IBM Turbonomic on AWS

Unlock the cloud at the lowest cost: AWS Cloud and IBM Turbonomic

As businesses embark on their cloud journey, the goal is often twofold: unlock scalability and keep cloud costs in budget. For many, however, the promise of agility and elasticity is hindered by a lack of insight into what is needed from an application resourcing perspective, leading to poor performance and bills that exceed budget.

In this paper, we'll show you how solutions from Amazon Web Services (AWS) – a reliable and scalable infrastructure platform in the cloud – work with IBM Turbonomic – an AI-powered application resource management (ARM) solution – to meet these challenges.

We'll introduce you to IBM Turbonomic and help you familiarize yourself with key concepts and provide an understanding of how you can optimize your cloud workloads while assuring application performance. You'll see how Turbonomic customers have been able to cut cloud spend by 33% or more and achieve a 470+% return on investment.*

Topics covered in this white paper include:

- About IBM Turbonomic including architecture and supply chains
- Creating targets to monitor workloads
- The four ways to optimize AWS resources with Turbonomic
- Migrating on-premises workloads to AWS

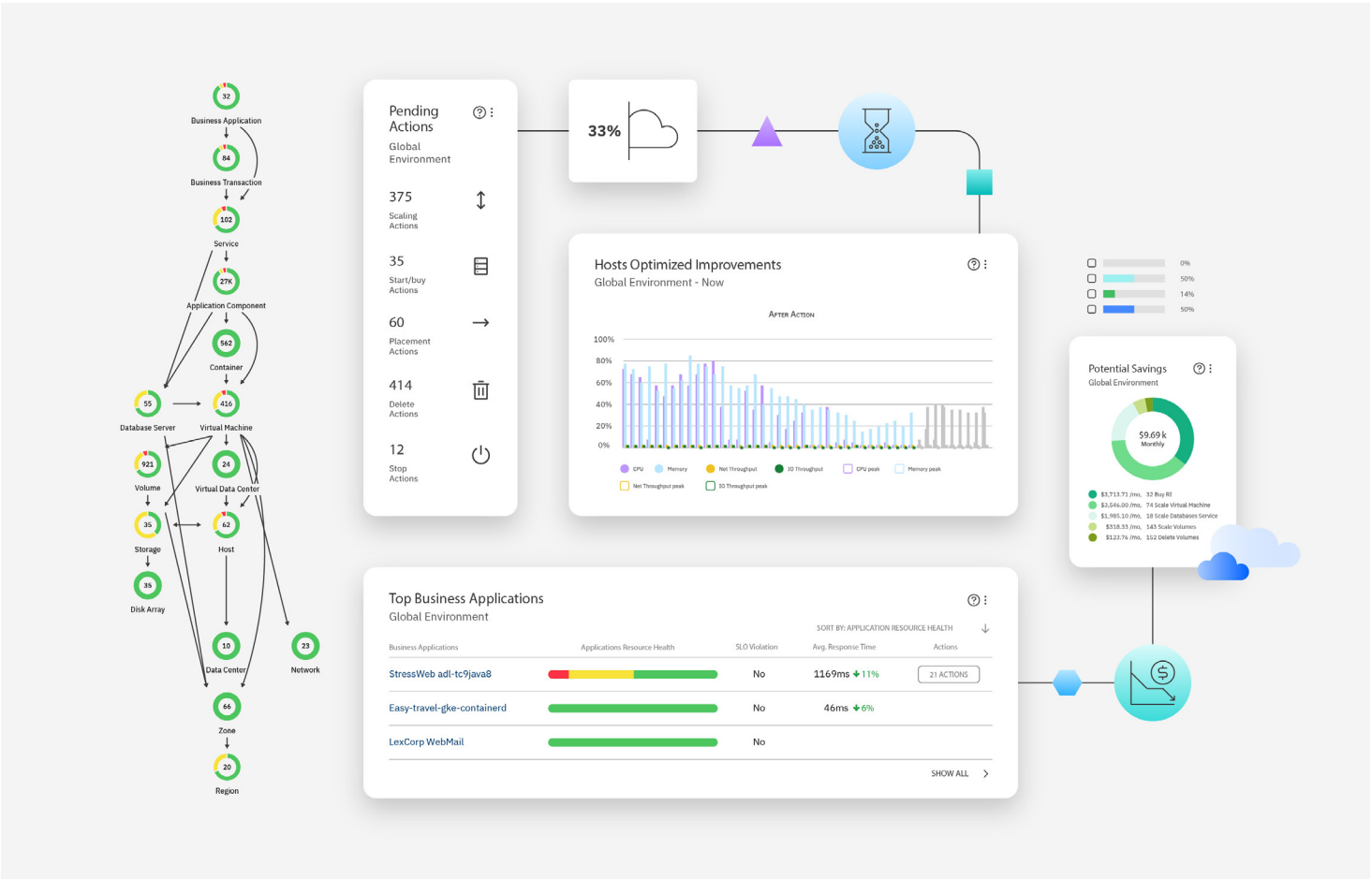

Figure 1. IBM Turbonomic offers full stack visualization of all your applications.

