



Best Practices for Kubernetes Cost Monitoring and Management

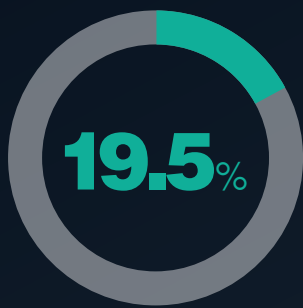
WHITE PAPER

CONTENTS

INTRODUCTION	3
MULTI-TENANCY MODELS FOR KUBERNETES RESOURCE SHARING	4
BEST PRACTICES FOR KUBERNETES COST MANAGEMENT	6
COST MANAGEMENT TOOLS	8
REQUIREMENTS FOR KUBERNETES COST MANAGEMENT	8
AVAILABLE KUBERNETES COST MANAGEMENT TOOLS	8
KUBECOST	8
CLOUD TOOLS	9
RAFAY'S COST MANAGEMENT SERVICE	10
BENEFITS OF RAFAY'S COST MANAGEMENT SERVICE	11
RAFAY COST MANAGEMENT DASHBOARD	12
RAFAY COST MANAGEMENT WORKFLOWS	13
ABOUT RAFAY KUBERNETES OPERATIONS PLATFORM PLATFORM	13

INTRODUCTION

Kubernetes provides enterprises with a fast and easy way to provision infrastructure resources and is often used in global, ever-changing, multi-tenant environments. As a result, Kubernetes resource consumption and spend can rapidly increase—soaring past budgets—making it incredibly difficult to not only view the organization's cloud spend across infrastructure, but to also report how spend is allocated across internal departments. To manage Kubernetes cloud costs appropriately, enterprises need a consolidated, real-time view of all Kubernetes cloud spend across clusters in public clouds and on-premise data centers with precise allocation and reporting for shared clusters and infrastructure.



JUST 19.5% REPORTED

being able to forecast
spend within 5% accuracy
found in collaboration with
the FinOps Foundation

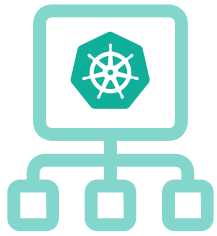
CNCF survey

Gaining visibility into enterprise Kubernetes cloud spend is difficult because organizations often allow multiple application teams or business units to run their modern application workloads in the same shared clusters to save the high cost of always procuring new clusters. However, shared clusters make it more challenging to evaluate what resources each team is actually utilizing and to properly allocate the cost of resources to each application team or business unit. This leads to Kubernetes cloud spend sticker shock at the end of each month.

In hybrid-cloud or multi-cloud environments, enterprises typically run clusters across multiple accounts with cloud vendors, as well as on-premise in their data centers, creating additional cost management silos, budget risk, and reporting nightmares for platform and FinOps teams.

A [2021 survey](#) conducted by CNCF in collaboration with the FinOps Foundation found that just 19.5% reported being able to forecast spend within 5% accuracy, and 53% say that it takes days for their FinOps teams to respond to unexpected cost increases.

By understanding and analyzing Kubernetes costs at the right levels of granularity, your organization can significantly reduce costs without impacting performance. This whitepaper explores three common models for allocating and sharing resources on Kubernetes clusters and discusses some of the tools and best practices for Kubernetes cost management.



Multi-Tenancy Models for Kubernetes Resource Sharing

Most organizations divide cluster resources among multiple teams to reduce costs and simplify administration. In these *multi-tenant* environments, resources are shared between multiple teams, each operating one or more applications. There are three common models for sharing cluster resources:

There are three common models for sharing cluster resources:

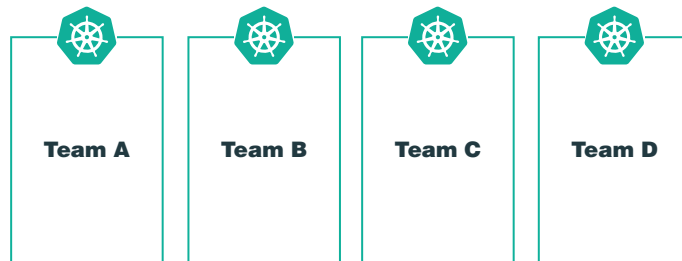
Hard multi-tenancy

Soft multi-tenancy
(node-group based)

Soft multi-tenancy
(namespace-based)

