来自 GOOGLE 的 DEVOPS 经验

孙宇聪 (sunyucong@gmail.com)
Coding.net CTO

CODING

CLOUD DEVELOPMENT

GOOGLE CLOUD PLATFORM

- Machine lifecycle management
  (> X clusters globally, > Y machines)

- Job Scheduling
- Borg , Omega
  (> X million jobs scheduled every week)

- GCE, GAE, …

YOUTUBE STREAMING

- Video transcoding, streaming, storage (> 1PB/month)
- Global CDN network (> 10K nodes, peaking 10Tbps egress).
- Olympic 2008 Live streaming

天下运维是一家

炮兵连中的炊事兵

开发人员的日常工作

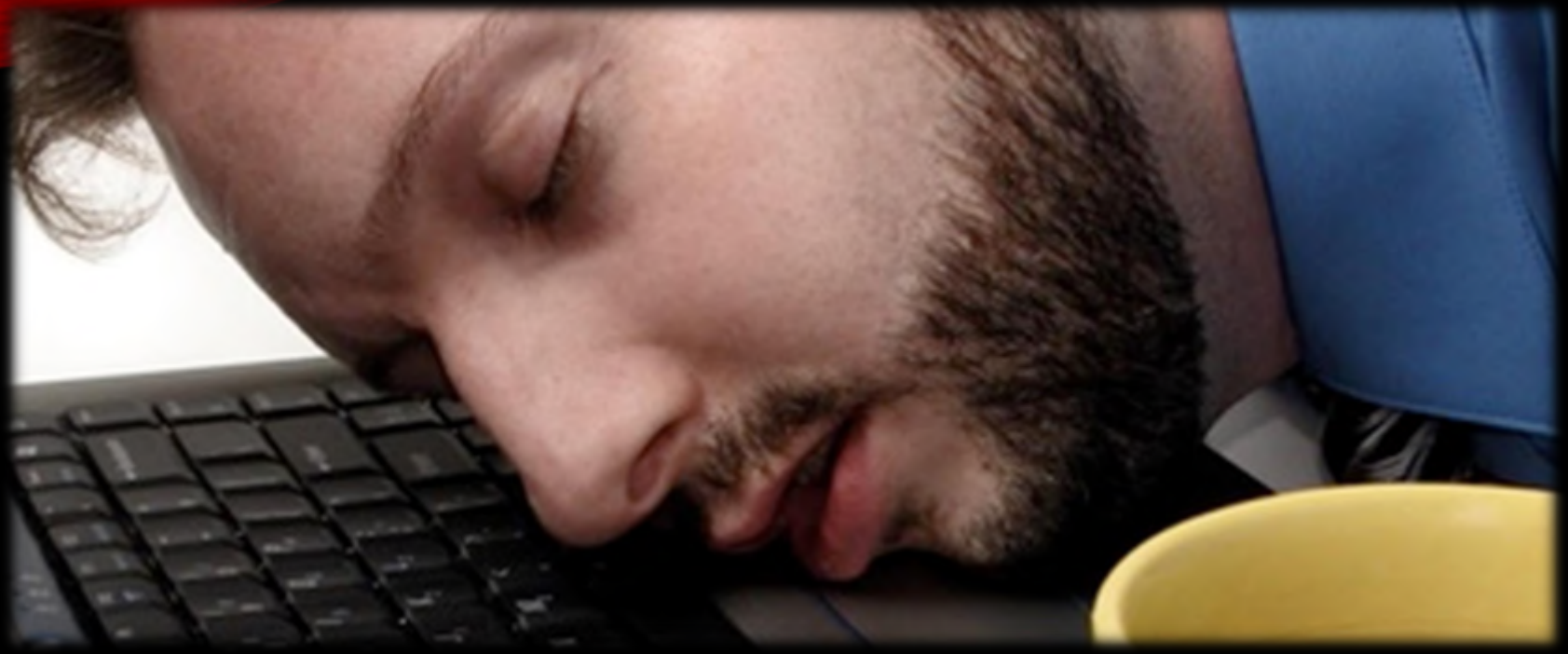