Why AWS and Turbonomic?

IBM Turbonomic offers a comprehensive, end-to-end view of your application and infrastructure stack, both in the data center and in the cloud (Figure 1). You gain real-time insights into the health and performance of applications to identify underutilized resources. This helps you right-size instances and leverage Reserved Instances (RIs) to facilitate application performance while minimizing costs. When integrated with AWS, customers can leverage AWS cloud native tools while assuring that workloads are continuously optimized to keep cloud consumption costs in line.

The result: the most efficient use of compute, storage, and network resources to your apps at every layer of the stack without overprovisioning. You'll only use what you need, resulting in a lower cloud bill and a stronger ROI.

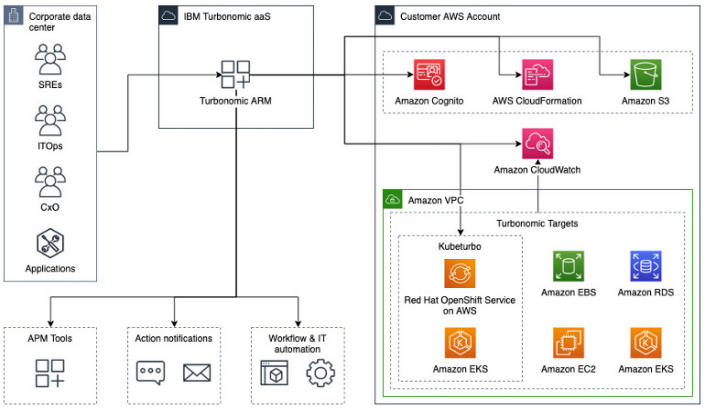

Figure 2. Reference architecture of IBM Turbonomic shows integration with AWS native tools.

Turbonomic architecture overview

Turbonomic is a containerized, microservices application (see Figure 2) available on the AWS Marketplace through self-managed or SaaS offers. It can be deployed on Amazon Elastic Kubernetes Service (Amazon EKS), Red Hat OpenShift Service on AWS (ROSA) or Amazon Elastic Compute Cloud (Amazon EC2).

Turbonomic uses AWS APIs to retrieve metadata and configuration details of AWS resources, including:

- Amazon EC2 instances
- Amazon EKS clusters
- Amazon Elastic Block Store (Amazon EBS) volumes
- Amazon Relational Database Service (Amazon RDS) databases
- Amazon Simple Storage Service (Amazon S3)

It collects performance metrics and monitoring data from Amazon CloudWatch to analyze resource utilization and make informed optimization decisions within your AWS environment.

When used alongside Application Performance Monitoring (APM) tools, like IBM Instana AppDynamics, Dynatrace, New Relic, etc., you can detect risks and correlate application performance with underlying AWS resources. Integration with Slack and email provides real-time notifications, and Amazon Cognito can be used for authentication.

To optimize your infrastructure and proactively prevent application performance issues, automation tasks – like provisioning resources or application deployments – can be done through Turbonomic Actions. In addition, further automation can be orchestrated by integrating AWS CloudFormation, Ansible, or Terraform if needed.

