The executive’s guide to SLOs

Why new customer experience paradigms require executives to think about service level objectives

Executive Summary

Modern applications are more complex and distributed than ever. And their performance is more critical to the business than ever. As a result, executives are under pressure to demonstrate the business impact of their organization. Fortunately, the same technologies being used to modernize and manage applications create opportunities to directly tie IT's impact to the business. IT and Platform teams have found new ways to measure the health of their environments, set expectations for their applications, and connect their efforts to business context, namely Service Level Objectives. Given the business-criticality of today's modern applications, leading executives must guide their organization's thinking towards defining SLOs in terms of business impact and customer experience.

In this paper, we will review common approaches and pitfalls when defining and implementing SLOs, and provide clear guidance on best practices on utilizing SLOs to drive better business outcomes.

Context

Today, "end-user satisfaction with application performance and reliability is critical for successful digital business operations."[1] Delivering a great end-user experience, however, is challenging as modern applications and the platforms and infrastructures they run on are more complex and distributed than ever. In fact, as seen in our 2021 State of the Multicloud Report, complexity was cited as one of the top challenges to achieving business goals.[2]

Along with the development of modern applications, "the shift to agile infrastructure — including hybrid IT, multicloud and containers" have left IT teams struggling to make sense of immense amounts of data and "has challenged the viability of traditional infrastructure monitoring tools."[3]

In response to these challenges, IT and Platform teams have found new ways to measure the health of their environments, set expectations for their applications, and connect their efforts to business impact. And with ever rising end-user expectations, it is vital that executives ensure their organizations use meaningful metrics when assessing the performance of their applications and their impact on the business.

1 Source: Worldwide Application Performance Management Software Market Shares, 2020 | June 2021, IDC #US47989021

2 Source: 2021 Turbonomic State of Multicloud Report, 2020 CNCF Survey

3 Source: Monitoring and Observability for Modern Services and Infrastructure| June 2020, Gartner G00720854

The Current Approach: Laborious & Reactive

A common approach IT and platform teams take to measure the health of their environments is through identifying and configuring Service Level Indicators, Service Level Agreements, and Service Level Objectives.

- A Service Level Indicator (SLI) is the metric(s) by which teams measure how a service performs in relation to the goal.
- A Service Level Agreement (SLA) is a promise of availability IT and platform teams make to customers or end-users.
- And a Service Level Objective (SLO) is the service level target or goal, measured with an SLI, that teams commit to reaching.

SLOs are especially important because they define the criteria and establish clear expectations for the performance of an organization's applications.

Today, most organizations undergo a time-consuming and laborious manual SLO configuration process. IT teams first must determine which services of an application directly impact the business and the experience of the end-user. Next, they must identify which metrics to use as service-level indicators. Some common SLI metrics are availability, latency, and transaction throughput. Once the appropriate SLIs are chosen, IT teams then need to determine their SLO target goal for that metric and a specific measurement period. After they complete this process, IT teams then must create error budgets for each SLO and link them to a threshold-based alert system. Many organizations take this threshold based approach because it is impossible for humans to monitor the performance of their applications 24/7/365.

Unfortunately, setting SLO thresholds does not solve for performance. This IT strategy is ineffective because it is too reactive for today's agile infrastructure with applications running in multicloud or containerized environments. If something goes wrong with a service and a threshold or alert is triggered, performance degradation has already occurred for that service resulting in a poor end-user experience.

IT teams have tried to improve on their system of thresholds/alerts by implementing Horizontal Pod Autoscaling (HPA) for applications running in containerized environments. However, HPA also fails to ensure a great end-user experience and prevent performance degradation. Similar to the SLO configuration process, in order to set horizontal autoscaling to meet resource demands, IT teams must identify metrics that best represent resource needs, configure goals and set thresholds, and test. This process needs to be repeated for every service of an application. With some applications having hundreds of different services, implementing HPA at scale is very difficult. Furthermore, different HPA policies still rely on thresholds and do not correlate or comply with one another, meaning that scaling one service could negatively affect another. Finally, this is not a one-off exercise and HPA scaling policies require continuous reconfiguration and monitoring to be effective.