What Developers do

运维的日常工作

What I do everyday

运维对自己的看法

How I see what I do



老板对运维的看法

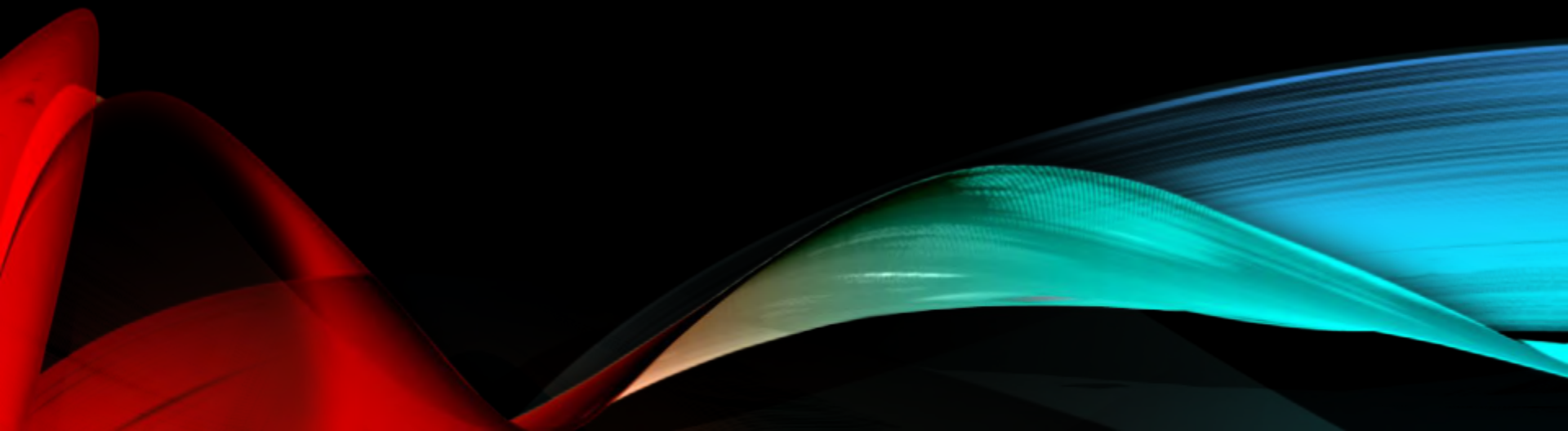
What my boss thinks I do

其实这才是我的日常工作

What I really do

SITE RELIABILITY ENGINEERING

Devops @ Google

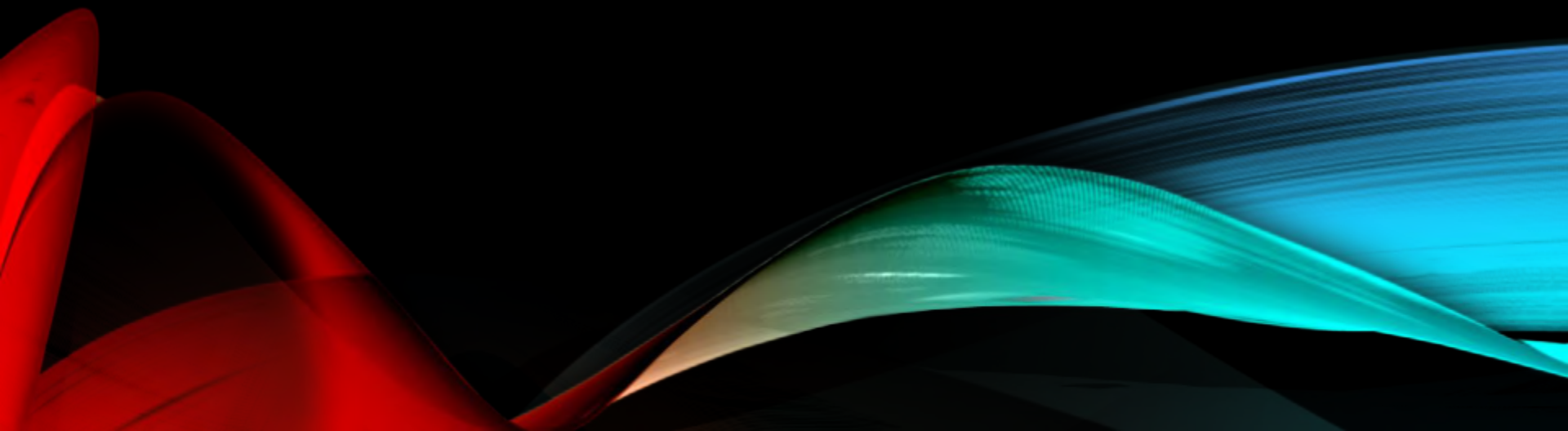


SRE

SITE	RELIABILITY	ENGINERRING
主导生产环境	确保业务连续性	码农
与业务共同成长	制定、监控 SLO	监控系统中毒用户
对最终用户负责	代码化、自动化、无人化	对未来负责

WHO: SRE

Developer tasked with Operation



双技能树的程序员

50-50 mix of software background and systems administration background.