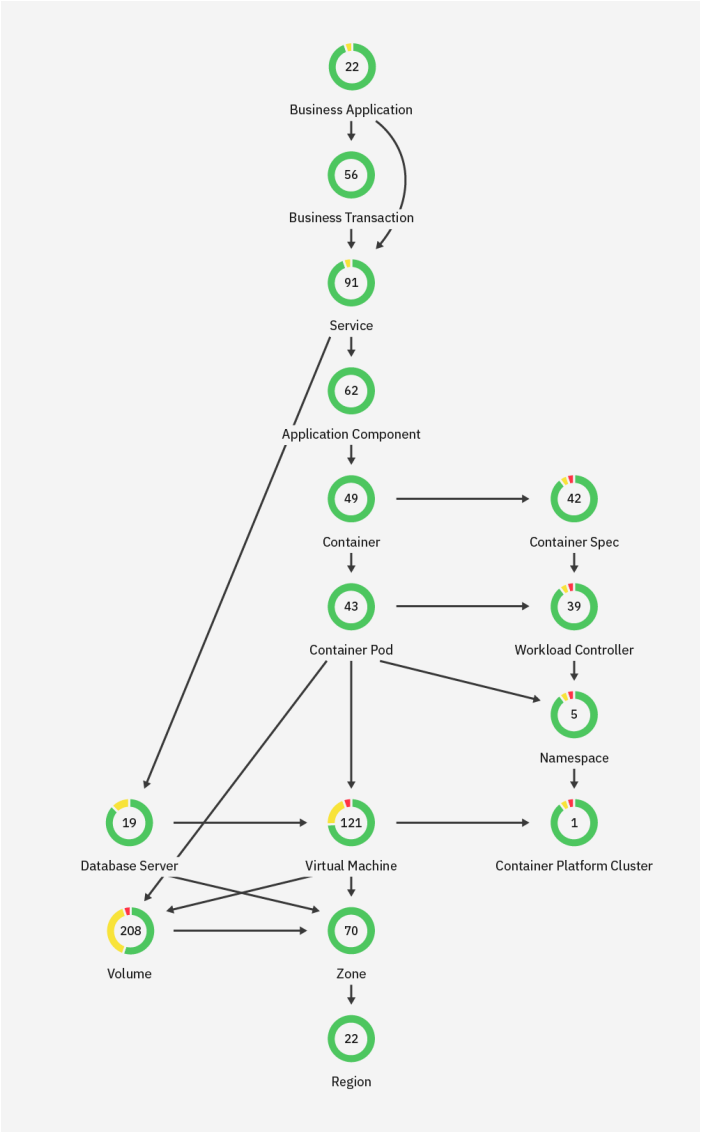

Figure 3. Supply chain of resources and resources mapping to give full view of resources across the entire stack.

Turbonomic supply chain

Turbonomic monitors your entire application stack to maintain environments in their desired states. To achieve this, Turbonomic models environments – from an application-driven perspective – as a market with buyers and sellers forming a supply chain (shown in Figure 3). This represents the relationships between your applications' resources, including components like containers, pods, namespaces, volumes, databases, virtual machines, and others.

Turbonomic's AI-based monitoring tool reviews and allocates a percentage of resources and assigns a value number to them. Each supply chain entity engages in a *supply and demand* relationship for resources, like CPU, memory, and storage. For instance, a host server is assigned physical space, power, and cooling from an AWS Availability Zone, while designating resources such as CPU and memory to EC2 instances. These instances can *apportion* resources (Memory and vCPU) to containers, and the containers can then *allocate* resources to applications.

AWS targets

Turbonomic uses targets to monitor workloads and execute actions on your environment. This enables admins to understand where applications are not properly resourced or are overprovisioned, or workloads that are left running. Turbonomic provides recommendations to correct and uniquely takes actions directly to properly optimize. This keeps costs in line while assuring applications are running smoothly no matter the demand.

To specify your AWS targets, simply provide AWS credentials for your accounts, with permissions allowing Turbonomic to discover, monitor, recommend, and execute actions on your AWS resources.

Create AWS Cost and Usage Reports (AWS CUR) and an Amazon S3 bucket for CUR, to display month-to-day spend and discover RIs utilization.

Enable memory metrics collection by installing AWS Systems Manager Agent (SSM Agent) and Amazon CloudWatch agent on your EC2 instances, granting Turbonomic the ability to recommend actions to optimize memory allocation in your instances.

After configuring your AWS targets, Turbonomic performs continuous analysis of CPU, memory, storage, and network metrics, along with historical data on resource usage and application performance to make recommendations and maintain your infrastructure's desired state.

AWS resource optimization using IBM Turbonomic

AWS customers are migrating to the cloud to leverage AWS native tools like Amazon EC2, EBS, RDS, and EKS to modernize their applications. IBM Turbonomic integrates together with these advanced features to enable the best utilization of application resources, enabling your applications to run fully optimized while keeping costs at a minimum. Let's look at how Turbonomic benefits each Amazon cloud tool.

Amazon EC2 optimization
Amazon Elastic Compute Cloud (EC2) is the most widely used service on AWS. It provides compute capacity in the cloud and has a wide range of virtual machines (VMs) known as EC2 instances. These instances run an operating system on top of resources like CPU, memory, and hard disk. With hundreds of different instance types at various price points, managing AWS EC2 pricing and usage can be a challenging task, but IBM Turbonomic can help identify which virtual machine instances are best suited for workloads.

Turbonomic collects continuous metrics (CPU, memory, network, and storage), instance details, and configuration details from Amazon EC2 instances via AWS API and Amazon CloudWatch. This data enables actions like scaling recommendations and instance type or size optimization.

