小米故障应急响应经验分享

2024年6月

https://sre-elite.com

个人介绍

王哲

小米云平台运维部

工作经历：

09-12,新浪网,平台架构部
12-21,奇虎360，云平台部
21-now，小米，云平台部

个人简介：

- 微博的第一版代码我上的线
- 从零构建360内部私有云平台Hulk
- 运维开发-产品经理-技术管理

目录

CONTENT

SREelite

01

小米故障应急面临的挑战

SREelite

小米的核心战略


小米集团战略
「人车家全生态」
Human X Car X Home
All Your Needs in one smart ecosystem

小米的核心战略


人、设备、智能服务之间
相互协作，共同进化
让智能跟随人的需求而流动
同时，人把智能带给生态，生态不断学习和进化，反过来服务人
小米澎湃OS
6.99亿
可连接智能设备*
200+
品类
95%+
覆盖用户生活场景
了解详情

小米面临的挑战

全球第三的手机制造厂商&全球领先的消费级AIoT平台

米家760万MAU

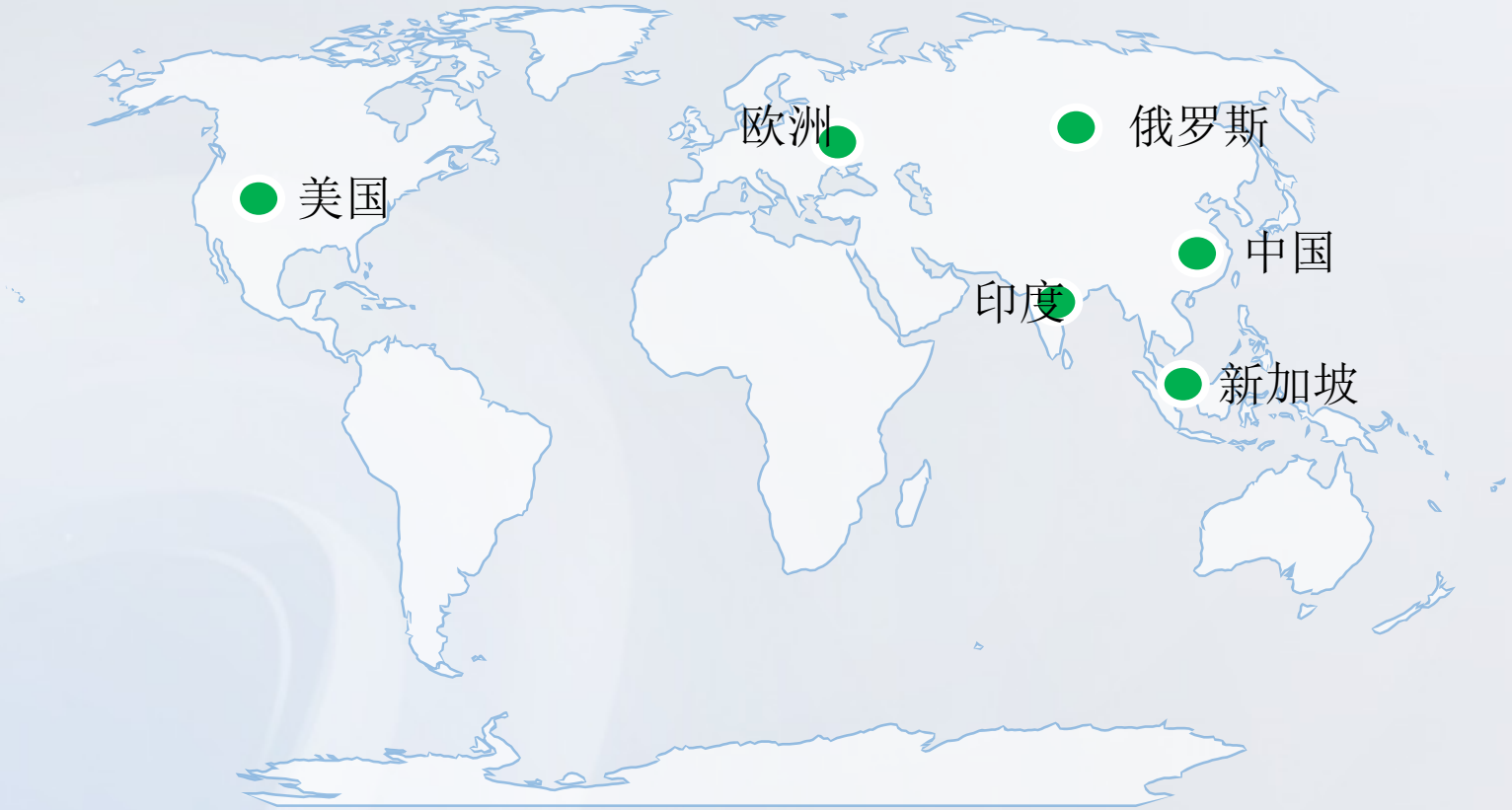
7亿联网设备

小爱1.14亿MAU

20+ 业务线	12个 核心机房	X万 服务器	100亿 日请求量	XPB 存储

小米面临的挑战

业务全球布局，数据本地存储
自有IDC，阿里云，金山云，腾讯云，火山云，AWS，Azure，GCP



02

构建哪些能力应对故障响应

SREelite

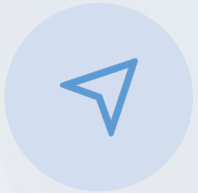
做好故障应急的基本能力拆解

故障发现

故障诊断

故障恢复

一分钟内发现问题



五分钟内定位问题

十分钟内控制损失

故障复盘

02-1

故障发现

SREelite

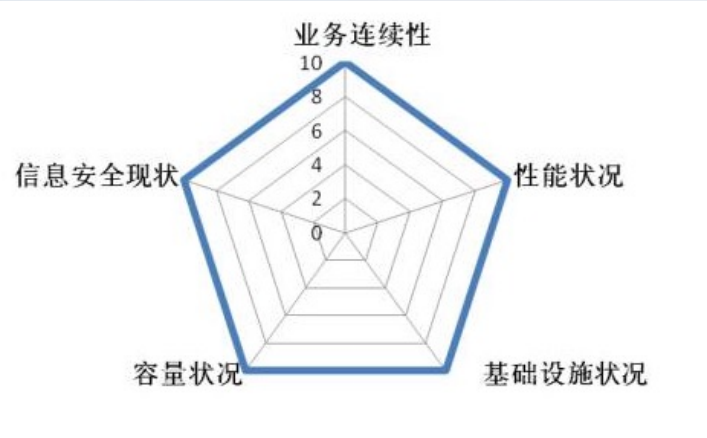