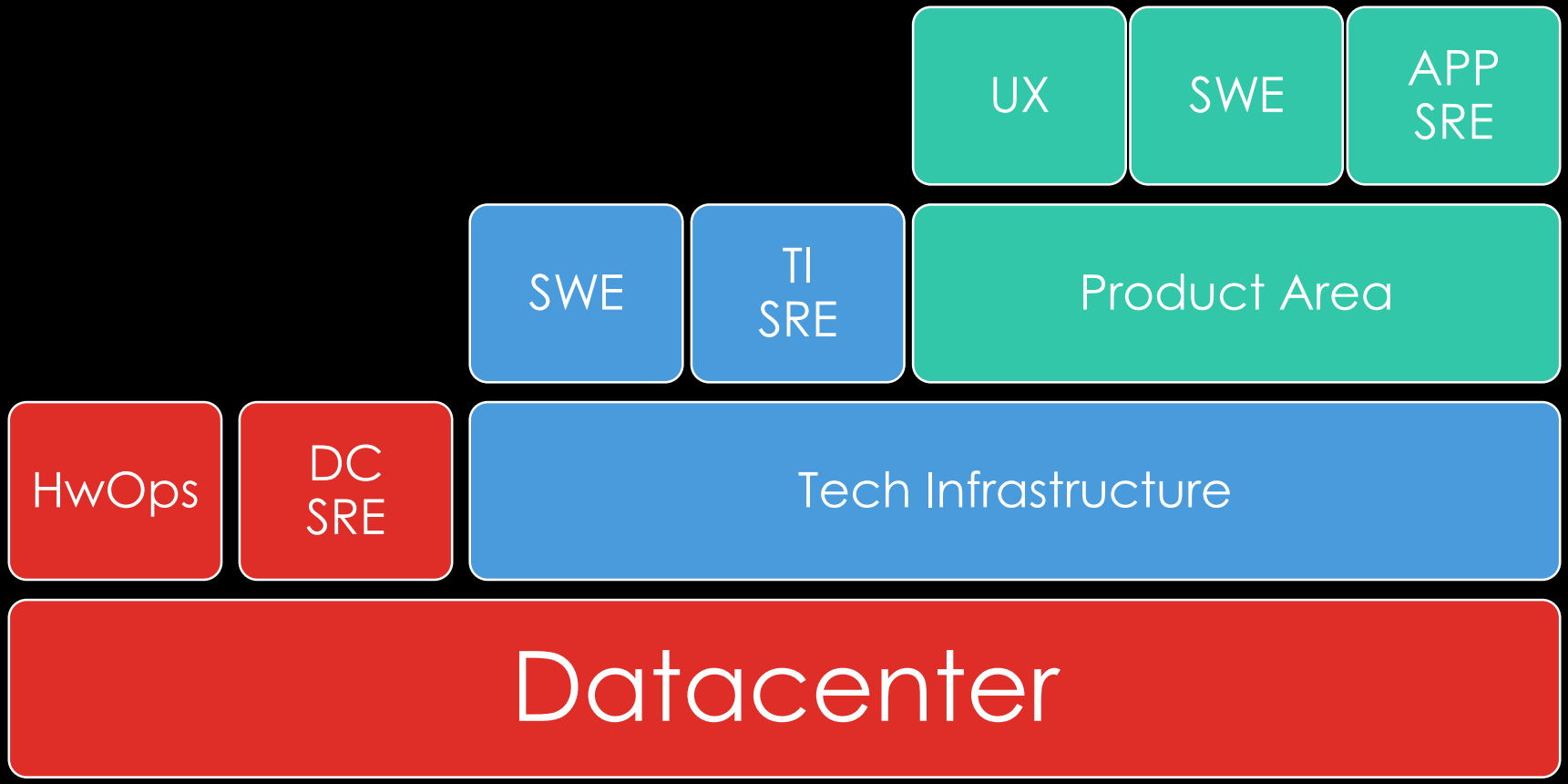
重度强迫症和处女座

“a team of people who fundamentally will not accept doing things over and over by hand. “ – Ben Treynor

该说不绝对说不

