

- 降本增效实战利器：Serverless 应用引擎 -

第 07 讲

如何通过压测工具+SAE 弹性能力轻松应对大促

代序 阿里巴巴高级开发工程师

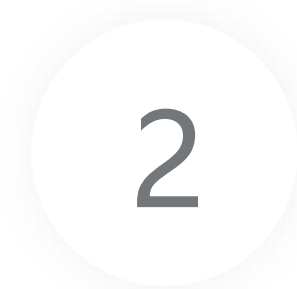


关注“Serverless”公众号
获取第一手技术资料



传统大促挑战

.....



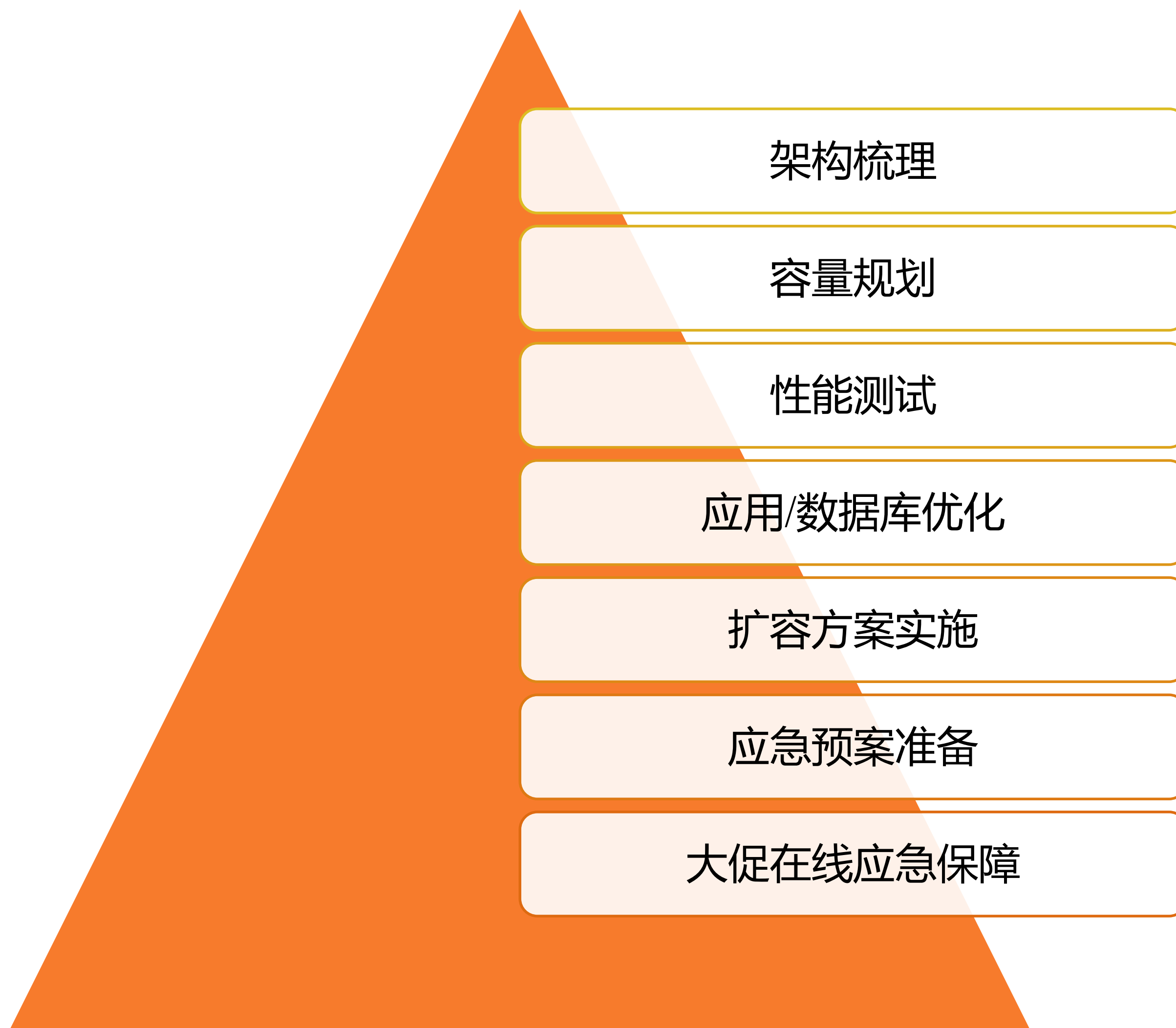
SAE 大促方案

.....