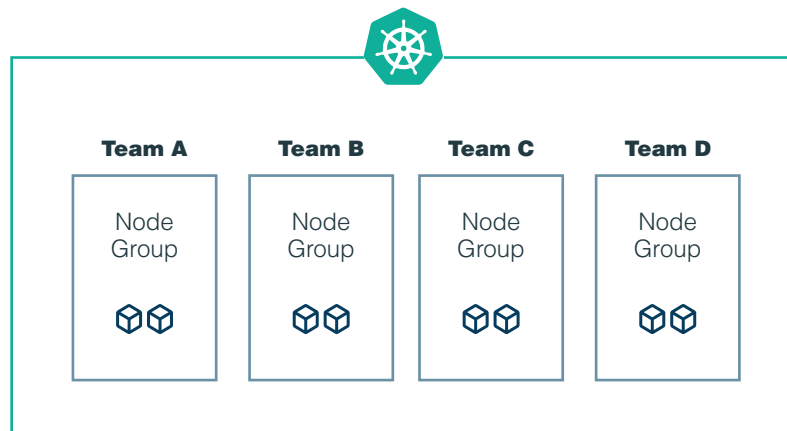
Hard multi-tenancy:

Each team runs separate clusters in dedicated accounts



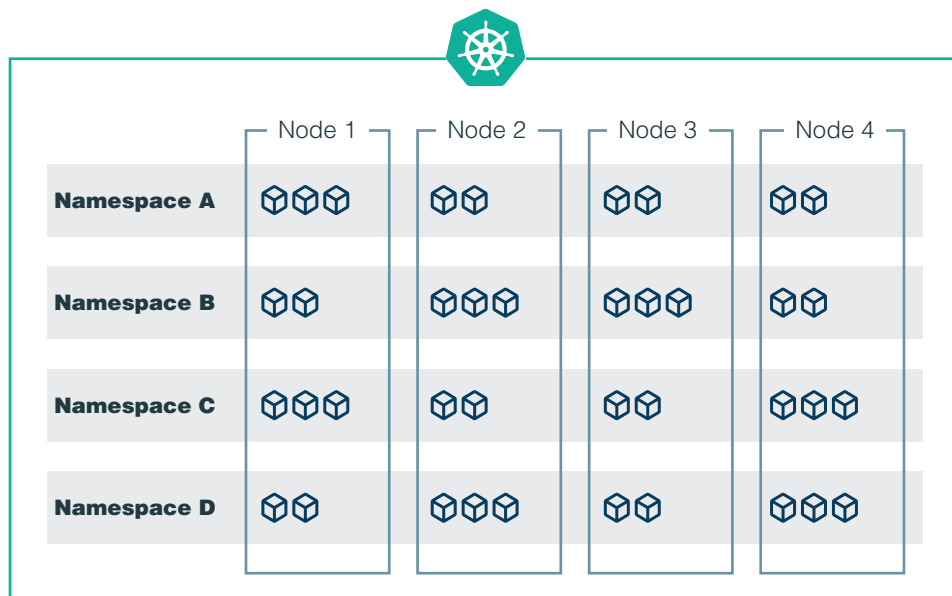
Soft multi-tenancy (node-group based):

Each team is allocated a dedicated node group in a shared cluster.



Soft multi-tenancy (namespace-based):

Each team is allocated a separate namespace in a shared cluster.



K8s Namespaces

Namespaces can be used in situations where a cluster has multiple apps and users. Each Kubernetes resource must have a unique name within its namespace, but not across namespaces. Namespaces cannot be nested inside one another.

Namespaces are a good approach to logically separate cluster resources between multiple users via resource quotas that limit aggregate resource consumption per namespace.

Of these options, hard-multi-tenancy is simplest in terms of cost visibility but the most expensive to implement. This is because, if you have four teams, you run a minimum of four separate clusters. In practice, each team will likely need additional clusters for different functions: development, testing, production, etc. so the total number of clusters is likely to be even higher, raising costs further.

Soft multi-tenancy based on node groups is the next simplest model in terms of cost visibility, since each team has a dedicated set of nodes. Since it's easy to identify the cost per node, it's easy to determine how much resources each team is consuming. However, this approach limits the number of nodes available to individual teams to run parallel instances of microservices. This has potential negative impacts on application performance and reliability.

