GOPS 2023
shenzhen

GOPS

GOPS 全球运维大会 2023

- XOps 风向标

深圳站

时间：2023年4月7日-8日　　地址：中国·深圳

指导单位：云计算开源产业联盟　FinOps产业推进方阵

主办单位：开放运维联盟 OOPSA Open OPS Alliance　高效运维社区 GreatOPS Community　DevOps时代

承办单位：华佑科技 HUAYOU TECHNOLOGY

SRE体系：快速修复故障的套路

张观石 《SRE原理与实践》作者

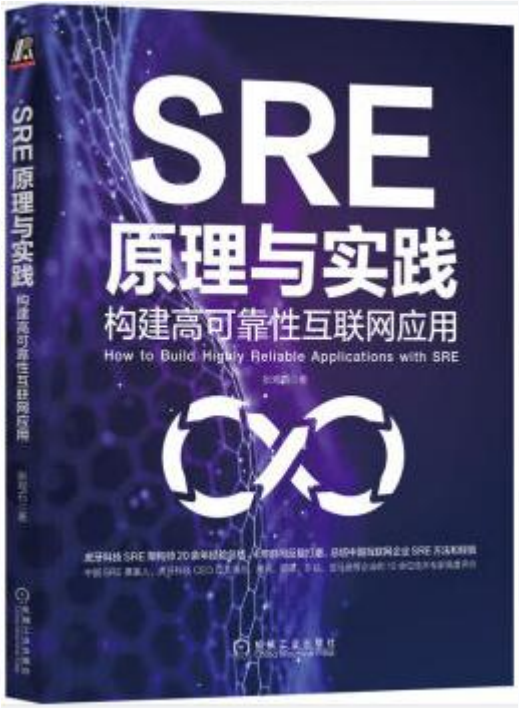
- ✓ 资深运维专家和架构师，拥有20年经验；
- ✓ 熟悉基于微服务架构的直播业务、音视频业务、海外直播业务的稳定的保障体系。熟悉混合多云架构、可观测性、预案、变更管控、AIOps等领域；
- ✓ 信通院分布式系统稳定性实验室高级技术专家，参与编写了信通院《信息系统稳定性保障能力建设指南》。

新书介绍

GOPS 2023 shenzhen

1. 读过的都说好
   - “可用于做SRE面试指南”
   - “用于指导实际项目开展”,读了3遍
   - 送朋友、送客户、送同事

2. 内容特点
   - SRE工程体系完整
   - 先进实战案例丰富



目录

CONTENTS

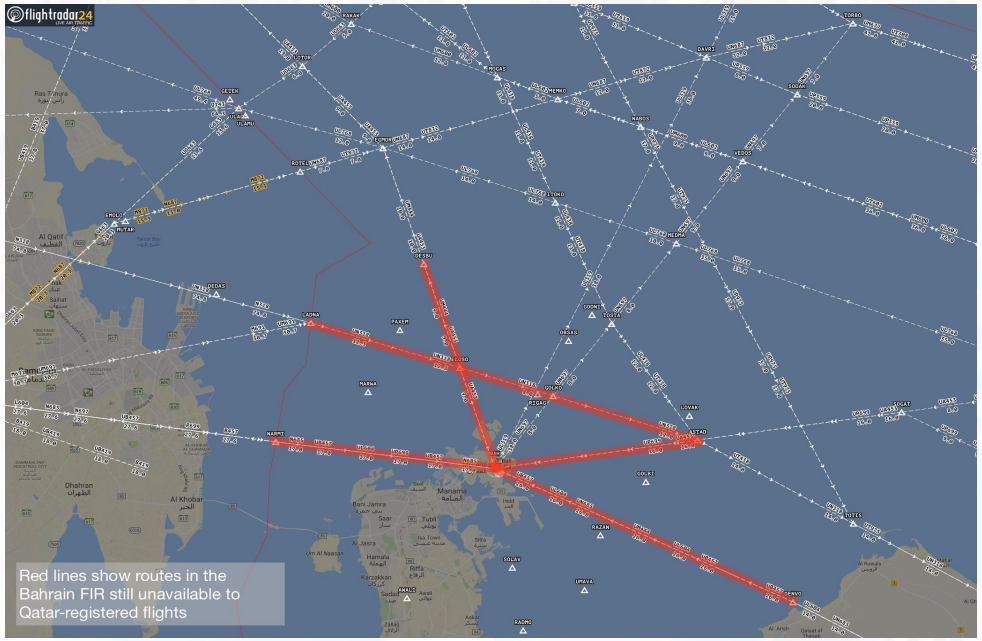
01

案例：3个惨案现场

故障案例1

- 背景：数据库M-S架构，正常主从是同步的。
- 故障描述：某天发现主从不同步了。
- 处理方法1：在修复同步问题时无意中删除了一个文件，DBA用了另外一个备份文件去替代。看起来是一样的文件，然后重启数据库。
- 结果：结果数据库系统启动不起来。