Turbonomic uses CPU effective capacity. That means it uses benchmark data from spec.org to build a catalog of CPU capacity instead of CPU processor speed to identify the highest performance processor model suitable the workload. This way, Turbonomic can help select the highest performance CPU model that can help lower cost by reducing core counts. Turbonomic also accounts for number of EBS storage devices attached, NICs attached, ephemeral storage, etc. while making the decision.

It also suggests moving workloads from On-Demand Instances to RIs, considering factors like RI inventory, pricing, discounts, disk count, quotas, and AWS Regions capacity for cost optimization.

Turbonomic can scale EC2 instances to enhance performance, reduce costs, or improve efficiency by using different instance types or sizes. Stop and Start actions, whether on-demand or scheduled, temporarily halt EC2 instances to lower cloud expenses.

Turbonomic supports "parking" actions for cloud resources. These actions *stop* your cloud resources for a given period to help you reduce your cloud expenses, and then *start* these resources later when you need them. You can enforce parking actions on demand or according to a schedule. VMs that become idle for a period (for example, after business hours) and those that undergo regular maintenance are ideal for parking. After discovering parkable VMs from a cloud provider, Turbonomic adds the VMs to the Parking page and shows their state (stopped or running) at the time of discovery. You can use the Parking page to execute parking actions or attach parking schedules to VMs. This can result in considerable cost savings for customers.

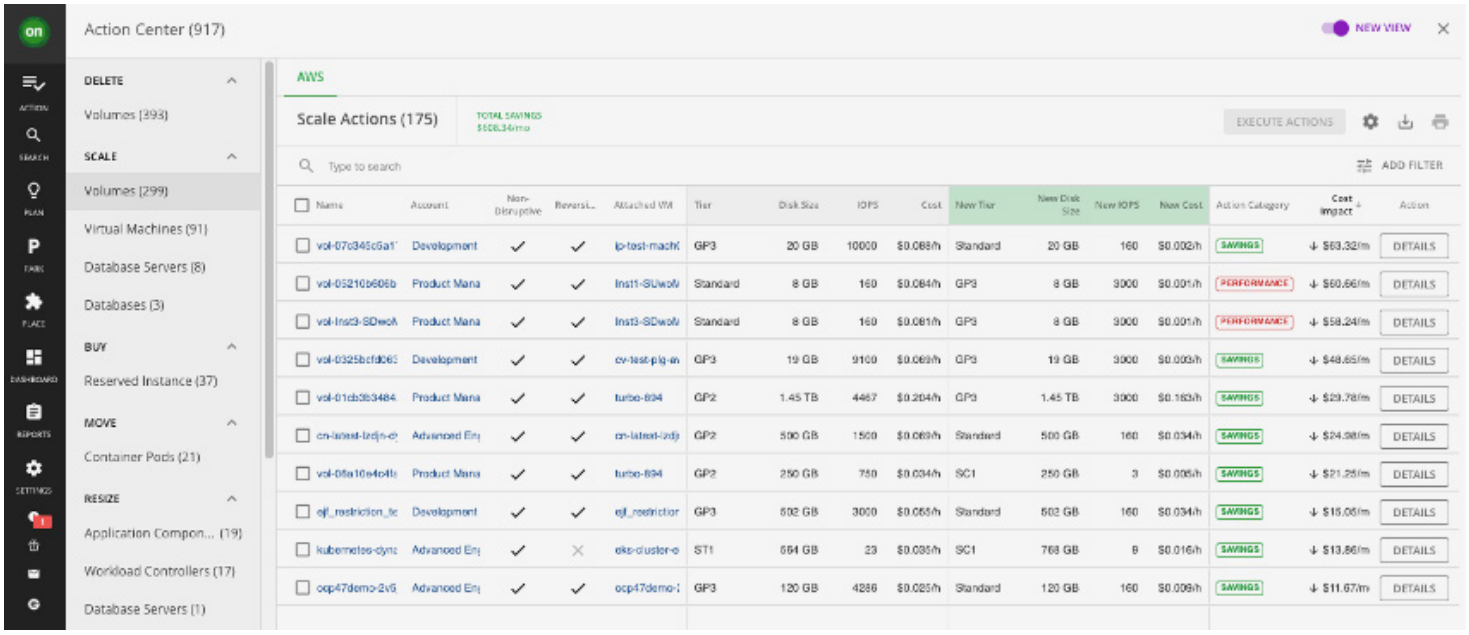

Figure 4. Turbonomic provides suggested actions to optimize Amazon EC2.

Turbonomic's Action Center (Figure 4) shows recommended actions. Information includes AWS account name, EC2 instance types, discount coverage, and On-Demand cost. You can select and initiate different actions.