Because it allows a team to spread microservices across all the nodes in a cluster and consume and release cluster resources more flexibly, soft multi-tenancy based on namespaces is frequently the most desirable option. However, cost visibility is difficult with this model. While namespaces provide a clear separation of resources, they require the collection of granular resource utilization metrics to tie costs to each namespace and each running service.



Best Practices for Kubernetes Cost Monitoring and Management

There are a number of best practices for any team that wants to implement Kubernetes cost management across a growing fleet of clusters. They include:

Choose the multi-tenancy model that fits your needs

The first step in Kubernetes cost management as noted above is understanding how your clusters are currently being shared—or how you want them to be shared—so that you can choose appropriate cost management tools. As noted earlier, hard multi-tenancy and soft multi-tenancy based on node groups are simpler in terms of cost visibility. Most enterprises choose soft multi-tenancy model based on namespaces. Be open to the possibility that a mixed model—hard for some needs, and soft multi-tenancy for others—may be the best option.

Provide visibility at the right level of granularity

Before you start considering cost management tools, it's also important to understand how you will need to slice and dice your cost data. As a rule of thumb, the more complex your multi-tenancy model, the greater granularity you will need. There are three levels of granularity (from simplest to most complex):

- **Infrastructure-level** You can visualize costs based on infrastructure components like servers (virtual or physical) and storage capacity.
- **Kubernetes-level** You can visualize costs at the level of Kubernetes constructs such as a namespace or deployment, or based on Kubernetes labels.
- **Organization-level** You can visualize costs based on organizational constructs such as organization, team, or project.

Providing visibility at the right level of granularity ensures you can meet your cost management goals without having to manually slice and dice the data to get the numbers you need. For example, if you're using hard multi-tenancy, you may be able to get all the data you need from infrastructure-level cost tools available from your cloud provider. If you're using soft multi-tenancy based on namespaces, you'll need Kubernetes-level and organization-level granularity.

Implement showback or chargeback

For most organizations, the purpose of Kubernetes cost management—at least initially—is to get teams to utilize resources more efficiently. This is accomplished using either *showback* or *chargeback*.



Unallocated Costs

Unallocated costs are costs due to the extra resources provisioned within a cluster to accommodate growth, spikes in demand, and resource failures.



Shared Costs

Shared costs are the costs due to the overhead of essential services such as monitoring, security, etc. that every production Kubernetes cluster must run.

With *showback*, groups and individuals can see how much the resources they use cost the organization. Ideally, they should also be able to see how efficiently those resources are being used so they can adjust usage accordingly. *Chargeback* takes this a step further, requiring payment for resources used.

In either case, it's essential to provide the right level of granularity to get the best results. Showing an individual or group how much the entire organization spends on Kubernetes infrastructure is far less motivating than showing them the cost of the resources they actually consume.

Allocate shared and unallocated costs

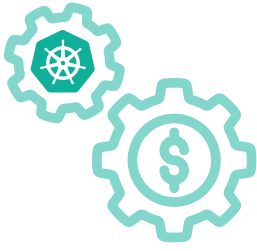
Your Kubernetes clusters, especially production clusters, can have significant overhead in terms of both *shared costs* and *unallocated costs*. Shared costs are for must-have cluster services like security and monitoring, while unallocated costs are the costs due to necessary overprovisioning.

Most production clusters are provisioned with “extra” resources to accommodate growth, spikes in demand, and resource failures. For instance, a cluster may have extra nodes so that if one node fails, there's enough headroom to ensure applications can keep running without constraint.

If you are operating anything other than a hard multi-tenancy model (where all resources in a cluster—allocated or not—are assigned to a single team), it can be important to consider how shared and unallocated costs are divided. The simplest approach is to divide these costs equally among tenants. A more advanced approach would be to allocate costs based on resource consumption. For example, if a cluster is used by three teams, with Team A consuming 80% of allocated resources and Team B and Team C each consuming 10% of allocated resources, it would obviously be unfair to Team B and Team C to expect them to each be responsible for a third of total shared and unallocated costs.

Retain data for the right amount of time

A final consideration is how long to retain your cost management data. Keeping data for a longer period can help you provide more accurate forecasts of future resource consumption so you can budget appropriately. On the other hand, the cost of retaining data at high granularity can increase quickly. Plan on starting with retention of six months of data, and adjust (up or down) as your needs become clearer and your cost management practices mature.



