does not have any applications configured. To do this, click the *Remove Cluster* icon.

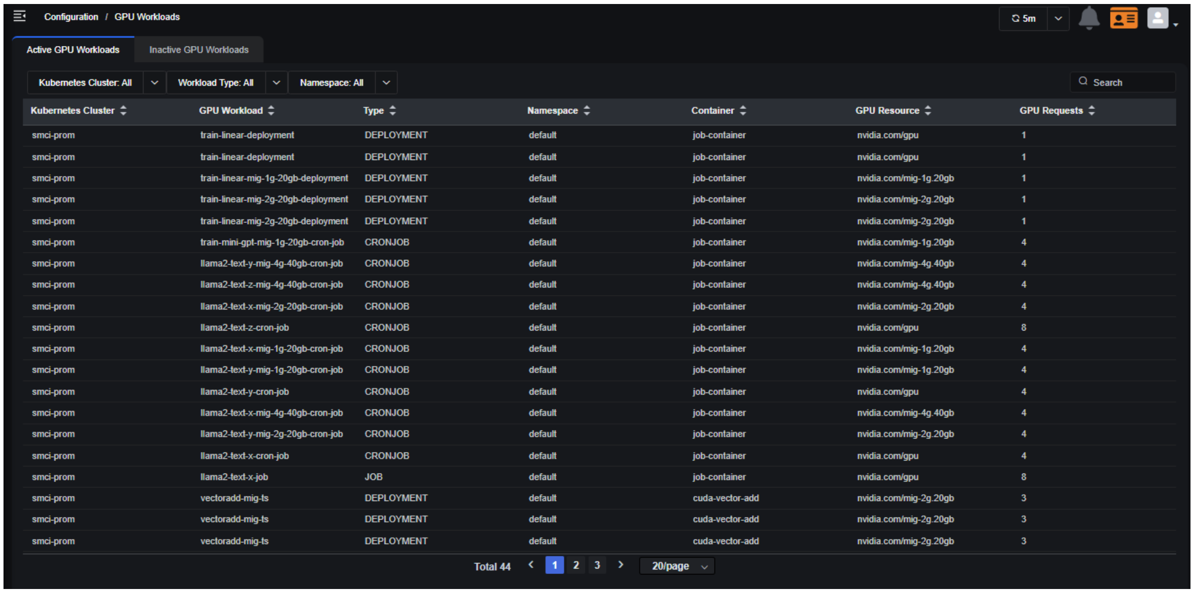
Related topics:

Terminology
Search/Sort Information in Tables
Configure Applications
Custom Price Books

Configuration – GPU Workloads

The *GPU Workloads* page displays all active GPU workloads as well as inactive GPU workloads in separate tabs. An active GPU workload becomes inactive if it is deleted from the Kubernetes cluster. GPU workloads are automatically discovered in a Kubernetes cluster that is managed by Federator.ai GPU Booster.

- *Active GPU Workloads* – This page displays all active GPU workloads from all clusters. Use drop-down menu to filter GPU workloads to specific cluster, namespace, or workload type.



Kubernetes Cluster: All | Workload Type: All | Namespace: All

Kubernetes Cluster	GPU Workload	Type	Namespace	Container	GPU Resource	GPU Requests
smci-prom	train-linear-deployment	DEPLOYMENT	default	job-container	nvidia.com/gpu	1
smci-prom	train-linear-deployment	DEPLOYMENT	default	job-container	nvidia.com/gpu	1
smci-prom	train-linear-mig-1g-20gb-deployment	DEPLOYMENT	default	job-container	nvidia.com/mig-1g.20gb	1
smci-prom	train-linear-mig-2g-20gb-deployment	DEPLOYMENT	default	job-container	nvidia.com/mig-2g.20gb	1
smci-prom	train-linear-mig-2g-20gb-deployment	DEPLOYMENT	default	job-container	nvidia.com/mig-2g.20gb	1
smci-prom	train-mini-gpt-mig-1g-20gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-1g.20gb	4
smci-prom	llama2-text-y-mig-4g-40gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-4g.40gb	4
smci-prom	llama2-text-z-mig-4g-40gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-4g.40gb	4
smci-prom	llama2-text-x-mig-2g-20gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-2g.20gb	4
smci-prom	llama2-text-z-cron-job	CRONJOB	default	job-container	nvidia.com/gpu	8
smci-prom	llama2-text-x-mig-1g-20gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-1g.20gb	4
smci-prom	llama2-text-y-mig-1g-20gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-1g.20gb	4
smci-prom	llama2-text-y-cron-job	CRONJOB	default	job-container	nvidia.com/gpu	4
smci-prom	llama2-text-x-mig-4g-40gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-4g.40gb	4
smci-prom	llama2-text-y-mig-2g-20gb-cron-job	CRONJOB	default	job-container	nvidia.com/mig-2g.20gb	4
smci-prom	llama2-text-x-cron-job	CRONJOB	default	job-container	nvidia.com/gpu	4
smci-prom	llama2-text-x-job	JOB	default	job-container	nvidia.com/gpu	8
smci-prom	vectoradd-mig-ts	DEPLOYMENT	default	cuda-vector-add	nvidia.com/mig-2g.20gb	3
smci-prom	vectoradd-mig-ts	DEPLOYMENT	default	cuda-vector-add	nvidia.com/mig-2g.20gb	3
smci-prom	vectoradd-mig-ts	DEPLOYMENT	default	cuda-vector-add	nvidia.com/mig-2g.20gb	3

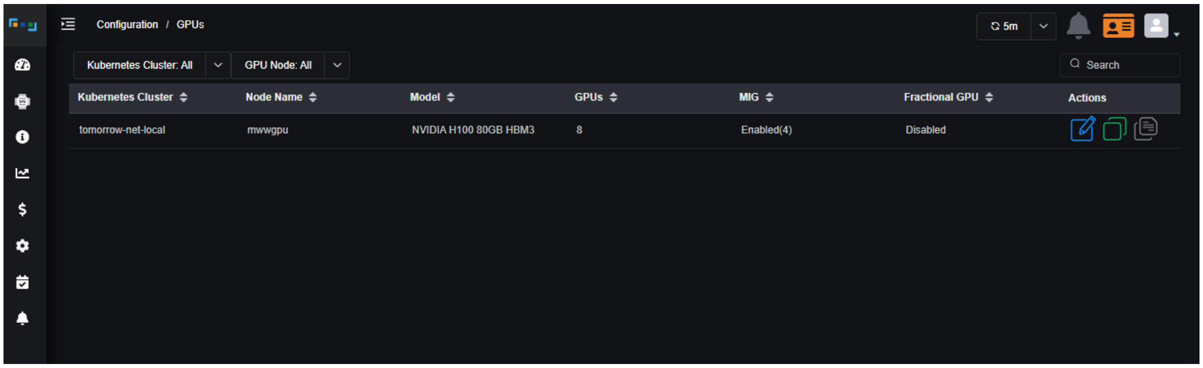
Total 44 ‹ 1 2 3 › 20/page

- *Inactive GPU Workloads* – This pages displays all inactive GPU workloads. To remove inactive workloads from the system permanently, click the “*Remove*” icon (⌫) next to an inactive workload entry.

Configuration – GPUs

Federator.ai GPU Booster supports *fractional* GPUs in Kubernetes. Administrators can use Federator.ai to enable fractional GPU mode for individual GPUs, allowing multiple workloads to share a single GPU. For example, one GPU workload can request 25% of an H100 GPU, while another requests 50% of the same GPU. This cabability increases the utilization of GPUs and accommodate more GPU workloads.

On the *Configuration / GPUs* page, a list of all GPU nodes, includingthe GPU models and the number of GPUs on each node, is displayed. It also shows whether Fractional GPU is enabled and the number of GPUs are configured with MIG (Multi-instance GPU).



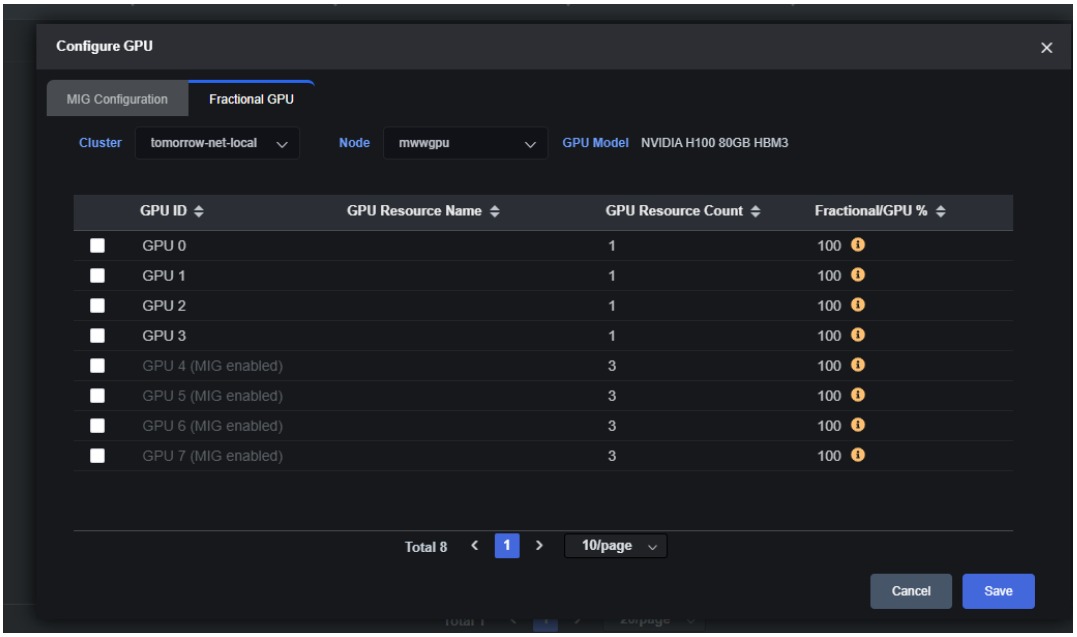
Kubernetes Cluster	Node Name	Model	GPUs	MIG	Fractional GPU	Actions
tomorrow-net-local	mwwgpu	NVIDIA H100 80GB HBM3	8	Enabled(4)	Disabled	✎ ⧉ 📋

Administrators can perform the following function when configure GPUs in a cluster.

Icon	Function
✎	Edit GPU settings in a GPU node.
⧉	Copy GPU settings in a GPU node.
📋	Paste the GPU settings from a GPU node to another GPU node of the same model and number of GPUs.

Enable/disable Fractional GPU

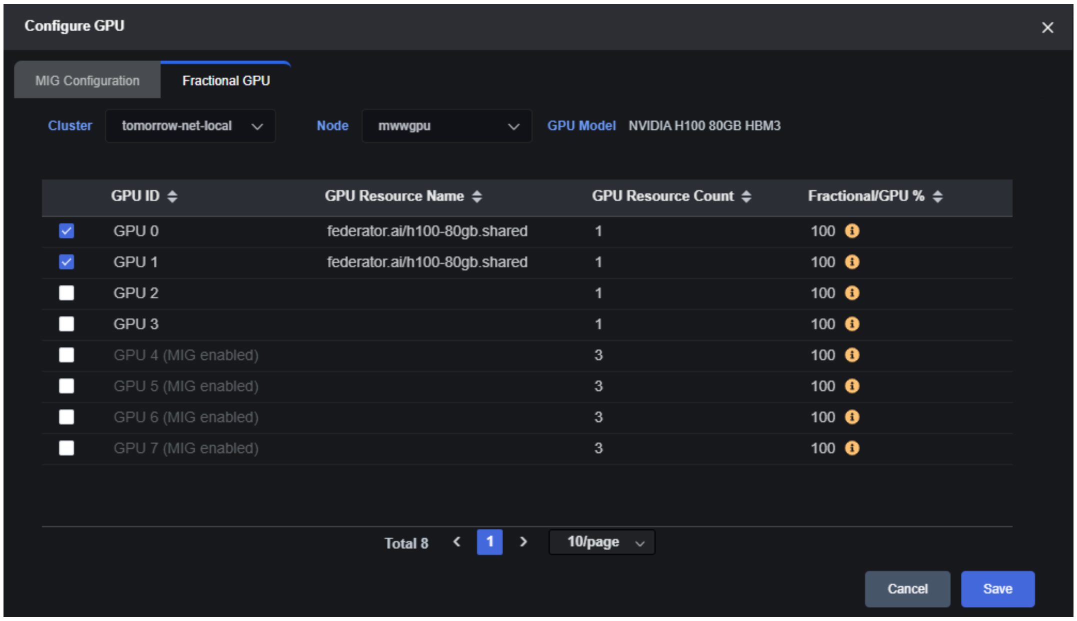
1. On the *Configuration / GPUs* page, click the *Edit* icon (✎) for the GPU node, a window displaying the current GPU configuration of the GPU node is shown.



Cluster: tomorrow-net-local | Node: mwwgpu | GPU Model: NVIDIA H100 80GB HBM3

	GPU ID	GPU Resource Name	GPU Resource Count	Fractional/GPU %
□	GPU 0		1	100
□	GPU 1		1	100
□	GPU 2		1	100
□	GPU 3		1	100
□	GPU 4 (MIG enabled)		3	100
□	GPU 5 (MIG enabled)		3	100
□	GPU 6 (MIG enabled)		3	100
□	GPU 7 (MIG enabled)		3	100

2. Check one or more GPUs to enable fractional GPU mode. In this example, this GPU node has 8 Nvidia H100 80GB GPUs with 4 (GPU 4-7) configured with MIG, each with 3 MIG partitions. Click on specific GPUs to enable fractional GPU feature.