故障发现

SREelite

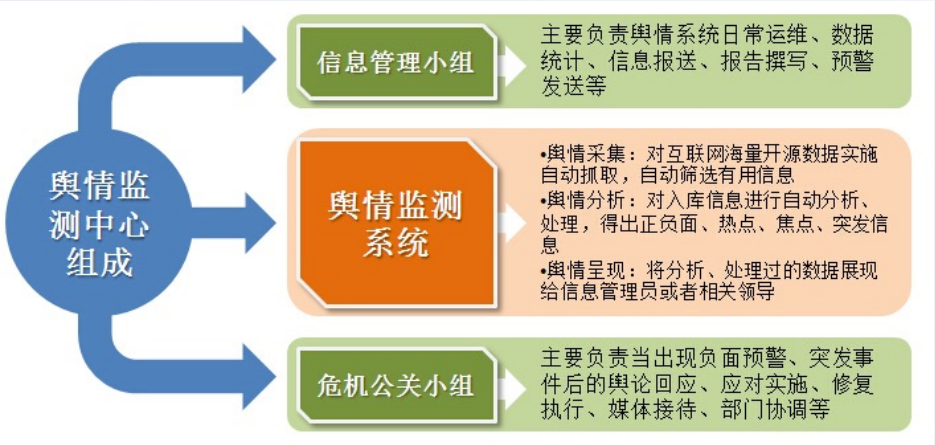
发现故障的几种手段

监控

巡检



上报



故障发现-监控

SREelite

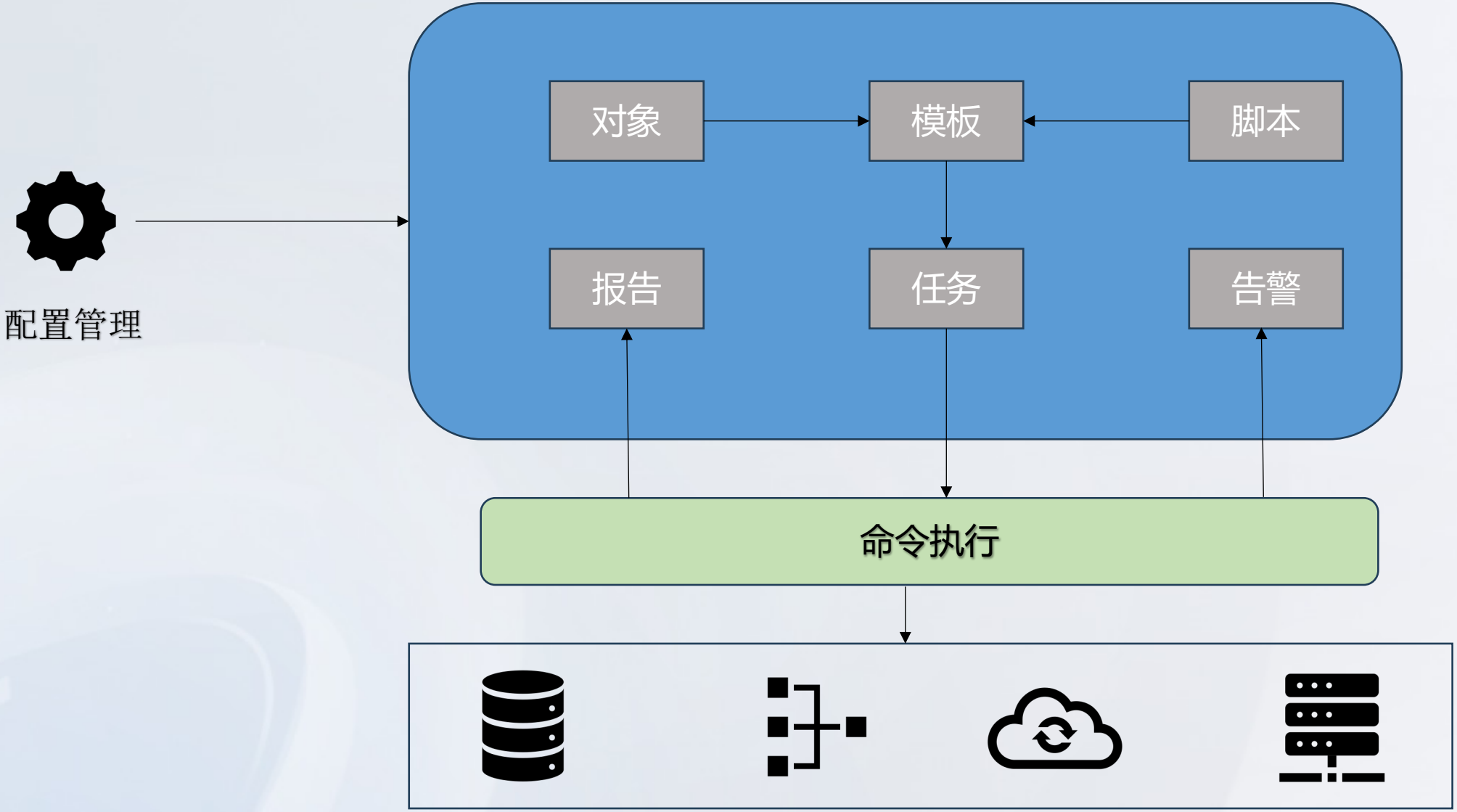
基础设施监控	•混合云环境下的IT基础架构信息实时监控（性能/可用性）
应用程序性能监控	•应用程序的功能/运行状况监控 •APM，跨微服务，主机，容器和Serverless
日志管理	•为应用，系统和云服务提供日志管理，可创建索引，查询可视化，报警 •支持机器学习，实现预测和阈值自定义等能力
用户体验监控	•模拟客户监控产品的体验和可用性 •真实用户监控
网络性能监控	•云或者混合环境的网络流量进行分析和可视化处理

协同

易用

整合

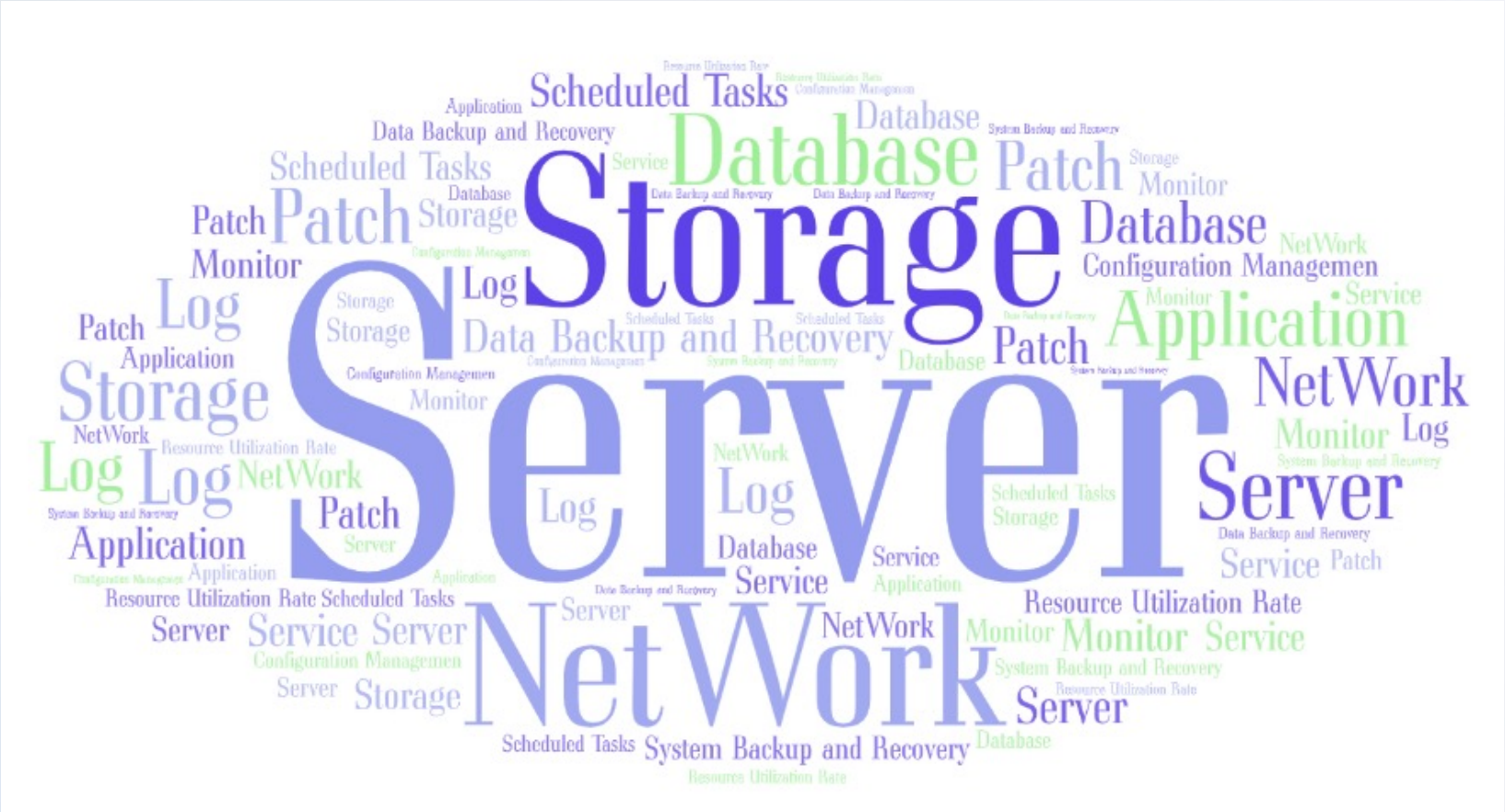
故障发现-巡检



故障发现-巡检

SREelite

巡检指标



故障发现-上报

关键词：自动化的监控不能绝对解决的问题，依赖终端用户的体验上报是很好的兜底手段

舆情监听
- 媒体信息收集
- 论坛讨论热点

客服反馈
- 用户投诉上报
- 客服统计数据
- VIP重点服务关注

运营反馈
- 新功能上线检查
- 日常操作反馈

02-2

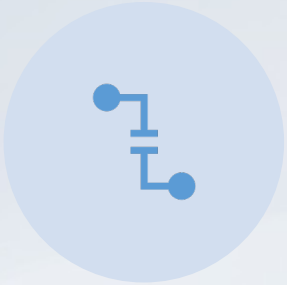
故障诊断

SREelite

https://sre-elite.com

故障诊断

关键词：故障恢复是第一位！



故障范围：

快速确认故障影响模块，逻辑功能依赖和上下游服务依赖

根因确认：

找到关键告警和分析入口，定位系统和责任人

影响评估：

基于故障范围，确认影响情况，决策信息同步周期和上报级别

组织能力：

值守，应急响应，升级流程，资源保障，信息同步...