Cost Management Tools

Excel is not an effective cost management tool for Kubernetes or any other IT environment. If you want to track and manage costs across your growing Kubernetes fleet, you need the right tools for the job. That means either identifying tools that meet your needs or creating your own.

Requirements for Kubernetes Cost Management

Before you set out to choose cost management tools, it's important to think about your requirements—in keeping with the best practices described earlier. The tool(s) you choose should ideally:

1. Work across all the platforms you use, on-premises, in the cloud, and at the edge.
2. Provide visibility at the right level of granularity for your organizational needs.
3. Require minimal installation and management.
4. Integrate with the tools and software you already use, including Kubernetes distributions, management tools, security tools like SSO, etc.
5. Provide a single UI to visualize costs across all of your Kubernetes infrastructure, allowing you to quickly and easily identify targets for “rightsizing.”
6. Retain both hourly and daily data for periods of time that meet or exceed your planning horizon.
7. Democratize access to cost metrics so that all teams can see the costs associated with resources they use and take steps to optimize costs.

The holy grail of cost management is a tool that is simple to use, requires minimal or no setup or ongoing management, provides visibility at the right level of granularity, automates regular tasks, and offers flexible and easy-to-use reporting.

Available Kubernetes Cost Management Tools

There are a number of solutions available that can help you solve your Kubernetes cost management challenges. Of these, the most widely known and used solution is Kubecost. Many teams also utilize the cost management tools provided by various cloud providers.



Cloud cost management tools can meet simple Kubernetes cost management needs.

For those with operations in hybrid or multiple clouds, companies will have to leverage multiple tools.

Kubecost

Kubecost can help you collect, analyze, and optimize Kubernetes costs, providing real-time visibility. It lets organizations breakdown costs at the Kubernetes level and the organizational level, allocate costs by defined aggregations, and view costs across multi-cluster, multi-cloud environments (with the right license). Kubecost works with Kubernetes deployments on AWS, Azure, Google Cloud Platform, and on premises. License options include:

- **Free level:** Monitors costs of individual clusters; 15 day metric retention.
- **Business level:** Unlimited clusters; priced in units of 100 nodes; 30-day metric retention; business support.
- **Enterprise level:** Unified multi-cluster view; unlimited metric retention, SSO/SAML integration, enterprise support.

In June 2022, Kubecost and partners also launched OpenCost, an open source cost standard for Kubernetes workloads based on Kubecost.

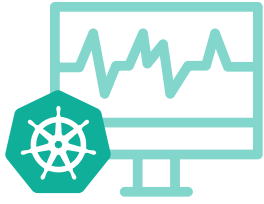
Here's how Kubecost measures up against the list of requirements above:

- 1. Platform support:** Kubecost covers major clouds and on-prem.
- 2. Granularity:** Kubecost offers Kubernetes-level and organization-level granularity. You'll need to define the right roles and role-based access to take full advantage of this.
- 3. Installation and management:** Kubecost is self-hosted, meaning it must be installed on each cluster. A SaaS version, Kubecost Cloud, is in limited availability as of October 2022.
- 4. Integration:** Kubecost integrates with tools such as PagerDuty and Slack. If you want SSO integration, you'll need the Enterprise version.
- 5. Visualize across clusters:** A multi-cluster view is available with the Enterprise version.
- 6. Data retention:** Data retention depends on the level you choose. Only Kubecost Enterprise gives you retention beyond 30 days.

Public Cloud Cost Management Tools

Public cloud vendors, including Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP) provide a reasonably detailed bill that will let enterprises calculate costs for running Kubernetes resources such as clusters and nodes. These capabilities are useful to deal with hard multi-tenancy and soft multi-tenancy based on node groups. However, if you plan to use soft multi-tenancy based on namespaces, companies will need finer-grained cost accounting.

Some public clouds such as AWS have begun incorporating Kubecost into their cost management software. However, because each cloud has its own implementation, it may complicate cost management when operating a fleet across multiple public clouds.



Rafay's Cost Management Service

The Rafay Cost Management Service is an easy-to-deploy SaaS service. It builds on the capabilities of OpenCost to allow your teams to understand the costs associated with operation of your Kubernetes fleet and take actions to optimize costs. You can view, analyze, and showback or chargeback the cost for your Kubernetes infrastructure and applications at multiple levels of granularity:

- **Kubernetes-level:** Namespace, Deployment, Services, Labels, etc.
- **Organization-level:** Organization, Project, Workspace

Rafay supports cloud platforms including AWS and Microsoft Azure, and it also works on-premises. It lets enterprises see resource costs associated with CPUs, memory, storage, etc., all in one place.

