美图故障管理体系搭建实践

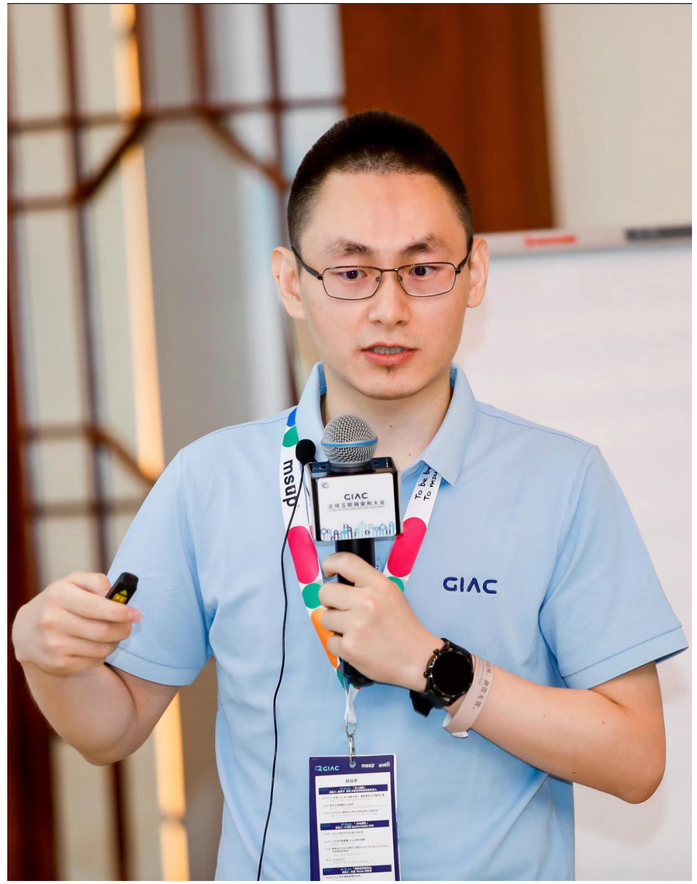
2024年6月

石鹏(东方德胜)@美图

SREelite

https://sre-elite.com

讲师介绍



石鹏(东方德胜)

高级运维经理

meitu美图

从业十余年，一直从事运维相关的工作。
2016年加入美图公司，现任美图SRE负责人，目前整体负责美图公司线上服务的稳定性保障工作。
曾多次参与或主导过美图公司多项基础设施、运维架构的调整和改造，在监控、灾备、故障管理、稳定性运营等方面有一定的经验积累和行业输出。

致力于推广SRE、稳定性运营相关的理念及实践，编著有「SRE系统建设指南」图谱，参与过业界多个SRE、DevOps相关案例集/期刊/标准/白皮书的编纂或供稿。

业界多个技术峰会的分享嘉宾、金牌讲师或出品人，中国信通院「稳定性保障实验室」认证专家，SRE精英联盟成员。

Part one

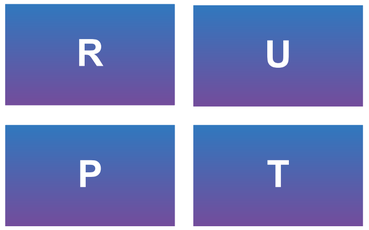
SRE的职责及困境

简介：VUCA时代下SRE/运维岗位所面临的挑战

VUCA时代

SREelite

VUCA时代



2018

- **R**apid (急剧)
- **U**npredictable (莫测)
- **P**aradoxical (矛盾)
- **T**angled (缠绕)

V
Volatility(易变性)

U
Uncertainty(不确定性)

1990s

C
Complexity(复杂性)

A
Ambiguity(模糊性)

2022

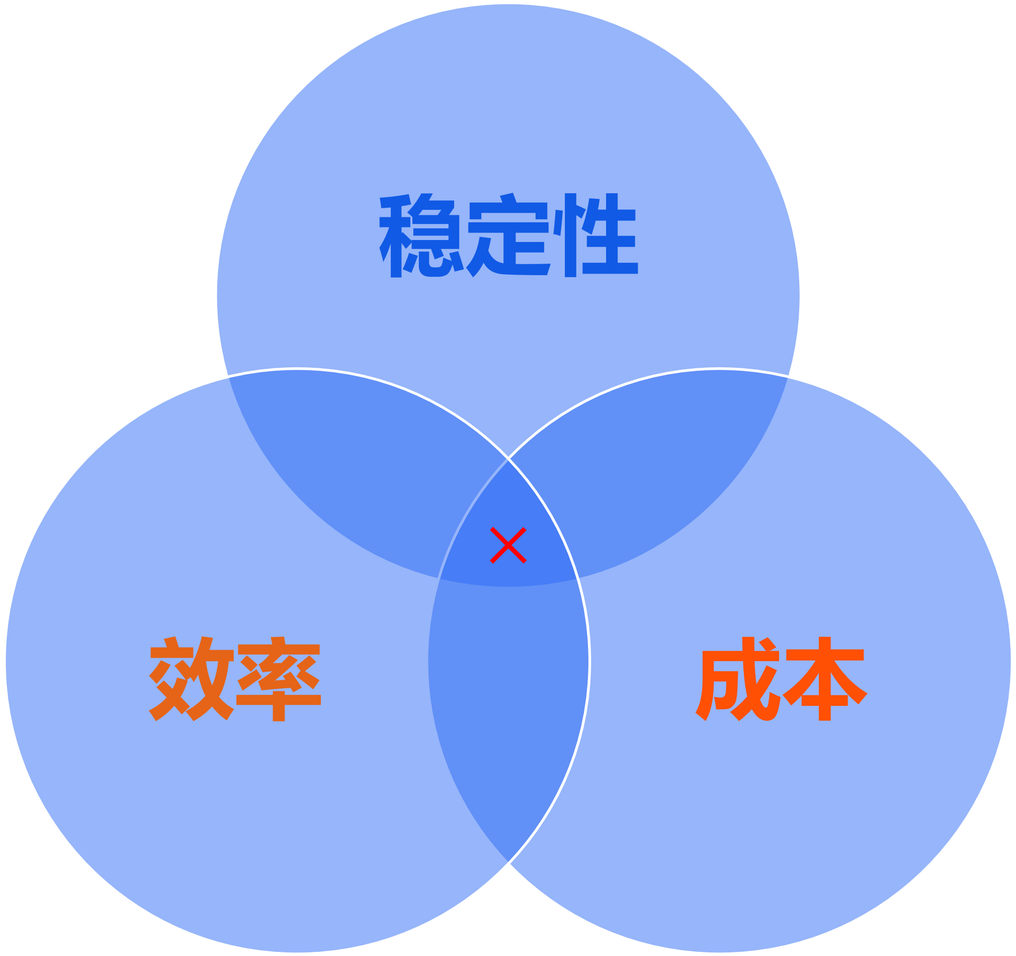
- **B**rittle (脆弱)
- **A**nxious (焦虑)
- **N**onliner (非线性)
- **I**ncomprehensible (不可理解)

https://www.vuca-world.org/vuca-bani-or-rupt/

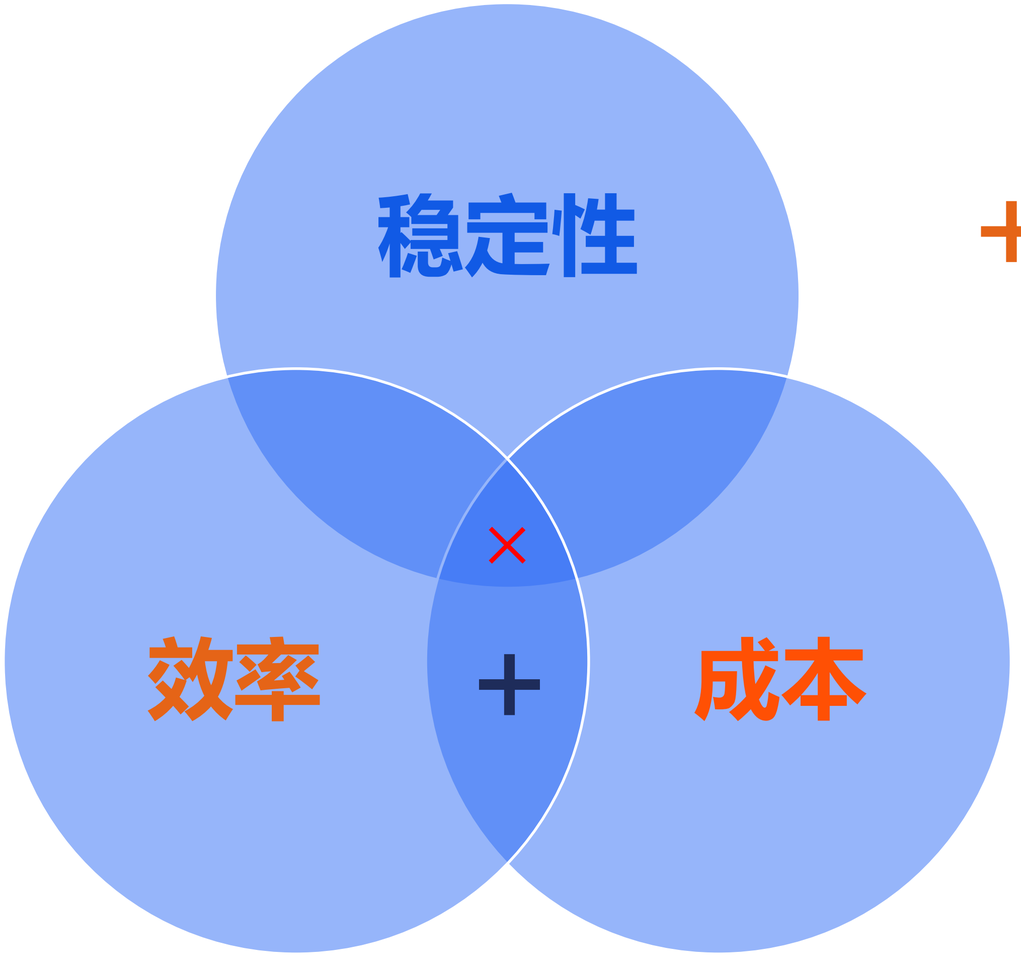
美图SRE的核心工作职责

- 岗位：产品SRE
- 职责：
  ① 保障线上服务的**稳定性**
  ② **建设**工具/平台/基础设施 提升**效率**
  ③ 用技术手段来控制、优化服务的**运行成本**
- 愿景：做美图服务最稳的大后方

SRE：寻求三个核心职责之间的平衡



三个核心职责 与 企业发展的关系



\+ 安全 → 安全生产

让企业活着

→ 降本增效

让企业获得优势

SRE的困境

SREelite

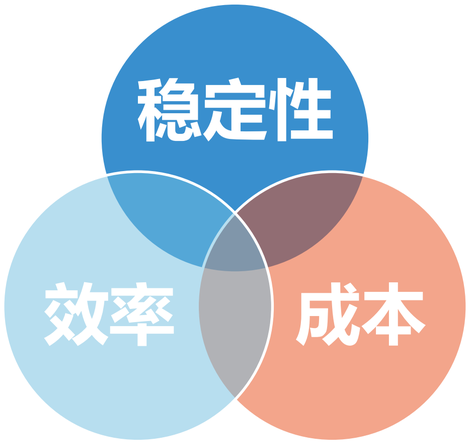
V U C A

人手不足

- 业务线众多且复杂
- 琐事过多
  - 低效的咨询或求助
  - 例行的工作事项
- 新人培养成本高
- 风吹草动 捉襟见肘

稳定性压力大

- 技术栈多样 更新快
- 服务发布频率高
- 历史技术债务堆积
- 告警消息满天飞
- 战战兢兢 如履薄冰
- 忙于救火 疲于应付



工具不称手

- 工具缺失或落后
- 工具建设进度缓慢
- 轮子众多 不成体系
- 工具间数据割裂
- 数据/行为的不一致性

成本管控压力大

- 组织部门多 成本数据集大
- 视角、目标优先级差异大
- 资源用量/利用率持续管控
- 资源容量动态规划和管控
- 工作繁复: 归集/分析/核算
- 成本持续优化的目标压力

危机四伏

忙于救火

疲于应付

恶性循环

缺乏幸福感

删库跑路...

