turbonomic

COST OPTIMIZATION

with

AMAZON WEB SERVICES

EXECUTIVE OVERVIEW

This document is intended to be used as a foundational component of the Powered by Turbonomic series for Amazon Web Services to assist cloud and data center administrators, architects and management teams in understanding the value and approach of the Turbonomic platform in conjunction with supported partner solutions. The reference architecture includes logical and physical design details for both Turbonomic and the associated technology solutions.

The Powered by Turbonomic Best Practices guides and Reference Architecture series are provided as validated solutions to assure application performance in both cloud and data center architectures. This document details the logical and physical interaction between Turbonomic and the underlying infrastructure to deliver the hybrid cloud management platform integration for the partner technology. Turbonomic in conjunction with AWS provides a distinct shared value which we will explore further in this white paper.

INTRODUCTION

Audience

The intended audience for this Powered by Turbonomic reference architecture guide are Cloud Operations, IT Operations Administrators, IT Operations Managers, Enterprise/Solution/Cloud/ Technical Architects, and systems integrators and other solutions providers who are planning to deploy the Turbonomic Platform in conjunction with Amazon Web Services.

Purpose

The overall framework illustrates the three key pillars of value where Turbonomic adds unmatched value and capability when used with AWS infrastructure:

- Assure Performance - Demand-driven placement, scaling, and movement of workloads to assure performance
- Maintain Compliance - Continuously enforce policies to maintain compliance
- Lower Cost - Leverage cloud elasticity to allocate and de-allocate the right resources at the right time

Use Cases

1. Migrate to Cloud – Migrate to the cloud on-time, and on-budget
2. Lower Costs – Lower hybrid cloud costs on-premises and in AWS
3. Assure Performance – Control workloads to assure performance and deliver Service Levels
4. Enforce Policies – Enforce compliance across a hybrid cloud environment

The Amazon Web Services Powered by Turbonomic Architecture is comprised of AWS cloud services and the Turbonomic Platform. This solution guide is focused on the core infrastructure components within AWS including EC2 (Elastic Compute Cloud) and EBS (Elastic Block Storage). Deployment methodologies are included for single region, multi-region, and hybrid cloud options.

THE CHALLENGE OF HYBRID CLOUD

Enterprise are facing incredible pressure to modernize and transform their IT to become a digital business; cloud is central to this transformation. Figuring out what workloads go on-premises, what goes in cloud, and how to manage the transition is challenging – and mistakes are costly. This is exacerbated even further once organization get to the cloud and need to manage the complexity.

The majority of customers are adopting hybrid cloud: a combination of on-premises and public cloud environments. The exponential complexity and scale that these environments create is

Gartner estimates that by 2020, 25% of large enterprises will implement "dynamic optimization" solutions (like Turbonomic) to manage public cloud costs and risks - up from less that 1% in 2016.

Workloads running both on-premises and in the cloud lead to enormous complexity. This presents a challenge in ensuring each workload can achieve a desired state: a state in which performance is assured while lowering costs, and maintaining continuous compliance. There are significant IT challenges that enterprises are facing as workloads and infrastructure are increasingly decoupled.

At the same time, this creates an opportunity to leverage the elasticity of public cloud. Enterprises want to know how to enable elasticity so that workloads can run in the right place at the right time – this cannot be done by humans. It requires software.

Our vision is enabling customers to control any workload on any infrastructure, anywhere, and at any time – and in doing so, achieve a desired state.

Gartner estimates that by 2020, 25% of large enterprises will implement "dynamic optimization" solutions (like Turbonomic) to manage public cloud costs and risks – up from less than 1% in 2016.

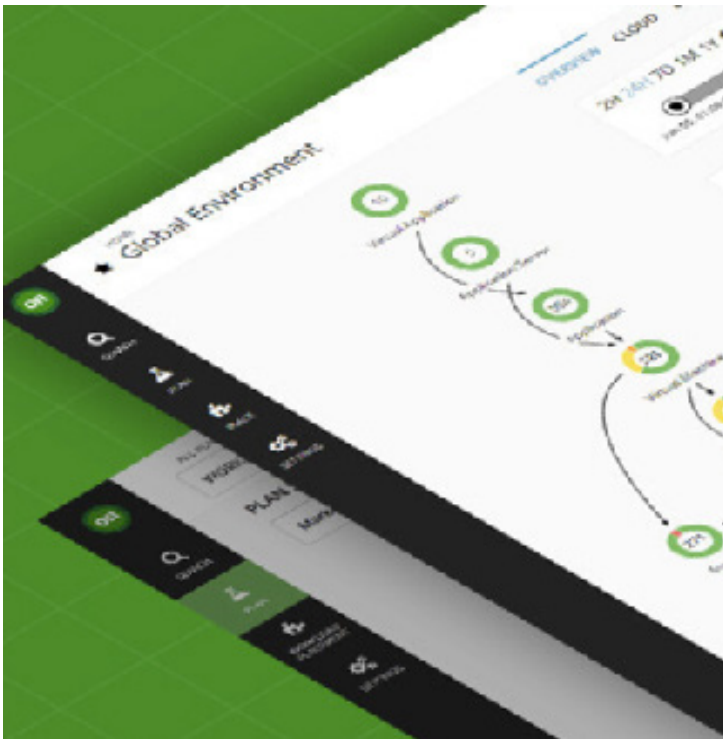
TURBONOMIC HYBRID CLOUD MANAGMENT PLATFORM

To solve the hybrid cloud management challenges, enterprises must start with a platform comprising of three core tenets:

- Abstraction – a data model of the hybrid cloud environment that abstracts away the limitless details and provides a common, semantically rich representation for introspecting and controlling the environment.
- Analysis – an intelligent analysis engine driven by the knowledge captured by the abstraction that makes continuous, real-time decisions to control the hybrid cloud environment in a desired state.
- Automation – orchestrated set of actions driven by the analytics engine to control any workload, on any infrastructure, anywhere all the time, controlling running workloads, deploying new workloads and planning for any future changes and trends.

The resulting system needs to be self-managing and self-regulating to be a truly autonomic system. This is how the Turbonomic approach has proven itself in hybrid cloud environments.

Turbonomic is the only vendor to deliver a platform to control the hybrid cloud, and the only platform where self-managed applications assure their performance autonomously through software and by software, with minimal human intervention only when absolutely needed.



Turbonomic abstracts the hybrid cloud as a Market of Buyers and Sellers, a Market of Service Entities that trade Commodities they consume from each other. Applications, Containers, VMs, Hosts, Zones in a public cloud, Storage, Networks, Disk Arrays, Switches, Etc., are all Services Entities. Compute resources/metrics (such as, memory, CPU, IO, Network, Swapping, Ready Queue, Ballooning, etc.), Storage resources /metrics (such as, IOPS, latency, storage amounts, thin and thick provisions, etc.), Network resources/metrics (such as, Flow, Buffers, etc.), Application metrics (such as, TPS, Response Time, etc.) are all Commodities traded by the Service Entities.

