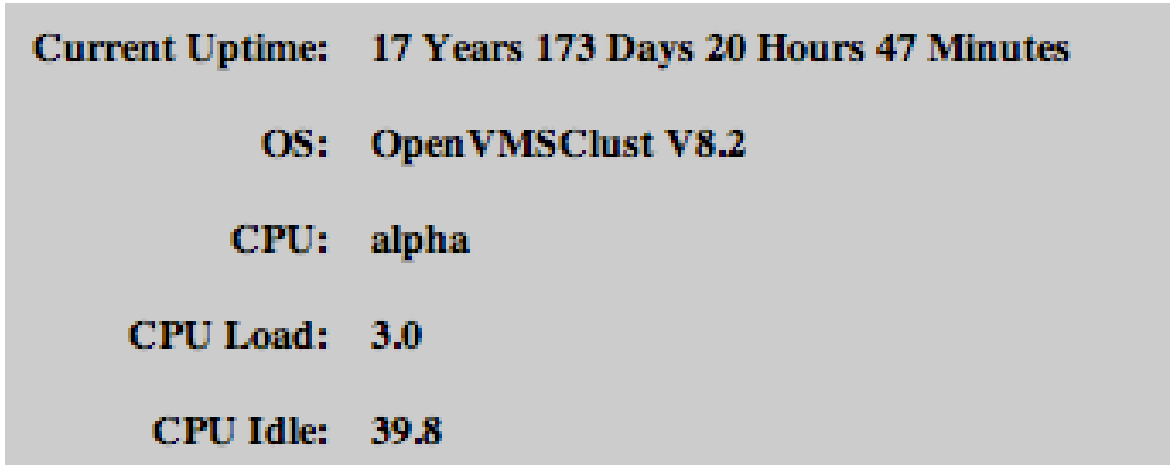

# 互联网系统的稳定性保证：微博的实践

新浪微博 洪小军

@XiaoJunHong

@微博平台架构

- 什么是稳定的系统?
  - 最稳定的操作系统 – OpenVMS




图片来源：uptimes-project.org
数据采集时间：2013年7月12日

- 什么是稳定的系统?
  - 最稳定的反馈控制系统 – 鸡头

图片来源：youtube

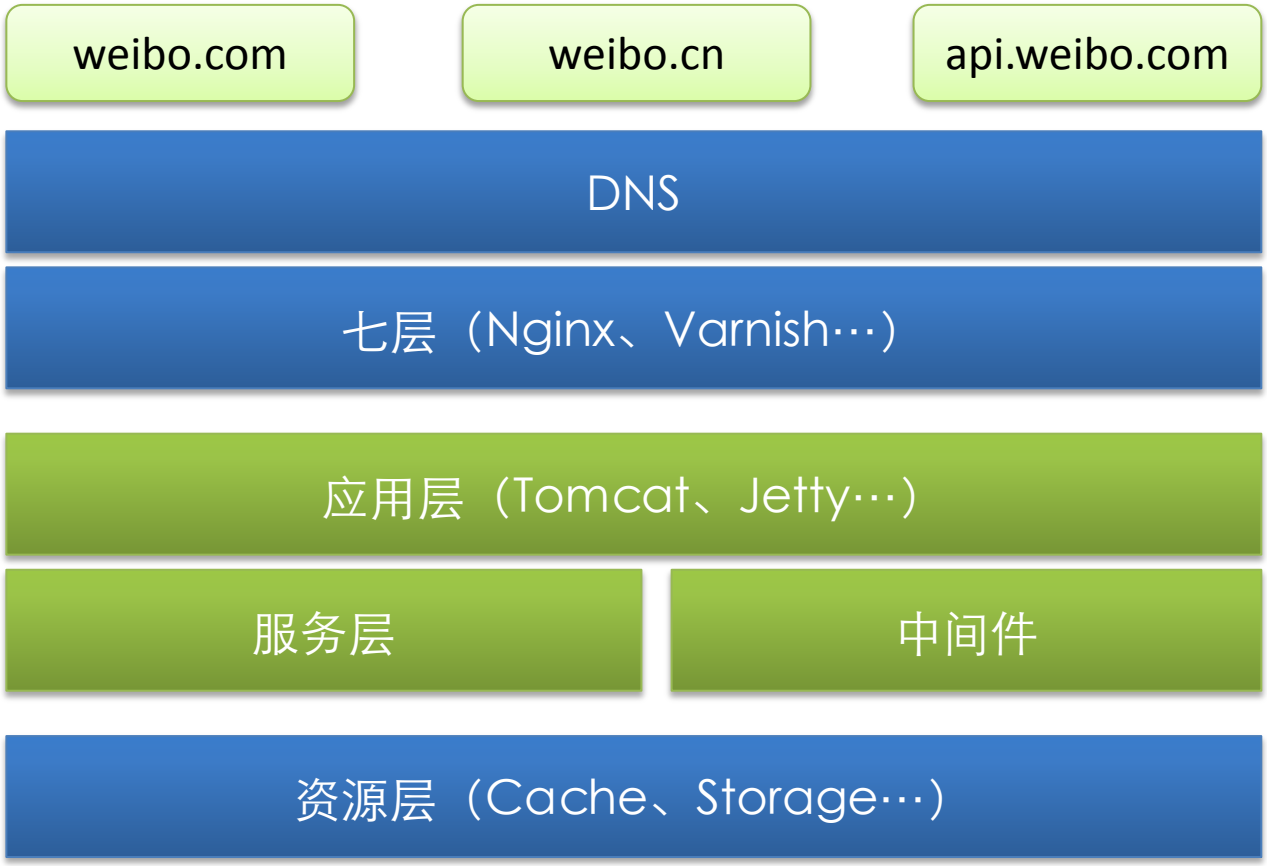
- 什么是稳定的互联网系统?
  - 少出问题
  - 快速解决
  - 清楚系统健康状况趋势