快速压测验证

传统大促活动方案与挑战



常见业务痛点与挑战

- 链路上下游问题，定位问题比较耗时
- 业务开发迭代快，需要常态化压测支持
- 预留资源成本高，需要频繁扩缩容



SAE：面向应用的 Serverless PaaS 平台

无需代码改动，快速上云，并借助 Serverless 能力，快速伸缩，降低运维成本保障业务 SLA

开发者工具/SaaS类服务集成

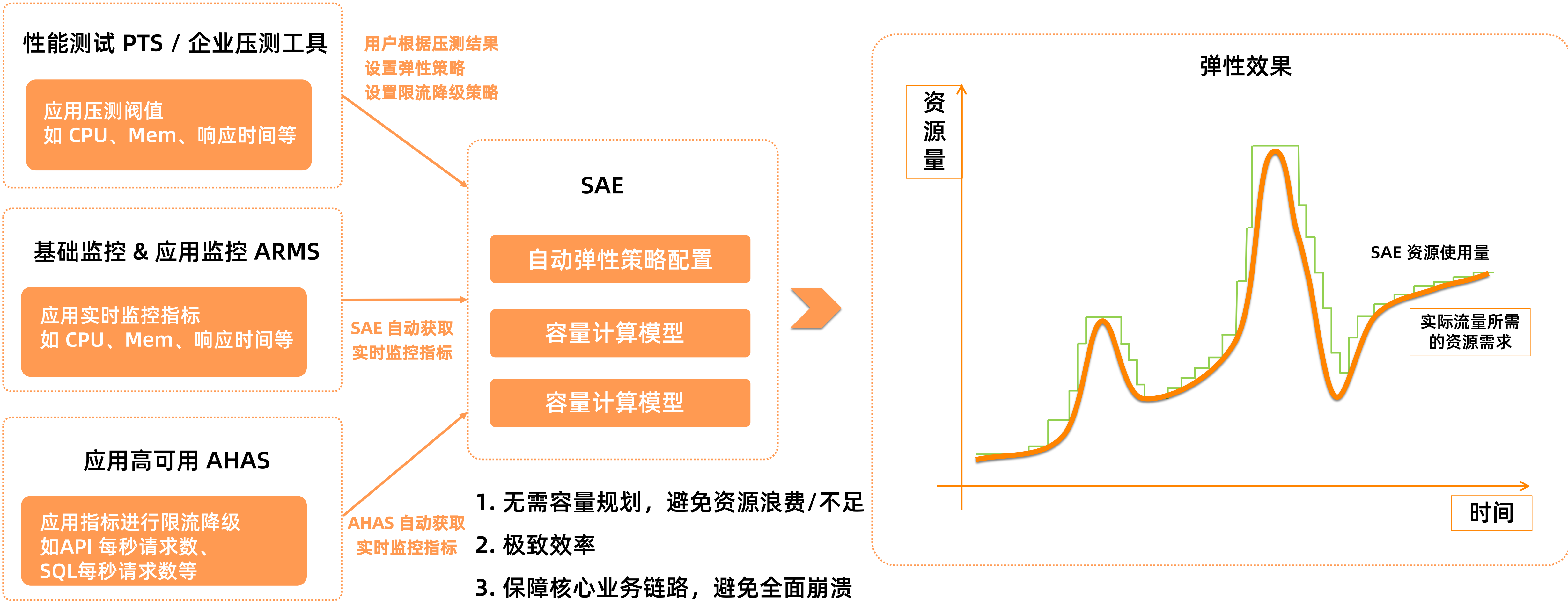


支持应用类型



SAE 精准容量、极致弹性的解决方案

一些比较大流量波动的在线业务（如电商大促，安防行业等），往往出现容量预估不准、弹性效率不及时，很难保证系统 SLA。
采用压测工具 + SAE 弹性后，无需容量规划，秒级自动弹性，轻松应对洪峰流量。