今天我们聚焦在「稳定性」聊一聊

今天我们聚焦在「稳定性」聊一聊

SREelite

Part two

稳定性运营

简介：从被动应对到主动出击

- 稳定性运营的一些实施准则
- 稳定性的度量和目标
- 稳定性运营的全景图

稳定性运营的一些实施准则

SREelite

SRE

Measuring and Managing Reliability

宏观框架、理论支撑

数据思维、度量先行

定期复盘、流程闭环

紧扣价值、持续输出

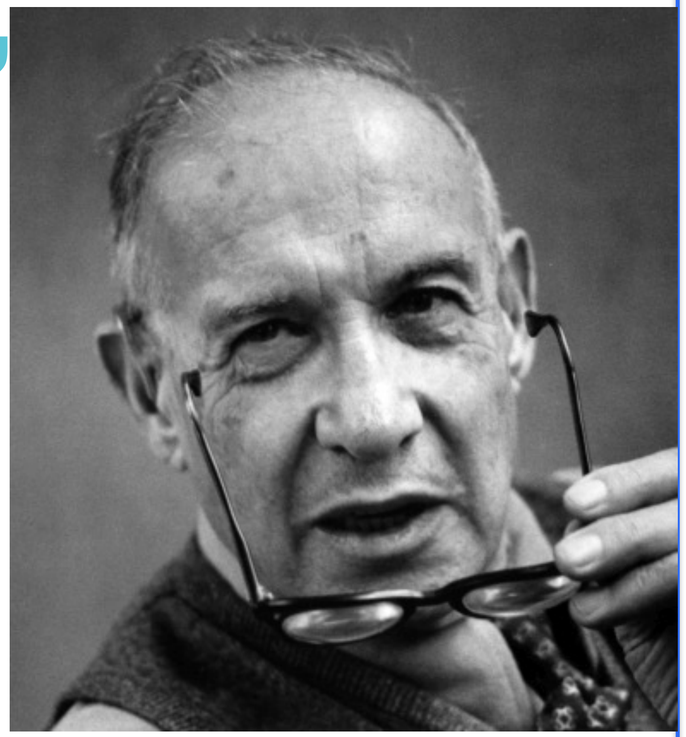
不设边界
主动出击

稳定性的度量&目标

If you can’t measure it,
you can’t improve it.

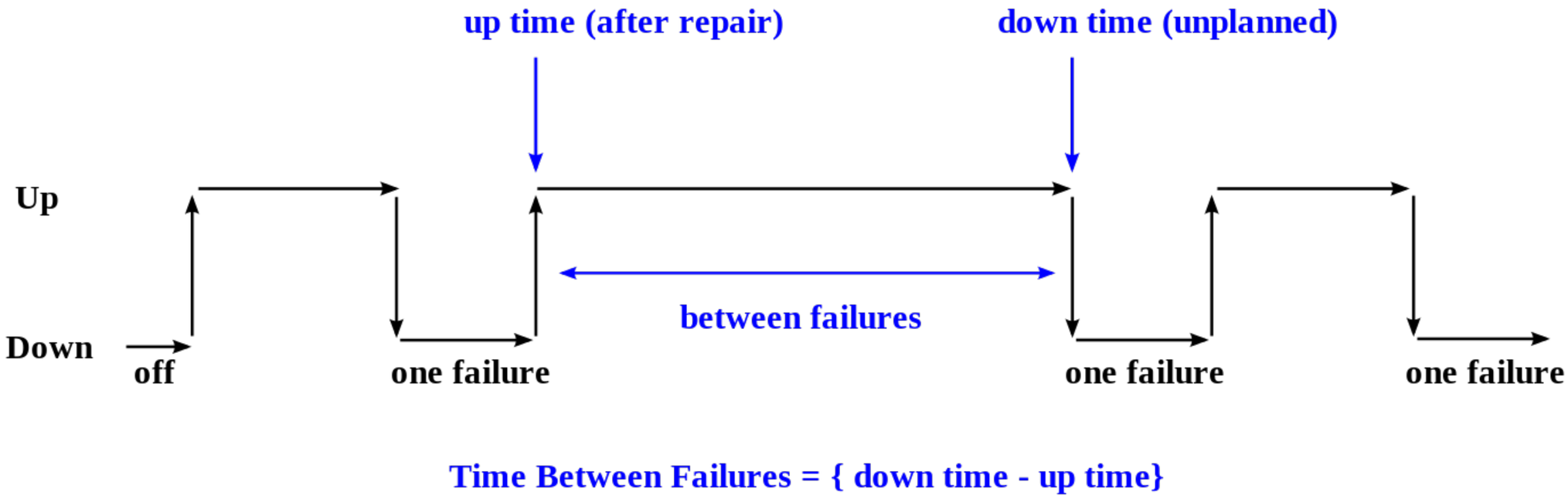
如果你不能度量它，
你就无法改进它。

---- Peter Drucker
彼得·德鲁克



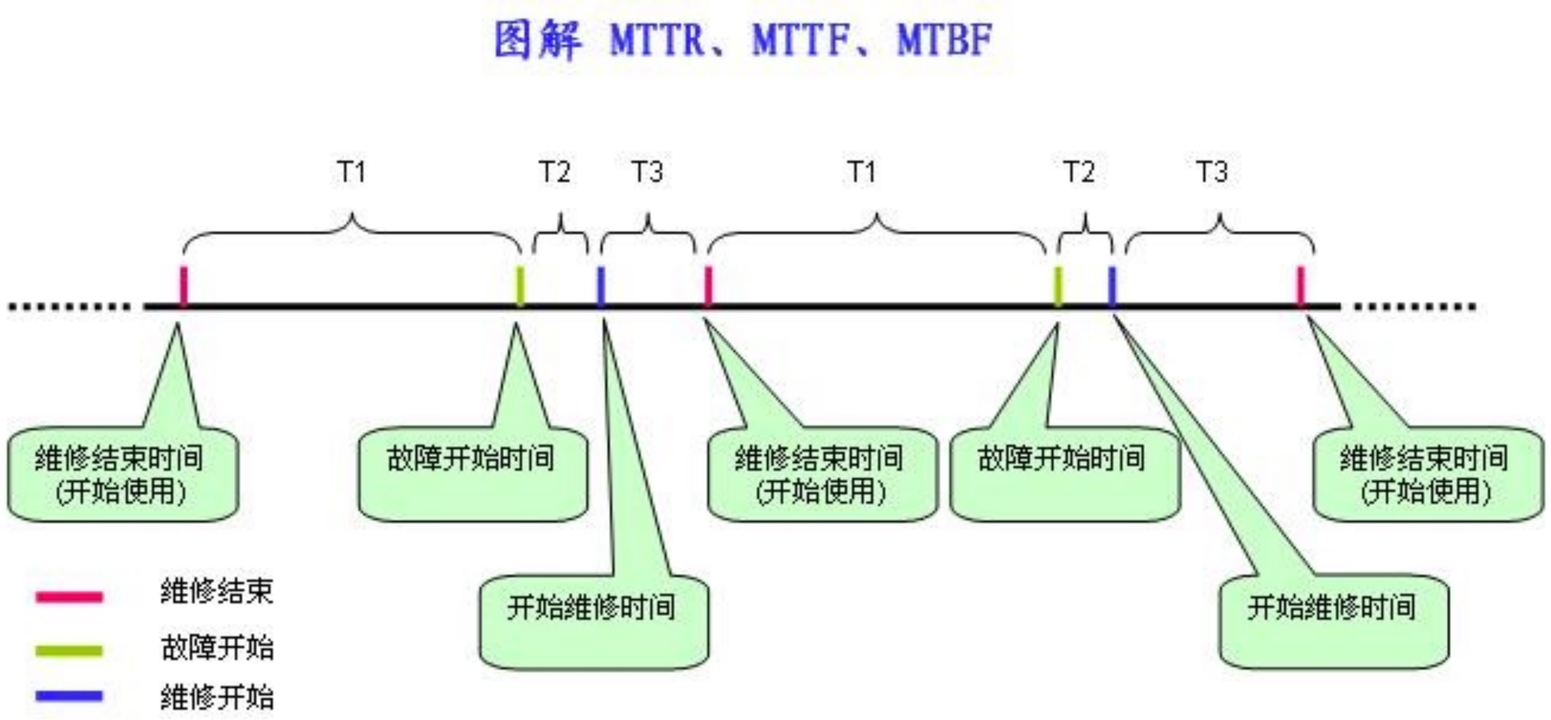
稳定性的度量&目标

- MTBF：平均故障间隔
- MTTR：平均修复时间
- MTTF：平均无故障时间



稳定性的度量&目标

- **MTBF：平均故障间隔** MTBF =∑(T2+T3+T1)/ N
- **MTTR：平均修复时间** MTTR =∑(T2+T3)/ N
- **MTTF：平均无故障时间** MTTF =∑T1/ N