- 影响稳定性的因素
  - 依赖的资源、服务异常
  - 网络、硬件故障
  - 流量异常突增
  - 代码bug
  - 各种“坑”
  - …….

- 构建稳定的系统
  - 少出问题：Design For Failure
  - 快速解决：容灾预案

- 构建稳定的系统 – Design For Failure
  - 分层隔离
  - SLA保证
  - 代码质量保证

- 分层隔离 – 分层模型


weibo.com
weibo.cn
api.weibo.com
DNS
七层（Nginx、Varnish…）
应用层（Tomcat、Jetty…）
服务层
中间件
资源层（Cache、Storage…）

- 分层隔离 – 隔离目标和原则
  - 保证异常出现时影响范围可控
    - ✓ 按主要接入方隔离
    - ✓ 按业务隔离
    - ✓ 按功能核心程度隔离

- 分层隔离 – 隔离方式
  - 物理隔离 | 逻辑隔离
  - 读写隔离

- SLA保证
  - 服务提供方：服务对外的SLA承诺
  - 服务消费方：对依赖资源或服务的SLA要求

- 服务SLA保证 – 超时控制
  - 依赖的资源或服务超时控制
  - 异步调用超时控制

- 服务SLA保证 – 谨慎重试
  - 异常场景下重试可能导致系统持续恶化
  - 对于写入场景存在数据异常风险

- 服务SLA保证 – 容量规划
  - 每季度至少一次例行性评估
  - 重大活动前容量评估
  - 监控系统黄色预警
  - 日常30%以上冗余
  - 资源或系统架构调整时需要重点关注

- 服务SLA保证 – Failover策略
  - 服务降级：保核心功能
  - 快速失败：保证不卡死
  - 流量限制：保正常请求

- 服务SLA保证
  - 超时控制
  - 谨慎重试
  - 容量规划
  - Failover策略

- 构建稳定的系统 – 容灾预案
  - IDC容灾
  - 限流
  - 降级
  - 紧急快速扩容

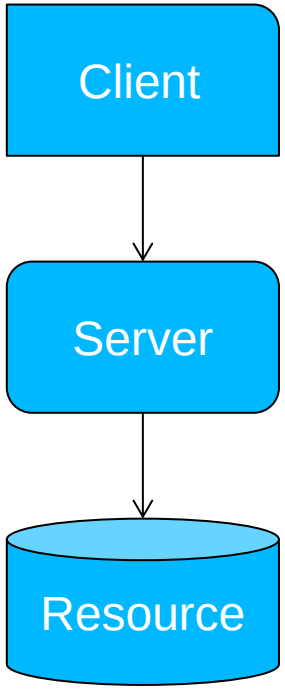
- 这些是否是有效的？是否有遗漏？
  - ✓ 在测试环境下已经做了充分测试！
  - ➢ 线上呢？等待异常出现时来验证系统是否经得起考验？


