turbonomic
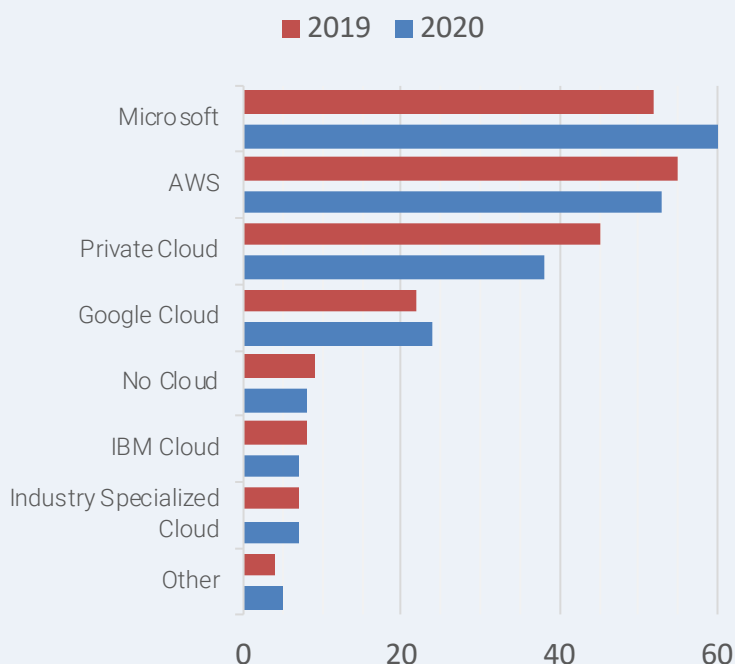
Application Driven
Cloud Optimization

A recent survey "2020 State of Multicloud" reported the following trends amongst respondents:

- The median number of clouds used: **two clouds**
- **Microsoft Surges in 2020**: Microsoft Azure took a significant lead in the survey, growing by 18% in the last year – 61% of respondents are using Azure.
- **Container Adoption Continues to Grow**: Up from 62% in 2019, 75% of respondents are on their journey to containers and cloud native, ranging from exploring, pre-production, early production, advanced production, or platform-first. 39% of respondents are running containers in production.
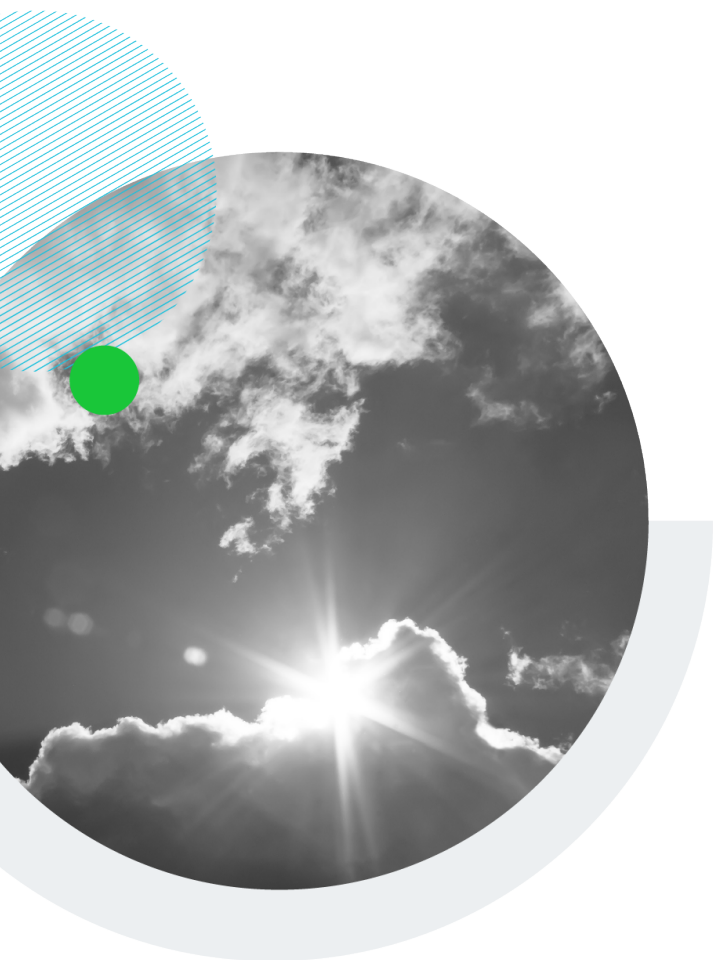
Introduction