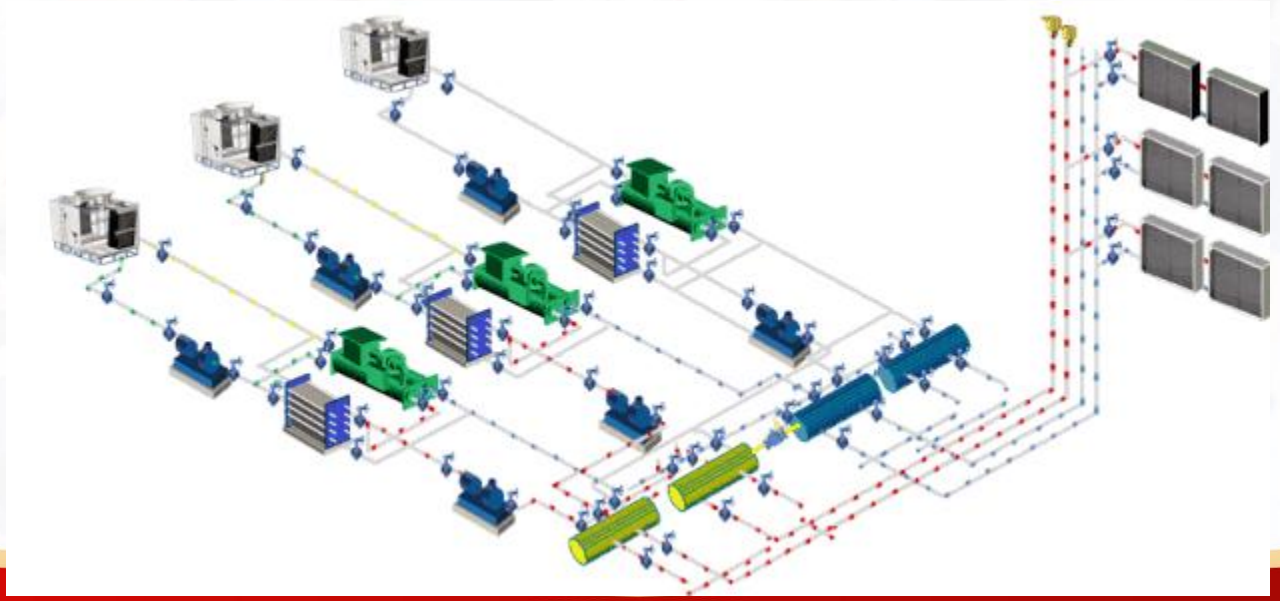
2023年1月12日 美国FAA NOTAM系统故障，全美12000个航班被延误或取消



故障案例2

- 背景：机房冷机4主+4备的架构，主机故障可以手工切备机。
- 故障描述：冷却系统缺水，导致4台主冷机服务异常。
- 处理预案1：冷机切到备机系统，发现缺水形成了气阻，备用冷机启动失败。
- 处理方法2：尝试一台台启动，阻力更小
- 结果：启动不起来，发现冷机设计为4台绑定一起重启，目的是为了批量操作方便。
- 紧急处理：只能远程与现场合作临时改代码逻辑、发布，解除群控逻辑。

某公有云AZ制冷故障，持续13小时



故障案例3

- 背景：业务产品和管控系统都在A、B。两机房容灾部署
- 故障：机房A挂了，大量迁移到机房B，用户集中迁移业务导致管控系统的并发增加，被限流；
- 预案：给管控系统扩容资源
- 问题：增加容量的管控系统的一个中间件被部署在故障机房A，扩容操作失败

