上海・2020

QCon全球软件开发大会

International Software Development Conference

Brought by InfoQ

上海・2020

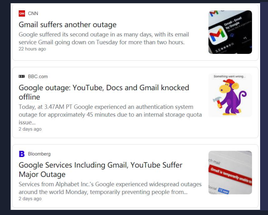
创建高可靠性系统的工程实践

宋涛 - 携程集团 商旅事业部 CTO

简单介绍

- 携程商旅CTO， 2017年加入携程，领导研发新一代机票引擎
- 此前在微软Windows团队，亚马逊云计算团队从事技术和管理工作
- 北京大学计算机毕业，加州大学计算机博士

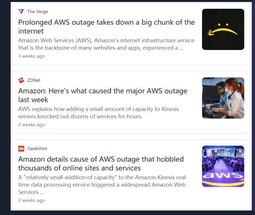
系统可靠性：互联网时代的挑战



2020/12/14

45

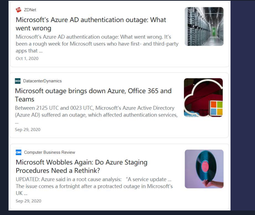
Global



2020/11/27

308

1 Region

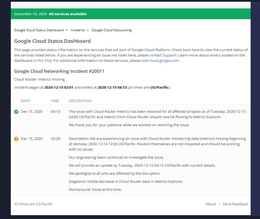


2020/10/27

150

Europe

系统可靠性：互联网时代的挑战

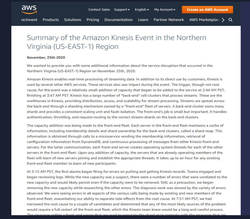


谷歌服务

内部工具

Auth服务

磁盘满了？

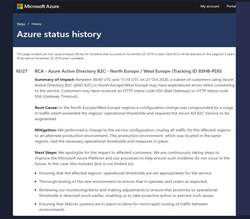


AWS用户

其它AWS服务

Kinesis服务

OS线程数耗尽



微软服务

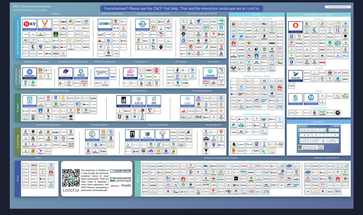
其它Azure服务

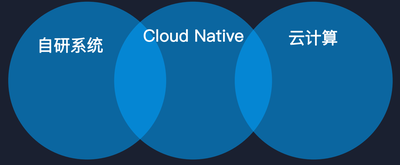
AD服务

流量尖刺

系统可靠性：互联网时代的挑战

- 高流量
- 低延时
- 数据一致性
- 快速上线





架构 开发 测试 运维

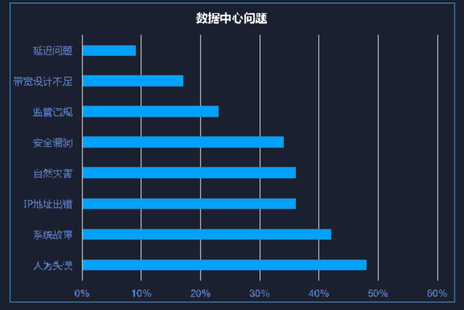
平生不识CNCF，便称架构师也枉然。

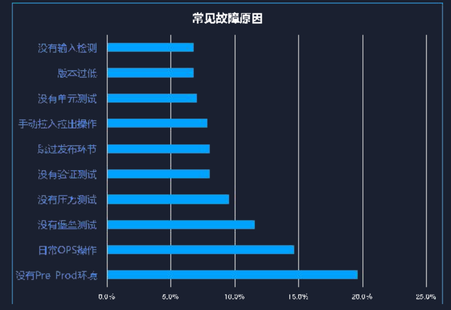
系统可靠性：互联网时代的挑战

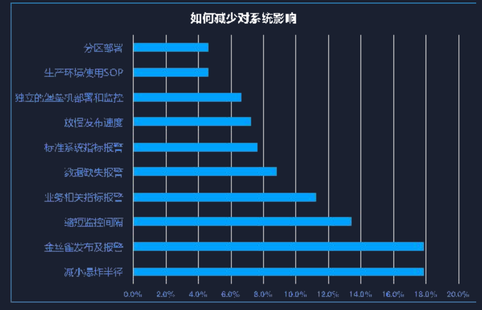
系统可靠性一般是指在规定的时间内和规定的工况下，系统完成规定功能的能力/概率。

可靠性 = f(发生概率, 持续时间, 影响半径)

系统可靠性：影响因素


数据中心问题
延迟问题
带宽设计不足
监管已知
安全漏洞
自然灾害
IP地址出错
系统故障
人为失误
0%
10%
20%
30%
40%
50%
60%


常见故障原因
没有输入检测
版本过低
没有单元测试
手动拉入或出操作
缺过发布环节
没有验证测试
没有压力测试
没有监控测试
日常OPS操作
没有Pre Prod环境
0.0%
5.0%
10.0%
15.0%
20.0%
25.0%