Today's digital transformation initiatives are measured by how quickly an enterprise delivers new services to market. Applications now make or break businesses, with customers interacting with enterprises through digital and mobile experiences. New generations of cloud native applications are highly distributed and employ microservices and container building blocks to enable parallel development, rapid introduction of new services, and personalized customer experiences. Cloud computing underpins digital transformation, enabling enterprises to leverage the agility and elasticity of public cloud services. Most organizations have adopted a multicloud strategy, and are starting to employ container platforms due to the portability and agility benefits they offer.

In this white paper, we review the latest trends in the IT industry's quest to assure application performance and provide always-on availability, anywhere, while also driving business innovation.

turbonomic

In today's competitive business climate, it is clear that applications drive revenue, profits, customer experience and customer retention. Application success is determined by an enterprise's ability to innovate through development teams and IT operations, but:

1. Cloud native development results in increased application complexity.
2. Cloud computing, containers, microservices and functions make application resourcing more complex.
3. Multicloud strategies introduce specific taxonomy, nomenclatures and service offerings
   for each cloud provider an enterprise invests in.

Industry Challenges in Dealing with IT Complexity

IT Operations, developers and business units are facing an unprecedented knowledge gap related to fast-moving technologies and eco-systems. In an effort to provide application performance assurance, the traditional approach to build in resource headroom has resulted in significantly oversized workloads, at the expense of massive CapEx and OpEx investment. Since overprovisioning is predicated on maintaining sufficient utilization overhead, monitoring tools focused on application performance and resource cost optimization are in wide use.

Application Performance Monitoring

APM platforms are widely used to make sure business critical applications are architected and written well. They provide insight into end user Digital Experience Monitoring (DEM) and application discovery, tracing and diagnostics to understand relationships between various elements of distributed application environments. In addition to providing application level visibility, APM platforms also provide visibility into levels of the IT stack in order to help pinpoint the source of IT issues that could potentially result in application performance degradation. As such, APM tools help facilitate root cause analysis and remediation of application performance degradation issues.

Infrastructure Resource Optimization

Infrastructure resource optimization tools have moved front and center as public and multicloud Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) markets climb into the tens of billions of dollars. Users of multicloud services are increasingly looking for ways to optimize public cloud cost, while assuring performance for their applications. Most cloud optimization tools, including native cloud provider tools, operate by looking at past cloud bills and current infrastructure utilization data. The objective of these tools is to reduce inefficiencies while avoiding resource congestion.

Through 2021, digital transformation initiatives will take large traditional enterprises, on average, twice as long and cost twice as much as anticipated.

Source: https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2020-and-beyond/

While features of monitoring tools vary, IT organizations have embraced a traditional approach for several decades: Resources are intentionally overprovisioned to accommodate for future growth and unforeseen demand fluctuations. A combination of monitoring resource utilization and generating threshold-based alerts is intended to determine appropriate resource adjustments to avoid or mitigate performance issues. In many enterprises, there are so many alerts being generated by different monitoring tools that a class of monitoring has emerged to filter out all but the most severe performance issues or risks, which are then surfaced for IT staff to investigate and remediate. All other alerts are ignored, as IT staff attempt to correct problems that may be negatively impacting the business.

It is clear that application resources today are not being successfully optimized by overprovisioning and monitoring. Why is this so?