OR

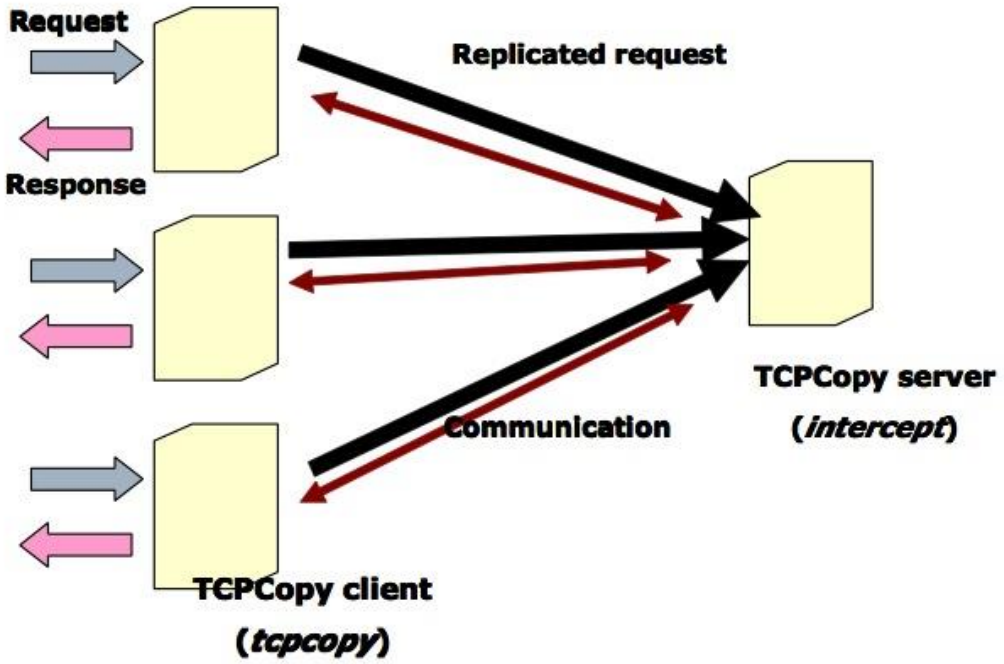
- 线上异常演练 – Touchstone系统
  - **确认**碰撞时安全气囊是打开的（Design For Failure）
  - 即使出现问题事后有**补救**措施（容灾预案处理）