Constraints are also Commodities. Network and Storage Configurations, as well as Business Constraints, such as Compliance, Licensing, etc., are all Commodities traded by the Service Entities. Note that every Service Entity is a Buyer and a Seller.

The Turbonomic analysis engine, the economic engine, uses the Market to control the data center in a Desired State, an equilibrium state where demand is satisfied by the supply, a state in which application performance is assured while the environment is utilized as efficiently as possible. Like in any market, prices are used to control the data center in equilibrium. The Sellers price the Commodities they provide/sell as a function of the Commodities' utilizations while the Buyers shop for the Commodities they consume/buy.

As a Buyer, a Service Entity shops around and decides where to consume the resources it consumes, while as a Seller, it continuously compares its revenue and expenses. When it makes money, it adds/provisions more of the inventory/resources it sells, while when it loses money, it removes/suspends some of the inventory/resources it sells. Each entity makes its own decisions and the hybrid cloud as a whole is controlled in an equilibrium where application performance is assured while maximizing the ROI of the available capacity, on and off prem.

Turbonomic automation uses the Economic Engine to deliver a unified, integrated, platform to control workloads running in the hybrid cloud, deploy new workloads, as well as plan for any future changes, projections and reservations. Turbonomic orchestrates the entire workload life cycles from conception to decommission.

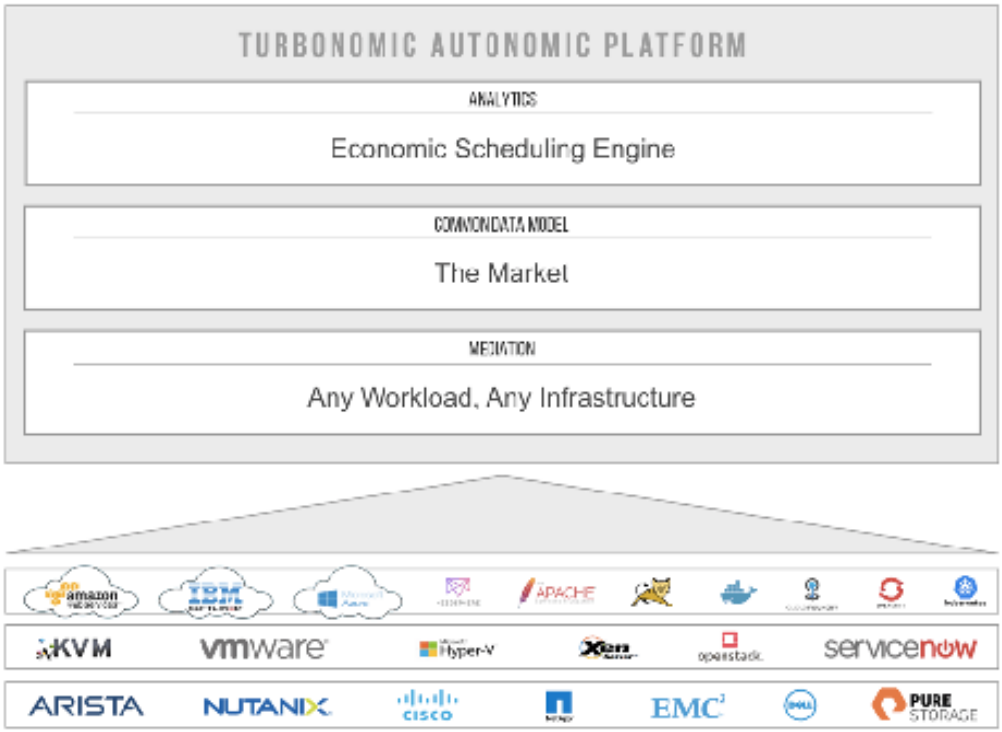
Turbonomic mediates with a broad range of platforms and systems across the entire IT stack and multi-cloud data centers, and maps them to the Market abstraction. Turbonomic mediates with:

- Workloads in Public Clouds (such as AWS, Azure, SoftLayer)
- Physical Compute (such as Cisco, HPE, Dell)
- Virtual Compute (such as ESX, Hyper-V, KVM, RHEV)
- Physical Network Devices (such as Cisco, Arista, Juniper)
- Virtual Networks (such as ACI, NSX)
- Physical Storage Devices (such as EMC VMAX, EMC VNX, NetApp, Pure, ExtremeIO, ScaleIO, HPE 3Par, HDS)
- Virtual Storage (such as vSAN)
- Converged and Hyperconverged Platforms (such as UCS, Nutanix, HPE)
- IaaS Platforms (such as OpenStack, vRealize Automation, CloudForms, System Center)
- Containers (such as Docker, Rocket)
- Containers as a Service (CaaS) Platforms (such as Kubernetes, Mesos)
- Platform as a Service (PaaS) Platforms (such as OpenShift, Cloud Foundry)
- Applications (such as VDI, WebSphere, WebLogic, JBOSS, MySQL, Microsoft Exchange)

By mediating with all of the above, Turbonomic is the only platform orchestrating resource allocation across the entire hybrid cloud and, as such, it is the only platform that assures application performance.

Turbonomic integrates (via an open REST API) with a variety of Service, Infrastructure, Cloud and Access Management systems to support an integrated full workload life cycle.

Turbonomic is the only platform to deliver an integrated Topology Scope, Data Collection and Management Function making it the only platform that truly solves the hybrid cloud management problem of how to assure performance while lowering costs and maintaining continuous compliance.



The market-based algorithms allow every entity to figure out placement, scaling, and start/stop actions for themselves — assuring application performance, while lowering costs and maintaining continuous compliance. Solving the challenge of hybrid cloud workload management in software allows IT teams and their environments to scale smarter.

OPTIMIZING AWS INFRASTRUCTURE

Using the Turbonomic Platform with AWS leverages the elastic, on-demand capabilities of AWS while delivering the industry leading application performance, cost optimization, and compliance enforcement capabilities of Turbonomic. This combined solution delivers performance for your application and cloud infrastructure while ensuring that you control costs, while maintaining compliance for your cloud and data center environments.

The entire value behind the on-demand consumption model is wrapped around leveraging supply and demand concepts to scale your capacity as needed. While the infrastructure in AWS provides the tooling and framework to expand and retract your infrastructure on-demand, the tooling does not provide a way to assure application performance while also reducing the cost of your cloud footprint, and while also maintaining many of your compliance requirements.

What is missing is a software platform that can continuously drive the right decision in real-time of what workload to run where and when, and how to properly configure the workloads, controling them in a desired state.

