华为云 APM 概述

目前，以用户为中心、以人为本的用户体验已成为应用的核心竞争力之一。但是随着应用复杂度不断提升、用户数量增多，海量业务下如何保障应用正常、如何快速完成问题定位、如何迅速找到性能瓶颈，已经成为应用运维的巨大挑战。

华为云应用性能管理服务（Application Performance Management，简称 APM）是实时监控并管理云应用性能和故障的云服务，提供专业的分布式应用性能分析能力，可以帮助运维人员快速解决应用在分布式架构下的问题定位和性能瓶颈等难题，为用户体验保驾护航。

APM 作为云应用诊断服务，适用于多种 Java 框架的应用。它包含了强大的分析工具，通过拓扑图、调用链、事务将应用状态、调用过程、用户对应用进行的操作可视化地展现了出来，以便您快速定位哪些过程出现了问题或者哪里是需要进行改善的性能瓶颈。

APM 架构特点

随着新技术、新方法的出现，能够快速、敏捷的兼容支持，实现对多层次、复杂技术、混合架构应用的监控分析的需求已经变得越来越迫切。APM 就是在移动化、云化、分布式等新型 IT 架构下，提供自动化、实时监控与分析能力的服务，它可主动运维，辅助优化，保障应用提供稳定持续的用户体验。

APM 通过多层次解耦、轻量化、独立扩展、框架无关的数据采集接入层封装底层技术细节，屏蔽技术变化对上层计算存储层架构的冲击，向上层提供可靠、稳定的数据分析格式，使得计算及存储层、表现层能够以稳定的数据视角，专注于分析计算和展示。

APM 应用场景

1. 通过视图来掌握应用状况

应用部署好以后，通常需要对应用的运行状态等进行监控，比如应用间关系、实例状态、调用次数、时延、错误等。

对此 APM 提供了应用概览、拓扑，针对应用关系、实例、调用进行可视化展示，实现了通过一张图即可完全掌握应用健康状况。

发高峰场景下，时延等关键 KPI 是否达标。

2. 透过数字来剖析事务状态

应用上线以后，运维人员需要掌握用户体验情况，关注用户体验差的地方并进行改进。如何才能迅速掌握用户体验情况呢？

APM 具有对事务实时监控、分析的能力，它监控了从 WEB 客户端或移动终端到服务端全栈业务流，然后对其进行分析，通过 Apdex 专业指标评估事务状态，让运维人员能够透过数字就可以掌握用户体验情况。

3. 通过链条来追踪问题根因

应用上线以后，基础监控可能已经满足不了业务需求，比如系统运行变慢却无法定位瓶颈所在，或者页面打开出错但是无法排查具体调用错误等。

对此，APM 提供了调用链，针对分布式系统的每一次调用进行精细化的监控和跟踪，帮助运维人员精准的找到系统瓶颈所在。

APM 基本概念

1. 拓扑

拓扑是对应用间调用关系和依赖关系的可视化展示（拓扑图）。拓扑图主要是由圆圈、箭

头连线、资源组成。每个圆圈代表一个应用，圆圈上每个分区代表一个实例。每个圆圈中的分数表示活跃的实例/总实例数。分数下的内容分别表示在当前所选的时间中应用的服务时延、应用被调用次数、错误数。

每个箭头连线代表一个调用关系。调用次数越多，连线越粗。连线上的数据表示吞吐量和整体时延。吞吐量即所选时间的调用次数。拓扑使用 Apdex 对用户应用性能满意度进行量化，并使用不同颜色对不同区间 Apdex 的值进行标识，方便用户快速发现问题，并进行定位。

2. 事务

表示一个从“用户请求 > webserver > DB > webserver > 用户请求”的完整过程，通常表现为一个 HTTP 请求。现实生活中，事务即一次任务，用户使用应用完成一项任务，比如电商应用程序中一次商品查询就是一个事务，一次支付也是一个事务。

3. 调用链

调用链跟踪、记录业务的调用过程，可视化地还原业务请求在分布式系统中的执行轨迹和状态，用于性能及故障快速定界。

4. 监控组

可以将某类相同业务的应用放到同个监控组中，并实现整个业务的应用性能管理。例如，可以将账户、产品、支付等应用，放入“商城”监控组中。

5. 采集探针

采集探针通过字节码增强技术进行调用埋点，生成数据。该数据后续会被 ICAgent 采集，之后 ICAgent 会将数据上报并呈现在界面中。开启了内存检测机制后，如果检测到实例内存过大时探针会进入休眠状态，停止数据采集。

6. ICAgent

ICAgent 是 APM 的采集代理，运行在应用所在的服务器上，用于实时采集探针所获取的

数据。

Apdex

Apdex 全称是 Application Performance Index，是由 Apdex 联盟开发的用于评估应用性能的工业标准。Apdex 标准从用户的角度出发，将对应用响应时间的表现，转为用户对于应用性能的可量化范围为 0-1 的满意度评价。

