第16章 服务质量指标与SLA

§16.1 SLA 与 QoS
§16.2 可用性指标
§16.3 可靠性指标
§16.4 服务性能指标
§16.5 可扩展性指标
§16.6 弹性指标

§16.1 SLA与QoS

- 云提供者发布的SLA
  - 描述了服务质量(quality-of-service, QoS) 特性、保证，以及
  - 一个或多个基于云的IT资源的限制。
- QoS特性通过服务质量指标来表达：
  - 可用性(availability)：可用性比率、停用时间
  - 可靠性(reliability)：平均故障间隔时间、可靠性比率
  - 性能(performance)：网络容量、存储容量、服务器容量、Web容量、实例启动时间、响应时间、完成时间
  - 可扩展性(scalability)：存储、服务器
  - 弹性(resiliency)：平均切换时间、平均系统恢复时间

服务质量指标

- SLA管理系统两个任务：
  - 周期性测量QoS指标来验证是否与SLA保证相符合
  - 收集与SLA相关的数据，用于各种类型的统计分析
- 服务质量指标要具有以下特点：
  - 可量化的(quantifiable)
  - 可重复的(repeatable)
  - 可比较的(comparable)
  - 容易获得的(easily obtainable)

可用性指标：可用性比率

- 可用性比率：运行时间的百分比
  - 例如，至少99.5%的运行时间
- 描述——服务运行时间的百分比
- 测量值——全部运行时间
- 频率——每周、每月、每年
- 云交付模型——IaaS、PaaS、SaaS

表16-1 可用性比率示例

可用性	不工作时间/周（秒）	不工作时间/月（秒）	不工作时间/年（秒）
99.5	3024	216	158112
99.8	1210	5174	63072
99.9	606	2592	31536
99.95	302	1294	15768
99.99	60.6	259.2	3154
99.999	6.05	25.9	316.6
99.9999	0.605	2.59	31.5

可用性指标：停用时间

- 停用时间指标
  - 最大和平均停用时间。
- 描述——一次停用的时长
- 测量值——停用结束的日期/时间-停用开始的日期/时间
- 频率——每次有事件发生
- 云交付模型——IaaS、PaaS、SaaS
- 示例——最长1小时、平均15分钟

可靠性指标

- 是一个与可用性紧密相关的特性
- 是IT资源在预先定义的条件下执行它所期望的功能而不发生故障的概率。
- 可靠性意在描述服务能够多久按照预期执行，这要求服务保持在运行和可用的状态。

可靠性指标：平均故障间隔时间

- 平均故障间隔时间指标
- 描述——两次相继发生的服务故障之间的平均时间
- 测量值——∑，正常运行持续时间之和/故障时间
- 频率——每月，每年
- 云交付模型——IaaS，PaaS
- 示例——平均90天

可靠性指标：可靠性比率

- 可靠性比率指标
  - 表示得到成功服务结果的百分比。
  - 测量在服务运行期间不致命的错误和故障的影响。
- 描述——在预先定义的情况下得到成功服务结果的百分比
- 测量值——成功过响应的总数/请求总数
- 频率——每周，每月，每年
- 云交付模型——SaaS
- 示例——至少99.5%

服务性能指标：网络、存储

- 是IT资源在预期的情况下执行其功能的能力。
- 网络容量指标
  - 描述——网络容量可测量的特性
  - 测量值——每秒以位为单位的带宽或吞吐量
  - 频率——持续的
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——10MB/s
- 存储设备容量指标
  - 描述——存储设备容量可测量的特性
  - 测量值——以GB为单位的存储大小
  - 频率——持续的
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——80GB的存储空间

服务性能指标：服务器

- 服务器容量指标
  - 描述——服务器容量可测量的特性
  - 测量值——CPU数量，CPU主频，RAM大小，存储空间
  - 频率——持续的
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——主频为1.7GHz的内核一个，16GB RAM，80GB存储空间

服务性能指标：Web

- Web容量指标
  - 描述——Web应用容量可测量的特性
  - 测量值—— 每分钟的请求速率
  - 频率——持续的
  - 云交付模型——SaaS
  - 示例——最大每分钟100000个请求

服务性能指标：实例启动时间

- 实例启动时间指标
  - 描述——初始化一个新的实例所需要的时间长度
  - 测量值—— 实例启动的日期/时间 – 开始请求的日期/时间
  - 频率——每次有事件发生时
  - 云交付模型——IaaS、PaaS
  - 示例——最多5分钟看，平均3分钟

服务性能指标：响应时间

- 响应时间指标
  - 描述——执行同步操作需要的时间
  - 测量值—— (请求的日期/时间-相应的日期/时间)/请求的总数
  - 频率——每周、每月、每年
  - 云交付模型——SaaS
  - 示例——平均5毫秒

服务性能指标：完成时间

- 完成时间指标
  - 描述——完成同步请求操作需要的时间
  - 测量值—— (请求的日期-响应的日期)/请求总数
  - 频率——每周，每月，每年
  - 云交付模型——PaaS、SaaS
  - 示例——平均1秒

可扩展性指标：存储（水平）

- 是与IT资源的弹性能力相关的
- 是指IT资源可以达到的最大容量，反映了它适应工作负载波动的能力。
- 存储可扩展性（水平）指标
  - 描述——为响应工作负载的增加而允许的存储设备容量的改变
  - 测量值——以GB为单位的存储空间大小
  - 频率——持续的
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——最大1000GB（自动扩展）

可扩展性指标：服务器（水平）

- 服务器可扩展性（水平）指标
  - 描述——为响应工作负载的增加而允许的服务器数量的改变
  - 测量值——资源池中的虚拟服务器的数量
  - 频率——持续的
  - 云交付模型——IaaS、PaaS
  - 示例——最少1个虚拟服务器，最大10个虚拟服务器（自动扩展）

可扩展性指标：服务器（垂直）

- 服务器可扩展性（垂直）指标
  - 描述——为响应工作负载的波动而允许的服务器容量的改变
  - 测量值——CPU个数、RAM大小等
  - 频率——持续的
  - 云交付模型——IaaS、PaaS
  - 示例——最多512核，512GRAM

弹性指标

- IT资源从运行问题中恢复的能力。
- 通常是基于在不同物理位置上的冗余和资源复制以及各种灾难恢复系统
- 可以在三个不同的阶段中实施弹性指标，来解决可能威胁到常规服务水平的挑战和事件：
  - 设计阶段(design phase)
  - 运行阶段(operation phase)
  - 恢复阶段(recovery phase)

弹性指标：平均切换时间

- 平均切换时间指标
  - 描述——完成从一个出现严重故障的虚拟服务器切换到位于不用地理区域内的复制实例上所需要的时间
  - 测量值——（切换完成日期/时间 – 故障发生日期/时间）
  - 频率——每月，每年
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——平均10分钟

弹性指标：平均系统恢复时间

- 平均系统恢复时间指标
  - 描述——弹性系统完整地执行一次从严重故障中恢复所需要的时间
  - 测量值——（恢复日期/时间 – 故障发生日期/时间）
  - 频率——每月，每年
  - 云交付模型——IaaS、PaaS、SaaS
  - 示例——平均120分钟

小结

- SLA vs. QoS
- 服务可用性指标
- 服务可靠性指标
- 服务性能指标
- 服务可扩展性指标
- 服务弹性指标