Details of each action are available from the console. You can see when an instance must be scaled down because it has underutilized vCPU, memory, and network throughput (see Figure 5).

Turbonomic also generates actions to scale applications horizontally to distribute load across multiple instances. This is accomplished in real time through automation and increases performance and reliability.

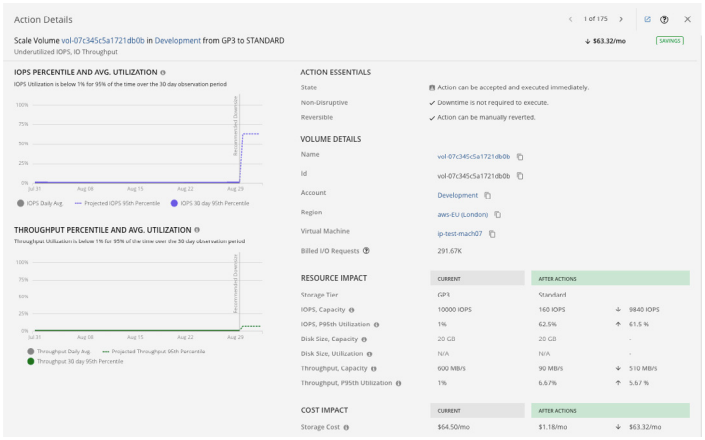

Figure 5. Turbonomic provides details of recommended optimization actions.

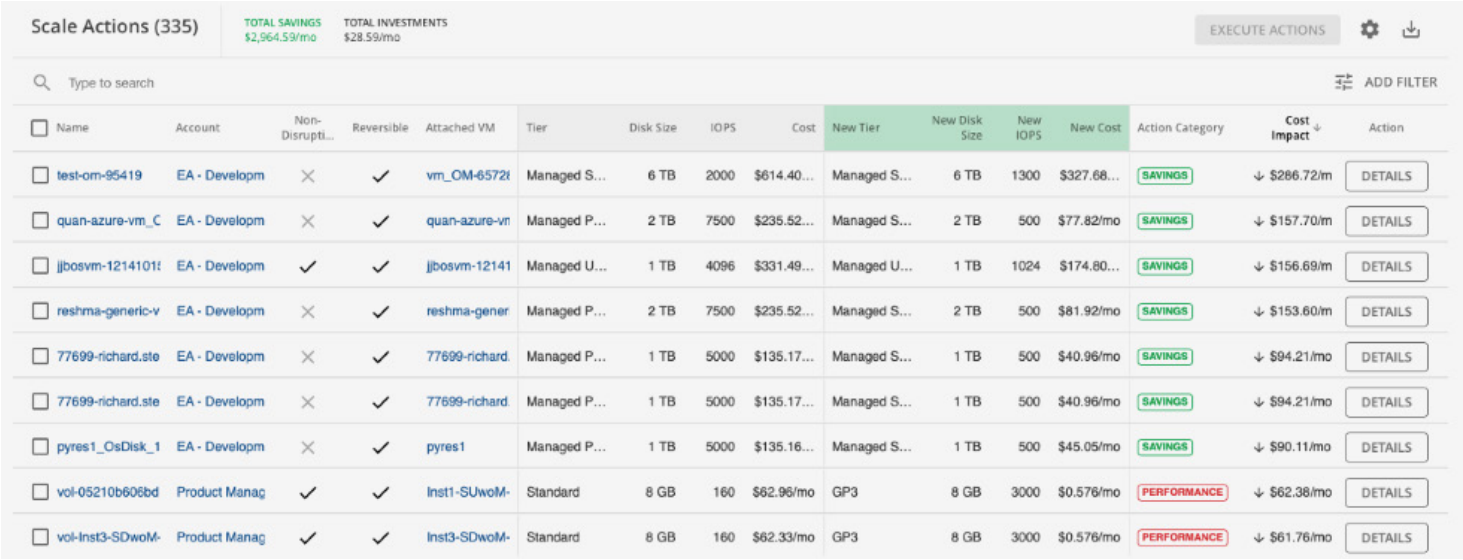

Figure 6. Suggested actions to optimize Amazon EBS.

Amazon EBS optimization

Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with EC2 instances. With IBM Turbonomic, you can accurately determine when to scale between cloud tiers to achieve optimal performance of both input/output operations per second (IOPS) and throughput while achieving cost efficiency.

Based on your policies, Turbonomic recommends and automatically configures Amazon EBS volume size and type to match your workload requirements. This is important because it adjusts resources in real-time, optimizing performance and cost to align with your goals.

You can also fine-tune settings such as I/O, IOPS, and bandwidth. Unlike some of the other clouds, all scaling actions on AWS storage are non-disruptive with zero downtime that allows the customers to execute the Turbonomic recommended Amazon EBS rightsize actions with zero risk.

