

SRE 是什么鬼

孙宇聪 (sunyucong@gmail.com)

Coding.net CTO

Geekbang>

极客邦科技

全球领先的技术人学习和交流平台

扫我，码上开启新世界



Geekbang>

InfoQ | EGO NETWORKS | StuQ

InfoQ

专注中高端技术人员
的社区媒体

EGO NETWORKS

EXTRA GEEKS' ORGANIZATION
高端技术人员
学习型社交网络

StuQ

实践驱动的IT职业
学习和服务平台



促进软件开发领域知识与创新的传播



实践第一 案例为主

时间：2015年12月18-19日 / 地点：北京·国际会议中心

欢迎您参加ArchSummit北京2015, 技术因你而不同



ArchSummit北京二维码



[北京站]

2016年04月21日-23日



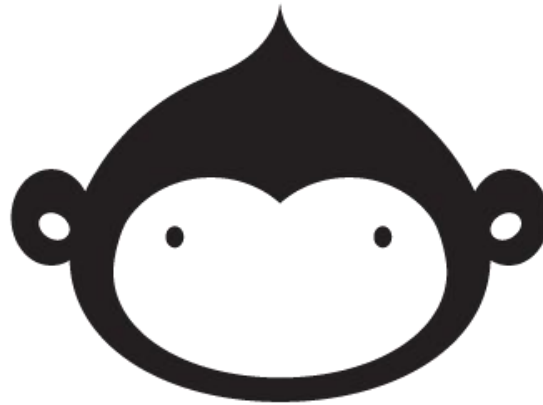
关注InfoQ官方信息
及时获取QCon演讲视频信息

Google SRE 07-14

- YouTube Streaming
 - Video transcoding, streaming, storage
(> 1PB/month)
- Global CDN network
(> 10K nodes, peaking 10Tbps egress).

Google SRE 07-14

- Google Cloud Platform
 - Machine lifecycle management
(> X clusters globally, > Y machines)
- Borg , Omega
(> X million jobs scheduled every week)



CODING
CLOUD DEVELOPMENT

SITE RELIABILITY ENGINEERING

说白了就是 DevOps 一回事

Site

- 生产线管理员
 - *Ensure user-visible uptime and service quality*
 - *Authority over production environment.*
- 跟网站一起成长
 - *Steep learning curve, mostly due to complexity*
 - *Continuous retraining, sites always being improved*
- 基础架构设施
 - *Specializations for shared infrastructure*
 - *Ensure those components have good reliability*

Reliability

- *it just works*
 - *Service Level Objective (SLO)*
 - *Monitoring/Deployment*
 - *Capacity Planning*
- *以一敌百*
 - *Team manages monitoring and develops automation*
 - *Implies use of scripting and data analysis tools*
 - *Most failures need automated recoveries in place*
- *救火队员和纵火犯合体*
 - *Elevated risk during convenient working hours*
 - *Learn of age mortality risk during preceding workday*
 - *Infant mortality ideally also avoids meals*

Engineering

- 码农
 - *Not administration*
- 报警系统重度（中毒）用户
 - *Holes may cause outage before notification occurs*
 - *Routinely use multiple layers, levels and viewpoints*
 - *Design the manual and automatic escalation paths*
- 对未来负责
 - *Responsible for enabling growth and scaling*
 - *Plan for requirements, identify inefficiencies*
 - *File bugs and, where appropriate, fix them too*

Who are SRE

- 跑偏了的程序员
 - 50-50 mix of software background systems engineering background.
- 重度强迫症和处女座
 - “a team of people who fundamentally will not accept doing things over and over by hand.” – Ben Treynor
- 脸皮厚
 - DEV / OPS

Eternal conflict

- **DEV**

- The incentive of the development team is to get features launched and to get users to adopt the product.

- **OPS**

- The incentives of a team with operational duties is to ensure that the thing doesn't blow up on their watch.