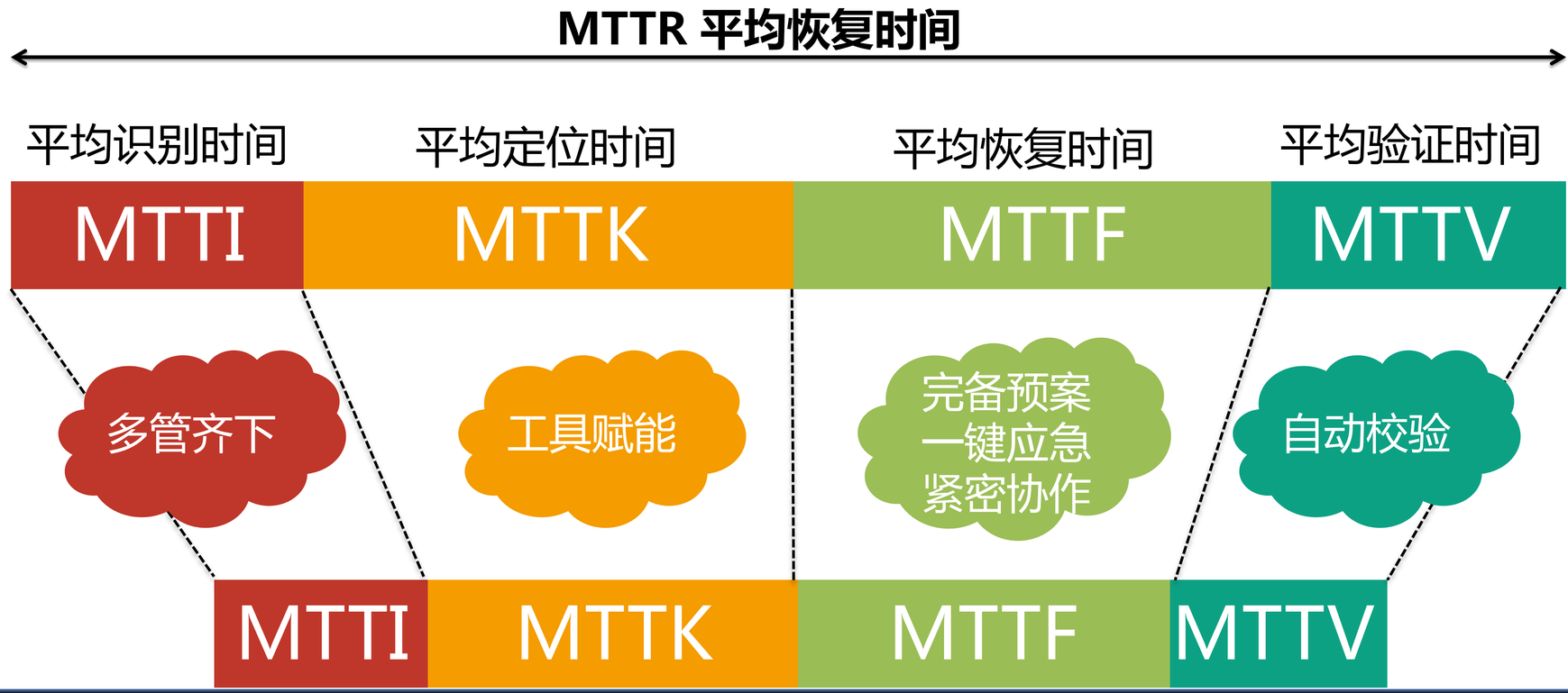
稳定性的度量&目标

- 细化MTTR

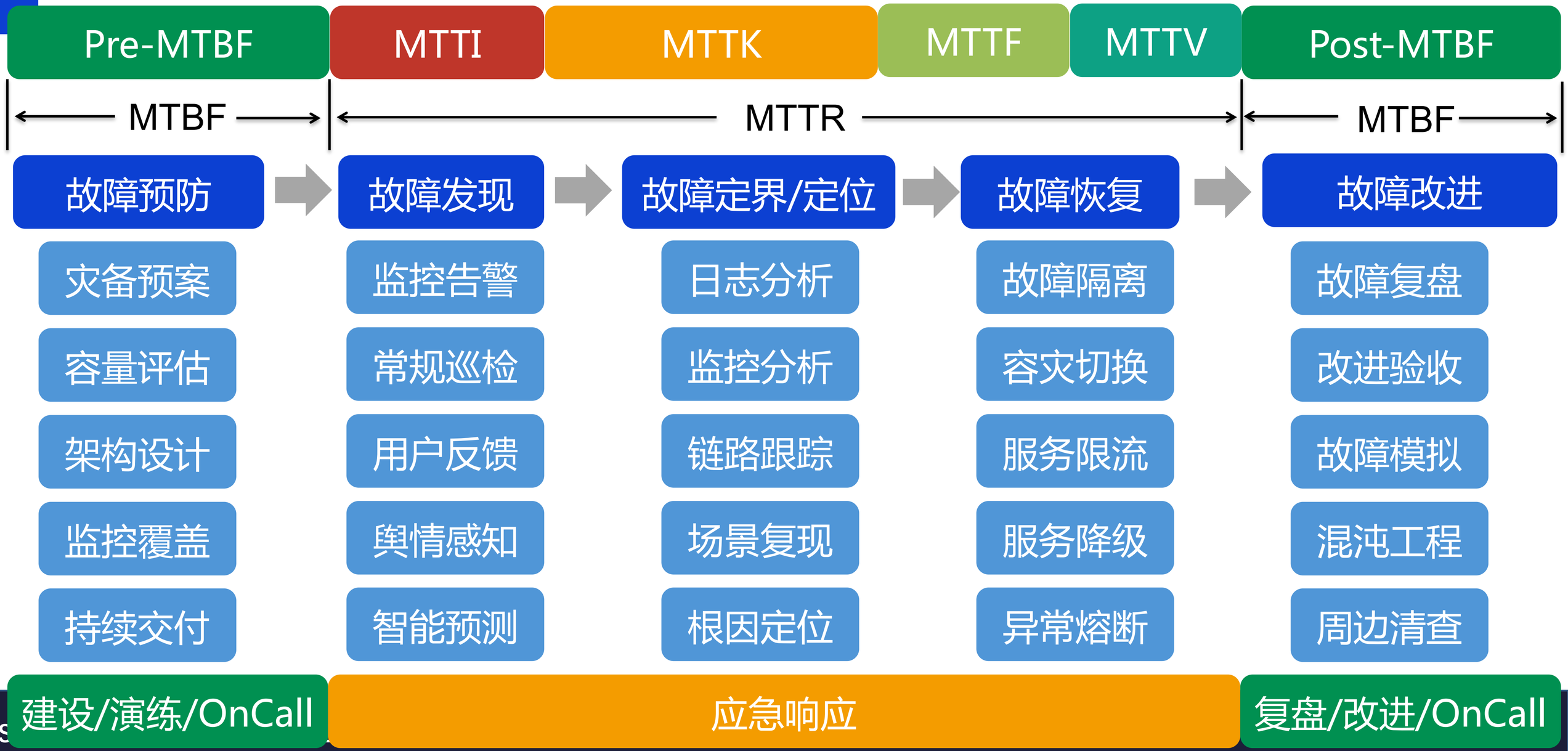
- 我们的目标

稳定性的度量&目标

- 如何达成目标



稳定性运营的全景图


Pre-MTBF
MTTI
MTTK
MTTF
MTTV
Post-MTBF
MTBF
MTTR
MTBF
故障预防
故障发现
故障定界/定位
故障恢复
故障改进
灾备预案
容量评估
架构设计
监控覆盖
持续交付
监控告警
常规巡检
用户反馈
舆情感知
智能预测
日志分析
监控分析
链路跟踪
场景复现
根因定位
故障隔离
容灾切换
服务限流
服务降级
异常熔断
故障复盘
改进验收
故障模拟
混沌工程
周边清查
建设/演练/OnCall
应急响应
复盘/改进/OnCall

Part three

故障治理

简介：故障管理的方法论

- 正确认识故障
- 故障生命周期管理
- 部分美图的实例演示

正确认识故障

SREelite

异常是常态

- 系统失效/异常的必然性
- 所有的干预手段都有代价

为何会发生

- 单机故障
- 负载变化
- 人为错误

常见的原因

- 配置变更
- 强依赖
- 时延增加
- 资源耗尽

From：FaceBook – Fail at Scale

故障生命周期管理-三段式拆解

故障生命周期管理-故障前

SREelite

监控覆盖	架构设计	容量评估
灾备预案	灾备演练	工具建设
OnCall	常规巡检	重点保障

故障生命周期管理-故障前：可观测建设

SREelite

Metrics, tracing, and logging

故障开始 → 故障恢复

故障排查/处理的主要过程

Metrics

➢告诉我们*有没有故障*

◆监控告警 / 常规巡检

- 告警通知
- 监控大盘
- 北极星指标
- 基础指标

Traces

➢告诉我们*故障在哪里*

◆链路跟踪

- 调用链路
- APM
- NPM
- RUM

Logs

➢准确告诉我们*故障原因*

◆日志分析 / 事件关联

- 系统日志
- 组件日志
- 应用日志
- Profile日志
- 变更事件

故障生命周期管理-故障前：可观测建设

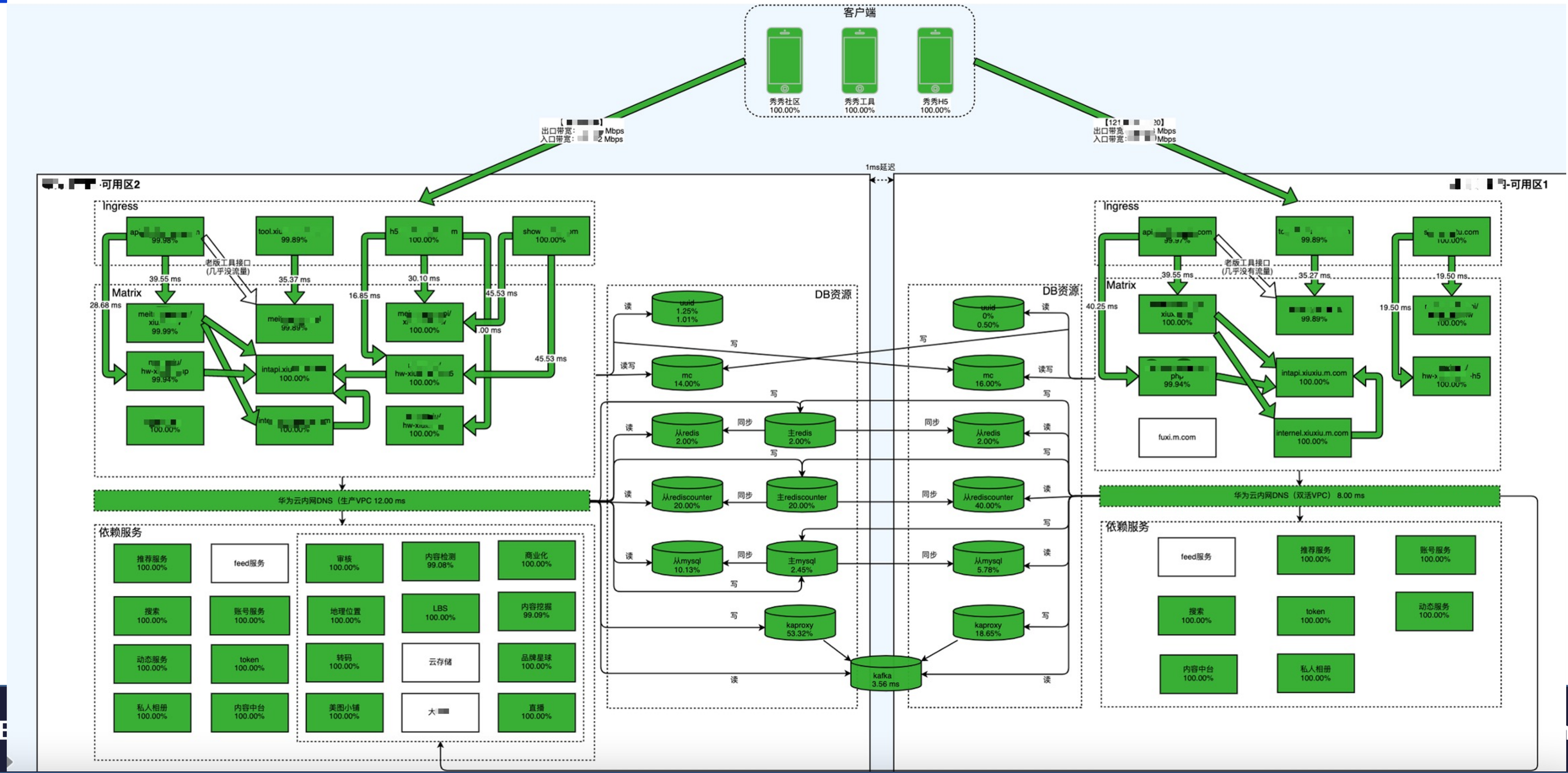
SREelite

用户端监控	流媒体监控	业务监控	服务监控	基础资源
• 网络质量&异常 • 内容&DNS劫持 • 崩溃&卡顿 • 返回码 • 响应时间 • 错误率 • 慢请求 • 请求吞吐量 • 组合分析	• 直播推流/拉流 • 点播拉流 • 主播监控 • 视频监控 • 直播/点播统计 • CDN质量 • CDN评分 • CDN日志	• 业务可用性 • 访问量/错误 • Profile监控 • 分布耗时 • Trace监控 • A/B Test监控 • 日志中心 • 事件监控	• DNS/ELB • 七层负载均衡 • SSL证书 • 进程/端口 • 后端资源 • 云PaaS服务 • SLA体系 • 产品运营指标	• 云IaaS监控 • 硬件监控 • 网络监控 • 专线监控 • TCP监控 • 容器监控 • 内核监控

第三方拨测	自研流媒体监控	InfluxDB套件	ElasticStack	OpenFalcon	eBPF
自研APM	自研CDN监控	SkyWalking	大数据流式处理套件	Prometheus	