Let's look at the four use-cases which illustrate how Turbonomic assures performance, lowers costs, while maintaining compliance for your AWS environment.

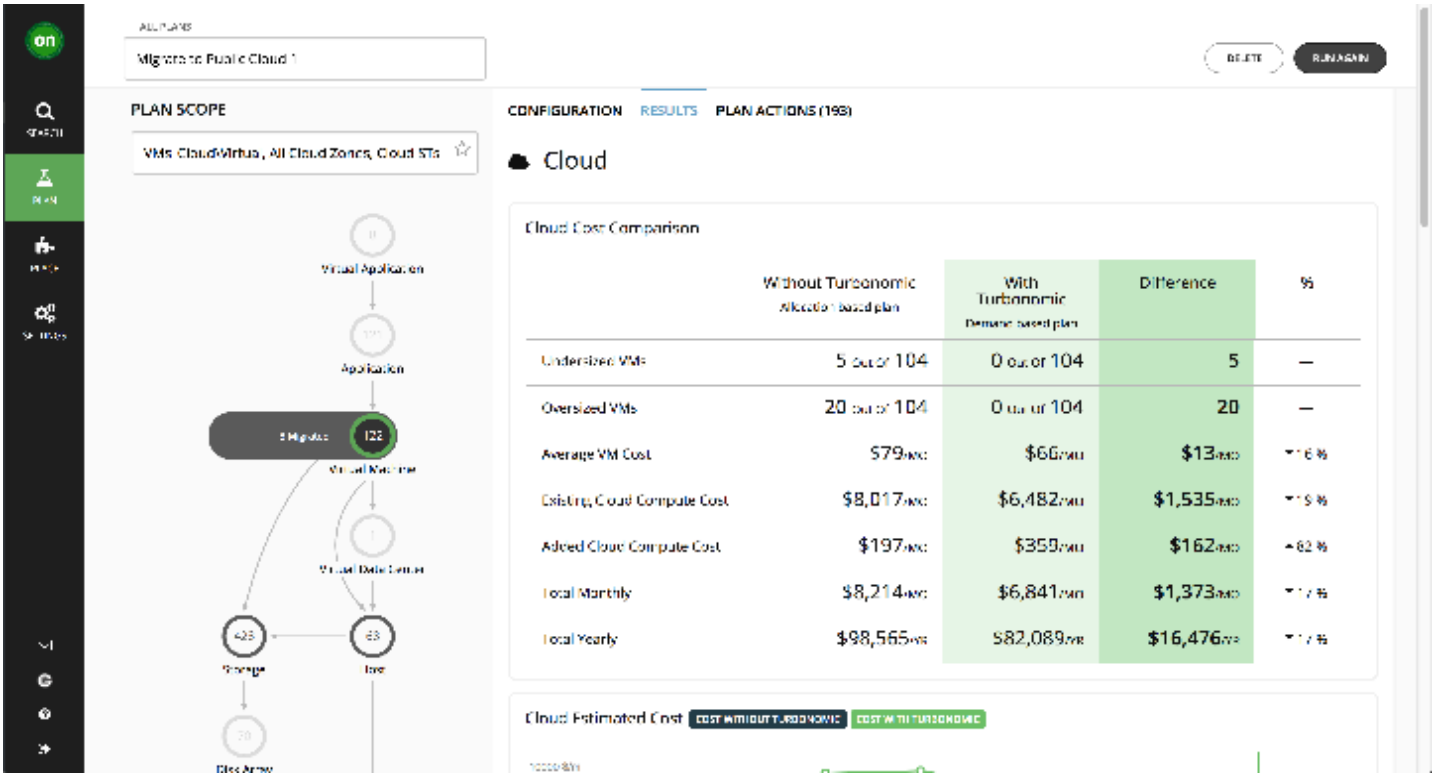
Migrate to Cloud

Migrating your applications to the AWS cloud can only be done in the best way when accounting for performance, cost, and compliance across your hybrid cloud infrastructure. Using Turbonomic allows you to plan the migration, properly place and scale the workloads, and retire your on-premises resources to enable you to migrate on-time, within budget, and while assuring both performance and compliance.

Migration Planning	Accurately matches what you consume on premises to what you need to consume in public cloud to assure performance while lowering cloud costs
Workload Placement	Leverage actual workloads on premises consumption and performance characteristics to place workloads in the right zones and regions to deliver reliable performance at the lowest cost
Workload Scaling	Automatically size workloads based on on-premises consumption for AWS or Azure
Resource Utilization	Identify compute and storage that can be safely retired or repurposed for on-premises workloads after migrating to the cloud

Using Turbonomic for your cloud migration ensures that workloads are sized to assure performance, while maintaining compliance, and while utilizing the lowest cost cloud infrastructure to deliver on your application requirements.

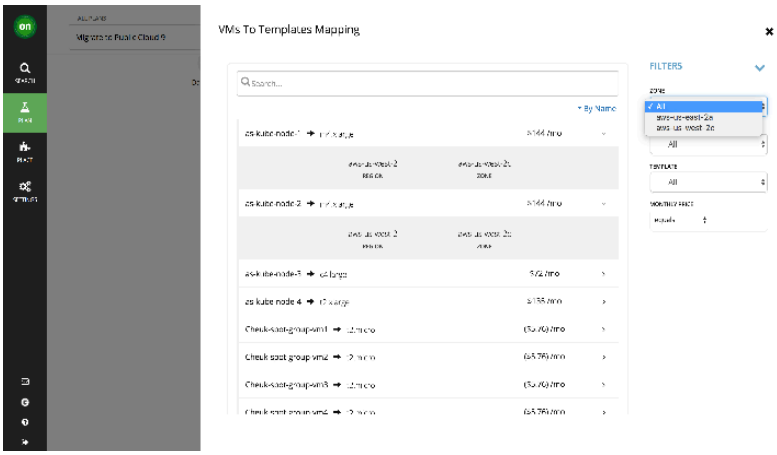
The cloud migration plan allows you to select any part of your on-premises environment to evaluate the move to the public cloud. Turbonomic uses real-time and historical performance to automatically choose the optimal cloud templates that will meet the demand of your application for CPU, memory, storage, and networking.



Plan results are delivered with a view of how a direct migration with the closest template match as a lift-and-shift migration, plus the results with which you would get by letting Turbonomic scale the environment before you migrate.

The entire cloud cost is provided as a part of the results to ensure that your migration is clearly defined by the value that Turbonomic delivers. Not only do you see the costs and savings that you would achieve with Turbonomic for the overall migration, but the detailed per-instance data is also provided.

Filters are available to select more granular views of the results, and the plans can be saved for later review. All data from the plans is accessible using the RESTful API, and plans can be initiated at any time through the Turbonomic interface or the API.