It has to do with the meaning of the term 'Optimize,' which is a continual balance of three IT mandates:

1. **Performance** – assure all workloads (VMs, DBs, containers, etc.) are getting the resources they need, when they need them, for optimal performance
2. **Policy Compliance** – be aware of and enforce policy compliance and other business constraints
3. **Cost** – efficiently reduce CapEx and OpEx

There is a reason that cost optimization is third on the list. While most organizations **want to** be cost-effective, application performance and business policy compliance are critical to sustaining the business and minimizing risk. The reason is straightforward: the financial impact of an under-performing business-critical application or compliance violation will usually be greater than the financial loss of not being very efficient. In the figure below, performance assurance and policy compliance are things organization **have to** do, even if they don't always do them well. The traditional IT approach of overprovisioning attempts to de-risk application performance, but sacrifices efficiency and cost.
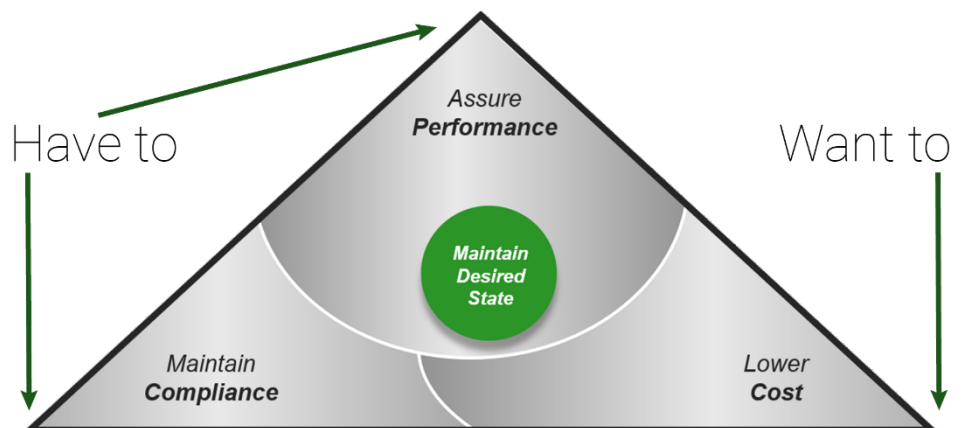
Why is a monitoring and alerting approach not sustainable at scale?

Most enterprises ignore thousands of alerts each day, while manually addressing a subset of those deemed to pose a risk of serious consequences. This means application performance is not always being assured with traditional approaches. As applications get more complex, and IT resource options introduce new complexities such as the rapid ascent of dynamic and ephemeral container environments, monitoring and alerting based approaches will prove untenable. They simply won't work for unpredictable application demand situations.

There is a clear and immediate need for a solution that can address the challenges related to multicloud adoption and containerization. In order to assure application performance, however, any solution platform must understand application requirements (demand) as well as the relationships between each layer of the IT stack, from the application components to the physical resources that are assembled to build containers and virtual elements in a cloud environment. This understanding of the full stack relationships is what is missing in traditional IT toolsets and approaches.

A Different Approach

The most important objective of IT infrastructure is to provide applications with the resources they need to deliver their services levels. A companion objective is to do this as cost efficiently as possible and adapt to changing environment and application demand scenarios by dynamically adjusting resources over time. Key capabilities include:

1. Application-Aware Optimization
2. Support On-Premises, Hybrid, and Multicloud Deployments
3. Full Stack Visibility and Control
4. Trustworthy & Automatable Actions
5. Enforce Business Policy Compliance

"Dynamic optimization technology for resources and cloud service is a technology capability that uses telemetry, algorithms, service and resource analytics, and policies to drive real-time automated actions that reduce waste, cost and risk exposure, while simultaneously improving service levels"[2]

2. Gartner Document – "Innovation Insight for Dynamic Optimization Technology for Infrastructure Resources and Cloud Services"

Application-Aware Optimization

Assuring application performance requires insight into the application's fluctuating demand and configuration to dynamically match demand to required resources (supply). Economic principles of supply and demand successfully balance demand in a variety of complicated market economies at global scale, including stock and stock futures exchanges, and commodities exchanges from gasoline to pork bellies.