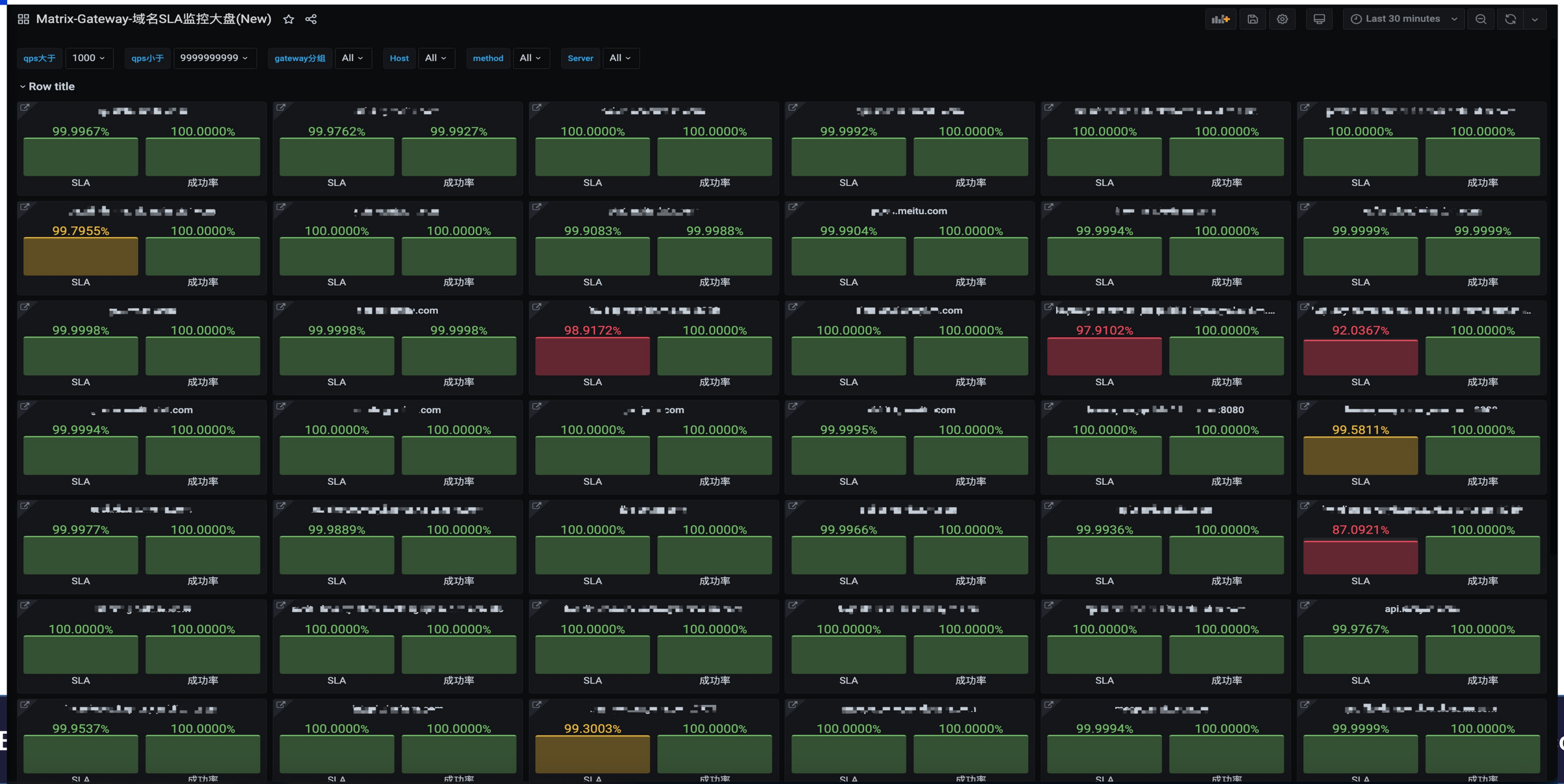
用户端质量监控体系 ⟷ 服务端质量监控体系

故障生命周期管理-故障前：可观测建设



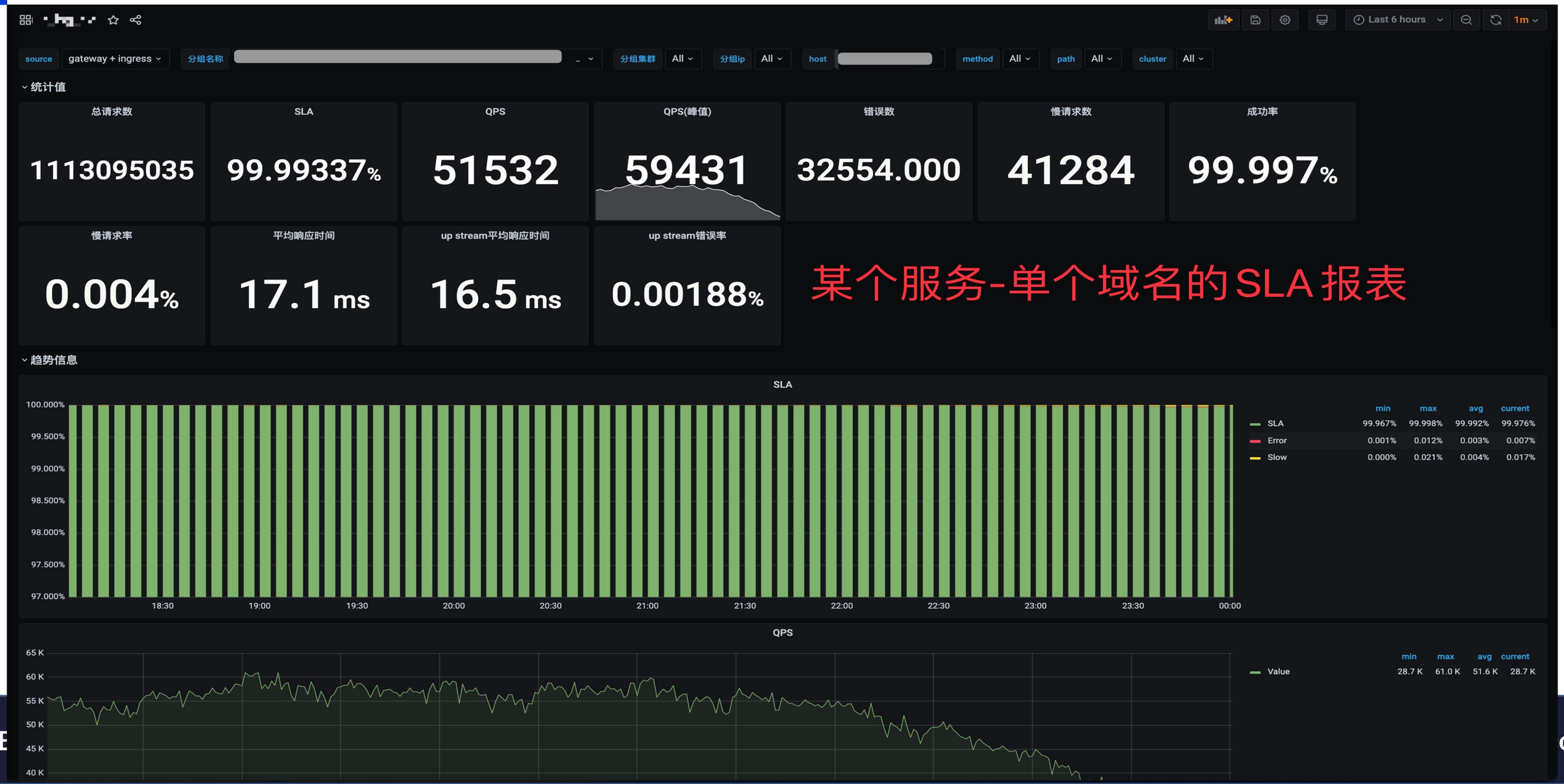
故障生命周期管理-故障前：可观测建设

SREelite



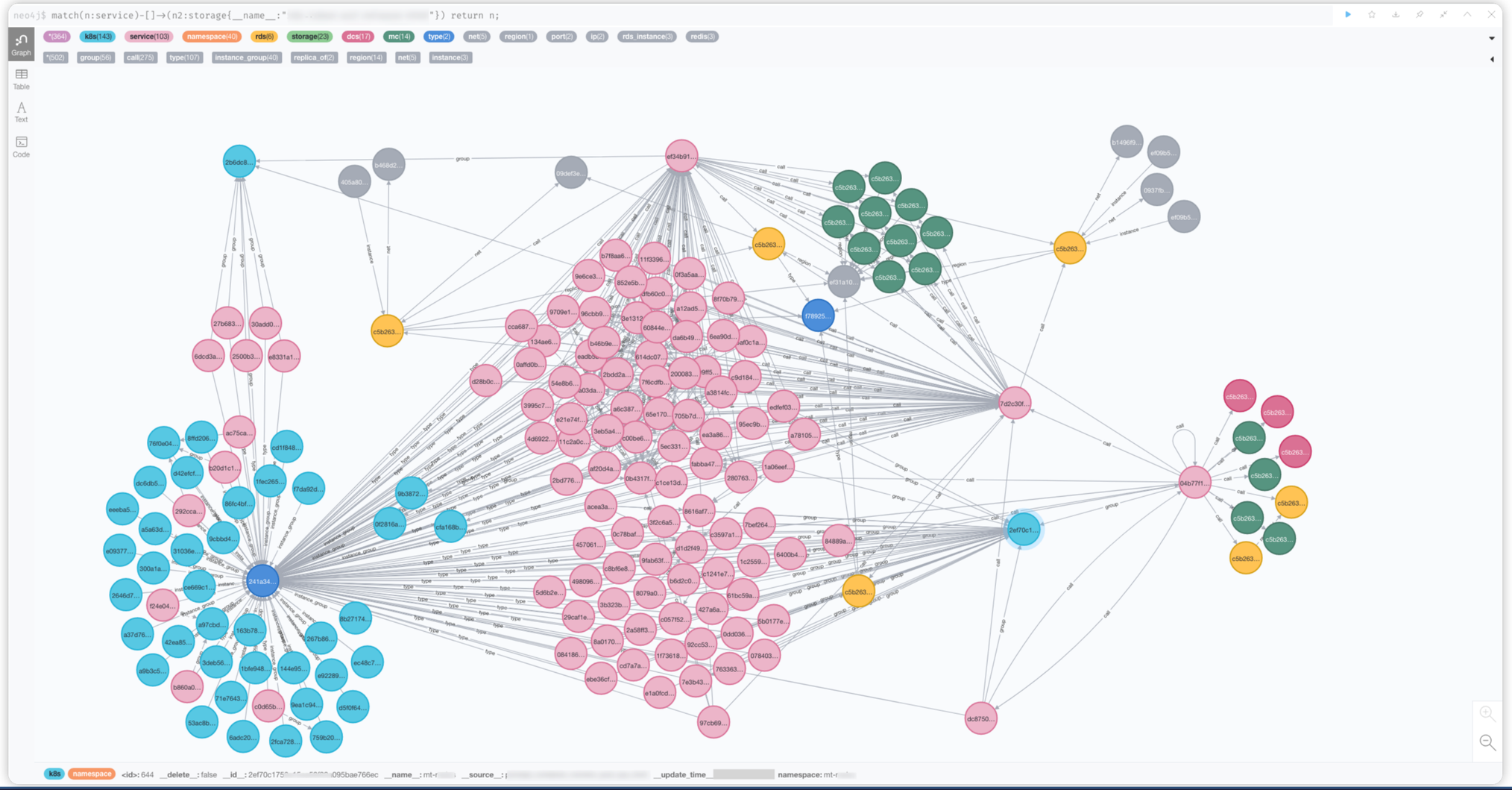
故障生命周期管理-故障前：可观测建设

SREelite

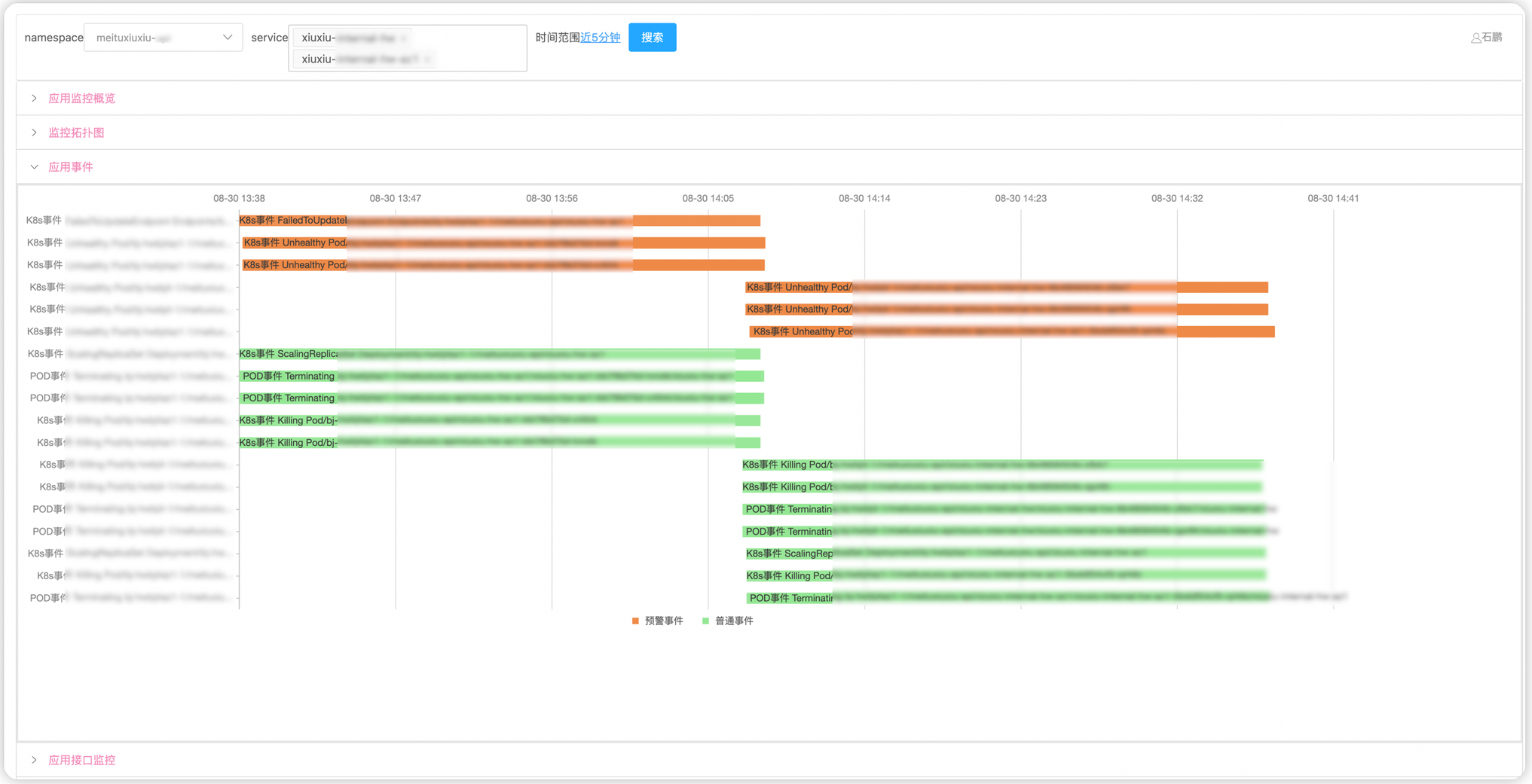


故障生命周期管理-故障前：可观测建设

SREelite



故障生命周期管理-故障前：可观测建设

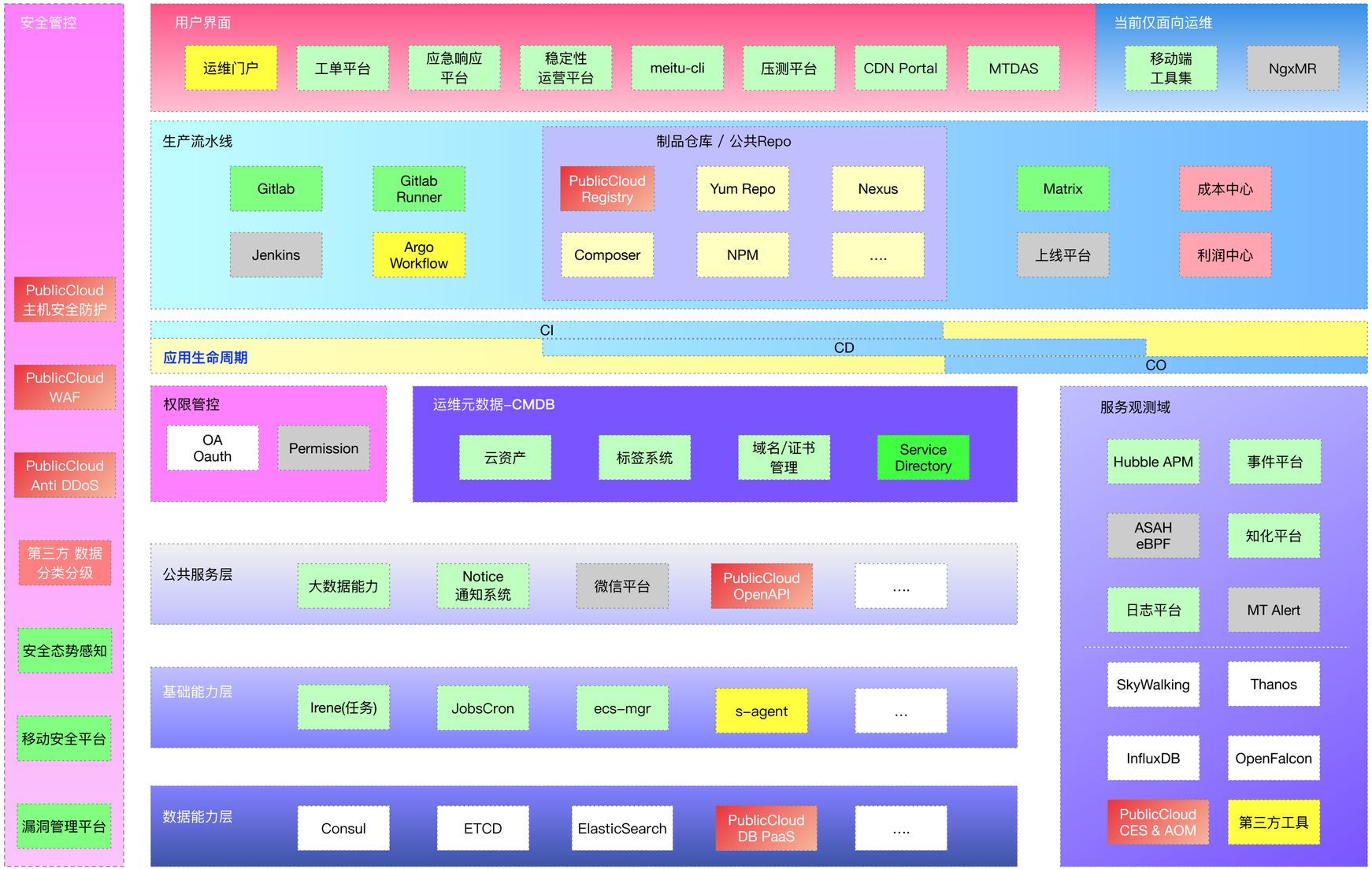


故障生命周期管理-故障前：灾备建设

SREelite

服务梳理	预案梳理	沙盘推演	预案落地	预案演练
请求链路	多级预案	部门协作	文档输出	无损演练
分段分层	各个击破	case推演	功能实现	轻损演练
周边依赖	智能调度	头脑风暴	架构适配	单点演练
架构风险	柔性设计	互相挑战	工具建设	组合演练

故障生命周期管理-故障前：基础能力建设



故障生命周期管理-故障中

SREelite

- 监控告警
- 日志分析
- 链路跟踪
- 关联事件
- 故障定界
- 预案匹配
- 故障隔离
- 容灾切换
- 降级熔断

故障生命周期管理-故障中：告警

SREelite

【[告警][级别-高]elb_elb流出流量5分钟突降30%】
忽略告警
time:2020-06-09 10:51:00
instance:9ee2443f-a[redacted]-adf94081c66f
__name__:elb:name_instance:m[redacted]3ps:rate5m
name:[redacted] proxy外网代理vip
value:-0.3497750513814819
warnName:elb流出流量突降30%

规则描述:elb流出流量5分钟突降30%
告警持续时间:1分钟

【[告警][级别-高]elb_elb流出流量5分钟突降30%】
忽略告警
time:2020-06-09 10:53:00
instance:9ee2443f[redacted]c-adf94081c66f