故障诊断 – 故障范围界定

关键词：SRE向上要理解业务功能模块，向下要精通云基础技术组件

故障诊断 – 根因确认

关键词：AIOPS最卷的就在这里

归纳分析（点：聚类）

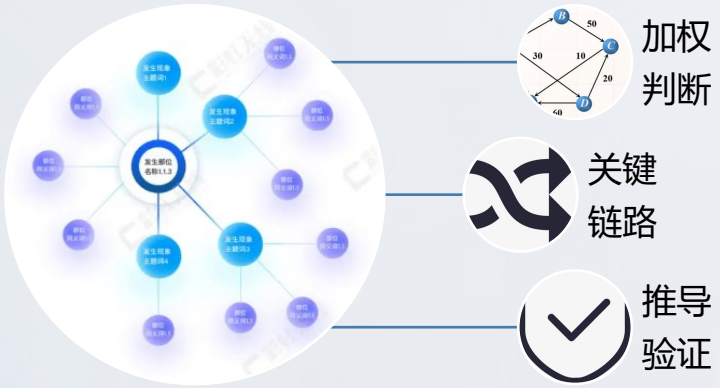
告警分析

日志分析

发布变更

配置变更

演绎推理（线和面：关联）



知识图谱

故障诊断 – 影响评估

SREelite

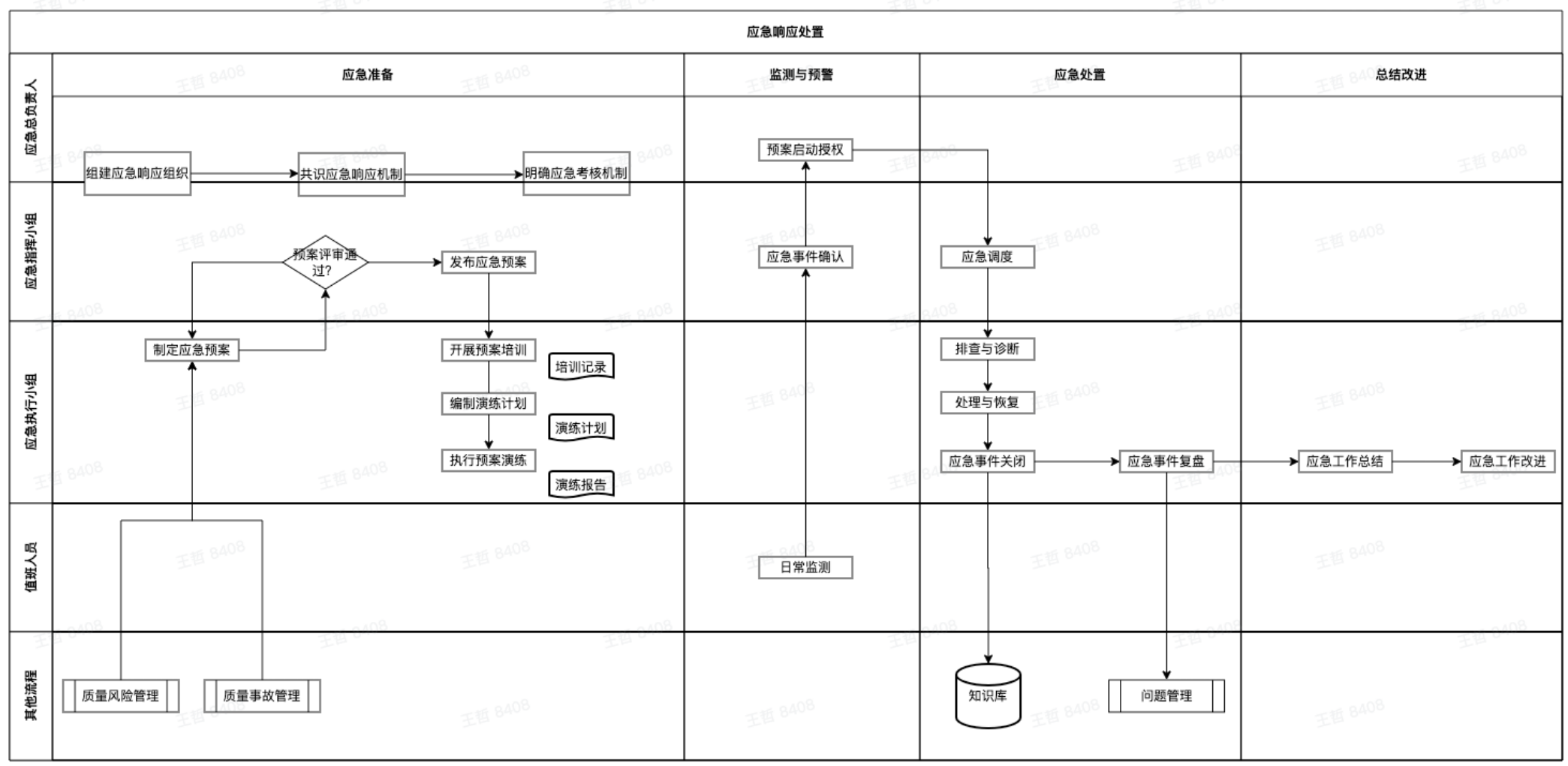
关键词：及时升级！

用户请求损失	用户无法使用功能造成的损失
影响用户数量	实际影响的用户数量
影响用户时长	精确的故障时长
客服来访数量	用户通过客服的报障数
用户反馈数量	用户通过反馈渠道的投诉反馈
直接经济损失	因为故障造成的经济损失
政府监管	抽检，召回，立案
市场舆情	维权事件，媒体报道，社会话题
售后维修	维修工单/销量
集团经营影响	产能，GMV金额，核心业务时长