Configure GPU ×

MIG Configuration | Fractional GPU

Cluster: tomorrow-net-local Node: mwwgpu GPU Model: NVIDIA H100 80GB HBM3

	GPU ID	GPU Resource Name	GPU Resource Count	Fractional/GPU %
☑	GPU 0	federator.ai/h100-80gb.shared	1	100
☑	GPU 1	federator.ai/h100-80gb.shared	1	100
□	GPU 2		1	100
□	GPU 3		1	100
□	GPU 4 (MIG enabled)		3	100
□	GPU 5 (MIG enabled)		3	100
□	GPU 6 (MIG enabled)		3	100
□	GPU 7 (MIG enabled)		3	100

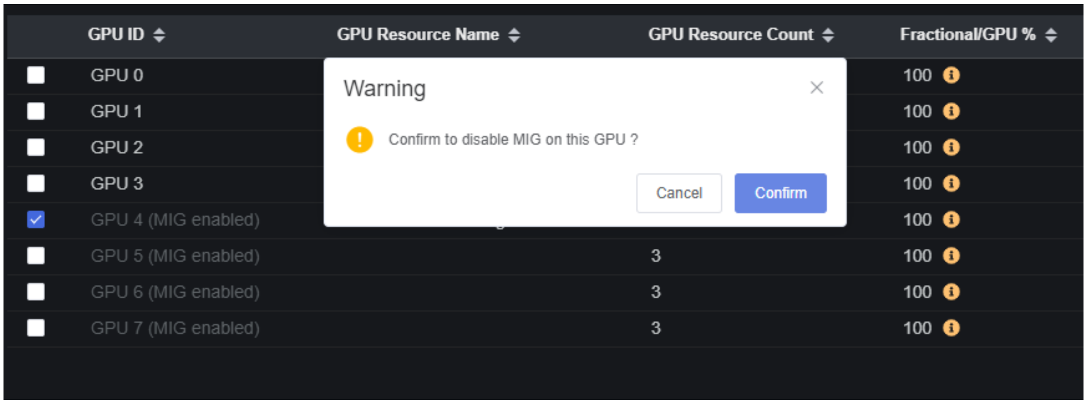
Total 8 < 1 > 10/page

Cancel Save

Note, 1% of a frational GPU refers to 1% of GPU memory resource. In the case of an H100 80GB GPU, 1% equates to 800MB of GPU memory. Therefore, a GPU workload requiring no more than 20GB of GPU memory should request 25% of a shared H100 GPU.

When a GPU is configured as a fractional GPU, Federator.ai GPU Booster creates a specific resource name that GPU workloads shoud use. In this example, the resource name "*federator.ai/h100-80gb.shared"* is generated. This naming convention indicates the GPU model and memory size, helping developers determine the percetange of the GPU resource to request for their workload.

A GPU cannot be configured with both MIG and fractional GPU at the same time. If an administrator chooses to enable fractional GPU on a GPU currently configured with MIG, a warning message will be displayed, confirming the need to disable MIG.



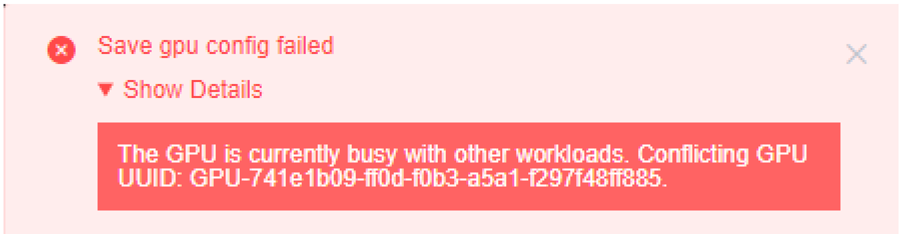
	GPU ID	GPU Resource Name	GPU Resource Count	Fractional/GPU %
□	GPU 0			100
□	GPU 1			100
□	GPU 2			100
□	GPU 3			100
☑	GPU 4 (MIG enabled)			100
□	GPU 5 (MIG enabled)		3	100
□	GPU 6 (MIG enabled)		3	100
□	GPU 7 (MIG enabled)		3	100

Warning ×

Confirm to disable MIG on this GPU ?

Cancel Confirm

3. Click *Save* to complete the GPU configuration. The configuration will take a couple of minutes to complete. Note: GPUs must free of workloads when configuration process. If there are any workloads currently running on the GPUs being configured, an error message will be displayed.



Save gpu config failed

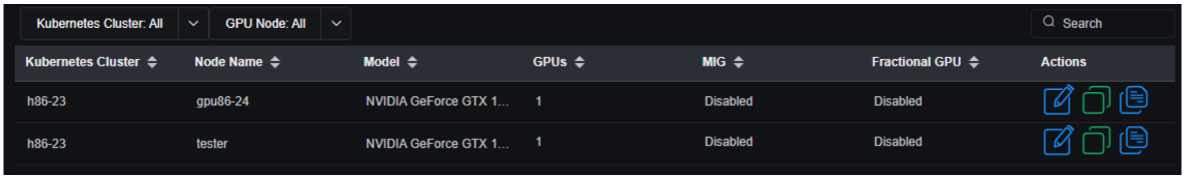
▼ Show Details

The GPU is currently busy with other workloads. Conflicting GPU UUID: GPU-741e1b09-ff0d-f0b3-a5a1-f297f48ff885.

Cloning GPU Configuration

It is possible to clone the configuration of a GPU node to one or more GPU nodes in the same cluster provided they have the same number of GPUs and the same GPU models. This features simplifies and speeds up the setup process of GPU configuration, especially in environments with mang GPU nodes.

1. On the *Configuration / GPUs* page, click the *Copy* icon (❐) to copy the GPU configuration of a GPU node.

   Kubernetes Cluster: All | GPU Node: All | Search

Kubernetes Cluster	Node Name	Model	GPUs	MIG	Fractional GPU	Actions
h86-23	gpu86-24	NVIDIA GeForce GTX 1...	1	Disabled	Disabled	✎ ❐ 📋
h86-23	tester	NVIDIA GeForce GTX 1...	1	Disabled	Disabled	✎ ❐ 📋

   A notification will be displayed on the upper right corner, indicating the copy acation is successful.

2. The *Paste* action icons (📋) for those GPU nodes that have the same GPU models and number GPUs in the same cluster will become active. Click on the *Paste* icon of those GPU nodes you want to copy, and you will be prompted to confirm the changes. Click *OK* to complete the action.

Using Fracational GPUs

When deciding the percentage of a GPU to be used for a workload, developers should estimate the GPU memory required by the workload in relative to the total memory of the target GPU. For instance, if a workload requires up to 20GB of GPU memory and targets an H100-80GB GPU, it should request 25% of the H100 fractional GPU. Please see the sample YAML file for a workload below.

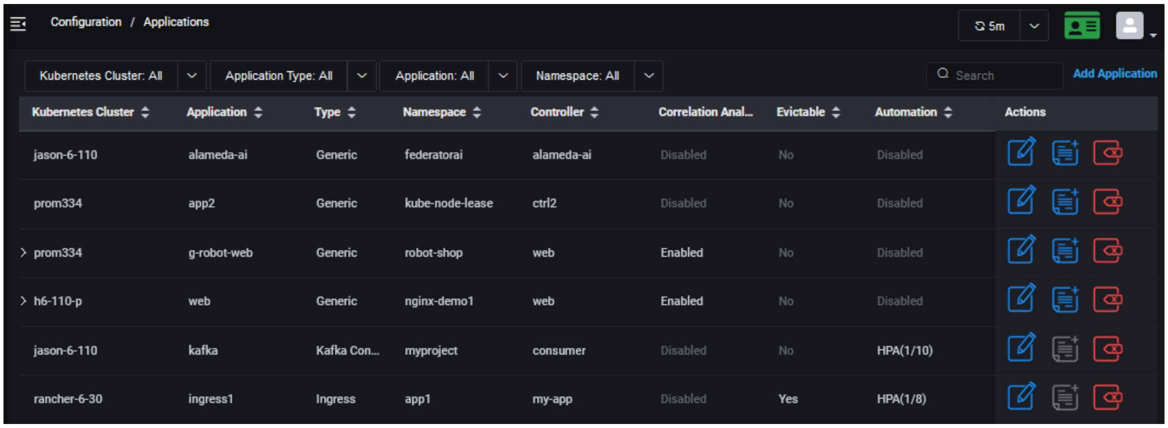
```
apiVersion: batch/v1
kind: Job
metadata:
labels:
   app: gpu-test-tensorflow-job
name: gpu-test-tensorflow-job
spec:
parallelism: 1
template:
   spec:
     containers:
     - name: gpu-test-tensorflow-job
       image: repo.prophetservice.com/tmp/gpu-test-tensorflow
       resources:
         limits:
           federator.ai/h100-80gb.shared: 25
     restartPolicy: Never
```

It's important to note that the fraction of the GPU resource requested pertains to GPU memory, not GPU compute power. Even if a workload requests 25% of a fractional GPU, it might use up to 50% of the GPU's compute power.

Federator.ai GPU Booster monitors workloads using fractional GPU resources. If a workload exceeds its requested fraction of GPU memory, it will be terminated to ensure that other workloads sharing the same GPU can continue to operate smoothly.

Configuration – Applications

The *Applications* page displays all Kubernetes applications being monitored by Federator.ai GPU Booster and allows you to manage them.



Kubernetes Cluster	Application	Type	Namespace	Controller	Correlation Anal...	Evictable	Automation	Actions
jason-6-110	alameda-ai	Generic	federatorai	alameda-ai	Disabled	No	Disabled	✎ ☰ ⌫
prom334	app2	Generic	kube-node-lease	ctrl2	Disabled	No	Disabled	✎ ☰ ⌫
› prom334	g-robot-web	Generic	robot-shop	web	Enabled	No	Disabled	✎ ☰ ⌫
› h6-110-p	web	Generic	nginx-demo1	web	Enabled	No	Disabled	✎ ☰ ⌫
jason-6-110	kafka	Kafka Con...	myproject	consumer	Disabled	No	HPA(1/10)	✎ ☰ ⌫
rancher-6-30	ingress1	Ingress	app1	my-app	Disabled	Yes	HPA(1/8)	✎ ☰ ⌫

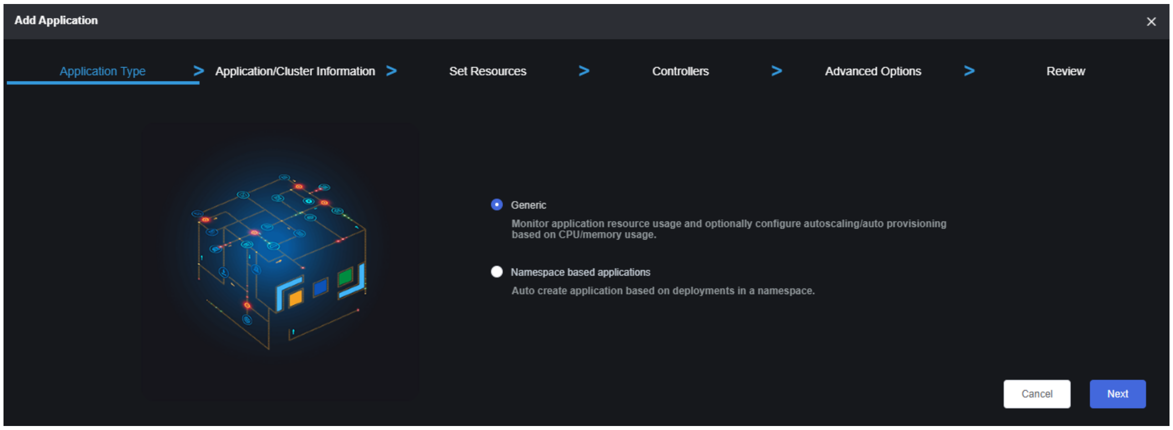
In addition to adding an application, you can perform the following functions for existing applications:

Icon	Function
✎	Edit an application.
☰	View the KEDA or resource integration script.
⌫	Remove an application.

Add an Application

Applications are configured via a wizard.

1. On the *Configuration / Applications* page, click *Add Application* and specify the type of application you are adding.



● Generic
Monitor application resource usage and optionally configure autoscaling/auto provisioning based on CPU/memory usage.

○ Namespace based applications
Auto create application based on deployments in a namespace.

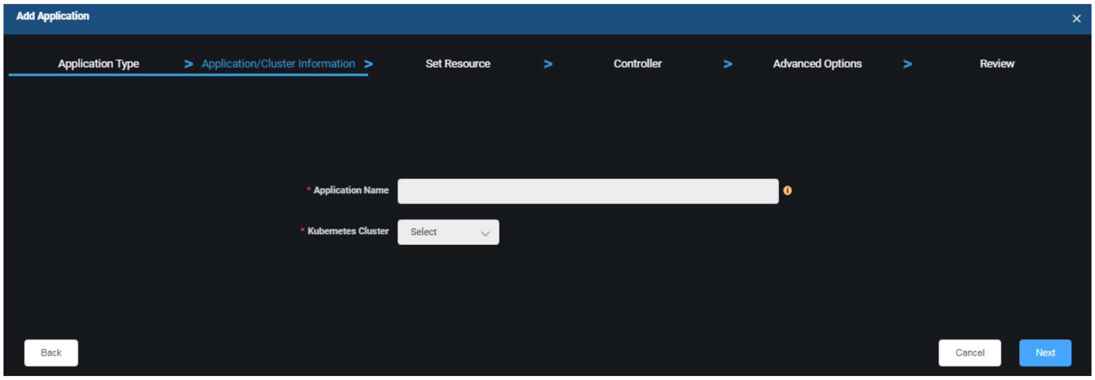
The application wizard offers two ways of configure an application to be monitored:

- configuring an application by selecting specific controllers/deployments from one or multiple namespaces; or
- configuring an application by selecting all controllers/deployments from a specific namespace.