某公有云AZ制冷故障，持续13小时

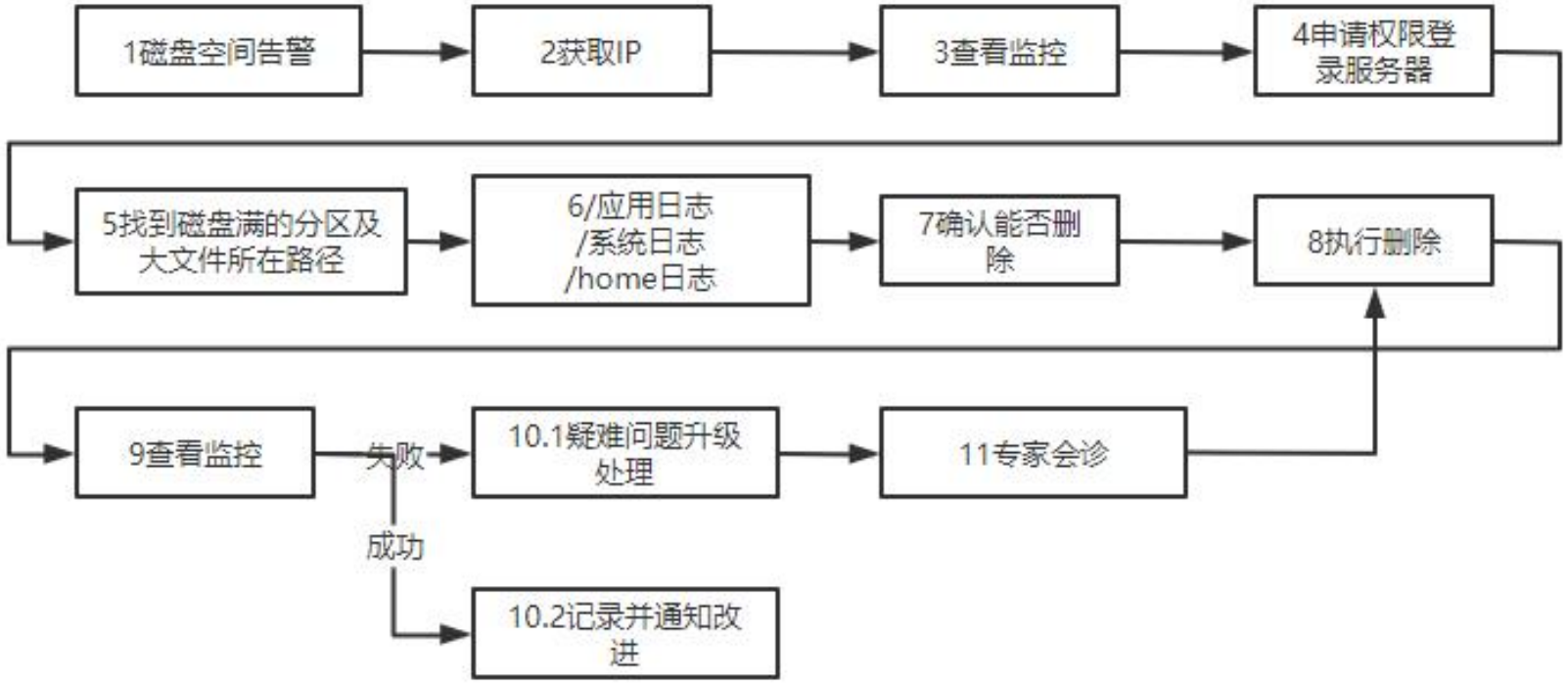
简单故障场景4

服务器磁盘被写满了，处理需要几步，需要多长时间

12月15日 周五 21:02

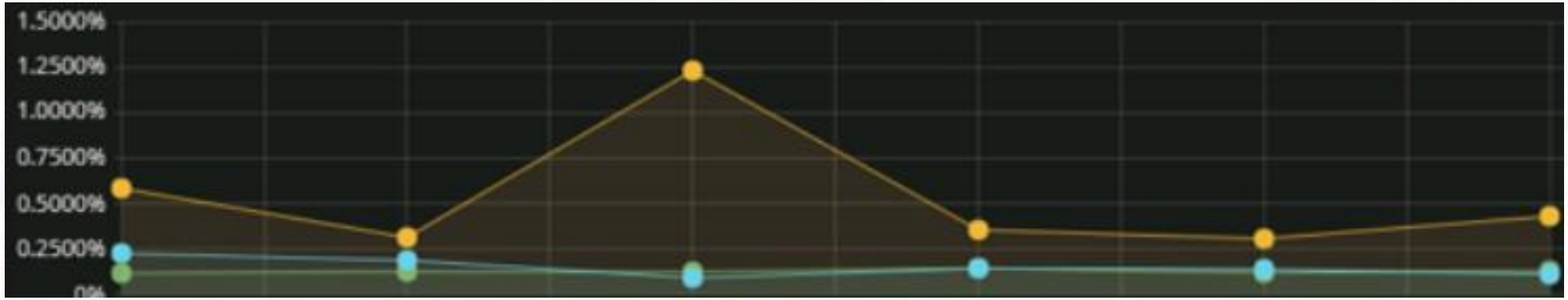
【广告门户】【告警项目】：广告门户项目 【告警主机】： 10.16.224.23 【告警详情】： 服务器10.16.224.23上 /data目录剩余空间少于20%

【广告门户】【告警项目】：广告门户项目 【告警主机】： 10.16.224.23 【告警详情】： 服务器10.16.224.23上 /data目录剩余空间少于20% 问题已恢复。



复杂故障场景5

- 直播平台大活动期间卡顿率上升1%



1. 怎么排查是哪部分、
2. 怎么定位是什么原因，什么维度
3. 怎么修复

故障修复的难点在哪?

GOPS 2023 shenzhen

系统复杂性

系统复杂、故障场景多、脆弱性因素多，防不胜防;
案例

涉及人员众多

涉及到众多人员、没有组织协同则混乱出错;
有时10几个团队人一起参与问题处理，指挥混乱、信息混乱
一个故障影响机房数百个产品和上千个系统

修复过程难

所用到的各方面能力，任何一环不能掉链子，以为有预案，关键时刻不工作。

发现难、定位难、修复难
案例:

02

快速修复故障的基本套路

>设计、预案、应急

系统可修复性设计要求

系统可被修

系统做了可被修复的设计

可感知、无状态、可切换/调度/容错/降级

有效的修复方案和工具

针对故障因素/场景设计修复方案

专门的修复工具，并打通依赖工具

有力保障能力：资源、人与流程

有接收故障，并执行处理的高效流程，预备资源，人的应急协同

可被修复的架构设计

- 设计便于修复的软硬件架构
  - 系统是可修复的（针对特定的故障场景已经有相应的修复设计）
  - 能自愈的尽量容灾自愈，不能自愈必须暴露接口

- 可修复的架构原则，架构风险治理
  - 标准化、无状态的软件架构
  - 多副本冗余的设计
  - 被隔离迁移、调度切换的能力

故障场景、 故障影响、预案是什么、故障预计修复时长
问研发：能不能把调度功能开放给运维？

各系统可被修复的架构设计&暴露API

GOPS 2023 shenzhen

微服务后台

节点屏蔽/删除
服务组扩缩容
变更系统回滚
业务降级

接入中台

接入服务切换
路由切换

音视频

直播间下行流屏蔽
切换线路
切换档位
主播切换上行线路

名字服务

节点屏蔽/删除
服务组扩容/缩容

运维通道

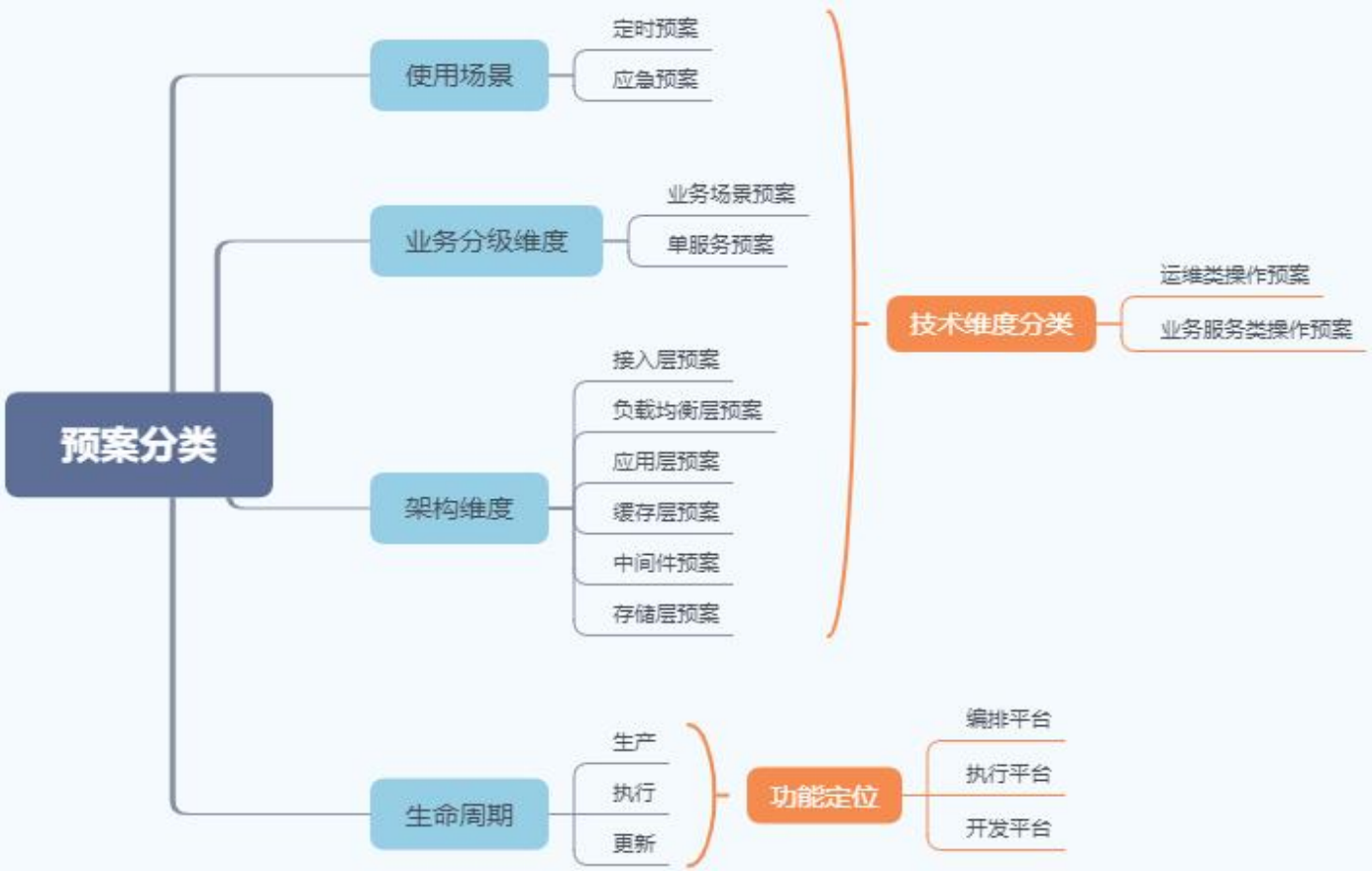