故障诊断 – 应急响应组织能力

SREelite

不同的层次都由不同的团队来负责运维管理
同层次不同的硬件/系统/应用都由不同的小组来负责运维管理



02-3

故障恢复

SREelite

故障恢复三板斧

SREelite

关键词：每一个动作前都应该知道预期的结果，每一步真正的操作后都需要充分检查

1

重启：

单个或多个机器上的服务出现响应问题，先重启就能先恢复，能恢复就能止损

2

回滚：

确认跟发布相关，回滚是最好的选择，最起码减少变量，更快定位

3

扩容：

确认硬件跑满，压力增大，立即扩容也是最佳选择之一

故障恢复三板斧升级版

SREelite

关键词：每一个动作前都应该知道预期的结果，每一步真正的操作后都需要充分检查

1

限流：

控制入口流量，减少超出后端承载量的请求，自动丢弃新来的请求

2

降级：

有意识的降低系统的部分功能和服务质量，确保系统的核心功能和关键服务继续运行

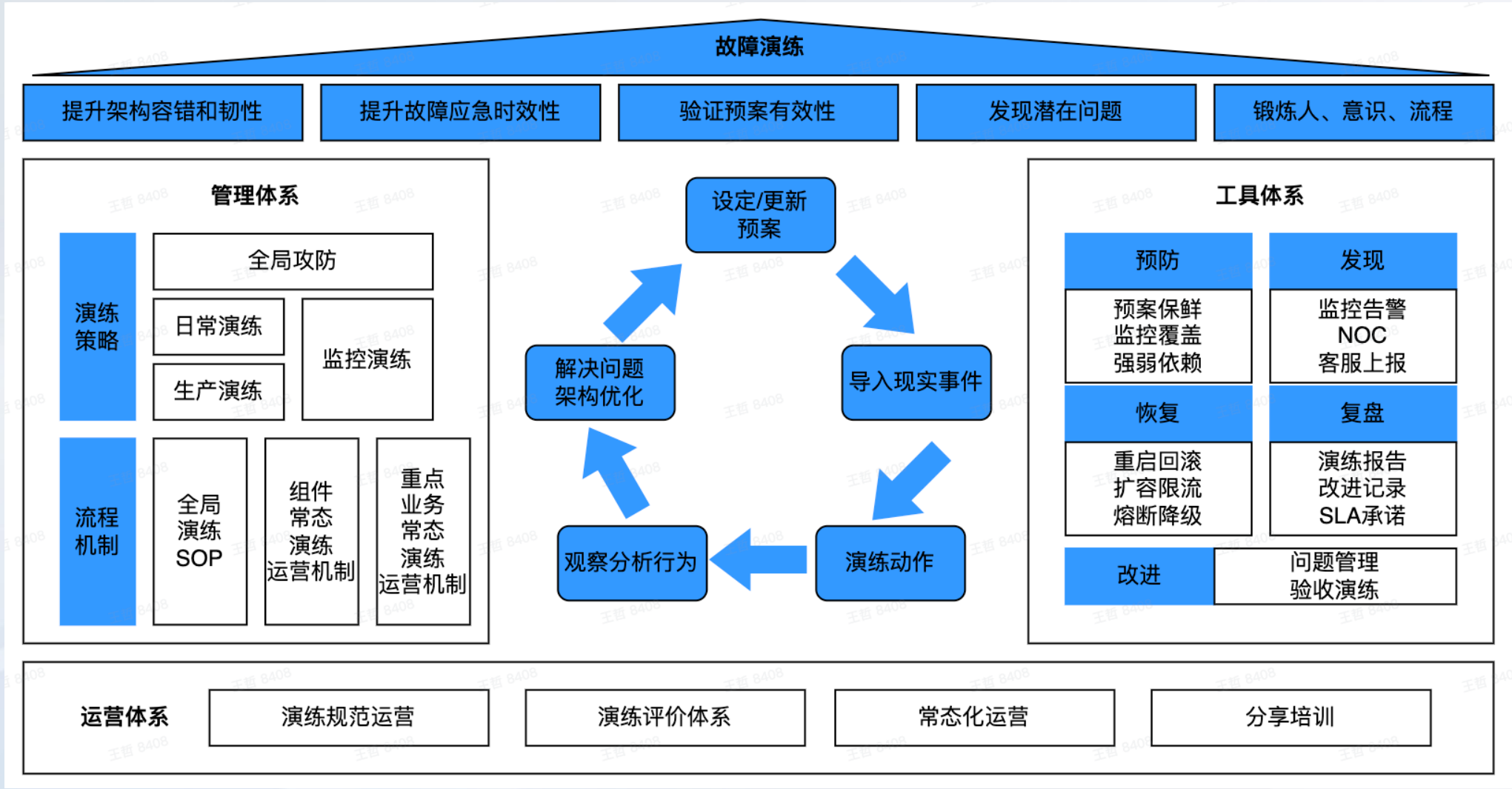
3

熔断：

依赖第三方服务异常，导致自身响应时长超长情况下快速返回失败而不是等待，阻止级联效应

故障恢复要以练养战

关键词：预案和爱情一样都有保鲜期



02-4

故障复盘

SREelite

https://sre-elite.com

故障复盘-思路

关键词：避免类似的问题重复发生

还原事实

- 过程中做了哪些动作
- 依据的流程机制标准是啥

问题反思

- 存在什么问题
- 做对了什么做错了什么

原因分析

- 主观原因
- 客观原因

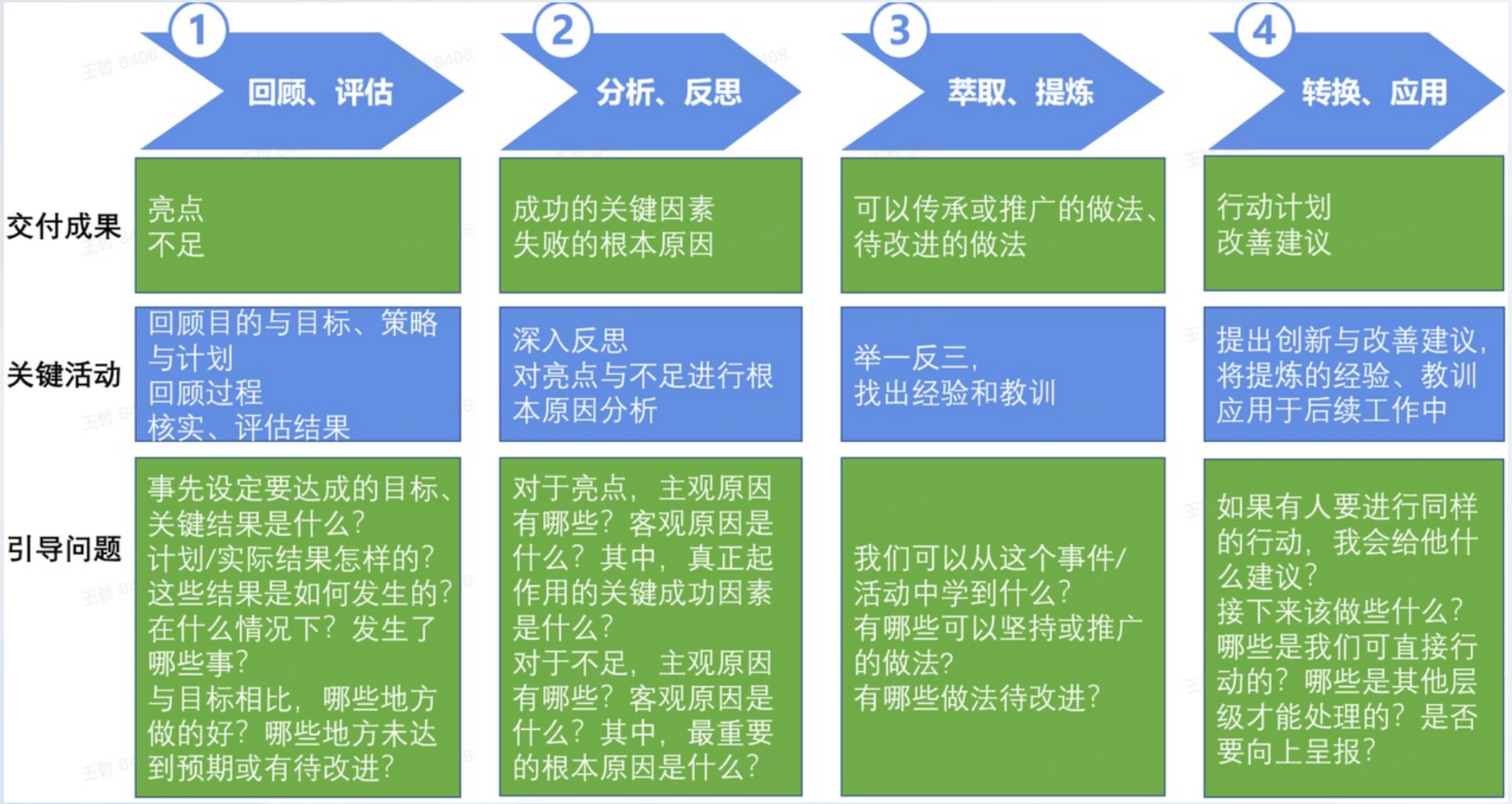
重来一遍

- 重来一遍怎么做
- 明确重点改进计划

故障复盘过程

SREelite

关键词：避免类似的问题重复发生

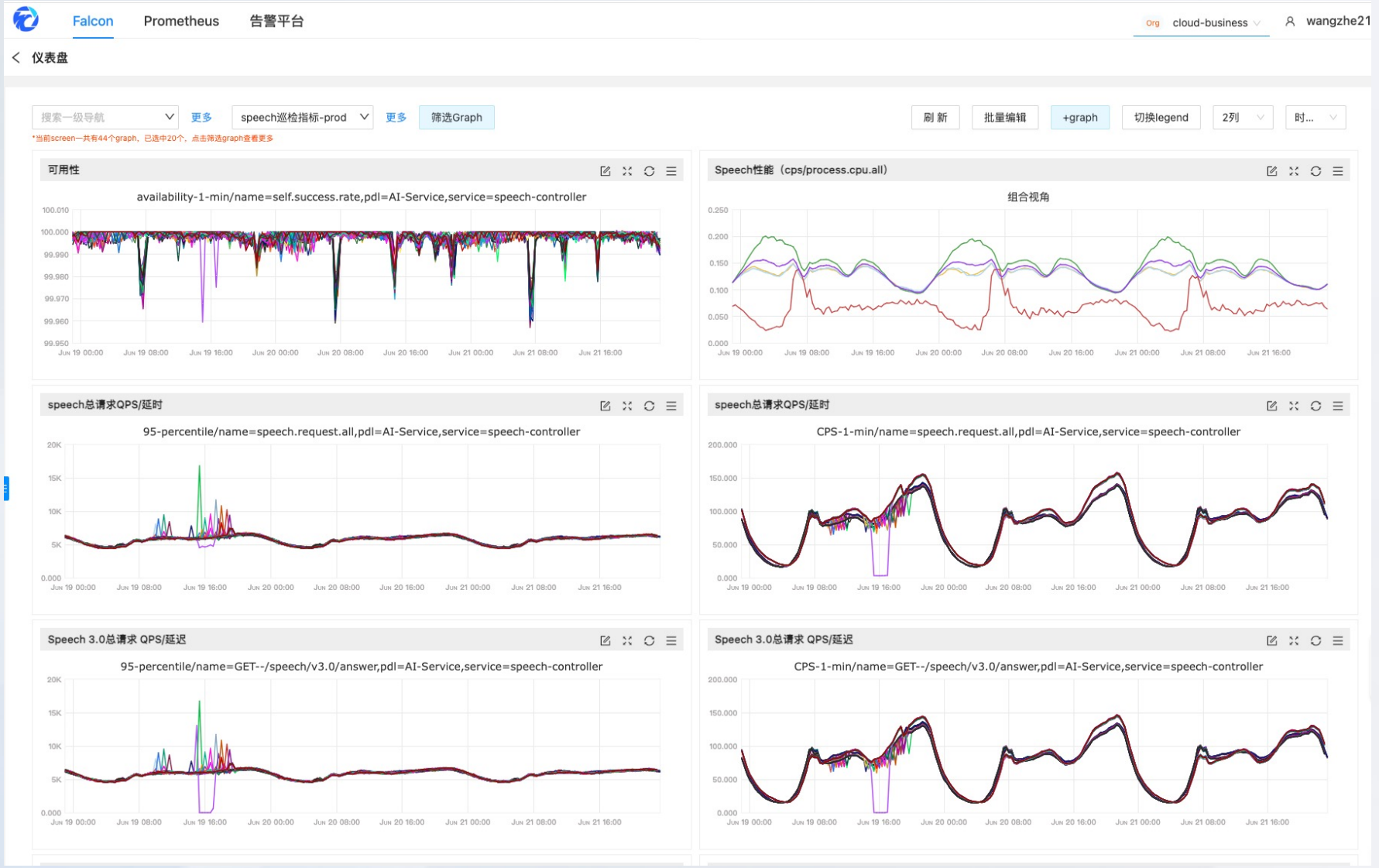


03

小米的一些实践经验

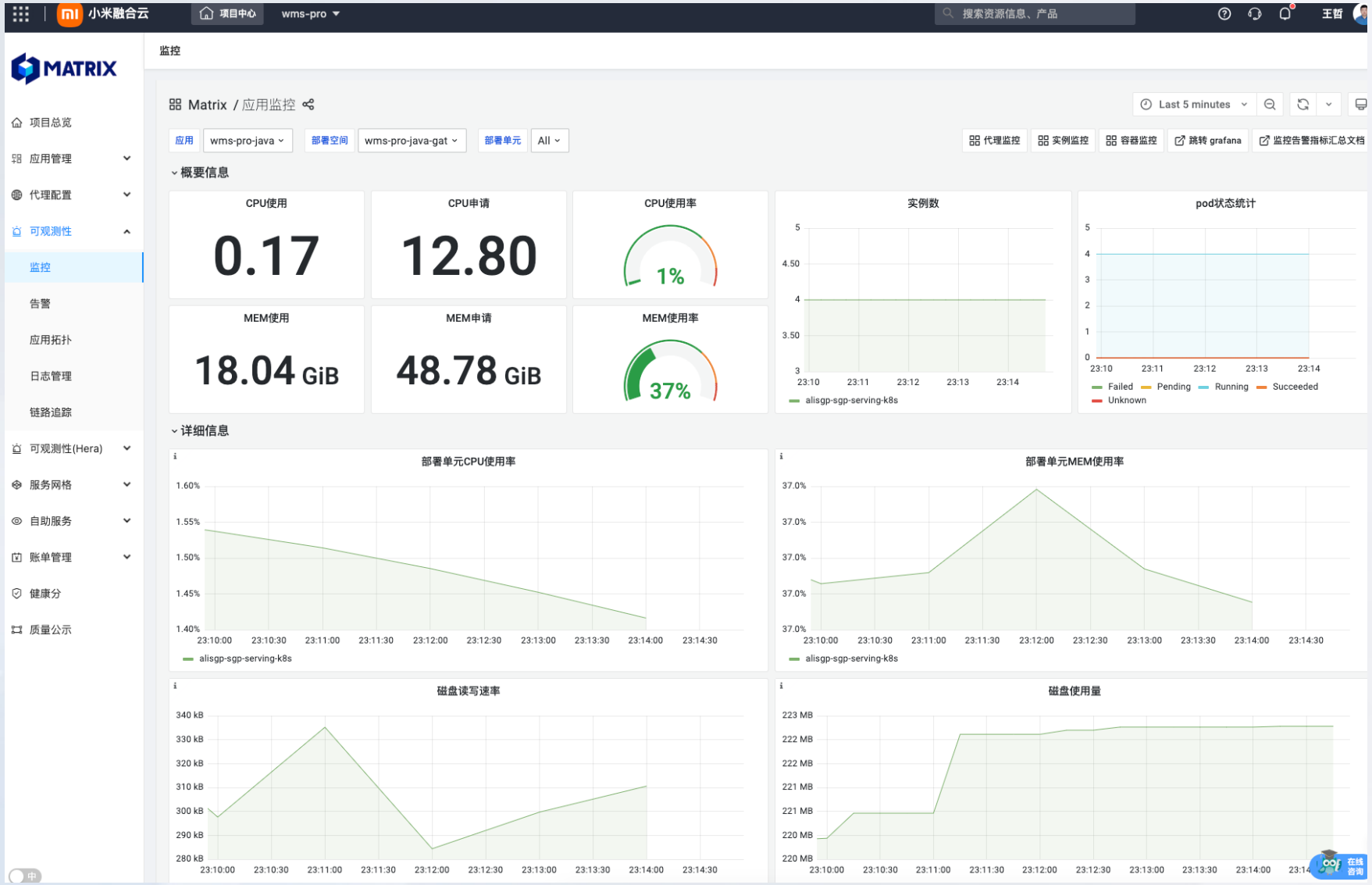
SREelite

监控


Falcon
Prometheus
告警平台
Org cloud-business
wangzhe21
仪表盘
搜索一级导航
更多
speech巡检指标-prod
更多
筛选Graph
刷 新
批量编辑
+graph
切换legend
2列
时...