Lower Costs

Along with performance and compliance comes the very important trade off with the cost of public cloud infrastructure. On-demand infrastructure provides the resources to meet the needs of the application, but there is a distinct trade off that we must be aware of. Throwing supply of resources at an application in the hope of adding performance is both ineffective and very expensive.

As described in the Cost Optimization with AWS guide:

> *"Unlike traditional IT infrastructure models, AWS is, by its very nature, elastic and on-demand. AWS provides mechanisms to programmatically scale up and down, as needed, or to implement lifecycle rules that archive or expire storage objects automatically. Consider these features and services to help you achieve a cost-optimized architecture."*

Using Turbonomic in your AWS environment ensures that you are delivering the best application experience using policy-driven SLA continuously in real-time while also ensuring that the cost of your AWS infrastructure is being operated within your budget requirements. Only the Turbonomic platform can provide the real-time capability to match the demand of the application to the supply of infrastructure for both performance and cost.

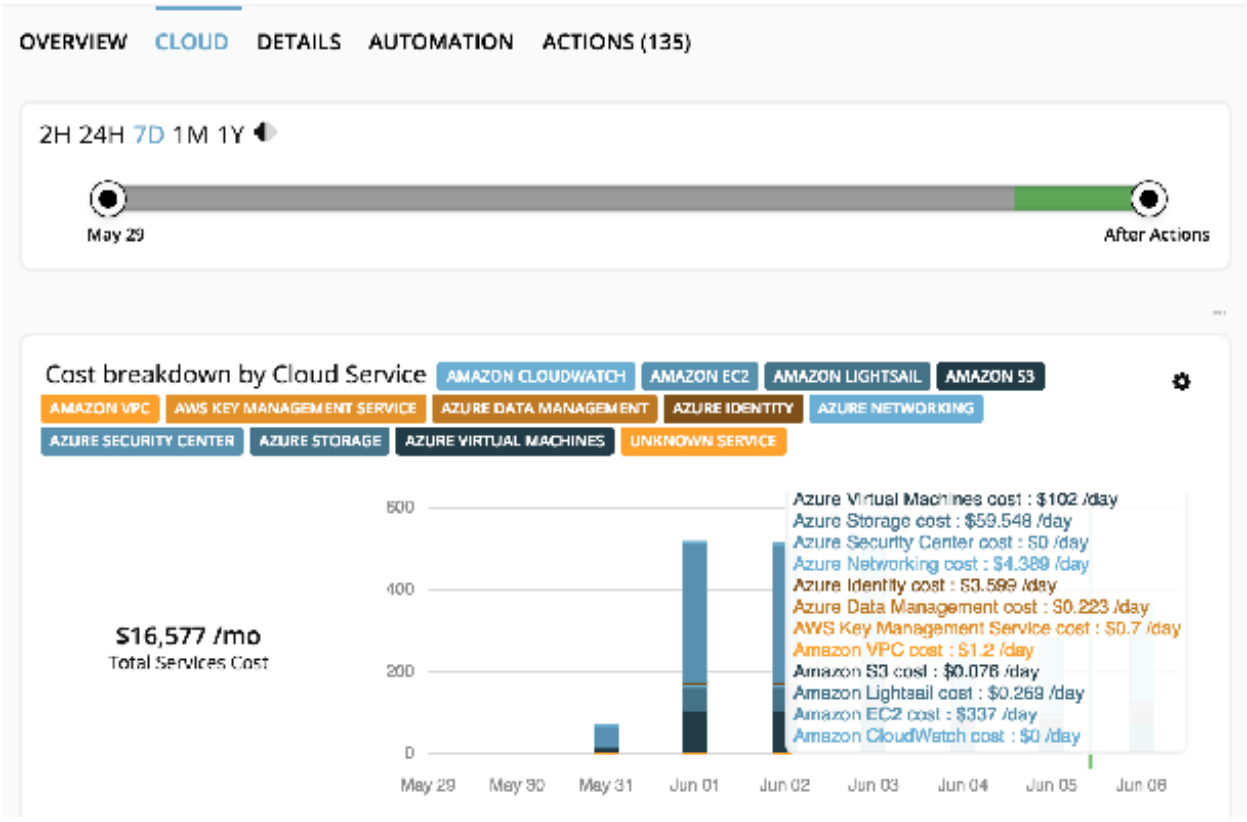
Rather than using fixed thresholds and human intervention to decide on when to scale and contract your infrastructure, Turbonomic dynamically scales your workloads to meet application demand while also staying within the constraints of your budget means. This extends the features natively available within AWS through the Turbonomic environment and uses our patented algorithm to provide the highest performance in the most cost-optimized way.

Turbonomic delivers the lowest cost to run your applications while also assuring performance and ensuring compliance not only for Day 1 deployment, but continuously for Day 2 and beyond.

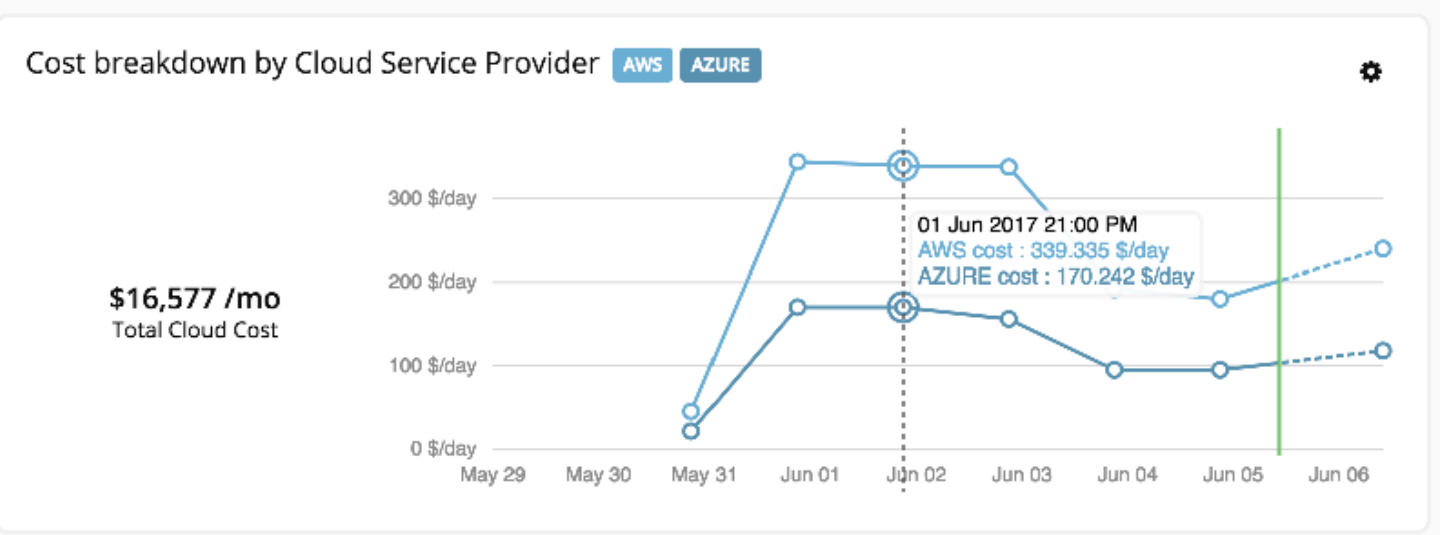
Reduce Costs	Automatically scale down workloads in AWS, Azure or on-premises reducing costs without impacting performance
Budget Control	AWS and Azure bills are aggregated across services, regions, accounts and lines of business and tracked in a single view against defined budget
Workload Specific Cost	AWS and Azure workload specific costs are continuously tracked and reported by regions, tags or custom groups
Cost Forecasting	Project future monthly AWS and Azure bills to accurately plan and budget
Cost Management	Analyze AWS and Azure expenses continuously to help track what is being spent and prevent unexpected bills