__name__:elb:name_instance:r[redacted]ut_Bps:rate5m
name[redacted]proxy外网代理vip
value:-0.39166423518381527
warnName:elb流出流量突降30%

规则描述:elb流出流量5分钟突降30%
告警持续时间:1分钟

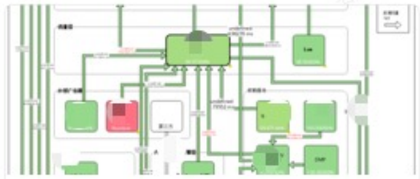
[redacted]健康大盘 BOT

[redacted]E健康大盘
[redacted]服务30分钟内RT状态[redacted]的prom监控

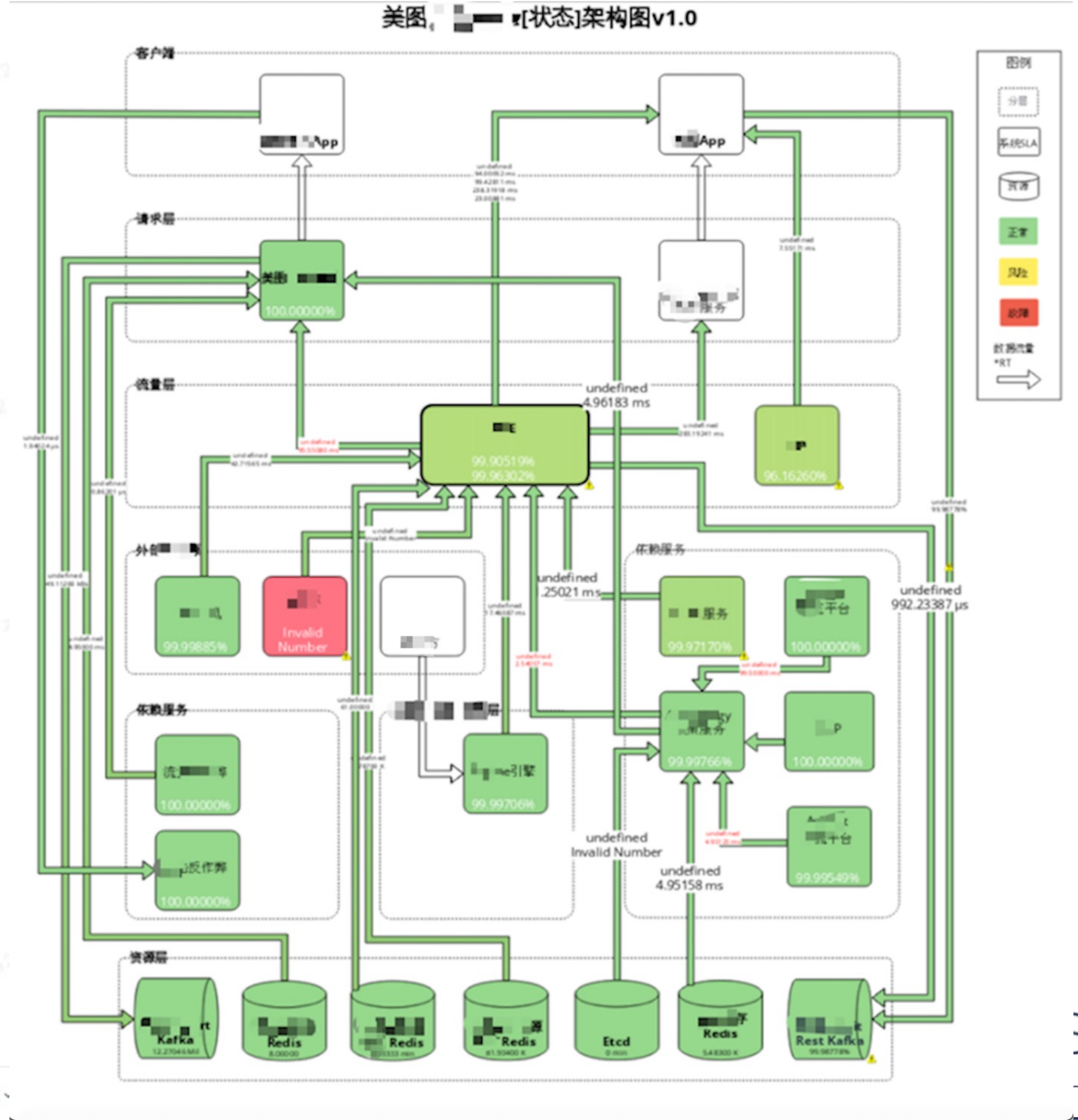
[redacted]健康大盘 BOT

[redacted]健康大盘
[redacted]服务30分钟内RT状态[redacted]服务内的prom监控

[redacted]健康大盘 BOT



[redacted]健康大盘
[redacted]E请求[redacted]服务接口RT大于2ms



故障生命周期管理-故障中：日志

SREelite

[redacted] hits Download: 500 1000 5000 10000 New Save Open Share Reporting Auto-refresh Last 15 minutes

Search... (e.g. status:200 AND extension:PHP) Options

Add a filter +

*:k8s_meit [redacted] -ap[redacted]

Selected Fields

? _source

Available Fields

Popular

t k8s.pod_name

t log.app_name

t log.level

t log.message

t log.type

@timestamp

t @version

t _id

t _index

_score

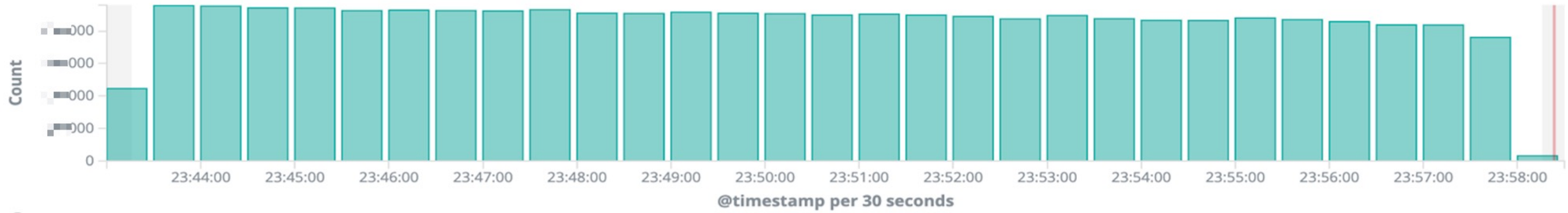
t _type

t k8s.cluster

t k8s.container_id

t k8s.container_name

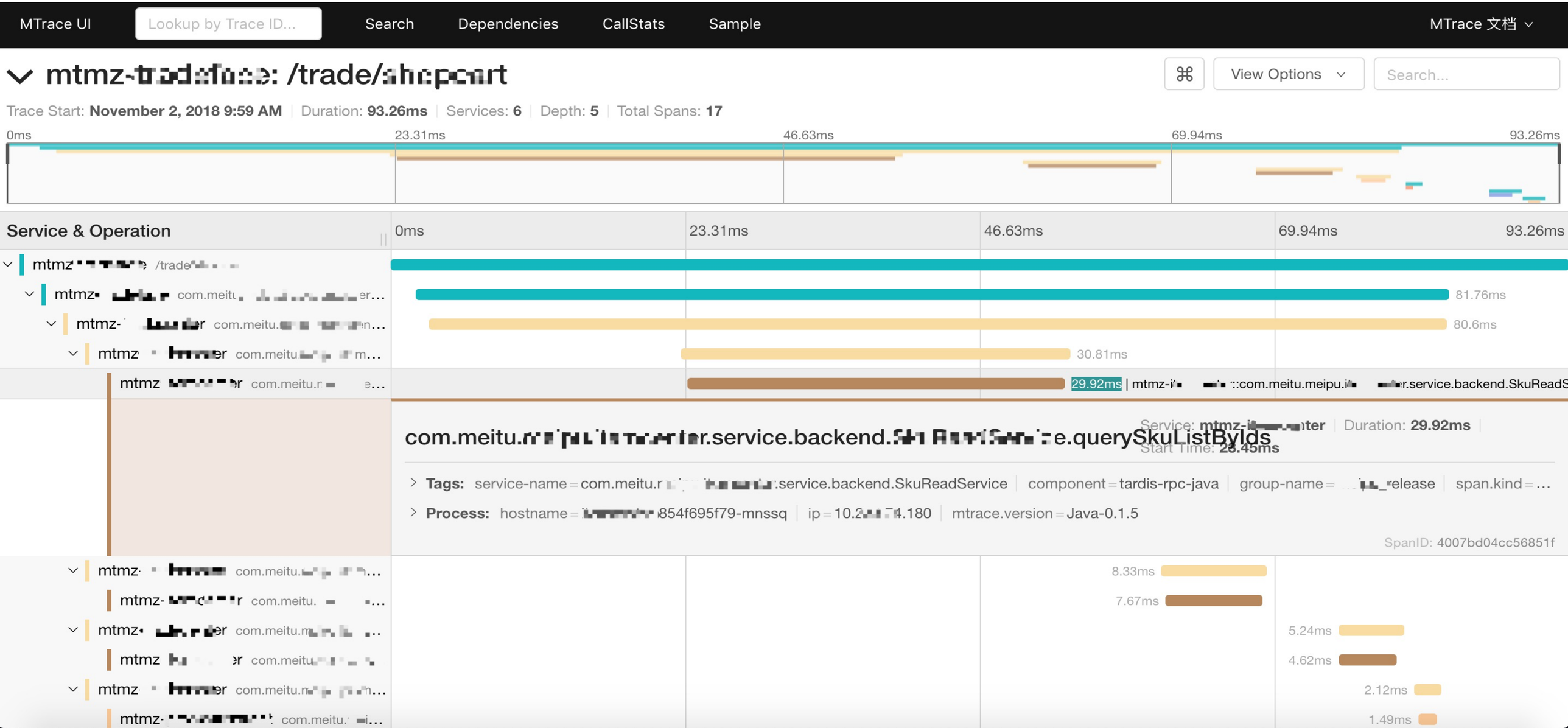
June 8th 2020, 23:43:15.821 - June 8th 2020, 23:58:15.821 — Auto