*当前screen一共有44个graph，已选中20个，点击筛选graph查看更多
可用性
availability-1-min/name=self.success.rate,pdl=AI-Service,service=speech-controller
Speech性能（cps/process.cpu.all）
组合视角
speech总请求QPS/延时
95-percentile/name=speech.request.all,pdl=AI-Service,service=speech-controller
speech总请求QPS/延时
CPS-1-min/name=speech.request.all,pdl=AI-Service,service=speech-controller
Speech 3.0总请求 QPS/延迟
95-percentile/name=GET--/speech/v3.0/answer,pdl=AI-Service,service=speech-controller
Speech 3.0总请求 QPS/延迟
CPS-1-min/name=GET--/speech/v3.0/answer,pdl=AI-Service,service=speech-controller

监控

SREelite


小米融合云
项目中心
wms-pro
搜索资源信息、产品
王哲
MATRIX
监控
项目总览
应用管理
代理配置
可观测性
监控
告警
应用拓扑
日志管理
链路追踪
可观测性(Hera)
服务网格
自助服务
账单管理
健康分
质量公示
Matrix / 应用监控
Last 5 minutes
应用
wms-pro-java
部署空间
wms-pro-java-gat
部署单元
All
代理监控
实例监控
容器监控
跳转 grafana
监控告警指标汇总文档
概要信息
CPU使用
0.17
CPU申请
12.80
CPU使用率
1%
实例数
alisgp-sgp-serving-k8s
pod状态统计
Failed
Pending
Running
Succeeded
Unknown
MEM使用
18.04 GiB
MEM申请
48.78 GiB
MEM使用率
37%
详细信息
部署单元CPU使用率
alisgp-sgp-serving-k8s
部署单元MEM使用率
alisgp-sgp-serving-k8s
磁盘读写速率
磁盘使用量