Cost is not only measured by template size, but also by region, and by availability zone. Limiting East-West network latency among the many dimensions of performance also ensures that the cost and efficiency are continuously optimized to meet the needs of the application and your cloud budget.

Turbonomic provides a top-level view in real-time of your complete cloud cost across your entire hybrid cloud:

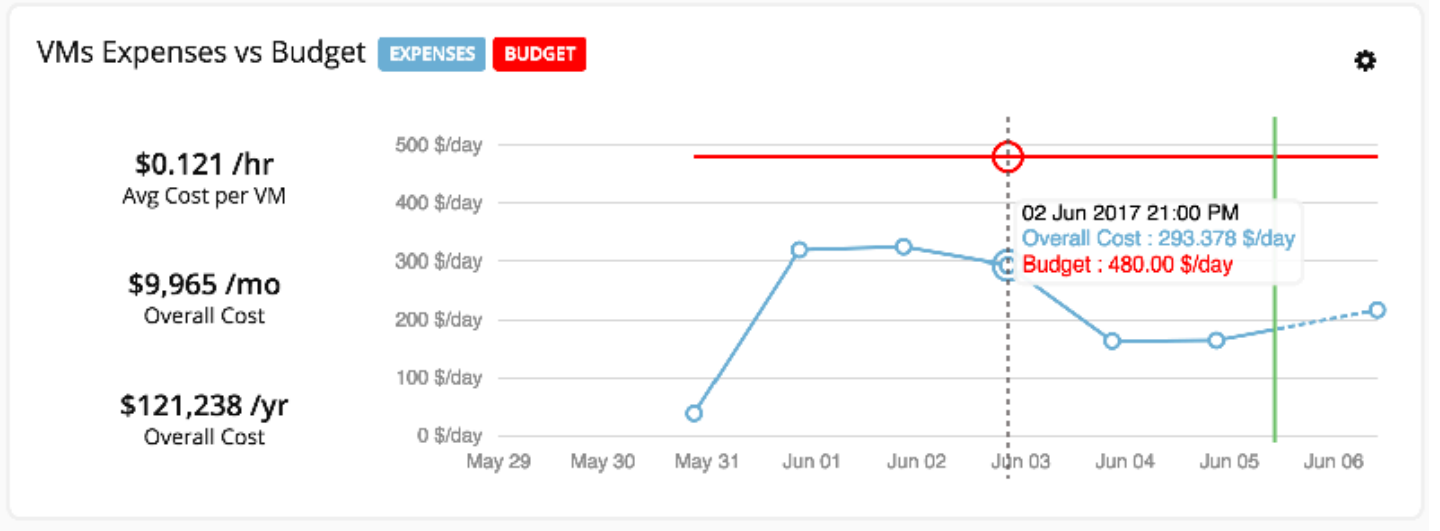


Cloud costs may be also viewed by cloud provider:



Within each provider, costs are also provided per each account within the provider. This is especially helpful when multiple departments have their own AWS credentials. Having the complete cost view for your team ensures that all decisions and actions taken around assuring performance and scaling your infrastructure are done using Turbonomic automation capabilities and with a complete view of the hybrid cloud infrastructure.

Each cloud account may also have a budget associated to ensure that actions with Turbonomic are made within the context of assuring performance while maintaining both compliance of your workloads and your budgets.



Assure Performance

Assuring performance in dynamic environments can only be done with the use of a real-time software platform. Performance is about delivering the best end-to-end experience for the application while also utilizing the underlying infrastructure as efficiently as possible while maintaining continuous compliance.

The most common and poor practice in IT today is the overprovisioning of virtual and physical hardware under the assumption that overhead provides performance and efficiency. In fact, it is quite the opposite.

Using every available dimension of performance measurement available with Turbonomic, the applications choose the appropriate template, region, and availability zone to assure the performance of the application when deploying to, or migrating to the AWS public cloud.

Turbonomic provides both performance assurance and the best hybrid cloud visibility for all your workloads.

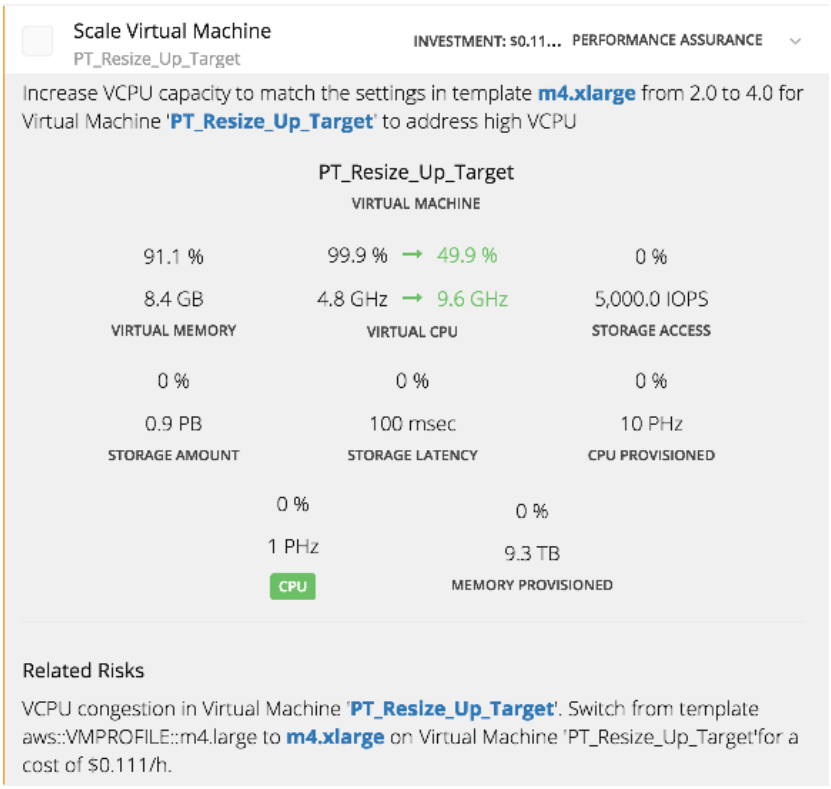
Hybrid Cloud Visibility	Single pane of glass for resource consumption across on-premises data centers, AWS and Azure environments
Cloud Workloads Visibility	Key performance metrics of workloads in AWS and Azure environments are racked, reported and trended, including Compute and Storage resources (CPU, Memory, IOPs and Latency), across cloud providers, regions, zones.
Auto-Scale	Automatically scale workloads across on-premises, AWS or Azure environments to assure performance
Demand Curation	Without agents, connect to the applications and use metrics collected (e.g. connenctions, heap, threads, response times, transaction rates) to ensure applications get the resources they need when they need them to align with Service Levels on-premises or in the cloud

