IBM **Turbonomic**

SLO 执行指南

为什么新的客户体验范式要求公司管理层设定正确的服务级别目标 (SLO)

执行摘要

现代应用程序比以往任何时候都更加复杂和分散，应用性能对业务的影响也变得更加重要和关键。因此，管理层的压力在于亟需证明其组织所应具备的业务影响力。幸运的是，用于实现应用现代化和管理的技术同样也创造了各种机会，可助力企业直接将 IT 部门的影响力与业务紧密结合起来。IT 和平台团队已找到新方式来衡量其运营环境的健康状况、设定针对各种应用的期望值，并将其效用与业务环境（即服务级别目标）相关联。当今的现代应用程序对业务起到至关重要的作用，因此一流的管理层必须引导其组织朝着可明确界定业务影响力和客户体验度的 SLO 方向进行思考。

本文将审视在定义和实施 SLO 时所需的通用方法，以及面临的常见陷阱，并为员工提供明确的最佳实践指导，以推动企业实现更出色的业务成果。

背景环境分析

现今，“最终用户在应用性能和可靠性方面可获得的满意度对成功实现数字化企业运营至关重要。”[1] 但是，提供出色的最终用户体验非常具有挑战性。因为现代应用和平台及其运行的基础架构比以往任何时候都更加复杂和分散。事实上，根据我们的 2021 年多云现状报告显示，复杂性已被列入实现业务目标的最重要挑战之一。[2]

现代应用的发展，以及“迁移至敏捷运行的基础架构（包括混合 IT、多云和容器）”让 IT 团队愈发难以理解海量数据，并转而“质疑传统基础架构监控工具的可行性。”[3]

为了应对这些挑战，IT 和平台团队已找到新的方式来衡量其运营环境的健康状况、设定针对各种应用的期望值，并将其效用与业务影响相关联。随着最终用户的期望越来越高，管理层必须确保其组织使用意义明确的度量标准来评估其应用性能和业务影响；这一点现已变得至关重要。

1 来源：2020 年应用性能管理软件之全球市场份额分析 | 2021 年 6 月，IDC #US47989021
2 来源：2021 年 Turbonomic 多云现状报告，2020年 CNCF 调查报告
3 来源：适用于现代化服务和基础架构的监控和可观察性 | 2020 年 6 月，Gartner G00720854

当前方法：费力且被动

IT 和平台团队用来衡量其环境的健康状况的一个常见方法是识别和配置服务级别指标 (SLI)、服务级别协议 (SLA) 和服务级别目标 (SLO)。

- **服务级别指标 (SLI) 是团队根据目标来衡量服务执行情况的度量标准。**
- **服务级别协议 (SLA) 是 IT 和平台团队向客户和最终用户做出的可用率承诺。**
- **服务级别目标 (SLO) 则是依据 SLI 所包含的指标来进行测量、且团队致力于实现的服务级别目标。**

SLO 特别重要之处在于它们对企业的应用程序性能明确定义了标准，并设定了清晰的期望值。

现今，大多数组织都在使用耗时且费力的人工 SLO 配置流程。IT 团队必须首先确定应用程序的哪些服务会直接影响业务和最终用户的体验。接着，他们必须确定使用哪些度量参数来作为服务级别指标。常见的 SLI 指标包括可用率、延迟率和事务吞吐量。选定适当的 SLI 指标后，IT 团队便需要确定其针对该度量标准的 SLO 目标和具体的测量周期。完成此流程后，IT 团队必须为每个 SLO 创建误差量，并将其与基于阈值的警报系统相连。许多组织都采用这种基于阈值的方法，因为人工无法每天 24 小时全年无间断监控应用程序的性能。

遗憾的是，设置 SLO 阈值并不能解决性能问题。这种 IT 策略效用甚低，因为对于现今敏捷运行的基础架构（应用程序在多云或容器化环境中运行）而言，它太被动了。如果某项服务出错导致触发阈值或警报，则表示该服务已出现了性能下降的问题，这就造成了糟糕的最终用户体验。

IT 团队已尝试通过对在容器化环境中运行的应用实施水平 Pod 自动缩放 (HPA) 策略，来改善其阈值/警报系统。但是，HPA 也未能保障出色的最终用户体验，无法阻止性能下降。与 SLO 配置流程类似，为了设置水平自动缩放以满足资源需求，IT 团队必须确定最能代表资源需求的度量标准，配置目标和设置阈值，并进行测试。需要对应用程序的每项服务重复此流程。由于有些应用程序拥有数百项不同的服务，大规模实施 HPA 非常困难。另外，不同的 HPA 策略依然依赖阈值，彼此不相关或需要相互趋于一致，这意味着缩放一项服务可能会对另一项服务产生负面的影响。最后，这不是一次性的措施，HPA 缩放政策需要执行连续运行的重新配置和监控才能取得效果。

如何考量 SLO...