2. Click *Next* to continue.

Generic Kubernetes Application

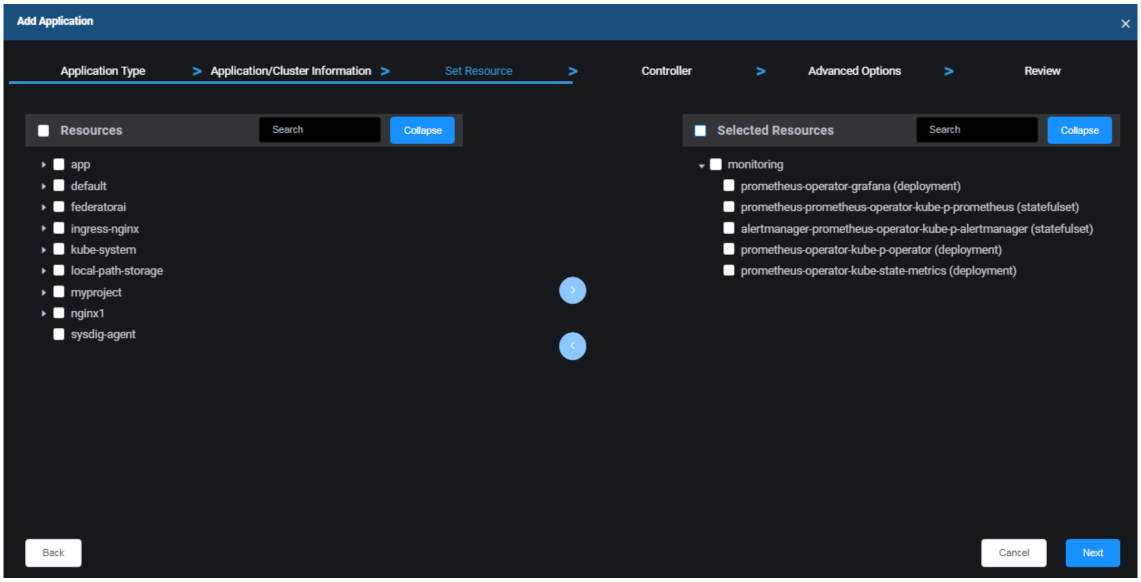
1. Specify an application name and Kubernetes cluster and then click *Next* to continue.



Application Name - Name of the application you want to manage. The name must be a maximum of 253 characters, contain only lowercase alphanumeric characters, "-", or ".", and start and end with an alphanumeric character. Note, an application is not a native Kubernetes object. You will define the controllers that are part of your application later.

Kubernetes Cluster – Select an existing cluster where this application resides.

2. Select one or more namespaces and controllers and then click *Next* to continue.



□ app
□ default
□ federatorai
□ ingress-nginx
□ kube-system
□ local-path-storage
□ myproject
□ nginx1
□ sysdig-agent

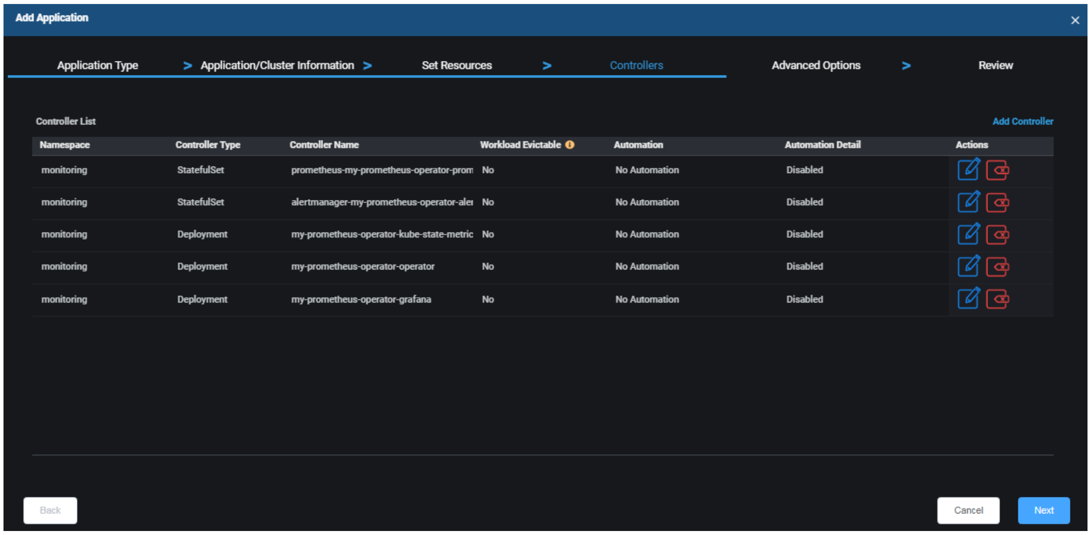
□ monitoring
□ prometheus-operator-grafana (deployment)
□ prometheus-prometheus-operator-kube-p-prometheus (statefulset)
□ alertmanager-prometheus-operator-kube-p-alertmanager (statefulset)
□ prometheus-operator-kube-p-operator (deployment)
□ prometheus-operator-kube-state-metrics (deployment)

Federator.ai GPU Booster automatically discovers all the namespaces and the controllers of the selected cluster

Expand a namespace to see the controllers below. By default, when you select a namespace, all of the controllers below are included. Click the *Add* icon › to move resources to the *Selected Resources* list. If you do not want to include a controller, you can de-select it. If it is already included in the *Selected Resources* list, select it and click the *Delete* icon ‹.

The controllers included under *Selected Resources* will be included the application being created.

3. Configure controllers. When done, click *Next* to continue.


Add Application
×
Application Type > Application/Cluster Information > Set Resources > Controllers Advanced Options > Review
Back
Cancel
Next

Controller List

Add Controller

Namespace	Controller Type	Controller Name	Workload Evictable	Automation	Automation Detail	Actions
monitoring	StatefulSet	prometheus-my-prometheus-operator-prom	No	No Automation	Disabled	
monitoring	StatefulSet	alertmanager-my-prometheus-operator-aler	No	No Automation	Disabled	
monitoring	Deployment	my-prometheus-operator-kube-state-metric	No	No Automation	Disabled	
monitoring	Deployment	my-prometheus-operator-operator	No	No Automation	Disabled	
monitoring	Deployment	my-prometheus-operator-grafana	No	No Automation	Disabled	

Each controller selected on the previous screen is listed with the default information. Click the *Edit Controller* icon to change information.

Controller Workload Evictable – Indicates if the controller can be interrupted if the node is shut down by a public cloud service provider. Evictable controllers are good candidates to be deployed in Spot instances.

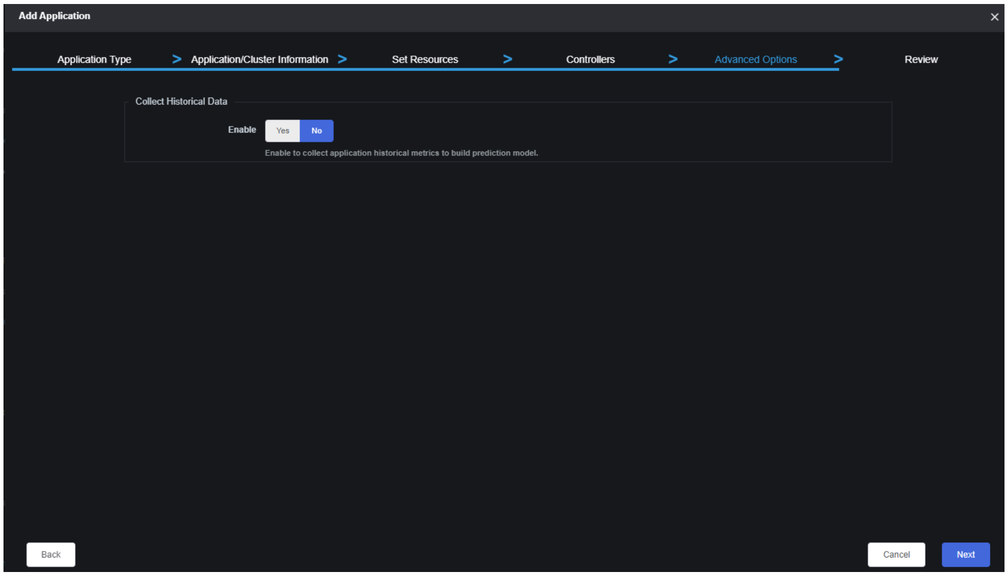
Automation – Indicates the type of automation used:

- *No Automation* –Federator.ai GPU Booster monitors resource usage only.
- *Auto Provisioning by Profile* – When selected, auto provisioning is based on the selected profile. This option can only be selected if Federator.ai GPU Booster is installed in this Kubernetes cluster.

  The resource integration script for auto provisioning is provided by the system (accessible from the *Integration Script* icon under *Actions*). When a script is run in a Kubernetes cluster, it queries Federator.ai GPU Booster for the most recent recommendations for this controller and applies the resource recommendations. Refer to Auto Provisioning Scripts for more information.

Automation Detail –If *Auto Provisioning by Profile* is selected, select a profile. If needed, you can create a profile.

4. Configure advanced options. When done, click *Next* to continue.



Collect Historical Data

Enable: Yes [✓ No]

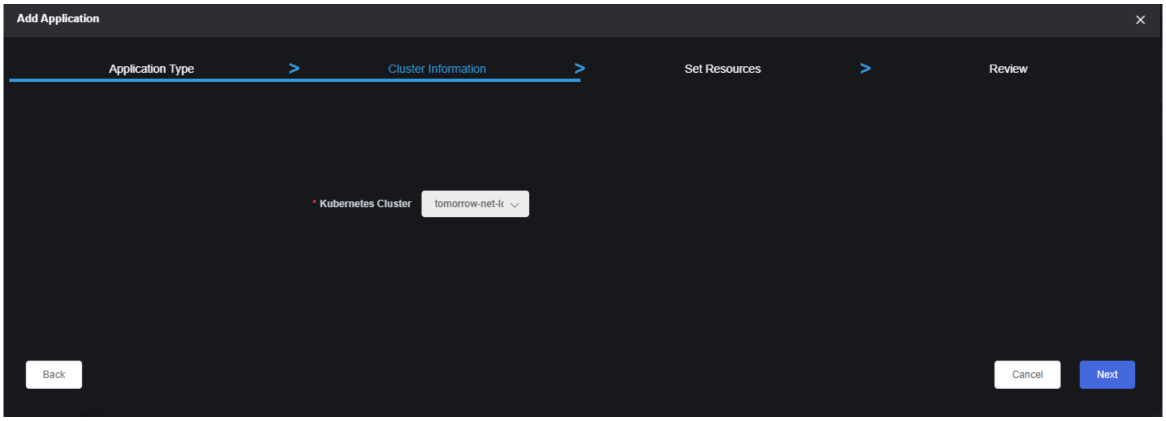
Enable to collect application historical metrics to build prediction model.

Collect Historical Data – Specify if you want the system to collect up to three months' worth of historical data for existing controllers of this application. This will enable weekly and monthly predictions, recommendations, and cost analysis for newly added applications. If less than three months' worth of data exists, the system will collect the maximum data that is available. Typically, collection of historical data will complete in about 2-3 hours. If you need to pause collection, you can edit application settings.

5. Review all information and click *Create* to create the application.

Namespace Based Applications

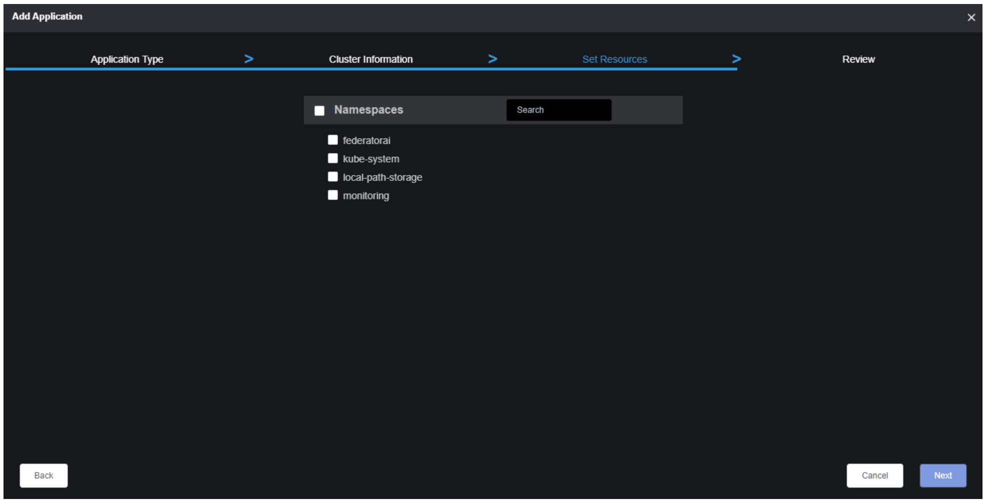
1. Select *Namespace based applications* and then click *Next* to continue.
2. Specify Kubernetes cluster and then click *Next* to continue.



* Kubernetes Cluster: tomorrow-net-lc

Kubernetes Cluster – Select an existing cluster where this application resides.