告警

SREelite

【PROBLEM】【P1】可用性 mibi.api.xiaomi.com 给商户提供的服务端接口服务

域名: mibi.api.xiaomi.com 米币.米币手机端APP接口 给商户提供的服务端接口服务
集群: cnbj4-talos
报警条件: 可用性 all(#3) 93.617 < 99.95
报警时间: 2022-09-22T23:50:00
报错数量: count: 6, rate: 6.383
http状态码: ('500', 6)
nginx分布(top 10):
('c3-miui-mibi-api04.bj', 4)
('c3-miui-mibi-api03.bj', 2)
路径分布(top 5):
('/j_security_check', 6)
客户端IP(top 5):
('121.62.63.178#湖北', 6)
后端服务(top 10):
('c3-miui-mibi-fe03.bj:9025#mibi-web', 5)
('c3-miui-mibi-fe04.bj:9025#mibi-web', 1)
看图 关联分析 编辑路径

处理人: @叶金鑫

认领 屏蔽2小时 屏蔽1天 封禁IP

巡检

SREelite

← 容量巡检

CPU机器 | MiCE容器 | Matrix容器 | GPU机器 | 主机信息(达标/过保/故障)

分组: 111 机器: 1178

cpu.busy | 近一周 | 分位选择 | 阈值: > 70 % | 均值: 29.8% | 机房选择

选择机房，最少选择1个机房

全选

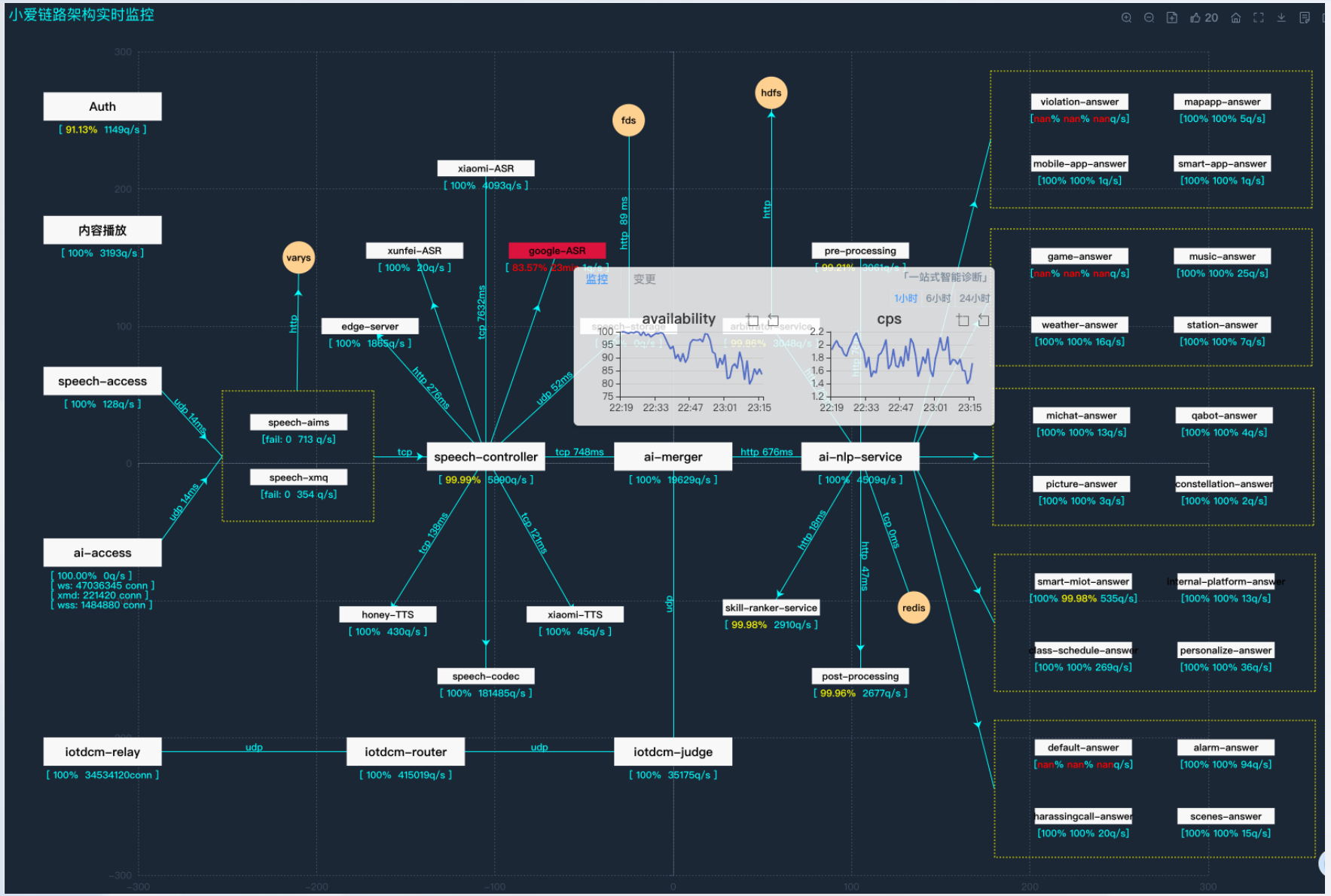
c3 c4 c5 qrc6

tj1 sh mos mb

or sgp fr

组名	用量	数量	c3	c4	c5	qrc6	tj1	sh	others
ai-archtest	cpu: 99.7% mem:85.4%	4	c[redacted]0.bj: 24.4%	c4-a[redacted]0.bj: 2.8%		[redacted] 8.0%			sgp1-ai
ai-ai-service-ksc...	cpu: 97.9% mem:38.8%	1					tj1-ai-ai-service-ksc-va2-2-10341-9640vxhd0.kscn: 97.9%		
ai-prod-mon-i1	cpu: 96.8% mem:57.2%	2	c3[redacted]-01.bj: 96.8% c3[redacted]-04.bj: 87.9%						
ai-prod-smartmiot	cpu: 84.1% mem:88.0%	31	c3-ai-prod-smartmiot01.bj: 52.3% c[redacted]3.bj: 52.2% c[redacted]2.bj: 48.2% c[redacted]8.bj: 47.7% c[redacted]5.bj: 47.1% c[redacted]2.bj: 46.9% c[redacted]7.bj: 46.6% c[redacted]0.bj: 46.0% c[redacted]0.bj: 45.4% c[redacted]6.bj: 42.8% ▾展示全部16台	c[redacted]bj: 84.1% c[redacted].bj: 55.5% c[redacted].bj: 51.5% c[redacted].bj: 51.0% c[redacted].bj: 50.7% c[redacted].bj: 48.7% c[redacted].bj: 48.2% c[redacted].bj: 47.9% c[redacted]bj: 46.8% c[redacted].bj: 45.9% ▾展示全部15台					
miot-camera-ec2...	cpu: 82.5% mem:17.5%	27					tj1-miot-camera-ec2-master07.kscn: 82.5% tj1-miot-camera-ec2-master03.kscn: 74.9% tj1-miot-camera-ec2-master06.kscn: 74.8% tj1-miot-camera-ec2-master02.kscn: 73.7% tj1-miot-camera-ec2-master05.kscn: 69.9% tj1-miot-camera-ec2-master04.kscn: 68.0% tj1-miot-camera-ec2-master08.kscn: 66.7% tj1-miot-camera-ec2-master01.kscn: 34.6% tj1-miot-camera-ec2-master00.kscn: 0.7%	sh1-miot-camera-ec2-master03.kscn: 77.1% sh1-miot-camera-ec2-master02.kscn: 69.0% sh1-miot-camera-ec2-master01.kscn: 64.7% sh1-miot-camera-ec2-master06.kscn: 63.3% sh1-miot-camera-ec2-master05.kscn: 59.5% sh1-miot-camera-ec2-master04.kscn: 59.2% sh1-miot-camera-ec2-master07.kscn: 57.4% sh1-miot-camera-ec2-master00.kscn: 53.3% sh2-miot-camera-ec2-master00.kscn: 0.4%	gz1-miot-camera-ec2-master05.kscn.idc.xiaomi.com: 65.1% gz1-miot-camera-ec2-master07.kscn.idc.xiaomi.com: 63.9% gz1-miot-camera-ec2-master02.kscn.idc.xiaomi.com: 61.6% gz1-miot-camera-ec2-master01.kscn.idc.xiaomi.com: 60.6% gz1-miot-camera-ec2-master03.kscn.idc.xiaomi.com: 59.4% gz1-miot-camera-ec2-master06.kscn.idc.xiaomi.com: 57.6% gz1-miot-camera-ec2-master08.kscn.idc.xiaomi.com: 55.9% gz1-miot-camera-ec2-master04.kscn.idc.xiaomi.com: 55.5% gz1-miot-camera-ec2-master00.kscn.idc.xiaomi.com: 1.2%