一图看懂组织结构

- *BOSS*
- 产品线
 - 小*BOSS*
 - 艺术类
 - 开发团队
- 生产线
 - *APP SRE*
 - *Infrastructure SRE*
 - 数据中心运营
 - 供应链

组织结构

- 以各产品线为核心，松散的学习型组织
 - *Get Incentives right.*
 - *SRE is a privilege, not a right.*
 - *Free to move, Free to leave bad service.*
- **SRE 要做什么 SRE 说了算**
 - *Production Readiness Review (PRR).*
 - *ROI matters most for SRE*
 - *SRE resource is limited*
 - *High marginal benefits work.*

Early phase

- *SRE gives guidance in automating routine tasks*
 - *Reduces workload by eliminating administrivia*
- *SRE points out errors, omissions in documents*
 - *Developer might then beg others for assistance*
- *SRE suggests additional long term monitors*
 - *These fill in coverage gaps and track performance*
 - *Administrators need sufficient, trustworthy monitoring*

Mature phase

- *The decisions become progressively longer term*
 - *Daily task workload for a site is getting reduced*
 - *Software improvements are tuning and analysis*
- *The developer still has a short term viewpoint*
 - *Working on the next release, fixing known bugs*
 - *The old live releases start to be a distraction*
 - *An obvious incentive to request site transfer to SRE*

ONCALL PHASE

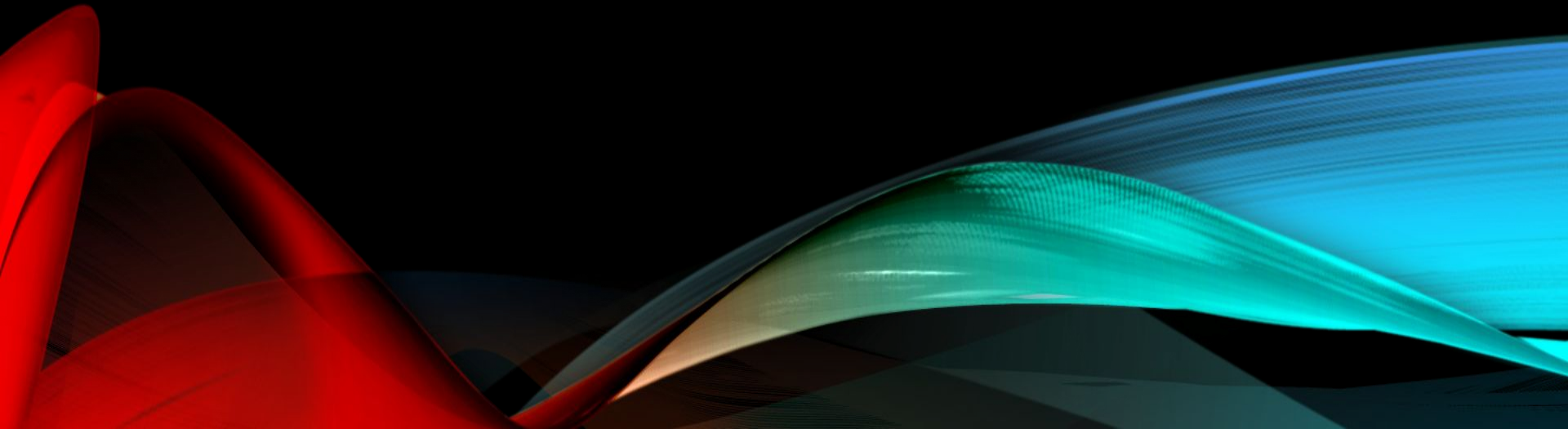
- *On call – more than quick fixes*
- *SRE team members take turns.*
 - *Fix any problem whose solution is not yet automated*
 - *Accumulate occurrence counts to identify priorities*
Document the effective diagnostics and solutions
- *The permanent solution takes a lot more time*
 - *File bug, develop patch, test, code review, submit*
 - *Schedule for integration, release and deployment*
 - *Why spend many hours or days doing all that?*

Deployment model

- Following the sun.
- Only one engineer responds to any given alert
 - Use a priority or escalation rule to avoid wasted effort
 - The other SREs on call are unlikely to be disturbed
- Redundancy everywhere!
 - What is the failure rate of your paging services?
 - Hopefully better than 10%, unlikely to achieve 1% eg:
5% with four way redundant paths is 99.999%

SRE Best practice

How to build a good SRE team.