明确界定衡量客户体验所需的 SLI 和 SLO。

行业专家指出：“重点关注管理系统和应用程序，确保优化最终用户体验，这属于重大优先事项，因为**要想取得数字业务的成功，就必须全力投入，保障实现快速性能和 100% 正常运行时间。”**[4] 由于快速性能和正常运行时间现已成为必须全力投入解决的事项，组织在评估应用性能时，就不能将时间浪费在收集那些无法直接体现最终用户体验的数据上。

收集恰当的数据并非易事。例如，虽然可用率是一个常用的度量指标，但它不是性能的直接延伸，因为应用程序即使在可用的情况下也可能会遭遇资源瓶颈和性能下降等问题。组织可使用许多不同的度量指标来测量性能，所以 IT 部门和业务线部门就收集和报告哪些数据达成一致意见非常重要。正如 Gartner 报告所示：“选择具有代表性和意义明确的 SLI 指标至关重要。在多数情况下，基于基础架构的度量指标（‘可用内存’或‘空闲工作程序节点百分比’）可能没有意义，因为服务用户一般不关心这些。最好选择可直接测量企业服务的用户体验所代表的 SLI 指标。”[5]

简单地说，SLO 应起到工具的作用，告知企业其应用程序是否在执行业务所需的各项功能。如果应用程序表现不佳，在应用适当级别已界定明确意义的 SLO 的组织就应知道具体采取何种行动，以便采取最有效的方式恢复应用性能。不同的组织需要测量其各业务线部门所特定具备的不同度量指标。但是，随着组织从单一应用架构向着更加现代和分散的方向发展，传统的性能指标（例如高内存和 CPU 使用率）变得越来越不起作用。相反，在界定 SLI/SLO 时，组织应寻求获取通常与响应时间和事务吞吐量等业务相关的度量参数。这些度量标准针对的是更直接的性能测量，例如 IT 团队可以通过定义事务吞吐量 SLO，确切知道每个 Pod 或 VM 的请求已得到何种服务。与根据使用率指标定义 SLO、并将其与响应时间和吞吐量等指标松散关联的度量方式相比，这类度量指标属于更直接的性能评估范畴。

4 来源：2020 年应用性能管理软件之全球市场份额分析 | 2021 年 6 月，IDC #US47989021
5 适用于现代化基础架构和应用监控的解决路径 | 2019 年 6 月，Gartner

虽然响应时间和交易吞吐量是评估客户体验和应用性能最常见的方式之一，但是这些度量指标并不适用所有组织。例如，属于虚拟桌面基础架构 (VDI) 服务提供商的组织不会想为事务吞吐量定义相关的 SLO，而只会希望为支持凭单定义 SLO。支持凭单会是更有效的 SLO 指标，因为它属于更直接的虚拟桌面性能测量方式。最终，需对每个不同的组织及其业务线部门明确界定各自特有的 SLI / SLO 指标。

必须保持持续分析，应对不断变化的应用需求。

现代应用和基础架构是富有弹性的，而资源需求时刻处于动态运行中。鉴于这种不断变化的本质，了解不同的来源和数据类型之间的关系变得非常困难。组织常常使用不同的工具跨团队监控不同的堆栈层，同时还要试图解决同一问题的不同方面。结果，缺乏协调会导致在寻找问题根源时，给出许多错误的线索。另外，还需保持持续收集数据的状态，否则 IT 和平台团队便只能猜测需要什么数据来解决问题。正如 Gartner 在一份 2020 年发布的监控和可观察性报告中所言：“必须采取措施持续收集潜在原因数据，但出于对某个症状的响应而启用这类措施，却又可能会导致彻底错过找到根本原因。”[6]

为了避免这种两难困境，组织需要一个能从应用堆栈的每一层汇聚和关联数据的中心存储库。通过这个系统，组织可以汇聚其持续收集的数据，并将其向上过滤至其 SLO。通过持续执行此分析，组织便能将一切情境化，知晓为了确保应用性能成功运行，必须达到何种目标；以及如果未能达到目标，则需要解决应用堆栈底层中的哪些问题。最终，实施能持续执行此分析的系统是解决现代应用和基础架构的动态资源需求的唯一途径。

动态资源分配自动化运行。

正如现代应用及其运行的基础架构所体现出的动态性质需要执行持续分析，管理层及其组织也应从中寻求实施持续自动化策略，以便完全利用 SLO 的优势并创建预防性系统来管理其应用程序。根据最近 IDC 发布的一份分析报告显示，组织要想在未来保持竞争力，就必须“考虑自动化在产品功能中的作用”。[7]