With real-time visibility into storage utilization, you can modify the size of EBS volumes. You can identify EC2 instances running low on storage and proactively increase volume sizes, before running out of space. Figure 6 shows scaling actions for Amazon EBS volumes that address performance issues or reduce costs.

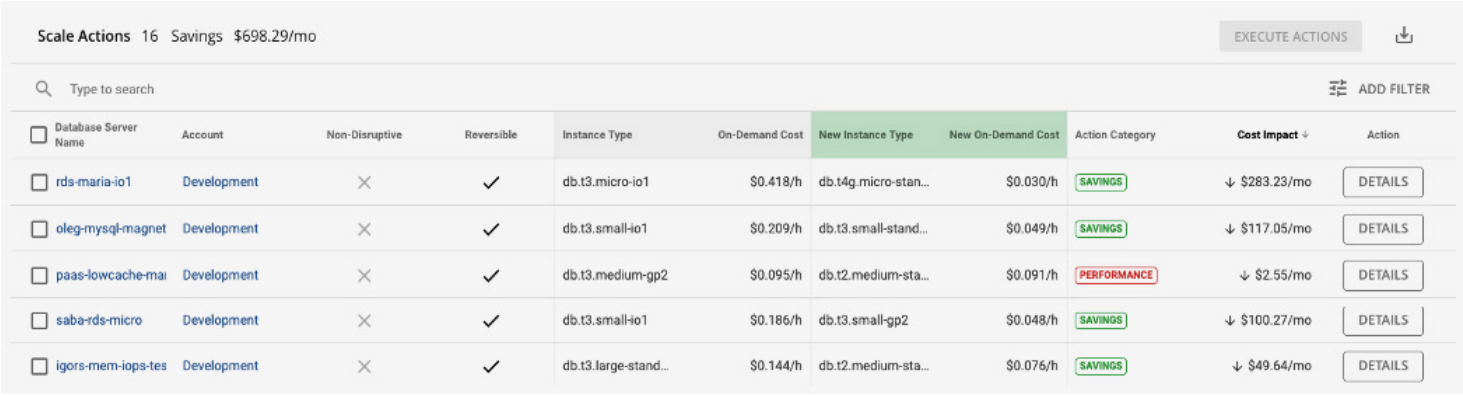

Figure 7. Suggested actions to optimize Amazon RDS.

Amazon RDS optimization

For Amazon Relational Database Service (Amazon RDS), Turbonomic's actions combine compute and storage scaling. It supports Amazon Aurora and Amazon RDS engines for MySQL, MariaDB, PostgreSQL, Oracle, and Microsoft SQL Server. To achieve your desired state, CPU, memory, DB Cache Hit Rate, storage size, and IOPS metrics are continuously monitored for scaling up or down.

Other actions include modifications to compute and storage tiers, storage size, Provisioned IOPS, or a combination of these. Some examples:

- If a performance risk for vCPU is identified, scale up to add vCPU resources
- To get more IOPS capacity and prevent congestion, increase storage amount
- To address underutilized compute resources, reduce vCPU allocation
- Scale down oversized instances, without impacting Amazon RDS performance
- Reduce Provisioned IOPS or move to different storage tiers

Database instances normally use all memory available to them. To make memory scaling decisions on Performance Insights-enabled Amazon RDS instances, Turbonomic considers the DB Cache Hit Rate metric. This permits instance type re-sizing, considering the cache hit rate and memory utilization ratio.

You can visualize all Amazon RDS scaling suggested actions, including information about disruptiveness and reversibility. This helps you understand if service downtime is required when performing scaling actions (shown in Figure 7).