3. Specify the namespace and then click *Next* to continue.



□ Namespaces

□ federatorai
□ kube-system
□ local-path-storage
□ monitoring

Namspaces – Select the namespace from which the application is based on. Note, the name of the application will be concatenation of namespace name and cluster name.

4. Review all information and click *Create* to create the application.

Manage Applications

You can do the following from the *Configuration / Applications* page:

- Edit application settings – Using the application wizard, you can add, edit, or remove a controller (generic application). For historical data collection of generic applications, you can:
  - Start the collection of up to three months' worth of historical data (from the current time) for existing controllers of the application, if it was not enabled when the application was added.
  - See the status of data collection, including the time period collected.
  - Pause/resume historical data collection that is in progress.
- Remove an application. To do this, click the *Remove Application* icon and confirm the removal.

Related topics:

Auto Provisioning

Terminology

Search/Sort Information in Tables

Application Analysis

Configuration – Auto Provisioning

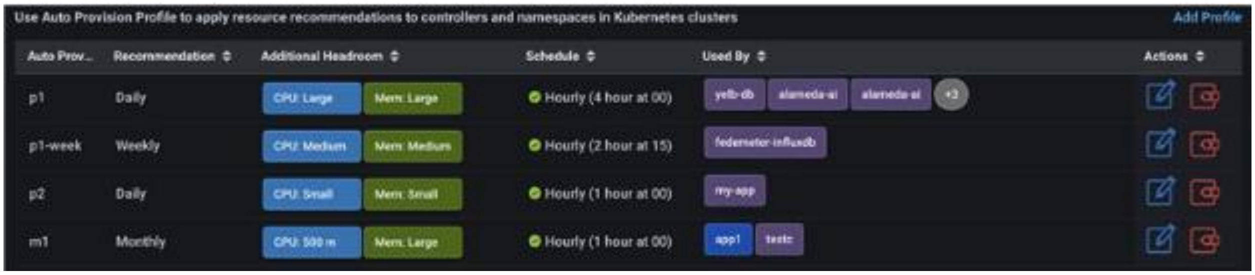
Federator.ai GPU Booster predicts CPU and memory usage for each application controller and application namespace in Kubernetes clusters and makes recommendations for the optimal amount of resources. Auto provisioning can automatically deploy resource recommendations to controllers and namespaces for generic applications based on a pre-defined profile.

An auto provisioning profile defines the conditions under which the resource recommendations will be automatically applied. It defines which recommendations to use (daily, weekly, or monthly), any adjustments to make on top of the system recommendations, and the schedule for when the resource recommendations should be applied.

If Federator.ai GPU Booster is installed in the same Kubernetes cluster as the application, you can assign auto provisioning profiles to controllers via the *Configuration / Applications* page or assign profiles to namespaces via the *Configuration / Clusters* page.

For remote clusters, you can copy a resource integration script to the remote cluster in order to run auto provisioning. Refer to Auto Provisioning Scripts below.

The *Auto Provisioning* page displays all the existing profiles and allows you to add, edit, and remove profiles.



Use Auto Provision Profile to apply resource recommendations to controllers and namespaces in Kubernetes clusters

Auto Prov...	Recommendation ⇕	Additional Headroom ⇕	Schedule ⇕	Used By ⇕	Actions ⇕
p1	Daily	CPU: Large Mem: Large	● Hourly (4 hour at 00)	yelb-db alameda-ai alameda-ai +3	
p1-week	Weekly	CPU: Medium Mem: Medium	● Hourly (2 hour at 15)	federator-influxdb	
p2	Daily	CPU: Small Mem: Small	● Hourly (1 hour at 00)	my-app	
m1	Monthly	CPU: 500 m Mem: Large	● Hourly (1 hour at 00)	app1 testc	

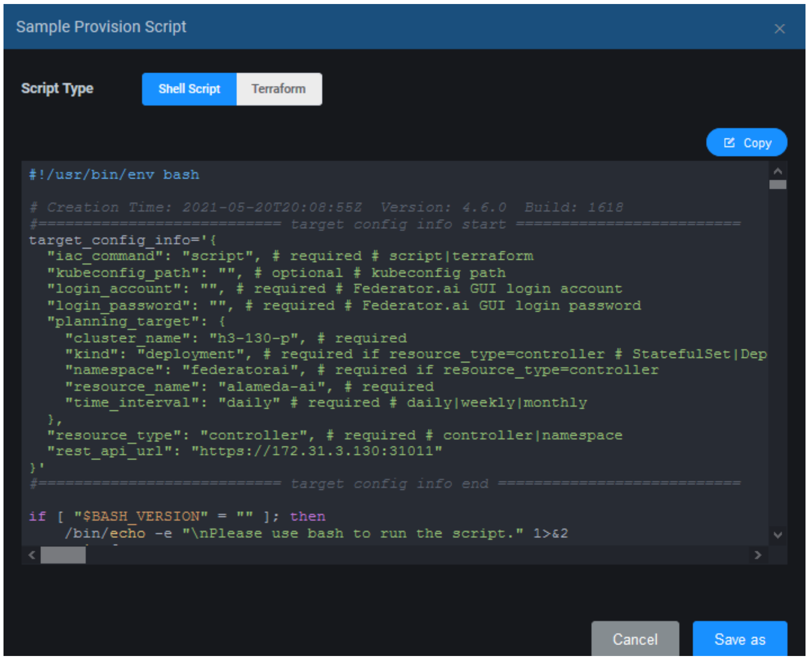
For each profile, you will see the frequency of the recommendations, CPU and memory adjustments, auto provisioning schedule, and which controllers and namespaces are using the profile. Purple represents a controller and blue represents a namespace.

Auto Provisioning Scripts

When you create an auto provisioning profile, the system generates a *resource integration script* that contains all the conditions in the profile.

For remote clusters, you can copy a resource integration script to the remote cluster in order to run auto provisioning. You can use a script associated with an auto provisioning profile or you can use the generic provisioning script provided by the system. This generic script uses system recommendations and does not have any adjustments or boundary (min/max) settings. When a resource integration script is run in a Kubernetes cluster, it queries Federator.ai GPU Booster for the most recent recommendations and applies them to a controller or a namespace.

You can find the scripts, save the scripts locally, and copy these scripts via the following pages: *Configuration / Applications, Configuration / Clusters* (when you edit a namespace), or *Planning / Kubernetes Workload Prediction* (when you are viewing information for controllers or namespaces).

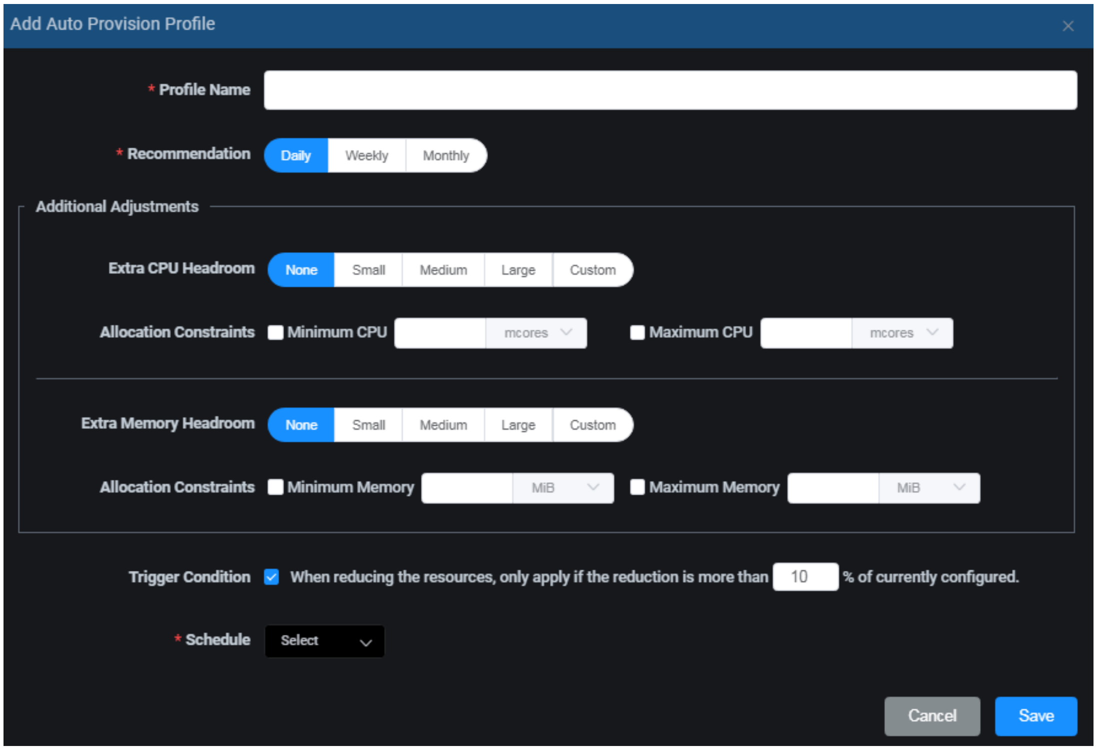


```
#!/usr/bin/env bash

# Creation Time: 2021-05-20T20:08:55Z  Version: 4.6.0  Build: 1618
#=========================== target config info start =========================
target_config_info='{
  "iac_command": "script", # required # script|terraform
  "kubeconfig_path": "", # optional # kubeconfig path
  "login_account": "", # required # Federator.ai GUI login account
  "login_password": "", # required # Federator.ai GUI login password
  "planning_target": {
    "cluster_name": "h3-130-p", # required
    "kind": "deployment", # required if resource_type=controller # StatefulSet|Dep
    "namespace": "federatorai", # required if resource_type=controller
    "resource_name": "alameda-ai", # required
    "time_interval": "daily" # required # daily|weekly|monthly
  },
  "resource_type": "controller", # required # controller|namespace
  "rest_api_url": "https://172.31.3.130:31011"
}'
#=========================== target config info end ===========================

if [ "$BASH_VERSION" = "" ]; then
    /bin/echo -e "\nPlease use bash to run the script." 1>&2
```

You can select a shell script to be run in a Kubernetes cluster where the application is run. For Terraform integration, you can select the Terraform script.

Add a Profile

1. On the *Configuration / Auto Provisioning* page, click *Add Profile*.



Add Auto Provision Profile

* Profile Name

* Recommendation: **Daily** | Weekly | Monthly

Additional Adjustments

Extra CPU Headroom: **None** | Small | Medium | Large | Custom

Allocation Constraints: □ Minimum CPU mcores □ Maximum CPU mcores

Extra Memory Headroom: **None** | Small | Medium | Large | Custom

Allocation Constraints: □ Minimum Memory MiB □ Maximum Memory MiB

Trigger Condition: ☑ When reducing the resources, only apply if the reduction is more than 10 % of currently configured.

* Schedule: Select

Cancel | Save

Profile Name – Specify a name for the profile.

Recommendation – Specify which system recommendations to use (daily, weekly, or monthly).

Adjustments: Extra Headroom – If desired, specify any adjustments to make on top of the system recommendations for CPU and memory. *Small* means that the adjustment is 10% more than the recommendation, *Medium* means 20%, and *Large* means 30%. You can also specify a custom adjustment (millicores or percent for CPU; MB, GB, or percent for memory).

Adjustments: *Allocation Constraints* – If desired, specify minimum and maximum limits for CPU and memory. Resources will not be deployed if above or below these boundaries.

Trigger Condition – Recommendations may trigger a reduction of resources. If desired, specify a percentage to limit reduction of resources that are currently configured. The application will be restarted when resources are reduced. Therefore, you should be careful not to make the difference too small, causing frequent application restarts.

Schedule – Specify when the resource recommendations should be applied. If you are using *Daily* recommendations, you may want to apply changes hourly, daily, or automatically at midnight. For *Weekly* recommendations, you may want to apply changes hourly, daily, weekly, or automatically (12:00 a.m. Sunday). For *Monthly* recommendations, you may want to apply changes daily, weekly, monthly, or automatically (12:00 a.m. on the first day of the month). Note that all times are local.

2. Click *Save* when you are done.

Manage Profiles

You can do the following from the *Configuration / Auto Provisioning* page:

- Edit a profile. To do this, click the *Edit Profile* icon.
- Remove a profile. You can only remove a profile if it is not being used by a namespace or controller. To do this, click the *Remove Profile* icon.

Related topic:

Applications
Terminology
Search/Sort Information in Tables
Configure Applications

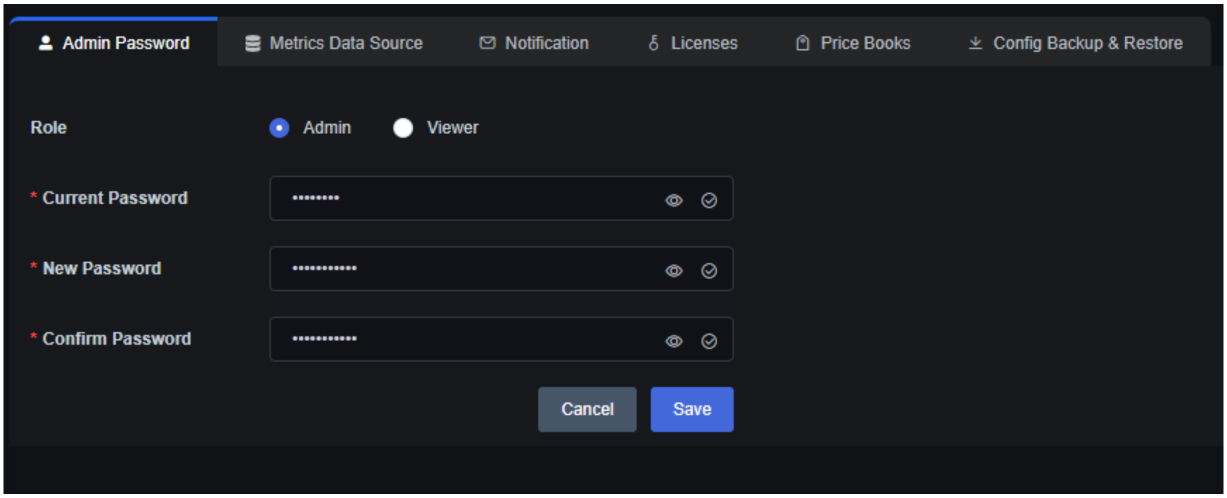
Configuration – System Settings

The *System Settings* page has tabs that allow you to:

- Change the admin password.
- Update metrics data source information.
- Set system notification.
- Manage the system license.
- Set policies and manage price books.
- Backup and restore the system configuration.