6 来源：适用于现代化服务和基础架构的监控和可观察性 | 2020 年 6 月，Gartner G00720854
7 来源：2020 年应用性能管理软件之全球市场份额分析 | 2021 年 6 月，IDC #US47989021

现代基础架构的发展昭示着应用程序的未来会更具韧性和弹性，但是现今许多组织依然很难维持其应用的运行性能。针对现代应用管理实施自动化是非常必要的，因为只有实现这一需求，才能获取应用韧性和弹性等重要优点。若无自动化，通过配置适当的 SLO 和持续收集数据而部署的解决方法就无法达到预期效果。倘若动态资源分配无法自动化运行，IT 和平台团队就无法确保应用性能：等到触发警报，以及人工执行资源分配决策时，应用性能下降的事件实际上就已经发生了。

此方法的根本问题在于：将预测建立在问题已经发生的事实之上。

自动化不是只对触发了阈值的事件做出响应那么简单。通过充分利用持续的数据收集活动，以及确定业务线部门产生的关键 SLI 和 SLO 指标等，组织便能通过软件生成应自动执行的可操作决策。如果组织采取必要的步骤和致力于部署自动化策略，便能打造一个真正富有弹性的环境，可主动管理您的应用资源分配机制，从而保障性能正常持续运行。这类自动化需要配置智能化系统来分析动态变化的环境，以及自动执行必要的决策链，以便在性能下降发生***之前***就解决问题。通过电子表格和警报流程完全不可能实现这种性能和弹性的保障。

为了摆脱被动的方法，就必须得到管理应用和基础架构的各团队利益相关方的完全认同。应用程序和产品所有者常常犹豫是否要将应用程序的控制权交付给自动化方案。犹豫原因在于对自动化方案缺乏信任，但这可以克服。实施自动化方案，需要企业 IT 组织实现文化转型。为了获得真正的应用弹性和韧性，组织必须能够信任要实施自动化的行动。当构建好富有意义的、以业务和应用为中心，并可直接与平台和基础架构的动态资源分配紧密相连的 SLO 机制，应用程序和产品所有者就会更容易信任自动化方案，并安心接受将在业务工作流程中完全实现自动化的预期。

谨记事项

随着组织越来越多地采用敏捷运行的基础架构（例如混合 IT、容器平台和多云等），应用程序只会变得越来越复杂。面对未来的这些发展，能够满足对客户体验和应用性能更高期望的组织将会蓬勃发展。可依照执行的最佳实践内容包括明确界定提高客户体验所需的 SLO 标准、持续分析多变的应用需求，以及确保动态资源分配实现自动化运行。为了实现数字化业务成功，组织必须采取必要的步骤来实施自动化，对应用程序执行动态资源分配，以满足不断变化的需求和业务 SLO 内容。

利用 Turbonomic 自动确保应用程序 SLO 顺利达标

Turbonomic 可将完成数据转化，付诸于行动，实现自动化防范应用性能风险，同时最大限度提升弹性运营。

针对任务关键型应用和基础架构完成现代化改造，这是一项可确立很多竞争优势的投资举措。但是，要想获得弹性、韧性和上市速度的优势，您需要配置软件来持续分析企业环境，并在正确的时间执行正确的资源分配决策，才能保障应用性能正常运行。利用 Turbonomic，您可以将应用响应时间、事务吞吐量或其他 SLI / SLO 与动态资源分配相关联。随着需求不断变化，Turbonomic 的动态资源分配机制将会保障应用性能持续正常的运行。

HPA 无法实现这一点。Turbonomic 使用自上而下的全栈分析，可动态确保您的 SLO 得到顺利实施。您只需设置好您的 SLO 标准，我们的 AI 驱动型软件会确保平台和底层基础架构提供满足那些 SLO 所需的资源，无论您的应用程序在何处运行。

无缝集成至业务工作流程。Turbonomic 与 Webhooks 相集成，确保企业能轻松地将 Turbonomic 行动注入应用程序生命周期、开发运维 (DevOps) 和基础架构管道、审批和审计工作流，以及通信流程等。

尽量减少手动工作：开发、开发运维和 SRE 无需设置阈值、约束或自动缩放等策略。软件会为您做出适当的资源分配决策，提供您能自动执行的行动。

无需支出过高的功能费用：无需依赖开发来做出资源分配决策（他们常出于安全考虑而过度配置，对吧?）我们的软件能完全根据应用程序需求，来确定资源分配服务所需的确切条件。

快速、轻松地规划增长：利用我们的软件模拟新服务的启用。确定支持新增长所需的确切需求。

立即试用 Turbonomic

turbonomic.com/try-SLO

国际商业机器（中国）有限公司
了解更多信息，欢迎访问我们的
中文官网：**https://www.ibm.com/cn-zh**

美国出品
2022 年 1 月