自定义脚本
任务脚本

通用功能

自定义api接入
基于事件消费的预案任务执行机制
监控内嵌查询
基于指标的判断功能
文档功能
IM@人功能
统一告警通知功能

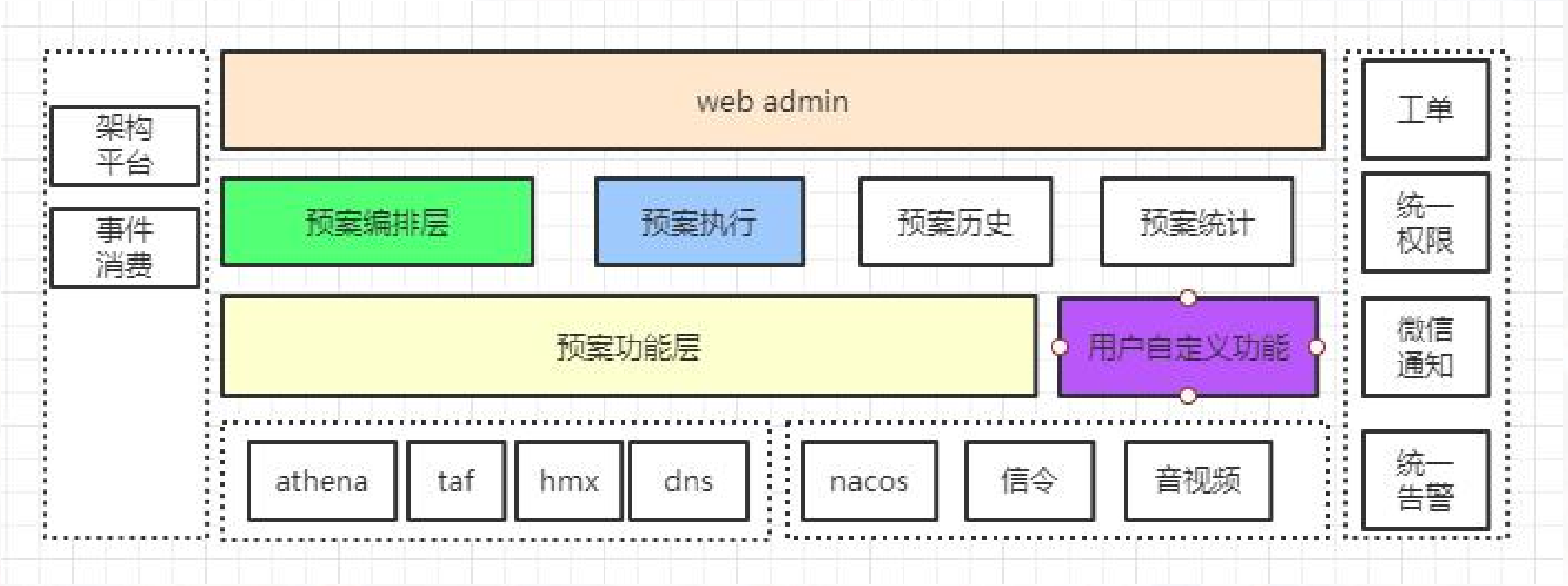
架构与预案结合

运维类操作、业务服务类操作



分层	系统服务	预案场景
接入层	信令层负载均衡	信令回源切换、信令 LB 切换、Set 切换、容量扩容、流控
	web 负载均衡	名字服务节点屏蔽、dns 解析屏蔽
	CDN 服务	切换源站、切厂商
应用层	微服务框架	摘除节点 队列方式的限流 对接微服务平台管控相关操作
	业务逻辑	服务级别、链路级别的降级、限流 服务级别快速扩容 链路级别快速扩容
	音视频	主播上行切换 观众下行切换 P2P 开关 主播网节点上下线 直播间流地址切换
缓存层	缓存中台服务	接入点摘除 主备切换
	Redis 集群	主从切换 故障节点切换
数据层	MySQL	接入点摘除 主备切换
运维管控层	运维通道	执行运维脚本
	容器平台	执行容器集群的几个重要管控指令
监控、告警	监控数据查询	封装监控数据查询，部分规则可进行判断输出
	监控图嵌入	嵌入监控图到预案文档，可以辅助决策
	打通统一告警服务	查询最近告警 短信通知 企业微信通知