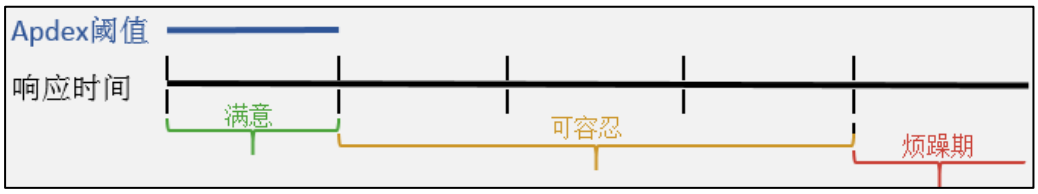
1. Apdex 的原理

Apdex 定义了应用响应时间的最优门槛为 T（即 Apdex 阈值，T 由性能评估人员根据预期性能要求确定），然后根据应用响应时间结合 T 定义了三种不同的性能表现：

Satisfied（满意）：应用响应时间低于或等于 T，比如 T 为 1.5s，则一个耗时 1s 的响应结果则可以认为是 satisfied 的。

Tolerating（可容忍）：应用响应时间大于 T，但同时小于或等于 4T。假设应用设定的 T 值为 1s，则 4*1=4s 为应用响应时间的容忍上限。

Frustrated（烦躁期）：应用响应时间大于 4T。



2. APM 如何计算 Apdex

APM 中，T 即自定义阈值中设置的阈值，应用响应时延即服务时延，Apdex 取值范围为 0～1，计算公式如下：

Apdex=（正常调用次数*1+慢调用次数*0.5）/总调用次数

其中，

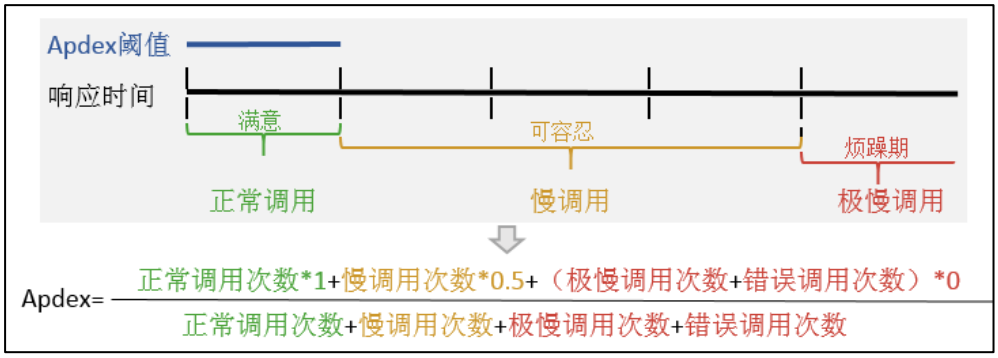
正常调用次数：表示调用耗时大于 0 小于 T 的成功调用数

慢调用次数：表示调用耗时大于等于 T 小于 4T 的成功调用数

极慢调用次数：表示调用耗时大于 4T 的成功调用数

总调用次数：正常调用次数+慢调用次数+极慢调用次数+错误调用次数

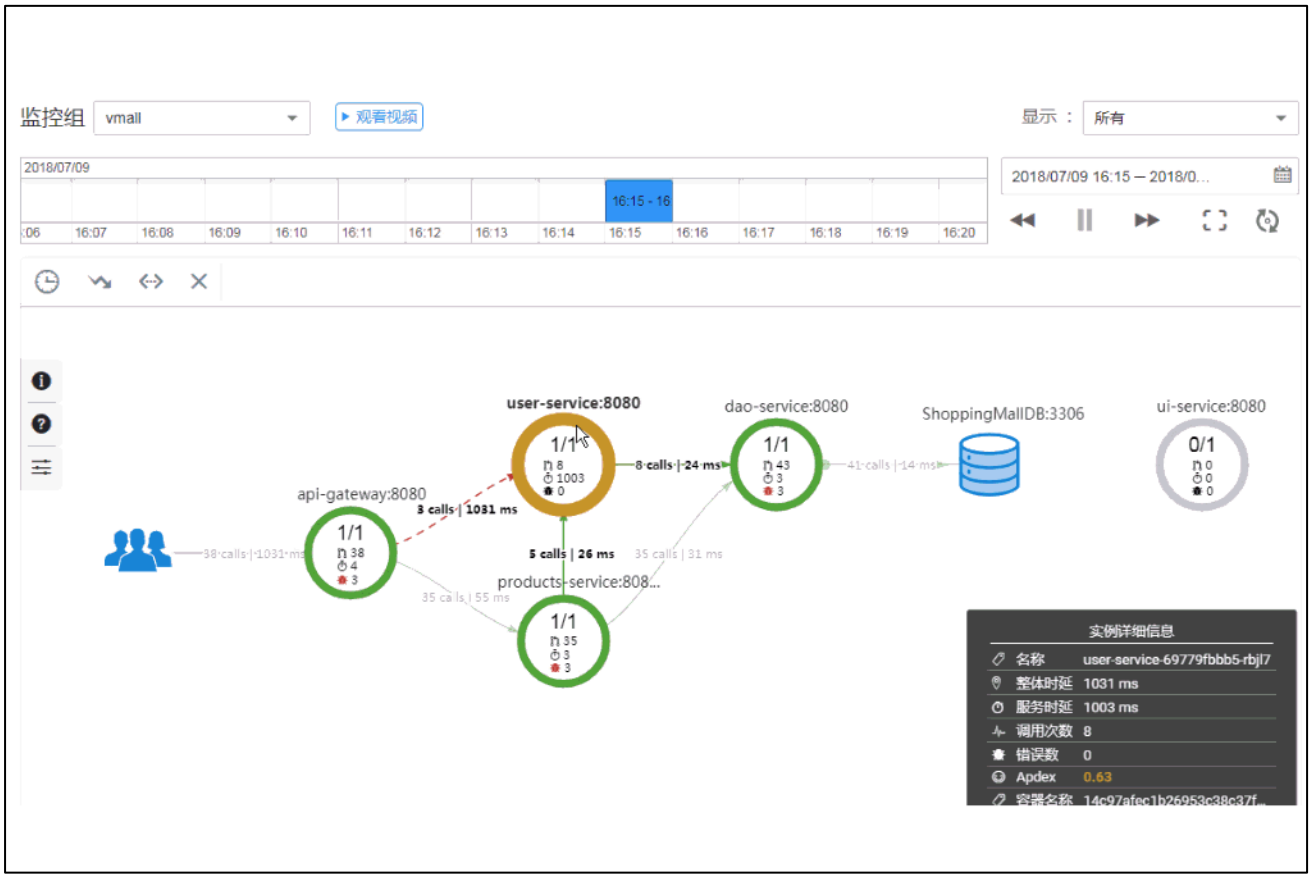
结合上述的 Apdex 原理，则计算公式表示如下：