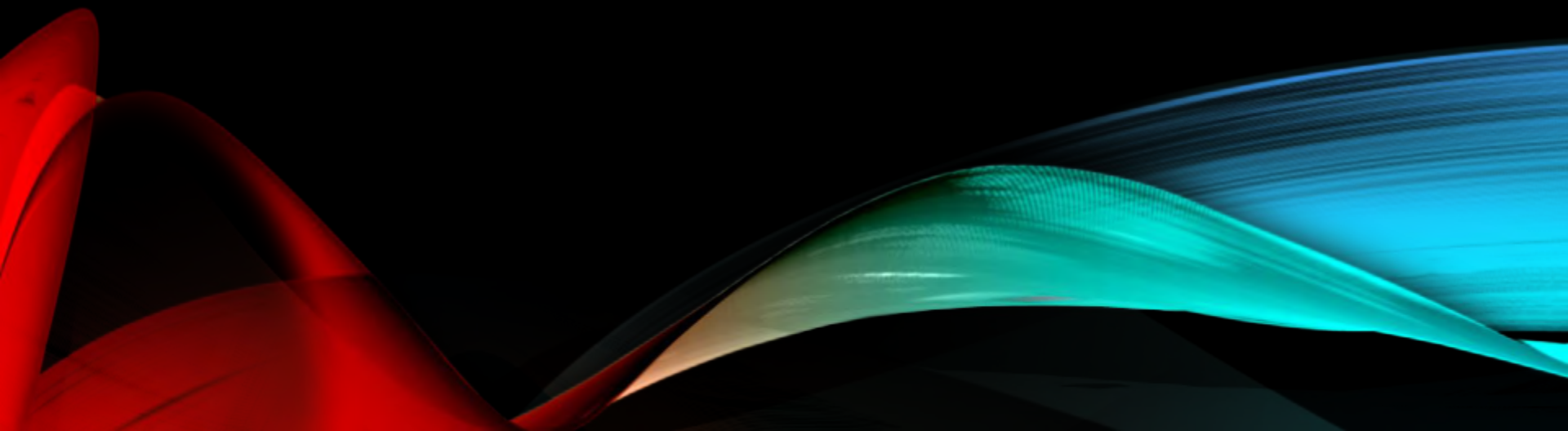
DEV / OPS eternal conflict

ORG CHART


UX
SWE
APP SRE
SWE
TI SRE
Product Area
HwOps
DC SRE
Tech Infrastructure
Datacenter