如何减少对系统影响
分区部署
生产环境使用SOP
独立的监控告警和监控
放慢发布速度
标准系统指标报警
数据数据报警
业务相关指标报警
增加监控间隔
金丝雀发布及报警
减小影响半径
0.0%
2.0%
4.0%
6.0%
8.0%
10.0%
12.0%
14.0%
16.0%
18.0%
20.0%

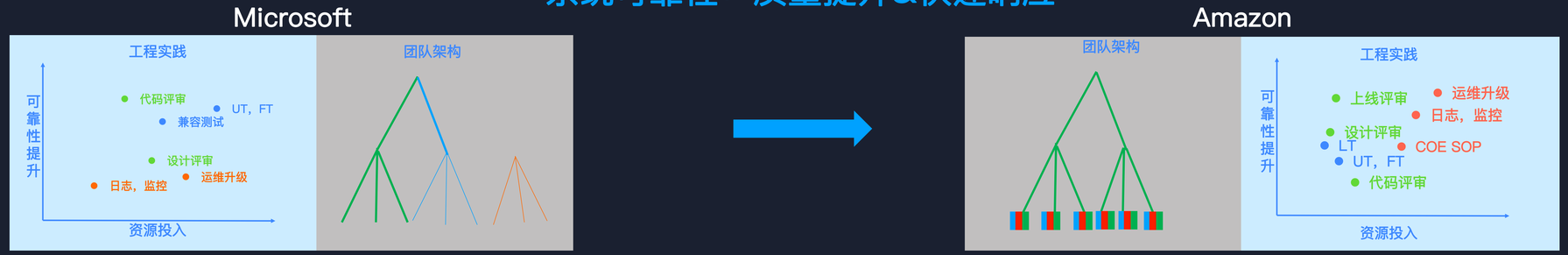
系统可靠性：质量提升&快速响应

设计	开发	测试	发布&OPS	SRE
• 面向失败的设计 • 冗余（副本，主从，灾备） • 隔离（DC，集群，渠道） • Design Review	• 编程规范 • 内部封装开源 • Code Review & Group Code Review	• UT，FT，LT，Fuzzy • Chaos	• 监控 • 报警 • SOP • 灰度发布 • 金丝雀发布	• Oncall • Escalation • 复盘

当质量提升陷入瓶颈或ROI变低时，

通过增强监控和快速响应能持续提升系统可靠性

系统可靠性：质量提升&快速响应



系统可靠性：RCA分享

09:37

09:39

09:41

09:48 有人误删...
09:57 无法冷启动
失败， 重试， 再失败
10:30 上CNN了
13:54 完全恢复

事件报警及时
响应升级迅速
数据备份恢复正常
团队应急有序

生产环境误操作
冷启动多年未试
失败重试缺少规范

TTD（事故监测时间）

TTR（事故恢复时间）

复盘整改

系统可靠性：携程商旅实践分享

Quality is not an act，it is a habit.”
–Aristotle

“品质不是一时的表现，是长久的习惯。”
– 亚里士多德

以自动机制培养良好的工程习惯
(Engineering Excellence)

Sonar，Doc，COE

“Good intentions don’t work，
mechanisms do.”
–Jeff Bezos

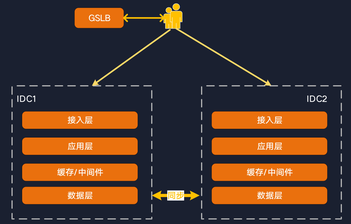
“良好的意愿是没有用的，建立机制才是关键！”

– 贝索斯

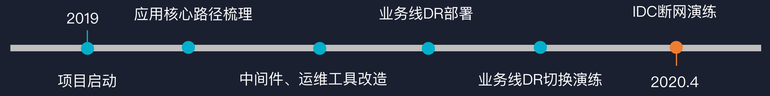
系统可靠性：携程商旅实践分享

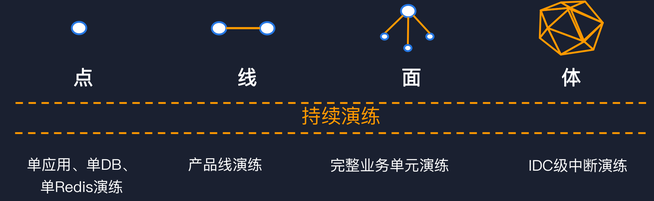
目标：5分钟检测，10分钟恢复

流量地球项目：“当一个太阳（IDC）发生不可用故障的时候，我们地球（业务）还可以继续活下去!”



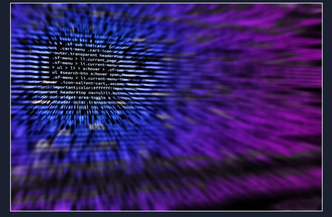
同城双活架构





技术展望

系统架构会更复杂，体量会更大



常规开发测试对可靠性的提升出现边界效应

快速反应 & 控制影响半径会成为关键

上海·2020

THANKS

上海·2020