How to think about SL0s...

Define SLIs and SLOs in terms of customer experience.
According to industry experts, “focusing on managing systems and applications to optimize the end-user experience is a major priority as fast performance and 100% uptime are table stakes for digital business success.”[4] Since fast performance and uptime are now table stakes, organizations cannot waste time collecting data that is not directly indicative of end-user experience when assessing application performance.

Collecting the right data is no easy task. For example, although availability is a commonly used metric, it is not a direct extension of performance because an application can be available but still suffer from resource bottlenecks and performance degradation. There are many different metrics organizations can use to measure performance, so it is important that IT and the line of business agree on what data to collect and report on. As Gartner puts it: “Choosing representative and meaningful SLIs is critical. In most cases, an infrastructure-based metric (“available memory,” or “% free worker nodes”) is unlikely to be meaningful because it is not often something that the users of a service care about. Choose SLIs that are direct measurements of the experience users have with your services.”[5]

Simply put, SLOs should be used as a tool that tells organizations whether their applications are doing what they need to do for their business. If applications are not performing, organizations that have a meaningful SLO defined at the right level of the application should know exactly what actions will be most impactful for restoring performance. Different organizations need to measure different metrics specific to their lines of business. However, as organizations move away from monolithic application architectures to be more modern and distributed, traditional performance indicators such as high memory and CPU usage become less useful. Instead, organizations should look to define SLIs / SLOs in terms of metrics that are typically relevant to business such as response-time and transaction throughput. These metrics are a more direct measurement of performance because, for example, by defining SLOs for transaction throughput, IT teams can know exactly how requests are being serviced for each individual pod or VM. This metric is a more direct assessment of performance than defining SLOs based on utilization metrics and loosely associating it to metrics like response time and throughput.

4 Source: Worldwide Application Performance Management Software Market Shares, 2020 | June 2021, IDC #US47989021 5 Solution Path for Modern Infrastructure and Application Monitoring | June 2019, Gartner

Although response time and transaction throughput are some of the most common ways to assess customer experience and application performance, these metrics do not work for all organizations. For example, an organization that is a VDI service provider would not want to define SLOs for transaction throughput but instead for support tickets. Support tickets would be a more impactful metric for an SLO because it is a direct measurement of the performance of virtual desktops. Ultimately SLIs / SLOs need to be defined specifically for each different organization and their line of business.

Continuous analysis of changing application demand is required.
Modern applications and infrastructures are elastic and resource demand is dynamic. Due to this constant change, understanding the relationships between different sources and types of data is very difficult. Organizations often use different tools to monitor different layers of the stack across different teams, all while trying to solve different sides of the same issue. Consequently, this lack of coordination results in many false leads to the root cause of a problem. Furthermore, data needs to be collected continuously because if it is not, IT and platform teams are left guessing what data they need to resolve an issue. As Gartner notes in a 2020 Monitoring and Observability report, "Collection of the potential cause data must take place continuously because enabling it in response to a symptom may miss the cause entirely."[6]

In order to avoid this dilemma, organizations need a central repository that can aggregate and correlate data from every layer of an application stack. With this system, organizations can bring their continuously collected data together and percolate it up to their SLOs. By doing this continuous analysis, organizations can then contextualize everything and know what objective they must reach for successful application performance. And if the goal is not met, what are the issues in the underlying layers of the application stack that need to be fixed. Ultimately, implementing systems that can do this continuous analysis is the only way to keep up with the dynamic resource needs of modern applications and infrastructure.