Turbonomic has taken an approach of applying economic principles to manage IT resources. Application Resource Management is accomplished by abstracting all layers of the IT stack, from discrete physical resources up through every layer to the application components, into a market economy supply chain. At each layer of the stack, analytics compare the available supply in the underlying layer to the demand being requested by the higher layers, to continuously make decisions about how to best meet demand. By making each layer application-ware, IT resources are being proactively driven to a desired state of assured performance rather than waiting for something to go wrong and attempting to resolve it. By driving to a continuous desired state, workloads of all types can become Self-Managing, anywhere in Real-Time (SMART).
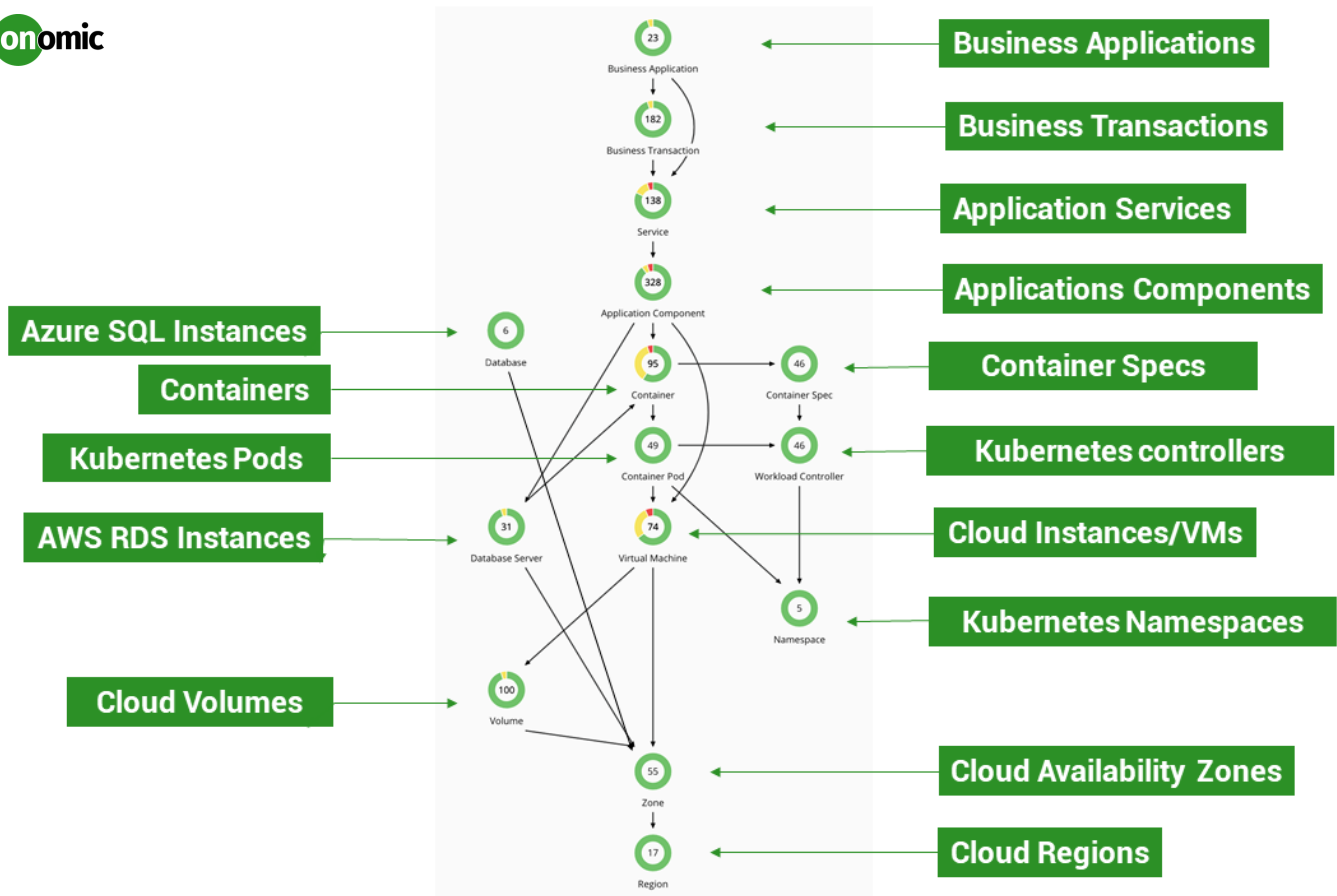
Full Stack Visibility and Control

Every environment consists of multiple stacks (or layers); at the top is the most important entity, the Application**. The purpose of IT operations is to make applications run well.**
Therefore, everything below the application layer should support the application.