WHAT: SRE

纵火犯、救火队员 、兼职开发



持证纵火

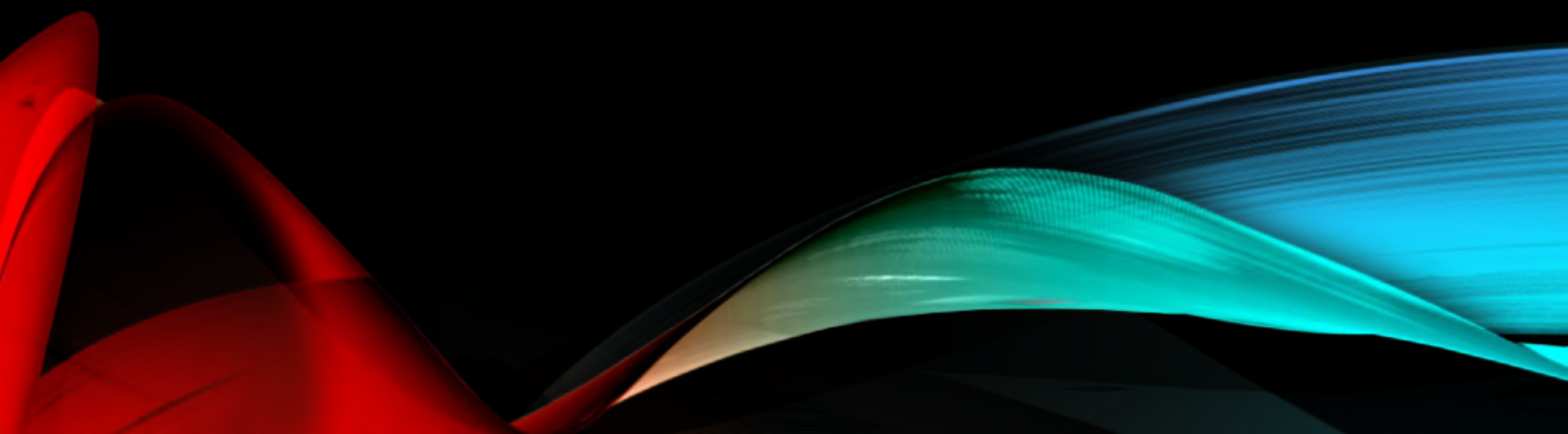
- Infant mortality risk (T0 - T+48H)
- Gradual Rollout
- Vital Signs