其计算结果表示应用的不同性能状态，即用户对应用的体验结果，采用不同的颜色表示，如：

Apdex 值	颜色	说明
0.75 ≤ Apdex ≤ 1	绿色	表示应用、实例或事务被调用时响应很快，用户体验较满意。
0.3 ≤ Apdex < 0.75	黄色	表示应用、实例或事务被调用时响应较慢，用户体验一般。
0 ≤ Apdex < 0.3	红色	表示应用、实例或事务被调用时响应极慢，用户体验较差。

3. Apdex 计算举例：



时延

1. TP99 时延

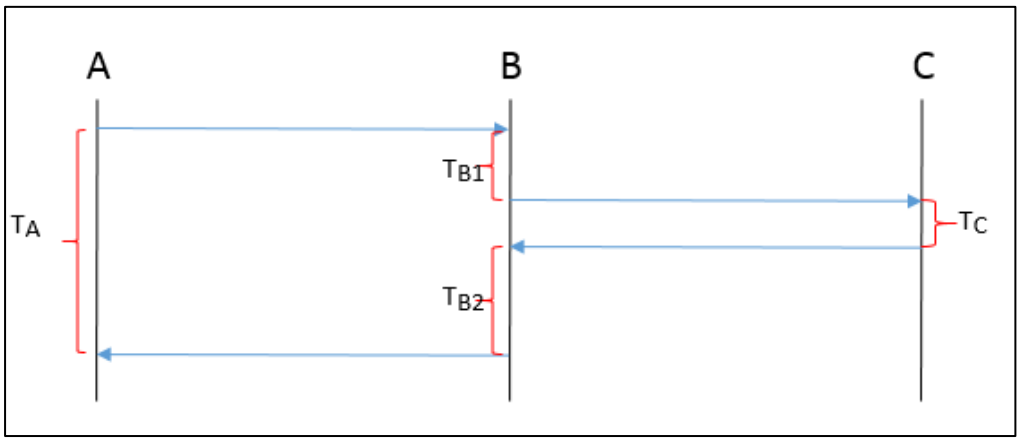
TP99 时延即满足百分之九十九的网络请求所需要的最低耗时。在 APM 中，所有的时延都是指 TP99 时延。

例如，有四次请求耗时分别为：10ms、100ms、500ms、20ms，那么 TP99 时延则是这么计算的：

4 次请求中，99%的请求数为 4*99%，进位取整得 4。也就是 99%的请求次数为 4 次。而满足这 4 次请求的最低耗时为 500ms，那么 TP99 时延为 500ms。

2. 整体时延/服务时延

时延指调用从发起请求到获得响应的耗时。APM 中，整体时延指整个请求的总耗时，服务时延指单个服务的耗时。例如，有服务 A、B、C，A 调用 B，B 调用 C，如下图示：



整体时延=TA，A 的服务时延=TA，B 的服务时延=TB1+TB2，C 的服务时延=TC。

了解更多信息，请访问 [应用性能管理 APM 主页]