Admin Password

To access this page, select *Configuration, System Settings, Admin Password* tab. There are two built-in administration accounts: *admin* and *viewer.* Account *admin* can perform both read/write operations on Federator.ai GPU Booster administration portal. Account viewer is restricted to read-only operations. Choose the correct *role* to change the password of either *admin* or *viewer* account. You must know the current password and *New Password* must match *Confirm Password*.



Role ● Admin ○ Viewer

* Current Password ••••••••

* New Password •••••••••••

* Confirm Password •••••••••••

Metrics Data Source

The *Metrics Data Source* page allows you to set the authentication values that are needed by clusters to access metrics from different data sources. To access this page, select *Configuration, System Settings, Metrics Data Source* tab. You will need to select a cluster type, data source, and cluster.

The current supported data sources is Prometheus.

- For the Prometheus open-source monitoring system, the *URL* is required but the *Token* is optional.

When you are done, click *Test Connection* to confirm that all information is correct.

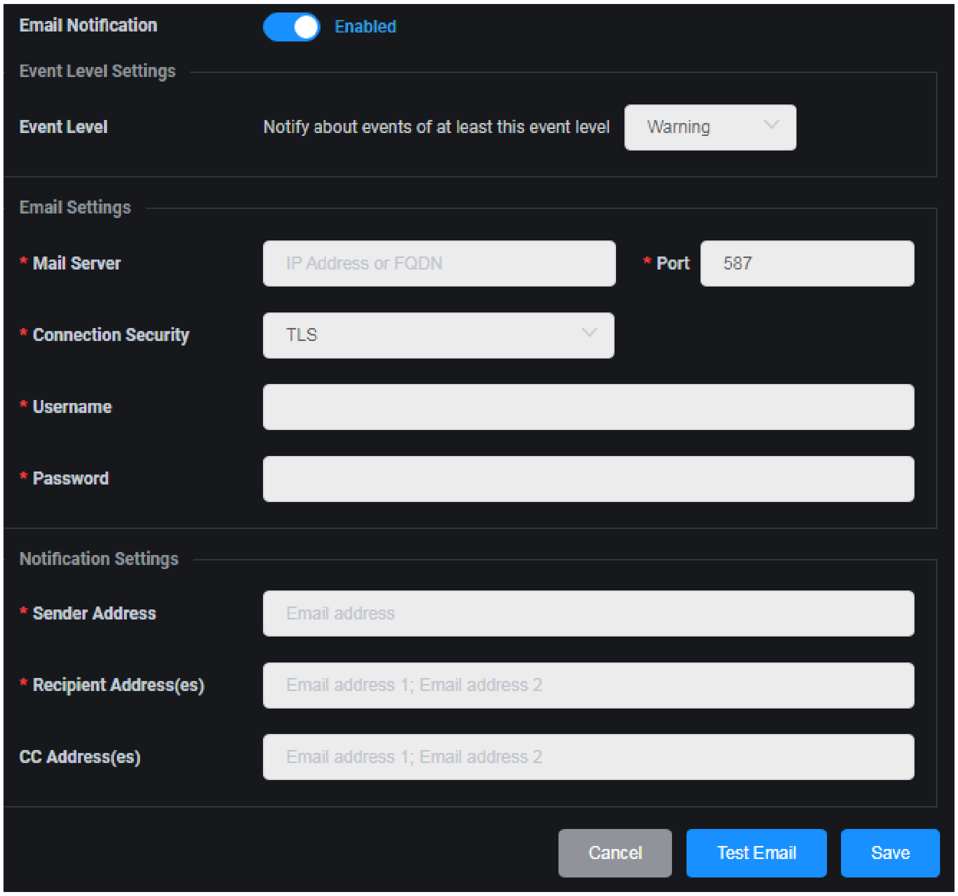
Notification

The *Notification* page allows you to configure email notifications to administrators when system errors and fatal issues occur. To access this page, select *Configuration, System Settings, Notification* tab.

Enable Notification

Follow the steps below to enable notification:

1. Toggle the *Email Notification* icon to *Enabled*.



Email Notification: Enabled

Event Level Settings

Event Level: Notify about events of at least this event level: Warning

Email Settings

* Mail Server: IP Address or FQDN   * Port: 587

* Connection Security: TLS

* Username

* Password

Notification Settings

* Sender Address: Email address

* Recipient Address(es): Email address 1; Email address 2

CC Address(es): Email address 1; Email address 2

Event Level – Select the minimum event level that should trigger notification. Higher levels will also trigger an email. For example, if you select *Error*, fatal events will also trigger an email.

Mail Server - Specify the mail server that should be used to send notification emails.

Port - Specify the mail server port that should be used.

Connection Security – Specify the protocol used by the mail server to secure email transmissions.

Username/Password - Specify the user account that will be used to log into the mail server.

Sender Address - Specify the email account that will be used in the “From” field of emails that are sent.

Recipient Address(es) - Specify the email address of the account that will receive emails. This will be used in the “To” field of emails. Separate multiple email addresses with semicolons.

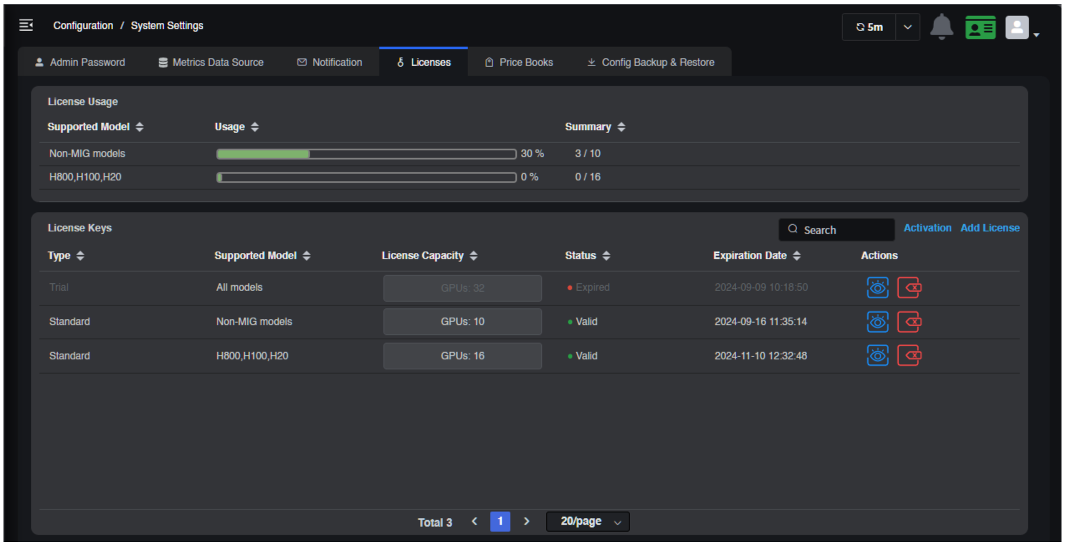
CC Address(es) - Specify any other email accounts that should receive emails. Separate multiple email addresses with semicolons.

2. Click *Test Email* to confirm that all information is correct.

3. Once the test emails are received, click *Save*.

Licenses

To access this page, select *Configuration, System Settings, Licenses* tab. The *Licenses* page displays your current system licenses, including license type, capacity, usage,status, and expiration date.   Federator.ai GPU Booster licenses are based on GPU models. Users need to obtain the correct license for each different GPU models in the system.



License Usage

Supported Model	Usage	Summary
Non-MIG models	30 %	3 / 10
H800,H100,H20	0 %	0 / 16

License Keys

Type	Supported Model	License Capacity	Status	Expiration Date	Actions
Trial	All models	GPUs: 32	● Expired	2024-09-09 10:18:50	
Standard	Non-MIG models	GPUs: 10	● Valid	2024-09-16 11:35:14	
Standard	H800,H100,H20	GPUs: 16	● Valid	2024-11-10 12:32:48	

There are two license types, *Trial*, and *Standard*. By default, a *Trial* license supporting upto 32 GPUs for 30 days is automatically applied when Federator.ai GPU Booster is installed. A trial license is applicable to all GPU models. Additional *Trial* license may be provided during Federator.ai GPU Booster evaluation if needed. Once a *Standard* license applied, the *Trial* license expires.

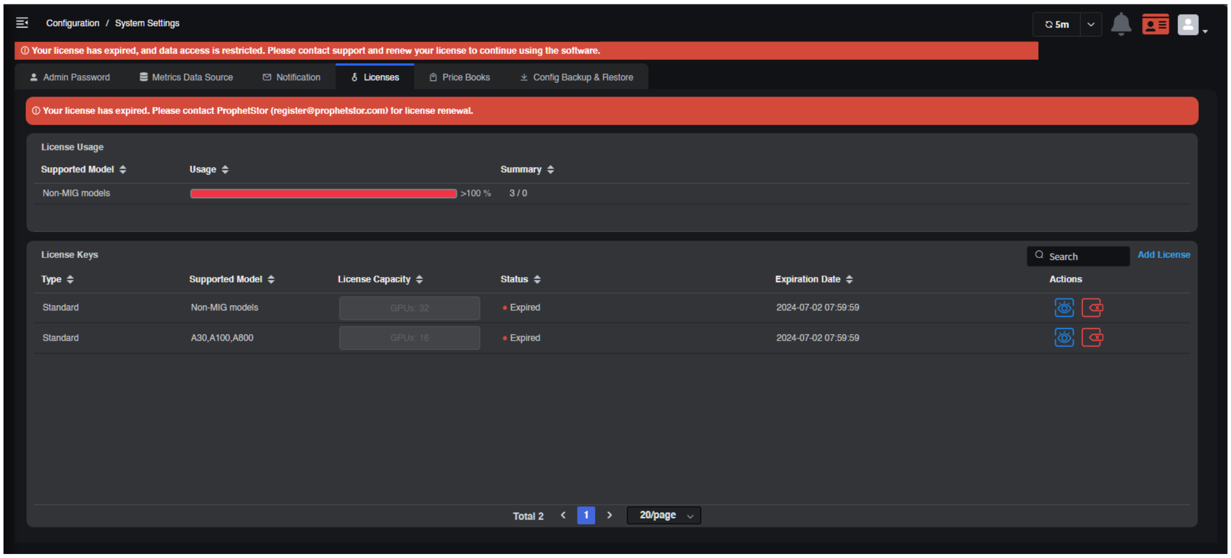
If you reach the number of licensed resources, there is a 30-day grace period, at which time the system prevents you from adding resources and some product functions, such as GPU management, predictions, recommendations, and cost analysis will stop. However, data collection will continue in the background until additional license capacity is purchased.

A *Standard* license must be activated within 30 days after installation. However, a license with a status of *Pending Activation* will still function as a valid license during that period. The status will be *Valid* once the license is activated.

The *Expiration Date* shows when the license expires. If the expiration of a license results in the system exceeding the license limit, predictions, recommendations, cost analysis, and GPU management will not work.

License Expiration

- Trial licenses – Expire on the expiration date displayed. A warning will be displayed when a trial license is expiring within 7 days. There is no grace period for an expired trial license. A trial license will be immediately expired when a standard license is added to the system.
- Standard licenses – Expire on the expiration date displayed. A warning will be displayed when a standard license is expiring within 7 days and will continue during the 30-day grace period after expiration. If a standard license expires and a remaining license does not have enough capacity, regular UI functions will stop.



ⓘ Your license has expired, and data access is restricted. Please contact support and renew your license to continue using the software.

ⓘ Your license has expired. Please contact ProphetStor (register@prophetstor.com) for license renewal.