You can create cost reports to track trends and efficiency across all or a subset of Kubernetes and Rafay constructs such as Projects, Clusters, Namespaces, Workloads, etc. The Cost Management Service provides real-time alerting and recurring reports, letting you gain cost control over monitored Kubernetes infrastructure and stay within budgeted limits by providing categorized cost allocation based on usage.

Benefits of Rafay's Cost Management Service

Rafay's Cost Management Service is fully integrated with the Rafay Kubernetes Operations Platform, ensuring that we adhere to the best practices described in this white paper while simplifying operations for organizations operating large fleets of Kubernetes clusters.

Benefits include:

Nothing to install: When you sign-up for the Rafay Cost Management Service, it automates all the hard work of installation, ensuring that the required components are installed correctly and identically on all clusters. Future software updates are similarly automated and straightforward.

Easy-to-use dashboards: You can visualize costs and usage and right-size your Kubernetes infrastructure with ease.

Extended data retention: Historical cost data is retained over a longer period of time, providing the data you need to forecast future costs. Default retention is 14 days of hourly data and 6 months of daily data.



The Rafay Cost Management Service

offers full-featured cost management within an integrated Kubernetes management and operations platform.

No licensing required: With the Rafay Cost Management Service, your Platform Team can avoid the challenges that come with licensing, installing, operating, and maintaining yet another product, from another vendor.

Integrated workflows: Rafay's RBAC model is automatically extended to all cost management dashboards providing users access to cost metrics based on the role granted to them. This increases transparency and reduces the time to execute on any 'closing the loop' cost optimization exercises.

For existing Rafay customers, you can utilize the roles and other organizational elements that you've already defined in conjunction with Rafay Cost Management, making it even simpler to get started.

Platform Integration: Rafay provides capabilities that enforce cost reduction policies at scale. For example, you can create policies that mandate definition of pod resource requests/limits, or templates that only allow use of Graviton-based instances on AWS, or only allow Spot instances for QA/Dev clusters.

Rafay Cost Management Dashboard

The Rafay Cost Management Dashboard provides a single place to unearth and view costs associated with Kubernetes resources. It provides cost overviews, along with an explorer that allows deep dives into costs. You can slice and dice the cost data on the dashboard across multiple dimensions or download the data for further analysis (see Figure 1).