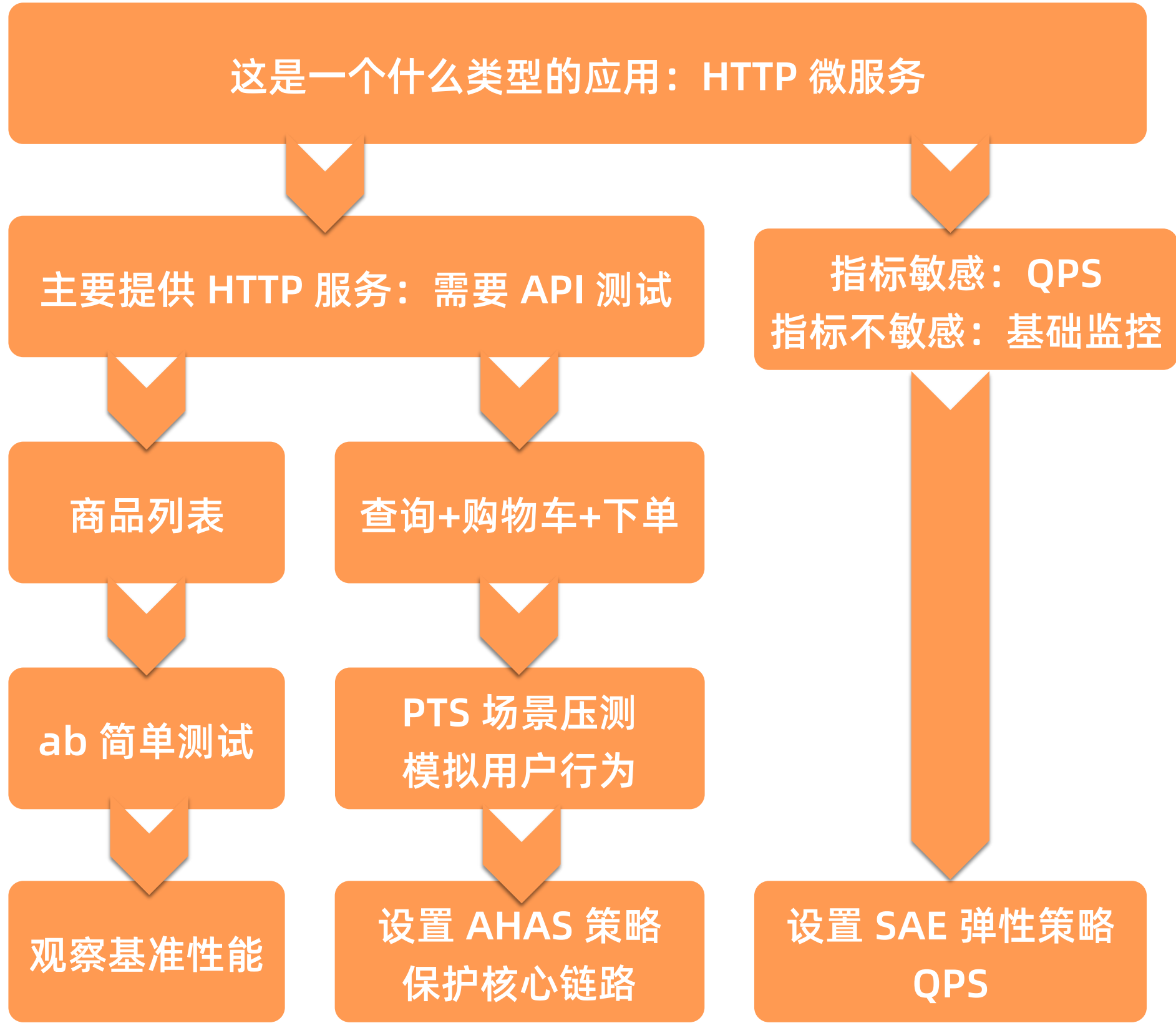
第一步：观察应用监控指标，大致拟定弹性/压测/限流降级

观察应用监控，对日常业务的监控指标，有一个大致的概念。下面以一个典型的电商类应用举例。

观察监控

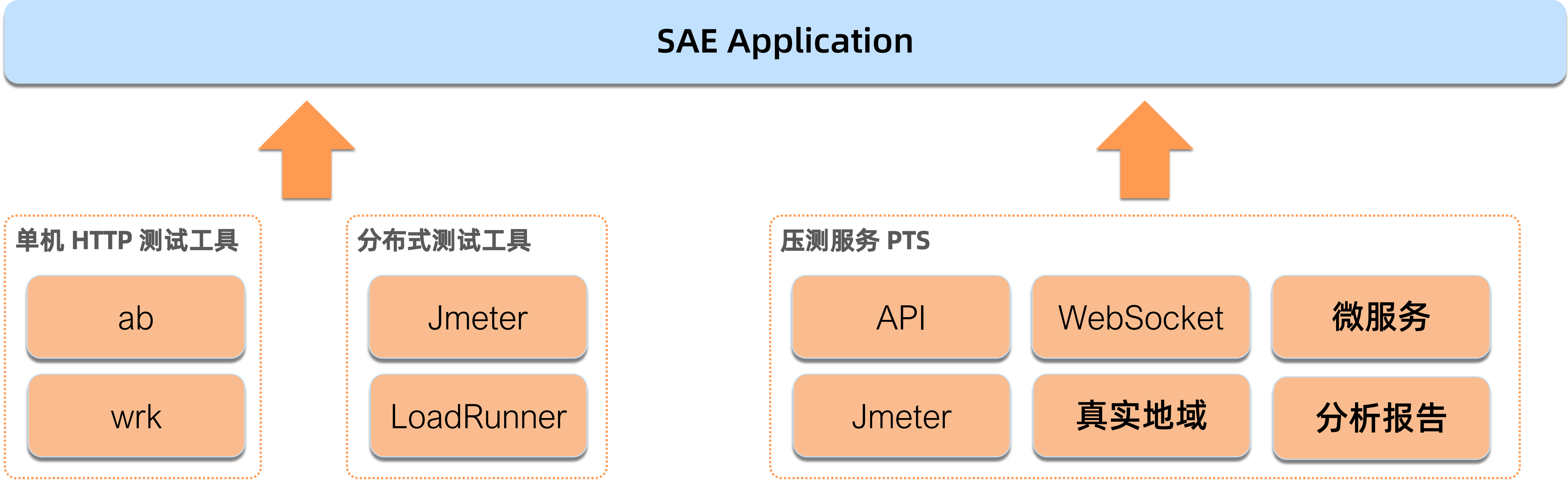


确定方案



第二步：选择合适的压测工具

根据业务诉求，可以选择快速使用的工具，或功能完整的压测工具。



优点：	简单快速	功能强大	功能/费用/成本的平衡；可以提供常态化压测
缺点：	只支持单机 不支持上下文	学习、部署、运维的 成本高	前期有一定学习成本

第三步：配置 SAE 弹性伸缩策略 + AHAS 限流降级策略

无需精准设置，选择一些合适的指标，配置 SAE 弹性伸缩策略，或额外配置 AHAS 限流策略 / ARMS 告警

API 链路

限流

核心 API

核心查询 SQL

弹性

QPS & RT

数据计算型应用

弹性

CPU & Memory

* 接口名称

com.sae.demo.IHelloService:sayHello(java.lang.String)

是否集群流控 ⓘ

☐

* 来源应用

default

统计维度 ⓘ

☒ 当前接口 ☐ 关联接口 ☐ 链路入口

用于接口调用流控。该接口被来源应用调用次数超过阈值时，会对当前接口来自于来源应用的请求进行流控

* 单机QPS阈值

800

流控效果 ⓘ

☒ 快速失败 ☐ 预热启动 ☐ 排队等待

常规流控方式。当前接口超过设置阈值的流量，直接返回默认流控信息，如文本/静态页面等。

是否开启

☒ 该规则打开，创建后即生效

触发条件

CPU使用率

内存使用率

应用QPS ✓ ⓘ

响应时间 (RT) ✓ ⓘ

应用QPS目标值

500

/秒

或 响应时间 (RT) 目标值

200

ms

SAE会自动伸缩实例数，无限接近您设置的监控指标目标值。

当应用监控指标的实际值小于目标值时，SAE会自动缩容实例。反之，SAE会自动扩容实例。[详情](#)

* 最大应用实例数

50