保证影响在预期可控的范围之内！

- Touchstone系统介绍
  - 引流线上真实流量
  - 异常场景搭建和模拟
  - 预案预演
  - 验证系统运行稳定性状况

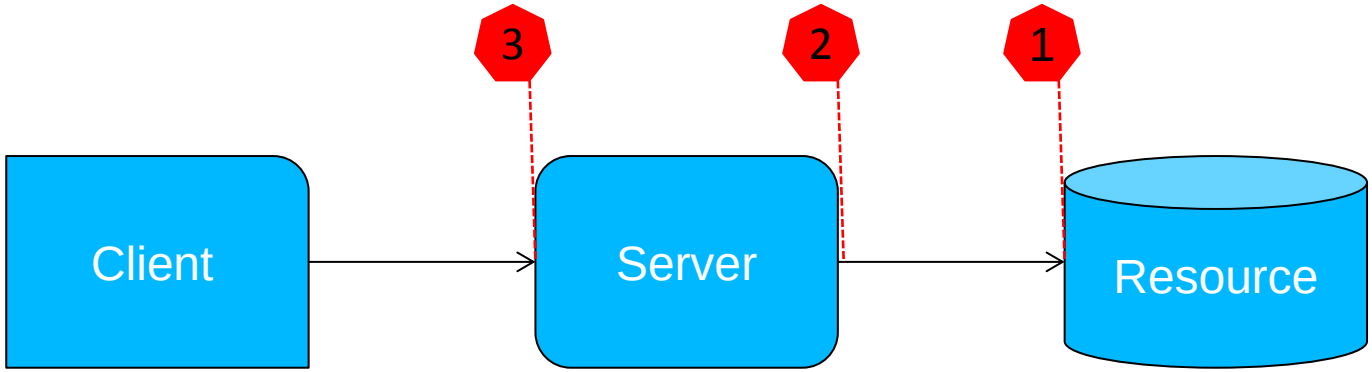


- 引流线上真实流量
  - 主要通过tcpcopy引流到演练服务器




图片来源：tcpcopy官方文档

- 异常场景搭建和模拟

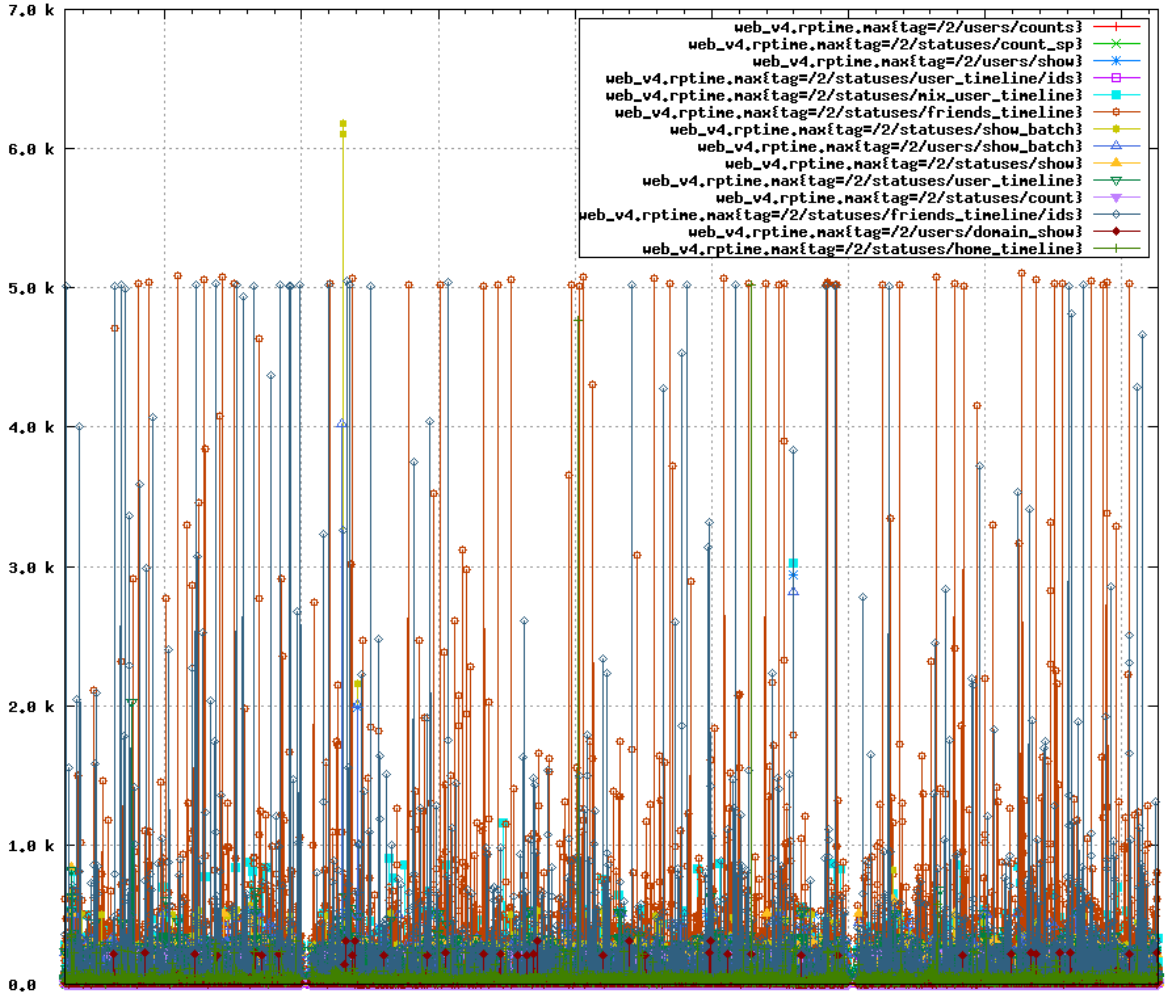



  - 资源或服务提供方搭建真实或对等异常场景
  - 使用linux tc模拟依赖资源或服务出现异常的场景
  - 接口字节码形式注入sleep代码模拟接口慢的场景