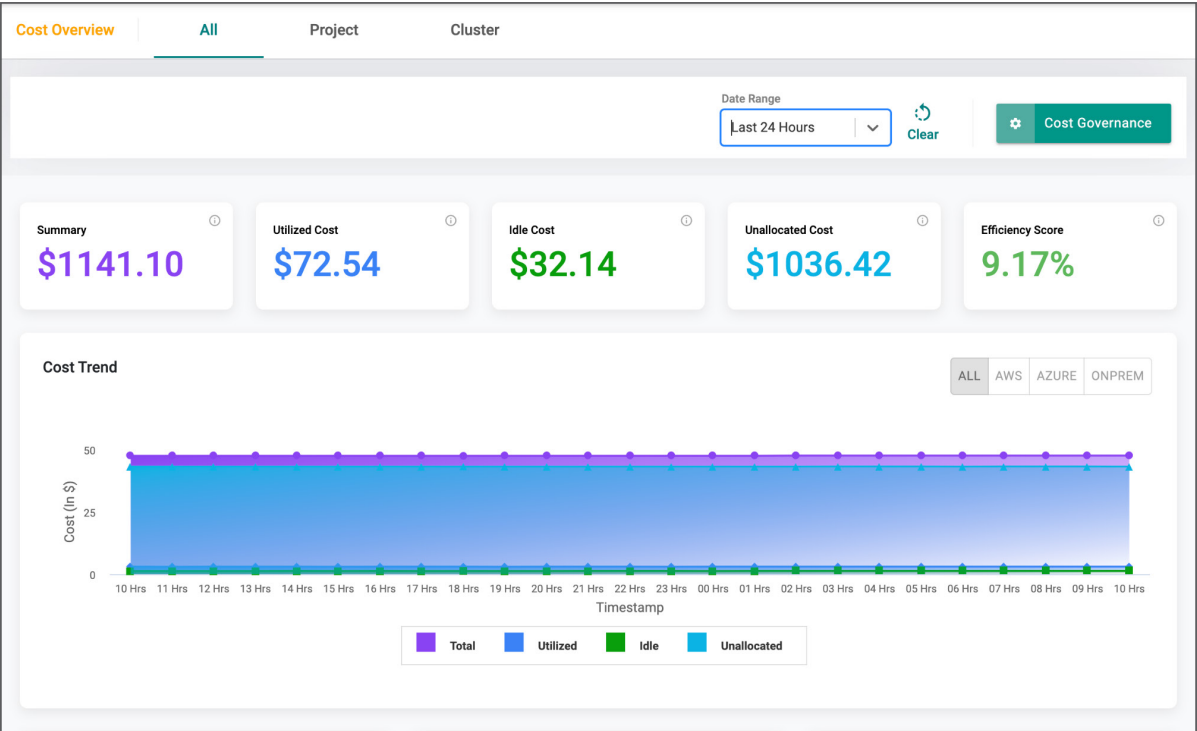
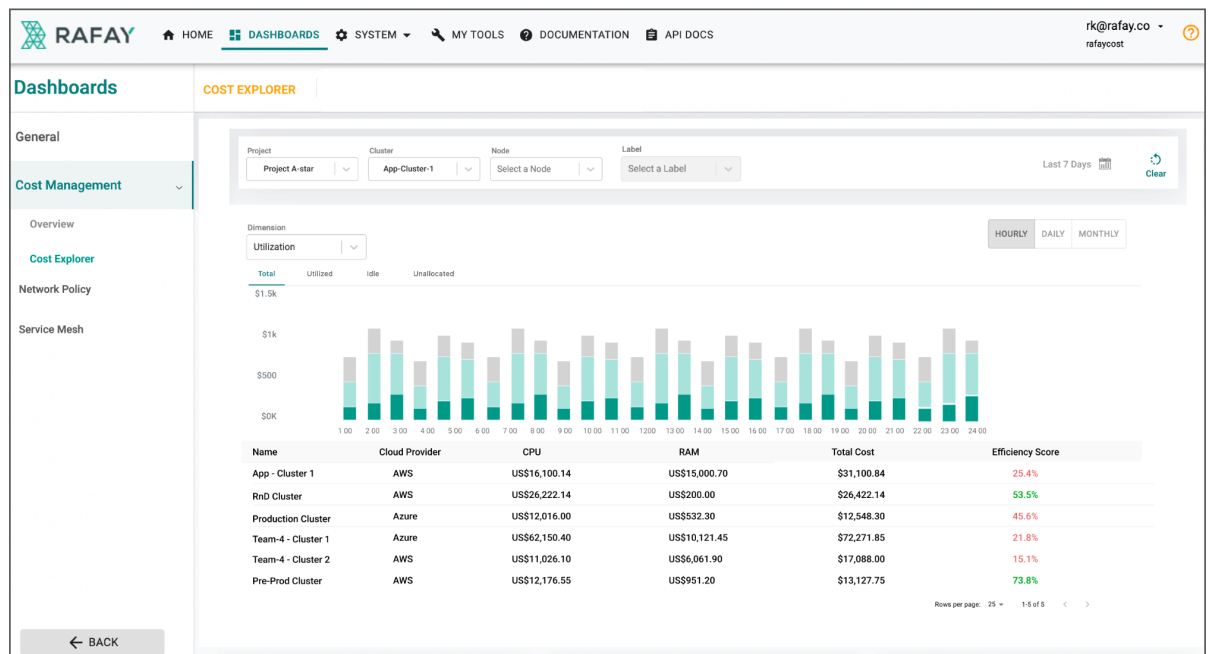


Figure 1 - Cost Management Dashboard

As seen below, data can be filtered based on time and at different levels of granularity such as Projects, Clusters, Namespaces, Labels, Workloads, and Cloud Providers.



The Rafay Cost Management Dashboard provides a single access point to cost data for all users, so they can observe, analyze, and report data. The Dashboard:

- Provides an appropriate summary for all personas
- Shows high-level trends
- Includes mechanisms to slice and dice cost data
- Enables users to correlate costs to corresponding resource utilization/ allocation
- Allows users to create and schedule reports
- Supports the creation of chargeback groups



Rafay Blueprints ensure that every cluster managed with Rafay has the proper add-ons, cost management controls, security, and other policies applied.