(可选范围：2~50)

* 最小应用实例数

10

(可选范围：1~50)

触发条件

CPU使用率 ✓

内存使用率 ✓

CPU使用率目标值

75

%

或 内存使用率目标值

90

%

SAE会自动伸缩实例数，无限接近您设置的监控指标目标值。

当应用监控指标的实际值小于目标值时，SAE会自动缩容实例。反之，SAE会自动扩容实例。[详情](#)

* 最大应用实例数

50

(可选范围：2~50)

* 最小应用实例数

10

(可选范围：1~50)

第四步：执行压测 – 观察结果 – 优化代码 – 调整策略配置

根据压测与监控结果，看是否有必要优化代码，或调整 SAE 弹性伸缩策略、AHAS 限流策略。

查看压测结果

数据信息

执行
2020-08-10 10:00
2020-08-10 10:00
总共

并发 (峰值/上限)
10000/10000

RPS (峰值/上限) ?
9070/80000

TPS (峰值)
9070

来源 (配置)
40/40分布详情

流量 (峰值/均值) ?
356.75KB/79.71KB

异常数 (请求/业务)
618 / 22万

总请求数
229750

配置信息

压力来源
阿里云内网

压测模式
并发模式

递增模式
手动模式

概览

明细

解读压测报告

压测之后的限流保护

操作记录

业务指标

串联链路 起始 / 最大并发	API名称	APIID	总请求数	平均TPS	请求成功率	业务成功率	平均响应时间
串联链路 10000/10000	API-商品查询		229750	1914.58	99.7(229132/229750) 详情	99.7(229132/229750)	5095.86ms 详情

查看监控异常



再次压测

问题优化

环境变量设置

设置容器运行环境中的一些变量，便于部署后灵活变更容器配置。

如何设置环境变量

类型	变量名称	变量值/变量引用
自定义	JAVA_OPTS	-Xmx3G -Xms3G -Xmn512M -XX:P
自定义		

谢谢观看

THANK YOU



关注“Serverless”公众号
获取第一手技术资料