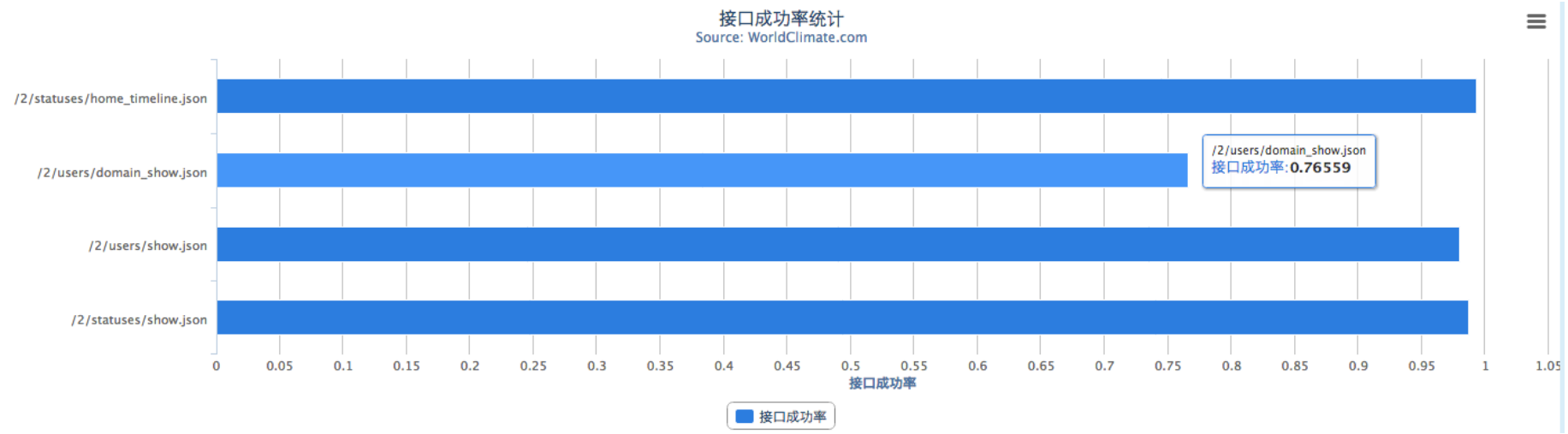
- 预案预演
  - 搭建异常的场景
  - 运维人员做相应预案操作
  - 验证系统运行稳定性状况

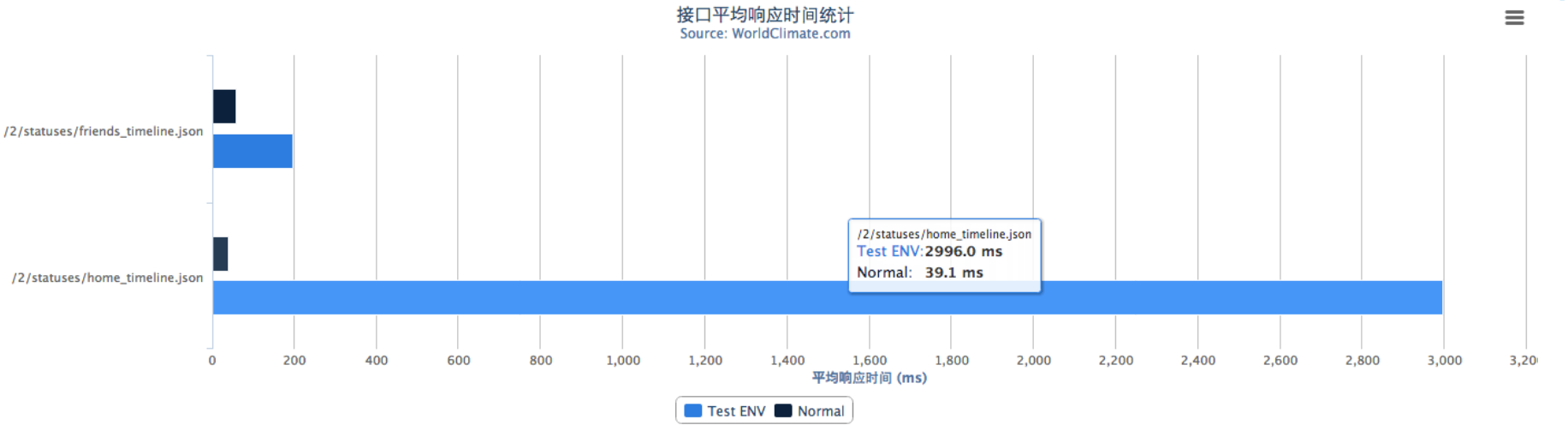
- 怎么判断系统是稳定的?
  - 接口层面
    - 分析状态码中4xx、5xx等比例
    - 响应时间是否在正常范围内
    - 是否满足SLA要求
    - 返回包大小（辅助手段）
    - 日志分析（辅助手段）
    - ……
  - 产品层面