救火队员

- On-call (5min)
- Business continuation
- Incident Management

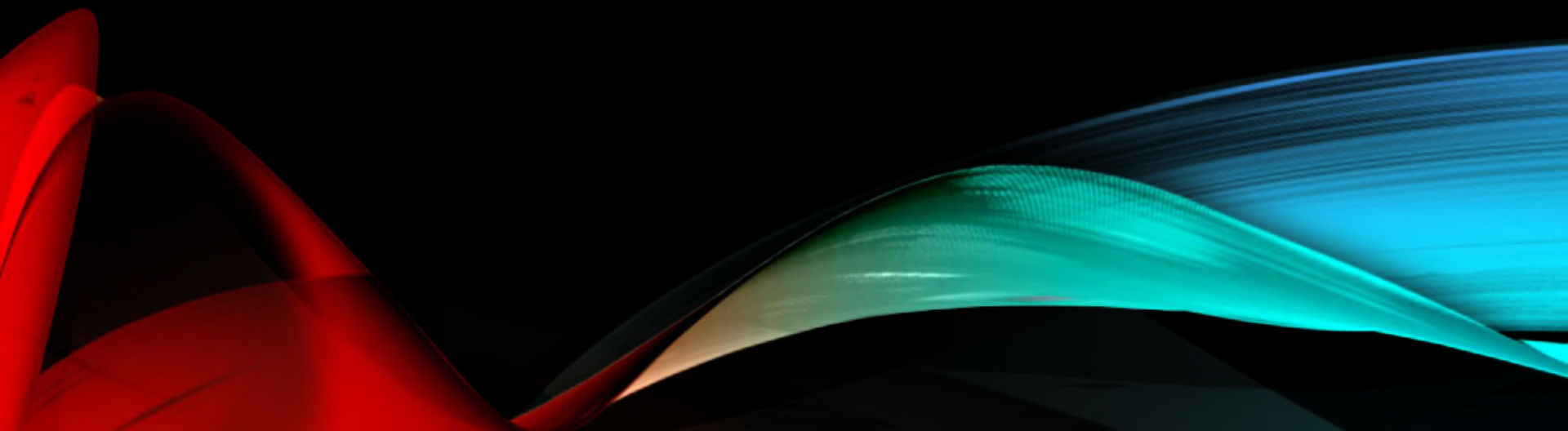
兼职开发

- 监控
- 可管理性、高可用性架构
- 无人化



HOW: SRE

如何做好社会主义接班人



社会主义初期阶段

- **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

业务成熟期

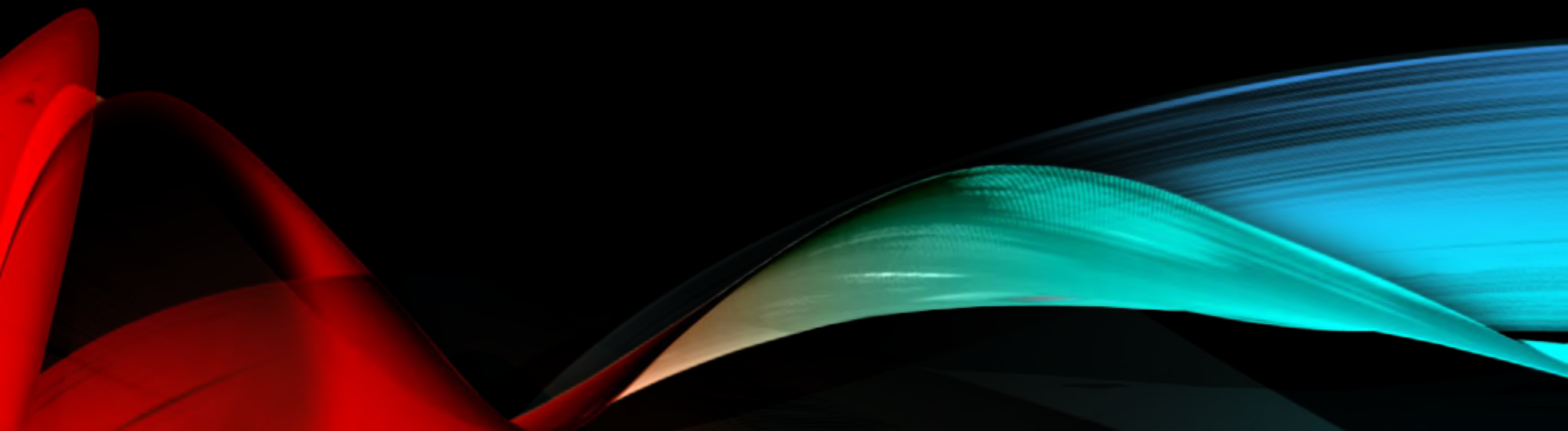
- **The decisions become progressively longer term**
  - Daily task workload for a site is getting reduced
  - Software improvements are tuning and analysis
- **The developer still has a short term viewpoint**
  - Working on the next release, fixing known bugs
  - The old live releases start to be a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

ONCALL 阶段

- **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?