License Usage

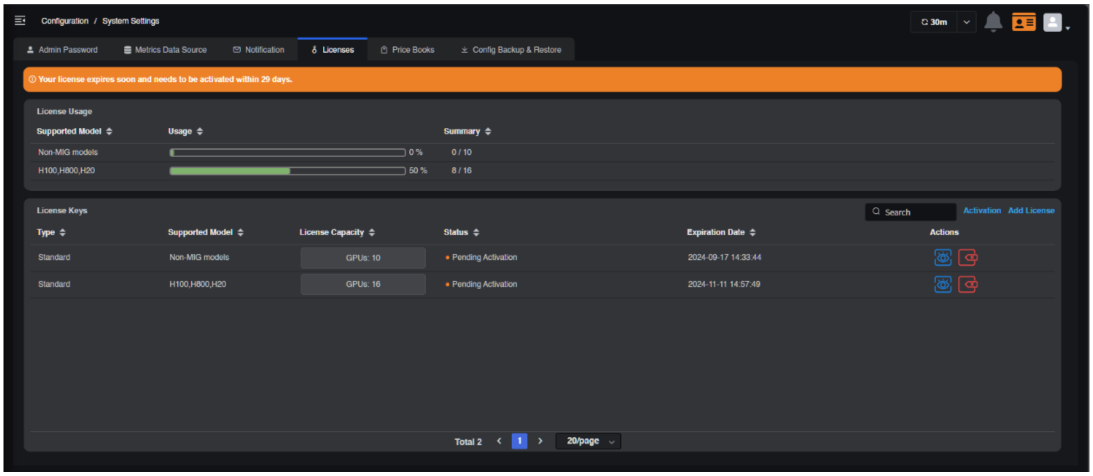
Supported Model	Usage	Summary
Non-MIG models	>100 %	3 / 0

License Keys

Type	Supported Model	License Capacity	Status	Expiration Date	Actions
Standard	Non-MIG models	GPUs: 32	● Expired	2024-07-02 07:59:59	
Standard	A30,A100,A800	GPUs: 16	● Expired	2024-07-02 07:59:59	

License Limits and Grace Periods

- A standard license is in *Pending Actication* state after it is first installed. It needs to be activated within 30 days. A warning message will be displayed to remind users to activate a standard license.



Your license expires soon and needs to be activated within 29 days.

License Usage

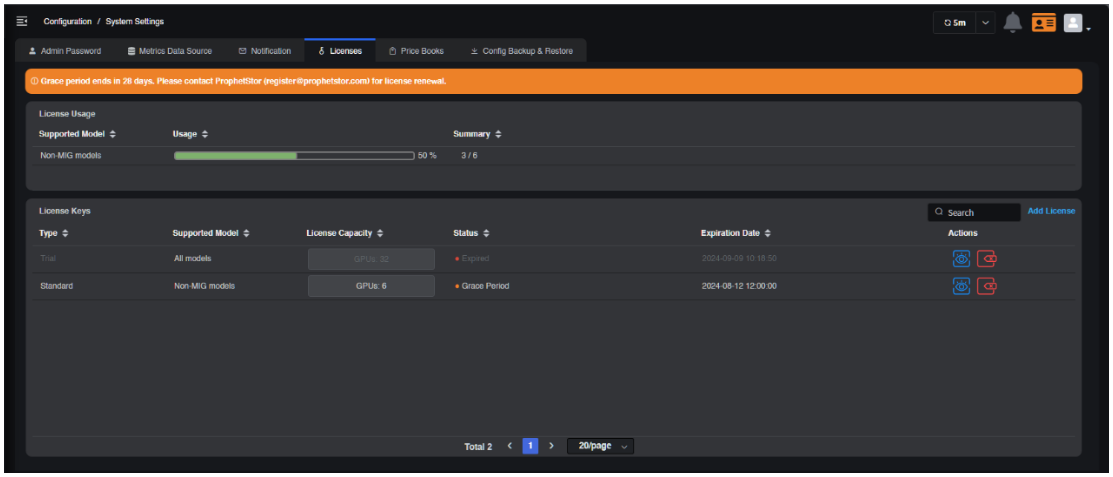
Supported Model	Usage	Summary
Non-MIG models	0 %	0 / 10
H100,H800,H20	50 %	8 / 16

License Keys

Type	Supported Model	License Capacity	Status	Expiration Date	Actions
Standard	Non-MIG models	GPUs: 10	Pending Activation	2024-09-17 14:33:44	👁 ⌫
Standard	H100,H800,H20	GPUs: 16	Pending Activation	2024-11-11 14:57:49	👁 ⌫

- If new GPU nodes are added to a cluster and there is not enough license capacity for new GPUs , there will be a 30-day grace period. After the grace period, normal UI functions will stop.

- There is a 30-day grace period after expiration of a standard license.



Grace period ends in 28 days. Please contact ProphetStor (register@prophetstor.com) for license renewal.

License Usage

Supported Model	Usage	Summary
Non-MIG models	50 %	3 / 6

License Keys

Type	Supported Model	License Capacity	Status	Expiration Date	Actions
Trial	All models	GPUs: 32	Expired	2024-09-09 10:18:50	👁 ⌫
Standard	Non-MIG models	GPUs: 6	Grace Period	2024-08-12 12:00:00	👁 ⌫

-

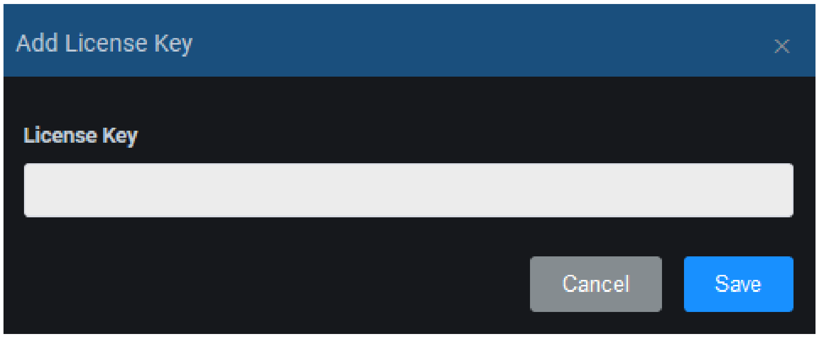
In addition to adding and activating licenses, you can perform the following functions from the *Licenses* page:

Icon	Function
👁	Show a license key.
⌫	Remove a license key.

Add a License

Follow the steps below to add and register a license:

1. On the *Licenses* page, click *Add License* and enter the license key.



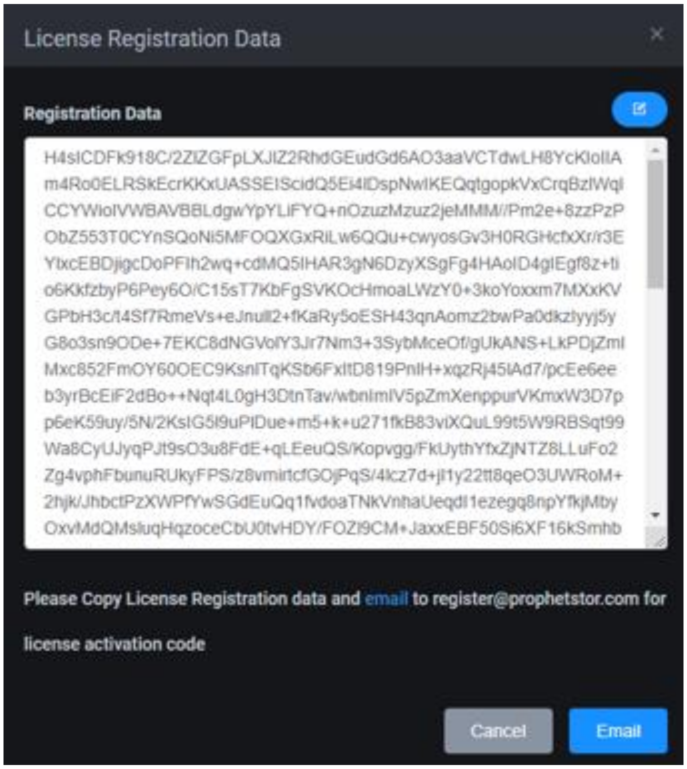
Add License Key

License Key

2. Click *Add*.

   A trial license is valid immediately after it is added. A standard license needs to be registered in order to be activated.

3. Email the registration data to register@prophetstor.com for a license activation code.



License Registration Data

Registration Data

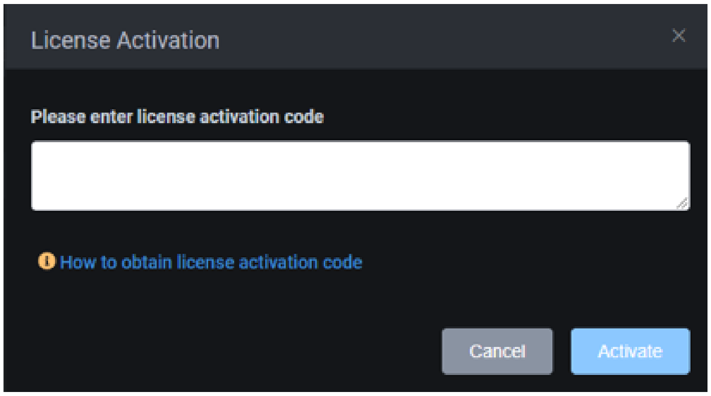
```
H4sICDFk918C/2ZlZGFpLXJlZ2RhdGEudGd6AO3aaVCTdwLH8YcKIoIIA
m4Ro0ELRSkEcrKKxUASSEIScidQ5Ei4IDspNwIKEQqtgopkVxCrqBzIWqI
CCYWIoIVWBAVBBLdgwYpYLiFYQ+nOzuzMzuz2jeMMM//Pm2e+8zzPzP
ObZ553T0CYnSQoNi5MFOQXGxRiLw6QQu+cwyosGv3H0RGHcfxXr/r3E
YlxcEBDjigcDoPFlh2wq+cdMQ5IHAR3gN6DzyXSgFg4HAoID4gIEgf8z+ti
o6KkfzbyP6Pey6O/C15sT7KbFgSVKOcHmoaLWzY0+3koYoxxm7MXxKV
GPbH3c/t4Sf7RmeVs+eJnull2+fKaRy5oESH43qnAomz2bwPa0dkzIyyj5y
G8o3sn9ODe+7EKC8dNGVolY3Jr7Nm3+3SybMceOf/gUkANS+LkPDjZmI
Mxc852FmOY60OEC9KsnITqKSb6FxItD819PnIH+xqzRj45IAd7/pcEe6ee
b3yrBcEiF2dBo++Nqt4L0gH3DtnTav/wbnImIV5pZmXenppurVKmxW3D7p
p6eK59uy/5N/2KsIG5I9uPIDue+m5+k+u271fkB83viXQuL99t5W9RBSqt99
Wa8CyUJyqPJt9sO3u8FdE+qLEeuQS/Kopvgg/FkUythYfxZjNTZ8LLuFo2
Zg4vphFbunuRUkyFPS/z8vmirtcfGOjPqS/4kz7d+jl1y22tt8qeO3UWRoM+
2hjk/JhbctPzXWPfYwSGdEuQq1fvdoaTNkVnhaUeqdl1ezegq8npYfkjMby
OxvMdQMsluqHqzoceCbU0tvHDY/FOZI9CM+JaxxEBF50Si6XF16kSmhb
```

Please Copy License Registration data and email to register@prophetstor.com for license activation code

   Click the *Email* button to launch your email program. Click the *Copy* icon to copy the registration text and paste it into your email before sending.

   The license status will now say *Pending Activation*. It must be activated within 30 days.

4. Once you have received the license activation code, click the *Activate License* icon and paste the activation code.


License Activation
×
Please enter license activation code
How to obtain license activation code
Cancel
Activate

5. Click *Activate*.

Manage Licenses

You can do the following from the *Configuration / Licenses* page:

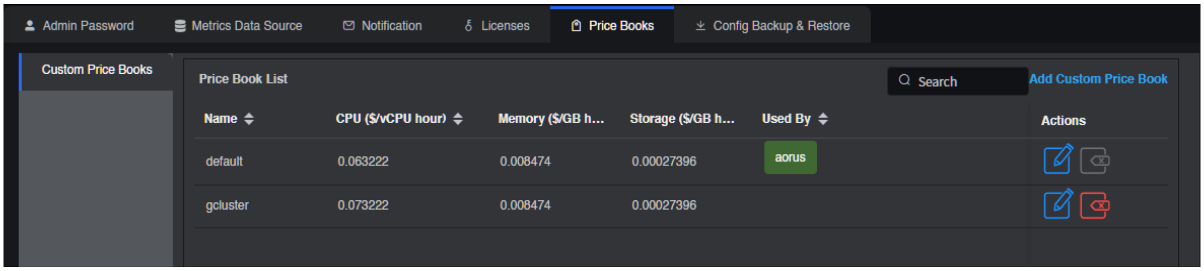
- View the key for your license. To do this, click the *Show License Key* icon.
- Remove the license. To do this, click the *Remove License Key* icon and confirm the removal.

Price Books

The *Price Books* page allows you to define price books for on-premises clusters. To access this page, select *Configuration, System Settings, Price Books* tab.

Custom Price Books

The *Custom Price Books* page allows you to define price books per cluster. Each price book contains hourly operating costs for CPU, memory, and storage. These numbers are used for calculating costs/savings for on-premises clusters.