Automate dynamic resourcing.
Just as the dynamic nature of modern applications and the infrastructure they run on require continuous analysis, executives and their organizations should also look to implement continuous automation if they want to fully leverage SLOs and create a preventative system for managing their applications. In a recent analyst report from IDC, they said that for organizations to stay competitive in the future they must "consider the role of automation in product capabilities."[7]

6 Source: Monitoring and Observability for Modern Services and Infrastructure | June 2020, Gartner G00720854
7 Source: Worldwide Application Performance Management Software Market Shares, 2020 | June 2021, IDC #US47989021

The development of modern infrastructure promised future application resiliency and elasticity, but many organizations today struggle to maintain the performance of their applications. Implementing automation in the management of modern applications is essential because it is necessary for reaping the benefits of resiliency and elasticity. Without automation, the resolutions made possible through configuring the right SLOs and collecting data continuously are not effective. IT and platform teams cannot assure application performance without automating dynamic resourcing: by the time an alert is triggered and a resource decision is manually executed, performance degradation has already occurred.

Ultimately, the problem with this approach is that it is predicated on the fact that a problem has already happened.

Automation can be more than simply a reaction to an event that is triggered by a threshold. By leveraging continuous data collection and identifying key SLIs andSLOs for your line of business, organizations can have software generate actionable decisions that should be automated. If, as an organization, you take the necessary steps and commit to automation, you can create a truly elastic environment that proactively manages how your applications are resourced hereby ensuring continuous performance. This type of automation requires an intelligent system that analyzes a dynamically changing environment and automates the necessary chain of decisions so that issues are resolved before performance degradation occurs. It is impossible to achieve this kind of performance and elasticity through a process of spreadsheets and alerts.

In order to move away from a reactive approach there must be complete buy-in from stakeholders across the various teams that manage the applications and those that manage the infrastructure. Application and Product Owners are often hesitant to relinquish control of their applications to automation. This reluctance is due to a lack of trust in automation. But it can be overcome. Automation requires a cultural transformation of your IT organization. To achieve true elasticity and resiliency, organizations must be able to trust in the actions that are to be automated. With meaningful business- and application-centric SLOs that are directly tied to dynamic resourcing in the platform and infrastructure it is easier for Application and Product Owners to trust automation and get comfortable with the expectation that automation will be fully implemented into business workflows.

Things to Keep in Mind

Applications are only going to become more complex as organizations increasingly adopt agile infrastructures such as hybrid IT, container platforms, and multicloud. With these future developments, organizations that meet the elevated expectations for customer experience and application performance will thrive. The best practices to follow are to define SLOs in terms of customer experience, continuously analyze changing application demand, and automate dynamic resourcing. In order to achieve digital business success, organizations must take the necessary steps to implement automation that dynamically resources applications to meet changing demand and business SLOs.

Automatically assure application SLOs with Turbonomic.

Turbonomic turns data into action, delivering automation that prevents application performance risks while maximizing elasticity.

Modernizing mission-critical applications and infrastructure is an investment with numerous benefits. But to reap the benefits of elasticity, resiliency, and speed to market, you need software that continuously analyzes your environment and executes the right resourcing decision at the right time to ensure application performance. With Turbonomic, you can correlate application response time, transaction throughput, or other SLIs / SLOs to dynamic resourcing. As demand fluctuates, Turbonomic's dynamic resourcing will ensure continuous application performance.

HPA doesn't cut it.
Turbonomic uses top-down, full-stack analysis to dynamically assure your SLOs. You set your SLOs and our AI-powered software ensures that the platform and underlying infrastructure provides the resources they need to meet those SLOs, wherever your applications run.

Seamlessly integrate into business workflows.
With Turbonomic's integration with Webhooks, you can easily inject Turbonomic actions into application lifecycles, DevOps & infrastructure pipelines, approval and audit workflows, and communication processes.

Minimize the manual labor.
Dev, DevOps, and SREs don't need to set thresholds, constraints, or autoscaling policies. Software makes the right resource decisions for you, providing actions you can actually automate.

Don't overspend on capacity.
No need to rely on Dev to make resourcing decisions (they often overprovision just to be safe, right?) Our software determines exactly what resources services need—all based on application demand.

Quickly and easily plan for growth.
Simulate the onboarding of new services with our software. Determine exactly what you need to support new growth.

To Try Turbonomic today, visit turbonomic.com/try-SLO.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2022

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.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.