预案及预案系统

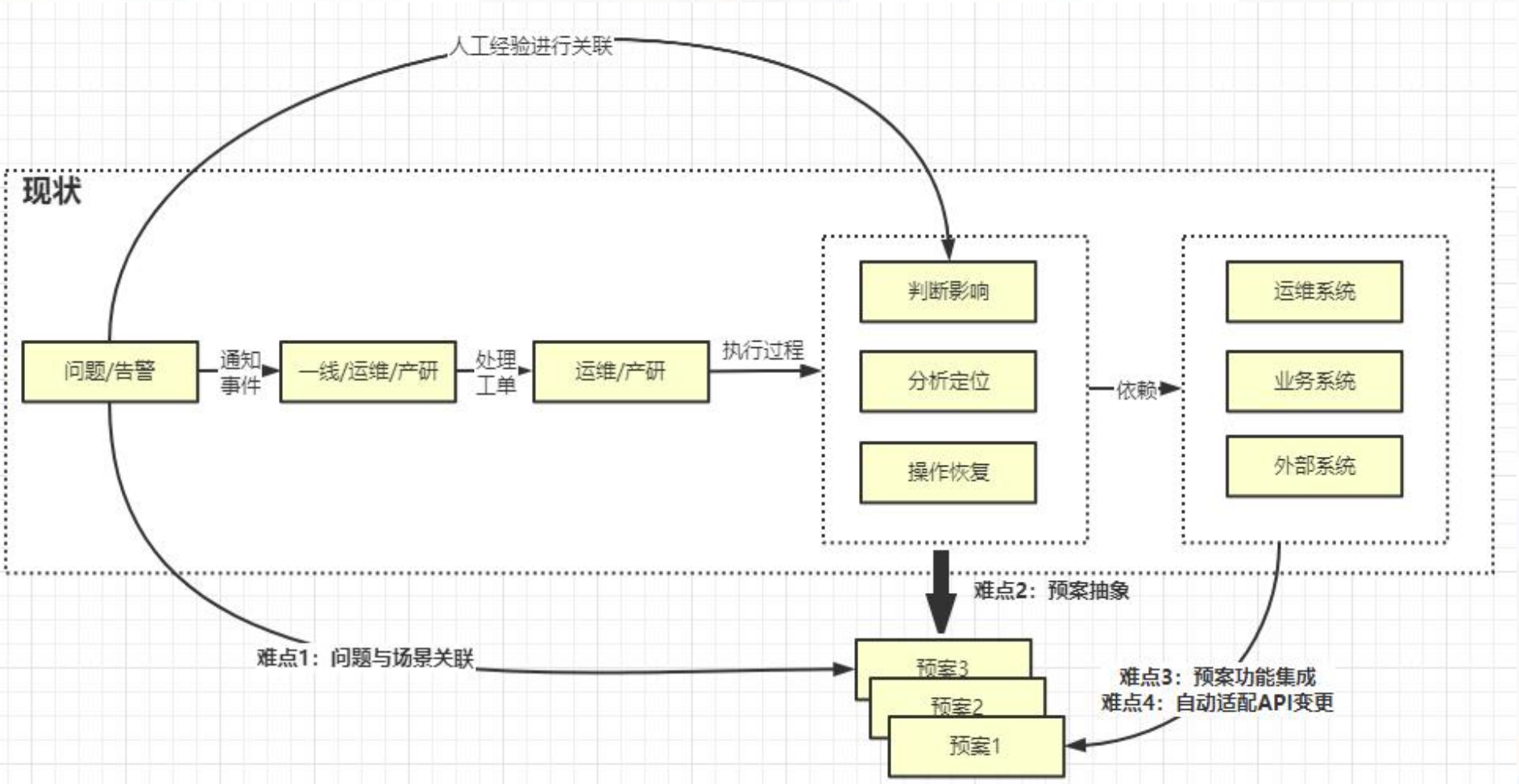


- 有修复的工具及其依赖的工具
- 有修复的人、及时协作，快速修复

高效执行，有力保障：预案不一定很复杂

1、问题本质原因：问题/故障解决依赖人的知识经验

2、核心要解决的：如何将处理经验通过技术手段固化成一个个可以被直接可执行的预案场景



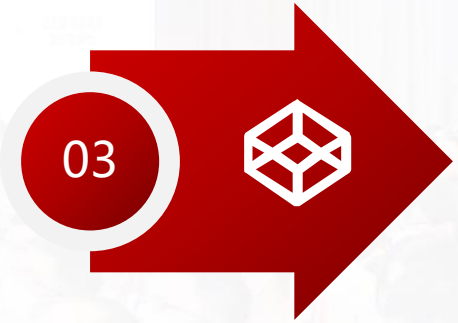
综合保障能力

人员保障

协作排查、修复、
指挥协同、发言

运维资源保障

紧急扩容资源
支撑工具



流程与制度保障

定期演练

一键到达：根因推荐与预案关联

告警等级: 灾难
告警来源: [redacted]心业务指标告警
AI告警名称: [redacted]黄金指标-订阅成功率【信令】ip去重 - V
时间点：2022-03-29 06:08:00
当前值（粒度：1.0分钟）：0.9898 - 下降异常
最近半小时曲线: http://alert-aiops.[redacted]s.com/1463-1648505420297.png
详情访问：http://bize[redacted]etails?id=1463&type=detail
相关根因：
platform=adr （当前值:0.9898，占总异常量的100.00%）
retcode=-1 （当前值:0.9898，占总异常量的100.00%）
_ip=[redacted].166 （当前值:0.9898，占总异常量的100.00%）
[redacted]黄金指标订阅信令成功率异常
预案：http://xy.huy[redacted]preplan/#/preplan/process?id=139

在告警信息中嵌入预案链接

数据库机器人 BOT

DMX故障切换
故障主机: [redacted]48.201
主机类型: MYSQL
实例数: 6
端口: 6879,7477,7478,7479,7480,7851
业务: 点播推[redacted],测试平台
故障发现时间: 202[redacted]3 07:59:40
二次检查完成时间: 202[redacted]-23 08:00:21(耗时41秒)
切换完成时间: 202[redacted]23 08:00:51(耗时30秒)
切换成功率: 100%

早监控早触发预案执行

预案来源

1. 企业内部/业界曾经发生的故障场景
2. 演练发现的故障场景
3. 通过技术分析、风险识别发现的潜在故障场景

预案功能设计：

1. 预案管理（增加录入、修改、删除、执行记录）
2. 基本任务（原子操作）管理：
    1. 可增加、删除、修改原子操作，
    2. 对接管控系统API、运维通道、软件集群，平台脚本编辑等
3. 预案编排：增加删除步骤、调整顺序，每一步对接基本任务或一些自行动作，参数传递
4. 预案执行：告警导入、页面引导、一键/逐步执行、每步结果显示、执行前通知、执行后通知，记录执行过程
5. 预案回退：部分支持灰度执行，也可回退，部分提供恢复现场功能
6. 预案统计分析：执行次数、时长、效果等
7. 其他功能：权限控制、执行历史、文档编辑、嵌入通知、嵌入监控、自动拉群等