- Touchstone系统 – 实时数据展示



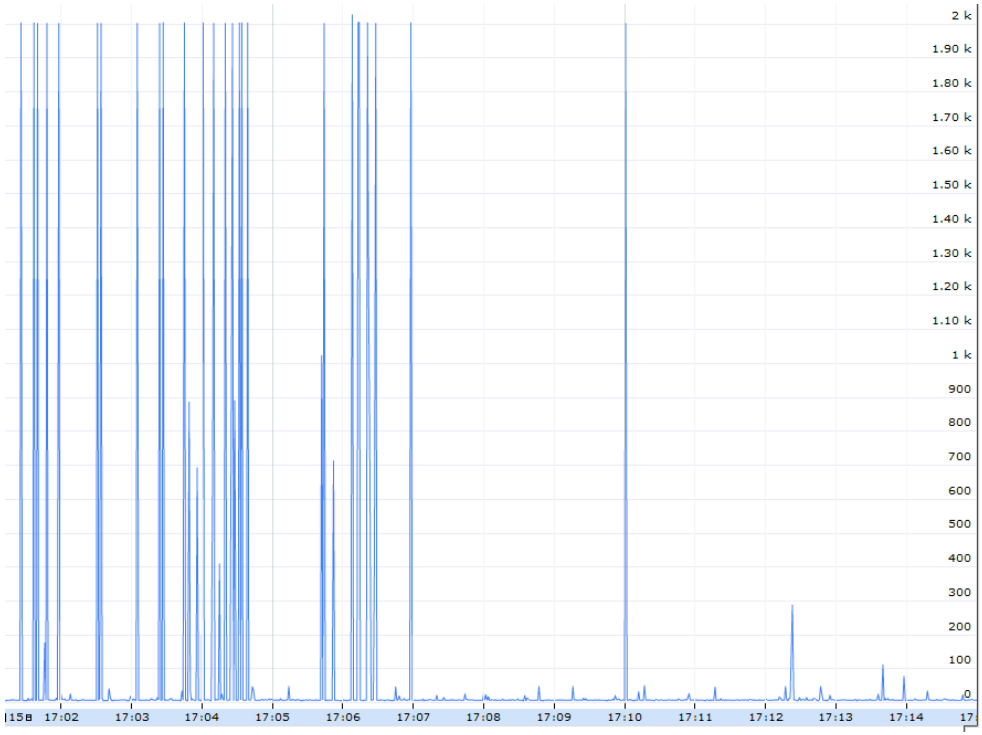
- Touchstone系统 – 报表输出



通过输出的异常数据，怎么分析系统问题点?

- 异常影响程度叠加放大
  - 描述：某组资源延迟400ms，但是接口整体持续延迟2s




串行化操作？

- 功能相关的接口同时受到影响
  - 描述：某资源异常，依赖此资源的功能都受较大影响
  - 依赖性的影响!
  - 思考:
    - 分层隔离中被隔离开的功能是否受到影响?
    - 是否存在非核心资源影响核心功能的情况?
    - 期望的SLA保证是否都生效?

- 大范围大量接口受到影响
  - 描述：某资源异常时，tomcat中大量接口出现503
  - 系统过载?
  - 容器保护策略失效?

- 某个接口的所有请求都受到影响
  - 描述：某个mysql slave节点异常，依赖此资源的接口全部受到影响
  - 资源单点部署？只有一套slave?

- Touchstone系统输出
  - 系统稳定性状况
  - 系统优化改进建议
  - 切实有效的处理预案

- 保证系统一直处于稳定状态
  - 周期性的演练测试
  - 新系统上线和重大改造前先进行演练测试

- 在线演练一些注意事项
  - 避免copy上行接口流量导致写请求被多次处理
  - 避免对后端造成很大压力
  - 避免写花缓存数据
  - 尽量选择在低峰和有工程师在场的时间段进行演练
  - 完善的监控报警机制

# QA

欢迎加入新浪微博!

@微博平台架构
@XiaoJunHong
xiaojun2@staff.sina.com.cn