Every resource in the stack can experience contention or issues that will impact the higher layers, all the way up to the application. For example, an instance on AWS EC2 assigned with an EBS volume not suitable for the application IOPS demand may negatively impact application performance despite having ample compute capacity.

Turbonomic offers full stack visibility and control into every environment it manages. Turbonomic understands the relationships between the layers, the available capacity of resources and fluctuating demand of the application, and is **able to execute actions at each layer** to assure application performance. Customers who deployed Turbonomic in their on-premises and cloud environments report:

1. Efficiency of running applications at desired SLOs and preventing application performance issues
2. Reduction in incidents and Mean-Time-To-Repair/Resolution (MTTR) of issues
3. Higher collaboration between often siloed teams (e.g., admin teams and developers, etc.)

Diagram labels (top to bottom):
- Business Applications → Business Application (23)
- Business Transactions → Business Transaction (182)
- Application Services → Service (138)
- Applications Components → Application Component (328)
- Azure SQL Instances → Database (6)
- Containers → Container (95)
- Container Specs → Container Spec (46)
- Kubernetes Pods → Container Pod (49)
- Kubernetes controllers → Workload Controller (46)
- AWS RDS Instances → Database Server (31)
- Cloud Instances/VMs → Virtual Machine (74)
- Kubernetes Namespaces → Namespace (5)
- Cloud Volumes → Volume (100)
- Cloud Availability Zones → Zone (55)
- Cloud Regions → Region (17)

Application-Aware, Automatable Actions

To become agile, **companies must embrace automation**, from automating the provisioning of infrastructure (for example, Infrastructure as Code) to automating workload and application deployment. While these examples of process automation are valuable for repetitive manual tasks, traditional IT resourcing has remained reactive and manual due to siloed resource alerts that arise with limited context in relation to application performance. The burden of resource resolution falls to IT staff. Process automation is useful but insufficient.

The potential of Artificial Intelligence for IT operations (AIOps) is to avoid issues rather than waiting for them to occur and requiring IT staff to resolve them. **Preventative measures must be determined and automated in application-aware software.** Customers must gain confidence that the decisions made by the software are safe to implement.
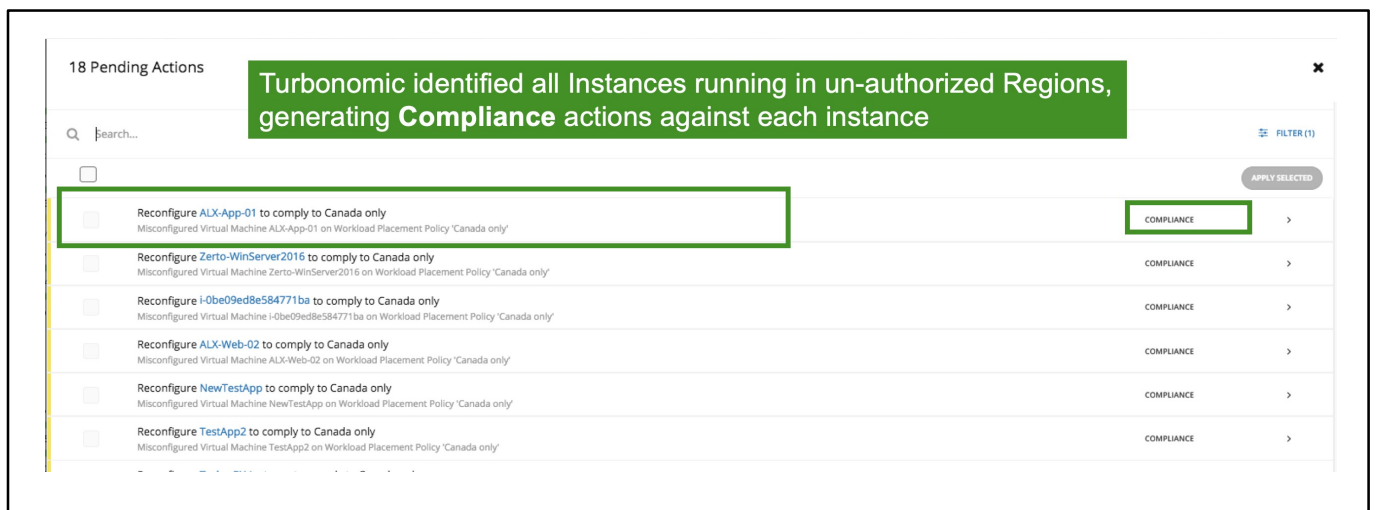
Turbonomic is deployed in many of the world's largest and most complex environments, where multiple teams and stakeholders initially reviewed the actions generated by Turbonomic before automation was turned on. Customers who automated our trustworthy actions reported the following benefits:

1. Improved application performance and reduction in the related cost to operate them
2. Improved IT staff productivity and abatement in employee churn
3. Adherence to business policies at scale (compliance)

Balancing Performance, Policies and Cost Efficiencies

While application performance is paramount to business success, enforcing policies are also top of mind. For example, a violation can result in a penalty or a fine by a governing authority or could contribute to application performance degradation or outage. Policies may range from where VMs can run (geolocation), to how they are built (e.g., approved storage and compute types), to how they are protected (high availability and business continuity) and any other constraints a business must follow. Cloud vendors have their own set of constraints, such as account quotas and backend limitations.

Turbonomic's Application Resource Management allows users to define business policies as well as importing rules from the targets it manages. Turbonomic not only operates within the constraints of the business policies but it will take actions when it detects a policy violation or configuration drift.

## Support For On-Premises, Hybrid, and Multicloud

The adoption of public cloud does not mean that all workloads will migrate to it. Datacenters are here to stay for the foreseeable future. **The need for a single platform that can manage, control and optimize all deployment models is critical**. Otherwise, companies must adopt multiple point-solution tools that address different layers in different environments, and multiple toolsets collectively add to IT complexity and cost.

Turbonomic is an API driven solution; with no agents used to connect to any on-premises, hybrid or public cloud targets. Turbonomic can be deployed on-premises or directly on the cloud in minutes, providing a flexible deployment architecture for wherever a customer is on their cloud journey.

Customers who deploy Turbonomic application resource management to manage their entire estate enjoy:

1. Automated application resource management assuring performance and improving efficiency across the entire stack
2. Reduction in management tools (and their licensing costs)
3. Increased cross-team collaboration (single source of truth) and IT staff productivity
4. A single platform to define and enforce business policies and constraints across all