03

套路有多深

> 深入故障规律，理解故障命脉

故障恢复的原则

研究规律、有效应对

按故障原因进行分类
针对原因设计对应预案

故障修复是工程

不仅靠运维
从架构设计、经验 沉淀
管控能力编程，决策执行

故障修复靠综合能力

不仅靠经验、靠预案
更需要系统协同
有力保障

应对之道

2 容量型：

- 提供更多的资源（扩容）、
- 把服务消耗的资源减少（优化、降级）

1 灾难型：容灾高可用

- 高可用设计
- 容灾切换，内部与外部切

3 变更型：

- 变更管控
- 人机可靠性

应对方法及案例：

灾难型：

部署架构高可用，混合云两地三中心
直播间上行和下行线路

变更型故障：

变更红线、变更管控系统

容量与负载故障型：

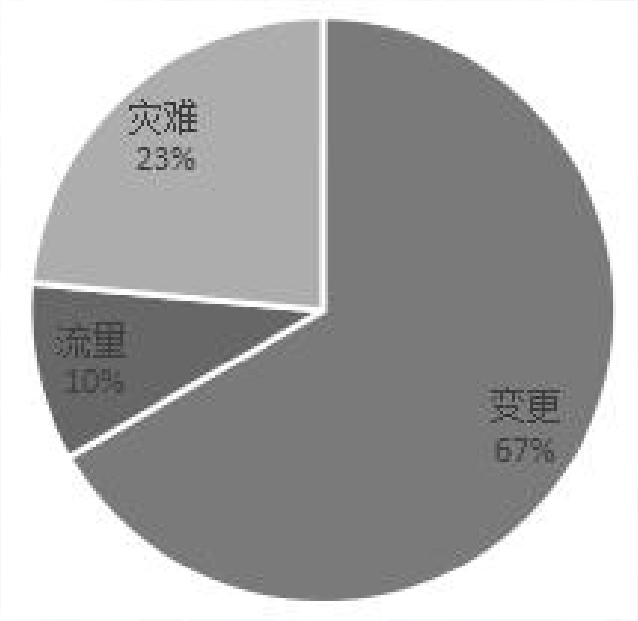
1、扩容
2、降级、熔断

应对案例：

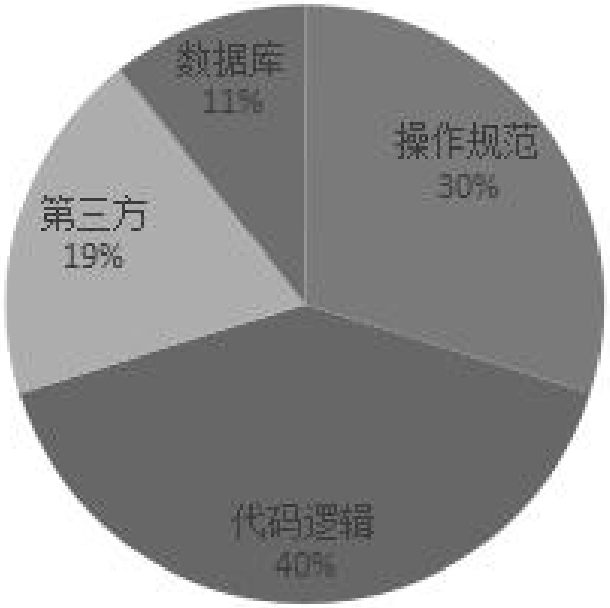
混合云弹性，预先弹性、一键扩容；
一键降级；

故障规律

- 故障分类及原因分类
  - 灾难型、容量负载型、变更型
    - 灾难型：服务器、机房、交换机、网络等单点不可用
    - 负载型：流量超出预期、性能下降造成资源不足
    - 变更型：应用发版、运维变更、软件基础设施变更、应急基础设施变更、配置变更