链路架构


小爱链路架构实时监控
Auth
[91.13% 1149q/s]
内容播放
[100% 3193q/s]
speech-access
[100% 128q/s]
ai-access
[100.00% 0q/s]
[ws: 47036345 conn]
[xmd: 221420 conn]
[wss: 1484880 conn]
iotdcm-relay
[100% 34534120conn]
varys
fds
hdfs
xiaomi-ASR
[100% 4093q/s]
xunfei-ASR
[100% 20q/s]
google-ASR
edge-server
[100% 1865q/s]
speech-aims
[fail: 0 713 q/s]
speech-xmq
[fail: 0 354 q/s]
speech-controller
[99.99% 5890q/s]
honey-TTS
[100% 430q/s]
xiaomi-TTS
[100% 45q/s]
speech-codec
[100% 181485q/s]
iotdcm-router
[100% 415019q/s]
iotdcm-judge
[100% 35175q/s]
ai-merger
[100% 19629q/s]
pre-processing
skill-ranker-service
[99.98% 2910q/s]
post-processing
[99.96% 2677q/s]
ai-nlp-service
[100% 4509q/s]
redis
监控
变更
「一站式智能诊断」
1小时 6小时 24小时
availability
cps
violation-answer
[nan% nan% nanq/s]
mapapp-answer
[100% 100% 5q/s]
mobile-app-answer
[100% 100% 1q/s]
smart-app-answer
[100% 100% 1q/s]
game-answer
[nan% nan% nanq/s]
music-answer
[100% 100% 25q/s]
weather-answer
[100% 100% 16q/s]
station-answer
[100% 100% 7q/s]
michat-answer
[100% 100% 13q/s]
qabot-answer
[100% 100% 4q/s]
picture-answer
[100% 100% 3q/s]
constellation-answer
[100% 100% 2q/s]
smart-miot-answer
[100% 99.98% 535q/s]
internal-platform-answer
[100% 100% 13q/s]
personalize-answer
[100% 100% 36q/s]
default-answer
[nan% nan% nanq/s]
alarm-answer
[100% 100% 94q/s]
scenes-answer
[100% 100% 15q/s]
http 89 ms
http
tcp 7632ms
http 276ms
udp 52ms
udp 14ms
udp 14ms
tcp
tcp 748ms
http 676ms
tcp 139ms
tcp 121ms
udp
http 18ms
http 47ms
tcp 0ms
udp
udp

磁盘自愈

SREelite

查看全部 cop.xiaomi_owt.AI_pdl.AI-Service

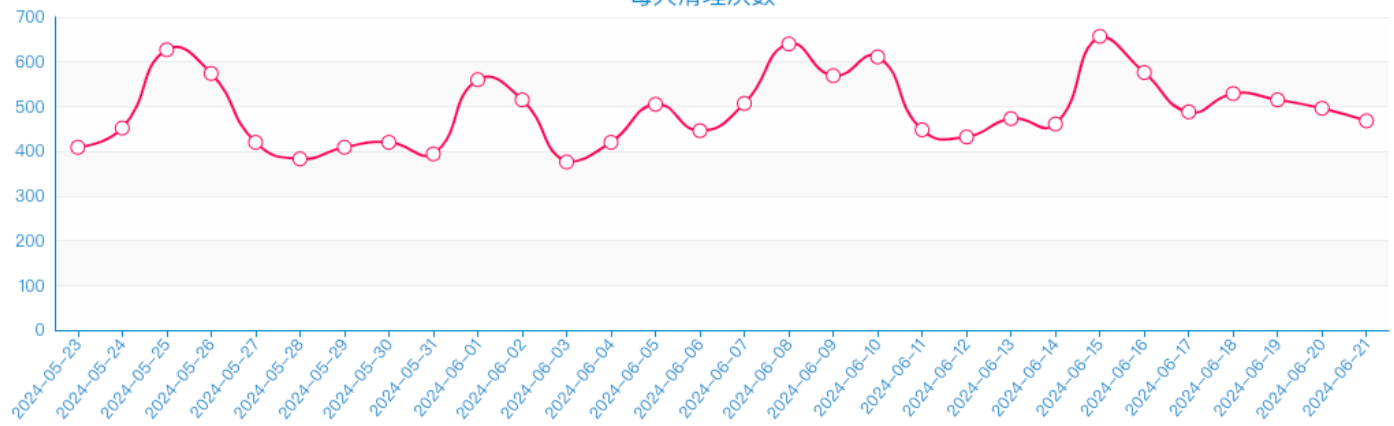
磁盘总量	磁盘剩余	覆盖机器	清理次数	清理空间
5.6 PB	3.4 PB / 60.7%	2381	531,543	118.6 PB

清理次数 Top 10

Hostname	Count
[redacted]	8611
[redacted]	8561
[redacted]	7557
[redacted]	7467
[redacted]	7423
[redacted]	7266
[redacted]	6924
[redacted]	6613
[redacted]	6140
[redacted]	5983