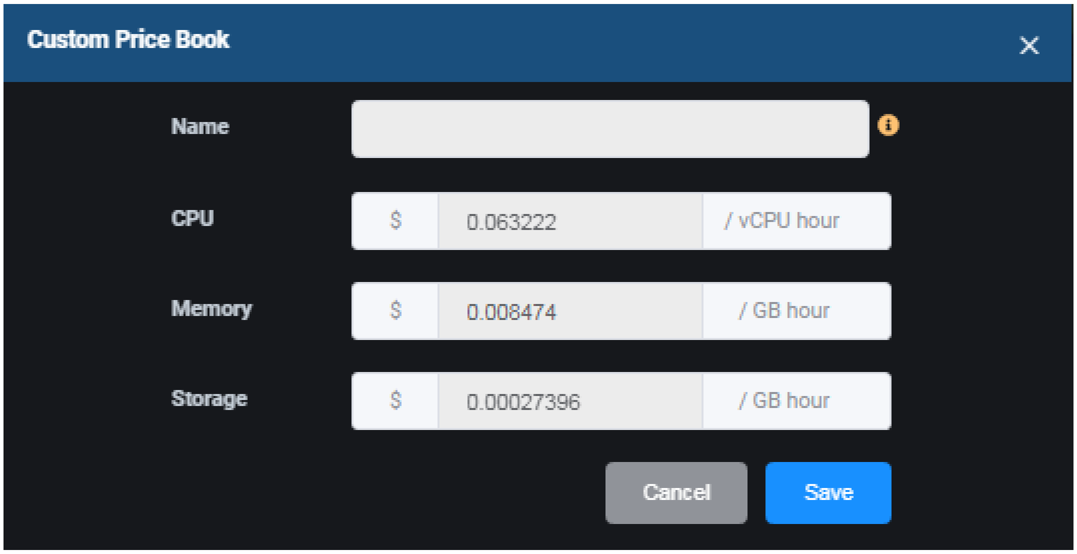


Price Book List

Name	CPU ($/vCPU hour)	Memory ($/GB h...	Storage ($/GB h...	Used By	Actions
default	0.063222	0.008474	0.00027396	aorus	
gcluster	0.073222	0.008474	0.00027396		

Add a Custom Price Book

1. On the *Custom Price Books* page, click *Add Custom Price Book*.



Custom Price Book

Name

CPU $ 0.063222 / vCPU hour

Memory $ 0.008474 / GB hour

Storage $ 0.00027396 / GB hour

Name – The name must have a maximum of 64 lowercase characters,"-", or "." allowed. The name must start and end with an alphanumeric character.

CPU/Memory/Storage - When determining your hourly costs, be sure to include electricity, cooling/heating, labor, hardware, etc.

2. Click *Save* when you are done.

   This custom price book can now be assigned to clusters from the *Configuration / Clusters* page.

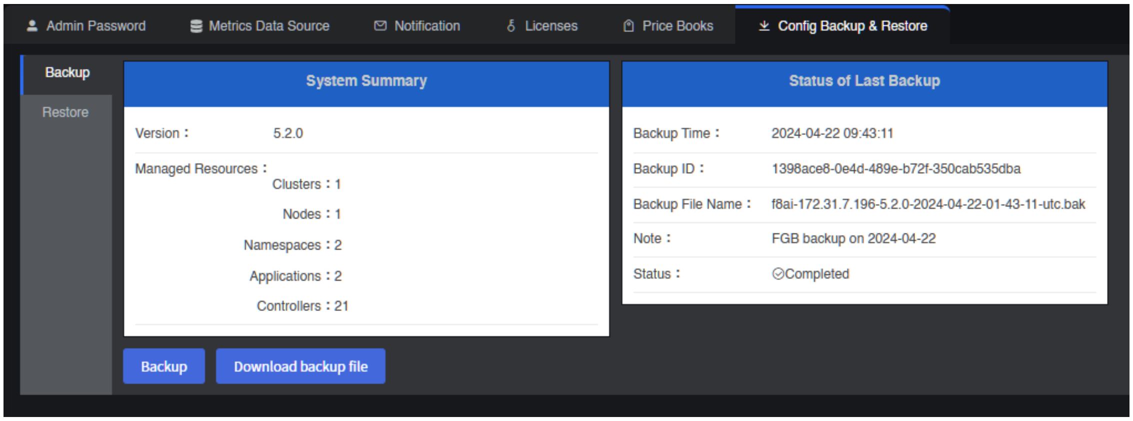
Manage Custom Price Books

You can do the following from the *Configuration / Custom Price Books* page:

- Edit a custom price book. To do this, click the *Edit Custom Price Book* icon.
- Remove a custom price book. You can only remove a custom price book if it is not being used by a cluster. To do this, click the *Remove Custom Price Book* icon.

Backup and Restore

Federator.ai GPU Booster enables you to back up and protect your system configuration and recover the configuration in case of disaster. The configuration that is backed up includes clusters, namespaces being monitored, applications, auto provisioning profile, and alert monitoring rules. To access this page, select *Configuration, System Settings, Config Backup & Restore* tab.



System Summary

Version : 5.2.0

Managed Resources :
Clusters : 1
Nodes : 1
Namespaces : 2
Applications : 2
Controllers : 21

Status of Last Backup

Backup Time :	2024-04-22 09:43:11
Backup ID :	1398ace8-0e4d-489e-b72f-350cab535dba
Backup File Name :	f8ai-172.31.7.196-5.2.0-2024-04-22-01-43-11-utc.bak
Note :	FGB backup on 2024-04-22
Status :	✓Completed

If a backup has been previously run, the *Status of Last Backup* window shows the time, ID, file name, status, and any note that was provided at the time of backup.

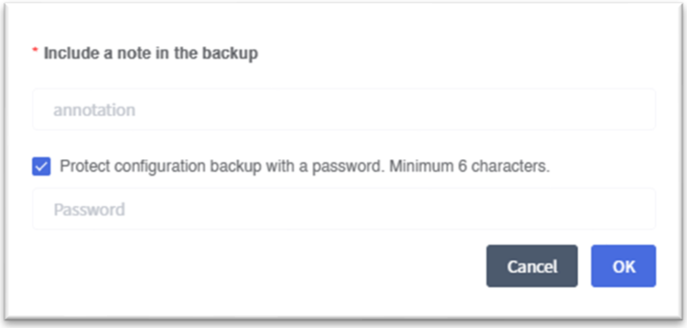
Back Up System Configuration

Follow the steps below to back up the system configuration:

1. On the *Backup* page, click *Backup*.

   The last backup performed will be deleted from the server. Confirm that you want this before proceeding. To make a copy of the last backup file, exit and download it before proceeding.

2. If desired, add a note to the backup.



* Include a note in the backup

annotation

☑ Protect configuration backup with a password. Minimum 6 characters.

Password

   This can be useful for noting relevant information, such as a note that the backup was performed prior to system hardware or operating system changes.

3. Click *OK* to start the backup.

   The status is displayed on the right side of the screen.

4. Once completed, click *Download backup file*.

   The most recent backup file is kept internally. It is a good idea to download and copy the file to a safe, ideally remote, location.

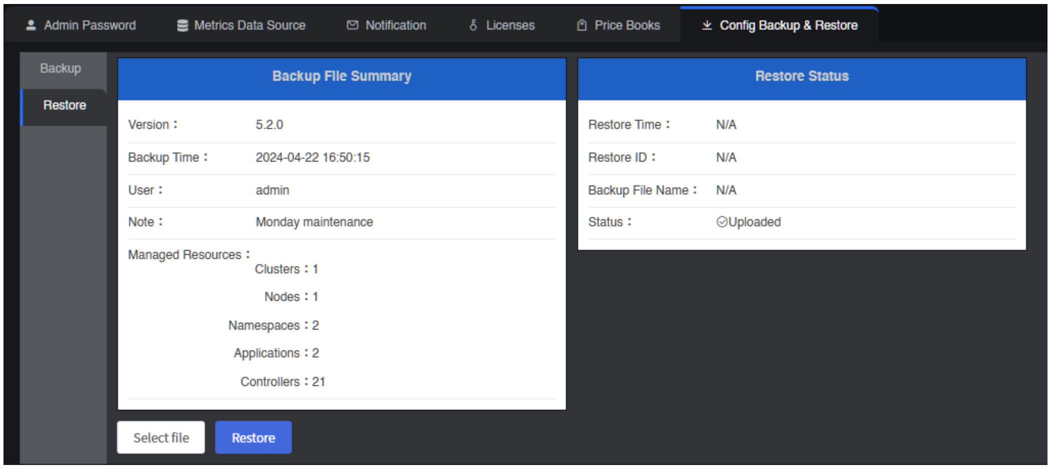
Recover System Configuration

Restoring a configuration is for disaster recovery purposes and should not be used in day-to-day operations. Changes made since the configuration was last saved will not be included in this restored configuration.

Follow the steps below to restore the system configuration:

1. On the *Restore* page, click *Select file*.

2. Locate the saved configuration file.



Version : 5.2.0
Backup Time : 2024-04-22 16:50:15
User : admin
Note : Monday maintenance
Managed Resources :
Clusters : 1
Nodes : 1
Namespaces : 2
Applications : 2
Controllers : 21

Restore Time : N/A
Restore ID : N/A
Backup File Name : N/A
Status : ⊘Uploaded

   Details about the backup appear, including any notes you may have added.

3. Confirm that this is the correct backup and click *Restore*.

   The status is displayed on the right side of the screen.

Related topics:

Terminology
Search/Sort Information in Tables
Configure Clusters

Events

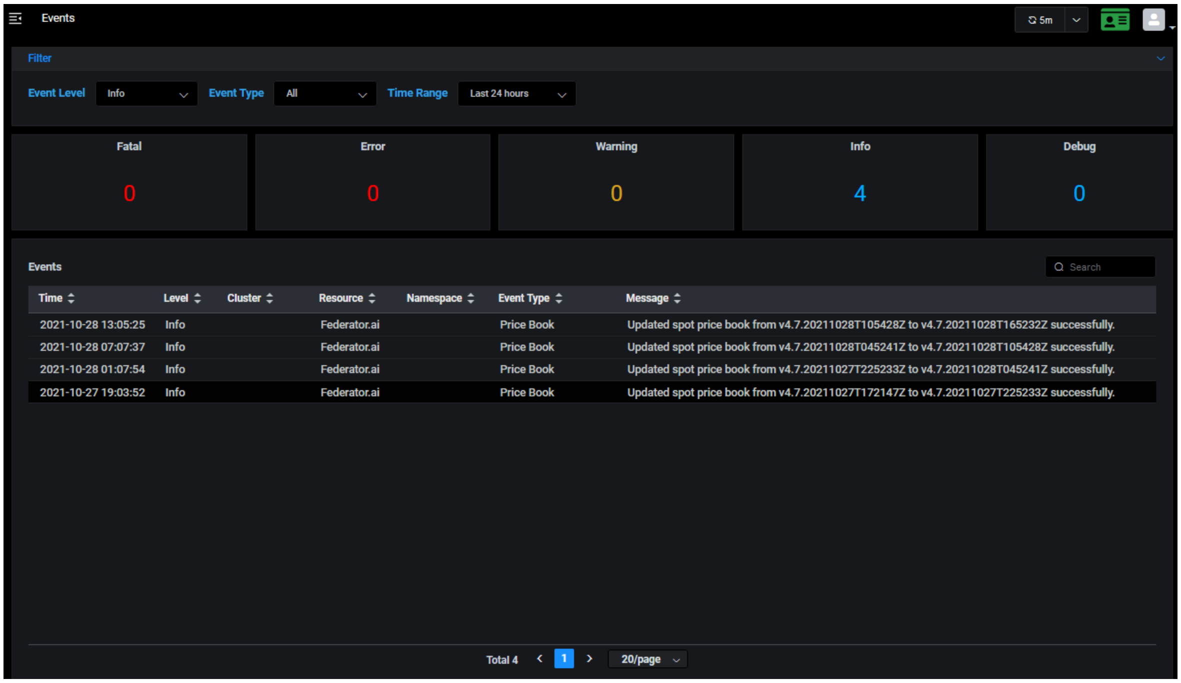
The *Events* page displays all system events that have occurred.

There are five levels of events:

- Fatal - Issues that may stop the system from operating properly.
- Error - Indicates that a failure has occurred.
- Warning - Indicates that something occurred that may require maintenance or corrective action; however, the system is still operational.
- Info - Day-to-day activities, which require no action.
- Debug - Detailed activities used for troubleshooting.

You can filter by the event level, event type, and the time frame to display. For example, if you select the warning level, all warnings, errors, and fatal events will be displayed for the specified time period.

To specify a custom range, select *Custom* under *Time Range* and then specify a date range.



Event Level: Info
Event Type: All
Time Range: Last 24 hours

Fatal	Error	Warning	Info	Debug
0	0	0	4	0

Events

Time	Level	Cluster	Resource	Namespace	Event Type	Message
2021-10-28 13:05:25	Info		Federator.ai		Price Book	Updated spot price book from v4.7.20211028T105428Z to v4.7.20211028T165232Z successfully.
2021-10-28 07:07:37	Info		Federator.ai		Price Book	Updated spot price book from v4.7.20211028T045241Z to v4.7.20211028T105428Z successfully.
2021-10-28 01:07:54	Info		Federator.ai		Price Book	Updated spot price book from v4.7.20211027T225233Z to v4.7.20211028T045241Z successfully.
2021-10-27 19:03:52	Info		Federator.ai		Price Book	Updated spot price book from v4.7.20211027T172147Z to v4.7.20211027T225233Z successfully.

Total 4 < 1 > 20/page

At the bottom of the page, the total number of events is shown, along with navigation to other pages. You can also determine how many events to show per page.