   environments (compliance and governance)
5. A single platform to connect to various other IT operations tools (APM tools, Self-Service CMBDs, Cloud orchestrators, Cloud chargeback/capacity management/backup tools, etc.) to make more accurate decisions and achieve closed loop application performance management.
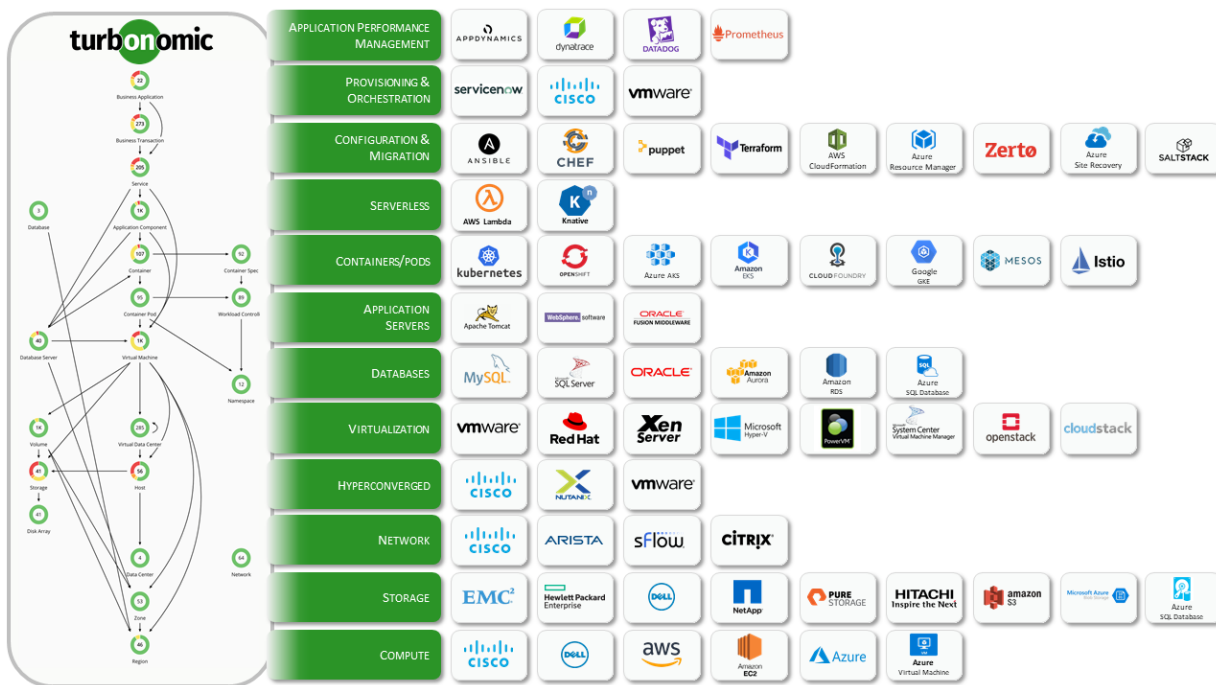
Improved Organizational Leverage

When issues occur in traditional monitoring environments, they are escalated to a "war-room" for triage and repair, usually involving multiple teams, often siloed, with various data points gathered from various sources. The result is an immense effort to connect the dots and determine the root cause, frequently leading to prolonged resolution timeline. Chasing problems cannot assure performance or deliver application driven performance optimization.

Turbonomic enables our customers to shift focus and time to driving business initiatives, spending less time and energy on "keeping the lights on", and minimizing common distractions that consume IT operations in more traditional environments.

Conclusion

Multicloud resource optimization represents a remarkably complex undertaking, introducing challenges on a scale that IT has never faced. Business leaders and IT operations alike must look for new ways to keep pace as the business adopts new technologies. Future planning must contemplate the speed at which these technologies evolve. The current generation of reactive tools that rely primarily on monitoring, threshold-based alerting, show back reports and manual remediation will not work at scale. A successful application driven cloud optimization solution must be able to provide continuous performance assurance while driving maximum cost efficiencies.

Application driven cloud optimization platforms must be application aware. IT staff will be increasingly called upon to focus on business innovation, and it will be up to intelligent AI software to manage the complexity of IT environments. Turbonomic has a proven approach of using economic principles to manage IT resources at scale in the world's most complex environments.

About Turbonomic

Turbonomic, leveraging Application Resource Management (ARM), is transforming the way people and organizations optimize their cloud, infrastructure, and network resources to support their business applications. Our mission is to be the go-to partner for assuring the performance of any application on any platform. We help manage the increasing complexity of the full stack of application resources to continually deliver the best end user experience. By delivering an integrated platform and empowering an ecosystem of strategic partners such as IBM, Cisco, Microsoft and AWS, we are assuring application and network performance for thousands of organizations including twenty-five percent of the Fortune 500. Our vision is that every application always gets the resources it needs, and users never have to wait for an application. To learn more, visit turbonomic.com.