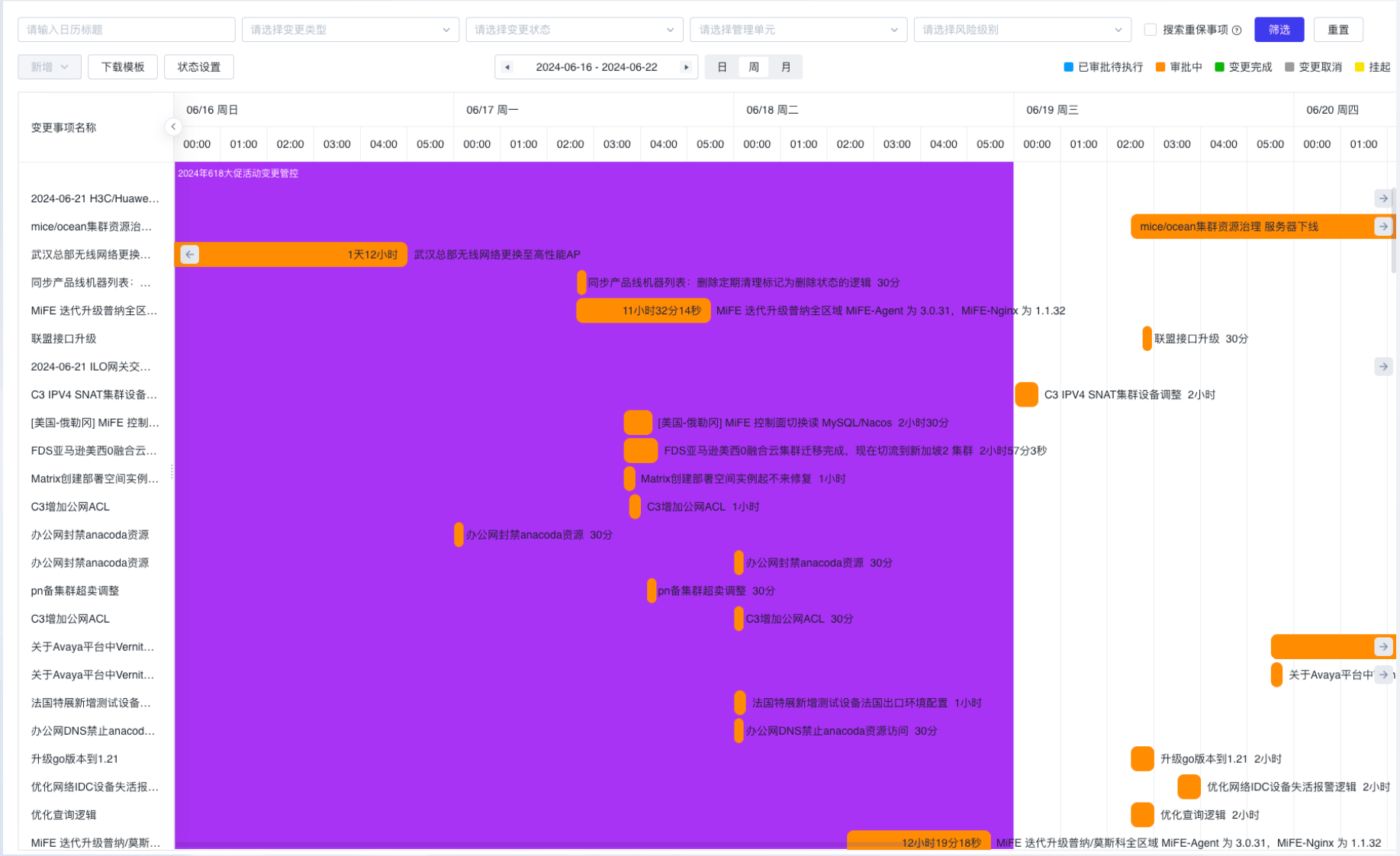
7天 30天

每天清理次数



变更日历

SREelite



故障上报

SREelite

故障处理

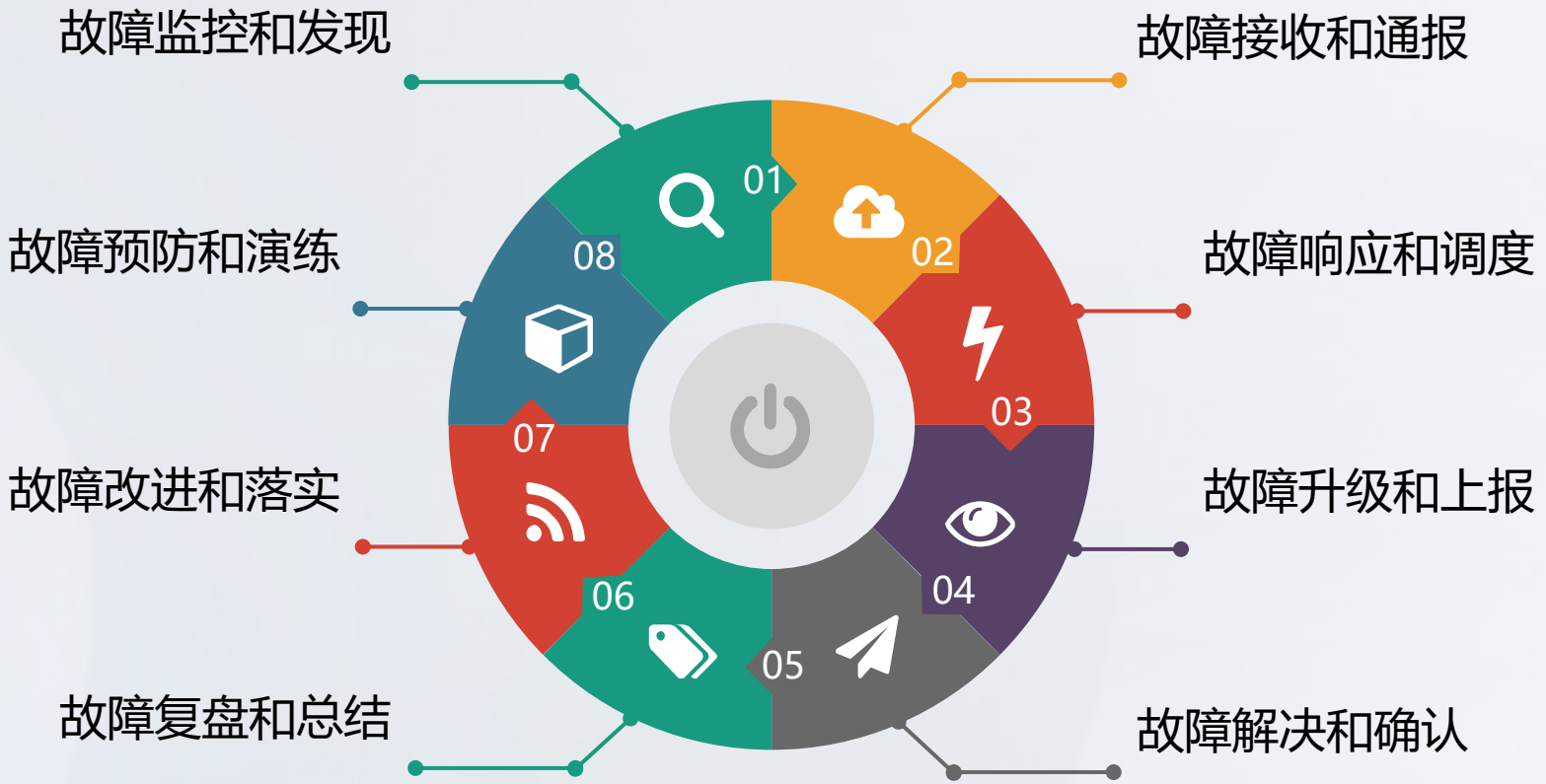
SREelite

统一故障上报入口，建立大部及公司级故障应急响应中心，**提高故障处理效率。**

从发现、响应、复盘、改善、预防进行全方位质量异常闭环管理，确保服务稳定性持续提升

故障管理目标

降低故障数量
缩短故障时长
降低故障对产品影响
预防同类故障再次发生
确保各平台和产品服务质
保障良好的用户体验
维护公司品牌形象



04

总结

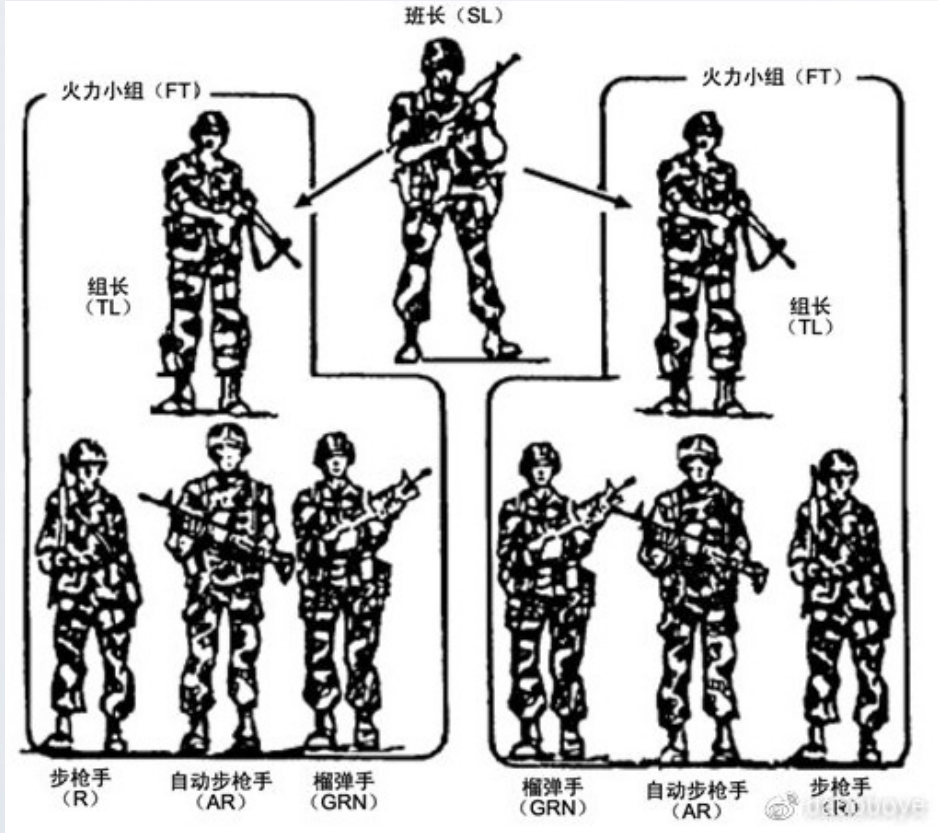
SREelite

总结

SREelite

故障应急 考较的是SRE体系构建的综合能力，包括丰富的信息采集，快速的反应和决策，稳定准确的处置措施，超强的资源整合，日常以练养战的流程体系以及经验丰富能力强大的专家团队

2004年在伊拉克的一个美国海军陆战队步兵班，里面有3挺M249。照片中有12人，第13人是拿相机的



Q&A

SREelite

https://sre-elite.com

“SRE精英联盟”概述

SREelite

2003年
Google 成立 SRE 团队

2016年
孙宇聪翻译出版首部 SRE 著作

2020年
赵成牵头成立 SRE研讨社群

SREelite
2023年11 月
SRE 精英联盟正式成立

《SRE 实践白皮书》首次正式发布

2024年2月

联盟首次直播

2024年2月29 日

联盟首次线下 Meetup

2024年4月13 日

https://sre-elite.com

《SRE实践白皮书》

- 经历数年，20 多位一线专家协作编写。
- 扫码下载 v1.0.1 。版本持续更新迭代中。
- 在官网 https://sre-elite.com/notice/下载最新版。

关注我们

SREelite

公众号

视频号

B 站

YouTube