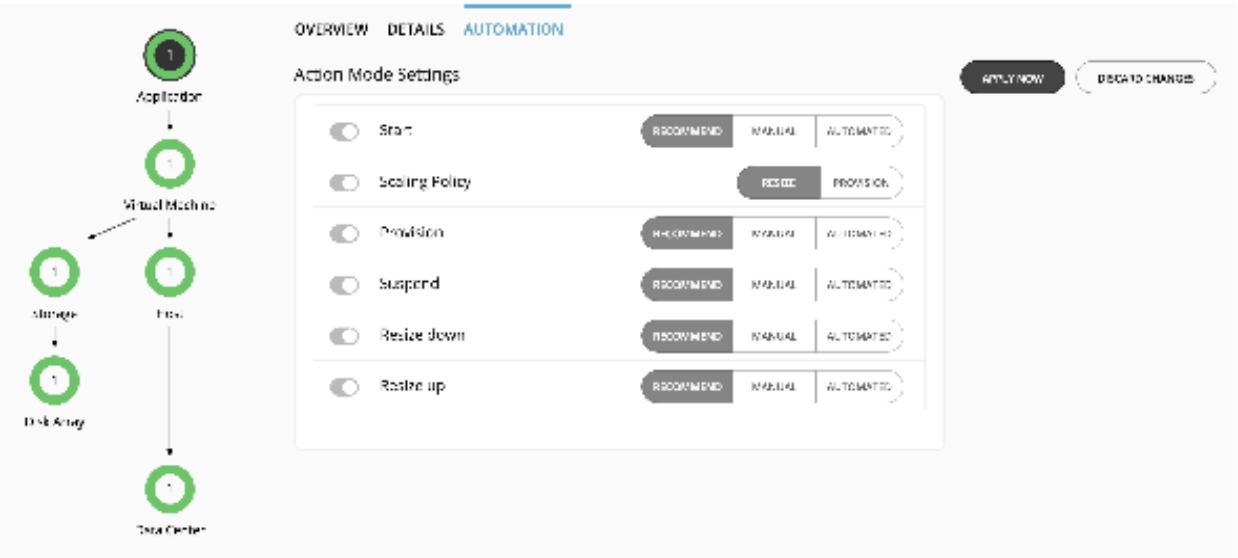
Turbonomic is the only platform with the ability to know when to scale, and to automatically scale your AWS infrastructure, to deliver on the three key pillars of assuring performance, maintaining compliance, and ensuring the lowest cost infrastructure while doing so.

Using Turbonomic will let you deliver on your application SLA using real-time performance requirements across CPU, memory, storage, and networking requirements while also allowing for placement and budget controls to drive the environment to its desired state.

Actions within Turbonomic may also be automated to assure performance continuously in real-time without needing intervention from your team.

Using Turbonomic will let you deliver on your application SLA using real-time performance requirements across CPU, memory, storage, and networking requirements while also allowing for placement and budget controls to drive the environment to its desired state.





Every part of your hybrid cloud infrastructure may have automation policies to allow for a granular control of which actions can be taken for you. All actions are logged both on the Turbonomic instance and within your infrastructure management product.

Enforce Policies

Assuring compliance on AWS infrastructure is also a key pillar of the Turbonomic Platform. Using Turbonomic policies will assure performance and optimize for the lowest cost for your AWS infrastructure while also maintaining compliance for use-cases including, but not limited to:

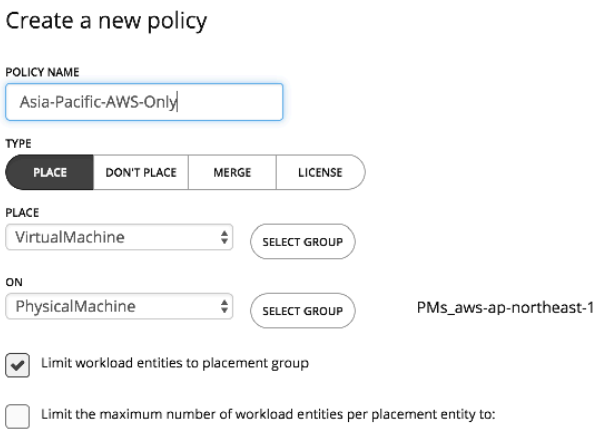
- Data sovereignty
- License compliance
- Regulatory restrictions

Among these use-cases, Turbonomic provides a policy-driven approach to ensure compliance while also assuring performance and lowering cloud costs in the context of your compliance requirements.

Anti-Affinity Enforcement	Designated HA workloads are spread across multiple regions and availability zones or data center, cluster and hosts on-premises complying with risk management specifications for mission critical applications
Data Sovereignty Enforcement	Workload placement and movement is limited to sanctioned cloud providers/regions or on-premises data centers and clusters
Placement Enforcement	Workload placement policies are imported (or defined) and enforced for initial placement and moves across hybrid environment

Turbonomic is the only platform with the ability to know when to scale, and to automatically scale your AWS infrastructure, to deliver on the three key pillars of assuring performance, maintaining compliance, and ensuring the lowest cost infrastructure while doing so.