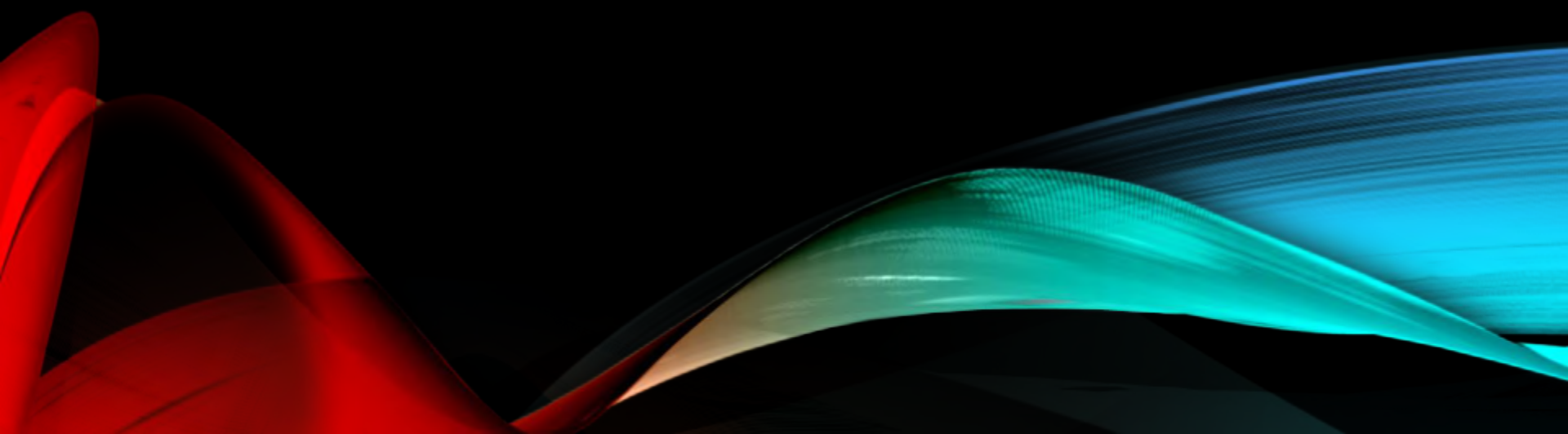
BEST PRACTICES: SRE

一些理念的分享



有成效的填坑

- 强调可管理性、降低复杂度
- SLO Budgeting
- Build Infrastructure

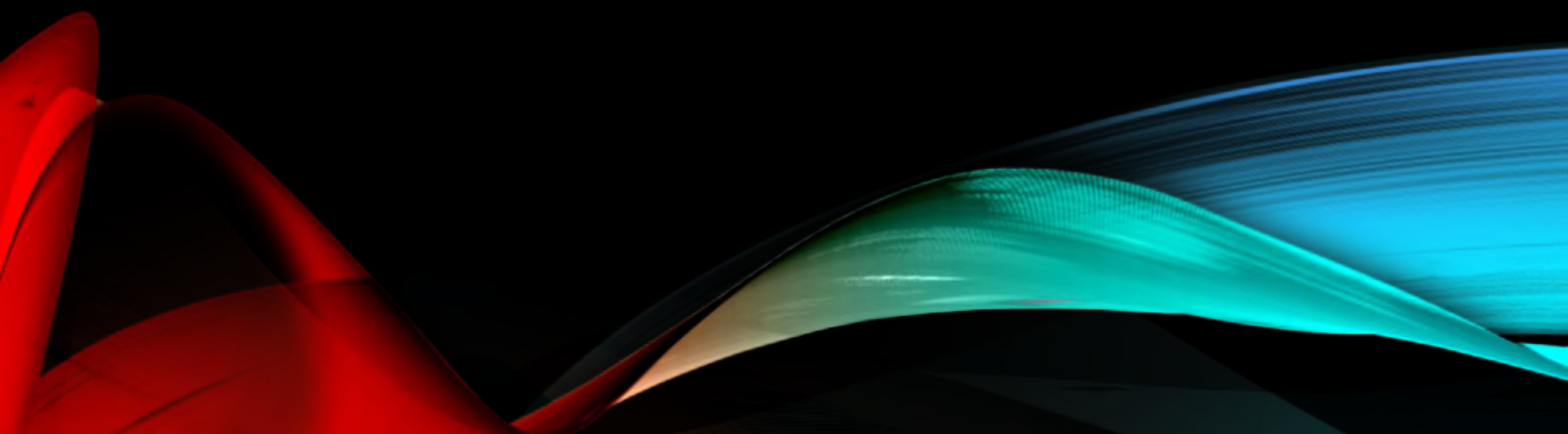


有计划的堆硬件

- N+M
- ~ 130% Peak
- 别再提灾备了

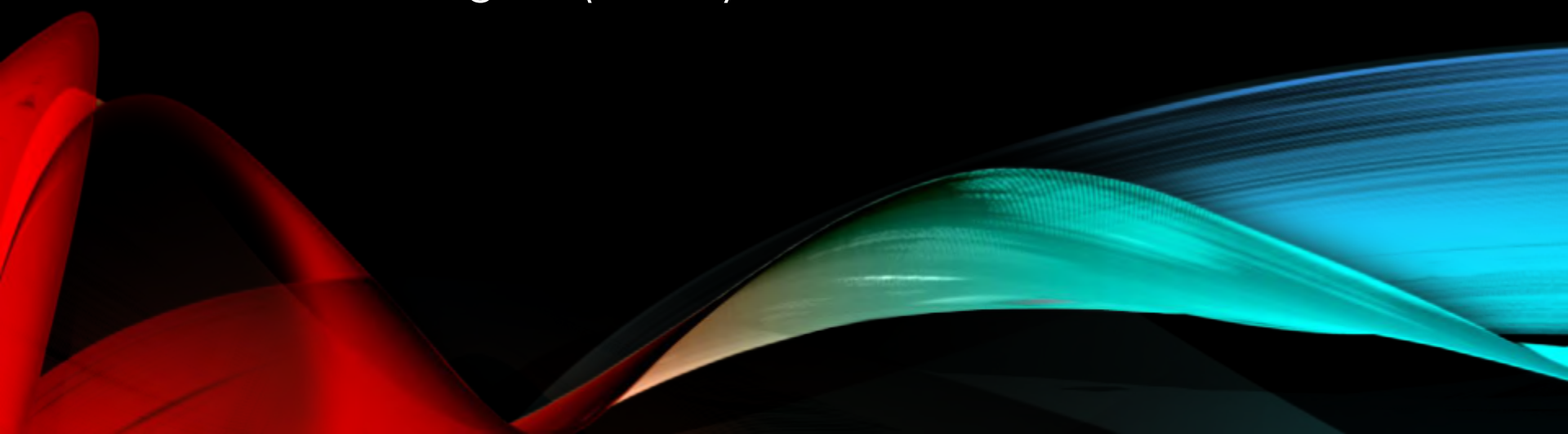
设计一套不会坏的系统

- 多级结构
- 可降级、有损运营、业务连续
- 无人运维



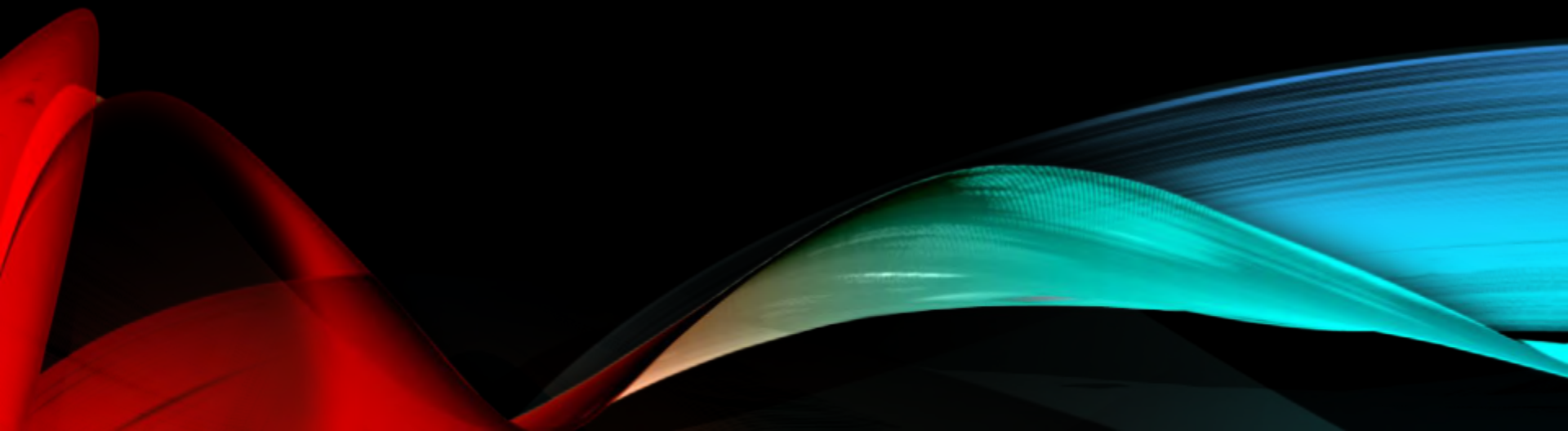
监控系统的三种输出

- 报警 - Alerts - O(H)
- 工单 - Tickets - O(D)
- 记录 – Log – O(Never)



ONCALL 的正确姿势

- 一个报警一个人处理，明确升级策略
- 解决线上问题为主，处女座强迫症为辅
- 做好总结：5why, Root cause, 自动化



做好实战演习

- Operation Readiness Drill 流程演习
- 场景演练：考验也是学习
- 运动式演练

多写事后总结

- Capture the facts
- Lessons Learned
- Action Plan

“A PROBLEM IS RESOLVED TO THE DEGREE THAT NO HUMAN BEING WILL EVER HAVE TO PAY ATTENTION TO IT AGAIN.”

http://www.codesimplicity.com/post/make-it-never-come-back/

灾难级别分类图

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

Q & A

扣钉 https://coding.net