Rafay Cost Profiles make it easy to account for custom pricing you've negotiated with a cloud provider as well as on-premises costs. The default cost profile for each supported cloud provider uses the provider's published costs. Cost profiles are deployed by associating them with Rafay Blueprints.

Rafay Cost Management Workflows

Rafay provides simple workflows that streamline operations and further increase the power of the Cost Management Service.

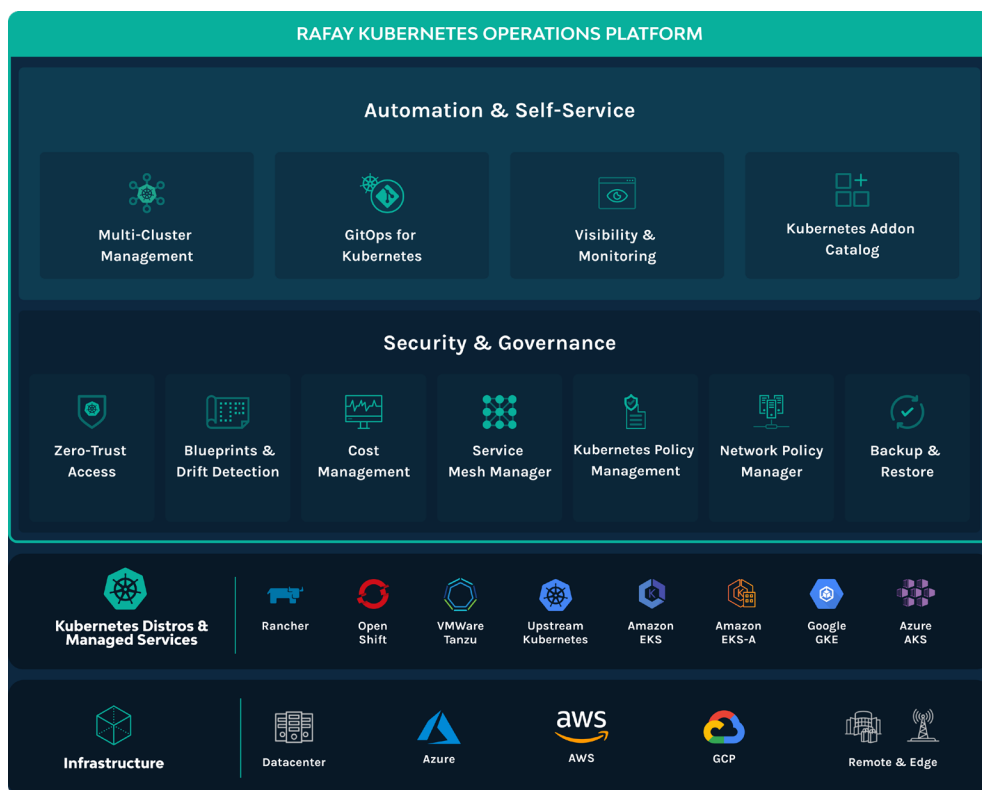
- **Rafay controller workflow:** Using the Rafay controller, you can create or update cost profiles, attach them to a Rafay Blueprint, then sync the blueprint to target clusters.
- **Kubernetes cluster workflow:** The Cost-Model agent in each cluster fetches pricing from the associated cloud provider. It combines that data with resource utilization to generate cost metrics that are exported to Rafay's time-series database and retained.
- **Cost governance workflow:** Rafay blueprints can contain specific cost governance requirements. Governance blueprints are used to set alerts, publish reports, and create chargebacks, automating most of the necessary cost governance tasks across a fleet of Kubernetes clusters.
- **Data pipeline workflow:** You can aggregate raw cost metrics based on defined governance parameters, create alerts, and send alerts to appropriate roles.

To learn more about implementing cost management across your entire Kubernetes fleet, visit [Rafay's Cost Management Service](#) landing page.



About Rafay Kubernetes Operations Platform

Rafay's [Kubernetes Operations Platform](#) helps companies accelerate their app modernization initiatives by eliminating the complexity and expense of building and operating Kubernetes and freeing up staff to focus on product innovation. With Rafay, platform and operations teams quickly roll out an enterprise-grade Kubernetes Operations platform with the automation, governance, security, and visibility features they need.



Sign up for a free trial today and streamline Kubernetes Operations with Rafay

Learn More About Rafay Systems