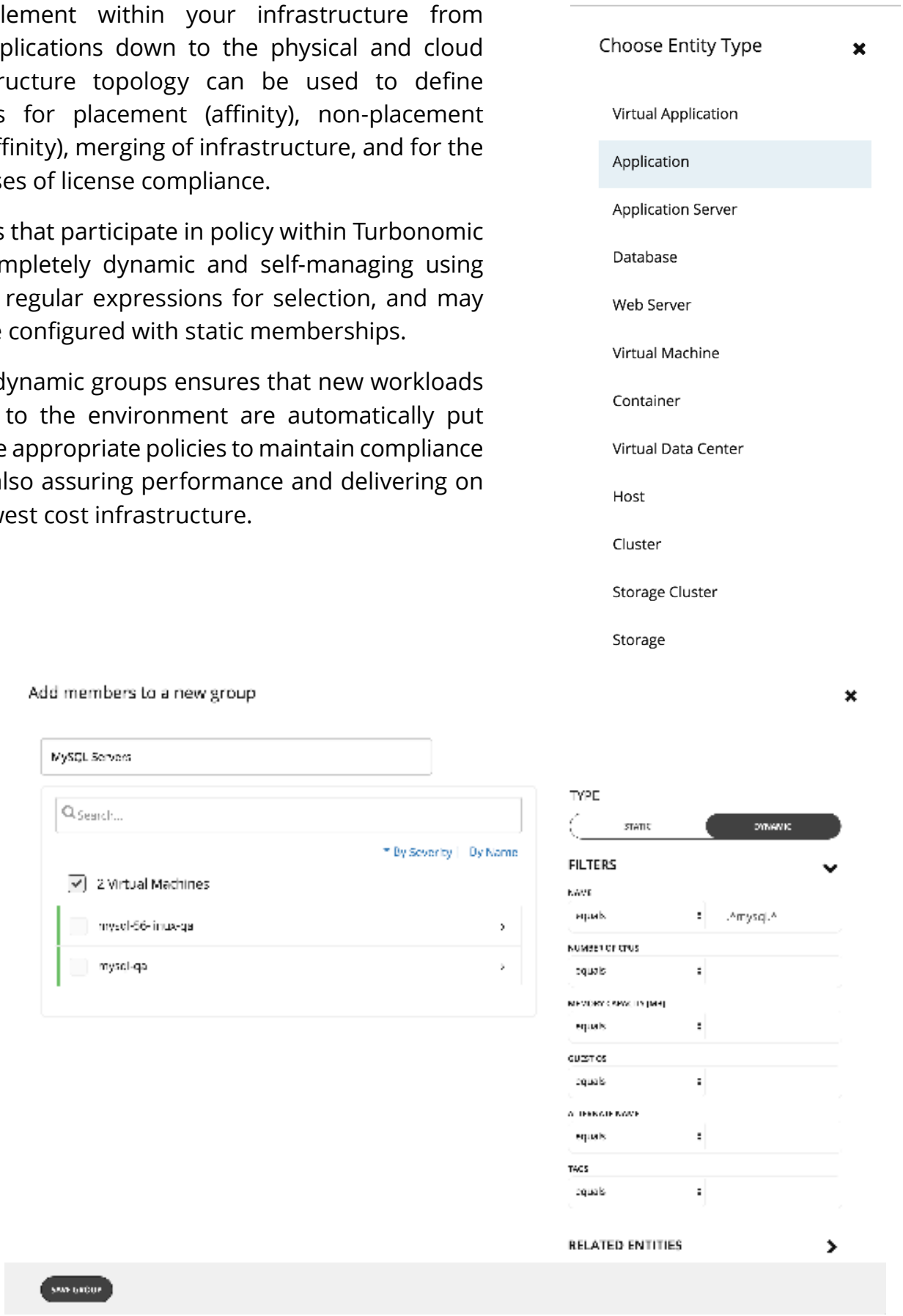
Turbonomic provides a policy framework to maintain compliance both on-premises and in the cloud. Here we can see a simple placement policy example that ensures workloads stay within a public cloud region which could be for data sovereignty or data locality requirements:



Any element within your infrastructure from the applications down to the physical and cloud infrastructure topology can be used to define policies for placement (affinity), non-placement (anti-affinity), merging of infrastructure, and for the purposes of license compliance.

Groups that participate in policy within Turbonomic are completely dynamic and self-managing using simple regular expressions for selection, and may also be configured with static memberships.

Using dynamic groups ensures that new workloads added to the environment are automatically put into the appropriate policies to maintain compliance while also assuring performance and delivering on the lowest cost infrastructure.



DEPLOYING TURBONOMIC FOR AWS ENVIRONMENTS

There are multiple methods to deploy Turbonomic to assure performance for your AWS infrastructure. Turbonomic is available as both an on-premises virtual instance that can be deployed into your existing data center hypervisor platform, and directly from the AWS Marketplace as a cloud-hosted instance..

Deploying Turbonomic On-Premises

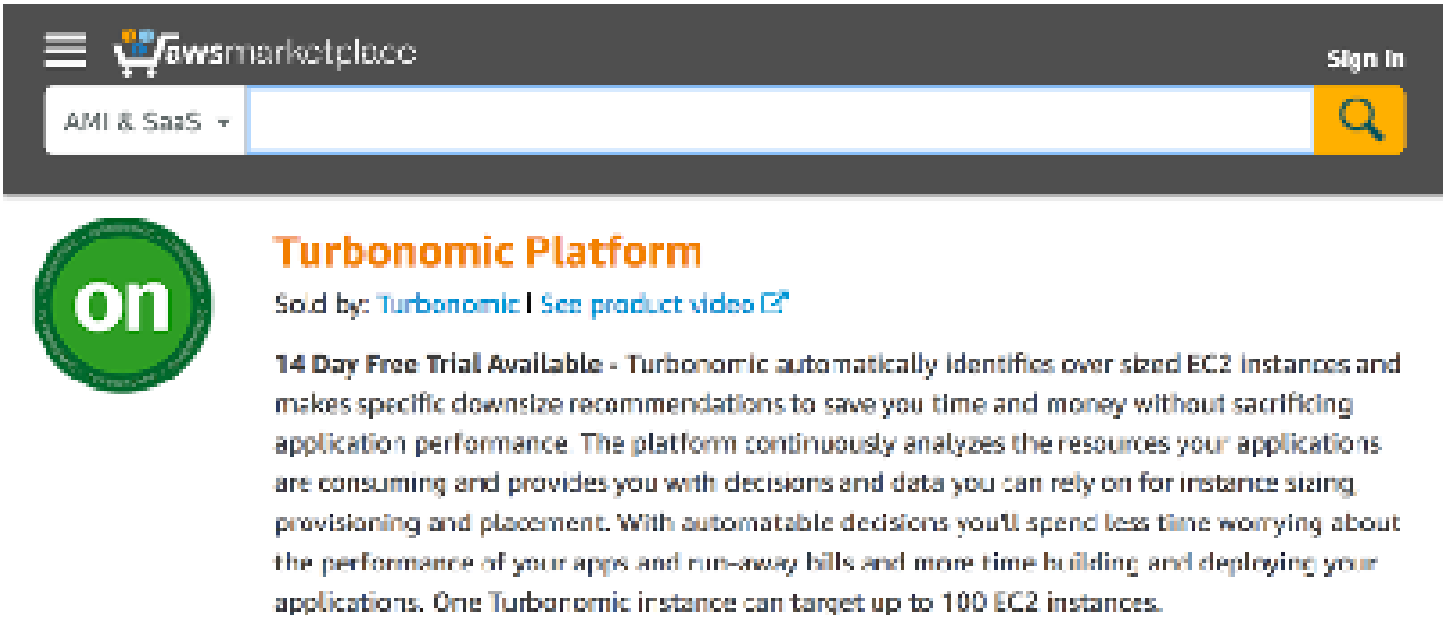
Leveraging your current on-premises environment to host your Turbonomic Autonomic Platform instance allows you to simply deploy with your current hypervisor and add AWS as an additional target.