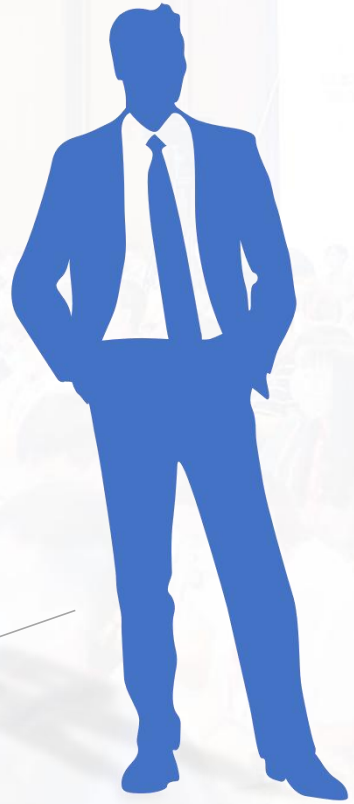


某年故障原因分类



变更型故障的细分原因

故障修复是工程



产研及架构师

- 改变软件系统架构
- 服务可配置开关能力
- 暴露可修复的功能

基础架构/中间件

- 实现基础设施、系统运维、组件的架构
- 提供可修复的功能

SRE

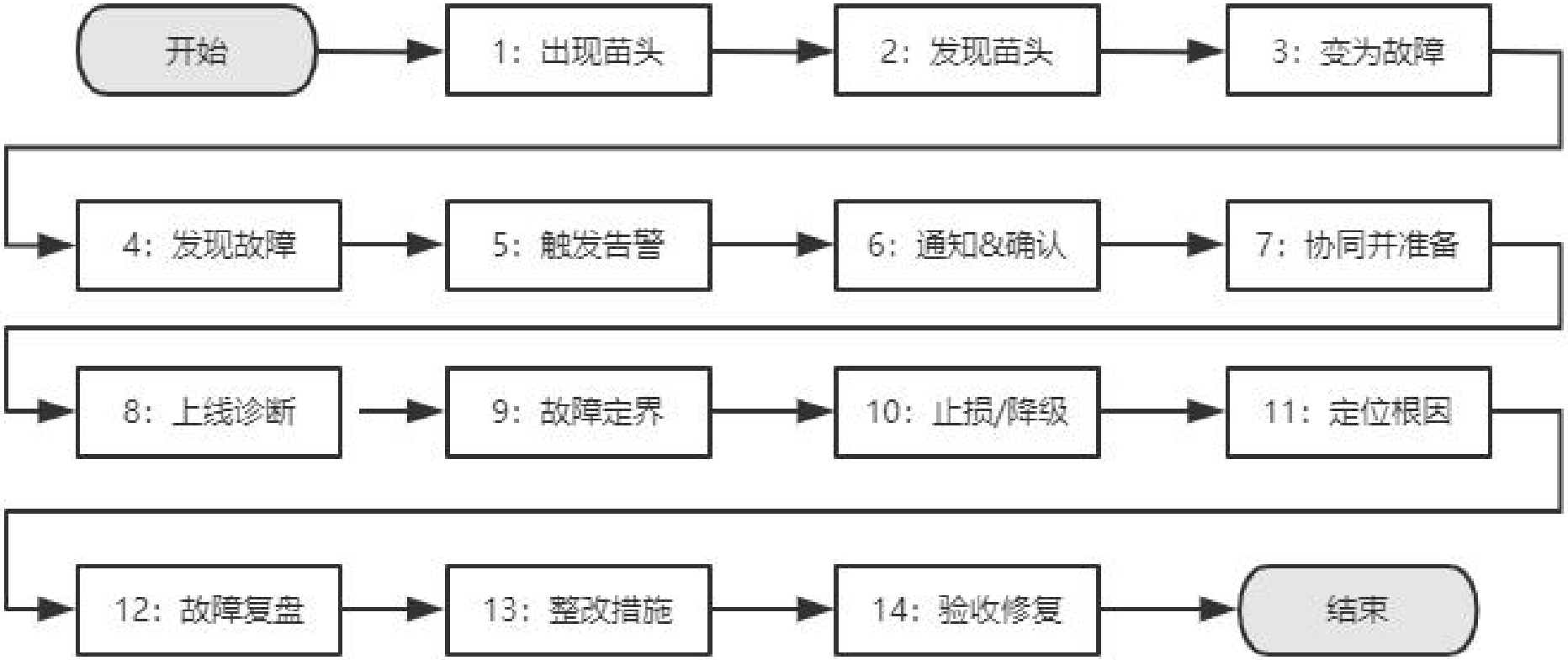
- 改变架构
- 编排能力开发预案

一线/NOC

预案最频繁的使用者

故障快速修复不是单个部门的事情，是研发、SRE、架构部门**共同目标**
预案平台是把系统各层技术能力加以集成，共同修复

故障生命周期:从苗头到修复的全过程



故障通知

通知谁,
通知渠道
通知内容

故障响应

确认响应
准备工作

定界定位

影响分析
定界分析
初因分析

快恢/止损

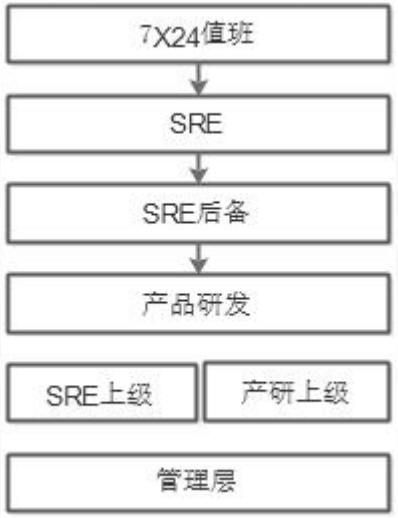
预案

应急修复

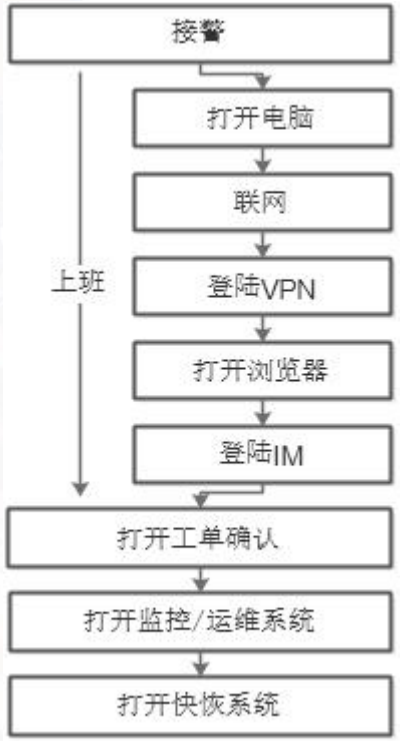
分工
升级流程
对内对外

故障复盘

应急修复故障的要点



通知环节最重要的是尽快通知能处理故障的人，提供简要关键信息，通知方式要便于转通知其他后续参与的人。



响应阶段考验人员规划、日常训练、人员责任心、组织安排，以及办公基础设施的完备性、响应相关系统的易用和便捷程度，准备步骤是否顺畅等。

应急修复故障原因排查顺序

原因类别	详细原因
是否有变更	应用程序变更：升级
	操作系统、系统服务、网络等基础变更有变更
	配置变更：应用配置，系统配置
	数据库/缓存/依赖等变更
	相关环境变更
是否有灾难事件	运营商、机房、机器、网络、机器内灾难事件 第三方、中台等被依赖服务灾难事件
业务量是否超过承载能力	是否突发流量：未预知的大活动
	异常流量：攻击、恶意访问等
	是否异常请求：用户行为可能发生变化
	工作负载是否有变化
趋势变化	耗时是否有普遍趋势变化 资源消耗（cpu/内存/IO）等是否有趋势变化 队列阻塞/锁是否有趋势变化

- 大胆假设
- 小心求证
- 迅速排除

04

怎么看套路有没有成效

>度量结果

度量成效：故障修复能力的度量

GOPS 2023 shenzhen

故障MTTR

- 单个故障的度量：
  - 修复过程时长
  - 故障分级分类
  - 修复能力级别
- 周期性度量：
  - 故障平均时长
  - 逐步提升、分析变化

过程能力

- 发现时长
- 响应时长
- 定界定位时长
- 修复时长
- 预案覆盖率
- 预案有效率

修复能力分级

- 20% 自愈越多越好
- 30% 预案平台一键修复
- 30% 多个步骤修
- 10% 按文档排查修复
- 10% 在线层层排查

总结

- 强调故障修复的工程化设计，故障修复也是个工程工作
  - 核心点：预案平台不是单个部门的事情，是研发、架构部门共同的目标。运维研发必须共同建设。
  - 支撑保障能力、管控系统的能力不能被忽视
- 研究故障规律，针对性设计故障修复预案
  - 灾难型、容量型、变更型
- 要持续度量，看到进步，更重要是看到短板和改进方向

以快速修复为目标，整合系统相关的技术栈各层能力，整合从运维、产研、值班、客服等团队协同，尽快速度修复故障。

GOPS 2023
shenzhen

Thanks

开放运维联盟

高效运维社区 **荣誉出品**

DevOps 时代