CMM maturity model

- Level 1 – Initial (Chaotic, Heroic)
- Level 2 – Repeatable
- Level 3 – Defined
- Level 4 – Managed
- Level 5 – Optimizing

如何有成效的填坑

OPS OVERLOAD

- *Reduce complexity*
 - *Less dependencies, configuration types, interfaces.*
 - *Knowledge sharing*
- *Assume there are no humans operating.*
 - *Refocus human involvement*
- *Quarterly Service Review*
 - *Hard cap 50% ops load.*
 - *Provide career path*

正确的和开发团队掐架

SLO Budgeting

- *Establish SLO Goal*
 - *Nothing is 100% reliable.*
- *Spend error budget*
 - *On bad releases, technical debt etc.*
 - *FREEZE on blown budget.*
- *No more arguing*
 - *Its' physics!*
 - *Self-policing Incentives.*
 - *Moral authority*

灾难级别分类 FAILURES

- Failover with minimal delay, near full quality service
- Failover with significant delay, near full quality service
- Partial or limited service, with good to medium quality
- Prevent crash with no or very limited service, low quality
- Crash without data loss or corruption
- Crash with data loss
- Crash with data loss, corruption, destruction

安全生产两大指标 MTBF / MTTR

- $Availability = f(MTBF, MTTR)$
- You can make it fail very rarely, or you are able to fix it really quickly when it does fail.
- Typically, no human will respond in less than two minutes to something that goes wrong.
- It is that the human correctly assesses the situation and takes the appropriate corrective actions, versus diagnosing incorrectly or taking ineffective steps.

如何设计不会坏的系统

Design: Failures

- *Defense in depth*
 - All the different layers of the system can/will fail..
 - User exp must not be affected.
 - No human involvement.
- *Graceful degradation*
 - Caching / Time shifting
 - Failover
 - Redundant Instances , $N + 2$
 - Localization of issue.

如何正确的花钱买机器

Capacity planning

- *what N+M do you run your services at?*
 - *"I don't know, because we've never assessed what the capacity of our service is."*
- *Tips*
 - *How to benchmark service,*
 - *How to measure its response to 100% or 130% of peak.*
 - *How much spare capacity you have at peak demand time*
- *Expect frequent outages and lots of emergencies.*

正确的实现监控系统 DESIGN: monitoring

- Alerts
 - which say a human must take action right now. Something that is happening or about to happen, that a human needs to take action immediately to improve the situation.
- Tickets.
 - A human needs to take action, but not immediately. You have maybe hours, typically, days, but some human action is required.
- Logging.
 - No one ever needs to look at this information, but it is available for diagnostic or forensic purposes. The expectation is that no one reads it.

实战演习

Wheel of misfortune

- Operational readiness drills.
- Tips
 - Picking a disaster
 - Role playing
 - Observe
- Drill people on the correct response to emergency situations until they don't have to think about it.
- Culturally compatible.

地球级别的灾难演习

D.I.R.T

- *Simulated disaster recovery.*
- *Total site loss.*
- *Incident management.*
- *Business continuation.*

POSTMOTERM

- *Blameless postmortem*
 - *About process and technology, not people*
- *Readable & shared to wide variety of readers.*
- *No over-engineering*

POSTMOTERM

- *Capture the facts*
 - *Impacted services and magnitude*
 - *Incident timeline*
 - *Key contact info*
 - *Data*
- *Root cause and trigger analysis*
 - *5 Whys*
 - *Why was this possible in the first place.*

POSTMOTERM

- Lessons learned
 - What went well
 - What went wrong
- Action Plan
 - Investigate, management of issue
 - Mitigation
 - Prevention.
- A problem is resolved to the degree that no human being will ever have to pay attention to it again.

Q & A

扣钉 <https://coding.net>