Related topics:

Terminology
Search/Sort Information in Tables

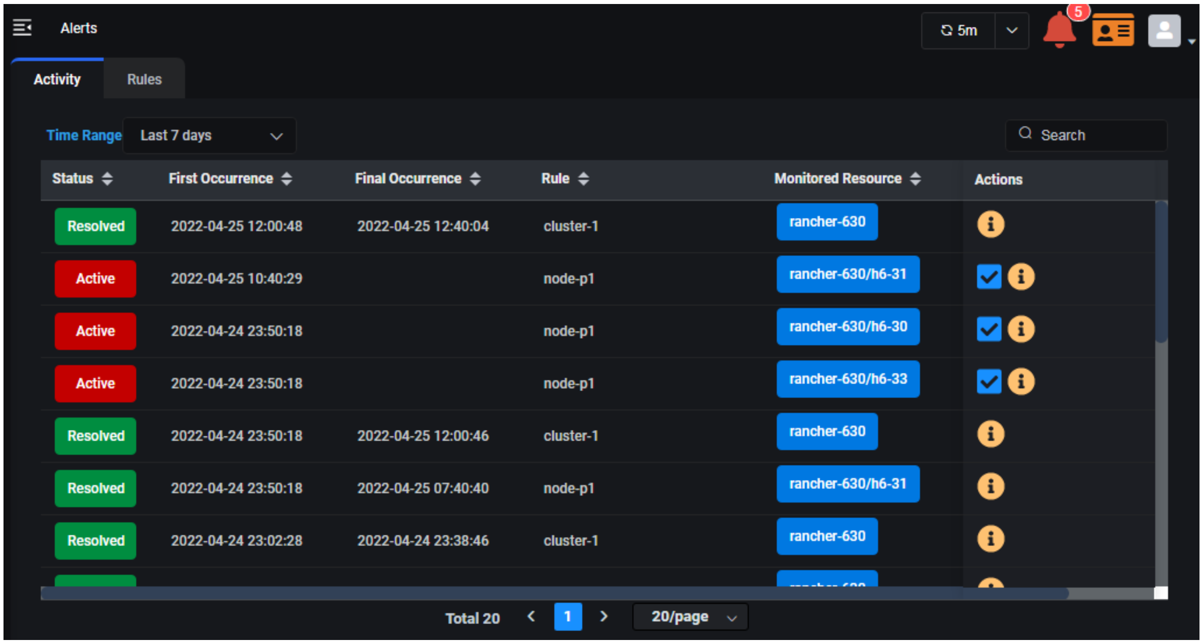
Alerts

The *Alerts* page has tabs that allow you to:

- View and manage alerts that have been generated when a configured rule is triggered.
- Create and manage rules.

Activity

The *Activity* page allows you to view alerts that have been generated when a configured rule is triggered. To access this page, select *Alerts, Activity* tab.



Time Range: Last 7 days

Status	First Occurrence	Final Occurrence	Rule	Monitored Resource	Actions
Resolved	2022-04-25 12:00:48	2022-04-25 12:40:04	cluster-1	rancher-630	ℹ
Active	2022-04-25 10:40:29		node-p1	rancher-630/h6-31	☑ ℹ
Active	2022-04-24 23:50:18		node-p1	rancher-630/h6-30	☑ ℹ
Active	2022-04-24 23:50:18		node-p1	rancher-630/h6-33	☑ ℹ
Resolved	2022-04-24 23:50:18	2022-04-25 12:00:46	cluster-1	rancher-630	ℹ
Resolved	2022-04-24 23:50:18	2022-04-25 07:40:40	node-p1	rancher-630/h6-31	ℹ
Resolved	2022-04-24 23:02:28	2022-04-24 23:38:46	cluster-1	rancher-630	ℹ

Total 20 ‹ 1 › 20/page

Each alert displays the following information:

- Status – Active or resolved. Alerts are resolved automatically based on a configured threshold or can be resolved manually by clicking the *Resolve* icon ☑ under *Actions*.
- First Occurrence – Time the alert was triggered.
- Final Occurrence – For resolved alerts; the last time the alert was triggered.
- Rule – Name of the rule.
- Monitored Resource – Resources being monitored by the rule.

At the bottom of the page, the total number of alerts is shown, along with navigation to other pages. You can also determine how many alerts to show per page.

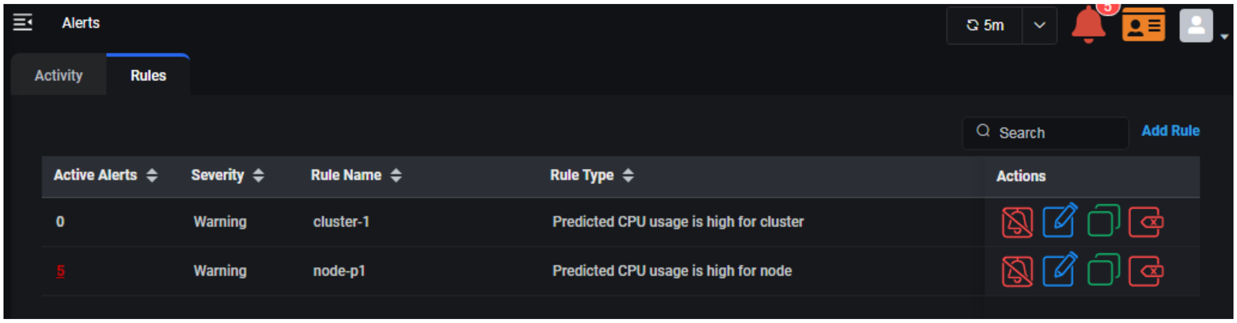
When a rule is created from the *Rules* page, monitoring begins immediately. When a rule condition is reached for first time (e.g., predicted daily CPU usage is greater than the configured trigger), an alert is generated. When the alert is generated, the *Activity* page will display the time of the *First Occurrence*. Monitoring occurs continuously and there is no update to the alert until the condition is manually resolved; each alert is only triggered once until it is manually resolved. At that time, the *Final*

Occurrence will be updated with current time. If the rule condition is reached again, a new alert is created.

Click the *Details* icon ℹ under *Actions* to view detailed information for an alert (active or resolved). Here you will see the condition that triggered the alert.

Rules

The *Rules* page allows you to create and manage rules for your system. To access this page, select *Alerts, Rules* tab*.*



Active Alerts	Severity	Rule Name	Rule Type	Actions
0	Warning	cluster-1	Predicted CPU usage is high for cluster	
5	Warning	node-p1	Predicted CPU usage is high for node	

Each configured rule displays the following information:

- Active Alerts – Number of active alerts for this rule. Click the number to jump to the Alerts / *Activity* Page to view just the active alerts for this rule.
- Severity – Type of alert, critical or warning.
- Rule Name – Name of the rule.
- Rule Type – Parameter monitored by the rule.

At the bottom of the page, the total number of rules is shown, along with navigation to other pages. You can also determine how many rules to show per page.

Rules allow you to monitor clusters, nodes, namespaces, applications, and controllers for a variety of conditions, including the need to increase or decrease CPU and memory allocations based on predicted usage.

Rules offer flexibility by allowing you to set trigger thresholds based on a specific value or a percentage. When a rule is triggered, an alert is generated. You can also set an auto-resolve threshold to resolve an active alert if the monitored value falls below the threshold. Generated alerts can be viewed on the *Activity* page. In addition, rules can be set to notify users via email. *Email Notification* must be configured to use this feature. It is set on the *Configuration*, *System Settings*, *Notification* tab.

In addition to adding rules, you can perform the following functions from the *Rules* page:

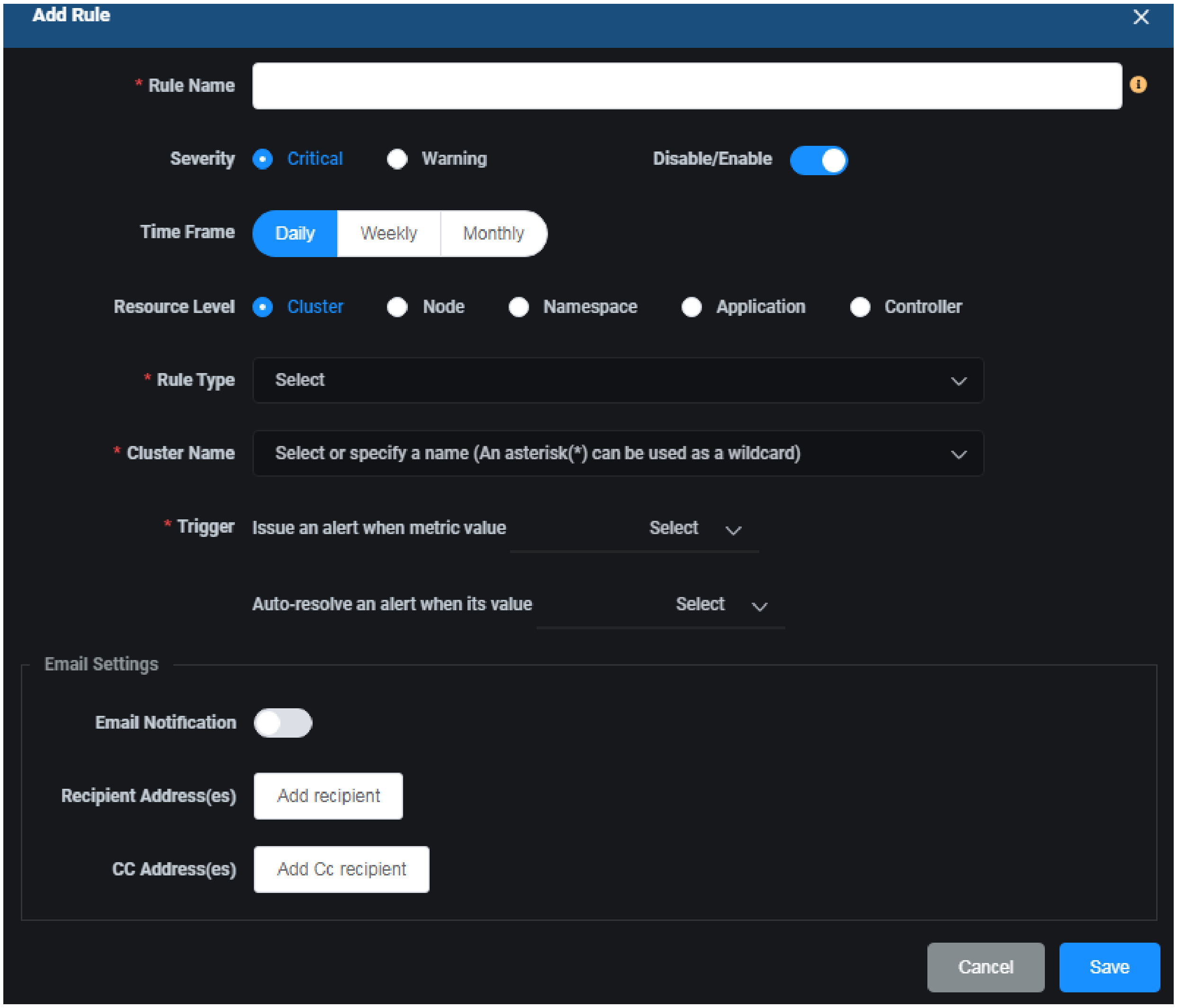
Icon	**Function**
🔔 🔕	Enable/disable a rule.

✎	Edit a rule.
⧉	Clone a rule.
⌫	Delete a rule.

Add a Rule

Follow the steps below to add a rule:

1. On the *Rules* page, click *Add Rule*.



Add Rule

* Rule Name

Severity ● Critical ○ Warning Disable/Enable (on)

Time Frame [Daily] Weekly Monthly

Resource Level ● Cluster ○ Node ○ Namespace ○ Application ○ Controller

* Rule Type Select

* Cluster Name Select or specify a name (An asterisk(*) can be used as a wildcard)

* Trigger Issue an alert when metric value Select

Auto-resolve an alert when its value Select

Email Settings

Email Notification (off)

Recipient Address(es) Add recipient

CC Address(es) Add Cc recipient

Rule Name – The rule name must have a maximum of 64 lowercase characters,"-", or "." allowed. The name must start and end with an alphanumeric character. It is recommended to create a meaningful rule name (e.g., cpu-under-provisioned-cluster1) so that it is easily recognizable on the *Activity* page.

Severity – The severity level for alerts can be *Critical* or *Warning*.

Disable/Enable – Enable the rule for monitoring to occur. Rules are enabled by default.

Time Frame – Specify the time frame for predictions for this rule (e.g., the weekly prediction for CPU determines whether the cluster trigger will be met.)

Resource Level – Specify what you want to monitor. Depending upon what you select, you will have fields for *Cluster Name*, *Node Name, Namespace Name, Application Name*, and/or *Controller Name.*

Rule Type – Specify the type of rule.

Cluster/Node/Namespace/Application/Controller Name – Select one from the drop-down list or specify a name. An asterisk (*) can be used as a wildcard to specify multiple clusters/nodes/namespaces/applications/controllers (for example, cluster *clus** would monitor *cluster1, cluster2, cluster3*, etc.). All resources in the selected clusters will be monitored.

Trigger – Specify the trigger threshold. The threshold can be based on a fixed value (e.g., greater than 5 cores) or, for clusters/nodes/namespaces/controllers, a percentage (e.g., greater than 95% of cluster capacity). Also, specify the threshold to automatically resolve an alert as a fixed value (e.g., less than 4 cores) or a percentage (e.g., less than 85% of cluster capacity).

Email Settings – Specify if you want to send alerts to users via email and the recipients. *Email Notification* must be configured to use this feature and is set on the *Configuration*, *System Settings*, *Notification* tab.

2. Click *Save* when you are done.

   To create a similar rule based on this rule, click the *Clone Rule* icon.

Manage Rules

You can do the following from the *Alerts / Rules* page:

- Enable/disable a rule. A disabled rule will not monitor resources. To do this, click the *Enable/Disable* icon.
- Edit a rule. To do this, click the *Edit Rule* icon.
- Clone a rule. Cloning a rule creates a new rule from an existing rule. You can modify the rule parameters, simplifying the ability to create multiple rules. To do this, click the *Clone Rule* icon, make the appropriate modifications and click *Save*.
- Delete a rule. To do this, click the *Delete Rule* icon and confirm the removal.

Related topics:

Terminology
Search/Sort Information in Tables
Email Notification