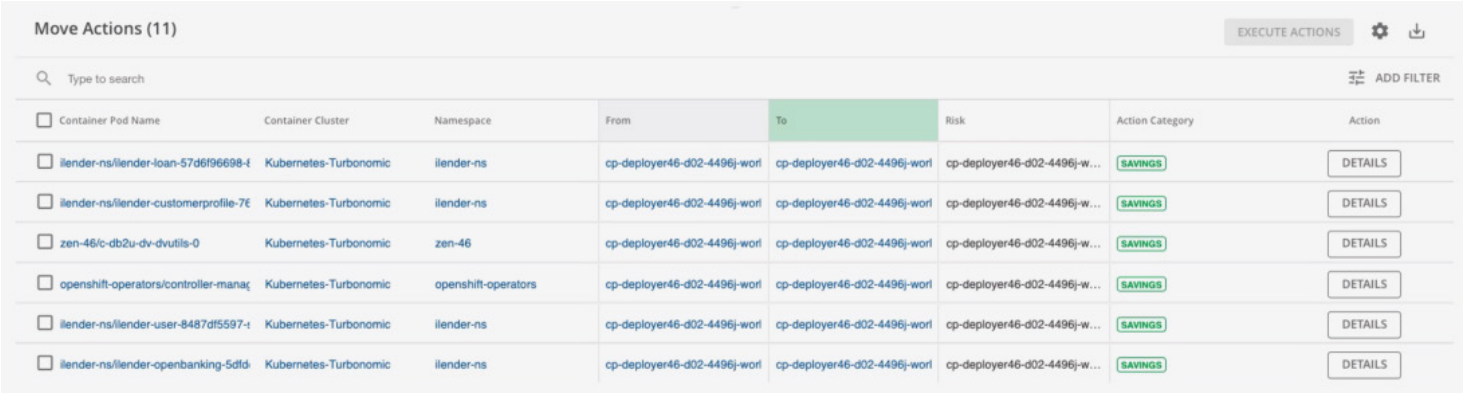

Figure 8. Suggested actions to optimize Amazon EKS.

Amazon EKS optimization

Through Kubeturbo, Turbonomic can discover the entire Amazon EKS infrastructure, including services, containers, container pods, container specs, namespaces, volumes, and nodes (VMs). This gives you visibility into the actions that impact the health of your container environment.

Turbonomic doesn't recommend actions for an Amazon EKS Cluster. Instead, it recommends actions for containers, pods, nodes, and volumes in the cluster, such as:

- Suspend application components, containers, or container pods due to node suspension
- "Resize Heap" actions are possible by integrating with application performance management (APM) tools, such as IBM Instana
- Resize container up or down
- Move pods across nodes
- Provision additional resources, like memory, vCPU, memory requests, vCPU requests
- Suspend nodes
- Change volume tier of nodes

Turbonomic shows these actions when you scope to a Container Cluster and view the Pending Actions chart (Figure 8).