Simply follow the instructions provided along with the download of the Turbonomic instance for your chosen hypervisor. Each hypervisor has requirements for minimum virtual hardware to operate your Turbonomic instance.

Support Hypervisor Technology	Storage Requirements	Memory	CPUs
VMware: vCenter versions 4.x - 6.0 running with ESX 3.x, 4,x, or 5.x	150GB or greater disk storage + swap space o match the RAM allocation (for example, 150GB + 16GB = 166GB)	16GB	2 vCPUs - 4 vCPUs preferred
Citrix: XenServer versions 5.6.x and 6.x	150GB or greater disk storage	16GB	2 vCPUs - 4 vCPUs preferred
Microsoft: Hyper-V as bundled with Windows 2008R2, or Hyper-V Server 2012	150GB or greater disk storage	16GB	2 vCPUs - 4 vCPUs preferred
Red Hat Enterprise Virtualization: RHEV 3.x	150GB or greater disk storage	16GB	2 vCPUs - 4 vCPUs preferred

Turbonomic may be downloaded for deployment at any time from the Turbonomic web site at https://turbonomic.com/downloads/ which provides the deployment package to match your hypervisor. One Turbonomic instance may be used to control any hypervisor and public cloud target.

Deploying Turbonomic on AWS



Connecting the AWS-hosted Turbonomic instance to your on-premises infrastructure will require a VPN to the VPC in which your Turbonomic instance is running. Once deployed and connected via the VPN, you simply add targets using the Settings option within Turbonomic.

Your Turbonomic instance must be updated to version 5.9 or later in order to get full functionality for cloud control.

NOTE: Running Turbonomic through the AWS Marketplace will incur AWS EC2 charges which are described fully in the Marketplace site. These sample prices are current as of the release of this document:

Once you have followed the instructions in the AWS Marketplace to deploy your Turbonomic instance and log in to the console, you may add your on-premises targets.

CONCLUSION

Using the Turbonomic Platform with AWS leverages the elastic, on-demand capabilities of AWS while delivering the industry leading application performance, cost optimization, and compliance capabilities of Turbonomic. Regardless of your hybrid cloud topology, Turbonomic will dynamically discover application workloads across your entire infrastructure continuously in real-time.

Using the real-time capabilities to deliver true autonomic performance for your hybrid cloud platforms, Turbonomic will ensure that you gain unmatched capabilities to:

- Assure real-time performance on-premises and in the cloud
- Dynamically scale and place workloads for maximum performance and minimized cost
- Intelligently plan cloud migrations and future infrastructure
- Ensure compliance for existing and new workloads

Customers benefit by maximizing the overall return on their investment by assuring performance, lowering cost, and maintaining continuous compliance. By curating workload demand, Turbonomic unleashes the full potential of the public cloud's elasticity.

We've encountered customers that have unknowingly overspent by $2.3 million. We have customers who have achieved a 30% reduction in cloud bills while assuring performance and being continually compliant.

Only Turbonomic can assure application performance while using a policy-based approach to optimize and lower your cloud costs and ensure compliance for your cloud application workloads on AWS infrastructure and across any hybrid cloud infrastructure.

REFERENCES

Turbonomic Green Circle Community Forum:
greencircle.vmturbo.com

Turbonomic Resource Library:
turbonomic.com/resources/

Complete Turbonomic Documentation:
docs.turbonomic.com/

AWS Well Architected Framework (WAF):
d0.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf

Cost Optimization with AWS:
https://d0.awsstatic.com/whitepapers/Cost_Optimization_with_AWS.pdf

Configuring AWS Billing for Turbonomic Cost Control:
greencircle.vmturbo.com/docs/DOC-4613

Configuring an AWS Target in Turbonomic:
greencircle.vmturbo.com/docs/DOC-3828

Configuring Turbonomic with Public Cloud (Master Page):
greencircle.vmturbo.com/docs/DOC-4003

WHAT IS TURBONOMIC

Turbonomic's Hybrid Cloud Management Platform enables heterogeneous environments to self-manage to assure the performance of any application in any cloud. Turbonomic's patented decision engine dynamically analyzes application demand and allocates shared resources in real time to maintain a continuous state of application health.

Launched in 2010, Turbonomic is one of the fastest growing technology companies in the virtualization and cloud space. Turbonomic's Hybrid Cloud Management Platform is trusted by thousands of enterprises to accelerate their adoption of virtual, cloud, and container deployments for all mission critical applications.

To learn how Turbonomic can help you and your organization, contact sales@turbonomic.com or visit turbonomic.com.

Turbonomic's Hybrid Cloud Management Platform delivers 30% faster application response time on 30% less infrastructure

ABOUT THE AUTHORS

Dr. Shmuel Kliger

Founder and President, Turbonomic

Shmuel is founder and president of Turbonomic and a recent Ernst and Young Entrepreneur of the Year finalist. He directs the technology teams encompassing development, product strategy and management, customer success and customer support. Prior to founding Turbonomic, Shmuel served as Vice President of Architecture and Applied Research in the EMC CTO Office. Before that, he was the Chief Technology Officer of EMC's Resource Management Software Group after joining EMC with the acquisition of SMARTS. Kliger was a Chief Technology Officer and co-founder of SMARTS.

Eric Wright

Principal Solutions Engineer and Technology, Turbonomic

Before joining Turbonomic, Eric Wright served as a systems architect at Raymond James in Toronto. As a result of his work, Eric was named a VMware vExpert and Cisco Champion with a background in virtualization, OpenStack, business continuity, PowerShell scripting and systems automation. He's worked in many industries, including financial services, health services and engineering firms.

Green Circle Community

The Green Circle Community, available at GreenCircleCommunity.com, is the central repository for open community content and support for Turbonomic solutions.

Powered by Turbonomic Program Support

Powered by Turbonomic Reference Architecture series are provided for fully supported solutions in conjunction with Turbonomic and the associated technology partners included within the solution. Support for Powered by Turbonomic Solutions is provided by the respective vendors partnered on each solution.