Time	_source
June 8th 2020, 23:58:06.361	@timestamp: June 8th 2020, 23:58:06.361 log.slowThreshold: 500 log.total_count: 176 log.type: MC log.interval5: 0 log.interval1: 175 log.interval2: 1 log.interval3: 0 log.interval4: 0 log.name: 10[redacted]41:31198 log.slow_count: 0 log.time: June 9th 2020, 07:58:06.361 log.category: profile log.timeMillis: 1,591,631,886,361 log.avg_time: 0.76 stream: stdout k8s.cluster: [redacted]-1 k8s.container_name: [redacted]w k8s.ns: [redacted]pi k8s.host: [redacted].0.228 k8s.pod_ip: [redacted].26.224 k8s.container_id: aa8630153a09 k8s.pod_name: [redacted]-2h
June 8th 2020, 23:58:06.358	@timestamp: June 8th 2020, 23:58:06.358 log.slowThreshold: 500 log.total_count: 238 log.type: MC log.interval5: 0 log.interval1: 238 log.interval2: 0 log.interval3: 0 log.interval4: 0 log.name: [redacted]70:31202 log.slow_count: 0 log.time: June 9th 2020, 07:58:06.358 log.category: profile log.timeMillis: 1,591,631,886,358 log.avg_time: 0.43 stream: stdout k8s.cluster: [redacted] k8s.container_name: x[redacted] k8s.ns: m[redacted]-api k8s.host: [redacted]28 k8s.pod_ip: 1[redacted]26.224 k8s.container_id: aa8630153a09 k8s.pod_name: [redacted]f85-2h
June 8th 2020, 23:58:06.275	@timestamp: June 8th 2020, 23:58:06.275 log.request: GE[redacted]t.json?ad_sdk_ver[redacted]t_bra nd[redacted]ans&clien[redacted]ient_networ[redacted]Fi&[redacted]B8%[redacted]%BF[redacted].ient_os=13.4.1&client_timestamp=1591

故障生命周期管理-故障中：Trace

SREelite



故障生命周期管理-故障中：故障处理的一些原则/建议

SREelite

- 统一目标：恢复优先，问题定界 > 根因定位
- 稳定心态：SRE一定要冷静，不要慌
- 流程机制：故障升级、War Room
- 故障现场：组织协调约定、信息通报机制

故障生命周期管理-故障后

SREelite

模拟复现

根因定位

整改修复

故障复盘

故障改进

预案完善

周边清查

经验固化

案例学习

SREelite

故障复盘-黄金三问

- 我们应该怎么做，才能更快地恢复业务？
- 我们应该怎么做，才能避免再次出现类似问题？
- 我们有哪些好的经验可以总结、提炼，并固化？

- One more thing，我们还能做些什么？

故障生命周期管理-故障后：故障报告

SREelite

故障报告模板

- 1、 故障标题或名称：
- 2、 所属业务部门：
- 3、 影响功能：
- 4、 故障级别：
- 5、 服务影响时长：
- 6、 故障原因：
- 7、 对用户的影响：
- 8、 责任部门：
- 9、 责任人：
- 10、故障原因分类
- 11、故障处理过程：
- 12、改进措施：
- 13、相关提案或文档：

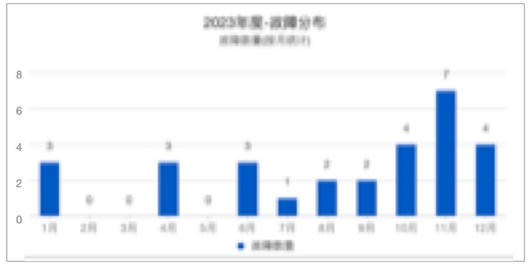
报告输出

故障报告归纳

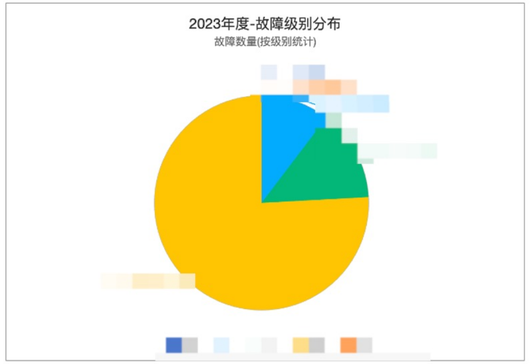
故障生命周期管理-周期回顾

2023年度 项目总结 -「O-稳定性:故障分析」

故障趋势变化

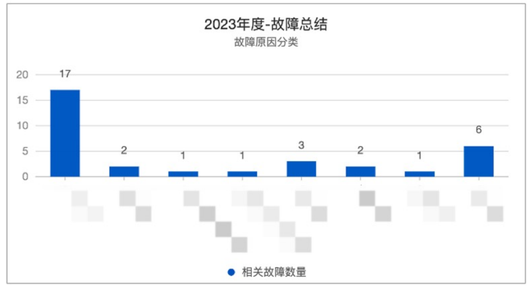


故障级别分布

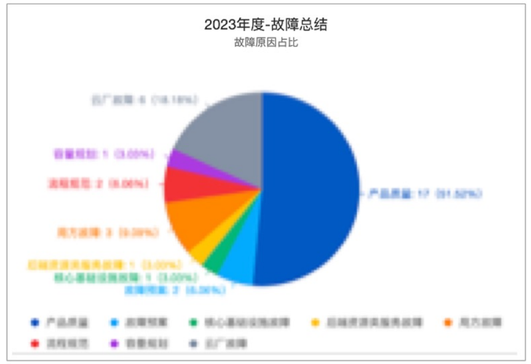


2023年度 故障原因分类

故障原因一级分类	故障原因二级分类	Sum of:相关故障数量
产品质量		
故障预案		
核心基础设施故障		
后端资源类服务故障		
局方故障		
流程规范		
容量规划		
云厂故障		
Total		



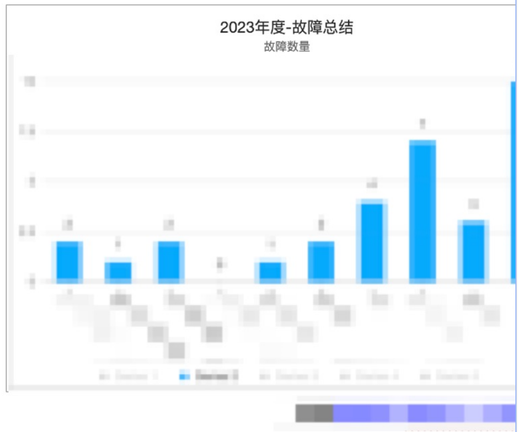
- 产品质量相关：
- 云厂/局方相关



2023年度 项目总结 -「O-稳定

故障	总分
全年考核分	
2023-全年 考核方故障总数(有扣分)	
2023-全年 剩余分	
2023-全年 总扣分	
达标率=剩余分/考核分*100%	

- 2023年度 故障分累计
- 2023 年度 故障统计
- 除【局方】外，美图所
- 【局方



Part four

美图SRE的实践

简介：

- 故障文化
- 故障管理框架
- 工具建设
- 故障案例

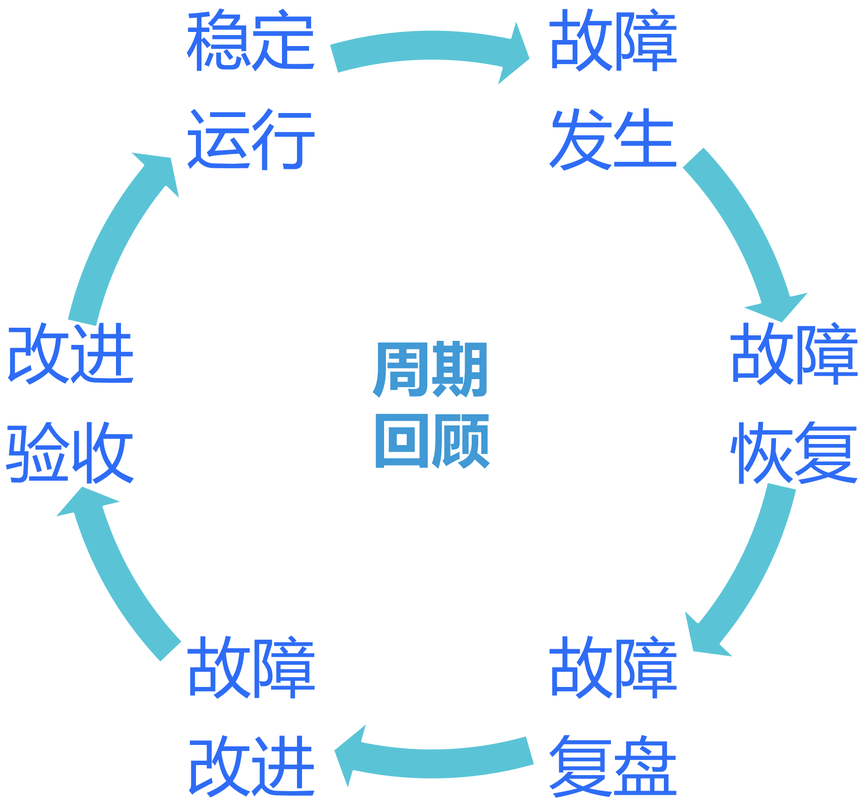
故障文化

SREelite

拥抱故障 卓越运维

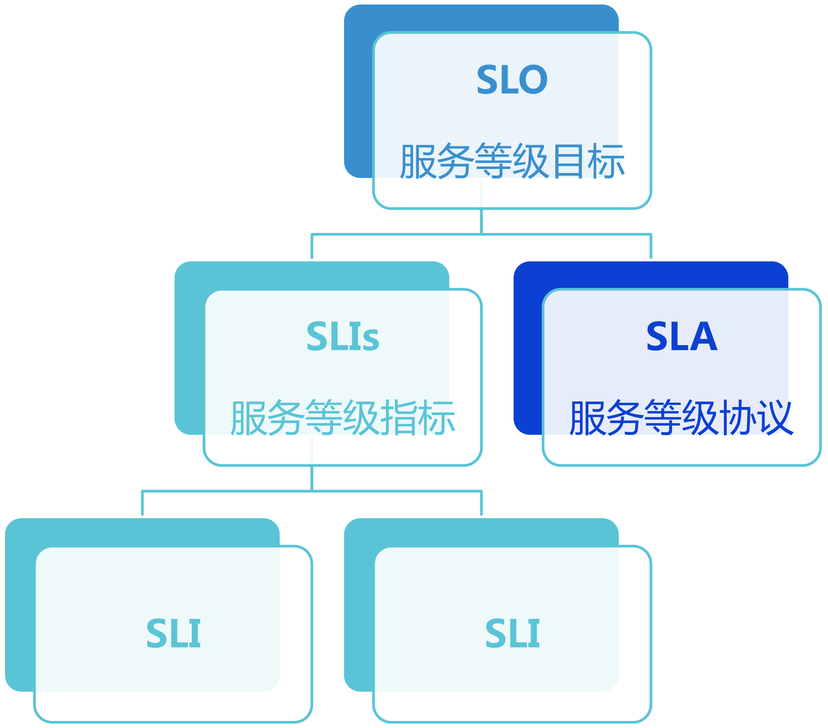
No-Blame Culture.