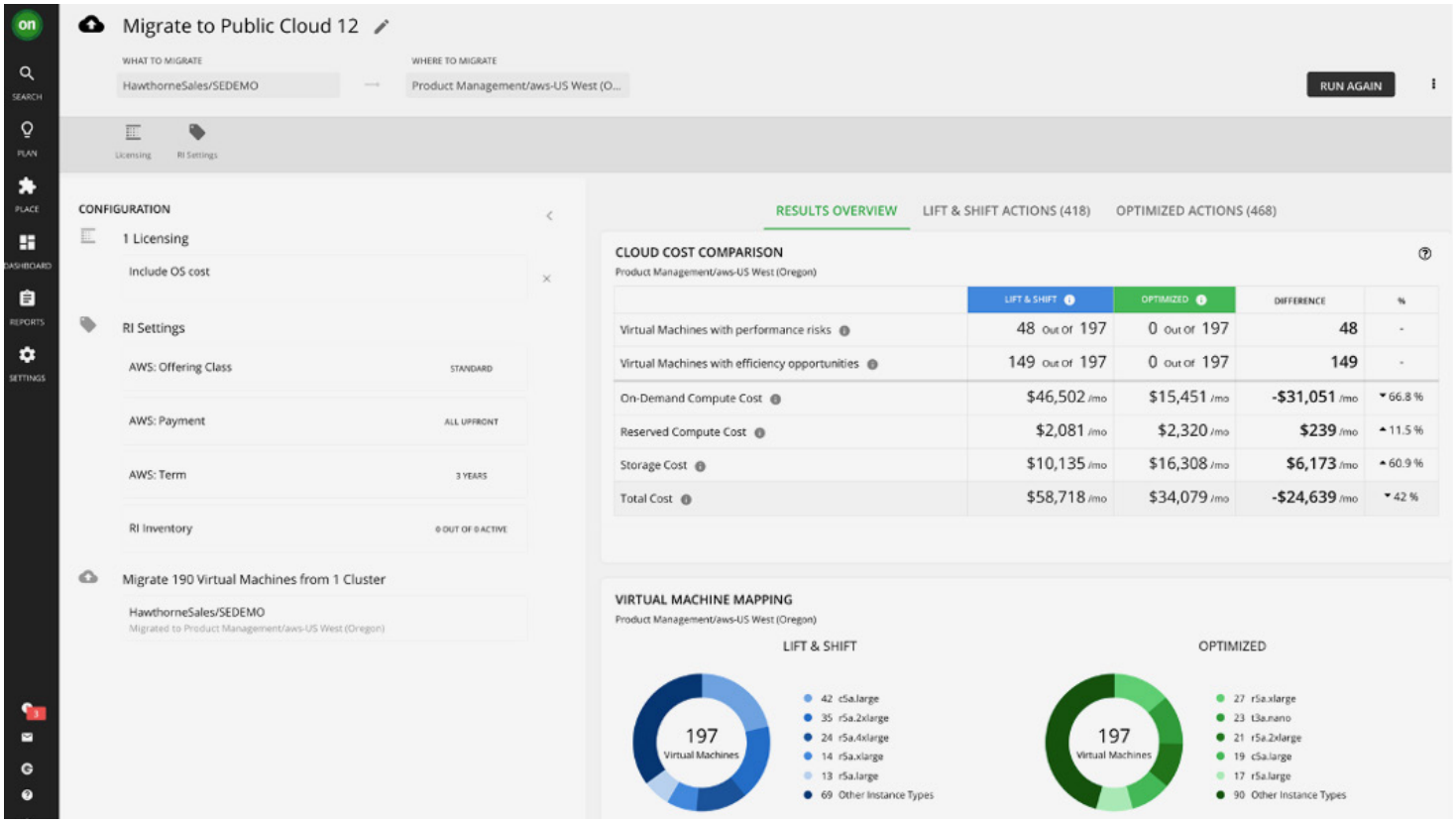

Figure 9. Turbonomic's migration planning dashboard provides easy-to-use cloud migration planning that allows you to optimize your cloud consumption from the start and simplify your cloud migration process.

How to migrate on-premises workloads to AWS

Turbonomic software offers intuitive cloud migration planning, allowing you to simplify and expedite your migration to AWS cloud. Turbonomic provides a "Migrate to Cloud" planning functionality that simulates on-premises virtual machine (VM) migration to AWS Cloud. Select your workloads and their migration destinations, along with discount pricing and licensing preferences. Turbonomic software then comes up with a cost-effective plan optimized for your cloud service needs.

It provides two migration options: Lift & Shift and Optimized. The Lift & Shift option suggests a like-to-like migration of workload from source to the target. The Optimized plan focuses on optimizing performance and costs by choosing the most suitable AWS resources for your VMs and their volumes. It recommends moving workloads from on-demand to RIs and purchasing more discounts plans. Costs are calculated based on negotiated billing, price adjustments, compute, service, and license costs (see Figure 9).

Conclusion

By combining IBM Turbonomic and AWS Cloud, your organization will simplify your move to the cloud and provide continuous application performance while keeping your cloud budget in line. Start today by trying IBM Turbonomic for free on AWS and see how you can unlock the cloud you need… and only what you need.

Next Steps

- Try Turbonomic Today
- AWS Marketplace: IBM Turbonomic
- IBM Turbonomic Product Tour
- Try IBM Turbonomic Sandbox
- Amazon AWS integration with IBM Turbonomic
- Product Documentation

IBM & AWS contributing authors

Debasis Das
Debasis is an IBM Cloud Solution Architect and part of IBM-AWS strategic partnership team. With 25 years of experience in IT, Debasis has focused on large scale application development in the areas of Middleware, Cloud, Kubernetes, Openshift, Automation, AIOPs and Integration. In his current role, Debasis works with partners and customers to drive the value proposition of IBM Software on AWS, through joint assets and solutions.

Eduardo Monich Fronza
Eduardo Monich Fronza is an AWS Partner Solutions Architect. His experience includes Cloud, solutions architecture, application platforms, containers, workload modernization and hybrid solutions. In his current role, Eduardo helps AWS partners and customers in their cloud adoption journey.

Paul Carley
Paul is an IBM Senior Product Marketing Manager for IBM Turbonomic. Paul has over 25 years' experience in marketing for technology, including mobility, virtualization, and sustainability.

Muralidhar Chavan
Muralidhar Chavan is an IBM Solution Architect and part of the IBM-AWS strategic partnership team. He has vast experience in delivering complex software solutions and advocating for IBM technology and products. Muralidhar is focused on the automation space, developing assets and solutions to demonstrate how IBM and AWS are better together.

Tejaswi Nemani
Tejaswi is an AWS Partner Solutions Architect. She is a certified AWS Cloud Practitioner and in her current role, Tejaswi helps AWS partners and customers in their cloud adoption journey.

*Based on Forrester TEI Study: "The Total Economic Impact of IBM Turbonomic Application Resource Management," Forrester, January 2022. Findings based on composite company results in a Forrester Total Economic Impact study commissioned by Turbonomic, an IBM company.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
November 2023

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

All client examples cited or described are presented as illustrations of the manner in which some clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions. Generally expected results cannot be provided as each client's results will depend entirely on the client's systems and services ordered. It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is compliant with any law or regulation.