故障管理框架



- 可用性体系/SLIs/SLO/SLA
- 故障定级/定性/定责
- 错误预算/故障分
- 组织结构支撑/故障委员会

故障管理框架-可用性体系：SLI/SLO/SLA



SLI

- 度量内容：
  - 能够度量服务质量的指标，如可用性、响应时间
- 度量范围：
  - 指标合理的区间，比如**响应时间<50ms，http状态码为非5xx**

SLO

- 核心接口响应时间<20ms的比例不低于99%
- 全站访问请求**非5xx的占比不低于99.95%**

SLA

- 涉及双方商务合约，违反将触发处罚条款
- **SLA = SLO + SLO未达成的后果**

故障管理框架-可用性体系：SLI的选择依据

SREelite

VALET 法则

- **Volume – 容量**
  → 服务承诺的最大容量是多少？(QPS、TPS、流量、连接数、吞吐)
- **Availability – 可用性**
  → 服务是否正常？(HTTP状态码2xx的占比)
- **Latency – 延迟**
  → 服务响应速度是否够快？(rt是否在预期范围内)
- **Errors – 错误率**
  → 错误率有多少？(HTTP状态码5xx的占比)
- **Tickets – 人工介入**
  → 是否需要人工介入处理？(人工修复)

故障管理框架-故障定级：通用标准

SREelite

美图故障等级衡量指标说明			
故障对服务功能的影响	**权重占比xx%**	**备注**	**级别分值(阶梯式)**
1级	主要功能完全不可用		100
2级	主要功能部分不可用		75
3级	次要功能完全 不可用		50
4级	次要功能部分不可用		25
故障影响时长	**权重占比xx%**	**备注**	**级别分值(阶梯式)**
1级	xx分钟以上		100
2级	xx分钟~xx分钟		75
3级	xx分钟~xx分钟		50
4级	xx分钟~xx分钟		25
故障发生所处时段	**权重占比xx%**	**备注**	**级别分值(阶梯式)**
1级	最活跃时段		100
2级	次活跃时段		75
3级	非活跃时段		50
4级	最不活跃时段		25
对用户的影响范围	**权重占比xx%**	**备注**	**级别分值(阶梯式)**
1级	影响用户占比xx%以上		100
2级	影响用户占比xx%~xx%		75
3级	影响用户占比xx%~xx%		50
4级	影响用户占比xx%以下		25

故障管理框架-故障定级：业务个性化标准

SREelite

- *eg:*
- *影响收入*
- *造成资损*
- *造成PR事件*

协商/映射

通用定级标准

故障管理框架-故障定性：有效分类

SREelite

代码质量	测试质量	流程规范	变更操作
容量规划	产品逻辑	硬件设备	预案失效
局方故障	云厂故障	第三方	……

故障管理框架-故障定责：判定原则

SREelite

- 高压线原则
- 健壮性原则
- 第三方默认无责
- 分段判定原则
- 自由裁量原则

故障管理框架-故障预算：故障分

故障定级规则
(加权评分)

- A 级：xx分以上
- B 级：xx分~xx分
- C 级：xx分~xx分
- D 级：xx分~xx分
- E 级：xx分以下
- (E级不算故障,归为问题)

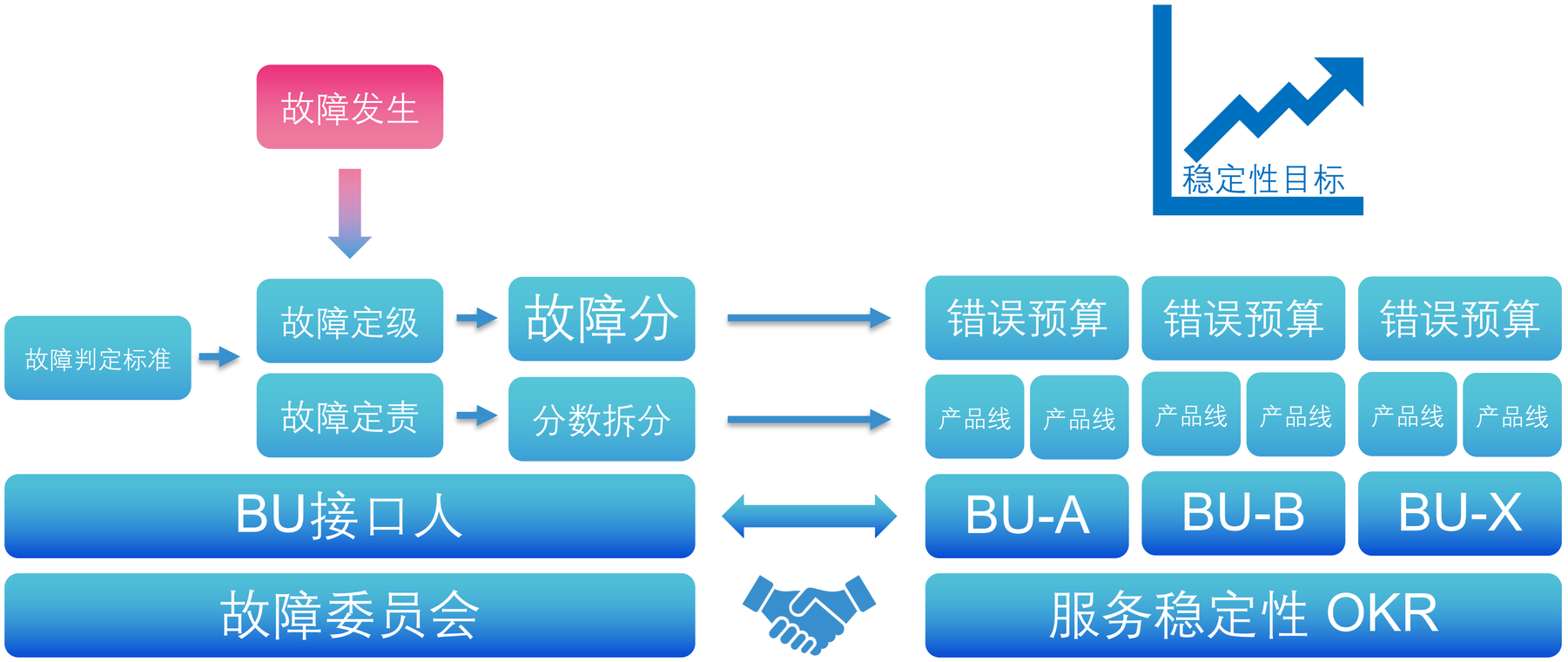
计分标准

故障级别	计分值
A级故障	3分
B级故障	1分
C级故障	0.3分
D级故障	0.1分
E级故障	不扣分

故障得分评估模板			
考核项	得分(0~100分)	权重(合计100%)	加权得分
对服务功能的影响			
故障影响时长			
故障所处时段			
对用户的影响范围			
累积得分			0~100分
故障级别			A/B/C/D/E

错误预算/故障分数(按OKR考核周期)			
产品线	BU	负责人	故障分预算
产品线A	BU-A	张三	0.4
产品线B	BU-A	李四	0.2
产品线C	BU-B	王五	0.4

故障管理框架-组织支撑&运作流程



故障管理框架-PPTV

明确OKR目标

用OKR目标的行政手段来保障各相关团队在稳定性方面的持续投入，用目标引导结果的达成。

多管齐下+敏捷行动

1. 数据支撑：监控平台、压测平台、稳定性运营平台
2. 服务干预和保障：应急响应平台、容器管理平台
3. 运维元数据平台建设：CMDB&SD
4. 各种自动化工具...

流程规范支撑

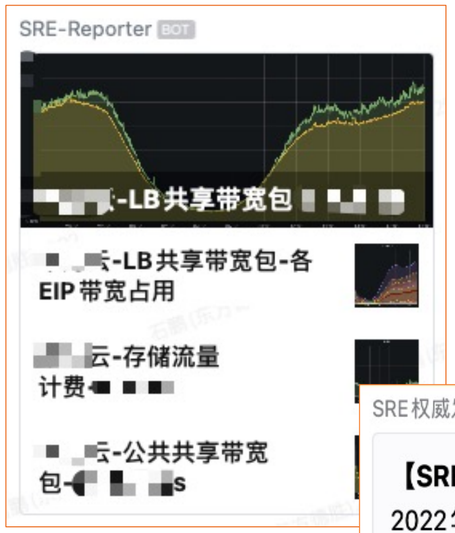
1. 制定完善的故障管理规范、故障管理流程，并将其逐步沉淀到工具中
2. 做好流程规范的宣贯，工具的使用引导和运营。

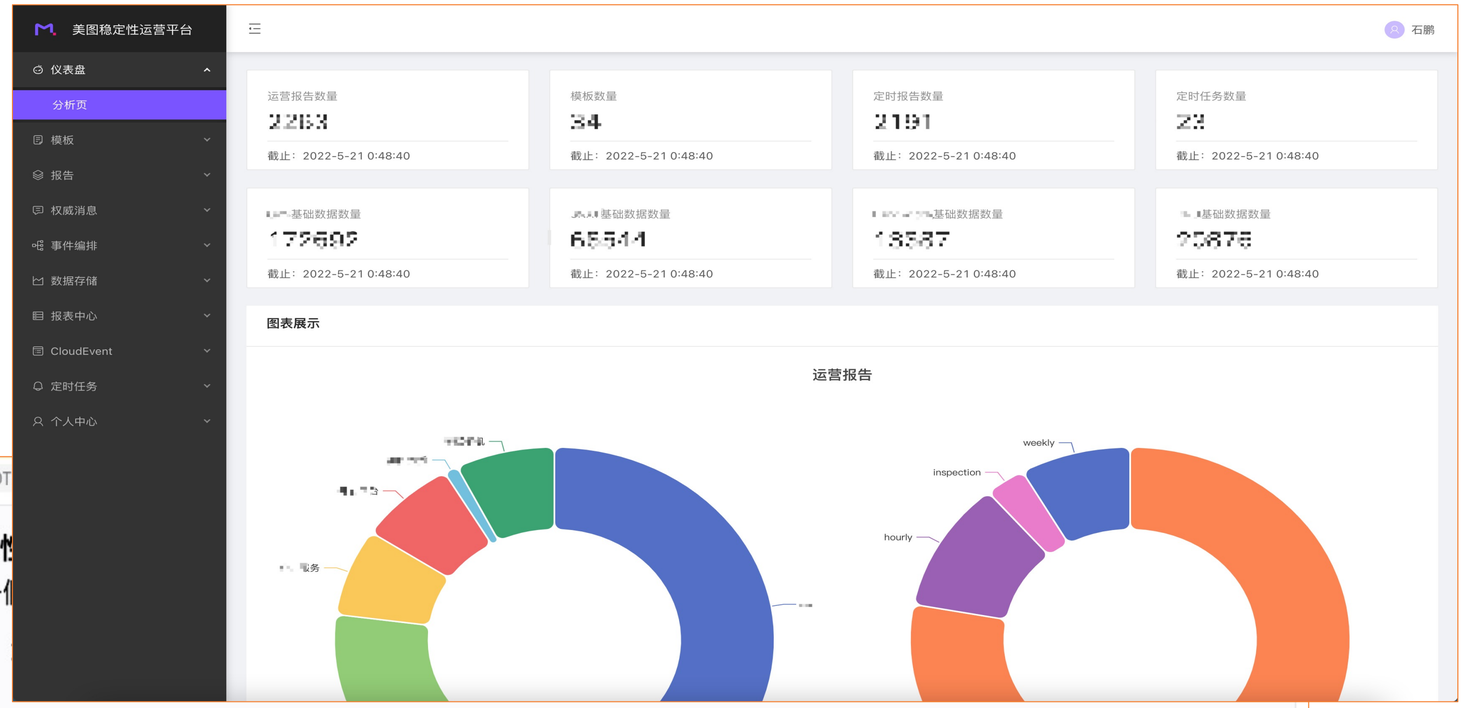
专人对接 保障落地

1. 建立中立的故障委员会，构建公开、公正、透明的故障管理环境
2. 明确接口人，保障各BU在稳定性管理方面的持续投入，保证更紧密的协作、消除因为信息不对称带来的壁垒和阻碍。

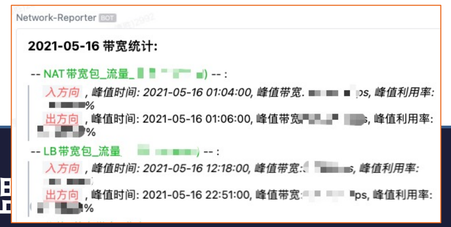
工具建设-稳定性运营平台

服务SLA / 基础设施巡检





资源用量统计



SRE权威发布 BOT

【SRE稳定性

2022年五一

时间范围：

运营周期概述:

1、五一假期内各服务整体稳定。

2、LB_共享带宽无压力，峰值在5月3日，为 Gbps。五一假期4天峰值都在 以上，假期峰值较平时峰值上涨 %。

3、容器集群运行正常，代理层有手工扩容，业务线弹性扩容，容器资源增加约 C。

4、主要业务线访问量均有明显上涨，秀秀社区较平时上涨 %，秀秀工具较平时上涨 %，广告增加 %，美颜相机增加 %

5、五一值班的稳定性事件记录在：链接

周期内重点运营活动/重大稳定性事件：五一假期

批注人：

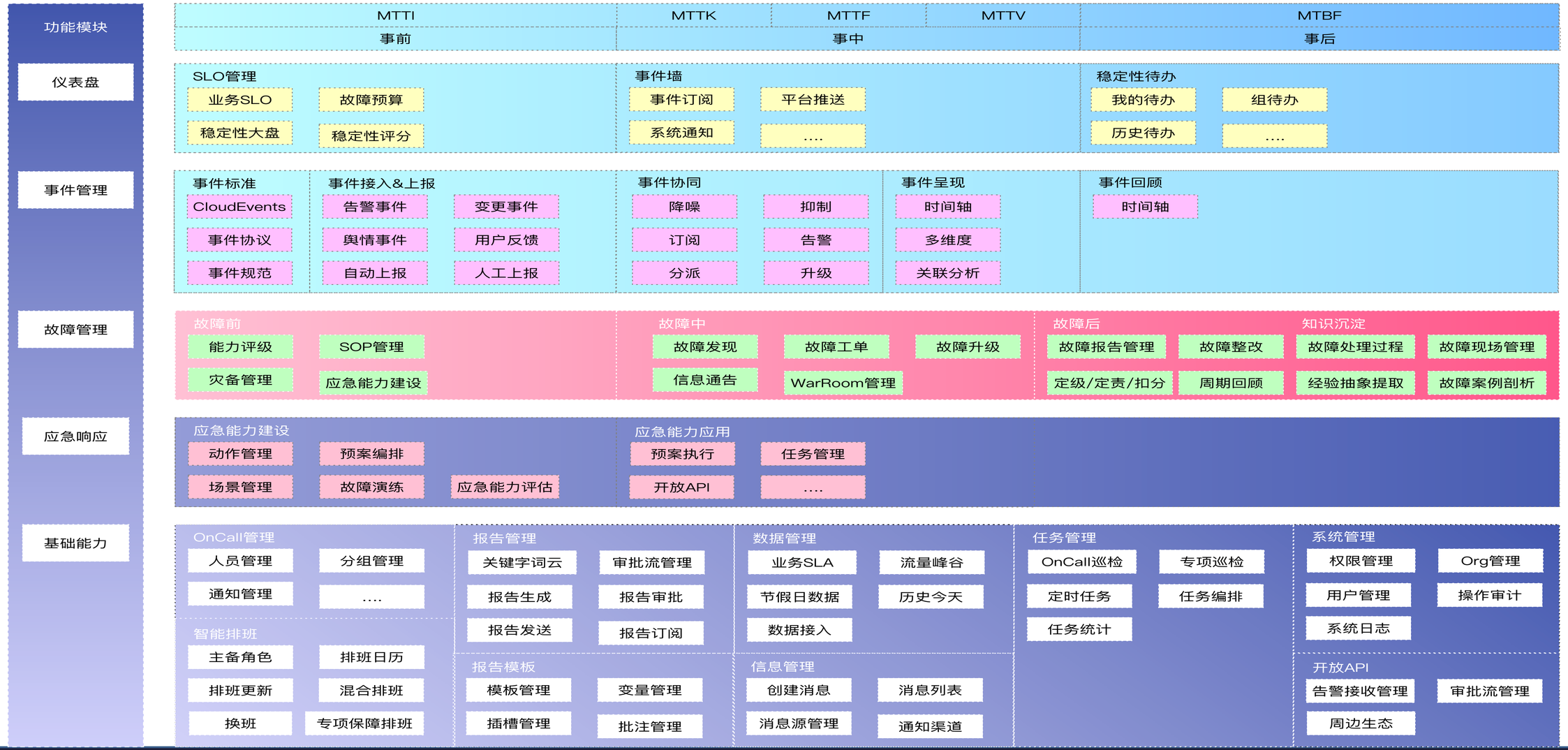
报告地址：点击我

From：美图稳定性运营平台

工具建设-稳定性运营平台

SREelite

新版稳定性运营平台(WIP)



工具建设-应急响应平台

SREelite

动作

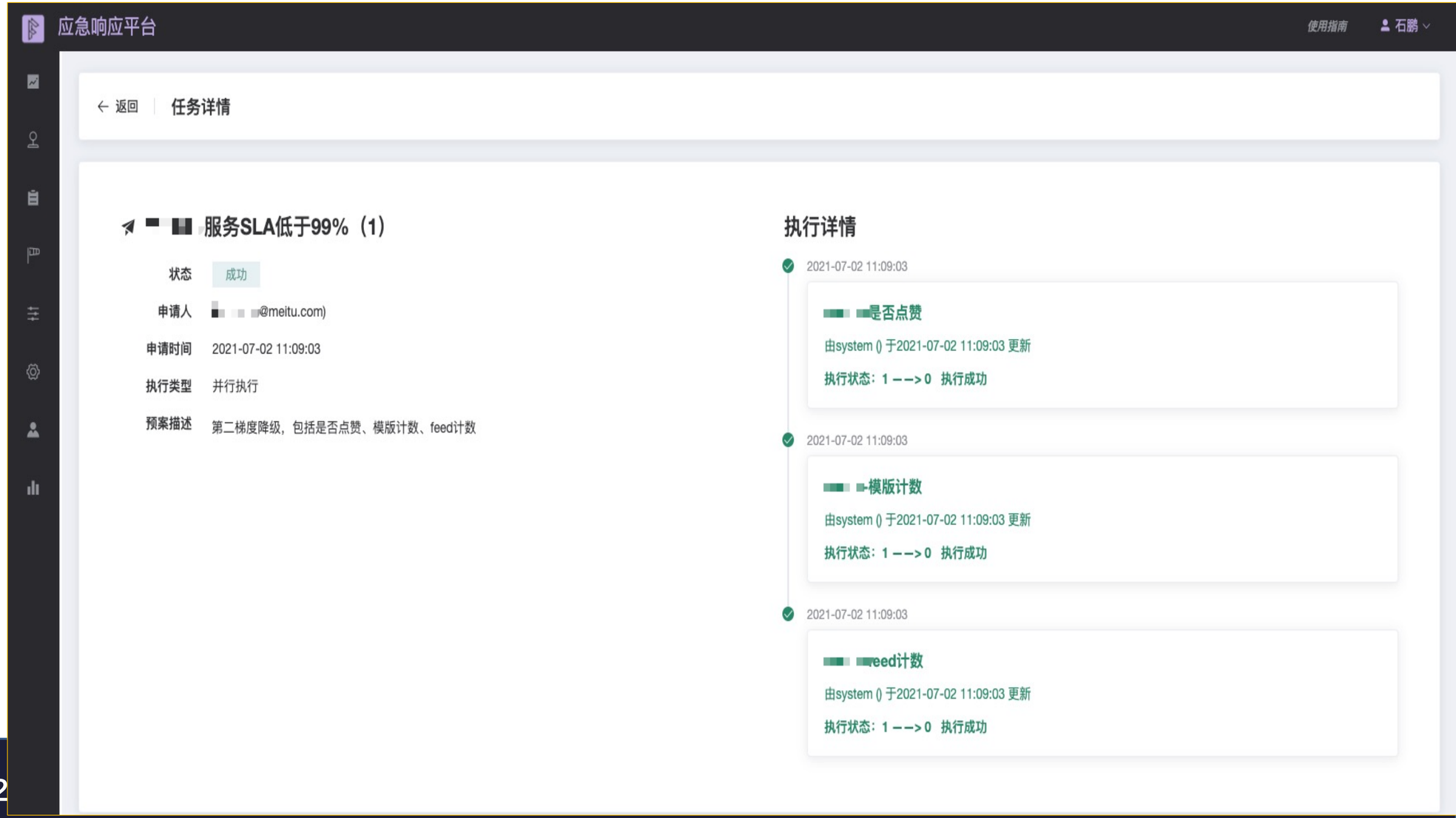
- 抽象服务干预手段
- 注册为原子性动作

预案

- 编排动作
- 串行/并行
- 依赖逻辑

场景

- 预案绑定
- 多级预案



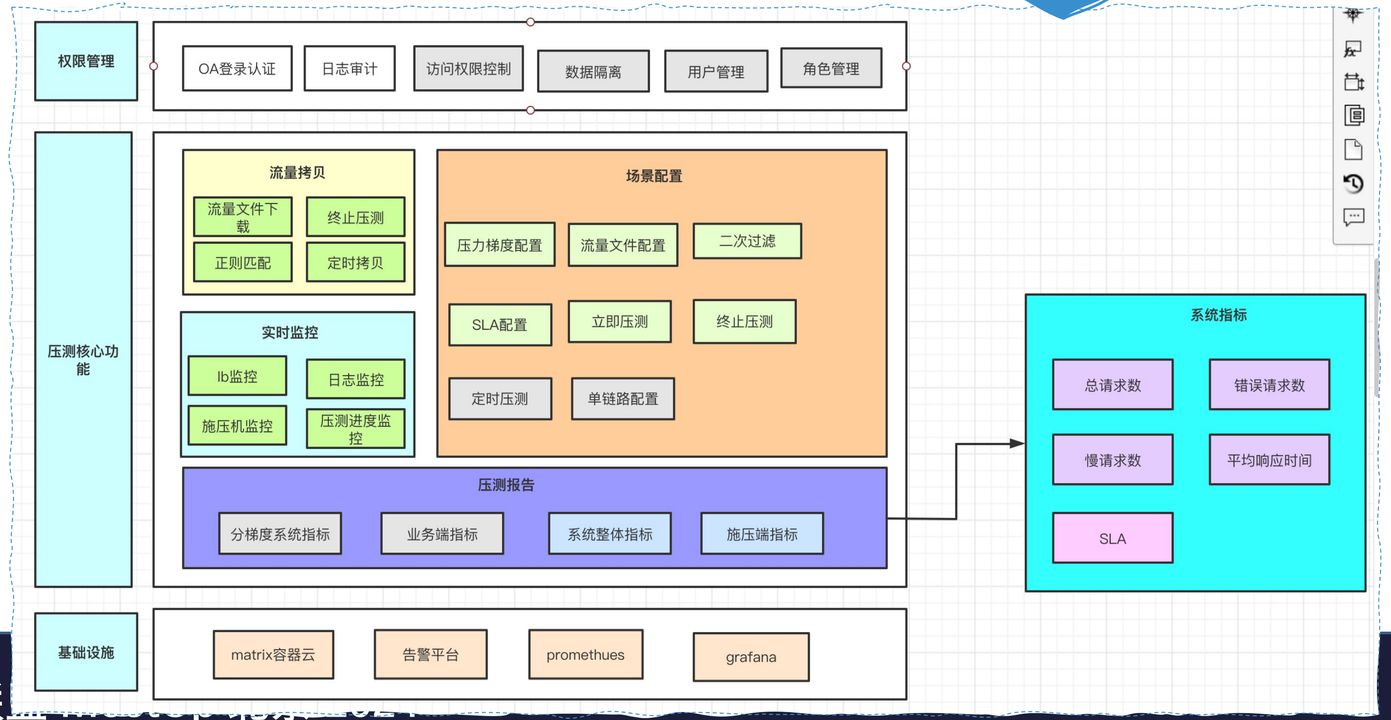
工具建设-全链路压测平台

原压测流程

1. 业务确定压测目标
2. 业务SRE拷贝流量，进行压测，并实时监控压测状态
3. 业务需要调整压力，通知SRE进行手动扩容
4. 压测过程中，如发现SLA指标异常，SRE需要手动停止压测
5. 问题排查完后，重新进行压测
6. 压测过程中，如发现单pod出现异常，SRE需要手动剔除/替换
7. 压测完成，人工撰写压测报告

痛点

1. 时间成本
2. 人员成本
3. 沟通成本
4. SLA风险



平台能力

1. 流量拷贝/回放
2. 流量过滤/修改
3. 压测梯度编排
4. 压测场景管理
5. 实时监控/全链路覆盖
6. SLA定义/自动止损
7. 压测报告自动生成

....

踩坑案例分享

踩坑案例分享

SREelite

案例1：七层负载均衡误退订故障

案例2：昨晚的故障

Part five

未来展望

简介：我所看到的几个发展趋势

我所看到的几个趋势

SREelite

云原生	•持续发展和深化
可观测	•被更多地认同和实践
混沌工程	•在更多的场景落地
Dev-X-Ops	•融合类的实践越来越多
AI Ops	•持续进化，更多落地

最后的话：如何面对汹涌的技术浪潮

SREelite

看清本质 拥抱变化 顺势而为

做好定位 葆有价值 泰然自若

云原生 容器 微服务 无服务 服务网格 低代码/零代码

可观测 GitOps ChatOps FinOps AI Ops XX Ops ....

THANK YOU

谢谢你的观看~

石鹏(东方德胜) @美图